THE JOURNAL OF PHYSICAL CHEMISTRY LETTERS
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCL

Letter

Divergence among Local Structure, Dynamics, and Nucleation Outcome in Heterogeneous Nucleation of Close-Packed Crystals

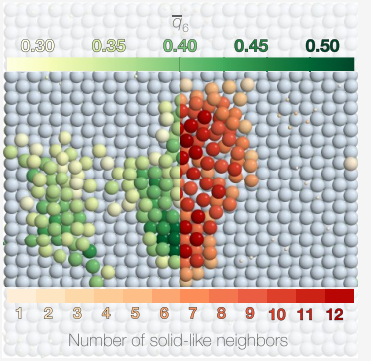
Tiago S. Domingues,[‡] Sarwar Hussain,[‡] and Amir Haji-Akbari*

Cite This: *J. Phys. Chem. Lett.* 2024, 15, 1279−1287

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Heterogeneous crystal nucleation is the dominant mechanism of crystallization in most systems, yet its underlying physics remains an enigma. While emergent interfacial crystalline order precedes heterogeneous nucleation, its importance in the nucleation mechanism is unclear. Here, we use path sampling simulations of two model systems to demonstrate that crystalline order in its traditional sense is not predictive of the outcome of the heterogeneous nucleation of close-packed crystals. Consequently, structure-based collective variables (CVs) that reliably describe homogeneous nucleation can be poor descriptors of heterogeneous nucleation. This divergence between structure and nucleation outcome is accompanied by an intriguing dynamical anomaly, wherein low-coordinated crystalline particles outpace their liquid-like counterparts. We use committor analysis, high-throughput screening, and machine learning to devise CV optimization strategies and present suitable structural heuristics within the metastable fluid for CV prescreening. Employing such optimized CVs is pivotal for properly characterizing the mechanism of heterogeneous nucleation in metallic and colloidal systems.



Crystallization is pivotal to many natural and industrial processes, ranging from microtubule assembly[1] and cloud microphysics,[2] to the production of semiconductors,[3] solar cells,[4] and pharmaceuticals.[5] Nonetheless, our understanding of its microscopic mechanism remains limited.[6] As a first-order transition, crystallization typically proceeds through nucleation and growth, with nucleation being the rate-limiting step under many scientifically and technologically important circumstances.[7] Unfortunately, most experimental techniques lack the spatiotemporal resolution necessary for probing the nucleation mechanism. Consequently, molecular simulations augmented with advanced sampling techniques have proven invaluable in probing nucleation in various single-component[8−14] and multicomponent[15−19] systems. Nucleation can occur either homogeneously or heterogeneously. Homogeneous nucleation is dominant at large thermodynamic driving forces as it involves crossing larger nucleation barriers but is mechanistically simpler, driven by intrinsic structural fluctuations in the fluid. In contrast, heterogeneous nucleation occurs in the presence of extrinsic impurities that facilitate freezing by decreasing the nucleation barriers. Its mechanism, however, is more complex, dominated by the interfacial properties of the metastable fluid. Developing a comprehensive framework to elucidate how such interfacial features impact the kinetics and mechanism of heterogeneous nucleation remains an ambitious pursuit.

The most vivid illustrations of this challenge are the Lennard-Jones (LJ) and hard sphere (HS) fluids, classic models for studying simple liquids. Both systems spontaneously assemble face-centered cubic (FCC) and hexagonally close-packed (HCP) crystals. While various aspects of both systems, including their homogeneous crystal nucleation,[20−26] have been extensively explored computationally, their heterogeneous nucleation remains surprisingly understudied, considered in only a few publications.[27−29] Here, we employ molecular dynamics (MD) simulations, jumpy forward flux sampling (jFFS),[30] and machine learning to unveil the complexities of heterogeneous crystal nucleation in these systems and to demonstrate that structure-based collective variables (CVs) that serve as reliable reaction coordinates (RCs) for homogeneous nucleation are inadequate for describing the progress of heterogeneous nucleation, even on simple surfaces. These transferability issues are exacerbated on more potent substrates and are accompanied by an intriguing dynamical anomaly, indicating a divergence between conventional structural order and dynamics and nucleation outcome. We use machine learning to systematically evaluate physics-based CVs and to tailor them specifically for the heterogeneous nucleation of close-packed crystals.

Within single-component systems, crystal nucleation is generally a single-step process, with the size of the largest crystalline nucleus often serving as the preferred RC. First, particles with local solid-like environments are identified on the

Received: December 20, 2023
Revised: January 5, 2024
Accepted: January 22, 2024

https://doi.org/10.1021/acs.jpclett.3c03561
J. Phys. Chem. Lett. 2024, 15, 1279−1287

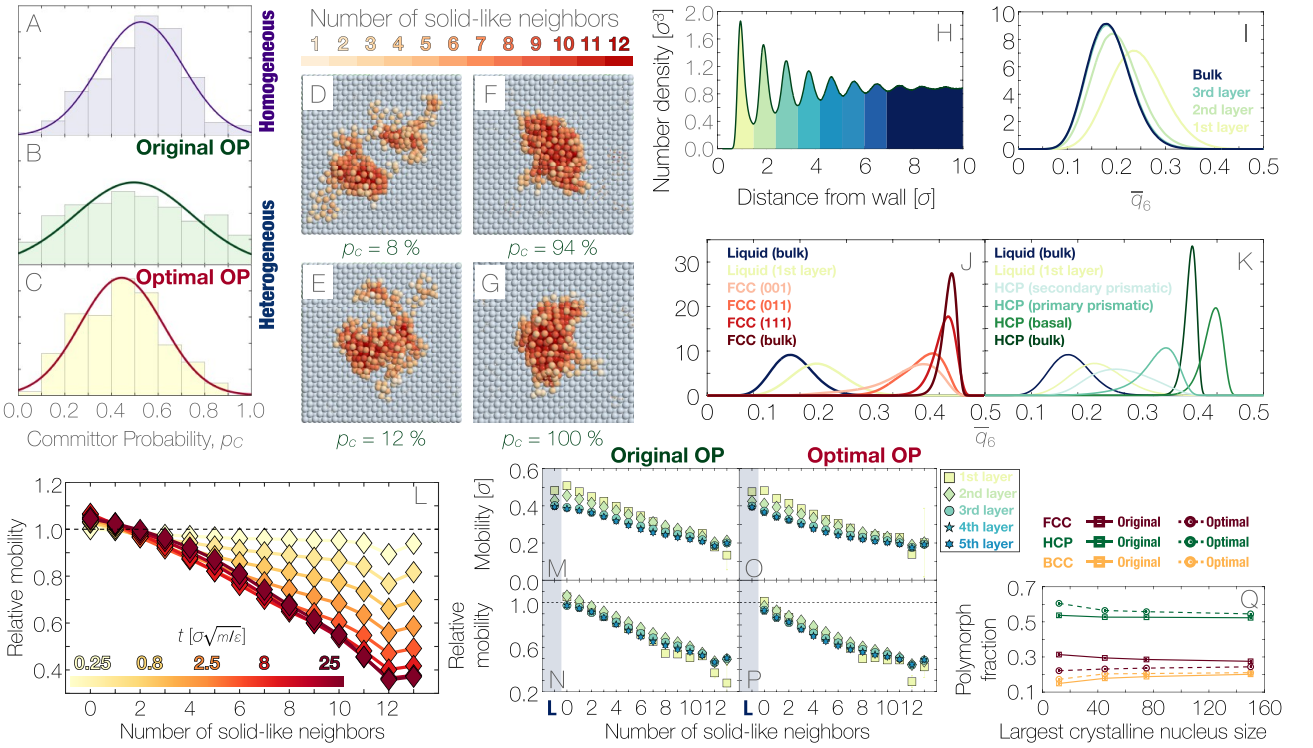


Figure 1. Committor probability distributions for (A) homogeneous and (B and C) heterogeneous nucleation in the LJ system using the (B) original and an (C) optimized CV (later shown to be optimal). The curves plotted in panels A−C are Gaussians, each with the same mean and standard deviation as the underlying histogram. Representative crystalline nuclei, showcasing (D and E) low and (F and G) high committor probabilities. The particles within the clusters are color-coded on the basis of the number of their solid-like neighbors. (H) Liquid density as a function of distance from the substrate. $\overline{q}_6$ profile of (I) the liquid across the bulk and the first few interfacial layers, as well as different crystallographic planes of (J) FCC and (K) HCP in the immediate vicinity of the substrate. Note that the secondary prismatic plane of HCP is not stable and undergoes partial melting. (L) Relative lateral mobility of solid-like particles (compared to their liquid-like counterparts) within the second liquid layer over various temporal windows. (M and O) Absolute and (N and P) relative mobilities of solid-like particles for the original (M and N) and optimal (O and P) CVs within the first five liquid layers over a temporal window $4\sigma\sqrt{m/\varepsilon}$. (Q) Polymorph composition along the nucleation pathway estimated by using the original and optimal CV.

basis of scalar invariants derived from Steinhardt's bond order parameters[31] (BOPs) (see section S1.3 of the Supporting Information). Within LJ and HS systems, the local $\overline{q}_6$ invariant effectively distinguishes the fluid from plausible crystalline phases (Figure S1). Subsequently, neighboring solid-like particles are clustered to form crystalline nuclei. The effectiveness of this approach, which involves multiple cutoffs and decision points, is typically assessed using the committor probability, $p_c(x)$, which is the probability that a trajectory initiated from configuration x reaches the crystalline phase (see section S1.5). A reliable CV is characterized[12,24,26] by its capacity to accurately predict p_c, where p_c values of configurations on its level sets must be narrowly distributed.

We first evaluate the effectiveness of a CV constructed according to these guidelines for homogeneous nucleation in the LJ system by computing the homogeneous nucleation rate at $T^* = 0.5$ and $p^* = 0$[32] and analyzing configurations gathered at a milestone with 50% survival probability (see eq S3). As illustrated in Figure 1A, the p_c distribution is narrow and unimodal, indicating the CV's adequacy for homogeneous nucleation. Subsequently, we assess the performance of the same CV for heterogeneous nucleation on a simple model surface, the 001 plane of a weakly attractive flexible FCC lattice at a reduced number density ($\rho_n^* = 1.13$). Analyzing the configurations obtained at a milestone with 50% survival probability, we observe a significantly broader p_c distribution (Figure 1B), spanning values from 0% to 100%. This demonstrates the inadequacy of the original CV in describing the progress of heterogeneous nucleation.

Upon visual inspection of low-p_c configurations (Figure 1D,E), we note the prevalence of fragmented crystalline nuclei consisting of smaller islands connected by narrow bridges composed of "low-coordinated" crystalline particles, i.e., solid-like particles with few other crystalline neighbors. In contrast, high-p_c configurations (Figure 1F,G) lack such islands and bridges. The inefficacy of the employed CV therefore appears to stem from its failure to exclude such small islands by including low-coordinate bridges within the crystalline nuclei. Upon visual inspection of such low-coordinated particles, we find them to be structurally crystalline (Figure S2). However, by examining particles' mobilities over intermediate time scales within the caging regime (where particles move by a fraction of σ), we observe low-coordinated solid-like particles within the first two liquid layers to move as rapidly as, or even faster than, their liquid-like counterparts (Figure 1M,N and Table S1). This subtle, yet statistically significant, anomaly is absent in the bulk (Figure S3A,B and Table S1), where crystalline particles consistently move more slowly than their liquid-like counterparts, irrespective of coordination number. Essentially, a divergence emerges between structure and dynamics. Interfacial particles with crystalline structures exhibit behavior akin to that of a disordered liquid. Notably, this divergence is not an artifact

https://doi.org/10.1021/acs.jpclett.3c03561
J. Phys. Chem. Lett. 2024, 15, 1279−1287

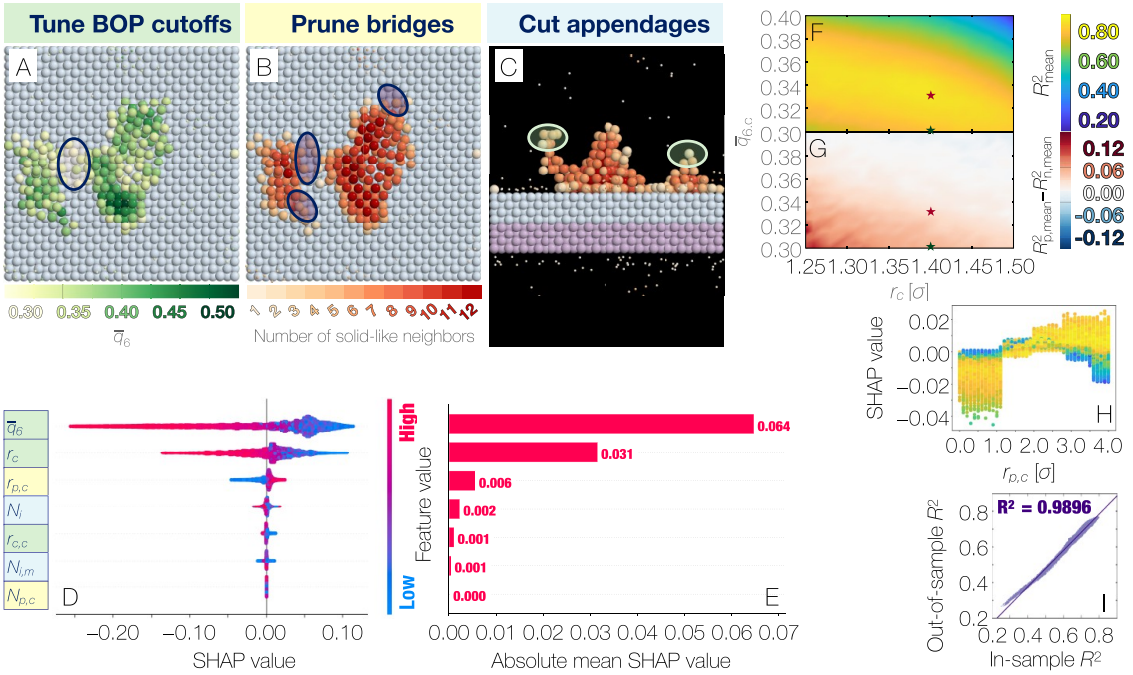


Figure 2. Three-pronged strategy for constructing good RCs for heterogeneous nucleation based on (A) adjustment of BOP cutoffs, such as r_c and $\overline{q}_{6,c}$, (B) pruning of bridges within the interfacial region, and (C) removal of low-coordinated and isolated crystalline extensions. Snapshots in panels A−C depict different renderings of the same crystalline nucleus using (A) the local $\overline{q}_6$ and (B and C) the number of solid-like neighbors. Highlighted transparent ovals identify regions targeted for removal by each strategy. (D) Beeswarm plots and (E) absolute mean SHAP values for the seven features pertinent to the implementation of strategies from panels A−C. Heat maps portraying (F) mean R^2 and (G) the change in R^2 ensuing from aggressive pruning as functions of (r_c, $\overline{q}_{6,c}$). Original and optimized CVs employed for panels B and C of Figure 1 are denoted by blue and red stars, respectively. (H) SHAP value plotted vs the pruning distance cutoff. Points are colored according to their R^2 value, with a color scale depicted in panel F and shared across panels. (I) Correlation between in-sample and out-sample R^2 values computed across all analyzed CVs.

of the observation window and persists over extended time scales (Figure 1L).

To enhance the effectiveness of CV, we pursue three strategies. The primary approach involves adjusting the original BOP cutoffs, primarily the $\overline{q}_6$ cutoff, motivated by our observation that low-coordinated particles within inter-island bridges also possess lower $\overline{q}_6$ values (Figure 2A). However, identifying an alternative $\overline{q}_6$ cutoff systematically is challenging due to considerable overlap between the $\overline{q}_6$ profiles of different crystallographic planes of FCC (Figure 1J) and HCP (Figure 1K) and that of the first liquid layer. Consequently, we adopt two other strategies in tandem. The second strategy, termed pruning (Figure 2B), involves excluding any solid-like particle possessing fewer than $N_{p,c}$ solid-like neighbors within a distance $r_{p,c}$ of the wall. The third strategy aims to cut crystalline appendages composed of low-coordinated solid-like particles uniformly across the simulation box by excluding any solid-like particle with fewer than N_i crystalline neighbors, unless one of those possesses a minimum of $N_{i,m} > N_i$ solid-like neighbors (Figure 2C). This strategy is motivated by our observation[32] of fragmented crystalline nuclei in homogeneous nucleation within the LJ liquid. A similar strategy has been employed in ice nucleation studies.[33] All of these strategies alter only the clustering criterion while maintaining the largest nucleus size as the chosen CV.

Implementing these strategies requires the specificity of values of seven distinct features. We use our physical intuition to assign reasonable values to every feature (listed in Table 1). Using the modified CV, a repeat rate calculation is performed. The resulting p_c distribution (Figure 1C) is notably narrower, suggesting improved efficacy of the modified CV. Moreover, no

Table 1. Employed Cutoff Values for the Optimized CV Used for Rate Calculations in the LJ and Hard Sphere Systems

cutoff	LJ	hard sphere
distance cutoff (r_c)	1.40σ	1.47σ
$\overline{q}_6$ cutoff	0.330	0.375
clustering distance cutoff ($r_{c,c}$)	1.30σ	1.28σ
pruning distance cutoff ($r_{p,c}$)	2.00σ	2.35σ
pruning neighbor threshold ($N_{p,c}$)	8	9
isolated neighbor threshold (N_i)	2	4
connection minimum threshold ($N_{i,m}$)	8	7

divergence between structure and dynamics is observed for the solid-like particles identified using the new CV, which move more slowly than their liquid-like counterparts regardless of their coordination number (Figure 1O,P).

Using the optimized CV yields a modest, but statistically inconclusive, increase in the computed nucleation rate (Table S3). This aligns with the known robustness of FFS to suboptimal CVs.[34] However, employing the optimized CV reveals distinct mechanistic insight. Examining configurations obtained at FFS milestones using the (less effective) original CV leads to an overestimation of FCC content and an underestimation of HCP and body-centered cubic (BCC) contents (Figure 1Q). This discrepancy arises because the islands excluded by the optimized CV predominantly exhibit a high FCC content. Essentially, relying on inadequate CVs may yield a flawed understanding of the nucleation mechanism, including the size, shape, and polymorphic composition of the critical nucleus.

The success of our three-pronged strategy raises three important questions. First, can an even more effective CV be

https://doi.org/10.1021/acs.jpclett.3c03561
J. Phys. Chem. Lett. 2024, 15, 1279−1287

constructed through an alternative combination of feature values? Second, which of these seven features exerts a statistically significant positive impact on CV efficacy? Lastly, do these transferability issues pertain solely to the LJ system, or do they extend to heterogeneous nucleation of close-packed crystals in other systems?

To tackle the first two questions, we systematically assess the influence of the seven features on CV performance by analyzing a collection of configurations with diverse p_c values (spanning from 0% to 100%). More than 250 000 distinct CVs, each representing a unique combination of cutoffs, are assessed using the R^2 values of the least-squares problem in eq S4, which quantifies the ability of an error functional committor model to predict individual p_c values.[35] Employing a machine learning approach,[36] a random forest model is trained on 80% of the data set, extracting Shapley (SHAP) scores for each feature across the remaining 20%. Feature significances are then assessed on the basis of the dispersion of SHAP score distributions illustrated in Figure 2D. Notably, distance, $\overline{q}_6$, and pruning cutoffs dominate CV performance, affirmed by their larger mean absolute SHAP scores (Figure 2E). Specifically, r_c and $\overline{q}_{6,c}$ explain 30% and 61% of the R^2 variability, respectively, consistent with their pivotal role in detecting crystallinity. The pruning distance cutoff, $r_{p,c}$, emerges as the third crucial feature, contributing 6% to data variability, whereas other features collectively explain 3% of the variability. Intriguingly, the associated pruning feature, $N_{p,c}$, lacks significance, indicating that removing particles with very low coordination numbers is sufficient for improving CV performance, while increasing the coordination threshold offers no improvement. This aligns with our observation that only solid-like particles with fewer than two crystalline neighbors exhibit heightened mobility (Figure 1M,N and Table S1). Contrary to our expectations, features linked to appendage removal (Figure 2C) exhibit no notable importance, and neither does using a clustering cutoff smaller than r_c. These findings underscore the vital role of feature selection in machine learning to evaluate intuition-driven strategies for enhancing CV performance.

If the important features are uncorrelated, examining SHAP score distributions at different values for each significant feature could help identify optimal cutoffs for that feature. Noticeable correlations among r_c, $\overline{q}_{6,c}$, and $r_{p,c}$ are, however, evident in Figure S4, hindering the identification of independent optimal cutoffs for each. Specifically, $R_{mean}{}^2(r_c, \overline{q}_{6,c})$ (Figure 2F) highlights strong correlations between $\overline{q}_{6,c}$ and r_c, indicating the presence of a tilted and curved optimality band.

The relatively low absolute mean SHAP value for $r_{p,c}$ might suggest minimal improvement from pruning, yet analyzing SHAP values against $r_{p,c}$ reveals an intriguing bifurcation (Figure 2H). Pruning with $r_{p,c} \lesssim 1.4\sigma$ yields consistently negative SHAP values regardless of CV effectiveness. This cutoff aligns with the first valley of the number density profile (Figure 1H), indicating that pruning will enhance CV efficacy only if the first liquid layer entirely falls within the pruning domain. This aligns with the considerable rightward shift of the $\overline{q}_6$ profile of the first liquid layer compared to the bulk (Figure 1I), necessitating its inclusion in the pruning domain. We therefore denote pruning with $r_{p,c} \geq 1.45\sigma$ as aggressive. Strong correlations between $r_{p,c}$ and $\overline{q}_{6,c}$, however, lead to the nonuniform efficacy of aggressive pruning (Figure 2G), which works best in the lower-left sector of the optimality band in Figure 2F. In other words, pruning wields the greatest impact when the unpruned CV lacks adequate selectivity, reinforcing CV robustness in scenarios where optimal r_c and $\overline{q}_{6,c}$ values are unknown.

An inherent risk in machine learning is to conduct screening and feature selection on nonrepresentative data sets. To preclude this possibility, we conduct a second screening in which each CV's R^2 is recalculated using a new set of 1000 configurations. The new R^2 values exhibit a perfect linear correlation with the older ones (Figure 2I), confirming that the efficacies of the assessed CVs remain unaffected by the screening data set.

Our systematic analysis confirms that the modified CV, initially constructed on the basis of physical intuition, falls within the optimality band in Figure 2F and within a zone where aggressive pruning is slightly effective (Figure 2G). Hence, enhancing its efficacy considerably with an alternative set of feature values appears to be improbable.

Next, we explored the applicability of these findings across various surfaces and systems. Initially, we examine heterogeneous crystal nucleation within the LJ liquid on a checkerboard surface with alternating attractive and repulsive patches (Figure S5A). Our SHAP analysis reveals similar important features, namely, $\overline{q}_{6,c}$, r_c, and $r_{p,c}$ (Figure S5D). Additionally, the optimality band of the $(r_c, \overline{q}_{6,c})$ space (Figure S5B) resembles that in Figure 2F. Furthermore, aggressive pruning enhances CV performance in the lower-left corner of the optimality band, aligning with our prior observations (Figure S5C). These findings validate the robustness and transferability of our insights across diverse nucleating surfaces, including those with significant chemical heterogeneity.

To confirm the broad applicability of these insights, we extend our analysis to the hard sphere system, which forms close-packed crystals identical to those of the LJ system but through a purely entropy-driven process. Using a CV similar to the LJ system but with a slightly different $\overline{q}_{6,c}$ (Figure S1B), we compute the homogeneous nucleation rate at a fluid packing fraction of 52% ($\beta P^* = 14.74$). By analyzing configurations with 50% survival probability, we observe a narrow and unimodal p_c distribution, confirming CV's efficacy for homogeneous nucleation (Figure 3A). We then consider heterogeneous nucleation on the 001 plane of an FCC lattice of hard spheres at $\rho_n^* = 1.3$, corresponding to a 2.7% lattice mismatch. Analyzing fluid configurations using the same CV yields large crystalline nuclei percolating across the periodic boundary (Figure 3D and Movie S1), indicating CV's probable inadequacy. This is confirmed by subsequent rate calculations and committor analysis, wherein an almost flat p_c distribution is obtained for configurations with 50% survival probability (Figure 3B). Similar to the LJ system, visually inspecting detected nuclei confirms CV's ability to identify genuine crystalline order (Figure 3D and Figure S6). We also observe a similar dynamical anomaly in which low-coordinated solid-like particles move as fast as or faster than those detected as disordered within the second and third fluid layers (Figure S7A,B and Table S4). Unlike the LJ system, however, pruning alone does not prevent the detection of percolating nuclei (Figure 3E), likely due to considerable ordering in the first two fluid layers, which is evident from the pronounced rightward shift of their $\overline{q}_6$ profiles (Figure 3G).

Considering these observations, identifying an optimal CV solely on the basis of physical intuition seems infeasible. Thus, we first conduct a systematic feature space analysis, screening more than 250 000 CVs and applying machine learning to evaluate their importance. Similar to the LJ system, $\overline{q}_{6,c}$, r_c, and $r_{p,c}$ are identified as pivotal features, explaining 72%, 21%, and

https://doi.org/10.1021/acs.jpclett.3c03561
J. Phys. Chem. Lett. 2024, 15, 1279−1287

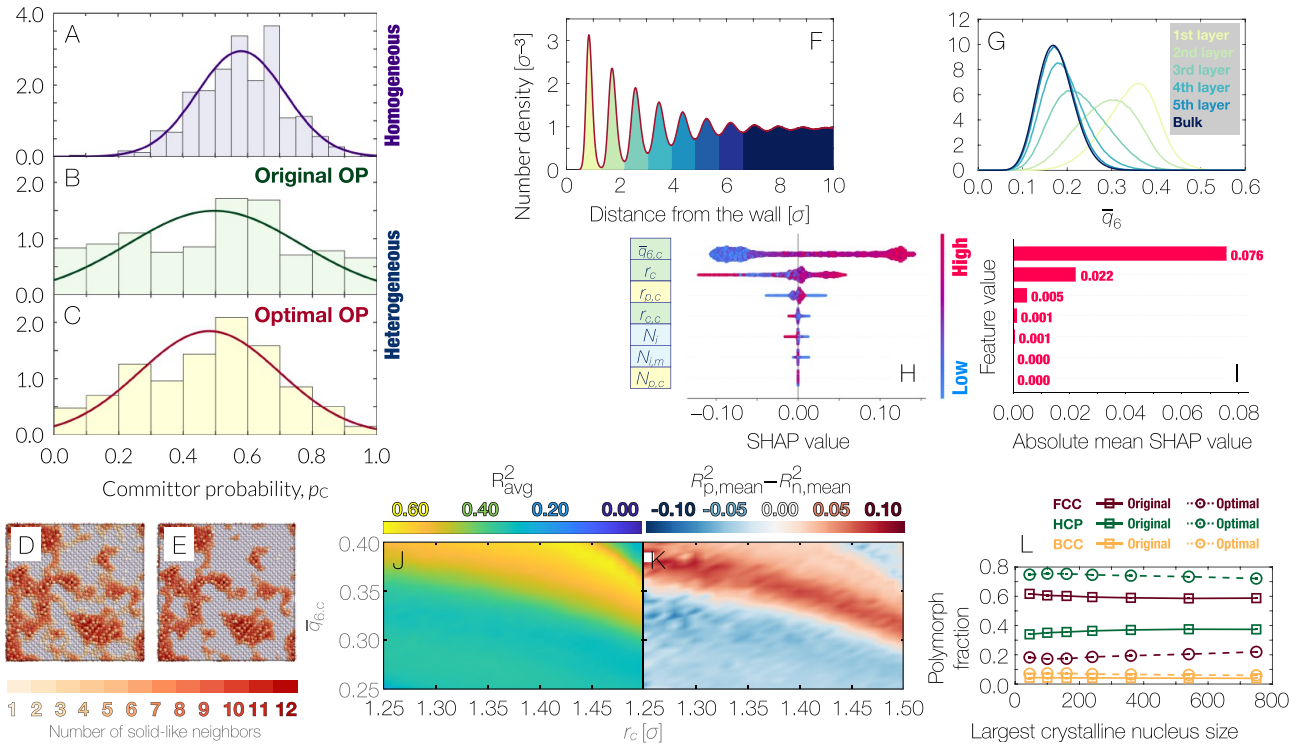


Figure 3. Committor probability distributions for (A) homogeneous and (B and C) heterogeneous nucleation in the HS system using (B) initial and (C) optimized CVs. The curves plotted in panels A−C are Gaussians, each with the same mean and standard deviation as the underlying histogram. Different representations of the same fluid configuration using the original CV (D) without and (E) with aggressive pruning. (F) Number density profile as a function of the distance from the substrate. (G) $\bar{q}_6$ profiles of different fluid layers. (H) Beeswarm plots and (I) absolute mean SHAP values for the seven features considered in this work. Heat maps illustrating (J) mean R^2 and (K) change in $R_{mean}{}^2$ after aggressive pruning across the $(r_c, \bar{q}_{6,c})$ space. (L) Polymorph composition along the nucleation pathway estimated using the original and the optimal CV.

5% of the R^2 variability, respectively (Figure 3H,I). Strong correlations exist among these features. Notably, the $(r_c, \bar{q}_{6,c})$ optimality band is narrower and more inclined (Figure 3J). Significant improvements due to aggressive pruning primarily occur in the band's bottom-left region (Figure 3K). Using the optimized CV eradicates the observed dynamical anomaly for low-coordinated solid-like particles (Figure S7C,D). These parallels with the LJ system suggest the potential universality of these transferability issues to heterogeneous nucleation of a close-packed crystal.

We use the top-performing CV from our analysis (given in Table 1) for a repeat rate calculation, yielding a statistically indistinguishable rate (Table S3). However, the p_c distribution narrows considerably for configurations with 50% survival probability (Figure 3C), albeit not by as much as the LJ system. This likely stems from pronounced interfacial ordering in the HS system, which can be addressed solely through layer-specific $\bar{q}_6$ cutoffs. Interestingly, polymorph compositions of crystalline nuclei exhibit stronger sensitivity to CV in the HS system. As depicted in Figure 3L, employing the original CV leads to a 3-fold overestimation of the FCC content (60% vs 19%) and a 2-fold underestimation of HCP content (36% vs 74%). Akin to the LJ system, using suboptimal CVs could misrepresent the nucleation mechanism, erroneously indicating the formation of FCC-rich nuclei.

These findings paint a consistent, qualitative picture. Using CVs optimized for homogeneous nucleation to describe heterogeneous nucleation on surfaces inducing interfacial ordering similar to that of the target crystal appears to be futile. Instead, it is necessary to enhance CV's selectivity, e.g., by adjusting BOP thresholds or pruning. Such modifications increase the threshold for local structural coherence, countering the premature order detection. An intriguing alternative for achieving similar efficacy involves generalizing neighbor averaging[37] of eq S2 by defining a collection of BOPs as

$$\mathbf{q}_l^{(k)}(i) = \frac{1}{N_b(i)+1}\left[\mathbf{q}_l^{(k-1)}(i) + \sum_{j=1}^{N_b(i)} \mathbf{q}_l^{(k-1)}(j)\right] \quad (1)$$

where $\mathbf{q}_l^{(1)} = \bar{\mathbf{q}}_l$. This process can be viewed as coarse graining, as it progressively incorporates structural information from an additional shell of nearest neighbors with each iteration.

Panels A and B of Figure S8 demonstrate that the fundamental character of the CV family endures after coarse graining. The $[r_c, q_{6,c}^{(2)}]$ domain maintains a curved, tilted optimality band (Figure S8A), while pruning retains potency in the band's lower-left corner (Figure S8B). What coarse graining does, however, is to diminish the R^2's dependence on auxiliary features as depicted in panels A and B of Figure 4. The coarse-grained CV (Figure 4B) exhibits reduced R^2 variability compared to the standard CV (Figure 4A), effectively reducing the significance of auxiliary features or, equivalently, the dimensionality of the important feature space. Indeed, our SHAP analysis reveals that $q_{6,c}^{(2)}$ and r_c explain 52% and 46% of R^2 variability, respectively, while $r_{p,c}$'s importance dwindles to a mere 1%. This underscores that pruning is not the sole means of inducing structural coherence. Despite the promise of coarse graining, however, pruning retains a unique advantage by enhancing CV resilience against suboptimal choices of BOP thresholds via harnessing variability.

https://doi.org/10.1021/acs.jpclett.3c03561
J. Phys. Chem. Lett. 2024, 15, 1279−1287

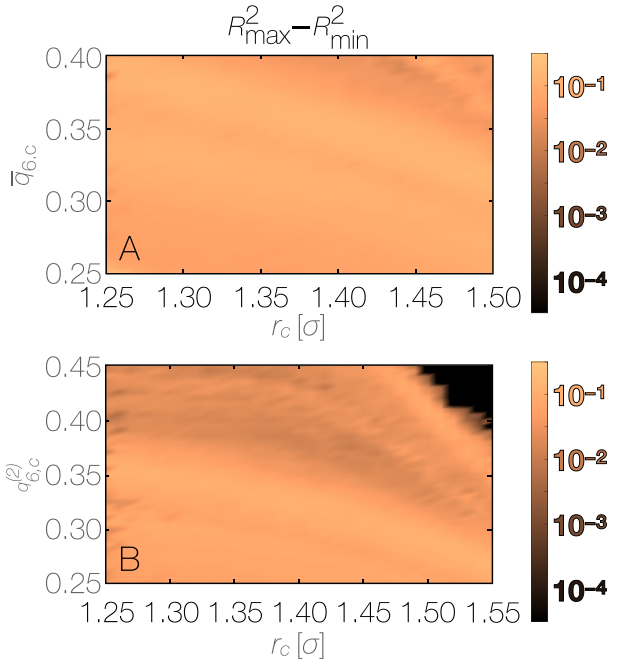


Figure 4. Heat maps of $R_{max}{}^2 - R_{min}{}^2$ across the (A) $(r_c, \bar{q}_{6,c})$ and (B) $[r_c, q^{(2)}_{6,c}]$ spaces for regular and coarse-grained CVs, respectively.

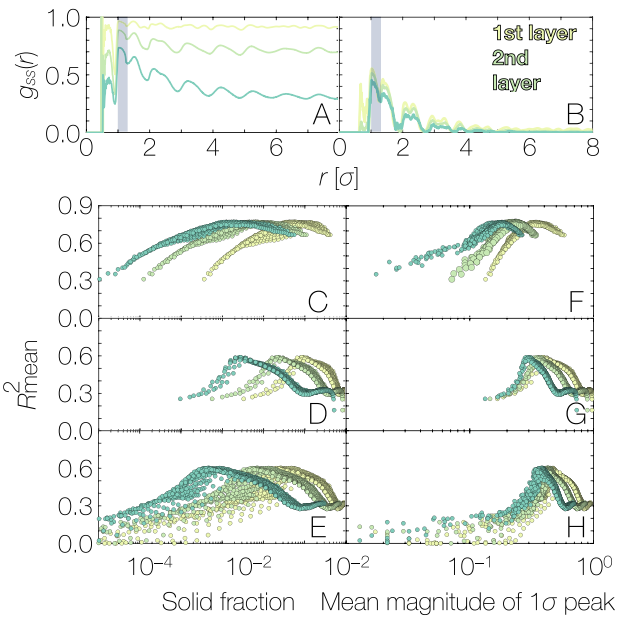


Figure 5. $g_{ss}(r)$, the solid−solid correlation function, computed for the first three interfacial layers of the hard sphere fluid using (A) a poor and (B) an effective coarse-grained CV. (C−E) Solid fractions and (F−H) mean magnitudes of the 1σ peak in the (C and F) LJ and (D, E, G, and H) hard sphere systems. For the hard sphere system, panels D and G depict the regular CV and panels E and H depict the coarse-grained CV.

These examples highlight the challenging nature of studying heterogeneous crystal nucleation, requiring customization and optimization of structure-based CVs through resource-intensive committor analysis and high-throughput screening. While feasible for simple model systems, this workflow might become prohibitively costly for complex systems. Therefore, identifying relevant mechanical observables within the metastable fluid for prescreening viable CVs could be extremely invaluable.

One such potential candidate is the solid fraction, representing the portion of interfacial particles labeled as solid-like by the CV. Heterogeneous nucleation occurs due to localized structural fluctuations at the interface, therefore necessitating non-zero solid fractions. However, an excessive solid fraction can hinder CV effectiveness, causing the detection of percolating nuclei similar to those depicted in panels D and E of Figure 3. Panels C−E of Figure 5 illustrate that each system and fluid layer features an optimal solid fraction for the most efficient CVs. However, this optimum varies by several orders of magnitude, depending on the system specifics, the distance from the interface, and the extent of coarse graining. This variation underscores that the solid fraction, while seemingly intuitive, cannot be used for robust prescreening.

Motivated by the localized nature of nucleation, alternative observables can be devised from solid−solid correlation function $g_{ss}(r)$, the probability that a particle at a lateral distance r from a solid-like particle is also solid-like. Panels A and B of Figure 5 depict representative $g_{ss}(r)$ profiles for the first three interfacial layers of the hard sphere fluid, computed for a poor (Figure 5A) and effective (Figure 5B) coarse-grained CV. Notably, as $r \to \infty$, $g_{ss}(r)$ converges toward the solid fraction within the corresponding layer. To capture relevant local correlations, we compute $g_{ss,\sigma}$, the average peak magnitude at $r \approx \sigma$, mirroring the first in-layer nearest neighbor shell. $g_{ss,\sigma} \approx 1$ will imply that nearly every first nearest neighbor of a solid-like particle will also be solid-like, while $g_{ss,\sigma} \approx 0$ will indicate complete isolation of solid-like particles. Both limiting scenarios are inconsistent with the spatially localized nature of nucleation. Therefore, a nonmonotonic relationship exists between $R_{mean}{}^2$ and $g_{ss,\sigma}$ (Figure 5F−H) akin to the solid fraction, yet the optimal $g_{ss,\sigma}$ values exhibit less variability within layers and across systems and CV variations, hovering within the range $[0.1-0.5]$, making it a better heuristic for CV prescreening. Further studies are needed to validate their broad applicability.

The findings herein prompt a profound paradox: Why do CVs capable of detecting crystalline order fail in predicting heterogeneous nucleation outcomes? A plausible avenue for reconciling this conundrum involves acknowledging that substrates that induce crystalline motifs within the interfacial fluid stabilize them solely within a finite-thickness region near the interface, even if such domains structurally resemble the crystalline phases assembled in the bulk. For nucleation to reach fruition, these ordered motifs must extend beyond their regions of stability and into the bulk via further structural fluctuations. Consequently, while nucleation is facilitated by their presence, its success relies on "orthogonal" structural fluctuations initiating at such pre-ordered islands.

This framework provides a coherent rationale for the CV refinement strategies employed here. Enhancing the coherence of the underlying BOP, via adjusting cutoffs or coarse graining, ensures the inclusion of only the pre-ordered domains surrounded by a corona of extending ordered patterns. This is quantitatively confirmed by observing a direct relationship between $\langle \bar{q}_6 \rangle$ and the number of solid-like neighbors (Figure S9). Similarly, pruning excludes bridges (or narrow islands) with a limited potential for fostering successful structural fluctuations.

In light of this conceptual picture, a more intricate procedure can be formulated for constructing refined CVs, starting with mapping a substrate's structural imprints within the fluid. Subsequently, regions exhibiting a consistent crystalline order could be permanently excluded from clustering. This approach

https://doi.org/10.1021/acs.jpclett.3c03561
J. Phys. Chem. Lett. 2024, 15, 1279−1287

would require a more extensive sampling of the metastable fluid. Furthermore, its rigorous implementation is possible only if the substrate induces permanent or fully wetting crystalline domains and becomes challenging when pre-ordered motifs are widespread but not all-encompassing, as seen for the surfaces considered here.

The observed pre-ordering in our study should not be misconstrued as surface freezing,[38] a phenomenon in which a finite-thickness crystalline film stabilizes at the interface above the melting temperature. This phenomenon typically leads to nearly barrier-free heterogeneous nucleation, which is not observed here. Furthermore, although interfacial solid fractions here are relatively large, they never approach unity, which is what would be expected in surface freezing.

Recently, machine learning methods have attracted attention for constructing CVs for rare events,[39,40] including nucleation.[41] The success of our physics-based approach in deriving effective CVs suggests limited potential improvements from the use of such strategies. However, one cannot dismiss their utility for surfaces with complex geometries and chemistries.

A method widely used for computing nucleation rates recently is seeding,[15,42,43] in which crystalline nuclei of varying sizes are introduced into the simulation box to determine the critical nucleus size based on their growth or shrinkage during MD simulations. Nucleation rates are then estimated by applying the classical nucleation theory (CNT). Seeding has been successfully applied across various systems[15,44,45] and has recently been extended to address heterogeneous nucleation.[46] Its success, however, relies both on CNT's validity and the availability of a suitable measure of nucleus size. Our findings caution against employing seeding for heterogeneous nucleation of close-packed crystals using traditional CVs, which might massively overestimate critical nucleus sizes and hence underestimate nucleation rates by orders of magnitude. Similar concerns apply to techniques such as umbrella sampling[47] and metadynamics[48] that are also highly sensitive to CV quality.

Traditional CVs can also fail in accurately describing homogeneous crystal nucleation from supersaturated solutions, as it is widely reported that the largest nucleus size is not always a suitable CV in that context.[19,49] This serves as another instance of divergence between traditional structural order and the nucleation outcome. In multicomponent systems, noticeable pre-ordering can arise, where solute aggregates with varying degrees of order will act as nucleation precursors. (Similar pre-ordering has also been reported in homogeneous nucleation within some metallic[12] systems.) This resembles the interfacial pre-ordering observed here, with the distinction that the latter is confined to the interface and does not typically require any barrier crossing. Therefore, refining the nucleus size definition, e.g., via a stricter crystallinity criterion, could yield robust CVs for both homogeneous and heterogeneous nucleation within mixtures and solutions.

Heterogeneous nucleation of tetrahedral crystals, such as ice, appears not to be affected by the transferability issues discovered here. CVs developed for homogeneous ice nucleation can be generally used in heterogeneous nucleation studies.[50,51] Because of directional interactions such as hydrogen bonding, tetrahedral liquids are less prone to excessive interfacial ordering, and the emerging domains are less likely to resemble bulk crystals. CV refinement is therefore unnecessary, except in extreme cases of interfacial ordering.[52] Further studies are needed to assess the true prevalence of CV breakdown in such systems.

In summary, our study highlights the limitations of traditional CVs for studying heterogeneous nucleation of close-packed crystals, especially with surfaces that induce significant pre-ordering in the fluid. We reveal an intriguing dynamical anomaly wherein the mobility of low-coordinated solid-like particles exhibit is higher than that of their disordered counterparts over intermediate time scales. To address this limitation, we customized conventional CVs, employing committor analysis and machine learning to systematically evaluate the associated feature space. Additionally, we introduce physics-based interfacial heuristics for CV pre-screening. These insights are critical for accurately simulating heterogeneous nucleation across different systems.

METHODS

All MD simulations are conducted using LAMMPS[53] within the isothermal−isobaric ensemble, with temperature and pressure controlled using the Nosé−Hoover thermostat[54] and the Parrinello−Rahman barostat.[55] In the LJ system, interactions between liquid and wall particles are specified by the LJ and Week−Chandlers−Andersen[56] (WCA) potentials. Hard spheres are represented using the pseudohard sphere potential.[57] To compute nucleation rates, we employ jFFS,[30] utilizing the size of the largest crystalline nucleus as the order parameter. To mitigate finite size effects in heterogeneous nucleation,[58] we choose system sizes conservatively, conducting simulations involving tens of thousands of liquid particles. CVs are assessed via a weighted mean squared error (MSE) approach[35] utilizing an error functional committor model inspired by earlier studies.[59,60] We explain our preference to use MSE, instead of alternatives such as likelihood maximization[12,24,26,60−62] in detail in section S2. Feature selection analysis is conducted employing random forest regression[63] followed by the estimation of Shapley scores.[36] For comprehensive information concerning system setup, MD simulations, rate calculations, CV screening, and feature selection, see the Supporting Information.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.3c03561.

Pre-ordering of the metastable hard sphere fluid prior to nucleation (Movie S1) (MOV)

Additional methodological details, including MD simulations and system setup, bond order parameters and rate calculations, CV screening and feature selection, mobility analysis, and polymorph characterization; a discussion of maximum likelihood estimators; Figures S1−S12; and Tables S1−S4 (PDF)

AUTHOR INFORMATION

Corresponding Author

Amir Haji-Akbari − *Department of Chemical and Environmental Engineering, Yale University, New Haven, Connecticut 06511, United States;* orcid.org/0000-0002-2228-6957; Email: amir.hajiakbaribalou@yale.edu

Authors

Tiago S. Domingues − *Department of Chemical and Environmental Engineering, Yale University, New Haven, Connecticut 06511, United States*

Sarwar Hussain − *Department of Chemical and Environmental Engineering, Yale University, New Haven, Connecticut 06511, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.3c03561

Author Contributions

‡T.S.D. and S.H. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

A.H.-A. gratefully acknowledges the support of National Science Foundation (NSF) Grants CBET-1751971 (CAREER Award) and CHE-2203527. This research was supported in part by NSF Grant PHY-1748958 to the Kavli Institute for Theoretical Physics (KITP). The authors thank P. G. Debenedetti, S. C. Glotzer, J. C. Palmer, and V. Molinero for insightful discussions. These calculations were performed on the Yale Center for Research Computing. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF Grant ACI-1548562.

■ REFERENCES

(1) Roostalu, J.; Surrey, T. Microtubule nucleation: beyond the template. *Nat. Rev. Mol. Cell Bio.* **2017**, *18*, 702−710.
(2) Knopf, D. A.; Alpert, P. A. Atmospheric ice nucleation. *Nat. Rev. Phys.* **2023**, *5*, 203−217.
(3) Wang, C.; Dong, H.; Jiang, L.; Hu, W. Organic semiconductor crystals. *Chem. Soc. Rev.* **2018**, *47*, 422−500.
(4) Battaglia, C.; Cuevas, A.; De Wolf, S. High-efficiency crystalline silicon solar cells: status and perspectives. *Energy Environ. Sci.* **2016**, *9*, 1552−1576.
(5) Chen, J.; Sarma, B.; Evans, J. M.; Myerson, A. S. Pharmaceutical crystallization. *Cryst. Growth Des.* **2011**, *11*, 887−895.
(6) Sosso, G. C.; Chen, J.; Cox, S. J.; Fitzner, M.; Pedevilla, P.; Zen, A.; Michaelides, A. Crystal nucleation in liquids: Open questions and future challenges in molecular dynamics simulations. *Chem. Rev.* **2016**, *116*, 7078−7116.
(7) De Yoreo, J. J.; Vekilov, P. G. Principles of crystal nucleation and growth. *Rev. Mineral. Geochem.* **2003**, *54*, 57−93.
(8) Yi, P.; Locker, C. R.; Rutledge, G. C. Molecular dynamics simulation of homogeneous crystal nucleation in polyethylene. *Macromolecules* **2013**, *46*, 4723−4733.
(9) Haji-Akbari, A.; Debenedetti, P. G. Direct calculation of ice homogeneous nucleation rate for a molecular model of water. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 10582−10588.
(10) Shibuta, Y.; Sakane, S.; Miyoshi, E.; Okita, S.; Takaki, T.; Ohno, M. Heterogeneity in homogeneous nucleation from billion-atom molecular dynamics simulation of solidification of pure metal. *Nat. Commun.* **2017**, *8*, 10.
(11) Bonati, L.; Parrinello, M. Silicon liquid structure and crystal nucleation from ab initio deep metadynamics. *Phys. Rev. Lett.* **2018**, *121*, 265701.
(12) Diaz Leines, G.; Rogal, J. Maximum Likelihood Analysis of Reaction Coordinates during Solidification in Ni. *J. Phys. Chem. B* **2018**, *122*, 10934−10942.
(13) Hussain, S.; Haji-Akbari, A. Role of nanoscale interfacial proximity in contact freezing in water. *J. Am. Chem. Soc.* **2021**, *143*, 2272−2284.
(14) Dijkstra, M.; Luijten, E. From predictive modelling to machine learning and reverse engineering of colloidal self-assembly. *Nat. Mater.* **2021**, *20*, 762−773.
(15) Knott, B. C.; Molinero, V.; Doherty, M. F.; Peters, B. Homogeneous nucleation of methane hydrates: Unrealistic under realistic conditions. *J. Am. Chem. Soc.* **2012**, *134*, 19544−19547.
(16) Kashchiev, D.; Cabriolu, R.; Auer, S. Confounding the paradigm: peculiarities of amyloid fibril nucleation. *J. Am. Chem. Soc.* **2013**, *135*, 1531−1539.
(17) Milek, T.; Zahn, D. Molecular simulation of Ag nanoparticle nucleation from solution: redox-reactions direct the evolution of shape and structure. *Nano Lett.* **2014**, *14*, 4913−4917.
(18) Liu, C.; Cao, F.; Kulkarni, S. A.; Wood, G. P.; Santiso, E. E. Understanding polymorph selection of sulfamerazine in solution. *Cryst. Growth Des.* **2019**, *19*, 6925−6934.
(19) Finney, A. R.; Salvalaglio, M. Multiple pathways in NaCl homogeneous crystal nucleation. *Faraday Discuss.* **2022**, *235*, 56−80.
(20) Rein ten Wolde, P.; Ruiz-Montero, M. J.; Frenkel, D. Numerical calculation of the rate of crystal nucleation in a Lennard-Jones system at moderate undercooling. *J. Chem. Phys.* **1996**, *104*, 9932−9947.
(21) Auer, S.; Frenkel, D. Prediction of absolute crystal-nucleation rate in hard-sphere colloids. *Nature* **2001**, *409*, 1020−1023.
(22) Moroni, D.; Ten Wolde, P. R.; Bolhuis, P. G. Interplay between structure and size in a critical crystal nucleus. *Phys. Rev. Lett.* **2005**, *94*, 235703.
(23) Trudu, F.; Donadio, D.; Parrinello, M. Freezing of a Lennard-Jones fluid: From nucleation to spinodal regime. *Phys. Rev. Lett.* **2006**, *97*, 105701.
(24) Beckham, G. T.; Peters, B. Optimizing nucleus size metrics for liquid−solid nucleation from transition paths of near-nanosecond duration. *J. Phys. Chem. Lett.* **2011**, *2*, 1133−1138.
(25) Filion, L.; Ni, R.; Frenkel, D.; Dijkstra, M. Simulation of nucleation in almost hard-sphere colloids: The discrepancy between experiment and simulation persists. *J. Chem. Phys.* **2011**, *134*, 134901.
(26) Jungblut, S.; Singraber, A.; Dellago, C. Optimising reaction coordinates for crystallisation by tuning the crystallinity definition. *Mol. Phys.* **2013**, *111*, 3527−3533.
(27) Cacciuto, A.; Auer, S.; Frenkel, D. Onset of heterogeneous crystal nucleation in colloidal suspensions. *Nature* **2004**, *428*, 404−406.
(28) Mithen, J.; Sear, R. Computer simulation of epitaxial nucleation of a crystal on a crystalline surface. *J. Chem. Phys.* **2014**, *140*, 084504.
(29) Espinosa, J. R.; Vega, C.; Valeriani, C.; Frenkel, D.; Sanz, E. Heterogeneous versus homogeneous crystal nucleation of hard spheres. *Soft Matter* **2019**, *15*, 9625−9631.
(30) Haji-Akbari, A. Forward-flux sampling with jumpy order parameters. *J. Chem. Phys.* **2018**, *149*, 072303.
(31) Steinhardt, P. J.; Nelson, D. R.; Ronchetti, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B* **1983**, *28*, 784.
(32) Hussain, S.; Haji-Akbari, A. How to quantify and avoid finite size effects in computational studies of crystal nucleation: The case of homogeneous crystal nucleation. *J. Chem. Phys.* **2022**, *156*, 054503.
(33) Reinhardt, A.; Doye, J. P.; Noya, E. G.; Vega, C. Local order parameters for use in driving homogeneous ice nucleation with all-atom models of water. *J. Chem. Phys.* **2012**, *137*, 194504.
(34) Hussain, S.; Haji-Akbari, A. Studying rare events using forward-flux sampling: Recent breakthroughs and future outlook. *J. Chem. Phys.* **2020**, *152*, 060901.
(35) Ma, A.; Dinner, A. R. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. CHem. B* **2005**, *109*, 6769−6779.
(36) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *arXiv* **2017**, DOI: 10.48550/arXiv.1705.07874.
(37) Lechner, W.; Dellago, C. Accurate determination of crystal structures based on averaged local bond order parameters. *J. Chem. Phys.* **2008**, *129*, 114707.
(38) Haji-Akbari, A.; Debenedetti, P. G. Perspective: Surface freezing in water: A nexus of experiments and simulations. *J. Chem. Phys.* **2017**, *147*, 060901.
(39) Nadler, B.; Lafon, S.; Coifman, R. R.; Kevrekidis, I. G. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Appl. Comput. Harmon. Anal.* **2006**, *21*, 113−127.
(40) Appeldorn, J. H.; Lemcke, S.; Speck, T.; Nikoubashman, A. Employing artificial neural networks to identify reaction coordinates and pathways for self-assembly. *J. Phys. CHem. B* **2022**, *126*, 5007−5016.

(41) Beyerle, E. R.; Zou, Z.; Tiwary, P. Recent advances in describing and driving crystal nucleation usingmachine learning and artificial intelligence. *Curr. Opin. Solid State Mater. Sci.* **2023**, *27*, 101093.
(42) Bai, X.-M.; Li, M. Calculation of solid-liquid interfacial free energy: A classical nucleation theory based approach. *J. Chem. Phys.* **2006**, *124*, 124707.
(43) Espinosa, J. R.; Vega, C.; Valeriani, C.; Sanz, E. Seeding approach to crystal nucleation. *J. Chem. Phys.* **2016**, *144*, 034501.
(44) Sanz, E.; Vega, C.; Espinosa, J.; Caballero-Bernal, R.; Abascal, J.; Valeriani, C. Homogeneous ice nucleation at moderate supercooling from molecular simulation. *J. Am. Chem. Soc.* **2013**, *135*, 15008−15017.
(45) Zimmermann, N.; Vorselaars, B.; Espinosa, J. R.; Quigley, D.; Smith, W. R.; Sanz, E.; Vega, C.; Peters, B. NaCl nucleation from brine in seeded simulations: Sources of uncertainty in rate estimates. *J. Chem. Phys.* **2018**, *148*, 222838.
(46) Yuan, T.; DeFever, R. S.; Zhou, J.; Cortes-Morales, E. C.; Sarupria, S. RSeeds: Rigid Seeding Method for Studying Heterogeneous Crystal Nucleation. *J. Phys. Chem. B* **2023**, *127*, 4112−4125.
(47) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.
(48) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562−12566.
(49) Bulutoglu, P. S.; Wang, S.; Boukerche, M.; Nere, N. K.; Corti, D. S.; Ramkrishna, D. An investigation of the kinetics and thermodynamics of NaCl nucleation through composite clusters. *PNAS Nexus* **2022**, *1*, pgac033.
(50) Lupi, L.; Peters, B.; Molinero, V. Pre-ordering of interfacial water in the pathway of heterogeneous ice nucleation does not lead to a two-step crystallization mechanism. *J. Chem. Phys.* **2016**, *145*, 211910.
(51) Lupi, L.; Hanscam, R.; Qiu, Y.; Molinero, V. Reaction coordinate for ice crystallization on a soft surface. *J. Phys. Chem. Lett.* **2017**, *8*, 4201−4205.
(52) Zhao, W.; Li, T. On the challenge of sampling multiple nucleation pathways: A case study of heterogeneous ice nucleation on FCC (211) surface. *J. Chem. Phys.* **2023**, *158*, 124501.
(53) Plimpton, S. J. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comp. Phys.* **1995**, *117*, 1−19.
(54) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695−1697.
(55) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.
(56) Weeks, J. D.; Chandler, D.; Andersen, H. C. Role of repulsive forces in determining the equilibrium structure of simple liquids. *J. Chem. Phys.* **1971**, *54*, 5237−5247.
(57) Jover, J.; Haslam, A.; Galindo, A.; Jackson, G.; Müller, E. Pseudo hard-sphere potential for use in continuous molecular-dynamics simulation of spherical and chain molecules. *J. Chem. Phys.* **2012**, *137*, 144505.
(58) Hussain, S.; Haji-Akbari, A. How to quantify and avoid finite size effects in computational studies of crystal nucleation: The case of heterogeneous ice nucleation. *J. Chem. Phys.* **2021**, *154*, 014108.
(59) Wedekind, J.; Strey, R.; Reguera, D. New method to analyze simulations of activated processes. *J. Chem. Phys.* **2007**, *126*, 134103.
(60) Peters, B. Recent advances in transition path sampling: accurate reaction coordinates, likelihood maximisation and diffusive barrier-crossing dynamics. *Mol. Simul.* **2010**, *36*, 1265−1281.
(61) Borrero, E. E.; Escobedo, F. A. Reaction coordinates and transition pathways of rare events via forward flux sampling. *J. Chem. Phys.* **2007**, *127*, 164101.
(62) Peters, B.; Trout, B. L. Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys.* **2006**, *125*, 054108.
(63) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.

https://doi.org/10.1021/acs.jpclett.3c03561
J. Phys. Chem. Lett. 2024, 15, 1279−1287