Fluctuations in Protein Aggregation: Design of Preclinical Screening for Early Diagnosis of Neurodegenerative Disease

Giulio Costantini,[1] Zoe Budrikis,[2] Alessandro Taloni,[1] Alexander K. Buell,[3] Stefano Zapperi,[1,2,4,5] and Caterina A. M. La Porta[6,*]

[1]*Center for Complexity and Biosystems, Department of Physics, University of Milano, via Celoria 16, 20133 Milano, Italy*

[2]*Institute for Scientific Interchange Foundation, Via Alassio 11/C, 10126 Torino, Italy*

[3]*Institute of Physical Biology, University of Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany*

[4]*CNR–Consiglio Nazionale delle Ricerche, Istituto di Chimica della Materia Condensata e di Tecnologie per l'Energia, Via R. Cozzi 53, 20125 Milano, Italy*

[5]*Department of Applied Physics, Aalto University, P.O. Box 11100, FIN-00076 Aalto, Espoo, Finland*

[6]*Center for Complexity and Biosystems, Department of Biosciences, University of Milano, via Celoria 26, 20133 Milano, Italy*

(Received 20 November 2015; revised manuscript received 27 June 2016; published 21 September 2016)

Autocatalytic fibril nucleation has recently been proposed to be a determining factor for the spread of neurodegenerative diseases, but the same process could also be exploited to amplify minute quantities of protein aggregates in a diagnostic context. Recent advances in microfluidic technology allow the analysis of protein aggregation in micron-scale samples, potentially enabling such diagnostic approaches, but the theoretical foundations for the analysis and interpretation of such data are, so far, lacking. Here, we study computationally the onset of protein aggregation in small volumes and show that the process is ruled by intrinsic fluctuations whose volume-dependent distribution we also estimate theoretically. Based on these results, we develop a strategy to quantify *in silico* the statistical errors associated with the detection of aggregate-containing samples. Our work explores a different perspective on the forecasting of protein aggregation in asymptomatic subjects.

DOI: 10.1103/PhysRevApplied.6.034012

I. INTRODUCTION

The presence of aberrant conformations of the amyloid β peptide and the protein α-synuclein is considered to be a key factor behind the development of Alzheimer's and Parkinson's diseases, respectively. The polymerization kinetics of these proteins has been shown to consist of nucleation and growth processes and to be strongly accelerated by the presence in solution of preexisting fibrils [1,2], thereby circumventing the slow primary nucleation of aggregates. It was found that surfaces, such as lipid bilayers [3,4] and hydrophobic nanoparticles [5], can accelerate the nucleation process dramatically. Indeed, in the case of α-synuclein, it was found that in the absence of suitable surfaces, the primary nucleation rate is undetectably slow [2]. Under certain conditions, the surfaces of the aggregates themselves appear to be able to catalyze the formation of new fibrils, leading to autocatalytic behavior and exponential proliferation of the number of aggregates [2,6,7]. This so-called secondary nucleation process is likely to play an important role in the spreading of aggregate pathology in affected brains [8], as the transmission of a single aggregate into a healthy cell with a pool of soluble protein might be sufficient for the complete conversion of the soluble protein into aggregates.

An intriguing idea is to exploit this observation to screen biological samples based on the presence of very low concentrations of aggregates for preclinical diagnosis of neurodegenerative diseases. Indeed, this has been achieved in the case of prion diseases in a methodology called prion-misfolding cyclic amplification [9], which is based on the amplification of aggregates through repeated cycles of mechanically induced fragmentation and growth. Recently, the applicability of this approach to the detection of aggregates formed from the amyloid β peptide has been demonstrated [10]. Furthermore, the autocatalytic secondary nucleation of amyloid β fibrils has been exploited to demonstrate the presence of aggregates during the lag phase of aggregation [11].

However, none of these methods currently allow one to easily determine the absolute number of aggregates in a given sample. One strategy to address this problem is to divide a given sample into a large number of subvolumes and determine for each of the subvolumes whether or not it contains an aggregate. Because of advances in microfluidic technology and microdroplet fabrication [12], it is now possible to monitor protein aggregation in micron-scale samples [13], a technique that could be used to design microarrays targeted for protein polymerization assays. To be successful, this program needs guidance from theory to

*caterina.laporta@unimi.it

quantify possible measurement errors due to false-positive and false-negative detection. The current understanding of protein polymerization is based on mean-field reaction kinetics that have proved successful to describe key features of the aggregation process in macroscopic samples [7,14,15]. This theory is, however, designed to treat the system in the infinite volume limit, where the intrinsic stochasticity of the nucleation processes cannot manifest itself, so that its applicability to small-volume samples is questionable. The importance of noise in protein aggregation is clearly illustrated in Ref. [16], who proposed and solved the master-equation kinetics of a model for polymer elongation and fragmentation, obtaining good agreement with experiments on insulin aggregation [13].

Here, we address the problem by numerical simulations of a three-dimensional model of diffusion-limited aggregation of linear polymers [17], including explicitly autocatalytic secondary nucleation processes [2,6,7]. A three-dimensional model overcomes the limitations posed by both mean-field kinetics [7,14,15] and master-equation approaches [16], which do not consider diffusion and spatial fluctuations. Most practical realizations of protein-aggregation reactions are not diffusion limited, due to the slow nature of the aggregation steps, caused by significant free-energy barriers [18]. This leads to the system being well mixed at all times and mean-field theories providing a good description. There are, however, cases both *in vitro* (e.g., when protein concentrations and ionic strengths are high, leading to gel formation [2]) and *in vivo* (due to the highly crowded interior of the cell), where a realistic description cannot be achieved without the explicit consideration of diffusion. Here, we use our model to study fluctuations in the aggregation process induced by small volumes and to provide predictions for the reliability of a seed-detection assay.

II. THREE-DIMENSIONAL MODEL

Simulations are performed using a variant of the protein-aggregation model described in Ref. [17], where individual protein molecules sit on a three-dimensional cubic lattice. The model considers primary nucleation due to monomer-monomer interaction, polymer elongation due to the addition of monomers to the polymer end points, and secondary nucleation processes in which the rate of monomer-monomer interaction is enhanced when the process occurs close to a polymer [see Fig. 1(a) for an illustration] [19]. In particular, monomers diffuse with rate k_D and attach to neighboring monomers with rate k_M (primary nucleation), but when a monomer is nearest neighbor to a site containing a polymer composed of at least n^* monomers, then the nucleation rate increases to $k_S \gg k_M$ (secondary nucleation). We do not consider polymer fragmentation, since this term is mostly relevant for samples under strong mechanical action [7], and some of the most important amyloid-forming proteins have been shown to exhibit aggregation kinetics dominated by secondary nucleation under quiescent conditions [2,7]. A monomer can attach to a polymer with rate k_H if it meets its end points. Polymers move collectively by reptation with a length-dependent rate k_R/i^2, where i is the number of monomers in the polymer (see Ref. [20], p. 89), and

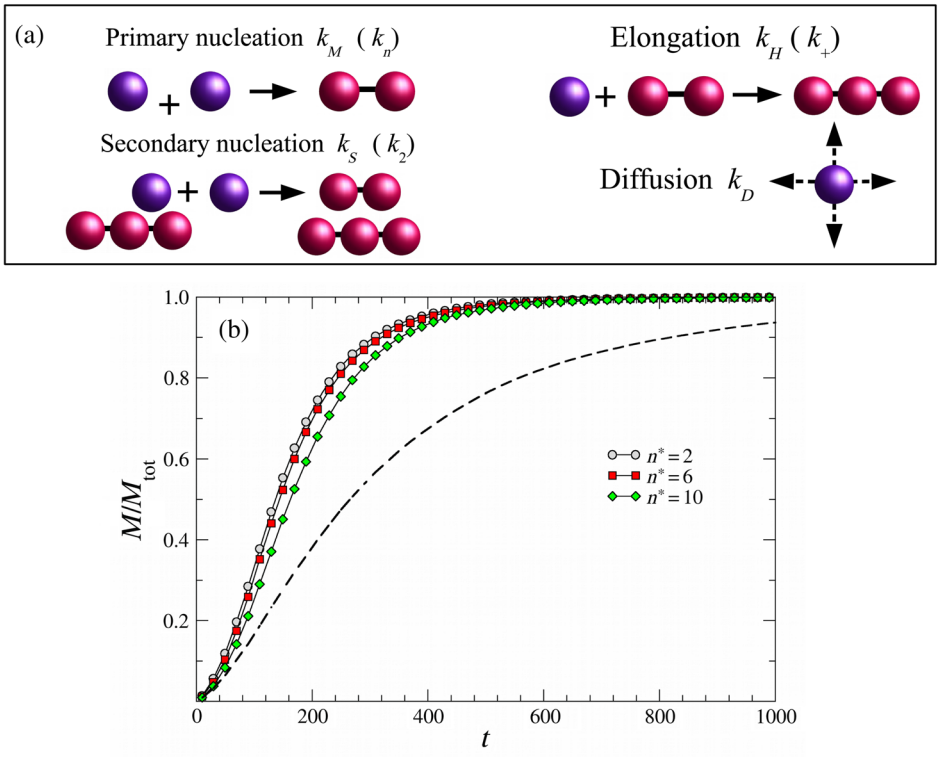

FIG. 1. (a) Schematic of the protein-aggregation model describing the main processes involved: primary nucleation occurring with rate k_M (and correspondingly k_n in the mean field), polymer elongation with rate k_H (k_+ in the mean field), secondary nucleation with rate k_S (k_2 in the mean field), and diffusion with rate k_D, which is not described by mean-field theory. (b) Simulations showing the dependence on the minimal polymer size n^* needed to catalyze secondary nucleation of the aggregation curve, the polymer mass fraction M/M_{tot}, obtained for $k_M = 4 \times 10^{-4}$, $k_D = 10^{-2}$, $k_S = 1$, $k_H = 10^4$, and $\rho = 0.16$. The dashed line is the curve obtained in the limit $n^* \to \infty$, or equivalently in the absence of secondary nucleation.

locally by end rotations, with rate k_E, and kink moves with rate k_K (for a review of lattice polymer models, see Ref. [20]). Simulations start with a constant number of N monomers in a cubic system of size $L = m_0 L_0$ (with m_0 being an integer), where L_0 is the typical monomer diameter, with periodic boundary conditions in all directions. We perform numerical simulations using the Gillespie Monte Carlo algorithm [21] and measure time in units of $1/k_S$ and rates in units of k_S. We explore the behavior of the model by varying independently both the monomer concentration $\rho \equiv N/m_0^3$ and the number of monomers N at fixed ρ, but also the rate constants. For the simulation results reported in the following, the rates describing polymer motion are chosen to be $k_E = k_R = k_K = 10^{-2}$, which is smaller than or equal to the diffusion rate of the monomers k_D.

As expected, secondary nucleation efficiently decreases the half time before rapid polymerization. We illustrate this by changing the critical polymer size n^* needed to induce secondary nucleation. We observe that the lower n^* is, the shorter the half time [see Fig. 1(b)]. Currently, no experimental data exist on the value of n^*, but it can be expected to be of a similar magnitude as the smallest possible amyloid fibril, defined as the smallest structure for which monomer addition becomes independent of the size of the aggregate and an energetic downhill event.

III. MEAN-FIELD THEORY

The progress of reactions observed experimentally in bulk systems can be well approximated by a mean-field model [7,14,15], without fragmentation or depolymerization of polymers. Such models are in contrast to our three-dimensional computational model, which describes also monomer diffusion and polymer motion due to reptation, kink motion, and end rotations, which are not treated by mean-field approximation. Despite this, it is still possible to fit polymerization curves resulting from three-dimensional simulation results through mean-field theory with effective diffusion-dependent parameters. The fact that both experimental and simulated polymerization curves are described by the same mean-field theory ensures that our model is appropriate to describe experiments. In the mean-field model, the evolution of the concentration f_j of polymers of length $j \geq n_c$, where n_c is the nucleation size, is given by [15]

$$\dot{f}_j(t) = k_n m(t)^{n_c} \delta_{j,n_c} + 2m(t)k_+ f_{j-1}(t) - 2m(t)k_+ f_j(t) + k_2 m(t)^{n_2} \sum_{i=n_c}^{\infty} i f_i(t) \delta_{j,n_2}, \quad (1)$$

where dots indicate time derivatives and $m(t)$ is the monomer concentration. The first term on the right-hand side represents an increase in the concentration of polymers of size n_c due to polymer nucleation by the aggregation of n_c monomers with rate constant k_n; this is a generalized version of the dimer formation with rate constant k_M in the 3D model. The second term represents an increase in the concentration of polymers of size j by the attachment of a monomer to a polymer of size $j-1$, with rate constant k_+. The third term is the corresponding loss of concentration of polymers of size j when they attach to a monomer. These two terms are the mean-field equivalent of the end-point attachment of monomers to polymers with rate constant k_H in the 3D model. The final term represents secondary nucleation, which in the mean-field model is described as an increase in concentration of polymers of size n_2 (the secondary nucleus size) occurring at a

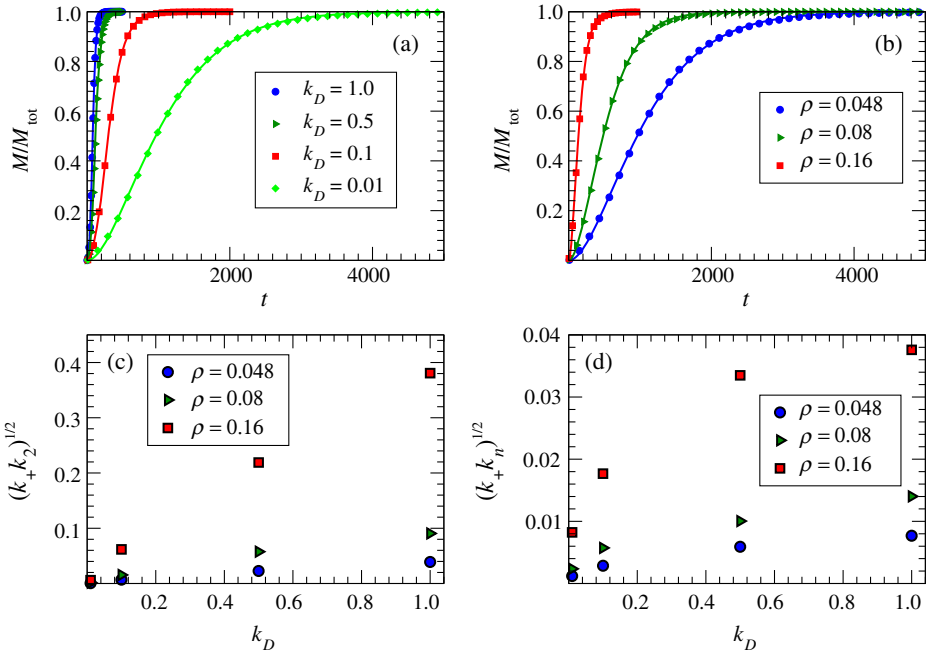

FIG. 2. Protein aggregation depends on monomer diffusion: (a) Simulations showing the diffusion dependence of the aggregation curve, the polymer mass fraction M/M_{tot}, obtained for $k_M = 4 \times 10^{-4}$, $k_S = k_H = 1$, $N = 2450$, and $\rho = 0.048$. The curves are well fit by mean-field theory (full lines) with effective parameters that depend on k_D. (b) Simulations of the density dependence of the polymer mass fraction for $N = 2450$, $k_M = 4 \times 10^{-4}$, $k_H = 1$, and $k_D = 10^{-2}$. Fits by mean-field theory are plotted as lines with effective parameters reported in panels (c) and (d). Time is measured in units of $1/k_S$. (c),(d) The effective mean-field parameters $\sqrt{k_+k_2}$ and $\sqrt{k_+k_n}$ obtained by fitting simulations performed for $k_M = 4 \times 10^{-4}$, $k_S = k_H = 1$ as a function of the concentration ρ and the diffusion rate k_D.

rate proportional to the mass of polymers and with a rate constant k_2. By conservation of mass, the evolution of the monomer concentration is

$$\dot{m}(t) = -\sum_{i=n_c}^{\infty} i\dot{f}_i(t). \tag{2}$$

The evolution of the number concentration $P(t) = \sum_{j\geq n_c} f_j(t)$ and mass concentration $M(t) = \sum_{j\geq n_c} j f_j(t)$ can be found by summing over j in (1). After some algebra, one obtains

$$\dot{P}(t) = k_2 M(t) m(t)^{n_2} + k_n m(t)^{n_c}, \tag{3}$$

$$\dot{M}(t) = 2k_+ m(t) P(t) + n_2 k_2 m(t)^{n_2} + n_c k_n m(t)^{n_c}. \tag{4}$$

Analytical approximation [7,14,15] of the system of equations gives a solution that depends on two parameters, $\lambda = \sqrt{2k_+ k_n m(0)^{n_c}}$ and $\kappa = \sqrt{2k_+ k_2 m(0)^{n_2+1}}$. We fit our data with the form given in Eq. (1) of Ref. [7], using a least-squares method. Each curve is fitted independently. Diffusion plays an important role in the aggregation progress, shifting the aggregation curves as shown in Figs. 2(a) and 2(b). For a considerable parameter range, however, the time evolution of the fractional polymer mass can be fitted by mean-field theory [lines in Figs. 2(a) and 2(b)] with effective parameters that now depend on the diffusion rate k_D [see Figs. 2(c) and 2(d)]. Similarly, mean-field theory describes the density dependence of the aggregation curves as shown in Fig. 2(b).

IV. FLUCTUATIONS IN PROTEIN AGGREGATION

Having confirmed that our computational model faithfully reproduces polymerization kinetics in macroscopic samples, we now turn to the main focus of the paper, the study of sample-to-sample fluctuations in small volumes, a feature that cannot be studied with mean-field kinetics. When the sample volume is reduced, we observe increasing fluctuations in the aggregation kinetics as shown in Figs. 3(a) and 3(b). These results are summarized in Fig. 3(c) showing the complementary cumulative distributions of half times for different monomer numbers N and constant monomer concentration,

$$S(t_{1/2}) \equiv \int_{t_{1/2}}^{\infty} P(x)dx, \tag{5}$$

where $P(x)$ is the probability density function and $t_{1/2}$ is defined as the half time of the polymerization curve (i.e., the time at which $M/M_0 = 1/2$).

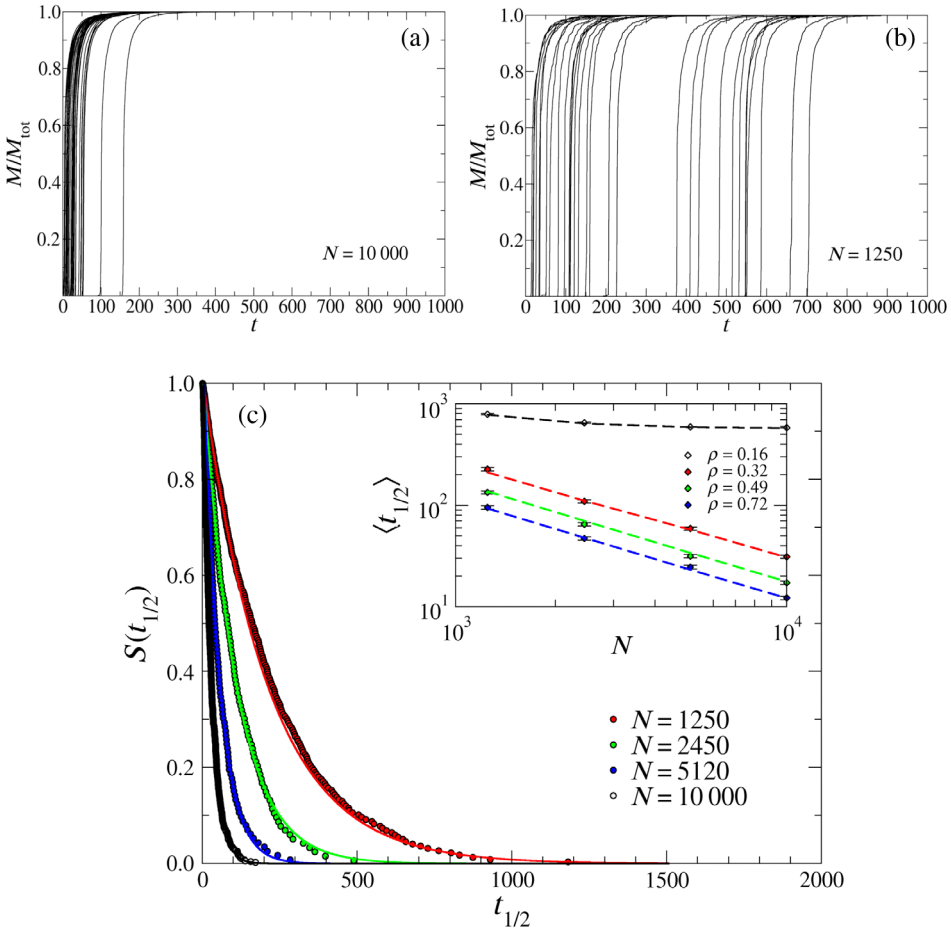

FIG. 3. Half-time sample-to-sample fluctuations are due to extreme value statistics. Different replicates of the simulations display wide fluctuations in half times, especially for small numbers of monomers. (a) Simulation results obtained for $N = 10\,000$ monomers at $\rho = 0.32$, $k_M = 4 \times 10^{-6}$, $k_S = 1$, $k_H = 10^4$, and $k_D = 10^{-2}$. The graph shows that the half time $t_{1/2}$ is very close to the nucleation time t_0 at which the curves depart from zero. (b) Same as panel (a), but with $N = 1250$. (c) The complementary cumulative distributions of half times obtained from simulations for different values of N are in agreement with the theory described in the text. The inset shows the average half times for different concentrations ρ as a function of N. The general trend is in agreement with experiments [13]. Time is measured in units of $1/k_S$. Symbols represent simulations, lines the theoretical predictions.

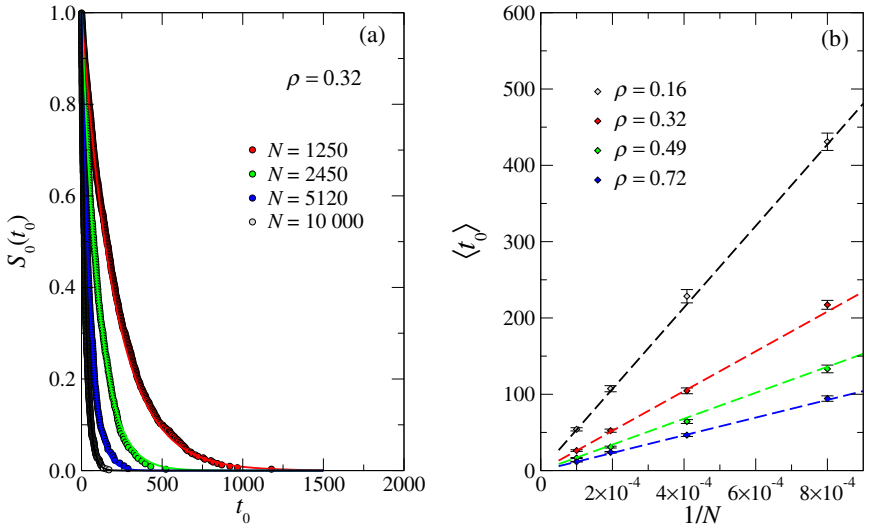

FIG. 4. (a) The complementary cumulative distribution functions $S(t_0)$ for four different monomer numbers N and density $\rho = 0.32$. The symbols correspond to the numerical simulations, while lines correspond to the theoretical predictions obtained from Eq. (6). (b) The average time $\langle t_0 \rangle$ as a function of $1/N$ for four different number densities. The theoretical predictions (dashed lines) are obtained from Eq. (12). Here, $k_D = 10^{-2}$ and $k_M = 4 \times 10^{-6}$.

The steepness of the aggregation curves in Figs. 3(a) and 3(b) suggests that, for $k_S \gg k_M$, fluctuations are mostly ruled by the time of the first primary nucleation event t_0, whose complementary cumulative distribution $S_0(t_0)$ can be estimated analytically as

$$S_0(t_0) = e^{-f_M k_M N t_0}, \tag{6}$$

where f_M is the average number of possible primary nucleation events per unit monomer (Figs. 4–6). We estimate that $f_M = 3\rho$ using a Poisson approximation, as we show in detail in the following section. Note that Eq. (6) displays a size dependence that is reminiscent of extreme value statistics $S_0(x, N) = \exp[-NF(x)]$, where $F(x)$ is a function that does not depend on N [22,23]. If $k_S \gg k_M$, the half time $t_{1/2}$ differs from the nucleation time by a weakly fluctuating time τ. This comes from the observation that, once the first primary nucleation event has happened, the polymerization follows rapidly, thanks to fast growth and secondary nucleation. This yields a weakly fluctuating delay $\tau(N, \rho) = t_{1/2} - t_0$, which in general depends on the number density ρ and on the number of monomers N. The distribution and average values of τ are reported in Fig. 7. The average value of τ decreases with ρ and displays only a smaller dependence on N. The distribution of τ is always peaked around its average, but whereas at small values of ρ the peak shifts with N while the standard deviation remains constant, for higher values of ρ only the standard deviation depends on N and the peak position does not change. Since the fluctuations in τ are much smaller than the fluctuations of t_0, we can safely assume that $t_{1/2} \simeq t_0 + \langle \tau \rangle$ for $t_0 \geq 0$, so the complementary cumulative distribution takes the form

$$S(t_{1/2}) = \begin{cases} 1 & t_{1/2} \leq \langle \tau \rangle \\ S_0(t_{1/2} - \langle \tau \rangle) & t_{1/2} > \langle \tau \rangle. \end{cases} \tag{7}$$

The predictions of Eqs. (6) and (7) are in agreement with numerical simulation results for $S_0(t_0)$ and $S(t_{1/2})$,

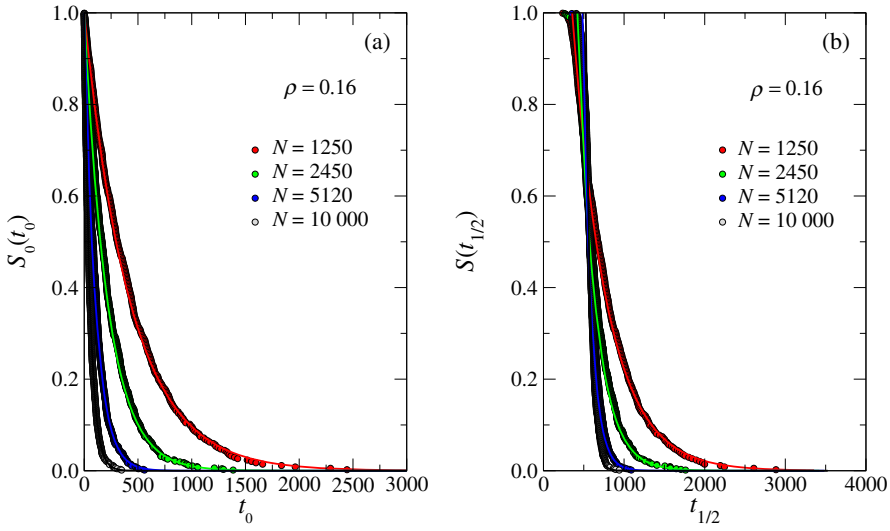

FIG. 5. The complementary cumulative distribution functions $S(t_0)$ (a) and $S(t_{1/2})$ (b) for four different monomer numbers N and density $\rho = 0.16$. The symbols correspond to the numerical simulations while the lines represent the theoretical predictions obtained from Eqs. (6) and (7). Here, $k_D = 10^{-2}$ and $k_M = 4 \times 10^{-6}$.

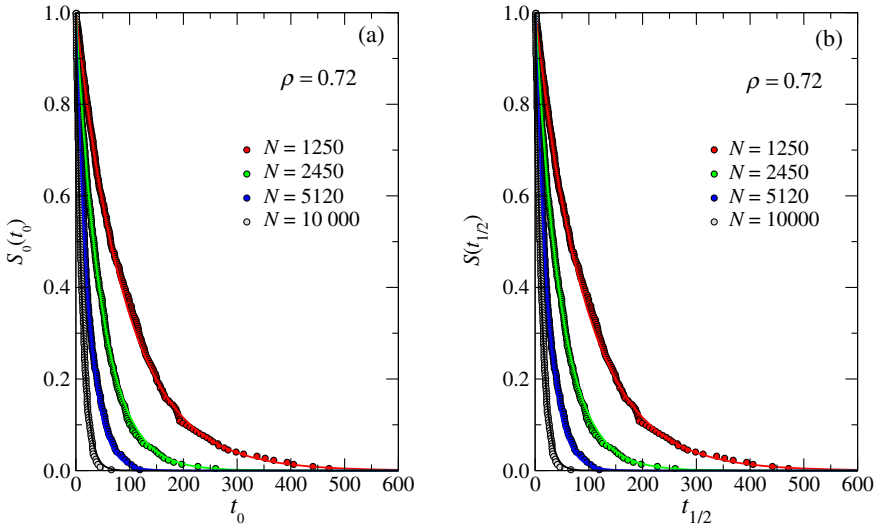

FIG. 6. The complementary cumulative distribution functions $S(t_0)$ (a) and $S(t_{1/2})$ (b) for four different monomer numbers N and density $\rho = 0.72$. The symbols correspond to the numerical simulations, while the lines represent the theoretical predictions obtained from Eqs. (6) and (7). Here, $k_D = 10^{-2}$ and $k_M = 4 \times 10^{-6}$.

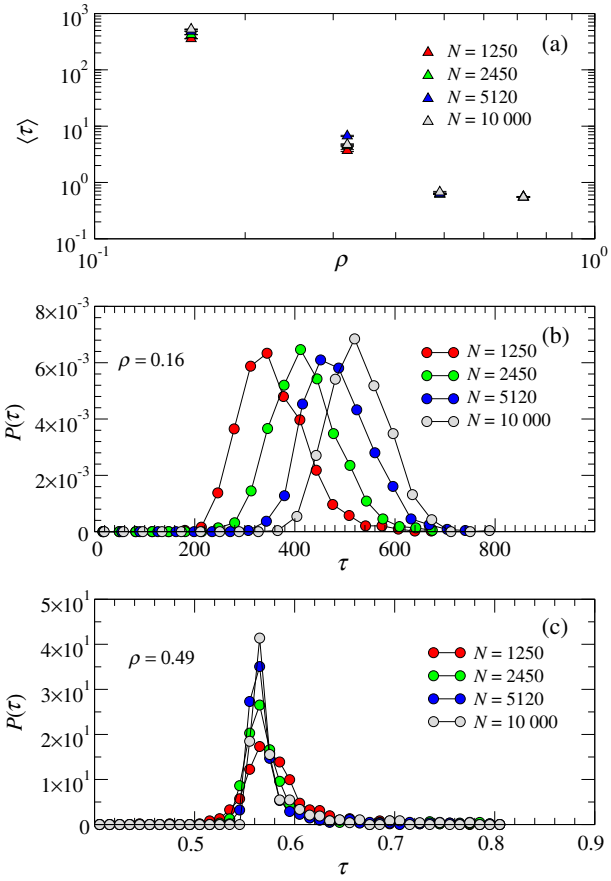

FIG. 7. (a) The mean delay time $\langle\tau\rangle$, obtained from numerical simulations, as a function of the number density ρ for four different monomer numbers N. For any N and ρ the averages are calculated over the different numerical simulation outcomes. The distributions of delay times τ as a function of the number of monomers N for (b) $\rho = 0.16$ and (c) $\rho = 0.49$. Here, $k_D = 10^{-2}$ and $k_M = 4 \times 10^{-6}$.

respectively [see Figs. 3(c) and 4]. In particular, the behavior of $S_0(t_0)$ is obtained without any fitting parameters, while $S(t_{1/2})$ only needs the estimate of the single parameter $\langle\tau\rangle$ (additional comparisons for different values of ρ are reported in Figs. 5 and 6). The corresponding average values of $\langle t_{1/2}\rangle$ are shown in the inset of Fig. 3(c) as a function of N [see also Fig. 4(b)].

V. THEORETICAL DERIVATION OF THE HALF-TIME DISTRIBUTION

In this section, we provide a detailed derivation of Eqs. (6) and (7) in the limit of relatively large diffusion when the system is well mixed. To this end, we consider a cubic lattice composed of m_0^3 nodes, in which N monomers are placed randomly at time $t = 0$. As illustrated in Fig. 8, each monomer i sits near $l^{(i)}$ neighboring monomers and

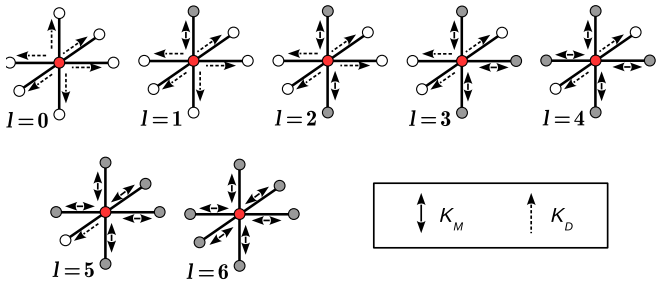

FIG. 8. A schematic representation of the possibilities of diffusion (dashed arrow) and aggregation (double arrow) for a monomer (red circle) placed in the center of the cubic lattice unit cell. The monomer partners for the dimerization (from $l = 1$ to $l = 6$) are colored in gray, while the empty sites are represented by white circles. The aggregation and diffusion rate are, respectively, k_M and k_D.

$6 - l^{(i)}$ neighboring empty sites, where $l^{(i)}$ is in general a fluctuating time-dependent quantity. In the model, each monomer i can either diffuse into one of the $6 - l^{(i)}$ empty sites or form a dimer with one of the $l^{(i)}$ neighboring monomers. Therefore, at any given time t the number of possible diffusion events in the system is $n_D(t) = \sum_i (6 - l^{(i)})$ and the number of possible aggregation events is $n_M(t) = 1/2 \sum_i l^{(i)}$, where the factor $1/2$ is needed to correct for the double counting of the number of monomer pairs. We can compute the time of first aggregation of N monomers using Poisson statistics, considering for simplicity the case in which the number of possible aggregation events n_M would not depend on time. In this case, the probability of having an aggregation event within Δt is $n_M k_M \Delta t$. We can then divide the time interval t_0 into n elementary time subintervals $\Delta t = (t_0/n)$. The rate of the aggregation event at t_0, i.e., the probability per unit time to have the first dimer formed after a time interval t_0 has elapsed, is given by the following expression:

$$\tilde{P}_0(t_0) = \lim_{n\to\infty} \left(1 - n_M k_M \frac{t_0}{n}\right)^{n-1} n_M k_M = n_M k_M e^{-n_M k_M t_0}. \tag{8}$$

As stressed above, n_M and n_D are, in principle, fluctuating quantities and therefore Eq. (8) is not valid. Yet, as shown in Fig. 9: n_M and n_D are both (i) stationary, (ii) ergodic, (iii) weakly fluctuating, and (iv) linearly dependent on N, on average. Hence, the probability $\tilde{P}_0(t_0)$ for a monomer to form a dimer at t_0 can reasonably be approximated by its ensemble average

$$P_0(t_0) \simeq \langle \tilde{P}_0(t_0) \rangle \simeq \langle n_M \rangle k_M e^{-f_M k_M N t_0}, \tag{9}$$

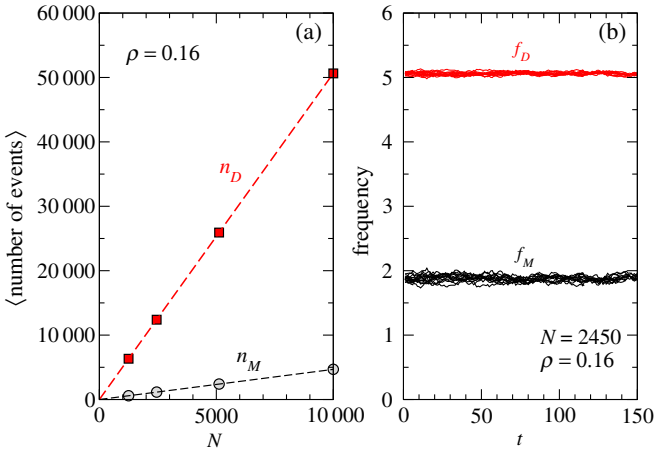

FIG. 9. (a) The number of possible primary nucleation and diffusion events (n_M and n_D, respectively) are a linear function of the number of monomers N. (b) The corresponding frequencies f_M and f_D fluctuate very little in time. Here, $k_D = 10^{-2}$, $k_M = 4 \times 10^{-6}$, $N = 2450$, and $\rho = 0.16$.

where we have replaced n_M by its average value $\langle n_M \rangle$ and defined $f_M \equiv (\langle n_M \rangle / N)$. From Eq. (9), we easily obtain the complementary cumulative distribution function

$$S_0(t_0) = \int_{t_0}^{\infty} d\tau P_0(\tau) = e^{-f_M k_M N t_0}, \tag{10}$$

recovering Eq. (6).

To conclude our calculation, we still need to evaluate f_M. To this end, we perform a discrete enumeration of the possible configurations of a single monomer, in the spirit of cluster expansions for percolation models. In particular, the six relevant configurations for a single monomer in contact with other monomers are reported in Fig. 8. The weight p_l of a configuration in which a monomer has l occupied neighbors is assumed to be given by the binomial distribution

$$p_l = \frac{6!}{l!(6-l)!} \rho^l (1-\rho)^{6-l}. \tag{11}$$

This single particle picture suggests that the average number of primary nucleation events per monomer f_M corresponds to the average number of nearest neighbors $\langle l \rangle$, divided by a factor of 2 since any nucleation event encompasses two particles. With a similar reasoning, we estimate $f_D = 6 - \langle l \rangle$, i.e., the average number of empty directions. Then, from the binomial distribution (11), we get $f_M = (6\rho/2)$ and $f_D = 6(1-\rho)$. Using these values in Eqs. (6) and (7), we obtain agreement with numerical simulations as illustrated in Figs. 5 and 6 [panels (a) and (b), respectively] for different values of ρ.

Finally, we calculate the averages of the first aggregation time and the half time as $\langle t_0 \rangle = \int_0^\infty dt_0 t_0 P_0(t_0)$ and $\langle t_{1/2} \rangle = \int_0^\infty dt_{1/2} t_{1/2} P(t_{1/2})$, where $P(t_{1/2}) = -[dS(t_{1/2})/dt_{1/2}]$. The expression for $\langle t_0 \rangle$ is given by

$$\langle t_0 \rangle = \frac{1}{3\rho k_M N}. \tag{12}$$

In Fig. 4(b) we show the perfect agreement of the theoretical estimate given by Eq. (12) with the numerical values of the average time for the first primary nucleation event as a function of $1/N$, for several densities. Notice that no fitting parameters are involved. The average time of $t_{1/2}$ follows from $\langle t_{1/2} \rangle = \langle t_0 \rangle + \langle \tau \rangle$: the inset of Fig. 3(c) confirms the agreement between the theoretical estimates (dashed lines) and the numerical values (symbols).

VI. STATISTICAL ANALYSIS OF SEED DETECTION TESTS

While the fluctuations we observe are intrinsic to the random nature of nucleation events, the ones usually encountered in bulk experiments are likely due to contamination or differences in initial conditions [3,24,25]. In those bulk systems (μl and larger), the number of protein

molecules involved in the aggregation process is extremely large, even at low concentrations, so that we can exclude intrinsic kinetic fluctuations. For instance, a volume of 100 μl at a concentration of 1 μM still contains 10^{14} monomers, leading to a large number (hundreds to thousands) of nucleation events per second for a realistic value of the nucleation rate [7,26]. However, if the relevant volumes are made small enough (pico- to nanoliters), the stochastic nature of primary nucleation can be directly observed. This has been exploited by aggregation experiments performed inside single microdroplets, where individual nucleation events could be observed, due to their amplification by secondary processes [13]. In these experiments, the average half time is found to scale with volume in a similar manner to what is shown in the inset of Fig. 3, thus in agreement with our simulations.

We are now in a position to use our model to design a test *in silico* to detect the presence of preformed polymers that act as seeds and nucleation sites for the secondary nucleation process, and that are thus amplified. As illustrated in Fig. 10(a), the test considers a set of small-volume samples containing protein solutions with a given concentration at time $t_0 = 0$. The aim of the test is to detect the samples containing at least one seed [case *B* in Fig. 10(a)]. In an ideal experiment, the size of the microdroplets would be adjusted so that most droplets contain no seeds, some contain one seed, and the proportion of droplets containing more than one seed is negligible. In practice, these conditions can be easily adjusted experimentally by progressively reducing droplet volumes until only a small fraction of them display aggregates. After a fixed time $t_1 \sim \langle \tau \rangle$, one can observe which samples contain macroscopic, detectable amounts of aggregates, enabling exact quantification of the number of aggregates present in the initial sample. Ideally, the test should be positive only when at least one seed is initially present, but given the large fluctuations intrinsic to the nucleation processes we demonstrate above, as well as the competition with *de novo* nucleation, there is a chance for false tests. In particular, a false-positive test occurs when an unseeded sample is found to contain aggregates, while a false-negative test corresponds to the case in which a seeded sample does not produce detectable amounts of aggregates within the time scale of the experiment.

In Fig. 10(b) we report the complementary cumulative distribution of aggregation half times $t_{1/2}$ as a function of

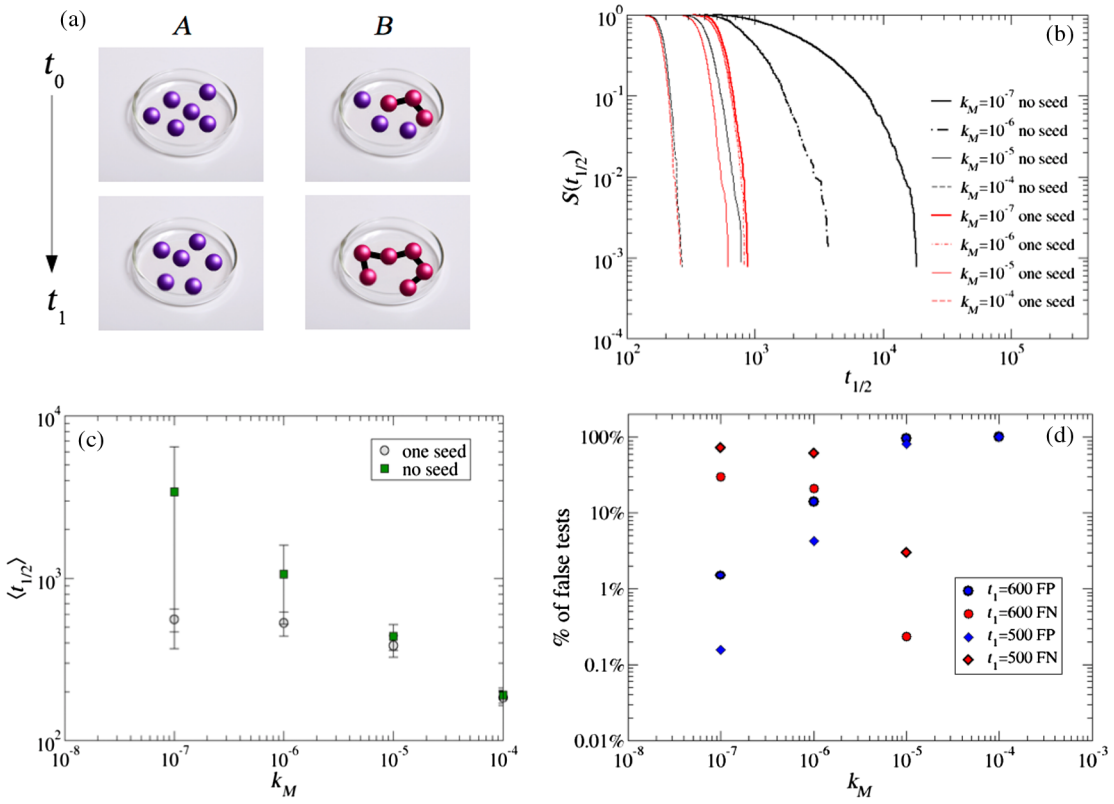

FIG. 10. Intrinsic fluctuations rule errors in the detection of protein-aggregation-prone samples. (a) A test to detect seeds for protein aggregation is based on small-volume sampling of protein solutions at time $t_0 = 0$ and on the assumption that only seeded samples (e.g., sample *B*) would form aggregates at time t_1. (b) Simulations allow us to compute the distribution of half times for samples with and without a seed as a function of the rate of primary nucleation k_M. Data are obtained sampling over a $n = 1200$ independent realization. (c) Average half time ($\pm$ standard deviation) as a function of the rate of primary nucleation k_M. (d) Fraction of false positives (FP) and false negatives (FN) for testing times $t_1 = 600$ and $t_1 = 500$. Time is measured in units of $1/k_S$, $N = 1250$.

the primary nucleation rate k_M for samples with or without seeds, in this case a single preformed trimer. For small values of k_M, seeded and unseeded samples yield distinct results, as also illustrated by the average half times reported in Fig. 10(c). As the value of k_M increases, however, the distributions become closer in the two cases. In Fig. 10(d), we quantify the fraction of false-positive and false-negative tests for two different testing times (e.g., $t_1 = 500$ and $t_1 = 600$). As expected, for large k_M errors are very likely and the test would not be reliable. For intermediate values of k_M, one can try to adjust t_1 to reduce possible errors with a caveat: decreasing t_1 reduces false-positive errors, but at the same time increases false negatives [Fig. 10(d)]. It is, however, possible to optimize t_1 so that both types of errors are minimized. In an experimental realization of such a setup, the most important system parameter that needs to be optimized for any given protein is the ratio of secondary nucleation rate to primary nucleation rate. Because of potentially different dependencies of these two rates on the monomer concentration [27], pH [2], and potentially other factors, such as temperature, salt concentration, etc., it is possible to fine tune this ratio and adjust it to a value that allows for an easy discrimination between droplets that do contain a seed aggregate and those that do not.

VII. CONCLUSIONS

In conclusion, we study protein polymerization in a three-dimensional computational model and elucidate the role of protein diffusion in the polymerization process. Most theoretical studies of protein aggregation neglect completely the role of diffusion and any other spatial effects. When the polymer diffusion and elongation rates are large enough, we recover the standard polymerization curves that can also be obtained from mean-field analytical treatments and that can be used to fit, for example, kinetic data of amyloid β aggregation [7]. It would be interesting to explore if, for small diffusion rates and small densities, mean-field kinetics would eventually fail to describe the results, but this is a challenging computational task. At low densities, diffusion could play an important role since a critical time scale would be set by the time needed by two monomers to meet before aggregating. This time scale can be estimated considering the time for a monomer to cover a distance $x_D \sim \rho^{-1/3}$, yielding $t_D \sim x_D^2/D \sim D^{-1}\rho^{-2/3}$. This time scale is not relevant for our simulations since at the relatively high densities we study, a considerable fraction of monomers are close to at least another monomer [see n_M in Fig. 9(a)]. Consequently, the distribution of the first aggregation time does not depend on the diffusion rate k_D [see Eq. (6)]. The half-time distribution, however, depends on diffusion even in this regime (Fig. 2).

Our simulations show intrinsic sample-to-sample fluctuations that become very large in the limit of small volumes and low aggregation rates. We show that the corresponding half-time distributions are described by Poisson statistics and display size dependence. As a consequence of this, the average half times scale as the inverse of the sample volume, in agreement with insulin aggregation experiments performed in microdroplets [13] and with calculations based on a master-equation approach [16]. We use this result to design and validate *in silico* a preclinical screening test based on a subdivision of the macroscopic sample volume that will ultimately allow the determination of the exact number of aggregates that are initially present. This is the first step to develop microarray-based *in vitro* tests for an early diagnosis of neurodegenerative diseases.

ACKNOWLEDGMENTS

G. C., Z. B., A. T., and S. Z. are supported by ERC Advanced Grant No. 291002 SIZEFFECTS. C. A. M. L. P. thanks the visiting professor program of Aalto University where part of this work was completed. S. Z. acknowledges support from the Academy of Finland FiDiPro program, Project No. 13282993. A. K. B. thanks Magdalene College, Cambridge and the Leverhulme Trust for support.

[1] Joseph T. Jarrett and Peter T. Lansbury, Seeding one-dimensional crystallization, of amyloid: A pathogenic mechanism in Alzheimer's disease and scrapie?, Cell **73**, 1055 (1993).

[2] Alexander K. Buell, Céline Galvagnion, Ricardo Gaspar, Emma Sparr, Michele Vendruscolo, Tuomas P. J. Knowles, Sara Linse, and Christopher M. Dobson, Solution conditions determine the relative importance of nucleation and growth processes in α-synuclein aggregation, Proc. Natl. Acad. Sci. U.S.A. **111**, 7671 (2014).

[3] Marie Grey, Christopher J. Dunning, Ricardo Gaspar, Carl Grey, Patrik Brundin, Emma Sparr, and Sara Linse, Acceleration of α-synuclein aggregation by exosomes, J. Biol. Chem. **290**, 2969 (2015).

[4] Céline Galvagnion, Alexander K. Buell, Georg Meisl, Thomas C. T. Michaels, Michele Vendruscolo, Tuomas P. J. Knowles, and Christopher M. Dobson, Lipid vesicles trigger α-synuclein aggregation by stimulating primary nucleation, Nat. Chem. Biol. **11**, 229 (2015).

[5] Robert Vacha, Sara Linse, and Mikael Lund, Surface effects on aggregation kinetics of amyloidogenic peptides, J. Am. Chem. Soc. **136**, 11776 (2014).

[6] Amy M. Ruschak and Andrew D. Miranker, Fiber-dependent amyloid formation as catalysis of an existing reaction pathway, Proc. Natl. Acad. Sci. U.S.A. **104**, 12341 (2007).

[7] Samuel I. A. Cohen, Sara Linse, Leila M. Luheshi, Erik Hellstrand, Duncan A. White, Luke Rajah, Daniel E. Otzen, Michele Vendruscolo, Christopher M. Dobson, and Tuomas P. J. Knowles, Proliferation of amyloid-β42 aggregates occurs through a secondary nucleation mechanism, Proc. Natl. Acad. Sci. U.S.A. **110**, 9758 (2013).

[8] W. Peelaerts, L. Bousset, A. Van der Perren, A. Moskalyuk, R. Pulizzi, M. Giugliano, C. Van den Haute, R. Melki, and

V. Baekelandt, α-synuclein strains cause distinct synucleinopathies after local and systemic administration, Nature (London) **522**, 340 (2015).

[9] Rodrigo Morales, Claudia Duran-Aniotz, Rodrigo Diaz-Espinoza, Manuel V. Camacho, and Claudio Soto, Protein misfolding cyclic amplification of infectious prions, Nat. Protoc. **7**, 1397 (2012).

[10] Natalia Salvadores, Mohammad Shahnawaz, Elio Scarpini, Fabrizio Tagliavini, and Claudio Soto, Detection of misfolded Aβ oligomers for sensitive biochemical diagnosis of Alzheimer's disease, Cell Rep. **7**, 261 (2014).

[11] Paolo Arosio, Risto Cukalevski, Birgitta Frohm, Tuomas P. J. Knowles, and Sara Linse, Quantification of the concentration of Aβ42 propagons during the lag phase by an amyloid chain reaction assay, J. Am. Chem. Soc. **136**, 219 (2014).

[12] Ashleigh B. Theberge, Fabienne Courtois, Yolanda Schaerli, Martin Fischlechner, Chris Abell, Florian Hollfelder, and Wilhelm T. S. Huck, Microdroplets in microfluidics: An evolving platform for discoveries in chemistry and biology, Angew Chem., Int. Ed. Engl. **49**, 5846 (2010).

[13] Tuomas P. J. Knowles, Duncan A. White, Adam R. Abate, Jeremy J. Agresti, Samuel I. A. Cohen, Ralph A. Sperling, Erwin J. De Genst, Christopher M. Dobson, and David A. Weitz, Observation of spatial propagation of amyloid assembly from single nuclei, Proc. Natl. Acad. Sci. U.S.A. **108**, 14746 (2011).

[14] Tuomas P. J. Knowles, Christopher A. Waudby, Glyn L. Devlin, Samuel I. A. Cohen, Adriano Aguzzi, Michele Vendruscolo, Eugene M. Terentjev, Mark E. Welland, and Christopher M. Dobson, An analytical solution to the kinetics of breakable filament assembly, Science **326**, 1533 (2009).

[15] Samuel I. A. Cohen, Michele Vendruscolo, Christopher M. Dobson, and Tuomas P. J. Knowles, Nucleated polymerization with secondary pathways. II. Determination of self-consistent solutions to growth processes described by non-linear master equations, J. Chem. Phys. **135**, 065106 (2011).

[16] Juraj Szavits-Nossan, Kym Eden, Ryan J. Morris, Cait E. MacPhee, Martin R. Evans, and Rosalind J. Allen, Inherent Variability in the Kinetics of Autocatalytic Protein Self-Assembly, Phys. Rev. Lett. **113**, 098101 (2014).

[17] Zoe Budrikis, Giulio Costantini, Caterina A. M. La Porta, and Stefano Zapperi, Protein accumulation in the endoplasmic reticulum as a non-equilibrium phase transition, Nat. Commun. **5**, 3620 (2014).

[18] Alexander K. Buell, Anne Dhulesia, Duncan A. White, Tuomas P. J. Knowles, Christopher M. Dobson, and Mark E. Welland, Detailed analysis of the energy barriers for amyloid fibril growth, Angew Chem., Int. Ed. Engl. **51**, 5247 (2012).

[19] Samuel I. A. Cohen, Michele Vendruscolo, Mark E. Welland, Christopher M. Dobson, Eugene M. Terentjev, and Tuomas P. J. Knowles, Nucleated polymerization with secondary pathways. I. Time evolution of the principal moments, J. Chem. Phys. **135**, 065105 (2011).

[20] Monte Carlo and Molecular Dynamics Simulations in Polymer Science, edited by K. Binder (Oxford University Press, New York, Oxford, 1995).

[21] Daniel T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, J. Comput. Phys. **22**, 403 (1976).

[22] E. J. Gumbel, *Statistics of Extremes* (Columbia University Press, New York, 2004).

[23] W. Weibull, *A Statistical Theory of the Strength of Materials* (Generalstabens litografiska anstalts foerlag, Stockholm, 1939).

[24] Lise Giehm and Daniel E. Otzen, Strategies to increase the reproducibility of protein fibrillization in plate reader assays, Anal. Biochem. **400**, 270 (2010).

[25] V. N. Uversky, J. Li, and A. L. Fink, Evidence for a partially folded intermediate in alpha-synuclein fibril formation, J. Biol. Chem. **276**, 10737 (2001).

[26] Alexander K. Buell, Christopher M. Dobson, and Tuomas P. J. Knowles, The physical chemistry of the amyloid phenomenon: Thermodynamics and kinetics of filamentous protein aggregation, Essays Biochem. **56**, 11 (2014).

[27] Georg Meisl, Xiaoting Yang, Erik Hellstrand, Birgitta Frohm, Julius B Kirkegaard, Samuel I. A. Cohen, Christopher M. Dobson, Sara Linse, and Tuomas P. J. Knowles, Differences in nucleation behavior underlie the contrasting aggregation kinetics of the Aβ40 and Aβ42 peptides, Proc. Natl. Acad. Sci. U.S.A. **111**, 9384 (2014).