Non-Markovian dynamics in ice nucleation

Pablo Montero de Hijes, Sebastian Falkner, Christoph Dellago

Non-Markovian dynamics in ice nucleation

Special Collection: Carlos Vega Festschrift

Pablo Montero de Hijes (iD) ; Sebastian Falkner (iD) ; Christoph Dellago ✉ (iD)

Check for updates

View Online

Export Citation

Articles You May Be Interested In

Efficient and exact sampling of transition path ensembles on Markovian networks

J. Chem. Phys. (July 2020)

Learning Markovian dynamics with spectral maps

J. Chem. Phys. (March 2024)

The fast committor machine: Interpretable prediction with kernels

J. Chem. Phys. (August 2024)

Non-Markovian dynamics in ice nucleation

View Online   Export Citation   CrossMark

Pablo Montero de Hijes,[1] 🆔 Sebastian Falkner,[2] 🆔 and Christoph Dellago[3,a] 🆔

AFFILIATIONS

[1] Faculty of Physics, University of Vienna, A-1090 Vienna, Austria

[2] Institute of Physics, University of Augsburg, 86159 Augsburg, Germany and Faculty of Physics, University of Vienna, A-1090 Vienna, Austria

[3] Faculty of Physics and Research Platform on Accelerating Photoreaction Discovery (ViRAPID), University of Vienna, A-1090 Vienna, Austria

Note: This paper is part of the Special Topic, Carlos Vega Festschrift.
[a] Author to whom correspondence should be addressed: christoph.dellago@univie.ac.at

ABSTRACT

In simulation studies of crystallization, the size of the largest crystalline nucleus is often used as a reaction coordinate to monitor the progress of the nucleation process. Here, we investigate, for the case of homogeneous ice nucleation, whether the nucleus size exhibits Markovian dynamics, as assumed in classical nucleation theory. Using 300 independent nucleation trajectories generated by molecular dynamics, we evaluate the mean recurrence time required to reach selected values of the largest nucleus size. Early recurrences consistently take longer than later ones, revealing a clear history dependence and thus non-Markovian dynamics. To identify the slow modes underlying this behavior, we analyze several structural descriptors of the nucleus, observing subtle but systematic differences between nuclei at early and late recurrences. By training a neural network on 2700 short trajectories to learn the committor, we identify relevant collective variables. Based on these features, symbolic regression provides a compact approximation of the committor, that is, an improved reaction coordinate, which we subsequently test for Markovian dynamics.

I. INTRODUCTION

Homogeneous ice nucleation is a fundamental process[1,2] whose microscopic mechanism remains only partially understood and continues to attract significant attention.[3–5] Classical nucleation theory (CNT)[6,7] is often used to describe this phenomenon, sometimes as part of computational approaches such as seeding.[8–10] According to this theory, crystallization proceeds via the formation of an ice nucleus in the supercooled liquid. A nucleation barrier hinders growth, and only nuclei that, by a rare fluctuation, reach the critical size corresponding to this barrier can grow further and complete crystallization. In its standard formulation, CNT relies on the Markovian dynamics of nucleus size;[11] however, generalizations that include memory effects have been derived.[12] In addition, CNT relies on several simplifying assumptions, such as the sphericity of the nucleus and the presence of a single crystalline phase. Moreover, CNT typically neglects subtleties in the thermodynamics of curved interfaces.[13–15] In crystalline solids, additional complications arise from strained states, defects, composition, or faceting. Although

it has been suggested that such complexities may be subsumed into suitably chosen bulk reference states,[16–19] their impact on the Markovian behavior of the process is unclear.

The assumptions made in CNT restrict the description of the nucleation process to a coarse-grained coordinate that omits many relevant slow degrees of freedom. Particularly in ice nucleation, assuming a single crystalline phase neglects the possible stacking of hexagonal (Ih) and cubic (Ic) ice[20–26] or the fivefold twin boundary defect identified in Ref. 27. In Refs. 28 and 29, structural and dynamical heterogeneities of supercooled water were identified as key factors influencing the ice nucleation process. Beyond ice, studies on colloids have shown that prestructured surface particles play an important role in nucleation—an effect that is not captured by the nucleus n alone.[30] For the Lennard-Jones (LJ) system, the dynamics of the largest nucleus size is manifestly non-Markovian,[12,31] and Moroni *et al.* demonstrated that crystallinity, shape, and structural composition (BCC/HCP/FCC) contribute importantly to the reaction coordinate.[32] Together, these studies show that crystallization cannot, in general, be reduced to one-dimensional kinetics along

18 March 2026 18:12:25

the largest cluster size alone and that memory effects arise naturally when essential slow variables are omitted. While mean first-passage-time approaches allow steady-state nucleation rates to be extracted without assuming Markovian dynamics along a chosen collective variable,[33,34] they do not address the origin or consequences of such memory effects. This type of behavior is not unique to crystallization; slow and history-dependent dynamics have also been observed in a variety of other nucleation processes.[35–37]

Here, using unbiased nucleation trajectories of the mW model obtained by straightforward molecular dynamics, we show that the dynamics projected onto the largest nucleus size n exhibit clear signatures of non-Markovian behavior. In particular, we show that the average recurrence time needed to reach a certain nucleus size strongly depends on the number of times that the particular nucleus size has been reached before, which is a clear memory effect. This strategy has been used previously to reveal pronounced memory effects in the nucleation dynamics of the LJ system.[31] We then analyze a broad set of structural descriptors and use a machine-learned committor to identify additional collective variables relevant to the reaction coordinate. Interestingly, an analysis of the dynamics of the approximated committor indicates that it displays signatures of non-Markovian behavior. The committor is considered to be the optimal reaction coordinate;[38,39] hence, our results suggest that even further features should be considered.

The remainder of the article is organized as follows. In Sec. II, we describe the simulation setup, the structural analysis of the crystalline nuclei, and the machine-learning workflow used to construct an approximate committor. In Sec. III, we first quantify the non-Markovian dynamics of the largest nucleus size (Sec. III A), then analyze structural features associated with this behavior (Sec. III B), introduce a machine-learned reaction coordinate based on the committor (Sec. III C), and finally compare the Markovian behavior of the largest nucleus size with that of the committor-based coordinate (Sec. III D). In Sec. IV, we summarize our main findings and discuss their implications for constructing improved reaction coordinates for ice nucleation and related rare-event processes.

II. METHODS

Molecular dynamics simulations of the coarse-grained mW water model[40] were carried out with the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS)[41] using the parametrization of Ref. 42. Each system contained $N = 12\,800$ molecules in a periodic cubic box. Simulations were performed at $T = 222$ K and $p = 1$ bar using a Nosé–Hoover thermostat and barostat. Under these conditions, spontaneous nucleation occurs on accessible timescales of nanoseconds, ensuring that no bias is required and allowing a clean assessment of the memory effects of the largest nucleus size.

We employed the CHILL+ algorithm[43] via OVITO[44] to classify each molecule as liquid, cubic ice (Ic), hexagonal ice (Ih), or others. Clusters of ice-like molecules (Ic plus Ih) were identified using a 3.5 Å cutoff distance. The number of molecules in the largest cluster ($n = n_{Ic} + n_{Ih}$) was then recorded along 300 independent nucleation trajectories. Figure 1 shows a snapshot of a nucleus that spontaneously formed. For each trajectory, the times t_n at which n crossed selected threshold values from below were determined. A recurrence time, τ_n, was defined as $t_n(i) - t_n(i - 1)$, where i indexes
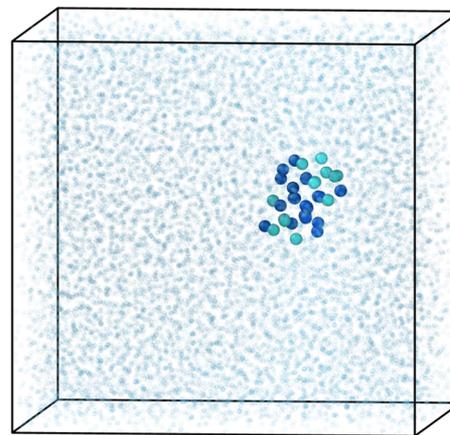

FIG. 1. Snapshot of a crystalline nucleus with Ih-like and Ic-like molecules shown in blue and cyan, respectively, embedded in the supercooled liquid, with molecules depicted partially transparently.

successive upward crossings of n. Figure 2(a) shows $n(t)$ for two example trajectories, including an inset clarifying how recurrence times are obtained. Recurrence times were grouped in blocks of four crossings, and the three recurrence intervals within each block were averaged. The block-averaged recurrence times were then averaged over all trajectories. The mean first-passage time to reach a certain n is shown in Fig. 2(b).

Structural features for the molecules of the largest nucleus were obtained with OVITO[44] and pyscal.[45] OVITO was used to compute the Ic/Ih composition, the radius of gyration R_g, the components of the gyration tensor (used to determine shape metrics such as the asphericity $\mathcal{K}$), and the mean and standard deviation of the Voronoi molecular volume $\mathcal{V}$ and coordination number $\mathcal{C}$. Pyscal was used to compute the tetrahedrality parameter of Uttormark et al.,[46] θ, as well as the Lechner–Dellago order parameters $\bar{q}_4$ and $\bar{q}_6$.[47] Molecular features were averaged among the nucleus molecules. Several additional features were derived from these quantities to facilitate the training of the neural network.

The training of the committor neural network follows the workflow described by Jung et al.[48] We employ a multilayer perceptron (see supplementary material for network architecture) and train on $N = 2700$ fleeting trajectories, each labeled $y_i = -1$ if state A (liquid, $n < 10$) or $y_i = 1$ if state B (solid, $n > 60$) was reached. The initial configurations for these trajectories are obtained by uniformly sampling configurations where n is between 10 and 60 molecules from the original 300 trajectories and then reinitializing the velocities. We define the predicted committor as $p_B(\mathbf{x}, \mathbf{w}) = 1/(1 + \exp(-q(\mathbf{x}, \mathbf{w})))$ with $\mathbf{x}$ being the normalized input collective variables and $\mathbf{w}$ being the network weights. The loss function for training the network is accordingly $L(\mathbf{w}|\theta) = \sum_{i=0}^{N} \log[1 + \exp(y_i q(\mathbf{x_i}, \mathbf{w}))]$. We train a committee of 10 neural networks with a batch size of 16 samples each for 100 epochs using a cosine annealing learning rate schedule between 10^{-2} and 10^{-6}. We estimate attribution values as in Jung et al.[48] by permutation of the input features, followed by symbolic regression using dCGP,[49] keeping the four most important features. In contrast to previous studies, we
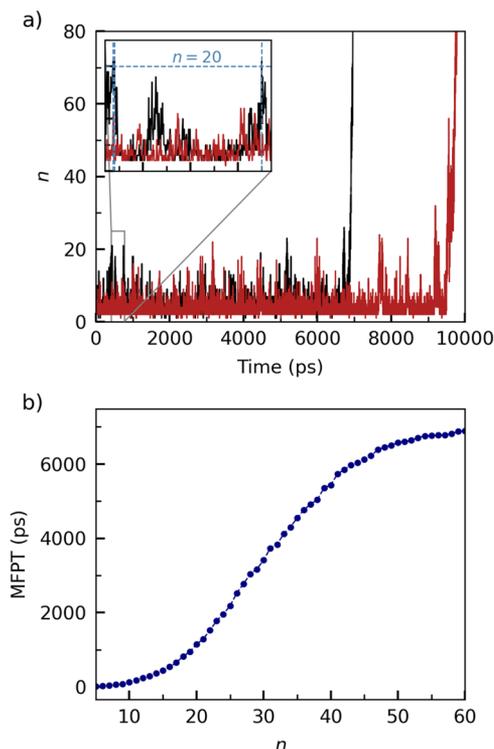
FIG. 2. (a) Time evolution of the largest cluster size n for two example trajectories, shown in black and red. The inset zooms in on one part of the trajectory where, if we define the threshold to be $n = 20$ (horizontal line), we find three crossing times t'_{20} (vertical lines), hence defining two recurrence times (the time difference between the second and first, and the third and second crossings). The second upward crossing occurs shortly after the first, causing the corresponding vertical lines to appear merged into a single thicker line. (b) Mean first-passage time (MFPT) with respect to n.

perform the regression with respect to the network output instead of the shooting data, as our initial dataset includes only one shot per point. The regression is run 8 times for 300 generations, allowing for summation, subtraction, multiplication, division, square, and cube operators. We then select the best-fitting expression that reasonably extrapolates for values outside of the training range (e.g., large cluster sizes should yield high committor probabilities).

III. RESULTS

A. Non-Markovian dynamics of the nucleus size

A Markovian description of nucleation, as postulated in CNT, assumes that the future evolution of the largest cluster size n depends only on its current value. To test this assumption, we analyze the statistics of upward crossings of selected thresholds of n. Figure 3 shows the mean recurrence times $\langle \tau_n \rangle$ for blocks of four consecutive crossings. If the dynamics projected onto n were Markovian, all blocks at a given n would yield identical recurrence times. Instead, the first block exhibits systematically longer recurrence times—up to nearly twice those of later blocks—demonstrating that trajectories reaching the same n retain memory of how they reached
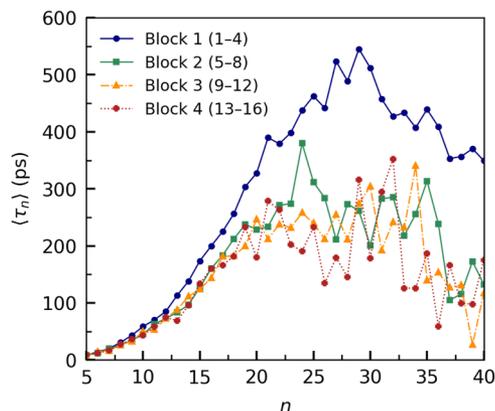

FIG. 3. Mean recurrence times $\langle \tau_n \rangle$ as a function of the largest nucleus size threshold, averaged over blocks of four consecutive crossings. The first block (solid line + circles in blue) reaches recurrence times that are about twice as long, revealing non-Markovian behavior. The non-Markovian behavior originates mostly in the first crossing of the first block, and the effect is larger around the critical cluster size.

that value. This block dependence constitutes a clear signature of non-Markovian behavior.

Further evidence for non-Markovian dynamics is provided in Fig. 4, which reports the average minimum and maximum values of the nucleus size explored between consecutive crossings of n. The first block clearly explores systematically smaller nuclei, both in their minima and maxima, compared to later blocks. The strong dependence of these excursion ranges on the crossing index indicates that the system retains some memory of previous states, at least in the initial stage of nucleation.

Together, these observations demonstrate unambiguously that the dynamics of the largest nucleus size, n, is intrinsically non-Markovian. Consequently, n alone cannot serve as a complete reaction coordinate for nucleation. Naturally, some arbitrariness
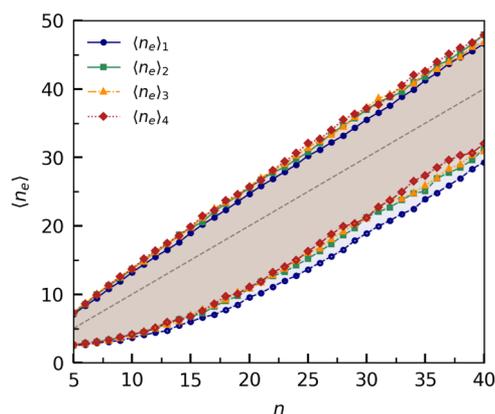
FIG. 4. Average maximum (solid) and minimum (dashed) values of n visited during excursions between threshold crossings per block, $\langle n_e \rangle$. The first block, $\langle n_e \rangle_1$, shows both maximum and minimum values that are systematically lower than those of later blocks, highlighting the non-Markovian dynamics of n.

is inherent in the definition of n, and one could argue that our specific criterion for identifying the largest nucleus might itself be responsible for the observed non-Markovian behavior—that is, a different definition of n might yield a coordinate with more Markovian dynamics. We acknowledge this possibility; however, we emphasize that our definition follows standard practices in the ice-nucleation community.[43,50–57] In studies where cubic ice is not considered, the largest cluster is identified using criteria based solely on $\bar{q}_6$,[10,58,59] and alternative definitions of n have also been used.[24,28] A systematic comparison of the dynamics of n for different definitions is an interesting subject for future work.

B. Structural features

What the largest nucleus size lacks to be Markovian may lie in subtle structural relaxations. To probe this hypothesis, we computed a series of structural descriptors for the molecules belonging to the largest ice nucleus at every time step of each trajectory. These descriptors were chosen to capture local structural motifs that may be overlooked when using only the largest nucleus size.

To account for the crystalline composition of the nucleus, we computed the ratio $n_{\mathrm{Ih(c)}}/n$, where $n_{\mathrm{Ih(c)}}$ is the number of hexagonal(cubic) ice-like molecules within the largest nucleus. To characterize tetrahedral arrangements, we evaluated the Uttormark tetrahedral order parameter θ for each molecule of the largest nucleus and averaged it over them, obtaining Θ. In this convention, lower values of θ (and hence of Θ) correspond to more tetrahedral local environments. To quantify local crystallinity, we computed the local order parameters $\bar{q}_4$ and $\bar{q}_6$ and averaged them over the molecules in the largest nucleus, yielding $\bar{Q}_4$ and $\bar{Q}_6$, respectively.

To characterize the overall geometry of the nucleus, we first computed its radius of gyration R_g. We then assessed the shape of the nucleus using the eigenvalues of the gyration tensor, denoted by $\lambda_1 \leq \lambda_2 \leq \lambda_3$. From these, we obtained the anisotropy $\mathcal{I}$,

$$\mathcal{I} = \frac{3}{2}\left(\frac{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}{(\lambda_1 + \lambda_2 + \lambda_3)^2} - \frac{1}{3}\right), \tag{1}$$

the acylindricity $\mathcal{J}$,

$$\mathcal{J} = \lambda_2 - \lambda_1, \tag{2}$$

and the asphericity $\mathcal{K}$,

$$\mathcal{K} = \lambda_3 - \frac{1}{2}(\lambda_1 + \lambda_2). \tag{3}$$

We also attempted to detect additional information on density inhomogeneities within the nucleus through the mean and standard deviation of the Voronoi molecular volumes, $\mathcal{V}_m$ and $\mathcal{V}_\sigma$, respectively. Analogously, for the Voronoi-based coordination number of the molecules in the nucleus, we defined $\mathcal{C}_m$ and $\mathcal{C}_\sigma$ as the mean and standard deviation of the coordination number distribution, respectively.

Finally, we combined these primary descriptors into a set of derived features that emphasize different physical aspects of the nucleus and might ease the training of the network. Specifically, we defined

$$\mathcal{E}_1 = n \cdot \bar{Q}_6, \tag{4}$$

$$\mathcal{E}_2 = \frac{\mathcal{K}}{R_g^2}, \tag{5}$$

$$\mathcal{E}_3 = \frac{\mathcal{V}_\sigma}{\mathcal{V}_m}, \tag{6}$$

$$\mathcal{E}_4 = \frac{\mathcal{C}_\sigma}{\mathcal{C}_m}, \tag{7}$$

$$\mathcal{E}_5 = \frac{1 + \mathcal{E}_3 + \mathcal{E}_4 - \bar{Q}_6}{3}, \tag{8}$$

$$\mathcal{E}_6 = \frac{n}{R_g^3}. \tag{9}$$

Here, $\mathcal{E}_1$ measures an "ordered" nucleus size, $\mathcal{E}_2$ combines shape anisotropy with nucleus compactness, $\mathcal{E}_3$ and $\mathcal{E}_4$ quantify relative fluctuations in Voronoi volume and coordination, $\mathcal{E}_5$ mixes density heterogeneity and local order into a single descriptor, and $\mathcal{E}_6$ is a measure of the effective number density of the nucleus.

Similar to the procedure used to compute $\langle \tau_n \rangle$, we evaluate the average value of each structural descriptor within each occurrence block for a given threshold n. This allows us to determine whether a particular structural feature changes between early and late occurrences of the same largest nucleus size. From this analysis, we see that several descriptors do not differ among blocks, but $\bar{Q}_6$, $\bar{Q}_4$, Θ, and $\mathcal{C}_m$ exhibit small but systematic differences between occurrence blocks, as shown in Fig. 5. Early occurrences of a given nucleus size tend to display slightly lower orientational order and reduced tetrahedral arrangement compared with later ones. We recall that a higher Θ means less tetrahedral arrangement. The coordination also increases modestly across blocks for the larger nuclei. These trends may suggest that the internal structure of clusters of a given size evolves during the trajectory and retains memory of the history of the system, consistent with the recurrence-time analysis.

It is worth stressing, however, that the interpretation of individual descriptors in this block analysis must be approached with caution. The differences observed between blocks do not, by themselves, determine how a given descriptor affects the nucleation probability, because its influence may be outweighed or even counterbalanced by other, more relevant variables. In particular, nuclei appearing in the first block could, in principle, carry a structural signature that would favor nucleation, yet at the same time exhibit other characteristics that have a much stronger and opposite impact on the nucleation probability.

Although the variations in any single feature are small, taken together, they suggest that nucleation proceeds along additional slow directions of configuration space that are not captured by the largest nucleus size alone. Identifying whether any of these descriptors plays a relevant role requires further analysis. In Subsection III C, we address this by performing a committor analysis based on extensive sets of short trajectories, followed by machine-learning classification and symbolic regression to determine which structural features are essential components of the reaction coordinate.

C. Machine-learned reaction coordinate

The previous analyses have shown that the largest nucleus size alone is not sufficient to predict the future behavior of the system.

<div style="text-align: right">18 March 2026 18:12:25</div>
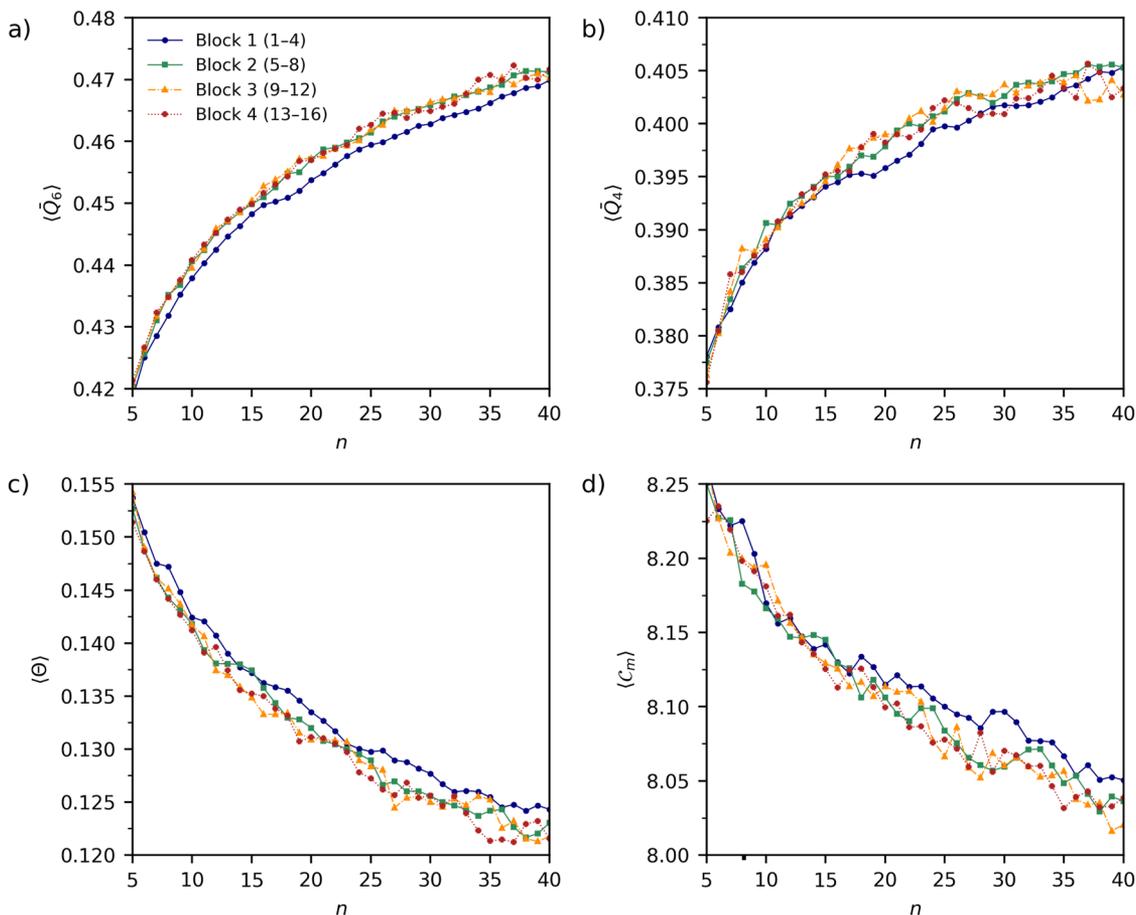
FIG. 5. (a) $\langle \bar{Q}_6 \rangle$, (b) $\langle \bar{Q}_4 \rangle$, (c) $\langle \Theta \rangle$, and (d) $\langle \mathcal{C}_m \rangle$ for the largest nucleus, averaged per occurrence block for each threshold value. The legend in panel (a) applies to all panels.

Therefore, we aim to find a better reaction coordinate to describe nucleation that minimizes non-Markovian behavior. In general, the ideal reaction coordinate for a rare event such as nucleation is the committor $p_B(\mathbf{x})$,[38,39] which describes the probability that a trajectory started from a certain configuration $\mathbf{x}$ transitions to the ice phase. While the ground-truth committor as a function of all Cartesian coordinates is analytically intractable and can only be obtained by resource-intensive simulations, Jung *et al.*[48] have shown that a neural network can approximate the committor as a function of a set of relevant collective variables. Leveraging these findings, we train a neural network on 2700 fleeting trajectories to predict the committor for ice nucleation to find additional important degrees of freedom, together with their contribution to the nucleation probability.

We first perform a correlation analysis on our broad set of collective variables by calculating their cross-correlations [Fig. 6(a)]. Based on these results, we removed variables that strongly correlated with the nucleus size, namely $\mathcal{E}_1$, n_{Ic}, and R_g, as they interfere with the later attribution analysis. After training the neural network, we see that it can successfully distinguish configurations from which the system tends to transition to ice or fall back to the liquid

state [Fig. 6(b)]. Each input feature of the network is then assigned an attribution corresponding to its importance to the predicted committor [Fig. 6(c)]. We observe that, although the nucleus size remains the most important parameter for predicting the committor probability by a clear margin, multiple other collective variables contribute to the network prediction.

Based on these findings, we perform symbolic regression to find a human-readable expression for the nucleation committor probability. We include as inputs the four most important features: the nucleus size n, the order parameter averaged over the nucleus molecules $\bar{Q}_6$, the number of hexagonal ice molecules n_{Ih}, and the tetrahedrality parameter averaged over the nucleus molecules Θ. Using these features, the best-fitting expression that also extrapolates well to larger nuclei is

$$p_B(\tilde{n}, \tilde{Q}_6, \tilde{\Theta}) = \text{sig}\left(0.216\tilde{n}^3 + 1.421\tilde{n} + 0.518\,\tilde{Q}_6 + 0.193\,\tilde{\Theta} + 0.803\right), \tag{10}$$

where $\text{sig}(x) = 1/[1 + \exp(-x)]$, and $\tilde{n}$, $\tilde{Q}_6$, and $\tilde{\Theta}$ are the normalized input features:

18 March 2026 18:12:25
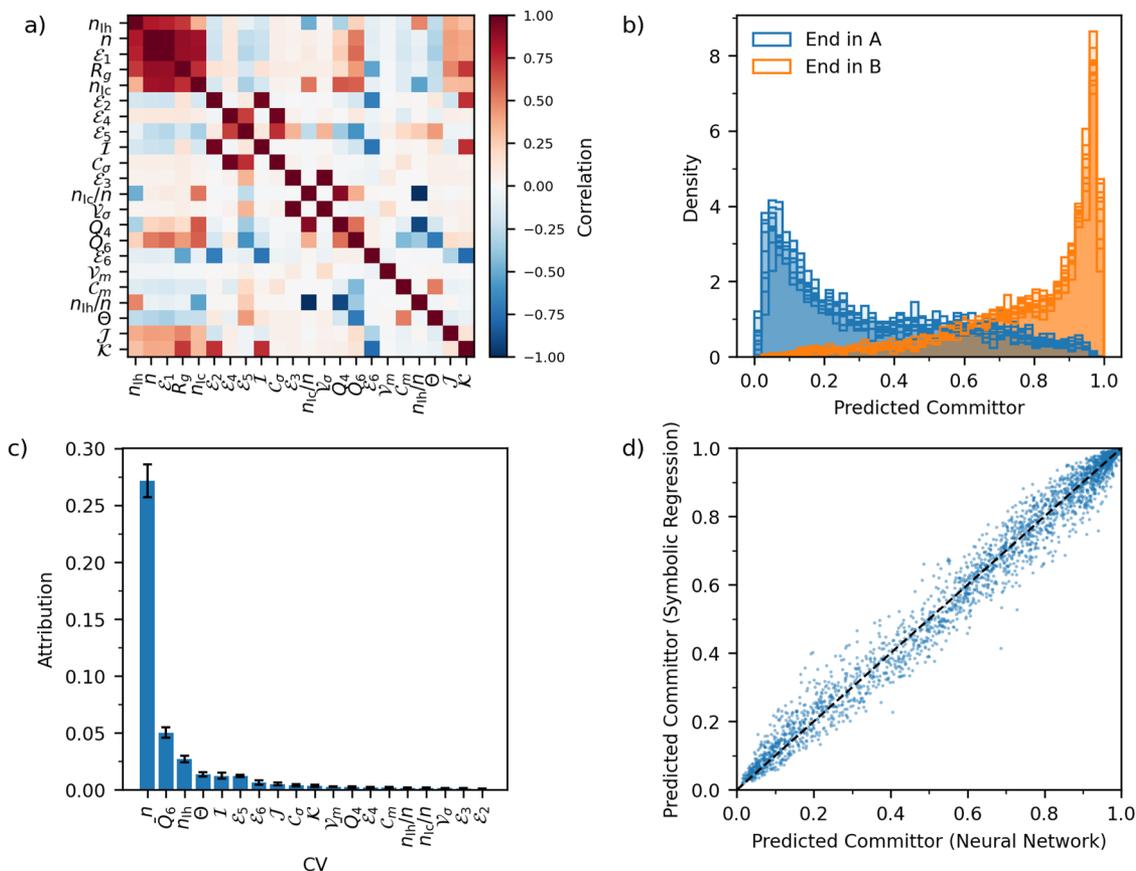
FIG. 6. (a) Correlation matrix for a selection of collective variables that describe the nucleus geometry. (b) Distribution of network-predicted committor values for trajectories that ended in the liquid (State A) or solid state (State B). The histograms for the ten neural networks in the committee are depicted on top of each other. (c) Network attribution values for each input collective variable obtained by permuting each feature separately and observing its impact on the loss function. Higher values indicate important features for the prediction of the committor. (d) Comparison of the neural network predicted committor with the committor obtained from symbolic regression.

$$\tilde{n} = \frac{n - 34.624}{14.722}, \tag{11}$$

$$\tilde{Q}_6 = \frac{\bar{Q}_6 - 0.467}{0.018}, \tag{12}$$

$$\tilde{\Theta} = \frac{\Theta - 0.125}{0.021}. \tag{13}$$

Notably, n_{Ih}, which has the third-highest network attribution, is not present in the fitted expression, most likely due to its correlation with the largest nucleus size. Still, the committor obtained via the simple analytical expression matches the network-predicted committor well [Fig. 6(d)].

Looking at the committor landscape as a function of these three parameters (Fig. 7), we see that an increasing nucleus size is still the best predictor for a successful phase transition. However, especially for critical nuclei with a committor of ~0.5, we also observe that the crystallinity of the nucleus, captured by the $\bar{Q}_6$ parameter, is essential for predicting the committor of the system. This is apparent
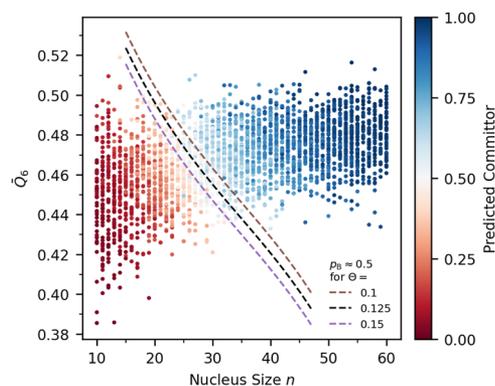
FIG. 7. Committor landscape as a function of the nucleus size n, the nucleus-averaged Lechner–Dellago order parameter $\bar{Q}_6$, and the nucleus-averaged tetrahedrality parameter Θ. Each point represents the initial point of a fleeting trajectory in the dataset and is colored according to the network-predicted committor. The dashed lines represent the $p_B \approx 0.5$ isocommittor lines for different tetrahedralities Θ.

in Fig. 7, where the $p_B \approx 0.5$ isocommittor line is tilted, indicating that for the same nucleus size (e.g., 27 molecules), the $\bar{Q}_6$ parameter can determine whether the nucleus tends to grow or vanish. In addition, the distilled committor expression indicates that a high Θ, corresponding to low tetrahedrality, generally increases the commitment probability to the ice state. Interestingly, the block analysis indicated that early crossings with increased recurrence times are correlated with low tetrahedrality. However, at the same time, these early crossings are associated with low $\bar{Q}_6$, which possibly negates the nucleation-promoting effect of low tetrahedrality.

D. Comparing memory effects between n and p_B

We further illustrate the non-Markovian behavior of the largest nucleus size n by examining the block-averaged values of p_B at upward crossings of n. If n were a Markovian reaction coordinate, all crossings of a given n would have the same average committor. Instead, as shown in Fig. 8(a), configurations belonging to the first occurrence block systematically exhibit lower values of p_B than those from later blocks. In other words, nuclei with the same size n are less likely to keep growing when they are encountered early in the trajectory. This dependence of the growth probability on the trajectory history confirms, once again, that n alone does not provide a Markovian description of the dynamics.

Having further established the non-Markovian dynamics of n, we now examine whether the approximate committor p_B offers a more Markovian representation. Since p_B incorporates additional structural information and correlates strongly with the probability that a configuration will proceed to the ice basin, one may expect it to reduce the memory effects observed when projecting the dynamics onto n. Figure 8(b) reports the mean recurrence time $\langle \tau_{p_B} \rangle$ required to return to a given threshold of p_B. As in the case of n, the first block of crossings (1–4) displays longer recurrence times than later blocks, indicating that memory effects persist. This is not surprising; the additional structural descriptors used to construct p_B are unlikely to span all remaining slow degrees of freedom, and therefore, even an improved reaction coordinate may not be perfectly Markovian.

The comparison is made explicit in Fig. 8(c), which shows the block-to-block difference $\Delta \langle \tau_{RC} \rangle_{1-2} = \langle \tau \rangle_{1-4} - \langle \tau \rangle_{5-8}$ for the normalized largest nucleus size and p_B from both the neural network and the symbolic regression. For n, we use the normalization $(n - n_A)/(n_B - n_A)$. Across most of the reaction pathway, $\Delta \langle \tau_{RC} \rangle_{1-2}$ is only slightly smaller for the p_B estimations than for n, confirming that p_B exhibits modestly improved Markovian behavior. The fact that the difference remains nonzero over the entire range shows that the committor-based mapping does not eliminate memory effects.

Overall, these results demonstrate that p_B provides a modest improvement over n in terms of Markovian behavior, yet it still does not constitute a fully memoryless reaction coordinate. Additional slow degrees of freedom—not captured by either n or the structural descriptors used to build p_B—continue to influence the dynamics. Nevertheless, the workflow established here provides a robust foundation for future efforts aimed at refining reaction coordinates for ice nucleation by explicitly assessing the memory effects of their projected dynamics. Beyond testing additional structural features, a possible direction is to incorporate descriptors of the surrounding
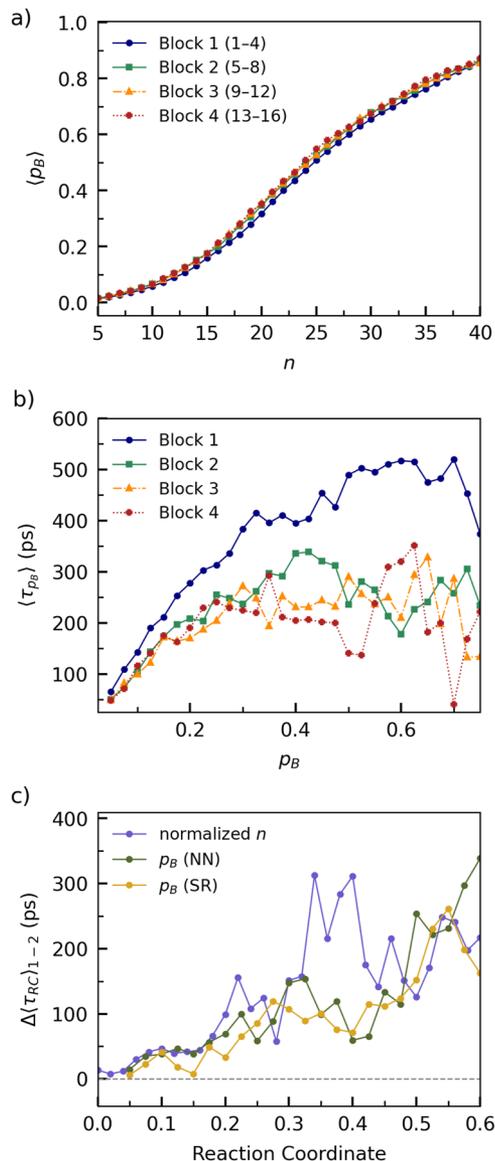
FIG. 8. (a) Recurrence-block-averaged committor $\langle p_B \rangle$ as a function of the largest nucleus size threshold, evaluated over successive blocks of four upward crossings. (b) Mean recurrence time $\langle \tau_{p_B} \rangle$ return to a given value of p_B, again averaged over blocks of four crossings. (c) Difference between the first and second blocks of recurrence times, $\Delta \langle \tau_{RC} \rangle_{1-2}$, for three reaction coordinates (RC): the normalized cluster size $(n - n_A)/(n_B - n_A)$ and both p_B from the neural network (NN) and symbolic regression (SR). The $\Delta \langle \tau_{RC} \rangle_{1-2}$ of p_B from symbolic regression agrees well with that from the neural network.

supercooled liquid, rather than focusing exclusively on the nucleus, together with molecular momenta as a proxy for local temperature. Such an approach may help capture liquid-state heterogeneities[28,29] and non-isothermal effects associated with latent heat,[37,60,61] both of which have been suggested to influence nucleation dynamics.

IV. CONCLUSIONS

In this work, we have examined whether the size of the largest ice nucleus, n, provides a Markovian description of homogeneous ice nucleation in supercooled water. Using 300 unbiased trajectories for mW water, we showed that the dynamics projected onto n exhibit clear and systematic history dependence: early crossings of a given n require significantly longer recurrence times than later ones. This behavior provides direct evidence that the nucleus size n alone is not a Markovian reaction coordinate for ice nucleation. Further support comes from analyzing the average minimum and maximum values of n explored between consecutive upward crossings; during the early occurrences, systematically smaller nuclei are visited, again reflecting history dependence.

We probed the origin of these memory effects by analyzing a broad set of structural descriptors of the nucleus. Although the differences are modest in magnitude, several features—including local crystalline order, tetrahedrality, and coordination—display systematic trends between early and late occurrences, indicating that structural relaxation proceeds along additional slow degrees of freedom that are not captured by the cluster size alone.

To identify which of these additional variables are most relevant, we trained a neural network to learn the committor from 2700 short trajectories. An attribution analysis based on this neural network showed that, in addition to the nucleus size, structural descriptors such as $\bar{Q}_6$ and tetrahedrality play a significant role in determining the nucleation probability. Using this information, symbolic regression yielded a compact, approximate analytical expression for the committor p_B, providing an interpretable and testable improved reaction coordinate.

Comparing the memory effects of n and p_B shows that the committor-based coordinate is modestly more Markovian, although memory effects persist. The residual memory suggests that the network still misses additional slow modes that influence the dynamics. The reason for this could simply be insufficient training data or possibly an incomplete set of relevant input features. For instance, parameters that characterize the surrounding liquid phase, which are neglected here, could play a role in the nucleation process.

Overall, our study establishes a quantitative framework for assessing the Markovian behavior of proposed reaction coordinates for ice nucleation. The workflow introduced here, which can also be applied to trajectories obtained with Transition Path Sampling[62] or other trajectory-based rare-event methods, provides the basis for constructing improved low-dimensional coordinates. The approach can easily be extended to more realistic water models with richer structural descriptions, as well as to entirely different nucleation processes.

SUPPLEMENTARY MATERIAL

See the supplementary material for a summary of the structural features and the neural network architecture.

ACKNOWLEDGMENTS

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Pablo Montero de Hijes: Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Sebastian Falkner**: Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Christoph Dellago**: Conceptualization (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Supervision (equal); Validation (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] K. Kelton and A. L. Greer, *Nucleation in Condensed Matter: Applications in Materials and Biology* (Elsevier, 2010), Vol. 15.

[2] P. G. Debenedetti, *Metastable Liquids: Concepts and Principles* (Princeton University Press, 2020).

[3] P. Bieber, T. M. Seifried, W. Bae, A. K. Bertram, and N. Borduas-Dedekind, "Contact line ice nucleation is the dominant freezing mechanism for water on macro- and microscopic polypropylene surfaces," Langmuir **41**, 31312 (2025).

[4] I. Sanchez-Burgos, A. R. Tejedor, C. Vega, M. M. Conde, E. Sanz, J. Ramirez, and J. R. Espinosa, "Homogeneous ice nucleation rates for mW and TIP4P/ICE models through Lattice Mold calculations," J. Chem. Phys. **157**(9), 094503 (2022).

[5] D. A. Knopf and P. A. Alpert, "Atmospheric ice nucleation," Nat. Rev. Phys. **5**(4), 203–217 (2023).

[6] S. Auer and D. Frenkel, "Numerical prediction of absolute crystallization rates in hard-sphere colloids," J. Chem. Phys. **120**(6), 3015–3029 (2004).

[7] D. Kashchiev, *Nucleation* (Elsevier, 2000).

[8] E. Sanz, C. Vega, J. R. Espinosa, R. Caballero-Bernal, J. L. F. Abascal, and C. Valeriani, "Homogeneous ice nucleation at moderate supercooling from molecular simulation," J. Am. Chem. Soc. **135**(40), 15008–15017 (2013).

[9] 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**(1), 015704 (2021).

[10] P. M. Piaggi, J. Weis, A. Z. Panagiotopoulos, P. G. Debenedetti, and R. Car, "Homogeneous ice nucleation in an ab initio machine-learning model of water," Proc. Natl. Acad. Sci. U. S. A. **119**(33), e2207294119 (2022).

[11] K. E. Blow, D. Quigley, and G. C. Sosso, "The seven deadly sins: When computing crystal nucleation rates, the devil is in the details," J. Chem. Phys. **155**(4), 040901 (2021).

18 March 2026 18:12:25

[12]A. Kuhnhold, H. Meyer, G. Amati, P. Pelagejcev, and T. Schilling, "Derivation of an exact, nonequilibrium framework for nucleation: Nucleation is a priori neither diffusive nor Markovian," Phys. Rev. E **100**(5), 052140 (2019).

[13]S. Kondo, "Thermodynamical fundamental equation for spherical interface," J. Chem. Phys. **25**(4), 662–669 (1956).

[14]J. S. Rowlinson and B. Widom, *Molecular Theory of Capillarity* (Courier Corporation, 2013).

[15]A. Tröster, M. Oettel, B. Block, P. Virnau, and K. Binder, "Numerical approaches to determine the interface tension of curved interfaces from free energy calculations," J. Chem. Phys. **136**(6), 064709 (2012).

[16]W. W. Mullins, "Thermodynamic equilibrium of a crystalline sphere in a fluid," J. Chem. Phys. **81**(3), 1436–1442 (1984).

[17]P. Montero de Hijes, K. Shi, E. G. Noya, E. Santiso, K. Gubbins, E. Sanz, and C. Vega, "The Young–Laplace equation for a solid–liquid interface," J. Chem. Phys. **153**(19), 191102 (2020).

[18]M. de Jager, C. Vega, P. Montero de Hijes, F. Smallenburg, and L. Filion, "Statistical mechanics of crystal nuclei of hard spheres," J. Chem. Phys. **161**(18), 184501 (2024).

[19]P. Montero de Hijes, K. Shi, C. Vega, and C. Dellago, "Comparing the mechanical and thermodynamic definitions of pressure in ice nucleation," J. Phys. Chem. Lett. (Published Online 2026).

[20]A. Reinhardt and J. P. K. Doye, "Free energy landscapes for homogeneous nucleation of ice for a monatomic water model," J. Chem. Phys. **136**(5), 054501 (2012).

[21]K. Thürmer and S. Nie, "Formation of hexagonal and cubic ice during low-temperature growth," Proc. Natl. Acad. Sci. U. S. A. **110**(29), 11757–11762 (2013).

[22]E. B. Moore and V. Molinero, "Is it cubic? Ice crystallization from deeply supercooled water," Phys. Chem. Chem. Phys. **13**(44), 20008–20016 (2011).

[23]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 **551**(7679), 218–222 (2017).

[24]H. Niu, Y. I. Yang, and M. Parrinello, "Temperature dependence of homogeneous nucleation in ice," Phys. Rev. Lett. **122**(24), 245501 (2019).

[25]M. Li, J. Zhang, H. Niu, Y.-K. Lei, X. Han, L. Yang, Z. Ye, Y. I. Yang, and Y. Q. Gao, "Phase transition between crystalline variants of ordinary ice," J. Phys. Chem. Lett. **13**(36), 8601–8606 (2022).

[26]M. Lin, Z. Xiong, and H. Cao, "Bridging classical nucleation theory and molecular dynamics simulation for homogeneous ice nucleation," J. Chem. Phys. **161**(8), 084504 (2024).

[27]T. Li, D. Donadio, G. Russo, and G. Galli, "Homogeneous ice nucleation from supercooled water," Phys. Chem. Chem. Phys. **13**(44), 19807–19813 (2011).

[28]M. Chen, L. Tan, H. Wang, L. Zhang, and H. Niu, "Exploring homogeneous ice nucleation through molecular dynamics simulations with *ab initio* accuracy," Phys. Rev. B **111**(13), 134116 (2025).

[29]M. Fitzner, G. C. Sosso, S. J. Cox, and A. Michaelides, "Ice is born in low-mobility regions of supercooled liquid water," Proc. Natl. Acad. Sci. U. S. A. **116**(6), 2009–2014 (2019).

[30]W. Lechner, C. Dellago, and P. G. Bolhuis, "Role of the prestructured surface cloud in crystal nucleation," Phys. Rev. Lett. **106**(8), 085701 (2011).

[31]S. Jungblut and C. Dellago, "Caveats of mean first-passage time methods applied to the crystallization transition: Effects of non-Markovianity," J. Chem. Phys. **142**(6), 064103 (2015).

[32]D. Moroni, P. R. Ten Wolde, and P. G. Bolhuis, "Interplay between structure and size in a critical crystal nucleus," Phys. Rev. Lett. **94**(23), 235703 (2005).

[33]J. Wedekind, R. Strey, and D. Reguera, "New method to analyze simulations of activated processes," J. Chem. Phys. **126**(13), 134103 (2007).

[34]P. Reimann, G. J. Schmid, and P. Hänggi, "Universal equivalence of mean first-passage time and Kramers rate," Phys. Rev. E **60**(1), R1 (1999).

[35]J. Kuipers and G. T. Barkema, "Limitations of a Fokker-Planck description of nucleation," Phys. Rev. E **82**(1), 011128 (2010).

[36]S.-T. Tsai, Z. Smith, and P. Tiwary, "Reaction coordinates and rate constants for liquid droplet nucleation: Quantifying the interplay between driving force and memory," J. Chem. Phys. **151**(15), 154106 (2019).

[37]J. C. Barrett, "Energy dependent decay rates of Lennard-Jones clusters for use in nucleation theory," J. Chem. Phys. **126**(7), 074312 (2007).

[38]G. Hummer, "From transition paths to transition states and rate coefficients," J. Chem. Phys. **120**, 516–523 (2004).

[39]E. Weinan, W. Ren, and E. Vanden-Eijnden, "Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes," Chem. Phys. Lett. **413**(1–3), 242–247 (2005).

[40]V. Molinero and E. B. Moore, "Water modeled as an intermediate element between carbon and silicon," J. Phys. Chem. B **113**(13), 4008–4016 (2009).

[41]A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. In't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen *et al.*, "LAMMPS–A flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," Comput. Phys. Commun. **271**, 108171 (2022).

[42]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**(1), 379 (2019).

[43]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**(29), 9369–9376 (2015).

[44]A. Stukowski, "Visualization and analysis of atomistic simulation data with OVITO–The Open Visualization Tool," Modell. Simul. Mater. Sci. Eng. **18**(1), 015012 (2009).

[45]S. Menon, G. Leines, and J. Rogal, "pyscal: A python module for structural analysis of atomic environments," J. Open Source Software **4**(43), 1824 (2019).

[46]M. J. Uttormark, M. O. Thompson, and P. Clancy, "Kinetics of crystal dissolution for a Stillinger-Weber model of silicon," Phys. Rev. B **47**(23), 15717 (1993).

[47]W. Lechner and C. Dellago, "Accurate determination of crystal structures based on averaged local bond order parameters," J. Chem. Phys. **129**(11), 114707 (2008).

[48]H. Jung, R. Covino, A. Arjun, C. Leitold, C. Dellago, P. G. Bolhuis, and G. Hummer, "Machine-guided path sampling to discover mechanisms of molecular self-organization," Nat. Comput. Sci. **3**(4), 334–345 (2023).

[49]D. Izzo and F. Biscani, "dcgp: Differentiable Cartesian genetic programming made easy," J. Open Source Software **5**(51), 2290 (2020).

[50]X. Zhu, M. Vandamme, Z. Jiang, and L. Brochard, "Molecular simulation of the confined crystallization of ice in cement nanopore," J. Chem. Phys. **159**(15), 154704 (2023).

[51]Y. Luo, Z. Zhang, Y. Li, D. Chen, Y. Mu, and Z. Yin, "Molecular insights into the microscopic promotion mechanism of CO_2 hydrate formation by the natural amino acid l-methionine for hydrate-based CO_2 sequestration," Fuel **405**, 136636 (2026).

[52]H. Zhang, C. Li, J. Chen, J. Wu, X. Lan, Z. Liang, and B. Deng, "A simulation of methane hydrate formation in nickel chloride solution," Fuel **340**, 127564 (2023).

[53]R. O. David, C. Marcolli, J. Fahrni, Y. Qiu, Y. A. Perez Sirkin, V. Molinero, F. Mahrt, D. Brühwiler, U. Lohmann, and Z. A. Kanji, "Pore condensation and freezing is responsible for ice formation below water saturation for porous particles," Proc. Natl. Acad. Sci. U. S. A. **116**(17), 8184–8189 (2019).

[54]Y. Qiu, N. Odendahl, A. Hudait, R. Mason, A. K. Bertram, F. Paesani, P. J. DeMott, and V. Molinero, "Ice nucleation efficiency of hydroxylated organic surfaces is controlled by their structural fluctuations and mismatch to ice," J. Am. Chem. Soc. **139**(8), 3052–3064 (2017).

[55]G. Roudsari, B. Reischl, O. H. Pakarinen, and H. Vehkamäki, "Atomistic simulation of ice nucleation on silver iodide (0001) surfaces with defects," J. Phys. Chem. C **124**(1), 436–445 (2019).

[56]P. Chen, R. Pillai, and S. Datta, "Nanoscale insights into vibration-induced heterogeneous ice nucleation," Nanoscale **17**(23), 14172–14182 (2025).

[57]M. H. Factorovich, P. M. Naullage, and V. Molinero, "Can clathrates heterogeneously nucleate ice?," J. Chem. Phys. **151**(11), 114707 (2019).

[58]P. Montero de Hijes, J. R Espinosa, C. Vega, and C. Dellago, "Minimum in the pressure dependence of the interfacial free energy between ice Ih and water," J. Chem. Phys. **158**(12), 124503 (2023).

[59] A. Haji-Akbari and P. G. Debenedetti, "Direct calculation of ice homogeneous nucleation rate for a molecular model of water," Proc. Natl. Acad. Sci. U. S. A. **112**(34), 10582–10588 (2015).

[60] J. Feder, K. C. Russell, J. Lothe, and G. M. Pound, "Homogeneous nucleation and growth of droplets in vapours," Adv. Phys. **15**(57), 111–178 (1966).

[61] J. Wedekind, D. Reguera, and R. Strey, "Influence of thermostats and carrier gas on simulations of nucleation," J. Chem. Phys. **127**(6), 064501 (2007).

[62] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, "Transition path sampling and the calculation of rate constants," J. Chem. Phys. **108**, 1964–1977 (1998).

18 March 2026 18:12:25