Biophysical Journal
Article

Biophysical Society

A kinetic model for the impact of packaging signal mimics on genome encapsulation

René de Bruijn,[1,2,*] Pieta Cornelia Martha Wielstra,[1] Carlos Calcines-Cruz,[3] Tom van Waveren,[1] Armando Hernandez-Garcia,[3] and Paul van der Schoot[1]

[1]Department of Applied Physics, Eindhoven University of Technology, Eindhoven, the Netherlands; [2]Institute for Complex Molecular Systems, Eindhoven University of Technology, Eindhoven, the Netherlands; and [3]Department of Chemistry of Biomacromolecules, Institute of Chemistry, National Autonomous University of Mexico, Mexico City, Mexico

ABSTRACT Inspired by recent experiments on the spontaneous assembly of virus-like particles from a solution containing a synthetic coat protein and double-stranded DNA, we put forward a kinetic model that has as main ingredients a stochastic nucleation and a deterministic growth process. The efficiency and rate of DNA packaging strongly increase after tiling the DNA with CRISPR-Cas proteins at predesignated locations, mimicking assembly signals in viruses. Our model shows that treating these proteins as nucleation-inducing diffusion barriers is sufficient to explain the experimentally observed increase in encapsulation efficiency, but only if the nucleation rate is sufficiently high. We find an optimum in the encapsulation kinetics for conditions where the number of packaging signal mimics is equal to the number of nucleation events that can occur during the time required to fully encapsulate the DNA template, presuming that the nucleation events can only take place adjacent to a packaging signal. Our theory is in satisfactory agreement with the available experimental data.

SIGNIFICANCE The rate and efficiency of the encapsulation of double-stranded DNA by synthetic coat proteins was recently found to be strongly enhanced by the presence of specifically positioned protein molecules on the DNA that mimic so-called packaging signals. We present a kinetic theory based on the initial stochastic nucleation and subsequent deterministic elongation of the protein coat with the aim of explaining these findings. We find that equidistantly placed nucleation sites that also act as diffusion barriers on the DNA have profound and non-trivial effects, and they can either slow down or speed up encapsulation depending on how fast nucleation is on the timescale of the elongation process. Our findings may contribute to the rational design of linear virus-like particles.

INTRODUCTION

Considering that the filamentous or rod-like morphology is the second most prevalent among all known viruses, we would not expect the packaging of a viral genome by coat proteins in a linear particle to be a non-trivial physics problem (1). Simply relying on the adsorption of coat proteins onto the genome is not sufficient to package it effectively, not even if the adsorbed proteins attract each other, e.g., by the presence of hydrophobic patches, by hydrogen bonding, or by ionic interactions mediated by multivalent ions, which in principle helps to increase the bound fraction of proteins (2,3). The reason for this is that the helical assembly of the proteins around its genome is in essence a one-dimensional process, even though the virus is a three-dimensional object. The helical arrangement of coat proteins in linear structures seems to be strongly preferred, presumably to produce (quasi-)equivalent environments to the proteins that themselves are not symmetrical objects (4–6). Quasi-one-dimensional assembly on a DNA (or RNA) template is dominated by entropy and hence prone to form "defects," i.e., naked sites on the template that make the template vulnerable to nucleases (3,7).

In an attempt to rationalize the successful and complete assembly of tobacco mosaic virus (TMV)—probably one of the most studied viruses to date—Kraft and collaborators relied on a so-called zipper model, in which assembly can only start by the binding of a protein at a preferred position on the polynucleotide (8). For TMV, this preferred location on the genome is called the origin of assembly sequence or OAS, but in the context of virology it is referred to as a packaging signal (9). The second crucial ingredient in

Submitted February 24, 2022, and accepted for publication May 24, 2022.
*Correspondence: r.a.j.d.bruijn@tue.nl
Editor: Jason Kahn.
https://doi.org/10.1016/j.bpj.2022.05.040

the model is a form of allostery, where the required conformational switching of the first bound coat protein catalyzes that of subsequently bound proteins, resulting in cooperative assembly (5). In this case, elongation occurs by subsequent binding to already bound proteins, and the genome is packaged in a zipper-like fashion (10).

The zipper model successfully explains how the free energy cost required for binding the first protein (or assembly of proteins) favors mixtures of completely packaged and naked templates over a mixture of partially covered templates, which would seem evolutionarily beneficial (8,10). Notably, the model does not only capture these equilibrium conditions but also captures the kinetics of the whole assembly process. Based on the insights provided by this model, Hernandez-Garcia and collaborators designed a tri-block polypeptide that functions as an artificial coat protein, packages both single- and double-stranded DNA molecules with high efficiency, and protects these against breakdown by nucleases (3). Synthetic virus-like particles made using these designer coat proteins have already been found to be promising candidates for therapeutic applications (11,12).

Crucial is the presence of the middle (assembly) block of the peptide that needs to undergo a conformational change from a disordered to β-sheet or β-roll configuration in order to bind to another protein and provides allosteric control over the assembly process (13,14). The kinetic version of the zipper model also almost quantitatively describes the assembly kinetics of the encapsulation of the DNA by the artificial coat proteins of Hernandez-Garcia and collaborators (10). This lends strong support for allosteric zippering as a generic mechanism for overcoming the detrimental impact of configurational entropy in quasi-one-dimensional self-assembly processes involving molecular units and a template (7).

While the artificial coat protein (referred to as C-S_{10}-B) does not have a specific preference for any nucleotide sequence, it commences the packaging of the DNA at one of the ends unless the DNA is sufficiently long, in which case assembly may start at any point along the chain (3). It seems that the ends act as (unintended) nucleation sites. Translational (or configurational) entropy can only render these ineffective if the template is sufficiently long and the entropy gain by random binding becomes appreciable.

In recent experiments, Calcines-Cruz and collaborators investigated whether introducing CRISPR-Cas12a proteins—which bind on prescribed positions on the DNA and act as barriers for the linear diffusion of C-S_{10}-B along the DNA—enhances packaging efficiency. These proteins were also chemically linked to the aforementioned assembly block (S_{10}) of the artificial coat protein, in which case the barriers would become a bona fide packaging signal (15). Remarkably, attaching both proteins to one or more positions along the DNA makes the assembly not only more efficient but also faster, and more so if the Cas12a proteins can actually bind the artificial coat proteins (via the S_{10} block). The effect becomes stronger with increasing number of bound Cas12a proteins albeit, as it turns out, with diminishing returns.

This is actually somewhat counter-intuitive, because breaking up a long chain into (seemingly) independent shorter portions should make the assembly less efficient, not more efficient, according to the *thermodynamic* zipper model (8,10,16). This, of course, calls into question the validity of the zipper model for the problem at hand. On the other hand, it would presume that thermodynamics strictly applies for the problem in hand, even though it could well be dominated by kinetics rather than thermodynamics. Indeed, it turns out that once assembled, the DNA-protein complexes and even protein-protein complexes that under appropriate conditions form in the absence of DNA are stable against dilution. This implies that they are easier to assemble than to disassemble (17). Incidentally, this is also true for icosahedral viruses (18). In principle, we would need to modify the kinetic zipper model of Kraft et al. (8), and its generalization by Punter et al. (10), to account for the influence of multiple "barriers" or packaging signal mimics, and find out whether the model survives confrontation with the experimental data of Calcines-Cruz and collaborators (15).

Here, we opt for a simpler version that has the same basic premise but which, in contrast to the original model, allows for analytical evaluation and relatively straightforward comparison with experiments. Our simplified model presumes a Poissonian stochastic nucleation process for the (equidistant) nucleation sites we define on a quasi-one-dimensional template. Once binding has taken place, the model assumes elongation to occur deterministically and irreversibly, i.e., with a fixed and constant rate. The nucleation and elongation rates are phenomenological parameters in our model. In reality, the nucleation sites can be distributed heterogeneously along the DNA (or RNA) template or can have different nucleation rates associated with them (19,20). We do not take this into account for the sake of simplicity, as we find that the current simplified description can already explain most of the observed phenomena.

As our model ignores any stochasticity of the elongation process, it ignores the statistical nature of the assembly and disassembly steps during the elongation stage of the growth of the protein coat encapsulating its cargo. This we deem appropriate if the thermodynamic driving force for assembly is sufficiently strong. In addition, we do not explicitly model how precisely proteins attach to the growing end, i.e., either directly from solution or by diffusion of proteins weakly bound to part of the template not yet encapsulated (21).

The model we present might perhaps seem too simple. This is, however, one of its main strengths. We find that our results depend on only a single dimensionless parameter, yet still show many of the essential features of the experiments. Indeed, despite these simplifications, the predictions of our simplified zipper model agree reasonably well with the experimental observations of (15). The model actually makes a number of testable predictions. First, according to the model, the average protein coverage of genetic material is a monotonically increasing function of the number of packaging signals for any nucleation and elongation rate. Second, and perhaps counter-intuitively, this causes either a decrease or an increase in the mean time for complete encapsulation with an increasing number of packaging signals. This depends on the ratio between the elongation rate and the nucleation rate. Third, we find that this mean time has an optimal value for some number of packaging signals, which is proportional the aforementioned ratio of rates.

In the remainder of this paper, we first introduce our nucleation-and-growth model for the self-assembly kinetics and put forward a dimensionless nucleation rate that acts as the sole relevant control variable. Our model predicts the existence of three temporal regimes in the encapsulation process. For early times, template coverage scales quadratically with time. For a sufficiently low or high nucleation rate, an intermediate linear regime emerges, while for the late stages we find the template coverage to approach exponentially the state of complete coverage. Next, we introduce the barriers or assembly signals into our model and illustrate how they influence the assembly kinetics. We show that if the dimensionless nucleation rate is sufficiently high, adding assembly signals increases the encapsulation rate. Finally, we compare our model with the experimental results of (15), and summarize our findings in the conclusions of the paper.

MATERIALS AND METHODS

Nucleation-and-growth model

Let each DNA strand present in the solution act as a quasi-one-dimensional template of (dimensionless) length L that can bind N_{DNA} proteins (Fig. 1 *A*). We assume that the (dimensionless) length that each protein covers is small, so $m = L/N_{\mathrm{DNA}} \ll L$, allowing us to treat the encapsulation process in the continuum limit. This approach is justified for $N_{\mathrm{DNA}} > 100$, below which we should take into account that proteins attach in discrete units. Mirroring the zipper model, we assume that the nucleation of the binding of proteins onto the template occurs at one of the free ends of the DNA template. This approach is known to be valid for DNA templates that are shorter than some length that is arguably set by the binding free energy (8,10).

Following established nucleation theory, we model this nucleation process as a Poisson process with a nucleation rate I (22). Hence, the nucleation site on the template nucleates at time t, i.e., it binds a single protein or cluster of proteins, with a probability density function given by

$$P(t) = Ie^{-It}. \tag{1}$$

We treat the nucleation rate I as a phenomenological parameter that can be obtained from direct comparison of our model with experimental data, and that in principle depends on the experimental conditions such as the protein concentration (10).

After the nucleation stage, proteins attach to the nucleus and the protein coat elongates. Since the critical protein nucleus consists of a single or a few proteins, its direct influence on the fraction of the template that is encapsulated can be neglected if the total number of proteins that fit on the DNA is very much larger (21). We introduce the elongation rate g, presumed to be constant during the complete process. This we justify by noting that we compare our theory with experiments that are conducted at a constant protein concentration by a constant flow of protein solution over an array of DNA strands in a setup where the fluid flow stretches the strands. The progress of their packaging can be visualized (Fig. 1 *B* and *C*) since the length Δ is proportional to the number of bound protein molecules (15).

For simplicity, as already disclosed, we treat the elongation process as if it were purely deterministic, which results in an encapsulated length $l(t_{\mathrm{i}}, t)$ of the DNA template that is a linear function of time,

$$l(t_{\mathrm{i}}, t) = (t - t_{\mathrm{i}})g \tag{2}$$

for $l(t_{\mathrm{i}},t) \le L$. Here, t_{i} is the time that nucleation occurs and $t > t_{\mathrm{i}}$ the time of observation. For a finite template of length L, the template is fully encapsulated for all $t \ge t_{\mathrm{i}} + L/g$, resulting in $l(t_{\mathrm{i}},t) = L$. We define the elongation time $t_{\mathrm{e}} = L/g$ as the time required to encapsulate the whole template if nucleation occurred at one of the free DNA ends. Experimentally, this happens to be the case as it is probably linked to the direction of the flow in the experimental setup (15). We discuss the validity of the (implicit) assumption that the kinetics are irreversible at the end of the paper.

RESULTS

Combining the nucleation and growth stages of the encapsulation process, we define the average template coverage $\langle\theta\rangle(t)$ as

$$\langle\theta\rangle(t) = \frac{\langle l\rangle(t)}{L} = \int_0^t \frac{l(t_{\mathrm{i}}, t)}{L} Ie^{-It_{\mathrm{i}}}\mathrm{d}t_{\mathrm{i}}, \tag{3}$$

where $t_{\mathrm{i}} < t$. Taking the finite length of the DNA into account, this can be straightforwardly shown to yield

$$\langle\theta\rangle(t) = \frac{\langle l\rangle(t)}{L} = \begin{cases} \dfrac{t}{t_{\mathrm{e}}} + \dfrac{1}{\gamma}\left(e^{-\gamma t/t_{\mathrm{e}}} - 1\right), & \text{if } t \le t_{\mathrm{e}}, \\ 1 + \dfrac{1}{\gamma}e^{-\gamma t/t_{\mathrm{e}}}(1 - e^{\gamma}), & \text{if } t > t_{\mathrm{e}}, \end{cases} \tag{4}$$

with the dimensionless nucleation rate $\gamma = It_{\mathrm{e}}$ as the only relevant parameter in our nucleation-and-growth model. We note that γ depends on the template size via the elongation time $t_{\mathrm{e}} = L/g$. Even though only a single nucleation event can occur per template, we can interpret γ as the expected number of nucleation events that could statistically have occurred during a time equal to t_{e}.

While a seemingly simple relation, the predictions of Eq. 4 are far from that. To highlight this, Fig. 2 shows the mean template coverage as function of the scaled time t/t_{e} for selected values of the parameter γ. The sigmoidal shape, more conspicuous for small values of γ, hints at the dynamics typical of a nucleated process, which, of course, is not entirely surprising.

From Fig. 2, we conclude that a number of distinctive growth regimes emerge and these depend on the value of γ. As shown in Fig. 2 *B*, we identify an early-time regime for $t \ll 1 \le t_{\mathrm{e}}$ where the template coverage scales quadratically

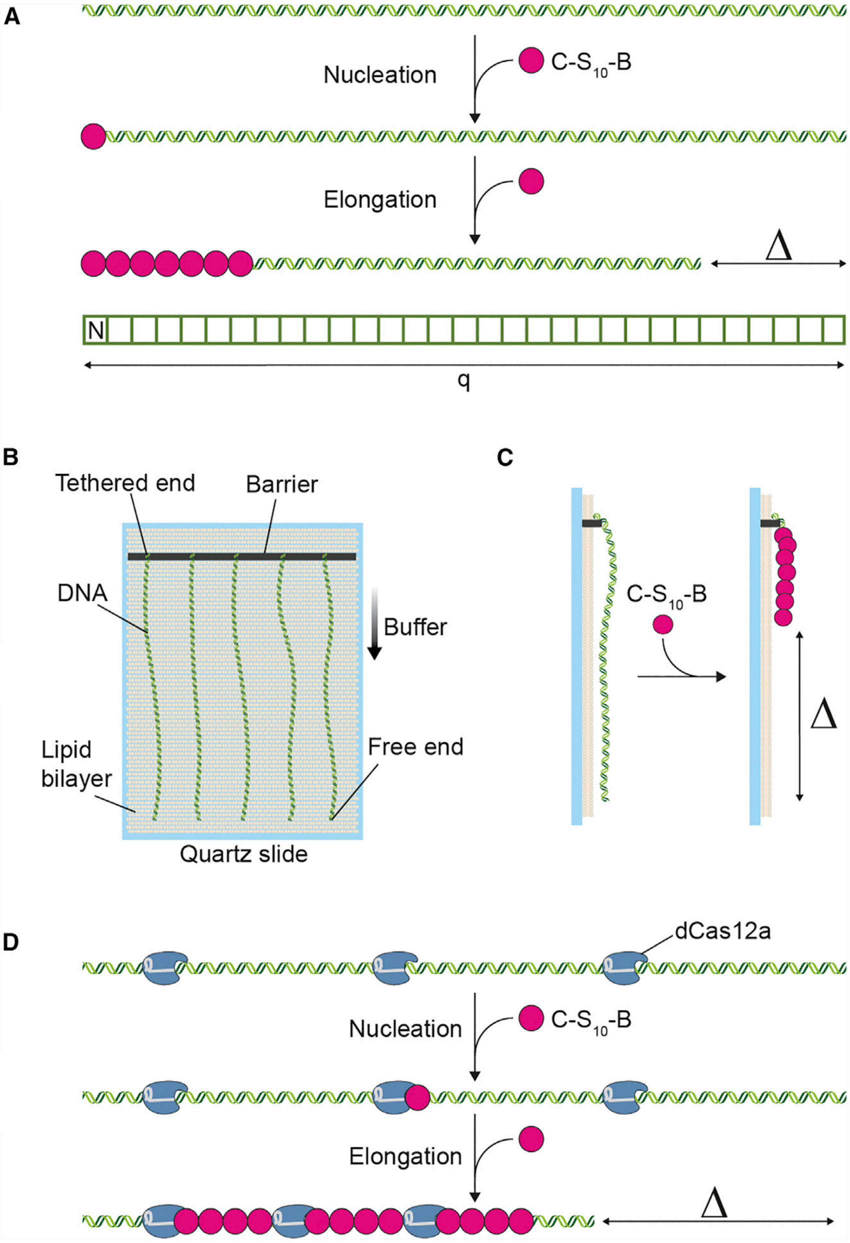

FIGURE 1 Schematic overview of our model. (*A*) The two-step mechanism for encapsulation of DNA by the coat protein is initiated by a stochastic nucleation process that occurs at the free end of the DNA. After this the protein coat can engage in the elongation process, which shortens the genetic material with the length Δ (an observable) by somehow folding it in the complex. In the lattice representation of our model, N represents the nucleation site at a free DNA end and q is the number of binding sites on the DNA. Schematic overview of the experimental setup is shown in (*B*) (top view) and (*C*) (side view). One of the DNA ends is tethered to a barrier and the buffer containing the C-S$_{10}$-B coat proteins flows along the DNA chains. Coat proteins attach from this buffer solution onto the DNA templates. (*D*) The packaging signal mimics (dCas12a) divide the DNA template into smaller independent subtemplates, each of which separately adheres to the nucleation-and-elongation mechanism. See (15) for more details on the experimental setup and results. Reprinted (adapted) with permission from (15). Copyright 2022, American Chemical Society. To see this figure in color, go online.

with time. Indeed, closer inspection of Eq. 4 shows that for this regime

$$\langle\theta\rangle(t) \propto \frac{1}{2}\frac{I}{t_e}t^2, \tag{5}$$

predicting that at early times encapsulation is dictated by the timescale $\tau_1 = \sqrt{t_e/I}$.

A linear intermediate growth regime emerges if γ is either distinctly smaller or larger than unity, corresponding to the situation that the timescales for nucleation and growth are clearly separated. For the case that nucleation is a (relatively) fast process and $\gamma \gg 1$, we find that the linear regime emerges for $It \gg 1$, yet $t \leq t_e$. This we explain from Eq. 4 by noting that in this case the exponential term becomes small and only the linear relation survives, i.e.,

$$\langle\theta\rangle(t) \propto \frac{1}{t_e}\left(t - \frac{1}{I}\right). \tag{6}$$

Hence, we are led to conclude that the relevant timescale must now be $\tau_2 = t_e$. We interpret the $1/I$ term as a lag time, because elongation can only commence after the nucleation event, which on average occurs after a time $1/I$.

For the case that nucleation is a (relatively) slow process and $\gamma \ll 1$ (not shown in Fig. 2 *B*), we find from Eq. 4 that a linear regime emerges for $t > t_e$ as

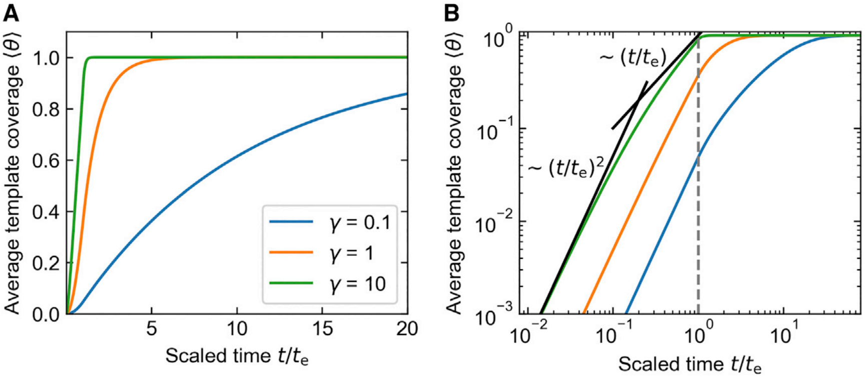

FIGURE 2 Average template coverage as function of the dimensionless time t/t_e, with t_e the elongation time. (*A*) Linear scale. Blue line, $\gamma = 0.1$; orange line, $\gamma = 1$; green line, $\gamma = 10$. (*B*) Log scale. Scaling included as guide for the eye. Gray dashed line separates initial regime ($t/t_\mathrm{e} \leq 1$) from exponential decay regime. To see this figure in color, go online.

$$\langle\theta\rangle(t) \propto It, \tag{7}$$

where we use that to linear order $1 - \exp\gamma \sim - \gamma$. The relevant timescale is $\tau_3 = 1/I$, as nucleation is the only relevant timescale here. This linear regime is characterized by the absence of a lag time.

Finally, for the case that $t/t_\mathrm{e} > 1$, we find this regime to be governed by a simple exponential relaxation to complete template coverage,

$$\frac{1 - \langle\theta\rangle(t)}{1 - \langle\theta\rangle(t_\mathrm{e})} = e^{-(It-\gamma)}, \tag{8}$$

where the relevant timescale is $\tau_4 = 1/I$. It transpires that our seemingly simple kinetic equations produce no fewer than three regimes, governed by four timescales.

These three growth regimes actually only manifest themselves if neither the elongation nor the nucleation process fully dominates the encapsulation kinetics. If this were the case, Eq. 4 reproduces what we would naively expect. For the case that elongation is much faster than nucleation, if $\gamma \to 0$, we find $\langle\theta\rangle = 1 - \exp(-It)$, noting that in Eq. 4 only the solution for $t > t_\mathrm{e}$ survives, that to linear order $1 - \exp\gamma \sim -\gamma$ and that $\gamma/t_\mathrm{e} = I$. This, in fact, is what we would expect, as it is equivalent to the (cumulative) probability that a protein has nucleated at the nucleation site before time t. For the elongation-limited case, where the nucleation process is much faster than the elongation process and $\gamma \to \infty$, Eq. 4 reduces to $\langle\theta\rangle(t) = t/t_\mathrm{e}$ for $0 \leq t \leq t_\mathrm{e}$ and unity for $t > t_\mathrm{e}$. This, obviously, is also in accord with what we expect for a deterministic growth process with a constant growth rate.

While the average template coverage provides relevant information about the encapsulation kinetics, it contains no direct information about the efficiency of complete encapsulation. Since genetic material is known to degrade if left unprotected, we also need to focus attention on the relevant statistics for complete template packaging. We can quantify the efficiency of complete encapsulation by considering the mean waiting time for this to happen. For any template to be completely encapsulated at time t, a nucleation event must have occurred at least a time t_e earlier. Hence, we conclude that the mean waiting time must obey

$$\langle t_\mathrm{w}\rangle = \int_{t_\mathrm{e}}^{\infty} t \times Ie^{-I(t-t_\mathrm{e})}\mathrm{d}t = t_\mathrm{e} + \frac{1}{I}. \tag{9}$$

This is actually a sensible prediction and in line with our earlier analysis, if we indeed interpret $1/I$ as a lag time and realize that after this lag time it takes another time t_e to fully encapsulate the template.

Finally, since the nucleation process is stochastic, we find that the standard deviation σ of the completion time depends only on the quantity I and obeys the simple relation $\sigma = 1/I$. From an experimental point of view, optimal control of the encapsulation process is characterized by both a short mean waiting time and a small standard deviation. Whereas the mean waiting time can be minimized by minimizing the elongation time or maximizing the nucleation rate, the standard deviation can only be made smaller by increasing the nucleation rate.

Packaging signals and packaging signal mimics

Attaching multiple packaging signals, or packaging signal mimics if we refer to protein molecules that mimic them, to the DNA template can have different effects depending on their functionality. The packaging signal mimics we consider bind at predesignated locations on the DNA and act as nucleation-inducing diffusion barriers. We refer to both the inert and the chemically active proteins as packaging signal mimics, as both were found to enhance the encapsulation kinetics, although to a different extent. Arguably, the proteins that act as packaging signal mimics only influence the assembly kinetics locally. Consequently, this suggests that these packaging signal mimics effectively subdivide the polynucleotide template into smaller, independent DNA templates ("subtemplates"). In the context of virology, a packaging signal usually only refers to sequences on the genetic material that have a higher binding affinity for the coat protein (8,9,23). Consequently, these would not subdivide the DNA template in independent subtemplates. We do not consider this important but subtle difference in the remainder of this paper, and we will not distinguish between the terms “packaging signal mimics” and “packaging signals.”

In principle, no distinction needs to exist between either side of a packaging signal, so growth can proceed in both directions. Bidirectional growth may cause changes in the geometry of encapsulated virus-like particles (VLPs) from linear to bend or star-like shapes, at least when multiple packaging signals (“origins of assembly”) are inserted in the RNA of TMV (23–25). We note, however, that the experiments of Calcines-Cruz et al. (15) were conducted in a flow cell, in which the flow direction breaks this symmetry (see Fig. 1 *B*). In practice, this means that the proteins move unidirectionally and that they tend to accumulate near one side of the packaging signal. This accumulation of weakly bound coat proteins can trigger the nucleation of a strongly bound state, akin to what happens at the free ends of the DNA (3). The other side of the packaging signal then acts as a barrier that halts the elongation process which progresses in the direction opposite to the flow direction (15).

To study the effect of the packaging signals on the encapsulation kinetics, we introduce in our model $n - 1$ additional, equidistantly placed packaging signals on the template. Thus, we have in total n nucleation sites if we include the one on one of the free ends. We presume that the nucleation rates at the preferred free end of the template and those at the packaging signals are the same, which, in effect, results in all n portions of the template being identical. Hence, in our model the free end of the DNA acts as a packaging signal. However, as already advertised in the introduction, the experimental findings show that specifically functionalized packaging proteins enhance the overall encapsulation rate, i.e., suggesting different nucleation rates for the end site and the others (15). We do not take this into account in our paper, but a discussion on the effect is presented in the supporting material. We shall return to this issue at a later stage in this article.

Given these assumptions, we find that the average coverage of any subtemplate adheres to Eq. 4, at least if we replace the template length L by that of the subtemplates L/n. Since all subtemplates are identical and independent, we conclude that the template coverage of the full template must be given by

$$\langle l \rangle(t) = n \times [\langle l \rangle(t)]_{L = L/n}, \tag{10}$$

yielding

$$\langle \theta \rangle(t) = \frac{\langle l \rangle(t)}{L} = \begin{cases} n\dfrac{t}{t_\mathrm{e}} + \dfrac{n}{\gamma}\left(e^{-\gamma t/t_\mathrm{e}} - 1\right), & \text{if } t \leq t_\mathrm{e}/n, \\ 1 + \dfrac{n}{\gamma} e^{-\gamma t/t_\mathrm{e}}\left(1 - e^{\frac{\gamma}{n}}\right), & \text{if } t > t_\mathrm{e}/n. \end{cases} \tag{11}$$

The effect of packaging signals on the average template coverage turns out to be equivalent to renormalizing both γ and t_e by $1/n$, as is evident from comparing Eqs. 4 and 11. This might suggest that an increase in the number of packaging signals should always produce a smaller (equal-time) template coverage, which would be in agreement with models and experiments of linear viruses, but contrasts with models and experiments on the effect of packaging signals on (icosahedral) viruses (see, for example, (9,19,25)). This, however, is not quite true, as we show in Fig. 3 for $\gamma = 10$ (note that the $n = 1$ case in Fig. 3 corresponds to the green curve in Fig. 2). From the figure we conclude that increasing the number of nucleation sites n actually makes the template saturate to full coverage faster, not more slowly.

Mathematically, this becomes evident if we rewrite Eq. 11 in terms of varying combinations of the time t and two timescales that emerge naturally, namely I^{-1} and t_e/n. The last timescale decreases with increasing n, arguably making the assembly faster if t_e is fixed by the length of the DNA only. A more intuitive explanation relies on realizing that it is the result of the competition between three effects. First, the overall nucleation rate on the template

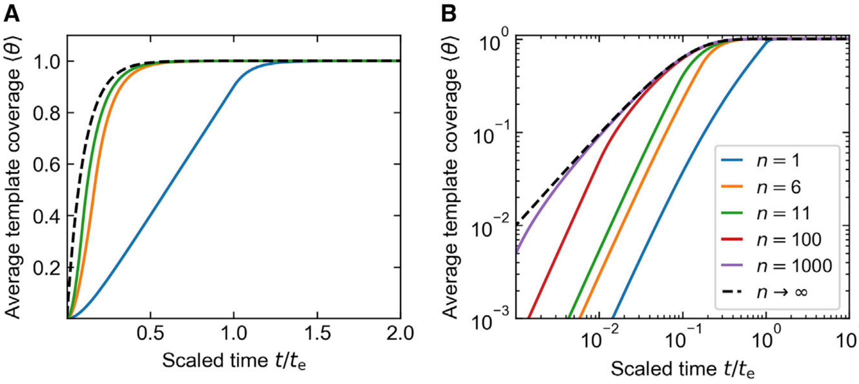

FIGURE 3 Average template coverage as function of the scaled time. Here, t_e is the elongation time, the dimensionless nucleation rate $\gamma = 10$, and the effect of packaging signals is included for 0 ($n = 1$), 5 ($n = 6$), 10 ($n = 11$), and $n \rightarrow \infty$ packaging signals in both graphs. Additionally, the cases for $n = 100$ (*red*) and $n = 1000$ are included in (*B*). (*A*) Linear scale. (*B*) Log scale. To see this figure in color, go online.

increases due to the presence of additional nucleation sites, which increases the average template coverage. Second, due to the additional nucleation sites, many different pathways now exist to achieve the same template coverage, which makes it more likely to occur. Third, the packaging signals block large parts of the template for elongation, which slows the whole process down. It turns out that the first two effects are always dominant, for Eq. 11 tells us that $\partial\langle\theta\rangle(t)/\partial n \geq 0$ for all values of γ, n, and t. We find that for large n there is a limit to the increase in average template coverage: for $n \rightarrow \infty$ we have $\langle\theta\rangle(t) = 1 - \exp(-It)$, which can be obtained from Eq. 11 in a similar fashion as we did before taking the limit $\gamma \rightarrow 0$ for the case that $n = 1$. This limit corresponds to the case where all sites on the DNA template are encapsulated by the nucleation process only, reducing the process to a (non-cooperative) Langmuir-type adsorption process albeit without any detachment reactions, which follows a simple exponential decay, as would any first-order reaction kinetics (21,26).

Figs. 2 and 3 exhibit the same regimes. However, the relevant timescales do turn out to vary with the number of packaging signals n; for the early-time regime the relevant timescale is $\sqrt{t_e/nI}$ and that for the intermediate linear regime we find either t_e/n if $\gamma/n \gg 1$ or I^{-1} if $\gamma/n \ll 1$. The timescale for the late-time exponential decay is not affected by the number of packaging signals. The $n = 100$ and $n = 1000$ curves approach the $n \rightarrow \infty$ limit very slowly and a quadratic regime persists only for $t/t_e < 1/n$, as can be expected from Eq. 11. All of this suggests that while the whole encapsulation process appears to be faster, this is mainly true for the short and intermediate times.

The number of packaging signals turns out to have a significant and non-trivial impact on the time required for complete encapsulation. As before, we measure the efficiency of complete encapsulation of the template using the mean waiting time for full coverage. The probability that the template is fully encapsulated at time t can be decomposed in the probability that $n - 1$ subtemplates have been encapsulated by time t, and the last subtemplate encapsulates at time t. Note that for any subtemplate to be fully encapsulated at time t, a nucleation event must occur at least a time t_e/n earlier.

Accounting for the n-fold degeneracy, which originates from the n identical pathways that now exist to reach full coverage, we find

$$\langle t_w\rangle = n\int_{t_e/n}^{\infty} t \times \left[1 - e^{-I\left(t-\frac{t_e}{n}\right)}\right]^{n-1} \times \left(Ie^{-I\left(t-\frac{t_e}{n}\right)}\right)\mathrm{d}t, \tag{12}$$

where the term between the square brackets represents the probability that a certain subtemplate is fully encapsulated at time t, and the term in round brackets denotes the probability that one of the templates encapsulates at time t. Carrying out the integral, we find it to reduce to

$$\langle t_w\rangle(n) = \frac{t_e}{n} + \frac{1}{I}H_n, \tag{13}$$

where $H_n = \sum_{j=1}^{n} 1/j$ is the so-called n^{th} harmonic number (27). We return to a discussion of the properties of H_n below. We associate $I^{-1}H_n$ with a lag time in a similar fashion as we did for the case $n = 1$ in the preceding section. It appears that this quantity also depends on the number of packaging signals on the template.

Finally, we find the standard deviation to remain inversely proportional to the nucleation rate, but now depends non-trivially on the number of packaging signals as

$$\sigma = \frac{1}{I}\sqrt{\sum_{k=1}^{n} k^{-2}}. \tag{14}$$

Since $\sum_{k=1}^{\infty} k^{-2} = \pi^2/6$, we conclude that $1 \leq \sqrt{\sum_{k=1}^{n} k^{-2}} \leq \pi/\sqrt{6} \approx 1.28$ (28). While the encapsulation process should logically become more stochastic upon the addition of nucleation sites, this is not mirrored by a concomitant increase in the standard deviation, which turns out to be essentially independent of the number of packaging signals n.

Surprisingly, while the mean template coverage always increases with increasing number of packaging signals, this is not the case for the mean waiting time, as can be seen in Fig. 4. In the figure, we present the waiting time-scaled to that of a single nucleation site, $\langle t_w\rangle(n)/\langle t_w\rangle(1)$,

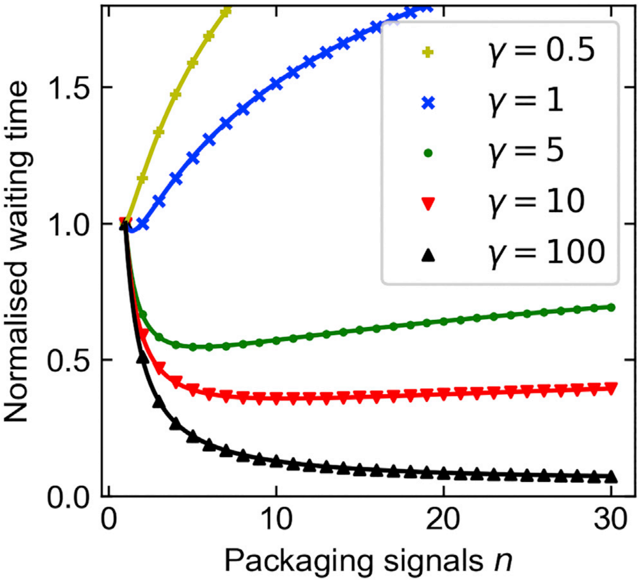

FIGURE 4 The normalized waiting time $t_w(n)/t_w(1)$ as function of the number of packaging signals for selected values $\gamma = 0.5$ (*yellow*, *plusses*), $\gamma = 1$ (*blue*, *crosses*), $\gamma = 5$ (*green*, *dots*), $\gamma = 10$ (*red*, *triangle-down*), and $\gamma = 100$ (*black*, *triangle-up*). Lines are the analytical continuation of Eq. 13. To see this figure in color, go online.

as a function of the number of nucleation sites n, for a range of values of the dimensionless nucleation rate γ. If this ratio attains a value smaller than unity, this indicates that the encapsulation process is more efficient and larger than unity if it is less efficient.

It transpires that only for $\gamma > 1$ the average encapsulation time decreases with increasing number of packaging signals. While this might seem inconsistent with our earlier observation that the average template coverage always increases with increasing n, this in fact is not so. To explain this, it is instructive to first focus on the case $\gamma = It_e < 1$, in which fewer than one nucleation event can occur in the relevant elongation timescale t_e. Adding an additional packaging signal then effectively halts the elongation process, and the encapsulation can only continue if another nucleation event occurs. Hence, partially covered templates have a larger lifetime but are easier to form as the additional packaging signals induce additional routes for partial encapsulation. Overall, this means that (on average) a larger fraction of the templates is covered, but mostly because a larger number of templates are partially encapsulated and fewer templates are completely encapsulated. Consequently, for $\gamma < 1$ the complete encapsulation is nucleation limited.

Now, for $\gamma > 1$, multiple nucleation events can occur within the elongation time t_e for the complete template. Adding packaging signals can then induce nucleation before they stall the overall elongation process, thereby speeding up the encapsulation process. In this case, partially encapsulated templates have a shorter lifetime than is the case without packaging signals, and thus the mean time for complete encapsulation decreases. In other words, for $\gamma > 1$ the encapsulation process must be elongation limited, not nucleation limited.

As is evident from Fig. 4, we find that for a given dimensionless nucleation rate $\gamma > 1$ the encapsulation efficiency quickly decreases to an optimal value, after which it slowly increases again. We can actually determine this optimal value from Eq. 13 by using an asymptotic expansion of the harmonic number H_n (27).

$$H_n \approx \ln n + \gamma_E + \frac{1}{2n} + \cdots, \tag{15}$$

with $\gamma_E \approx 0.577$ the Euler-Mascheroni constant. We now treat n as a continuous and not as a discrete variable, and optimize Eq. 13 to obtain

$$n_{opt} \approx \frac{1}{2} + It_e = \frac{1}{2} + \gamma \tag{16}$$

for the optimal value of the number of nucleation sites n_{opt}.

Neglecting the constant value of $1/2$, this results in an optimum encapsulation efficiency if the nucleation lag time for a single nucleation site is approximately equal to the elongation time of a subtemplate, i.e., $1/I \approx t_e/n$. It has to be noted, however, 1) that the decrease in the waiting time with increasing value of n is significant only for the first few packaging signals, and 2) that the optimum is very shallow, as Fig. 4 clearly shows. Indeed, for, say, $\gamma = 10$, any value between approximately 6 and (at least) 30 packaging signals yields a normalized waiting time that is essentially indistinguishable from its true optimum $n = 10$; see Eq. 13.

From Eqs. 13, 15 and 16, we find the normalized mean waiting time corresponding to the optimal value of the number of nucleation sites to obey the approximate relation

$$\min\left(\frac{\langle t_w\rangle(n)}{\langle t_w\rangle(n = 1)}\right) \approx \frac{\langle t_w\rangle(n = \gamma)}{\langle t_w\rangle(n = 1)} \propto \frac{C + \ln\gamma}{\gamma + 1}, \tag{17}$$

with $C = 1 + \gamma_E \approx 1.577$ a constant, where we neglect a term of the order γ^{-2}. Since the optimum in the normalized waiting time is very shallow, discreteness effects are small, and treating n as a continuous variable does not lead to a significant error. The combination of the optimum value for n, and the corresponding normalized waiting time, suggests a route to optimize the encapsulation process experimentally. Obviously, this presumes that we can control the dimensionless nucleation rate γ. We discuss the implications of this result for the role that packaging signals may have in the assembly of naturally occurring viruses at the end of this article.

Generalizing these results to predictions of times associated with partial template coverage turns out to be highly non-trivial on account of the possibility of elongation processes happening simultaneously on the different subtemplates. This makes the calculation increasingly complex the larger the number of packaging signals on the template. Such a calculation would, however, still be highly interesting, as it is experimentally observable and produces information on intermediates. Actually, partial coverage times might be easier to measure if the experimentally accessible time window for measurements is, for whatever reason, too short to measure complete encapsulation.

Obviously, for most practical applications we are only interested in fully encapsulated DNA templates. Still, we can ask ourselves the question as to the extent of correlation between the mean waiting time for partial and that for complete encapsulation. To deal with this problem, we supplement our analytical calculations with kinetic Monte Carlo (kMC) simulations, from which the mean waiting time for partial template coverage can be straightforwardly extracted. This also allows us to verify our analytical predictions and investigate how stochasticity expresses itself in this problem.

Kinetic Monte Carlo simulations

We set up our kMC simulations (detailed below) in such a way that they mimic the basic ingredients of our analytical theory, in which case we can simply regard our kMC results as a (near-)exact representation of our analytical model. We let the DNA template be represented by a one-dimensional lattice with a fixed number of $N_{\mathrm{DNA}} = 10^4$ lattice sites. This is slightly larger than the expected number of proteins ($N_{\mathrm{DNA}} \approx 8083$) required to fully encapsulate a DNA strand with a length of 48.5 kbp, which we determine by considering that the complete capsid is charge neutral (3,10,15). We note that our results do not depend on the value for N_{DNA}, as long as it is sufficiently large ($N_{\mathrm{DNA}} > 100$). Correct kinetic parameters for different N_{DNA} can be obtained by simply rescaling the number of proteins on the DNA.

In our simulations, we introduce the nucleation reaction with a rate k_{nuc} and an elongation reaction with rate k_{e}. Both reactions are stochastic, yet we do not allow for unbinding of proteins in the elongation process to keep as close to the model as possible. Thus, the overall kinetics remain based on the same model description, and the nucleation reactions can commence only on predesignated nucleation sites that act as the packaging signals. We invoke the well-known Gillespie algorithm (29–31), which can yield exact trajectories for stochastic reactions. In applying the algorithm, we calculate two random numbers at each kMC step to perform a single reaction. The first random number we use to select the reaction with a probability proportional to the reaction rate, and the second we use to select the time step for the kMC step, sampled from a Poisson distribution. These two steps we repeat until the DNA template is fully encapsulated. We refer the reader to the review paper of Gillespie (31) and the supporting material for more information.

We are able to map our simulations onto the parameter space of our theory by defining $t_{\mathrm{e}}^{\mathrm{kMC}} = N_{\mathrm{DNA}}/k_{\mathrm{e}}$ and $\gamma^{\mathrm{kMC}} = k_{\mathrm{nuc}} t_{\mathrm{e}}^{\mathrm{kMC}} = N_{\mathrm{DNA}} k_{\mathrm{nuc}}/k_{\mathrm{e}}$. This implies that the kinetic parameters of our kMC simulations and our model parameters must be related via $k_{\mathrm{e}} = N_{\mathrm{DNA}}/t_{\mathrm{e}}$ and $k_{\mathrm{nuc}} = I = \gamma/t_{\mathrm{e}}$. To obtain good statistics, all averages are calculated using 20,000 independent trajectories. As we show in the supporting material, by setting $t_{\mathrm{e}}^{\mathrm{kMC}} = t_{\mathrm{e}}$ and $\gamma^{\mathrm{kMC}} = \gamma$, the kMC results for the mean template coverage (Fig. S1) and mean waiting time for complete encapsulation (Fig. S2) are, for all intents and purposes, indistinguishable from the predictions of our analytical theory.

Consequently, we conclude that the stochastic nature of the elongation rate in the simulations does not appreciably influence the average observables, and that we can indeed treat the averages obtained from the kMC simulation as a (near-)exact representation of our analytical model. This can be understood intuitively if we interpret the standard deviation for a single reaction, which is inversely proportional to the reaction rate, as a measure for the stochasticity. Based on Eq. 14, we argue that the standard deviation associated with many of such reactions is essentially equal to that of a single reaction. Hence, we expect that the stochastic nature of the elongation rate becomes apparent only if k_{el} is of equal or larger order than k_{nuc}, which in our case corresponds to $\gamma^{\mathrm{kMC}} \gtrsim N_{\mathrm{DNA}} = 10^4$.

Our kMC results for the mean waiting time $\langle t_w \rangle$ as a function of the partial *targeted coverage* $\theta = N/N_{\mathrm{DNA}}$ are summarized in Fig. 5. We present the $\langle t_w \rangle$, scaled to the elongation time t_{e}, as a function of the targeted coverage θ for the selected values of $\gamma = 0.1$, 1, and 10. For small values of γ, the curves turn out to be step-like. This happens if the time between consecutive nucleation events is larger

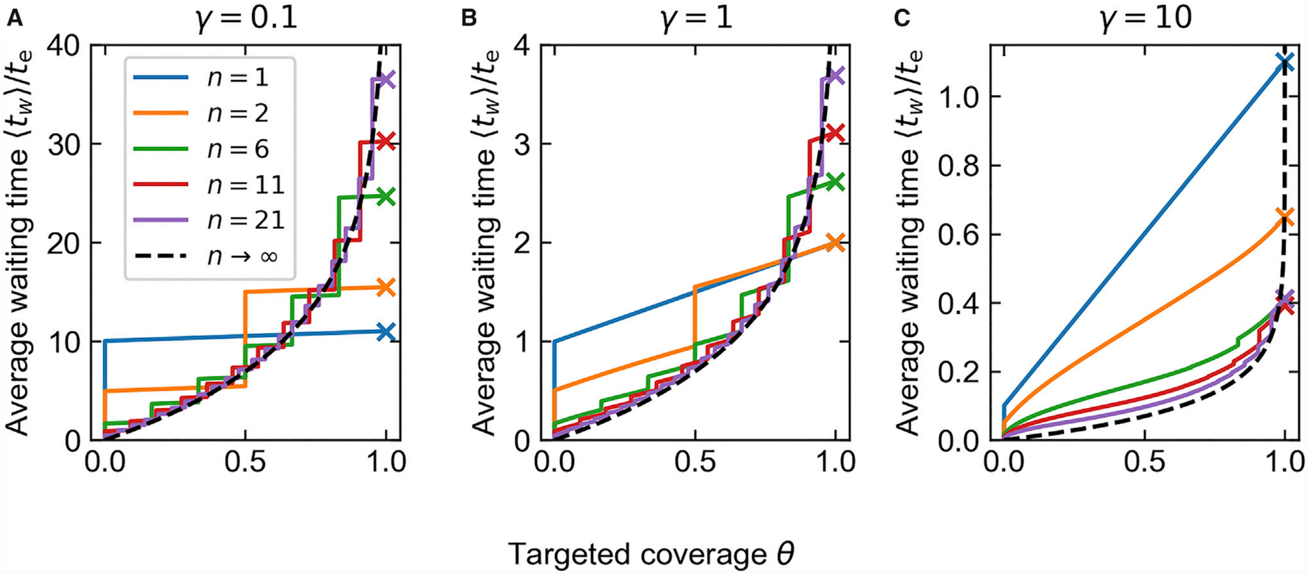

FIGURE 5 The mean waiting time to encapsulate a given fraction of the DNA template as a function of the target coverage θ (not to be mistaken for the average template coverage $\langle \theta \rangle$). We include kMC results for $n = 1$ (*blue*), $n = 2$ (*orange*), $n = 6$ (*green*), $n = 11$ (*red*), and $n = 21$ (*purple*), and the exact asymptotic result $n \to \infty$ (*black*, *dashed*). Crosses represent our analytical result and are color matched to the simulation curves. (*A*) $\gamma = 0.1$; (*B*) $\gamma = 1$; (*C*) $\gamma = 10$. To see this figure in color, go online.

than the time required to encapsulate a subtemplate. We discuss this in more detail below.

For the case of $n = 1$ only a single jump occurs and the mean waiting time increases linearly afterward with a slope of unity, which holds for all finite values of γ. The standard deviations are not shown for the sake of clarity, but, considering that the kMC results are essentially equal to our analytical theory, we argue that the dominant contribution is from the nucleation processes only. Hence, they must be proportional to $1/\gamma$.

In Fig. 5 we have indicated our analytical prediction for complete coverage, Eq. 13, with crosses, all of which are within the 95% confidence intervals of our simulations (not shown). The prediction for the asymptotic limit for $n \rightarrow \infty$, i.e., for the case that the template is encapsulated by the nucleation process only, is equal to $\langle t_{\mathrm{w}} \rangle / t_{\mathrm{e}} = -1/\gamma \log(1-\theta)$ and is shown in Fig. 5 too (see also the supporting material).

The first relevant observation we make from Fig. 5 is that for small values of γ all curves appear to be discontinuous, i.e., characterized by a sequence of steps. The reason for this is that, if nucleation events are rare, the encapsulation of a subtemplate is finished before a next nucleation event happens. Only the "jump" associated with the first nucleation event we expect to be infinitely sharp, because now the template changes from empty to not empty. The remaining "jumps" occur as a sequence of smaller steps, which are too small to be observed in Fig. 5. In the continuum limit, this would translate to the edges of the jumps to be slightly rounded and the slopes to be large but finite. For large values of γ, this step-like behavior only appears to be relevant for large n and then only for large values of θ.

It turns out that we can understand the presence of these jumps if we focus attention on the mean waiting time for the nucleation events, which simplifies the description considerably. As shown in the supporting material, the mean waiting time for the k^{th} nucleation event, given that n packaging signals are present on the DNA template, obeys the relation

$$\frac{\langle t_{\mathrm{w}} \rangle}{t_{\mathrm{e}}}(k,n) = \frac{1}{\gamma}\left(H_n - H_{n-k}\right), \tag{18}$$

with H_x again the harmonic number. If the time between two consecutive nucleation events is sufficiently large, the encapsulation of the template may stall. This happens if the (mean) time between two consecutive nucleation events is larger than the elongation time for a subtemplate, t_{e}/n. Hence, we expect that a jump in the mean waiting time is present at the k^{th} nucleation event only if

$$k > n\left(1 - \frac{1}{\gamma}\right). \tag{19}$$

This shows that for any $\gamma \leq 1$ jumps are always present, whereas for $\gamma > 1$ they are only present if the template coverage is sufficiently large. Comparison with Fig. 5 shows that this condition is only approximately valid for $\gamma > 1$, likely because several elongation processes can occur concurrently.

The second relevant observation we make from Fig. 5 is that the asymptotic limit $n \rightarrow \infty$ appears to produce a limiting lower boundary for the average waiting time for most values of the template coverage, i.e., if we exclude the presence of the jumps. As is perhaps to be expected, we find that for sufficiently small partial template coverage, adding packaging signals always decreases the mean waiting time compared with the case where no packaging signals are present. Only for large template coverage does this result in an increase in the mean waiting time upon adding further nucleation sites.

We can, broadly, distinguish the crossover of these behaviors by considering the intersection between the $n = 1$ and $n \rightarrow \infty$ curves, the former of which is given by $\langle t_{\mathrm{w}} \rangle / t_e = 1/\gamma + \theta$, which we obtain from Eq. 9 upon replacing the encapsulation time for the complete template t_{e} with that of a partial template θt_{e}. The intersection point $\theta = N/N_{\mathrm{DNA}}$ is defined by $1/\gamma + \theta = -1/\gamma \log(1-\theta)$. The solution of this equation is $\theta = 1 + 1/\gamma W_0(-\gamma e^{-1-\gamma})$, with $W_0(z)$ the principal branch of the Lambert W-function (28). From this we find that the limiting value of this intersection is given by $\theta = N/N_{\mathrm{DNA}} = 1 - 1/e \approx 0.63$ for $\gamma \rightarrow 0$ and by $\theta = 1$ for $\gamma \rightarrow \infty$. This suggests that adding a sufficient number of packaging signals increases the encapsulation efficiency for any value of γ if $\theta < 1 - 1/e$, although this is not necessarily accurate for small values of n due to the presence of the jumps shown in Fig. 5. Surprisingly, this means that the mean time to achieve, say, one-half encapsulation cannot be used to quantify the effect of packaging signals on the relative efficiency to completely encapsulate the DNA template (15).

Still, a prediction for the mean waiting time for one-half encapsulation would be interesting, if only because this observable has been determined experimentally (15). Lacking an analytical expression that takes the number of packaging sites into account, we can make a prediction of the shortest possible mean waiting time to encapsulate one-half of the DNA template. In this case, we deduce from our kMC results that the shortest mean waiting time corresponds the $n \rightarrow \infty$ limit. Hence, we obtain

$$\frac{\left\langle t_{\frac{1}{2}} \right\rangle (n \rightarrow \infty)}{\left\langle t_{\frac{1}{2}} \right\rangle (n = 1)} = \frac{\ln 2}{\gamma/2 + 1}, \tag{20}$$

and $\langle t_{\frac{1}{2}} \rangle (n) / \langle t_{\frac{1}{2}} \rangle (n = 1)$ must be larger than this value for any finite n. The merit of this estimate is that it depends on γ in a simple manner, providing us with a relatively straightforward and direct method to determine a lower bound on the model parameter γ from experiments, as we will show at the end of the next section.

At this point, the question arises as to how reasonable and accurate our model is in describing actual experimental results. We answer this question by comparing our theoretical and kMC simulation results with the experiments by Calcines-Cruz et al. (15) in the following section.

Comparison with experiments

Ideally, a comparison with experiments allows us not only to validate our model but also to determine values of the model parameters γ and t_e. As we shall see next, we find that our model agrees reasonably well with the available experimental data, albeit that the values of the model parameters that we extract by curve fitting do not necessarily reflect our perhaps somewhat naive expectations. Regardless, considering the simplicity of our model, we find the agreement between theory and experiment encouraging.

As already mentioned, the experiments of Calcines-Cruz and collaborators (15) are conducted in a flow cell. This results in the DNA templates being stretched and supplied by a solution with a constant concentration of C-S_{10}-B coat proteins. Proteins that attach to the templates cause these templates to reduce in length, which are believed to be fully coated when the length of the DNA strands reduces to about one-third of their initial length if no packaging signals are being used, or one-fourth if packaging signals are present on the DNA. Hence, the apparent length of the DNA can be tracked as a function of time and can be converted to a time evolution of the fraction of the template encapsulated. For each measurement 30 min of observation time is taken, irrespective of whether the DNA templates are fully or only partially encapsulated in that amount of time.

We compare our model with several types of measurement by Calcines-Cruz et al. (15) who obtain the template coverage, averaged over 25 templates, for three cases: 1) undecorated DNA templates, i.e., without additional packaging signals, in contact with a C-S_{10}-B coat-protein solution at a concentration ranging from 10 to 300 nM; 2) for a fixed C-S_{10}-B concentration of 25 nM, where either 5 or 10 "bare" dCas12a proteins are aimed to be attached at specific, equidistant positions on the DNA template; or 3) for a fixed C-S_{10}-B concentration of 10 nM, where five functionalized dCas12a proteins can be attached to the DNA templates.

For the second case, the bare dCas12a-proteins are believed to act only as diffusion barriers on the DNA template, while in the third case, the functionalized dCas12a protein binds specifically to the silk motifs of the engineered coat proteins. We will not attempt to curve fit the data on the latter type, as the assumption of equal binding strength must be strongly violated: the ends are likely to bind much less strongly than the packaging signals. We return to this below.

In the remainder of this section we first compare our model with the experimental results for the mean template coverage for undecorated DNA templates. Second, we compare our model with the cases that 5 or 10 dCas12a proteins can at least theoretically be bound to the DNAs, focusing on both the mean template coverage and the mean waiting time for one-half encapsulation.

In Fig. 6 we present the experimental data for the case without additional packaging signals for six different protein concentrations together with the confidence intervals. It should, in principle, be possible to obtain our model parameters, the elongation time t_e and nucleation rate I, by a curve fit using, e.g., a non-linear least-squares method (32). Unfortunately, we find this procedure to be highly sensitive to the initial parameter estimates, and it does not necessarily converge unless a very good estimate of the (unknown) model parameters is supplied. Such a high sensitivity for the estimate of an initial parameter is not uncommon in non-linear least-squares methods (32).

To work around this, we opt for a different method to find the (in some sense) best estimates for our model parameters. We first define the global residual $R(t_e, I)$ that quantifies how well our theoretical model compares with the experimental results. Since the standard errors are not constant, we use a weighted sum of squared residuals $R = \sum_{j=1}^{N_{\mathrm{data}}} \sigma_j^{-2} (\langle\theta\rangle_{\mathrm{exp.}}(t_j) - \langle\theta\rangle_{\mathrm{model}}(t_j, t_e, I))^2$, where N_{data} is the total number of data points, $\langle\theta\rangle_{\mathrm{exp.}}(t_j)$ represents the experimental outcome at time t_j, $\langle\theta\rangle_{\mathrm{model}}(t_j, t_e, I)$ is our model value at this time, and σ_j is the standard error at time t_j (32). We discard the first two data points that have a very small standard deviation in order to avoid that these points weigh overly heavily in our curve fitting (see the supporting material for a discussion). Next, we discretize the parameter space spanned by t_e and I and determine the global residual for a set of t_e and I values.

We limit our region to $0 < t_e \leq 300$ (min) and $0 < I \leq 33$ (min^{-1}), and find this to be sufficiently large. (Here, "min" stands for "minutes" in time.) Since our residual R is the same as the one commonly used in weighted least-squares methods, we expect that the values for t_e and I corresponding the smallest value of R are now the same as those that would have been obtained by a properly converged weighted least-squares method (32). This is obviously only true if the global minimum for R is actually within the region that we investigate. Contour plots of the residuals as function of $I(= \gamma / t_e)$ and t_e can be found in the supporting material.

From these contour plots we typically find that there are two types of minimum: for small values for t_e and I we find a relatively deep and well-defined minimum, whereas for larger values of t_e and I a large region with a nearly identical residual R is observed. The presence of these plateau-like local minima might explain why a direct curve fit is very sensitive to the initial parameter estimates. Nevertheless, in all cases we can distinguish the global minimum for the residual R, and we set the model parameters I and t_e to the values corresponding to this smallest residual.

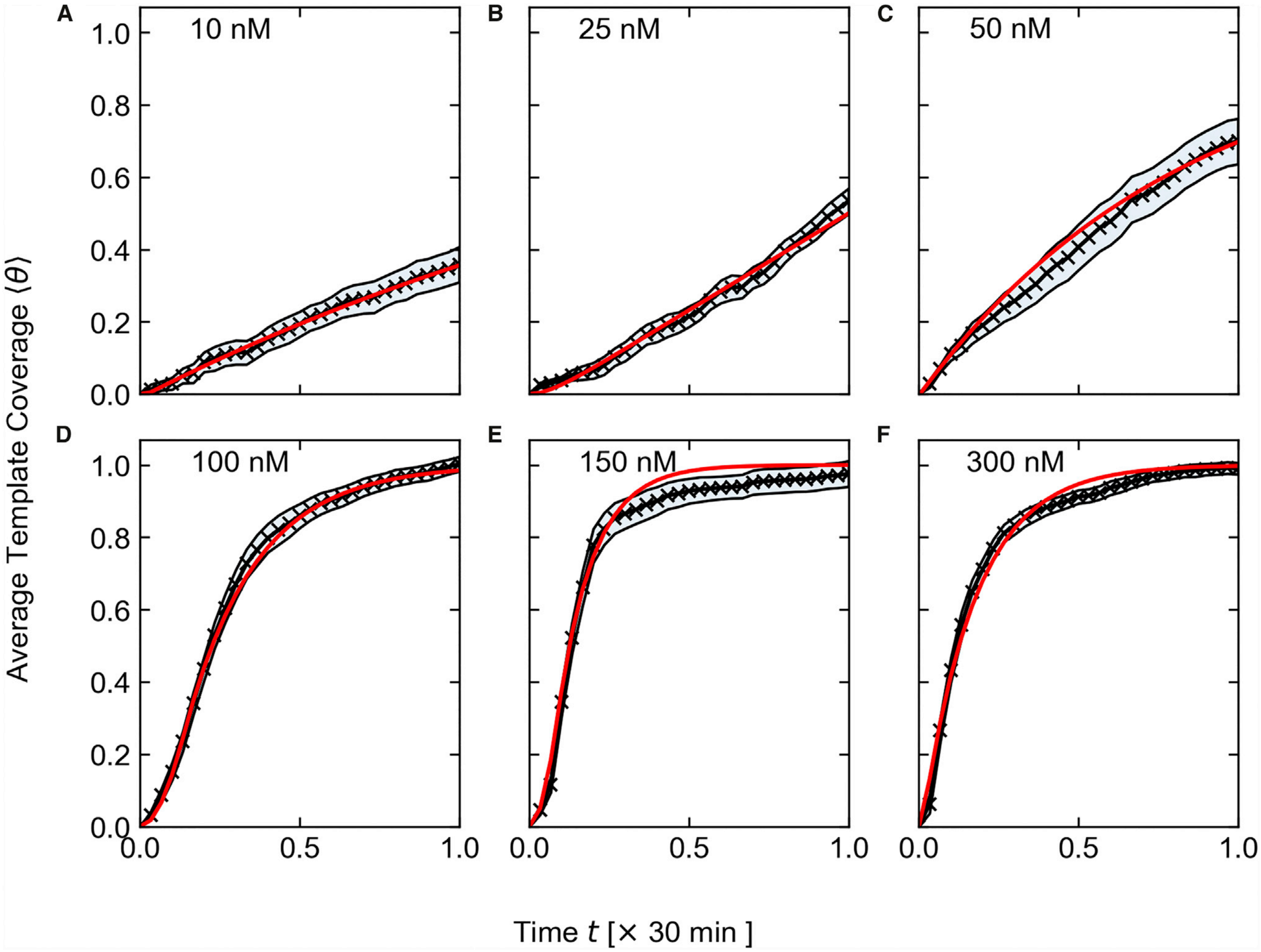

FIGURE 6 Black (*crosses*): experimentally obtained average template coverage $\langle\theta\rangle$ as a function of the time t in units of 30 min from 25 DNA templates (15). The shaded area represents the 95% confidence interval (15). Red (*drawn line*): model fit to the average template coverage. Different curves show results for different concentrations of C-S_{10}-B proteins of (*A*) 10 nM, (*B*) 25 nM, (*C*) 50 nM, (*D*) 100 nM, (*E*) 150 nM, and (*F*) 300 nM. To see this figure in color, go online.

As shown in Fig. 6, the overall agreement between theory and experiments is good, although the theoretical prediction for $\langle\theta\rangle$ is not always within the 95% confidence intervals for all times, in particular for the high concentrations. Our curve fits for low C-S_{10}-B concentrations of 10, 25, and 50 nM are significantly less reliable considering that only partial curves are available, hence the fit parameters should be interpreted with caution. According to the model, the relevant timescale for early times is equal to $\sqrt{t_e/I}$, indicating that we cannot determine t_e and I independently if the experimental data set is limited to the early stages of the assembly process. This is also mirrored by the contour plot for the residual in Fig. S4, where we find the residual for the global minimum to be only slightly smaller than those of the other minima. All sets of parameters are presented in Table 1, including the normalized square deviation V, with $V = 0$ representing a perfect fit. We define the square deviation as $V = \sum_{j=1}^{N_{\text{data}}} (\langle\theta\rangle_{\text{exp.}}(t_j) - \langle\theta\rangle_{\text{model}}(t_j, t_e, I))^2 / \langle\theta\rangle_{\text{exp.}}(t_j)^2$, which is scaled to the mean template coverage instead of the standard deviation and acts as a measure for the goodness of the fit.

In the fitting, we obviously presume that the assumption of irreversible kinetics is valid for all concentrations. Based on the equilibrium zipper model, however, reversible effects might become relevant at the lowest concentrations. Earlier studies put the binding free energy for elongation at about -17 to -18 $k_B T$, resulting in a dissociation constant for elongation $K_D \approx 10^{-8}$M (10). Incidentally, this is identical for the binding of coat protein of TMV to its RNA (2). Our treatment of the kinetics as irreversible should therefore be (approximately) valid for a C-S_{10}-B concentration well above 10 nM. Note that even below the critical concentration, even thermodynamically, the DNA templates are either naked or fully encapsulated due to the cooperative character of the assembly (8). In our experimental result we find all templates to be (at least) partly encapsulated within the observation time, suggesting that the assembly is limited by kinetics, not thermodynamics. Hence, if reversibility

TABLE 1 Parameter estimates for t_e, I, and $\gamma = It_e$ for increasing coat-protein concentration C-S_{10}-B if no packaging signals are attached to the DNA template

[C-S_{10}-B] (nM)	I (min^{-1})	t_e (min)	γ (–)	V
10	0.015	1.3	0.02	0.48
25	0.5	56	29	1.29
50	0.04	0.2	0.01	0.27
100	0.2	4.1	0.6	0.27
150	0.3	2.7	0.7	0.42
300	0.2	0.6	0.1	1.2

The scaled residual V represents the goodness of the fit, and smaller values reflect a better fit. See also the main text.

does indeed become relevant, it should result mainly in different values for I and t_e. We return to this at the end of this article.

We expect the model parameters I and t_e to both depend on the C-S_{10}-B concentration. Focusing on the high concentration ($\geq$ 100 nM) cases only, we find this to be the case: see Table 1. In general, the elongation time t_e tends to decrease with increasing concentration, albeit not consistently so. This global tendency of the elongation time to decrease with increasing concentration has to be expected, for a larger concentration means that the (thermodynamic) driving force for elongation must be larger. Indeed, the elongation time t_e should be inversely proportional to the C-S_{10}-B concentration because the elongation rate g corresponds with a quasi-first-order reaction. From t_e we can extract the elongation reaction rates (k_{el}/[C-S_{10}-B]) as 2×10^{10} min^{-1} M^{-1}, 2×10^{10} min^{-1} M^{-1}, 9×10^{10} min^{-1} M^{-1} for the 100 nM, 150 nM, and 300 nM cases, respectively. These values are not identical but, considering the simplicity of our model, reasonably close. We note that these values are higher than those found in previous experiments (0.7×10^8 min^{-1} M^{-1}), but this is not entirely surprising given that 1) we expect that the experimental conditions should make the encapsulation of DNA more efficient, and 2) we excluded coat nucleation at random intermediate positions from our model, yet this is observed (10).

Contrasting with this is the nucleation rate, I, which does not seem to vary much with increasing concentration. If we accept that the nucleation rate does indeed not depend much on the concentration at high protein concentration, this would indicate that the nucleation process is limited by some intermediate reaction-limited process rather than a diffusion-limited process, arguably similar to Michaelis-Menten kinetics (33).

Of course, we need to be cautious not to overinterpret the outcome of our curve fitting, firstly because of the simplicity of our model and secondly because the fits are based on a limited set of experimental data; and thirdly, our model ignores aspects that we know take place, such as that nucleation events at random spots on the DNA template do also occur, even though these are not included in our model (21). Nevertheless, the overall agreement is remarkably good, indicating that our model describes most of the relevant underlying physics.

We next compare theory and experiment for the case that packaging signals have been attached to the DNA templates, which are in contact with a C-S_{10}-B concentration of 25 nM. The designated binding sites are chosen such as to in theory produce 5 or 10 regularly placed on the DNA for the dCas12a proteins to bind onto. This does not mean that all these binding sites actually do carry a dCas12a protein in the experiments (15). Experimentally, only lower bounds of 1.2 ± 0.5 and 2.6 ± 1.1 dCas12a proteins could be verified to be attached to the DNA in the cases of 5 or 10 available binding sites, respectively (15). We expect actual values in between the lower bounds and the theoretical maximum.

Ignoring this inconvenience for the moment, and presuming that the theoretical number of packaging signal mimics is equal to the actual one, we show in Fig. 7 the experimental average template coverages and the curves corresponding to the best model parameters found in our procedure. The values of these parameters we present in Table 2, noting that our model fits actually produce values for I and t_e/n, which only fixes t_e if the value of n is known. Hence, we quote the values of t_e/n rather than those of t_e (see also below). We further note that Fig. 7 *A*, representing the case for zero dCas12a binding sites, is identical to Fig. 6 *B* and that we add it for the sake of reference.

The first thing we observe is that, indeed, as announced, adding packaging signal mimics speeds up the packaging of the DNA, and that our model is able to describe this. For five packaging signals we find very good agreement with our analytical model, where our curve fit is (mostly) within the 95% confidence intervals of the experiments. For 10 packaging signals agreement is somewhat less good, and we find that any nucleation rate between approximately unity and 30 per minute yields a residual R between 20 and 40 (see also Table 2 and the supporting material). This wide range in optimal parameter estimates implies that the fit in this particular case is not very accurate. The model curve given in Fig. 7 *C* is that for parameter values $I = 10$ min^{-1} and $t_e/n = 26$ min, which has the lowest residual $R \approx 20$ for all values tested. Although a wide range of parameters appears to fit almost equally well, all give a value of $\gamma = It_e \gg 1$. In this case, Eq. 11 predicts that the mean template coverage $\langle\theta\rangle(t)$ is essentially linear until it reaches the maximum value of unity. Although the linear regime is properly captured within our model, the late-time regime clearly is not, as Fig. 7 *C* shows.

The latter might actually not be surprising. By construction, in our model all packaging signal mimics are equidistantly positioned on the DNA template. As already announced, it is likely that not all of the equidistant dCas12a binding sites have a packaging signal mimic attached to it. This would effectively result in inhomogeneously distributed packaging signal mimics over equidistant binding sites. As we show in the supporting material, this has a pronounced

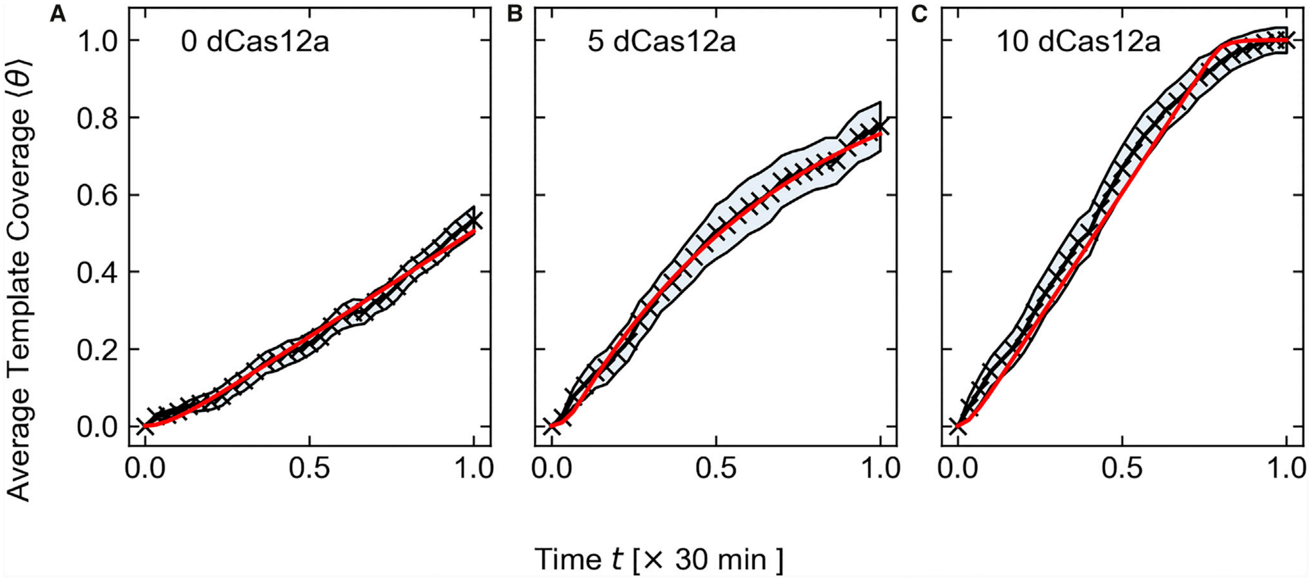

FIGURE 7 Black (*crosses*): experimentally obtained mean template coverage $\langle \theta \rangle$ as function of time obtained from 25 DNA templates for [C-S_{10}-B] = 25 nM (15). The shaded area represents the 95% confidence interval. Red (*drawn line*): best model fit to the average template coverage. (*A*) 0, (*B*) 5, and (*C*) 10 dCas12a proteins attached to the DNA. For the parameter values, see Table 2. To see this figure in color, go online.

effect on the late-time average template coverage only and slows down the assembly kinetics. For early times, no distinction exists between n equidistant or (on average) $\langle n \rangle$ inhomogeneously distributed packaging signals. The reason for this is that the whole encapsulation process is governed by the local conditions only. Therefore, for the average template coverage we can only distinguish between n uniform and n non-uniform subtemplates, if in the latter case the elongation process can encapsulate a larger part of the DNA template than would be possible in the former case. As a result of this, the effect of inhomogeneously distributed bound dCas12a molecules becomes apparent only for later times. This would explain why our model does not properly capture both short- and late-time behavior shown in Fig. 7 *C*.

If we compare the three cases with the different number of dCas12a binding sites in Fig. 7 and Table 2, agreement turns out to be not quite satisfactory. Indeed, within the validity of our model, we would expect that nucleation rate I and the elongation time t_e to not depend on n. From Table 2 we conclude that this is clearly not the case if we insert the theoretical values of $n = 1,\ 6,\ 11$, we find for t_e values of approximately 56, 15, and 250 min, so our model appears to not be internally consistent with the data. The origins of this discrepancy may be due to 1) the nucleation rates not being equal on the free DNA end and the packaging signals, or 2) the value of n not being equal to the maximum value and, because of this, the barriers not being actually equidistant in the experiments. We return to these points at the end of the article.

Finally, we compare the experimental mean waiting time for one-half encapsulation both qualitatively and quantitatively with our model calculations. Our kMC calculations show that this mean waiting time decreases with increasing value of the number of nucleation sites n for fixed values of γ and t_e. This is clearly in agreement with the experimental results of $\langle t_{\frac{1}{2}} \rangle = 31 \pm 6$ min, $\langle t_{\frac{1}{2}} \rangle = 15 \pm 7$ min, and $\langle t_{\frac{1}{2}} \rangle = 10 \pm 3$ min for 0, 5, and 10 dCas12a binding sites, respectively (15). We have no direct analytical prediction of how the mean one-half encapsulation time depends on the number of packaging signals, but we did conclude from Fig. 5 that the shortest mean waiting time for one-half encapsulation corresponded to the case where the number of packaging signals n becomes very large.

In that case, Eq. 20 describes the relation between $\langle t_{\frac{1}{2}} \rangle (n \to \infty) / \langle t_{\frac{1}{2}} \rangle (n = 1)$ and γ, and can be used to derive bounds for the value of γ. Note that the experimental result

TABLE 2 Parameter estimates for the nucleation rate *I*, the reduced elongation time t_e/n, and the elongation time t_e, presuming that all dCas12a binding sites are occupied and the dimensionless nucleation rate γ/n for the case that 0, 5, or 10 dCas12a binding sites are present on the DNA template

dCas12a binding sites	I (min^{-1})	t_e/n (min)	t_e (min)	γ/n (–)	γ (–)	V
0	0.5	56	56	29	29	1.29
5	0.05	2.5	15	0.125	0.7	0.7
10	1 – 30	24 – 26	264 – 286	24 – 720	264 – 7920	1.0

Here, n denotes the number of nucleation sites on the template. The concentration of coat proteins is fixed at [C-S_{10}-B] = 25 nM. The scaled residual V represents the goodness of the fit, and smaller values reflect a better fit.

$\langle t_{\frac{1}{2}} \rangle = 10 \pm 3$ min for 10 dCas12a binding sites must now be an upper bound for $\langle t_{\frac{1}{2}} \rangle (n \to \infty)$. It follows from Eq. 20 that this results in a lower bound for the value of γ, which makes sense as a larger value for γ in our model correlates to a faster encapsulation process. In this way we find the lower bound $\gamma > 2 \pm 2$. Consequently, the parameter γ is likely to be larger than unity, and we conclude from Eq. 13 and Fig. 4 that in this particular case the packaging signals should also decrease the mean time for complete encapsulation.

It is important to note that this value is not completely consistent with the parameter estimations we found comparing the average template coverage. Here, we found that for the case of five dCas12a, $\gamma = 0.7$ is slightly smaller than unity, contrasting with our conclusion that γ is likely larger than that (see Table 2). However, since the number of packaging signals n is unknown, and the fit for five dCas12a is obtained using partial template coverage only, we refrain from overinterpreting this apparent discrepancy.

CONCLUSIONS

To summarize, we have developed a nucleation-and-growth model to explain the recent experimental observation that attaching so-called packaging signal mimics onto DNA templates can significantly increase the encapsulation rate of the DNA by designer proteins that mimic virus coat proteins (15). These packaging signals are synthetic proteins designed with CRISPR-Cas techniques that insert themselves on predestined positions along the DNA and can act as inert barriers or as actual binding sites for the coat proteins. In both cases, the encapsulation rate increases with the number of packaging signals, although we have limited our study to the former.

Our one-dimensional model is based on the experimental observation that the protein coat can only elongate after an initial nucleation process. Packaging signals are modeled as diffusion and elongation barriers that behave as additional nucleation sites. For the inert packaging signals, we understand this to originate from the local accumulation of coat proteins near such packaging signals in the experiments, resulting from the flow of protein solution across the stretched DNA chains. In the absence of packaging signals and for DNA molecules that are in some sense sufficiently short, nucleation happens at the far end of the DNA molecule, which then acts as an effective packaging signal.

In the absence of additional packaging signals, we find that there are three growth regimes for the mean template coverage, i.e., for the fraction of DNA packaged by the coat proteins. These include a quadratic early regime, a linear intermediate regime, and an exponentially decaying late-time regime. This indicates that our model shows a surprisingly rich behavior considering that it is essentially determined by a single dimensionless parameter in the form of the dimensionless nucleation rate. If additional packaging signals are attached to the DNA template, we find the same three regimes, but the relevant timescales for the early-time and intermediate regime now depend on the number of packaging signals.

We find that the mean template coverage does always increase with increasing number of packaging signals. Interestingly, this is not mirrored by the mean waiting time for complete encapsulation, which we find to depend non-trivially on the dimensionless nucleation rate and the number of packaging signals. For sufficiently large nucleation rates, we find that this mean waiting time decreases rapidly only with the first few packaging signals, after which it depends only weakly on the number of packaging signals. A clear but shallow optimum exists in the mean waiting time as a function of the number of packaging signals, which is important for experimentally optimizing genome packaging using packaging signals.

The mean waiting time for complete encapsulation might, however, not always be easily accessible experimentally. Our kMC simulations show, however, that the mean time for partial encapsulation is not always correlated in a simple way with the mean waiting time for complete encapsulation. In fact, based on the results presented in Figs. 4 and 5, we find that for a template coverage below approximately 63%, adding a sufficient number of packaging signals always decreases the mean waiting time with increasing number of packaging signals, even if the mean waiting time for complete encapsulation increases. Hence, there is some reason for caution when interpreting experimental data on encapsulation kinetics if near completion of the process is not achieved.

The quantitative agreement between our kinetic theory and the experiments is not quite perfect. Still, we find that the qualitative agreement is certainly encouraging. Generally, our highly idealized model is able to explain many of the experimental results of Calcines-Cruz and co-workers (15), suggesting that it does include most of the relevant physics. Still, that our model cannot self-consistently explain all experimental observations suggests that our model is incomplete. We note that some of the discrepancies might also be associated with the limited experimental data set, spanning only the first 30 min of the co-assembly. This becomes especially relevant for low concentrations of coat protein, as only a small portion of the DNA can encapsulate within this time window. A better understanding of the disagreement between our calculations and the experiments by Calcines-Cruz et al. (15) would therefore also require data on the encapsulation kinetics spanning a longer time.

The first and arguably one of the main simplifications in the current model is the assumption that the nucleation sites at the free end of the DNA and at the packaging signals are equivalent, i.e., have the same nucleation rate. That this is not necessarily true was already shown by the experimental observation that a specifically functionalized dCas12a protein changes the overall encapsulation rate (15). This can, in principle, be incorporated relatively straightforwardly in the mean template coverage using Eq. 10 by requiring that the nucleation rate on a single subtemplate differs from that on the other $n - 1$ subtemplates. Preliminary results show that while the average template coverage changes in a fairly trivial way, this is certainly not the case for the mean waiting time (see supporting material). Yet the effect remains minor if the ratio of the nucleation rates deviates not too much from unity.

A second effect absent in our model is that not all target sites on the DNA that can bind a packaging signal are actually bound to one. Although we argue that this should have only little influence on the early-time regime, the late-time regime should be expected to be affected significantly. Indeed, it is especially in this late-time regime that our model deviates significantly from the experimental results.

The third effect that we did not incorporate into the model is coat-protein nucleation at arbitrary places on the DNA template. As already alluded to, this has been observed experimentally for sufficiently long DNA templates (15). We expect, however, that this is mostly relevant for DNA strands without packaging signals, as the additional nucleation sites make it less likely that this process occurs. We intend to remedy all three of these simplifications in future work.

It seems sensible to remind the reader why we opted for a description based on irreversible kinetics to explain the experiments of Calcines-Cruz et al. (15) rather than one based on microscopic reversibility that eventually produces a thermodynamically consistent coverage. We recall that the main motivation for our model is the experimental observation that disassembly of encapsulated DNAs is exceedingly slow when exposed to protein-free buffer solution and is not observed within the experimental time (unpublished data). This is even the case for naked capsids formed in absence of DNA, which, naively, should be less stable than capsids that contain DNA (17). This contrasts with the predictions of a kinetic model based on the equilibrium zipper model, which shows that disassembly must commence immediately (8).

A purely thermodynamic model also predicts that adding packaging signal mimics to the DNA template actually decreases the ensemble-averaged template coverage. The implication is that from a thermodynamic point of view the overall assembly becomes less, not more, efficient with increasing number of nucleation sites. That this is indeed so is actually a highly non-trivial result attributable to a competition between two opposing effects. Although additional nucleation sites introduce a large entropy gain for partially encapsulated templates, the additional nucleation sites also introduce additional energy penalties for DNA encapsulation. It turns out that the energy penalty always dominates over the entropy gain. Albeit based on an equilibrium theory, we believe that lower mean coverage with increasing number of nucleation sites also implies slower dynamics, because the thermodynamic driving force for encapsulation becomes smaller. This is also in disagreement with experimental observations shown in Fig. 7.

As a final note, we would like to emphasize that our work might also be relevant for understanding the role of packaging signals in naturally occurring viruses, even though we designed our model specifically to understand the effect of specifically designed packaging signals on the assembly kinetics of filamentous VLPs. Our simple model shows that adding packaging signals can speed up but also slow down the assembly kinetics of complete viruses and that this depends on how fast nucleation is on the timescale of elongation (see Fig. 4). This might perhaps explain the diversity in the number of packaging signals and their binding strength in viruses. It could simply be the kinetically optimal way to encapsulate the genetic material for a given interaction strength between the coat protein and the genetic material (34).

For viruses with a more complex capsid geometry, such as the icosahedral geometry, self-assembly must become a higher-than-one-dimensional process to enable the formation of complete capsids. Whether our model has any bearing on the impact of packaging signals in these viruses remains to be seen. We note, however, that in some modeling approaches the assembly process essentially reduces this from a three-dimensional to a quasi-one-dimensional process, albeit with additional geometric constraints (9,19). The complex models of Twarock and collaborators indicate that including a larger number of high-affinity packaging signals is not necessarily beneficial for the encapsulation process (19). This agrees with the results from our simple kinetic model.

All of this confirms that in molecular self-assembly, where hysteresis plays a very important role, establishing what assembled state is the most prevalent is not necessarily only dictated by thermodynamics but most certainly also by kinetics (35–37).

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2022.05.040.

AUTHOR CONTRIBUTIONS

R.d.B., A.H.-G., and P.v.d.S. designed the research. P.C.M.W. and R.d.B. carried out the theoretical calculations and analyzed the data. C.C.-C. and R.d.B. designed, conducted, and analyzed the kMC calculations. T.v.W. and P.v.d.S. analyzed the effect of packaging signals mimics within the thermodynamic zipper model. R.d.B. and P.v.d.S. wrote the article.

ACKNOWLEDGMENTS

R.d.B. and P.v.d.S. acknowledge funding by the Institute of Complex Molecular Systems at Eindhoven University of Technology. A.H.-G. acknowledges funding from a UC MEXUS-CONACyT collaborative grant.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPORTING CITATIONS

Reference (38) appears in the supporting material.

REFERENCES

1. Zanotti, G., and A. Grinzato. 2021. Structure of filamentous viruses. *Curr. Opin. Virol.* 51:25–33. https://doi.org/10.1016/j.coviro.2021.09.006.
2. Kegel, W. K., and P. van der Schoot. 2006. Physical regulation of the self-assembly of tobacco mosaic virus coat protein. *Biophys. J.* 91:1501–1512. https://doi.org/10.1529/biophysj.105.072603.
3. Hernandez-Garcia, A., D. J. Kraft, …, R. de Vries. 2014. Design and self-assembly of simple coat proteins for artificial viruses. *Nat. Nanotechnol.* 9:698–702. https://doi.org/10.1038/nnano.2014.169.
4. Caspar, D. L. D., and A. Klug. 1962. Physical principles in the construction of regular viruses. *Cold Spring Harb. Symp. Quant. Biol.* 27:1–24. https://doi.org/10.1101/sqb.1962.027.001.005.
5. Caspar, D. 1980. Movement and self-control in protein assemblies. Quasi-equivalence revisited. *Biophys. J.* 32:103–138. https://doi.org/10.1016/s0006-3495(80)84929-0.
6. Stubbs, G., and A. Kendall. 2012. Helical viruses. *In* Viral Molecular Machines. M. G. Rossmann and V. B. Rao, eds. Springer.
7. Janssen, P. G. A., S. Jabbari-Farouji, …, A. P. H. J. Schenning. 2009. Insights into templated supramolecular polymerization: binding of naphthalene derivatives to ssDNA templates of different lengths. *J. Am. Chem. Soc.* 131:1222–1231. https://doi.org/10.1021/ja808075h.
8. Kraft, D. J., W. Kegel, and P. van der Schoot. 2012. A kinetic zipper model and the assembly of tobacco mosaic virus. *Biophys. J.* 102:2845–2855. https://doi.org/10.1016/j.bpj.2012.05.007.
9. Stockley, P. G., R. Twarock, …, R. Tuma. 2013. Packaging signals in single-stranded RNA viruses: nature's alternative to a purely electrostatic assembly mechanism. *J. Biol. Phys.* 39:277–287. https://doi.org/10.1007/s10867-013-9313-0.
10. Punter, M. T. J. J. M., A. Hernandez-Garcia, …, P. van der Schoot. 2016. Self-assembly dynamics of linear virus-like particles: theory and experiment. *J. Phys. Chem. B.* 120:6286–6297. https://doi.org/10.1021/acs.jpcb.6b02680.
11. Jekhmane, S., R. De Haas, …, R. De Vries. 2017. Virus-like particles of mRNA with artificial minimal coat proteins: particle formation, stability, and transfection efficiency. *Nucleic Acid Ther.* 27:159–167. https://doi.org/10.1089/nat.2016.0660.
12. Cárdenas-Guerra, R. E., D. S. Moreno-Gutierrez, …, A. Hernandez-Garcia. 2020. Delivery of antisense DNA into pathogenic parasite Trypanosoma cruzi using virus-like protein-based nanoparticles. *Nucleic Acid Ther.* 30:392–401. https://doi.org/10.1089/nat.2020.0870.
13. Razzokov, J., S. Naderi, and P. van der Schoot. 2014. Prediction of the structure of a silk-like protein in oligomeric states using explicit and implicit solvent models. *Soft Matter.* 10:5362. https://doi.org/10.1039/c4sm00384e.
14. Razzokov, J., S. Naderi, and P. van der Schoot. 2018. Nanoscale insight into silk-like protein self-assembly: effect of design and number of repeat units. *Phys. Biol.* 15:066010. https://doi.org/10.1088/1478-3975/aadb5e.
15. Calcines-Cruz, C., I. J. Finkelstein, and A. Hernandez-Garcia. 2021. CRISPR-guided programmable self-assembly of artificial virus-like nucleocapsids. *Nano Lett.* 21:2752–2757. https://doi.org/10.1021/acs.nanolett.0c04640.
16. Kittel, C. 1969. Phase transition of a molecular zipper. *Am. J. Phys.* 37:917–920. https://doi.org/10.1119/1.1975930.
17. Vargas, E. C., M. A. C. Stuart, …, A. Hernandez-Garcia. 2019. Template-free self-assembly of artificial de novo viral coat proteins into nanorods: effects of sequence, concentration, and temperature. *Chem. Eur. J.* 25:11058–11065. https://doi.org/10.1002/chem.201901486.
18. Chevreuil, M., L. Lecoq, …, G. Tresset. 2020. Nonsymmetrical dynamics of the HBV capsid assembly and disassembly evidenced by their transient species. *J. Phys. Chem. B.* 124:9987–9995. https://doi.org/10.1021/acs.jpcb.0c05024.
19. Twarock, R., R. J. Bingham, …, P. G. Stockley. 2018. A modelling paradigm for RNA virus assembly. *Curr. Opin. Virol.* 31:74–81. https://doi.org/10.1016/j.coviro.2018.07.003.
20. Rein, A. 2019. RNA packaging in HIV. *Trends Microbiol.* 27:715–723. https://doi.org/10.1016/j.tim.2019.04.003.
21. Marchetti, M., D. Kamsma, …, W. H. Roos. 2019. Real-time assembly of viruslike nucleocapsids elucidated at the single-particle level. *Nano Lett.* 19:5746–5753. https://doi.org/10.1021/acs.nanolett.9b02376.
22. Cahn, J. W. 1995. The time cone method for nucleation and growth kinetics on a finite domain. *Mater. Res. Soc. Symp. Proc.* 398:425. https://doi.org/10.1557/proc-398-425.
23. Gallie, D. R., K. A. Plaskitt, …, A. Wilson. 1987. The effect of multiple dispersed copies of the origin-of-assembly sequence from TMV RNA on the morphology of pseudovirus particles assembled in vitro. *Virology.* 158:473–476. https://doi.org/10.1016/0042-6822(87)90225-x.
24. Eber, F. J., S. Eiben, …, C. Wege. 2015. RNA-controlled assembly of tobacco mosaic virus-derived complex structures: from nanoboomerangs to tetrapods. *Nanoscale.* 7:344–355. https://doi.org/10.1039/c4nr05434b.
25. Buzón, P., S. Maity, and W. H. Roos. 2020. Physical virology: from virus self-assembly to particle mechanics. *Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol.* 12:e1613. https://doi.org/10.1002/wnan.1613.
26. Cingil, H. E., E. B. Boz, …, J. Sprakel. 2017. Illuminating the reaction pathways of viromimetic assembly. *J. Am. Chem. Soc.* 139:4962–4968. https://doi.org/10.1021/jacs.7b01401.
27. Guo, B.-N., and F. Qi. 2011. Sharp bounds for harmonic numbers. *Appl. Math. Comput.* 218:991–995. https://doi.org/10.1016/j.amc.2011.01.089.
28. Olver, F. W., D. W. Lozier, …, C. W. Clark. 2010. NIST Handbook of Mathematical Functions. Cambridge University Press.
29. Gillespie, D. T. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* 22:403–434. https://doi.org/10.1016/0021-9991(76)90041-3.
30. Gillespie, D. T. 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81:2340–2361. https://doi.org/10.1021/j100540a008.
31. Gillespie, D. T. 2007. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* 58:35–55. https://doi.org/10.1146/annurev.physchem.58.032806.104637.
32. Strutz, T. 2011. Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and beyond. Vieweg+ Teubner.
33. Engel, P. C. 1981. Enzyme Kinetics: The Steady-State Approach. Chapman & Hall.
34. Comas-Garcia, M. 2019. Packaging of genomic RNA in positive-sense single-stranded RNA viruses: a complex story. *Viruses.* 11:253. https://doi.org/10.3390/v11030253.
35. Harkness, V. R. W., N. Avakyan, …, A. K. Mittermaier. 2018. Mapping the energy landscapes of supramolecular assembly by thermal hysteresis. *Nat. Commun.* 9:3152. https://doi.org/10.1038/s41467-018-05502-z.
36. Singh, S., and A. Zlotnick. 2003. Observed hysteresis of virus capsid disassembly is implicit in kinetic models of assembly. *J. Biol. Chem.* 278:18249–18255. https://doi.org/10.1074/jbc.m211408200.
37. Michaels, T. C., A. Šarić, …, T. P. Knowles. 2018. Chemical kinetics for bridging molecular mechanisms and macroscopic measurements of amyloid fibril formation. *Annu. Rev. Phys. Chem.* 69:273–298. https://doi.org/10.1146/annurev-physchem-050317-021322.
38. Ross, S. 2014. A First Course in Probability. Pearson.