Nonclassical Nucleation Pathways in Stacking-Disordered Crystals

Fabio Leoni[*]
Department of Physics, Sapienza University of Rome, P.le Aldo Moro 5, 00185 Rome, Italy

John Russo[†]
Department of Physics, Sapienza University of Rome, P.le Aldo Moro 5, 00185 Rome, Italy
and School of Mathematics, University of Bristol, Bristol BS8 1UG, United Kingdom

(Received 24 June 2020; revised 28 January 2021; accepted 23 April 2021; published 9 July 2021)

The nucleation of crystals from liquid melt is often characterized by a competition between different crystalline structures or polymorphs and can result in nuclei with heterogeneous compositions. These mixed-phase nuclei can display nontrivial spatial arrangements, such as layered and onionlike structures, whose composition varies according to the radial distance, and which so far have been explained on the basis of bulk and surface free-energy differences between the competing phases. Here we extend the generality of these nonclassical nucleation processes, showing that layered and onionlike structures can emerge solely based on structural fluctuations even in the absence of free-energy differences. We consider two examples of competing crystalline structures, hcp and fcc forming in hard spheres relevant for repulsive colloids and dense liquids, and the cubic and hexagonal diamond forming in water relevant also for other group 14 elements such as carbon and silicon. We introduce a novel structural order parameter that combined with a neural-network classification scheme allows us to study the properties of the growing nucleus from the early stages of nucleation. We find that small nuclei have distinct size fluctuations and compositions from the nuclei that emerge from the growth stage. The transition between these two regimes is characterized by the formation of onionlike structures, in which the composition changes with the distance from the center of the nucleus, similar to what is seen in the two-step nucleation process.

DOI: 10.1103/PhysRevX.11.031006

Subject Areas: Chemical Physics,
Condensed Matter Physics, Soft Matter

I. INTRODUCTION

Nucleation is a discontinuous transition in which clusters of molecules self-assemble due to fluctuations that are very localized in space and time to form a growing nucleus. It is a crucial phenomenon in many fields of natural science [1–3], going from the planetary scale to nanoscale. During the nucleation process of many materials, including several metals, minerals, and polymers, different crystalline phases called polymorphs can nucleate. The structure of the growing nucleus in such materials can depend on many, eventually size-dependent [4,5], effects, such as energy and entropy competition, or frustration. Understanding the selection mechanism of polymorphs is fundamental to predict the structure of the growing nucleus, with applications ranging from Earth's weather and climate forecast, especially in relation to the formation of nanometer-sized ice crystallites in clouds [6–11], to the pharmaceutical industry, where the physical and chemical properties of the drug molecules can change with the eventual crystallization of unwanted polymorph forms [12]. For example, the molecule for aspirin (acetylsalicylic acid), one of the most widely consumed medications, has two polytypic crystalline forms [13].

Here we study the nucleation of polytypes, a specific type of polymorph where the crystalline structures have the same projection along a specific direction and differ only in the way the planes perpendicular to that direction are stacked onto each other. Some of the most common crystalline structures formed in metals are polytypic, notably the hcp (hexagonal-close-packed) and fcc (face-centered-cubic) crystalline structures, and the hexagonal and cubic diamond forms.

We consider the formation of polytypes in two important systems: the hard-sphere (HS) model and the coarse-grained mW model of water. They are representative of a wide class of materials, repulsive colloids, and dense liquids [14,15] for HS and tetrahedrally bonded materials (like water and group 14 elements such as carbon and

[*]fabio.leoni@uniroma1.it
[†]john.russo@uniroma1.it

silicon) for mW [1,16]. They crystallize in two different polytypes: either fcc or hcp for HS, and either cubic ice (I_c) or hexagonal ice (I_h) for water. Importantly, in both cases the difference in all thermodynamic relevant quantities (such as free-energy difference, nucleation barrier, and solid or melt surface tension) between the competing polytypes are negligibly small (within $10^{-3}k_BT$ per particle for all cases) [1,5,17–21]. For example, in the mW water model [16] the stacking fault between the ice I_c and I_h has been estimated as low as 0.16 ± 0.05 mJ m^{-2} at $T = 218$ K [4]. In this way, the nucleation mechanisms for both systems are determined not by bulk free-energy properties or by details of their interactions, but by general principles, which we aim to elucidate in the present work.

One of the main difficulties in studying polymorph composition is assigning the local environments surrounding a particle to a particular phase, distinguishing between amorphous (liquid) structures and crystalline ones [22]. Several methods for local structure identification have been developed so far. Contrary to common belief, the method employed to classify a single particle as belonging to a specific polymorph can sensibly alter the measured composition of the nucleus [23,24]. In the present work, we compare some of the more representative methods found in the literature and introduce new methods which allow us to find fundamental properties of the nucleus growing during the early stage of homogeneous nucleation, and in particular, to find evidences of a two-step nucleation pathway.

Departures from single-step nucleation have been recently observed in the nucleation of polymorphs [25–31] in systems as different as hard-particle fluids [32], colloids [33], salt solutions [34], and calcium carbonate formation [35–37].

Two-step nucleation mechanisms involve one disordered phase (the melt) and (at least) two ordered crystalline phases. Two-step nucleation often produces layered structures, where the composition changes radially within the nucleus, in such a way that the most stable polymorph form is closer to the center and is "wetted" by the metastable form on the surface. This structure, often referred to as the *onion* structure, is one of the hallmarks of two-step nucleation pathways that have been theoretically predicted via classical nucleation theory (CNT) [25,38], density-functional theory [39–42], phase-field models [43], two-dimensional lattice models [44], and molecular simulations [45].

These two-step nucleation pathways are generally explained via free-energy differences between the two crystalline phases, and in particular, with the different surface free energy of the crystals with respect to the melt [44]. Instead, in the systems under consideration, the polytypes in competition have the same bulk free-energy properties, and classical theory would predict in this case a homogeneous composition of the nuclei. We observe that, due to finite-size fluctuation effects, onionlike structures are formed also under these conditions and that well-separated free-energy channels corresponding to the competing polymorphs can be distinguished, extending the phenomenology of structured nuclei to this large family of crystals.

The outline of the article is as follows: In Sec. II, we describe the methods for local structure identification employed in the present work to study the properties of nuclei forming during the homogeneous nucleation of HS and mW water. In Sec. III, we describe the model systems we simulate—HS and mW water. In Sec. IV, we compare the properties of nuclei as obtained by using the methods described in Sec. II. Finally, in Sec. V, we present concluding remarks.

II. PARTICLE IDENTIFICATION

A. Order parameters

Common widespread methods used in the literature to identify local structures usually employ one- or two-dimensional order-parameter (OP) maps, which involve the comparison of the local environment of a particle with different reference structures. For this reason, thresholds are usually introduced to establish which reference structure the particle under investigation belongs to. Steinhardt or bond-orientational-order (BOO) parameters in their averaged form $\bar{q}_l$ [46] (see the Appendix) are the standard choice as OPs, where fourfold ($l = 4$) and sixfold ($l = 6$) are often the only symmetries considered. Other methods involve the study of topological properties of the bond network, such as the common neighbor analysis (CNA) method. For waterlike systems, the CNA method considers also second-nearest neighbors and is named extended CNA (ext CNA). In the Appendix, we describe some of the most representative low-dimensional OPs employed in the present study for local structure identification (for a comprehensive review on common OPs, see Ref. [22]). We also present some tests aiming to determine the accuracy of the different methods in controlled situations (see the Appendix, Sec. VIII). Since previous *low-dimensional* OPs have produced different results when applied to the model systems studied here [5,23,24,47–51], we consider a high-dimensional OP based on 30 BOO (see the Appendix). In the following, we drop the number 30 and use only the acronym BOO to refer to this method. However, the degeneracy of an OP like BOO, for which the same OP value can correspond to different local environments [52], could result, in some specific application, in a suboptimal performance due to misidentification. For example, the use of Steinhardt OPs, especially of those related to the spherical harmonic with angular momentum $l = 3$ and $m = 2$ (Y_{32}), which is the only one with tetrahedral geometry, to distinguish I_c from I_h in water has been already questioned in previous works [53]. In order to resolve also the issue related to the degeneracy of the OP, in the following section, we introduce a novel

lossless order parameter for the characterization of local environments.

B. Local interdistance

Here we introduce a novel order parameter for the characterization of local environments that is built according to the following two principles. First, the OP is *high dimensional*: Increasing the dimensionality of the order-parameter space allows us to easily increase the separation between the different populations of the local environments we want to discriminate between. Second, the OP is *lossless*: With this, we mean that no information is lost by going from the real-space coordinates of the particles in the environment under consideration to its order-parameter representation; in other words, from the OP it is possible to reconstruct the original positions of the particles, except for translations, rotations, or particle-index permutations. This method is based on the distances between all possible pairs obtained from a particle and its neighbors a feature which leads to the lossless property of this OP. Indeed, the problem to establish whether to the set of all possible interdistances between a number of points corresponds only one points configuration dates back to the problem of the uniqueness in the x-ray analysis of crystal structures [54], in which case, only very few specific exceptions are known.

The new order parameter is inspired by the permutation-invariant vector of Refs. [55,56] and the deep potential molecular-dynamics method of Ref. [57] and is constructed in the following way: For each particle i, we make a list of its first (f_i^j) and second (s_i^k) nearest neighbors, with $j = 1 \cdots N$ and $k = 1 \cdots M$, where N and M are the numbers of first- and second-nearest neighbors, respectively. We then compute all the $(N+M+1)(N+M)/2$ possible distances $d_{pq} = |\vec{r}_p - \vec{r}_q|$ between particle p and particle q with $p, q = 1, 2, \ldots, N+M+1$ and $p \neq q$ and subdivide them in the following groups. For HS, we group the d_{pq} in five categories: (i, f_i^j) (12 terms), (i, s_i^k) (six terms), $(f_i^{j'}, f_i^{j''})$ (66 terms), $(s_i^{k'}, s_i^{k''})$ (15 terms), and (f_i^j, s_i^k) (72 terms). In mW water, we group the d_{pq} in six categories where now f_i^j and s_i^k are the first and second energetic neighbors of particle i. The six categories are (i, f_i^j) (four terms), $(f_i^{j'}, f_i^{j''})$ (six terms), $(f_i^j, s_i^{k_j})$ (12 terms), (i, s_i^k) (12 terms), $(s_i^{k'}, s_i^{k''})$ (66 terms), and (f_i^j, s_i^k) (36 terms) where $s_i^{k_j}$ is a second neighbor of particle i which is also the first neighbor of particle f_i^j. The number of terms in each category is obtained by considering $N = 12$ and $M = 6$ for HS, while $N = 4$ and $M = 12$ for mW water. These values for N and M are related to the number of first and second neighbors in the crystalline structures forming in these models.

The distances in each group are then sorted in ascending order. This makes the OP invariant under particle-index permutations. Since in the neural network (NN) we use the sigmoid as the activation function (see Sec. II C), which works better with inputs between -1 and 1, we normalize the grouped and sorted distances $d_{pq}^{g,s}$ for the average local environment radius r_0 (considering the first neighbors' shell for HS and up to the second shell for mW water), and subtract from it the total normalized interdistances $\langle d_{pq}^{g,s}/r_0 \rangle_{\text{out}}$ (considering all outputs of the NN). Finally, the order parameter we introduce here, which we name LID (local interdistance), is the vector obtained from the union of all the groups: $d_{pq}^{g,s,n} = d_{pq}^{g,s}/r_0 - \langle d_{pq}^{g,s}/r_0 \rangle_{\text{out}}$.

To emphasize the advantages of LID, we compare its results with the ones obtained via either a low-dimensional method, i.e., CNA, or via a high-dimensional (but not lossless [53]) order parameter constructed as an array of 30 different BOO parameters (built from spherical harmonic invariants of order up to $l = 12$; see the Appendix).

C. Neural-networks classification scheme

To partition a multidimensional OP space in different volumes, each one associated with the local environment of a crystalline structure or liquid phase, we use artificial NNs. In condensed matter, NNs have been used for potential energy surface calculations [57,58] to construct accurate molecular force fields [59], improve potential energy of coarse-grained models for water [49], or for identification and classification of local ordered or disordered structures using supervised [60–62] and unsupervised [63–68] methods. Ideally, unsupervised learning allows us to cluster high-dimensional OP space into sets corresponding to different structures before they have been identified [63–65,67]. If all possible structures present in the system are known *a priori*, supervised learning is a powerful method to identify local structures not requiring any threshold chosen *ad hoc* and being less sensitive to hyperparameters. We choose here supervised training, in which the NN is first trained against sample configurations of the phases we are interested in identifying. For the training, we use bulk configurations prepared at coexistence conditions, where thermal fluctuations in the solid phases are maximized. For the HS system, this choice corresponds to preparing bulk fcc, hcp, bcc, and fluid configurations at pressure $P = 11.54$ (in conventional reduced units) [69] and running event-driven molecular-dynamics simulations and using each local environment as a training sample. In detail, the training set for HS is obtained by ten different realizations of fcc, hcp, and bcc crystals at the melting point ($\phi = 0.545$) and 20 different realization of the liquid phase at the freezing point ($\phi = 0.494$), all composed of $N \sim 10\,000$ particles. The training set for mW water is obtained by running Monte Carlo simulations at ambient pressure of ten different realizations of I_c, I_h at the melting temperature $T_m = 275$ K, and of ice 0 at its melting temperature $T_m = 244$ K (being metastable, it has a lower melting temperature), and 20 different realization of the liquid phase at the melting temperature $T_m = 275$ K, all composed

of $N = 5376$ particles. We notice here that we do not train the NN against surfaces, as this requires external criteria in order to be defined (such as the Gibbs dividing surface), and we are interested only in bulklike local environments. We choose a single-layer feed-forward network topology. As descriptors, we consider the 30-dimensional (both for HS and mW water) BOO OP, and the 171-dimensional (for HS) and the 136-dimensional (for mW water) LID OP. The hidden layer (HL) for BOO (both for HS and mW water) is composed of eight nodes. We obtain the same performance varying the number of nodes in the HL from four to 20, indicating that the network is quite robust. The HL in LID is composed of 30 nodes (both for HS and mW water). Also, in this case we observe the same performance of the network for a wide range of nodes in the HL. We initialize the weights following the Xavier method [70] consisting of setting random weights from a normal distribution with zero mean and variance equal to 2 divided by the sum of the number of nodes in the input layer (IL) and the output layer (OL). We consider the sigmoid, or logistic function, as the activation function for both the IL-HL and HL-OL. The OL is composed of four nodes, which correspond to the four possible phases identified during the homogeneous nucleation of HS and mW water at the thermodynamic conditions considered here. As the error or loss function, we take the overall mean-square error between the actual and the target output. We minimize the error using the stochastic gradient descent (for a critical discussion, see Ref. [71]) and update the weights following the back-propagation approach [72]. The performance of the NN is higher than 98% in all cases. The absence of overfitting is verified by obtaining the same performance considering both the test and the training set. For all cases, we set the learning rate to $\alpha = 0.01$, while the number of epochs is 50 for BOO and 100 for LID for both HS and mW water.

D. Nucleus identification

After all particles in the system are classified as belonging to a specific phase, in order to identify clusters of solid particles, we use the same method employed in Ref. [73]: Two solid particles are considered to belong to the same cluster if their distance is smaller than the value of the first minimum of the radial distribution function of the liquid (which turns out to be approximately $1.5\sigma_{\mathrm{HS}}$ in HS and approximately $1.5\sigma_{\mathrm{mW}}$ in mW water). After a solid particle is added to a cluster, the enumeration needed to distinguish the different clusters is obtained by using the Hoshen-Kopelman algorithm [74].

Other methods used to identify neighbors are the Voronoi construction, which is parameter-free but computationally expensive and sensitive to thermal fluctuations [75,76], and the solid-based nearest-neighbor (SANN) algorithm [75], which is parameter-free and more robust against thermal fluctuations with respect to the Voronoi construction but occasionally can include second shell neighbors in the first shell neighbors [66].

III. MODEL SYSTEMS

A. Homogeneous nucleation of hard spheres

Here we consider nonoverlapping hard spheres, the reference model for systems with excluded volume interactions [77]. We perform event-driven molecular simulation of $N = 100\,000$ hard spheres at constant volume V using the open-source event-driven particle simulator DYNAMO [78]. The phase diagram of hard spheres is a function of the volume fraction $\phi = Nv/V$, which is the fraction of the box volume V covered by the N spheres, each sphere having volume $v = (\pi/6)\sigma_{\mathrm{HS}}^3$. σ_{HS} is the sphere diameter. We consider 100 different trajectories simulating different initial configurations of supersaturated fluids at volume fraction $\phi = 0.535$, between the freezing $\phi = 0.494$ and the melting $\phi = 0.545$ value. Each configuration of supersaturated fluid is obtained using a Monte Carlo method whose moves consisting of the expansion of the spheres' diameters are rejected if at least two spheres' volumes overlap. $\phi = 0.535$ is close enough to the melting value for the supersaturated fluid to nucleate easily but far enough to avoid multiple critical nuclei from growing and eventually merging together. Indeed, in all of the 100 different trajectories simulated, we always observe one critical nucleus growing within the maximum number of collisions simulated, which is 10^{10}, i.e., an average of 2×10^5 collisions per particle. We also perform simulations for $\phi = 0.54$ and observe multiple nuclei growing and merging during the nucleation process.

B. Homogeneous nucleation of mW water

The mW model of water is a popular coarse-grained representation of water, where the molecule is replaced by a single site having both two-body and three-body interactions [1,16]. We perform Monte Carlo simulations of $N = 4000$ mW particles in the NPT ensemble at pressure $P = 0$ Pa and temperature $T = 204$ K. At these thermodynamic conditions, the mW water model spontaneously nucleates within the maximum time simulated. We consider 100 different trajectories simulating different initial configurations of supercooled fluid. Each supercooled fluid configuration is obtained using the same Monte Carlo method employed to get supersaturated HS fluid.

A system with directional tetrahedral interaction has the potential to offer additional insights into nucleation pathways, as, in principle, it can involve many polymorphic structures [79]. We focus here on the stable ice I polytypes (the cubic form I_c and the hexagonal form I_h), and on the metastable ice 0 structure [80,81]. We choose this polymorph as it is currently the only known structure to satisfy all the following criteria: It has the lowest free energy outside the stable cubic and hexagonal (ice I) structures [82],

it is the simplest structure that can be built by deformation of the diamond crystal while preserving to a large degree a highly regular fourfold coordination for the sites [83], and it can stack coherently (without breaking of bonds between grains) with the diamond crystal [4]. These structures have never been observed as fully formed crystals, and instead, we focus on clusters of molecules whose nearest-neighbor environment is close to those found in the bulk ice 0 crystal. It has been recently shown that these clusters have a lower energy than their stable ice *I* counterparts up to cluster sizes of around 40 water molecules [4].

IV. RESULTS

A. Hard spheres

1. Nucleus composition

In Fig. 1, we show the results from the homogeneous nucleation of the HS system obtained from 100 independent event-driven molecular-dynamics [78] trajectories of 10^5 particles at the volume fraction $\phi = 0.535$. The snapshots in Figs. 1(a), 1(b), and 1(c) compare the same simulation configuration of large-scale grains, which are colored according to the classification output of the CNA, BOO, and LID order parameters, respectively. The color indicates the detected phase: blue, green, red, respectively, for the fcc, hcp, and bcc local environments. Already from a quick visual inspection, we see that both the CNA and BOO methods have a lower resolution in the details of grains respect to LID whenever there is a high degree of hcp and fcc stacking. On a quantitative level, Figs. 1(d)–1(f) report, for the same order parameters, the average fraction of the different polymorphs within the largest nucleus as a function of the nucleus size n. All methods do not detect any bcc in the nucleus, as was already found in Ref. [84]. Both the fcc and hcp fractions instead grow linearly (volume scaling) with n. If we define the ratio $r = n_{\mathrm{fcc}}/n_{\mathrm{hcp}}$ (where n_{fcc} and n_{hcp} are the number of particles in the fcc and hcp phase, respectively), we see that the low-dimensional method CNA gives a value ($r = 1.31 \pm 0.05$) that is incompatible with both multidimensional methods: $r = 1.07 \pm 0.05$ for BOO and $r = 1.00 \pm 0.05$ for LID. A ratio $r \sim 1$ is indeed expected during the growth stage given the low-free-energy difference between the fcc and hcp phases and the fact that the crystals are polytypes; i.e., they can stack onto each other with considerable entropy gain [5]. Both multidimensional methods agree within the

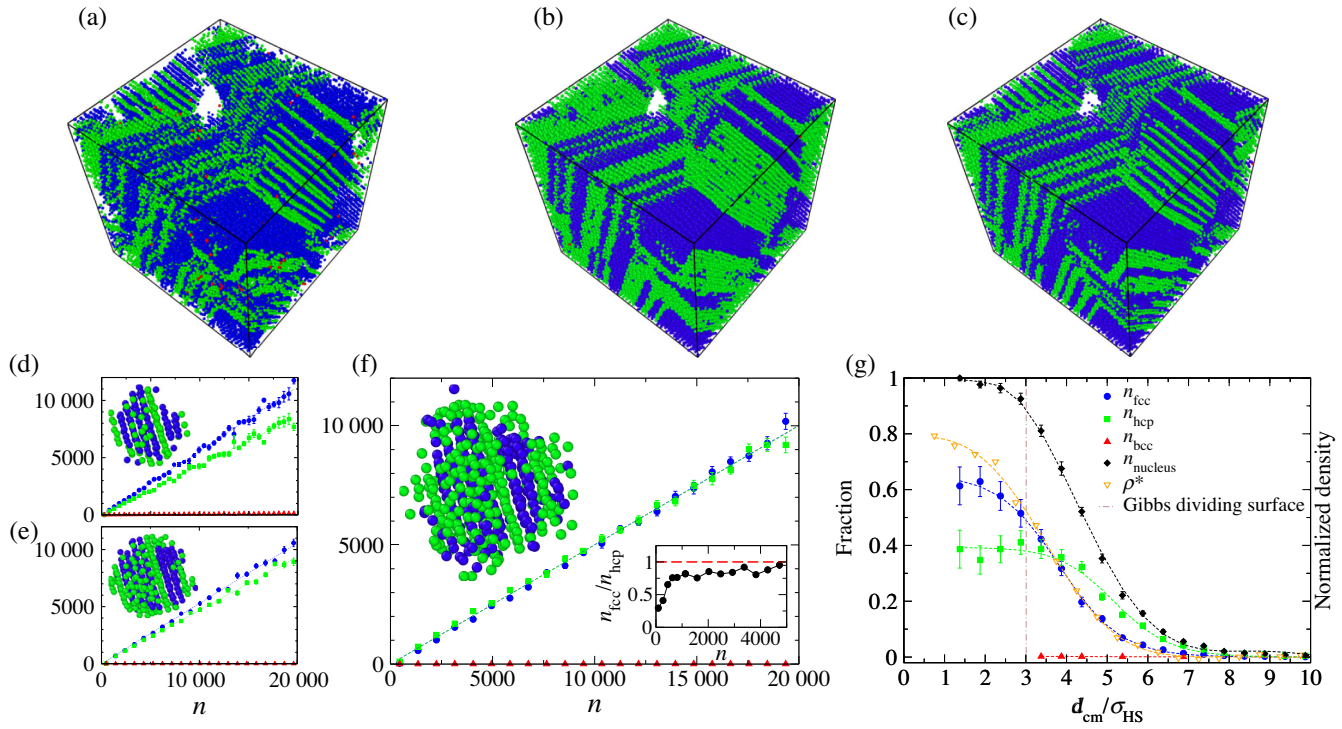

FIG. 1. Homogeneous nucleation of hard spheres. (a)–(c) Snapshots of crystalline grains spanning the simulation box. Particle structures from the same configuration are identified using the following methods: CNA (a), BOO (b), and LID (c). In all panels, the colors associated with fcc, hcp, and bcc structures are blue, green, and red, respectively. We show average composition of the main cluster as identified by CNA (d), BOO (e), and LID (f). Insets in (d), (e), and (f) show a typical nucleus composed of 188, 398, and 502 particles, respectively. The inset to the right in (f) shows the average ratio $r = n_{\mathrm{fcc}}/n_{\mathrm{hcp}}$ between the number of particles composing the nucleus in the fcc and hcp phase using LID for local structure identification. (g) Average radial fractional composition of the main cluster (for clusters of size $500 \leq n \leq 550$) as identified by LID. $d_{\mathrm{c.m.}}$ is the distance from the center of mass of the cluster and σ_{HS} the hard-sphere diameter. Dashed fitting lines are a guide for the eyes. Snapshots obtained using OVITO [85].

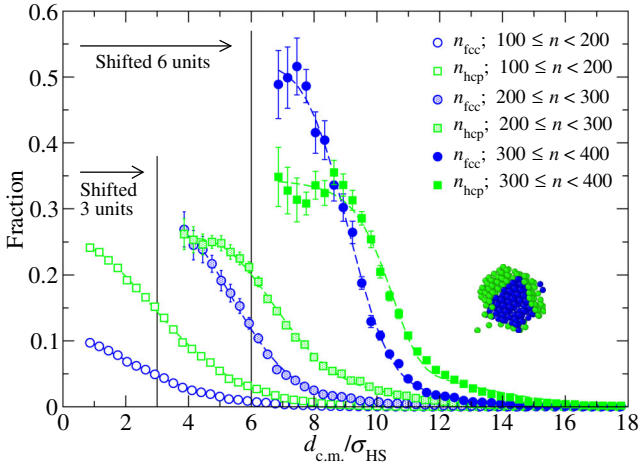

FIG. 2. Average radial fractional composition of the main cluster of hard spheres for different range of cluster size as identified by LID. $d_{c.m.}$ is the distance from the center of mass of the cluster and σ_{HS} the hard-sphere diameter. Curves for $200 \le n < 300$ and $300 \le n < 400$ are shifted along the x axis of three and six units, respectively. Dashed fitting lines are a guide for the eyes. The snapshot shows the section of a typical nucleus of size $n \simeq 400$ particles. fcc and hcp particles are in blue and green, respectively.

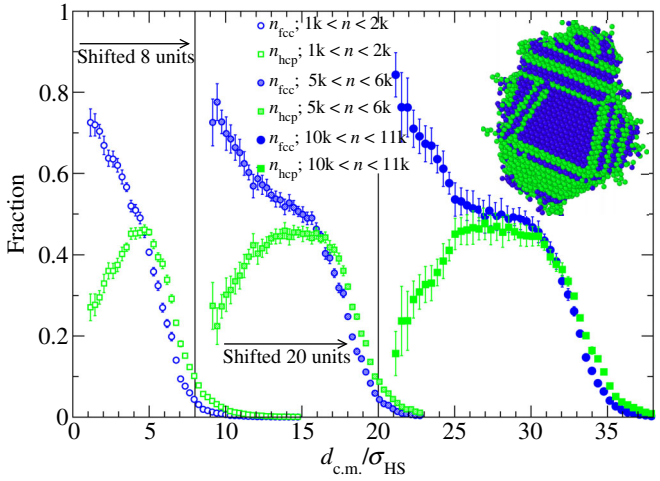

FIG. 3. Average radial fractional composition of the main cluster of hard spheres for different range of cluster size as identified by LID. $d_{c.m.}$ is the distance from the center of mass of the cluster and σ_{HS} the hard-sphere diameter. Curves for $5000 < n < 6000$ and $10\,000 < n < 11\,000$ are shifted along the x axis of eight and 20 units, respectively. The snapshot shows the section of a typical nucleus of size $n \simeq 11\,000$ particles. fcc and hcp particles are in blue and green, respectively.

error. The snapshots in Figs. 1(d)–1(f) show a nucleus identified by the different order parameters from the same configuration. We note that the number of particles identified as crystalline varies considerably depending on the method: 188 particles for CNA, 398 for BOO, and 502 for LID. The multidimensional methods that use the NN detect larger nuclei as they are trained with configurations of crystal structures at melting, thus, including as many thermal fluctuations as possible without breaking the crystal order. LID, as we confirm below for the mW model, is particularly effective even for distorted local environments.

In Fig. 1(g), we focus on the LID method and show both the composition (full symbols) and density (open symbols) profiles as a function of the distance from the center of mass of the nucleus, averaged over nuclei of size $500 \le n \le 550$. This size is chosen to be well above the critical nucleus size: From a mean-first-passage time of the nucleating trajectories (see Sec. IV B 1 for a theoretical description in the case of mW water), we estimate the critical size to be $n_c \sim 180$ for the LID order parameter, meaning that the profiles in Fig. 1(g) are for nuclei which are 3 times this size. The figure reveals two important characteristics of two-step nucleation pathways. The first is the decoupling between the density and structural order fields. The open symbols represent the normalized density $\rho^* = (\rho - \rho_f)/(\rho_x - \rho_f)$, such that the values 0 and 1 are assigned, respectively, to the bulk density of the fluid and crystal phases; $\rho = 1/\langle v \rangle$ and $\langle v \rangle$ is the average specific volume computed via a Voronoi tessellation. As is seen here, and contrary to CNT assumptions, the nucleus reaches only about 80% of its bulk density close to the center. Recently, for HS it has been confirmed that using CNT in combination with bulk quantities yields inaccurate results in the description of nucleation [86]. The second characteristic is the difference in profiles for the fcc (blue symbols) and hcp (green symbols) polytypes. While the fcc phase is found more abundantly near the center of mass of the nucleus, hcp has a relative higher concentration toward the surface with the fluid. This is the *onion* structure mentioned before. In the next section, we examine in more detail the nucleation pathway of these structures.

2. Onionlike structures

The imbalance between the two polytypes, fcc and hcp, is measured with the ratio $r = n_{fcc}/n_{hcp}$, which we plot in the right inset of Fig. 1(f) as a function of the cluster size n. The ratio is not constant: It shows a predominance of hcp for small values of n, which then converges toward a homogeneous composition as the size n increases. As we note in Fig. 1(g), at sizes above the critical value, nuclei are also not homogeneous, with the fcc phase being more abundant on average toward the center of the nucleus.

To understand the appearance of onionlike structures, in Fig. 2 we plot the average radial fractional composition of crystalline clusters for different sizes, ranging from precritical nuclei to nuclei just above the critical size. The figure confirms that there is a transition between spatially uniform nuclei ($n \lesssim 200$) where hcp is the majority component, to larger nuclei where the core becomes more abundant in fcc and the outer layers in hcp. Visual inspection of these nuclei reveals the presence of a fcc-rich core surrounded by

stacking faults. There are two reasons for the size-dependent stability of fcc cores. The first one is that fcc is a cubic crystal, and thus can form stacking disorder along four independent directions (along the 1,1,1 planes) instead of only one direction as in the case of the hcp crystal (which has hexagonal symmetry, and where the only stacking direction is the one perpendicular to the basal plane). The inset of Fig. 2 shows a snapshot from the formation of these structures: a fcc core (blue particles) developing stacking faults in two directions (green hcp particles). The second reason is that the intersection of the stacking planes growing in different directions creates fivefold coherent grain boundaries from which the crystal can go radially maintaining a fcc-rich core. These grain boundaries were first observed in Ref. [87] for HS particles.

These observations are confirmed by looking at the radial fractional composition for large clusters plotted in Fig. 3. With increasing size, the core of the nuclei retains the fcc-rich character, while the surface develops an intermediate plateau with equimolar composition. This region is due to random stacking along one or multiple 1,1,1 planes that emanate from the nucleus core. An example of this process is shown in the inset of Fig. 3. The preference for hcp in the outermost part of the surface of nuclei shown by clusters of all sizes can be explained with the preference of clusters of tetrahedra in the liquid phase to coalesce via their faces in order to form locally dense aggregates [88], and this prevalent tetrahedral arrangement is compatible with the hcp phase.

We now investigate the transition between precritical homogeneous nuclei and onionlike structures. To characterize the change in structure, we compute the gyration tensor $S_{\alpha\beta}$,

$$S_{\alpha\beta} = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(r_\alpha^i - r_\alpha^j)(r_\beta^i - r_\beta^j), \tag{1}$$

where $\alpha, \beta = x, y, z$, and r_α^i is the α component of the position vector of the particle i belonging to the cluster. The eigenvalues of $S_{\alpha\beta}$ are also called principal moments and can be written as the ordered elements $\lambda_x^2 \le \lambda_y^2 \le \lambda_z^2$, and the radius of gyration is defined as $R_g = \sqrt{\mathrm{Tr}(S)} = \sqrt{\lambda_x^2 + \lambda_y^2 + \lambda_z^2}$.

In Fig. 4, we plot the normalized histograms

$$F(n, x) = -\log P(n, x), \tag{2}$$

where $P(n, x)$ is the reduced probability distribution function taken from our simulations data, with n being the size of the nucleus, and $x = R_g$ (radius of gyration) in the left panel and $x = f = n_{\mathrm{fcc}}^c/n^c$ (see definition in the following) in the right panel.

We first examine $x = R_g$ (left panel). Up to the critical nucleus size, $F(n, x)$ coincides with the potential of mean force for the two reaction coordinates n and $x = R_g$. The dashed blue line indicates the critical nucleus size, while the red dashed lines are, from left to right, power laws with $R_g \sim n^{2/3}$ (surface scaling) and $R_g \sim n^{1/3}$ (volume scaling), respectively. The figure shows that there is a clear distinction between precritical clusters, with large surface fluctuations, and postcritical clusters. Large surface fluctuations for small nuclei are compatible with previous experimental observations on repulsive colloids [33,89]. The majority of precritical nuclei are not compact enough for barrier crossing, and the path with the smallest barrier selects nuclei from the population with a small radius of gyration (compact nuclei). This transition occurs in correspondence to the nucleus size where onionlike structures start to appear. Indeed, stacking and defects like grain boundaries, which favor the formation of fcc in the inner part of nuclei, can take place only when they are compact enough.

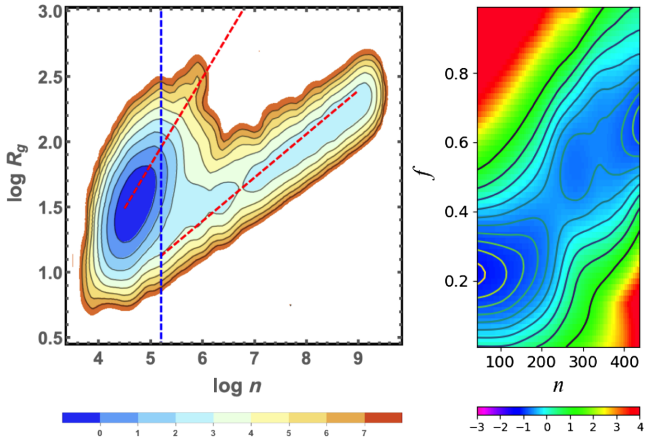

FIG. 4. Histogram plots, $F(n, x) = -\log P(n, x)$, where $P(n, x)$ is the reduced probability distribution function for the variables n (size of the nucleus), and $x = R_g$ (radius of gyration) in the left panel and $x = f$ (fcc composition ratio of the core) in the right panel.

In the right panel of Fig. 4, the reaction coordinate $x = f = n_{\mathrm{fcc}}^c/n^c$ is given by the fraction of fcc particles in the core of a nucleus, where the core is defined by a sphere of radius $3\sigma_{\mathrm{HS}}$ centered in the barycenter. The results for different values of the sphere diameter are qualitatively similar. This choice in the computation of f allows us to better highlight the transitions in the core for the small cluster sizes we consider here. From the plot, we can see a distinction between the fcc-core-poor ($f < 0.5$) basin at small n, and the fcc-core-rich ($f > 0.5$) basin at large n. Lines represent contour lines. The saddle point is found at a value of n close to the estimated value of the critical nucleus.

Overall, Fig. 4 shows that crystal nuclei that pass the nucleation barrier are more compact and have a higher fcc content compared to precritical nuclei.

B. mW water

1. Critical nucleus

First of all, we estimate the size of the critical nucleus by using the mean-first-passage theory [90,91]. This theory allows us to estimate the average time at which the growing nucleus overcomes the nucleation barrier and then to estimate the critical nucleus size n_c. The mean-first-passage time $t_{\mathrm{FP}}(n)$, which gives the average time after which a nucleus of size n appears first in the system, is given by

$$t_{\mathrm{FP}}(n) = \frac{1}{2kV}\{1 + \mathrm{erf}[Z(n - n_c)]\}, \tag{3}$$

where k is the nucleation rate, erf is the error function, and $Z = \sqrt{-\Delta G''(n_c)/(2\pi K_B T)}$ is the Zeldovich factor. $\Delta G''$ is the second derivative of the formation free energy of nuclei. n_c corresponds to the value of n where the curvature of $t_{\mathrm{FP}}(n)$ changes its sign. In Fig. 5, we show $t_{\mathrm{FP}}(n)$ versus the nucleus size n. From it, we can see a big variation in the estimation of n_c from the different methods compared here. To summarize these results, ext CNA, ext CNA 1st, CHILL+, and BOO give a small value for n_c going from 4 to 20. $\bar{q}_{12}$ and LID give values for n_c close to each other, that is, 41 and 47, respectively, while $\bar{q}_4\bar{q}_6$ gives a value for n_c which is very sensitive to the protocol employed to compute it (see the Appendix).

2. Composition ratio of the nucleus

After estimating n_c, we analyze the composition of the main cluster obtained from the different identification methods. In Fig. 6, we show the ratio $r = n_{Ic}/n_{Ih}$ between the number of particles belonging to the main cluster which are associated with the cubic ice (n_{Ic}) and those associated with the hexagonal ice (n_{Ih}) versus the normalized nucleus size n/n_c, where n_c is the critical size of the nucleus given by the method under consideration. We do not show the ratio of particles of the nucleus identified as ice 0 because only some of the methods analyzed here include it between the possible crystal phases.

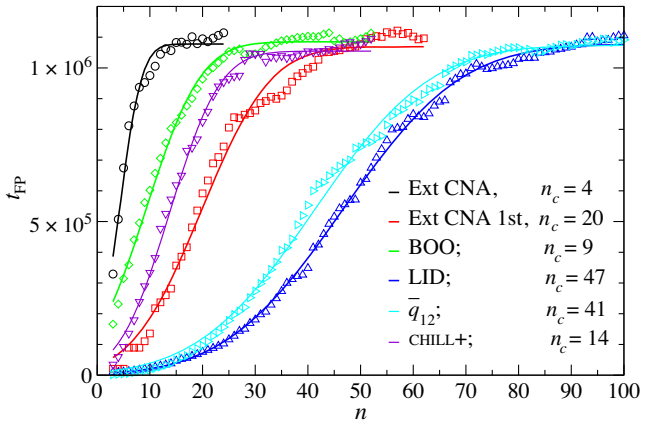

FIG. 5. Average first passage time t_{FP} as a function of the nucleus size computed using different methods for local structure identification (see legend).

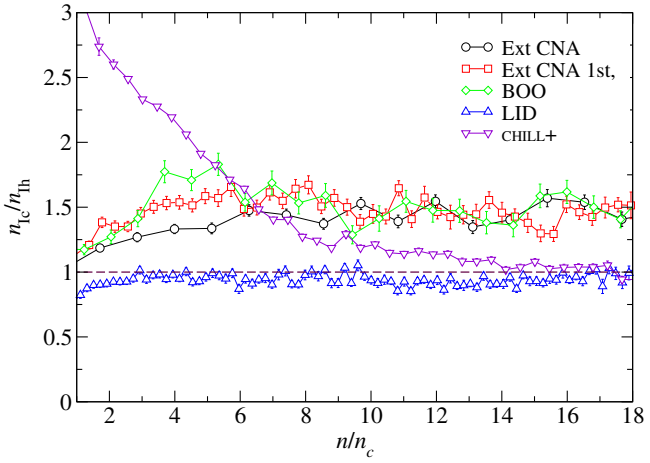

FIG. 6. Average ratio $r = n_{I_c}/n_{I_h}$ between the number of particles composing the nucleus in the cubic phase (I_c) and the hexagonal phase (I_h) using different methods for local structure identification. r is plotted against the nucleus size (n) normalized by the critical nucleus size of a specific method i (n_c^i).

As shown by Prestipino in Ref. [24], the $\bar{q}_4\bar{q}_6$ method can give predictions on the composition of the nucleus completely differently by changing the protocol used to compute and partition this two-dimensional OP. Here we consider different protocols obtaining different values for the ratio r and report the results separately in the Appendix.

Ext CNA and ext CNA 1st give a preference to I_c with a value of the ratio r between 1.3 and 1.4. BOO predicts a value $r \sim 1.4$ for small normalized nucleus size, while for larger normalized nucleus size, it drops to values closer to 1 ($r \sim 1.1$). CHILL+ has a strong imbalance toward ice I_c for small sizes and reaches $r \sim 1$ only for large nucleus size. Only LID measures $r \sim 1$, except for small cluster size, where the hexagonal ice becomes predominant, a similar behavior to what we observed for hard spheres (see Sec. IV A 2). As shown in the Appendix Sec. VIII, the ratio $r \sim 1$ given by LID and CHILL+ at large nucleus size is not observed in other methods. The larger value of r of these methods comes from the fact that they perform well only near the center of the nucleus, which comprises a majority of cubic ice environments, and perform worse near the surface, where the hexagonal environments are more abundant than cubic ones.

3. Radial composition of the nucleus

Figure 7 shows the average radial composition for nuclei of size $150 \leq n \leq 200$ obtained using LID. We find the same nucleation property that also characterizes HS nuclei: While the overall average composition is the same between the stable ice I_c and I_h polytypes ($r \sim 1$), the cubic diamond is more abundant than the hexagonal diamond

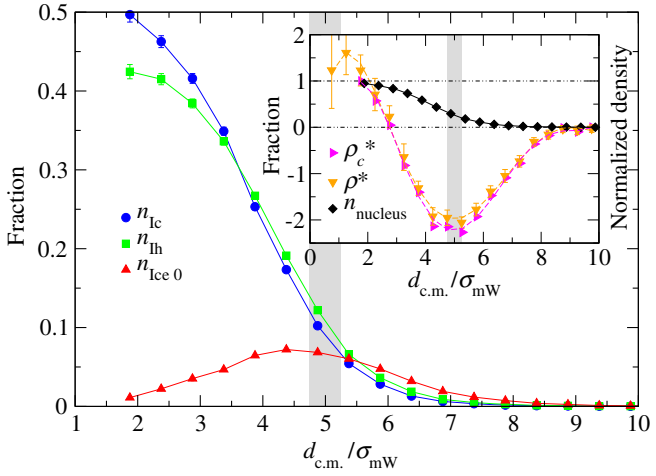

FIG. 7. Homogeneous nucleation of mW water. Average radial fraction composition of the main cluster (for clusters of size $150 \leq n \leq 200$) as identified by LID. $d_{c.m.}$ is the distance from the center of mass of the cluster and σ_{mW} the mW water molecule diameter. The inset shows the normalized density profiles (colored symbols) for the same nuclei considered in the main panel, while black symbols represent the total fraction of all crystalline particles.

near the center of the nucleus. But the mW model also offers additional insights respect to the HS system. LID is the only method to detect the presence of ice-0-like structures (red symbols), whose growth is slower than the volume growth of both the ice I polytypes (see Fig. 20). Figure 7 indeed confirms the presence of a small population of 0-like environments which peaks toward the surface of the nucleus. An independent confirmation of this unusual surface behavior of mW water can be seen in the inset of Fig. 7 where we plot (orange symbols) the normalized density $\rho^* = (\rho - \rho_f)/(\rho_x - \rho_f)$, where $\rho_x = 0.985\ \mathrm{g/cm^3}$ is the bulk density of the ice I phase, and $\rho_f = 0.980\ \mathrm{g/cm^3}$ is the density of the bulk liquid phase at the thermodynamic conditions considered here. Importantly, the density of ice 0 ($\rho = 0.953\ \mathrm{g/cm^3}$) is lower than both the metastable liquid and ice I crystals at the same conditions. Indeed, we observe that, instead of monotonically increasing from ρ_f at the surface toward ρ_x at the center of the nucleus, the density profile has a very pronounced density minimum toward the surface of the nucleus. The location of this minimum (which is computed independently from any structural order parameter, if not for the location of the center of mass) corresponds exactly to the location of the maximum in the ice 0 population (a gray vertical band is drawn in Fig. 7 to highlight the location of both). To further support the association between the density minimum and the presence of a population of low-density local structures, we independently compute the local density of particles associated with each environment, and in the inset of Fig. 7, we plot the density ρ_c^* obtained by weighting these local densities with the fractional compositions obtained from LID (main panel). We see that ρ_c^* exactly mirrors ρ^*, showing that we obtain a good partial density decomposition. These results offer an even stronger case for the onionlike structure of growing nuclei, which in the case of water appears to be multistep.

The presence of onionlike structures and their radial composition is not explained by the small free-energy differences between the bulk phases. In fact, we observe that the cubic crystals (fcc and I_c) are found more abundantly near the center of the nucleus, while their hexagonal counterparts (hcp and I_h) are found more abundantly toward the surface. In terms of bulk free energies, instead the stable phases are fcc and ice I_h in hard spheres and mW water, respectively. To account for the ordering of the phases, one needs to consider the free-energy cost of structural fluctuations, which is size dependent. It has been observed that small finite-size clusters of the cubic phase gain relative stability compared to the hexagonal phase thanks to the entropy associated with stacking disorder [5,87,92] and the low energetic cost of their grain boundaries [4].

We repeat here the analysis of small (precritical) clusters that we perform for HS (see Sec. IV A 2). In Fig. 8, we show the average radial fraction composition of the main cluster for two size range. For clusters of size in the range $20 \leq n < 50$, the nucleus is composed of a mixture of I_c, I_h, and ice 0 with predominance of ice 0 and then of I_h. Going from precritical to immediately critical clusters, that is, for clusters of size in the range $50 \leq n < 100$, the onionlike structure starts to appear with ice 0 forming a peak which shifts toward the outer layers for increasing cluster size. Also, for mW water, as seen for HS, there is a selection of more compact clusters at the onset of the postcritical regime.

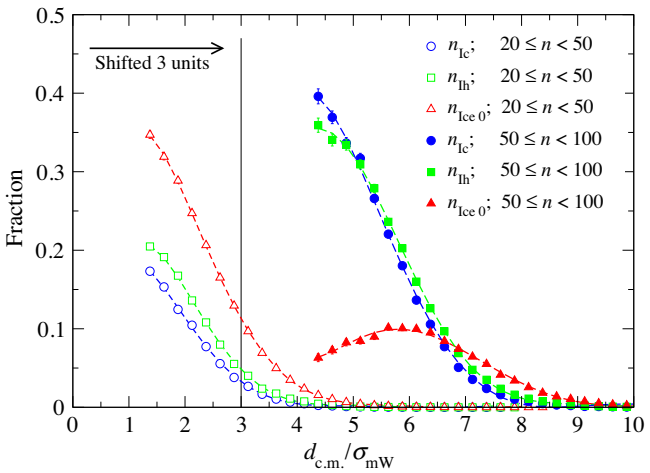

FIG. 8. Average radial fraction composition of the main cluster for clusters of size in the range 20–50 and 50–100 (shifted along the x axis of three units) as identified by LID. $d_{c.m.}$ is the distance from the center of mass of the cluster and σ_{mW} the mW water molecule diameter.

4. Equilibrium trajectories

To exclude the possibility that the nucleation pathway is due to the nonequilibrium nature of nucleation events at high supercooling, we apply the same analysis to trajectories obtained from umbrella sampling (US) simulations. Umbrella sampling, and other techniques such as metadynamics or forward flux sampling, are usually employed in homogeneous nucleation to enhance the sampling of a crystalline cluster [4,92–95]. In order to test the LID OP against homogeneous nucleation in mW water, which would confirm its ability to capture the local crystalline phases I_c, I_h, and ice 0, we bias the umbrella sampling simulations by using LID as a reaction coordinate. For performance reasons, here we construct LID by considering spatial first and second neighbors of a local particle, as done for hard spheres, instead of energetic neighbors. US simulations are performed with $N = 10\,000$ mW particles at ambient pressure and $T = 218$ K.

In Fig. 9, we show the average composition of the main cluster nucleated with US simulations for clusters of size $50 \leq n \leq 100$, as identified by the LID OP with spatial neighbors. Different from the spontaneous nucleation pathways analyzed before, US simulations allow us to study the structure of the nuclei in equilibrium. Moreover, it allows us to study nucleation at higher temperatures (where spontaneous nucleation would not be observed). Despite these differences, we get a very similar result to that obtained by using LID from spontaneous nucleation (see Fig. 7): I_c particles are more concentrated near the center of mass of nuclei, whereas I_h particles are slightly more abundant near the surface (note that small differences in the fraction composition between phases are magnified when computing the number of particles in a crystalline phase composing the nucleus because it depends on the square of their distance from the center of mass), and ice 0 particles concentrated around the surface of nuclei. In the inset of Fig. 9, we show, as in Fig. 7 for spontaneous nucleation simulations, the total fraction of crystalline particles (black diamonds), the normalized density $\rho^* = (\rho - \rho_x)/(\rho_f - \rho_x)$ (orange downward triangles), where $\rho_f = 0.995$ g/cm^3, $\rho_x = 0.983$ g/cm^3, and $\rho_{\text{ice0}} = 0.952$ g/cm^3 at the present thermodynamic conditions, and the normalized density ρ_c^* (magenta rightward triangles) computed by weighting the local densities of each phase with their fractional compositions obtained from LID (main panel). Note that here the linear transformation applied to ρ in order to get a normalized density ρ^* differs from the one used in Fig. 7 for the swap of ρ_f with ρ_x because at $T = 218$ K $\rho_f > \rho_x$, while at $T = 204$ K it is the opposite (see Ref. [4]).

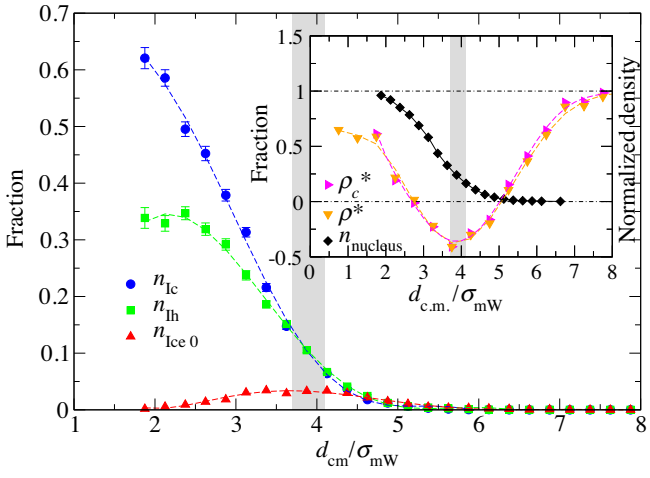

FIG. 9. Homogeneous nucleation of mW water from LID-biased umbrella sampling. Average radial fraction composition of the main cluster (for clusters of size $50 \leq n \leq 100$) as identified by LID with neighbors identification from spatial condition (see text). $d_{\text{c.m.}}$ is the distance from the center of mass of the cluster and σ_{mW} the mW water molecule diameter. The inset shows the normalized density profiles (colored symbols) for the same nuclei considered in the main panel, while black symbols represent the total fraction of all crystalline particles. Dashed fitting lines are a guide for the eyes.

Similar to the HS case (right panel of Fig. 4), in Fig. 10 we show the normalized histograms of $F(n, f) = -\log P(n, f)$. The reaction coordinate f is defined in the same way as for HS where the radius of the sphere defining the core is now $3\sigma_{\text{mW}}$. Figure 10(a) shows direct molecular simulations, while Figs. 10(b) and 10(c) are the result of US simulations, again at $T = 218$ K, in which we initialize the configurations using the seeding technique [96] with nuclei in the I_{sd} and I_c phases, respectively. I_{sd} is the stacking-disordered phase. For details on the simulation procedure, see Ref. [4]. From Fig. 10(a), we can see the presence of the two basins, the I_c core poor ($f < 0.5$) at small n, and the I_c core rich ($f > 0.5$) at large n, separated

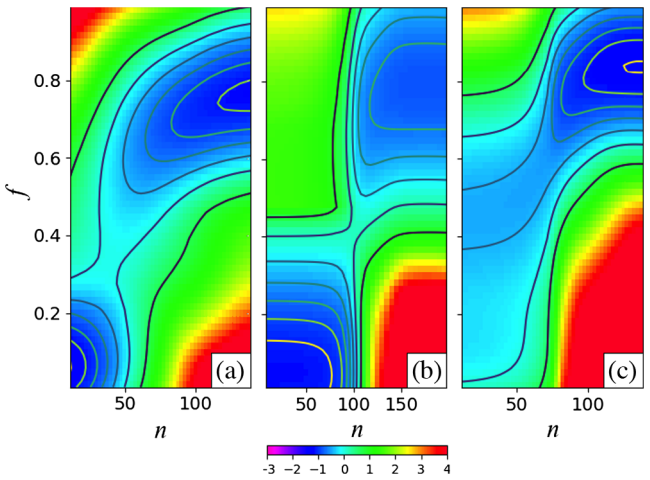

FIG. 10. 2D plots of $F(n, f) = -\log P(n, f)$ for direct molecular simulations (a), and US simulations with configurations initialized using the seeding technique to have a nucleus in the I_{sd} (b) or I_c (c) phase.

by the saddle point located in correspondence to the critical nucleus (at $T = 218$ K n_c is close to approximately 100). The US simulations [Figs. 10(b) and 10(c)] offer a view on the equilibrium landscape of the nucleation process for the formation of different nuclei: I_{sd} nuclei in Fig. 10(b) and I_c nuclei in Fig. 10(c). The potential of the mean force for the I_{sd} nucleation shows two channels: one corresponding to I_c-core-poor nuclei at small n and one with I_c-core-rich nuclei at large n. The overall process in this case is similar to the one observed in direct simulations [Fig. 10(a)]. The potential of the mean force for the I_c nuclei in Fig. 10(c) displays a process devoid of the I_c-core-poor basin, showing the existence of well-separated nucleation channels [44].

5. Dynamical behavior

To study the dynamical behavior of the growing nucleus, we compute how many particles attaching to the nucleus change their phase and how many do not during the entire dynamical process as a function of the nucleus size. In particular, we trace the evolution of particles in the main cluster in reverse time: For each trajectory, we count how many particles of the main cluster, which are in a specific phase at the end of the dynamics, are still found to be in that phase at the time when they attached to the cluster as a function of the cluster size n at that time. In Fig. 11, we show the conditional probability that a particle in a cluster of size n will always stay in the I_c (I_h) phase during the whole dynamics, indicated with black circles (green diamonds), and the conditional probability that at the end of the dynamics it will be in the opposite phase, indicated with red squares (blue triangles). In this case, we use the LID method to identify the local structure around each particle.

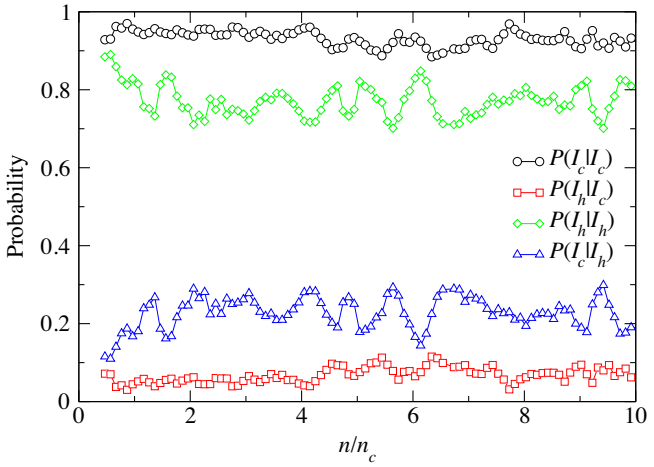

FIG. 11. Conditional probability that a particle attaching to a nucleus of size n will stay in the same phase until the end of the dynamics (black circles for I_c and green diamonds for I_h) or change phase at the end of the dynamics (red squares for I_c which transforms into I_h, and blue triangles for I_h which transforms into I_c).

From Fig. 11, we can see that for critical clusters (that is, for $n/n_c > 1$), on average a particle appearing in the main cluster of size n with phase I_c (I_h) will stay in that phase for the whole dynamics with a conditional probability $p(I_c|I_c) \simeq 0.93$ [$p(I_h|I_h) \simeq 0.77$]. Also, the probability of starting with a phase and ending with the other phase is not symmetric: Particles appearing in the main cluster of size n with phase I_c (I_h) will end up to be in the I_h (I_c) phase with a conditional probability $p(I_h|I_c) \simeq 0.07$ [$p(I_c|I_h) \simeq 0.23$].

We see that hexagonal local environments (more abundant on the surface) are more likely to change to cubic local environments as they get incorporated into the nucleus during the growth stage. To confirm that this transformation occurs on the surface, i.e., soon after local environments become crystalline, in Fig. 12 we compute the probability distribution of the time between the first appearance of the crystalline environment (black diamond symbols for I_c and red triangles for I_h) and its last phase transformation. We see that transformations occur exponentially fast in time following the same curve for both transformations and are thus surface events.

6. Precursors

Here we investigate the nature of the density decrease in proximity of the surface of the nucleus as found in the radial compositions of Figs. 7 and 9. In the upper panel of Fig. 13, we show the size n of clusters identified by LID as a function of Monte Carlo (MC) steps for a specific trajectory. The horizontal dashed red line corresponds to the critical nucleus size $n_c = 47$. We define t^* as the time when the nucleus has the critical size $n = n_c$ for the last time during the growth process (vertical dashed orange line in the figure). In the lower panel of Fig. 13, we show the system density ρ as a function of the MC steps for the same trajectory considered in the upper panel. The horizontal dashed blue lines are obtained from density averages over a short time interval and highlight that ρ decreases in

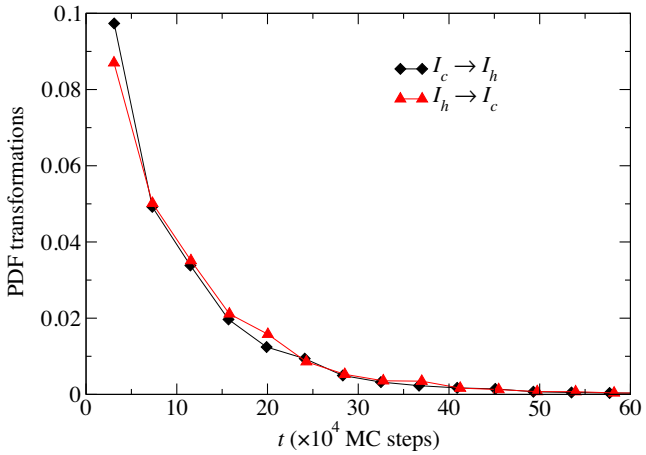

FIG. 12. Probability distribution function (PDF) of phase transformation versus time t in 10^4 MC steps units.

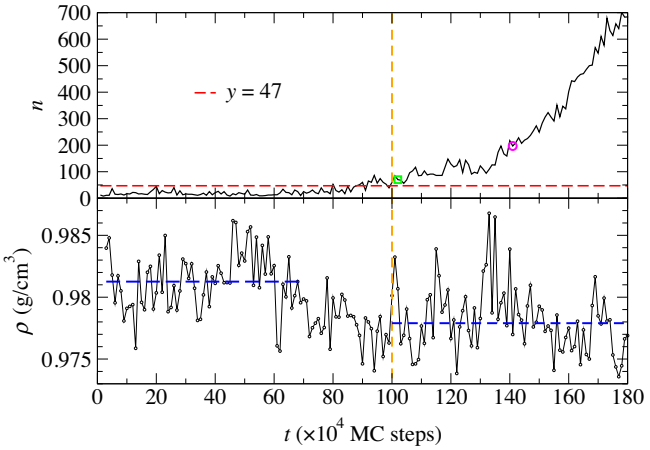

FIG. 13. Upper panel: size of the main cluster n, as identified by LID, versus time t in 10^4 MC steps units for a specific nucleation trajectory. The horizontal dashed red line corresponds to the critical nucleus size $n_c = 47$, as identified by LID. The vertical dashed orange line corresponds to the largest time at which $n = n_c$. The green square and violet circle correspond to the points (102,70) and (141,197), respectively. Lower panel: system density ρ versus time t in 10^4 MC steps units for the same nucleation trajectory considered in the upper panel. The vertical dashed orange line corresponds to the largest time at which $n = n_c$. The horizontal dashed blue lines are obtained from density averages over a short time interval and are a guide for the eyes.

correspondence to the formation of the critical nucleus. At the thermodynamic conditions we consider here ($P = 0$ Pa, $T = 204$ K), the density of ice I, ice 0, and liquid phase is

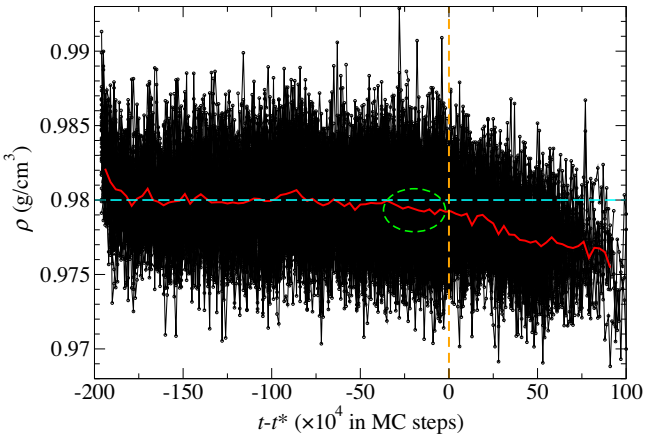

FIG. 14. System density ρ of all nucleation trajectories as a function of $t - t^*$ in 10^4 MC steps units, where t^* is the last time (vertical dashed orange line) at which the main cluster in the system (identified by using LID) has critical size ($n = n_c$). t^* is different for every trajectory i. The horizontal dashed cyan line represents the bulk liquid density ($\rho_L = 0.980$ g/cm^3 at the present thermodynamic conditions). The red line is the average of all densities. The green dashed circle highlight a precritical precursor region.

$\rho = 0.984, 0.953, 0.980$ g/cm^3, respectively (see Ref. [4]). Different from classical predictions for which the formation of a crystalline nucleus should correspond to an increase in density at the present thermodynamic conditions, here we see the opposite. As we discuss in Sec. IV B 3, this density decrease can be explained by the formation of ice-0-like local structures in correspondence to the nucleus surface (see Fig. 7).

The same trend is observed in all nucleating trajectories i: In Fig. 14, we plot the densities as a function of the time from t_i^*. In Fig. 14, the red line is the average density $\langle \rho(t - t_i^*) \rangle$, the horizontal dashed cyan line shows the bulk liquid density $\rho_L = 0.980$ g/cm^3 at $T = 204$ K, the vertical dashed orange line corresponds to $t = t^*$ for each trajectory i, and the dashed green circle indicates precritical nuclei. From Fig. 14, we see that the average density steadily decreases from precritical precursor regions.

V. CONCLUSIONS

In two-step nucleation, an intermediate phase is in size-dependent competition with the stable phase (fcc vs hcp in hard spheres or cubic ice vs hexagonal ice in mW water) [44,97,98]. Here we consider this phenomenology in the particular case of polytype nucleation. We study the microscopic nucleation pathway in systems characterized by a competition of different polytypes, whose bulk free-energy properties do not discriminate between them. Even in systems where no classical argument for a two-step process is expected, we find a selection of critical clusters with a compact structure that leads to the formation of onionlike structures, thus considerably extending the number of systems showing this type of nucleation mechanism [30,39–43]. In particular, our results highlight the role of structural fluctuations in nucleation phenomena [30,44].

Our results hinge on the development of a novel order parameter for local structure identification which is multidimensional and lossless, and is shown to successfully characterize these complex nucleation pathways and to identify local structures with high accuracy. A proper polymorph decomposition, for example, is essential in the determination of the nucleation rate [20]. We believe that the generality and flexibility of our method makes it suitable for the study of a large range of systems showing characteristic ordered or disordered signatures, such as defects or interfaces in crystalline or amorphous materials.

ACKNOWLEDGMENTS

We acknowledge support from the European Research Council Grant No. DLV-759187. We thank A. Attanasi and M. Mosayebi for useful discussions. This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol [99].

APPENDIX: LOCAL STRUCTURE IDENTIFICATION METHODS

1. Common neighbor analysis

The CNA method [100] assigns a structure type to every particle based on a nearest-neighbor graph accounting for the bond connectivity among neighbors of a given particle. Particles are considered to be neighbors if they are closer to each other than a specific cutoff. In the present work, for HS we employ the adaptive CNA (ACNA) method [101], in which an optimal cutoff radius is automatically computed for each individual particle. A major disadvantage of CNA is that no structure type is assigned to particles with unknown signatures, and it is sensitive to thermal fluctuations [102].

2. Extended common neighbor analysis

In order to assign a cubic or hexagonal diamond structure type to a water oxygen atom, information on the position of its second-nearest neighbors (i.e., second shell) are needed. In the diamond structure, nearest-neighbor oxygen atoms do not have common neighbors, and the second and third shells are not well separated. In order to apply the CNA method to identify diamond structures, the ext CNA has been introduced in Ref. [103]. In the software OVITO [85], it is available as the "identify diamond structure" function. In the ext CNA, the CNA method is applied to the 12 second-nearest neighbors of a central particle, which are found as the first neighbors of the first four neighbors of the central particle under consideration. We refer to this CNA method to identify particles in the mW water model. We also consider the method we name ext CNA 1st (available as option in OVITO), which includes in the ice I structures also particles being first neighbors of a particle classified as ice I by the ext CNA method. These additional particles have four first neighbors positioned on the right lattice sites of the relative ice I structure, but at least one of its second-nearest neighbors is off lattice.

3. Polyhedral template matching

This method is based on the topology of the local particle environment [102]. It makes use of the convex hull formed by a fixed number of neighboring particles, which are identified using a Voronoi-based method. The planar graph representing the convex hull is used to classify structures. Polyhedral template matching (PTM) is less sensitive to thermal fluctuations with respect to ACNA, but it still requires the definition of reference structures. In Fig. 15, we show the nucleus spanning the simulation box displayed in the main text in Fig. 1, but here identified by using (from left to right) ACNA, PTM, BOO up to second shell, BOO up to first shell, and LID. We notice that BOO up to the second shell shows problems in distinguishing parallel layers of alternating phases when they are close to each other, while BOO up to the first shell improves the identification of those parallel layers of alternating phases, even though it gives similar results on the composition of polytypes with respect to BOO up to second shell.

4. CHILL+

CHILL+ [104] classifies cubic ice, hexagonal ice, and clathrate hydrate structures in water. It is based on the identification of staggered and eclipsed bonds: Since an oxygen atom in crystalline ice is 4-coordinated (first neighboring shell), if we consider two neighboring oxygen atoms, we can look at the cluster of eight atoms composed by these two and their first neighbors. Looking at the atoms along the axis of the bond between the first two atoms, if all six neighboring atoms are visible, we have a staggered bond, while if we see three neighboring atoms, we have an eclipsed bond. Because of the presence of thermal fluctuations and other effects distorting bonds, as for other methods comparing local environments to a reference structure, thresholds to establish if a bond is close enough to the perfect staggered or eclipsed bond and then being identified with it have to be introduced. In particular, if the bond order parameter q_{3m} is between 0.25 and -0.35, the bond is eclipsed, while if it is less than -0.8, the bond is staggered. The crystalline structure associated with an oxygen atom depends on the number of eclipsed and staggered bonds. For example, hexagonal ice has one eclipsed and three staggered bonds, while cubic ice has all four bonds staggered. This method is specific for water.

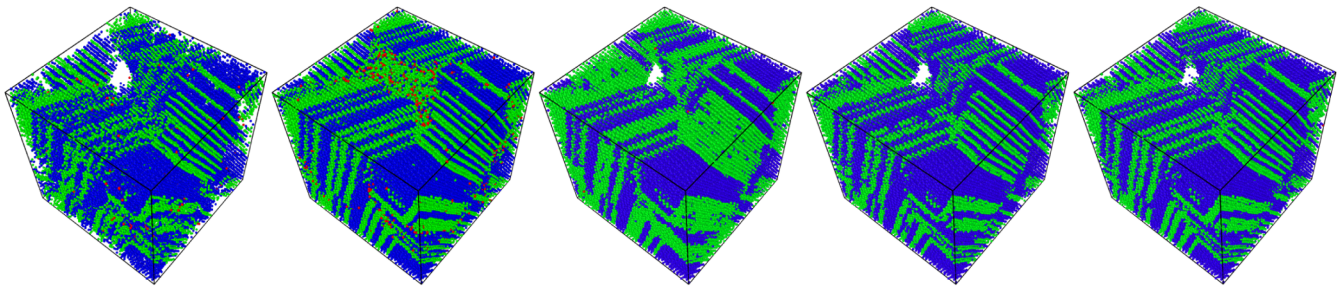

FIG. 15. Homogeneous nucleation of hard spheres. Particles structure from the same configuration (snapshot) are identified using the following methods (from left to right): ACNA, PTM, BOO up to second shell, BOO up to first shell, and LID. In all panels, the colors associated with fcc, hcp, and bcc structures are blue, green, and red, respectively. The calculation of ACNA and PTM, and snapshots visualization are obtained using OVITO [85].

5. Bond orientational order

Steinhardt or BOO parameters $q_l(i)$ and $w_l(i)$ describe local order (as seen from particle i) in terms of spherical harmonics of order l. They are based on the following complex vector $q_{lm}(i)$ associated with the particle i,

$$q_{lm}(i) = \frac{1}{N_b(i)} \sum_{j=1}^{N_b(i)} Y_{lm}(r_{lm}), \tag{A1}$$

where $N_b(i)$ is the number of neighbors of particle i, l is an integer, and m is an integer running from $m = -l$ to $m = l$, $Y_{lm}(r_{ij})$ are the spherical harmonics, and r_{ij} is the position vector from particle i to j, and on the averaged $\bar{q}_{lm}(i)$ defined as

$$\bar{q}_{lm}(i) = \frac{1}{N_b(i)+1} \sum_{k \in \{i, N_b(i)\}} q_{lm}(k), \tag{A2}$$

where the sum runs over the $N_b(i)$ plus the particle i. The local bond order, or Steinhardt, parameters $q_l(i)$ are defined as

$$q_l(i) = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |q_{lm}(i)|^2}. \tag{A3}$$

The parameter corresponding to a specific value of l captures a specific crystal symmetry. All $q_l(i)$ depend on the angles formed by neighboring particles and are independent of a reference frame. The averaged Steinhardt OPs $\bar{q}_l(i)$ are defined as

$$\bar{q}_l(i) = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |\bar{q}_{lm}(i)|^2}. \tag{A4}$$

Cubic Steinhardt OPs $w_l(i)$ are defined as

$$w_l(i) = \frac{\sum_{m_1+m_2+m_3=0} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} q_{lm_1}(i) q_{lm_2}(i) q_{lm_3}(i)}{\left(\sum_{m=-l}^{l} |q_{lm}(i)|^2\right)^{3/2}}, \tag{A5}$$

where the term in parentheses is the Wigner $3j$ symbol, while the cubic averaged Steinhardt OPs $\bar{w}_l(i)$ are defined as

$$\bar{w}_l(i) = \frac{\sum_{m_1+m_2+m_3=0} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} \bar{q}_{lm_1}(i) \bar{q}_{lm_2}(i) \bar{q}_{lm_3}(i)}{\left(\sum_{m=-l}^{l} |\bar{q}_{lm}(i)|^2\right)^{3/2}}. \tag{A6}$$

Here we consider $\bar{q}_{12}$ only for the identification of the solid nucleus without distinguishing polytypes, which has been used in other works [4,105], and two methods based on BOO OP for the identification of all phases: (i) the standard $\bar{q}_4\bar{q}_6$ map, in which case, the choice of the protocol to compute and partition the map can strongly affect its application and (ii) a group of 30 BOO, as described in the following, to considerably increase the dimensionality of the order-parameter space which allows us to easily increase the separation between the different populations of local environments we want to discriminate between. The OP we use as input for the NNs is composed of the following 30 BOO: $q_l(i)$ with $l = 3, 4, \ldots, 12$, $\bar{q}_l(i)$ with $l = 3, 4, \ldots, 12$, $w_l(i)$ with $l = 4$, 6, 8, 10, 12, and $\bar{w}_l(i)$ with $l = 4$, 6, 8, 10, 12. There are different ways to obtain first and second shells of neighbors in order to compute the BOO, like the SANN algorithm [75] or using a fixed cutoff (see Sec. II D). Here we consider the first neighbors shell as composed by the N particles closer to the particle under investigation, and the second neighbors shell as composed by the M particles closer to the particle under investigation, excluding the first N particles. As N and M, we consider $N = 12$ and $M = 6$ for HS, while $N = 4$ and $M = 12$ for mW water. These values for N and M are related to the number of first and second neighbors in the crystalline structures forming in these models.

6. $\bar{q}_4\bar{q}_6$ sensitivity to protocols

This method for local structure identification is very popular, but, as we discuss in the main text, it is very sensitive to the way in which it is computed and to the thresholds used to partition the map. Here we show that, when applied to the determination of the nucleus size and its composition of mW water, the $\bar{q}_4\bar{q}_6$ method can give very different results.

First of all, in order to define the neighbors of a particle i, two approaches are usually employed: considering the n_n particles closer to particle i or considering all the n_{cut} particles found at a distance from particle i smaller than r_{cut}. Once the $\bar{q}_4\bar{q}_6$ map is computed, it can be partitioned in different ways.

In Fig. 16, we show the $\bar{q}_4\bar{q}_6$ map obtained by considering n_{cut} neighbors with $r_{\mathrm{cut}} = 1.43\sigma_{\mathrm{mW}}$, and the particles phase is associated with fluid if $\bar{q}_6 < 0.415$, otherwise they are crystalline, and in particular, in the phase I_c if $\bar{q}_4 > 0.425$, and I_h otherwise (orange dashed lines correspond to these thresholds). This method named LD-A in Ref. [24] (where LD stands for Lechner and Dellago) does not discriminate between the liquid phase and ice 0 (black and red dots corresponding to the fluid phase and ice 0, respectively, overlap and then cannot be distinguished; see Fig. 16).

In Fig. 16, we show also another choice of thresholds to partition the $\bar{q}_4\bar{q}_6$ map that we name LD-A2: Particles are fluid if $\bar{q}_6 < 0.385$, otherwise, they are crystalline, and in particular, associated with the phase I_c if $2\bar{q}_4 > \bar{q}_6 + 0.35$, and I_h otherwise (dark green dash-dotted lines correspond to these thresholds). This choice of thresholds allows us to

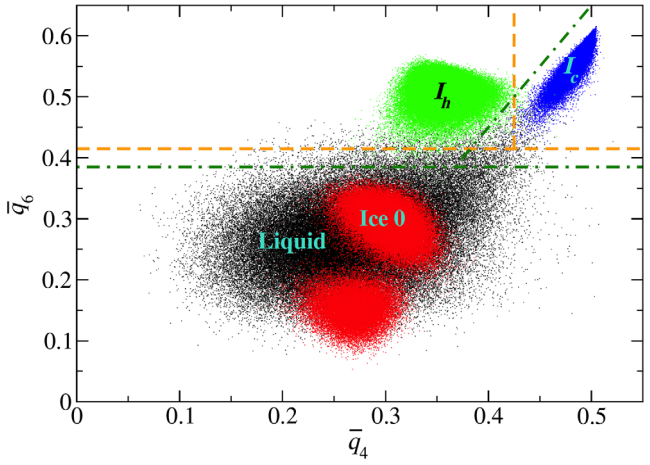

FIG. 16. $\bar{q}_4\bar{q}_6$ map calculation and partition following method LD-A(orange dashed lines), and LD-A2 (dark-green dot-dashed lines). Each dot corresponds to the $\bar{q}_4$, $\bar{q}_6$ coordinates associated with a particle of the following systems composed of $N = 5376$ mW particles at melting: I_c (blue), I_h (green), ice 0 (red), and liquid water (black).

better partition the $\bar{q}_4\bar{q}_6$ map at melting (not shown here) with respect to LD-A.

In Fig. 17, we show another method to obtain the $\bar{q}_4\bar{q}_6$ map, where the number of neighbors is fixed to $n_n = 16$, and the threshold is the following: If $\bar{q}_4 < 0.105$, particles are fluid, while crystalline in the opposite case. Crystalline particles are classified as ice 0 if $\bar{q}_6 < 0.11$, and ice I in the opposite case. Ice I particles are classified as I_c if $\bar{q}_4/0.36 + \bar{q}_6/0.45 > 1$, and I_h otherwise. This method named LD-B in Ref. [24] allows us to discriminate between the liquid phase and ice 0. In all cases, the $\bar{q}_4\bar{q}_6$ map is computed at the nucleation temperature $T = 204$ K and pressure $P = 0$ Pa.

In Fig. 18, we show the average first passage time t_{FP} described in Sec. IV B 1, as a function of the nucleus size n

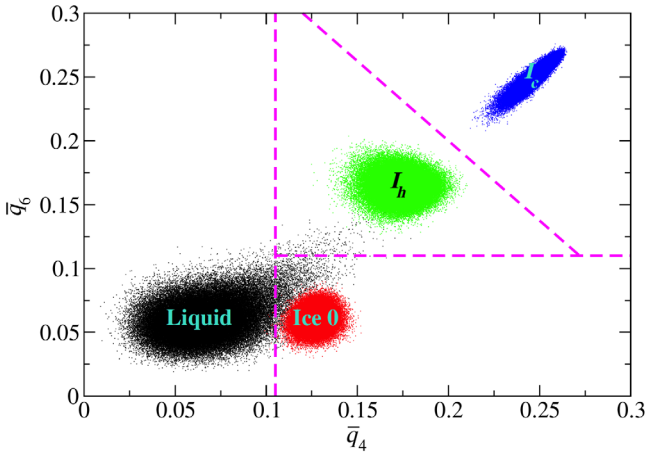

FIG. 17. $\bar{q}_4\bar{q}_6$ map calculation and partition following method LD-B.

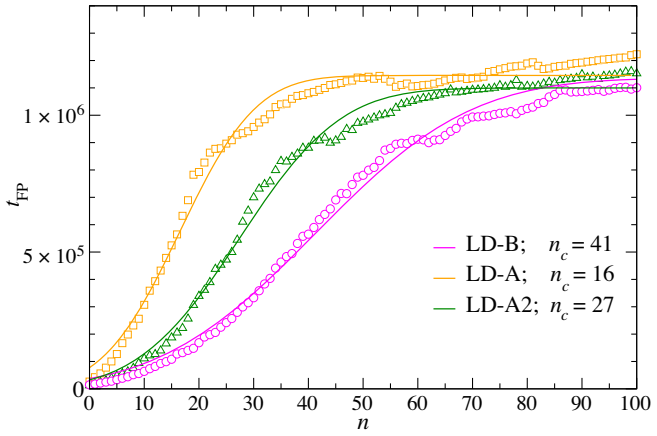

FIG. 18. Average first passage time t_{FP} as a function of the nucleus size computed using the $\bar{q}_4\bar{q}_6$ methods LD-A (orange squares), LD-A2 (dark-green triangles), and LD-B (magenta circles).

obtained applying the three different methods considered here to compute and partition the $\bar{q}_4\bar{q}_6$ map. We can notice the big variation in the value of the critical nucleus size n_c estimated from the different methods.

In Fig. 19, we show the ratio r between the number of particles n_{I_c} in the cubic phase and the number of particles n_{I_h} in the hexagonal phase found in the nucleus as a function of its size n divided by the critical nucleus size n_c applying the three different methods considered here to get the $\bar{q}_4\bar{q}_6$ map. As we find for the average first passage time, in this case also, each method gives a different estimation of r (averaging only on the stationary part, that is, excluding small cluster size): 0.94, 1.07, and 0.07 for LD-A, LD-A2, and LD-B, respectively. Even though LD-B is able to

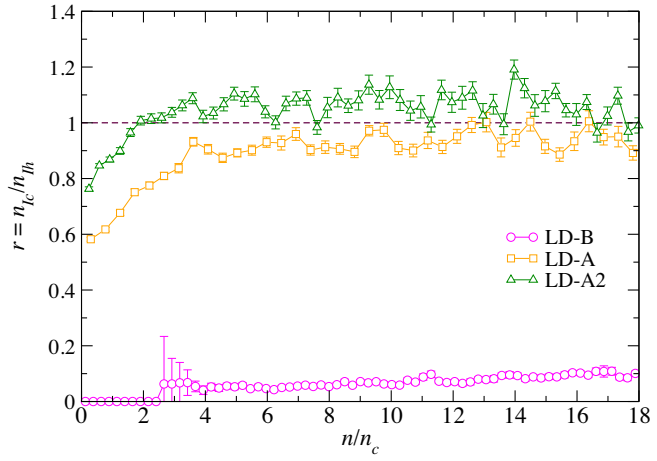

FIG. 19. Average ratio $r = n_{I_c}/n_{I_h}$ between the number of particles composing the nucleus in the cubic phase (I_c) and the hexagonal phase (I_h) using the $\bar{q}_4\bar{q}_6$ methods LD-A (orange squares), LD-A2 (dark-green triangles), and LD-B (magenta circles). r is plotted against the nucleus size n normalized by the critical nucleus size n_c for each specific method.

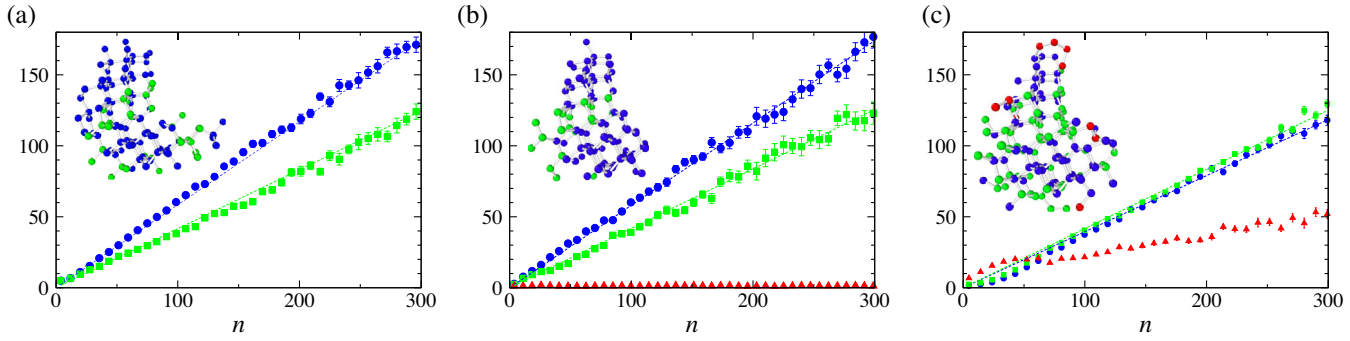

FIG. 20. Homogeneous nucleation of mW water: average composition of the main cluster as identified by (a) CNA up to first neighbors, (b) BOO, and (c) LID. Insets in (a), (b), and (c) show a typical nucleus composed of 179 (66 if disregarding I_c and I_h first neighbors), 104, and 144 particles, respectively. In all panels, the colors associated with I_c, I_h, ice 0 structures are blue, green, and red, respectively.

discriminate between the liquid phase and ice 0, it is strongly biased toward the hexagonal phase.

7. Composition of mW nuclei

Similar to Fig. 1, we show in Fig. 20 the average fractional composition as a function of the nucleus size for mW molecules at ambient pressure and temperature $T = 204$ K as identified by ext CNA 1st [Fig. 20(a)], BOO [Fig. 20(b)], and LID [Fig. 20(c)].

8. Benchmark

Considering the wide variation of results on the nucleus properties predicted by different methods adopted in the literature, some of which we analyze here, it would be desirable to find benchmarks to evaluate the accuracy and reliability of each. Here we propose a simple test in which we know by construction the phase of each particle belonging to the nucleus and we use different methods to identify them. We consider a cluster composed of particles of both phases ice I_c and I_h obtained from a perfect lattice of stacking ice with alternated layers of I_c and I_h at a density $\rho = 0.982$ g/cm^3 corresponding to the temperature $T = 235$ K at equilibrium conditions (see Ref. [4]). We obtain a cluster of size $n = 200$ following the minimum energy rule described in Ref. [4]. Then, we let the cluster equilibrate in contact with a liquid phase of density $\rho = 1.002$ g/cm^3 corresponding to equilibrium conditions at the temperature $T = 235$ K, using fixed-topology MC simulations (see Ref. [4]) which allow for bonds elongation up to a maximum cutoff (set to 1.3 Å), while keeping the topology fixed.

Since we know the phase (I_c or I_h) of each particle composing the cluster, using different methods, we identify each particle phase and compare this prediction with its true value. We distinguish particles of the cluster as belonging to different regions depending on the number of their first neighbors (FN) and the sum of first neighbors of first neighbors (FN2) in the following way: For all regions under consideration FN = 4, while FN2 = 16, 15, 14, 13 for regions 1, 2, 3, and 4, respectively. Only particles belonging to region 1 have a fully formed second shell.

In Table I, we show (second column) the average percentage of particles of the cluster belonging to each region (first column), and the percentage of particles correctly identified as I_c, or incorrectly identified as I_h or as liquid phase L for the different methods (columns 3 to 17). In Table II, we show the same results, but for the identification of I_h. For example, for clusters of size $n = 200$ considered here, particles belonging to region 1 are on average only 16.4% of the total. These results are obtained by averaging over 20 different clusters realized by using the minimum energy rule and ten different evolution times. The identification method $\bar{q}_4\bar{q}_6$ is strongly

TABLE I. Benchmark of different methods for the identification of particles in the I_c phase present in nuclei of size $n = 200$. L refers to the liquid phase. Except for the number indicating the region, all the other numbers refer to percentages.

Test I_c		Ext CNA			Ext CNA 1st			CHILL+			BOO			LID		
Region	Particles	I_c	I_h	L	I_c	I_h	L	I_c	I_h	L	I_c	I_h	L	I_c	I_h	L
1	16.4	96.78	0.00	3.22	98.43	1.27	0.30	97.10	0.00	2.90	93.86	3.78	2.36	99.54	0.18	0.28
2	7.25	10.28	0.00	89.72	11.73	73.50	14.77	75.56	0.14	24.30	64.15	17.58	18.27	67.52	1.55	30.93
3	7.8	0.60	0.00	99.40	34.22	45.58	20.20	57.76	1.53	40.71	63.51	9.52	26.97	10.93	1.58	87.49
4	9.05	0.12	0.00	99.88	16.37	60.04	23.60	17.69	1.18	76.13	30.17	5.48	59.35	4.15	0.72	95.08

TABLE II. Benchmark of different methods for the identification of particles in the I_h phase present in nuclei of size $n = 200$. L refers to the liquid phase. Except for the number indicating the region, all the other numbers refer to percentages.

Test I_h		Ext CNA			Ext CNA 1st			CHILL+			BOO			LID		
Region	Particles	I_c	I_h	L	I_c	I_h	L	I_c	I_h	L	I_c	I_h	L	I_c	I_h	L
1	16.4	0.00	96.01	3.99	3.06	96.37	0.58	0.00	97.11	2.89	4.53	93.63	1.84	0.51	98.85	0.64
2	7.25	0.00	4.94	95.06	68.31	24.26	7.43	0.00	76.70	23.30	19.63	73.40	6.97	5.98	38.84	55.18
3	7.8	0.00	1.65	98.35	71.32	10.76	17.92	0.00	64.50	35.50	11.43	65.48	23.09	0.64	8.75	90.61
4	9.05	0.00	0.00	100.00	41.29	21.38	37.33	1.67	9.69	78.64	19.68	57.66	22.66	0.41	2.27	97.32

affected by the choice of the protocol used to compute and partition it (see Ref. [24] and the Appendix 6), and then it is not shown in the tables.

From Tables I and II, we can see, for example, that the method ext CNA correctly identifies the cubic and hexagonal ice particles in region 1 approximately 96% of the time. When considering other regions, the percentage of correct identification quickly goes to zero for increasing region label, that is, for more and more incomplete second shells, in which case, particles are more likely associated with a liquid phase. This behavior is reflected in the very small value of the critical cluster size $n_c = 4$ obtained with this method (see Fig. 5). In the case of ext CNA 1st, the performance in region 1 is similar to the method ext CNA, while particles in other regions are mainly identified as crystalline. However, as we also note by snapshots inspection, in regions 2–4, ext CNA 1st misidentifies crystalline particles, often associating the I_c phase with I_h particles and vice versa. This result is not surprising considering that ext CNA 1st is likely to associate to the first neighbors of a particle in the I_c (I_h) phase the same I_c (I_h) phase (see the Appendix 2), and nuclei tested in the present benchmark are composed of alternating layers of the I_c and I_h phases. For this reason, when using the "identify diamond structure" function of OVITO, it would be important to specify if also first neighbors or even second neighbors of crystalline particles are included in the method to compute quantities like, for example, the cubicity which gives a measure of the amount of I_c with respect to I_h composing the nucleus. Finally, BOO shows a good identification rate with limited misidentifications, while LID and CHILL+ give the best performance with extremely low misidentifications.

A conservative way to rate the performance of a method from these benchmarks is to evaluate the percentage of particles correctly identified in region 1 (particles with a fully formed second shell) and the misidentification for increasing region label. From these considerations, we conclude that ext CNA is too conservative, missing many crystalline particles of the nucleus, while ext CNA 1st is affected by an important misidentification of crystalline particles with incomplete second shells. BOO shows low misidentification of crystalline particles. On the other hand, LID and CHILL+ are the two methods with the lowest misidentification, with LID showing the best performance for identification of crystalline phases in region 1.

In order to evaluate the influence of thermal fluctuations on the particle identification methods, we repeat the previous benchmark, but this time considering rigid clusters (no bonds elongation) equilibrated with the liquid phase. Also in this case, we observe a similar behavior of the different methods.

9. Correlation between precursors and OP

For each particle i, we compute the Euclidean distance d_i^{LID} between the vector LID at a specific time and the LID associated with the perfect crystalline structure. Here we consider as reference the LID signal associated with I_c, as very similar results are obtained with respect to I_h (not shown). In the following, we show the value of d_i^{LID} associated with each particle of a sample at two specific times (see Fig. 13) at which the nucleus has a size of $n = 70$ (Figs. 21 and 22) and $n = 197$ (Fig. 22) (see the green square for $n = 70$ and the violet circle for $n = 197$ in

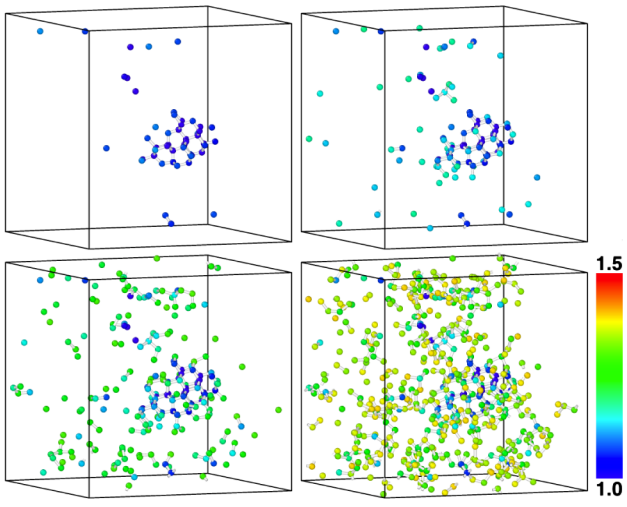

FIG. 21. Snapshots of a nucleation trajectory at time $t = 102$ in 10^4 MC steps units to which it corresponds the presence of a nucleus identified with LID of size $n = 70$ (see Fig. 13). Colors are assigned to particles whose distance d_i^{LID} is smaller than 1.1 (top left panel), 1.2 (top right panel), 1.3 (bottom left panel), or 1.4 (bottom right panel).

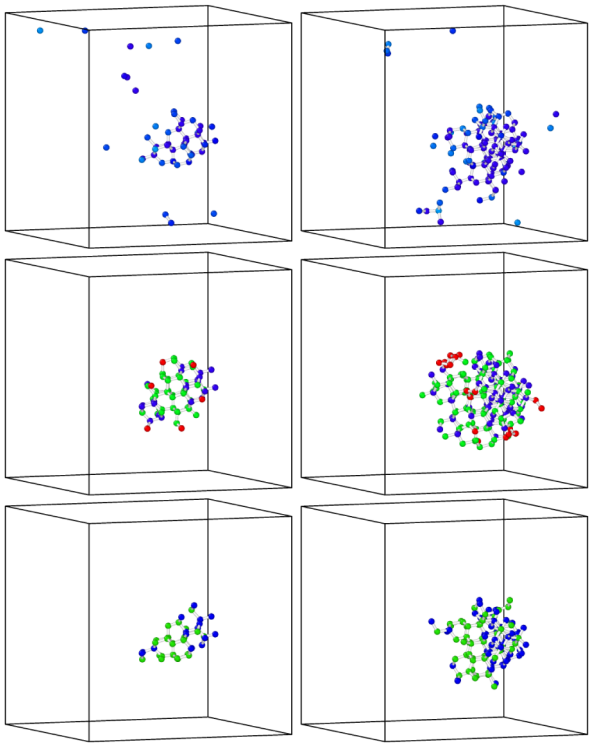

FIG. 22. Snapshots of a nucleation trajectory at time $t = 102$ (left panels) and $t = 141$ (right panels) in 10^4 MC steps units showing the presence of a nucleus identified with particles whose distance d_i^{LID} is smaller than 1.1 (upper panels), LID (middle panels; blue, green, and red for I_c, I_h, and ice 0, respectively) giving a nucleus of size $n = 70$ (left panel) and $n = 197$ (right panel) (see Fig. 13), CHILL+ (lower panels; blue and green for I_c and I_h, respectively).

Fig. 13). In Fig. 21, we show snapshots corresponding to the nucleus of size $n = 70$, where particles i with a distance d_i^{LID} smaller than 1.1, 1.2, 1.3, and 1.4 (from left to right and from top to bottom) are shown with a color code going from 1.0 (blue) to 1.5 (red). The field d_i^{LID} correlates with crystalline structures present in the system (see top left snapshot in Fig. 21), and in particular, with the main cluster as detected by other methods (see Fig. 22).

In Fig. 22, from top to bottom, we show particles with the Euclidean distance $d_i^{\mathrm{LID}} < 1.1$ (see Fig. 21 for color codes), particles belonging to the main cluster as identified by using LID (blue for I_c, green for I_h, and red for ice 0), and particles belonging to the main cluster as identified by using the CHILL+ algorithm (blue for I_c and green for I_h). The left (right) column in Fig. 22 refers to a snapshot of the nucleation trajectory shown in Fig. 13 at the time $t = 102$ ($t = 141$) in 10^4 MC steps units. From Fig. 22, we can see that d_i^{LID} correlates very well with the nucleus identified by LID and CHILL+, and that the latter method, apart from not providing ice 0 particles, finds a smaller nucleus, as expected from its ability to estimate a smaller value of the critical nucleus respect to LID (see Fig. 5).

[1] E. B. Moore and V. Molinero, *Structural Transformation in Supercooled Water Controls the Crystallization Rate of Ice*, Nature (London) **479**, 506 (2011).

[2] T. Bartels-Rausch, V. Bergeron, J. H. E. Cartwright, R. Escribano, J. L. Finney, H. Grothe, P. J. Gutierrez, J. Haapala, W. F. Kuhs, J. B. C. Pettersson *et al.*, *Ice Structures, Patterns, and Processes: A View across the Icefields*, Rev. Mod. Phys. **84**, 885 (2012).

[3] G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen, and A. Michaelides, *Crystal Nucleation in Liquids: Open Questions and Future Challenges in Molecular Dynamics Simulations*, Chem. Rev. **116**, 7078 (2016).

[4] F. Leoni, R. Shi, H. Tanaka, and J. Russo, *Crystalline Clusters in mW Water: Stability, Growth, and Grain Boundaries*, J. Chem. Phys. **151**, 044505 (2019).

[5] L. Lupi, A. Hudait, B. Peters, M. Grünwald, R. Gotchy Mullen, A. H. Nguyen, and V. Molinero, *Role of Stacking Disorder in Ice Nucleation*, Nature (London) **551**, 218 (2017).

[6] Y. J. Kaufman, D. Tanré, and O. Boucher, *A Satellite View of Aerosols in the Climate System*, Nature (London) **419**, 215 (2002).

[7] B. J. Murray, D. A. Knopf, and A. K. Bertram, *The Formation of Cubic Ice under Conditions Relevant to Earth's Atmosphere*, Nature (London) **434**, 202 (2005).

[8] S. Sastry, *Ins and Outs of Ice Nucleation*, Nature (London) **438**, 746 (2005).

[9] K. Sassen, *Dusty Ice Clouds over Alaska*, Nature (London) **434**, 456 (2005).

[10] R. J. Herbert, B. J. Murray, S. J. Dobbie, and T. Koop, *Sensitivity of Liquid Clouds to Homogeneous Parametrizations*, Geophys. Res. Lett. **42**, 1599 (2015).

[11] R. A. Shaw, A. J. Durant, and Y. Mi, *Heterogeneous Surface Crystallization Observed in Undercooled Water*, J. Phys. Chem. B **109**, 9865 (2005).

[12] A. Y. Lee, D. Erdemir, and A. S. Myerson, *Crystal Polymorphism in Chemical Process Development*, Annu. Rev. Chem. Biomol. Eng. **2**, 259 (2011).

[13] P. Vishweshwar, J. A. McMahon, M. Oliveira, M. L. Peterson, and M. J. Zaworotko, *The Predictably Elusive Form II of Aspirin*, J. Am. Chem. Soc. **127**, 16802 (2005).

[14] L. Berthier and G. Tarjus, *Nonperturbative Effect of Attractive Forces in Viscous Liquids*, Phys. Rev. Lett. **103**, 170601 (2009).

[15] A. K. Bacher, T. B. Schrøeder, and J. C. Dyre, *Explaining Why Simple Liquids Are Quasi-Universal*, Nat. Commun. **5**, 5424 (2014).

[16] V. Molinero and E. B. Moore, *Water Modeled as an Intermediate Element between Carbon and Silicon*, J. Phys. Chem. B **113**, 4008 (2009).

[17] S. Pronk and D. Frenkel, *Can Stacking Faults in Hard-Sphere Crystals Anneal Out Spontaneously?*, J. Chem. Phys. **110**, 4589 (1999).

[18] T. L. Malkin, B. J. Murray, C. G. Salzmann, V. Molinero, S. J. Pickering, and T. F. Whale, *Stacking Sisorder in Ice I*, Phys. Chem. Chem. Phys. **17**, 60 (2015).

[19] A. Zaragoza, M. M. Conde, J. R. Espinosa, C. Valeriani, C. Vega, and E. Sanz, *Competition between Ices Ih and Ic in Homogeneous Water Freezing*, J. Chem. Phys. **143**, 134504 (2015).

[20] B. Cheng, C. Dellago, and M. Ceriotti, *Theoretical Prediction of the Homogeneous Ice Nucleation Rate: Disentangling Thermodynamics and Kinetics*, Phys. Chem. Chem. Phys. **20**, 28732 (2018).
[21] D. Quigley, *Communication: Thermodynamics of Stacking Disorder in Ice Nuclei*, J. Chem. Phys. **141**, 121101 (2014).
[22] H. Tanaka, H. Tong, R. Shi, and J. Russo, *Revealing Key Structural Features Hidden in Liquids and Glasses*, Nat. Rev. Phys. **1**, 333 (2019).
[23] E. Allahyarov, K. Sandomirski, S. U. Egelhaaf, and H. Löwen, *Crystallization Seeds Favour Crystallization Only during Initial Growth*, Nat. Commun. **6**, 7110 (2015).
[24] S. Prestipino, *The Barrier to Ice Nucleation in Monatomic Water*, J. Chem. Phys. **148**, 124505 (2018).
[25] D. Kashchiev, P. G. Vekilov, and A. B. Kolomeisky, *Kinetics of Two-Step Nucleation of Crystals*, J. Chem. Phys. **122**, 244706 (2005).
[26] D. Erdemir, A. Y. Lee, and A. S. Myerson, *Nucleation of Crystals from Solution: Classical and Two-Step Models*, Acc. Chem. Res. **42**, 621 (2009).
[27] P. G. Vekilov, *The Two-Step Mechanism of Nucleation of Crystals in Solution*, Nanoscale **2**, 2346 (2010).
[28] T. Schilling, H. J. Schöpe, M. Oettel, G. Opletal, and I. Snook, *Precursor-Mediated Crystallization Process in Suspensions of Hard Spheres*, Phys. Rev. Lett. **105**, 025701 (2010).
[29] T. K. Haxton, L. O. Hedges, and S. Whitelam, *Crystallization and Arrest Mechanisms of Model Colloids*, Soft Matter **11**, 9307 (2015).
[30] J. Russo and H. Tanaka, *Nonclassical Pathways of Crystallization in Colloidal Systems*, MRS Bull. **41**, 369 (2016).
[31] G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen, and A. Michaelides, *Crystal Nucleation in Liquids: Open Questions and Future Challenges in Molecular Dynamics Simulations*, Chem. Rev. **116**, 7078 (2016).
[32] S. Lee, E. G. Teich, M. Engel, and S. C. Glotzer, *Entropic Colloidal Crystallization Pathways via Fluid-Fluid Transitions and Multidimensional Prenucleation Motifs*, Proc. Natl. Acad. Sci. U.S.A. **116**, 14843 (2019).
[33] P. Tan, N. Xu, and L. Xu, *Visualizing Kinetic Pathways of Homogeneous Nucleation in Colloidal Crystallization*, Nat. Phys. **10**, 73 (2014).
[34] H. Jiang, P. G. Debenedetti, and A. Z. Panagiotopoulos, *Nucleation in Aqueous NaCl Solutions Shifts from 1-Step to 2-Step Mechanism on Crossing the Spinodal*, J. Chem. Phys. **150**, 124502 (2019).
[35] D. Gebauer, A. Völkel, and H. Cölfen, *Stable Prenucleation Calcium Carbonate Clusters*, Science **322**, 1819 (2008).
[36] E. M. Pouget, P. H. H. Bomans, J. A. C. M. Goos, P. M. Frederik, G. de With, and N. A. J. M. Sommerdijk, *The Initial Stages of Template-Controlled* $CaCo_3$ *Formation Revealed by Cryo-TEM*, Science **323**, 1455 (2009).
[37] R. P. Sear, *The Non-Classical Nucleation of Crystals: Microscopic Mechanisms and Applications to Molecular Crystals, Ice and Calcium Carbonate*, Int. Mater. Rev. **57**, 328 (2012).
[38] D. Kashchiev, *Classical Nucleation Theory Approach to Two-Step Nucleation of Crystals*, J. Cryst. Growth **530**, 125300 (2020).
[39] G. I. Tóth, T. Pusztai, G. Tegze, G. Tóth, and L. Gránásy, *Amorphous Nucleation Precursor in Highly Nonequilibrium Fluids*, Phys. Rev. Lett. **107**, 175702 (2011).
[40] K. Barros and W. Klein, *Liquid to Solid Nucleation via Onion Structure Droplets*, J. Chem. Phys. **139**, 174505 (2013).
[41] M. Santra, R. S. Singh, and B. Bagchi, *Nucleation of a Stable Solid from Melt in the Presence of Multiple Metastable Intermediate Phases: Wetting, Ostwald's Step Rule, and Vanishing Polymorphs*, J. Phys. Chem. B **117**, 13154 (2013).
[42] J. F. Lutsko, *How Crystals Form: A Theory of Nucleation Pathways*, Sci. Adv. **5**, eaav7399 (2019).
[43] S. Tang, J. Wang, B. Svendsen, and D. Raabe, *Competitive bcc and fcc Crystal Nucleation from Non-Equilibrium Liquids Studied by Phase-Field Crystal Simulation*, Acta Mater. **139**, 196 (2017).
[44] D. James, S. Beairsto, C. Hartt, O. Zavalov, I. Saika-Voivod, R. K. Bowles, and P. H. Poole, *Phase Transitions in Fluctuations and Their Role in Two-Step Nucleation*, J. Chem. Phys. **150**, 074501 (2019).
[45] C. Desgranges and J. Delhommelle, *Can Ordered Precursors Promote the Nucleation of Solid Solutions?*, Phys. Rev. Lett. **123**, 195701 (2019).
[46] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, *Bond-Orientational Order in Liquids and Glasses*, Phys. Rev. B **28**, 784 (1983).
[47] 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).
[48] J. Taffs, S. R. Williams, H. Tanaka, and C. P. Royall, *Structure and Kinetics in the Freezing of Nearly Hard Spheres*, Soft Matter **9**, 297 (2013).
[49] H. Chan, M. J. Cherukara, B. Narayanan, T. D. Loeffler, C. Benmore, S. K. Gray, and S. K. R. S. Sankaranarayanan, *Machine Learning Coarse Grained Models for Water*, Nat. Commun. **10**, 379 (2019).
[50] H. Niu, Y. I. Yang, and M. Parrinello, *Temperature Dependence of Homogeneous Nucleation in Ice*, Phys. Rev. Lett. **122**, 245501 (2019).
[51] F. Martelli, N. Giovambattista, S. Torquato, and R. Car, *Searching for Crystal-Ice Domains in Amorphous Ices*, Phys. Rev. Mater. **2**, 075601 (2018).
[52] E. A. Lazar, J. Han, and D. J. Srolovitz, *Topological Framework for Local Structure Analysis in Condensed Matter*, Proc. Natl. Acad. Sci. U.S.A. **112**, E5769 (2015).
[53] A. V. Brukhno, J. Anwar, R. Davidchack, and R. Handel, *Challenges in Molecular Simulation of Homogeneous Ice Nucleation*, J. Phys. Condens. Matter **20**, 494243 (2008).
[54] A. L. Patterson, *Ambiguities in the X-Ray Analysis of Crystal Structures*, Phys. Rev. **65**, 195 (1944).
[55] G. A. Gallet and F. Pietrucci, *Structural Cluster Analysis of Chemical Reactions in Solution*, J. Chem. Phys. **139**, 074101 (2013).

[56] S. Pipolo, M. Salanne, G. Ferlat, S. Klotz, A. M. Saitta, and F. Pietrucci, *Navigating at Will on the Water Phase Diagram*, Phys. Rev. Lett. **119**, 245701 (2017).
[57] L. Zhang, J. Han, H. Wang, R. Car, and E. Weinan, *Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics*, Phys. Rev. Lett. **120**, 143001 (2018).
[58] J. Behler and M. Parrinello, *Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces*, Phys. Rev. Lett. **98**, 146401 (2007).
[59] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, *Machine Learning of Accurate Energy-Conserving Molecular Force Fields*, Sci. Adv. **3**, e1603015 (2017).
[60] P. Geiger and C. Dellago, *Neural Networks for Local Structure Detection in Polymorphic Systems*, J. Chem. Phys. **139**, 164105 (2013).
[61] V. Bapst, T. Keck, A. Grabska-Barwińska, C. Donner, E. D. Cubuk, S. S. Schoenholz, A. Obika, A. W. R. Nelson, T. Back, D. Hassabis, and P. Kohli, *Unveiling the Predictive Power of Static Structure in Glassy Systems*, Nat. Phys. **16**, 448 (2020).
[62] F. Martelli, F. Leoni, F. Sciortino, and J. Russo, *Connection between Liquid and Non-Crystalline Solid Phases in Water*, J. Chem. Phys. **153**, 104503 (2020).
[63] M. Spellings and S. C. Glotzer, *Machine Learning for Crystal Identification and Discovery*, Am. Instit. Chem. Eng. J. **64**, 2198 (2018).
[64] W. F. Reinhart, A. W. Long, M. P. Howard, A. L. Ferguson, and A. Z. Panagiotopoulos, *Machine Learning for Autonomous Crystal Structure Identification*, Soft Matter **13**, 4733 (2017).
[65] W. F. Reinhart and A. Z. Panagiotopoulos, *Multi-Atom Pattern Analysis for Binary Superlattices*, Soft Matter **13**, 6803 (2017).
[66] 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 (2018).
[67] E. Boattini, M. Dijkstra, and L. Filion, *Unsupervised Learning for Local Structure Detection in Colloidal Systems*, J. Chem. Phys. **151**, 154901 (2019).
[68] C. S. Adorf, T. C. Moore, Y. J. Melle, and S. C. Glotzer, *Analysis of Self-Assembly Pathways with Unsupervised Machine Learning Algorithms*, J. Phys. Chem. B **124**, 69 (2020).
[69] E. G. Noya, C. Vega, and E. de Miguel, *Determination of the Melting Point of Hard Spheres from Direct Coexistence Simulation Methods*, J. Chem. Phys. **128**, 154507 (2008).
[70] X. Glorot and Y. Bengio, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (PMLR, Sardinia, 2010), p. 249, http://proceedings.mlr.press/v9/glorot10a.html.
[71] Y. Zhang, A. M. Saxe, M. S. Advani, and A. A. Lee, *Energy-Entropy Competition and the Effectiveness of Stochastic Gradient Descent in Machine Learning*, Mol. Phys. **116**, 3214 (2018).
[72] C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford, 1995).
[73] 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 (1996).
[74] J. Hoshen and R. Kopelman, *Percolation and Cluster Distribution. I. Cluster Multiple Labeling Technique and Critical Concentration Algorithm*, Phys. Rev. B **14**, 3438 (1976).
[75] 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).
[76] F. Saija, S. Prestipino, and P. V. Giaquinta, *Scaling of Local Density Correlations in a Fluid Close to Freezing*, J. Chem. Phys. **115**, 7586 (2001).
[77] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids: With Applications to Soft Matter* (Academic Press, San Diego, 2013).
[78] M. N. Bannerman, R. Sargant, and L. Lue, *DYNAMO: A Free $\mathcal{O}(n)$ General Event-Driven Molecular Dynamics Simulator*, J. Comput. Chem. **32**, 3329 (2011).
[79] E. A. Engel, A. Anelli, M. Ceriotti, C. J. Pickard, and R. J. Needs, *Mapping Uncharted Territory in Ice from Zeolite Networks to Ice Structures*, Nat. Commun. **9**, 2173 (2018).
[80] 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 (2014).
[81] G. S. Bordonskiy and A. O. Orlov, *Signatures of the Appearance of Ice 0 in Wetted Nanoporous Media at Electromagnetic Measurements*, JETP Lett. **105**, 492 (2017).
[82] B. Slater and D. Quigley, *Crystal Nucleation: Zeroing In on Ice*, Nat. Mater. **13**, 670 (2014).
[83] A. Mujica, C. J. Pickard, and R. J. Needs, *Low-Energy Tetrahedral Polymorphs of Carbon, Silicon, and Germanium*, Phys. Rev. B **91**, 214104 (2015).
[84] S. Auer and D. Frenkel, *Prediction of Absolute Crystal-Nucleation Rate in Hard-Sphere Colloids*, Nature (London) **409**, 1020 (2001).
[85] A. Stukowski, *Visualization and Analysis of Atomistic Simulations Data with OVITO—The Open Visualization Tool*, Model. Simul. Mater. Sci. Eng. **18**, 015012 (2010).
[86] D. Richard and T. Speck, *Crystallization of Hard Spheres Revisited. II. Thermodynamic Modeling, Nucleation Work, and the Surface of Tension*, J. Chem. Phys. **148**, 224102 (2018).
[87] B. O'Malley and I. Snook, *Crystal Nucleation in the Hard Sphere System*, Phys. Rev. Lett. **90**, 085702 (2003).
[88] A. V. Anikeenko and N. N. Medvedev, *Polytetrahedral Nature of the Dense Disordered Packings of Hard Spheres*, Phys. Rev. Lett. **98**, 235504 (2007).
[89] U. Gasser, E. R. Weeks, A. Schofield, P. N. Pusey, and D. A. Weitz, *Real-Space Imaging of Nucleation and Growth in Colloidal Crystallization*, Science **292**, 258 (2001).
[90] J. Wedekind and D. Reguera, *Kinetic Reconstruction of the Free-Energy Landscape*, J. Phys. Chem. B **112**, 11060 (2008).
[91] J. Russo, A. C. Maggs, D. Bonn, and H. Tanaka, *The Interplay of Sedimentation and Crystallization in Hard-Sphere Suspensions*, Soft Matter **9**, 7369 (2013).

[92] T. Li, D. Donadio, G. Russo, and G. Galli, *Homogeneous Ice Nucleation from Supercooled Water*, Phys. Chem. Chem. Phys. **13**, 19807 (2011).

[93] D. Quigley and P. M. Rodger, *Metadynamics Simulations of Ice Nucleation and Growth*, J. Chem. Phys. **128**, 154518 (2008).

[94] A. Reinhardt, J. P. K. Doye, E. G. Noya, and C. Vega, *Local Order Parameters for Use in Driving Homogeneous Ice Nucleation with All-Atom Models of Water*, J. Chem. Phys. **137**, 194504 (2012).

[95] A. Reinhardt and J. P. K. Doye, *Note: Homogeneous TIP4P/2005 Ice Nucleation at Low Supercooling*, J. Chem. Phys. **139**, 096102 (2013).

[96] V. Bianco, P. M. de Hijes, C. P. Lamas, E. Sanz, and C. Vega, *Anomalous Behavior in the Nucleation of Ice at Negative Pressures*, Phys. Rev. Lett. **126**, 015704 (2021).

[97] P. G. Debenedetti, *Metastable Liquids: Concepts and Principles* (Princeton University Press, Princeton, NJ, 1996).

[98] K. Kelton and A. L. Greer, *Nucleation in Condensed Matter: Applications in Materials and Biology* (Elsevier, New York, 2010).

[99] http://www.bris.ac.uk/acrc/

[100] J. D. Honeycutt and H. C. Andersen, *Molecular Dynamics Study of Melting and Freezing of Small Lennard-Jones Clusters*, J. Phys. Chem. **91**, 4950 (1987).

[101] A. Stukowski, *Structure Identification Methods for Atomistic Simulations of Crystalline Materials*, Model. Simul. Mater. Sci. Eng. **20**, 045021 (2012).

[102] P. M. Larsen, S. Schmidt, and J. Schiøtz, *Robust Structural Identification via Polyhedral Template Matching*, Model. Simul. Mater. Sci. Eng. **24**, 055007 (2016).

[103] E. Maras, O. Trushin, A. Stukowski, T. Ala-Nissila, and H. Jónsson, *Global Transition Path Search for Dislocation Formation in Ge on Si(001)*, Comput. Phys. Commun. **205**, 13 (2016).

[104] A. H. Nguyen and V. Molinero, *Identification of Clathrate Hydrates, Hexagonal Ice, Cubic Ice, and Liquid Water in Simulations: The CHILL+ Algorithm*, J. Phys. Chem. B **119**, 9369 (2015).

[105] H. Tanaka, R. Shi, H. Tong, and J. Russo, *Revealing Key Structural Features Hidden in Liquids and Glasses*, Nat. Rev. Phys. **1**, 333 (2019).