Spanning tree model and the assembly kinetics of RNA viruses

Inbal Mizrahi,[1] Robijn Bruinsma,[1,2] and Joseph Rudnick[1]

[1]*Department of Physics and Astronomy, University of California, Los Angeles, California 90095, USA*
[2]*Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, USA*

(Received 9 August 2021; accepted 17 May 2022; published 12 October 2022)

Single-stranded RNA (ssRNA) viruses self-assemble spontaneously in solutions that contain the viral RNA genome molecules and viral capsid proteins. The self-assembly of *empty* capsids can be understood on the basis of free energy minimization. However, during the self-assembly of complete viral particles in the cytoplasm of an infected cell, the viral genome molecules must be selected from a large pool of very similar host messenger RNA molecules and it is not known whether this also can be understood by free energy minimization. We address this question using a simple mathematical model, the spanning tree model, that was recently proposed for the assembly of small ssRNA viruses. We present a statistical physics analysis of the properties of this model. RNA selection takes place via a kinetic mechanism that operates during the formation of the nucleation complex and that is related to Hopfield kinetic proofreading.

DOI: 10.1103/PhysRevE.106.044405

I. INTRODUCTION

Many single-stranded RNA (ssRNA) viruses, such as the polio and common cold viruses, are able to self-assemble spontaneously into infectious viral particles ("virions") in solutions that contain appropriate concentrations of viral capsid proteins and RNA molecules [1,2]. For these viruses, assembly is believed to be a purely passive process driven by free energy minimization. Early work by Klug on the tobacco mosaic virus (TMV) [3] indicated that RNA genome molecules ("gRNA") act as *templates* that direct the viral assembly process. Klug proposed a physical model for viral assembly in which the repulsive electrostatic interactions between positively charged groups of the capsid proteins are just strong enough to overcome competing attractive hydrophobic interactions between the proteins, thus preventing the self-assembly of empty capsids under physiological conditions. If then viral RNA molecules are added to the solution, the negative charges of the RNA nucleotides neutralize some of the positive charges of the capsid proteins tilting the free energy balance towards assembly [4–7].[1]

gRNA molecules must compete for packaging with a large pool of, quite similar, host messenger RNA (mRNA) molecules [8]. For the case of influenza, the number of gRNA molecules inside an infected cell is less than 10^4 [9] while the total number of host mRNA molecules is in the range of 3.6×10^5. For the case of the HIV-1 virus, the number of gRNA molecules may be as low as 10^2. Like other ssRNA molecules, gRNA molecules have a treelike "secondary structure" produced by Watson-Crick base pairing between complementary RNA nucleotides of the primary sequence of RNA nucleotides [10]. The redundancy of the genetic code allows for the possibility of "silent" (or synonymous) mutations that can alter the secondary structure of the molecule without altering the structure of the proteins encoded by the nucleotide sequence [11]. gRNA molecules appear to have undergone different forms of evolutionary adaptation that increased the packaging probability. On the one hand, gRNA molecules have short sequences, known as *packaging signals* (PS), with specific affinity for the capsid proteins of the virus [12–18]. On the other hand, the global topology of gRNA molecules differs from that of generic mRNA molecules: they are longer while their secondary structure is significantly more branched and compact. Compactness reduces the radius of gyration of the RNA molecules in solution and hence the free energy cost of compacting the RNA molecules prior to encapsidation [19].

The physical aspects of ssRNA packaging have been extensively studied experimentally, theoretically, and by numerical studies of model systems [4–6,20–35]. Many of the theoretical studies have focused on the minimization of the free energy of assembled virions. This produced global measures for RNA selectivity in terms of their length and the compactness of the RNA molecules. On the other hand, experimental studies of the self-assembly of *empty* capsids [36–38] were interpreted in terms of a kinetic nucleation-and-growth scenario, where the energetically uphill formation of a "nucleation complex," composed of a small number of capsid proteins, is followed by an energetically downhill "elongation process" that ends with the closure of the capsid.

This nucleation complex may be compared to the critical nucleus of the classical theory of nucleation and growth theory as applied to empty capsid assembly [39,40]. The TMV assembly scenario proposed by Klug is actually a nucleation-and-growth scenario with a nucleation complex composed of a single PS associating plus a disk of proteins. The subsequent elongation proceeds by the addition of more protein disks until the end of the genome molecule is reached. The contribution

[1] For a quantitative treatment, see Ref. [21].

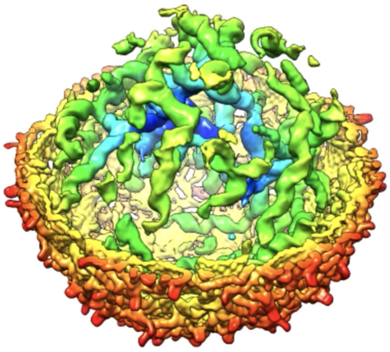

FIG. 1. CryoEM asymmetric reconstruction of the MS2 virion. The viral genome molecules (in green and blue) associate reproducibly mostly with one half of the capsid shell (from Ref. [50]).

of the initial PS to the assembly free energy is probably very small, which suggests that free energy minimization might not be a good method to understand how TMV RNA is being selected. Recent observations on the assembly kinetics of individual MS2 viruses (a small ssRNA bacteriophage virus) reported a wide distribution of timescales [41], which is consistent with a nucleation-and-growth scenario. Next, for the case of the assembly of the HIV-1 retrovirus (see [42] and references therein), RNA selectivity was found to depend on the cooperative action of a cluster of PS located at the 5′ end of the gRNA molecule, known as the ψ sequence. This sequence is about 100 nucleotides long, which is again small compared to the total length of the HIV-1 genome of about 10^4 nucleotides. HIV-1 gRNA selection appears to take place during the nucleation stage of the assembly process when this ψ sequence interacts with a small group of capsid proteins. Changing the RNA sequence of the non-ψ part of the genome molecules does not affect selectivity. The PS of HIV-1 was shown to provide no significant thermodynamic advantage to the gRNA molecules over nonviral RNA molecules of the same length [43]. As for the MS2 case, the HIV-1 assembly process is characterized by a broad distribution of timescales.

Important information about the kinetics of viral co-assembly can be gleaned as well from purely structural studies. Reconstruction of packaged genome molecules using "icosahedral averaging" [44] showed that the interior surface of the icosahedral capsids of the *Nodaviridae* [45,46]) is decorated by paired RNA strands lining the edges of the "capsomers" (pentameric or hexameric groupings of capsid proteins). Recent progress in cryoelectron tomography has made it possible to reconstruct the way individual ss-RNA genome molecules are packaged inside spherical capsids without having to resort to icosahedral averaging ("asymmetric reconstruction" [47,48]). An important example is again the MS2 virus. It was found that sections of the RNA genome rich in PS reproducibly associate with roughly *half* of the interior surface of the capsid [49,50], as shown in Fig. 1. The remaining capsid proteins do not associate in a reproducible manner with the gRNA. These results can be interpreted as evidence for a well-defined nucleoprotein complex held together by a particular section of the viral RNA molecule that is rich in packaging sequences while the subsequent downhill elongation process is driven by generic electrostatic interactions. In contrast, the asymmetric reconstruction of the CCMV and BMV plant viruses produced only a very small amount of reproducible RNA-protein association [48]. Interestingly, this same group of viruses is much less selective than MS2. In fact, CCMV capsid proteins promiscuously select BMV gRNA over their own CCMV gRNA, while they package a wide variety of nonviral ssRNA molecules and even non-RNA polyelectrolytes [51,52].

The nucleation-and-growth scenario provides a possible framework for RNA selection in which the PS reduce the height of the assembly energy activation barrier. This selectively increases the packaging probability of viral RNA molecules as compared to the packaging of host mRNA molecules. Since the production rate of virions depends *exponentially* on the height of activation energy barrier, a limited number of PS could have a disproportionally large effect on the RNA selectivity. Such a mechanism might be called "selective nucleation."[2] The action of PS would be similar in this view to that of enzymes or catalysts that increase the rate of a chemical reaction by reducing the height of an energy activation barrier. This selective nucleation scenario should be contrasted with the selection mechanism based on RNA compactness that was discussed earlier

The aim of this article is to explore the physics of selective nucleation for the case of a recently proposed mathematical model of RNA-directed viral assembly, the "*spanning tree model.*" This model allows for tens of thousands of competing RNA secondary structure configurations that have the same final assembly energy but different assembly energy barriers and assembly pathways. In this model, there is practically no selectivity under conditions of thermodynamic equilibrium, which allows us to focus on selective nucleation. The model, which is sufficiently simple so its kinetics can be determined by numerical integration of a set of coupled master equations, is itself a generalization of an earlier model for the assembly of *empty* dodecahedral capsids by Zlotnick [53–55]. The Zlotnick model obeys a nucleation-and-growth assembly scenario [56] and it has been used to carry out productive simulations of the packaging of linear genome molecules [7,57]. Using the spanning tree model, we investigate how long kinetic specificity can persist in the face of a final state of thermal equilibration in which there is no selectivity.

The spanning tree model is introduced in Sec. II, followed by a topological and geometrical classification of the model genome molecules. Next is a discussion of minimum energy assembly pathways and of the structural properties of the partial assemblies. In Sec. III, we introduce the nonlinear master equation for the assembly kinetics. Numerical integration of the master equation is used to determine the characteristic time scales of the assembly kinetics and to study packaging competition between different classes of genome molecules for different levels of supersaturation and RNA-to-protein mixing ratios. In Sec. IV we examine a *two-stage* packaging scenario to compare the results with a collective assembly scenario. In the concluding Sec. V we summarize our results, discuss experimental predictions, and return to the question of

[2] Selective nucleation was proposed by Rouzina in the context of the assembly of retroviruses.

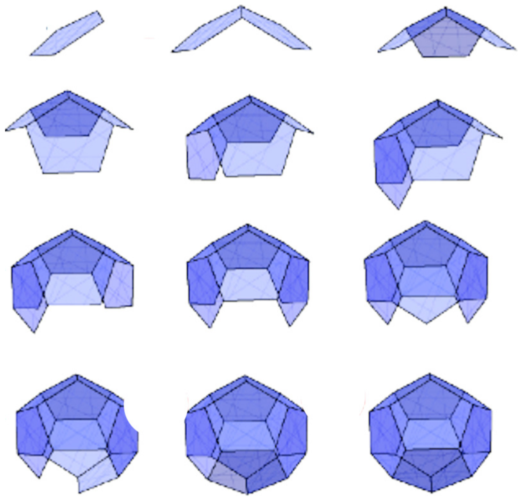

FIG. 2. Empty capsid assembly pathway. The figure shows one of the minimum-energy pathways for the assembly of a dodecahedral shell composed of 12 pentamers with adhesive edges. Note that the assembly intermediates all are compact structures.

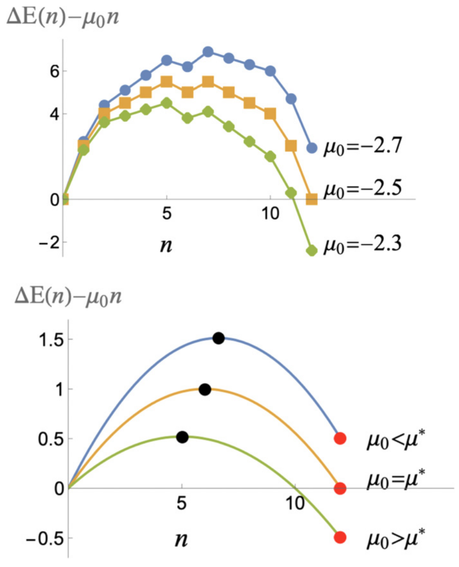

FIG. 3. Top: Energy profiles of a minimum-energy assembly pathway of the Zlotnick model. Blue dots: The chemical potential μ_0 is slightly below μ^*, the value of the chemical potential for assembly equilibrium. Orange squares: μ_0 is equal to μ^*. Green diamonds: μ_0 is slightly above μ^*. Bottom: Assembly energy profiles according to the continuum theory of nucleation and growth [39]. Solid red dots: energy minima. Solid black dots: energy maxima.

kinetic selection by PS versus thermodynamic selection based on RNA compactness.

II. SPANNING TREE MODEL

A. Empty capsid assembly

The Zlotnick model treats the capsid as a dodecahedral shell composed of 12 pentamers. The 60 proteins of the shell correspond to the capsid of a minimal "$T = 1$" virus. Assembly is driven by attractive edge-edge interactions between the pentamers. A *minimum-energy assembly pathway* can be defined as a pentamer-by-pentamer addition sequence where each added pentamer is placed in a location that minimizes the free energy of the partial shell. An example of one of the very many ($\simeq 10^5$) degenerate minimum-energy assembly pathways is shown in Fig. 2. The assembly energy $\Delta E(n)$ of a partial shell composed of n pentamers is defined to be $\Delta E(n)/E_0 = -(n_1 + n\mu_0)$. Here, E_0 is the magnitude of the edge-to-edge binding energy between pentamers. This binding energy can be estimated by comparison with thermodynamic assembly studies of empty capsids, giving a value for E_0 of about $4.3k_bT$ [53]. In the following, energy parameters will be expressed in units of E_0. Next, n_1 is the number of shared pentamer edges of the partial shell and μ_0 is the pentamer chemical potential at a certain reference concentration. The assembly energy of a complete capsid equals $\Delta E(12)/E_0 = -(30 + 12\mu_0)$ for all minimum energy assembly pathways. Assembly equilibrium is the state where the chemical potential of a pentamer in solution is the same as the energy of a pentamer that is part of a capsid. This is the case if $\Delta E(12) = 0$ so if the reference chemical potential equals $\mu^* = -\frac{5}{2}$.

Figure 3 (top) shows the minimum energies of the n-pentamer partial assemblies of Fig. 2 for three different values of the reference chemical potential near μ^*. All minimum energy assembly pathways of the Zlotnick model have the same value of n_1 for given n so this minimum energy assembly pathway is highly degenerate. For $\mu_0 < \mu^*$, the absolute energy minimum is at $n = 0$ while for $\mu_0 > \mu^*$ the absolute minimum is at $n = 12$, the assembled capsid. The assembly activation energy barrier of a profile is the height of the maximum. The location n^* of the maximum under equilibrium conditions corresponds to a half-filled shell, shifting to lower values as μ_0 increases. For comparison, Fig. 3 (bottom) shows the assembly energy of a spherical cap growing into a spherical shell [39], which produces the standard nucleation-and-growth profile. The initial rise of $\Delta E(n)$ with n is due to the fact that the line energy of the perimeter of the cap increases with n for n less than six while the subsequent drop of $\Delta E(n)$ is due to the fact that the line energy decreases as a function of n for n larger than six, when the perimeter starts to shrink.

B. Spanning trees and their classification

The second part of the definition of the model concerns the representation of RNA molecules. The RNA molecules are assumed to be compacted into dodecahedra of identical size whose shape matches the interior of the dodecahedral capsid of the Zlotnick model. The molecules differ only in terms of a "PS section" that is in contact with the capsid. This section is assumed to have a secondary structure in the form of a tree graph with 20 nodes that cover all the vertices and 19 links, leaving 11 of the 30 edges of the dodecahedron uncovered. This choice, motivated by the organization of the RNA genome molecules of the *Nodaviridae* [45,46], allows for considerable mathematical simplification. The interaction between the 19 links of the tree and the capsid constitute the specific interactions while the interaction of the 11 remaining edges with the capsid constitute the generic contacts.

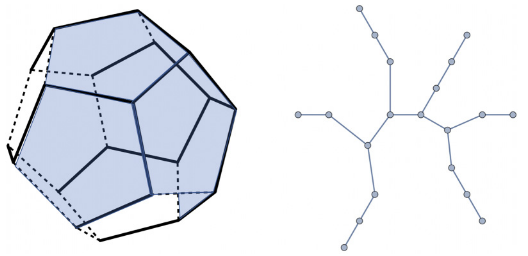

FIG. 4. Left: Spanning tree with nodes located on the vertices of a dodecahedron. The solid lines are the links of the spanning tree, while the dashed lines indicate edges of the dodecahedron that are not links of the spanning tree. A maximum of six pentamers, shown in blue, can be placed on the dodecahedron such that each pentamer is wrapped by four links. Right: Planar graph of the spanning tree.

Tree graphs are defined as collections of nodes connected by links such that there is one and only one path of links connecting any pair of nodes [58]. A *spanning tree graph* of a polyhedron is defined as a tree graph whose nodes are located on the vertices of the polyhedron with just enough links to connect the nodes together in a tree structure [59]. For a dodecahedron there are of the order of 10^5 spanning trees that represent the different possible PS configurations of the spanning tree model. Figure 4 (left) shows on the left an example of a spanning tree graph of the dodecahedron. The projection of this spanning tree on the plane is shown on the right. Now place a pentamer on the spanning tree. Each pentamer interacts with five edges of the dodecahedron, but by drawing different spanning trees one can convince oneself that a pentamer can interact with no more than four links of a spanning tree. If the interaction of pentamer edges with the links of the spanning tree is energetically favorable, as we will assume, then a cluster of pentamers minimizes the interaction free energy between pentamers and spanning tree by maximizing the number of pentamers in contact with four links of the spanning tree. The figure shows that a maximum of six pentamers can be positioned in this fashion. We will say that the *wrapping number* of this tree structure is $N_P = 6$. The maximum wrapping number for a spanning tree of the dodecahedron is eight while the minimum is two. The same spanning tree can in general be distributed over a dodecahedron in different ways, resulting in different wrapping numbers. The wrapping number is thus not a topological characteristic of the secondary structure.

The *compactness* of a spanning tree is a measure of the probability that two pentamers placed on the dodecahedron are able to share an edge. The *maximum ladder distance* (or MLD) is a frequently used measure of the compactness of a secondary structure [19,60]. The MLD of a spanning tree graph is defined here as the maximum number of links separating any pair of nodes. In graph theory, the ladder distance between two nodes of a tree graph is called the "distance" while the MLD is known as the "diameter" of a tree graph [58]. The MLD of the tree molecule shown in Fig. 4 is 9. It can be demonstrated that the smallest possible MLD for a spanning tree of the dodecahedron is 9 (see Appendix A) while the largest possible MLD of a spanning tree is 19. Minimum MLD spanning trees resemble Cayley trees while maximum MLD spanning trees are *Hamiltonian paths*. These are walks without self-intersection that visit all vertices of a polyhedron [17,61]. In the absence of interactions, the solution radius of gyration of a branched polymeric molecule increases with the MLD as a power law [62]. A systematic comparison between the genomic RNA molecules of RNA viruses confirms that they have significantly lower MLDs than randomized versions of the same molecules [19,60]. In Ref. [11] the authors showed that this result holds also when the randomization only amounts to synonymous mutations. It should be emphasized however that in our paper the MLD concept is applied only to the PS section of twenty links.

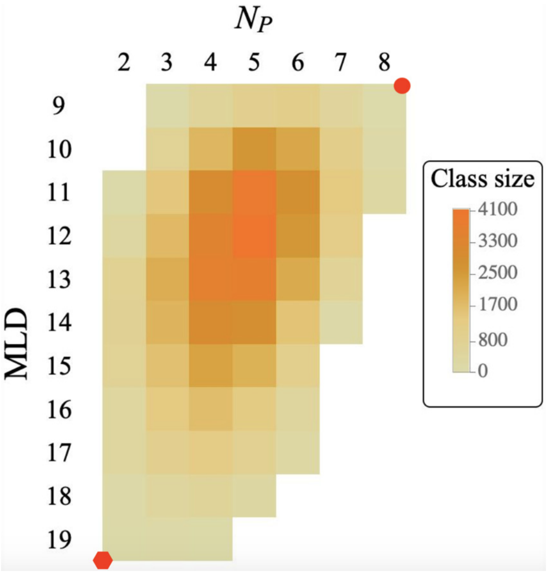

FIG. 5. Heat map of the number of spanning trees for different wrapping and MLD numbers. The circular marker indicates spanning trees that are maximally adapted for packaging with MLD = 9 and $N_P = 8$, while the polygonal marker shows spanning trees that are minimally adapted for packaging with MLD = 19 and $N_P = 2$.

The wrapping number and the maximum ladder distance are the two characteristics that we will use to classify spanning trees. Figure 5 is a plot of the class size, i.e., the number of spanning trees for a given pair of wrapping and MLD numbers. The class size has a pronounced maximum for spanning trees with MLD and N_P numbers in the range of MLD = 12 and $N_P = 5$. The typical class size is about 4×10^3 in this range. If different spanning trees would have the same *a priori* probability, then spanning trees with MLD and N_P numbers in this range would be overwhelmingly more probable than spanning trees that are maximally adapted for packaging, with MLD = 8 and $N_P = 9$ (circular dot), or spanning trees that are minimally adapted, with MLD = 19 and $N_P = 2$ (polygonal dot). Under such conditions, the logarithm of the class size could be viewed as a configurational entropy. The configurational entropy of an annealed branched polymer composed of 19 monomers *not* constrained to be a spanning tree is known to approximately $19 - \mathrm{MLD}^2/19$ [62], which has a maximum at the smallest possible MLD. Evidently, the demand that a tree molecule also is a spanning tree of a dodecahedron significantly alters the configurational entropy.

The plot would seem to give the impression that there are no spanning trees that are maximally adapted for packaging, with MLD = 8 and $N_P = 9$ (circular dot), nor spanning trees that are minimally adapted, with MLD = 19 and $N_P = 2$ (polygonal dot). In actuality, there is a relatively small but finite number of such spanning trees, though this is not visible in the heat map. Appendix B shows projections of the class size on the MLD axis and of the class size on the N_P axis, which make it clear that there are spanning tree realizations for all allowed MLD and N_P numbers.

C. Assembly energy profiles

The next step is to construct the minimum energy assembly pathways and energy profiles for the spanning tree model. The initial state is a spanning tree molecule folded over the edges of a mathematical dodecahedron with no pentamers. Different spanning trees are assumed to have the same folding energy prior to the binding of pentamers. Next, pentamers are placed on the dodecahedron, one after the other. The energy of a cluster of n pentamers associated with a spanning tree is defined as

$$\Delta E(n)/E_0 = -[n_1 + n_2\epsilon + n_3(1+2\epsilon) + n\mu_0]. \quad (2.1)$$

Here, $n_1 \leqslant 11$ is the number of edges shared between two pentamers that are not covered by a spanning tree link. The corresponding affinity is denoted by minus E_0, as for the Zlotnick model. Next, n_2 is the number of pentamer edges that are in contact with a link of the spanning tree while the edge is *not* shared with another pentamer. The dimensionless number ϵ is here the ratio of the affinity of a pentamer edge with a spanning tree link over the affinity between two pentamer edges that are not in contact with a link. Finally, n_3 is the number of spanning tree links that lie along a pentamer edge that *is* shared with another pentamer. Interactions between edges and spanning tree links are assumed to be additive so the bond energy of such a link is $-(1+2\epsilon)E_0$. The assembly energy of a complete particle is equal to $\Delta E/E_0 = -[19(1+2\epsilon) + 11 + 12\mu_0]$ for all spanning trees, with, as before, μ_0 the reference pentamer chemical potential. Importantly, the assembly energy of complete particles does not depend on the class of spanning trees. Note that compared with the Zlotnick model ϵ is the only new energy parameter.

Examples of minimum-energy assembly profiles are shown in Fig. 6. The top figure shows the assembly energy profiles of $N_P = 8$, MLD = 9 and of $N_P = 2$, MLD = 19 spanning trees, both for $\epsilon = 0.5$. The reference chemical potential is close to that of assembly equilibrium ($\mu^* \simeq -4.083$). The activation energy barrier of $N_P = 8$, MLD = 9 spanning trees is about $2E_0$ lower than that of the $N_P = 2$, MLD = 19 spanning trees. The curious horizontal flatness of the middle section of the energy profile of $N_P = 8$, MLD = 9 spanning trees is accidental [when a pentamer is added on maximal wrapping sites making two new contacts with adjacent pentamers then the energy change in units of E_0 equals $-(2+4\epsilon+\mu_0) = 0$ for $\epsilon = 0.5$ and $\mu_0 = 4$]. Because for $N_P = 2$, MLD = 19 trees there are no adjacent maximal wrapping sites, the energy profile continues to increase with n over this interval for n less than five.

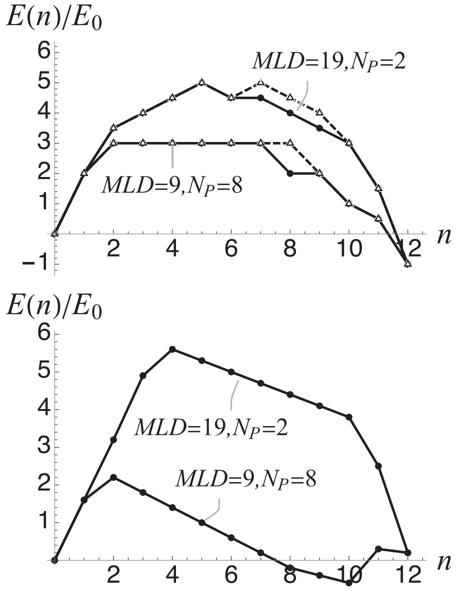

FIG. 6. Top: Minimum-energy assembly profiles for $N_P = 8$, MLD = 9 and $N_P = 2$, MLD = 19 spanning trees. The affinity ratio is $\epsilon = 0.5$ and the reference chemical potential is $\mu_0 = -4.0$. Energies are expressed in units of the overall scale E_0. Bottom: Same except that $\epsilon = 1.1$ and the reference chemical potential is $\mu_0 = -6.0$.

The bottom figure shows the same case except that $\epsilon = 1.1$ while the reference chemical potential is raised to $\mu_0 = -6.0$ in order to maintain a state that is close to assembly equilibrium. The difference between the assembly energy barriers has increased to about four units of E_0. This is expected since increasing ϵ increases the energy contrast between pentamer bonds that are and that are not lined by an RNA link. However, the minimum energy state of the $N_P = 8$, MLD = 9 spanning tree is now at $n = 10$. This means that the minimum energy state is a particle with two missing pentamers! This type of breakdown of the nucleation-and-growth assembly scenario becomes increasingly frequent as ϵ is raised beyond 0.5.

An important observation is that different spanning trees with the same N_P and MLD numbers *nearly all have the same energy profiles*. The minor changes of the energy profiles of the very few exceptional cases are shown in Fig. 6 (top). Since molecules with the same N_P and MLD have practically always the same assembly energy profiles, we can treat the group of spanning trees with the same N_P and MLD as belonging to a class of molecules characterized by a particular minimum energy assembly profile. The very large original degeneracy of the Zlotnick model thus has been lifted in terms of these different classes. Since the class size is now the number of configurations associated with a certain minimum energy assembly pathway, one might view the logarithm of the class size as the entropy of a particular assembly pathway.

D. Assembly trajectories

Next, we explored the space of all possible minimum energy assembly *trajectories*. The results are shown in Fig. 7.

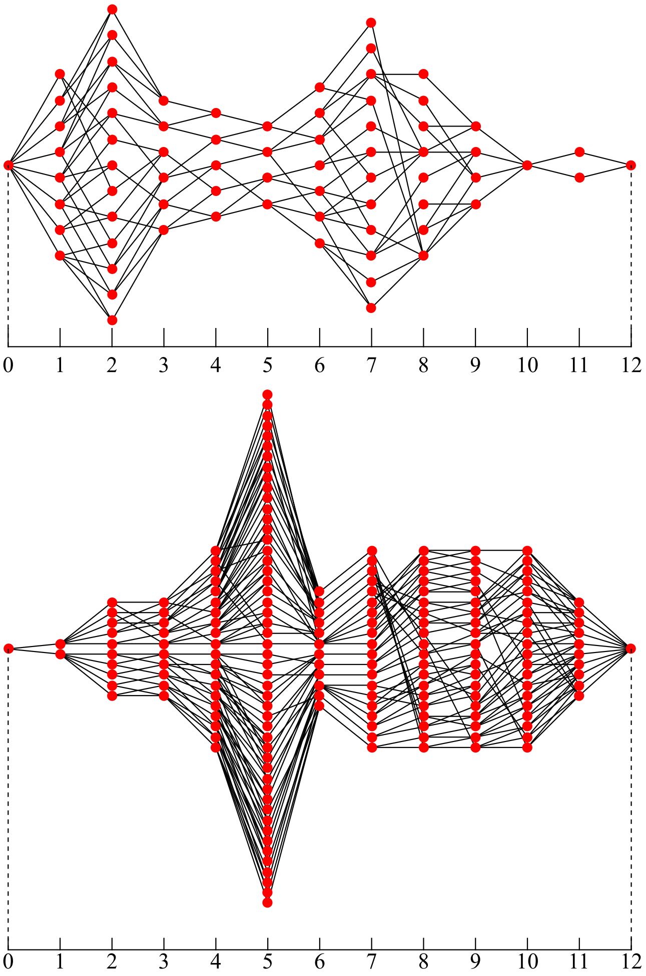

FIG. 7. Minimum-energy assembly trajectories of $N_P = 8$, MLD $= 9$ spanning trees (top) and $N_P = 2$, MLD $= 19$ spanning trees (bottom). Each dot marks a physically distinct intermediate structure with, from left to right, $n = 0, 1, \dots, 12$ pentamers. Every possible path from $n = 0$ to $n = 12$, including back steps, represents a possible minimum energy assembly pathway.

Each dot marks a physically distinct assembly intermediate. Assemblies related by a symmetry operation of the dodecahedron are being treated as the same. Assembly intermediates can be assigned "coordinates" (n, i) with $n = 0, 1, \dots, 12$ the number of pentamers of the intermediate state and with $i = 1, 2, \dots, m(n)$ ranging over the distinct n-pentamer states $m(n)$ the number of distinct n-pentamer intermediates. The multiplicities of the $N_P = 8$, MLD $= 9$ spanning trees and the $N_P = 2$, MLD $= 19$ spanning trees are shown in Fig. 8. The multiplicity of the $n = 5$ assembly intermediates of the $N_P = 2$, MLD $= 19$ spanning trees is about 10^2 times larger than that of the $N_P = 8$, MLD $= 9$ spanning trees.

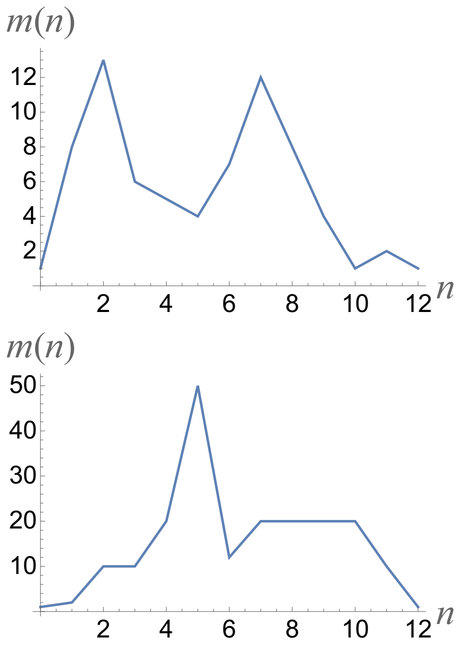

FIG. 8. Multiplicities $m(n)$ of the $N_P = 8$, MLD $= 9$ spanning trees (top) and the $N_P = 2$, MLD $= 19$ spanning trees (bottom).

A black line linking two intermediate states in Fig. 7 indicates that the two states can be interconverted by addition or removal of a pentamer. Assembly of viral particles can be viewed as a net "current" flowing from the $n = 0$ source state to the $n = 12$ sink state along all possible paths across the network linking the initial state to the final state. Under conditions of thermodynamic equilibrium, the current across each individual link should be zero according to the principle of detailed balance. Assuming that the assembly energy profiles of spanning trees with the same N_P and MLD are all the same allows us to define a "low temperature" Boltzmann distribution for the assembly intermediates of RNA molecules with a particular N_P and MLD:

$$P_B(n) \propto \exp -\Delta F(n), \tag{2.2}$$

where $\Delta F(n) = \beta \Delta E(n) - \ln m(n) - n \ln c_f(eq)$. The second term includes the entropic free energy associated with the multiplicity $m(n)$ of n-pentamer assemblies. The third term is the correction to the pentamer solution chemical potential for the general case that the equilibrium concentration of free pentamers $c_f(eq)$ differs from the reference concentration (which is our unit of concentration). With "low temperature" we mean here that we only include n-pentamer assembly intermediates that minimize the assembly energy for given n. In Appendix C we discuss the equilibrium phase behavior obtained from this Boltzmann distribution. It turns out to be typical of that of self-assembling systems in general. There is a critical pentamer concentration ("CAC") below which viral particles do not form and above which the particle concentration increases linearly with the pentamer concentration. The equilibrium phase diagram is only very weakly dependent on the MLD and N_P numbers so there is practically no selectivity under equilibrium conditions.

If ϵ is significantly increased above 0.5, in order to increase selectivity, then kinetic traps appear in the assembly energy profile while assembly intermediates start to undergo structural transitions, as discussed in Appendix D. Also, completed particles no longer are minimum free energy states, as we saw earlier. For this reason we will maintain in the following $\epsilon = 0.5$ as a reasonable choice.

III. KINETICS AND PACKAGING COMPETITION

In order to construct the kinetics, we start by characterizing the graph of the assembly pathways of spanning tree with given N_P and MLD in terms of an *adjacency matrix* $A_n^{i,j}$ for the intermediate states shown in Fig. 7. This adjacency matrix equals one if a link connects intermediate state (n, i) to state $(n+1, j)$ and zero if there is no link. Next, we define for each intermediate (n, i) of the network a time-dependent occupation probability $P_{i,n}(t)$. The assembly process is assumed to be Markovian with the probabilities $P_{i,n}(t)$ evolving in time according to the master equation [63]

$$\frac{dP_{i,n}(t)}{dt} = \sum_j \left\{A_{n-1}^{j,i} W_{n-1,n} P_{j,n-1}(t) + A_n^{i,j} W_{n+1,n} P_{j,n+1}(t)\right\} - P_{i,n}(t) \sum_j \left\{A_{n-1}^{j,i} W_{n,n-1} + A_n^{i,j} W_{n,n+1}\right\}. \tag{3.1}$$

Here, $W_{n,n+1}$ is the on rate for the transition of an intermediate with n pentamers to one with size $n+1$ by the addition of a pentamer while $W_{n,n-1}$ is the off rate at which a pentamer is removed from an assembly of size n. Physically, the assumption of Markovian kinetics means that the rate of change $P_{i,n}(t)$ is completely determined by the occupation probabilities at time t. We will assume a simplified diffusion-limited chemical kinetics [64] in which the addition or removal of a pentamer to an assembly of size n is treated as a bimolecular reaction with an on rate that has the form of a kinetic Monte Carlo algorithm:

$$W_{n,n+1} = \lambda c_f(t) \begin{cases} e^{-\Delta\Delta E_{n,n+1}} & \text{if} \quad \Delta E(n+1) > \Delta E(n), \\ 1 & \text{if} \quad \Delta E(n+1) < \Delta E(n). \end{cases} \tag{3.2}$$

The concentration $c_f(t)$ of free pentamers is in general time dependent, and different from the reference concentration because assembly of capsids reduces the concentration of free pentamers. The on rates are thus time dependent as well. Next, $\Delta\Delta E_{n,n+1} = \Delta E(n+1) - \Delta E(n)$ is the energy cost of adding a pentamer while λ is a base rate that depends on molecular quantities such as diffusion coefficients and reaction radii but not on the pentamer and RNA concentrations. The inverse of λ is the fundamental timescale of the kinetics. In the following, time will be expressed in units of $1/\lambda$. If $\Delta\Delta E_{n,n+1}$ is negative, then the on rate is equal to this base rate while if $\Delta\Delta E_{n,n+1}$ is positive, then the base rate is reduced by the Arrhenius factor $e^{-\Delta\Delta E_{n,n+1}}$.

The off-rate entries $W_{n+1,n}$ are determined in part by the condition that in the long-time limit the occupation probabilities $P_{i,n}(t)$ must approach the equilibrium Boltzmann distribution (2.2). This imposes the condition of detailed balance $\frac{W_{n,n+1}}{W_{n+1,n}}|_{t\to\infty} = \frac{P_B(n+1)}{P_B(n)} = c_f(eq)\, e^{\Delta\Delta E_{n,n+1}}$. Separately we also demand, on physical grounds, that the off rates for the release of a protein from a cluster should be independent of the solution concentration of free pentamers. Both conditions are satisfied by imposing

$$\frac{W_{n,n+1}}{W_{n+1,n}} = c_f(t) e^{\Delta\Delta E_{n,n+1}}. \tag{3.3}$$

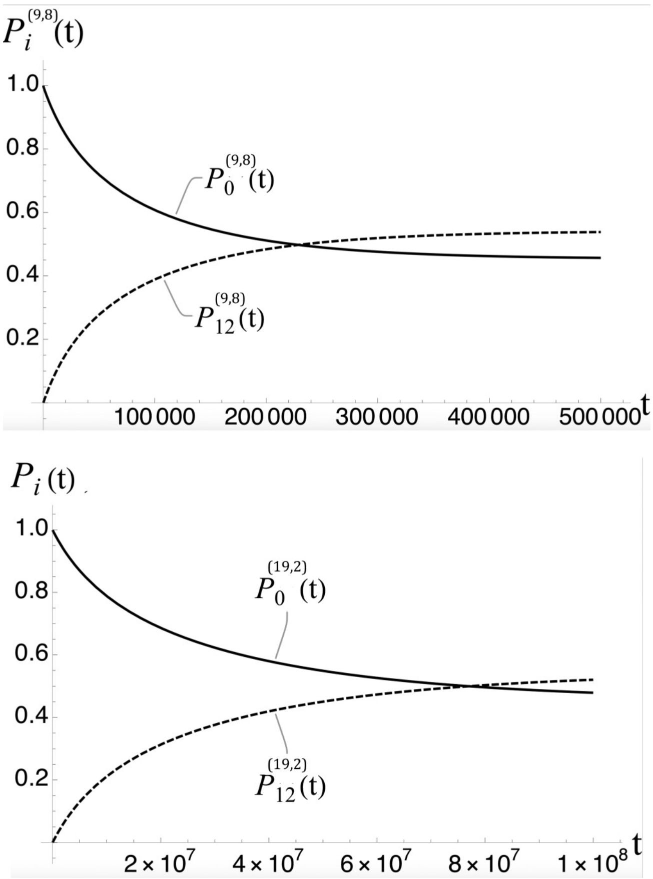

FIG. 9. Top: Packaging kinetics of MLD = 9 and $N_P = 8$ spanning trees. Parameter values are $E_0 = 4k_bT$ for $\epsilon = 0.5$, $c_0 = 1$, $D = 0.5$, and $\mu_0 = -4$. Bottom: Packaging kinetics of MLD = 19, $N_P = 2$ spanning trees for the same parameters.

The definition of the kinetics is completed by noting that ratio of the free pentamer concentration $c_f(t)$ over the total time-independent pentamer concentration c_0 is determined by pentamer number conservation:

$$c_f(t)/c_0 = 1 - (D/12) \sum_{n=0}^{12} \left(\sum_{i=1}^{m(n)} n P_{i,n}(t) \right). \tag{3.4}$$

Here, $D \equiv 12 r_t / c_0$, with r_t the total RNA concentration, is the RNA to protein *mixing ratio*. If $D = 1$, then there are exactly enough pentamers to encapsidate all spanning trees, which corresponds to the *stoichiometric ratio*. Because all occupation probabilities enter in the relation for $c_f(t)$ that itself enters in all 13 equations, the rate equations form a coupled set of nonlinear differential equations. In the following subsections these equations are solved by numerical integration using *Mathematica*.

A. Timescales

Figure 9 shows the packaging kinetics of RNA molecules that represent maximal (top), respectively, minimal (bottom) packaging adaptation. For the overall energy scale we used a value $E_0 = 4k_bT$, close to that of the pentamer-pentamer affinity of the Zlotnick model for empty capsids. The ratio ϵ between RNA and pentamer to pentamer and pentamer interaction was set to $\epsilon = 0.5$. Next, the total pentamer concentration was set to $c_0 = 1$ and the mixing ratio to $D = 0.5$, which means that there are twice as many pentamers as would be necessary to package all RNA molecules. Finally, we set the reference chemical potential to $\mu_0 = -4$, which is close to the assembly equilibrium chemical potential.

In both cases, thermal equilibrium is reached in the $t \to \infty$ limit. The fractions of RNA molecules being packaged are roughly equal (about 66%), reflecting the fact that the two classes of molecules have the same assembly energy. The remaining small difference is due to the fact that the entropy of the assembled particles for the two classes is not exactly the same because of differences between the energy cost of removing one or two pentamers under the action of thermal fluctuations. The reason that a significant fraction of RNA molecules are *not* being packaged, despite the fact that there are more than enough pentamers to package all RNA molecules, reflects the fact that the chemical potential is close to assembly equilibrium. The global shapes of the time dependence of the occupation probabilities are similar but the timescales are quite different, about 10^5 time units for MLD = 9 and $N_P = 8$ spanning trees and about 10^7 time units for MLD = 19, $N_P = 2$ spanning trees. This difference reflects the fact that the assembly activation barrier is about $2E_0$ larger (so about $8k_bT$) for the MLD = 19, $N_P = 2$ spanning trees.

In order to compute relaxation times, one first completes the definition of the rate matrices by introducing the diagonal entries $W_{n,n} = -\sum_{m\neq n} W(m,n)$. The resulting matrix $W_{m,n}$ now has column elements adding to zero. Using this completed transition matrix, the master equation can be rewritten in the form of the matrix equation $\frac{d\mathbf{P}}{dt} = \mathbf{WP}$. This looks like a linear rate matrix equation but because the concentration of free pentamers is self-consistently dependent on all occupation probabilities through Eq. (3.4), the rate matrix $\mathbf{W}$ itself depends on the occupation probabilities. However, in the long-time limit when the system is close to thermal equilibrium, one can replace the occupation probabilities in Eq. (3.4) by the equilibrium Boltzmann probabilities. The equation $\frac{d\mathbf{P}}{dt} = \mathbf{WP}$ then does reduce to a linear rate matrix equation, which can be solved by standard matrix diagonalization methods. The eigenvalues of the rate matrix are the late time decay rates of the various modes that correspond to the eigenvectors. The thermalization time t_r is the inverse of the smallest eigenvalue of $W_{m,n}$. This gives $t_r \simeq 3.26 \times 10^5$ for the MLD = 9, $N_P = 8$ spanning trees and $t_r \simeq 3.4 \times 10^7$ for the MLD = 19, $N_P = 2$ spanning trees, consistent with the numerical results.

This thermalization time can be compared with the early-time assembly *delay time* t_d, i.e., the time lag between the establishment of solution assembly conditions and the first appearance of assembled viral particles. We obtain t_d from the intersection of the tangent to $P_{12}(t)$ at the point of maximum slope with the horizontal axis (see Fig. 10). For the case of the MLD = 9 and $N_P = 8$ class of spanning trees, this gives about 8.5 time units, so four to five orders of magnitude smaller than the thermalization time. Other classes have comparable delay times. Measured delay times for the assembly of empty capsids are in the range of minutes [36–38], which indicates that the time unit $1/\lambda$ is in the range of 1–10 s. The thermalization time under conditions of assembly equilibrium would then be in the range of 200 h for MLD = 9, $N_P = 8$ spanning trees and two orders of magnitude longer for the MLD = 9, $N_P = 8$ spanning trees. In actuality, *in vitro* assembly experiments are carried out on supersaturated solutions. When the reference pentamer chemical potential μ_0 is raised to −3.6, which corresponds to a moderate level of supersaturation, the thermalization time reduces to a, more reasonable, 8.3×10^3 time units while the delay time remains about the same.

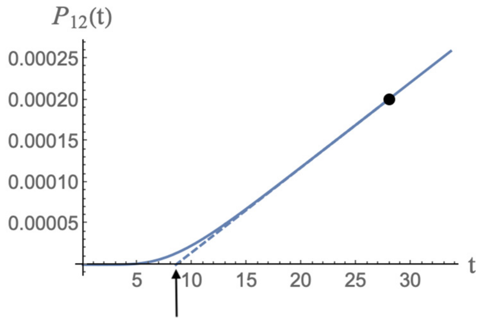

FIG. 10. Definition of the delay time as the intersection of the tangent to the assembly curve $P_{12}(t)$ with maximum slope with the time axis. The parameters are MLD = 9, $N_P = 8$, $\epsilon = 0.5$, $c_0 = 1$, $D = 0.5$, and $\mu_0 = -4$.

B. Packaging competition

The kinetic equations can be extended to the case of packaging competition in a solution that contains equal amounts of two types of spanning trees, say (1) and (2), that are competing for pentamers. The two occupation probabilities $P_{i,n}^{(1,2)}(t)$ both obey a set of 13 master equations for the respective 13 occupation probabilities. These two sets of equations are then coupled because the same free pentamer concentration appears in both sets of equations. This free pentamer condition is, in turn, determined by the condition of pentamer number conservation, which now takes the form

$$c_f(t)/c_0 = 1 - (D/24)\sum_{n=0}^{12}\left(\sum_{i=1}^{m_n^{(1)}} nP_{i,n}^{(1)}(t) + \sum_{i=1}^{m_n^{(2)}} nP_{i,n}^{(2)}(t)\right). \tag{3.5}$$

In Fig. 11, we show the outcome of a packaging competition experiment with the same total amount of RNA molecules and pentamers as before but now with half of the RNA molecules being MLD = 9 and $N_P = 8$ spanning trees and the other half MLD = 19, $N_P = 2$ spanning trees. The top and middle sections of Fig. 11 show that the fraction packaged MLD = 9, $N_P = 8$ molecules is significantly larger than the fraction of packaged MLD = 19, $N_P = 2$ molecules up to about 10^6 time units. About 80% of the MLD = 9, $N_P = 8$ molecules are packaged at that time, as against only a few percent of the MLD = 19, $N_P = 2$ spanning trees. For later times,

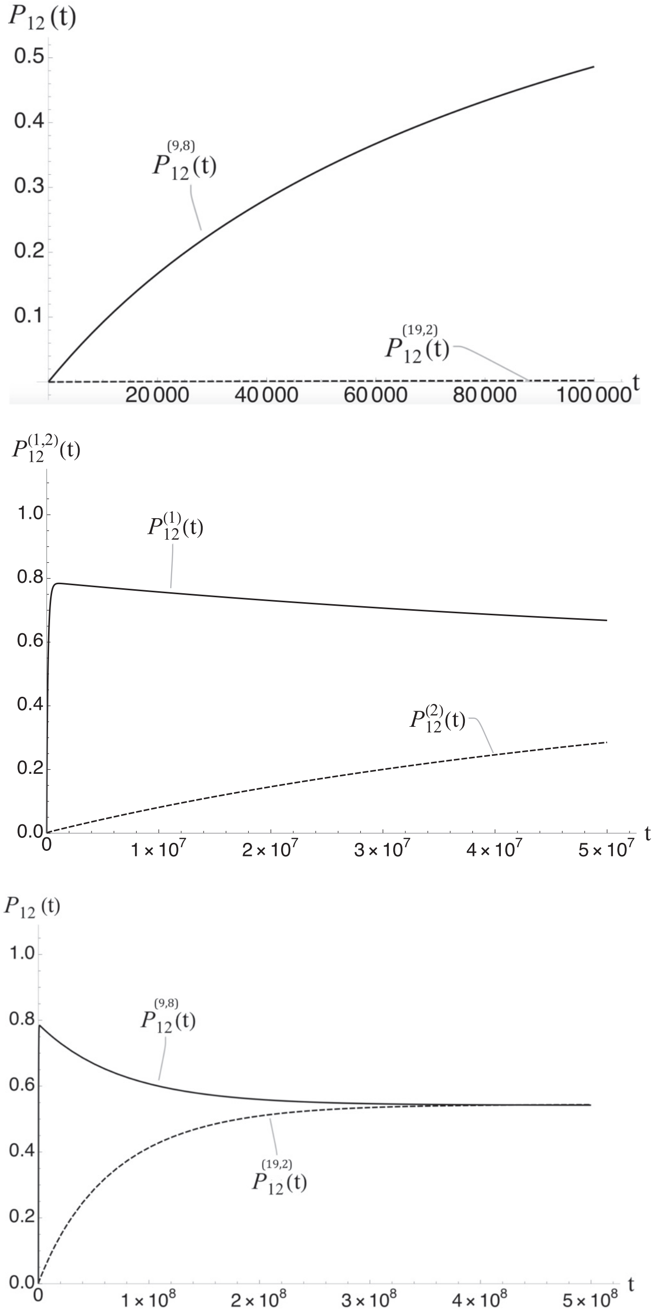

FIG. 11. Packaging competition between MLD $= 9$, $N_P = 8$ spanning trees and MLD $= 19$, $N_P = 2$ spanning trees with the same parameters as the previous figure. Top: timescale 10^5 units; middle: timescale 10^7; bottom: timescale 10^8 units.

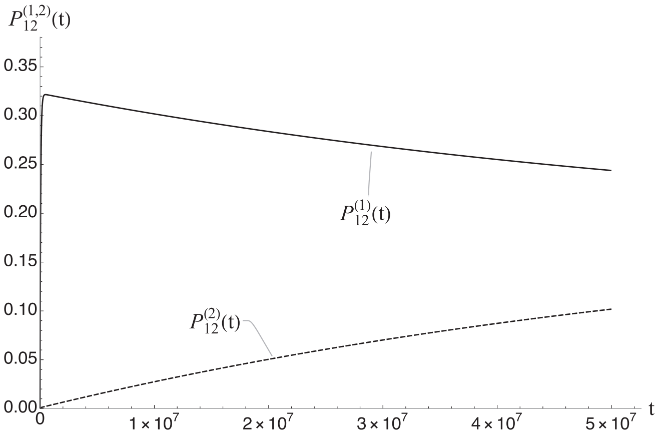

FIG. 12. Packaging competition between MLD $= 9$, $N_P = 8$ spanning trees and MLD $= 19$, $N_P = 2$ spanning trees with the same parameters as the previous figure except that the mixing RNA to protein mixing ratio has been increased to $D = 2$.

the packaged fraction of MLD $= 9$, $N_P = 8$ spanning trees starts to decrease slowly. This means that fully assembled MLD $= 9$, $N_P = 8$ particles are *disassembling*. The fraction of packaged MLD $= 19$, $N_P = 2$ spanning trees increases correspondingly: pentamers freed up by disassembly of MLD $= 9$, $N_P = 8$ spanning trees are being used to feed the assembly of the MLD $= 19$, $N_P = 2$ spanning trees. The bottom figure shows the eventual approach to thermal equilibrium when the packaging fractions of the two classes are nearly the same. The fact that during packaging competition the disassembly of completed particles is an essential step in the final approach to thermal equilibrium will play a key role in the following.

We now can explore under which conditions kinetic selectivity is maximized, both in terms of the ratio between packaging fractions and in terms of the late time persistence of the selectivity. One key quantity turns out to be the mixing ratio D. For the $D = 0.5$ value that we have been using until now, there was a significant excess of pentamers. We reasoned that if the early packaging of MLD $= 9$ and $N_P = 8$ particles would deplete the available pentamers then increasing D might "starve" the subsequent assembly of MLD $= 19$, $N_P = 2$ spanning trees, which could extend the time interval over which the packaging of MLD $= 9$, $N_P = 8$ spanning trees dominates. We tested this for the case of $D = 2$. In that case there would be only enough pentamers to package half of all RNA molecules. Figure 12 shows what happens. The fraction of packaged MLD $= 19$, $N_P = 2$ spanning trees at 5×10^7 time units does decrease, from about 0.25 to about 0.12, but the fraction of packaged MLD $= 9$, $N_P = 8$ spanning trees *also* decreases, from about 0.7 to about 0.27. The increase of the mixing ratio increased only marginally the packaging fraction ratio.

We then reasoned that the starvation effect would be enhanced under conditions of supersaturation since that reduces the concentration of free pentamers. In Fig. 13 we show the result of reducing the reference chemical potential from $\mu_0 = -4.0$ to -3.4. The top and middle figures show the minimum-energy assembly profiles of the two classes. As expected, the assembly activation energy has decreased significantly, by about $3E_0$ for the first class and by about $5E_0$ for the second class. Both are now about $2E_0$. Because the two activation energy barriers are similar, one might expect that the kinetic selectivity actually is *weakened* by supersaturation. The bottom figure shows that the opposite is true: supersaturation greatly increases packaging selectivity! On a timescale of about 5×10^7 time units, the fraction of MLD $=$ 9, $N_P = 8$ spanning trees has increased from about 0.32 back up to about the 0.81 of Fig. 11. The packaging ratio is high and, importantly, the growth rate of the packaged fraction

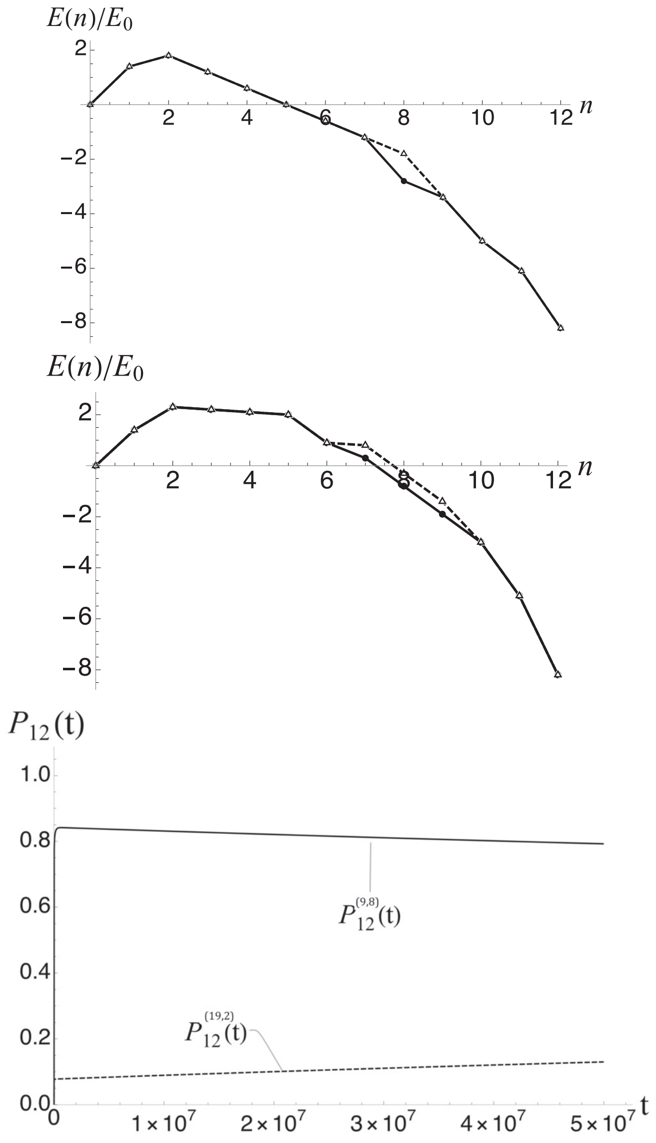

FIG. 13. Top and middle figures: Effect on the assembly energy profiles of reducing the reference chemical potential to a state of supersaturation with $\mu_0 = -3.4$. Bottom: Packaging competition between MLD = 9, $N_P = 8$ spanning trees and MLD = 19, $N_P = 2$ spanning trees for $\mu_0 = -3.4$. Other parameters are the same as those of Fig. 12.

of MLD = 19, $N_P = 2$ particles at 5×10^7 time units has reduced from about 10^{-8} inverse time units to about four parts in 4×10^{-10} inverse time units. The disassembly of complete particles has greatly slowed down. The reason is that supersaturation increases the energy barrier for the disassembly of completed particles to about $10E_0$ (for both classes, see Fig. 13). Following the early assembly of the MLD = 9, $N_P = 8$ spanning trees, there were few free pentamers left in solution since for $D = 2$ there are just enough pentamers to package the MLD = 9, $N_P = 8$ spanning trees. Then, in the absence of much disassembly of the MLD = 9, $N_P = 8$ particles, MLD = 19, $N_P = 2$ spanning trees indeed are starved of pentamers. Note that this "monopoly mechanism" could never work under conditions of assembly equilibrium since there would always be a significant fraction of free pentamers in that case.

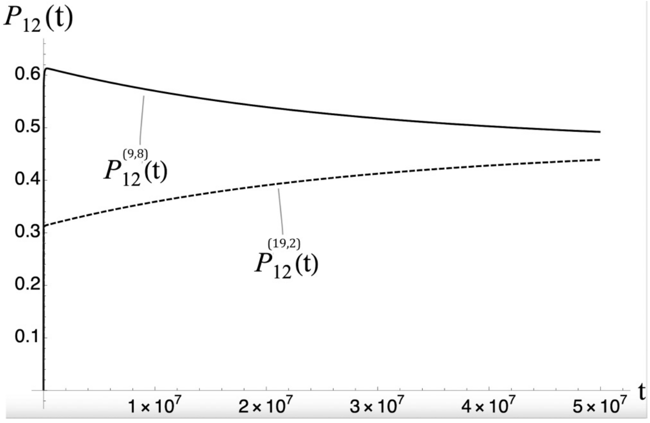

FIG. 14. Packaging competition between MLD = 9, $N_P = 8$ spanning trees and MLD = 12, $N_P = 5$ spanning trees for $\mu_0 = -3.4$. The parameters are the same as those of Fig. 13.

This result can be understood by noting that the height of an activation energy barrier does not fully characterize the rate of barrier crossing. For the MLD = 9, $N_P = 8$ spanning trees, the $n = 2$ state does function as a true transition state since there is a substantial energy drop for the $n = 3$ state (and larger states) as well as for the $n = 1$ state (see Fig. 13, top). However, for the MLD = 19, $N_P = 2$ trees, the $n = 2$ state is not really a transition state as the whole interval between $n = 2$ and $n = 5$ has roughly the same energy (see Fig. 13, middle). The probability that a cluster of size $n = 3$, 4, and 5 can "fall back" to the $n = 1$ state remains quite large. Kinetic selection in favor of the MLD = 9, $N_P = 8$ spanning trees thus remains quite effective under conditions of moderate supersaturation even if the barrier heights are similar.

How about packaging competition between a maximally adapted MLD = 9, $N_P = 8$ spanning tree and other spanning trees? Figure 14 shows the case where MLD = 9, $N_P = 8$ spanning trees compete with MLD = 12, $N_P = 5$ spanning trees. The kinetic selectivity in favor of the MLD = 9, $N_P = 8$ spanning trees is reduced to only a factor of 2. Recall that if all possible spanning trees would have the same *a priori* probability then, in a solution containing all possible spanning trees, the overwhelming majority of spanning trees would have MLD and N_P numbers in the range of MLD = 12, $N_P = 5$. The entropic bias in favor of spanning trees in the MLD = 12, $N_P = 5$ range would completely erase the kinetic selectivity effect. The kinetic selection mechanism "works" only for competition between two different spanning trees with comparable concentrations.

IV. TWO-STAGE ASSEMBLY

The master equation was based on a protein-by-protein assembly scenario. This is not the only option: numerical simulations of coarse-grained model systems [57,65] reported that a collective assembly process, called the *en-masse* scenario, also can take place. This scenario was encountered for higher values of protein-genome affinity as compared to the protein-protein affinity. The genome molecules initially are in a swollen state due to, say, electrostatic self-repulsion between the negative charges, and free of capsid proteins. Next, when this swollen genome molecule starts to capture capsid proteins, a disordered nucleoprotein condensate forms. As the number of captured proteins increases, the condensate shrinks as both positive capsid protein charges and negative genome charges are increasingly neutralizing each other. Finally, some form of spatial ordering of the capsid proteins produces a viral particle (order-disorder transitions where a spherically symmetric condensate develops broken rotational symmetry can be described by Landau theory applied to viral assembly [66]). Other modes of collective assembly have been proposed as well [67]. Experiments on the encapsidation of linear *double-stranded* genome molecules by capsid proteins [68] have been interpreted according to this en-masse scenario as have assembly studies of the CCMV virus [69].

Can we use the spanning tree model to test for RNA selectivity in an en-masse scenario? Because in the spanning tree model the genome molecule is compactified right from the start, it cannot "capture" the transition from a swollen to a collapsed state. However a fascinating *in vitro* assembly experiment that mimics the en-masse scenario was carried out in Ref. [70]. During a first stage, the pH level was set at a level at which the protein-RNA affinity was large with respect to the protein-protein affinity. Disordered and incomplete assemblies were observed. In a second stage, the pH level was set at a level such that the protein-protein affinity was increased with respect to the protein-RNA affinity. The disordered condensates of the first stage transformed into viruslike particles. The selectivity for such a two-stage assembly scenario *can* be examined within the model by performing two subsequent assembly calculations. During the first stage, the assembly energy profile is set to zero in order to mimic a state dominated by entropy. This disordered state is then used as the *initial state* for a second assembly calculation, but now with the energy parameters set at the values we used earlier (such as those of Fig. 13).

Figure 15 shows the first-stage occupation probabilities of the same two classes as before (see Fig. 13). The first-stage occupation probabilities are time independent and consistent with a equilibrium Boltzmann distribution determined by the multiplicities of the assembly intermediates. Next, Fig. 16 shows what happens when at $t = 20\,000$ the energy parameters are reset to the values of Fig. 13. There is an instant and complete reorganization. The intermediate-sized clusters produced during the first stage disappear, leaving behind fully assembled particles plus free RNA molecules of both classes. Right at $t = 20\,000$ the fraction of packaged MLD $= 19$, $N_P = 2$ molecules exceeds the fraction of MLD $= 9$, $N_P = 8$ molecules, which is a consequence of the fact that before the reorganization there were more clusters on the MLD $= 19$, $N_P = 2$ molecules with $n = 4$ and 5 (see Fig. 15). After the reorganization, these clusters rapidly slid down the slope of the assembly curve of Fig. 13 (top and center) towards completion. For later times there is some continued particle assembly because there is a substantial concentration of free pentamers and unoccupied RNA molecules at $t = 20\,000$. Because the width of the assembly barrier of MLD $= 9$, $N_P = 8$ molecules is significantly smaller than that of MLD $=$ 19, $N_P = 2$ molecule (see again Fig. 13) this leads nearly exclusively to formation of additional MLD $= 9$, $N_P = 8$ particles and thus some kinetic selectivity in favor of the MLD $= 9$, $N_P = 8$ molecules. The selectivity produced by the second stage of the assembly process is, however, quite weak compared to that of the one-stage assembly process discussed in the previous subsection. It should be noted that the two-stage assembly scenario very much speeds up the formation of assembled particles: if assembly speed instead of selectivity would be a central aim, then two-stage assembly could well be more efficient than one-stage assembly.

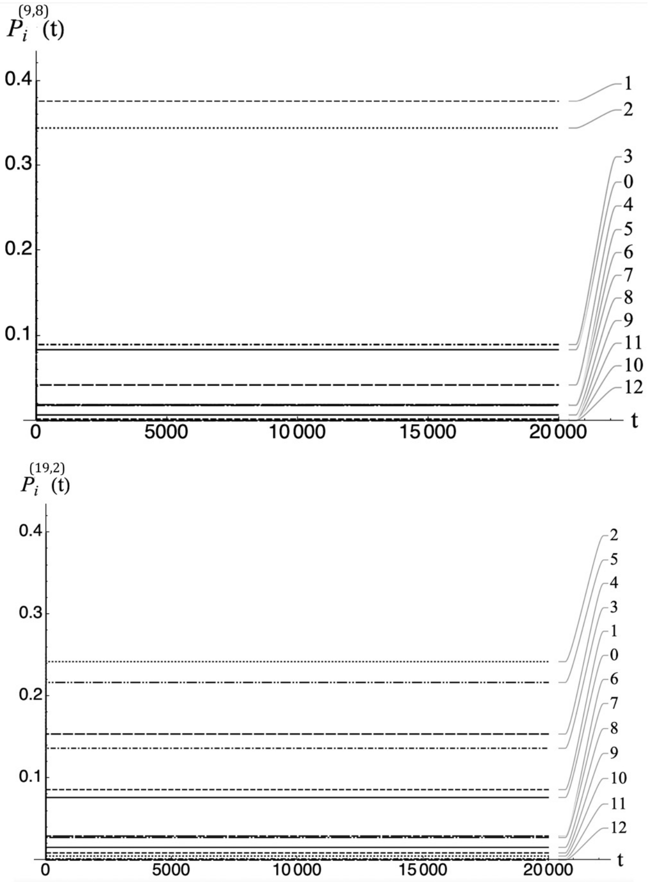

FIG. 15. First-stage assembly competition between MLD $= 9$, $N_P = 8$ and MLD $= 19$, $N_P = 2$ molecules. The two figures show the first-stage occupation probabilities of the MLD $= 9$, $N_P = 8$ molecules and of the MLD $= 19$, $N_P = 2$ molecules.

V. CONCLUSION

In the Introduction we asked about the physical properties of a gRNA selection mechanism that operates during the formation of a nucleation complex. The spanning tree model, a drastic simplification in which the huge number of possible RNA configurations is reduced to the, roughly, 10^5 different spanning trees of a dodecahedron, was used to address this question. The spanning trees, all of which have 19 links, are supposed to represent a cluster of packaging signals located on the outer surface of a condensed RNA molecule, as inspired by the asymmetric reconstruction of the packaged RNA of the MS2 virus. Different spanning trees were classified according to their maximum ladder distance and their wrapping number

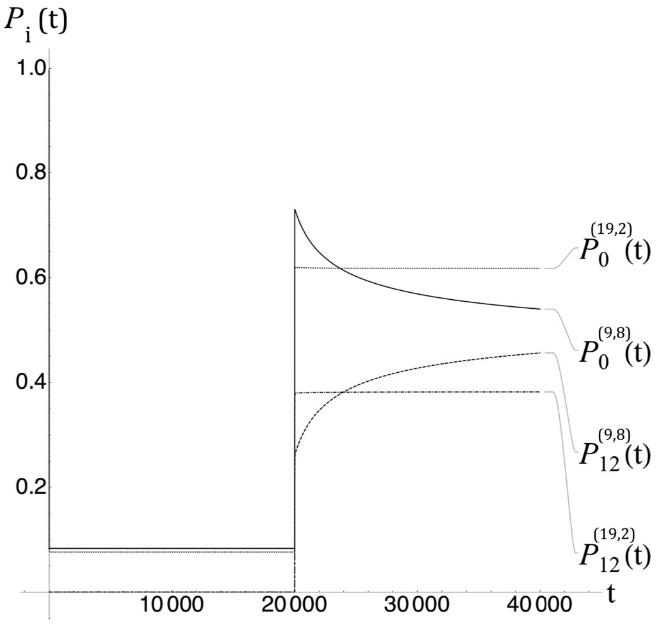

FIG. 16. First- and second-stage occupation probabilities. At $t = 20\,000$, the energy parameters were reset to the values of Fig. 13. For the second stage, only the occupation probabilities of unoccupied ($n = 0$) and fully occupied ($n = 12$) particles are shown as the occupation probabilities for assembly intermediates during the second stage are negligible. The fraction of packaged particles containing MLD = 19, $N_P = 2$ molecules is time independent while the fraction of packaged particles containing MLD = 9, $N_P = 8$ molecules increases moderately with time.

such that different spanning trees belonging to the same class have the same (or nearly the same) minimum energy assembly profile. By carrying out numerical packaging competition experiments, we found that spanning trees that belong to a class with small MLD and large N_P are kinetically selected over spanning trees with large MLD and small N_P numbers. This selectivity is eventually lost when the system reaches a state of thermodynamic equilibrium. With only a modest amount of "tweaking" of the system parameters, it was possible to enhance quite significantly the kinetic selectivity and its persistence.

The model is not a realistic description of any particular virus, so direct quantitative validation is not possible. There are, however, general consequences that, we believe, will carry over to more realistic models and that can be tested experimentally: (i) RNA selectivity should be significantly more efficient under conditions of moderate supersaturation as compared to conditions of assembly equilibrium; (ii) RNA selectivity should degrade progressively as the strength of the affinity between capsid proteins is decreased (for example, by changes in the pH); (iii) RNA selectivity should be significantly weakened under conditions of collective assembly. Finally, the basic assumptions that underlie the model imply that there should be a strong correlation between the degree of RNA selectivity and the degree of internal order of the packaged RNA.

The notion that RNA selection is a kinetic effect conflicts with the common assumption that assembled viral particles are in a state of full thermodynamic equilibrium. Is the model even consistent with what is already known about viral self-assembly? Virions are typically assembled in the cytoplasm of infected cells where there is a significant concentration of capsid proteins. Assembled virions are then released from infected cells into environments that have virtually no free capsid proteins. Capsid proteins freed in a solution that does not contain capsid proteins should, for reasons of entropy, have a lower chemical potential than capsid proteins that are still part of viral particles, so virions would be expected to disassemble. This is, of course, not what is observed. Even empty capsids do not disassemble spontaneously in (capsid) protein-free solutions unless there is also a significant change in the thermodynamic parameters of the environment (e.g., a change in pH or salinity). Virions should be considered to be in a state of *constrained* thermodynamic equilibrium in which spontaneous disassembly does not take place on laboratory timescales. In terms of the spanning tree model, this justifies a focus on kinetic selection on timescales shorter than the final thermodynamic equilibration time. Recall that final equilibration for packaging competition calculations was signaled by the spontaneous disassembly of assembled particles (see Fig. 11).

In the Introduction it was noted that considerations based on free energy minimization indicate that the compactness of an RNA molecule should be an important criterion for packaged selection. How does this relate to the spanning tree model? Assume that an RNA molecule has a configuration of PS that significantly reduces the assembly free energy barrier because the MLD of the cluster of PS is low while the wrapping number is high. Now assume also that the MLD *of the whole molecule* is too large to allow for packaging. Under those conditions, this molecule might initially be kinetically selected but completion of the assembly is not possible because the free energy cost of compacting the whole molecule is too high. This aspect of the assembly of a viral particle was suppressed in the spanning tree molecule by the assumption that all molecules were compactified into identical dodecahedra. The only difference between the tree molecules was the organization of the PS sequence that decorates the outer surface of the dodecahedron.

The spanning tree model can be extended to include this aspect by letting the equilibrium size of the initial dodecahedron depend on the compactness (i.e., the MLD) of the whole molecule *and* by letting the dodecahedron be *deformable*. When a pentamer attempts to bind to the dodecahedron then, in general, it has to do some extra work by locally deforming the dodecahedron so one of the facets of the dodecahedron matches the size of the pentamer. As more and more pentamers adhere to the dodecahedron it will shrink (or swell), which means that this extra work per pentamer is progressively reduced. Now, the assembly free energy of particles with different overall MLD indeed *will* be different. Note that the cooperativity of this effect will sharpen the onset of viral assembly as a function of the pentamer chemical potential. It would be interesting to investigate how in such a model kinetic selectivity would compete with selectivity based on free energy minimization.

There are numerous examples in cell and molecular biology where kinetic selection is more effective than selection based on equilibrium thermodynamics. A well-known case is the fidelity of DNA duplication during cell division, which is much higher than what is expected based on thermodynamic equilibrium considerations. During DNA duplication, nuclease activity attempts to break the bond between newly formed base pairs [71,72]. Since mispairing is associated with weaker bonds, the fraction of Watson-Crick paired bonds that survive the challenge intact is much larger than that of the mispaired bonds. Kinetic selection by active challenging is known as the Hopfield proofreading mechanism [73] and it intrinsically consumes free energy. Does kinetic proofreading apply to the spanning tree model and, if so, what is the corresponding free energy source for the proofreading? Consider the initial formation of a small RNA-associated pentamer assembly intermediate where an MLD $= 9$, $N_P = 8$ assembly is viewed as proper pairing and MLD $= 19$, $N_P = 2$ assembly as a form of mispairing. Thermal fluctuations challenge both states. Under the conditions of supersaturation shown in Fig. 13, MLD $= 9$, $N_P = 8$ assemblies of size three or larger will slide down the assembly energy profile and form assembled particles on a quasi-irreversible basis but a small MLD $= 19$, $N_P = 2$ assembly located somewhere on the wide and flat activation barrier is repeatedly challenged against disassembly by thermal fluctuations. There is a high probability it will disassociate. The free energy source that allows for a quasipermanent increased selectivity is provided here by the supersaturated environment.

ACKNOWLEDGMENTS

We would like to thank A. Grosberg for drawing our attention to spanning trees and I. Rouzina for introducing us to the concept of selective nucleation. We would like to thank R. Twarock, C. Knobler, and W. Gelbart for reading a first draft and commenting on it. We also benefited from discussions with C. Lin, Z. Gvildys, and W. Vong. R.B. would like to thank the NSF-DMR for continued support under CMMT Grant No. 1836404.

APPENDIX A: DEMONSTRATION THAT THE SMALLEST MLD FOR SPANNING TREES ON THE DODECAHEDRON IS NINE

We begin by noting that for every vertex on the dodecahedron there is a vertex on the opposite side of the polyhedron that is a ladder distance five away. That is, getting from one of the two vertices to the other requires traversing at least five edges. Figure 17 shows such a path. For each such pairs of vertices there are 12 minimal paths.

Now, assume that there is a spanning tree with MLD 8. In such a case, we can pick out a path of ladder distance eight in that tree. All other elements of the tree will consist of trees that branch out from that path. Figure 18 is a figurative depiction of the path along with the longest allowed branch sprouting off each vertex on that path. The likelihood of branching off those "side branch" paths is ignored; such branching does not alter the argument below.

Consider first the central vertex on the ladder distance 8 path, labeled 5 in Fig. 18. The side path with ladder distance 4 is the longest that can attach to it. A longer path increases the MLD of the tree. Clearly, there is no possibility of reaching a point a ladder distance 5 from vertex 5 along any path with ladder distance 4, so the path shown cannot connect the central vertex to the vertex a distance 5 away from it. Next, consider the two sites flanking the central vertices, labeled 4 and 6. Attached to each is the longest possible path branching out from them, Such a path has ladder distance 3. If either of these paths reached to the vertex a ladder distance 5 away from the central vertex, then there would be a ladder distance 4 (or less) path from that vertex through one of the flanking vertices to the maximally separated vertex, and we know that no such path exists. We can continue this argument to encompass all allowed paths sprouting from vertices on the chosen path. Thus, there is a vertex on the dodecahedron that cannot be a part of the MLD 8 tree containing this path. Consequently, no tree with MLD 8 can be a spanning tree on the dodecahedron. The argument above can clearly be applied to the possibility of a spanning tree with MLD less than 8. That there is a spanning tree with MLD 9 is readily established by construction.

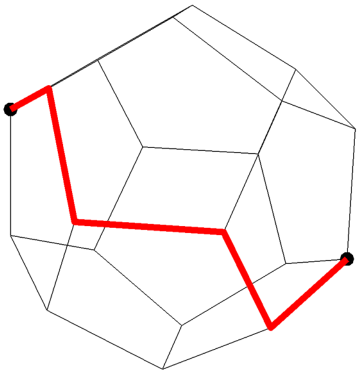

FIG. 17. Two maximally separated vertices on the dodecahedron and one of the 12 shortest paths consisting of five edges that join them.

APPENDIX B: PROJECTED MULTIPLICITIES

Figure 19 is a plot of the number of spanning trees of the dodecahedron as a function of the MLD for all possible values of the N_P number. Two spanning trees of the dodecahedron that are related by a symmetry operation of the dodecahe-

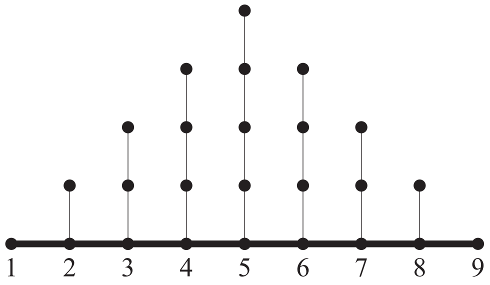

FIG. 18. A ladder distance 8 path in the hypothetical MLD 8 spanning tree on the dodecahedron. The path is shown as a thick line, and the nine vertices are labeled for easy reference. The thinner vertical lines represent longest allowed paths branching off the ladder distance 8 path.

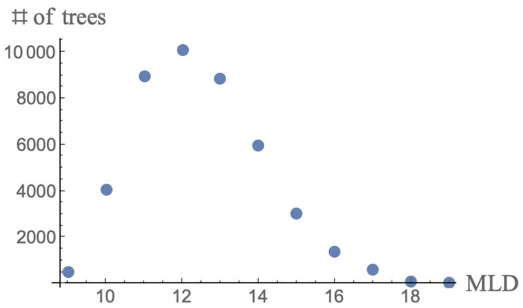

FIG. 19. The number of spanning trees on the dodecahedron as function of the maximum ladder distance (MLD).

dron are treated as the same. The converse, the distribution of wrapping numbers for all possible values of the MLD, is shown in Fig. 20. The same spanning tree can be distributed over the dodecahedron in different ways with different wrapping numbers. The wrapping number is thus not a topological characteristic of the secondary structure.

APPENDIX C: BOLTZMANN DISTRIBUTION

In this Appendix we discuss the equilibrium phase behavior of the model for a Boltzmann distribution:

$$P_n = \frac{\exp -\Delta F(n)}{Z}. \tag{C1}$$

Here $\Delta F(n) = \beta\Delta E(n) - \ln m(n) - n \ln c_f$ is the dimensionless free energy and $Z = \sum_{n=0}^{12} \exp -\Delta F(n)$ the partition sum. We will assume a solution containing only one class of RNA molecules with total concentration r_t as well as pentamers with a total concentration c_0. The concentration r_n of particles containing n pentamers is then $r_t P_n$. Finally, c_f is the concentration of free pentamers and r_f the concentration of unoccupied RNA molecules. Because $m(12) = 1$

$$r_{12}/r_t = \frac{\exp -\Delta E(12)}{Z}, \qquad r_f/r_t = \frac{1}{Z} \tag{C2}$$

for $c_f = 1$ (i.e., the reference concentration). Using these relations, it can be checked that

$$\frac{c_f^{12} r_f}{r_{12}} = K \tag{C3}$$

with $K = \exp \Delta E(12)$. This relation has the form of the *law of mass action* (LMA) of physical chemistry with K the dissociation constant.

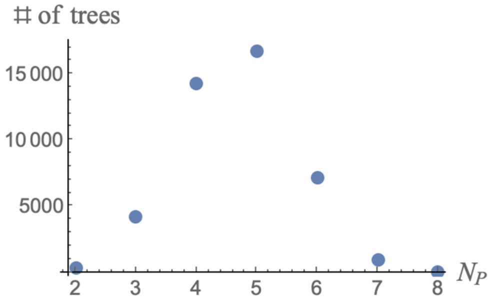

FIG. 20. The number of spanning trees on the dodecahedron as a function of the wrapping number N_P.

Conservation of tree molecules requires that $r_f + \sum_{n=1}^{12} r_n = r_t$, which is ensured if the probabilities sum to one $\sum_{n=0}^{12} P_n = 1$. Next, conservation of pentamer molecules requires that

$$c_f/c_0 = 1 - (D/12)\sum_{n=1}^{12} nP_n \tag{C4}$$

with $D = 12r_t/c_0$ the mixing ratio. Recall that if $D = 1$, then there are exactly enough pentamers to encapsidate all tree molecules. Because the concentration of free pentamers depends on the Boltzmann distributions of all aggregate sizes, the occupation probabilities for different values of n are coupled. This means that the concentration of assembled particles cannot be obtained from the LMA by itself, but would need to be complemented by similar relations for the concentrations of assembly intermediates.

If the intermediate occupation probabilities can be neglected with respect to P_0 and P_{12}, then the conservation law for RNA molecules reduces to $P_0 \simeq (1 - P_{12})$ and that of pentamers to $c_f \simeq c_0(1 - DP_{12})$. Inserting these two relations into the LMA equation (C3) produces a closed-form expression for the concentration c_f of free pentamers and hence of assembled particles:

$$\left(\frac{c_f}{c_0}\right)^{12}\left(\frac{D - 1 + \frac{c_f}{c_0}}{1 - \frac{c_f}{c_0}}\right) = \left(\frac{K}{c_0^{12}}\right). \tag{C5}$$

Because K depends only on the assembly energy of complete particles, Eq. (C5) does not depend on the class of spanning tree molecules.

A standard diagnostic for self-assembly processes are plots of the concentration of free monomeric building blocks and of assembled particles as a function of the total concentration of building blocks [74]. Such a plot is shown in Fig. 21 for MLD $= 9$, $N_P = 8$ molecules. The dots show the concentrations of free pentamers in solution and of pentamers that are part of an assembled particle as a function of the total pentamer concentration c_0 computed from Eq. (C5). For low pentamer concentrations, nearly all pentamers are free in solution and the concentration of free pentamers is close to the total concentration c_0. As c_0 increases, the concentration of free pentamers stops increasing and then saturates. Now, the concentration of pentamers that are part of an assembled particle starts to increase, proportional to c_0. The transition point between these two regimes is around $c_0 = 0.2$. This point is known in the soft-matter physics literature as the *critical aggregation concentration* (or CAC) [74].

If one again neglects assembly intermediates then it follows from Eq. (C5) that the relation $c_0(D)$ for 95% occupancy is a hyperbola in a plane with c_0 and D as coordinates:

$$c_0(D) \simeq \frac{1}{(1 - D\,P_{12})}\left(\frac{KP_{12}}{1 - P_{12}}\right)^{1/12} \tag{C6}$$

with $P_{12} = 0.95$. The hyperbola diverges at $D = 1/P_{12}$, which is close to one for a 95% packaging fraction. It shifts to smaller values of D as the pentamer concentration c_0 is

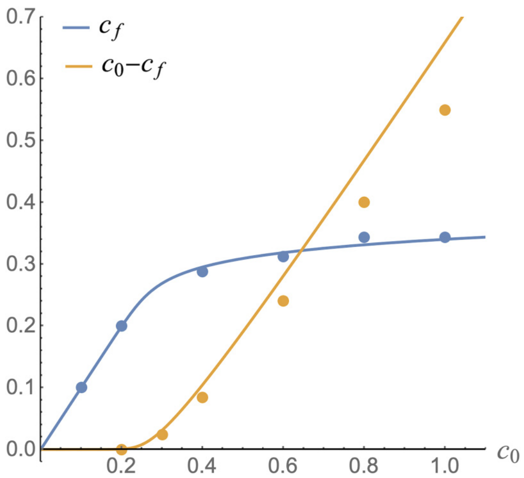

FIG. 21. Equilibrium self-assembly diagram for $N_P = 8$, MLD $= 9$ molecules for $\epsilon = -0.2$, $\mu_0 = -2.5$, and $D = 1$. Horizontal axis: total pentamer concentration c_0. Vertical axis: either the free pentamer concentration c_f (blue) or the concentration $c_0 - c_f$ of pentamers that are associated with a tree molecule (ochre). Solid lines: solution of Eq. (C5).

reduced with $c_0(D)$ always larger than $K^{1/12}$. One can use Eq. (C5) to obtain the limiting behaviors. For $(\frac{c_0^{12}}{K})$ small compared to one, the equation has a solution with c_f close to c_0 given by

$$\frac{c_f}{c_0} \simeq 1 - \frac{Dc_0^{12}}{K} \tag{C7}$$

which corresponds to the linear portion of the blue curve. For $(\frac{c_0^{12}}{K})$ large compared to one and mixing ratio D larger than one, the equation has a solution with c_f independent of c_0:

$$c_f \simeq \left(\frac{K}{D-1}\right)^{1/12} \tag{C8}$$

which corresponds to the flat part of the blue curve. For $(\frac{c_0^{12}}{K})$ large compared to one but a mixing ratio D less than one, the equation has a different solution with c_f independent of c_0:

$$\frac{c_f}{c_0} \simeq 1 - D + \frac{K}{c_0^{12}} \frac{D}{(1-D)^{1/2}}. \tag{C9}$$

There is thus a change in regimes around $D = 1$. For the special case that $D = 1$, the LMA equation reduces to

$$\left(\frac{c_f}{c_0}\right)^{13} \left(\frac{1}{1 - \frac{c_f}{c_0}}\right) \simeq \left(\frac{K}{c_0^{12}}\right). \tag{C10}$$

A second way to display self-assembly measurements under equilibrium conditions is in the form of a quasi-phase-diagram that shows the dominant type of assembly as a function of thermodynamic parameters.[3] For viral assembly, the protein and RNA concentrations are a natural choice for such a phase diagram. For the case of the spanning tree model, we will use the pentamer concentration c_0 and the mixing ratio D as thermodynamic parameters. The blue dots in Fig. 22 show points in a c_0 vs D diagram where 95% of the spanning trees are fully encapsidated. To the right of the blue dots, most tree molecules are encapsidated and coexist with excess free pentamers. To the left of the blue dots, most pentamers are part of assembled particles and coexist with excess free tree molecules. The blue dots can be viewed as "optimal mixing states" that minimize excess free pentamers and excess tree molecules. For high pentamer concentrations, the line of blue dots approaches $D = 1$, the stoichiometric ratio.

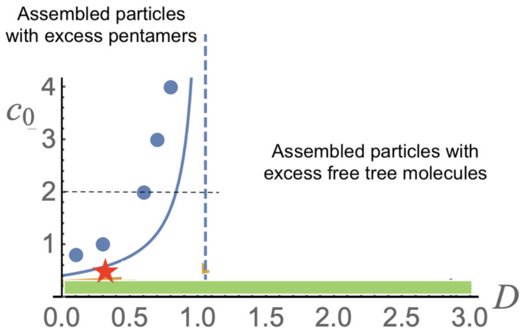

FIG. 22. Quasi-phase-diagram for $\epsilon = 0.2$ and $\mu_0 = -2$. Horizontal axis: depletion factor D. Vertical axis: pentamer concentration c_0. Blue dots: points where 95% of the genome molecules have been packaged according to the Boltzmann distribution. Solid blue line: computed from Eq. (C6). In the green sector there is practically no capsid assembly. The red star marks a possible operating point for viral assembly inside infected cells, just above the CAC, under conditions of excess pentamers in solution.

APPENDIX D: STRUCTURAL TRANSITIONS

For $\epsilon \lesssim 1$, the pentamers are most often placed on minimum-energy sites where they make the maximum number of edge-to-edge contacts with previously placed pentamers. The resulting assembly intermediates are compact pentamer clusters, similar or the same to the ones shown in Fig. 2 for the Zlotnick model. An example is shown in Fig. 23. On the other hand, for $\epsilon \gtrsim 1$, the first N_P pentamers typically are placed on maximum wrapping sites. This indicates the possibility for a *structure transition* of assembly intermediates as a function of ϵ. For example, for small ϵ six-pentamer clusters have fivefold symmetry with one central pentamer sharing its five edges with five other pentamers that each share three edges with their neighbors (see the $n = 6$ state of Fig. 11). On the other hand, for large ϵ a minimum energy $n = 6$ cluster of class (1) $N_P = 8$, MLD $= 9$ spanning trees has the six pentamers placed on the six available maximum wrapping sites of the $N_P = 6$ spanning tree (see Fig. 4). By moving just one pentamer, the two structures can be transformed into one another. This transition takes place at $\epsilon = -1$. For class (2), with $N_P = 2$, the transition is more dramatic. For small ϵ, the $n = 4$ pentamer cluster is a compact structure with a twofold symmetry axis, the same as the $n = 4$ structure shown in Fig. 2. On the other hand, the $n = 4$ minimum energy structure for large ϵ shown in Fig. 24 is completely different.

[3] Since a virus is a system of limited size, true phase transitions are not possible.

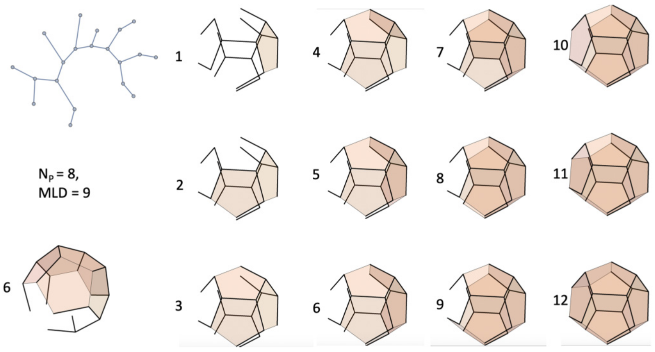

FIG. 23. Assembly pathway for a $N_P = 8$, MLD = 9 spanning tree for the case of small ϵ. The first five pentamers can be placed on sites that maximize both the number of pentamer-pentamer contacts and pentamer-spanning tree link contacts. The sixth pentamer, shown separately with a different perspective, makes only two spanning tree link contacts, which allows it to still have three pentamer-pentamer contacts. Note the similarity with Fig. 2.

This linear arrangement of pentamers has an interesting feature. Allow tree links to rotate freely around the nodes of the tree and allow pentamers to swivel freely around shared edges. The pentamers of the empty-capsid partial assemblies of Fig. 2 would, for $n > 2$, not be able to move with respect to each other without breaking pentamer-pentamer bonds. The empty-capsid partial assemblies can be said to be mechanically rigid. The same holds for the $n = 6$ structure of Fig. 4 and other small ϵ partial assemblies. However, this is not the case for the four-pentamer structure shown in Fig. 24: if this structure were allowed to fluctuate freely, then the four pentamers could freely swivel along the three shared edges. Structural transitions of this type as a function of ϵ become more common for larger values of the MLD.

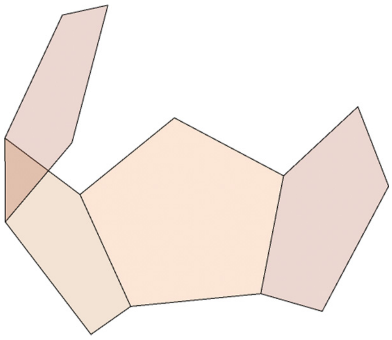

FIG. 24. The minimum-energy $n = 4$ assembly state of a class (2) molecule for $\epsilon = 1.2$.

[1] H. Fraenkel-Conrat and R. C. Williams, Reconstitution of active tobacco mosaic virus from its inactive protein and nucleic acid components, Proc. Natl. Acad. Sci. USA **41**, 690 (1955).

[2] P. J. G. Butler and A. Klug, Assembly of a virus, Sci. Am. **239**, 62 (1978).

[3] A. Klug, The tobacco mosaic virus particle: Structure and assembly, Philos. Trans. R. Soc. London, Ser. B **354**, 531 (1999).

[4] P. van der Schoot and R. Bruinsma, Electrostatics and the assembly of an RNA virus, Phys. Rev. E **71**, 061928 (2005).

[5] C. Forrey and M. Muthukumar, Electrostatics of capsid-induced viral RNA organization, J. Chem. Phys. **131**, 105101 (2009).

[6] R. F. Garmann, M. Comas-Garcia, M. S. T. Koay, J. J. L. M. Cornelissen, C. M. Knobler, and W. M. Gelbart, Role of electrostatics in the assembly pathway of a single-stranded RNA virus, J. Virol. **88**, 10472 (2014).

[7] J. D. Perlmutter and M. F. Hagan, The role of packaging sites in efficient and specific virus assembly, J. Mol. Biol. **427**, 2451 (2015).

[8] N. J. Dimmock, A. J. Easton, and K. N. Leppard, *Introduction to Modern Virology* (Blackwell, Malden, MA, 2001).

[9] T. Frensing, S. Y. Kupke, M. Bachmann, S. Fritzsche, L. E. Gallo-Ramirez, and U. Reichl, Influenza virus intracellular replication dynamics, release kinetics, and particle morphology during propagation in mdck cells, Appl. Microbiol. Biotechnol. **100**, 7181 (2016).

[10] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, J. Mol. Biol. **288**, 911 (1999).

[11] L. Tubiana, A. Lošdorfer Božič, Cristian Micheletti, and Rudolf Podgornik, Synonymous mutations reduce genome compactness in icosahedral ssRNA viruses, Biophys. J. **108**, 194 (2015).

[12] E. Frolova, I. Frolov, and S. Schlesinger, Packaging signals in alphaviruses, J. Virol. **71**, 248 (1997).

[13] G. Basnak, V. L. Morton, O. Rolfsson, N. J. Stonehouse, A. E. Ashcroft, and P. G. Stockley, Viral genomic single-stranded RNA directs the pathway toward a $T = 3$ capsid, J. Mol. Biol. **395**, 924 (2010).

[14] D. H. J. Bunka, S. W. Lane, C. L. Lane, E. C. Dykeman, R. J. Ford, A. M. Barker, R. Twarock, S. E. V. Phillips, and P. G. Stockley, Degenerate RNA packaging signals in the genome of satellite tobacco necrosis virus: Implications for the assembly of a $T = 1$ capsid, J. Mol. Biol. **413**, 51 (2011).

[15] P. G. Stockley, R. Twarock, S. E. Bakker, A. M. Barker, A. Borodavka, E. Dykeman, R. J. Ford, A. R. Pearson, S. E. V. Phillips, N. A. Ranson *et al.*, Packaging signals in single-stranded RNA viruses: Nature's alternative to a purely electrostatic assembly mechanism, J. Biol. Phys. **39**, 277 (2013).

[16] E. C. Dykeman, P. G. Stockley, and R. Twarock, Building a viral capsid in the presence of genomic RNA, Phys. Rev. E **87**, 022717 (2013).

[17] E. C. Dykeman, P. G. Stockley, and R. Twarock, Packaging signals in two single-stranded RNA viruses imply a conserved assembly mechanism and geometry of the packaged genome, J. Mol. Biol. **425**, 3235 (2013).

[18] N. Patel, E. C. Dykeman, R. H. A. Coutts, G. P. Lomonossoff, D. J. Rowlands, S. E. V. Phillips, N. Ranson, R. Twarock, R. Tuma, and P. G. Stockley, Revealing the density of encoded functions in a viral RNA, Proc. Natl. Acad. Sci. USA **112**, 2227 (2015).

[19] A. M. Yoffe, P. Prinsen, Ajaykumar Gopal, C. M. Knobler, W. M. Gelbart, and A. Ben-Shaul, Predicting the sizes of large RNA molecules, Proc. Natl. Acad. Sci. USA **105**, 16153 (2008).

[20] D. Q. Zhang, R. Konecny, N. A. Baker, and J. A. McCammon, Electrostatic interaction between RNA and protein capsid in cowpea chlorotic mottle virus simulated by a coarse-grain RNA model and a Monte Carlo approach, Biopolymers **75**, 325 (2004).

[21] W. K. Kegel and P. van der Schoot, Physical regulation of the self-assembly of tobacco mosaic virus coat protein, Biophys. J. **91**, 1501 (2006).

[22] V. A. Belyi and M. Muthukumar, Electrostatic origin of the genome packing in viruses, Proc. Natl. Acad. Sci. USA **103**, 17174 (2006).

[23] T. Hu, R. Zhang, and B. I. Shklovskii, Electrostatic theory of viral self-assembly, Phys. A (Amsterdam) **387**, 3059 (2008).

[24] B. Devkota, A. S. Petrov, S. Lemieux, M. B. Boz, L. Tang, A. Schneemann, J. E. Johnson, and S. C. Harvey, Structural and electrostatic characterization of pariacoto virus: Implications for viral assembly, Biopolymers **91**, 530 (2009).

[25] M. F. Hagan, A theory for viral capsid assembly around electrostatic cores, J. Chem. Phys. **130**, 114902 (2009).

[26] T. Jiang, Z. G. Wang, and J. Z. Wu, Electrostatic regulation of genome packaging in human hepatitis B virus, Biophys. J. **96**, 3065 (2009).

[27] A. Šiber, R. Zandi, and R. Podgornik, Thermodynamics of nanospheres encapsulated in virus capsids, Phys. Rev. E **81**, 051919 (2010).

[28] C. L. Ting, Jianzhong Wu, and Zhen-Gang Wang, Thermodynamic basis for the genome to capsid charge relationship in viral encapsidation, Proc. Natl. Acad. Sci. USA **108**, 16986 (2011).

[29] P. Ni, Z. Wang, X. Ma, N. C. Das, P. Sokol, W. Chiu, B. Dragnea, M. F. Hagan, and C. C Kao, An examination of the electrostatic interactions between the *n*-terminal tail of the coat protein and RNA in brome mosaic virus, J. Mol. Biol. **419**, 284 (2012).

[30] A. Šiber, A. L. Bozic, and R. Podgornik, Energies and pressures in viruses: Contribution of nonspecific electrostatic interactions, Phys. Chem. Chem. Phys. **14**, 3746 (2012).

[31] R. J. Ford, A. M. Barker, S. E. Bakker, R. H. Coutts, N. A. Ranson, S. E. V. Phillips, A. R. Pearson, and P. G. Stockley, Sequence-specific, RNA–protein interactions overcome electrostatic barriers preventing assembly of satellite tobacco necrosis virus coat protein, J. Mol. Biol. **425**, 1050 (2013).

[32] R. Zhang and P. Linse, Icosahedral capsid formation by capsomers and short polyions, J. Chem. Phys. **138**, 154901 (2013).

[33] G. Erdemci-Tandogan, J. Wagner, P. van der Schoot, R. Podgornik, and R. Zandi, RNA topology remolds electrostatic stabilization of viruses, Phys. Rev. E **89**, 032707 (2014).

[34] J. Kim and J. Z. Wu, A thermodynamic model for genome packaging in hepatitis B virus, Biophys. J. **109**, 1689 (2015).

[35] K. Bond, I. B. Tsvetkova, J. Che-Yen Wang, M. F. Jarrold, and B. Dragnea, Virus assembly pathways: Straying away but not too far, Small **16**, 2004475 (2020).

[36] P. E. Prevelige, D. Thomas, and J. King, Nucleation and growth phases in the polymerization of coat and scaffolding subunits into icosahedral procapsid shells, Biophys. J. **64**, 824 (1993).

[37] G. L. Casini, D. Graham, D. Heine, R. L. Garcea, and D. T. Wu, *In vitro* papillomavirus capsid assembly analyzed by light scattering, Virology **325**, 320 (2004).

[38] M. Medrano, M. Ángel Fuertes, A. Valbuena, P. J. P. Carrillo, A. Rodríguez-Huete, and M. G. Mateu, Imaging and quantitation of a succession of transient intermediates reveal the reversible self-assembly pathway of a simple icosahedral virus capsid, J. Am. Chem. Soc. **138**, 15385 (2016).

[39] R. Zandi, P. van der Schoot, D. Reguera, W. Kegel, and H. Reiss, Classical nucleation theory of virus capsids, Biophys. J. **90**, 1939 (2006).

[40] R. F. Bruinsma, G. J. L. Wuite, and W. H. Roos, Physics of viral dynamics, Nat. Rev. Phys. **3**, 76, (2021).

[41] R. F. Garmann, A. M. Goldfain, and V. N. Manoharan, Measurements of the self-assembly kinetics of individual viral capsids around their RNA genome, Proc. Natl. Acad. Sci. USA **116**, 22485 (2019).

[42] M. Comas-Garcia, S. A. K. Datta, L. Baker, R. Varma, P. R. Gudla, and A. Rein, Dissection of specific binding of HIV-1 gag to the 'packaging signal' in viral RNA, Elife **6**, e27055 (2017).

[43] N. Jouvenet, S. M. Simon, and P. D. Bieniasz, Imaging the interaction of HIV-1 genomes and gag during assembly of individual viral particles, Proc. Natl. Acad. Sci. USA **106**, 19114 (2009).

[44] T. S. Baker, N. H. Olson, and S. D. Fuller, Adding the third dimension to virus life cycles: Three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs, Microbiol. Mol. Biol. Rev. **63**, 862 (1999).

[45] M. Tihova, K. A. Dryden, T. V. L. Le, S. C. Harvey, J. E. Johnson, M. Yeager, and A. Schneemann, Nodavirus coat protein imposes dodecahedral RNA structure independent of nucleotide sequence and length, J. Virol. **78**, 2897 (2004).

[46] J. M. Johnson, D. A. Willits, M. J. Young, and A. Zlotnick, Interaction with capsid protein alters RNA structure and the pathway for in vitro assembly of Cowpea chlorotic mottle virus, J. Mol. Biol. **335**, 455 (2004).

[47] R. I. Koning, J. Gomez-Blanco, I. Akopjana, J. Vargas, A. Kazaks, K. Tars, J. María Carazo, and A. J. Koster, Asymmetric cryo-em reconstruction of phage ms2 reveals genome structure *in situ*, Nat. Commun. **7**, 12524 (2016).

[48] C. Beren, Y. Cui, A. Chakravarty, X. Yang, A. L. N. Rao, C. M. Knobler, Z. Hong Zhou, and W. M. Gelbart, Genome organization and interaction with capsid protein in a multipartite RNA virus, Proc. Natl. Acad. Sci. USA **117**, 10673 (2020).

[49] E. C. Dykeman, N. E. Grayson, K. Toropova, N. A. Ranson, P. G. Stockley, and R. Twarock, Simple rules for efficient assembly predict the layout of a packaged viral RNA, J. Mol. Biol. **408**, 399 (2011).

[50] X. Dai, Z. Li, M. Lai, S. Shu, Y. Du, Z. Hong Zhou, and R. Sun, *In situ* structures of the genome and genome-delivery apparatus in a single-stranded RNA virus, Nature (London) **541**, 112 (2017).

[51] R. D. Cadena-Nava, Y. Hu, R. F. Garmann, Benny Ng, A. N. Zelikin, C. M. Knobler, and W. M. Gelbart, Exploiting fluorescent polymers to probe the self-assembly of virus-like particles, J. Phys. Chem. B **115**, 2386 (2011).

[52] R. D. Cadena-Nava, M. Comas-Garcia, R. F. Garmann, A. L. N. Rao, C. M. Knobler, and W. M. Gelbart, Self-assembly of viral capsid protein and RNA molecules of different sizes: Requirement for a specific high protein/RNA mass ratio, J. Virol. **86**, 3318 (2012).

[53] A. Zlotnick, To build a virus capsid - an equilibrium-model of the self-assembly of polyhedral protein complexes, J. Mol. Biol. **241**, 59 (1994).

[54] D. Endres and A. Zlotnick, Model-based analysis of assembly kinetics for virus capsids or other spherical polymers, Biophys. J. **83**, 1217 (2002).

[55] A. Zlotnick, Distinguishing reversible from irreversible virus capsid assembly, J. Mol. Biol. **366**, 14 (2007).

[56] A. Y. Morozov, R. F. Bruinsma, and J. Rudnick, Assembly of viruses and the pseudo-law of mass action, J. Chem. Phys. **131**, 155101 (2009).

[57] J. D. Perlmutter, M. R. Perkett, and M. F. Hagan, Pathways for virus assembly around nucleic acids, J. Mol. Biol. **426**, 3148 (2014).

[58] B. Bollobás, *Modern Graph Theory*, Vol. 184 (Springer, New York, 2013).

[59] R. L. Graham and P. Hell, On the history of the minimum spanning tree problem, Ann. Hist. Comput. **7**, 43 (1985).

[60] L. T. Fang, W. M. Gelbart, and A. Ben-Shaul, The size of RNA as an ideal branched polymer, J. Chem. Phys. **135**, 155105 (2011).

[61] J. Rudnick and R. Bruinsma, Icosahedral Packing of RNA Viral Genomes, Phys. Rev. Lett. **94**, 038101 (2005).

[62] A. M. Gutin, A. Y. Grosberg, and E. I. Shakhnovich, Polymers with annealed and quenched branches belong to different universality classes, Macromolecules **26**, 1293 (1993).

[63] Nicolaas Godfried Van Kampen, *Stochastic Processes in Physics and Chemistry*, Vol. 1 (Elsevier, Amsterdam, 1992).

[64] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.106.044405 for the derivation of the expression used for the on-rates Wn, $n+1$ used in the main text (based on Ch. 8 of the lecture notes "Non-Equilibrium Statistical Mechanics", UIUC, by Klaus Schulten).

[65] J. D. Perlmutter and M. F. Hagan, Mechanisms of virus assembly, Annu. Rev. Phys. Chem. **66**, 217 (2015).

[66] J. Rudnick and R. Bruinsma, Invariant theory and orientational phase transitions, Phys. Rev. E **100**, 012145 (2019).

[67] R. Zandi, B. Dragnea, A. Travesset, and R. Podgornik, On virus growth and form, Phys. Rep. **847**, 1 (2020).

[68] M. G. M. van Rosmalen, D. Kamsma, A. S. Biebricher, C. Li, A. Zlotnick, W. H. Roos, and G. J. L. Wuite, Revealing in real-time a multistep assembly mechanism for sv40 virus-like particles, Sci. Adv. **6**, eaaz1639 (2020).

[69] S. Panahandeh, S. Li, L. Marichal, R. L. Rubim, G. Tresset, and R. Zandi, How a virus circumvents energy barriers to form symmetric shells, ACS Nano **14**, 3170 (2020).

[70] R. F. Garmann, M. Comas-Garcia, Ajaykumar Gopal, C. M. Knobler, and W. M. Gelbart, The assembly pathway of an icosahedral single-stranded RNA virus depends on the strength of inter-subunit attractions, J. Mol. Biol. **426**, 1050 (2014).

[71] H. Echols and M. F. Goodman, Fidelity mechanisms in DNA replication, Annu. Rev. Biochem. **60**, 477 (1991).

[72] B. Tippin, P. Pham, and M. F. Goodman, Error-prone replication for better or worse, Trends Microbiol. **12**, 288 (2004).

[73] J. J. Hopfield, Kinetic proofreading: A new mechanism for reducing errors in biosynthetic processes requiring high specificity, Proc. Natl. Acad. Sci. USA **71**, 4135 (1974).

[74] S. Safran, *Statistical Thermodynamics of Surfaces, Interfaces, and Membranes* (CRC Press, Boca Raton, FL, 2018).