Combinatoric analysis of heterogeneous stochastic self-assembly

Maria R. D'Orsogna,[1,2] Bingyu Zhao,[3] Bijan Berenji,[1,2] and Tom Chou[2,4]
[1]*Department of Mathematics, CSUN, Los Angeles, California 91330-8313, USA*
[2]*Department of Biomathematics, UCLA, Los Angeles, California 90095-1766, USA*
[3]*Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912, USA*
[4]*Department of Mathematics, UCLA, Los Angeles, California 90095-1555, USA*

(Received 21 April 2013; accepted 17 July 2013; published online 1 August 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_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 using kinetic Monte-Carlo simulations in terms of the density of seeds, the total mass, and the maximum cluster size. In the limit of slow detachment, we also find new analytic expressions and recursion relations for the cluster densities at intermediate times and at equilibrium. Our analytic and numerical findings are compared with those obtained from classical mass-action equations and the discrepancies between the two approaches analyzed. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4817202]

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 larger clusters; however, in many cases, the growth process may be catalyzed or even triggered by a "seed" such as an impurity particle or 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]

Self-assembly arises in numerous systems in the natural sciences and engineering. For example, 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.[3] Conversely, the formation of large aggregates of insulin and other proteins is problematic in drug preparation, delivery, and storage.[4] Polymerization of various proteins and polypeptides into amyloid fibers is also implicated in the emergence of neurodegenerative disorders such as Parkinson's, Alzheimer's, and prion diseases.[5] 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, molecular stoichiometry typically prevents clusters from growing indefinitely. After a maximum size N is reached, the self-assembly process is completed. Examples of self-assembly constrained by a maximum cluster size include ligand-receptor binding such as oxygen binding to a single hemoglobin protein ($N = 4$),[8] self-assembly of membrane peptides to form pores ($N \approx 6$–8),[9] self-assembly of capsid proteins around RNA/DNA templates to form viral capsids ($N \sim 100$–1000),[10] or assembly of clathrin triskelion proteins to form the clathrin-coated pits that arise in endocytosis ($N \sim 25$–50).[11] Macroscopic examples of self-assembly in finite-sized systems can also be easily realized using, e.g., capillary-force assisted self-assembly.[12]

We describe our problem as a self-assembly process in a closed system with a total of M particles that can bind N_s seeds, each of which can accommodate a maximum of N particles as shown in Fig. 1. Given the discreteness of the system and possible finite size effects, we will consider a discrete stochastic treatment and our results will be compared to those derived from classical mean-field equations. The classical mass-action equations for heterogeneous nucleation under fixed $\{M, N_s, N\}$ were previously analyzed in Ref. 13, 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 Ref. 13.

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.[14] 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.

The goal of the current paper is to investigate any possible discrepancies between stochastic and mean-field results within the context of homogeneous nucleation. To this end, we perform numerical simulations and, where possible, derive

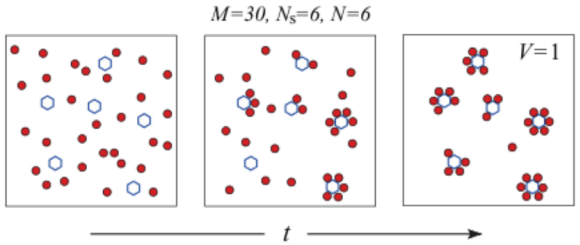

FIG. 1. A schematic of the heterogeneous self-assembly process in a closed system. The open hexagons represent seed particles on which the monomers (filled circles) aggregate. In this example, the total mass, the number of seed particles, and the maximum cluster size are $M = 30$, $N_s = 6$, and $N =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

analytical results. We find that, although less dramatic than in the case of homogeneous nucleation, subtle discrepancies between the two methods also arise in the heterogeneous case. Similar to homogeneous nucleation, these discrepancies are smallest for $M \approx NN_s/2$ (excluding the trivial cases when $M \approx N_s N$ and $M \approx 0$).

In Sec. II, we will give a brief overview for the classical mass actions equations for heterogeneous nucleation, as derived in Ref. 13. In Sec. 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 Sec. IV. We end with some brief considerations on our results in Sec. V.

II. MASS-ACTION KINETICS

In this section, we briefly recapitulate results from Ref. 13 that will be used for comparison with the stochastic results that will be later derived in Sec. III. We derive mass-action equations for a system of total fixed number M of bound and unbound monomers and a fixed number N_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 concentration 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

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

where the number of free monomers is constrained by

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

and where conservation of seeds requires

$$N_s = \sum_{j=0}^{N} c_j(t). \tag{3}$$

Initial conditions are chosen so that $m(t = 0) = M$, $c_0(t = 0) = N_s$, $c_{k>0}(t = 0) = 0$. Equations (1) are analogous to the Becker-Döring (BD) equations commonly used to describe homogeneous nucleation.[14] 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 $\sigma \equiv M/(NN_s)$. In the case of irreversible binding ($q = 0$), the choice $\sigma \geq 1$ implies that all seeds are fully occupied at $t \to \infty$ so that $c_N(t \to \infty) = N_s$, $c_{k\neq N}(t \to \infty) = 0$, and $m(t \to \infty) = M - NN_s$. 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^* were found to be[13]

$$\frac{c^*_{k<N}(\xi)}{N_s} = \frac{\xi^k e^{-\xi}}{k!}, \quad \frac{c^*_N(\xi)}{N_s} = 1 - \sum_{j=0}^{N-1} \frac{\xi^j e^{-\xi}}{j!}, \tag{4}$$

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 ($q > 0$), we will focus on the small $\varepsilon \equiv q/p$ regime, where monomers bind strongly to 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

$$\frac{c_k^{\rm eq}}{N_s} \equiv \frac{(z-1)z^k}{z^{N+1} - 1}, \tag{5}$$

where z satisfies

$$\begin{aligned}
&\left(\frac{\varepsilon z}{N_s N} - \sigma\right)(z-1)(z^{N+1} - 1) + z^{N+2} \\
&\quad - \left(1 + \frac{1}{N}\right) z^{N+1} + \frac{z}{N} = 0
\end{aligned} \tag{6}$$

and $0 < \varepsilon \equiv q/p \ll 1$. Since ε multiplies the highest power of the fugacity z in Eq. (6), $\varepsilon \to 0^+$ constitutes a singular limit. When $\sigma < 1$, not all binding sites can be filled, and approximations for z can be found for $\sigma \ll 1$ and $\sigma \approx 1/2$. In Ref. 13, we performed numerical estimates of Eq. (6) showing that for $1/2 < \sigma < 1$, $z > 1$, and $c_{k+1}^{\rm eq} > c_k^{\rm eq}$, implying that larger clusters tend to be favored. For example, consider the case $M = 5$, $N_s = 2$, and $N = 3$, for which $\sigma = 5/6$. Equations (1) yield increasing values of $c_k^{\rm eq}$: $c_0^{\rm eq} = 0.063$, $c_1^{\rm eq} = 0.173$, $c_2^{\rm eq} = 0.473$, $c_3^{\rm eq} = 1.291$. On the other hand, for $\sigma < 1/2$, $z < 1$, $c_{k+1}^{\rm eq} < c_k^{\rm eq}$, and smaller clusters are favored. For $\sigma = 1/2$, $z = 1$, and all cluster sizes are equally populated.

Note that even if $q \to 0^+$ (or $\varepsilon \to 0^+$), the equilibration values $c_k^{\rm eq}$ arising in the $t \gg q^{-1}$ regime can be quite different from the metastable values c_k^* obtained by directly setting $q_k = q = 0$ in Eqs. (1). On the other hand, in the case of excess

monomers ($\sigma > 1$), all binding sites will be nearly always filled and

$$c_k^{\rm eq} \approx \frac{N_{\rm s}}{(N_{\rm s}N)^{N-k}} \frac{\varepsilon^{N-k}}{(\sigma-1)^{N-k}} + O(\varepsilon^{N-k+1}). \tag{7}$$

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. In Sec. III, we derive the discrete master equation associated with the heterogeneous self-assembly process. We will find expected cluster sizes and compare them with their corresponding values found from mass-action kinetics.

III. MASTER EQUATION FOR HETEROGENEOUS SELF-ASSEMBLY

We now introduce the master equation for our discrete heterogeneous self-assembly. Denote by $P(\{n\};\, t) \equiv P(m|n_0,\, n_1, \ldots,\, n_N;\, t)$ the probability distribution function for the system to be in a state with m free monomers, n_0 unbound seeds, and n_i $(1 \le i \le N)$ 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 Sec. II for the attachment and detachment rates p_k and q_k, respectively, we can write the full master equation as

$$\begin{aligned}\dot{P}(\{n\};t) = &-m\sum_{i=0}^{N-1} p_i n_i P(\{n\};t) - \sum_{i=1}^{N} q_i n_i P(\{n\};t)\\ &+(m+1)\sum_{i=0}^{N-1} p_i(n_i+1)W_*^+ W_i^+ W_{i+1}^- P(\{n\};t)\\ &+\sum_{i=1}^{N} q_i(n_i+1)W_*^- W_{i-1}^- W_i^+ P(\{n\};t),\end{aligned} \tag{8}$$

where we have implicitly 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

$$\begin{aligned}&W_*^+ W_i^+ W_{i+1}^- P(\{n\};t)\\ &\quad\equiv P(m+1|n_0,\ldots,n_i+1,n_{i+1}-1,\ldots,n_N;t).\end{aligned} \tag{9}$$

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,

$$M = m + \sum_{k=1}^{N} k n_k \tag{10}$$

and a total cluster number constraint

$$N_{\rm s} = \sum_{k=0}^{N} n_k. \tag{11}$$

Equations (10) and (11) 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

$$P(\{n\};t=0) = \delta_{m,M}\delta_{n_0,N_{\rm s}}\delta_{n_1,0}\cdots\delta_{n_N,0}, \tag{12}$$

where $\delta_{i,j}$ is the Kronecker delta function such that $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. In order to compare results arising from Eq. (9) to the ones derived from the mean-field Eqs. (1) we define the mean number of clusters of size k as

$$\langle n_k(t)\rangle = \sum_{\{n\}} n_k P(\{n\};t). \tag{13}$$

The values of $\langle n_k(t)\rangle$ derived from the full stochastic treatment in Eq. (13) 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 multiplying Eq. (9) by n_k and by summing over all possible states to give

$$\begin{aligned}\langle \dot{n}_0(t)\rangle &= -\langle m n_0\rangle + \varepsilon\langle n_1\rangle,\\ \langle \dot{n}_k(t)\rangle &= \langle m n_{k-1}\rangle - \langle m n_k\rangle - \varepsilon(\langle n_k\rangle - \langle n_{k+1}\rangle),\\ \langle \dot{n}_N(t)\rangle &= \langle m n_{N-1}\rangle - \varepsilon\langle n_N\rangle,\end{aligned} \tag{14}$$

where we have rescaled time in units of p^{-1} and $\langle m n_k\rangle \equiv \sum_{\{n\}} m n_k P(\{n\};\, t)$ represent monomer-cluster correlations. If we further assume that the monomer and cluster numbers are uncorrelated so that $\langle m n_k\rangle = \langle m\rangle\langle n_k\rangle$, and identify $m \equiv \langle m\rangle$ and $\langle n_k\rangle \equiv c_k$, Eqs. (14) 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 $\langle m n_k\rangle \neq \langle m\rangle\langle n_k\rangle$. 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. Consider again the simple case of $M = 5$, $N_{\rm s} = 2$, $N = 3$, corresponding to $\sigma = 5/6$ that we have considered at the end of Sec. II. Here, there are nine possible configurations $(m|n_0,\, n_1,\, n_2,\, n_3)$ 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,

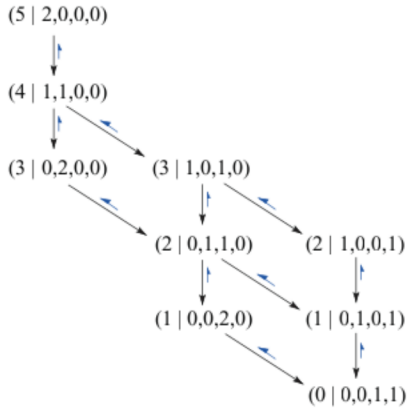

FIG. 2. State space for a self-assembling system consisting of $M = 5$ total monomers, $N_{\rm s} = 2$ seeds, and a maximum cluster size of $N = 3$. In this example, $\sigma = M/(N_{\rm s}N) = 5/6$.

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 ordinary differential equations for $P_k(t)$ we use Eq. (13) to construct the expected equilibrium cluster numbers and find for $\varepsilon = 10^{-4}$, $\langle n_0(t \to \infty)\rangle = 0$, $\langle n_1(t \to \infty)\rangle = 0.0001$, $\langle n_2(t \to \infty)\rangle = 0.99995$, $\langle n_3(t \to \infty)\rangle = 0.99995$. Clearly, these mean equilibrium cluster numbers $\langle n_k^{\rm eq}\rangle$ are often quite different from the corresponding concentrations $c_k^{\rm eq}$ found from the mass-action equations in Sec. II.

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

$$Z = \sum_{j=0}^{[\frac{M}{N}]} \sum_{k=0}^{[\frac{M-jN}{N-1}]} \sum_{\ell=0}^{[\frac{M-jN-k(N-1)}{N-2}]} \cdots 1, \qquad (15)$$

where $[\cdot]$ indicates the integer part and where there are N sums to be performed with their respective indices subject to the constraints $0 \le j + k + \ell + \cdots \le N_s$ and $M - Nj - (N-1)k - (N-2)\ell - \cdots \le N_s$. As can be verified numerically, the sum increases dramatically even for moderate values of $\{M, N_s, N\}$. For such larger systems, kinetic Monte-Carlo (KMC) simulations of the stochastic process described by Eq. (9) can be straightforwardly performed. We discuss our numerical and analytical results, as well as how they relate to mean cluster concentrations derived from mass-action equations, in Sec. IV.

IV. RESULTS AND DISCUSSION

We first show results obtained by simulating the stochastic process described by the master equation (9), where we rescale time in units of p^{-1}. In order to compare our simulated stochastic results to those obtained from mass-action kinetics we plot $\langle n_k(t)\rangle$ together with the solutions $c_k(t)$ of the mass-action equations (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 equation (9). We consider a system with $q/p = \varepsilon = 10^{-5}$ and $N_s = 10$ 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. For $M < N_s N$, 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 do exist and depend primarily on $\sigma \equiv M/(N_s N)$.

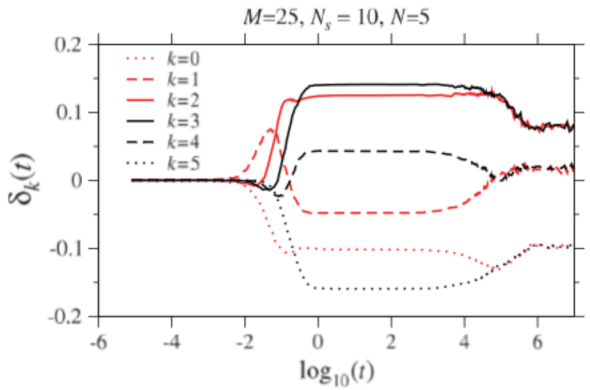

FIG. 4. The difference $\delta_k(t) \equiv \langle n_k(t)\rangle - c_k(t)$ for $k = 0, 1, 2, 3, 4, 5$. In this example, $M = 25$, $N = 5$, and $N_s = 10$.

The discrepancies $\delta_k(t) \equiv \langle n_k(t)\rangle - c_k(t)$ are plotted in Fig. 4 as a function of log (t) for various k. Note that the errors in this case are predominantly negative for small k, increase with cluster size k, then become negative again for the largest k. These qualitative dependences of the discrepancy on cluster sizes are also observed for different values of N, N_s.

In order to more efficiently analyze the discrepancies between exact solutions and those from mass-action kinetics, we now develop some analytic approaches. To estimate the metastable cluster numbers $\langle n_k^*\rangle$, we preclude detachment by setting $\varepsilon = 0$. As shown in the previous work,[13] 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

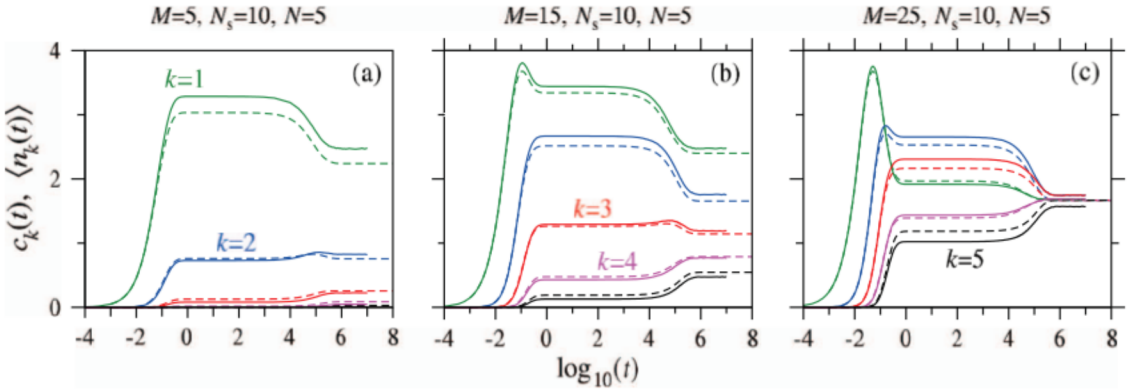

FIG. 3. Mean cluster sizes $\langle n_k(t)\rangle$ obtained from averaging 10^5 KMC simulations of the stochastic process in Eq. (9) with $N = 5$, $N_s = 10$, and $\varepsilon = 10^{-5}$. The dashed curves represent solutions from BD equations for comparison. (a) $M = 5$ corresponding to $\sigma = 0.1$, (b) $M = 15$ corresponding to $\sigma = 0.3$, and (c) $M = 25$ ($\sigma = 1/2$).

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. While results for the simple case of unlimited capacity ($N = \infty$) are well-known, to the best of our knowledge, expressions for the finite capacity N case have not been previously derived. Using simple combinatorial arguments based on the inclusion-exclusion principle illustrated in the Appendix, we find

$$\langle n_k^* \rangle = \frac{b_{\{k,M,N_s\}} N_s^M + \sum_{i=1}^{[\frac{M}{N+1}]} (-1)^i \binom{N_s}{i} \sum_{j_1=N+1}^{M} \cdots \sum_{j_i=N+1}^{M-\sum_{\ell=1}^{i-1} j_\ell} b_{\{k,M-\sum_{\ell=1}^{i} j_\ell, N_s-i\}} \frac{(N_s-i)^{M-\sum_{\ell=1}^{i} j_\ell} M!}{(M-\sum_{\ell=1}^{i} j_\ell)! \prod_{\ell=1}^{i} j_\ell!}}{N_s^M + \sum_{i=1}^{[\frac{M}{N+1}]} (-1)^i \binom{N_s}{i} \sum_{j_1=N+1}^{M} \cdots \sum_{j_i=N+1}^{M-\sum_{\ell=1}^{i-1} j_\ell} \frac{(N_s-i)^{M-\sum_{\ell=1}^{i} j_\ell} M!}{(M-\sum_{\ell=1}^{i} j_\ell)! \prod_{\ell=1}^{i} j_\ell!}}, \tag{16}$$

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

$$b_{\{k,M,N_s\}} = N_s \binom{M}{k} \left(1 - \frac{1}{N_s}\right)^{M-k} \left(\frac{1}{N_s}\right)^k. \tag{17}$$

We can also map the problem onto a Tonks gas and use similar techniques.[15]

Expressions for the true equilibrium cluster numbers $\langle n_k^{\rm eq}(t \gg \varepsilon^{-1}) \rangle$ can be constructed using detailed balance among the lowest free energy states, just as done for the homogeneous case.[14] 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 $M = 5$, $N_s = 10$, and $N = 4$ the states without any free monomers 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). These states are indirectly connected via intermediate states with population of order ε by detaching one particle from an existing cluster and reattaching the monomer to another cluster. 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 equilibrium states with $m = 0$ to find their relative weights. Upon normalizing, one can derive the exact probability for each state to occur. In the above case of $M = 5$, $N_s = 10$, $N = 4$, 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

$$\begin{aligned}
\langle n_0^{\rm eq} \rangle &= \tfrac{2130}{332} = 6.416, \quad c_0^{\rm eq} = 6.5927, \\
\langle n_1^{\rm eq} \rangle &= \tfrac{525}{332} = 1.5813, \quad c_1^{\rm eq} = 2.2673, \\
\langle n_2^{\rm eq} \rangle &= \tfrac{275}{332} = 0.8283, \quad c_2^{\rm eq} = 0.7797, \\
\langle n_3^{\rm eq} \rangle &= \tfrac{75}{332} = 0.2259, \quad c_3^{\rm eq} = 0.2682, \\
\langle n_4^{\rm eq} \rangle &= \tfrac{15}{332} = 0.04518, \; c_4^{\rm eq} = 0.0922.
\end{aligned} \tag{18}$$

As expected, the $\langle n_k^{\rm eq} \rangle$ agree with results from our KMC simulations, but differ significantly from results derived from the mass-action equations, shown in the right column.

One can extend the detailed balance method to larger systems, however state space becomes increasingly larger as $\{M, N_s, N\}$ increase and the enumeration process much more difficult. Therefore, we have implemented a computational algorithm that determines the allowable transitions among the various states $(n_0, n_1 \ldots, n_N)$ via single monomer detachment to an intermediate state, followed by reattachment, under the fixed seed number constraint. In our algorithm, we first enumerate all possible states for a given set of $\{M, N_s, N\}$. Next, we determine the set of all allowable transitions, and determine the probabilities between various states by detailed balancing.

Using these combinatoric approaches, we can easily compute $\langle n_k^* \rangle$ and $\langle n_k^{\rm eq} \rangle$ as functions of the total mass M, and the number of seeds N_s. In Fig. 5, we plot c_k^*, $\langle n_k^* \rangle$, $c_k^{\rm eq}$, and $\langle n_k^{\rm eq} \rangle$. Note that the equilibrium pair values of $c_k^{\rm eq}$ and $c_{N-k}^{\rm eq}$, and $\langle n_k^{\rm eq} \rangle$ and $\langle n_{N-k}^{\rm eq} \rangle$ are symmetric with respect to each other about $M/2$ (in the $\varepsilon \to 0^+$ limit) due to simple particle-hole symmetry considerations. As is evident from Fig. 5, except for very small systems, the difference between $c_k(t)$ and $\langle n_k(t) \rangle$ are largely quantitative.

In order to systematically quantify the overall difference between the mass-action cluster size estimates $c_k(t)$ and the stochastic exact value $\langle n_k(t) \rangle$, we introduce the squared discrepancy $\Delta(t)$ which measures the relative error averaged over

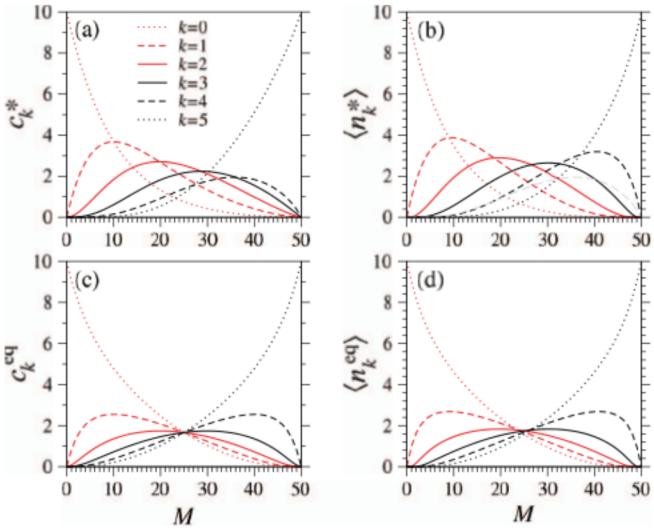

FIG. 5. Expected metastable and equilibrium cluster numbers calculated as a function of M using numerical methods and combinatoric algorithms. Here, as in Fig. 3, $N_s = 10$, $N = 5$. For metastable concentrations, (a) and (b), the differences between the mean-field and exact results are small except for the largest cluster sizes $k = 4, 5$. For comparison, $c_4^*(M)$ is shown by the dashed grey curve in (b). (c) and (d) show c_k^{eq} and $\langle n_k^{eq} \rangle$, respectively. The differences between c_k^{eq} and $\langle n_k^{eq} \rangle$ are more subtle but are generally most noticeable for $\sigma \sim 0.1, 0.9$. All densities are symmetric in $k \leftrightarrow N - k$ about $M = N_s N/2$, with c_k^{eq}, to order ε, forming an isosbestic point at $M = N_s N/2$.

all k clusters,

$$\Delta(t) \equiv \frac{1}{N_s^2(N+1)} \sum_{k=0}^{N} |\langle n_k(t) \rangle - c_k(t)|^2. \qquad (19)$$

In Fig. 6, we plot $\Delta(t)$ for different values of M for $N_s = 10$, $N = 5$ (the same parameters as used in Figs. 3–5). The error increases in time and is seen to be largest for $M = 5$ and $M = 45$. The behavior in the error $\Delta(t)$ suggests that correlations in the cluster densities are nonmonotonic in M, and that mean-field approximations are more accurate for certain masses M.

To explore the dependence of the error on the system mass, we restrict ourselves to the metastable ($1 \ll t \ll \varepsilon^{-1}$) and equilibrium ($t \gg \varepsilon^{-1}$) regimes. We denote the errors in

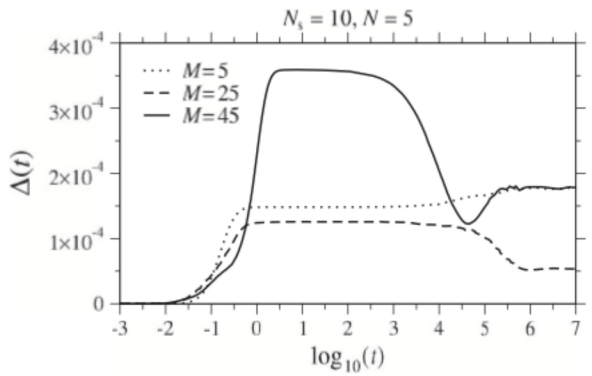

FIG. 6. Plot of $\Delta(t)$ ($N = 5$, $N_s = 10$) for $M = 5$, 25, 45. These curves were generated from KMC simulation of the stochastic process and numerical evaluation of Eqs. (1) for $c_k(t)$. Note that the error in the metastable regime ($1 \ll t \ll \varepsilon^{-1}$) is largest for $M = 45$, while at equilibrium ($t \gg \varepsilon^{-1}$) the errors are largest for $M \approx 5, 45$.

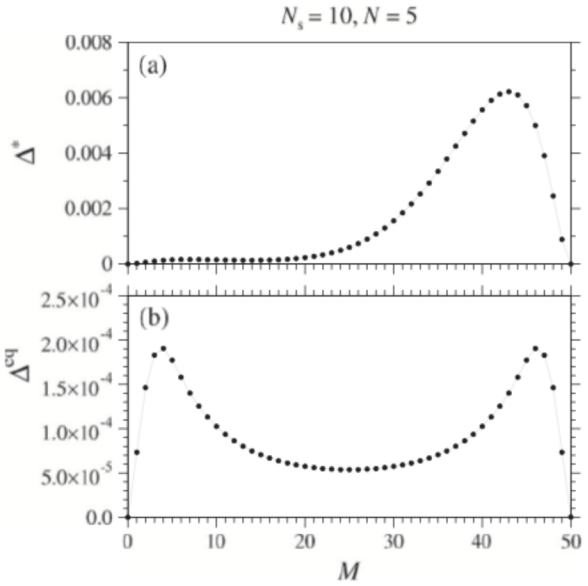

FIG. 7. The overall error of mass-action kinetics. (a) The averaged error in the metastable regime Δ^*. (b) The averaged error Δ^{eq} in the equilibrium limit $t \gg \varepsilon^{-1}$.

these regimes Δ^* and Δ^{eq}, respectively. In Fig. 7(a), we use Eq. (16) to compute $\langle n_k^* \rangle$ and Eq. (4) to find c_k^*, and plot Δ^* according to Eq. (19). Values for different sets of $\{M, N_s, N\}$ are plotted. Note that Δ^* vanishes as $M \to 0$ and $M \to N_s N$ as expected. We find that the maximum error typically occurs for $\sigma \sim 0.9$. In Fig. 7(b), we used Eqs. (5) and (6) to find c_k^{eq} and our combinatoric algorithm to compute $\langle n_k^{eq} \rangle$ in the construction of Δ^{eq}. Here, both c_k^{eq} and $\langle n_k^{eq} \rangle$ are particle-hole "symmetric" as clearly shown in Fig. 5. Therefore, the error Δ^{eq} is a symmetric function about $M = N_s N/2$, and is typically maximal near $\sigma \sim 0.1, 0.9$.

The error in the metastable regime Δ^* typically has a strong peak near $\sigma = 0.9$. The difference between c_k^* and $\langle n_k^* \rangle$ is most pronounced for large k as shown in Figs. 5(a) and 5(b). Because the dynamics are irreversible, small M populating an initially large number of sites can be accurately described by mass-action. However, when M is large and clusters of maximum size are populated, a correlation between larger clusters is induced and mass-action becomes less accurate, resulting in larger errors in c_{N-1}^*, c_N^* near $\sigma \approx 0.9$, giving rise to the qualitative shape of $\Delta^*(M)$.

The behavior of the equilibrium error Δ^{eq} indicates that the effects of correlations are dominant when there are few particle or holes in the system. As can be seen from Figs. 5(c) and 5(d), it is clear that the aggregate difference between c_k^{eq} and $\langle n_k^{eq} \rangle$ is greatest at the lobes near small and large M, giving rise to the double-peaked error function Δ^{eq}. When $M \approx N_s N/2$, all cluster sizes are equally (mean-field), or nearly equally (exact solution) populous, leading to a minimum in Δ^{eq}.

V. CONCLUSIONS

We formulated and analyzed a model of heterogeneous self-assembly in a finite-sized system. The system is initiated with N_s seeds, where each can bind monomers one at a time

with unit rate. Monomers can also detach with rate ε. A maximum cluster size of N and a total mass M are imposed. Analytic results in the $\varepsilon \to 0$ limit were found for the mean-field, mass-action model. The full stochastic problem was treated using KMC simulations, as well as combinatoric analyses in metastable regime ($1 \ll t \ll \varepsilon^{-1}$) and at equilibrium ($t \gg \varepsilon^{-1}$). These results allowed us to efficiently analyze the error arising from cluster number correlations not captured by the mass-action model.

By comparing the mass-action equation derived cluster concentrations $c_k(t)$ with the expected cluster numbers $\langle n_k(t)\rangle$ derived analysis (KMC or combinatorics) of the stochastic model, we find general patterns for the cluster-averaged squared error Δ as a function of the mass ratio $\sigma = M/(N_s N)$. During the metastable phase of self-assembly ($1 \ll t \ll \varepsilon^{-1}$) the overall error Δ^* is maximal for $\sigma \sim 0.9$, which at equilibrium, the error $\Delta^{\rm eq}$ symmetrically peaks at $\sigma \sim 0.1$ and $\sigma \sim 0.9$. These qualitative behaviors are evident in Fig. 7, but also hold for general parameter values N and N_s. We have not considered the effects of cluster-size dependent attachment and detachment rates, however, we expect that the double-peaked symmetry properties at equilibrium to hold to order ε even when $q_k = O(\varepsilon)$ as long as $p_k = p$ is constant.

Our analytic results may provide a useful means for estimating occupancies of macromolecular assemblies in confined systems such as ligand-receptor complexes and cellular transcription machinery. The results also confirm that, unlike in the case of homogeneous self-assembly,[14] heterogeneous self-assembly provides an additional constraint on the number of seeds or clusters, rendering the mean-field, mass-action description qualitatively good. However, the overall discrepancy between the mass-action and exact combinatoric analysis was found to be nonmonotonic and is maximal especially near $M/(N_s N) \sim 0.9$. Finally, although the expected cluster numbers can be qualitatively described using mass-action equations, calculation of other statistical quantities such as the first passage time to maximum cluster formation will require a full stochastic analysis.[16]

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation through Grant Nos. DMS-1021850 (M.R.D.) and DMS-1021818 (T.C.). M.R.D. was also supported by an ARO MURI Grant No. W1911NF-11-10332, while T.C. was also supported by ARO Grant No. 58386MA.

APPENDIX: COMBINATORICS IN THE METASTABLE REGIME

In this appendix, we illustrate the steps taken to derive Eq. (16), for general $\{M, N_s, N\}$. 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," since this is a well known topic in combinatorics.[17] It is straightforward to see that our heterogeneous cluster size distribution at equilibrium must reduce to the "balls in bins" results in the limit of $\varepsilon = 0$ when no detachment is allowed. Although the problem is well defined, to the best of our knowledge there are no known results for the expected cluster numbers under an additional maximum capacity constraint.

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

$$b_{\{k,M,N_s\}} = N_s \binom{M}{k} \left(1 - \frac{1}{N_s}\right)^{M-k} \left(\frac{1}{N_s}\right)^{k}. \qquad \text{(A1)}$$

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. Equation (A1) 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 $\sigma \le 1/N_s$. We can now use Eq. (A1) to find the average number of bins occupied by k balls $b_{\{k,M,N_s,N\}}$ 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 $b_{\{k,M,N_s,N\}}$ is to consider the general distribution without constraints given by Eq. (A1) 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 $N < M \le 2N + 1$

$$b_{\{k,M,N_s,N\}} = \frac{b_{\{k,M,N_s\}} N_s^M - \sum_{j=N+1}^{M} b_{\{k,M-j,N_s-1\}} N_s (N_s - 1)^{M-j} \binom{M}{j}}{N_s^M - \sum_{j=N+1}^{M} N_s (N_s - 1)^{M-j} \binom{M}{j}}. \qquad \text{(A2)}$$

Here, $b_{\{k,M,N_s\}}$ is the average number of bins of size k for M balls in N_s bins, not subject to any constraints. Similarly, $b_{\{k,M-j,N_s-1\}}$ is the average number of bins of size k for $M - j$ particles in $N_s - 1$ bins, not subject to any constraints. Both are given by Eq. (A1). Note that if $M \le N$, Eq. (A2) reduces to the unconstrained distribution in Eq. (A1).

In Eq. (A2), the sum that appears in the numerator is to isolate and discard configurations with bin occupancy of size $j \geq N+1$. Since at the most one bin can exceed capacity the remaining $M-j$ balls are distributed without constraints among the other $N_s - 1$ 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. (A2) represents $\langle n_k(t \to \infty)\rangle$ for $1/N_s < \sigma \leq 2/N_s + 1/(N_s N)$ and $\varepsilon = 0$ and may be used to approximate $\langle n_k(1 \ll t \ll \varepsilon^{-1})\rangle$ for $\varepsilon \to 0^+$. For instance, when $M = 5$, $N_s = 2$, and $N = 3$, Eq. (A2) yields the following approximations for $\langle n_k^*\rangle$: $\langle n_0\rangle \approx \langle n_1\rangle \approx 0$ and $\langle n_2\rangle \approx \langle n_3\rangle \approx 1$. These values coincide with those obtained at the end of Sec. III. Using Eq. (A2), mean, metastable cluster numbers in systems of any size can be readily approximated to within order ε. For $M = 6$, $N_s = 3$, $N = 4$, we find $\langle n_0\rangle \approx 5/23$, $\langle n_1\rangle \approx 18/23$, $\langle n_2\rangle \approx 24/23$, $\langle n_3\rangle \approx 16/23$, and $\langle n_4\rangle \approx 6/23$.

We can now extend this result to larger values of $M > 2N$, by invoking an exclusion-inclusion principle. Given general $\{M, N_s, N\}$, 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 $b_{\{k,N_s,N\}}$, as done above for Eq. (A2) 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. (A2)—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 $j_1, j_2 \geq N+1$ particles, thus beyond capacity, and by $M - j_1 - j_2$ particles distributed within capacity among the remaining $N_s - 2$ bins. We thus pick two bins from the N_s that are available, $j_1 \geq N+1$ from the M population, and $j_2 \geq N+1$ from the $M - j_1$ left. We find that the collective weight of these configurations, for all possible $j_1, j_2 \geq N+1$ is

$$\sum_{j_1=N+1}^{M} \sum_{j_1=N+1}^{M-j_1} \binom{N_s}{2} \binom{M}{j_1} \binom{M}{j_2} (N_s - 2)^{M-j_1-j_2}.$$

This is the extra term that appears in the denominator of Eq. (16) for $[M/(N+1) = 2]$. The numerator will contain the distribution of the remaining particles within the remaining bins associated to these configurations, $b_{\{k,M-j_1-j_2,N_s-2\}}$, with their proper weights.

The same enumeration process can be iterated for general $\{M, N_s, N\}$ 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. (16), which can be easily verified, for example, in the case $2N + 1 \leq M \leq 3$. In particular, Eq. (16) reduces to Eq. (A2) for $[M = (N+1)] = i = 1$ and to Eq. (A1) 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 (Elsevier, 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] N. E. Chayen, E. Saridakis, and R. P. Sear, "Experiment and theory for heterogeneous nucleation of protein crystals in a porous medium," Proc. Natl. Acad. Sci. U.S.A. **103**, 597–601 (2006).

[4] 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).

[5] J. Q. Trojanowski and V. M. Lee, "Fatal attractions of proteins: A comprehensive hypothetical mechanism underlying Alzheimer's disease and other neurodegenerative disorders," Ann. N.Y. Acad. Sci. **924**, 62–67 (2000).

[6] F. Librizzi and C. Rische, "The kinetic behavior of insulin fibrillation is determined by heterogeneous nucleation pathways," Protein 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] C. Walsh, *Enzymatic Reaction Mechanisms* (W. H. Freeman & Co., 1978).

[9] 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**, 11292–11302 (2010).

[10] I. G. Johnston, A. Louis, and J. P. K. Doye, "Modelling the self-assembly of virus capsids," J. Phys.: Condens. Matter **22**, 104101 (2010); A. Zlotnick, "To build a virus capsid: An equilibrium model of the self assembly of polyhedral protein complexes," J. Mol. Biol. **241**, 59–67 (1994).

[11] 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 (2000); A. Banerjee, A. Berezhkovskii, and R. Nossal, "Stochastic model of clathrin-coated pit assembly," Biophys. J. **102**, 2725–2730 (2012).

[12] P. W. K. Rothemund, "Folding DNA to create nanoscale shapes and patterns," Nature (London) **440**, 297–302 (2006).

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

[14] M. R. D'Orsogna, G. Lakatos, and T. Chou, "Stochastic self-assembly of incommensurate clusters," J. Chem. Phys. **136**, 084110 (2012).

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

[16] R. Yvinec, M. R. D'Orsogna, and T. Chou, "First passage times in homogeneous nucleation and self-assembly," J. Chem. Phys. **137**, 244107 (2012).

[17] C. A. Charalambides, *Enumerative Combinatorics* (CRC Press, Boca Raton, FL, 2002).