Supporting Information

Fiber formation seen through the high-resolution computational microscope

by Tomasz K. Piskorz, Vasudevan Lakshminarayanan, Alex H. de Vries, and Jan H. van Esch

CONTENTS

S1. Computational methods

a. Molecular dynamics simulations

Single simulations are done using a modified version of GROMACS,[1,2] which allows simulations with the CHARMM Drude force-field[3] (GROMACS version 2016-dev-20170105-c53d212, which is accessible through git repository git://git.gromacs.org/gromacs.git). We use the CHARMM Drude force-field, which in our earlier work enabled stable CTA fiber simulation.[4] The CHARMM Drude force-field explicitly models electronic polarizability by the inclusion of Drude oscillators, which are small charge-carrying particles connected to atoms. For Drude particles, we used standard parameters described by Lemkul at et.[3] Polarizable force fields are computationally demanding. However, the recent implementation of extended Lagrangian dynamics with a dual Nose-Hoover thermostat allows one to perform simulation efficiently.[3] Since GROMACS has implemented only a thermostat and no barostat for this efficient use of the force-field, most of the simulations are run at constant volume.

Systems were set up in triclinic or cubic simulation boxes under periodic boundary conditions. The compositions of the studied systems are presented in **Table S1**. Before running the production simulation, the energy was minimized using the steepest descent algorithm. Then a short simulation (10,000 steps with a timestep of 1fs) in the NPT ensemble is performed using the V-rescale thermostat[5] at 298.15 K (with coupling time 0.1 ps) and isotropic Parinello-Rahman barostat[6] at 1.0 bar (with coupling time 1.0 ps and a compressibility 4.5 10^{-5} bar^{-1}) with the self-consistent field treatment in which the positions of the Drude oscillators are relaxed to the potential energy minimum at each simulation step. After this equilibration, the volume was kept constant. The production simulations run were done with a dual Nose-Hoover thermostat developed by Lemkul et al.[3], with a coupling constant of 0.005 ps for Drude particles and 0.1 ps for all other particles. All simulations were run at 298.15 K. The equations of motion were solved numerically using extended Lagrangian dynamics with a timestep of 1 fs.

Table S1. Composition of the systems studied in this work.

System		Number of CTA molecules	Number of water molecules (SWM4-NDP model[7])	Concentration [M]
Primary nucleation		8	1549	0.29
Elongation	Attachment	17 (16 in fiber, 1 free)	3450	0.31
	Diffusion	9 (8 in fiber, 1 free)	1490	0.34
Secondary nucleation		16	1620	0.55
Bundling	Two fibers	16	2600	0.34
	Three fibers	24	2350	0.57

b. The framework

A single simulation starting from randomly dispersed CTA molecules in solution does not lead to the formation of fiber, even a short one, within 500 ns,[4] indicating that nucleation might be relatively slow on the simulation time scale. To tackle this issue, we implemented the conformational resampling procedure.[8] In the conformational resampling approach, many short simulations are run, which are analyzed to distinguish several different states by a similarity of conformations. Similar conformations are then grouped into states. The next generation of simulations is run from states which are least visited. In this manner, a large part of conformational space can be explored. The graphical explanation is presented in **Figure S1**. The main advantage of such a framework is that all simulations are unbiased (in the sense that there is no additional term added to the Hamiltonian). That gives a big advantage: on several occasions during the procedure, we realized that our criteria for distinguishing states or for defining the fiber state are not precise enough. Since all the trajectories are unbiased, in the next

iteration, we simply recalculate all states by the modified criteria, and therefore we could still use all previously run simulations.

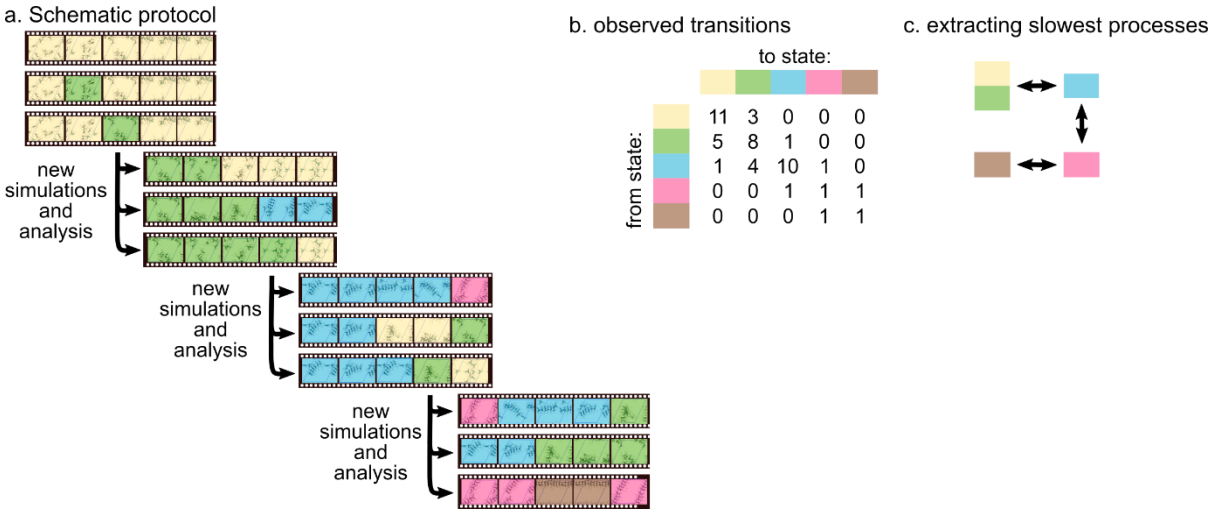

Figure S1. Overview of the conformational resampling simulation protocol. (a) Consecutive sets of simulations are run, and snapshots are analyzed to assign each frame as belonging to a distinguishable state, here represented by coloring each frame. In this work states are characterized by the number of molecules in an ordered stack (details in Section S§2). Each new set of simulations starts from snapshots taken from the collection obtained thus far, biasing the choice in favor of members of less-visited states. (b) Transitions between observed states are counted and collected in a transition matrix, which leads to a Markov State Model that reveals the slowest kinetic pathways between states, (c) grouping states that rapidly interconvert (here yellow and green) into one Markov state.

Here we describe in detail the procedure. First, we run the first generation of simulations from randomly distributed gelator molecules in water. After the first run, we run several scripts that work in parallel. Each script takes all trajectories of simulations finished thus far and for every frame measures the size of the largest ordered cluster (Section S§2). The histogram of these sizes is measured. Most often, we would like to start the simulation with a state which is the least visited. Therefore, we choose the state for the next simulation with a probability inversely proportional to the number of counts for that state in the histogram. Consequently, the least visited state is the most probable as the start of the next simulation, but it still leaves a chance to start from another state. The new simulation is started from a randomly chosen conformation belonging to the selected state. After the simulation is finished, the scripts again analyze all trajectories (also the ones created by other scripts: this is the point where all simulations collaborate with each other) and search for the least visited state. A flow chart of the simulation script of the entire procedure is visually presented in **Figure S2**. We ran 12 jobs running in parallel, from which 8 were performing 10 ns simulations, and 4 were performing 1 ns simulations.

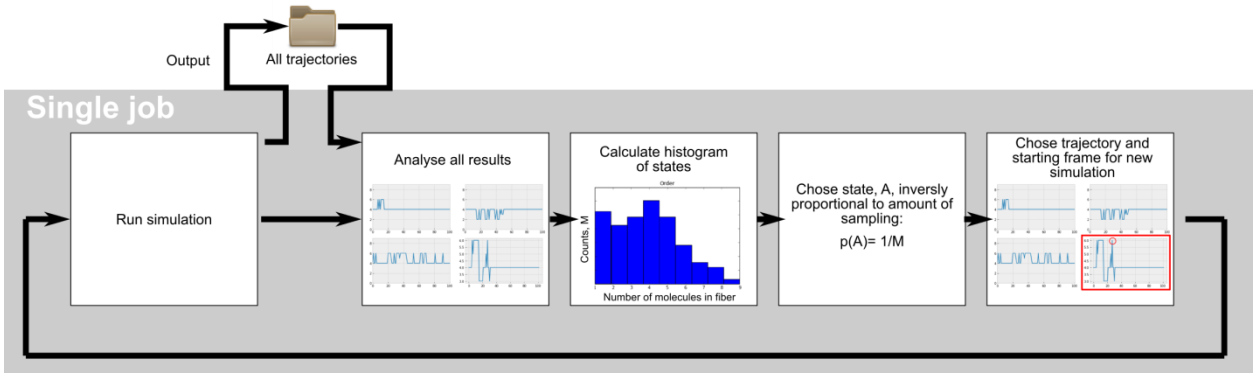

Figure S2. Single script from the framework follows the scheme: (i) run simulation, (ii) combine and analyze all trajectories present in the folder (also created by other scripts), (iii) calculate the distribution of states, (iv) choose rarely visited state (with

probability inversely proportional to the occurrence of the state), (v) choose randomly trajectory in which this state occurs and from this trajectory choose randomly starting frame for simulation with the chosen state, (vi) go to point (i).

S2. Analysis

Cluster analysis. Cluster analysis was done by creating an adjacency matrix A representing adjacent molecules (i.e., molecules in the neighborhood). Two molecules are considered in the neighborhood, according to the function:

$$A_{ij} = \begin{cases} 1, & \sigma\left(\left|\overrightarrow{r_{ij}}\right|\right)K\left(\angle\left(\overrightarrow{n_i}, \overrightarrow{n_j}\right)\right)K\left(\angle\left(\overrightarrow{n_i}, \overrightarrow{r_{ij}}\right)\right) > 0.5 \\ 0, & \sigma\left(\left|\overrightarrow{r_{ij}}\right|\right)K\left(\angle\left(\overrightarrow{n_i}, \overrightarrow{n_j}\right)\right)K\left(\angle\left(\overrightarrow{n_i}, \overrightarrow{r_{ij}}\right)\right) < 0.5 \end{cases} \quad \text{(S1)}$$

Where $\overrightarrow{r_{ij}}$ is the vector connecting the centers of the cyclohexane rings of molecules i and j, $\overrightarrow{n_i}$ is a normal vector to the plane created by the cyclohexane ring of molecule i. $\sigma(r), K(\theta)$ are the switching functions for distance and angle, respectively. These functions are defined as:

$$\sigma(r) = \frac{1 - \left(\frac{r - d_0}{r_0}\right)^6}{1 - \left(\frac{r - d_0}{r_0}\right)^{12}} \quad \text{(S2)}$$

$$K(\theta) = \begin{cases} 1, & \theta < b \; or \; \theta > b \\ 2 - \left|\frac{\theta}{b}\right|, & 0 < 2 - \left|\frac{\theta}{b}\right| < 1 \\ 0, & \left|\frac{\theta}{b}\right| < 2 \; and \; \left|\frac{\theta - \pi}{b}\right| < 2 \\ 2 - \left|\frac{\theta - \pi}{b}\right|, & 0 < 2 - \left|\frac{\theta - \pi}{b}\right| < 1 \end{cases} \quad \text{(S3)}$$

Examples of these functions are presented in **Figure S3** along with data from a trajectory of a long ordered fiber and of an unorganized cluster. We calibrated both functions to give a positive answer for a long ordered fiber, for which we know that all molecules are in the neighborhood. This resulted in parameters $r_0 = 0.2\,nm$, $d_0 = 0.48\,nm$ and $b = 0.35$. Graphical explanation of this function is presented in **Figure S4**. As a result of the clustering algorithm, we obtain the adjacency matrix A, from which we can measure the largest connected cluster.

Ordered clusters. The size of the largest ordered or connected cluster was determined as the largest connected component of graph created from the adjacency matrix A.

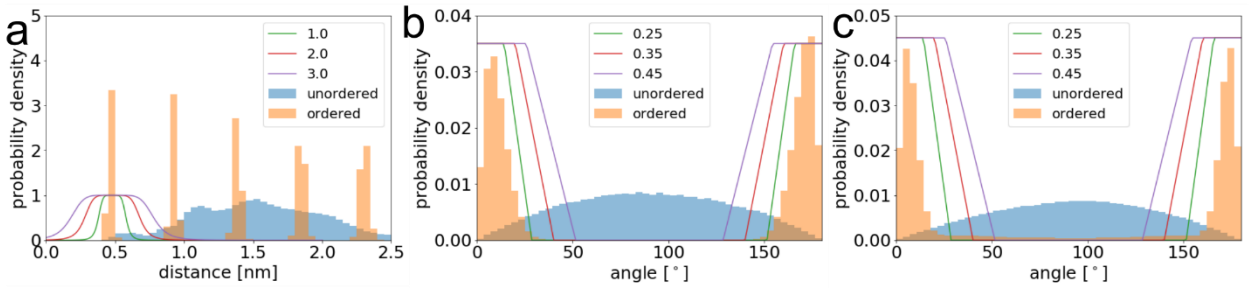

Figure S3. (a) Plot shows switching function, $\boldsymbol{\sigma(r)}$, for different parameters $\boldsymbol{r_0}$. (b) Angle switching functions for different parameters $\boldsymbol{b}$ (normalized to 0.035); the histograms show the data for the angle between the normal vectors of the cyclohexane rings $\angle(\overrightarrow{\boldsymbol{n_i}}, \overrightarrow{\boldsymbol{n_j}})$. (c) Angle switching functions (normalized to 0.045) for different parameters $\boldsymbol{b}$; the histograms show the data for the angle between the normal vector of one cyclohexane ring and the vector connecting centers of two cyclohexane rings $\angle(\overrightarrow{\boldsymbol{n_i}}, \overrightarrow{\boldsymbol{r_{ij}}})$. Histograms show the data for long fibre (orange) and unordered system (blue).

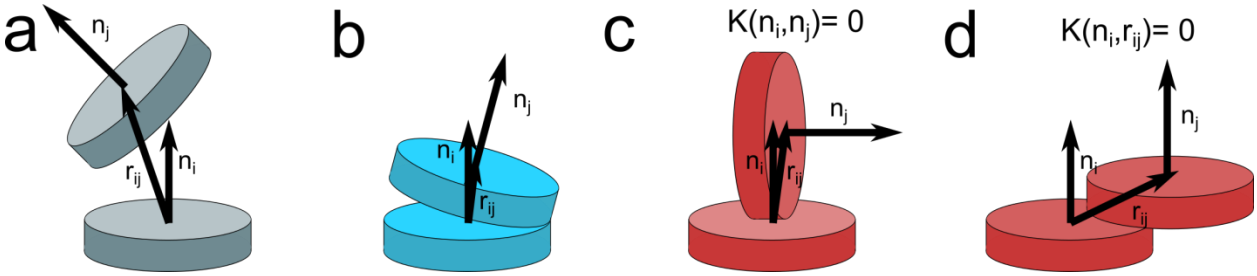

Figure S4. Schematic representation of possible conformations of two molecules. (a) Two molecules are considered in the neighborhood if the distance between them is in a certain range and the angle between normal vectors $\overrightarrow{\boldsymbol{n_i}}$ and $\overrightarrow{\boldsymbol{n_j}}$ is in a certain range, and the angle between the normal vector $\overrightarrow{\boldsymbol{n_i}}$ and the vector connecting centers of two rings $\overrightarrow{\boldsymbol{r_{ij}}}$ is in a certain range. (b) Example of two molecules in the neighborhood, the distance between centers, $|\overrightarrow{\boldsymbol{r_{ij}}}|$, is small, and angles $\angle(\overrightarrow{\boldsymbol{n_i}}, \overrightarrow{\boldsymbol{n_j}})$ and $\angle(\overrightarrow{\boldsymbol{n_i}}, \overrightarrow{\boldsymbol{r_{ij}}})$ are small. (c) Example of two molecules not in the neighborhood, although molecules might be in the appropriate distance range, the angle $\angle(\overrightarrow{\boldsymbol{n_i}}, \overrightarrow{\boldsymbol{n_j}})$ is far from 0° or 180°. (d) Example of two molecules not in the neighborhood, although the distance and angle between two normal vectors of two molecules might be in range, the angle between the normal and the vector connecting two centers $\angle(\overrightarrow{\boldsymbol{n_i}}, \overrightarrow{\boldsymbol{r_{ij}}})$ is far from 0° or 180°.

Unordered clusters. Analysis of unordered clusters was done similarly. Firstly, the adjacency matrix B was constructed, where components were obtained from the distance cut-off function:

$$B_{ij} = \begin{cases} 1, & \sigma(|\overrightarrow{r_{ij}}|) > 0.5 \\ 0, & \sigma(|\overrightarrow{r_{ij}}|) < 0.5 \end{cases} \tag{S4}$$

In this case, only the distance switching function, $\sigma(r)$, is used (none of the angles between two molecules are considered). The parameters d_0 and r_0 used in the switching function are the same for ordered and unordered clusters. The size of the largest unordered cluster is determined as the largest connected component from the adjacency matrix B.

Dipole moment. Dipole moments were calculated using *gmx dipoles*, which is part of the GROMACS package.

Hydrogen bonds. Hydrogen bonds were calculated using *hbonds* packages from VMD. Only hydrogen bonds between amide groups were calculated, and standard parameters were used, i.e., the distance between oxygen and nitrogen must be less than 0.3 nm, and the angle oxygen-hydrogen-nitrogen must be less than 20°.

S3. Markov State Modeling

We have analyzed results using a Markov State Model (MSM) using pyEMMA.[9] Prior to analysis of the system with 8 molecules in which we study primary nucleation, we have excluded all trajectories leading to an infinite fiber (i.e., a fiber crossing periodic boundary conditions) since we did not observe disassembly in any of them. The system leading to the formation of the infinite fiber is therefore not ergodic and does not fulfill MSM assumptions. In the case of the system with 16 molecules in which we study secondary nucleation, this issue did not arise.

The number of ordered molecules has been sufficient as a measure to sample formation of the fiber, but it turned out not to be enough for a Markov State Model, because it resulted in a model which does not fulfill Markov's assumption (see **Figure S5**). Therefore, we have used a more detailed measure for the Markov model. We have measured a vector containing eight elements $v = (a_1, \dots, a_8)$, whose i-th element describes the number of the ordered neighbors of molecule i, that is $a_i = \sum_j B_{ij}$, where $B_{ij} = \sigma\left(\left|\overrightarrow{r_{ij}}\right|\right)K(\angle(\overrightarrow{n_i}, \overrightarrow{n_j}))K(\angle(\overrightarrow{n_i}, \overrightarrow{r_{ij}}))$ (see equations S1-S3). As a result, every value a_i can have a value [0,2] (more than two neighbors is not possible due to the design of the measure that is tailored to detect linear stacks; note that the a_i are real (non-integers numbers) in contrast to the elements of the adjacency matrix). However, such a vector would depend on the numbering of the molecules and two exactly the same systems with different numbering of molecules would be described by different vectors.[10] A simple trick to ensure that the obtained vector is invariant under numbering is to order the elements of the vector from smallest to largest (**Figure S6**).[11]

Using this measure, we were able to construct a Markov State Model for fiber formation. For primary nucleation, we have divided space by assigning every vector to cluster centers. We have defined cluster centers as all possible combinations vectors with integer elements: $v = (0, \dots, \underbrace{0}_{i}, 1, \dots, \underbrace{1}_{j}, 2, \dots, \underbrace{2}_{8-i-j})$, where $i \in \{0, \dots, 8\}$ and $j \in \{0, \dots, i\}$, which resulted in 45 cluster centers, from which only 35 were observed in the simulations (**Table S2**). We have also tested K-means clustering with 2000 clusters (results not shown), but the obtained system was not fulfilling the Chapman-Kolmogorov test.[12] An additional advantage of assigning clusters is that results are easy to interpret. After clustering, we estimated implied timescale (see **Figure S7a**). Then we estimated Bayesian Markov state model[13] using lag-time 87 frames and validate it using the Chapman-Kolmogorov test (see **Figure S7b**). The resulting model has been coarse-grained for 13 states using hidden Markov model.[14] **Figure S7c** shows the result of MSM analysis. It is worth noting that results presented in **Figure S7c** give essentially insights into the kinetics of the process and could be used to calculate transition rates.

The coarse-grained states have been labeled by their most populated cluster center(s) (**Table S2)**. Most of them have one dominant cluster center, which populates over 90% of the state with an exception for 2+2/2+2+2. The labels are used on the graph of transition path sampling presented in Figure 2a in the main text.

Table S2. Compositions of coarse-grained states for primary nucleation. Labels of the states have been chosen as the state with a population of over 90% of the state. Note that vector [1 1 1 1 2 2 2 2], labeled "6+2", is equivalent to 2 stacks, one of size 6 (4 molecules with two neighbors and 2 with one (ends)) and one of 2.

Label	The population of the cluster center in the coarse-grained state	Vector describing the cluster center, *v*
6+2	93.8%	[1. 1. 1. 1. 2. 2. 2. 2.]
	5.2%	[1. 1. 1. 2. 2. 2. 2. 2.]
	1.1%	[1. 1. 1. 1. 1. 2. 2. 2.]
5+2	95.2%	[0. 1. 1. 1. 1. 2. 2. 2.]
	4.1%	[0. 1. 1. 1. 2. 2. 2. 2.]
	0.6%	[0. 1. 1. 1. 1. 1. 2. 2.]
	0.1%	[0. 2. 2. 2. 2. 2. 2. 2.]
4+2	95.2%	[0. 0. 1. 1. 1. 1. 2. 2.]
	3.8%	[0. 0. 1. 1. 1. 2. 2. 2.]
	1.0%	[1. 1. 1. 1. 1. 1. 2. 2.]
8	99.9%	[1. 1. 2. 2. 2. 2. 2. 2.]
	0.1%	[2. 2. 2. 2. 2. 2. 2. 2.]
	0.1%	[1. 2. 2. 2. 2. 2. 2. 2.]
1	100.0%	[0. 0. 0. 0. 0. 0. 0. 0.]
	0.0%	[0. 0. 0. 0. 0. 2. 2. 2.]
5	100.0%	[0. 0. 0. 1. 1. 2. 2. 2.]
6	100.0%	[0. 0. 1. 1. 2. 2. 2. 2.]
	0.0%	[0. 1. 2. 2. 2. 2. 2. 2.]
	0.0%	[0. 0. 1. 2. 2. 2. 2. 2.]
2+2 or 2+2+2	87.1%	[0. 0. 0. 0. 1. 1. 1. 1.]
	8.5%	[0. 0. 1. 1. 1. 1. 1. 1.]
	2.7%	[0. 0. 0. 0. 0. 1. 1. 1.]
	1.5%	[0. 0. 0. 1. 1. 1. 1. 1.]
	0.1%	[0. 1. 1. 1. 1. 1. 1. 1.]
	0.0%	[1. 1. 1. 1. 1. 1. 1. 1.]
3+2	93.3%	[0. 0. 0. 1. 1. 1. 1. 2.]

	3.6%	[0. 1. 1. 1. 1. 1. 1. 2.]
	2.3%	[0. 0. 0. 0. 1. 1. 1. 2.]
	0.8%	[0. 0. 1. 1. 1. 1. 1. 2.]
7	100.0%	[0. 1. 1. 2. 2. 2. 2. 2.]
4	98.1%	[0. 0. 0. 0. 1. 1. 2. 2.]
	1.9%	[0. 0. 0. 1. 1. 1. 2. 2.]
3	100.0%	[0. 0. 0. 0. 0. 1. 1. 2.]
2	98.6%	[0. 0. 0. 0. 0. 0. 1. 1.]
	1.4%	[0. 0. 0. 0. 0. 0. 0. 1.]

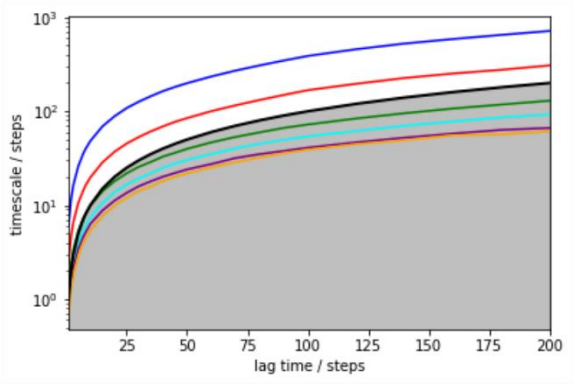

Figure S5. Implied timescales for the MSM for fiber formation (primary nucleation) based only on the number of ordered molecules. Implied timescales do not converge to a constant value. (A constant value of implied timescales means that the process is independent of the choice of lag time.)

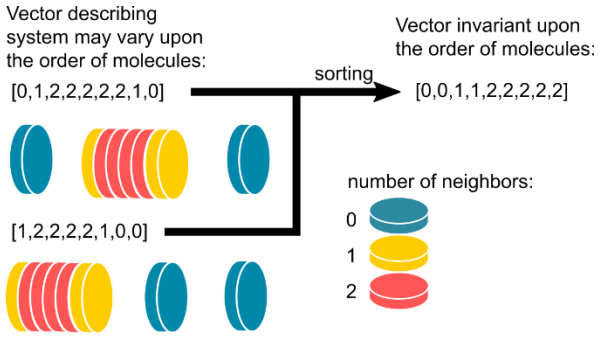

Figure S6. Example of a trick to tackle indistinguishability of molecules: sorting values of the obtained vector.

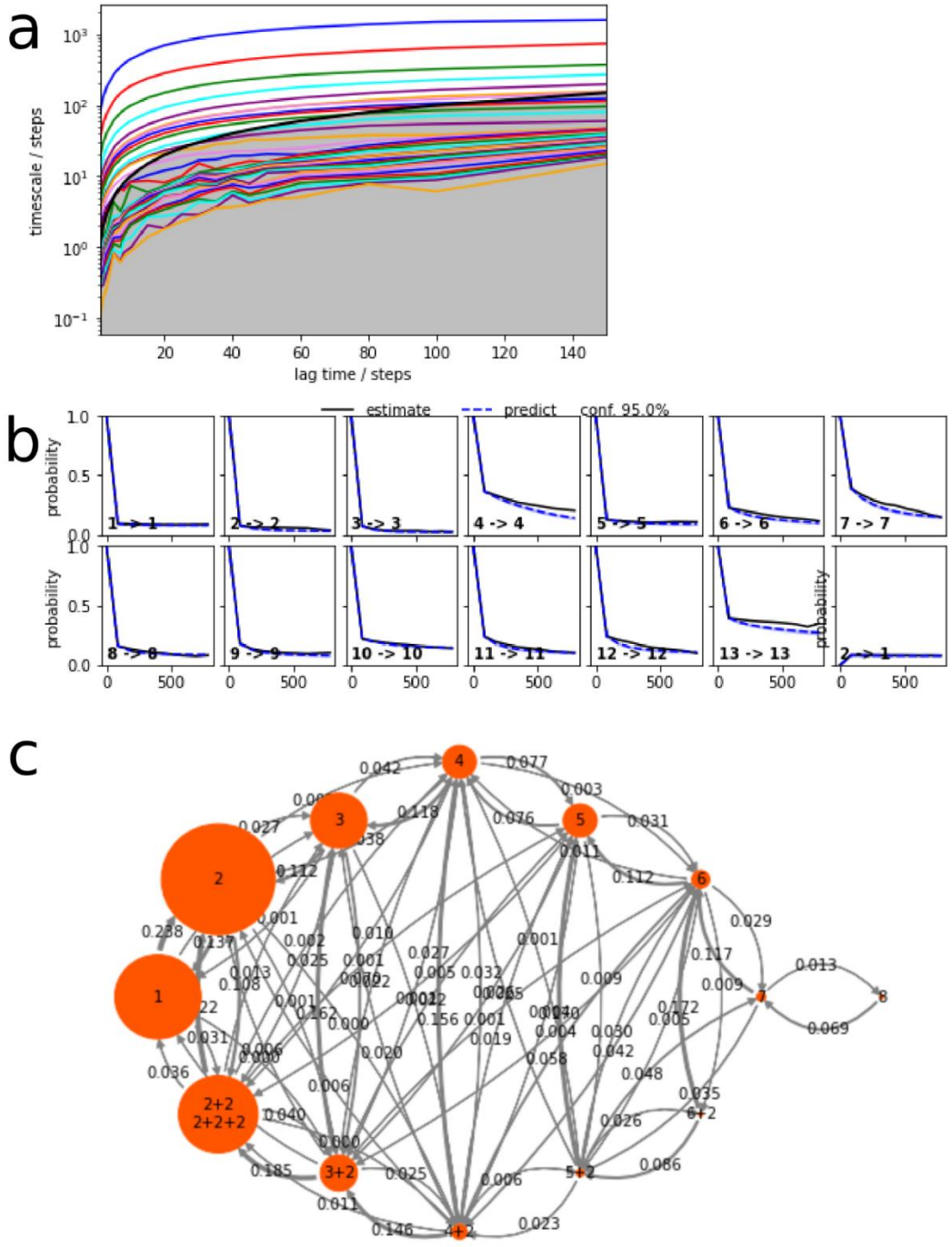

Figure S7. Markov State Model of primary nucleation, using all simulations except the ones in which an infinite fiber is formed. This data is presented in summarized form in the main text, Figure 2. (a) Implied timescales, (b) Results of Chapman-Kolmogorov test. For clarity, we show only the test for self-transitions. (c) Graph representing the coarse-grained model.

A similar analysis has been done for secondary nucleation. One difference that we made was to use all the frames of the simulation, also those forming infinite fiber. In contrast to primary nucleation, states classified as infinite fiber disassembled into an 8-mer fiber. Therefore, the model with these trajectories still is ergodic, and we could use these trajectories.

The results of Markov state model analysis are presented in **Figure S8**. As noted in the main text, the model does not fulfill the Markov assumption, which can be seen in the results of the Chapman-Kolmogorov test (see **Figure S8b**). Similarly, like for primary nucleation, the coarse-grained states have been labeled by their most populated cluster center(s) (**Table S3**).

The newly formed fiber was oriented with its macrodipole (mostly due to aligned amide dipoles) parallel to that of the original fiber, suggesting that there is no strong preferential orientation of the fiber via macrodipole stabilization.

A similar analysis has been done for secondary nucleation (**Table S5)**.

Table S3. Compositions of coarse-grained states for secondary nucleation. Labels of the states have been chosen as the state with a population of over 90% of the state. Note that the last eight elements of the vectors (indicated by red color) are representing existing fiber in the system, and elements are always 2.

Label	The population of the cluster center in the coarse-grained state	Vector describing the cluster center, *v*
inf	90.5%	[2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
	9.5%	[1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
6+2	71.2%	[1. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
	28.8%	[0. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
4+2	98.3%	[0. 0. 1. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
	1.6%	[0. 0. 0. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
	0.1%	[1. 1. 1. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
2+2/2+2+2	89.8%	[0. 0. 0. 0. 1. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2.]
	9.4%	[0. 0. 1. 1. 1. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2.]
	0.8%	[1. 1. 1. 1. 1. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2.]
	0.0%	[0. 0. 0. 0. 0. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
3+2	97.2%	[0. 0. 0. 1. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
	1.2%	[0. 0. 0. 1. 1. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2.]
	1.0%	[0. 1. 1. 1. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
	0.5%	[0. 0. 1. 1. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
5+2	98.2%	[0. 1. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
	1.3%	[0. 0. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
	0.3%	[1. 1. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
	0.2%	[0. 1. 1. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
0	100.0%	[0. 0. 0. 0. 0. 0. 0. 0. 2. 2. 2. 2. 2. 2. 2. 2.]
	0.0%	[0. 0. 0. 0. 0. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
8	93.7%	[1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]

	6.3%	[1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
5	100.0%	[0. 0. 0. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
4	100.0%	[0. 0. 0. 0. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
6	100.0%	[0. 0. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
3	99.6%	[0. 0. 0. 0. 0. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
	0.4%	[0. 0. 0. 0. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
2	98.1%	[0. 0. 0. 0. 0. 0. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2.]
	1.2%	[0. 0. 0. 0. 0. 1. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2.]
	0.7%	[0. 0. 0. 0. 0. 0. 0. 1. 2. 2. 2. 2. 2. 2. 2. 2.]
7	99.8%	[0. 1. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
	0.1%	[0. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]
	0.0%	[0. 1. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2. 2.]

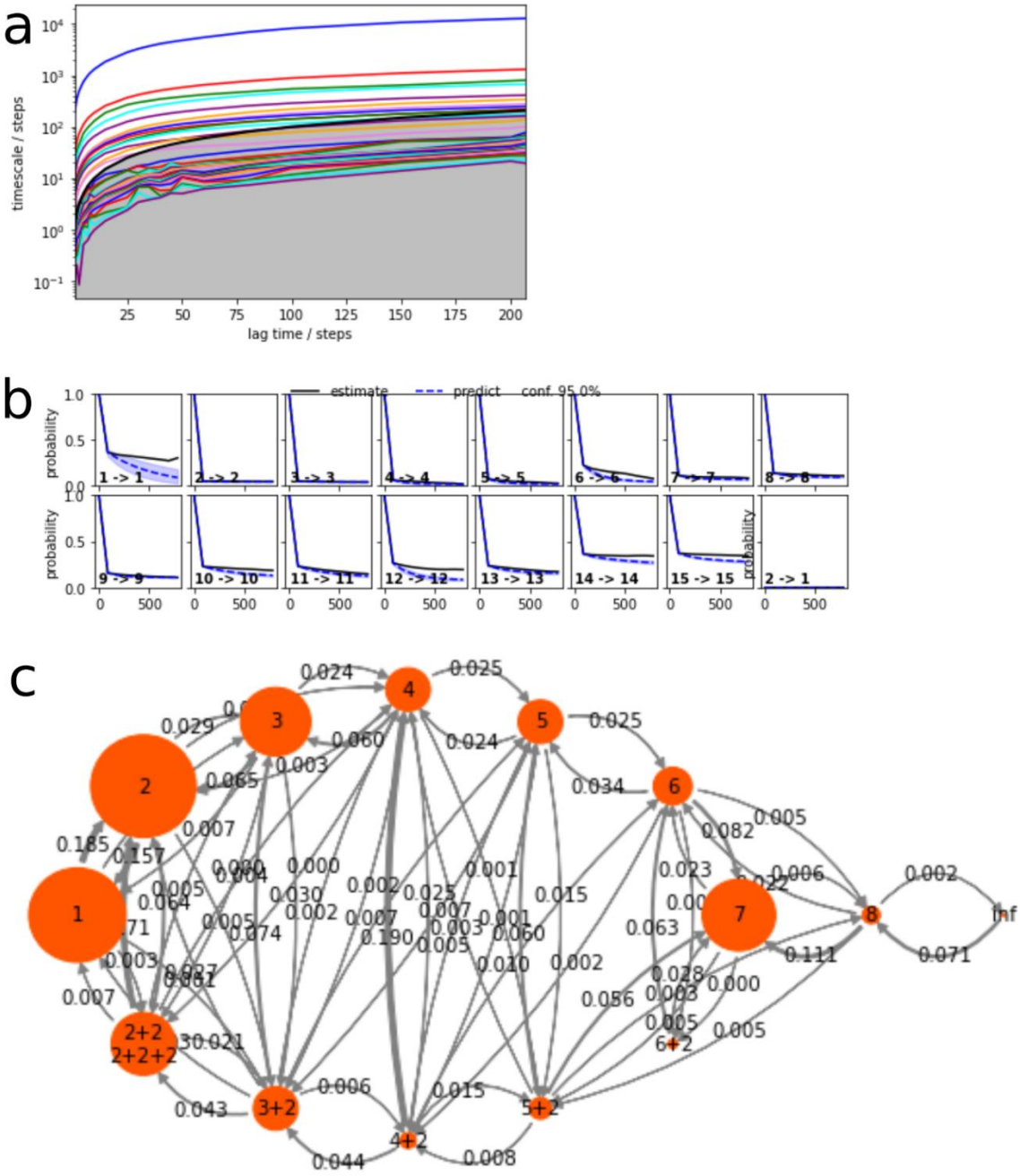

Figure S8. Markov State Model of secondary nucleation simulations. This data is presented in summarized form in the main text, Figure 4. (a) Implied timescales. (b) Results of the Chapman-Kolmogorov test. For clarity, only self-transitions are shown. The test is not fulfilled for self-transition of states 1 and 12. (c) Graph representing coarse-grained Markov model.

S4. Pathway analysis

We applied transition path theory (TPT) [15] to identify the most probable pathways and the net flux between states based on the hidden Markov models. The net flux and mean first passage time are presented in the main text. Here, we give insight into different pathways of the mechanism by pathway decomposition. For clarity, we use a 0.8 fraction of total flux. All results are presented in **Table S4**. For clarity, we show only the first three in the main text. However, the additional three pathways affirm the pattern explained in the main text: pathways follow mostly monomer association with sporadic dimer association.

Table S4. Pathway decomposition of the primary nucleation process. Analysis using transition path theory from randomly distributed molecules to ordered fiber. The path is shown with the labels of coarse-grained states (**Table S2**).

percentage	path
38.6	(1)->(2)->(3)->(4)->(5)->(6)->(7)->(8)
19.3	(1)->(2)->(2+2 or 2+2+2)->(3+2)->(5)->(5+2)->(7)->(8)
16.7	(1)->(2+2 or 2+2+2)->(3+2)->(4+2)->(6)->(7)->(8)
9.4	(1)->(2)->(4)->(5)->(6)->(7)->(8)
8.8	(1)->(2)->(4)->(4+2)->(6)->(6+2)->(8)
7.2	(1)->(2)->(2+2 or 2+2+2)->(3)->(3+2)->(4)->(6)->(8)

Figure S9 represents the changes in the magnitude of the total dipole moment of all CTA molecules and number of H-bonds per molecule along the main pathway.

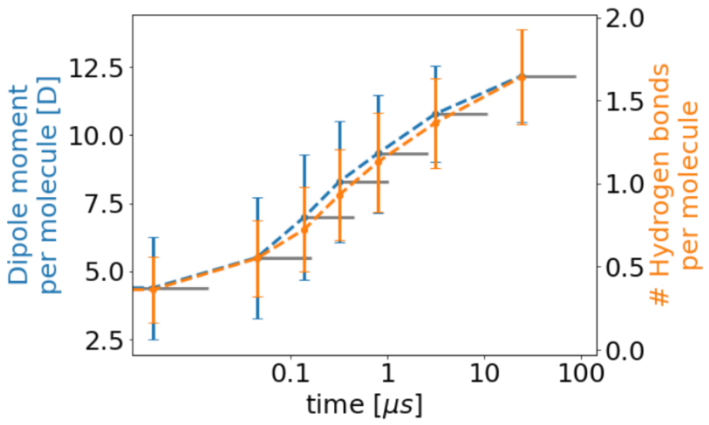

Figure S9. Progress of the magnitude of the total dipole moment of all CTA molecules (blue) and number of CTA-CTA amide H-bonds per CTA molecule (orange). The gray horizontal lines represent the 0.95 confidence intervals of the first passage time to the state.

Table S5. Pathway decomposition of the secondary nucleation process. Analysis using transition path theory from randomly distributed molecules to ordered fiber. The path is shown with the labels of coarse-grained states (see **Table S3**).

percentage	path
61.4	(1)->(2)->(3)->(4)->(5)->(6)->(7)->(8)
19.3	(1)->(2)->(2+2 or 2+2+2)->(3+2)->(5)->(5+2)->(7)->(8)

10.7	(1)->(2)->(3+2)->(5)->(6)->(8)
8.6	(1)->(3)->(4)->(4+2)->(5+2)->(7)->(8)

S5. Stability of the ordered stack.

To study the stability of the ordered stack, we have performed series of simulations of ordered stacks of sizes 2, 3, 5, 8, 12, and 16 in water (the list of simulations is present in **Table S6**). An example snapshot of a dodecamer is presented in **Figure S10b**). Most of the simulations lead to dissociation of one molecule (**Figure S10a**), but we have also observed one fragmentation (**Figure S10c**).

Table S6. List of simulation of ordered stack in water.

Simulation	Number of independent simulations	Time
Dimer	15	10 ns
Trimer	15	10 ns
Pentamer	33	10 ns
Octamer	29	10 ns
Dodecamer	12	10 ns
Hexadecamer	12	10 ns

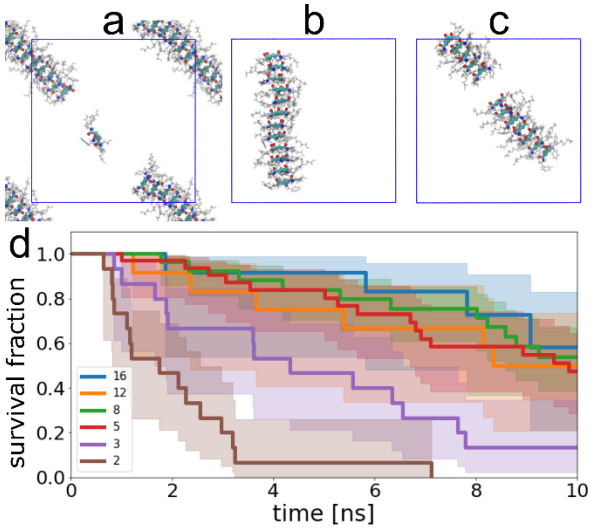

Figure S10. Example snapshots of simulations starting from an ordered stack (b); most of the simulations lead to dissociation of one the molecules (a), but we have also observed one fragmentation (c). The same snapshots (but without simulation box) are presented in the main text, Figure 2e. (d) Survival fraction of the ordered cluster with standard deviation error. The same figure (but without standard deviation error) is presented in the main text, Figure 2d.

S6. Elongation

(i) Molecular dynamics simulations

Adsorption and desorption on the fiber. During the course of ten independent simulations of 15 ns starting from an infinite fiber and one molecule in solution, in each case, the single molecule adsorbs

on the surface of the fiber within the first 6 ns. Later, three processes of desorption can be observed (see **Figure S11a**).

Directionality of diffusion on the fiber. By analyzing the diffusion in the direction of the main axis (z) of the fiber, we have not observed a preferential direction of the movement. **Figure S11b** shows the progress of z-coordinate in 10 independent simulations/molecules analyzed from the moment when molecule adsorbs on the surface of the fiber. We also calculated the autocorrelation function of these coordinates, and it is presented in **Figure S11c**. The autocorrelation shows the average change of the z-coordinate after a particular time, called lag-time. It shows that on average, molecules diffuse more into one direction than the other, but the difference is just a result of limited statistics (see the standard deviation error). We were wondering if molecules orient during the diffusion. Therefore we have calculated directions of dipole moments. However, we have not observed preferential orientation; the dipole moment of the fiber is always pointing in the direction of the z-axis (see **Figure S11d**), whereas the dipole moment of the free molecule is pointing in a random direction (see **Figure S11e**).

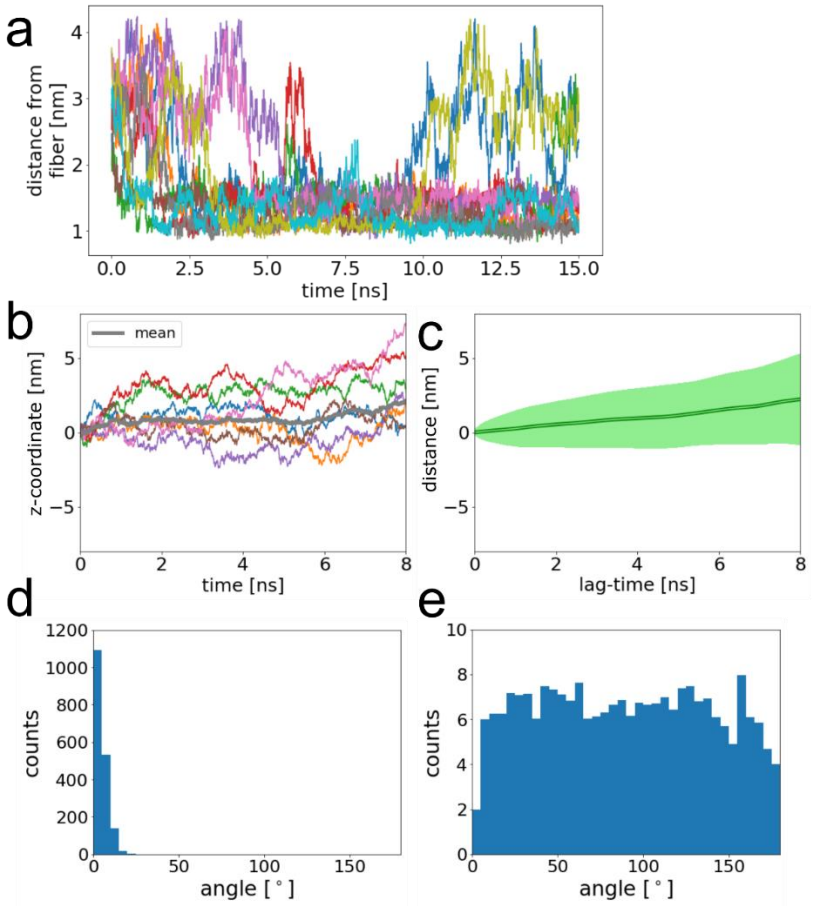

Figure S11. (a) The distance of the free molecule from the center of the fiber (different color represent traces for different simulation). During the first 7.5 ns, all molecules adsorb on the surface. Later three desorption events can be observed. (b-d) The directionality of diffusion of free molecule in a direction parallel to the axis of the fiber. The analysis was performed from the moment the molecule adsorbs on a fiber until the end of the simulation. Moreover, only trajectories in which the molecule does not desorb were taken into account. (b) Progress of z-coordinate for ten different trajectories and its (c) mean autocorrelation function. (d) Histogram of the angle between vector of the dipole moment of molecules creating fiber and z-axis (corrected for the geometrical $\boldsymbol{sin\phi}$ factor[16]). (e) Histogram of the angle between vector of the dipole moment of the free molecule and z-axis (corrected for the geometrical $\boldsymbol{sin\phi}$ factor[16]).

(ii) Elongation: simple models

Let's create simple models of two proposed models of elongation: one for finding the end of a fiber and a second for finding the side of a fiber. Let's imagine that we have a system with one free molecule and

$n-1$ molecules in fiber, and the total concentration is equal to c; the volume of such a system is then equal to $V = n/c$. Let's assume that a single molecule is a disc and has a volume $V_1 = R_1 \cdot h$.

Model without crawling. The fiber has two ends, therefore the probability of finding the end of the fiber by the molecule is equal to:

$$p = \frac{2V_1}{V} = \frac{2V_1 \cdot c}{n}. \tag{S5}$$

This implies that for large fibers, that is $n \to \infty$, we see that $p \to 0$.

Model with crawling. If a molecule can adsorb on a surface, then the probability of finding the surface is equal to $h(n-1)(R2-R1)$, where *h* is the height of one molecule, and *R2* and *R1* are the outer and inner radii of a cylinder surrounding the fiber, from which the adsorbed molecule cannot escape (which can be calculated from the histogram shown **Figure 3c** in the main text). Therefore, the probability of a molecule to attach to the side of the fiber is equal to:

$$p = \frac{h(n-1)(R2-R1)}{V} = \frac{h(n-1)(R2-R1)c}{n}. \tag{S6}$$

This implies that for large fibers, that is $n \to \infty$, we see that $p \to \frac{h(R2-R1)}{c}$ $(const.)$.

S7. cryo-TEM imaging

The gel was prepared by suspending 11.24 mg of CTA in 1221.52 mg of water, heating the mixture until dissolved, and then allowing it to cool. Diluted gel suspensions were first vitrified using Leica Vitribot. About 4μL of the sample was cast onto a plasma treated Cu 200 mesh Quantifoil™ grid and blotted against filter paper for 3s. The grid was then plunged into liquid ethane maintained at -185°C using liquid nitrogen to achieve sample vitrification. The grid was then transferred to a Cryo-storage container from where they were loaded into JEOL-1400 electron microscope via a Gatan single tilt Cryo holder. Images were recorded at an acceleration voltage of 120 kV at low dosage conditions.

For high-resolution images, vitrified samples were loaded into a JEOL JEM3200-FSC microscope, and images were recorded at an acceleration voltage of 300kV at low dosage conditions.

Sample pictures of the fibers are presented in **Figure S12**−**Figure S16**. Histograms of widths of bundles were calculated using the imegeJ.[17] Pictures were rotated in a way that they align with the y-axis. Then the profile was obtained by summing the grey values in rows from a rectangular area. The widths of the fibers were calculated as distances between consecutive minima. The process is presented in **Figure S17**.

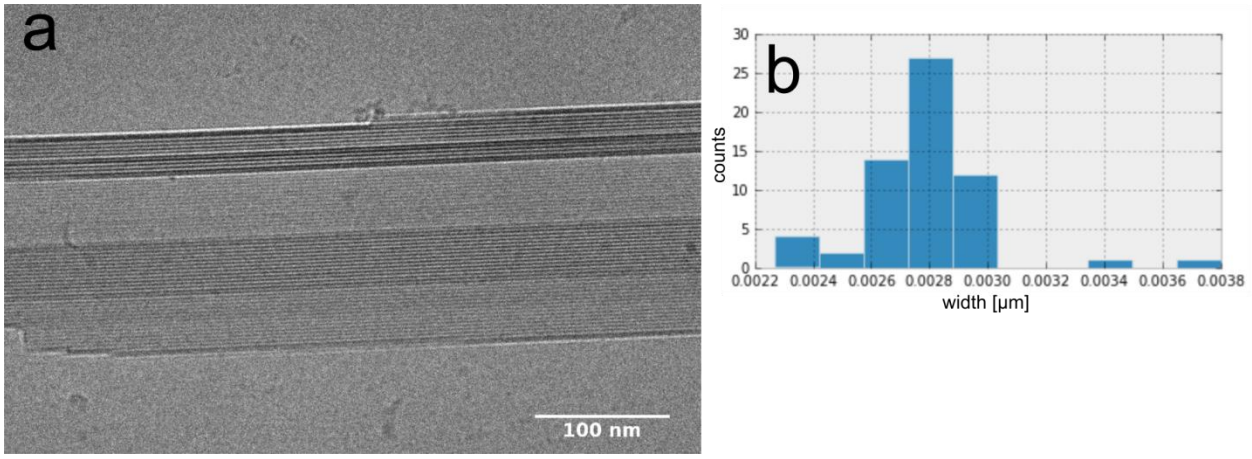

Figure S12. (a) Sample cryo-TEM picture of a bundle and (b) histogram of the widths.

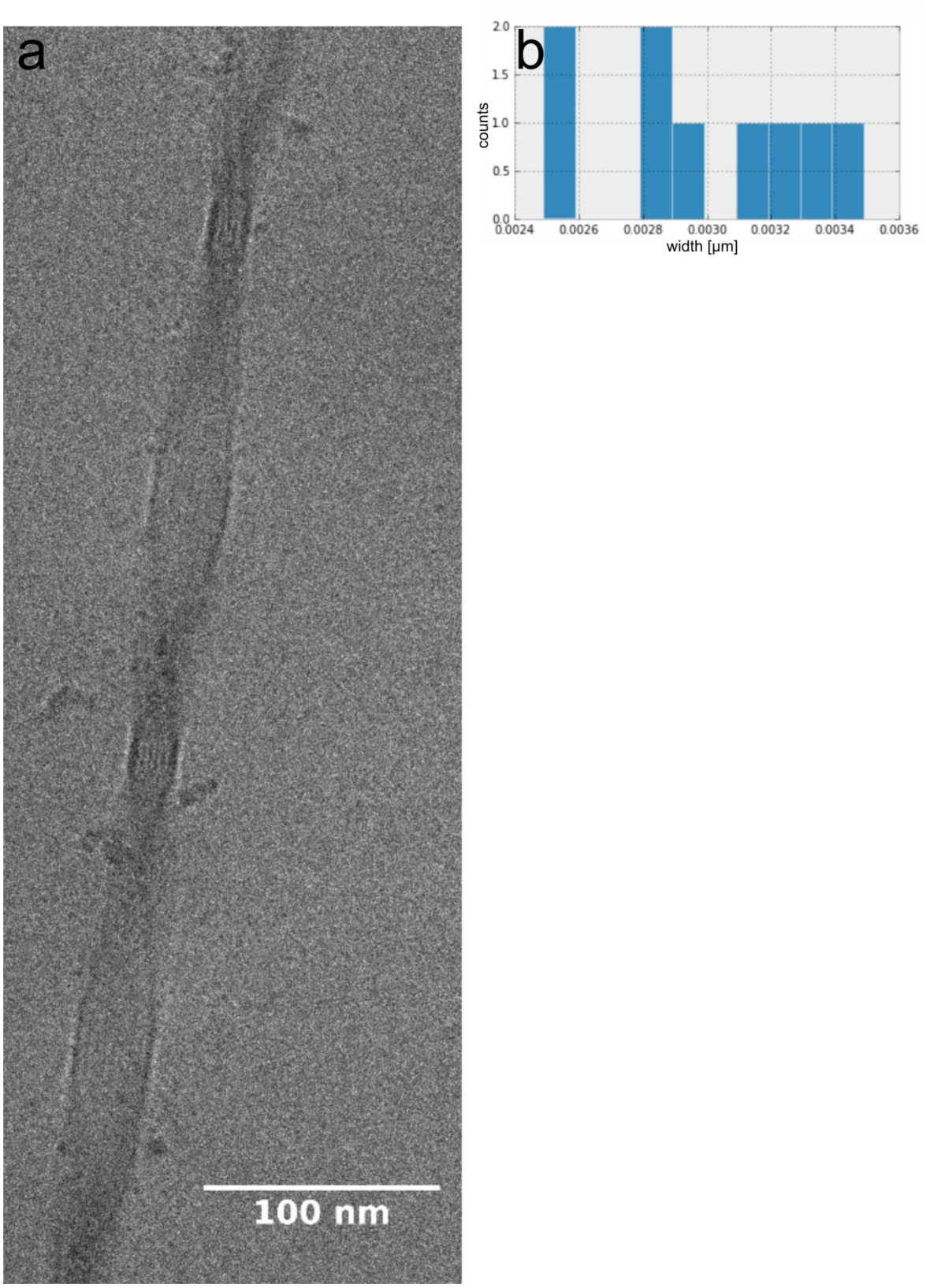

Figure S13. (a) Sample cryo-TEM picture of a bundle and (b) histogram of the widths.

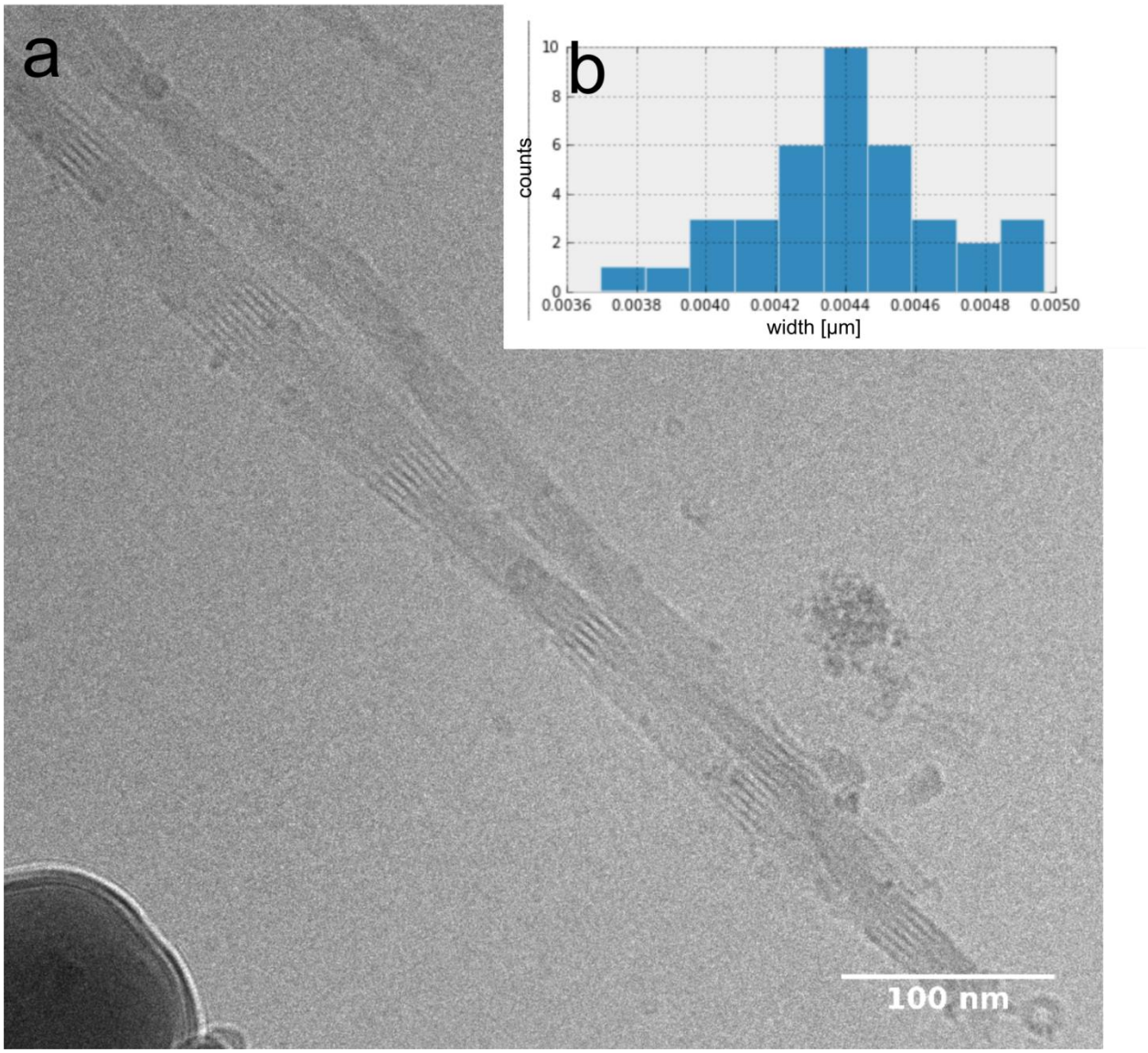

Figure S14. (a) Sample cryo-TEM picture of a bundle and (b) histogram of the widths.

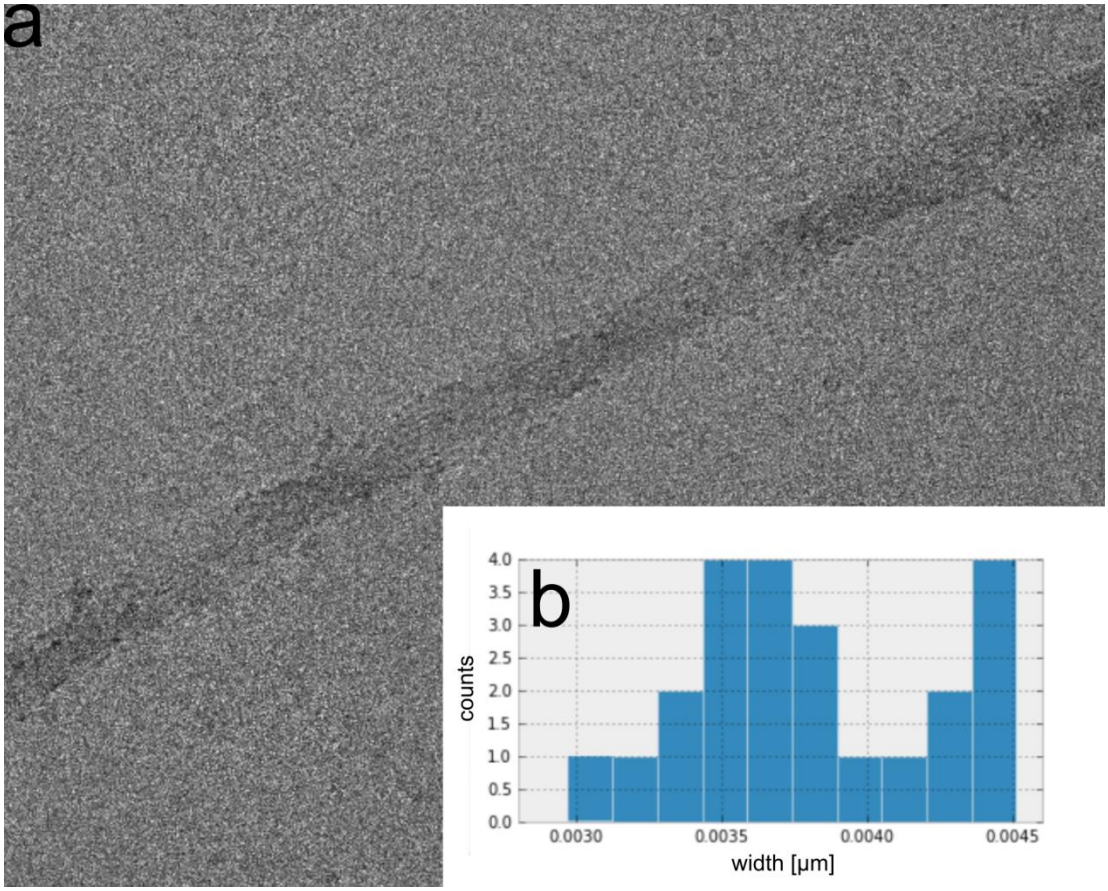

Figure S15. (a) Sample cryo-TEM picture of a bundle and (b) histogram of the widths.

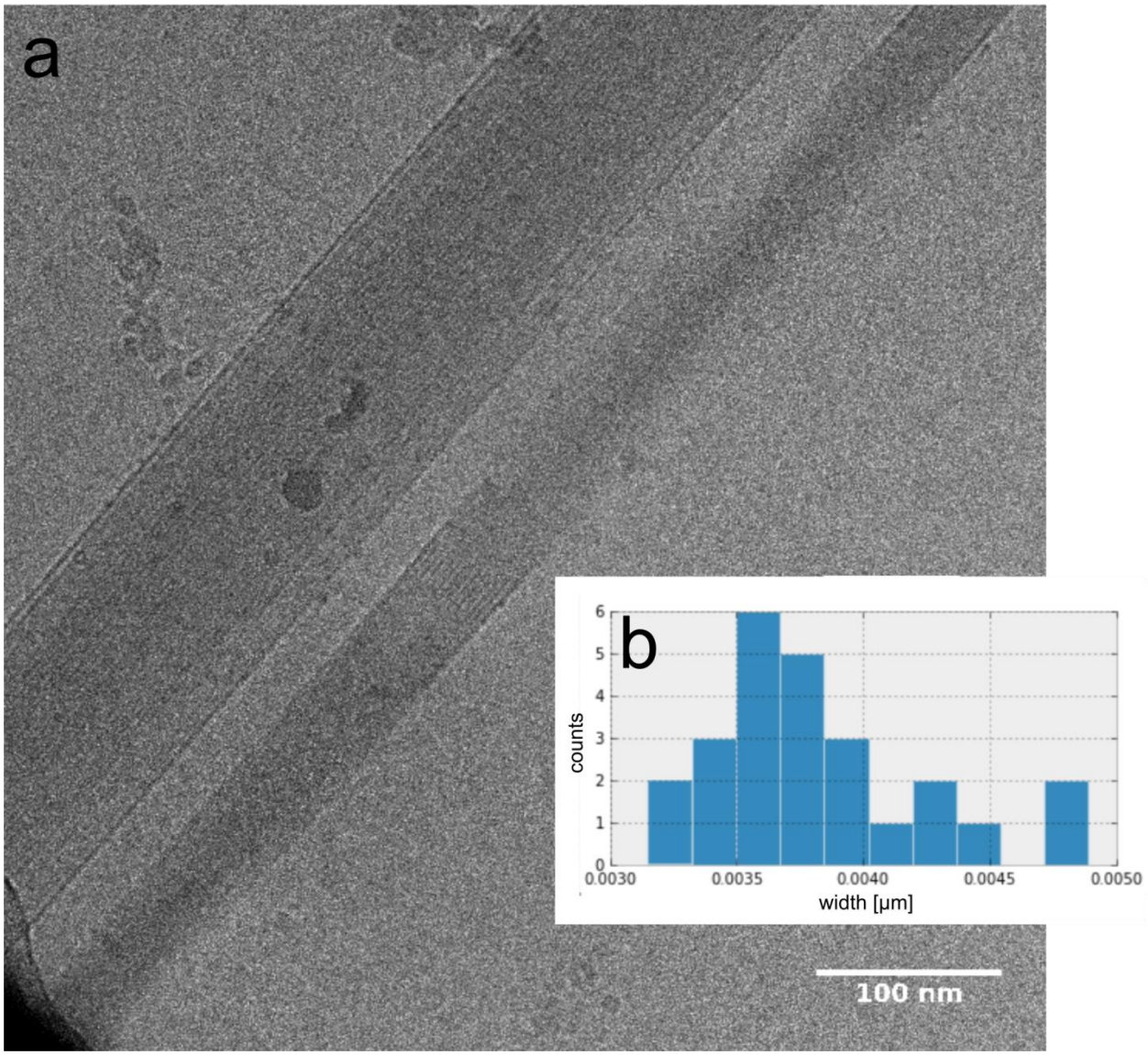

Figure S16. (a) Sample cryo-TEM picture of a bundle and (b) histogram of the widths.

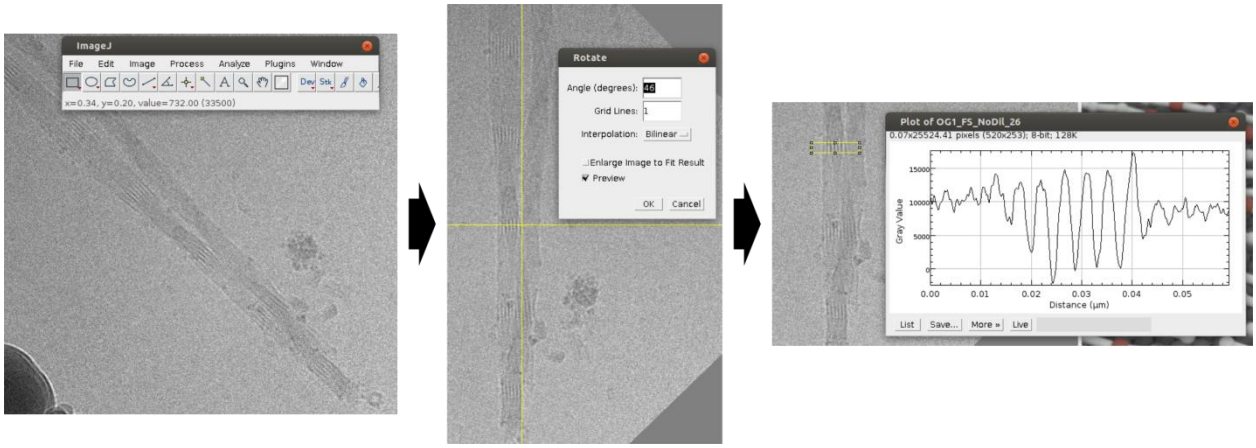

Figure S17. Image is processed using ImageJ. Image is rotated that the axis of the fibre is in y direction. Then the gray value is are summed in rows resulting in a profile of the fibre.

S8. Bundling

(i) Bundling of parallel and antiparallel fibers

To check the bundling of two fibers, we have run 20 independent simulation: 16 starting from parallel configuration of fibers and four from antiparallel. The results suggest no difference between bundling of parallel or antiparallel fiber (see individual traces of distance between cores of fibers and their final distribution in **Figure S18**).

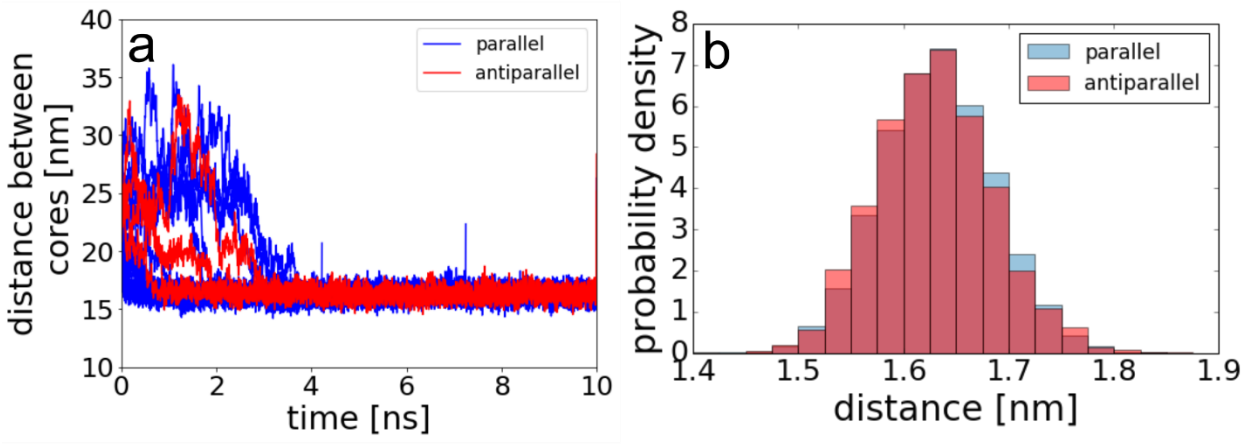

Figure S18. Distances of cores of two fibres oriented parallel or antiparallel at the beginning of the simulations) for 20 independent simulations. (a) Progress of the distances over time. (b) Final distribution of the distances (starting from 1ns).

(ii) Four fibers bundling

We have also performed simulations of four fibers bundling. We have run 31 simulations of 10 ns starting from fibers located in the plane of a square (**Figure S19a,b**). However, the size of the simulation box was too small to see bundling of just the four fibers, in the sense that many final structures crossed periodic boundary conditions and are therefore representing infinitely long bundles (see **Figure S19c**). Therefore the result here can be treated only semi-quantitatively. In all simulations, we observed strong affinity between fibers. Most importantly, we observed that in all simulations bundles create linear structures and not cyclic ones (which would have a rhombic structure). This could mean two things: bundles have a shape of a ribbon, or their structures consist of more than four fibers. If the fibers tend to form cyclic structures, it would mean that they create tubes with water molecules trapped in it.

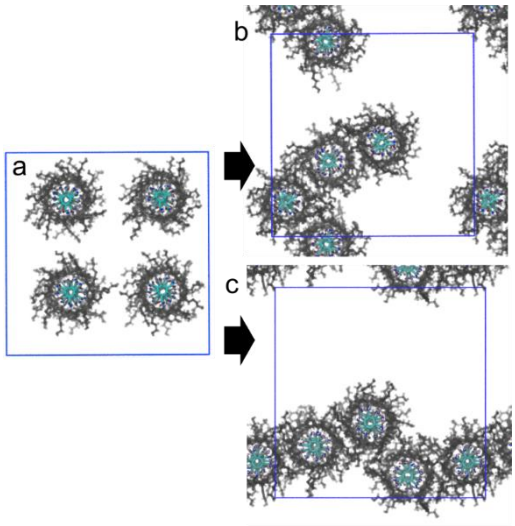

Figure S19. Bundling of four fibers. (a) Snapshot of starting configuration. (b) A final bundle that does not cross the PBC. (c) A final bundle which crosses PBC.

Captions to the videos

SI_video_nucleation.mov: The video shows the trajectory of the formation of fiber from randomly distributed molecules, of which snapshots are presented in Figure 2a in the main text. Snapshots are saved every 50 ps, and trajectory smoothing is applied for clarity. Cores (cyclohexane ring and amide groups) are indicated by cyan color, and the side branches are indicated by semi-transparent grey.

SI_video_secondary_nucleation.mov: The video shows the trajectory of the formation of fiber from randomly distributed molecules with a presence of existing fiber. Snapshots of this video are presented in Figure 4a in the main text. Snapshots are saved every 50 ps, and trajectory smoothing is applied for clarity. Cores (cyclohexane ring and amide groups) of free additional molecules are indicated by cyan color, cores of existing fiber by orange color, and all the side branches are indicated by semi-transparent grey.

References

1 D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen, *J. Comput. Chem.*, 2005, **26**, 1701–1718.

2 M. J. Abraham, T. Murtola, R. Schulz, S. Pall, J. C. Smith, B. Hess and E. Lindah, *SoftwareX*, 2015, **1**–**2**, 19–25.

3 J. A. Lemkul, B. Roux, D. Van Der Spoel and A. D. Mackerell, *J. Comput. Chem.*, 2015, **36**, 1473–1479.

4 T. K. Piskorz, A. H. De Vries and J. H. Van Esch, *J. Chem. Theory Comput.*, 2022, **18**, 431–440.

5 G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, 014101.

6 M. Parrinello and A. Rahman, *Phys. Rev. Lett.*, 1980, **45**, 1196–1199.

7 W. Yu, P. E. M. Lopes, B. Roux and A. D. MacKerell, *J. Chem. Phys.*, 2013, **138**, 034508.

8 R. Harada, T. Nakamura, Y. Takano and Y. Shigeta, *J. Comput. Chem.*, 2015, **36**, 97–102.

9 M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Perez-Hernandez, M. Hoffmann, N. Plattner, C. Wehmeyer, J. H. Prinz and F. Noe, *J. Chem. Theory Comput.*, 2015, **11**, 5525–5542.

10 E. Hong, K. Lee and W. Wenzel, *Int. J. Biol. Biomed. Eng.*, 2007, **1**, 50–52.

11 U. Sengupta, M. Carballo-Pacheco and B. Strodel, *J. Chem. Phys.*, 2019, **150**, 115101.

12 J. H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schtte and F. Noé, *J. Chem. Phys.*, 2011, **134**, 174105.

13 B. Trendelkamp-Schroer, H. Wu, F. Paul and F. Noé, *J. Chem. Phys.*, 2015, **143**, 174101.

14 F. Noé, H. Wu, J. H. Prinz and N. Plattner, *J. Chem. Phys.*, 2013, **139**, 184114.

15 F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich and T. R. Weikl, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 19011–19016.

16 J. Kroon, J. A. Kanters, J. G. C. M. van Duijneveldt-van De Rijdt, F. B. van Duijneveldt and J. A. Vliegenthart, *J. Mol. Struct.*, 1975, **24**, 109–129.

17 C. T. Rueden, J. Schindelin, M. C. Hiner, B. E. Dezonia, A. E. Walter, E. T. Arena and K. W. Eliceiri, *BMC Bioinformatics*, 2017, **18**, 1–26.