Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/5/4/eaav7399/DC1

Supplementary Materials for

How crystals form: A theory of nucleation pathways

James F. Lutsko*

*Corresponding author. Email: jlutsko@ulb.ac.be

Published 5 April 2019, *Sci. Adv.* **5**, eaav7399 (2019)
DOI: 10.1126/sciadv.aav7399

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/5/4/eaav7399/DC1)

Movie S3 (.mov format). Cross section of solid cluster density as it evolves along the nucleation pathway for the example presented in the main text showing the log of the density.
Movie S4 (.mov format). Cross section of solid cluster density as it evolves along the nucleation pathway for the example presented in the main text showing the density on a linear scale.

I. DETERMINING THE MLP

A. Summary of the calculation

The Most Likely Path (MLP) is determined in the weak noise limit by performing gradient descent starting at the critical cluster. The goal is to determine a path in the space of densities that approximates the steepest decent path obtained from

$$\frac{\partial}{\partial t} n_t(\boldsymbol{r}) = D\nabla \cdot n_t(\boldsymbol{r}) \nabla \frac{\delta F[\hat{n}_t]}{\delta n_t(\boldsymbol{r})} \tag{1}$$

where the parameter t here has no physical meaning - it is just a parameter characterizing movement along the path. We are free to transform to any other invertible parameterization of the form $s(t) \Leftrightarrow t(s)$ in which case this becomes

$$\frac{\partial}{\partial s} n_s(\boldsymbol{r}) = D\left(\frac{dt(s)}{ds}\right) \nabla \cdot n_s(\boldsymbol{r}) \nabla \frac{\delta F[\hat{n}_s]}{\delta n_s(\boldsymbol{r})} \tag{2}$$

In general, steepest descent paths for a collection of variables $x^{(a)}$ are obtained from generic models of the form

$$\frac{\partial}{\partial s} x_s^{(a)} = D g^{(ab)}(\mathbf{x}_s) \frac{\partial V(\mathbf{x}_s)}{\partial x_s^{(b)}} \tag{3}$$

where V is an energy-surface on which the steepest descent occurs and $g_{(ab)}(\mathbf{x})$ is a metric, which determines the geometry of the surface (note that $g^{(ab)}$ is the matrix inverse of $g_{(ab)}$) . In the present case, the variables are a continuous field (the density), energy surface is determined by the free energy functional $F[n]$, the gradient becomes a functional derivative and the metric is the operator $\nabla \cdot n(\mathbf{r})\nabla$ as discussed elsewhere. The two elements needed to implement these calculations, a free energy functional, $F[n]$, and a method for tracing out the steepest descent path, are discussed in the next Sections.

B. The free energy functional

The free energy functional used here was presented in detail in Ref. [50] and here we only recall some elements concerning its structure. In general, it is written as a sum of several terms

$$F[n] = F^{id}[n] + F^{hs}[n] + F^{mf}[n] + F^{\phi}[n] \tag{4}$$

where the ideal gas contribution is

$$F^{id}[n] = \int \left(n(\boldsymbol{r}) \ln \Lambda^3 n(\boldsymbol{r}) - n(\boldsymbol{r}) \right) d\boldsymbol{r} \tag{5}$$

where Λ is the thermal wavelength. The "hard-sphere" contribution, $F^{hs}[n]$, is more complex and given in detail in Ref. [50], the mean-field contribution has the form

$$F^{mf}[n] = \frac{1}{2} \int n(\boldsymbol{r}_1) n(\boldsymbol{r}_2) w(r_{12}) d\boldsymbol{r}_1 d\boldsymbol{r}_2 \tag{6}$$

where $w(r_{12})$ is the attractive part of the interatomic potential and finally the contribution due to external fields

$$F^{\phi}[n] = \int n(\mathbf{r})\phi(\mathbf{r})d\mathbf{r} \tag{7}$$

where $\phi(\mathbf{r})$ could represent the presence of walls as well as gravitational, electric etc. fields. In the following, the last three contributions will sometimes be grouped together and collectively referred to as the excess contribution, $F^{\text{ex}} = F^{\text{hs}} + F^{\text{mf}} + F^{\phi}$. Note that one can evaluate the ideal gas contribution directly in which case Eq.(1) can be written as

$$\frac{\partial}{\partial t} n_t(\mathbf{r}) = D\nabla^2 n_t(\mathbf{r}) + D\nabla \cdot n_t(\mathbf{r}) \nabla \frac{\delta F^{ex}[\hat{n}_t]}{\delta n_t(\mathbf{r})} \tag{8}$$

In the limit of low densities and no external fields, the excess term becomes negligible and this becomes a simple diffusion equation - this is the reason that the coefficient D is identified unambiguously as the coefficient of tracer diffusion. In the present application, the equations are very stiff due to the fact that in the solid state, the density is highly localized at the lattice positions with very strong gradients. This, combined with the diffusive nature of the dynamics, requires some care.

Our implementation is based on a discretization of space on a cubic calculational grid with spacing Δ. The grid points are $\mathbf{r}_{ijk} = i\Delta\mathbf{x} + j\Delta\mathbf{y} + k\Delta\mathbf{z}$ and the density at the grid points is $n^{ijk}{}_t = n_t(\mathbf{r}_{ijk})$. In the following, a compressed notation will sometimes be used in which the three cartesian indices are referred to by a single "super-index" which will always be indicated by a capitalized letter, e.g. $n_t^I = n_t^{i_x i_y i_z}$ The present calculations are based on those of Ref. [50] and, in particular, the same discretized form of the free energy functional is used here. This functional depends on all of the density points which are referred to collectively as $\mathbf{n}_t$. Thus the discretized free energy functional is written as $F(\mathbf{n}_t)$.

In order to be consistent with this discretized free energy, one cannot use Eq(8) as the starting point as the discretization of the ideal gas term would not be consistent with $F(\mathbf{n}_t)$. Instead, we introduce centered finite differences into Eq.(1) giving

$$\frac{\partial}{\partial t} n_t^I = \frac{D}{\Delta^2} \sum_{\alpha} \left\{ \frac{n_t^{I+\hat{\mathbf{e}}_\alpha} + n_t^I}{2} \frac{1}{\Delta^3} \left(\frac{\partial F}{\partial n_t^{I+\hat{\mathbf{e}}_\alpha}} - \frac{\partial F}{\partial n_t^I} \right) - \frac{n_t^I + n_t^{I-\hat{\mathbf{e}}_\alpha}}{2} \frac{1}{\Delta^3} \left(\frac{\partial F}{\partial n_t^I} - \frac{\partial F}{\partial n_t^{I-\hat{\mathbf{e}}_\alpha}} \right) \right\} \tag{9}$$

where the unit vectors are $\mathbf{e}_1 = \mathbf{x}$, $\mathbf{e}_2 = \mathbf{y}$ and $\mathbf{e}_3 = \mathbf{z}$. The factor of $1/\Delta^3$ comes from the discretization of the functional derivative

$$\frac{\delta F[n]}{\delta n(\mathbf{r})} \to \frac{1}{\Delta^3} \frac{\partial F(\mathbf{n})}{\partial n^I} \tag{10}$$

For example, evaluating this for the discretized ideal gas contribution

$$F^{id}(\mathbf{n}) = \sum_I \left(n^I \ln \Lambda^3 n^I - n^I \right) \Delta^3 \tag{11}$$

gives

$$\frac{1}{\Delta^3}\frac{\partial F^{id}}{\partial n^I} = \ln n^I \Lambda^3 \tag{12}$$

as one would expect. Note that the contribution of this term to the discretized dynamical equation is

$$\frac{D}{\Delta^2}\sum_\alpha \left\{ \frac{n_t^{I+\hat{\mathbf{e}}_\alpha} + n_t^I}{2}\left(\ln n_t^{I+\hat{\mathbf{e}}_\alpha} - \ln n_t^I\right) - \frac{n_t^I + n_t^{I-\hat{\mathbf{e}}_\alpha}}{2}\left(\ln n_t^I - \ln n_t^{I-\hat{\mathbf{e}}_\alpha}\right)\right\} \tag{13}$$

compared to what one gets discretizing after taking the functional derivative

$$\frac{D}{\Delta^2}\sum_\alpha \left\{ n_t^{I+\hat{\mathbf{e}}_\alpha} + n_t^{I-\hat{\mathbf{e}}_\alpha} - 2n_t^I \right\} \tag{14}$$

The connection between the two is easily established by starting with the former and writing

$$n_t^I = \frac{n_t^{I\pm\hat{\mathbf{e}}_\alpha} + n_t^I}{2} - \frac{n_t^{I\pm\hat{\mathbf{e}}_\alpha} - n_t^I}{2} = \frac{n_t^{I\pm\hat{\mathbf{e}}_\alpha} + n_t^I}{2}\left(1 - \frac{n_t^{I\pm\hat{\mathbf{e}}_\alpha} - n_t^I}{n_t^{I\pm\hat{\mathbf{e}}_\alpha} + n_t^I}\right) \tag{15}$$

and expanding using $\frac{n_t^{I\pm\hat{\mathbf{e}}_\alpha} - n_t^I}{n_t^{I\pm\hat{\mathbf{e}}_\alpha} + n_t^I}$ as a small parameter. The diffusive result is obtained at first order but this spoils the accuracy of the centered-finite differences which are of second order. Hence, it is necessary to use the more complex form.

C. Finding critical clusters

To find a critical cluster for an open system, we use a trick described in Ref. [50] where it was shown that the stable clusters that are easily found for a closed system are critical clusters when the system is open based on the following reasoning. In a closed, finite system, with a constant number of particles, a post-critical cluster is thermodynamically driven to grow but, if it is denser than the mother phase, the growth must stop at some point since there is only a finite amount of material present. This final state is obviously stable as long as the system is closed. To minimize the Helmholtz free energy at constant particle number, one can introduce a Lagrange multiplier, λ, and minimize

$$L = F[n] - \lambda\left(\int n(\mathbf{r})\,d\mathbf{r} - N\right) \tag{16}$$

where N is the specified number of particles. Doing so gives the Euler-Lagrange equations

$$\begin{aligned} \frac{\delta F[n]}{\delta n(\mathbf{r})} &= \lambda \\ N &= \int n(\mathbf{r})\,d\mathbf{r} \end{aligned} \tag{17}$$

Solving these equations will yield a density field, $n^*(\mathbf{r})$ and a number, λ^* which satisfy these equations. Now, consider an open system at chemical potential μ. The stationary points of the grand-canonical free energy functional satisfy the Euler-Lagrange equation

$$\frac{\delta F[n]}{\delta n(\mathbf{r})} = \mu \tag{18}$$

so clearly, for $\mu = \lambda^*$, one solution is the field $n^*(\mathbf{r})$. If, physically, there is no finite stable cluster in the open system, this must clearly be a critical cluster. Thus, by finding a stable minimum in the closed system (a fairly easy numerical task) we automatically get the critical cluster for a corresponding open system (a much more difficult task to tackle directly). An exception is the case of crystallization where, as shown in the text, there are in fact metastable finite clusters (having to do with the freezing of shells of material onto the cluster) so that the procedure could isolate one of these rather than the critical cluster itself. In any case, for the purposes here, this presents no problem.

Two methods were used to generate the crystalline clusters. The first was described in Ref. [50] and consists of starting with a "droplet" which is a sphere with a uniform, liquid-like density in its interior and a vapor outside. The free energy funcitonal is then minimized with respect to the density at fixed temperature and number of particles. Once an equilibrium droplet is determined, the temperature is lowered and the system is again minimized. At sufficiently low tempertures (typically, around $k_B T=0.3\varepsilon$), crystalline structures form spontaneously. The temperature is then raised and the density minimized until the desired temperature is reached. Obviously, at sufficiently high temperatures, the crystalline structure melts and the droplet is recovered.

The second method begins with the zero-temperature minimum energy clusters tabulated in the Cambridge Cluster Database [62] and the by Xuegang Shao and co-workers [63]. For a given N, the zero-temperature structure is used to generate a density field by putting a normalized Gaussian at each particle coordinate, R_j, and adding in a constant background, n_b

$$n_I = n_b + \sum_{j=1}^{N} \left(\frac{\alpha}{\pi}\right)^{3/2} \exp\left(-\alpha\left(\mathbf{r}_j - \mathbf{R}_j\right)^2\right). \tag{19}$$

The value $\alpha = 40$ was used in all cases and the resulting density field was then used as the initial value to minimize the free energy functional at constant particle number. Although the structures obtained for the same conditions are somewhat different using the two procedures, see Ref. [50], no qualitative difference in the nucleation pathway was observed in the calculations.

D. String Method

We have implemented the determination of the steepest descent path using the string method [36]. Our implementation begins with the initial and final states of the path, $n_{\text{initial}}(\mathbf{r})$ and $n_{\text{final}}(\mathbf{r})$. A number $N + 1$ of images along an initial path are generated as

$$n_n^I = \left(1 - \frac{n}{N}\right) n_{\text{initial}}^I + \frac{n}{N} n_{\text{final}}^I, \quad n = 0, \ldots, N \tag{20}$$

where the linear interpolation ensures that, for the case of a closed system, the mass of all of the images is the same. Furthermore, the euclidean distance between successive images, defined as

$$s\left(\mathbf{n}_n, \mathbf{n}_{n+1}\right) = \sqrt{\sum_I \left(n_{n+1}^I - n_n^I\right)^2} = \frac{1}{N}\sqrt{\sum_I \left(n_{\text{final}}^I - n_{\text{initial}}^I\right)^2} \tag{21}$$

is the same for each pair.

The outline of the string method is as follows:

- Relax each image (except the end points) by integrating the equations of motion a time δ.
- Calculate the Euclidean distances between the images
- Interpolate between the images so that the distance between each successive pair of images is again equal.
- Test for convergence: if not converged, return to first step.

1. Relaxation step

The first step of the string method requires that each image be propagated in time according to the dynamical equation. Thus we must integrate

$$\frac{\partial}{\partial t} n_t^I = R^I\left(\mathbf{n}_t\right) \tag{22}$$

and, as stated above, this requires care due to the stiffness of the equations. First, it should be noted that even though the ideal gas term does not give simple diffusion, it is very similar to a diffusive processes. Thus, we choose to rewrite the evolution equations as

$$\frac{\partial}{\partial t} n_t^I = \frac{D}{\Delta^2}\sum_\alpha \left\{ n_t^{I+\hat{\mathbf{e}}_\alpha} + n_t^{I-\hat{\mathbf{e}}_\alpha} - 2n_t^I \right\} + R'^I\left(\mathbf{n}_t\right) \tag{23}$$

where

$$R'^I\left(\mathbf{n}_t\right) = R^I\left(\mathbf{n}_t\right) - \frac{D}{\Delta^2}\sum_\alpha \left\{ n_t^{I+\hat{\mathbf{e}}_\alpha} + n_t^{I-\hat{\mathbf{e}}_\alpha} - 2n_t^I \right\} \tag{24}$$

At this point, the scheme is based on that of Nie et al [64, 65]. Let us consider first a closed system with periodic boundaries. In this case, the dynamical equations are transformed by making a discrete Fourier transform. Defining

$$\tilde{n}_t^K = \sum_I \exp\left(2\pi i \left(\frac{i_x k_x}{N_x} + \frac{i_y k_y}{N_y} + \frac{i_z k_z}{N_z} \right) \right) n_t^I \tag{25}$$

for $0 \le k_x < N_x$, etc. and making use of the periodic boundaries, it is easily seen that the dynamical equations become

$$\frac{\partial}{\partial t} \tilde{n}_t^K = \Lambda_K \tilde{n}_t^K + \tilde{R}'^K\left(\mathbf{n}_t\right) \tag{26}$$

with

$$\Lambda_K = \frac{4D}{\Delta^2}\left\{ \left(\cos\left(\frac{2\pi k_x}{N_x} \right) - 1 \right) + \left(\cos\left(\frac{2\pi k_y}{N_y} \right) - 1 \right) + \left(\cos\left(\frac{2\pi k_z}{N_z} \right) - 1 \right) \right\}. \tag{27}$$

This is just a linear transformation of the original system of equations and has the effect of diagonalizing the diffusive term. This allows the diffusive part of the right hand side to be integrated exactly giving

$$\tilde{n}^K_{t+\delta} = e^{\Lambda_K \delta}\tilde{n}^K_{t+\delta} + e^{\Lambda_K (t+\delta)}\int_t^{t+\delta} e^{-\Lambda_K \tau}\tilde{R}'^K\left(\mathbf{n}_\tau\right)d\tau \tag{28}$$

Finally, making the approximation

$$\tilde{R}'^K\left(\mathbf{n}_\tau\right) = \left(1-\frac{\tau - t}{\delta}\right)\tilde{R}'^K\left(\mathbf{n}_t\right) + \frac{\tau - t}{\delta}\tilde{R}'^K\left(\mathbf{n}_{t+\delta}\right) \tag{29}$$

and performing the integral gives an implicit set of equations for $\mathbf{n}_{t+\delta}$. This is therefore a generalization of the Crank-Nicholson scheme. These are solved by iteration in the obvious way and in general only four or five iterations are necessary to achieve a solution. Using this method, we are able to integrate the equations with a time step on the order of $10^{-2}\frac{\Delta^2}{D}$ with little difficulty. If it turns out that the iterative procedure is not converging or if it generates negative densities (in real space) then the time step is decreased and the relaxation step re-started.

Open systems are more complicated to implement. The solution used here is to hold the density constant on the boundary of the computational cell. This is a somewhat artificial constraint which, on the one hand, allows the number of particles within the cell to increase or decrease without constraint but on the other hand is clearly artificial since, e.g., one cannot convert the entire simulation cell to a uniform dense phase. Practically, if the simulation cell is sufficiently large that the cluster is far from the boundaries, this should have no effect on the nucleation stage of the process. Even in the post-critical growth stage, the artifice of the boundaries should only be important if the cluster grows sufficiently large.

The effect of holding the density constant on the boundaries is that the finite difference equations, Eq.(22) are modified with the right hand side being zero for points *I* on the boundary. These equations are no longer diagonalized by the discrete Fourier transform used for the open system, but rather, they are diagonalized by a sine transform

$$\tilde{n}^K_t = \sum_I \sin\left(i\pi\left(\frac{i_x k_x}{N_x} + \frac{i_y k_y}{N_y} + \frac{i_z k_z}{N_z}\right)\right) n^I_t . \tag{30}$$

The rest of the implementation is then the same.

Finally, we allow each image to have a different time step. This is no problem as it simply corresponds to using the rescaled-time version of the equations, Eq.(2) and allows us to adjust the time step according to the nature of the local density in each image.

2. Interpolation step

The distances between each pair of images is used to construct the overall distance of each image from the first one as

$$\alpha_{n+1} = s\left(\mathbf{n}_n ; \mathbf{n}_{n+1}\right) + \alpha_n \tag{31}$$

The goal of the interpolation is to replace image *n* by one for which the distance from the end first image is $\frac{n}{N}\alpha_N$ i.e. to give uniformly distributed images. The original string method made use of cubic spline interpolation but we found this unsuitable for two reasons. The first is that it sometimes produces negative densities. This can be resolved by interpolating something like the log of the density rather than the density itself. However, this runs into the second problem which is that in a closed system, mass must be conserved and this is not guaranteed for such a scheme. Hence,

to resolve both problems, I have used simple linear interpolation. For each distance *(n/N)-α_N* an index *m* is identified such that $\alpha_m \le (n/N)\alpha_N \le \alpha_{m+1}$ and then the new density is

$$\mathbf{n}_n = \frac{\alpha_{m+1} - \frac{n}{N}\alpha_N}{\alpha_{m+1} - \alpha_m}\bar{\mathbf{n}}_m + \frac{\frac{n}{N}\alpha_N - \alpha_m}{\alpha_{m+1} - \alpha_m}\bar{\mathbf{n}}_{m+1} \tag{32}$$

Clearly, if the two density distributions $\bar{\mathbf{n}}_m$ and $\bar{\mathbf{n}}_{m+1}$ are always positive, than the result is also. Furthermore, the overall mass is conserved. Since linear interpolation is not necessarily very accurate, the actual distances between the interpolated images are somewhat different than what was intended. To correct for this, the procedure is iterated 5 times at which point the desired distances are obtained.

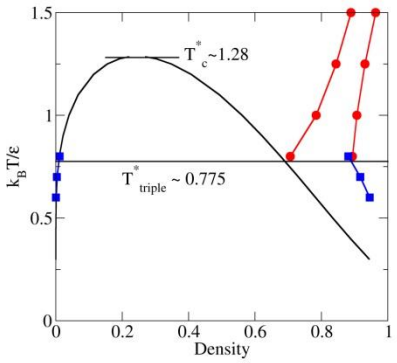

Fig. S1. Phase diagram as presented in (*50*).

3. Convergence criterion

Convergence is monitored using a criterion based on that used in Ref. [36]. Let $N_i^J = n_i^J \Delta V$ be the number of molecules in cell *i* of image *J* and let $N^J = \sum_i N_i^J$ be the total number of molecules in image *J*. The convergence is monitored by computing the "velocities"

$$\Delta^J(t+\delta) = \frac{\sigma^2}{D}\frac{1}{\delta}\frac{\sqrt{\sum_i \left(N_i^J(t+\delta) - N_i^J(t)\right)^2}}{N^J} \tag{33}$$

which is a measure of the rate of change of the image and our convergence criterion is $\max_J\left(\Delta^J(t)\right) < \epsilon$ for some threshold . Typically, this quantity was on the order of 1 at the start of the calculation and the convergence threshold was typically taken to be$\epsilon = 10^{-4}$. Some calculations were run with$\epsilon = 10^{-6}$ to verify that the former criterion was sufficient.

II. CALCULATIONAL DETAILS

The potential and numerical lattice

The calculations reported here are for particles interacting via a Lennard-Jones potential

$$v_{\mathrm{LJ}}(r) = 4\varepsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right), \tag{34}$$

which is cutoff and shifted at $R_c = 3\sigma$ so that the potential used was

$$v(r) = v_{\mathrm{LJ}}(r) - v_{\mathrm{LJ}}(R_c), \quad r < R_c \tag{35}$$

and zero for $r > R_c$. The phase diagram has been calculated using the DFT and is shown in Fig. 1. It displays a weak-solution, vapor-like, phase, a strong-solution, liquid-like , phase and an (FCC) crystalline phase.

For the calculations reported in the text, the calculational cell extended 25σ in each Cartesian direction and the calculational grid had spacing $\sigma/7$ giving a total of 175^3 grid points in total. For the DFT calculations, all details are as described in Ref. [50].

Thermodynamic input

The free energy of the uniform (weak or strong) is determined by evaluating the DFT functional for a constant density giving the grand-canonical free energy per unit volume at inverse temperature β and chemical potential μ as

Table S1. Thermodynamic quantities at weak-solution/dense-solution liquid-liquid coexistence.

Temperature k_BT/ε	0.475
Concentration of dense solution at coexistence $n_l^{(coex)}\sigma^3$	0.847195
Concentration of weak solution at coexistence $n_v^{(coex)}\sigma^3$	1.06228×10^{-4}
Surface tension $\beta\gamma\sigma^2$	2.07115
Chemical potential $\beta\mu^{(coex)}$	−9.15268
Grand canonical free energy per unit volume $\beta\omega^{(coex)}\sigma^3$	-1.06081×10^{-4}

Table S2. Thermodynamic quantities used in liquid-liquid calculations.

Temperature k_BT/ε	0.475
Concentration of weak solution in calculation $n_v\sigma^3$	3.39446×10^{-4}
Concentration of dense solution with same chem potential $n_l\sigma^3$	0.87541
Chemical potential $\beta\mu$	−7.997
Dense solution free energy per unit volume $\beta\omega_l\sigma^3$	−0.996107
Weak solution free energy per unit volume $\beta\omega_v\sigma^3$	-3.37952×10^{-4}

being

$$\beta\omega(n;\beta,\mu) = n\ln n - n + n\frac{\eta(4-3\eta)}{(1-\eta)^2} + \frac{1}{2}an^2 - \mu n \tag{36}$$

where the packing fraction is $\eta=(\pi/6)nd^3$ and the effective hard-sphere diameter is calculated using the Barker-Henderson prescription

$$d = \int_0^{r_0}\left(1 - e^{-\beta(v(r)-v(r_0))}\right)dr \tag{37}$$

with $r_0 = 2^{1/6}\sigma$ being the minimum of the potential. The contribution of the attractive part of the potential is the third term where

$$a = \int w(r)\,d\mathbf{r} \tag{38}$$

with $w(r) = v(r_0)$ for $r < r_0$ and $w(r) = v(r)$ for $r > r_0$. To be completely consistent with the numerical DFT calculations, this integral must be performed using the same numerical integration scheme as used in the case of inhomogeneous systems, notwithstanding the fact that the integral is easy to perform analytically.

The thermodynamic data at coexistence used to make the comparison to CNT for the dense-solution droplet with 290 particles is easily determined from the equation of state and is given in Table I and in particular one finds that the supersaturation is $S=(n_v-n_v^{coex})/\ n_v^{coex}$ =2.2.

Supersaturation and free energies of the critical droplets

The supersaturation versus the size of the critical clusters is shown in fig. S2.

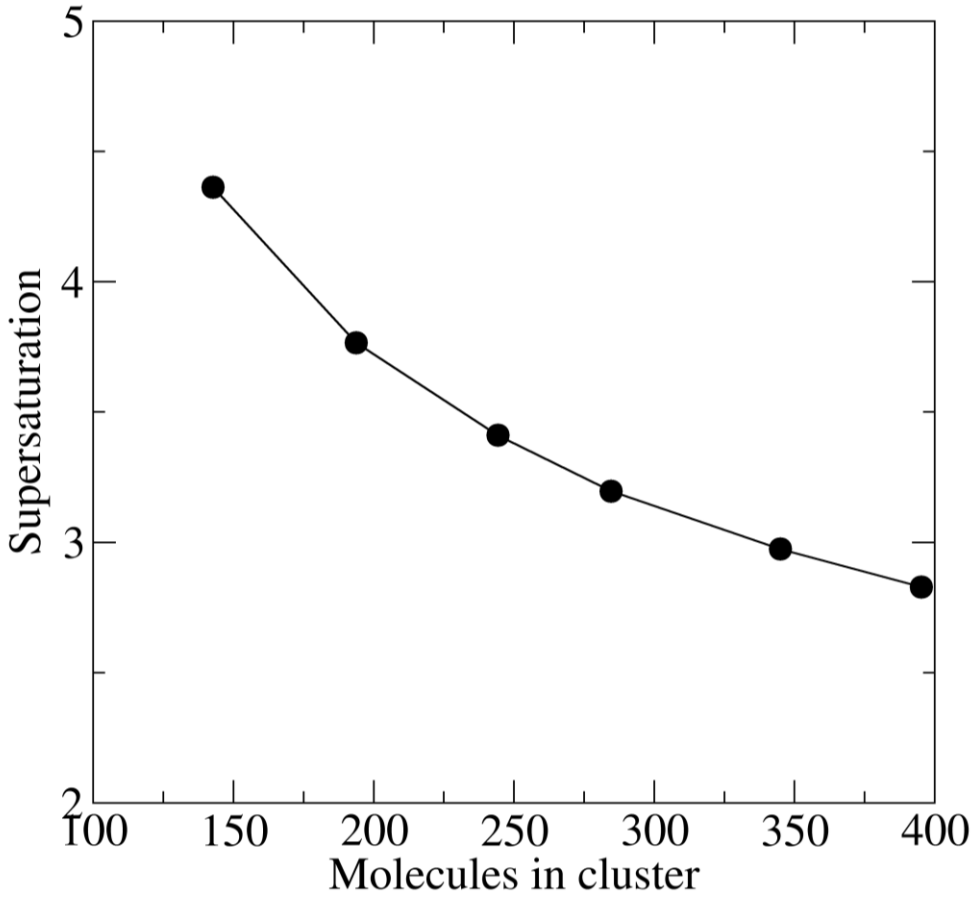

Fig. S2. Supersaturation as a function of the excess number of particles for liquid critical droplets.

III. EFFECT OF A PLATEAU ON THE NUCLEATION RATE

In order to provide some intuition as to the effect of a plateau on the nucleation rate, consider a one-dimensional model in which a particle diffuses under the influence of a potential, $\beta\Delta F(x)$. Assume that the potential has a minimum at $x = 0$ and a maximum at $x = X_0$ where the minimum at $x = 0$ represents the initial phase and the maximum at X_0 is the critical cluster. If the reaction coordinate, x, is the radius of the cluster this maps directly onto CNT but here there is no need to be so specific. Assuming that x moves diffusively under the effect of white noise, the mean first passage time to pass over $x = X_0$ (which is inversely related to the nucleation rate) is given by

$$\langle \tau \rangle_x = \frac{1}{D} \int_x^{X_0} e^{\beta \Delta F(z)} \left(\int_0^z e^{-\beta \Delta F(y)} dy \right) dz. \tag{39}$$

This expression is completely general, given the boundary conditions. Let us suppose that instead of X_0 being a maximum, it is the beginning of a plateau, so that $\beta\Delta F(x) = \beta\Delta F(X_0) \equiv \Delta\beta F_0$ for $X_0 \le x \le X_1$ and for $x > X_1$ the free energy decreases rapidly. Then, the relevant mean first passage time is that to reach X_1 starting at $x = 0$ which is

$$\langle \tau \rangle_x = \frac{1}{D} \int_x^{X_1} e^{\beta \Delta F(z)} \left(\int_0^z e^{-\beta \Delta F(y)} dy \right) dz \tag{40}$$

or, separating the domains

$$\begin{aligned} \langle \tau \rangle_x &= \frac{1}{D} \int_x^{X_0} e^{\beta \Delta F(z)} \left(\int_0^z e^{-\beta \Delta F(y)} dy \right) dz + \frac{1}{D} \int_{X_0}^{X_1} e^{\beta \Delta F(z)} \left(\int_0^z e^{-\beta \Delta F(y)} dy \right) dz \\ &= \frac{1}{D} \int_x^{X_0} e^{\beta \Delta F(z)} \left(\int_0^z e^{-\beta \Delta F(y)} dy \right) dz + \frac{1}{D} \int_{X_0}^{X_1} e^{\beta \Delta F(z)} \left(\int_0^{X_0} e^{-\beta \Delta F(y)} dy + \int_{X_0}^z e^{-\beta \Delta F(y)} dy \right) dz \\ &= \frac{1}{D} \int_x^{X_0} e^{\beta \Delta F(z)} \left(\int_0^z e^{-\beta \Delta F(y)} dy \right) dz + \frac{1}{D} (X_1 - X_0) e^{\beta \Delta F_0} \left(\int_0^{X_0} e^{-\beta \Delta F(y)} dy \right) + \frac{1}{2D} (X_1 - X_0)^2 \end{aligned} \tag{41}$$

To make contact with the standard analysis, I assume that for $x < X_0$, $F(x)$ has the same form as is assumed not only in the simplest version of CNT but also in the multidimensional analyses of Langer, Talkner, etc., namely that for x less than, but near, X_0 one has that $F(x) \simeq F_0 - (1/2)(x-X_0)^2 |F''(X_0)|$ so that, assuming also $\beta F_0 \gg$ the first term can be approximated by

$$\begin{aligned} \frac{1}{D} \int_x^{X_0} e^{\beta \Delta F(z)} \left(\int_0^z e^{-\beta \Delta F(y)} dy \right) dz &\simeq \frac{1}{D} \int_x^{X_0} e^{\beta \Delta F_0 - \frac{1}{2}(z-X_0)^2 |F''(X_0)|} \left(\int_0^{X_0} e^{-\beta \Delta F(y)} dy \right) dz \\ &\simeq \sqrt{\frac{\pi}{2|\beta F''(X_0)|}} e^{\beta \Delta F_0} \left(\int_0^{X_0} e^{-\beta \Delta F(y)} dy \right) \end{aligned} \tag{42}$$

The largest contribution from the remaining integral comes from y near zero so expanding again

$$\int_x^{X_0} e^{\beta \Delta F(z)} \left(\int_0^z e^{-\beta \Delta F(y)} dy \right) dz \simeq \frac{\pi}{2\sqrt{\beta F''(0) |\beta F''(X_0)|}} e^{\beta \Delta F_0} \tag{43}$$

The same analysis of the second integral gives the result

$$\langle \tau \rangle_x = \frac{\pi}{2D\sqrt{\beta F''(0)\left|\beta F''(X_0)\right|}} e^{\beta \Delta F_0} + \frac{1}{D}(X_1 - X_0)\sqrt{\frac{\pi}{2\beta F''(X_0)}} e^{\beta \Delta F_0} + \frac{1}{2D}(X_1 - X_0)^2 \tag{44}$$

For $\beta\Delta F_0 \gg 1$ the last term is negligible leaving

$$\langle \tau \rangle_x = \frac{\pi}{2D\sqrt{\beta F''(0)\left|\beta F''(X_0)\right|}} e^{\beta \Delta F_0} \left(1 + (X_1 - X_0)\sqrt{\frac{2}{\pi}\left|\beta F''(0)\right|}\right) \tag{45}$$

so the effect of a plateau is to increase the mean first passage time by a quantity that depends on the length of the plateau but that in all likelihood is small compared to the exponential prefactor.

IV. ADDITIONAL RESULTS

Comparison to CNT

The CNT (i.e. capillary model) free energy for an open system is

$$\beta\Delta\Omega = (\beta\omega_l - \beta\omega_v)\frac{4\pi}{3}R^3 + \beta\gamma\left(1 + \frac{l}{R}\right)4\pi R^2 \tag{46}$$

where *l* is the Tolman length (which is equal to zero in the simplest models). Using $\Delta N = (4\pi/3)R^3(n_l - n_v)$ this can also be written as

$$\Delta\Omega = \frac{\beta\omega_l - \beta\omega_v}{n_l - n_v}\Delta N + \beta\gamma(4\pi)^{1/3}\left(\frac{3}{n_l - n_v}\right)^{2/3}\Delta N^{2/3} + l\beta\gamma(4\pi)^{2/3}\left(\frac{3}{n_l - n_v}\right)^{1/3}\Delta N^{1/3} \tag{47}$$

When the Tolman length is zero, the radius and energy barrier are

$$\begin{aligned} R_c(l=0) &= -\frac{2\beta\gamma}{\beta\omega_l - \beta\omega_v} \\ \beta\Delta\Omega_c &= \frac{16\pi}{3}\frac{(\beta\gamma)^3}{(\beta\omega_l - \beta\omega_v)^2} \end{aligned} \tag{48}$$

while in general

$$R_c(l) = R_c(0)\left(\frac{1 + \sqrt{1 + \frac{1}{2}R_c(0)l}}{2}\right) \tag{49}$$

For the numbers given in the tables, these formula yield

$$\begin{aligned} \beta\Delta\Omega &= -1.44\Delta N + 10.95\Delta N^{2/3} + 16.88\Delta N^{1/3} \\ R_c\left(l=0\right) &= 4.16\sigma \\ N_c\left(l=0\right) &= 264 \end{aligned} \tag{50}$$

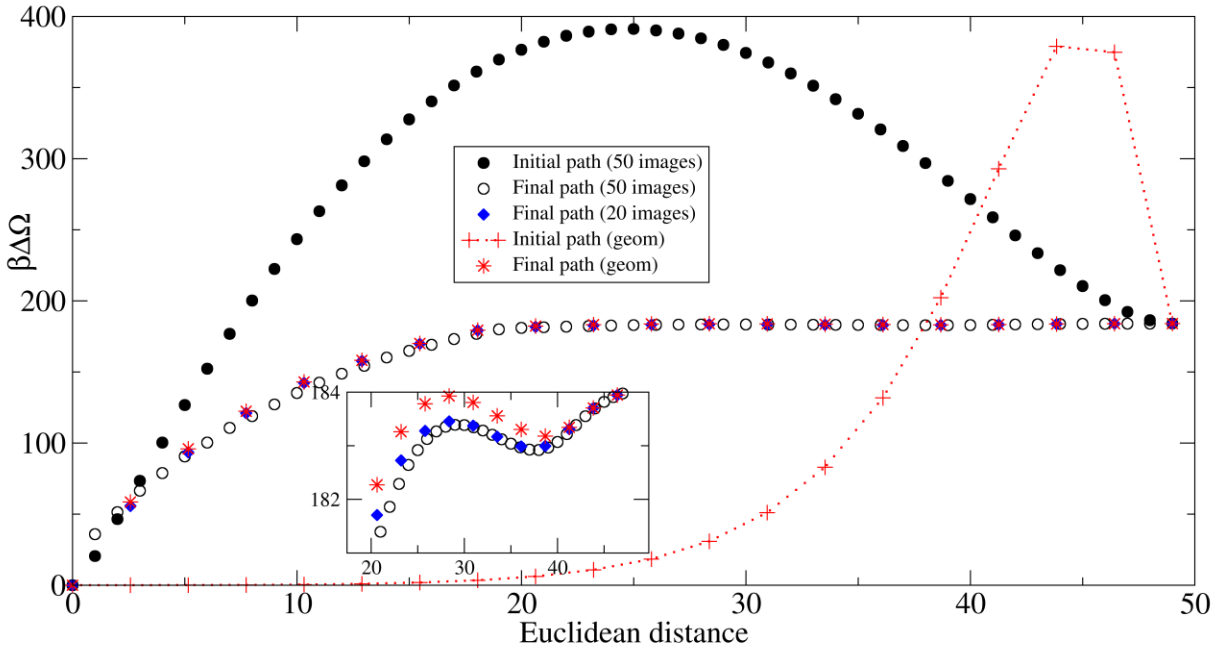

Fig. S3. Initial and final paths for crystallization with a final cluster of 286 molecules using 50 images and a final path using 20 images showing the excess free energy as a function of Euclidean distance along the path (arbitrarily scaled so that the distance between images is equal to one). Also shown are results beginning with a different initial condition based on geometric, rather than arithmetic, interpolation between the initial and final states (labeled "geom"). The free energies are broadly consistent over the entire path and, as can be seen in the inset, this is true even for the weak structure as the solid cluster begins to grow.

Cluster radius and density

The radius of the clusters was taken to be

$$R^2 = \frac{5}{3}\frac{\int \Delta n(\mathbf{r}) r^2 d\mathbf{r}}{\int \Delta n(\mathbf{r}) d\mathbf{r}} \tag{51}$$

where $\Delta n(\mathbf{r}) = n(\mathbf{r}) - n_0$ and n_0 is the value of the fixed density on the boundaries of the cell. The numerical prefactor was chosen because it gives the correct result for a spherical cluster with a constant interior density. The average cluster density was then defined as the average density of all points within the ball of radius R.

Effect of initial condition and stability of the calculations

In order to demonstrate how much the final paths differ from the initial guess, the free energies along the initial and final paths for crystallization of the 286 molecule cluster are shown in fig. S3. Note that the initial images are constructed so as to be equally spaced according to the Euclidean distance so that it is the relaxation that generates the movement away from the initial path. A sequence of snapshots of the process are shown in fig. S4 which can be compared to the figures in the main text to illustrate again the difference in the pathways. The figure also shows results for an initial path based on a completely different construction (geometric rather than linear interpolation between the end points). The latter converges to the same results as in the other calculations thus illustrating the robustness of the pathways.

Additional Pathways

Figures S5 and S6 show several nucleation pathways plotted as excess free energy and excess cluster mass, respectively. It is seen that the resolution of the process varies widely with the critical cluster occuring as the final image in some cases and very early along the path in others. In the latter cases, further resolution could be obtained by repeating the calculation using the obtained critical cluster as the endpoint. The example used in the main text is the system of 290 particles and it was chosen because it is one of the best resolved. In all cases, a plateau exists in the cluster mass during which period the ordering takes place.

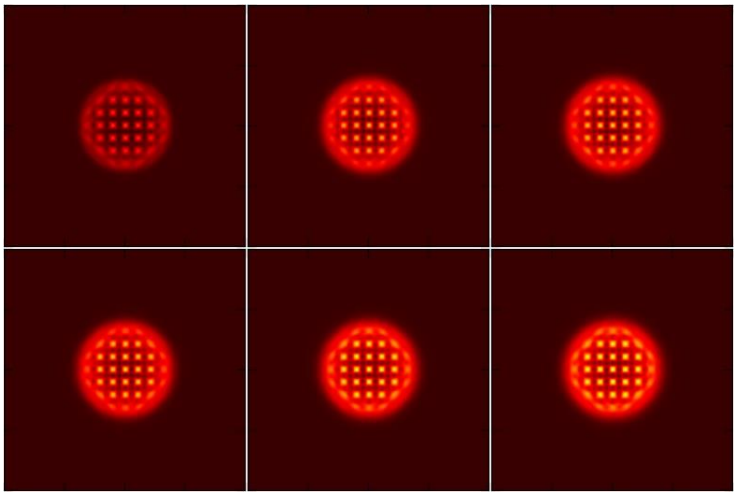

Fig. S4. Log of the density for slices of images from the initial guess to the pathway. The images correspond to the insets in Fig. 4 in the main text and were made with the same color scale so that they can be directly compared with that (final) result. Note that for the initial path, the solid structure is clearly visible even for the first step along the pathway whereas for the final result shown in the main text, the structure is liquid-like over the first half of the

pathway.

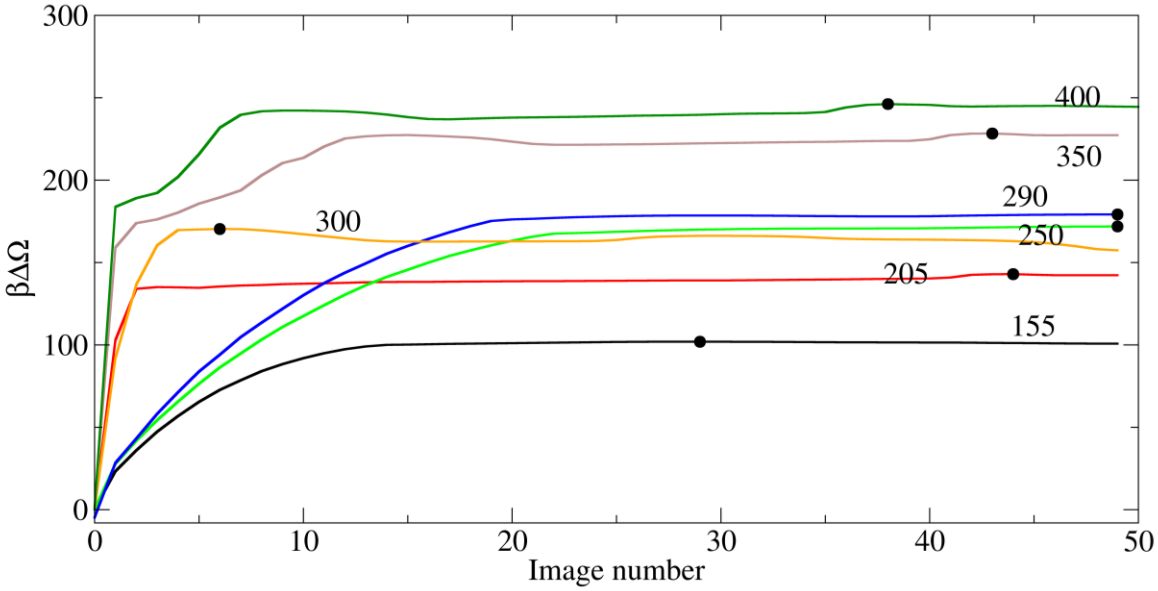

Fig. S5. Nucleation pathways for several systems displayed in terms of the free energy along the paths. Each curve is labeled with the total number of particles in the final image. The black dots indicate the maximal excess free energy, i.e. the critical cluster.

Figures S7 and S8 show the size and excess free energy of the critical clusters as a function of supersaturation. Note that in most cases, the critical cluster is obtained from the pathways and is not the end point. By chance, two systems (having 250 and 300 particles, respectively, in the final images) have almost exactly the same supersaturation. Remarkably, the siwe and energy of the critical clusters obtained for the two cases are almost identical even though in the former case, the critical cluster is the final image, which was held fixed during the path determination, whereas in the latter case, it is dynamically generated near the middle of the pathway. The close agreement in this case is a strong indication of the robustness of the procedure used to generate the paths.

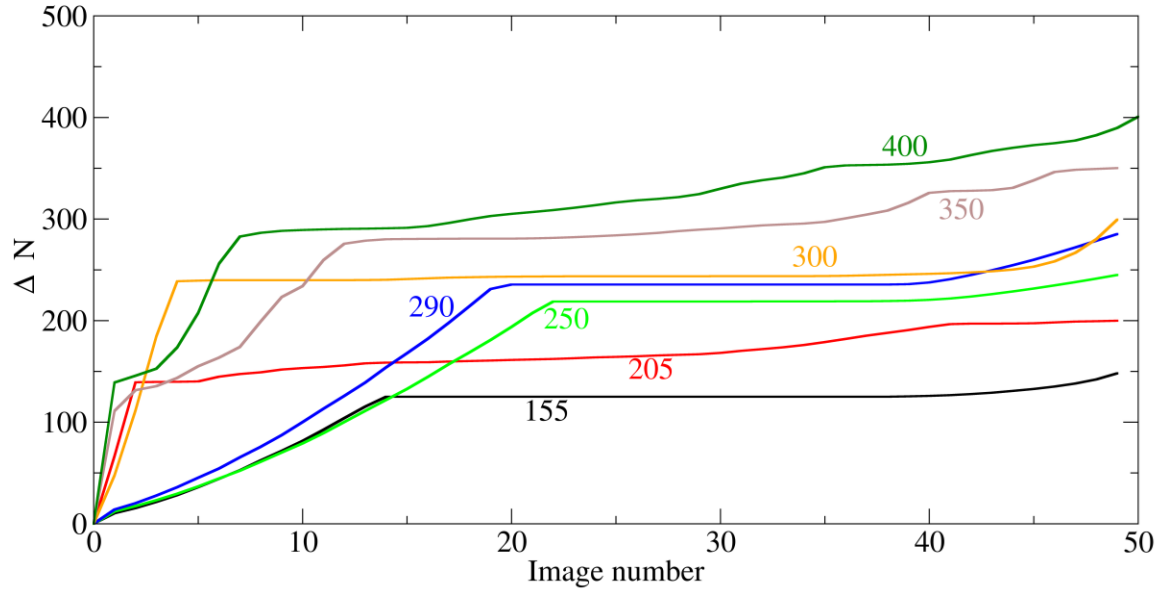

Fig. S6. Nucleation pathways for several systems displayed in terms of the number of molecules in the clusters along the paths. Each curve is labeled with the total number of particles in the final image.

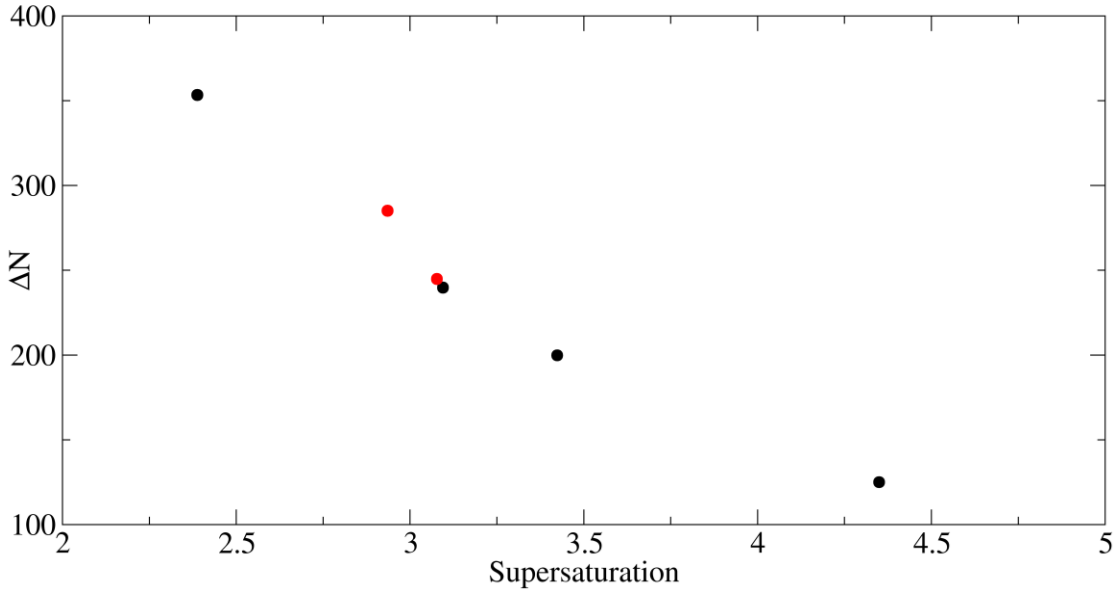

Fig. S7. Excess number of particles in the critical clusters as a function of supersaturation. For the black points, the final cluster was obtained starting with minimum energy clusters as described above wheras the red points used clusters that spontaneously formed by quenching liquid clusters in DFT calculations.

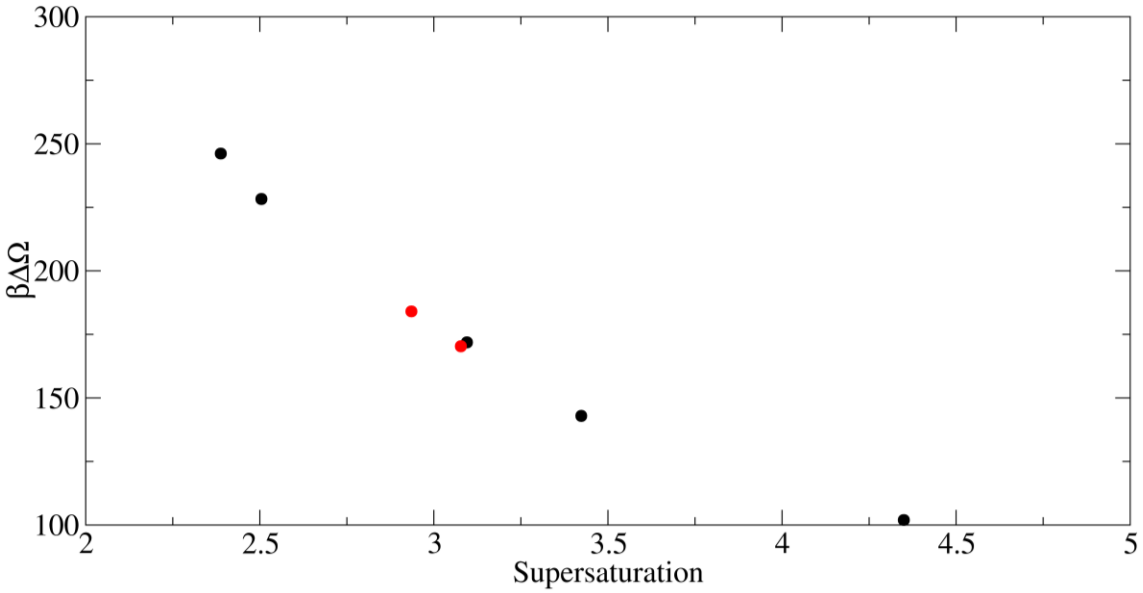

Fig. S8. Excess free energies in the critical clusters as a function of supersaturation. For the black points, the final cluster was obtained starting with minimum energy clusters as described above wheras the red points used clusters that spontaneously formed by quenching liquid clusters in DFT calculations.