The Journal of Chemical Physics

A variational approach to assess reaction coordinates for two-step crystallization

Cite as: J. Chem. Phys. **158**, 094503 (2023); https://doi.org/10.1063/5.0139842
Submitted: 23 December 2022 • Accepted: 13 February 2023 • Accepted Manuscript Online: 13 February 2023 • Published Online: 06 March 2023

A. R. Finney and M. Salvalaglio

AIP Publishing

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

A variational approach to assess reaction coordinates for two-step crystallization

Cite as: J. Chem. Phys. 158, 094503 (2023); doi: 10.1063/5.0139842
Submitted: 23 December 2022 • Accepted: 13 February 2023 •
Published Online: 6 March 2023

A. R. Finney[a] and M. Salvalaglio[b]

AFFILIATIONS

Thomas Young Centre and Department of Chemical Engineering, University College London, London WC1E 7JE, United Kingdom

Note: This paper is part of the JCP Special Topic on Nucleation: Current Understanding Approaching 150 Years After Gibbs.
[a]**Electronic mail:** a.finney@ucl.ac.uk
[b]**Author to whom correspondence should be addressed:** m.salvalaglio@ucl.ac.uk

ABSTRACT

Molecule- and particle-based simulations provide the tools to test, in microscopic detail, the validity of *classical nucleation theory*. In this endeavor, determining nucleation mechanisms and rates for phase separation requires an appropriately defined reaction coordinate to describe the transformation of an out-of-equilibrium parent phase for which myriad options are available to the simulator. In this article, we describe the application of the variational approach to Markov processes to quantify the suitability of reaction coordinates to study crystallization from supersaturated colloid suspensions. Our analysis indicates that collective variables (CVs) that correlate with the number of particles in the condensed phase, the system potential energy, and approximate configurational entropy often feature as the most appropriate order parameters to quantitatively describe the crystallization process. We apply time-lagged independent component analysis to reduce high-dimensional reaction coordinates constructed from these CVs to build Markov State Models (MSMs), which indicate that two barriers separate a supersaturated fluid phase from crystals in the simulated environment. The MSMs provide consistent estimates for crystal nucleation rates, regardless of the dimensionality of the order parameter space adopted; however, the two-step mechanism is only consistently evident from spectral clustering of the MSMs in higher dimensions. As the method is general and easily transferable, the variational approach we adopt could provide a useful framework to study controls for crystal nucleation.

I. INTRODUCTION

Crystal nucleation marks the emergence of long-range order in a parent liquid or gas phase that may only display short-range symmetry at the scale of constituent monomers. In particle or molecular systems, the size of the critical nucleus—the smallest collection of monomers with crystalline order that can lead to bulk crystals—is typically many orders of magnitude smaller than Avogadro's number.[1] Combined with the fact that nucleation is a rare event, this makes investigating nucleation mechanisms *in situ* particularly challenging.

Computer simulations employing Molecular Dynamics (MD) algorithms have provided significant insights into crystallization pathways, especially since the advent of methods to enhance the sampling of rare events.[2] To monitor the crystallization process and establish nucleation kinetics in these types of simulations, a suitable reaction coordinate (RC) is needed to reduce the $6N$ (N being the number of monomers) dimensional phase space to just a handful of collective variables (CVs) that completely capture the emergence of long-range order.[3,4] All other degrees of freedom can be ignored when determining *relative* nucleation rates.

Classical Nucleation Theory (CNT) adopts the size of an embryo of a new thermodynamic phase, usually its radius, as an RC for phase transformation.[5] With respect to crystallization, the number of monomers in the new phase is a more appropriate metric for this size, given the highly faceted and non-spherical geometry of crystals, even at small sizes. As several studies have demonstrated, however, a one-dimensional RC can be unsuitable to describe the evolution of a crystallizing system.[6–9] For example, our own work demonstrates that a two-dimensional RC, quantifying both the size of emerging clusters and their crystalline order, is helpful to describe the formation of crystals from metastable solutions.[10–12] Still, no obvious definition for these variables can

have consequences for understanding nucleation mechanisms and predicting crystallization rates.[3,4,13]

Common CVs to approximate RCs for crystallization are functions of the positional coordinates of a collection of particles. These must be continuous and differentiable if used in biased enhanced sampling schemes, but typically this is not a prerequisite for analysis purposes. A simple example CV used in this context is the first-sphere coordination number; however, this typically fails to capture the local symmetry of a crystal lattice and, therefore, might not be suitable to distinguish dense amorphous phases and crystal polymorphs. Bond orientational order parameters can achieve this by, for example, making use of spherical harmonic functions to quantify the relative position of monomers in a coordination sphere with respect to one another.[14] Alternatively, if a reference structure is known, one can compute the relative distance between particles in simulations and this reference in topography space, or perform topological graph analyses, with nodes in the graph representing monomers, to identify crystal structures.[15,16] Accurate classification of monomers at crystal surfaces and defects is challenging in all of these methods due to under-coordination at these sites.

No generally applicable procedure exists to choose order parameters to study multi-step crystal nucleation; this often comes down to chemical/physical intuition on behalf of the researcher. A useful review on the topic was provided by Peters,[3] who remarks "[h]uman intuition remains the best source of trial coordinates and mechanistic hypotheses, and there is no procedure for having an epiphany." There are, however, methods available to test the suitability of the RC. These include, for example, likelihood maximization[17] and committor analyses.[18]

The Variational Approach to Markov Processes (VAMP)[19] is a generalized version of the Variational Approach to Conformational Dynamics (VAC)[20] that has been successfully applied to determine suitable RCs in systems with stochastic dynamics, including protein folding and problems associated with molecular and crystallization kinetics.[21,22] Here, we apply VAMP to test the suitability of thousands of potential RCs defined by combining sets of CVs typically used to study crystallization pathways in monoatomic solids. VAMP allows us to quantify the effectiveness of the RCs to capture the slow dynamic modes associated with crystallization and identify which combinations of CVs best describe emerging order. To this aim, we perform simulations of metastable colloid suspensions that undergo crystallization, use VAMP to identify the most suitable combination of CVs for every set of dimensionality to define RCs, perform dimensionality reduction using TICA, and construct Markov state models to quantify kinetics and transformation mechanisms. In Sec. II, we provide a brief overview of the salient features of the methods employed, with an emphasis on VAMP. For a more involved discussion, including associated Markov modeling methods, see Refs. 19, 23, and 24.

II. THEORETICAL BACKGROUND

Projections of the highly nonlinear evolution of a system in phase space onto a low-dimensional representation are often employed to understand physicochemical processes. When analyzing transitions in nonlinear dynamical systems, the Koopman operator, $\mathcal{K}$, which is linear in a space of infinite observables, completely describes the time evolution of a system. If a system occupies states in phase space at $\mathbf{x}_1$ at time t, $\mathcal{K}$ is an operator that acts on the function g to determine the expectation of the system being in states at $\mathbf{x}_2$ at $t+\tau$ (τ being some lag-time), given the conditional probability density of states, $p(\mathbf{x}_1, \mathbf{x}_2)$

$$[\mathcal{K}g](\mathbf{x}) = \int p(\mathbf{x}_1, \mathbf{x}_2) g(\mathbf{x}_2) d\mathbf{x}_2 = \mathbb{E}[g(\mathbf{x}_{t+\tau})]. \tag{1}$$

Its spectral decomposition, therefore, completely characterizes (meta)stable states and transitions between them.[20] With a finite number of functions characterizing the time evolution for the process of interest, a good approximation of $\mathcal{K}$ is the propagator, $\mathbf{K}$, which in principle allows the determination of the transition probabilities and timescales associated with crystallization in closed thermodynamic systems (where the partition function is bounded by the finite number of particles in the simulations).

Given a set of functions of the configurational space of the system of interest, $\mathbf{f}(\mathbf{r}) = (f_1(\mathbf{r}), f_2(\mathbf{r}), \ldots, f_n(\mathbf{r}))$, i.e., CVs that project the full $3N$ configurational coordinates of N atoms/particles in a system, $\mathbf{r}$, onto an n-dimensional RC, the time-dependent Markovian dynamics can be predicted according to

$$\mathbb{E}[\mathbf{f}(\mathbf{r}, t+\tau)] \approx \mathbf{K}^{\top} \mathbb{E}[\mathbf{f}(\mathbf{r}, t)], \tag{2}$$

where $\mathbb{E}$ is the expectation value evaluated for an average trajectory and $\mathbf{f}$ is the array of CVs that approximates the eigenfunctions characterizing transitions between (meta)stable states. VAMP can be applied to optimize the dimensionality and choice of $\mathbf{f}$. This involves computing the time-dependent covariance matrices,

$$\begin{aligned}
\mathbf{C}_{00} &= \frac{1}{T-\tau} \sum_{t=0}^{T-\tau} [\mathbf{f}(\mathbf{r}, t) - \overline{\mathbf{f}_0}(\mathbf{r})][\mathbf{f}(\mathbf{r}, t) - \overline{\mathbf{f}_0}(\mathbf{r})], \\
\mathbf{C}_{11} &= \frac{1}{T-\tau} \sum_{t=\tau}^{T} [\mathbf{f}(\mathbf{r}, t) - \overline{\mathbf{f}_1}(\mathbf{r})][\mathbf{f}(\mathbf{r}, t) - \overline{\mathbf{f}_1}(\mathbf{r})], \\
\mathbf{C}_{01} &= \frac{1}{T-\tau} \sum_{t=0}^{T-\tau} [\mathbf{f}(\mathbf{r}, t) - \overline{\mathbf{f}_0}(\mathbf{r})][\mathbf{f}(\mathbf{r}, t+\tau) - \overline{\mathbf{f}_1}(\mathbf{r})],
\end{aligned} \tag{3}$$

where $\overline{\mathbf{f}_0}(\mathbf{r})$ and $\overline{\mathbf{f}_1}(\mathbf{r})$ are the mean CV values in $t = 0 \to (T-\tau)$ and $t = \tau \to T$ time windows, respectively. $\mathbf{K}$ is simply given by $\mathbf{C}_{00}^{-1}\mathbf{C}_{01}$, and singular values of the half-weighted propagator provide a metric to determine how well $\mathbf{f}(\mathbf{r})$ approximates the eigenvectors that capture the slow modes for crystallization,

$$R_2(\mathbf{f}, \mathbf{r}) = \| \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \|^2, \tag{4}$$

R_2 is known as the VAMP-2 score and is used in this work to judge the appropriateness of RCs to describe two-step colloid crystallization. Cross-validation[25] is carried out to ensure that the dynamical model is not overfitted. This is done by partitioning the CV time series into training and test data, building the model on the training subset, and validating it with the remaining data. It is then possible to compute R_2 from multiple partitions of training and test data.[19]

In Secs. III and IV, we use VAMP to identify the best combinations of CVs to determine the crystallization kinetics. In stochastic dynamical systems, one can construct models based on Markovian dynamics to map the evolution of systems in these low-dimensional coordinates to extract mechanisms and timescales for the processes

of interest. First, we use time-lagged independent component analysis (TICA)[26,27] to project the CVs onto one or two components by solving

$$\mathbf{C}(\tau)\mathbf{f}(\mathbf{r},t) = \lambda \mathbf{C}_{00}\mathbf{f}(\mathbf{r},t), \tag{5}$$

where λ are eigenvalues that determine the slowness of the transitions in the system dynamics.

After TICA, we partition the sampled data into discrete partitions $\{s_2, s_2, \ldots, s_N\}$. The probability weights for each partition are a function of the stationary distribution, μ,

$$\pi_i = \int_{\mathbf{x}\in s_i} \mu(\mathbf{x}) d\mathbf{x}, \tag{6}$$

so that combining simulation trajectories that sample reasonably well the local TICA space centered on different partitions allows us to determine the relative probability for the system to occupy these partitions and evaluate free energy differences. In addition, counting the transitions between partitions in the TICA trajectories allows us to evaluate $\mathbf{K}$,

$$\mathbf{K}(\tau)\boldsymbol{\pi} = \boldsymbol{\pi}, \tag{7}$$

and, therefore, provides kinetic information for the crystallizing system.

III. COMPUTATIONAL DETAILS

We performed simulations using the LAMMPS (v. 7Aug2019) MD simulator.[28] To prepare the initial configurations, 388 spherical particles were randomly assigned to a $10 \times 10 \times 10$ face-centered cubic lattice (where the reduced lattice density was $0.005\sigma^{-3}$ and the lattice constant was $9.283\,177\sigma$) in a simulation cell with a reduced particle density, $\rho^* = 0.000\,485$. The resulting cubic simulation cell lengths were $92.831\,77\sigma$ and ~90% of the lattice sites were vacant. Particle interactions were modeled using a colloid/Yukawa[29,30] potential to simulate van der Walls attraction and electrostatic repulsion between colloid particles in simulations adopting three-dimensional periodic boundaries. The pair potential coefficients with the force field implemented in LAMMPS were $A^* = 53$, $d^* = 5$, and $B^* = 20$, representing the Hamaker constant, particle diameter, and prefactor of the Yukawa potential, respectively, which approximate a surrounding electrolyte solution as a continuum field. Interactions were truncated at a reduced distance of 12.5σ. Particle velocities were assigned at random from a Maxwell–Boltzmann distribution with mean reduced temperature, $T^* = 2$. The simulations were performed for 2×10^7 steps with a timestep $\Delta t^* = 0.005$, during which particle velocity rescaling was carried out every 100 steps to maintain a constant temperature, and particle positions were recorded every 100 steps for subsequent analyses. We performed 1000 simulations where the initial velocity assignment was randomized, but all other simulation details remained the same. While condensation was observed in all simulations, only 11 of the simulations resulted in crystallization, as indicated by a potential energy per particle threshold: $E^* < -10\varepsilon$. It was these crystallizing trajectories that were used for analyses of RCs.

A total of 19 CVs were computed either during time integration or by post-processing simulation trajectories using the PLUMED software (v. 2.5.1).[31] These are useful indicators for phase separation and/or crystallization and are potential order parameters that can be used to construct RCs. Table I provides the list of CVs and their labels adopted herein. The CVs can be classified into one of three categories: (i) average properties of all particles in the system regardless of their local environment (ene, ent, cn.mean, Q4.mean, and Q6.mean); (ii) average properties of all particles in the system according to their local structure (q4.mean, laQ4.mean, q6.mean, and laQ6.mean); and (iii) total numbers of particles according to some geometric criteria of their local structure (ncl, ncs, ncnq4, ncnq6, nclust1, non, fcc, hcp, bcc, and ico). Please see Sec. S1 in the supplementary material for a detailed description and mathematical definition of the CVs.

TABLE I. Collective variables (CVs) computed in this work.

CV	Label
Mean first-sphere coordination number	cn.mean
Number of particles in a condensed phase (CN > 3)	ncl
Number of particles in a solid-like phase (CN > 6)	ncs
Number of particles in the largest cluster[34]	nclust1
Mean Q4 Steinhardt bond order[14]	Q4.mean
Mean local Q4 bond order	q4.mean
Number of coordinated particles with local Q4 < 0.3	ncnq4
Local average[35] Q4 bond order	laQ4.mean
Mean Q6 Steinhardt bond order[14]	Q6.mean
Mean local Q6 bond order	q6.mean
Number of coordinated particles with local Q6 > 0.7	ncnq6
Local average[35] Q6 bond order	laQ6.mean
Pair entropy function[36]	ent
System potential energy	ene
Number of particles not identified as fcc/hcp/bcc/ico[a]	non
Number of face-centered cubic particles[a]	fcc
Number of hexagonal close-packed particles[a]	hcp
Number of body-centered cubic particles[a]	bcc
Number of icosahedral particles[a]	ico

[a]Evaluated using polyhedral template matching (PTM).[15]

The VAMP and Markov state model (MSM) analyses were performed using the PyEMMA (v. 2.5.11)[32] and deeptime (v. 0.4.1)[33] Python libraries. See the Data Availability section for information on how to access interactive Python notebooks used in this work and to download input files used to perform simulations and generate the CV time series data.

IV. RESULTS

A. Two-step colloid crystallization

In all 1000 independent simulations, an initial phase separation resulted in a finite-sized droplet of a condensed disordered phase in pseudo-equilibrium with a diluted vapor-like phase characterized by a significantly lower density cf. the initial one. In all simulations, the emergent phase was liquid-like: colloid particles in the dense liquid droplet (DLD) were highly mobile, and there was a frequent exchange between monomers in the DLD and the surrounding low-density phases. The condensation is indicated by a change in the average reduced potential energy per particle from an initial value

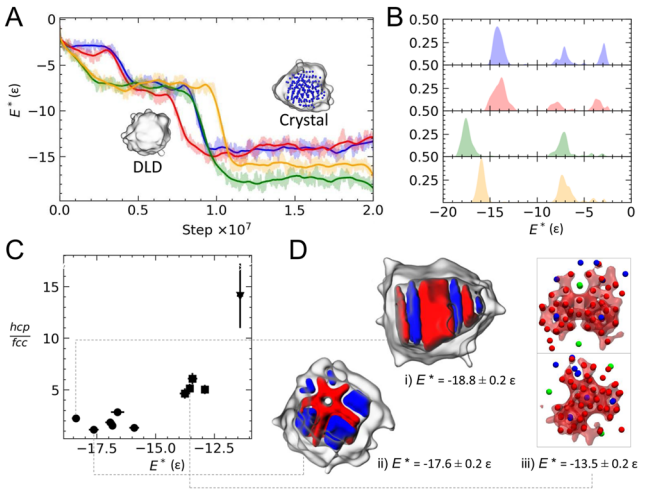

FIG. 1. (a) Mean reduced potential energy per particle, E^*, as a function of step number in four different crystallizing trajectories, as indicated by the colors. Raw data are shown by the shaded regions, while the solid lines result from a third-order polynomial Savitzky–Golay filtering[37] of the raw values using a window length of 2×10^4 trajectory frames. The inset shows snapshots of the largest cluster (indicated by the white transparent surface) in the trajectory shown by the blue curve. The blue spheres highlight colloid particles with CN > 6. (b) Probability densities for E^* from the trajectory data in A, as indicated by the colors. (c) Average hcp/fcc fraction in crystals as a function of the average potential energy per particle in the final 5000 frames of each trajectory. Error bars indicate uncertainties of one standard deviation in the data. Circle, square, and triangle data points highlight different clusters of crystal structures. (d) Snapshots of example crystal structures were taken from the end of simulation trajectories. Red domains/spheres indicate hcp-like particles, blue domains/spheres are fcc-like particles, green spheres are icosahedral-like particles, and the transparent white surface indicates all other types of particles that reside at the crystal surface. In iii, the crystal transparent surface is removed, and the fcc domain is made transparent for clarity. The same structure is projected perpendicular and parallel to the plane formed by icosahedral particles in the top and bottom boxes, respectively.

of $E^* \approx -2\varepsilon$ that reaches a plateau corresponding to $E^* \approx -7\varepsilon$, as shown in Fig. 1 for four example crystallizing trajectories. The time for this transition varies, as expected for an activated process of condensation.

Analyzing the behavior of E^* throughout the trajectories represented in Fig. 1(a) reveals the presence of one additional step change in E^*, marking the emergence of a crystal phase within the DLD. Crystallization occurred in ~1% of simulations, with long-range order consistently emerging within the DLDs, indicative of a two-step crystallization pathway. This behavior is not unexpected. Indeed, two-step crystallization was identified both in simulations and experiments in a range of systems, demonstrating that this pathway to crystals is more prolific than once assumed.[2,38]

In the seminal work of ten Wolde and Frenkel,[6] simulations indicated that colloid crystallization occurs in dense fluids when the simulation conditions approach those associated with the fluid–fluid critical point. In their work, the free energy landscape for crystallization was projected onto a two-dimensional RC characterizing the total size of monomer clusters and the size of the crystalline regions in clusters. While the lowest energy crystallization pathway evolved with a near–linear correlation in the two RC variables away from the critical point, large amorphous clusters emerge before the onset of crystalline order close to this point. It is important to note, though, that this *roundabout* pathway to crystals involves a single energy barrier in the 2D RC space and is not necessarily consistent with the observations in this work, where two activated events are involved in the crystallization of the initial fluid.

To consider the proximity of the initial system conditions to the fluid–fluid critical point in our model, we performed an additional 15 simulations. Each of these was prepared using the same random distribution of particles on a sparse face centered cubic lattice with $\rho^* = 0.000\,485$, but where the reduced temperature was $T^* = 1.8 - 2.2$. From the densities of the emerging DLDs and nanocrystals in (pseudo-)equilibrium with vapor phases (see Sec. S2 in the supplementary material for details), we constructed a $T^* - \rho^*$ phase diagram, shown in Fig. S1 in the supplementary material. This indicates that the simulations in this study at $T^* = 2$ are initiated in the immiscible region of the phase diagram and close to the (upper) vapor–fluid critical temperature, $T_c^* \approx 2.05$; hence, $T^*/T_c^* \approx 0.976$.

As for the nucleation of the DLD, the times associated with the nucleation of a crystalline domain within the DLD are stochastically distributed and are marked by a significant variation in E^* [see Fig. 1(a)]. While an escape probability could be built based on the time taken to observe such a sudden change in E^*, given the limited statistics, alternative methods to evaluate crystal nucleation times are necessary. They will be discussed in Subsections IV B and IV C.

Another noteworthy observation from the crystallizing trajectories is that despite crystallization conditions being consistent throughout the entire set of simulations, the structures spontaneously emerging from crystal nucleation appear to differ. In Fig. 1(b), the density of energy states representing the crystal in equilibrium with a low-density vapor phase is misaligned in different simulations. Some systems have a much lower average potential energy than others, despite the crystal phase emerging relatively early on in the trajectories. These different crystals, characterized by different potential energy levels, result from stacking faults introduced during the rapid propagation of order in the DLD.

By performing polyhedral template matching (PTM),[15] we can estimate the numbers of colloids in the single crystals with face-centered cubic (fcc), hexagonal close-packed (hcp), body-centered cubic (bcc), and icosahedral (ico) local symmetries. Considering the latter stages of trajectories, PTM points to nanocrystals rich in fcc and hcp local environments, while negligible levels of bcc are found. Figure 1(c) shows the relative hcp/fcc content as a function of the system potential energy (which at equilibrium is dominated by the potential energy of the crystal). Crystals with lower hcp/fcc content have the most negative potential energy. Figure 1(d) provides example snapshots for three crystals with system potential energies provided inset. These crystals contain large domains of fcc and hcp particles, though, as is clear in Fig. 1(d-ii and iii), defects are apparent. The crystal in Fig. 1(d-i) displays planar fcc and hcp domains with a large hcp core, while the crystal in Fig. 1(d-ii) displays a fivefold hcp symmetric axis with hcp protrusions encompassing fcc domains.

The lowest energy crystals form a cluster in the data in Fig. 1(c) at the more negative end of E^*. None of these crystals contain icosahedral particles. A higher energy cluster of points centered around $E^* \approx -13.5\varepsilon$, however, include crystals, all of which contain two

to three icosahedral particles. Figure 1(d-iii) provides an example structure where three icosahedral particles introduce a trifold symmetry in the crystal structure. This motif was a common feature of the crystals in this cluster and seemed to minimize the extent to which fcc and hcp domains grow. For example, 31% of particles were identified as fcc or hcp in the lowest energy crystals, while this was 18% in the higher energy clusters, on average. A more poorly crystalline structure was found for the system where $E^* = -11\varepsilon$ and four icosahedral particles emerge in the solid, associated with a very limited propagation of the crystal lattice: 10% of particles in this system can be recognized as matching a crystal structure at the end of the trajectory, and these were nearly all hcp-like.

Different types of defects emerging consistently possibly suggest similar growth patterns for crystals in the DLDs. The force field used in this work was chosen for efficiency purposes: in order to sample multiple crystallizing trajectories, crystallization must occur over a reasonable simulation timescale. The result, however, is that even in very small crystals, relaxation of the crystal structure does not readily occur; hence, the defects are locked into the final crystal in the steady state that was sampled.

B. Crystallization CVs and VAMP

In Sec. IV A, E^* and the number of crystal-like particles evaluated using PTM were used to describe phase separation and different crystal structures. These, however, are only some of the possible CVs that can be used to monitor and describe the crystallization process. As described in Sec. III, we computed a total of 19 CVs, listed in Table I, which may provide good metrics to monitor the evolution of a crystallizing system, such as the one adopted in this work.

The concatenated CV time series obtained from 11 crystallizing trajectories are shown in Fig. S2 in the supplementary material. In addition, the CV histograms for one of the trajectories are provided in Fig. S3 in the supplementary material. The two-step nucleation process is clearly identifiable in cn.mean, ncl, Q4.mean, laQ4.mean ent, nlcust1, and ene CVs, where step changes in these variables separate time windows where the data are approximately constant within noise. Some CVs are better suited to identify crystal phases from amorphous ones, and these include ncs, ncnq4, ncnq6, q6.mean, non, fcc, and hcp CVs. Other CVs are best suited to identify condensed phases from vapor phases, such as q4.mean, Q6.mean, and laQ6.mean. Finally, CVs that do not clearly differentiate the probability distributions of states between at least two phases observed in the trajectories are bcc and ico.

In order to determine which CVs best describe the crystallization dynamics, we constructed RCs containing all possible combinations of CVs as well as those containing a single CV. The number of possible RCs is given by $2^{\max(n)} - 1$, where n is the number of CVs and, therefore, the maximum number of dimensions in any RC. For this analysis, we did not include the ico and bcc CVs; hence, the maximum n was 17, providing a total of 131 071 RCs. Figure 2(a) provides a histogram for the number of RCs according to the dimensionality of the order parameter space. Using VAMP, R_2 scores were evaluated for each of these RCs with a lag-time, $\tau = 20\,000$ simulation steps (this equates to a simulation time, $t^* = 100$). Analysis of a range of τ values indicated that the R_2 scores were relatively insensitive to the choice of τ up to around $\tau = 50\,000$. We also chose not to limit the total number of dynamic processes for the given τ; hence,

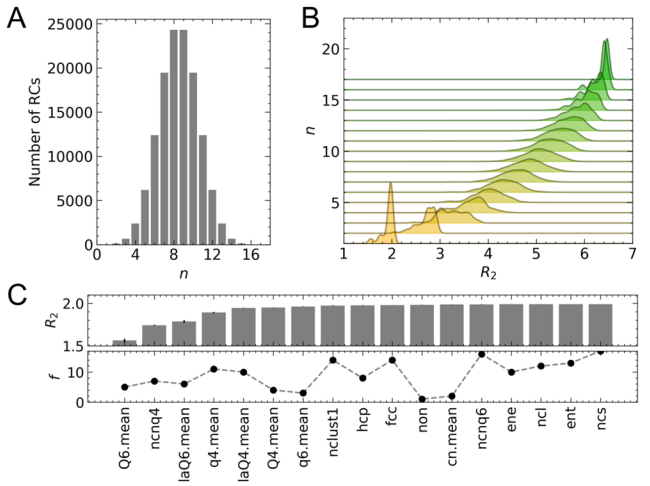

FIG. 2. (a) Total number of RCs, i.e., combinations of the CVs for every RC dimension, n, ranging from 1 to 17. (b) Probability densities for the R_2 scores for all RCs according to their dimensionality on the y axis. The distributions were evaluated using Gaussian kernel density estimation with a bandwidth of 0.05. (c) Top: R_2 scores for RCs constructed from one CV. Error bars highlight uncertainties in the scores. Bottom: The number of CV occurrences, f, in the highest ranking RCs when n = 1–17.

all eigenvalues are used to compute the VAMP-2 scores. In this analysis, the trajectory that led to a poorly crystalline solid was neglected, and the $\mathbf{f}(\mathbf{r}, t)$ arrays were constructed using absolute, normalized CV values, such that CVs range from zero to one. The distributions for these rescaled CVs from the combined trajectories are provided in Fig. S4 in the supplementary material.

Figure 2(b) provides the distributions of R_2 scores for all of the RCs. For monodimensional RCs ($n = 1$), the R_2 scores are provided for each CV in Fig. 2(c) (top panel). All R_2 scores are greater than the minimum of one, which would indicate invariant sampling of the RC. The best scoring CV is ncs with $R_2 = 1.991 \pm 0.001$; within statistical uncertainties, however, ncs, cn.mean, ncl, ncnq6, ent, and ene are equal. Not all of these CVs were identified as best suited to follow the two-step mechanism, but they all identify the emergence of a crystalline phase. Apart from cn.mean, ene, and ent, the 10 highest scoring CVs are determined by counting the number of particles according to the density or symmetry of their local coordination environment. When a monodimensional RC is used to study crystal nucleation, such as in CNT-based seeding methods,[39–41] CVs quantifying the size of the emerging phase based on local structure are adopted. Our analysis here validates that these features (i.e., total numbers of particles with solid-like first-sphere coordination numbers or particles with high local coordination symmetries reminiscent of the crystal) are good indicators for the slow dynamics of the system. This is consistent with the results from likelihood maximization (and validated using committor analysis), which identified that the best 1D reaction coordinate to study crystallization in Lennard–Jones liquid was a product of the nucleus size and the local Q_6 CVs.[42] Monodimensional RCs based on fourth-order Steinhardt parameters, as well as Q6.mean, in our work, are low-ranking indicators for crystallization; this is perhaps unsurprising in the case

of Q_4-based CVs, given that crystals display fcc and hcp particle packing.

As discussed in Sec. I, some simulation studies of crystallization adopt RCs constructed from two CVs to investigate pathways in systems where crystalline order emerges from amorphous clusters.[6,7,9–12] In such cases, the RCs characterize cluster size/density and relative cluster crystalline order in orthogonal degrees of freedom to evaluate pathways from supersaturated solutions to crystals. Provided this context, we consider 2D combinations of CVs that rank highly in the VAMP analysis. Given the 136 possible combinations of CVs used to propose a two-dimensional RC candidate here, three scored equally highly; these were {ncl, ncnq6}, {ncs, ncnq6}, and {ncnq6, ene}, where $\overline{R_c} = 2.939 \pm 0.006$. These were followed by a second tier set with $\overline{R_c} = 2.929 \pm 0.011$: {ncnq6, nclust1}, {ncl, fcc}, {cn.mean, ncnq6}, {ene, fcc}, {cn.mean, fcc}, {ncnq6, ent}, and {ent, non}.

Generally, the highest ranking 2D RCs combine CVs, one of which distinguishes well the two-step pathway and another which clearly identifies the emergence of crystalline order. It is notable that the ncnq6 variable appears in six of the ten highest scoring RCs. This is the only CV that is zero in the absence of a crystalline phase and perfectly resolves any degeneracy between disordered and ordered clusters. As in the case of the monodimensional RCs, many of the CVs listed above scale with the size of emerging phases. In our previous work on NaCl crystallization, we adopted an RC using two CVs to characterize the size of dense ion clusters and the level of crystalline order in these regions to follow crystallization where multiple pathways to crystals are evident, including those where order emerges in liquid-like intermediates.[12] The closest RC analog in this work to the one adopted previously is {ncl, ncnq6}, which is among the highest scoring set of 2D CVs and indicates that, for the specific problem at hand, a choice driven by observation and intuition was able to identify a good set of candidate CVs.

Across the entire range of n, adding more descriptors for collective particle features leads to shifting of the R_2 distributions to higher values [see Fig. 2(b)]: $\log(\tilde{R}_2) = -0.03(\log n)^2 + 0.46 \log n + 0.3$, where $\tilde{R}_2$ indicates the median, and the coefficient determining the fit is 0.99. In the case of the highest ranking RCs, R_2 converges to a maximum around 6.5 when $n = 15$–17; here, $\log(R_2) = -0.3(\log n)^2 + 0.79 \log n + 0.3$. Thus, adding more descriptors for crystallization increases the VAMP-2 score and provides RCs that more accurately capture the slow modes. Given the small increases to $\max(R_2)$, however, for large values of n, it is possible to trade off computational efficiency with accuracy to determine the kinetics for these transitions.

The best performing RCs when $n = 1$–17 tend to comprise CVs such as ncs (i.e., CVs that identify the size of crystalline regions) as well as ene and ent (see Table S2 in the supplementary material). Indeed, the highest scoring CVs in monodimensional RCs feature in the highest ranking multidimensional RCs, as shown in Fig. 2(c) (bottom). While this observation is general, there are notable exceptions in the case of cn.mean and non. This is perhaps not surprising, given that the time-dependent ncs and ncl values are highly correlated with cn.mean. Similarly, fcc and hcp time series are highly correlated with non. Generally, however, the analysis of the full spectrum of possible CV combinations identified some CVs as better than others at monitoring a two-step crystallizing system. Our analysis supports the conclusions from previous simulation studies demonstrating that CVs that better characterize the *local* symmetry in the first-coordination sphere and those that quantify the size of emerging phases are the best candidates to describe and follow the crystallization process.[35,41,43–45]

In Subsections IV C 1 and IV C 2, we further assess the performance of RCs obtained by combining different CVs by constructing Markov state models and using them to compute nucleation rates, mechanisms, and associated free energies of the relevant (meta)stable states.

C. Markov state models

With knowledge of the VAMP-2 scores, it is interesting to see how the rates for crystal nucleation compare when evaluated using RCs constructed from the highest scoring CV combinations. In this section, we build Markov state models (MSMs) for all of the highest scoring RCs for $n = 1$ to $n = 17$ to identify (meta)stable states and transitions between them. Furthermore, in order to compare the system representation across RCs with different dimensionality, we use TICA to project the high-dimension RCs onto just two coordinates that best separate the states of interest. As TICA quantifies the variance in the crystallization kinetics, the dimensionality reduction produces a reaction coordinate where the distance between (meta)stable states is a function of the system's time evolution.[27]

1. MSM from $n = 17$ RC data

We begin by discussing the general features arising from a Bayesian MSM constructed using the highest ranking RCs following the protocol below. In this case, the RC comprised 17 CVs; the best performing RCs when $n = 14$–17 had the same R_2 score within statistical uncertainties. As for the calculation of R_2 values, we used the absolute, scaled CV coordinates to build the MSM. To reduce the uncertainty in the estimate of the slowest timescales, following an initial analysis, we complemented the set of ten reactive trajectories discussed in Subsections IV A and IV B with an additional five simulations, four of which produced a DLD and another that led to a crystal with mean $E^* = -17.282 \pm 0.157\varepsilon$. The addition of these data resulted in no qualitative differences in the MSMs but did facilitate a more accurate determination of crystallization kinetics—the slowest implied timescales were 1.078–2.198×10^7 steps and 1.164–1.789×10^7 steps within a 95% confidence interval (CI) before and after including the additional simulation trajectories. Note that in what follows, we report the timescales and rates in terms of numbers of trajectory steps, with each step corresponding to $\Delta t^* = 0.005$.

First, we applied TICA to project the time-dependent 17 CV values from 15 simulation trajectories onto the $\{\phi_1, \phi_2\}$ TICA RC where $\tau = 10\,000$ steps in the evaluation of the time-lagged components. Although the calculation of VAMP-2 scores was rather insensitive to the choice of τ when $\tau < 50\,000$ steps, identifying the fastest processes in the system dynamics requires a smaller value of the lagtime (as discussed below, the implied timescales from the model shown in Fig. S6 in the supplementary material indicate that this was a reasonable choice to distinguish all of the relevant transitions). The concatenated ϕ_1 and ϕ_2 trajectories resulting from TICA are provided in Fig. S5 in the supplementary material. These indicate that ϕ_2 clearly separates different crystal states resulting

from crystallization. On the other hand, ϕ_1 shows distinct time windows where dense amorphous phases i.e., DLDs, are present.

Figure 3(a) shows the cumulative sampled probability density of states as a function of $\{\phi_1, \phi_2\}$. A small peak for states in the VP is observed at $\phi_1 \approx -1.4$, $\phi_2 \approx -1.8$, while a much more pronounced peak at $\phi_1 \approx -0.6$, $\phi_2 \approx 0.6$ accounts for microstates in the DLD. Two to three broad peaks highlighted on the plot are observed for states where crystals are present. The wide distribution of the states highlights the slow time evolution of the crystals during the simulations.

To construct the MSM, the sampled configurations were mapped onto a discrete set of partitions in the $\{\phi_1, \phi_2\}$ space using regular space clustering with a minimum distance of $\phi = 0.2$ between cluster centers. The attribution of microstates to a partition was carried out by Voronoi tessellation of the sampled data.[46] Partition centers are shown in Fig. 3(a). This procedure generates trajectories describing transitions between discrete partitions, which can be used to construct a transition matrix. We confirmed that the resulting 119 partitions were fully connected in the MSM, and all transitions between partitions were used to determine kinetic information from the fully connected network of partitions.

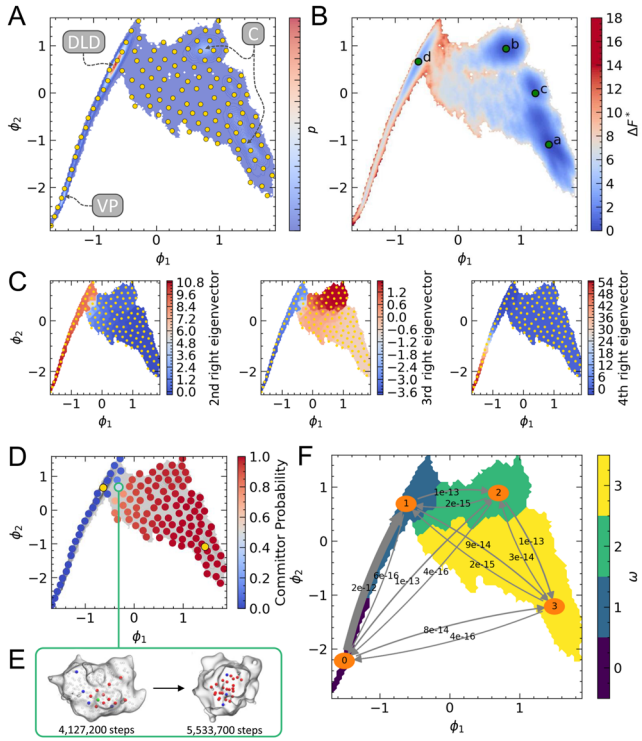

FIG. 3. Results from a Bayesian MSM constructed from 17 CVs evaluated for 15 independent simulation trajectories projected onto two TICA coordinates: ϕ_1 and ϕ_2. (a) The relative probability density (highlighted by the color scale) of the sampled states in the 2D TICA RC, with (meta)stable vapor phase (VP), dense liquid droplet (DLD), and crystal peaks (C) indicated by the arrows. The positions for 119 partition centers used in the construction of the MSMs are overlaid and shown as yellow circles. (b) Relative free energies (ΔF^*) in units of $k_B T^*$ computed from the Bayesian weighted stationary distribution projected onto the sampled states; labels a, b, c, and d identify the four lowest energy minima in the landscape when determined using the approach discussed in the text. (c) Projections of right eigenvectors 2 – 4 onto the sampled states and with partitions also highlighted. (d) Committor probabilities, highlighting the probabilities for partitions to commit to either the DLD or C_3 basins indicated by the yellow circles. (e) Snapshots of microstates from a single trajectory, which are associated with the partition in D where the committor probability is ≈0.5 (see Fig. 1 caption for a description of the representation). (f) Transitions between the four PCCA+ (meta)stable states: VP (0), DLD (1), C_2 (3), and C_4 (4). The labels indicate the rates, also provided in Table II, and the width of the arrows indicates the fastest transitions.

The probability density weights associated with partitions determine the stationary distribution of states, which can be Boltzmann-inverted to generate the free energy landscape in $\{\phi_1, \phi_2\}$, provided in Fig. 3(b). The landscape indicates a narrow reactive pathway associated with the VP to DLD transition, corresponding to the condensation process. Instead, the path from the DLD to different crystal states is less constrained in the RC space. The model accurately determines the relative stability of the different crystalline nuclei observed in simulations. Crystals with the lowest potential energy, in fact, correspond to the global minimum in the free energy landscape [see point a in Fig. 3(b)]. The two metastable states corresponding to local minima of the free energy (determined using moving 20×20 windows in a 150×150 grid of the RC space) and identified by labels b and c, also represent crystal nuclei associated with a $\Delta F^* = 0.26$ and $1.17 k_B T^*$, respectively. The difference in the free energies between a DLD and the most stable crystal, $\Delta F^* \approx 3 k_B T^*$ and, though not shown in the Figure, ΔF^* for the basin representing VP microstates is $6 k_B T^*$ (the energies are shifted so that at the global minimum, $F^* = 0$). Therefore, the ranking of relative stabilities of the VP, DLD, and crystals, which can be qualitatively inferred by observing the dynamic trajectories, is captured well by the MSM.

The eigenfunctions of the MSM approximate the transitions between (meta)stable states in the system. The eigenvalues associated with these functions determine their importance when predicting the time evolution of the system. Figure S6 in the supplementary material provides the implied timescales for the 12 slowest transitions as a function of different τ values, which are computed as $t_i = -\tau / \ln|\lambda_i(\tau)|$ and where λ_i is the eigenvalue for process i. These implied timescales all increase as a function of τ but plateau when $\tau \approx 100$ steps; indeed, this analysis was used to identify the appropriate value of τ in a series of trial and improvement cycles. Several orders of magnitude separate the implied timescales for the slowest and fastest transitions, which is not surprising given the reaction under consideration.

By projecting the right eigenfunctions with the largest eigenvalues onto the TICA RC, we can visualize the slowest modes in the system. Figure 3(c) shows that the slowest transition is from crystalline partitions to the VP; the second slowest process is one that goes from the VP to higher energy crystals via the lowest energy crystal partitions; while the third slowest process is a transition from condensed phases to a VP. It is important to note that, though we never observe these transitions in the forward reactive trajectories generated in the simulations, the construction of the MSM through the partitioning of states in energy minima and transition state regions means that we can predict these slow modes. Provided the free energy landscape in Fig. 3(b) and physical intuition, these slowest transitions are to be expected with the model assumption of ergodicity. To test the accuracy of the model pre-

dictions, we performed a Chapman–Kolmogorov test[46] using four (meta)stable states, the results for which are shown in Fig. S7 in the supplementary material. This test evaluates the left- and right-hand sides of the equation $\mathbf{T}(k\tau) = \mathbf{T}^k(\tau)$, where $\mathbf{T}$ is the transition matrix and k is the number of trajectory steps we adopt in the calculation. The results in Fig. S7 in the supplementary material indicate that the model predictions and estimates from the data are consistent.

The second and third eigenvectors in Fig. 3(b) highlight the approximate transition between amorphous and crystalline states. To explore this more accurately for the forward transition associated with the onset of crystalline order within the DLD, we performed a committor probability analysis considering the DLD and the most stable crystal minimum as end states, as shown by the yellow circles in Fig. 3(d). The partitions in the Figure are colored blue to red according to their probability of committing to the crystal basin. The transition state ensemble projection onto $\{\phi_1, \phi_2\}$ corresponds to the region of CV space approximately identifying the isocommittor. In particular, the partition highlighted by a green circle in Fig. 3(d) has a committor probability of 0.5, providing the closest approximation of the transition state (TS) associated with the crystal nucleation transition. Figure 3(e) provides snapshots of the dense phase at the beginning and end of a portion of a single trajectory crossing the TS partition, where the crystal-like particles are identified using PTM. It is clear that the number and local density of the particles with crystal-like local environments increases [highlighted by the colors in Fig. 3(e)]. Moreover, their arrangement appears to become more ordered, in line with what is expected for the second step in a two-step crystallization mechanism.

To determine the rates in two-step crystallization, we performed a spectral clustering of the partitions using Robust Perron Cluster Cluster Analysis (PCCA+)[47] to cluster partitions according to the eigenvectors of the transition matrix associated with the MSM; Fig. S8 in the supplementary material highlights the weights for each partition assignment to states. The assignment of partitions to four (meta)stable states, ω, using this approach is shown in Fig. 3(f), and the fraction of microstates associated with $\omega = 0$, 1, 2, and 3 was 0.0004, 0.0104, 0.2487, and 0.7405, respectively. These states represent the VP, DLD, C_2, and C_3 in order of increasing ω, where C_2 are crystals with higher potential energy and C_3 are the more stable crystals.

The mean first passage times (MFPTs) between (meta)stable states can be determined from the transition matrix of the MSM. Rates computed from these MFPTs and their uncertainties determined within a 95% CI are provided in Table II, which indicate that the fastest transition is the condensation of the VP to form a DLD. The fastest transitions following this are the emergence of order in the DLD to form higher energy crystals (C_2) and the transformation of C_2 to C_3 crystals. We did not observe this latter transition during the simulations, as already discussed, and so the quantitative predictions of the model here should be further tested. The slowest transitions are those already identified as transformations of crystals to the VP and DLD phases. Faster transitions occur from the VP to crystals; however, the distribution of states indicates that the system must first go via the DLD. Indeed, following the forward reaction, the MSM indicates that the crystallization pathway proceeds according to VP → DLD → C_2 → C_3, and, in general, the predictions of the MSM are consistent with the pathways and relative kinetics of transitions that are to be expected for a two-step crystallizing system.

TABLE II. Transitions between the (meta)stable states in a 2D TICA RC constructed using 17 CVs, ranked according to their rates calculated from MFPTs. 95% confidence intervals (CI) in the rates are also provided and the units are $(\sigma^3 \text{ steps})^{-1}$.

	Transition	Rate	95% CI
1	VP → DLD	1.85×10^{-12}	$1.47 - 2.36 \times 10^{-12}$
2	$C_2 \rightarrow C_3$	1.18×10^{-13}	$0.87 - 1.58 \times 10^{-13}$
3	DLD → C_2	1.13×10^{-13}	$0.85 - 1.43 \times 10^{-13}$
4	VP → C_2	9.59×10^{-14}	$0.77 - 1.16 \times 10^{-13}$
5	DLD → C_3	8.58×10^{-14}	$0.68 - 1.05 \times 10^{-13}$
6	VP → C_3	7.55×10^{-14}	$6.12 - 9.07 \times 10^{-14}$
7	$C_3 \rightarrow C_2$	3.41×10^{-14}	$2.47 - 5.26 \times 10^{-14}$
8	C_2 → DLD	1.77×10^{-15}	$1.08 - 2.67 \times 10^{-15}$
9	C_3 → DLD	1.72×10^{-15}	$1.06 - 2.57 \times 10^{-15}$
10	DLD → VP	5.92×10^{-16}	$3.76 - 9.22 \times 10^{-16}$
11	C_2 → VP	3.98×10^{-16}	$2.64 - 5.75 \times 10^{-16}$
12	C_3 → VP	3.96×10^{-16}	$2.62 - 5.71 \times 10^{-16}$

2. MSMs for n = 1–17 RCs

The approach laid out above for $n = 17$ can be applied to describe the mechanisms and compute the crystallization rates with other combinations of CVs. Therefore, we constructed MSMs for all of the $n = 2$–16 highest R_2 scoring CV combinations. To ensure a fair comparison of model results, minimal changes were made during the construction of MSMs; hence, we first projected the CVs onto two TICA coordinates using $\tau = 10\,000$ steps and computed the stationary distributions and transition matrices by sampling discrete partitions in a TICA 2D RC space. As before, we ensured that the value of τ and the partitioning of states led to converged implied timescales for the slowest modes, along with fully connected partitions, as well as model predictions for transitions that were consistent with sampled data in Chapman–Kolmogorov tests of the constructed MSMs.

In general, the MSMs when $n = 2$–16 identified the same qualitative and quantitative features identified and discussed for $n = 17$. Some notable differences were that the relative free energy differences between the (meta)stable states fluctuate within $\sim 2k_BT^*$, particularly for smaller values of n. Nevertheless, when $n = 6$–17, ΔF^* for the DLD to crystal transition converges to $-3.1 \pm 0.3k_BT^*$. For our purposes, however, amorphous phases were always higher in energy than crystalline states, and the forward reaction, i.e., VP → crystal, was always predicted to be significantly faster than the reverse reaction in all of the MSMs. In addition, while there was some reordering of the implied timescales for state-to-state transitions, the relative ranking of the transitions involved in the formation of crystals from the VP was consistent throughout, regardless of the choice of CVs used to construct the RC. This is a good sign of the robustness of the kinetic models to capture the slow transitions with reasonable choices for the CVs that can describe the time-dependent structural evolution of the system.

Figure S9 in the supplementary material provides the state maps evaluated when $n = 2$–17 from each Bayesian MSM and

PCCA + to identify the (meta)stable states. These reflect the small change in the assignment of partitions to (meta)stable states and the order of state-to-state transitions when $n > 5$. As n decreases, the extent of ϕ_1 and ϕ_2 in the 2D TICA RC tends to increase, resulting in more partitions of the sampled data. Despite this, the fraction of partitions assigned to the VP state decreased. It was necessary to increase the minimum distance between partitions from $\phi = 0.2$ to $\phi = 0.22$ in the regularly spaced clustering algorithm when $n = 2$–5. Also, when there is a reduction in the number of partitions for the VP phase, the expansion of the TICA coordinate range is concomitant with a broadening of the crystal state regions (see the relative areas for ω_0/ω_1 and ω_2/ω_3 in Fig. S9 in the supplementary material). Another interesting feature is that when $n \leq 4$, Fig. S9 in the supplementary material shows that the TICA projection of states is reflected in ϕ_2 such that VP microstates are found when ϕ_2 is at positive values, unlike in Fig. 3.

In the case of $n = 2$, the area in the state map for the VP and DLD is very small, and the number of partitions representing these states is substantially decreased cf. $n = 17$ (though the fraction of states in amorphous phases remains constant at around 0.011). Due to the small number of partitions, particularly in the VP region, spectral clustering can capture only three (meta)stable states, where ω_0, in this case, includes all of the non-crystalline microstates. This model was constructed using a 2D TICA projection of {ncs, ncnq6} where both CVs are designed to identify the emergence of crystals.

Equally highly scoring R_2 2D RCs were {ncl, ncnq6} and {ene, ncnq6}; hence, we constructed Bayesian MSMs using 2D TICA projections of these sampled CVs. As shown in Fig. S10 in the supplementary material, the RCs constructed from ncl and ene (along with ncnq6) are very similar to those constructed from the highest scoring CV pair. Applying PCCA+ to the partitions, however, does result in four (meta)stable states with a separation of the VP and DLD, and with a forward transition between amorphous states on the order of 1×10^{12} $(\sigma^3\text{steps})^{-1}$, consistent with MSMs built from RCs capturing greater numbers of degrees of freedom. Here, the rates for crystal nucleation in the DLD were 0.98–1.87 $\times 10^{13}$ and $1.13 - 1.89 \times 10^{13}$ $(\sigma^3\text{steps})^{-1}$, respectively, with 95% statistical confidence in the values, compared to the rate of $0.78 - 1.39 \times 10^{13}$ $(\sigma^3\text{steps})^{-1}$ predicted in the case of {ncs, ncnq6}. Although there is overlap in the rate predictions, {ncs, ncnq6} results in a slower nucleation rate.

It is possible to compute a VAMP-2 score for the time series of TICA coordinates that we label R_2^{TICA}. The R_2^{TICA} for the three reaction coordinates {ncl, ncnq6}, {ncs, ncnq6}, and {ene, ncnq6}, were 1.964 ± 0.002, 1.961 ± 0.005, and 1.96 ± 0.002, respectively, reflecting the earlier observation that all CV combinations are able to capture the slow variations in the underlying system dynamics. Despite this, the mechanistic insight provided by the three models differs, and this is somewhat sensitive to the method used to discretize trajectories and identify (meta)stable states. Care should be taken, therefore, when assessing model outcomes.

In the case of $n = 1$—ncs provided the highest R_2 scoring CV—only one TICA coordinate (ϕ_1) was used to construct a Bayesian MSM. For this reason, we used a minimum distance $\phi_1 = 0.04$ to generate partitions during regular space clustering, resulting in 59 partitions. The probability density of states in ϕ_1 is provided in Fig. 4(a), with the VP, DLD, C_2, and C_3 crystal states clearly apparent in the TICA projection of ncs. Figure 4(b) provides

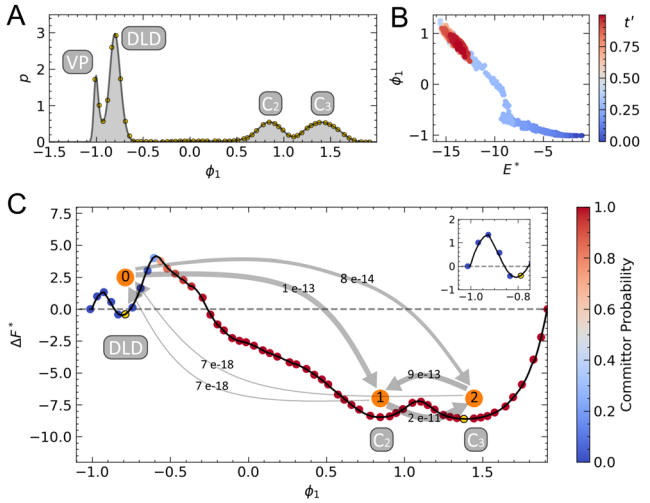

FIG. 4. Results from a Bayesian MSM constructed from the ncs CV evaluated for 15 independent simulation trajectories and projected onto a TICA coordinate, ϕ_1. (a) Probability densities for ϕ_1 states with partition centers used in the MSM indicated by the yellow circles. Peaks indicate the (meta)stable states that are labeled. (b) A single trajectory plotted as a function of ene and ϕ_1; the color scale indicates the scaled simulation time. (c) Relative free energy (in units of $k_B T^*$) as a function of ϕ_1 with the inset frame showing ΔF^* at small values of ϕ_1. The circles indicate the partition centers, colored according to their probability to commit to the yellow partition center in the C_3 minimum from the DLD minimum. States identified using PCCA+ are labeled 0, 1, and 2, with arrows between the states indicating the rate for transitions computed using MFPTs.

the time-dependent trace in {ene, ϕ_1} for a single crystallizing trajectory, highlighting how the TICA coordinate values are correlated with CV values. In the case of ene, there is a clear non-linearity in the data, as was also observed, e.g., for {ncnq6, ϕ_1} and {ent, ϕ_1}, while {ncs, ϕ_1} shows a near-perfect linear correlation in the coordinates. In all of these high R_2 scoring CVs, the distribution of TICA coordinates clearly distinguishes crystal and non-crystalline microstates.

Figure 4(c) provides the free energy profile, determined from a 1D MSM using ncs data, aligned such that $\Delta F^* = 0$ for the vapor. A small energy barrier separates the VP from the DLD, while a more pronounced energy barrier separates the DLD from the crystalline states. In the latter, there is good agreement between the position of the maximum in ΔF^* and the partition committor probabilities to commit to either the DLD or C_3. It is clear that the choice of CV affects the relative weights associated with states and, therefore, their ΔF^* values, since this is a function of the stationary distribution computed using Bayesian MSM weights. For example, a 1D MSM constructed using a 1D TICA reduction of the ncl CV data provides a free energy profile shown in Fig. S11 in the supplementary material; here, ΔF^* between the minimum representing the VP and DLD is around $2.5 k_B T^*$, compared with $\sim 0.5 k_B T^*$ from Fig. 4(c). Qualitatively, the order in the stability of the VP, DLD, C_2, and C_3 is consistent across the entire $n = 1$–17 range, but the ΔF^* values for DLD → crystal change from ~1 to $\sim 8 k_B T^*$ depending on the choice of CVs and the level of reduction in the dimensionality. Despite this,

the free energy difference between the VP and crystals is approximately consistent with the ΔF^* when additional CVs are included in the constriction of MSMs ($\Delta F^* \approx -8k_BT^*$ in the monodimensional RC for the forward reaction, compared to $\Delta F^* = -5.8 \pm 0.5k_BT^*$ when $n = 6$–17).

Where the 1D MSM does provide consistent quantitative information with MSMs constructed using additional CV dimensions, is in the overall crystallization rate. When the highest scoring CVs were used to construct the MSM, we found that the transition rate from the DLD to crystals was around $1 \times 10^{-13}(\sigma^3\text{steps})^{-1}$. As for the $n = 2$ case, spectral clustering provided only one amorphous state, centered in the DLD and marked by 0 in Fig. 4(c), to determine state-to-state transitions—this was a general observation for MSMs constructed for all of the high scoring 1D RCs. The rates indicate that the $C_2 \rightarrow C_3$ transition is the fastest between the three identified (meta)stable states.

A 1D representation of the free energy pathways to crystals from the supersaturated vapor phase demonstrates how the picture for crystal nucleation differs from one expected for a single-step transformation of the vapor to a crystal following established nucleation theories based on the earliest ideas of Gibbs.[5] The two energy barriers may be perceived as a clear departure from a CNT-based model for phase separation; however, one can interpret the two barriers as two distinct steps, each of which can be reasonably well described using CNT-based theories. The interfacial tension used to predict the crystal nucleation barrier in CNT must account for the crystal lattice's emergence in a DLD. This phase separation process is consistent with Ostwald's rule of stages,[48] where the first product from nucleation is a thermodynamic phase with chemical potential closest to the parent phase, and which subsequently undergoes further transformations to more stable states. Furthermore, the pathway is distinct from those where amorphous intermediates do not occupy states representing a depression in the free energy landscape.[8]

Multi-step crystallization pathways are known experimentally for colloidal systems,[44,49] and pathways to crystals via amorphous intermediates were reported for other crystallizing systems.[38,50] These pathways may also include intermediate crystal phases; however, we believe that the different crystal minima in our work are the result of the stacking faults already discussed and not thermodynamically distinct phases at equilibrium, which are a feature of, e.g., binary colloid mixtures.[43,45]

In all of the MSMs constructed, the slowest process to crystallization is the second step, i.e., the emergence of order in the liquid. Figure 5(a) provides the nucleation rates for this step, J, which are roughly constant as a function of n. Assuming that the highest dimension CV description of the crystallization dynamics is the best choice to predict the kinetics as indicated by the higher VAMP-2 score, the most significant departure in the mean rates computed using MFPTs occurs when $n \approx 9$–13. However, the clear overlap of the 95% confidence intervals of the rate estimates allows us to confidently determine the rates within the same order of magnitude, indicating that all of the TICA RCs, constructed from a basis of high-scoring CVs of increasing dimensionality, predict consistent nucleation times.

Figure 5(b) shows how DLD → crystal nucleation rates change as a function of the R_2^{TICA} scores. The TICA RCs that have a smaller R_2^{TICA} value are in the range $n = 9$–13, where J is higher,

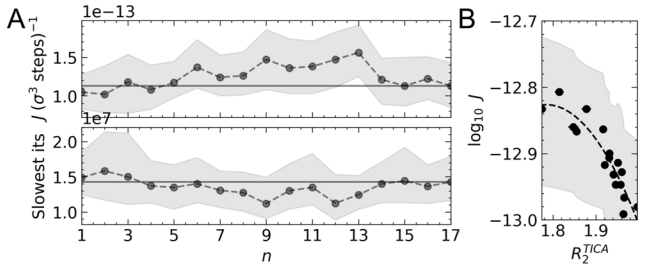

FIG. 5. (a) Top: Crystal nucleation rate, J, determined from the MFPT for transitions from an amorphous phase to a crystalline one in the Bayesian MSMs as a function of n, the number of CVs used to generate the TICA trajectories. Bottom: The slowest implied timescales from the MSM as a function of n. Solid lines mark the mean data for the $n = 17$ case. (b) Logarithm of the rates in A are plotted against the R_2^{TICA} score for the TICA RC. The dashed line is a fit to the data. Shaded regions indicate a 95% CI in the mean points throughout.

and the slowest implied timescales in the model [see Fig. 5(a)] show a departure from the solid line marking the mean values for $n = 17$, while the highest R_2^{TICA} score was for $n = 1$. The dashed line in the Figure is a fit to the data with functional form $\log_{10}(J) = -3.89(R_2^{TICA})^2 + 13.85(R_2^{TICA}) + 25.18$. This indicates that the RCs, which best capture the slowest dynamics in the system, also predict slower mean rates for the nucleation of crystals in the DLD. It is important to reiterate, however, that the uncertainties mean that the crystallization rates are predicted consistently in the MSMs, regardless of the CVs chosen to characterize the process and the projection of these onto their time-lagged independent components.

V. CONCLUSIONS

VAMP analysis of the CV time series data from crystallizing trajectories indicates that CVs characterizing the size of emerging phases in the system often feature in the highest scoring CV combinations that best describe the slow dynamics for crystallization. These CVs, along with, e.g., system potential energy and the configurational entropy, often feature in high VAMP-2 scoring crystallization RCs, provide validation that the characterization of these processes, often adopted in simulations,[2,4,6,11,12,41] provide good CVs to reduce the high-dimension configuration space to a handful of relevant degrees of freedom and extract kinetic information. In more complex systems, it may be necessary to incorporate additional CVs into the RC to describe how, for example, non-spherical monomers (perhaps with internal degrees of freedom), explicit solvent, and impurities/additives affect nucleation. As there is no standard procedure to choose the best CVs to gain thermodynamic and kinetic information, trials of suitable functions to define (collective) molecular features must be performed. CV accuracy can be affirmed using the analyses described in this work and elsewhere.[3,4] Generally, the distribution of CV values representing the reactant, product, and any intermediate states in a multi-step reaction pathway must be clearly distinguishable in CV space; hence, a multi-modal probability density of states should be apparent in the reaction coordinate. This is no guarantee, however, that the reaction coordinate is a good one to determine mechanisms and rates.

The fact that the nucleation rates for the emergence of crystalline order in dense liquid intermediates are consistent regardless of the number of CVs used to construct MSMs and determine timescales for these transitions, is a testament to the robustness of kinetic models constructed from CV combinations with a high VAMP-2 score. A general conclusion from our analyses is that despite kinetic information being remarkably consistent in the MSMs constructed using TICA projections of n = 1–17 CVs, quantitative thermodynamic information *and* mechanistic insight are only accurately gained when a sufficiently large number of CVs are considered.

From the majority of the MSMs constructed in this work, we were able to identify a crystallization pathway progressing from the vapor phase to crystals via a dense liquid intermediate, with committor probabilities, spectral analysis, and stationary distributions all indicating two bottlenecks to the formation of crystals: the condensation of the vapor to the liquid and rearrangement of particles in the liquid to form a crystal lattice, with the latter representing the rate-determining step. Each of these two steps could, in principle, be described using their respective thermodynamic driving forces for nucleation, which form the basis of CNT. However, while a straightforward application of Gibbs' theory for nucleation might be possible to characterize the first step, the capillary approximation is likely to fail for crystal nucleation in the liquid, where we observed a population of nuclei with different defect densities and local crystalline arrangements.

The model agreement across the range of dimensionalities (n = 1–17) of CV spaces, and particularly the consistent prediction of nucleation rates, is a remarkable result that highlights the value of selecting combinations of crystallization CVs that, for every n, maximize the VAMP-2 score. We believe this approach is general and sufficiently transferable to support the study of other crystallization or dissolution processes (where these events can be observed within reasonable simulation timescales) or to guide the choice of CVs used in enhanced sampling simulations of nucleation processes.

SUPPLEMENTARY MATERIAL

See the supplementary material for a detailed description of the collective variables (S1), the phase diagram of the colloidal system studied (S2), a summary of the highest scoring CV combinations (S3), and additional figures (S4).

ACKNOWLEDGMENTS

The authors acknowledge funding from an EPSRC Program Grant (Grant No. EP/R018820/1) that funds the Crystallization in the Real World consortium. The authors thank members of the consortium for useful discussions. The authors acknowledge the use of the UCL Myriad High Throughput Computing Facility (Myriad@UCL), and associated support services, in the completion of this work.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

A. R. Finney: Conceptualization (equal); Data curation (lead); Formal analysis (equal); Investigation (lead); Methodology (equal); Software (lead); Visualization (lead); Writing – original draft (lead); Writing – review & editing (equal). **M. Salvalaglio**: Conceptualization (equal); Data curation (supporting); Formal analysis (equal); Funding acquisition (lead); Investigation (supporting); Methodology (equal); Project administration (lead); Resources (lead); Software (supporting); Supervision (lead); Visualization (supporting); Writing – original draft (supporting); Writing – review & editing (equal).

DATA AVAILABILITY

LAMMPS and PLUMED input files, shell scripts used to automate data generation, and interactive Python notebooks used in the analyses are available for download at https://github.com/aaronrfinney/VAMP-MSM. PLUMED input files used in this work are also available via PLUMED-NEST (https://www.plumed-nest.org [51]), the public repository for the PLUMED consortium, using the project ID: plumID:22.044.

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

REFERENCES

[1] J. W. Mullin, *Crystallization*, 4th ed. (Butterworth-Heinemann, 2001).
[2] G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen, and A. Michaelides, "Crystal nucleation in liquids: Open questions and future challenges in molecular dynamics simulations," Chem. Rev. **116**, 7078–7116 (2016).
[3] B. Peters, "Reaction coordinates and mechanistic hypothesis tests," Annu. Rev. Phys. Chem. **67**, 669–690 (2016).
[4] 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**, 040901 (2021).
[5] D. Kashchiev, *Nucleation: Basic Theory with Applications* (Butterworth Heinemann, 2000).
[6] P. R. t. Wolde and D. Frenkel, "Enhancement of protein crystal nucleation by critical density fluctuations," Science **277**, 1975–1978 (1997).
[7] H. Jiang, P. G. Debenedetti, and A. Z. Panagiotopoulos, "Nucleation in aqueous NaCl solutions shifts from 1-step to 2-step mechanism on crossing the spinodal," J. Chem. Phys. **150**, 124502 (2019).
[8] D. Kashchiev, "Classical nucleation theory approach to two-step nucleation of crystals," J. Cryst. Growth **530**, 125300 (2020).
[9] P. S. Bulutoglu, S. Wang, M. Boukerche, N. K. Nere, D. S. Corti, and D. Ramkrishna, "An investigation of the kinetics and thermodynamics of NaCl nucleation through composite clusters," PNAS Nexus **1**(2), 1–11 (2022).
[10] M. Salvalaglio, M. Mazzotti, and M. Parrinello, "Urea homogeneous nucleation mechanism is solvent dependent," Faraday Discuss. **179**, 291–307 (2015).
[11] M. Salvalaglio, C. Perego, F. Giberti, M. Mazzotti, and M. Parrinello, "Molecular-dynamics simulations of urea nucleation from aqueous solution," Proc. Natl. Acad. Sci. U. S. A. **112**, E6–E14 (2015).
[12] A. R. Finney and M. Salvalaglio, "Multiple pathways in NaCl homogeneous crystal nucleation," Faraday Discuss. **235**, 56–80 (2022).
[13] N. E. R. Zimmermann, B. Vorselaars, J. R. Espinosa, D. Quigley, W. R. Smith, E. Sanz, C. Vega, and B. Peters, "NaCl nucleation from brine in seeded simulations: Sources of uncertainty in rate estimates," J. Chem. Phys. **148**, 222838 (2018).
[14] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, "Bond-orientational order in liquids and glasses," Phys. Rev. B **28**, 784–805 (1983).
[15] P. M. Larsen, S. Schmidt, and J. Schiøtz, "Robust structural identification via polyhedral template matching," Modell. Simul. Mater. Sci. Eng. **24**, 055007 (2016).

[16]N. F. Francia, L. S. Price, J. Nyman, S. L. Price, and M. Salvalaglio, "Systematic finite-temperature reduction of crystal energy landscapes," Cryst. Growth Des. **20**, 6847–6862 (2020).
[17]B. Peters and B. L. Trout, "Obtaining reaction coordinates by likelihood maximization," J. Chem. Phys. **125**, 054108 (2006).
[18]P. L. Geissler, C. Dellago, and D. Chandler, "Kinetic pathways of ion pair dissociation in water," J. Phys. Chem. B **103**, 3706–3710 (1999).
[19]H. Wu and F. Noé, "Variational approach for learning Markov processes from time series data," J. Nonlinear Sci. **30**, 23–66 (2020).
[20]F. Noé and F. Nüske, "A variational approach to modeling slow processes in stochastic dynamical systems," Multiscale Model. Simul. **11**, 635–655 (2013).
[21]Y.-Y. Zhang, H. Niu, G. Piccini, D. Mendels, and M. Parrinello, "Improving collective variables: The case of crystallization," J. Chem. Phys. **150**, 094509 (2019).
[22]A. Mardt, L. Pasquali, H. Wu, and F. Noé, "VAMPnets for deep learning of molecular kinetics," Nat. Commun. **9**, 5 (2018).
[23]*An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, edited by G. R. Bowman, V. S. Pande, and F. Noé (Springer, 2013).
[24]F. Noé and C. Clementi, "Kinetic distance and kinetic maps from molecular dynamics simulation," J. Chem. Theory Comput. **11**, 5002–5011 (2015).
[25]R. T. McGibbon and V. S. Pande, "Variational cross-validation of slow dynamical modes in molecular kinetics," J. Chem. Phys. **142**, 124105 (2015).
[26]C. R. Schwantes and V. S. Pande, "Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9," J. Chem. Theory Comput. **9**, 2000–2009 (2013).
[27]G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction," J. Chem. Phys. **139**, 015102 (2013).
[28]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, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton, "LAMMPS—A flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," Comput. Phys. Commun. **271**, 108171 (2022).
[29]R. Everaers and M. R. Ejtehadi, "Interaction potentials for soft and hard ellipsoids," Phys. Rev. E **67**, 041710 (2003).
[30]S. A. Safran, *Statistical Thermodynamics of Surfaces, Interfaces, and Membranes* (CRC Press, 2018).
[31]G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, "PLUMED 2: New feathers for an old bird," Comput. Phys. Commun. **185**, 604–613 (2014).
[32]M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, "PyEMMA 2: A software package for estimation, validation, and analysis of Markov models," J. Chem. Theory Comput. **11**, 5525–5542 (2015).
[33]M. Hoffmann, M. Scherer, T. Hempel, A. Mardt, B. de Silva, B. E. Husic, S. Klus, H. Wu, N. Kutz, S. L. Brunton, and F. Noé, "Deeptime: A Python library for machine learning dynamical models from time series data," Mach. Learn.: Sci. Technol. **3**, 015009 (2022).
[34]G. A. Tribello, F. Giberti, G. C. Sosso, M. Salvalaglio, and M. Parrinello, "Analyzing and driving cluster formation in atomistic simulations," J. Chem. Theory Comput. **13**, 1317–1327 (2017).
[35]W. Lechner and C. Dellago, "Accurate determination of crystal structures based on averaged local bond order parameters," J. Chem. Phys. **129**, 114707 (2008).
[36]P. M. Piaggi, O. Valsson, and M. Parrinello, "Enhancing entropy and enthalpy fluctuations to drive crystallization in atomistic simulations," Phys. Rev. Lett. **119**, 015701 (2017).
[37]A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," Anal. Chem. **36**, 1627–1639 (1964).
[38]J. De Yoreo, "A perspective on multistep pathways of nucleation," in *Crystallization via Nonclassical Pathways Volume 1: Nucleation, Assembly, Observation, and Application*, edited by X. Zhang (ACS Publications, 2020), Vol. 1358, pp. 1–17; available at https://pubs.acs.org/doi/book/10.1021/bk-2020-1358.
[39]B. C. Knott, V. Molinero, M. F. Doherty, and B. Peters, "Homogeneous nucleation of methane hydrates: Unrealistic under realistic conditions," J. Am. Chem. Soc. **134**, 19544–19547 (2012).
[40]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**, 15008–15017 (2013).
[41]N. E. R. Zimmermann, B. Vorselaars, D. Quigley, and B. Peters, "Nucleation of NaCl from aqueous solution: Critical sizes, ion-attachment kinetics, and rates," J. Am. Chem. Soc. **137**, 13352–13361 (2015).
[42]G. T. Beckham and B. Peters, "Optimizing nucleus size metrics for liquid–solid nucleation from transition paths of near-nanosecond duration," J. Phys. Chem. Lett. **2**, 1133–1138 (2011).
[43]E. Pretti, H. Zerze, M. Song, Y. Ding, R. Mao, and J. Mittal, "Size-dependent thermodynamic structural selection in colloidal crystallization," Sci. Adv. **5**, eaaw5912 (2019).
[44]J. R. Savage and A. D. Dinsmore, "Experimental evidence for two-step nucleation in colloidal crystallization," Phys. Rev. Lett. **102**, 198302 (2009).
[45]H. Fang, M. F. Hagan, and W. B. Rogers, "Two-step crystallization and solid–solid transitions in binary colloidal mixtures," Proc. Natl. Acad. Sci. U. S. A. **117**, 27927–27933 (2020).
[46]J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation," J. Chem. Phys. **134**, 174105 (2011).
[47]S. Röblitz and M. Weber, "Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification," Adv. Data Anal. Classif. **7**, 147–179 (2013).
[48]W. Ostwald, "Studien über die bildung und umwandlung fester körper," Z. Phys. Chem. **22U**, 289–330 (1897).
[49]T. H. Zhang and X. Y. Liu, "How does a transient amorphous precursor template crystallization," J. Am. Chem. Soc. **129**, 13520–13526 (2007).
[50]P. G. Vekilov, "The two-step mechanism of nucleation of crystals in solution," Nanoscale **2**, 2346 (2010).
[51]M. Bonomi, G. Bussi, C. Camilloni, G. Tribello, P. Banáš, A. Barducci, M. Bernetti, P. Bolhuis, S. Bottaro, D. Branduardi, R. Capelli, P. Carloni, M. Ceriotti, A. Cesari, H. Chen, W. Chen, F. Colizzi, S. De, M. De La Pierre, D. Donadio, V. Drobot, B. Ensing, A. Ferguson, M. Filizola, J. Fraser, H. Fu, P. Gasparotto, F. Gervasio, F. Giberti, A. Gil-Ley, T. Giorgino, G. Heller, G. Hocky, M. Iannuzzi, M. Invernizzi, K. Jelfs, A. Jussupow, E. Kirilin, A. Laio, V. Limongelli, K. Lindorff-Larsen, T. Löhr, F. Marinelli, L. Martin-Samos, M. Masetti, R. Meyer, A. Michaelides, C. Molteni, T. Morishita, M. Nava, C. Paissoni, E. Papaleo, M. Parrinello, J. Pfaendtner, P. Piaggi, G. Piccini, A. Pietropaolo, F. Pietrucci, S. Pipolo, D. Provasi, D. Quigley, P. Raiteri, S. Raniolo, J. Rydzewski, M. Salvalaglio, G. Sosso, V. Spiwok, J. Šponer, D. Swenson, P. Tiwary, O. Valsson, M. Vendruscolo, G. Voth, and A. White, "Promoting transparency and reproducibility in enhanced molecular simulations," Nat. Methods **16**, 670–673 (2019).