Molecular mechanisms of protein aggregation from global fitting of kinetic models

Georg Meisl[1], Julius B Kirkegaard[1], Paolo Arosio[1], Thomas C T Michaels[1], Michele Vendruscolo[1], Christopher M Dobson[1], Sara Linse[2] & Tuomas P J Knowles[1]

[1]Department of Chemistry, University of Cambridge, Cambridge, UK. [2]Department of Biochemistry and Structural Biology, Lund University, Lund, Sweden. Correspondence should be addressed to T.P.J.K. (tpjk2@cam.ac.uk).

Published online 7 January 2016; doi:10.1038/nprot.2016.010

The elucidation of the molecular mechanisms by which soluble proteins convert into their amyloid forms is a fundamental prerequisite for understanding and controlling disorders that are linked to protein aggregation, such as Alzheimer's and Parkinson's diseases. However, because of the complexity associated with aggregation reaction networks, the analysis of kinetic data of protein aggregation to obtain the underlying mechanisms represents a complex task. Here we describe a framework, using quantitative kinetic assays and global fitting, to determine and to verify a molecular mechanism for aggregation reactions that is compatible with experimental kinetic data. We implement this approach in a web-based software, AmyloFit. Our procedure starts from the results of kinetic experiments that measure the concentration of aggregate mass as a function of time. We illustrate the approach with results from the aggregation of the β-amyloid (Aβ) peptides measured using thioflavin T, but the method is suitable for data from any similar kinetic experiment measuring the accumulation of aggregate mass as a function of time; the input data are in the form of a tab-separated text file. We also outline general experimental strategies and practical considerations for obtaining kinetic data of sufficient quality to draw detailed mechanistic conclusions, and the procedure starts with instructions for extensive data quality control. For the core part of the analysis, we provide an online platform (http://www.amylofit.ch.cam.ac.uk) that enables robust global analysis of kinetic data without the need for extensive programming or detailed mathematical knowledge. The software automates repetitive tasks and guides users through the key steps of kinetic analysis: determination of constraints to be placed on the aggregation mechanism based on the concentration dependence of the aggregation reaction, choosing from several fundamental models describing assembly into linear aggregates and fitting the chosen models using an advanced minimization algorithm to yield the reaction orders and rate constants. Finally, we outline how to use this approach to investigate which targets potential inhibitors of amyloid formation bind to and where in the reaction mechanism they act. The protocol, from processing data to determining mechanisms, can be completed in <1 d.

INTRODUCTION

Interactions within and between proteins are fundamental to the biological function of all organisms, yet the same interactions can lead to the formation of aberrant protein aggregates, in particular amyloid fibrils. Such aggregates are now implicated in a range of increasingly prevalent and currently incurable human disorders, including Alzheimer's and Parkinson's diseases[1–7]. As a consequence, this process is increasingly at the focus of widespread efforts to understand its fundamental nature and to define the key mechanisms through which proteins convert from their normally functional forms into pathogenic aggregates.

An established approach for elucidating the mechanisms of chemical reactions is based on the comparison of integrated rate laws and experimental kinetic measurements. This strategy has been very successful over the past 50 years in areas as diverse as organic synthesis and enzyme action[8]. In the context of filamentous protein aggregation, this approach has revealed important mechanistic insights, including the mechanism of sickle-hemoglobin polymerization[9] and the differences underlying actin and tubulin nucleation[10]. Recent advances in reaction network theory have brought even the complicated pathways associated with amyloid formation within reach of chemical kinetics[11]. However, the application of the conventional methods of chemical kinetics to this key area of science has been hampered by the absence of methods that allow the routine global fitting of these complex rate equations to extensive sets of experimental data in the manner that is required to draw definitive mechanistic conclusions. As a result, the mechanistic information contained in experimental measurements of the kinetics of protein aggregation has generally not been fully exploited—for example, to advance drug discovery programs[12–14].

Thus, although monitoring the kinetics of the aggregation has been an important tool in the study of aggregation-prone proteins, kinetic analysis is often performed in a limited manner: many models used are phenomenological, simply reproducing the curve shapes but not linking the parameters to underlying microscopic processes. Here we present a protocol (**Figs. 1** and **2**) that allows this problem to be alleviated and to make full use of kinetic descriptions derived from a model of the underlying microscopic reactions that make up the aggregation network. The fitted parameters are therefore meaningful and correspond to physical properties of the system, such as nucleus sizes, rate constants of individual reactions and saturation concentrations[9–11,15–19].

Another key problem in the analysis of protein aggregation data is that the individual sigmoidal curves obtained in a typical aggregation experiment contain little information on their own, and they can be fitted by a range of different reaction schemes without yielding information on, or discriminating between, the different underlying molecular mechanisms, which is a classic example of overfitting (**Fig. 3**). In this protocol, we address this fundamental issue by analyzing a large data set of multiple kinetic traces at different reagent concentrations simultaneously, with a single rate law, thereby yielding strong mechanistic constraints[11,20]. This approach, however, requires the routine global

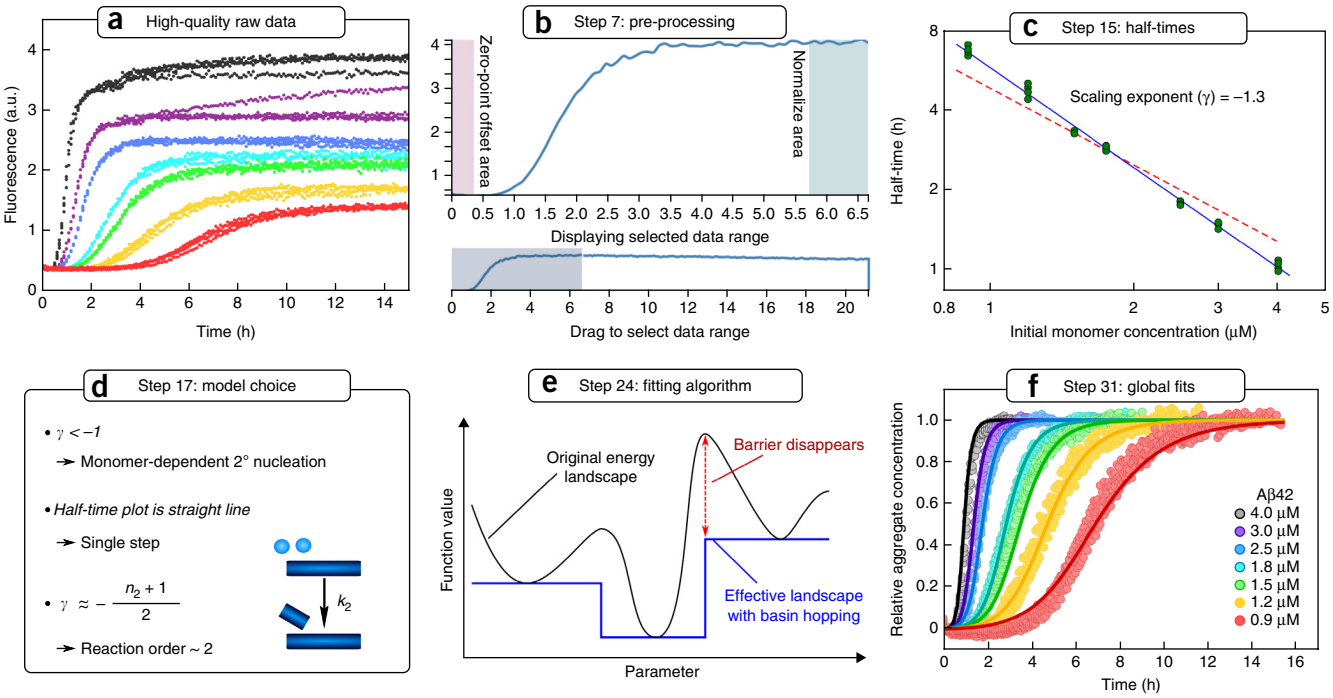


Figure 1 | Key steps of the protocol. This figure illustrates the key parts of the protocol, using data for the aggregation of Aβ42 (ref. 28). (**a**,**b**) After raw data are acquired (**a**), they are processed and normalized (**b**). (**c**) The half-times are extracted and the scaling exponent is calculated. (**d**) The red line in **c** shows the scaling exponent expected for a fragmentation model, the misfits of which are shown in **Figure 3**, highlighting the power of a half-time analysis as a guide to the correct model. (**e**,**f**) By smoothing the energy landscape and avoiding local minima (**e**), a basin-hopping algorithm then helps to find the best fit (**f**).

fitting of complex rate equations to extensive sets of experimental data, which can pose substantial difficulties; rough fitting landscapes (i.e., an abundance of local minima) can result in convergence issues in which the algorithm converges to a local minimum rather than the global best fit. To this end, we have developed a fitting platform, AmyloFit, which is freely accessible online (http://www.amylofit.ch.cam.ac.uk/), and it enables robust global data analysis without the need for extensive programming or detailed mathematical knowledge.

Here we present a step-by-step protocol, using this fitting platform, which details all aspects of analyzing protein aggregation kinetic data, including quality control, model selection, global fitting, prediction and verification (**Figs. 1** and **2**; see also **Box 1** for terminology). The data are securely stored on our servers, thus allowing users to access their data from any location and to easily analyze them in collaboration with other researchers.

Application of the method

As of the time of publication, the software includes equations describing the aggregation of proteins into linear assemblies, motivated by the prevalence of amyloid fibrils in numerous diseases[1,9] as well as the functional roles of linear protein filaments[10,21]. However, the protocol outlined is general, and it will also be applicable, for example, to the formation of disc-shaped aggregates or aggregates of other dimensionalities once the required kinetic rate laws for the relevant systems have been derived.

This protocol and our software are specifically aimed at the study of protein aggregation and the effect of various intrinsic and extrinsic factors on the aggregation behavior in terms of the underlying microscopic reactions. As such, they can be used to investigate the mechanistic origin of the different behavior of disease-related mutations and other sequence variations of a peptide (e.g., of the two major forms of Aβ (ref. 18), whose aggregation is associated with Alzheimer's disease); of the effects of variations of solution conditions, such as ionic strength or pH, on the individual microscopic processes; and of the system's response to other factors, such as mechanical stress.

Moreover, our protocol can also be used to study how various compounds affect each of the microscopic steps of the aggregation reaction, for example, to help determine which molecular species or step in the reaction is affected by a potential inhibitor of amyloid formation[12]. Currently, this analysis would introduce such a compound as a perturbation to the rate constants. Models that include the presence of other compounds explicitly are in development; they will be added to AmyloFit in the future, and they will then allow a global analysis with the compound concentration being an additional degree of freedom that can be varied, modeled and fitted to yield robust mechanistic constraints.

Limitations of the method

This protocol describes the kinetic analysis of a measurement of the total concentration of aggregates of all sizes over time. In the future, as kinetic data on aggregation intermediates, such as oligomers[22], or on the size distribution of aggregates becomes more easily accessible, the models can be extended within this kinetic framework to take this additional information into account. As a protocol on kinetic analysis, it is not suitable for the study of processes that do not influence the kinetics significantly (i.e., equilibrium effects). For example, the dissociation of monomers from fibril ends is slow, with negligible effects on the kinetics

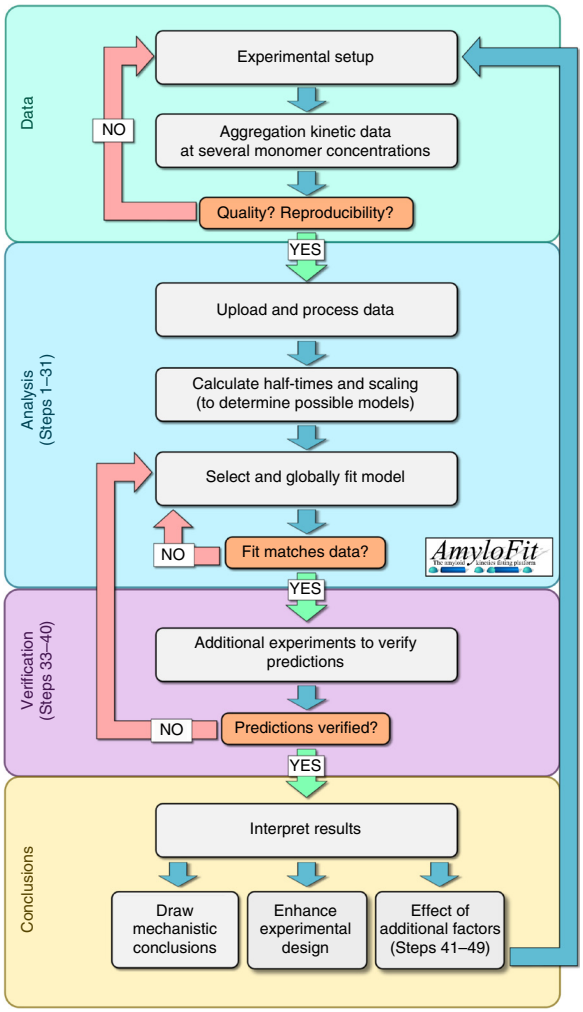


Figure 2 | Flowchart of the protocol. The chart outlines the workflow of the protocol. The core part, the analysis of the kinetic data, can largely be performed on our fitting platform AmyloFit.

of aggregation, but this process is relevant in order to maintain detailed balance[23] once the reaction has reached completion—i.e., fibrils and soluble monomers, present at very low concentrations at this stage, are in equilibrium.

Finally, a limitation of kinetic analysis in general is that its main output consists of reaction orders and rate constants; these may provide structural constraints on the nature of the reacting species, but they do not directly yield information on their conformations. For example, this protocol will allow one to establish that growing fibrils elongate by the addition of monomeric species, rather than oligomers; however, it will not contain any information about, for example, what conformational change accompanies the monomer's incorporation into the fibril.

Experimental design

Data quality control. A chemical kinetic analysis relies heavily on quantitative fitting of specific mathematical models to well-defined experimental data. In this context, the importance of data quality control cannot be emphasized enough: the data must be reproducible (see **Fig. 4**), and it is also absolutely essential that boundary conditions (such as the purity of the protein) be carefully controlled and that experimental assumptions be thoroughly checked (for example, to ensure that the measured fluorescence of a reporter dye is indeed proportional to aggregate concentration). Only if these conditions are fulfilled can the results of a kinetic analysis yield meaningful insights into the mechanism of a reaction.

Obtaining qualitative constraints. Because of the complexity of the aggregation process and the increasing number of models, a practical way to achieve a qualitative insight into an aggregation reaction and to obtain constraints on possible mechanisms is desired for narrowing down the number of models that are suitable for fitting. In this context, a convenient, representative quantity is the half-time—i.e., the time at which half the protein that is present initially in soluble form has aggregated.

AmyloFit can determine the half-times of each curve generated using the software; to do this, normalized data are used and the half-time is defined as the time at which the signal has reached half its final plateau value. The algorithm for extracting half-times is outlined in section 2.1 of the **Supplementary Methods**.

The dependence of the half-time, $t_{1/2}$, on the initial monomer concentration, m_0, is captured by $t_{1/2} \sim m_0^{\gamma}$, where γ is the scaling exponent. By using the rate laws for the time evolution of aggregate mass, γ can be related to the reaction orders (i.e., to the monomer dependence of the dominant processes) for each of the models. In practice, it is useful to plot the half-time versus the monomer concentration on a double logarithmic plot, as $\log(t_{1/2}) = \gamma \log(m_0) + \text{constant}$ (**Fig. 1c**). The slope of this plot therefore gives the scaling exponent, and any deviations of the points from a straight line indicate that γ is dependent on the monomer concentration: physically, a variation of γ with monomer concentration means that the monomer dependence of the dominant mechanism of aggregation is changing with monomer concentration. Very generally, a negative curvature in the double logarithmic plots (i.e., when the slope becomes steeper at higher monomer concentrations, and therefore the process more monomer dependent) is indicative of competition between several processes in parallel. A positive curvature, in contrast (i.e., a flattening of the curve at higher monomer concentrations, and thus a decrease in monomer dependence), suggests the presence of a saturation effect in a serial process (**Fig. 5** and **Supplementary Methods**), or, in rare cases, at monomer concentrations close to solubility, it can be due to a change in nucleus size[9]. The behavior of half-times with varying monomer concentration is, therefore, a good first guide to narrowing down the number of possible models, because it limits the number of acceptable reaction networks by determining the reaction order of the dominant process and probing for competition or saturation effects. The model for fitting needs to be chosen to reflect these findings (**Fig. 1d**), as well as taking into account other information (e.g., the mechanism of aggregation for a similar protein under similar conditions).

Different models. The kinetic schemes are all derived from the underlying series of molecular steps, including nucleation and growth processes, that are inherent in all filamentous self-assembly reactions[10]. The system is described by two quantities: the aggregate mass concentration, $M(t)$ (this is the concentration of monomers one would obtain by re-dissolving all aggregates), and the aggregate number concentration, $P(t)$ (this is the total concentration of aggregates of any size). Accordingly, all

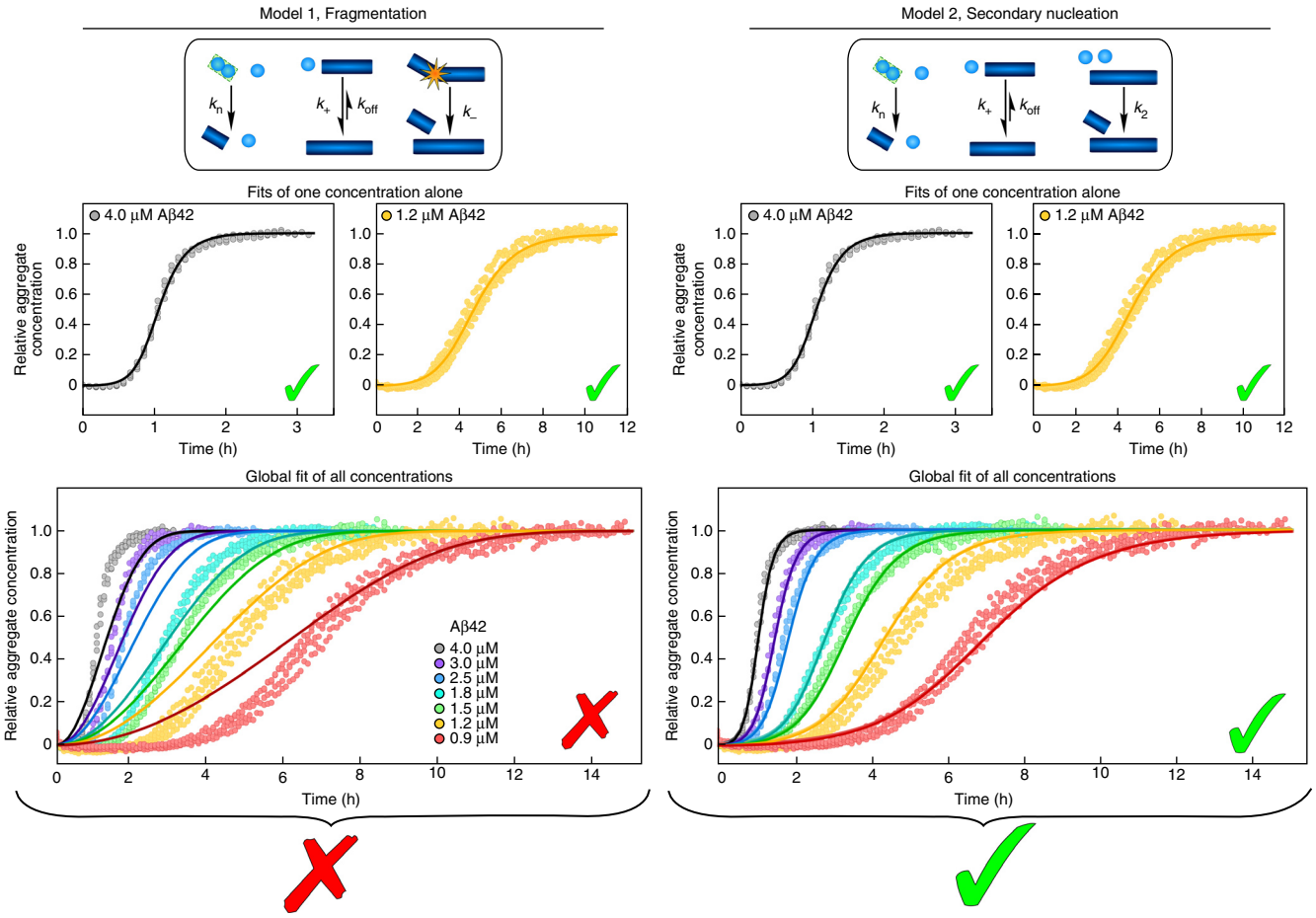


Figure 3 | The power of global fitting. Two models of distinct physical processes (primary nucleation, elongation and fragmentation on the one hand and primary nucleation, elongation and surface-catalyzed nucleation on the other hand) are fitted to the same data set (Aβ42, used in Cohen *et al.*[28]), first fitting each concentration individually, and then using global fitting. The fits of single concentrations are equivalent for the two models, and thus they do not allow one to distinguish which one is more viable. In contrast, the global fits clearly show that the aggregation is consistent with surface-catalyzed secondary nucleation, and not with fragmentation of fibrils.

microscopic processes can be grouped into two classes, namely processes that change the aggregate mass and processes that change the aggregate number (**Fig. 6**). We consider three processes that affect the aggregate number: (i) Primary nucleation, which depends only on the concentration of free monomers, has a reaction order of n_c and proceeds with rate constant k_n. Homogeneous nucleation in solution is an example of this kind of process. (ii) Secondary nucleation, which depends both on the concentration of free monomers and on the concentration of aggregate mass, has a reaction order of n_2 in monomer and 1 in aggregate mass and proceeds with rate constant k_2. Surface-catalyzed nucleation on existing aggregates is an example of this process. (iii) Fragmentation, which depends only on the concentration of aggregate mass and proceeds with rate constant k, has a reaction order of 1 in aggregate mass.

The change in aggregate mass by contrast is dominated by the elongation process, in which free monomers add to the growth competent ends of existing fibrils. Although nucleation processes also produce aggregate mass, this contribution is usually negligible in comparison with elongation; indeed, if nucleus formation was a significant contribution to the overall aggregate mass, the average fibril length would be expected to be close to the nucleus size.

These processes were initially all modeled as single-step reactions, but recent experiments have shown that their multi-step nature may become apparent through saturation effects at high monomer concentrations, both in the case of secondary nucleation[18] and elongation[24,25]. We therefore also include models that treat secondary nucleation and elongation as multi-step reactions, with Michaelis-Menten–like kinetics: in a first monomer-dependent step, monomeric peptides bind to existing fibrils, and then in a second, monomer-independent step they rearrange and extend the fibril in the case of elongation, or rearrange and detach as a newly formed nucleus in the case of secondary nucleation. This generalized description introduces a new parameter in the form of the Michaelis constant K_M for secondary nucleation and K_E for elongation, which determines the monomer concentration at which saturation effects become important.

Currently, the program brings together integrated rate laws for a range of general descriptions of filament formation derived from the classical models of linear polymerization[10,26], through to the inclusion of secondary nucleation[9], and recently derived rate laws valid for the full course of the reaction[11,18,20] (**Fig. 6**).

Box 1 | Terminology

Nuclei: Nuclei are the smallest aggregates for which addition of monomers is more favorable than loss of monomers. They behave in the same way as larger aggregates within our models.

Primary nuclei: Nuclei that are formed from monomers, whose formation does not require the presence of aggregates (cf., secondary processes).

Secondary processes: Processes that produce new aggregates with a rate that also depends on the concentration of existing aggregates, an auto-catalytic pathway. The associated positive feedback is the origin of the sudden increase in aggregation rate after long lag-times seen in many systems.

(Un)seeded: The system is referred to as 'seeded' if preformed aggregates are added to or are present in a solution of monomer at the beginning of the aggregation experiment. The seed concentration refers to the mass concentration of the preformed aggregates in the reaction solution at the beginning of the aggregation, M_0. Unseeded experiments refer to experiments in which there is no aggregate mass present initially.

Global fitting: Fitting of experimental data that depend on several changing variables, e.g., time and concentration, simultaneously to a single rate law.

Half-time: The time at which half the protein mass present has been converted into aggregates. This quantity will usually be used in unseeded experiments.

Scaling: The exponential dependence of the half-time on the monomer concentration is given by the scaling exponent γ. Usually, this will be a negative number, indicating that the half-times decrease (i.e., the reaction speed increases) as the monomer concentration increases.

Global fitting. One of the core elements that makes this protocol a reliable method for aggregation data analysis is the fact that we fit large data sets, under a variety of conditions, usually a number of monomer concentrations, simultaneously to a single rate law. This approach enforces a relationship between experimental curves in which the free fitting parameters, such as rate constants and reaction orders, must be equal for all curves. Fitting under these restrictions is called 'shared parameter fitting' or 'global fitting'. It ensures that the microscopic model has the correct dependency not only on time but also on the other parameters that are varied, such as the monomer concentration. Only in this manner can we obtain sufficient constraints to distinguish between the various complex models describing different aggregation reaction networks.

The fitting process minimizes the mean squared residual error (MRE), given by

$$\frac{1}{N}\sum_{i=0}^{N}\left(y_i - f(t_i)\right)^2$$

where N is the number of data points, y_i is the measured value at time point t_i and $f(t_i)$ is the model function evaluated at that time point. Mathematically, this represents a search for the global minimum on an n-dimensional energy landscape, where n is the number of free fitting parameters. In the case of large data sets and complex equations, as encountered here, these energy landscapes can be rough, and if the initial guesses for the fitting parameters are far from their values at the global minimum a simple gradient descent minimization can easily get trapped in a local minimum. Equally, a Monte Carlo algorithm, which relies on a random search of parameter space and specific criteria for accepting new parameter values, can also struggle to overcome barriers, and it may take a long time to converge. In order to mitigate these issues, our software makes use of a basin-hopping algorithm during the fitting process[27]: this method relies on coupling a simple gradient descent algorithm to a Monte Carlo step. After each Monte Carlo step, a simple minimization is performed and the minimized value is recorded as the energy for the Monte Carlo algorithm. The energy landscape then sampled by the algorithm is much smoother (i.e., there are fewer local minima as barriers between adjacent minima have effectively been removed (**Fig. 1e**)) than in the other algorithms using only Monte Carlo or simple minimization by itself. The smoothed landscape enables the algorithm to sample a much larger region of phase space to find the global minimum—i.e., the best fit to the data (**Fig. 1f**). The details of this algorithm are described in Meisl *et al.*[18].

Model verification. Finally, it is important to verify the model that is deemed to be the most suitable in the fitting process by making predictions based on this model and verifying them through further experiments. A simple test is the variation of another system parameter—for example, the initial fibril concentration. Such additional data introduce a new degree of freedom that the

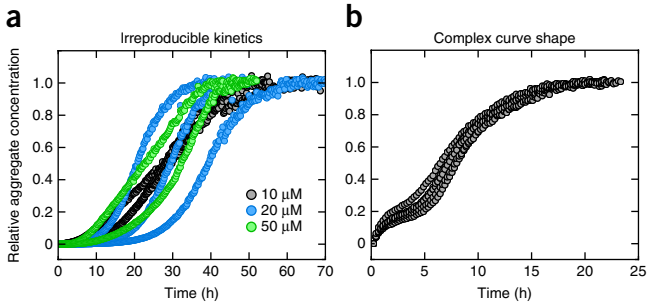


Figure 4 | Data unsuitable for analysis. (**a**) The data are not sufficiently reproducible; the variation between repeats is larger than the difference between the different concentrations. (**b**) The data are reproducible, but they display complex, biphasic behavior, which cannot be described by our nucleation/growth models. Such complex behavior can be an intrinsic, and hence relevant, property of the aggregating protein, but often it is due to poor control of initial conditions and insufficient sample purity, as was the case in the data in **b**. Further experimental optimization and quality control in both systems yielded data of sufficient quality for fitting.

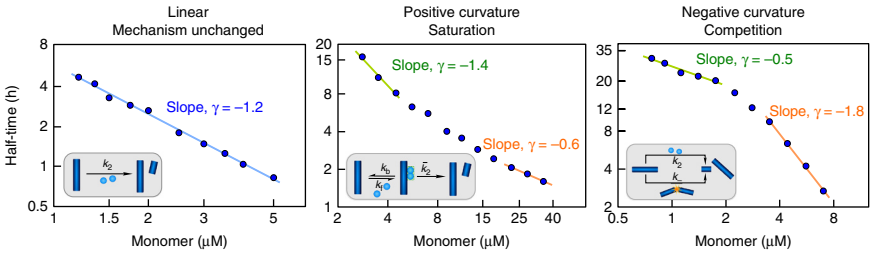
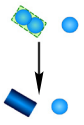
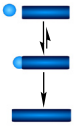
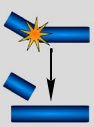
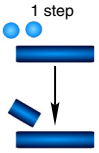
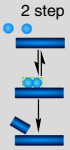


Figure 5 | Half-times as a guide to mechanisms. The curvature of the double-logarithmic plots and the value of their slopes give insights into which aggregation mechanisms are dominant. Linear plots indicate that the dominant mechanism does not change at different monomer concentrations, positive curvature indicates the presence of saturation effects in the dominant mechanism and negative curvature indicates that competition of processes in parallel is present. The detailed ranges of scaling exponents for each model are given in **Figure 6**. A flowchart to help decide on probable models using scaling exponents is shown in the **Supplementary Methods**. The data displayed in the plots are, from top to bottom, Aβ42 (used in Cohen *et al.*[28]), Aβ40 (used in Meisl *et al.*[18]) and Aβ42 at low ionic strengths (S.L. and X. Yang, unpublished data).

model has to reproduce correctly. In addition, other parameters can be investigated—e.g., the concentration of oligomeric species and how it is affected by the addition of seeds[28] or the fibril length distribution. Depending on the type of experiment, the analysis of the new data may not follow this protocol, but predictions about the outcome, based on the fitting of the original data, can still be made and verified.

Investigating effects of mutations, binders or solution conditions. Once a mechanism has been established and verified, it may be used as a basis to investigate the specific effect of various factors on the microscopic processes of protein aggregation. Factors such as mutations in the protein sequence and changes in solution conditions (pH, ionic strength) may alter the dominant mechanism of aggregation and require a new independent analysis; however, they can still be analyzed using the same strategy as above. By contrast, the effect of the addition of compounds that are able to bind to species present in the system may change the fundamental nature of the underlying reaction network, by introducing new steps that were not accounted for in the aggregation models of pure protein. Under such circumstances, a full analysis of the system will require the global fitting of models that explicitly include the presence of potential binders, at a number of monomer and binder concentrations. These models are currently in development, and they will allow a full analysis of the effects in future versions of AmyloFit. However, under certain conditions, such as a fast binding of the compound compared with the rate of aggregate growth, the full model yields the same solution as the model of pure protein aggregation, in which the simple rate constants are replaced with effective rate

			1° Nucleation	Elongation 1 step	Elongation 2 step	Fragmentation	2° Nucleation 1 step	2° Nucleation 2 step
								
			k_n, n_c	k_+	k_+, k_E	k_-	k_2, n_2	k_2, n_2, K_M
		Scaling	Curvature in $t_{1/2}$ plots					
Model names	Nucleation elongation*	$-\frac{n_c}{2}$	None	✓	✓			
	Secondary nucleation**	$-\frac{n_2+1}{2}$	None	✓	✓		✓	
	Fragmentation	$-\frac{1}{2}$	None	✓	✓	✓		
	Fragmentation and 2° nucleation	$-\frac{1}{2}$ to $-\frac{(n_2+1)}{2}$	–ve	✓	✓	✓	✓	
	Multistep 2° nucleation	$-\frac{(n_2+1)}{2}$ to $-\frac{1}{2}$	+ve	✓	✓		✓	✓
	Saturating elongation	$-\frac{n_c}{2}$ to $-\frac{(n_c-1)}{2}$	+ve	✓	✓	✓		
	Saturating elongation and 2° nucleation	$-\frac{(n_2+1)}{2}$ to $-\frac{n_2}{2}$	+ve	✓	✓	✓	✓	
	Saturating elongation and fragmentation	$-\frac{1}{2}$ to 0	+ve	✓	✓	✓		

*Oosawa and Asakura **Ferrone *et al.*

Figure 6 | Mechanisms and models. A list of all the different mechanisms of aggregation considered in this protocol is given, together with the models from which they are derived. Many of the simpler models are limits of the more complicated ones—e.g., nucleation elongation is a special case of all other models, when secondary processes become negligible. The scaling exponents that are given are therefore only valid if all processes in the model are in fact contributing significantly to the aggregation reaction. The solid ticks signify the mechanisms that are involved in the modeled process. The grayed-out ticks signify that this mechanism is a limit of another mechanism included in the model—e.g., one-step elongation is the limit of two-step elongation at low monomer concentrations. The first two models, marked by stars, are based on the models of nucleated polymerization, which were developed by Oosawa and Asakura[10], and nucleated polymerization with secondary nucleation, developed Ferrone *et al.*[9]. The remaining models were developed by Knowles and colleagues[11,16,18]. +ve, positive; –ve, negative.

constants that incorporate the effect of the binder. Therefore, the presence of the compound can be modeled as a perturbation to the kinetics of the system in the absence of the compound. However, it is important to note that this perturbative analysis can yield important insights into what process is affected by the presence of the binder[12], but it does not contain information about the specifics of the binding itself, such as the rate constants.

In summary, this protocol enables researchers to convert experimental kinetic data measured in aggregation assays into rate constants for the underlying microscopic processes, and to identify the mechanisms that are compatible with the experimental measurements. Not only is this procedure crucial for the mechanistic interpretation of data, but the fitting of initial data sets can also aid in a rational approach to the design of additional kinetic experiments and suggest additional tests for mechanistic conclusions. The online nature of our software, including personal accounts, allows for easy cloud storage and sharing of data, as well as for collaborative analysis by various groups in any location. We believe that this strategy will help to bring together experimental and theoretical approaches to understanding the nature and mechanism of protein aggregation and its links to human disease.

MATERIALS

EQUIPMENT

- Data of the total aggregate concentration versus time; see **Box 2** for advice on designing the wet-lab experiments such that the data will meet the quality requirements and the **Supplementary Data** file for a sample data set
- A computer with either Firefox or Google Chrome web browser installed
- A working Internet connection
- A spreadsheet software package, such as Microsoft Excel or OpenOffice Calc, to aid in producing the correctly formatted input files
- An account at http://www.amylofit.ch.cam.ac.uk (Registration and use are free of charge.)

EQUIPMENT SETUP

Software Open your browser, go to http://www.amylofit.ch.cam.ac.uk and log in with your account. While you are on the website please use only the links provided; do not use your browser's back and forward buttons. Note also the help pop-ups, which are small circular blue buttons with a question mark; they contain useful hints and instructions on all aspects of the program.

Box 2 | Data requirements and quality control

The details of the protocols for the purification of specific peptides or proteins peptide and for the aggregation experiments depend on the specific system under study, and they are not part of this general analysis protocol. In our laboratories we have obtained good results for the analysis of Aβ using recombinant proteins and the method in Walsh *et al.*[30] (for details also see TROUBLESHOOTING), although for some batches additional gel filtration steps are needed. However, synthetic Aβ is known to behave differently and in a less reproducible manner[31], which is most likely due to the presence of impurities in the form of proteins with very similar sequence, which cannot be removed easily. Regardless of the specific purification protocol used, some general requirements have to be fulfilled in order to obtain data suitable for the determination of kinetic parameters by fitting. The importance of these data quality controls cannot be overemphasized; the conclusions of a kinetic analysis are only meaningful if assumptions are verified and the system is carefully controlled. Good data are characterized by their reproducibility (we recommend at least triplicate repeats), low signal-to-noise ratio and sigmoidal shape (or at least the shape of the latter part of a sigmoid in the case of seeded experiments), with a clear plateau at the completion of the aggregation reaction.

Purity and controlling initial conditions

Protein aggregation reactions are complex networks that involve positive-feedback loops, which mean that they can be very sensitive to small variations in conditions. It is of crucial importance to ensure high purity of all reagents and to take full control of environmental factors including pH, ionic strength, salt composition, inertness of surfaces, minimized and controlled air-water interfaces, absence of co-solvents and so on.

One important factor is the purity of the monomeric protein, not only in terms of the absence of other molecular species but also in terms of the absence of small aggregates. It is especially important to ensure the latter, as the presence of even a small quantity of preformed aggregates can often catalyze the aggregation reaction and it can markedly alter the observed kinetics. Crucial factors include mixing and keeping solutions at low temperatures until the start of the aggregation reaction, avoiding unnecessary waiting and handling times, to minimize losses and to avoid inducing aggregation before time zero.

For seeded experiments (i.e., experiments beginning from monomer supplemented with a known concentration of seeds), the careful control of initial conditions becomes even more important: Seeds can strongly catalyze the aggregation reaction; therefore, the dead-times need to be as short as possible. This can be achieved by keeping all monomer samples on ice while adding the seeds to their respective wells, finally adding monomer using a multichannel pipette and then commencing the readout as quickly as possible. Moreover, the efficiency of seeds depends not only on their mass concentration but also on the number of ends (i.e., on the average length). As the average length may be easily influenced by the specifics of the seed preparation, the same batch of seeds should therefore be used when comparing seeded experiments differing, for example, in seed or monomer concentration, or another solution parameter.

Poor control of initial conditions and purity may be reflected in the irreproducibility of the data, as shown in **Figure 4a**, or complex behavior that cannot be described by our models of linear polymerization, as shown in **Figure 4b**, in which the data display biphasic behavior. Such complex behavior could also be an intrinsic property of the aggregation under perfectly controlled conditions, in which case our models of linear polymerization may not be sufficient to describe the behavior. However, more commonly such complex behavior is due to the presence of an impurity or poor control of conditions, so very thorough quality control is required to rule out this possibility. A repeat of the entire experiment, using a new batch of protein and purifying it separately, is particularly important in this context.

(continued)

Box 2 | (Continued)

Linear scaling of the experimental signal with aggregate concentration

Initially, before any extensive kinetic studies are performed, the method used for detecting the total aggregate concentration needs to be verified and optimized. A popular method to detect amyloid fibrils is the use of dyes, such as thioflavin T, whose fluorescence intensity increases upon binding to the fibrils. Two factors need to be considered when using dyes: first, the dye needs to report quantitatively the aggregate mass, which can be checked by aggregating the protein at a number of monomer concentrations and determining the plateau value of the fluorescence intensity at completion of the aggregation reaction. The type and concentration of the dye should be optimized to give a linear scaling of the total protein concentration with the value of the fluorescence intensity at the plateau, in the concentration region relevant to the aggregation reactions. This is achieved by setting up test aggregation reactions at several peptide or protein concentrations and at several dye concentrations and plotting the signal after reaching the plateau against the peptide or protein concentration. The dye concentration range that gives a proportional response is chosen (see also next point about checking monomer concentrations at the plateau).

Second, the dye should not interfere with the aggregation reaction, which may be tested by using an alternative, dye-free detection method, e.g., circular dichroism spectroscopy, and by comparing the aggregation kinetics in the presence and absence of dye. Another approach is to compare the kinetic curves of experiments to which the dye is added at different time points[18,32].

Note that even small changes in conditions in terms of pH, ionic strength or additives may have significant effects on dye binding and fluorescence, so these checks need to be performed for every new system. In addition, it should be ensured that the technique is indeed detecting all aggregate mass and that there is no major build-up of invisible species—i.e., aggregate species that do not react with the dye. This can, for example, be achieved by comparing the depletion of monomer with the appearance of aggregates, at different concentrations.

Very generally, regardless of the method of aggregate detection used, the curves obtained need to be proportional to the total mass concentration of aggregates. As a further measure to ensure a reliable conversion of the signal into an aggregate concentration, we recommend that all aggregation experiments be run to completion, i.e., until the plateau value is reached, as this plateau can then be used for normalization.

Monomer concentrations at equilibrium

In most cases, aggregation reactions are performed at concentrations well above the solubility of the monomeric protein; therefore, the free monomer concentration at completion of the aggregation reaction is low or negligible. However, for unknown systems, this assumption needs to be verified by measuring the free monomer concentration remaining in solution when the plateau of the aggregation reaction has been reached[33].

Choosing relevant conditions

One important decision is the choice of concentration regime that should be studied. The accessible concentration range may be limited by the experimental techniques used to detect aggregate mass, but ideally a regime relevant in the context of the importance of the system under study should be chosen. It is also important to ensure that concentrations are low enough for to avoid substantial crowding and viscosity effects, as these are not considered in the models. Moreover, concentrations should be chosen to be low enough to avoid effects such as gelation and the formation of higher-order structures, which are not part of our description of the formation of linear aggregates. The range of monomer concentrations that needs to be covered to obtain a meaningful mechanism often depends on the accuracy of the measurement compared with the magnitude of the monomer dependence. In addition, saturation or competition effects often occur over a large range of monomer concentrations, which need to be sampled fully to determine these effects. As a rough guide, a variation of one order of magnitude in monomer concentration should in most cases yield sufficient constraints to distinguish between models provided that the data are of good accuracy. Spacing concentrations evenly in logarithmic space (e.g., every concentration 20% lower than the previous one), using serial dilutions, usually gives good separation of aggregation curves, as the half-times scale exponentially with concentration (for details also see TROUBLESHOOTING).

In the past, many systems have been studied under shaking or other forms of agitation, but in recent work we have shown that such conditions can in fact alter the aggregation mechanism substantially—e.g., by increasing the fragmentation processes[28]. Therefore, care needs to be taken when considering aggregation data under shaking conditions, as mechanical stress introduces an additional effect and may skew the aggregation mechanism to one that is not relevant for an in vivo situation. If at all possible, we therefore recommend that aggregation be studied initially under quiescent conditions, in order to obtain the mechanism in the absence of external mechanical stimuli.

PROCEDURE

Part 1: data upload ● TIMING ~30 min, depending on the size of the data set

1| *Data format*. Export your data of total aggregate concentration (for example, in units of fluorescence, if the data were obtained by recording thioflavin T fluorescence) versus time into a spreadsheet. The first column should contain the time points, and the consecutive columns should contain the measured values, with each column corresponding to a repeat or to an experiment under different conditions. If the data are initially obtained in rows, transpose your data, e.g., by copying

the data to the clipboard and then using 'Edit'→'Paste Special'→'Transpose' in Excel or OpenOffice. We advise you to use the first row to name the columns with descriptive headers—for example: 'time' | 'monomer 5uM' | 'monomer 6uM' | —as these headers will be used as the default name for the curves once they are uploaded into the program.
▲ **CRITICAL STEP** Because the data layout only has one column for time, you cannot combine data that were recorded at different time points in one file. If your data consist of several experiments that do not share the same time axis (i.e., that were not recorded at exactly the same time points), make several files to upload individually.
▲ **CRITICAL STEP** A note on units: upload the data in the units of time in which you would like to receive your output. The program will only give dimensions (e.g., time or concentration); it will not give specific units. Hence, the units of time in the final output will be the same as the units of the time column in the initial input. The units of protein concentration of the curves are irrelevant if the data are normalized for fitting.

2| Save the data as a plain text file, with columns separated by 'tab'. In Windows, this can be achieved by simply highlighting the relevant cells in the spreadsheet and then copy-pasting them into the plain text editor, Notepad. (As detailed below, one can also choose to upload comma-separated data (.csv file) instead, if this option is selected in the drop-down menu 'Data Format Options'.) The file may only contain numerical values, except in the first row, which is assumed to be the header (unless selected otherwise in the 'Data Format Options'). It may not contain empty cells—i.e., all rows and columns must be of equal length (**Supplementary Methods** and **Supplementary Data**).

Starting a new session

3| Now open your browser and log into your AmyloFit account at http://www.amylofit.ch.cam.ac.uk. To keep different experiments and studies separate, the use of different sessions for each is encouraged. To start a new session, click on the 'Change Session' link in the top right, and then enter a name for the new session and press 'Create Session'. Scroll down to the newly created session and click 'Load', which will take you to the new session.

4| In the new session, press 'Add data'. On the new page, use the 'Browse' button to select the previously formatted data file from your file system. If the data do not contain a row for headers, or if they are comma-separated rather than tab-separated, select the corresponding options in the 'Data Format Options' drop-down menu on the left, and then click 'Load new data'.
? TROUBLESHOOTING

Data pre-processing

5| After pressing 'Load new data', you will be taken to the pre-processing page, where you can select the relevant region of your data, discard failed experiments and normalize your data. You will now be asked to pre-process each curve (i.e., column) separately. At the top of the page, the number of curves remaining in the uploaded file is displayed. Decide which pre-processing steps you need to do.

6| (Optional) On the top left, specify a new name for the curve. The default name is of the form <file name>:<column header>.

7| Select the region of data that is relevant for analysis. The data can be found in the bottom right-hand panel; use the mouse to drag the selection window. A zoomed-in version of the selected region is displayed in the plot above it. By default, the whole data range is selected. Discarding single points is not possible.
▲ **CRITICAL STEP** The protein should generally be kept on ice until the start of the measurement, and often the measurement is performed above room temperature (e.g., at 37 °C). Therefore, there is an initial period (usually not more than 5 min) during which the sample temperature equilibrates within the plate reader. During this time, the signal may vary unexpectedly—e.g., it may decrease because of differing fluorescence efficiencies at different temperatures. When selecting the data, make sure not to include the part of the curve that has been affected by this variance.
? TROUBLESHOOTING

8| Below the name, specify the parameters for offset or baseline correction: choose the number of points to average over in order to obtain the background value, and set the value at which this background is meant to be, in units of normalized data. In unseeded experiments, the initial aggregate concentration (and therefore also the offset value) is 0—i.e., one will not expect any signal at the beginning of the experiment, so any signal present initially is background.
▲ **CRITICAL STEP** In a seeded aggregation, a substantial amount of aggregate mass may be present at time zero, and this needs to be taken into account during normalization. To illustrate this point, consider an aggregation reaction containing as much aggregate mass as free monomer at time zero—i.e., 50% seeds—then the offset value would be set to 0.5. The number of points to average over depends on the level of noise, and it should be chosen to be sufficiently large to give a representative average. The region chosen is displayed in red in the top right-hand plot. Choose 0 points in order not to include any offset (i.e., $y_{baseline} = 0$ in equation (1)).

9| Next, choose the number of points over which to average in order to obtain the plateau value; this can be found under the menu point 'Endpoint normalisation'. There should be enough points to give a representative average, but not to include any signal before the plateau is reached. In general, all aggregation experiments should be run to completion (i.e., until the plateau value is reached) to give a reliable conversion of the signal into an aggregate concentration. The region chosen is displayed in green in the top right-hand plot.

10| To upload the curve with these parameters, press 'Submit'. If the data are of too poor quality to be analyzed, press 'Discard' to discard this one curve. To discard all remaining curves that are not yet uploaded and to return to the main fitting page, press 'Finish and return to fitter'.

Upon upload, the program normalizes the data using the equation

$$y_{\mathrm{norm},i} = \left(1 - M_{0,\mathrm{frac}}\right)\frac{y_i - y_{\mathrm{baseline}}}{y_{\mathrm{plateau}} - y_{\mathrm{baseline}}} + M_{0,\mathrm{frac}} \quad (1)$$

where y_i is the original value of the ith data point, $y_{\mathrm{norm},i}$ is its normalized value, y_{baseline} is the average value of the data at the baseline, y_{plateau} is the average value of the data at the plateau and $M_{0,\mathrm{frac}}$ is the relative initial concentration of aggregates (i.e., a value between 0 and 1). The original data are also saved, so after the upload is complete you can return to this pre-processing page and change any of the options, by clicking the 'Edit' link next to the desired curve on the main fitter page.

? TROUBLESHOOTING

Merging and sorting data

11| Once all data have been uploaded and pre-processed, you will be taken back to the main fitter page. The left column displays all uploaded data, with various options for each curve. Below are the save and load buttons, followed by the model selection section. The right-hand column shows a plot of the uploaded data, with various customization and download options, followed by the fitting section, which will display the fitting results once the data have been fitted.

Now, before analysis, all repeats of the same experiment should be merged. Select all repeats of an experiment by ticking the box on the right of the curve name, and then click the merge button. Merging will result in them being treated as one curve; they will have the same color and take up only one slot. Curves can be unmerged again at any point, and indeed to edit curves they have to be unmerged.

12| After merging your curves, you may want to change their order, so it reflects better the trend in, e.g., the monomer concentration used. The curves are colored in order, from a spectrum going from black (purple) to red, to help recognize trends. To change the location of a curve, use the arrow keys next to the curve to move it up or down the list. Alternatively, you can also press the 'Change curve order' button and enter the desired position for each curve.

13| You may want to look at a subset of the data that have been uploaded. In this case, use the boxes on the left of the curves to (un)select them as desired. This feature can be used to plot or fit to only some of the uploaded curves. All ticked curves will be used in the plotting and fitting; by default, all curves are ticked. After (un)selecting curves, press the update plot button to update the plot with the new selection of data.

Part 2: analysis ● TIMING 2–3 h

▲ CRITICAL Normalized data are needed for determining the half-times of each curve. See **Supplementary Methods** (section 2.1) for the algorithm for extracting the half-times.

14| *Half-times and scaling*. To determine the half-times, press the 'Half time plotter' link next to the model selection. For each individual curve, you now need to specify the quantity that is varied between experiments. When determining the mechanism of aggregation of a protein, rather than, e.g., the effect of a binder, this quantity will be the monomer concentration, as its effect on the half-time can give insights into the dominant mechanism of aggregation. However, the dependence of the half-time on other system parameters—e.g., the concentration of seeding material or a potential inhibitor (see Step 45)—can also be investigated simply by inputting this quantity instead of the monomer concentration.

15| After entering these values and de-selecting any unwanted data as in Step 13, press the 'Make Plot' button. A log(half-time) versus log(varied quantity) plot will then be produced and a straight line will be fitted to this plot, with the slope (i.e., the scaling exponent) displayed at the top of the plot. The half-times and the plot can be downloaded using the buttons underneath the plot.

? TROUBLESHOOTING

16| Take particular note of the scaling exponent, as well as any curvature in the double logarithmic plots, as this information will be important in the next step, choosing a model (**Fig. 6**): if the points lie on a straight line, then there is no sign of the dominant mechanism changing with monomer concentration (**Fig. 5**). In contrast, a variation in the slope indicates a change in the dominant mechanism. In general, a positive curvature in the plot indicates a saturating process, where the monomer dependence decreases with increasing monomer concentration as the system becomes saturated. An example is the saturation of secondary nucleation (model: 'multistep secondary nucleation'). Negative curvature indicates the competition of processes in parallel, where the more monomer-dependent process dominates at higher monomer concentrations. This is the case for a system with both secondary nucleation and fragmentation (model: 'fragmentation and secondary nucleation').

The relevant values or range of values of the scaling exponent are given in **Figure 6** and in the help pop-ups for each of the models on the website. Section 3 of the **Supplementary Methods** contains a flowchart that should also help in interpreting the scaling exponent.

Choosing models

17| Several microscopic models are pre-implemented in the software. Consult **Figure 6** to see which models are available and what microscopic process (reaction mechanism) each is modeling. Schemes for these processes are also included in the help pop-ups at the beginning of the model selection section of AmyloFit. The equations that are being fitted for each of these models can be found in section 3 of the **Supplementary Methods**.

18| Determine whether any secondary processes are present. In the presence of secondary processes, existing aggregates catalyze the formation of new aggregates, leading to positive feedback, which manifests itself experimentally as a sudden increase in aggregate mass. The curve shapes are often an indication of the presence of secondary nucleation: secondary process–dominated aggregation gives curves with a flat lag phase and a sudden increase (the curves at early times are exponential), and by varying the monomer concentration these curves are shifted sideways. Aggregation reactions in the absence of secondary processes instead tend to show a more gradual increase with time (the curves at early times are polynomials); the effect of an increase in monomer concentration is mainly to make the curves steeper. If you are unsure about the presence of secondary processes, include both models with and without secondary processes in the fitting. Furthermore, seeded experiments can be performed to distinguish clearly between primary and secondary dominated aggregation reactions. These are part of the model verification process, but they can also be performed at this stage (see Step 36). If there is no evidence for secondary processes, use the model 'nucleation elongation' or 'saturating elongation' (**Fig. 6**).

If secondary processes are present, the choice of the model becomes more complex. In that case, to decide on a model, use the information obtained in the half-time analysis in Step 16, **Figure 6**, and the flowchart in the **Supplementary Methods**. ▲ **CRITICAL STEP** Please select the simplest model that obeys the constraints of the half-time analysis. Note that often the simpler models are contained within the more complex ones (e.g., 'nucleation elongation' is a special case of 'secondary nucleation', which again is a limit of 'multistep secondary nucleation'). Fitting of a more complex model will in most cases give smaller mean errors in the fit, but the value of using a more complex model should be judged on the grounds of whether the improvement to the fits is significant compared with experimental errors and the data spread. One should also evaluate the significance of any additional included processes, by checking their rates as outlined in Steps 32 and 33. In summary, if a simple model reproduces the data it should be used instead of a more complex one.

19| Narrow down which models are appropriate by taking into account whether aggregation was seeded or unseeded. Most of the models have a version that explicitly states 'unseeded'. Use the unseeded models for experiments starting from free monomer without preformed aggregates to avoid overfitting. In an unseeded experiment, the initial mass and number concentrations of fibrils are zero, which often leads to a substantial simplification of the equations. Some parameters will become dependent—e.g., the elongation rate constant and the nucleation rate constant can no longer be determined separately, but rather the behavior depends only on the product of the two. By choosing the 'unseeded' model, the combined parameters, rather than the individual ones, will be fitted to take this into account. (See **Supplementary Methods** for details regarding the difference in equations for seeded and unseeded models.) For advanced users, we also provide a custom model feature that allows users to fit the data to any equation of their choice. Details can be found in the help pop-ups and the manual.

20| Once you have chosen a model, select it from the menu and press 'Load Model'.

Parameters and initial guesses

21| Each model is described using an equation that depends on various parameters, some of which are known, such as the initial monomer concentration, and some of which have to be determined in the analysis, such as the elongation rate constant. Given the parameters and the model, we can use the corresponding equation to calculate the aggregate

Box 3 | Parameter types

Parameter type: Description

Global fit: The parameter will be fitted globally—i.e., the fitting algorithm finds the one value that best describes all curves in the data set.

Fit: The parameter will be fitted individually to each curve—i.e., the fitting algorithm finds a different value for each curve. This option may be relevant if the conditions differ for each curve. As an example, consider a compound that inhibits only primary nucleation, which is added at different concentrations (see Step 46); then, one would let primary nucleation be fitted individually, but fit all other rate constants globally (as they are not affected by the inhibitor). It is important to note that this option increases the number of parameters and the complexity of fits significantly. It also removes much of the power of global fitting by (partly) fitting each curve separately, so it should be avoided unless there is explicit reason to use this parameter type, as in this example.

Group fit: Curves can be grouped together, and one value of the parameter will be fitted to each group of curves. This option may be relevant if groups of experiments are performed under different conditions but still have some parameters in common. Again as an example, consider a compound that inhibits only primary nucleation, which is added at different concentrations. For each given inhibitor concentration, several experiments at a different monomer concentration are performed. Then one can let primary nucleation be fitted within the groups of constant inhibitor concentration, but one should fit all other rate constants globally (as they are not affected by the inhibitor). Selecting this option lets a box appear next to each curve to select the group to which it belongs.

Const: A parameter of known value, whose value is different for the different curves. This parameter will not be fitted. Selecting this option lets a box appear next to each curve to input the known value.

Global Const: A parameter of known value, whose value is the same for every curve. This parameter will not be fitted.

mass concentration at any point in time. Select a parameter type for each parameter associated with the model that you have chosen. The parameter type tells the fitter whether a parameter is known and, if not, in what way it should be fitted. The five parameter types are given in **Box 3**, and they allow any possible combination of shared and independent parameters for global fitting.

In most cases, parameters will either be fitted globally or be constants: rate constants are usually unknown and independent of the concentration of reactants; hence, they should be 'global fit' parameters. Often, experiments at several different initial monomer concentrations, m_0, will be performed, in which case m_0 should be a constant, 'const'. **Box 4** summarizes all parameters, their physical interpretation and our recommendations for their type.

22| Input the known values of the constants and the initial guesses for each of the fitting parameters. The default initial guesses are approximately based on Aβ42, at pH 8, in units of hours and mol/liter. They should be adapted accordingly for other proteins or units of time and concentration.

▲ **CRITICAL STEP** All parameters need to be positive; no initial guesses for fitting parameters may be 0.

23| *Seeded experiments*. If preformed aggregates are added to the reaction, there are two quantities that characterize the seeds added: their number concentration, P_0 and their mass concentration, M_0. Although M_0 is usually easy to determine (e.g., from the total amount of monomer that was aggregated fully in order to produce these seeds), the determination of the number concentration can be more problematic; there are three options (A, B and C) for addressing this issue.

(A) Measure P_0

(i) Estimate P_0 by experimentally determining the average dimensions of the fibrils; the fibril dimensions can then be used to estimate the average number of monomers per fibril (**Box 4**) and thereby relate the mass to the number concentration, but these estimates tend to be rather inaccurate (often only to within a factor of 10).

(B) Fit P_0

(i) Set the fibril number P_0 as a fitting parameter (global if the same seed stock was used in all experiments). However, this approach may result in a different issue: the fact that the speed at which seed fibrils elongate depends in the same way on both the elongation rate constant (k_+) and the number of growth-competent fibril ends (P_0) means that k_+ and P_0 may be coupled and cannot be fitted separately.

(C) Set P_0 to arbitrary value

(i) To avoid overfitting issues associated with k_+ and P_0 being coupled, set P_0 to a constant value (e.g., M_0/10,000, which would correspond to an average fibril size of 10,000 monomers) and fit k_+. Once fitted, compute the value of $k_+ P_0$; do not use k_+ alone. The absolute value of $k_+ P_0$ is not an insightful parameter in itself; however, it can be used to compare different experiments, if the same sample of seeds (i.e., the same, albeit unknown, value of P_0) was used.

Box 4 | Parameters

Parameter: Description

m,(m_0): (Initial) monomer concentration. $m(t)$ is the concentration of free, nonaggregated monomer, called m_0 at the beginning of the aggregation reaction. Usually, m_0 is a known parameter.

M, (M_0): (Initial) fibril mass concentration. $M(t)$ is the mass concentration of aggregates—i.e., the equivalent monomer concentration if the aggregates were re-dissolved. Its value at the beginning of the reaction is M_0, which is 0 in the case of an unseeded aggregation reaction. Usually, M_0 is a known parameter, which is determined from the amount of monomer that was fully aggregated to produce the aggregates with which to seed the reaction.

P, (P_0): (Initial) fibril number concentration. $P(t)$ is the number concentration of aggregates, proportional to the number concentration of growth competent ends, which are the points at which the aggregate can elongate. Its value at the beginning of the reaction is P_0, which is 0 in the case of an unseeded aggregation reaction. P is linked to M by the average fibril length, L, via $M/P=L$. P is difficult to measure directly, but it can be estimated from M by using the average fibril length. This may be a known or a fitting parameter.

k_n: Primary nucleation rate constant. This appears as $k_n m^{nc}$ in the rate of formation of primary nuclei. It has units of time^{-1} concentration^{-nc+1}. Usually, this is a global fitting parameter.

n_c: Reaction order of primary nucleation. This appears as $k_n m^{nc}$ in the rate of formation of primary nuclei. Its simple interpretation is that of a nucleus size; however, this interpretation is only valid if the reaction is a simple single-step process. It is unit-less and typically has a value between 0 and 5. Usually, it will be a global fitting parameter.

k_+: Elongation rate constant. This appears as $2k_+mP$ in the rate of formation of new aggregate mass. It has units of time^{-1} concentration^{-1}. Usually, it will be a global fitting parameter.

K_E: Michaelis constant for elongation. This appears as $2k_+mP/(1+m/K_E)$ in the rate of aggregate mass formation, for two-step elongation. K_E has units of concentration, and it gives the monomer concentration at which elongation is half saturated—i.e., the elongation step proceeds at half its maximum speed. Usually, this will be a global fitting parameter.

k_{off}: Depolymerization rate constant. This appears as $2k_{off}P$ in the rate of aggregate mass formation, and it is the rate at which monomers are lost from fibril ends. It has units of time^{-1}. This may be a global fitting parameter. However, in most cases, it is negligibly small and hence set to be a constant.

k_-: Fragmentation rate constant. This appears as k_-M in the rate of formation of new growth competent ends from fragmentation. It has units of time^{-1}. This form of the fragmentation rate assumes that an aggregate is equally likely to break anywhere along its length, with the time scale of breaking given by $1/k_-$. Usually, this is a global fitting parameter.

k_2: Secondary nucleation rate constant. This appears as $k_2 m^{n2}M$ in the rate of formation of secondary nuclei. It has units of time^{-1} concentration^{-n2}. Usually, this is a global fitting parameter.

n_2: Reaction order of secondary nucleation. This appears as $k_2 m^{n2}M$ in the rate of formation of secondary nuclei. Its simple interpretation is that of a nucleus size; however, this interpretation is only valid if the reaction is a simple single-step process. It is unit-less and typically has a value between 0 and 5. Usually, it will be a global fitting parameter.

K_M: Michaelis constant for secondary nucleation. This appears as $k_2 m^{n2}M/(1+m^{n2}/K_M)$ in the rate of formation of secondary nuclei through a two-step process. K_M has units of concentration^{n2} and $K_M^{1/n2}$ gives the monomer concentration at which secondary nucleation is half saturated; i.e., the nucleation step proceeds at half its maximum speed. Usually, this will be a global fitting parameter.

Fitting

24| The procedure for fitting is summarized in **Figure 7**. After entering all parameter values and types, press the 'Fit' button, below the plot on the right.

The number of basin hops, which may be thought of as the number of randomizations of the initial guess, can be set. To do this, enter the desired number of basin hops into the field next to the 'Fit' button. We suggest starting with a low value (3 by default), to make sure that the initial guesses give reasonable results and do not cause a crash of the fitting algorithm. If the fit runs to completion without errors, the number of basin hops should be increased and the fit should be rerun to ensure convergence, simply by entering a higher number and pressing 'Fit' again.

? TROUBLESHOOTING

25| Take note of the mean residual error (MRE) to compare the goodness of different fits or models to the same data set.

26| If a good fit has been achieved (i.e., the data are reproduced to within experimental error), move on to Step 31. If the fits are a poor match to the data, first increase the number of basin hops and determine whether or not this leads to an improvement in the fit. The number of basin hops should be increased until the fit no longer improves, which is an indication that the fits have converged. If increasing the number of basin hops does not improve the fit and the fit is still a poor match to the data, there may be two reasons: first, initial guesses are so far away from the global minimum that even a large number of basin hops cannot find this global minimum—i.e., the fits have not converged. Second, the model (for this specific choice

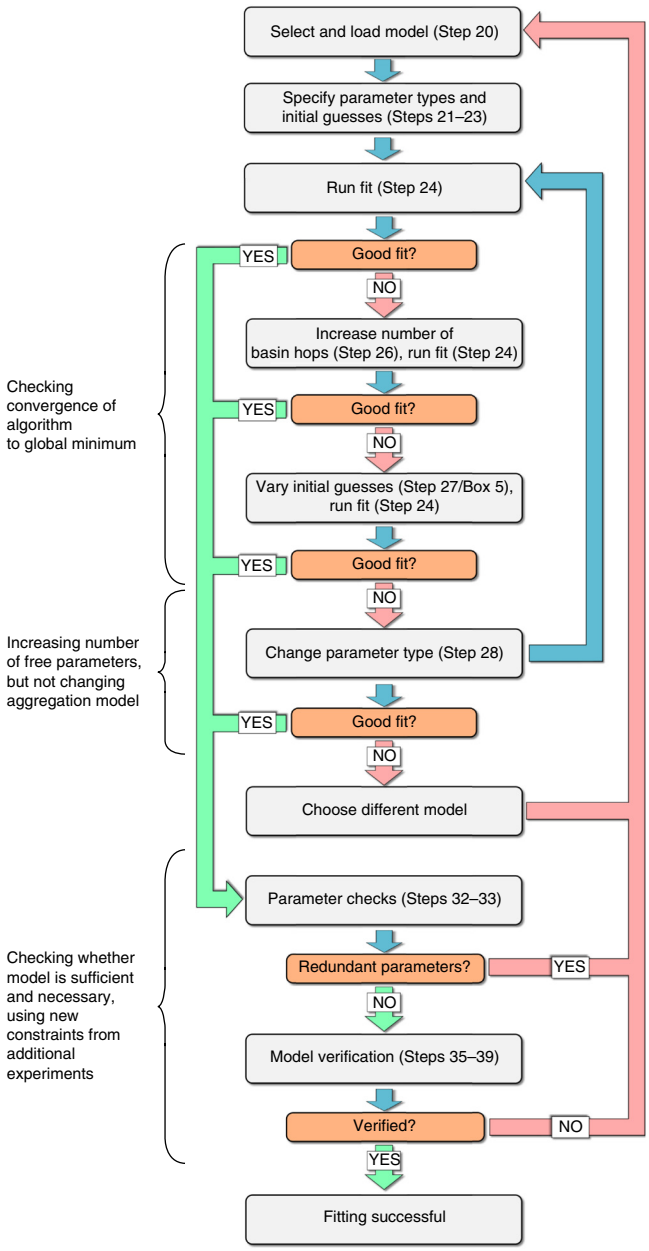


Figure 7 | Fitting flowchart. This flowchart summarizes Steps 20–39. The procedure ensures convergence of the fits (Steps 26–28), by increasing iterations and varying initial conditions. Then, it tests whether the complexity of the model used was necessary (Steps 32 and 33) and whether the model was sufficient to make valid predictions (Steps 35–39).

of parameter types) is not able to reproduce the data. If the fits are very different from the data and happen on a different time scale, in a different part of the plot, there is usually an issue with initial guesses. If there is some overlap with the data and the time scales are similar, but the half-times and the curve shapes are not accurately reproduced, either issue 1 or issue 2 may be the reason.

27| If a good fit has been achieved, move on to Step 31; otherwise, vary the initial guesses as explained in **Box 5** and repeat Steps 24–27.

If neither increasing the number of basin hops nor changing the initial guesses improves the fit, the chosen model and parameter types are likely to be unsuitable for reproducing the data; move on to Step 28.

28| This step is performed if the chosen model does not fit the data. If some parameter types were set to be constant ('const') such as, for example, the nucleus size, try setting them to 'global fit' and rerunning the fit (Steps 24–27). This should be done one parameter at a time. Changing the type to 'fit', rather than 'global fit', should be avoided; it will usually lead to much better fits, but it is effectively like fitting each curve individually, thereby removing the feature that makes this analysis so robust in the first place (**Fig. 3**). If no good fit is found by varying the parameter type, move on to Step 29.

29| This step is performed if the chosen model cannot reproduce the data. Repeat the fitting from Step 20 with a different model. If you have chosen the simplest model consistent with the half-time analysis and other constraints in the previous fitting attempt, this new model will be a more complex one, potentially including another nucleation mechanism or a saturation step. For a completely unknown system, you may have to repeat this procedure (Steps 20–28) several times.

In general, it should be emphasized that a mechanism can never be 'proved' by a good fit. The most one can hope for is that all experimental data are consistent with the model and that the model yields valid predictions, as outlined in Steps 35–40. However, a mechanism can be disproved by a kinetic analysis; in that sense, reporting models that are invalid is often almost as important as reporting the model that fits best. If you are unable to fit the data by the end of Step 28, this can be reported as a valid misfit (i.e., the model is in disagreement with the data). If none of the models is able to fit your data, move on to Step 30.

30| If none of the allowed models (considering any constraints such as the independently established presence/absence of secondary processes) yield an acceptable fit, the monitored aggregation process cannot be described within the framework of our simple linear polymerization models. This may be due to the fact that the aggregation does indeed proceed via a more complex mechanism, which cannot be mapped onto our simpler models. However, a poor fit can also be caused by the presence of experimental artifacts, such as sedimentation affecting the fluorescence signal, or by issues with experimental precision and reproducibility (**Fig. 4**). If the latter situation is suspected, perform further optimization of the experimental procedures and begin again from Step 1.

Box 5 | Fitting strategies and initial guesses

Mean residual error (MRE)

The MRE, which is displayed at the beginning of the fitting results, provides an objective way to judge the goodness of fits. The MRE is the quantity that is being minimized in the fitting algorithm; it is obtained by calculating the sum of the squared residuals, divided by the number of data points minus the number of free parameters:

$$MRE = \frac{1}{(N - N_p)} \sum_{i=0}^{N} \left(y_i - f(t_i, a)\right)^2$$

where t_i and y_i are the dependent and measured variables in the data (e.g., time and fluorescence intensity), $f(t,a)$ is the model function to be fitted, with a being a vector of the parameters and N_p the number of free parameters in the fit. The MRE should not be used to compare fits of different sets of data.

Varying initial guesses

In most cases, convergence of fits should be good if the initial guesses for the rates are within a few orders of magnitude of the correct values. Varying initial guesses is best done by considering what change in the modeled curves would lead to a better reproduction of the data: For example, if the fitted curves never leave the baseline, the entered rate constants are too low. If on the other hand the fitted curves reach the plateau immediately at the start of the reactions, the entered rate constants are too high. If the half-times match but the increase in the fitted curves is too sudden, the rate of the secondary process was probably too high. If the half-times are in the right region but the fitted curves are too spread out, the scaling exponent of the fitted model was too high. By setting all parameters to constants and varying their values, you can develop a better feel for how each one affects the curve shapes.

When changing initial guesses, proceed by varying one initial guess at a time. Steps of two orders of magnitude for the rate constants should be sufficient. Use smaller steps for the reaction orders. Reaction orders usually take values between 0 and 5.

Fixing parameters to aid convergence

If one is experiencing particular difficulty with fits failing to converge, one possible strategy is to fix some fitting parameters as constants to reduce the degrees of freedom of the fit. When a converged fit has been found, the fixed parameters are allowed to vary again in a new fit, using the values from the previous fit as initial guesses. Most commonly, this is used on reaction orders: reaction orders are initially fixed to a constant value and the fit is performed with only the rate constants being fitted. Once a converged fit has been found, the fitted values of the rate constants are used as initial guesses for a new fit. In this new fit, the reaction orders are then also fitted.

Initial guess for reaction orders

Reaction orders enter the rates as exponentials and therefore they can affect the behavior of the fits much more strongly than the rate constants. For this reason, it is important to take care when varying the initial guess of reaction orders, to avoid crashes of the fitting algorithm. One very useful strategy in this context is to compensate changes in initial guess of the reaction orders by changes in the initial guess of the corresponding rate constant. To illustrate this concept, consider primary nucleation: primary nuclei are produced at the rate $k_n m^{nc}$, so this product is the physically relevant quantity. Let us assume that a fit has been performed yielding $k_n = 1$ and $n_c = 1$, and the curves are similar to the experimental curves; however, the monomer dependence, which is determined by n_c, is not correctly reproduced. To rectify this problem, we want to increase the reaction order n_c to 3. The monomer concentration was of the order of µM, so in the first fit the rate of formation of nuclei was given by $k_n m^{nc} = 10^{-6}$. To help the fits converge with a new initial guess of $n_c = 3$, the overall rate of formation of primary nuclei should remain similar. However, simply setting $n_c = 3$ without adjusting k_n would give a rate of formation of primary nuclei of $k_n m^{nc} = 10^{-18}$, 12 orders of magnitude smaller than the value obtained in the first fit. We therefore compensate by re-adjusting the initial guess for k_n as well. The new fit is initiated with $k_n = 10^{12}$ and $n_c = 3$, which gives the same rate of formation of primary nuclei as in the first fit, but now also better reproduces the monomer dependence. This principle applies equally to other processes.

Advanced: redundancy and overfitting

As mentioned in Step 19, in unseeded experiments the rate constants of elongation and nucleation processes become dependent (i.e., they only appear as products of elongation × nucleation). Therefore, there is not enough information to determine their individual values any more. This effect was explicitly considered in the program by the inclusion of unseeded models, so if the protocol for the standard analysis of aggregation mechanisms is followed there will be no issues with dependent parameters. One other case in which parameters become dependent, which may be relevant for the analysis of the effect of binders, is the aggregation at only a single monomer concentration: if only data at one monomer concentration are analyzed, reaction orders and the corresponding rate constants become dependent. Consider, for example, primary nucleation. The rate at which primary nuclei are formed is given by $k_n m^{nc}$, which is the physically relevant quantity. The reaction order n_c describes how this rate depends on the monomer concentration. However, if the monomer concentration is not varied, $k_n m^{nc}$ remains constant and there are two parameters k_n and n_c to describe this one constant, making them dependent. In general, if dependent parameters are identified, one of them should be set as a constant and only one should be fitted, to avoid overfitting.

(continued)

Box 5 | (Continued)

Advanced: use fitted parameters when switching models

Similarly to the strategy of adjusting the initial guesses of reaction orders, this relies on the idea that one should consider the effect on physically relevant parameters, such as the rate of formation of nuclei, when changing models or initial guesses. To illustrate this point, consider the following: the data have been fitted to a fragmentation model. The fits are in the correct region of the plot, but they do not reproduce the spread of curves very well; therefore, we want to fit a secondary nucleation model instead. Secondary nucleation and fragmentation produce secondary nuclei at the rates $k_2\, m^{n2} M$ and $k_{-}\, M$, respectively. Therefore, the initial guess for the secondary nucleation model should be chosen such that $k_2\, m^{n2} \sim k_{-}$. This idea can be extended to other processes, by considering the differential equations describing their behavior (see **Supplementary Methods**, section 3 for the relevant differential equations). Note that using the fitted parameters when switching models only makes sense if the fits of the first model approximately reproduced the data.

31| If a good fit to the data has been found, press the 'Save Backup' button to save the parameters. By pressing the 'Load Backup' button, the model and initial guesses can be loaded again at any later point. In order to obtain errors on the fitted parameters, repeat the fit by pressing the 'Fit with Errors' button. This may take considerably longer than fits without errors; hence, the search for an acceptable fit should always be performed without errors.

The errors obtained in this way are a measure for how much one can vary the parameters without substantially changing the goodness of the fit. In other words, they quantify how strongly the data constrain the parameters of the chosen model to the values obtained from the best fit. This quantity is calculated by varying the parameters about their best fit values and determining the effect of the variation of a certain parameter on the accuracy of the fit. Further details on the algorithm can be found in Meisl *et al.*[18] (http://www.pnas.org/content/suppl/2014/06/16/1401564111.DCSupplemental/pnas.201401564SI.pdf). However, this is only one aspect that determines the accuracy of the fitted parameters with respect to the real values. One will also have to take into account errors in the data and inaccuracies in the measured quantities, such as the concentrations. In order to estimate all these errors, separate fits to a repeat of the entire experiment should be compared. In general, in a kinetic analysis of this kind, the rate constants are unlikely to be more accurate than a factor of 2. Reaction orders, on the other hand, may be determined more accurately, because of the strong effect that slight variations in their values can have on the kinetics.

Results and interpretation

32| Once a converged fit has been obtained, check the parameter values to make sure that the complexity of the model used was indeed necessary. This step constitutes the application of Occam's razor, and it is crucial to ensure that none of the fitted rate constants for any of the processes are so small that their contribution to the overall aggregation behavior is insignificant. If that is the case, the model used was unnecessarily complex.

Depending on the model, there may be up to three different processes producing new aggregates in parallel, so it is necessary to check whether the inclusion of all processes was required by comparing the rates at which each process produces new aggregates:

$$\begin{aligned} \frac{dP_{\text{prim\_nuc}}}{dt} &= k_n m_0^{n_c} \\ \frac{dP_{\text{sec\_nuc}}}{dt} &= k_2 m_0^{n_2} M \\ \frac{dP_{\text{multistep\_sec\_nuc}}}{dt} &= k_2 M \frac{m_0^{n_2}}{1 + m_0^{n_2} / K_M} \\ \frac{dP_{\text{frag}}}{dt} &= k_{-} M \end{aligned} \tag{2}$$

For unseeded experiments, there is no aggregate mass present initially ($M = 0$), so primary nucleation always has to be present to produce an initial concentration of aggregates. However, it may or may not remain the dominant process of formation of new aggregates over the time course of the reaction. To estimate the relative importance of secondary processes, compute the concentration of aggregates, M_{crit}, at which the two processes contribute equally. For simplicity, define the fraction $F_{\text{crit}} = M_{\text{crit}} / m_0$ as

$$F_{\text{crit}} = \frac{1}{m_0} \frac{\text{prim}}{\text{sec} / M} = \frac{k_n m_0^{n_c - 1}}{\frac{1}{M} \frac{dP_{2^o}}{dt}} \tag{3}$$

Where dP_{sec_nuc}/dt refers to the rate of the secondary process as given in equation 3). For example, in the case of secondary nucleation, this would be $F_{crit} = k_n m^{nc-1-n2}/k_2$. If F_{crit} is >0.1—i.e., if the secondary process is only comparable to the primary one when 10% of monomers have aggregated—then that particular secondary process is negligible compared with primary nucleation. Equivalently, different secondary processes can be compared by looking at the ratio of the corresponding rates.

33| In addition to checking the importance of nucleation processes, check the relevance of saturation effects, if the chosen model included any. This can be achieved by comparing the Michaelis constant, K_M, in the case of multistep secondary nucleation, or K_E, in the case of saturating elongation, with the sampled monomer concentrations. If the saturation effects only become effective above the highest monomer concentrations sampled (i.e., $m_{max} < K_E$ for elongation or $m_{max} < K_M^{1/n2}$ for multistep secondary nucleation), then the nonsaturating model should be used instead. Importantly, if the Michaelis constant is not within the range of sampled monomer concentrations, its absolute value will be inaccurate (because the saturation effects are not experienced in the sampled monomer concentration range). In that case the values should be quoted as 'larger than the largest concentration sampled' or 'smaller than the smallest concentration sampled'.
▲ **CRITICAL STEP** In general, if any process is found to be insignificant, the fitting should be re-done with the simpler model not including this process, starting from Step 20. The fitted parameters of the more complex model may be used as initial guesses for the simpler one (**Box 5**).

34| Once all checks have been completed, download all fit details, data points, points for the fit and residuals (data minus fit at each point) as a zip file via the 'Download Data and Fit' button. The files are all in text format, tab-separated. Simply copy-pasting them into your spreadsheet program should display them in a clearly formatted manner. The plots can be customized and downloaded via the 'Plot Download and Options' menu underneath the plot. The curves to be displayed can be chosen on the left (see Step 13).

Part 3: model verification and conclusions ● TIMING several days to weeks
35| *Experimental system for model verification*. The addition of preformed aggregates at the beginning of the aggregation reaction is one of the best methods to test the validity of the chosen model, as there is a clear prediction of its effect. To obtain seeds, simply use the aggregated material obtained at the plateau of the aggregation reactions and dilute it to the required concentration. Sonication of the aggregates may be necessary to allow for uniform mixing during dilution. Note that sonication can also significantly increase the efficiency of seeds by breaking them into smaller pieces.

36| Adding preformed seeds can be used as a qualitative way to check for the presence of secondary nucleation processes: add a small quantity of seed material (usually <1%) at the start of the unseeded reaction. If the half-times are substantially shortened by the addition of these seeds, this is an indication that feedback mechanisms, such as secondary nucleation, have a dominant role in the aggregation process.

37| To sample the elongation process separately from the other processes, add a high concentration (usually >30%) of preformed seeds. In the presence of high concentrations of preformed aggregates, the initial behavior will be determined purely by the elongation of existing seeds. Specifically, the initial gradient is given by $dM/dt = 2k_+m_0P_0$. Perform seeded experiments, with the same concentrations of the same seed stock at different monomer concentrations, and plot the initial gradient versus m_0. The slope of this gives k_+P_0 and thereby the monomer dependence of the elongation rate. Moreover, curvature of this plot is evidence for the presence of saturation effects in elongation[18]. In addition to verifying the chosen model, these data can also be used to obtain additional information on the aggregation in the form of estimates of the elongation rate constant k_+ (see Step 23A and **Box 4** on how to estimate P_0). The estimate of k_+ can in turn be used to estimate the other rate constants, which are often only obtained in the form of products from the unseeded fits (see Step 19).

38| Seeded experiments can also be used in a quantitative analysis to verify the chosen model: add a known concentration of preformed aggregates at the beginning of the aggregation reaction and then analyze the data in the same way as the unseeded data (see Step 40).
▲ **CRITICAL STEP** When performing these experiments, it is important to combine the monomer and the seed material just before the start of the measurement, as even during short dead times a large number of aggregates may already form from the seed fibrils. In general, seeded experiments tend to be less accurate and less reproducible because of the addition of another possible source of error in the form of the seed concentration and composition. Reproducibility between different batches of seed material is low because the average length of seed fibrils is difficult to control and it can significantly influence the kinetics. Always use the same seed stock within one set of experiments.

39| Perform further, independent experiments to verify predictions based on the model that fits the original data. Many other experiments can be envisioned, including the use of selective isotope or IR labels in monomers or fibrils to pinpoint the origin of new aggregates[12,28] or the determination of the concentration of fibrils before they are visible in standard thioflavin T measurements[29].

Analysis of additional data ● TIMING 1–2 h

40| For kinetic data of the total aggregate mass, in the presence of preformed aggregates as seeds, repeat Steps 1–13 to upload the data, taking care to normalize any seeded experiments to the correct initial concentration of aggregate mass. Steps 14–19 become redundant. The fitting of the new data (Steps 20–27) should be performed with the same model and parameter types as the analysis of the original data. The fits can be performed either by fitting all parameters or by fixing the rate constants and reaction orders to the values that have been determined previously, in order to test the predictions of the model based on the original data. If the chosen model is valid, good fits should still be achieved and the fitted rates should be comparable to those obtained in the unseeded experiments.

Part 4 (optional): effect of mutations, binders and conditions

41| After the aggregation mechanism of the original protein sequence or conditions has been determined, one may want to investigate how a variation in different factors may alter the aggregation mechanism. How you will do this will depend on whether the variation will result in a different aggregation system. For variations in which the aggregation system can be considered the same (e.g., if the binding of the compound is fast compared with the aggregate growth processes), proceed to Step 42.

A variation in solution conditions, shaking or a protein sequence mutation, however, effectively constitutes a different aggregation system. In this situation, the mechanism determined for the original mutant or solution conditions can be used as a guide to identify probable models and good initial guesses, but the analysis is effectively independent of the original one. To investigate the effect of solution conditions or mutations, repeat the analysis from Step 1. In this case, it is very important to remember to repeat all the data quality control steps and dye concentration optimizations under the new conditions.

(Optional) Effect of small concentrations of potential binder (experimental) ● TIMING several days to weeks

42| (Optional) The approach outlined in the main text will indicate which processes the potential binder affects, as long as these effects do not alter the aggregating system and only affect a single process. Perform aggregation experiments at a fixed concentration of protein, with increasing amounts (in small sequential steps) of the compound to be tested for its effect on aggregation. Ideally, this is done at several different protein concentrations spanning the range of concentrations used in the determination of the original mechanism in the absence of the compound. The following analysis is similar to the original one, Steps 1–31, but it is modified in some parts.

(Optional) Effect of binder (analysis) ● TIMING 1–2 h

43| The data quality control procedure needs to be performed again. In particular, check the dye scaling and monomer concentration at the plateau level. This is to ensure, for example, that effects on dye binding are not misinterpreted as binding of the compound.

44| Follow Steps 1–13 for data upload as before. Treat each set of data at one protein monomer concentration separately.

45| Steps 14–19 of the original analysis are not relevant here. The half-times as a function of inhibitor concentration (not as a function of monomer concentration!) may be determined in the same manner as before, by entering the inhibitor concentrations as the varied quantities, rather than the monomer concentration as in Step 14. The half-times can then be used to visualize the effect of binder; however, they cannot be interpreted in the same way as for the pure protein in terms of reaction orders.

46| For the fitting, use the model determined in the global fits of pure protein, set all parameters but one to 'global constant' and use the values determined in the global analysis without binder for these parameters. Set the one remaining parameter to 'fit' and use the value determined in the global analysis without binder as the initial guess, and then fit the data (Steps 24–27). This will find the best fit, varying only one parameter between different binder concentrations. Do not set more than one parameter to 'fit'; too many degrees of freedom will result in overfitting and render the results meaningless.

47| Repeat Step 46 for each independent (see below) fitting parameter and note whether or not the data can be well fitted. Note that changes of nucleation rate constants and reaction orders become indistinguishable, if only a single monomer concentration is used (**Box 5**). Therefore, variations in the curves at one monomer concentration can equally well be described by changes in the reaction orders and changes in the rate constants. However, as the scope of this analysis is merely to determine which process is affected, this does not present an issue.

48| If a good fit to all curves is achieved by fitting one parameter, the effect of the binder can be rationalized by assuming that it affects the corresponding microscopic process.

Plot the fitted parameter versus binder concentration to determine whether a trend can be observed. If the binder does indeed interfere with the corresponding microscopic process, one would expect a monotonic dependence of the value of the fitted parameter on the binder concentration. Compare these results at the different protein monomer concentrations to ensure that they agree qualitatively (i.e., which process the presence of the binder affects). If no good fit is achieved, the binder leads to more complex behavior than can be captured by this perturbative approach, and this simplified kinetic analysis is insufficient.

49| This perturbative analysis should be used as an indication of which part of the aggregation mechanism is likely to be affected by the presence of the binder. Perform additional experiments in order to further support these findings. They could, for example, include a study of the effect on the concentration of oligomers[12], or an investigation of the efficiency of adding preformed seeds to the reaction. The latter example would, e.g., apply if the binder is predicted to shift the system from a secondary nucleation to a primary nucleation–dominated mechanism or vice versa, as these two mechanisms can easily be distinguished through their different seeding efficiencies (see Step 36).

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**.

TABLE 1 | Troubleshooting table.

Step	Problem	Possible reason	Solution
All	Buttons or menus do not work, and the page is frozen	There is an issue with the loaded page	Try reloading the page by using the 'Back to main fitter' link or your browser's refresh button Log out and back into your account Restart your browser Reminder: do not use your browser's back and forward buttons on AmyloFit
4	'Format not understood' error message	The format of the input data did not follow the guidelines	Check for the following possible causes: There were one or more empty cells in the original spreadsheet of data, copied into the text file. This problem can often be missed if the empty cell is at the end of a row There was a non-numerical value in the file (except for the first row only numerical values are allowed) The options in 'Data Format Options' were not chosen correctly—i.e., the wrong separator was chosen, or the first row was selected to be data, although it contains headers
7	Plateau cannot be determined	Data are of poor quality or the time of measurement is too short to reach the plateau	Discard this curve. If no plateau value can be found, the data cannot be normalized. Depending on the cause, either run your experiments for longer or address noise issues in the measurement
10	Data are not reproducible	Impurities are present or environmental conditions are not sufficiently controlled	Several steps can be taken to improve sample purity and control of conditions: Use recombinant rather than synthetic protein (even small amounts of material with mismatched sequence can significantly affect kinetics) Purify the protein until no contaminant proteins are seen on silver-stained SDS-PAGE and no small-molecule impurities are seen in ^{1}H-NMR spectra Isolate the monomer and store it as identical aliquots, and then isolate the monomer again right before beginning the kinetic experiments to remove any oligomers that may have formed during storage Combine monomer and seeds as late as possible, and keep the samples on ice until start of measurement. Make sure to control the presence of active surfaces; use nonbinding plates and degas the buffers to avoid air bubbles

(continued)

TABLE 1 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
15	Half-times are extracted incorrectly	Data are very noisy	For very noisy data (>30% of increase in signal) noise may be misinterpreted as the curves reaching the half-time. Data with this level of noise should not be used for fitting
24	Invalid initial guess error	The initial guess was invalid for one or more of the parameters	The error could be because of the initial guess not being a number, being a negative number or being 0 if the parameter type was not chosen to be 'Const' or 'Global const'. For any fitted parameter, the initial guess needs to be >0, for constants the value needs to be ≥0
	Fit gives 'NaN' (not a number) error message	The initial guess resulted in a non-numerical value of the fitted function (e.g., division by 0, logarithm of a negative number)	This is an issue with the initial guesses, and some models are more susceptible to this issue than others (because of the functional forms of the equations that are being fitted). Try varying initial guesses until there is no longer an error message, and then proceed with the fitting as usual. Be particularly careful with varying reaction orders (**Box 5**)
	Fits do not complete, and the counter for the number of iterations is not changing	There has been an issue with the fitting algorithm or the communication with the server	Press the 'Stop' button, restart the fit. If the problem persists, reload the page, using the 'Back to main fitter' link, adjust initial parameters and restart the fit
	Parameters change, but fit does not	You are probably overfitting your data	This will be the case if two or more of the fitted parameters are coupled (**Box 5**). For example, if both k_n and k_+ are fitted in an unseeded experiment, this results in overfitting because only their product is defined. To avoid this issue, use the correct model (unseeded) or fix one of the fitting parameters to a constant value
	Fits look jagged	Numerical computation issue	The fits can sometimes have a jagged appearance at early times. This is due to rounding errors when extreme differences in the magnitudes of the calculated values are involved. It is usually associated with large reaction orders and can be avoided by adjusting initial guesses

● TIMING

Experimental setup optimization and data acquisition: up to weeks or months
Steps 1–13, data upload: 15 min plus 10 min for every 100 data sets
Steps 14–34, data fitting and analysis: 2–3 h
Steps 35–39, model verification (experimental): several days to weeks
Step 40, model verification (analysis): 1–2 h
Step 41 (optional), variation in solution conditions or mutations: repeat of the entire protocol, Steps 1–41
Step 42 (optional), effect of binder (experimental): several days to weeks
Steps 43–49 (optional), effect of binder (analysis): 1–2 h

ANTICIPATED RESULTS

The analysis will determine which models are in direct disagreement with the experimental data and therefore which mechanisms can be discarded as possible explanations for the system under study. Moreover, the analysis should result in a model that reproduces all experimental data, as shown in the fit in **Figures 1f** and **3**, and yields the rate constants and reaction orders of the microscopic processes involved in the aggregation reaction. This information can be used to compare and contrast the behavior of the system studied to that of other previously studied aggregating systems and also gives insight into the dominant mechanism of the generation of new aggregates, which in turn can be a guide toward what processes or species a possible inhibitor should target.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS This work was supported by grants from the Swedish Research Council (VR) and its Linneaus Centre Organizing Molecular Matter (S.L.), the European Research Council (S.L. and T.P.J.K.), the Cambridge Home and EU Scholarship Scheme (G.M.), the Frances and Augustus Newman Foundation (T.P.J.K.) and the Biotechnology and Biological Sciences Research Council (T.P.J.K.), St. John's College Cambridge (T.C.T.M. and T.P.J.K.) the Marie Curie Intra-European Fellowship scheme (P.A.), and the Engineering and Physical Sciences Research Council (J.B.K.). We thank the members of the Knowles and Linse research groups for their input and testing of the program, in particular X. Yang, R. Gaspar, T. Mueller, P. Flagmeier and G.R. McInroy.

AUTHOR CONTRIBUTIONS G.M., T.P.J.K. and M.V. conceived the project; G.M. and J.B.K. wrote the software; G.M., J.B.K., S.L., C.M.D. and T.P.J.K. wrote the paper; G.M., P.A. and T.C.T.M. designed the analysis in the presence of binders.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

1. Knowles, T.P.J., Vendruscolo, M. & Dobson, C.M. The physical basis of protein misfolding disorders. *Phys. Today* **68**, 36 (2015).
2. Chiti, F. & Dobson, C.M. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* **75**, 333–366 (2006).
3. Dobson, C.M. Protein folding and misfolding. *Nature* **426**, 884–890 (2003).
4. Aguzzi, A. & Haass, C. Games played by rogue proteins in prion disorders and Alzheimer's disease. *Science* **302**, 814–818 (2003).
5. Aguzzi, A. & O'Connor, T. Protein aggregation diseases: pathogenicity and therapeutic perspectives. *Nat. Rev. Drug Discov.* **9**, 237–248 (2010).
6. Hardy, J. & Selkoe, D.J. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* **297**, 353–356 (2002).
7. Hu, X. *et al.* Amyloid seeds formed by cellular uptake, concentration, and aggregation of the amyloid-beta peptide. *Proc. Natl. Acad. Sci. USA* **106**, 20324–20329 (2009).
8. Fersht, A.R. *Structure and Mechanism in Protein Science*. (W.H. Freeman, 1999).
9. Ferrone, F.A., Hofrichter, J. & Eaton, W.A. Kinetics of sickle hemoglobin polymerization. II. A double nucleation mechanism. *J. Mol. Biol.* **183**, 611–631 (1985).
10. Oosawa, F. & Asakura, S. *Thermodynamics of the Polymerization of Protein* (Academic Press, 1975).
11. Knowles, T.P.J. *et al.* An analytical solution to the kinetics of breakable filament assembly. *Science* **326**, 1533–1537 (2009).
12. Cohen, S.I.A. *et al.* The molecular chaperone brichos breaks the catalytic cycle that generates toxic Aβ oligomers. *Nat. Struct. Mol. Biol.* **22**, 207–213 (2015).
13. Arosio, P., Vendruscolo, M., Dobson, C.M. & Knowles, T.P.J. Chemical kinetics for drug discovery to combat protein aggregation diseases. *Trends Pharmacol. Sci.* **35**, 127–135 (2014).
14. Arosio, P., Meisl, G., Andreasen, M. & Knowles, T.P.J. Preventing peptide and protein misbehavior. *Proc. Natl. Acad. Sci. USA* **112**, 5267–5268 (2015).
15. Ruschak, A.M. & Miranker, A.D. Fiber-dependent amyloid formation as catalysis of an existing reaction pathway. *Proc. Natl. Acad. Sci. USA* **104**, 12341–12346 (2007).
16. Cohen, S.I.A. *et al.* Nucleated polymerization with secondary pathways. I. Time evolution of the principal moments. *J. Chem. Phys.* **135**, 065105 (2011).
17. Cohen, S.I.A., Vendruscolo, M., Dobson, C.M. & Knowles, T.P.J. 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).
18. Meisl, G. *et al.* Differences in nucleation behavior underlie the contrasting aggregation kinetics of the aβ40 and aβ42 peptides. *Proc. Natl. Acad. Sci. USA* **111**, 9384–9389 (2014).
19. Abelein, A., Graslund, A. & Danielsson, J. Zinc as chaperone-mimicking agent for retardation of amyloid β peptide fibril formation. *Proc. Natl. Acad. Sci. USA* **112**, 5407–5412 (2015).
20. Cohen, S.I.A., Vendruscolo, M., Dobson, C.M. & Knowles, T.P.J. From macroscopic measurements to microscopic mechanisms of protein aggregation. *J. Mol. Biol.* **421**, 160–171 (2012).
21. Fowler, D.M., Koulov, A.V., Balch, W.E. & Kelly, J.W. Functional amyloid—from bacteria to humans. *Trends Biochem. Sci.* **32**, 217–224 (2007).
22. Cremades, N. Direct observation of the interconversion of normal and toxic forms of α-synuclein. *Cell* **149**, 1048–1059 (2012).
23. Michaels, T.C.T. & Knowles, T.P.J. Role of filament annealing in the kinetics and thermodynamics of nucleated polymerization. *J. Chem. Phys.* **140**, 214904 (2014).
24. Scheibel, T., Bloom, J. & Lindquist, S.L. The elongation of yeast prion fibers involves separable steps of association and conversion. *Proc. Natl. Acad. Sci. USA* **101**, 2287–2292 (2004).
25. Esler, W.P. *et al.* Alzheimer's disease amyloid propagation by a template-dependent dock-lock mechanism. *Biochemistry* **39**, 6288–6295 (2000).
26. Oosawa, F. & Kasai, M. A theory of linear and helical aggregations of macromolecules. *J. Mol. Biol.* **4**, 10–21 (1962).
27. Wales, D.J. & Doye, J.P.K. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* **101**, 5111–5116 (1997).
28. Cohen, S.I.A. *et al.* Proliferation of amyloid-β42 aggregates occurs through a secondary nucleation mechanism. *Proc. Natl. Acad. Sci. USA* **110**, 9758–9763 (2013).
29. Paolo Arosio, P., Cukalevski, R., Frohm, B., Knowles, T.P.J. & Linse, S. Quantification of the concentration of aβ42 propagons during the lag phase by an amyloid chain reaction assay. *J. Am. Chem. Soc.* **136**, 219–225 (2014).
30. Walsh, D.M. *et al.* A facile method for expression and purification of the Alzheimer's disease-associated amyloid β-peptide. *FEBS J.* **276**, 1266–1281 (2009).
31. Finder, V.H., Vodopivec, I., Nitsch, R.M. & Glockshuber, R. The recombinant amyloid-beta peptide Aβ1-42 aggregates faster and is more neurotoxic than synthetic Aβ1-42. *J. Mol. Biol.* **396**, 9–18 (2010).
32. Cukalevski, R. *et al.* Role of aromatic side chains in amyloid β-protein aggregation. *ACS Chem. Neurosci.* **3**, 1008–1016 (2012).
33. Hellstrand, E., Boland, B., Walsh, D.M. & Linse, S. Amyloid β-protein aggregation produces highly reproducible kinetic data and occurs by a two-phase process. *ACS Chem. Neurosci.* **1**, 13–18 (2010).