Combinatoric and mean-field analysis of heterogeneous self-assembly

Bingyu Zhao[1], Bijan Berenji[2], Tom Chou[2], Maria R. D'Orsogna[2,3]
[1]*Division of Applied Mathematics, Brown University, Providence, RI 02912*
[2]*Depts. of Biomathematics and Mathematics, UCLA, Los Angeles, CA, 90095-1766 and*
[3]*Dept. of Mathematics, CSUN, Los Angeles, CA 91330-8313*

(Dated: March 15, 2013)

We analyze a fully stochastic model of heterogeneous nucleation and self-assembly in a closed system with a fixed total particle number M, and a fixed number of seeds $N_{\rm s}$. Each seed can bind a maximum of N particles. A discrete master equation for the probability distribution of the cluster sizes is derived and the corresponding cluster concentrations are found in terms of the density of seeds, the total mass, and the maximum cluster size. The heterogeneous stochastic self-assembly process is also analyzed using kinetic Monte-Carlo simulations. Our analytic and numerical findings are compared with those obtained from classical mass–action equations. We analyze the discrepancies between the stochastic and mass-action results as a function of model parameters.

PACS numbers: 82.60.Nh,02.30.Hq,05.70.Ln

I. INTRODUCTION

The self-assembly of molecules and macroscopic particles into larger units is a common process in materials science and cell biology[1]. In homogeneous nucleation identical components are able to spontaneously self-assemble to form; however, in many cases, the growth process may be catalyzed or even triggered by "seeds" such as an impurity, an exogenous particle or a boundary. Such seeds tend to lower the free energy barrier for particle aggregation so that heterogeneous nucleation is typically more commonly observed than homogeneous nucleation[2].

Within structural biology, a long standing issue has been that of identifying a "universal nucleant" to induce the rapid growth of protein crystals suitable for X-ray diffraction to determine the protein's 3D structure[4]. Conversely, the formation of large aggregates of insulin and other proteins is problematic in drug preparation, delivery and storage[5]. Polymerization of various proteins and polypeptides into amylodid fibers are also implicated in the emergence of neurodegenerative disorders such as Parkinson's, Alzheimer's and prion diseases. The typical mechanism through which proteins self assemble in all these biological examples is by monomers slowly forming an intermediate size fiber of few units, which then acts a nucleation site for accelerated absorption of further units[6,7].

In this paper we will be concerned with systems where heterogeneous self-assembly occurs in small compartments of finite volumes, such as cells and organelles. This assumption is appropriate when particle aggregation is much faster than the typical times for monomers or seeds to be synthesized or degraded. Moreover, stoichiometry typically prevents clusters from growing indefinitely. After a maximum size is reached, the self-assembly process is completed or the dynamics changes. Examples of such maximum binding limits include oxygen binding to to a single hemoglobin protein ($N = 4$), self-assembly of membrane peptides to form pores ($N \approx 6-8$)[8], self assembly of capsid proteins to form virus capsids ($N \sim 100-1000$)[9], or assembly of clathrin triskelion proteins to form the clathrin-coated pits that arise in endocytosis ($N \sim 25-50$)[10].

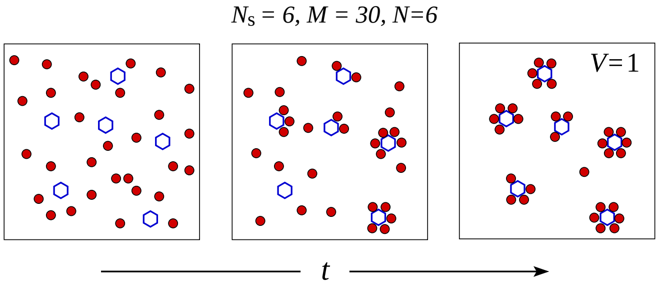

FIG. 1: A schematic of the heterogeneous self-assembly process in a closed system. The open hexagons represent seed particles on which the monomers (filled circles) aggregate. In this example, the total mass, the number of seed particles, and the maximum cluster size are $M = 30$, $N_{\rm s} = 6$, and $N = 6$, respectively.

Thus, our problem is described by a self-assembly process in a closed system with a total of M particles that can bind $N_{\rm s}$ seeds, each of which can accommodate a maximum of N particles. Given the discreteness of the system and possible finite size effects, we will consider a stochastic treatment and our results will be compared to those derived through classical mean–field equations.

The classical mass–action equations for heterogeneous nucleation under fixed $\{M, N_{\rm s}, N\}$ were previously analyzed in (Chou and D'Orsogna, 2011), where both limits of reversible and irreversible monomer attachment and detachment were considered. In this paper we will present the corresponding master equation for the probability distribution of cluster sizes, and from these, derive mean cluster concentrations to be directly compared to those obtained in (Chou and D'Orsogna, 2011).

We have performed a similar comparison in the case of homogeneous nucleation where stochastic and mean–field treatments were shown to yield remarkably different results at equilibrium, especially when M and N were of the same order of magnitude and in the limit of small detachment[15]. The origin of the discrepancy was identified in the non-commensurability between M and N, so that when M was not a multiple of N, finite-size effects not captured by the mass action equations could be quite striking in the stochastic system.

In the heterogeneous problem, we will show that more subtle discrepancies between results derived from the stochastic and mean-field approaches arise. In the next section we will give a brief overview for the classical mass actions equations for heterogeneous nucleation, as derived in (Chou and D'Orsogna, 2011). In Section III we will introduce the corresponding Master equation and derive the average cluster sizes for comparison with the mean–field values. Analytical and numerical results are discussed in Section IV and V, respectively.

II. MASS-ACTION KINETICS

In this section we briefly recapitulate results from (Chou and D'Orsogna, 2011) that will be used for comparison with the stochastic results that will be later derived in Section III. We derive mass–action equations for a system of total fixed number M of bound and unbound monomers and a fixed number $N_{\rm s}$ of seeds where each seed can accommodate at the most N monomers. Fragmentation and aggregation processes that do not involve monomers are neglected.

Following conventional notation, we denote by $c_k(t)$ the number of clusters of size k at time t. The attachment of monomers to a cluster of size k depends on an intrinsic rate p_k and on the total number of free monomers $m(t)$, while detachment from clusters occurs at a rate q_k. The mass–action equations for $c_k(t)$ are thus written as

$$\begin{aligned}\dot{c}_0 &= -p_0 m(t) c_0 + q_1 c_1,\\ \dot{c}_k &= -p_k m(t) c_k - q_k c_k + p_{k-1} m(t) c_{k-1} + q_{k+1} c_{k+1},\\ \dot{c}_N &= -q_N c_N + p_{N-1} m(t) c_{N-1}, \end{aligned} \qquad (1)$$

where the number of free monomers is contrained by

$$m(t) \equiv M - \sum_{k=1}^{N} k c_k(t), \qquad (2)$$

and where conservation of seeds requires

$$N_{\rm s} = \sum_{j=0}^{N} c_j(t). \qquad (3)$$

Initial conditions are chosen so that $m(t=0) = M, c_0(t=0) = N_{\rm s}, c_{k>0}(t=0) = 0$. Eqs. 1 are analogous to the Becker-Döring equations commonly used to describe homogeneous nucleation[15]. Here, we restrict ourselves to the case of constant detachment rates that are much smaller than the constant monomer attachment rates ($q_k = q \ll p_k = p$. We will analyze results for both the reversible limit and the strictly singular, irreversible limit $q = 0$.

The long-time behavior of this process will depend critically on whether there is an excess or deficiency of monomers. An important parameter will be the quantity $\sigma \equiv M/NN_{\rm s}$. When $\sigma < 1$, there is an excess of seeds and, in the irreversible case ($q = 0$), monomers will be depleted before all seeds can be completely filled. At $t \to \infty$, a distribution of partially completed clusters will arise. For cases where $\sigma \geq 1$ and detachment is slow, all seeds will be populated to nearly full capacity N with approximately $M - NN_{\rm s}$ free, unattached monomers remaining. While the specific cluster size distributions depend independently on $M, N_{\rm s}$, and N, we found that their overall qualitative features are most sensitive to the magnitude of the combination $\sigma \equiv M/(N_{\rm s}N)$.

In the strictly irreversible case of $q = 0$, the choice $\sigma \geq 1$ implies that all seeds are fully occupied at $t \to \infty$ so that $c_N(t\to\infty) = N_{\rm s}$, $c_{k\neq N}(t \to \infty) = 0$ and $m(t \to \infty) = M - NN_{\rm s}$. However, if $\sigma < 1$, a finite time t^* exists at which the pool of free monomers is depleted, $m(t^*) = 0$, and the system stops evolving. The final concentrations $c_k^*/N_{\rm s}$ were found to be[11]

$$\frac{c_{k<N}(\xi)}{N_{\rm s}} = \frac{\xi^k e^{-\xi}}{k!}, \quad \frac{c_N(\xi)}{N_{\rm s}} = 1 - \sum_{j=0}^{N-1} \frac{\xi^j e^{-\xi}}{j!}, \qquad (4)$$

where ξ is determined by the real root of the transcendental equation $\xi^N e^{-\xi} + (N-\xi)\Gamma(N,\xi) = (1-\sigma)N\Gamma(N)$.

For the case of reversible binding, we will stay focussed on the small $\varepsilon \equiv q/p$ regime, where monomers bind strongly to the clusters. In this limit, the concentrations $c_k(t)$ first approach values close to c_k^* before slow detachment eventually allows monomer redistribution and equilibration to a new cluster size distribution after a time scale $t \gg q^{-1}$. These *equilibrium* cluster concentrations, $c_k^{\rm eq}$, can be found by keeping $q > 0$ and setting the left hand side of Eqs. 1 to zero. Upon solving the resulting algebraic equations along with Eqs. 2 and 3, we find

$$\frac{c_k^{\rm eq}}{N_{\rm s}} \equiv \frac{(z-1)z^k}{z^{N+1}-1} \qquad (5)$$

where $\varepsilon \equiv q/p$ and z satisfies

$$\left(\frac{\varepsilon z}{N_{\rm s}N} - \sigma\right)(z-1)(z^{N+1}-1) + z^{N+2} - \left(1 + \frac{1}{N}\right) z^{N+1} + \frac{z}{N} = 0.$$

Note that since ε multiplies the highest power of the fugacity z in Eq.6, that $\varepsilon \to 0^+$ constitutes a singular limit.

When $\sigma < 1$, not all binding sites can be filled, and approximations for the root z can be found for $\sigma \ll 1$ and $\sigma \approx 1/2$. In (Chou and D'Orsogna, 2011), we performed numerical estimates of Eq. 6 for $\sigma < 1$ showing that for $1/2 < \sigma < 1$, Eq. 6 yields $z < 1$ implying $c_{k+1}^{\rm eq} > c_k^{\rm eq}$ and that smaller clusters tend to be favored. On the other hand, for $\sigma < 1/2$, $z > 1$, $c_{k+1}^{\rm eq} < c_k^{\rm eq}$ and larger clusters are favored. For $\sigma = 1/2$, $z = 1$ and all cluster sizes are equally populated.

In the excess seed case, for $\sigma < 1$, only the metastable values c_k^* attained during the reversible dynamics for $\varepsilon \to 0^+$ are well approximated by the irreversible results, while the later equilibration values $c_k^{\rm eq}$ obtained by taking the $\varepsilon \to 0^+$ limit at $t \to \infty$ can be quite different from the metastable values c_k^* obtained by directly setting $\varepsilon = 0$ in Eqs. 1. As an example, we can consider the case $M = 5$, $N_{\rm s} = 2$ and $N = 3$, for which $\sigma = 5/6$ and Eqs. 1 yield increasing values of $c_k^{\rm eq}$: $c_0^{\rm eq} = 0.063$, $c_1^{\rm eq} = 0.173$, $c_2^{\rm eq} = 0.473$, $c_3^{\rm eq} = 1.291$.

When there are excess monomers ($\sigma = M/(N_{\rm s}N) > 1$), all binding sites will be nearly always filled and

$$c_k^{\rm eq} \approx \frac{N_{\rm s}}{(N_{\rm s}N)^{N-k}} \frac{\varepsilon^{N-k}}{(\sigma - 1)^{N-k}} + O(\varepsilon^{N-k+1}). \qquad (6)$$

Here the difference between reversible and irreversible binding kinetics vanishes since $c_{k\neq N}^{\rm eq} \approx c_{k\neq N}^* \to 0$ and $c_N^{\rm eq} \approx c_N^* \to N_{\rm s}$ in the $\varepsilon \to 0^+$ limit.

III. MASTER EQUATION FOR HETEROGENEOUS SELF-ASSEMBLY

We now introduce the master equation for our discrete heterogeneous self-assembly. Denote by $P(\{n\};t) \equiv P(m|n_0, n_1, ..., n_N; t)$ the probability distribution function for the system to be in a state with m free monomers, n_0 unbound seeds, and n_i $(1 \leq i \leq N)$ seeds with i bound monomers. Since each seed can bind at most N particles, the sequence is arrested at n_N. Using the same notation as in the previous section for the attachment and detachment rates p_k and q_k respectively, we can write the full master equation as

$$\begin{aligned}\dot{P}(\{n\};t) \; = \; & -m\sum_{i=0}^{N-1} p_i n_i P(\{n\};t) - \sum_{i=1}^{N} q_i n_i P(\{n\};t) \qquad (7)\\ & +(m+1)\sum_{i=0}^{N-1} p_i (n_i+1) W_*^+ W_i^+ W_{i+1}^- P(\{n\};t) + \sum_{i=1}^{N} q_i (n_i+1) W_*^- W_{i-1}^- W_i^+ P(\{n\};t),\end{aligned}$$

where we have implicity assumed that $P(\{n\};t) = 0$ if, for any i, $n_i < 0$ or $m < 0$. The $W_i^\pm$ and $W_*^\pm$ terms represent the unit raising or lowering operators on the number n_i of clusters of size i and on the number of free monomers m, respectively. For example, the operator $W_*^+ W_i^+ W_{i+1}^-$ acting on state $P(\{n\}, t)$ is defined as

$$\begin{aligned}W_*^+ W_i^+ W_{i+1}^- P(\{n\};t) \; \equiv & \qquad (8)\\ & P(m+1|n_0, ..., n_i+1, n_{i+1}-1, ..., n_N; t).\end{aligned}$$

As in our analysis of the mass-action kinetics, we will assume that monomer binding and unbinding occur at constant, cluster size-independent rates p and q, respectively, and rescale time in units of p^{-1}. The form of the master equation is identical to that in Eq. 7 except with the replacements $t \to p^{-1}t$, $p_k \to 1$, and $q_k \to \varepsilon \equiv q/p$. Note that the stochastic dynamics described by Eq. 7 obeys total mass conservation

$$M = m + \sum_{k=1}^{N} k n_k, \qquad (9)$$

and a total cluster number constraint

$$N_{\rm s} = \sum_{k=0}^{N} n_k. \qquad (10)$$

Eqs. 9 and 10 are the discrete counterparts to the mass-action equation constraints Eqs. 2 and 3. We assume that all the monomers are free at $t = 0$ so that

$$P(\{n\}; t=0) = \delta_{m,M}\delta_{n_0,N_{\rm s}}\delta_{n_1,0}\cdots\delta_{n_N,0}. \qquad (11)$$

In order to compare results arising from Eq. 7 to the ones derived from the mean-field Eqs. 1 we define the mean number of clusters of size k as

$$\langle n_k(t)\rangle = \sum_{\{n\}} n_k P(\{n\};t). \tag{12}$$

The values of $\langle n_k(t)\rangle$ derived from the full stochastic treatment in Eq. 12 are the direct counterparts to the mean–field approximation to $c_k(t)$ found by solving Eqs. 1. This can be most easily seen by by multiplying Eq. 7 by n_k and by summing over all possible states to give

$$\begin{aligned}
\langle \dot{n}_0(t)\rangle &= -\langle mn_0\rangle + \varepsilon\langle n_1\rangle \\
\langle \dot{n}_k(t)\rangle &= -\langle mn_k\rangle + \langle mn_{k-1}\rangle - \varepsilon(\langle n_k\rangle - \langle n_{k+1}\rangle) \\
\langle \dot{n}_N(t)\rangle &= \langle mn_{N-1}\rangle - \varepsilon\langle n_N\rangle,
\end{aligned} \tag{13}$$

where $\langle mn_k\rangle \equiv \sum_{\{n\}} mn_k P(\{n\};t)$ represent monomer-cluster correlations. If we further assume that the monomer and cluster numbers are uncorrelated ($\langle mn_k\rangle = \langle m\rangle\langle n_k\rangle$), and identify $m \equiv \langle m\rangle$ and $\langle n_k\rangle \equiv c_k$, Eqs. 13 reduce to the mass–action equations (Eqs. 1).

Differences between the expected cluster numbers derived from stochastic and mean-field approaches arise from nonvanishing correlations $\langle mn_k\rangle - \langle m\rangle\langle n_k\rangle \neq 0$. One approach for determining exact cluster numbers involves enumerating the possible states of the system by elements of the probability vector $\mathbf{P}$, and solving a large set of coupled ordinary differential equations $\dot{\mathbf{P}} = \mathbf{AP}$. Here, the transition matrix $\mathbf{A}$ is to be constructed from the rates of entering and exiting each configuration. This approach is feasible only for small values of $\{M, N_{\rm s}, N\}$ where the number of distinguishable configurations is manageable. For example, consider the simple case of $M = 5, N_{\rm s} = 2, N = 3$, which corresponds to $\sigma = 5/6$. Here, there are nine possible configurations as shown in Fig. 2, which we enumerate in the order $(5|2,0,0,0)$, $(4|1,1,0,0)$, $(3|0,2,0,0)$, $(3|1,0,1,0)$, $(2|0,1,1,0)$, $(2|1,0,0,1)$, $(1|0,0,2,0)$, $(1|0,1,0,1)$, $(0|0,0,1,1)$, so that $P_1(t) \equiv P(\{5|2,0,0,0\},t)$. After solving the nine coupled ODEs for $P_k(t)$ we use Eq. 12 to construct the expected equilibrium cluster numbers and find for $\varepsilon = 10^{-4}$, $\langle n_0(t\to\infty)\rangle = 0$, $\langle n_1(t\to\infty)\rangle = 0.0001$, $\langle n_2(t\to\infty)\rangle = 0.99995$, $\langle n_3(t\to\infty)\rangle = 0.99995$.

In principle, one can construct the transition matrix $\mathbf{A}$ for general values of $\{M, N_{\rm s}, N\}$, but its dimensionality rapidly increases with increasing system size. For a given set of $\{M, N_{\rm s}, N\}$ the total number of configurations is

$$\sum_{j=0}^{[\frac{M}{N}]}\ \sum_{k=0}^{[\frac{M-jN}{N-1}]}\ \sum_{\ell=0}^{[\frac{M-jN-k(N-1)}{N-2}]} \cdots 1, \tag{14}$$

where $[\cdot]$ indicates the integer part and where there are N sums to be performed with their respective indeces subject to the constraints $0 \leq j+k+\ell+\cdots \leq N_{\rm s}$ and $M - Nj - (N-1)k - (N-2)\ell - \cdots \leq N_{\rm s}$. As can be verified numerically, the sum increases dramatically even for moderate values of $\{M, N_{\rm s}, N\}$. For such larger systems, kinetic Monte-Carlo (KMC) simulations of the stochastic process described by Eq. 7 can be straightforwardly performed. We discuss our numerical and analytical results, as well as how they relate to the classical Becker–Döring mean cluster sizes, in the next section.

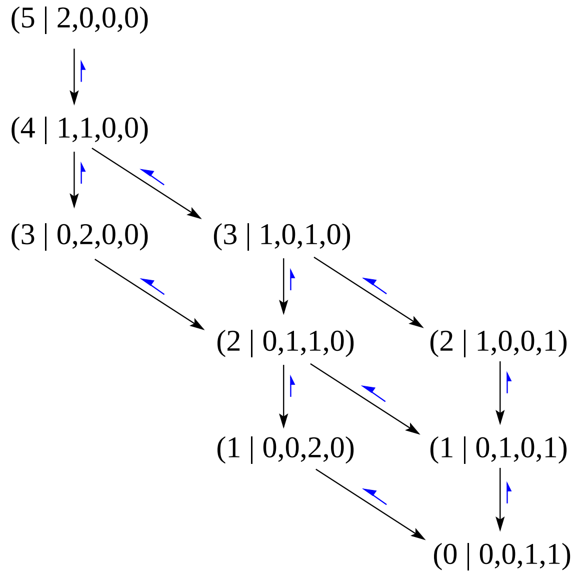

FIG. 2: State space for a self-assembling system consisting of $M = 5$ total monomers, $N_{\rm s} = 2$ seeds, and a maximum cluster size of $N = 3$. In this example, $\sigma = M/(N_{\rm s}N) = 5/6$.

IV. RESULTS AND DISCUSSION

We first show results for obtained by simulating the stochastic process described by the Master equation 7. In order to compare our simulated stochastic results from those obtained from mass-action kinetics we plot $\langle n_k(t)\rangle$ together with the the solutions $c_k(t)$ of the Becker–Döring Eqs. 1.

In Fig. 3, we compare mean cluster numbers derived from numerical solutions of Eqs. 1 with those derived from KMC simulations of the process described by the Master Eq. 7. We consider a system with $N_{\rm s} = 10$ seeds that can bind up to $N = 5$ monomers. The mean clusters numbers $\langle n_k(t)\rangle$ are plotted as a function of time for increasing total mass M ((a)-(d)). For $M < N_{\rm s}N$, we find the expected intermediate metastable configuration that lasts on order of $t \sim 1/\varepsilon$, before reorganizing into an equilibrium configuration. Upon comparing KMC simulations with the mass-action results, we find that generally, both methods give qualitatively similar results. However, deviations of mass-action kinetics from the simulated and exact results depend primarily on $\sigma \equiv M/(N_{\rm s}N)$.

Our simulations show that correlations between

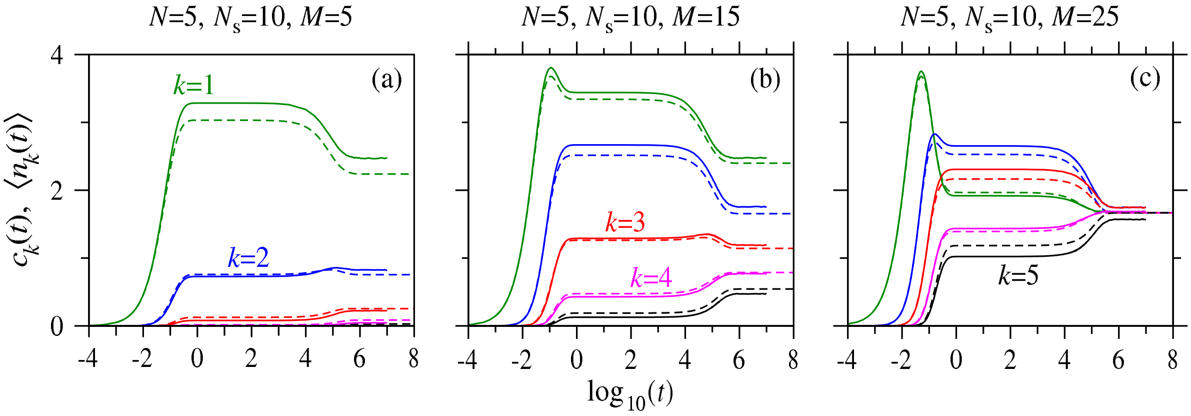

FIG. 3: Mean cluster sizes $\langle n_k(t)\rangle$ obtained from averaging 10^5 KMC simulations of the stochastic process in Eq. 7 with $N = 5$, $N_\mathrm{s} = 10$, and $\varepsilon = 10^{-5}$. The dashed curves represent solutions from BD equations for comparison. (a) $M = 5$ corresponding to $\sigma = 0.1$. (b) $M = 15$ corresponding to $\sigma = 0.3$, and (c) $M = 25$ ($\sigma = 1/2$).

monomers and larger clusters are relatively more important when M is very small or almost close – but not equal – to the total available cluster vacancy. This to be expected, since in both these cases of $\sigma \gtrsim 0$ and $\sigma \lesssim 1$ the number of monomers is incommensurate with the total cluster vacancy and in addition to fully completed clusters there will be a few spurious monomers that may dramatically affect the total cluster distribution, underlining finite size effects, as illustrated above in the case of $M = NN_\mathrm{s} - 1$. On the other hand when $\sigma \sim 1/2$ the number of spurious monomers is large enough for a broader distribution to emerge, closer to the Becker–Döring results.

Note that the trend predicted by the Becker–Döring equations, of smaller clusters being more populated than larger ones for $0 < \sigma < 1/2$ is reflected and amplified in our stochastic system. Similarly for $1/2 < \sigma < 1$ larger clusters are favored according to our Becker–Döring results, as confirmed and magnified in our stochastic simulations.

In order to more efficiently analyze the discrepancies between exact solutions and those from mass-action kinetics, we now develop some analytic approaches. For the metastable cluster numbers $\langle n^*\rangle$, we preclude detachment by setting $\varepsilon = 0$. As shown in previous work[11], the final quenched configurations in this case will depend on the initial cluster distribution. Analytic progress can be made by using combinatoric analyses for the related "urn" problem where the number of ways to distribute M balls into N_s bins is enumerated. Here, we must also consider a maximum capacity for each bin of N balls. We can also map the problem onto a Tonks gas and use similar techniques[12]. Using simple combinatorial arguments based on the inclusion-exclusion principle illustrated in the Appendix, we find

$$\langle n_k^*\rangle = \frac{b_{\{k,M,N_\mathrm{s}\}} N_\mathrm{s}^M + \displaystyle\sum_{i=1}^{[\frac{M}{N+1}]} (-1)^i \binom{N_\mathrm{s}}{i} \sum_{j_1=N+1}^{M} \cdots \sum_{j_i=N+1}^{M-\sum_{\ell=1}^{i-1} j_\ell} b_{\{k,M-\sum_{\ell=1}^{i} j_\ell, N_\mathrm{s}-i\}} \frac{(N_\mathrm{s}-i)^{M-\sum_{\ell=1}^{i} j_\ell} M!}{(M-\sum_{\ell=1}^{i} j_\ell)! \prod_{\ell=1}^{i} j_\ell!}}{N_\mathrm{s}^M + \displaystyle\sum_{i=1}^{[\frac{M}{N+1}]} (-1)^i \binom{N_\mathrm{s}}{i} \sum_{j_1=N+1}^{M} \cdots \sum_{j_i=N+1}^{M-\sum_{\ell=1}^{i-1} j_\ell} \frac{(N_\mathrm{s}-i)^{M-\sum_{\ell=1}^{i} j_\ell} M!}{(M-\sum_{\ell=1}^{i} j_\ell)! \prod_{\ell=1}^{i} j_\ell!}}, \quad (15)$$

where $b_{\{k,M,N_\mathrm{s}\}}$ is the average number of clusters of size k assuming M particles can be distributed in N_s seeds without any constraints

$$b_{\{k,M,N_\mathrm{s}\}} = N_\mathrm{s} \binom{M}{k} \left(1 - \frac{1}{N_\mathrm{s}}\right)^{M-k} \left(\frac{1}{N_\mathrm{s}}\right)^{k}. \quad (16)$$

Eq. 15 leads to the same results as the previously-found recursion relation[15].

Expressions for the true equilibrium cluster numbers $\langle n_k^\mathrm{eq}(t \gg \varepsilon^{-1})\rangle$ must be constructed using detailed balance among the lowest free energy states, just as done for the homogeneous case[15]. For small detachment rates and $\varepsilon \to 0^+$, the lowest energy states are those containing no free monomers. We can enumerate such states and find their relative weights by invoking the appropriate, single-monomer connecting states, and applying detailed balance. For example, in the specific case of $M = 5$, $N = 10$ and $N = 4$ the states that carry the most weight are $(0|8,1,0,0,1)$, $(0|8,0,1,1,0)$, $(0|7,2,0,1,0)$, $(0|7,1,2,0,0)$, $(0|6,3,1,0,0)$ and $(0|5,5,0,0,0)$. The states are connected via intermediate states of order ε built by detaching one particle from the existing clusters and reattaching it to any of the N_s seeds. For instance, detachment from the cluster of size four of state $(0|8,1,0,0,1)$ leads to state $(1|8,1,0,1,0)$ with a free monomer, that may reattach to any of the eight free seeds to create state $(0|7,2,0,1,0)$. Similarly, any one of the two monomers can detach from the latter state, leading to the single-monomer configuration $(1|8,1,0,1,0)$. This free monomer can then attach to the trimer and lead to the state $(0|8,1,0,0,1)$. Detailed balance among the two states with $m = 0$ leads to $8\varepsilon P(0|8,1,0,0,1) = 2\varepsilon P(0|7,2,0,1,0)$, so that $4P(0|8,1,0,0,1) = P(0|7,2,0,1,0)$. Similar arguments can be applied to all equilibration states to find their relative weights. Upon normalizing, one can derive the exact probability for each state to occur. In the above case of $M = 5$, $N_{\rm s} = 10$, $N = 4$ and $\varepsilon \to 0^+$, we find $P(0|8,1,0,0,1) = 15/332$, $P(0|8,0,1,1,0) = 15/332$, $P(0|7,2,0,1,0) = 60/332$, $P(0|7,1,2,0,0) = 60/332$, $P(0|6,3,1,0,0) = 140/332$ and $P(0|5,5,0,0,0) = 42/332$. These weights lead to

$$\begin{array}{ll}
\langle n_0^{\rm eq}\rangle = \frac{2130}{332} = 6.416, & c_0^{\rm eq} = 6.5927 \\
\langle n_1^{\rm eq}\rangle = \frac{525}{332} = 1.5813, & c_1^{\rm eq} = 2.2673 \\
\langle n_2^{\rm eq}\rangle = \frac{275}{332} = 0.8283, & c_2^{\rm eq} = 0.7797 \\
\langle n_3^{\rm eq}\rangle = \frac{75}{332} = 0.2259, & c_3^{\rm eq} = 0.2682 \\
\langle n_4^{\rm eq}\rangle = \frac{15}{332} = 0.04518, & c_4^{\rm eq} = 0.0922
\end{array} \tag{17}$$

As expected, these agree with results from our KMC simulations, but differ significantly from results derived from the Becker–Döring mass-action equations, shown to the right.

One can extend the detailed balance method to larger systems, however state space becomes increasingly larger as $\{M, N_{\rm s}, N\}$ increase and the enumeration process much more difficult. Therefore, we have implemented a computational algorithm that determines the allowable transitions between various states $(n_0, n_1..., n_N)$, allowing for single monomer detachment as an intermediate state, and reattachment, under the fixed seed number constraint. First, we enumerate all possible states for a given set of $\{M, N_{\rm s}, N\}$. Next, we determine the set of all allowable transitions, and determine the probabilities between various states by detailed balancing. As there is degeneracy in the system of equations for the probabilities between the various states, as becomes manifest for large M, we consider only a linearly independent set of equations. The probabilities between the various states are determined using detailed–balancing arguments, as discussed in Section ?. We solve the linear system of equations exactly using LU-decomposition, and after normalizing to the total probability of the various states, we determine the equilibrium weights $\langle n_k^{\rm eq}\rangle$ by weighting by the probability of the various enumerated states.

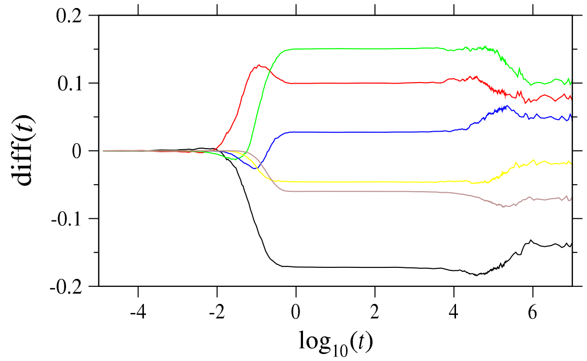

FIG. 4: The relative error $\Delta(t)$ as a function of time, obtained from averaging 10^5 KMC simulations of the stochastic process in Eq. 7, and from numerically solving Eq. 1. Parameters used were $\varepsilon = 10^{-4}$, $N_{\rm s} = 10$, and $N = 5$. Different curves represent systems with total mass of $M = 5, 10, 15$, and 25, corresponding to $\sigma = 0.1, 0.2, 0.3, 0.5$.

In order to systematically quantify the discrepancies between the mass-action cluster size estimates $c_k(t)$ and the stochastic exact value $\langle n_k(t)\rangle$, we introduce the system-size-averaged variation in for the expected cluster numbers:

$$\Delta(t) \equiv \frac{1}{N+1}\sum_{k=0}^{N}\left|\frac{\langle n_k(t)\rangle}{N_{\rm s}} - \frac{c_k(t)}{N_{\rm s}}\right|^2. \tag{18}$$

$\Delta(t)$ provides the relative error averaged over all k clusters.

In Fig. 4 we plot $\Delta(t)$ for different values of M for $N_{\rm s} = 10, N = 5$ (the same parameters as used in Fig. 3). The error vanishes at initial times but increases during evolution the nucleation process.

In Fig. 5(a) we plot the relative error in the metastable and equlibrium regimes as a function of σ. In Fig. 5(a), we used Eq. 15 to compute $\langle n_k^*\rangle$ and Eq. 4 to find c_k^*, and constructed Δ^* according to Eq. 18. Values for different sets of $\{M, n_{\rm s}, N\}$ are plotted. Note that Δ^* vanishes as $M \to 0$ and $M \to N_{\rm s}N$ as expected. We find that the maximum error typically occurs for σ
$sim 0.8 - 0.9$. In Fig. 5(b), we used Eqs. 5 and 6 to find $c_k^{\rm eq}$ and KMC simulations to estimate $\langle n_k^{\rm eq}\rangle$ in the construction of $\Delta^{\rm eq}$. Here, do to particle hole symmetry, the error is a symmetric function about $M = N_{\rm s}N/2$, but is also typically maximal near $\sigma \sim 0.1, 0.9$.

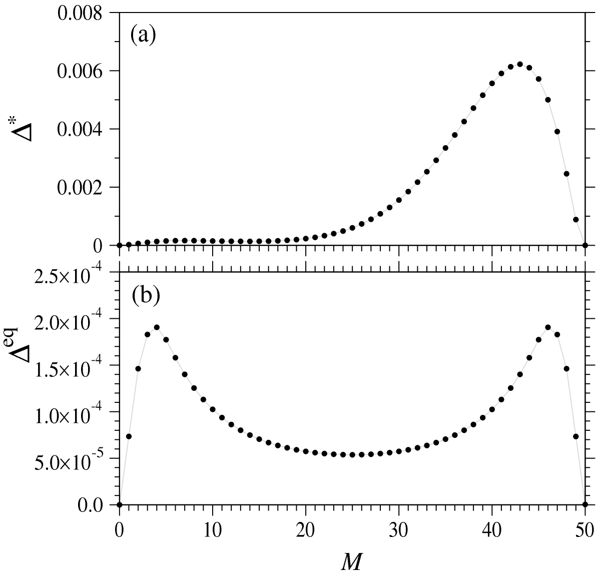

FIG. 5: The overall error of mass-action kinetics. (a) The error during the metastable regime Δ^*. (b) The error Δ^{eq} in the equilibrium limit $t \gg \varepsilon^{-1} \gg 1$

V. CONCLUSIONS

We have derived the fully stochastic Master equations associated with heterogeneous self-assembly in a closed system with a maximum cluster size constraint. Results for cluster concentrations derived from classical mass-action equations were computed and compared with corresponding results from analysis of the discrete problem.

VI. APPENDIX

In this Appendix we illustrate the steps taken to derive Eq. 15, for general $\{M, N_{\mathrm{s}}, N\}$. We will frame our discussion by referring to M as balls and to N_{s} as bins within the context of the "balls in bins" problem with finite capacity N, since this is a well known topic in combinatorics. It is straightforward to note that our heterogeneous cluster size distribution at equilbrium must reduce to the "balls in bins" results in the limit of $\varepsilon = 0$ when no detachment is allowed. Altough the problem is well defined, to the best of our knowledge there are no known results for the average occupancy distribution under the limited capacity constraint.

We start by considering the case of $M \le N$. Here, bins will never be filled to capacity so that $b_{\{k,M,N_{\mathrm{s}}\}}$, the average number of bins occupied by k balls without contraints and assuming there are M balls to distribute, is given by the well known result

$$b_{\{k,M,N_{\mathrm{s}}\}} = N_{\mathrm{s}} \binom{M}{k} \left(1 - \frac{1}{N_{\mathrm{s}}}\right)^{M-k} \left(\frac{1}{N_{\mathrm{s}}}\right)^{k}. \qquad (19)$$

The above expression is derived by noting that out of M possible balls k must occupy one specific bin out of a total of N_{s}, while the other $M-k$ balls must occupy a different one. Eq. 19 is also the mean cluster size $\langle n_k(t \to \infty)\rangle$ at equilibrium for our heterogeneous problem, in the limit of $\varepsilon = 0$ for $M \le N$, or equivalently $\sigma \le 1/N_{\mathrm{s}}$.

Using the particle-hole duality, we may also write a mirror expression for $M' = NN_{\mathrm{s}} - M \le N$, or equivalently for $\sigma \ge 1 - 1/N_{\mathrm{s}}$ so that

$$b_{\{N-k,M,N_{\mathrm{s}}\}} = \binom{NN_{\mathrm{s}} - M}{k} \left(1 - \frac{1}{N_{\mathrm{s}}}\right)^{NN_{\mathrm{s}}-M-k} \left(\frac{1}{N_{\mathrm{s}}}\right)^{k}. \qquad (20)$$

We can now use Eq. 19 to find the average number of bins occupied by k balls $b_{\{k,M,N_{\mathrm{s}},N\}}$ where each bin cannot exceed capacity N and assuming there are M balls to distribute. We first consider the case of $N < M \le 2N + 1$ where there will be at the most one bin occupied to capacity. One possible way of evaluating $b_{\{k,M,N_{\mathrm{s}},N\}}$ is to consider the general distribution without constraints given by Eq. 19 for general M, N_{s} and discard from this evaluation any configurations with bins where there are more than N balls present. We do this by enumerating all possible configurations in the unconstrained case, given by N_{s}^M since all balls can be placed in any of the N_{s} bins, subtracting the contribution of all configurations that exceed bin capacity and renormalizing by the total number of configurations within the constraint. We thus find, for $N < M \le 2N + 1$

$$b_{\{k,M,N_{\mathrm{s}},N\}} = \frac{b_{\{k,M,N_{\mathrm{s}}\}} N_{\mathrm{s}}^M - \sum_{j=N+1}^{M} b_{\{k,M-j,N_{\mathrm{s}}-1\}} N_{\mathrm{s}} (N_{\mathrm{s}} - 1)^{M-j} \binom{M}{j}}{N_{\mathrm{s}}^M - \sum_{j=N+1}^{M} N_{\mathrm{s}} (N_{\mathrm{s}} - 1)^{M-j} \binom{M}{j}}. \qquad (21)$$

Here, $b_{\{k,M,N_{\mathrm{s}}\}}$ is the average number of bins of size k for M balls in N_{s} bins, not subject to any constraints. Similarly $b_{\{k,M-j,N_{\mathrm{s}}-1\}}$ is the average number of bins of size k for $M - j$ particles in $N_{\mathrm{s}} - 1$ bins, not subject to any constraints. Both are given by Eq. 19. Note that if $M \le N$, Eq. 21 reduces to the unconstrained distribution in Eq. 19.

In Eq. 21, the sum that appears in the numerator is to isolate and discard configurations with bin occupancy of size $j \ge N+1$. Since at the most one bin can exceed capacity the remaining $M-j$ balls are distributed without constraints among the other $N_{\rm s}-1$ bins. The denominator is a normalizing factor calculated on the total number of viable states under the capacity constraint.

Within our heterogeneous nucleation framework, Eq. 21 represents $\langle n_k(t \to \infty)\rangle$ for $1/N_{\rm s} < \sigma \le 2/N_{\rm s} + 1/NN_{\rm s}$ and $\varepsilon = 0$. Eq. 21 can can also be recast using the particle-hole duality with $N < M' = NN_{\rm s} - M \le 2N+1$, or equivalently $1 - 2/N_{\rm s} - 1/NN_{\rm s} \le \sigma < 1 - 1/N_{\rm s}$. For instance, in the case $M = 6$, $N_{\rm s} = 3$, $N = 4$ Eq. 21 yields $\langle n_0\rangle = 5/23$ $\langle n_1\rangle = 18/23$, $\langle n_2\rangle = 24/23$, $\langle n_3\rangle = 16/23$, and $\langle n_4\rangle = 6/23$.

We can now extend this result to larger values of $M > 2N$, by invoking the exclusion-inclusion principle.

Given general $\{M, N_{\rm s}, N\}$, at the most there can be $[M/(N+1)]$ clusters that exceed capacity, where $[\cdot]$ denotes the integer part. We will progressively eliminate the contribution of all of them from the unconstrained evaluation of $b_{k,N_{\rm s},N}$, as done above for Eq. 21 when $[M/(N+1)] = 1$.

Assume, for instance that $[M/(N+1)] = 2$. In this case, there can be at the most two bins that exceed capacity. We must then eliminate from the configurations that led to Eq. 21 – where we have only included the possibility that one bin and one bin only exceeds capacity – the ones where a second bin may be filled beyond capacity.

These configurations are characterized by two bins populated by $j_1, j_2 \ge N+1$ particles, thus beyond capacity, and by $M - j_1 - j_2$ particles distributed within capacity among the remaining $N_{\rm s} - 2bins$. We thus pick two bins from the $N_{\rm s}$ that are available, , $j_1 \ge N+1$ from the M population, and $j_2 \ge N+1$ from the $M - j_1$ left. We find that the collective weight of these configurations, for all possible $j_1, j_2 \ge N+1$ is

$$\sum_{j_1=N+1}^{M} \sum_{j_1=N+1}^{M-j_1} \binom{N_{\rm s}}{2}\binom{M}{j_1}\binom{M}{j_2}(N_{\rm s}-2)^{M-j_1-j_2} \quad (22)$$

This is the extra term that appears in the denominator of Eq. 15 for $[M/(N+1) = 2]$. The numerator will contain the distribution of the remaining particles within the remaining bins associated to these configurations, $b_{k,M-j_1-j_2,N_{\rm s}-2}$, with their proper weights.

The same enumeration process can be iterated for general $\{M, N_{\rm s}, N\}$ and for increasing values of $[M/(N+1)]$. At every step of the iteration, we need to subtract configurations from the previous terms, resulting in an alternating series. A careful evaluation results in Eq. 15, which can be easily verified, for example, in the case $2N+1 \le M \le 3$. In particular Eq. 15 reduces to Eq. 21 for $[M = (N+1)] = i = 1$ and to Eq. 19 for $[M = (N+1)] = i = 0$.

[1] K. F. Kelton and A. L. Greer, *Nucleation in condensed matter. Applications in materials and biology* (Pergamon Materials Series, The Netherlands, 2010)

[2] B. R. Novak, E. J. Maginn, and M. J. McCready, Comparison of heterogeneous and homogeneous bubble nucleation using molecular simulations, *Phys. Rev. B* **75** 085413 (2007)

[3] T.T. Yeh, T.E. Hsieh and H.P.D.Shieh, A method to enhance the data transfer rate of eutectic Sb-Te phase change recording media, *J. Appl. Phys.* **9**8 023102 (2005)

[4] N. E. Chayen, E. Saridakis and R. P. Sear, Experiment and theory for heterogeneous nucleation of protein crystals in a porous medium, *Proc. Nat. Acad. Scie.* **1**03 597–601 (2005)

[5] J. Brange, Physical stability of proteins. In *Pharmaceutical formulation development of peptides and proteins* edited by S. Frokjaer and L. Hovgaard (Taylor and Francis, London 2000)

[6] F. Librizzi and C. Rische, The kinetic behavior of insulin fibrillation is determined by heterogeneous nucleation pathways, *Prot. Sci.* **14** 3129–3134 (2005)

[7] G. A. Barabino, M. O. Platt and D. K. Kaul, Sickle Cell Biomechanics, *Annu. Rev. Biomed. Eng.* **12** 345–367 (2010)

[8] M. M. Javadpour and M. D. Barkley, Self-assembly of designed antimicrobial peptides in solution and micelles, *Biochemistry*, **36**, 9540–9549, (1997); W. Soliman, S. Bhattacharjee, and K. Kaur, Adsorption of an Antimicrobial Peptide on Self-Assembled Monolayers by Molecular Dynamics Simulation, *J. Phys. Chem. B*, **114**, 1129211302, (2010);

[9] Iain G. Johnston, Ard A. Louis and Jonathan P. K. Doye, Modelling the self-assembly of virus capsids, *Journal of Physics: Condensed Matter*, **22**, 104101, (2010); A. Zlotnick, To Build a Virus Capsid: An Equilibrium Model of the Self Assembly of Polyhedral Protein Complexes *Journal of Molecular Biology*, **241**, 59-67, (1994).

[10] B. Greene, S.-H. Liu, A. Wilde, and F. M. Brodsky, Complete Reconstitution of Clathrin Basket Formation with Recombinant Protein Fragments: Adaptor Control of Clathrin Self-Assembly, *Traffic*, **1**, 69-75, (2002); A. Banerjee, A. Berezhkovskii and R. Nossal, Stochastic Model of Clathrin-Coated Pit Assembly, *Biophys. J.*, **102**, 2725-2730, (2012).

[11] T. Chou and M. R. D'Orsogna, Coarsening and accelerated equilibration in mass-conserving heterogeneous nucleation, *Phys. Rev. E* **84** 011608 (2011)

[12] M. R. D'Orsogna and T. Chou, Interparticle gap distributions on one-dimensional lattices, *J. Phys. A* **38** 531 (2005)

[13] R. A. Usmani, Inversion of a tridiagonal Jacobi matrix, *Lin. Alg. Appl.* **212** 413–414 (1994)

[14] R. Yvinec, M. R. D'Orsogna and T. Chou, *J. Chem. Phys.* (2012)

[15] M. R. D'Orsogna, G. Lakatos and T. Chou, Stochastic self-assembly of incommensurate clusters, *J. Chem. Phys.* **126** 084110 (2012)