Self-Assembly Dynamics of Linear Virus-Like Particles: Theory and Experiment

Melle T.J.J.M. Punter,[†] Armando Hernandez-Garcia,[‡] Daniela J. Kraft,[*,¶]

Renko J. de Vries,[‡] and Paul van der Schoot[§]

[†]*FOM institute AMOLF, The Netherlands*

[‡]*Wageningen University and Research Centre, The Netherlands*

[¶]*Leiden University, The Netherlands*

[§]*Eindhoven University of Technology, The Netherlands*

[§]*Utrecht University, The Netherlands*

E-mail: kraft@physics.leidenuniv.nl

Abstract

We study experimentally and theoretically the self-assembly kinetics of linear virus-like particles (VLPs) consisting of double-stranded DNA and virus-like coat proteins. The polynucleotide acts as a self-assembly template for our proteins with engineered attractive protein-DNA and protein-protein interactions that imitate the physicochemical functionality of virus coat proteins. Inspired by our experimental observations, where we find VLPs to grow from one point onwards, our model presumes a nucleation step before subsequent sequential cooperative binding from one of the ends of the polynucleotide. By numerically solving the pertinent reaction rate equations, we investigate the assembly dynamics as a function of the ratio between the number of available binding sites and proteins in the solution, i.e., the stoichiometry of the molecular building blocks. Depending on the stoichiometry, we find monotonic or non-monotonic assembly kinetics. If the proteins in the solution vastly outnumber the binding sites on all the polynucleotides, the assembly kinetics are strictly monotonic and the assembled fraction increases steadily with time. However, if the concentration of proteins and binding sites is equal, we find an overshoot in the concentration of fully covered polynucleotides. We compare our model with length distributions of two types of VLP measured by AFM imaging and find satisfactory agreement, suggesting that a relatively simple model may be useful in describing the assembly kinetics of chemically complex systems. We furthermore re-evaluate data by Hernandez-Garcia et al.[1] to include the effect of a finite protein concentration previously ignored. By fitting our model to the experimental data, we are able to pinpoint the sum of the protein-protein and protein-DNA interaction free energy, the binding rate of a protein to the DNA, as well as the nucleation free energy associated with switching a protein from the solution to the bound conformation. The values we find for the virus-like particles are comparable to virus capsid binding energies of linear and spherical viruses.

Introduction

Virus particles are arguably amongst the most complex objects in condensed matter physics yet among the simplest in biology. Their ability to deliver their genetic material into susceptible cells has triggered interest in the development of virus-like particles (VLPs) with the same capability but without the potential health risks.[2–6] Such particles are envisioned to be useful in gene therapy for the delivery of therapeutic nucleic acids and drugs, and to target tumor cells, not least because of their biocompatibility and biodegradability.[5,7,8]

Generally, viruses self-assemble in solutions containing polynucleotides and coat proteins, implying that the former become protected by a protein mantle usually of spherical or linear shape. Typically, this self-assembly involves the formation of a supramolecular complex through the cooperative binding of coat proteins to the polynucleotides, a process that has been difficult to mimic using synthetic biology[9–11]. Recently, however, Hernandez-Garcia et al.[1] found effective design principles for synthetic coat proteins that give rise to the self-assembly of linear VLPs. These designer coat proteins are composed of three simple and distinctive polypeptide blocks that independently provide a basic functionality required to form linear VLPs, see Figure 1a. The first block, C, is a hydrophilic and long random coil sequence (approx. 400 amino acids) that provides remarkable colloidal stability to the formed VLPs. The middle block S_n consists of n repeats of the self-assembly silk-like octapeptide GAGAGAGQ.[12] This silk-like block has been demonstrated to provide cooperative encapsulation of DNA with length repetition of $n \geq 10$.[1] The third block, B, is a stretch of twelve-lysines that mediates the electrostatic interaction with the phosphate backbone of the DNA. In this study we use two protein constructs with different self-assembly block length: C-S_{10}-B and C-S_{14}-B. Using AFM imaging Hernandez-Garcia et al. determined the self-assembly kinetics of the designer coat proteins on DNA by measuring at different times the length of the self-assembled protein part around a 2.5kbp linear dsDNA chain, see Figure 1b. These self-assembled protein-DNA lengths were fitted successfully to a model originally put forward by Kraft et al.[13] to describe the self-assembly of tobacco mosaic virus (TMV).

This model presumes a vast excess of proteins, while in the actual experiments the protein concentration was a limiting factor. It is unclear to what extent the parameters obtained through this fit are robust to accounting for the correct stoichiometry.

To address this issue we revisit the so-called nucleated Zipper model and focus on the influence of a finite protein concentration on the kinetics of the assembly by a numerical evaluation of the pertinent reaction rate equations. We show that the concentration of fully assembled particles, the fraction of occupied binding sites on the DNA and the concentration of uncovered polynucleotides can vary non-monotonically as a function of time. This contrasts with what we find when there is a vast excess of proteins in solution, where these quantities always evolve monotonically with time. From our calculations we can identify the predominant kinetics for different relative concentrations of protein and DNA. We are able to quantify the temporal evolution of the concentration of fully assembled particles during assembly and compare it to the concentration in equilibrium. Furthermore, we conduct additional experiments with two synthetic proteins of different silk-domain repeat lengths and approximately equal protein concentrations, and fit our improved model to the assembly data. In the article by Hernandez-Garcia et al. it was hypothesized that the length of the silk-domain positively correlates with the cooperativity of the binding process. We here investigate this hypothesis by studying the assembly of proteins comprised of ten and fourteen silk strands and re-evaluate the earlier experiments of Hernandez-Garcia et al.[1]

From our curve fits we are able to extract the sum of the protein-protein and protein-DNA interaction free energies, the rate constant of binding a protein to the DNA, as well as the nucleation free energy associated with switching the protein from the solution to the bound conformation. Agreement between theory and experiment is satisfactory, confirming that the assembly is not extremely sensitive to the relative protein and DNA concentrations. This suggests that a relatively simple model may be useful in describing the assembly kinetics of chemically complex systems.

The remainder of this paper is organised as follows: first, we summarise the equilibrium

properties of the Zipper model of Kraft et al.,[13] because this will be helpful in understanding the non-monotonic assembly kinetics. Next, we derive the reaction rate equations for arbitrary relative protein concentrations and show that the concentration of fully assembled virus-like particles, the fraction of occupied binding sites on the DNA molecules and the concentration of free polynucleotides can change non-monotonically as a function of time. We present a diagram of the predominant kinetic regime as a function of the overall and relative protein concentration. Finally, we present the curve fits to our experimental data and the previously published data of Hernandez-Garcia et al.[1]

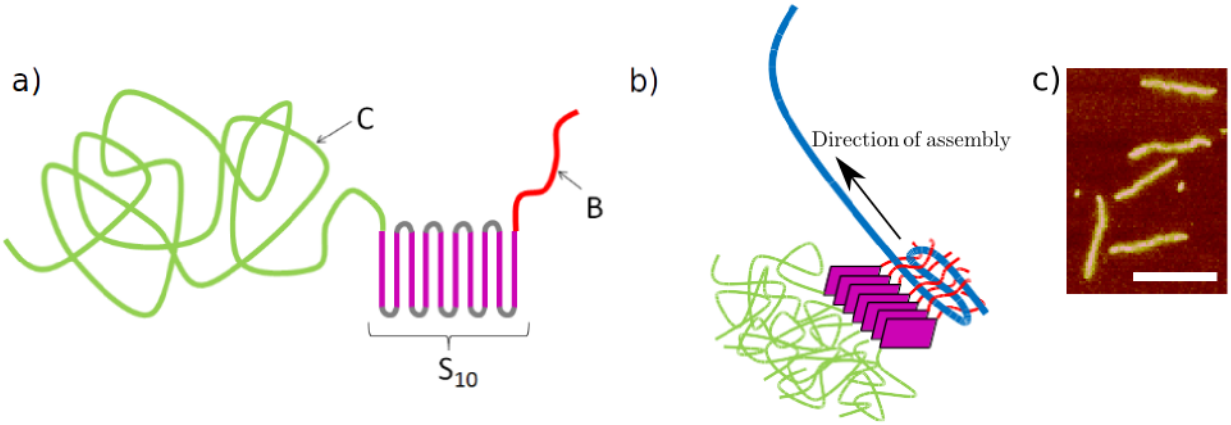

Figure 1: **a**) Schematic representation of a C-S_{10}-B protein consisting of three blocks: a hydrophilic random coil sequence C that prevents the different protein-DNA complexes in the solution from aggregating, a silk-like sequence S_n consisting of a variable number of silk strands $S_n=(GAGAGAGQ)_n$ that putatively dictates the cooperativity of the protein binding to the DNA, and an oligolysine binding block $B=K_{12}$ that binds non-sequence specifically to the dsDNA through electrostatic interactions. **b**) If the protein binds to the DNA it causes the DNA to be compacted by a factor of about three presumably to realise charge neutralization of the DNA by the binding blocks B. The arrow indicates the direction of protein addition. **c**) An AFM image of observed rod-like protein-DNA complexes which are fully covered with protein, the scale bar is 300 nm.

Equilibrium concentrations

In this section we introduce the model used to describe the equilibrium concentrations, i.e., the concentrations in the long time limit, of the different supramolecular molecules

arising in a solution of polynucleotides and coat proteins. In particular, we focus on the concentration of fully assembled particles, the concentration of naked polynucleotides and the fraction of occupied binding sites. This will provide a reference frame for the next section in which we show that these quantities evolve non-monotonically with time for finite protein concentrations.

We consider a dilute solution of polynucleotides and coat proteins. The polynucleotides are linear molecules that have q binding sites for proteins and become encapsulated through this binding process. In the case of TMV, the polynucleotide is a single-stranded RNA molecule[13] and in the experiments we conducted here and in earlier work[1] double-stranded DNA was employed. We describe the fraction of polynucleotides with n bound proteins in thermodynamic equilibrium by $P_{eq}(n)$, where $n = 0, 1, 2, ..., q$. We denote the scaled concentration of free proteins in solution, i.e., those not bound to a polynucleotide, as $s_{eq} = \rho_{P,eq}/\phi_c$, with $\rho_{P,eq}$ the dimensionless free protein concentration and ϕ_c the dimensionless critical concentration, all given as mole fractions. The critical concentration, $\phi_c = e^{\epsilon+g}$, is the concentration below which no significant binding of proteins occurs, with $\epsilon < 0$ the attractive interaction free energy between two proteins cooperatively bound to the polynucleotide and $g < 0$ the attractive interaction free energy between a bound protein and the polynucleotide. Both energies are in units of the thermal energy $\mathrm{k_B T}$.

The free energy contributions capture the effects of what in reality involves microscopically complex processes, which we we do not need to consider in all their detail. For a discussion of this approach in the context of the assembly of icosahedral viruses, we refer to, e.g., Muthukumar et al.[14] and Hagan.[15] The attractive interaction free energy between a bound protein and the polynucleotide, g, for instance, accounts not only for the free energy gain due to the electrostatic interaction between the positively charged proteins and the negatively charged DNA, but also for the entropy cost of any steric interactions of the protein with the DNA and the elastic free energetic cost of bending the DNA in the compact compound structure.

Furthermore, because the solution is dilute we can link the concentration of free proteins to the (dimensionless) chemical potential $\mu_P < 0$ of the free proteins in solution by $\rho_{P,eq} = e^{\mu_P}$. If the entropy cost of removing a protein from solution is lower than the energy gain of binding that protein to the polynucleotide, i.e., if $\epsilon + g < \mu_P$, binding occurs. In this case, the protein density $\rho_{P,eq}$ exceeds the critical protein density ϕ_c and the relative density of free proteins is larger than unity. Therefore, s_{eq} is a measure for the probability of binding a protein cooperatively to the polynucleotide.

The nucleated Zipper model of Kraft et al.[13] considers consecutive binding of proteins to the polynucleotide from one end onward, as if a zipper is pulled up. Originally, the theory was devised specifically to model the *in vitro* assembly of TMV. The RNA of TMV has a sequence of nucleotides near one of its ends that has been identified as an origin of assembly (OAS). Self-assembly of TMV starts by the binding of proteins at this OAS and the binding of subsequent proteins happens by zipping toward the end of the RNA.[13] Notwithstanding that the DNA that we use in our experiments does not have an OAS, we can still apply the nucleated Zipper model because we find nucleation to commence (almost) always at one of the ends of a DNA molecule. A plausible cause of this is that the end of the DNA may act as an *effective* OAS for the engineered proteins due to steric effects.

Indeed, as already discussed, our triblock proteins consist of an oligolysine sequence, a silk-like sequence and a long collagen-like sequence that in solution has a random-coil structure. In our experiments, protein binding is driven by aspecific Coulomb interactions between the positively charged oligolysine block and the negatively charged DNA, and in principle the first protein to bind would not have a preference for any specific position along the backbone of the DNA. However, due to the random coil structure of the collagen-like sequence, binding at one of the ends must be more favourable than anywhere else due to a larger amount of free space available to the random coil block to explore its configuration space. Plausibly, the DNA molecule excludes a larger volume to the random coil of a protein in the central portion than near the ends, which in our view explains the preference for the

ends.

From statistical mechanics, the fraction of polynucleotides with n bound proteins can be derived within the nucleated Zipper model as

$$P_{eq}(n) = \begin{cases} \frac{1}{\Xi} & \text{if } n = 0, \\ \frac{\sigma s_{eq}^n}{\Xi} & \text{if } 0 < n \leq q, \end{cases} \tag{1}$$

where Ξ is a normalization constant and $\sigma = e^{-h+\epsilon}$ a measure for the cooperativity of the binding of the proteins to the polynucleotide. The quantity $h > 0$ is the free energy cost of conformational switching of the first protein upon binding to the template. The assumption of allostery implies that subsequently bound proteins do not require this conformational free energy cost.[16] In essence, the nucleated Zipper model is a reduction of the one-dimensional Ising model in the limit of infinite interaction strength and therefore fundamentally different from, for example, the Langmuir adsorption model where all binding sites are independent. In essence, the nucleated Zipper model is a reduction of the one-dimensional Ising model in the limit of infinite interaction strength.[17] It is therefore fundamentally different from, for example, the Langmuir adsorption model where all binding sites are independent,[18] or Ising model-based approaches that allow correlated adsorption[19–21]. We furthermore note that other works on closed assembly onto spherical templates[22] are fundamentally different from the linear case discussed here due to the more complex assembly pathways and the higher dimensionality of the structures formed.

The energy penalty for nucleation is captured by the cooperativity parameter σ: a small value of sigma reflects a large nucleation energy barrier. Because all partially coated polynucleotide states rely on nucleation, the probability $P_{eq}(n)$ for a polynucleotide having bound at least one protein, $n > 0$, scales linearly with σ. Furthermore, the probability $P_{eq}(n)$ strongly depends on the scaled concentration of free proteins, and, by mass action, is proportional to s_{eq}^n. Therefore, for $s_{eq} > 1$, complete coverage of the polynucleotides is most likely. To

quantify the degree of coverage, we calculate the average number of bound proteins per polynucleotide as

$$\langle\theta\rangle_{eq} = \frac{1}{q}\sum_{n=0}^{q} nP_{eq}(n). \tag{2}$$

The probability distribution function, and hence all of its moments including the mean occupation number, are a function of the overall concentration of proteins $\phi_P \equiv N_P/N_t$ and of the concentration of polynucleotides (templates) $\rho_T \equiv N_T/N_t$, with $N_t = N_P + N_T + N_s$ the sum of the number of proteins, polynucleotides and solvent molecules. In dilute solution, $N_t \approx N_s$. Mass conservation requires that the scaled concentration of free proteins s_{eq} obeys the following equality

$$s_{eq} = S(1 - \lambda\langle\theta\rangle_{eq}), \tag{3}$$

where $S \equiv \phi_P/\phi_c$ is the overall protein concentration scaled to the critical assembly concentration and hence S is a measure for the supersaturation, and $\lambda \equiv q\rho_T/\phi_P$ is the ratio of the number of available binding sites on all polynucleotides $q\rho_T$ in the solution and the number of coat proteins ϕ_P, i.e., the stoichiometric ratio of the solution.

It is immediately clear from equations 1 and 2 that the mean occupancy of the binding sites, $\langle\theta\rangle_{eq} = \langle\theta\rangle_{eq}(s_{eq}, \sigma, q)$, must be a function of the quantities s_{eq}, σ and q. This implies that equation (3) needs to be solved self-consistently for the scaled concentration of free proteins as a function of the experimentally controllable parameters, $s_{eq} = s_{eq}(S, \lambda, \sigma, q)$. The values of the cooperativity parameter σ and the number of binding sites per template q can, at least in principle, be controlled through the design of the polynucleotide and the capsid protein. The stoichiometric ratio λ is determined by the concentration of polynucleotides and proteins, and the mass action variable S is set by the concentration and exact design of the coat proteins through the critical aggregation concentration ϕ_c.

Given the above considerations, we present in Figure 2 the mean fraction of bound sites $\langle\theta\rangle_{eq}$, the fraction of naked polynucleotides $P_{eq}(0)$, and the fraction of completely encapsulated polynucleotides, $P_{eq}(q)$, as a function of the scaled protein concentration S for a

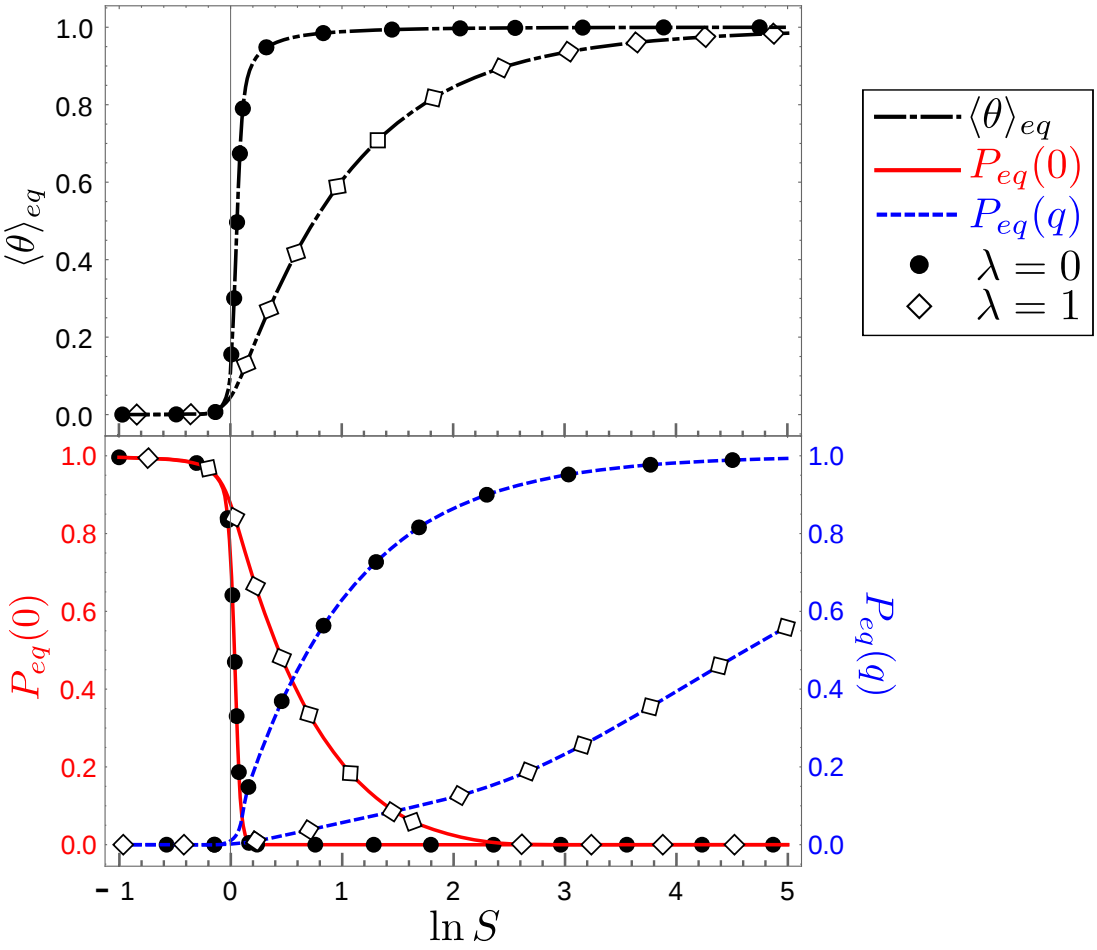

Figure 2: Influence of the total protein concentration on the equilibrium state. Top: the fraction of occupied binding sites $\langle\theta\rangle_{eq}$ on the polynucleotide molecules. Bottom: the fraction of naked polynucleotides $P_{eq}(0)$, and the fraction of fully covered polynucleotides $P_{eq}(q)$. All are given as a function of the ratio of the total protein concentration to the critical concentration S, for a stoichiometric ratio λ, that is, the ratio between the number of available binding sites and proteins in the solution, of zero and unity.

cooperativity of $\sigma = e^{-5}$, and polynucleotides consisting of $q = 51$ binding sites. We compare results for two different values of the stoichiometry, $\lambda = 0$ and $\lambda = 1$. We will arbitrarily choose the values $\sigma = e^{-5}$ and $q = 51$ throughout this section and the next. We have verified that other values do not give qualitatively different results.

Figure 2 shows that for $S < 1$ practically all polynucleotides are naked, implying that the assembly is indeed highly cooperative on account of the small value of the cooperativity parameter σ. If the concentration of proteins exceeds the critical concentration and $S > 1$, the polynucleotides become encapsulated to a larger degree. Interestingly, the effective level of cooperativity, measured, e.g., by how strongly the fraction of occupied sites or the number of fully covered templates increases with increasing concentration, not only depends on the value of σ but also on the stoichiometry of the mixture λ. The larger λ, the fewer proteins are present in solution and the more difficult it is to encapsulate the templates. The effective cooperativity of binding even at a seemingly ideal stoichiometry of $\lambda = 1$ is much smaller than when protein is present in vast excess of the number of binding sites. Full coverage, i.e., $P_{eq}(q) = 1$ is only obtained at much higher protein concentration S and the protein concentration required to reach saturation, $\langle\theta\rangle_{eq} \approx 1$, differs by an order of magnitude. More detailed information on the influence of the stoichiometric ratio λ on the fraction of occupied binding sites $\langle\theta\rangle_{eq}$ on the polynucleotide molecules and the fraction of naked polynucleotides $P_{eq}(0)$ is shown in Figure 3 for a high, $S = e^2 \approx 7.4$, and a low, $S = e^{0.4} \approx 1.5$, value of the scaled protein concentration.

For values of λ between zero and unity, the fraction of bound sites $\langle\theta\rangle_{eq}$ remains largely unaffected, especially in the case of high protein concentration. If $\lambda > 1$, that is, if the number of proteins is smaller than the number of binding sites, mass conservation dictates that $\langle\theta\rangle_{eq} < 1/\lambda$, see Eq. (3). In Figure 3 this upper bound for the fraction of bound sites $1/\lambda$ is indicated by the thin black dotted line. The fraction of complete VLPs, $P_{eq}(q)$, starts decreasing at much lower values of λ than the fraction of occupied binding sites $\langle\theta\rangle_{eq}$, which means that fully formed VLPs are more strongly affected by the precise stoichiometric ratio

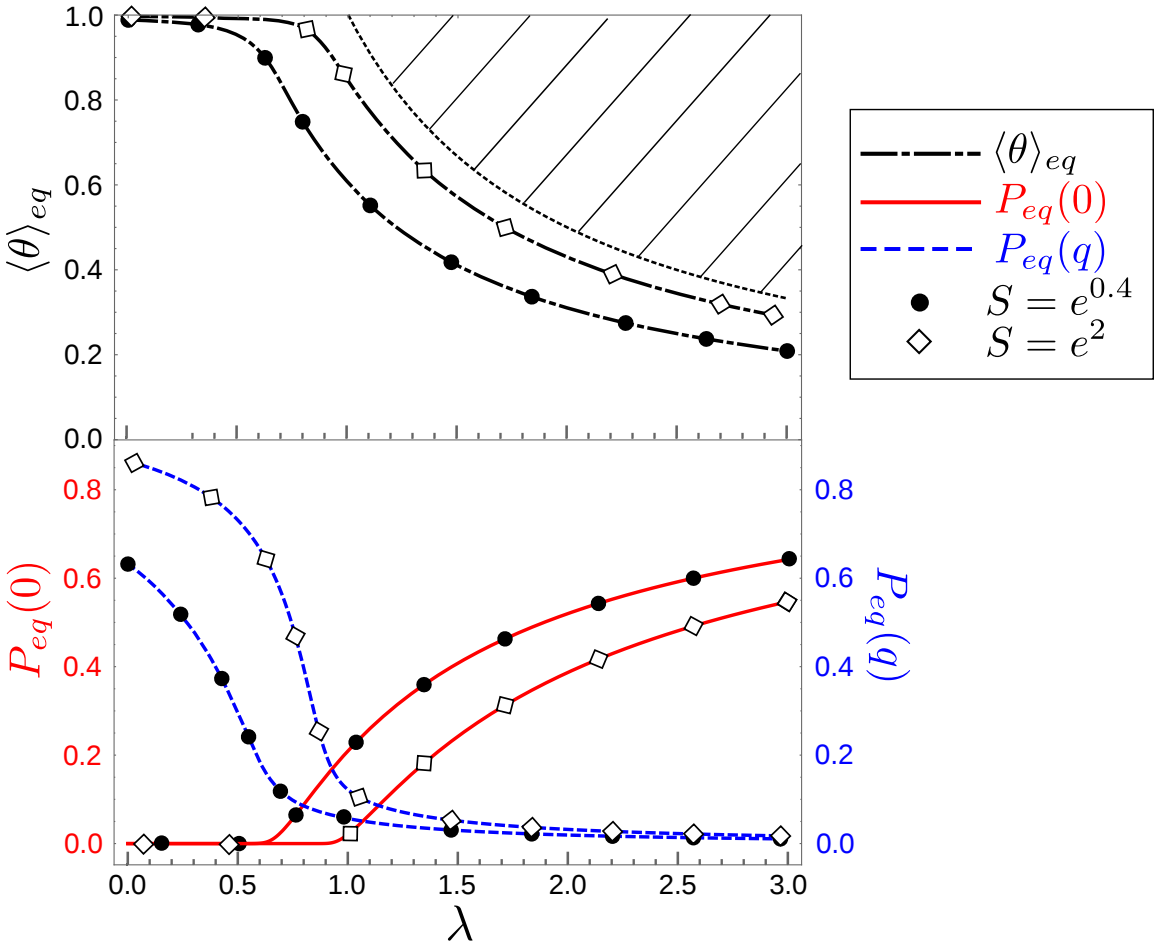

Figure 3: Influence of stoichiometry λ on the equilibrium state. Top: the fraction of occupied binding sites $\langle\theta\rangle_{eq}$ on the polynucleotide molecules. Bottom: the fraction of naked polynucleotides $P_{eq}(0)$, and the fraction of fully covered polynucleotides $P_{eq}(q)$. They are given as a function of the ratio of the number of available binding sites and proteins, i.e., the stoichiometric ratio λ, for a low ratio of the overall protein concentration and the critical concentration, $S = e^{0.4} \approx 1.5$, and a high ratio, $S = e^{2} \approx 7.4$.

than the fraction of occupied binding sites.

For a stoichiometric ratio of unity, the sum of the fraction of fully covered particles $P_{eq}(q)$ and the fraction of naked polynucleotides $P_{eq}(0)$ is much smaller than $\langle\theta\rangle_{eq}$. This implies that complexes are dominated by incomplete intermediates, that is, a large fraction of polynucleotides is neither completely encapsulated nor naked. From the perspective of the virus or virus-like particles this is an undesirable state, as it leaves the polynucleotide exposed to attack by enzymes including nucleases.

Finally, a lower cooperativity σ implies a larger value of the scaled protein concentration S at which $\langle\theta\rangle_{eq}$ starts to deviate appreciably from zero. Furthermore, we find that while a larger value of the number of binding sites per polynucleotide q results, for example, in a sharper transition from naked polynucleotides ($\langle\theta\rangle_{eq} = 0$) to fully encapsulated polynucleotides ($\langle\theta\rangle_{eq} = 1$) at a stoichiometry of zero, generally speaking σ has no influence on the *qualitative* dependence of $\langle\theta\rangle_{eq}$, $P_{eq}(q)$ and $P_{eq}(0)$ on the stoichiometry and the scaled protein concentration.

With these equilibrium predictions in the back of our mind, we consider in the next section the fraction of occupied binding sites $\langle\theta\rangle_{eq}$, the fraction of naked nucleotides $P_{eq}(0)$, and the fraction of fully formed VLPs $P_{eq}(q)$ as a function of time, extending the earlier work of Kraft et al.[13] towards arbitrary stoichiometries $\lambda \geq 0$. As we shall see, our insight in the equilibrium properties of the model outlined in this section will be helpful in explaining the non-monotonic time dependence of, e.g., the fraction occupied binding sites and the fraction completely encapsulated templates, that we find under certain combinations of template and protein concentrations.

Assembly dynamics

To describe the temporal evolution of the concentrations of all the supramolecular species in a solution containing proteins and polynucleotides, that is, polynucleotides with $n =$

$0, 1, 2, ..., q$ proteins bound, we view the binding of the proteins to the templates as a set of chemical reactions and write down the reaction rate equations. We consider the following set of reactions to occur

$$\mathrm{P} + \mathrm{NUC_{m-1}} \quad \underset{\mathrm{k_-(m)}}{\overset{\mathrm{k_+(m)}}{\rightleftharpoons}} \quad \mathrm{NUC_m}, \text{ with } m = 1, 2, ..., q, \tag{4}$$

where P denotes a capsid protein and $\mathrm{NUC_m}$ a polynucleotide to which m proteins are bound. We neglect the binding of protein oligomers to the polynucleotide and any non-cooperative binding of proteins to the polynucleotide. To simplify the kinetic equations we make similar assumptions as Kraft et al.[13]: we presume the binding rate of proteins to the polynucleotide to be equal for every reaction, such that $k_+(m) = k_+$. Furthermore, by imposing that the system approaches thermodynamic equilibrium in the long-time limit, $t \to \infty$, we calculate the equilibrium constants as

$$K_m \equiv \frac{k_+}{k_-(m)} = \begin{cases} \frac{\sigma S}{\phi_P} = e^{-h-g}, & \text{if } m = 1, \\ \frac{S}{\phi_P} = e^{-\epsilon-g}, & \text{if } 1 < m \leq q, \end{cases} . \tag{5}$$

with, as previously defined, $S = \phi_P/\phi_c$ the scaled protein concentration, ϕ_P the total protein concentration, $\phi_c = e^{\epsilon+g}$ the critical concentration and $\sigma = e^{-h+\epsilon}$ the cooperativity parameter. Here, again, the dimensionless free energy $h > 0$ accounts for the conformational switching of the first protein bound, $\epsilon < 0$ is the protein-protein interaction free energy, and $g < 0$ the protein-nucleotide interaction free energy. By defining the scaled binding rate of proteins to the DNA $k'_+ \equiv k_+\phi_P$, we can express all backward rate constants in terms of the quantities k'_+, σ and S and the equations in dimensionless time $\tau = k'_+ t$. Finally, we define the relative fraction of polynucleotides having n proteins bound as $f(n, \tau) \equiv P(n, \tau)/P_{eq}(n)$, with $P(n, \tau)$ the fraction of polynucleotides with n proteins bound at time τ and $P_{eq}(n)$ its equilibrium value calculated in the previous section.

We obtain the following set of kinetic equations

$$\begin{aligned}
\frac{\partial f(0,\tau)}{\partial \tau} &= -\frac{s_{eq}}{S}\Big(y(\tau)f(0,\tau) - f(1,\tau)\Big), && (6)\\
\frac{\partial f(1,\tau)}{\partial \tau} &= -\frac{s_{eq}}{S}\Big(y(\tau)f(1,\tau) - f(2,\tau)\Big) + \\
&\quad \frac{1}{\sigma S}\Big(y(\tau)f(0,\tau) - f(1,\tau)\Big), && (7)\\
\frac{\partial f(n,\tau)}{\partial \tau} &= -\frac{s_{eq}}{S}\Big(y(\tau)f(n,\tau) - f(n+1,\tau)\Big) + \\
&\quad \frac{1}{S}\Big(y(\tau)f(n-1,\tau) - f(n,\tau)\Big), && (8)\\
\frac{\partial f(q,\tau)}{\partial \tau} &= \frac{1}{S}\Big(y(\tau)f(q-1,\tau) - f(q,\tau)\Big), && (9)\\
\frac{dy(\tau)}{d\tau} &= -\frac{\lambda}{q\Xi}\Big[\Big(y(\tau)f(0,\tau) - f(1,\tau)\Big) + \\
&\quad \sigma\sum_{n=1}^{q-1} s_{eq}^{n}\Big(y(\tau)f(n,\tau) - f(n+1,\tau)\Big)\Big], && (10)
\end{aligned}$$

where $2 \leq n \leq q-1$, $y(\tau) = s(\tau)/s_{eq}$, $s(\tau) = \rho_P(\tau)/\phi_c$ is the scaled concentration of free proteins at time τ, s_{eq} is the scaled concentration of free proteins in equilibrium, calculated from Eq. (3), and Ξ is again the normalisation from Eq. (1). In the long-time limit, if $\tau \to \infty$, we have $y(\tau) \to 1$ and $f(n,\tau) \to 1$, such that all equilibrium equations are satisfied and all concentrations are constant. We cannot solve this set of non-linear coupled differential equations analytically albeit that in some limits analytical approximations can be made. We will not dwell on this here. Therefore, we solve the equations numerically and refer to the Supporting Information for more details on the numerical methods underlying the results presented in this section. We assume $P(0,0) = 1$ throughout this paper, implying that only naked polynucleotides are brought into the solution of coat proteins at time $\tau = 0$. Because $\sum_{n=0}^{q} P(n,\tau) = 1$, this defines all initial conditions.

Next we discuss in more detail how the fraction of occupied binding sites $\langle\theta\rangle(\tau)$, the fraction of naked polynucleotides $P(0,\tau)$ and the fraction of fully covered particles $P(q,\tau)$ evolve as a function of time for typical values of the ratio of the overall protein concentration to the critical concentration S, and the ratio of the number of binding sites and proteins

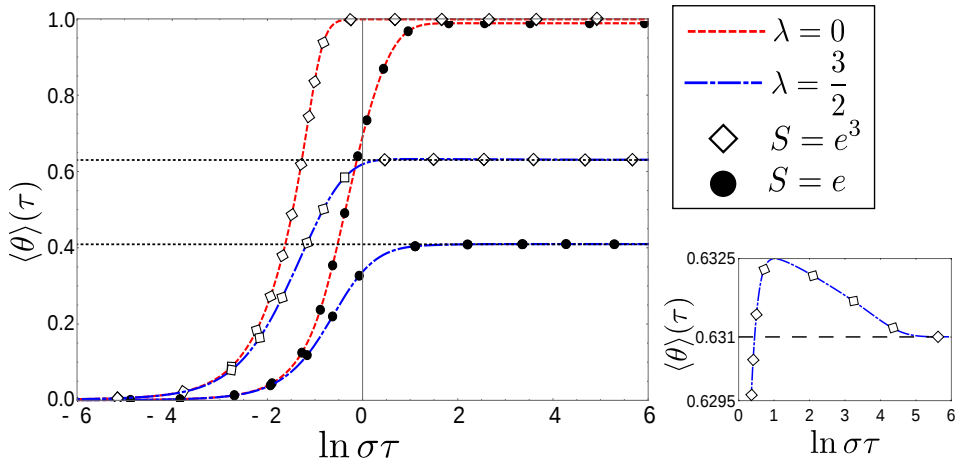

Figure 4: The mean fraction of occupied binding sites $\langle\theta\rangle(\tau)$ as a function of dimensionless time τ scaled to the cooperativity of binding the first protein onto the polynucleotide σ, the ratio of the number of available binding sites and proteins λ, and the ratio of the overall protein concentration to the critical concentration S. Results are shown for a vast excess of proteins, $\lambda = 0$, a shortage of proteins, $\lambda = 3/2$, a low protein concentration, $S = e \approx 1.7$, and a high protein concentration, $S = e^3 \approx 20$. Inset: for $\lambda = 3/2$ and $S = e^3$, the fraction of bound sites $\langle\theta\rangle(\tau)$ exhibits non-monoticity in the form of a (small) overshoot.

λ, i.e., the stoichiometric ratio of the solution. We monitor under what conditions the non-monotonic growth of these fractions occurs, that is, for what values of S and λ what kind of non-monotonic time evolution presents itself.

Figure 4 shows the mean fraction of occupied binding sites $\langle\theta\rangle(\tau)$ as a fraction of dimensionless time τ scaled to the cooperativity $\sigma = e^{-5} \approx 0.007$, for a number of binding sites per polynucleotide of $q = 51$. We scale the dimensionless time τ to the cooperativity σ because for small values of σ the nucleation reaction is the rate limiting step, as can be seen from Eq. (5). Thus, we expect the typical time to reach saturation proportional to $1/\sigma$. The time evolution of $\langle\theta\rangle(\tau)$ is given for a stoichiometry λ of zero and 3/2, i.e., for a vast excess and shortage of proteins, as well as for a low, $S = e \approx 1.7$, and for a high, $S = e^3 \approx 20$, scaled protein concentration S.

The increase of the average occupation $\langle\theta\rangle(\tau)$ is independent of λ for $\ln \sigma\tau \leq -3$. The long-time value of $\langle\theta\rangle(\tau)$, however, does depend on λ, as we can expect from the equilibrium results of Figure 3. Therefore, the stoichiometry λ only influences the mean fraction of

occupied binding sites after some initial period. This is not surprising, of course, because for small times the solutions is not yet depleted of proteins binding to the template. The lag time we find to decrease with increasing S. This is to be expected, for higher protein concentration induce a stronger thermodynamic driving force for assembly.

Finally, as shown in the inset, we observe for large supersaturation $S = e^3$ and a shortage of proteins expressed in a stoichiometric ratio of $\lambda = 3/2$ the first example of non-monoticity: a tiny - only 0.15% - but long-lived overshoot of the fraction of occupied sites $\langle\theta\rangle(\tau)$.

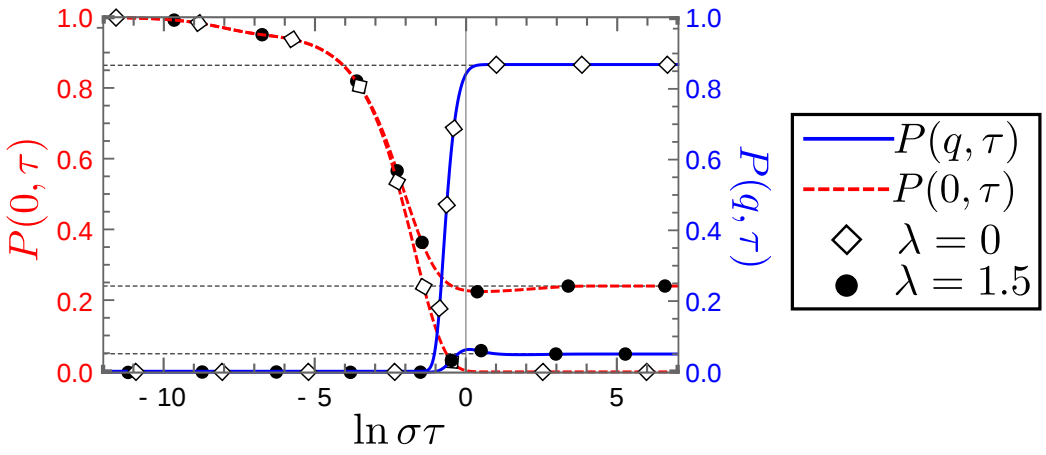

Figure 5: The fraction of naked polynucleotides $P(0,\tau)$ and the fraction of fully formed VLPs $P(q,\tau)$ as a function of dimensionless time τ scaled to the cooperativity σ for a scaled protein concentration $S = e^2 \approx 7.4$, and a ratio of the number of available binding sites and proteins λ of zero and 1.5, representing an excess and a shortage of proteins. For short times $P(q,\tau)$ and $P(0,\tau)$ do not depend appreciably on λ, while in equilibrium ($\ln \sigma\tau \gg 1$) they differ significantly, see also Figure 3. Notice that for $\lambda = 1.5$ $P(q,\tau)$ exhibits an overshoot at intermediate times and $P(0,\tau)$ an undershoot.

Next, we consider in Figure 5 the fraction of naked polynucleotides $P(0,\tau)$ and the fraction of fully encapsulated VLPs $P(q,\tau)$ as a function of $\sigma\tau$ for a cooperativity of $\sigma = e^{-5}$, a number of binding sites per nucleotide of $q = 51$, a scaled protein concentration $S = e^2$, and for an excess as well as shortage concentration of proteins, setting stoichiometry at values of $\lambda = 0$ and $\lambda = 1.5$. The dotted lines indicate the equilibrium values of the various curves.

Echoing what we find for the fraction occupied sites in Figure 4, we find that the decrease of the fraction of naked polynucleotides is, for all intents and purposes, independent of the

stoichiometry λ up to a time $t \approx 1/\phi_P k_+$. This is the typical time needed for one successful nucleation per each polynucleotide, and corresponds to $\ln \sigma\tau \approx -5$ because we set $\sigma = e^{-5}$ and $\tau \equiv \phi_P k_+ t$. Moreover, the probability of subsequent cooperative binding of proteins to a polynucleotide is, after nucleation has occurred, initially much larger than in the long time limit, at least for $\lambda = 1.5$.

The reason for this is straightforward to understand. Note first that if the number of available binding sites is larger than the number of proteins, the equilibrium concentration of free proteins is close to the critical concentration, whereas the initial free protein concentration is approximately equal to the overall protein concentration. The initial sensitivity to the value of λ and favourable cooperative binding causes the initial evolution for a solution of $\lambda = 1.5$ to be similar to that of a solution of $\lambda = 0$. This means that the undershoot in the concentration of naked polynucleotides $P(0, \tau)$ and the overshoot in the concentration of fully formed VLPs $P(q, \tau)$ for $\lambda = 1.5$ are a consequence of the correction to the initial evolution of the solution in order to acquire the correct long time limit of the respective concentrations.

From Figures 4 and 5 we expect that the evolution of $\langle\theta\rangle(\tau)$, $P(0, \tau)$ and $P(q, \tau)$ exhibits non-monoticity only if the ratio of the number of binding sites and proteins is greater than zero, $\lambda > 0$. This condition is required but not sufficient. We investigated for what values of the stoichiometry λ and the scaled protein concentration S we find non-monotonic time evolution in either of the three quantities. The results are summarised in Figure 6 for a cooperativity of $\sigma = e^{-5} \approx 0.007$ and a number of binding sites per polynucleotide of $q = 51$. The region of overshoots in $P(q, \tau)$ does not overlap with the region of undershoot in $P(0, \tau)$ and overshoots in $\langle\theta\rangle(\tau)$, except for a narrow boundary region, where for example the parameter values of the evolution shown in Figure 5 are located.

For small values of the stoichiometry we obtain a region of monotonic time evolution evolution, and the boundary of this region with the $P(q, \tau)$ overshoot region correlates with a large sensitivity of the equilibrium value of fully encapsulated polynucleotides $P_{eq}(q)$ to

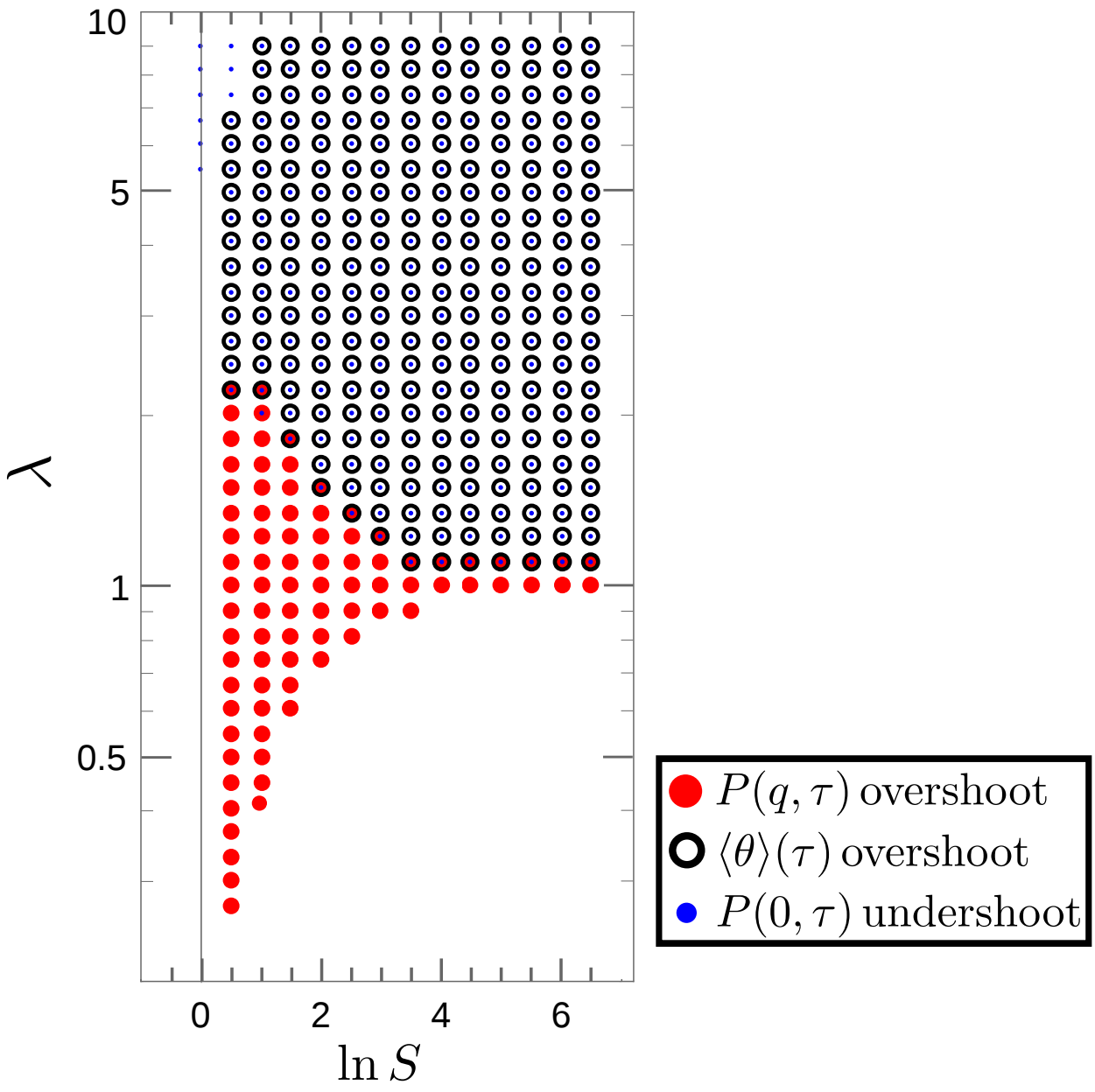

Figure 6: Non-monotonic evolution of the mean fraction of occupied binding sites $\langle\theta\rangle(\tau)$, the fraction of fully formed VLPs $P(q,\tau)$ and the fraction of naked polynucleotides $P(0,\tau)$, as a function of the ratio of the number of available binding sites and proteins λ and the ratio of the protein concentration to the critical concentration S. The location of the boundary between the overshoots in $\langle\theta\rangle(\tau)$ and the monotonic region for low λ is correlated to a large sensitivity of the equilibrium fraction of fully formed VLPs with respect to the stoichiometry, as can be seen in Figure 3 for $S = e^2$.

the stoichiometry λ, as can be seen in Figure 3 for the case of $S = e^2 \approx 7.4$. The crossovers between the $P(q, \tau)$ overshoots and the other non-monoticities, however, do not seem to be correlated to the sensitivity of any equilibrium concentration to the control variable, and their cause remains unclear.

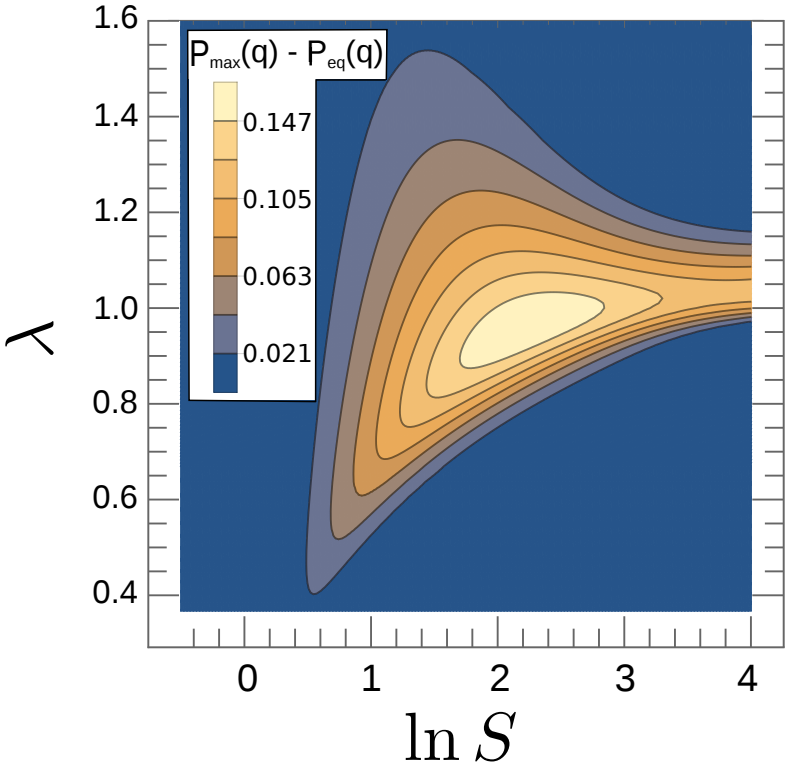

Figure 7: The difference in the maximum fraction of fully encapsulated VLPs during assembly $P_{max}(q)$ and the equilibrium fraction of fully formed particles $P_{eq}(q)$ as a function of the ratio of the number of available binding sites to the number of proteins λ and the ratio of the overall protein concentration to the critical concentration S. The maximum difference on the raster is 0.165 at $S = e^{2.2}$ and $\lambda = 0.96$.

Of interest is not only under what conditions of stoichiometry λ and protein concentration S what kind of dynamical regime prevails. Also of interest is what the *magnitude* of the over- or undershoot is. As an example, we present in Figure 7 the magnitude of the overshoot in $P(q, \tau)$, defined as the difference between the maximum fraction of fully encapsulated VLPs during the assembly $P_{max}(q)$ and the equilibrium fraction $P_{eq}(q)$, as a function of λ and S. The maximum overshoot that we find is $P_{max}(q) - P_{eq}(q) = 0.165$ at $S = e^{2.2} \approx 9$ and

$\lambda = 0.96$, at least for a cooperativity of $\sigma = e^{-5}$ and a number of available binding sites per polynucleotide of $q = 51$. This is a sizeable overshoot, considering that for those conditions $P_{eq}(q) = 0.17$.

In the next section we compare our theory to the experiments on dsDNA and two kinds of coat protein under conditions of non-zero stoichiometry. We pinpoint the parameters which enter in our kinetic model: the nucleation free energy associated with switching a protein from the solution to the bound conformation σ, the sum of the protein-protein and protein-DNA interaction free energy $\epsilon + g$, and the binding rate of a protein to the DNA k_+.

Comparison to experiment: procedure

In the previous section we studied the time evolution of the concentrations of the different supramolecular species that according to our model assemble in a solution of proteins and polynucleotides under conditions of a non-zero stoichiometric ratio. The question arises how realistic the model actually is. To put our model to the test, we aim to compare our model predictions with our experimental findings. For that purpose, we first describe our experimental procedures and the curve-fitting methodology to obtain thermodynamic and kinetic information on the temporal evolution of the length distribution of protein-DNA aggregates of two kinds of protein. Then, we will draw a quantitative comparison with our model and also re-evaluate the earlier experiments of Hernandez-Garcia et al.[1]

In our experiments we measure by means of AFM imaging, at different points in time, the length of the significantly thickened part of DNA strands caused by binding of proteins to DNA. See Figure 8a. Details on the proteins and the DNA follow below. The experimental procedure can be summarized as follows: proteins, dsDNA and stabilizing agents in stock solutions are pipetted into a microtube to obtain the desired concentration of, and ratio between, protein and DNA molecules. We vortex the solution for a few seconds and incubate it at room temperature (approximately 20 °C); the start of incubation is taken to be $t = 0$

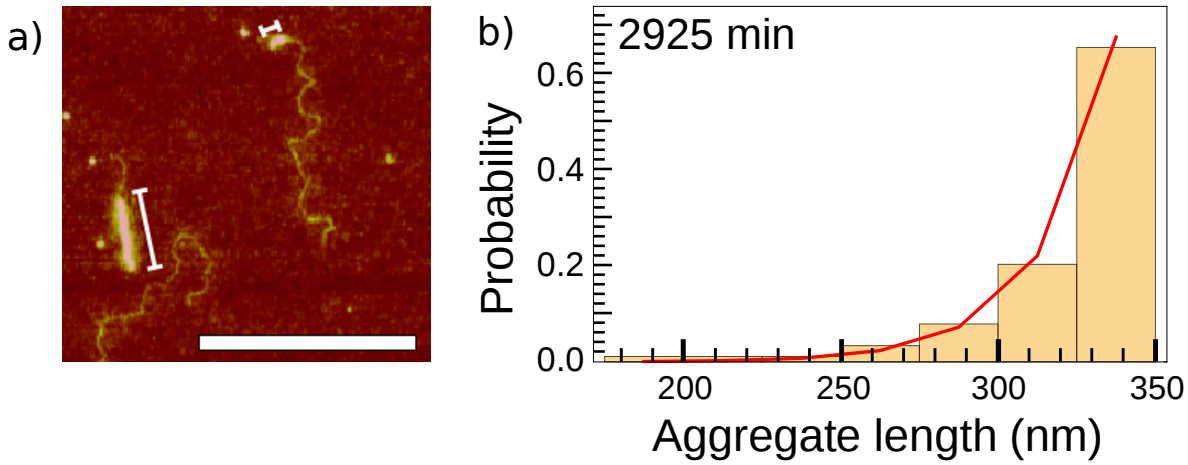

Figure 8: **a**) AFM image of the protein-DNA aggregates present in a sample of dsDNA and coat proteins. The bright rod-like complexes show the thickened part of the DNA encapsulated by protein (indicated by white bars) and the thin curved structures correspond to the uncovered parts of DNA molecules. The scale bar corresponds to 500 nm. **b**) The measured (yellow bars) and fitted (red curve) probability distribution of the length of the aggregates arising in a solution of dsDNA and coat proteins with fourteen silk strands (C-S_{14}-B) after 2925 minutes of self-assembly. The measured probability distribution is obtained by dividing the range of the measured aggregate lengths in intervals of 25 nm, and by calculating what fraction of length measurements lie in a given length interval. The fit is constructed from the equilibrium probability of finding a nucleated dsDNA in the solution with n proteins attached, by letting $n = q$ correspond to the largest measured aggregate length, where q is the number of available binding sites per DNA molecule, and by summing over the probabilities whose corresponding lengths fall within a length interval.

of the experiment.

At multiple points in time we take a sample of 3-5 μL from the solution, deposit it on a silicon surface and let it rest there for about two minutes. Thereafter, we rinse the surface with 1 mL of pure water (MQ-water) to remove salts and non-absorbed particles. The excess of water is soaked up laterally with a tissue and the surface is dried with N_2 steam. To correct for the possible binding of proteins to DNA before $t = 0$ and during deposition of the sample on the silicon surface, and to account for any other disturbances of the sample during rinsing and drying, we add an offset time t_0 to the sample times, which we determine afterwards by fitting our model to the measured length distributions.

The dried sample is imaged using atomic force microscopy (AFM). From the AFM images, we measure the length of the thickened part of the aggregates, which is assumed to be proportional to the number of proteins bound to the DNA molecule and will be referred to as the aggregate length. A typical part of an AFM image is shown in Figure 8a, in which also the length of the thickened parts of the protein-DNA aggregates has been indicated by white bars. We identify this length as the encapsulated length. Distributions of encapsulated aggregate lengths for different incubation times are obtained by determining the encapsulated lengths of around 50 protein-DNA aggregates from a sample of the solution. An example of such an experimentally determined distribution is shown in Fig. 8b. These experimental distributions can be compared with theoretical predictions using our theory in a way that we describe next.

For the length measurement we only consider the thickened part of the aggregates and ignore the presence of naked DNA molecules. This is inevitable also because of the waiting time and the limited field of view in the experiments. Hence, we need to compare the experimental findings to a theoretical probability distribution of nucleated polynucleotides $P_{nuc}(n, \tau)$, defined as

$$P_{nuc}(n, \tau) = \frac{P(n, \tau)}{1 - P(0, \tau)}, \tag{11}$$

for $1 \leq n \leq q$, with $P(n, \tau)$ the probability of finding a DNA molecule with n proteins bound

at dimensionless time $\tau = \phi_P k_+ t$, q the number of available binding sites per DNA molecule, and k_+ the binding rate of proteins to the DNA.

To compare the theoretical distribution of partially covered polynucleotides to the experimental length distribution, we construct a histogram for each sample by first dividing the range of measured aggregate lengths in intervals, and, second, by calculating what fraction of the measured aggregate lengths falls in each interval. Details can be found in the Supporting Information. We scale the theoretical distribution to the experimental one, by letting $n = q$ correspond to the largest measured aggregate length, where q is the number of available binding sites per DNA molecule. We then collect all theoretically obtained assembly lengths into size categories corresponding to the experimental ones. By comparing the experimental distribution to the curve fits, we determine the value of the fitting parameters. We optimise the fit by eye because the statistical uncertainty in the measured probability distribution is large. This is due to the relatively small number of counts per category, which are of the order 10, at most. For more details on the production of the proteins, the employed AFM imaging technique, and the processing of the raw data, the reader is referred to the Supporting Information.

We set our model parameters to mirror the experimental conditions of the number of available binding sites per DNA molecule q and the ratio of the number of available binding sites and the proteins in the solution, i.e., the stoichiometric ratio λ. By taking into account the approximate charge neutrality of a fully formed VLP,[1] we can calculate the number of available binding sites per DNA molecule q. Each dsDNA molecule consists of 2.5×10^3 base pairs, which implies a net charge of -5.0×10^3 elementary charges. The capsid proteins have a positive charge of +12 elementary charges each. Therefore, the number of available binding sites is $q = 5000/12 \approx 417$. Given the molar concentration of protein and DNA: $c_P \simeq \phi_P \cdot c_{H_2O}$ and $c_{DNA} \simeq \rho_T \cdot c_{H_2O}$ with c_{H_2O} the molar concentration of water molecules, 55.6 M, we can calculate the stoichiometric ratio as $\lambda \equiv q\rho_T/\phi_P = qc_{DNA}/c_P$.

By curve fitting, we extract the values for the ratio of the total protein concentration to

the critical concentration, i.e., the scaled protein concentration S, the scaled concentration of free proteins in equilibrium, s_{eq}, the cooperativity parameter, σ, the binding rate of the proteins to the DNA, k_+, and the offset time, t_0, which we add to the time of all sample measurements to take into account the uncertainties inherent to the experimental procedure. For both kinds of protein, that we refer to as C-S_{10}-B and S_{14}, the position of the peak in the distribution at 2 min determines the magnitude of the binding rate of proteins to the DNA k_+. For the case of the C-S_{10}-B proteins, the offset time t_0 is set by the position of the peak at 15 min, and the cooperativity σ through the width of the distribution at 8 min. Moreover, for the C-S_{14}-B proteins the position and form of the peak at 15 min determines both the offset time and the cooperativity.

Finally, to determine the scaled concentration of free proteins s_{eq}, we consider the length distribution of the (partially) assembled polynucleotides at long times when equilibrium presumably has been reached. In particular, we calculate the equilibrium distribution of (partially) assembled polynucleotides $P_{nuc,eq}(n)$ by using the fact that in equilibrium $P(n,\tau) \rightarrow P_{eq}(n)$ and invoking Eqs. 1 and 11. As $P_{nuc,eq}(n)$ only concerns nucleated polynucleotides, it depends only on s_{eq} and q, and we are able to determine s_{eq} from the fit.

With the cooperativity parameter σ found, we can obtain the scaled protein concentration S by inserting the cooperativity parameter σ, the stoichiometry λ and the scaled concentration of free proteins s_{eq} into the mass conservation equation, see Eq. (3). Moreover, after calculating the dimensionless concentration protein, ϕ_P, we can extract the critical concentration, $\phi_c = \phi_P/S = e^{\epsilon+g}$. To find ϕ_P we consider its definition, $\phi_P = N_P/N_t$, where $N_t = N_s + N_P + N_D$ is the sum of the number of solvent, DNA and protein molecules dissolved. Since $N_{solvent}$ is about seven orders of magnitude larger than N_P and N_D, we write

$$\phi_P \approx \frac{N_P}{N_s} = \frac{c_P}{c_s}, \tag{12}$$

with $c_i = N_i/VN_A$, V the volume of the solution and N_A Avogadro's number. From the dimensionless concentration of proteins ϕ_P we calculate the dimensionless critical concentra-

tion ϕ_c.

In the next section curve fits to the evolution of the length distribution of aggregates of dsDNA and two kinds of protein are presented as well as a refit of the experimental data of Hernandez-Garcia et al.,[1] where we take into account the non-zero stoichiometry of the experiments instead of the presumed zero stoichiometry of that work.

Comparison to experiment: curve fits

Hernandez-Garcia et al.[1] reported on the design of artificial coat proteins using simple polypeptide domains which have physicochemical functionalities resembling those of viral coat proteins. As illustrated in Figure 1a, the coat proteins were made of three blocks: an oligolysine binding block B=K_{12} that binds non-sequence specifically to the dsDNA through electrostatic interactions, a silk-like sequence S_n consisting of a variable number of silk strands S_n=$(GAGAGAGQ)_n$ that putatively dictates the cooperativity of the protein binding to the DNA, and a hydrophilic random coil sequence C that prevents the different complexes from aggregating.

Hernandez-Garcia et al.[1] demonstrated that the effective, compact encapsulation of the dsDNA depends strongly on the number of silk sequences per protein. They found that the triblock proteins $C-S_{10}-B$ and $C-S_{14}-B$ that we consider here, assembled into rod-like VLPs. The silk-strands fold into a beta-solenoid, as shown in the work of Zhao et al.[23] Different solenoids stack on top of each other due to hydrogen bonding, forming the rod-shaped silk core of the VLP. Hernandez-Garcia et al.[1] also demonstrated that charge neutralization of the DNA by the binding blocks dictates that the DNA is compacted by a factor of around three, illustrated in the schematic model of Figure 1b.

To investigate the influence of the number of silk strands per protein, we present curve fits following the procedure outlined in the previous section to the time evolution of the length distribution of two kinds of protein: one with ten silk strands (C-S_{10}-B) and the other with

fourteen silk strands (S_{14}). This enables us to probe the sensitivity of our model parameters (among which the cooperativity σ) to the number of silk-like sequences in the proteins used. Our experiments have been carried out at (approximately) the same stoichiometry, allowing us to probe the influence of the number of silk sequences. The production method of the proteins is the same as Hernandez-Garcia et al.[1] previously reported on, details of which can be found in the Supporting Information. The dsDNA molecules we use are NoLimits^{TM} individual DNA fragments consisting of 2.5×10^3 base pairs.

Furthermore, we re-evaluate the data of Hernandez-Garcia et al.[1] to probe how sensitive the fitting parameters of our model are to invoking a presumed zero stoichiometry instead of the actual stoichiometry. We present the curve fits to our new data and those of Hernandez-Garcia et al. separately in the next three sections.

Curve fitting to the C-S_{10}-B assembly data

For the coat protein with ten silk strands we use a molar concentration of dsDNA of $c_{DNA} = 0.65\,\mathrm{nM}$ and a protein concentration of $c_P = 2695\,\mathrm{nM}$, giving a stoichiometry of $\lambda = 0.101$ and a dimensionless protein concentration of $\phi_P = 4.84 \times 10^{-8}$. The times at which samples were taken are t = 2, 8, 15, 25, 85, 180, 360, 480, 1500 and 3000 min. From the measurement at t = 3000 min we find the scaled concentration of free proteins in equilibrium to be $s_{eq} = 1.03$ and using s_{eq}, we fit our theoretical model to the temporal evolution of the experimental length distribution, see Figure 9.

By optimising the curve fit we find a cooperativity of $\sigma = 0.004$, implying that the nucleation free energy associated with switching a protein from the solution to the bound conformation is $h - \epsilon \approx 5.5\,k_BT$. The binding rate of a protein to the DNA is $k_+ = 4.1 \times 10^9\,\mathrm{min}^{-1}$, and the offset time, taking into account possible disturbances of the samples, is $t_0 = 7\,\mathrm{min}$. These three parameters are determined by the position of the peak at 2 min, the peak at 15 min and the width of the distribution at t = 2, 8 and 15 min. We find the scaled protein

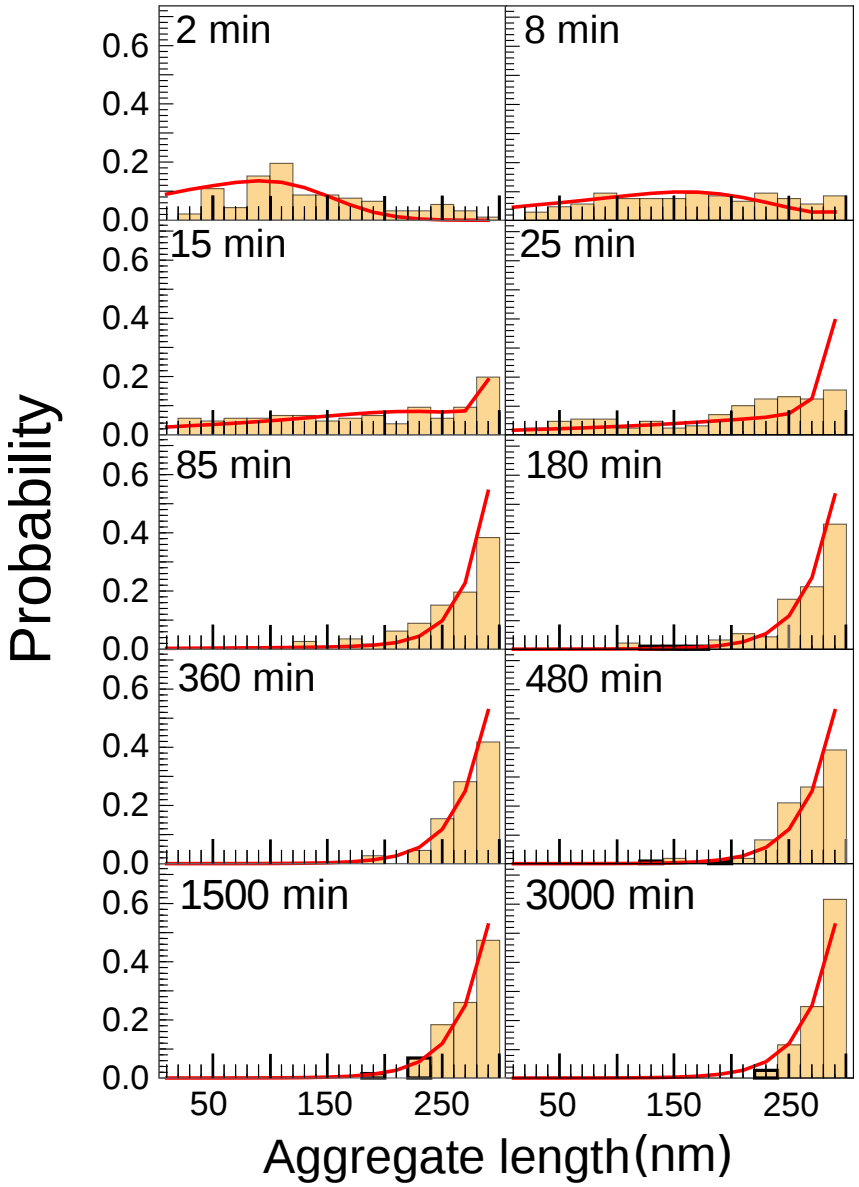

Figure 9: The measured (yellow bars) and fitted (red curves) probability distributions of the length of the aggregates arising in samples drawn from a solution of 2.5 kbp dsDNA and coat proteins with ten silk strands (C-S_{10}-B) at multiple stages of the self-assembly process. The fit is determined by the position of the peak at 2 min, the width of the distribution at 8 min, the position of the peak at 15 min and the equilibrium distribution at 3000 min.

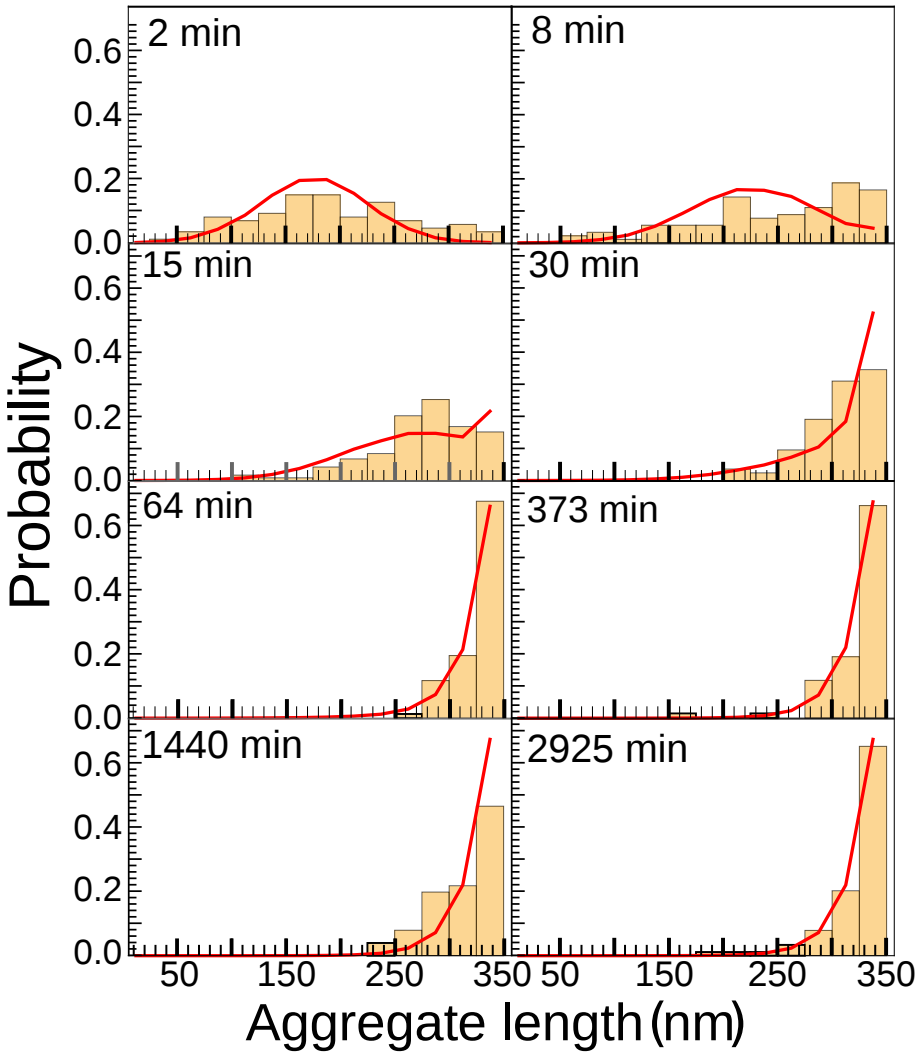

Figure 10: The measured (yellow bars) and fitted (red curves) probability distributions of the aggregate lengths in a solution of 2.5 kbp dsDNA and coat proteins with fourteen silk strands (C-S_{14}-B) at different times in the self-assembly process. The fit is determined by the position of the peak at 2 min, the position and height of the peak at 15 min and the equilibrium distribution at 2925 min.

concentration to be $S = 1.13$ and obtain for the sum of the protein-protein and protein-DNA interaction free energy a value of $\epsilon + \mathrm{g} \approx -17\,\mathrm{k_B T}$.

These are reasonable values of the energetic parameters, for the self-assembly of VLPs is a thermodynamically driven process involving non-covalent bonding of proteins to the DNA. Furthermore, the free energies found by Hernandez-Garcia et al.[1] for the self-assembly of proteins with ten silk strands, with $\mathrm{h} - \epsilon \approx 5.3\,\mathrm{k_B T}$ and $\epsilon + \mathrm{g} \approx -18\,\mathrm{k_B T}$, are reasonably close to the values we find here, although they presume in their curve fitting a stoichiometry of zero while the actual value was $\lambda = 0.324$. This shows that the value of the energetic parameters they find are robust.

Curve fitting the C-S_{14}-B assembly data

To probe the influence of the number of silk strands on the binding energies we conduct an assembly experiment with coat proteins having fourteen silk strands and compare the fitted parameter values with those of a protein with ten silk strands. We conduct the experiment with C-S_{14}-B proteins at a molar concentration of dsDNA of $\mathrm{c_{DNA}} = 0.65\,\mathrm{nM}$ and a concentration of proteins of $\mathrm{c_P} = 2016\,\mathrm{nM}$, implying a stoichiometric ratio of $\lambda = 0.134$ and a dimensionless protein concentration of $\phi_P = 3.63 \times 10^{-8}$. The times at which samples were taken are t = 2, 8, 15, 30, 64, 373, 1440, 2925 min. From the measurement of the length distribution at t = 2925 min we obtain a scaled concentration of free proteins of $s_{eq} = 1.04$; see Figure 8b.

In Figure 10 we present the fit of our model to the temporal evolution of the experimental length distribution. From this fit we find the cooperativity to be equal to $\sigma = 0.05$, giving a nucleation free energy associated with switching a protein from the solution to the bound conformation of $\mathrm{h} - \epsilon \approx 3.0\,\mathrm{k_B T}$. Comparing this nucleation free energy to that for our proteins with ten silk strands, which we found to be equal to $\mathrm{h} - \epsilon \approx 5.5\,\mathrm{k_B T}$, we conclude that a protein with a larger number of silk strands must have associated with it a *smaller*

nucleation free energy. This ties in with data for fibril-forming proteins with much longer silk blocks (S_{24} and S_{48}), for which it was found that fibril and hence hydrogel formation was much more rapid for proteins with longer silk blocks.[24]

Returning to our experimental system, the cause of the lower nucleation energy for the larger protein (C-S_{14}-B) may be that it has a larger number of configurations in the bound conformation, implying that the C-S_{14}-B proteins have an entropically more favourable switch to the bound conformation than the C-S_{10}-B proteins.

The binding rate of a protein to the DNA is $k_+ = 3.6 \times 10^9\,\mathrm{min}^{-1}$, which is smaller than for proteins with ten silk strands ($k_+ = 4.1 \times 10^9\,\mathrm{min}^{-1}$). As noted above, we have a small number of measured aggregates per interval of the length distribution, giving rise to a significant uncertainty in the value of the fitted parameters, and therefore we do not find the binding rate of protein to the DNA to differ significantly. Moreover, we find the offset time, which takes into account possible disturbances of the sample, to be $t_0 = 12\,\mathrm{min}$, which is comparable to but larger than the value we obtained for our experiment with C-S_{10}-B proteins, $t_0 = 7\,\mathrm{min}$. The reason for this difference is not clear. The C-S_{14}-B proteins might be more sensitive to the process of rinsing and drying because of their size compared to the C-S_{10}-B proteins.

Finally, to determine the fitting parameters we focused on the position of the peak at 2 min and 15 min. The peak at 15 min pinned down the values of both σ and t_0. From these fitting parameters we obtain a scaled protein concentration of $S = 1.19$ and the sum of the protein-protein and protein-DNA interaction free energy as $\epsilon + g \approx -10\,k_B T$.

Re-evaluation of the data of Hernandez-Garcia et al.[1]

Hernandez-Garcia et al.[1] measured the length of the different aggregates arising in a solution of S_{10} proteins and dsDNA at times t = 10, 60, 350, 1485, 2880, 7440 minutes. Their curve fit, with a presumed zero stoichiometry, provided a nucleation free energy associated with

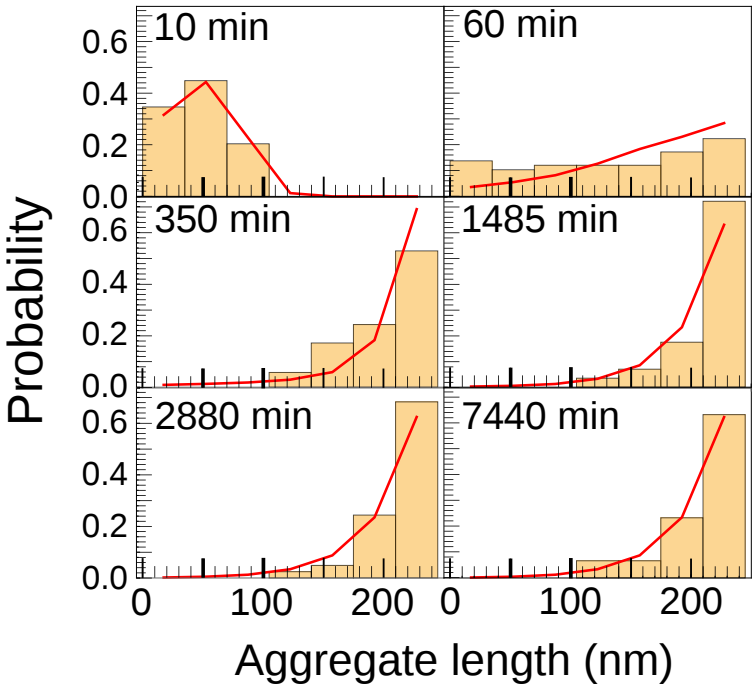

Figure 11: The data of Hernandez-Garcia et al.[1] (yellow bars) and the fitted (red curves) probability distributions of the aggregate lengths in a solution of 2.5 kbp dsDNA and coat proteins with ten silk strands (C-S_{10}-B) at different times in the self-assembly process. The fit is made under conditions of non-zero stoichiometry, as opposed to the fit in the earlier work of Hernandez-Garcia et al., and is determined by the position of the peak at 10 and 60 minutes, and the increase of the distribution in the last three bins in the measurement at 60 minutes.

switching a protein from the solution to the bound conformation of $h - \epsilon \approx 5.3\,k_B T$ and a sum of the protein-protein and protein-DNA interaction free energy of $\epsilon + g \approx -18\,k_B T$. To evaluate the sensitivity of the interaction free energies to the presumed stoichiometry in the model we fit their data with our model under conditions of a non-zero stoichiometry.

The molar concentration of proteins is $c_P = 830\,nM$ and the concentration of DNA molecules is $c_{DNA} = 0.65\,nM$. This implies a stoichiometric ratio of $\lambda = 0.324$ and a dimensionless protein concentration of $\phi_P = 1.49 \times 10^{-8}$. From the presumed equilibrium measurement at $t = 7440$ minutes we extract the scaled concentration of the free proteins in equilibrium to be $s_{eq} = 1.016$. The fit of our kinetical model to the temporal evolution of the experimental length distribution of the data of Hernandez-Garcia[1] is presented in Figure 11.

From this fit we find a cooperativity of $\sigma = 0.005$, implying the nucleation free energy associated with switching a protein from the solution to the bound conformation to be $h - \epsilon \approx 5.3\,k_B T$. The binding rate of proteins to the DNA is $k_+ = 2.2 \times 10^9\,min^{-1}$, and the offset time is $t_0 = 5\,min$. These three parameters have been determined by focusing on the position of the peak at 10 and 60 minutes, and the increase of the distribution in the last three bins in the measurement at 60 minutes. We calculate the scaled protein concentration as $S = 1.40$ and find the sum of the protein-protein and protein-DNA interaction free energy to be $\epsilon + g \approx -18\,k_B T$.

In summary, the values we find under conditions of non-zero stoichiometry are in agreement with the values obtained by Hernandez-Garcia et al.[1], implying the free energy parameters of our model to be robust with respect to the stoichiometry.

Conclusion and discussion

In this paper we experimentally measure and theoretically describe the self-assembly kinetics of synthetic coat proteins and polynucleotides using the nucleated Zipper model under conditions of finite protein concentrations, that is, a non-zero stoichiometry. We find non-

monotonic temporal evolution of the mean fraction of occupied binding sites, the fraction of fully encapsulated polynucleotides and the fraction of free polynucleotides. This is in contrast to the earlier results of Kraft et al.[13] for the case of zero stoichiometry, that is, an infinite excess of proteins in solution. We pinpoint under what conditions what kind of non-monotocity prevails and quantify the level of overshoot in the fraction of fully encapsulated polynucleotides.

We conducted assembly experiments with dsDNA and two engineered proteins in solution. The proteins include a silk-like domain with ten (C-S_{10}-B) and fourteen repeat units (S_{14}). The number of repeat units was previously hypothesized to be directly related to the level of cooperativity of the binding process. From the fit of our model to the temporal evolution of the length distribution of the aggregates we find for the C-S_{10}-B proteins a nucleation free energy, associated with switching a protein from the solution to the bound conformation, of $h - \epsilon = 5.5\,\mathrm{k_B T}$, a binding rate of proteins to the DNA of $k_+ = 4.1 \times 10^9\,\mathrm{min}^{-1}$, and for the sum of the protein-protein and protein-DNA interaction free energy, i.e., the binding free energy $\epsilon + g = -18\,\mathrm{k_B T}$. For the C-$S_{14}$-B protein we obtain $h - \epsilon = 3.0\,\mathrm{k_B T}$, $k_+ = 3.6 \times 10^9\,\mathrm{min}^{-1}$ and $\epsilon + g = -10\,\mathrm{k_B T}$, respectively.

The sum of protein-protein and protein-DNA binding free energies for the proteins with fourteen silk strands is smaller by a factor of almost 2 compared to that for the proteins with ten silk strands. This is surprising because the two proteins feature identical binding domains and, moreover, the larger silk block is expected to lead to a stronger protein-protein interaction. The same reduction factor appears when comparing the sum of the configurational and protein-protein free interaction energies, $h-\epsilon$, between the two proteins. This implies that also the sum of the protein-DNA binding and the configurational free energies, $g+h$, differ by almost the same factor or, in other words, that the protein with the larger silk domain is less prone to bind to DNA than that with the shorter one. A possible cause for this is steric hindrance by the larger silk block.

In addition, we re-analysed the data of Hernandez-Garcia et al.[1] accounting for the finite

protein concentration. We find that the free energies are not very sensitive to the exact stoichiometry, at least for sufficiently small values of this quantity, supporting the previously obtained values. We do note, however, that each size category of the histogram data only contains about ten measurements and hence that the measured data have a significant uncertainty. Furthermore, the fitting requires assigning values to no fewer than four free parameters, which in itself poses a challenge in establishing accurate values. This implies that to be able to observe the predicted non-monotonicity in the assembly dynamics, experiments need to be performed at much larger stoichiometries and with a significantly larger set of data points than presented here.

Acknowledgement

We gratefully acknowledge funding through VENI grant 680-47-431 by the Netherlands Organisation for Scientific Research (NWO). A.H.-G. is financially supported by the Dutch Polymer Institute (DPI), project #698 SynProt and the Consejo Nacional de Ciencia y Tecnología (CONACyT), México.

Supporting Information Available

Numerical methods - protein production - protein amino acid sequence - imaging method - raw data processing. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Hernandez-Garcia, A.; Kraft, D. J.; Janssen, A. F.; Bomans, P. H.; Sommerdijk, N. A.; Thies-Weesie, D. M.; Favretto, M. E.; Brock, R.; de Wolf, F. A.; Werten, M. W.

et al. Design and Self-Assembly of Simple Coat Proteins for Artificial Viruses. *Nat. Nanotechnol.* **2014**, *9*, 698–702.

(2) Glasgow, J.; Tullman-Ercek, D. Production and Applications of Engineered Viral Capsids. *Appl. Microbiol. Biotechnol.* **2014**, *98*, 5847–5858.

(3) Hernandez-Garcia, A.; Werten, M. W.; Stuart, M. C.; de Wolf, F. A.; de Vries, R. Coating of Single DNA Molecules by Genetically Engineered Protein Diblock Copolymers. *Small* **2012**, *8*, 3491–3501.

(4) Machida, K.; Imataka, H. Production Methods for Viral Particles. *Biotechnol. Lett* **2015**, *37*, 753–760.

(5) Unzueta, U.; Saccardo, P.; Domingo-Espín, J.; Cedano, J.; Conchillo-Solé, O.; García-Fruitós, E.; Céspedes, M. V.; Corchero, J. L.; Daura, X.; Mangues, R. et al. Sheltering DNA in Self-Organizing, Protein-Only Nano-Shells as Artificial Viruses for Gene Delivery. *Nanomed. Nanotechnol. Biol. Med.* **2014**, *10*, 535–541.

(6) Zhou, J. C.; Soto, C. M.; Chen, M.-S.; Bruckman, M. A.; Moore, M. H.; Barry, E.; Ratna, B. R.; Pehrsson, P. E.; Spies, B. R.; Confer, T. S. Biotemplating Rod-Like Viruses for the Synthesis of Copper Nanorods and Nanowires. *J. Nanobiotechnol.* **2012**, *10:18*, 1–12.

(7) Mastrobattista, E.; van der Aa, M. A.; Hennink, W. E.; Crommelin, D. J. Artificial Viruses: a Nanotechnological Approach to Gene Delivery. *Nat. Rev. Drug Discovery* **2006**, *5*, 115–122.

(8) Mou, Q.; Ma, Y.; Jin, X.; Zhu, X. Designing Hyperbranched Polymers for Gene Delivery. *Mol. Syst. Des. Eng.* **2016**, *DOI:10.1039/c5me00015g*, 1–15.

(9) Lim, Y.-B.; Lee, E.; Yoon, Y.-R.; Lee, M. S.; Lee, M. Filamentous Artificial Virus from a Self-Assembled Discrete Nanoribbon. *Angew. Chem. Int. Ed.* **2008**, *120*, 4601–4604.

(10) Aoyama, Y.; Kanamori, T.; Nakai, T.; Sasaki, T.; Horiuchi, S.; Sando, S.; Niidome, T. Artificial Viruses and their Application to Gene Delivery. Size-Controlled Gene Coating with Glycocluster Nanoparticles. *J. Am. Chem. Soc.* **2003**, *125*, 3455–3457.

(11) Janssen, P. G.; Meeuwenoord, N.; van der Marel, G.; Jabbari-Farouji, S.; van der Schoot, P.; Surin, M.; Tomović, Z.; Meijer, E. W.; Schenning, A. P. ssPNA Templated Assembly of Oligo(p-phenylenevinylene)s. *Chem. Commun.* **2010**, *46*, 109–111.

(12) Krejchi, M. T.; Atkins, E. D. T.; Waddon, A. J.; Fournier, M. J.; Mason, T. L.; Tirrell, D. A. Chemical Sequence Control of Beta-Sheet Assembly in Macromolecular Crystals of Periodic Polypeptides. *Science* **1994**, *265*, 1427–1432.

(13) Kraft, D. J.; Kegel, W. K.; van der Schoot, P. A Kinetic Zipper Model and the Assembly of Tobacco Mosaic Virus. *Biophys. J.* **2012**, *102*, 2845–2855.

(14) Muthukumar, M.; Ober, C. K.; Thomas, E. L. Competing Interactions and Levels of Ordering in Self-Organizing Polymeric Materials. *Science* **1997**, *277*, 1225–1232.

(15) Hagan, M. Modeling Viral Capsid Assembly. *Adv. Chem. Phys.* **2014**, *155*, 1–68.

(16) Caspar, D. L. Movement and Self-Control in Protein Assemblies. *Biophys. J.* **1980**, *32*, 103–138.

(17) Punter, M. T. The Role of Assembly Signals in the Self-Assembly of Linear Viruses. M.Sc. thesis, Utrecht University, July 2015.

(18) Janssen, P. G.; Jabbari-Farouji, S.; Surin, M.; Vila, X.; Gielen, J. C.; de Greef, T. F.; Vos, M. R.; Bomans, P. H.; Sommerdijk, N. A.; Christianen, P. C. et al. Insights into Templated Supramolecular Polymerization: Binding of Naphthalene Derivatives to ssDNA Templates of Different Lengths. *J. Am. Chem. Soc.* **2009**, *131*, 1222–1231.

(19) McGhee, J.; von Hippel, P. Theoretical Aspects of DNA-Protein Interactions: Cooperative and Non-Cooperative Binding of Large Ligands to a One-Dimensional Homogeneous Lattice. *J. Mol. Biol.* **1974**, *86*, 469–489.

(20) Jabbari-Farouji, S.; van der Schoot, P. Competing Templated and Self-Assembly in Supramolecular Polymers. *Macromolecules* **2010**, *43*, 5833–5844.

(21) Jabbari-Farouji, S.; van der Schoot, P. Theory of Supramolecular Co-Polymerization in a Two-Component System. *J. Chem. Phys.* **2012**, *064906*, 1–14.

(22) Mahalik, J. P.; Muthukumar, M. Langevin dynamics simulation of polymer-assisted virus-like assembly. *J. Chem. Phys.* **2012**, *136*, 1–13.

(23) Zhao, B.; Cohen Stuart, M. A.; Hall, C. K. Dock 'n Roll: Folding of a Silk-Inspired Polypeptide into an Amyloid-Like Beta Solenoid. *Soft Matter* **2016**, *12*, 3721–3729.

(24) Beun, L. H.; Storm, I. M.; Werten, M. W.; de Wolf, F. A.; Cohen Stuart, M. A.; de Vries, R. From Micelles to Fibers: Balancing Self-Assembling and Random Coiling Domains in pH-Responsive Silk-Collagen-Like Protein-Based Polymers. *Biomacromolecules* **2014**, *15*, 3349–3357.

Graphical TOC Entry

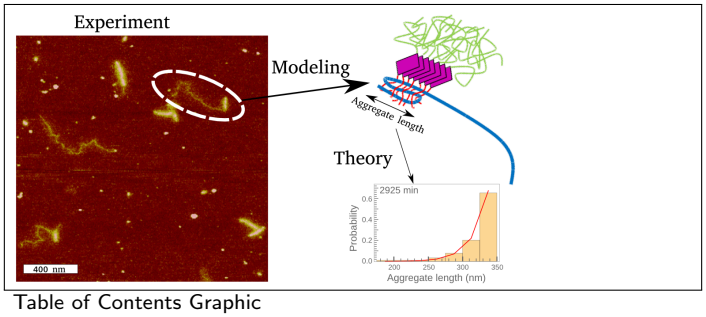

Table of Contents Graphic