Bridging the gap between atomistic and macroscopic models of homogeneous nucleation

Bingqing Cheng[1, a)] and Michele Ceriotti[1]
Laboratory of Computational Science and Modeling, Institute of Materials,
École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

Macroscopic theories of nucleation such as classical nucleation theory envision that clusters of the bulk stable phase form inside the bulk metastable phase. Molecular dynamics simulations are often used to elucidate nucleation mechanisms, by capturing the microscopic configurations of all the atoms. In this paper, we introduce a thermodynamic model that links macroscopic theories and atomic-scale simulations and thus provide a simple and elegant framework for testing the limits of classical nucleation theory.

Keywords: nucleation, atomistic simulation, thermodynamics, statistical mechanics
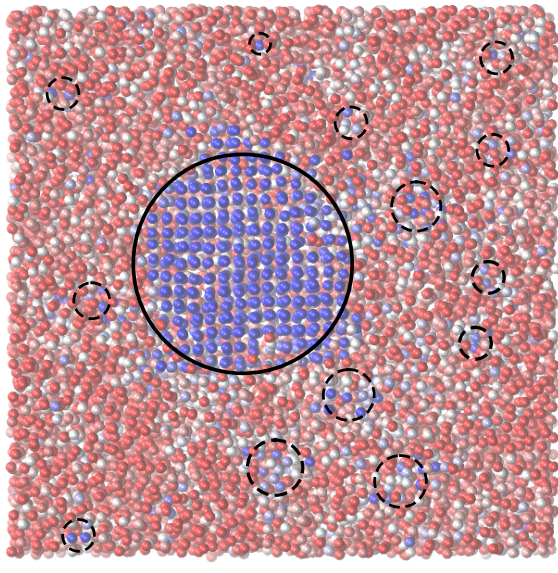
FIG. 1. A snapshot of an under-cooled liquid system that has a large solid cluster (solid circle), and many smaller clusters (dashed circles). Atoms are colored based on the value of a local order parameter so solid-like atoms are colored in blue, while liquid-like atoms are colored in red. Details on the underlying simulation are given in Section II.

Nucleation is a key step in bulk phase transitions[1–4]. This process plays a crucial role in natural phenomena and in technological applications, from the formation of clouds[4] to self-assembly[5], and from casting to the growth of thin films[6,7]. One of the simplest models to rationalize nucleation is classical nucleation theory (CNT), which assumes that the stable phase forms by accretion of nanoscopic nuclei. These clusters are unstable when they are smaller than a critical size $n^\star$, and at any given time the metastable phase contains multiple sub-critical clusters of the stable phase (Figure 1). The average number of clusters containing n atoms in the system is given by:

$$\langle p_n \rangle \propto \exp(-\beta G(n)), \qquad (1)$$

where $\beta = 1/k_B T$, and $G(n)$ is the free energy excess associated with a single cluster of size n. In the context of homogeneous nucleation, CNT further assumes that $G(n)$ can be expressed as the sum of a bulk and a surface term, i.e.

$$G(n) = \mu n + \sigma v^{\frac{2}{3}} n^{\frac{2}{3}}, \qquad (2)$$

where μ is the chemical potential difference between the stable and the metastable phases, σ is the effective interfacial free energy, and v is the molar volume of the bulk stable phase.

Investigating experimentally the nature and behavior of the unstable subcritical nuclei is extremely difficult. Therefore, in the last two decades, a consider-
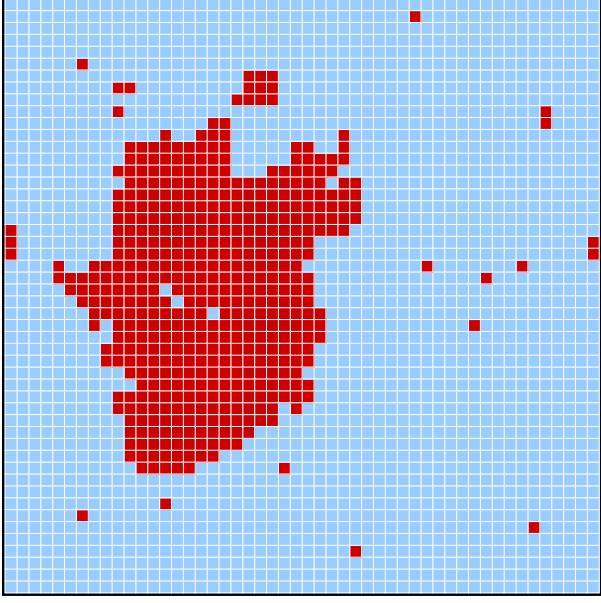
FIG. 2. A snapshot of a two-dimensional square-lattice Ising model undergoing a transition between "up" (blue) and "down" (red) ferromagnetic phases. Details on the underlying simulation are given in Section II.
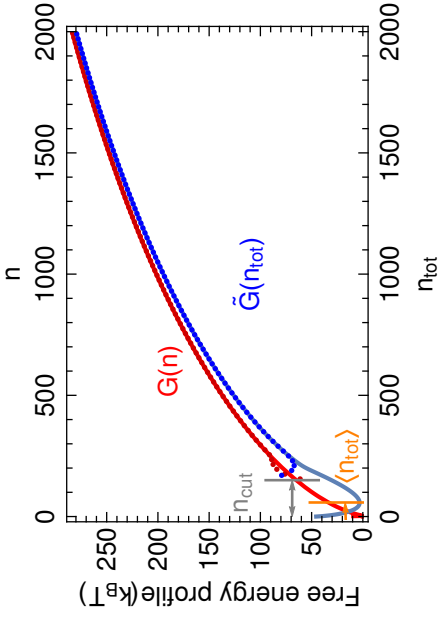

FIG. 3. The red line represents the free energy profile $G(n)$ of a single cluster. The blue line shows the exact $\tilde{G}(n_{\text{tot}})$ of the atomistic system with multiple clusters. The red dots and the blue dots indicate $G(n)$ and $\tilde{G}(n_{\text{tot}})$ that are approximated using Eq. (5), respectively. The grey and yellow vertical lines indicate n_{cut} and $\langle n_{\text{tot}}\rangle$, respectively.

macroscopic theories of nucleation. To achieve this, we first investigate a multiple cluster model that we use as a proxy for an idealized atomistic system. Then, we develop a thermodynamic framework that is consistent with the multiple cluster model, requires fewer assumptions, and is fully applicable to the atomistic systems simulated in molecular dynamics or Monte Carlo studies. For the sake of clarity, we will develop our theoretical framework making reference to the case of solidification from the melt (Figure 1), but our results are general, and we present an application to a two-dimensional Ising model (Figure 2) to demonstrate that it can be applied to all sorts of activated phase transition processes.

I. A THERMODYNAMIC MODEL OF ATOMIC-SCALE NUCLEATION EVENTS

A. An idealized multiple-cluster model

where $G(n)$ is the free energy of a single cluster of size n, and N_s is the number of nucleation sites. N_s is proportional to the total number of particles in the system, and guarantees the appropriate scaling with system size. $G(n)$ can take any form as long as it is a monotonically increasing function for n smaller than the size of the critical nucleus, but for this idealized system we use Eqn. (2) with the parameters reported in the Supplementary Material. The free energy profile for this equation with these parameters is shown in red in Figure 3.

As we have assumed that the cluster size distribution follows Eqn. (3), we can derive the following expression for the free energy profile of the whole system as a function of the *total* number of solid atoms, $\tilde{G}(n_{\text{tot}})$,

$$e^{-\beta\tilde{G}(n_{\text{tot}})} = \sum_{p_1=0}^{\infty}\sum_{p_2=0}^{\infty}\cdots\sum_{p_{n^\star}=0}^{\infty}\delta\left(\sum_{n=1}^{\infty}np_n - n_{\text{tot}}\right)\prod_{n=1}^{n^\star}P(n,p_n), \tag{4}$$

the most probable configuration for the system was composed of several small clusters. However, as n_{tot} gets larger, it typically contained one large solid cluster accompanied by many smaller ones. As we discuss in detail in the Supplementary Material, one can thus define a cutoff size n_{cut}, such that for $n_{\text{tot}} \gg n_{\text{cut}}$ it is orders of magnitude more likely to have precisely one cluster with size $n > n_{\text{cut}}$ than to have several or none of such large clusters. At $n_{\text{tot}} \gg n_{\text{cut}}$, the largest cluster can be interpreted as a standalone solid cluster with n atoms, associated with a probability $P(n, 1)$. The rest of clusters in the background can be treated as a separate bulk liquid system that follows the same distributions (Eqn. (3)) as the original whole system. Under such treatment, at $n_{\text{tot}} \gg n_{\text{cut}}$ the expression for $\tilde{G}(n_{\text{tot}})$ can be simplified tremendously as

$$\exp(-\beta\tilde{G}(n_{\text{tot}})) = \sum_{n=n_{\text{cut}}}^{n_{\text{tot}}} N_s e^{-\beta G(n)} e^{-\beta\tilde{G}(n_{\text{tot}}-n)},$$ (5)

considering also that $P(n, 1) \approx N_s\lambda(n)$ for large n. The blue dots in Figure 3 correspond to the approximate $\tilde{G}(n_{\text{tot}})$ that can be computed using Eqn. (5). These points overlap perfectly with the exact values at $n_{\text{tot}} \gtrsim n_{\text{cut}} + \langle n_{\text{tot}} \rangle$. What is more, Eqn. (5) suggests that there is a one-to-one mapping between the $\tilde{G}(n_{\text{tot}})$ for the whole system and the $G(n)$ for a single cluster with $n \gtrsim n_{\text{cut}}$. As such, Eq. (5) allows us to calculate $G(n)$ from a knowledge of $\tilde{G}(n_{\text{tot}})$, without any information on the sizes of individual clusters in each snapshot. The $G(n)$ that is reconstructed from Eq. (5) using a fixed-point numerical scheme is indicated using red dots in Figure 3, and overlaps perfectly with the exact $G(n)$.

B. A probabilistic definition of Gibbs dividing surfaces

Our discussion thus far demonstrates that the average distribution of cluster sizes can be extracted from the distribution of the total number of atoms assigned to the stable phase. Unfortunately, in actual atomistic simulations it is impossible to assign individual atoms or molecules to either of the two phases without additional empirical assumptions. In what follows, we will therefore use the concept of Gibbs dividing surface,

signed to the interface. It is then useful to construct a reference system, in which the solid part and the liquid part maintain their bulk properties up to the dividing surface. With a dividing surface in place, the free energy and the properties of any two-phase system can be naturally decomposed into a term corresponding to the reference bulk system and an excess term associated with the interface.

To remove the degree of freedom associated with the choice of a dividing surface, it is customary to select a surface such that there is no surface excess of a certain extensive quantity Φ, i.e. such that the real system and the reference system exhibit the same value for the chosen extensive quantity. This extensive quantity could be the volume occupied by that region, its internal energy or its entropy for example. More often than not, it is also convenient to use an extensive order parameter $\Phi = \sum_i \phi_i$, where the atomic order parameter ϕ_i is calculated based on the local environment of each of the particles in the system. The zero surface excess condition can be schematically expressed as

$$\Phi_{\text{sl}}(n_s, n_l) = \Phi_{\text{ref}}(n_s, n_l),$$ (6)

where $\Phi_{\text{sl}}(n_s, n_l)$ and $\Phi_{\text{ref}}(n_s, n_l)$ stands for the values of the extensive quantity Φ of the real solid-liquid system and the reference system that both comprise n_s solid atoms and n_l liquid atoms.

Now consider a solid-liquid system comprising a total of N atoms. In our previous work[22] we argued that an ideal reference system for any microstate of such a system can be constructed based on the instantaneous value of Φ in that microstate. This reference system comprises a bulk solid that has n_s atoms, and a bulk liquid that has $N - n_s$ atoms. To find the value of n_s in this system one simply applies the deterministic mapping $\Phi_{\text{ref}}(n_s, n_l) = \phi_s n_s + \phi_l(N - n_s)$ where ϕ_s and ϕ_l are the average value for the order parameter of each atom in the solid and liquid respectively. This mapping corresponds to a Gibbs dividing surface between the two phases that has zero excess for the extensive variable Φ.

However, it is important to realize that an instantaneous extensive quantity of a finite piece of bulk solid or bulk liquid can fluctuate even at fixed thermodynamic conditions. As such, the extensive quantity Φ of a reference system that has a bulk solid part with n_s

is the value of Φ for that microstate. This distribution, $\rho_s (\Phi|n)$, can be regarded as the conditional probability for observing Φ in a system consists of a bulk solid region containing n atoms. Since this solid region should mimic the bulk solid part in the reference system, strictly speaking its shape should be delimited by the the Gibbs dividing surface and the boundaries of the reference system. It is worth noting, however, that in the cases we considered in this study the shape of the solid region has little impact on $\rho_s (\Phi|n)$, as long as a compact shape is chosen. An analogous distribution $\rho_l (\Phi|n)$ can be derived for a bulk liquid region that contains n atoms and also has a shape determined by the dividing surface and the boundaries of the reference system.

Contrary to the multiple cluster model discussed above, "solid" and "liquid" in this case indicate well-defined thermodynamic states. The bulk solid state encompasses all the possible configurations for a system of solid, which can contain point defects, other crystal defects, and even small molten pools. Similarly, a bulk liquid comprises local crystalline orderings and sub-critical solid clusters.

The finite width of the distributions for $\rho_s (\Phi|n)$ and $\rho_l (\Phi|n)$ in Eqn. (7) ensures that the value of Φ for a given microstate cannot be used to determine the composition of a reference system with absolute certainty. Instead, we can compute the distribution of Φ for a reference system composed of n_s solid atoms and n_l liquid atoms using

$$\rho_{\mathrm{ref}} (\Phi|n_s, n_l) = \int \mathrm{d}\varphi\, \rho_s (\varphi|n_s)\, \rho_l (\Phi - \varphi|n_l) . \quad (8)$$

According to the concept of the Gibbs dividing surface, if the reference and the actual system both have n_s solid atoms and n_l liquid atoms, they should also both have the same distribution for Φ. In other words, the zero-excess condition in Eqn. (6) take a probabilistic form,

$$\rho_{\mathrm{sl}} (\Phi|n_s, n_l) \equiv \rho_{\mathrm{ref}} (\Phi|n_s, n_l) . \quad (9)$$

Only when $\rho_s (\Phi|n)$ and $\rho_l (\Phi|n)$ are both δ functions at any given n, Eqn. (8) is reduced to a deterministic mapping between Φ and n_s analogous to that introduced in Ref.[22].

have configurations that contain sub-critical nuclei of large sizes, such as the one illustrated in Figure 1. Those configurations can have a value of Φ approaching those typically encountered for a solid sample. As we have discussed in the multiple cluster model, configurations that contain a large number of solid-like atoms are overwhelmingly likely to comprise one and only one cluster of size larger than n_{cut} and a liquid-like background. By a similar logic, a configuration with a value of Φ that has enough solid-like characteristics can be interpreted as a single solid cluster larger than n_{cut} and the surrounding liquid.

Consider a single solid cluster that has $n_s > n_{\mathrm{cut}}$ atoms together with a liquid background of $N - n_s$ atoms. The values of the extensive quantity for this combination of phases follow the distribution $\rho_{\mathrm{sl}} (\Phi|n_s, N - n_s)$. Inside the undercooled bulk liquid, the average population for solid clusters of size n_s can be expressed as $\langle p_{n_s} \rangle = N_s \exp(-\beta G(n_s))$, where N_s is the number of nucleation sites that can often be considered as the number of atoms or molecules or lattice sites in homogeneous nucleation, and $\exp(-\beta G(n_s))$ the probability that a nucleus of size n_s has grown around a nucleation site in the metastable liquid. In other words, $G(n_s)$ represents the free energy excess associated with a solid cluster that has n_s atoms relative to the bulk liquid. Notice also that for $n_s > n_{\mathrm{cut}}$, the average population $N_s \exp(-\beta G(n_s))$ is also the probability of observing a solid cluster of size n_s in the bulk liquid system. Based on these considerations, and using the law of total probability, the probability distribution for Φ in such systems follows

$$e^{-\beta \tilde{G}(\Phi)} = \int_{n_{\mathrm{cut}}}^{n^{\star}} \mathrm{d}n_s\, \rho_{\mathrm{sl}} (\Phi|n_s, N - n_s)\, N_s\, e^{-\beta G(n_s)} . \quad (10)$$

This expression is valid for values of Φ that satisfy $\rho_{\mathrm{sl}} (\Phi|n, N - n) \approx 0$ for all $n < n_{\mathrm{cut}}$, so that the system can be considered to have a single cluster of size larger than n_{cut}.

In principle, $G(n_s)$ can be determined from Eqn. (10), as both $\tilde{G}(\Phi)$ and $\rho_{\mathrm{sl}} (\Phi|n_s, N - n_s)$ can be computed from simulations. However, to avoid the numerical instabilities in the direct deconvolution process, we cast the problem as a fixed-point iteration. The average Φ value for a system containing
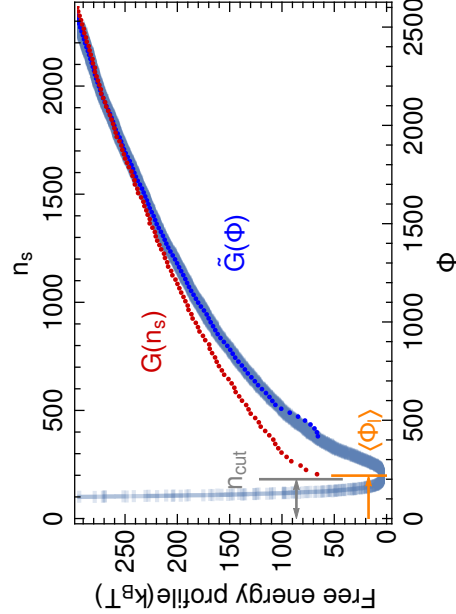
FIG. 4. The solid blue line is the free energy profile $\tilde{G}(\Phi)$, with statistical errors indicated by the error bars. The red and the blue dots indicate the reconstructed curves for $G(n_{\rm s})$ and $\tilde{G}(\Phi)$, respectively. The grey and yellow vertical lines indicate $n_{\rm cut}$ and the average extensive quantity $\langle\Phi_l\rangle$, respectively.

dition on $G(n_{\rm s})$

$$G(n_{\rm s}) = \tilde{G}(\bar{\Phi}(n_{\rm s})) + \frac{1}{\beta}\log N_s$$

$$+\frac{1}{\beta}\log\int_{n_{\rm cut}}^{n^\star}{\rm d}n\,\rho_{\rm sl}\left(\bar{\Phi}(n_{\rm s})|n,N-n\right)e^{-\beta[G(n)-G(n_{\rm s})]}, \quad (12)$$

which can be solved iteratively starting from the initial guess $G_0(n_{\rm s}) = \tilde{G}(\bar{\Phi}(n_{\rm s}))$, and plugging the old guess onto the right-hand side to obtain a new estimate at each iteration. Upon convergence, $G(n_{\rm s})$ is an estimate of the free energy for a solid cluster containing $n_{\rm s}$ atoms relative to the bulk liquid.

It is worth stressing that the cluster size $n_{\rm s}$ and the associated free energy $G(n_{\rm s})$ in Eqn. (10) are still dependent on the choice of Φ, because the reference system is defined based on a Gibbs dividing surface that has zero excess for the extensive quantity Φ. Due to the diffuse nature of the physical interface, a different choice for the extensive variable can result in a different location of the Gibbs dividing surface and a different reference system. However, as extensively discussed in our previous work, as long as one uses one extensive quantity and its associated reference system consistently throughout the analysis, no ambiguity will arise in the value of the nucleation barrier[22].

solid and the liquid phases, and use that knowledge to convert $\tilde{G}(\Phi)$ of the whole system into $G(n_{\rm s})$, using the iterative expression in Eqn. (12). Here we also assume the total number of atoms in the system N to be the number of nucleation sites N_s, although any other choice would simply amount to a vertical shift of the free-energy curve. This curve of $G(n_{\rm s})$ plotted as the red dots in Figure 4, corresponds to the free energy for a single cluster relative to the bulk liquid. In order to demonstrate the convergence of the conversion, we used the computed $G(n_{\rm s})$ to reconstruct the free energy profile $\tilde{G}(\Phi)$ using Eqn. (10). As shown in Figure 4, the reconstructed $\tilde{G}(\Phi)$ is indistinguishable from that obtained directly from the simulation.

As suggested by the many similarities between Figure 4 and Figure 3, the multiple cluster model and the thermodynamic model are very closely related. In the Supplementary Material we show that under a few additional assumptions Eqn. (10) is exactly the same as Eqn. (5), with Φ taking the role of $n_{\rm tot}$. Eqn. (10) serves a dual purpose: it converts the extensive quantity Φ into an estimate for the overall solid fraction and it singles out the free-energy excess for the largest

II. APPLICATIONS OF THE FLUCTUATING REFERENCE FRAMEWORK

A. Solidification of a Lennard-Jones system

To demonstrate how this thermodynamic framework can be applied to an atomistic simulation of a phase transition, we simulated the processes of homogeneous solidification for a Lennard-Jones system of 23328 atoms at $T = 0.58$[23–25] – corresponding to a moderate undercooling relative to the melting temperature of $T_m = 0.6185$.[22–24] We performed 12 independent biased sampling runs using the well-tempered metadynamics protocol with adaptive Gaussians[26–29]. The sample input files can be found in the Supplementary Material. We used a collective variable $\Phi = \sum_i S(\kappa(i))$ in the biased simulations, where
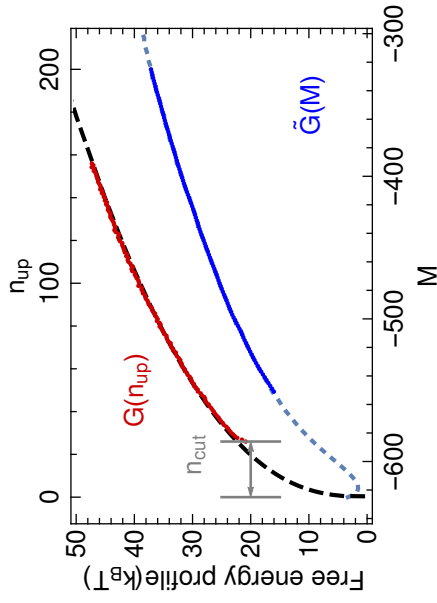
FIG. 5. The dotted blue line is the free energy profile $\tilde{G}(M)$ as a function of the magnetization, for a square-lattice Ising model with size $L = 25$ and periodic boundary conditions, simulated at $T = 1.5$. The solid red and the blue lines indicate the reconstructed curves for $G(n_{up})$ and $\tilde{G}(M)$, respectively. The dashed black line is the prediction from Eqn. (14). Note that $G(n_{up})$ and $\tilde{G}(M)$ are discrete functions, as n_{up} can only have integer values and M can only be odd integers in this system. The grey vertical lines indicate the choice of n_{cut} for this system.

extends over all its nearest neighbors in the lattice. We used a periodic square lattice with side $L = 25$, and performed a Monte Carlo simulation with biased sampling[30] at the temperature $T = 1.5$, well below the critical temperature $T_c = 2.269$. We started the simulation with all spins down, and also restricted the sampling to the states with negative total magnetization.

The snapshot in Figure 2 (from a simulation with $L = 50$) shows a large cluster with positive magnetization embedded in the phase with spins down. Note that one can observe spontaneous fluctuations of opposite spins not only in the negatively-magnetized background, but also inside the large nucleating cluster, underscoring the ambiguity in defining cluster sizes by counting the number of contiguous spins with the same orientation.[13] In contrast, our thermodynamic model does not rely on any clustering algorithm to identify nuclei of the different phases, but instead only focuses on the total magnetization $M = \sum_i s_i$ as a macroscopic order parameter to characterize the overall state of the system. The dotted blue line in Figure 5 indicates the free energy profile $\tilde{G}(M)$. Taking a reference state with a circular Gibbs dividing surface, we computed the probability distribution $\rho\left(M | n_{up}, n_{down}\right)$ from unbiased simulations of the bulk phases as shown in the Supplementary Material. Using the iterative expression in Eqn. (12), and assuming the number of nucleation sites N_s is the total size of the lattice, L^2, we obtained $G(n_{up})$, which corresponds to the free energy for a single positively-magnetized cluster relative to the bulk negatively-magnetized phase. $G(n_{up})$ is plotted as the solid red line in Figure 5, together with the free energy profile $\tilde{G}(M)$ reconstructed using Eqn. (10), that matches perfectly the directly computed free-energy curve, signaling the convergence of self-consistent iterations.

In Ref.[31], the computed nucleation free energy profile of the 2D Ising model was found to agree well with the expression

$$G(n) = 2\sqrt{\pi n}\sigma + \tau k_B T \ln n + d, \tag{14}$$

where $\sigma = 1.20585$ is the temperature-dependent interfacial free energy for this Ising model that can be computed analytically[32], $\tau k_B T \ln n$ accounts for the

chosen to discriminate between the solid and the liquid, be it built upon local descriptors, or a traditional thermodynamic quantity such as the total volume, the energy or the magnetization. By characterizing the fluctuations of this extensive quantity we can rigorously define, in a probabilistic manner (Eqn. 8), a reference state consistent with a zero-excess Gibbs dividing surface that encloses a single cluster of the stable phase.

Our method is applicable to all types of phase transitions – from solidification, to precipitation or condensation – and it can be combined with any sampling method one chooses to accelerate nucleation[3,8-18]. From such simulations the free-energy for the overall system as a function of any extensive quantity can be computed, and then converted into the free energy of a single cluster relative to the metastable bulk. By avoiding the need of singling out atom-size clusters

[32] V. Shneidman, K. Jackson, and K. Beatty, The Journal of chemical physics **111**, 6932 (1999).

[33] M. Bonomi and M. Parrinello, Phys. Rev. Lett. **104**, 190601 (2010).

protocols and a commented sample input file.

ACKNOWLEDGMENTS

[1] D. W. Oxtoby, Journal of Physics: Condensed Matter **4**, 7627 (1992).

[2] J. Schmelzer, G. Röpke, and V. B. Priezzhev, *Nucleation theory and applications* (Wiley Online Library, 2005).

[3] P. Yi and G. C. Rutledge, Annual review of chemical and biomolecular engineering **3**, 157 (2012).

[4] G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen, and A. Michaelides, Chemical reviews (2016).

[5] P. Jonkheijm, P. van der Schoot, A. P. Schenning, and E. Meijer, Science **313**, 80 (2006).

[6] W. Boettinger, S. Coriell, A. Greer, A. Karma, W. Kurz, M. Rappaz, and R. Trivedi, Acta Materialia **48**, 43 (2000).

[7] J. Venables, G. Spiller, and M. Hanbucken, Reports on Progress in Physics **47**, 399 (1984).

[8] P. R. ten Wolde, M. J. Ruiz-Montero, and D. Frenkel, The Journal of chemical physics **104**, 9932 (1996).

[9] P. R. ten Wolde and D. Frenkel, Physical Chemistry Chemical Physics **1**, 2191 (1999).

[10] S. Auer and D. Frenkel, Nature **409**, 1020 (2001).

[11] D. Moroni, P. R. Ten Wolde, and P. G. Bolhuis, Physical review letters **94**, 235703 (2005).

[12] F. Trudu, D. Donadio, and M. Parrinello, Physical review letters **97**, 105701 (2006).

[13] L. Maibaum, Physical review letters **101**, 256102 (2008).

[14] W. Lechner, C. Dellago, and P. G. Bolhuis, Physical review letters **106**, 085701 (2011).

[15] S. Prestipino, A. Laio, and E. Tosatti, Physical review letters **108**, 225701 (2012).

[16] M. Salvalaglio, C. Perego, F. Giberti, M. Mazzotti, and M. Parrinello, Proceedings of the National Academy of Sciences **112**, E6 (2015).

[17] J. McCarty, O. Valsson, and M. Parrinello, Journal of chemical theory and computation **12**, 2162 (2016).

[18] P. M. Piaggi, O. Valsson, and M. Parrinello, Faraday Discussions (2016).

[19] Y. Lifanov, B. Vorselaars, and D. Quigley, The Journal of Chemical Physics **145**, 211912 (2016), http://dx.doi.org/10.1063/1.4962216.

[20] J. W. Gibbs, H. A. Bumstead, W. R. Longley, *et al.*, *The collected works of J. Willard Gibbs*, Vol. 1 (Longmans, Green

Bridging the gap between atomistic and macroscopic models of homogeneous nucleation
Supplementary Material

Bingqing Cheng[1, *] and Michele Ceriotti[1]

[1]*Laboratory of Computational Science and Modeling, Institute of Materials,*
École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
(Dated: December 30, 2016)

CLUSTER DISTRIBUTIONS UNDER THE MULTIPLE CLUSTER MODEL

In the multiple cluster model described in the main text, the probability of cluster formation follows

$$\langle p_n \rangle \propto \exp(-\beta G(n)), \tag{1}$$

with

$$G(n) = \mu n + \sigma v^{\frac{2}{3}} n^{\frac{2}{3}}, \tag{2}$$

and the effective interfacial free energy σ is equal to the specific interfacial free energy γ times a geometrical constant Ω. For our toy system, we selected parameters $k_B = 1$, $T = 0.58$, $\mu = -0.0562\epsilon$, $\gamma = 0.354\epsilon/\sigma^2$, $\Omega = 4.836$, $v^{\frac{2}{3}} = 1.035\sigma^2$, and ϵ and σ are the Lennard-Jones energy and length units. These parameters were selected to mimic those obtained in the calculations we performed for the Lennard-Jones system [1].

The probability distribution of the number of clusters that has n atoms (denoted by k_n) are approximated by Poisson distributions, i.e.

$$\begin{aligned} P(n, k_n) &= \lambda(n)^{k_n} \exp\left(-\lambda(n)\right)/k_n! \,, \\ \lambda(n) &= N_{\mathrm{s}} \exp(-\beta G(n)), \end{aligned} \tag{3}$$

and we assumed $N_{\mathrm{s}} = 700$.

The free energy profile $\tilde{G}(n_{\mathrm{tot}})$ as a function of total number of atoms in all the clusters can be computed by enumerating all the possible combinations such as

$$\exp\left(-\beta \tilde{G}(n_{\mathrm{tot}})\right) = \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} \ldots \sum_{k_{n^\star}=0}^{\infty} \delta\left(\sum_{n=1}^{\infty} n k_n - n_{\mathrm{tot}}\right) \prod_{n=1}^{n^\star} P(n, k_n), \tag{4}$$

where $\delta(x = 0) = 1$ and $\delta(x \neq 0) = 0$, and $n^\star$ is the size of the critical nucleus. If a cutoff size n_{cut} is selected, one can also compute the distribution of n_{tot} under the constraint that no cluster in the system is of size larger than n_{cut}, i.e.

$$\exp\left(-\beta \tilde{G}^{(0)}(n_{\mathrm{tot}})\right) = \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} \ldots \sum_{k_{n_{\mathrm{cut}}}=0}^{\infty} \delta\left(\sum_{n=1}^{\infty} n k_n - n_{\mathrm{tot}}\right) \prod_{n=1}^{n_{\mathrm{cut}}} P(n, k_n), \tag{5}$$

On the other hand, if one restricts the system to have exactly one cluster larger than n_{cut}, the distribution of n_{tot} can be written as

$$\exp\left(-\beta \tilde{G}^{(1)}(n_{\mathrm{tot}})\right) = \sum_{n_b=n_{\mathrm{cut}}}^{n_{\mathrm{tot}}} P(n_b, 1) \exp(-\beta \tilde{G}^{(0)}(n_{\mathrm{tot}} - n_b)) \tag{6}$$

Furthermore, if one restricts the system to have two clusters that have sizes larger than n_{cut}, the distribution of n_{tot} can be written as

$$\exp\left(-\beta \tilde{G}^{(2)}(n_{\mathrm{tot}})\right) = \sum_{n_a=n_{\mathrm{cut}}}^{n_{\mathrm{tot}}} \sum_{n_b=n_{\mathrm{cut}}}^{n_a} P(n_a, 1) P(n_b, 1) \exp(-\beta \tilde{G}^{(0)}(n_{\mathrm{tot}} - n_a - n_b)) \tag{7}$$

We selected $n_{\mathrm{cut}} = 150$, computed $\tilde{G}(n_{\mathrm{tot}})$, $\tilde{G}^{(0)}(n_{\mathrm{tot}})$, $\tilde{G}^{(1)}(n_{\mathrm{tot}})$, and $\tilde{G}^{(2)}(n_{\mathrm{tot}})$, and plotted them in Figure 1. It can be seen that at $n_{\mathrm{tot}} > n_{\mathrm{cut}} + \langle n_{\mathrm{tot}} \rangle$, $\tilde{G}^{(1)}(n_{\mathrm{tot}})$ overlaps with $\tilde{G}(n_{\mathrm{tot}})$. In addition, at $n_{\mathrm{tot}} > n_{\mathrm{cut}} + \langle n_{\mathrm{tot}} \rangle$ it is many orders of magnitude less likely to have zero or two clusters of size larger than n_{cut} compared with only having one such cluster. Therefore, at $n_{\mathrm{tot}} > n_{\mathrm{cut}} + \langle n_{\mathrm{tot}} \rangle$ the system effectively only contains one large cluster with size larger than n_{cut}, and $\tilde{G}^{(1)}(n_{\mathrm{tot}})$ is a good approximation of $\tilde{G}(n_{\mathrm{tot}})$. Note that n_{cut} does not need to be decided exactly. As long as n_{cut} is reasonably large, the conclusions above apply.
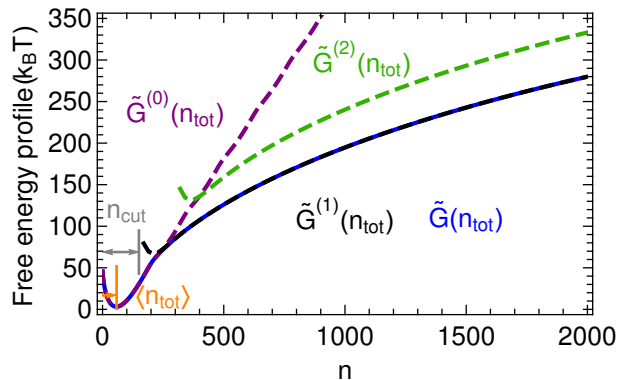
FIG. 1: Free energy profiles $\tilde{G}(n_{tot})$, $\tilde{G}^{(0)}(n_{tot})$, $\tilde{G}^{(1)}(n_{tot})$, and $\tilde{G}^{(2)}(n_{tot})$ of the atomistic system of multiple clusters. At $n_{tot} < n_{cut}$, $\tilde{G}^{(0)}(n_{tot})$ is exactly equal to $\tilde{G}(n_{tot})$. At $n_{tot} > n_{cut} + \langle n_{tot} \rangle$, $\tilde{G}^{(1)}(n_{tot})$ lies on the top of $\tilde{G}(n_{tot})$.

THE CONNECTION BETWEEN THE MULTIPLE CLUSTER MODEL AND THE THERMODYNAMIC MODEL

In the main text, we introduced a thermodynamic model to describe the metastable liquid system, and argued that for sufficiently large Φ (here for simplicity we assume that the average order parameter in the solid phase is larger than in the liquid), the probability distribution of Φ in the liquid should follow

$$\rho_l(\Phi|N) = \int_{n_{cut}}^{n^\star} \mathrm{d}n\, \rho(\Phi|n, N-n)\, N_s \exp(-\beta G(n)). \tag{8}$$

On the other hand, we also employed the multiple cluster model, and suggested that at $n_{tot} \gg n_{cut}$,

$$\exp(-\beta \tilde{G}(n_{tot})) = \sum_{n_b = n_{cut}}^{n_{tot}} N_s \exp(-\beta G(n)) \exp(-\beta \tilde{G}(n_{tot} - n_b)). \tag{9}$$

It is instructive to highlight the connection between Eqn. (8) of the thermodynamic model and Eqn. (9) of the multiple cluster model. By inserting the definition for $\rho(\Phi|n, N-n)$, and using the approximation $\rho_l(\Phi|N-n) \approx \rho_l(\Phi|N)$, (which can be regarded as equivalent to the assumption that the presence of other clusters do not affect the $P(n, k_n)$ in the multiple cluster model), one obtains

$$\rho_l(\Phi|N) = \int \mathrm{d}\varphi \rho_l(\Phi - \varphi|N)\, N_s \exp(-\beta G(\varphi)), \tag{10}$$

where we defined the free energy for a cluster characterized by an order parameter φ,

$$\exp(-\beta G(\varphi)) = \int_{n_{cut}}^{n^\star} \mathrm{d}n \rho_s(\varphi|n) \exp(-\beta G(n)). \tag{11}$$

Eqn. (10) corresponds to a continuous version of Eqn. (9), in which the extensive order parameter Φ is used to characterize the solid fraction in the metastable liquid and the size of the largest solid cluster.

CONDITIONAL PROBABILITY DISTRIBUTIONS $\rho_s(\Phi|n)$ AND $\rho_l(\Phi|n)$ IN THE LENNARD-JONES SYSTEM

For the Lennard-Jones system described in the main text, assuming the Gibbs dividing surface separating the solid cluster and the bulk liquid has a spherical shape, we can compute the two probability distributions $\rho_s(\Phi|n)$ and $\rho_l(\Phi|n)$ from simulations of the bulk phases.

For example, take a molecular dynamics simulation run of a bulk solid system that has N atoms. From each snapshot, one can randomly select a volume that has a spherical shape, as illustrated in Figure 2. One can then count the number of atoms n in that volume, as well as the extensive quantity Φ associated with those n atoms. After accumulating enough statistics for (n_i, Φ_i), one can build the conditional probability distribution $\rho_s(\Phi|n)$ by

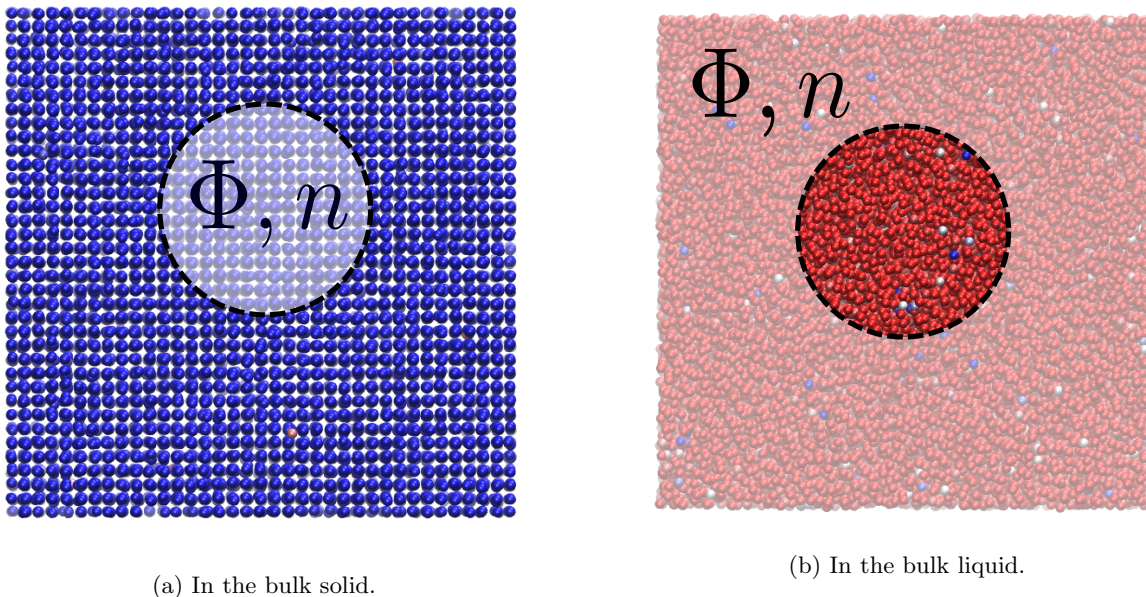$$\rho_s(\Phi|n) = \langle \delta(\Phi_i - \Phi)\delta(n_i - n) \rangle / \langle \delta(n_i - n) \rangle, \tag{12}$$

(a) In the bulk solid.

(b) In the bulk liquid.

FIG. 2: A schematic illustration of selecting a contiguous portion of n atoms in a system.

(a) $\overline{\Phi_s}$ in the bulk solid.

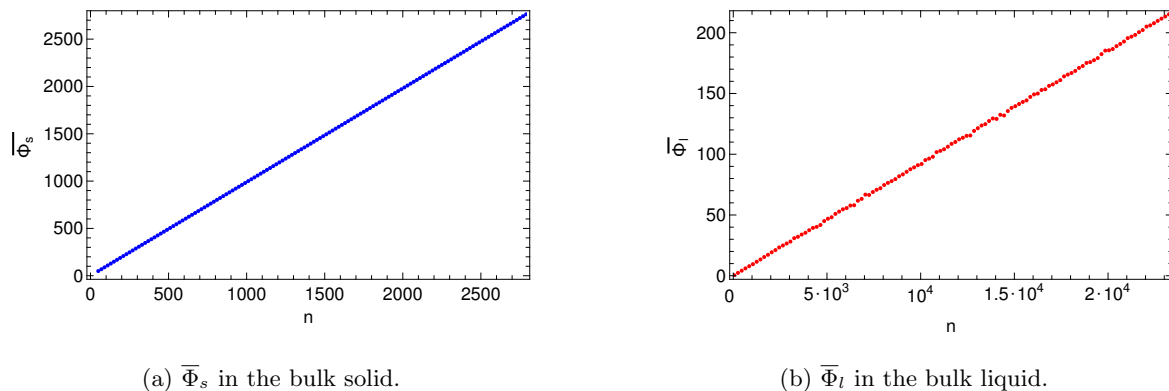(b) $\overline{\Phi_l}$ in the bulk liquid.

FIG. 3: The average extensive quantity Φ for different system sizes.

where $\langle \ldots \rangle$ denotes the ensemble average at the given thermodynamic condition. Furthermore, $\rho_s(\Phi|n)$ can also be computed from biased simulations, by statistical re-weighting

$$\rho_s(\Phi|n) = \langle \delta(\Phi_i - \Phi)\delta(n_i - n)\exp(\beta V_i)\rangle / \langle \delta(n_i - n)\exp(\beta V_i)\rangle, \tag{13}$$

where V_i is the instantaneous external bias for each snapshot.

For the bulk solid system, the average extensive quantity

$$\overline{\Phi_s}(n) = \int \mathrm{d}\Phi \rho_s(\Phi|n)\,\Phi \tag{14}$$

at each n was computed and plotted in the left panel of Figure 3. It can be seen that scales linearly with system size, and thus can be expressed as $\overline{\Phi}_s = n\phi_s$, where ϕ_s is the average value of order parameter of each atom in the bulk solid. The average Φ in bulk liquid $\overline{\Phi_l}(n) = \int \mathrm{d}\Phi \rho_l(\Phi|n)$ is also shown in the right panel of Figure 3. Similarly, in bulk liquid $\overline{\Phi_l} = n\phi_l$, where ϕ_l is the average value of order parameter of each atom in the bulk solid. From simulations we determined $\phi_s = 0.99000$ and $\phi_l = 0.00936$. We have also computed the variances of Φ in the bulk solid and the bulk liquid at different n, using the equations

$$\mathrm{Var}(\Phi_s) = \int \mathrm{d}\Phi (\Phi - n\phi_s)^2 \rho_s(\Phi|n) \tag{15}$$

and

$$\mathrm{Var}(\Phi_l) = \int \mathrm{d}\Phi (\Phi - n\phi_l)^2 \rho_l(\Phi|n). \tag{16}$$

(a) $\mathrm{Var}(\Phi_s)$ in bulk solid.

(b) $\mathrm{Var}(\Phi_l)$ in bulk liquid.
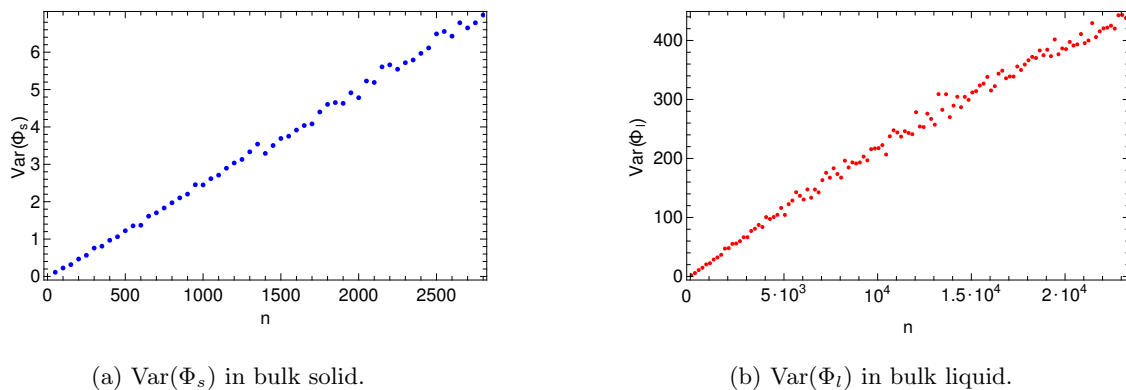
FIG. 4: The variance of the extensive quantity Φ at different sizes.

(a) $f_s(\Phi - n\phi_{\mathrm{s}} \mid n)$ in bulk solid.

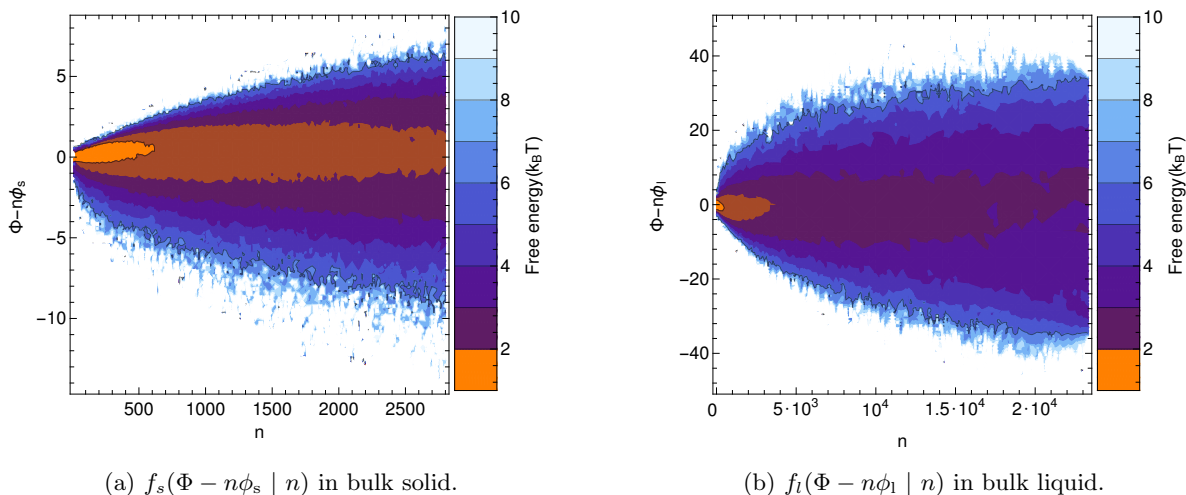(b) $f_l(\Phi - n\phi_{\mathrm{l}} \mid n)$ in bulk liquid.

FIG. 5: The free energies associated with the quantities $\Phi - n\phi_{\mathrm{s}}$ and $\Phi - n\phi_{\mathrm{l}}$ at different system size.

The computed variances of Φ are plotted in Figure 4, which shows that the variances of Φ in both phases also scale linearly with the number of the atoms. Furthermore, in Figure 5 we plot the quantities $f_s(\Phi - n\phi_{\mathrm{s}} \mid n)$ and $f_l(\Phi - n\phi_{\mathrm{l}} \mid n)$, which are defined as

$$f_s(\Phi - n\phi_{\mathrm{s}} \mid n) = -\log(\rho_s(\Phi \mid n)), \qquad (17)$$

and

$$f_l(\Phi - n\phi_{\mathrm{l}} \mid n) = -\log(\rho_l(\Phi \mid n)). \qquad (18)$$

The parabolic shapes of $f_s(\Phi - n\phi_{\mathrm{s}} \mid n)$ and $f_l(\Phi - n\phi_{\mathrm{l}} \mid n)$ suggest that $\rho_s(\Phi|n)$ and $\rho_l(\Phi|n)$ follow Gaussian-like distributions centered at $n\phi_{\mathrm{s}}$ and $n\phi_{\mathrm{l}}$, respectively. All in all, the analyses above suggest that $\rho_s(\Phi|n)$ and $\rho_l(\Phi|n)$ can be approximated as the distributions of the sum of n independent Gaussian-distributed random variables.

THE PROBABILITY DISTRIBUTION $\rho(M|n_{up}, n_{down})$ OF SQUARE-LATTICE ISING MODEL

For the square-lattice Ising model with L^2 lattice described in the main text, the probability distribution $\rho(M|n_{up}, n_{down})$ can be determined as

$$\rho(M|n_{up}, n_{down}) = \int \mathrm{d}M' \rho_{up}(M'|n_{up}) \rho_{down}(M - M'|n_{down}). \qquad (19)$$

In general $\rho_{up}(M|n) = \rho_{down}(-M|n)$ because of the symmetry in the system, therefore only one of them needs to be characterized.

We took a Monte Carlo simulation of the Ising lattice, and from each snapshot we can randomly selected a circle and then counted the number of up spins n as well as the magnetization M inside (and outside) the circle.
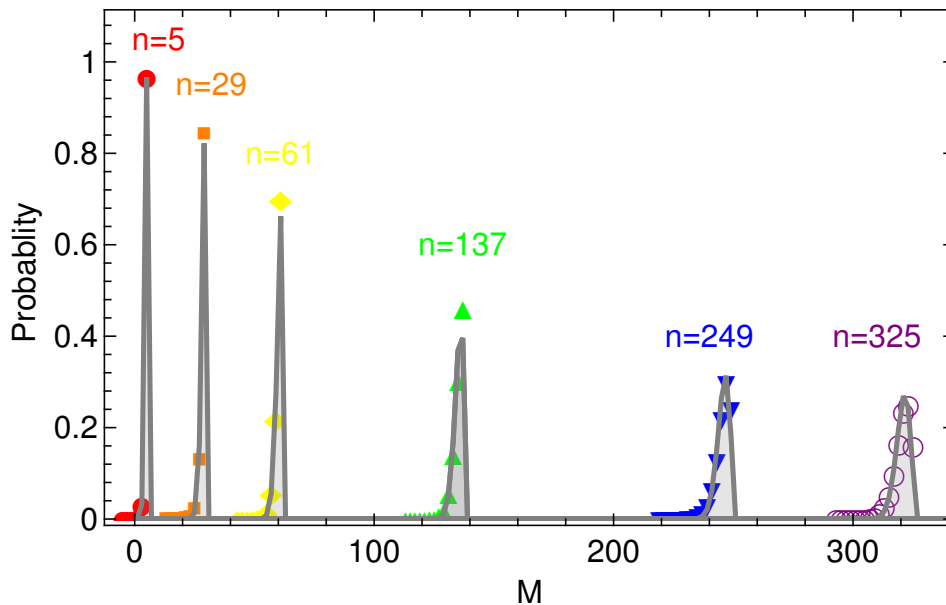
FIG. 6: The probability distributions $\rho_{up}(M|n)$ at different values of n. The gray lines indicate the predictions from the Binomial model.

Whether the n lattices were selected inside or outside the circle was found to have negligible effect. We plot some of the probability distributions $\rho_{up}(M|n)$ at different values of n in Figure 6.

We also found a Binomial model that is able to describe the behavior of $\rho_{up}(M|n)$ very well at all values of n. In this Bionomial model, we assume each spin has a probability of $p = 0.00674$ to flip to the opposite site in the sea of up spins. As such, $\rho_{up}(M|n)$ can be approximated as

$$\rho_{up}(M|n) \approx \binom{n}{k}p^k(1-p)^{n-k}, \ k = (M-n)/2. \tag{20}$$

In Figure 6, we also plotted the predictions of this model, which find good agreement with the distributions computed directly from the simulations.

Sample LAMMPS input file

Here is a sample LAMMPS input file for running the molecular dynamics simulations.

```
atom_style      atomic
units           lj
dimension       3
boundary        p p p
processors      * * 1

read_data       ./data.lj    # Coordinate file
```

Atoms of type 1 are the Lennard-Jones particles with the truncated pairwise potential [2–4]. The atom of type 2 is a fixed ghost atom that does not interact with its surroundings. It is used as reference point in some of the collective variables.

```
group real type 1
velocity real create 0.6 RandomSeedHere dist gaussian

pair_style table linear 5000
pair_coeff 1 1 tr-lj.table TR_LJ 2.5
pair_coeff * 2 tr-lj.table NULL 2.5

neighbor        1.0 bin

timestep 0.004
```

An initial equilibration of 20'000 steps is performed following by a complete melting of the system. During these processes a highly efficient Generalized Langevin thermostat [5] is employed.

```
fix 1 real press/berendsen iso 0.0 0.0 0.5
fix 2 real gle 6 0.84 0.84 12rrr5 smart2.A every 2
run 80000
unfix 2
fix 3 real gle 6 0.58 0.58 12rrr5 smart2.A every 5
run 20000
unfix 3
```

> After the equilibration a metadynamics simulation in the NPT ensemble, using a Nosé-Hoover barostat and thermostat is started.

```
fix 4 real npt temp 0.58 0.58 0.2 iso 0.0 0.0 0.5
fix 5 all plumed   plumedfile  plumed.dat     outfile p.log

thermo 500
thermo_style one
dump coord all xyz 5000 traj.xyz

run            20000000
write_restart restart.lj
```

Sample PLUMED input file

> Here we provide a sample input file for PLUMED. This input was used to compute the free energy profile during homogenous nucleation.

```
UNITS NATURAL
RESTART
```

> Record the volume of the simulation cell.

```
cell: CELL
```

> We first compute the local order parameter κ using the cubic harmonic function "FCCUBIC". The sigmoid switching function "SMAP" is used to do a non-linear mapping on κ. We collect $\Phi = \sum S(\kappa(i))$ - the global order parameter used to identify n_s^{Φ} - using the keyword "MORE_THAN". The value of "cub.morethan" is proportional to the number of atoms in the fcc crystal.
> The system contains a total of 23328 real atoms, The first atom is a "ghost" atom that stays stationary at the center of the supercell. This ghost atom is only used as a reference point.

```
FCCUBIC ...
LABEL=cub
SPECIES=2-23329 SWITCH={CUBIC D_0=1.2 D_MAX=1.5} ALPHA=27 MEAN
MORE_THAN1={SMAP R_0=0.45 D_0=0.0 A=8 B=8}
... FCCUBIC
```

> Finally the instantaneous values of the CV and the biases are recorded. These are used in the re-weighting processes to construct the FES, e.g.
>
> $$G(\Phi) = -\frac{1}{\beta} ln \left(\frac{\int dt e^{\beta V_{tot}(t)} \delta(\Phi)}{\int dt e^{\beta V_{tot}(t)}} \right), \qquad (21)$$
>
> where $\beta = 1/k_B T$, and $V_{tot}(t)$ is the total external bias including the metadynamics bias and the restraint biases at time t.

```
METAD ...
LABEL=metad
ARG=cub.morethan-1
PACE=1000 HEIGHT=0.3 SIGMA=600 FILE=HILLS
TEMP=0.58 BIASFACTOR=400
ADAPTIVE=DIFF SIGMA_MAX=300 SIGMA_MIN=0.1
... METAD

fixmax: UPPER_WALLS ARG=cub.morethan-1 AT=2500 KAPPA=0.1

# monitor the two variables and the metadynamics bias potential
PRINT STRIDE=10 ARG=cell.ax,cub.*,metad.bias,fixmax.bias FILE=COLVAR


ENDPLUMED
```

---

* bingqing.cheng@epfl.ch

[1] B. Cheng, G. A. Tribello, and M. Ceriotti, Physical Review B **92**, 180102 (2015).

[2] R. L. Davidchack and B. B. Laird, The Journal of chemical physics **118**, 7651 (2003).

[3] S. Angioletti-Uberti, M. Ceriotti, P. D. Lee, and M. W. Finnis, Phys. Rev. B **81**, 125416 (2010).

[4] R. Benjamin and J. Horbach, The Journal of chemical physics **141**, 044715 (2014).

[5] M. Ceriotti, G. Bussi, and M. Parrinello, J. Chem. Theory Comput. **6**, 1170 (2010).