Effect of the charge distribution of virus coat proteins on the length of packaged RNAs

Yinan Dong, Siyu Li ,[*] and Roya Zandi
Department of Physics and Astronomy, University of California, Riverside, California 92521, USA

Single-stranded RNA viruses efficiently encapsulate their genome into a protein shell called the capsid. Electrostatic interactions between the positive charges in the capsid protein's N-terminal tail and the negatively charged genome have been postulated as the main driving force for virus assembly. Recent experimental results indicate that the N-terminal tail with the same number of charges and same lengths packages different amounts of RNA, which reveals that electrostatics alone cannot explain all the observed outcomes of the RNA self-assembly experiments. Using a mean-field theory, we show that the combined effect of genome configurational entropy and electrostatics can explain to some extent the amount of packaged RNA with mutant proteins where the location and number of charges on the tails are altered. Understanding the factors contributing to the virus assembly could promote the attempt to block viral infections or to build capsids for gene therapy applications.

I. INTRODUCTION

Viruses have optimized the feat of packaging of their negatively charged genomes into a protein shell called the capsid, often built from a large number of one or a few different kinds of protein subunits [1]. Under many *in vitro* conditions, coat proteins of several single-stranded RNA (ssRNA) viruses can spontaneously encapsulate all types of anionic cargos including their native genome, linear polymers, and heterologous and nonviral RNAs [2–6]. The capsid proteins of several RNA viruses contain an unstructured positively charged N-terminal domain that extends toward the center of the capsid and interacts with the viral genome; see Fig. 1 [7]. Although the specific sequence of the viral RNA plays an important role in packaging [8,9], it is now well established that the electrostatic interaction between N-terminal tails and RNA is the main driving force for the formation of viral particles and their stability [10–12].

Self-assembly studies of various ssRNA viruses have revealed that the amount of RNA packaged depends directly on the number of positive charges on the N-terminal tails of capsid proteins. Many experiments show that mutant virions with less positive charges on N-terminal domain encapsidate lower amounts of RNA and mutants with increased positive charges package more [13,14]. For example, the experimental studies of Sivanandam *et al.* show that the deletions of even one single positively charged residue of the satellite tobacco mosaic virus N-terminal domain results in the formation of virus particles with a reduced amount of viral RNAs [13]. Belyi and Muthukumar as well as Hu *et al.* [15,16] also examined the relation between the total number of positive charges in the tails and the length of the encapsidated RNA in various viruses and found a strong relation between them.

Of particular interest is the self-assembly experiments of Ni *et al.* who specifically focused on the brome mosaic virus

(BMV) and systematically investigated the role of electrostatics on the amount of RNA packaged [14]. The N-terminal domain of BMV capsid proteins is composed of 26 residues, eight of which are positively charged. The genome of BMV consists of four RNA molecules: RNA1 (3.2 kb), RNA2 (2.9 kb), RNA3 (2.1 kb), and RNA4 (0.9 kb). While RNA3 and RNA4 co-assemble together in one capsid, RNA1 and RNA2 are each encapsidated separately. Quite interestingly, the total length of encapsidated genome is more or less the same in each capsid. The BMV capsids of these three types are virtually identical, i.e., have $T = 3$ icosahedral structures consisting of 180 copies of the same protein with the same mechanical properties [17]; see Fig. 1. We note that the structural index T, introduced by Casper and Klug, defines the number of protein subunits in viral shells, which is 60 times the T number [18]. Thus $T = 1$ and $T = 3$ capsids have 60 and 180 protein subunits, respectively.

To gain more insight into the effect of electrostatic interactions, Ni *et al.* made several mutants to increase the number of charges on N-terminal domains. A summary of their experimental results is presented in Fig. 2. In one case, they inserted eight residues including four positively charged ones after residue 15 ($2H_{15}$). They also examined the impact of the length of the N-terminal without adding more positive charges but by introducing six alanines and two threonines, which are neutral ($2HA_{15}$). To examine whether the position of the insertions has an impact on the amount of packaged RNA, they repeated the aforementioned experiments but introduced insertions after residue 7 and constructed $2H_7$ and $2HA_7$. Furthermore, to exclusively examine the effect of the increasing charges while keeping the length of the N-terminal tail the same as the wild-type one, they replaced four uncharged residues along the tail with four arginines (4R), each containing one positive charge. They found that in all cases, the structure of capsids was almost the same even though the amount of encapsidated RNA was different.

The spectroscopic analysis of the experiments of Ni *et al.* reveals that as the number of charges on the N-terminal

*Present address: Department of Materials Science and Engineering, Northwestern University, Evanston, Illinois 60208, USA.
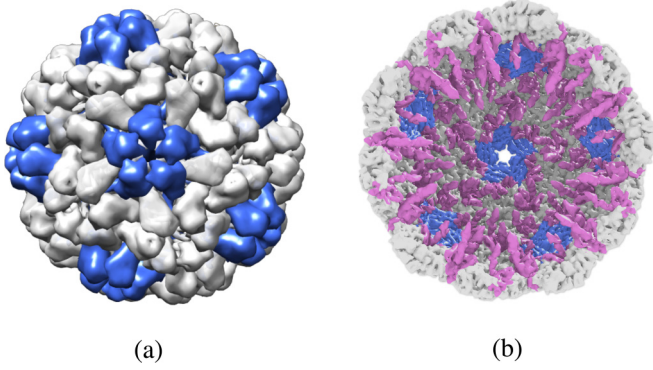
(a)   (b)

FIG. 1. (a) A $T = 3$ icosahedral shell with 180 protein subunits. The darker (blue) color shows the pentamers. The structure is similar to the BMV capsid. (b) The interior of a $T = 3$ viral shell with N-terminal domains (pink tails) extended toward the center of the capsid. Each N-terminal domain contains eight positive charges, not shown in the figure. The structure in (a) is reproduced using UCSF Chimera packages [43].

increases, the higher amount of nucleotides per capsid is packaged [14]. Nevertheless, it appears that the amount of encapsidated RNA increase does depend on other factors than the number of positive charges on the N-terminals. While the experiments clearly indicate that electrostatics plays a major role in RNA packaging, it is not obvious whether electrostatics can explain all the effects observed in Fig. 2. Many theoretical and experimental studies have already shown that the length of packaged RNA increases with the number of charges in N-terminal tails [11,13,14], but how the amount of RNA encapsidated depends on the distribution and location of charges on the N-terminals has remained elusive.

In this paper we show that electrostatics is indeed able to explain at least to some extent many observed effects relevant to RNA packaging. Using the mean-field theory, we show that the charge discreteness, the location, and the distance between
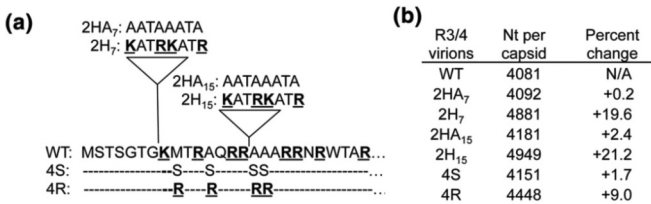

FIG. 2. (a) Schematic of the sequences of N-terminal tails of six mutants used in the experiments of Ni *et al.* [14]. The mutants are denoted by $2HA_7$, $2H_7$, $2HA_{15}$, $2H_{15}$, 4S, and 4R. The triangles denote the location of the insertions. For $2HA_7$ and $2HA_{15}$, eight neutral amino acids are inserted into the N-terminal. For $2H_7$ and $2H_{15}$, four neutral and four positive amino acids (with boldface and underlined) are inserted. The four positive amino acids are two lysines (K) and two arginines (R), leading to the increased length of N-terminal regions and also 720 additional positive charges per capsid. For 4S and 4R, the length of N-terminals remains the same. In the case of 4S four neutral amino acids (MAAA) are replaced with another four neutral amino acids and for 4R mutants, four neutral amino acids (MAAA) are replaced with four positively charged arginines (R). (b) Spectroscopic analysis of the number of nucleotides per virion.

the charges along the N-terminal tails have a huge impact on the optimal number of nucleotides packaged. Consistently with the experiments of Ni *et al.* we find that the optimal amount of packaged RNA depends on the location of charges within the peptide sequence and increases nonlinearly with the total number of positive charges on the capsid.

The paper is organized as follows. In the next section, we introduce the model and derive the equations that we will employ later. In Sec. III, we present our results corresponding to the nonuniform charge distribution along the N-terminal tails of BMV coat proteins. Section IV discusses the impact of the length and sequence of amino acid N-terminal tails on the the length of encapsidated genomes, and finally, we present our conclusion and summarize our findings.

II. METHOD

To explore the impact of N-terminal charge distribution on the length of packaged RNA, we model RNA as a negatively charged flexible polymer. Many experiments show that RNA acts effectively as a branched polymer in solution [19,20]. Due to the relatively weak strength of RNA base-pairing, the number of branch points of RNA can easily be modified through the interaction with the positive charges of virus coat proteins. Thus, we focus on the case of annealed branched polyelectrolyte, which allows the degree of branching of RNAs, a statistical quantity, to be modified [21]. Using the mean-field theory, we calculate the free energy of the RNA confined into a spherical shell that interacts attractively with the positive charges residing on the N-terminal domains of the capsid proteins. Under the ground-state dominance approximation [22,23] where only the dominating contribution to the polymer partition function is considered, the free energy of the genome-capsid complex in a salt solution is [11,24–27]

$$
\beta F = \int d^3r \left\{ \frac{a^2}{6} |\nabla \Psi(\mathbf{r})|^2 + W[\Psi(\mathbf{r})] \right.
$$
$$
- \frac{\beta^2 e^2}{8\pi \lambda_B} |\nabla \Phi(\mathbf{r})|^2 - 2\mu \cosh[\beta e \Phi(\mathbf{r})]
$$
$$
\left. + \beta \tau \Phi(\mathbf{r}) \Psi^2(\mathbf{r}) \right\} + \int d^2r [\beta \rho(\mathbf{r}) \Phi(\mathbf{r})], \quad (1)
$$

where β is the inverse of temperature in the units of energy, a is the Kuhn length of the polymer, e is the elementary charge, μ is the density of monovalent salt ions, and τ is the linear charge density of chain. The Bjerrum length $\lambda_B = e^2\beta/4\pi\epsilon$ is about 0.7 nm for water at room temperature. The dielectric permittivity of the medium ϵ is assumed to be constant [28]. See Ref. [29] and the Appendix of Ref. [27] for a step by step derivation of Eq. (1), in the absence and presence of electrostatic interactions, respectively.

The field $\Psi(\mathbf{r})$ is the monomer density field and $\Phi(\mathbf{r})$ is the electrostatic potential. The density of positive charges on the N-terminal tails of capsid proteins is denoted by $\rho(\mathbf{r})$. The first term in Eq. (1) is the entropic cost of deviation from a uniform chain density. The last two lines of Eq. (1) are associated with the electrostatic interactions between the chain segments, the capsid, and the salt ions at the level of Poisson-Boltzmann theory [24,30–32]. The term $W[\Psi]$ represents the free energy density associated with the annealed branching of the polymer

including the self repulsion of the polyelectrolyte [33–35],

$$W[\Psi] = -\frac{1}{\sqrt{a^3}}\left(f_e\Psi + \frac{a^3}{6}f_b\Psi^3\right) + \frac{1}{2}\upsilon\Psi^4, \quad (2)$$

where f_e and f_b are the fugacities of the end and branched points of the annealed polymer, respectively [29], and υ is the effective excluded volume for each monomer. Note that the stem-loop or hair-pin configurations of RNA are counted as end points in this model. The quantity $\frac{1}{\sqrt{a^3}}f_e\Psi$ indicates the density of end points and $\frac{\sqrt{a^3}}{6}f_b\Psi^3$ the density of branch points. The expectation numbers of end and branched points, N_e and N_b, are related to the fugacities f_e and f_b, and can be written as

$$N_e = -\beta f_e \frac{\partial F}{\partial f_e} \quad \text{and} \quad N_b = -\beta f_b \frac{\partial F}{\partial f_b}. \quad (3)$$

There are two additional constraints in the system. The first one corresponds to the fact that the total number of monomers (Kuhn lengths) inside the capsid is fixed [36,37],

$$N = \int d^3\mathbf{r}\ \Psi^2(\mathbf{r}). \quad (4)$$

We impose this constraint through a Lagrange multiplier, E, introduced below. Second, there is a relation between the number of the end and branched points,

$$N_e = N_b + 2, \quad (5)$$

as there is only a single polymer in each capsid and no closed loops within the secondary structure of an RNA are allowed. The polymer is linear if $f_b = 0$, and the number of branched points increases with increasing value of f_b. For our calculations, we vary f_b and find f_e through Eq. (3) and Eq. (5). To this end, f_e is not a free parameter.

Extremizing the free energy with respect to the fields $\Psi(\mathbf{r})$ and $\Phi(\mathbf{r})$, subject to the constraint that the total number of monomers inside the capsid is constant [Eq. (4)], we obtain three self-consistent nonlinear coupled equations for the interior and exterior of the capsid,

$$\frac{a^2}{6}\nabla^2\Psi(\mathbf{r}) = -E\Psi(\mathbf{r}) + \tau\beta\Phi_{\text{in}}(\mathbf{r})\Psi(\mathbf{r}) + \frac{1}{2}\frac{\partial W}{\partial\Psi}, \quad (6a)$$

$$\nabla^2\Phi_{\text{in}}(\mathbf{r}) = \frac{1}{\lambda_D^2}\sinh[\Phi_{\text{in}}(\mathbf{r})] - \frac{\tau}{2\lambda_D^2\mu\beta e^2}\Psi^2(\mathbf{r})$$
$$- \frac{1}{2\lambda_D^2\mu\beta e^2}\rho(\mathbf{r}), \quad (6b)$$

$$\nabla^2\Phi_{\text{out}}(\mathbf{r}) = \frac{1}{\lambda_D^2}\sinh[\Phi_{\text{out}}(\mathbf{r})], \quad (6c)$$

where $\lambda_D = 1/\sqrt{8\pi\lambda_B\mu}$ is the (dimensionless) Debye screening length and E is the Lagrange multiplier implementing the fixed monomer number inside capsid. The polymer concentration in the exterior of the capsid is considered to be zero, $\Psi = 0$. Equations (6) along with the constraints shown in Eqs. (4) and (5) represent a set of coupled nonlinear differential equations that, subject to appropriate boundary conditions, can only be solved numerically for the unknown parameters f_e and E and fields $\Psi(\mathbf{r})$ and $\Phi(\mathbf{r})$.

The boundary conditions for the two coupled differential equations (6b) and (6c) can be obtained by minimizing the
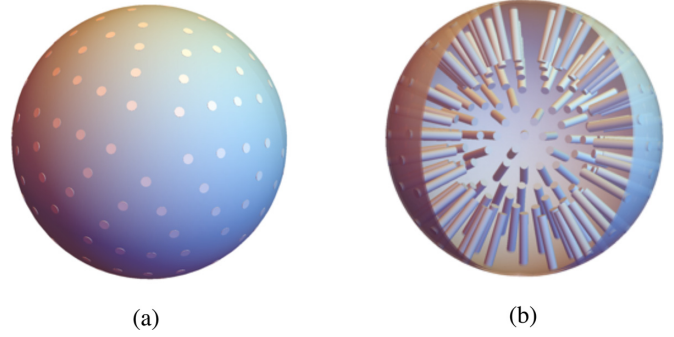

FIG. 3. (a) The white circles indicate the locations of N-terminals on a $T = 3$ capsid. (b) 3D view of inside of a $T = 3$ capsid with 180 protruded regions representing N-terminals. There are eight positive charges on each cylinder (N-terminal tail) in a wild-type BMV capsid. The positive charges are not shown in the figure.

free energy with respect to the $\Phi(\mathbf{r})$ field on the surface of the capsid and are

$$\hat{n}\cdot\nabla\Phi_{\text{in}}(\mathbf{r})|_{r=R} = \hat{n}\cdot\nabla\Phi_{\text{out}}(\mathbf{r})|_{r=R},$$
$$\Phi_{\text{in}}(\mathbf{r})|_{r=R} = \Phi_{\text{out}}(\mathbf{r})|_{r=R},$$
$$\Phi_{\text{out}}(\mathbf{r})|_{r=\infty} = 0. \quad (7)$$

We employ Dirichlet boundary condition $\Psi(\mathbf{r})|_{r=R} = 0$ for the monomer density field at the capsid wall. Because of the symmetric monomer distribution, we set $\partial_r\Psi(\mathbf{r})|_{r=0} = 0$. We emphasize that the derivations of all equations given in this section can be found in the Appendix of Ref. [27]. A more detailed derivation of the partition function and free energy for branched polymers can be found in Ref. [29].

N-terminal tails

Figure 1 shows a $T = 3$ structure with 180 N-terminal tails extending into the interior of the capsid, distributed with icosahedral symmetry. Because of the repulsion between the positive charges residing on the N-terminal tails, and the fact that RNA wraps around them, we assume that the N-terminal tails take an extended configuration. To this end, we model the N-terminal tails of BMV capsids as solid cylinders; see Fig. 3(b). We note that the charged tails are placed inside the capsid, and we will use the same boundary conditions for them as those given in Eq. (7) at the surface.

In the next section we will examine the impact of different charge distributions along N-terminal domains on the optimal genome length, which we will compare with the experimental results presented in Fig. 2. Since most of the positive charges are residing on the N-terminal tails, we consider that the charges of the coat proteins are only distributed in the cylindrical regions with no charges on the capsid wall.

For simplicity, we first consider a $T = 1$ capsid with only two positive charges on each of its 60 N-terminal tails and then focus on the $T = 3$ capsid of BMV.
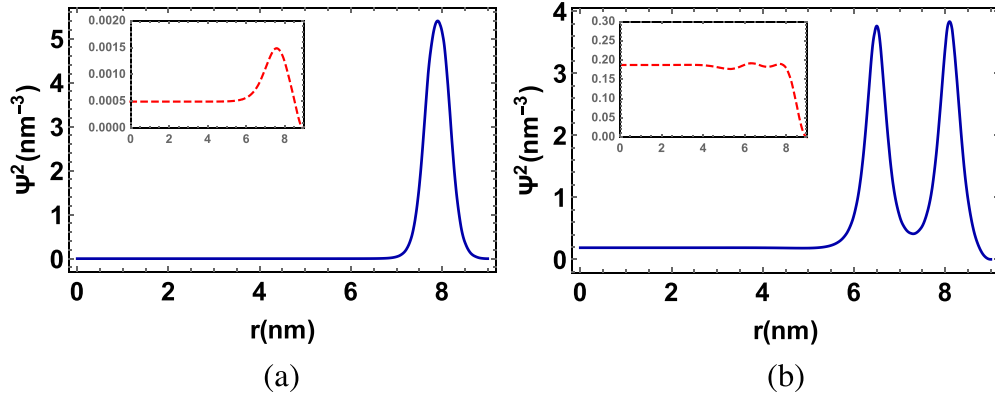
FIG. 4. Genome density profile inside a $T = 1$ capsid as a function of the distance from the capsid center. The solid lines in the figure show the profiles along N-terminal tails, but the dashed graphs correspond to the direction without N-terminal tails (inset). (a) The plot illustrates the profile when the distance between the two charges is 0.2 nm with the total number of monomers $N = 658$. (b) The plot corresponds to the profile when the distance between the two charges is 1.4 nm with $N = 680$. See Fig. 5(a) for a schematic view of charge distributions. The length of the tail is 4 nm, and the size of each charged region is 0.2 nm. The polymers are branched with $f_b = 3.86$. The other parameters are salt concentration $\mu = 500$ mM, the capsid radius $R = 9$ nm, and the total charge on N-terminals $Qc = 120$.

III. RESULTS

A. A capsid with 60 tails ($T = 1$)

To obtain the optimal length of encapsidated genome in a $T = 1$ shell, we numerically solve the nonlinear coupled differential equations (6a), (6b), and (6c), subject to the constraints given in Eqs. (4) and (5). We operate on the nonlinear coupled differential equations with the finite element method and deal with the convergence issue employing the Newton method [38–40].

After finding the solutions for the fields $\Psi(\mathbf{r})$ and $\Phi(\mathbf{r})$ we insert them into Eq. (1) to obtain the free energy of the polymer-capsid complex, F [26,27,41]. To obtain the encapsidation free energy, we need to calculate the free energy of a polymer free in solution and that of a positively charged shell and then subtract them both from the polymer-capsid complex free energy, F, given in Eq. (1). The capsid self-energy $[F(N = 0)]$ due to the electrostatic interactions is calculated through Eqs. (6) and (7) in the limit as $N \to 0$ and should be explicitly subtracted from the polymer-capsid complex free energy, F. We emphasize that the focus here is on the solution conditions in which the capsid proteins can self-assemble in the absence of the genome. We also note that previous works have shown that the free energy associated with a free chain (both linear and branched) is negligible under most experimental conditions [25,27].

The results of our numerical calculations are given in Fig. 4 as a plot of the polymer concentration profile vs r, the distance from the center of the shell for a branched polymer with the radius of capsid $R = 9$ nm at $\mu = 500$ mM salt concentrations. The total number of charges in the capsid is $Qc = 120$ with two charges on each N-terminal tail. The length of N-terminal is 4 nm and the size of each charge is 0.2 nm [see Fig. 5(a)].

Figure 4(a) shows the genome profile if the distance between the two positive charges along the N-terminal tails is 0.2 nm while Fig. 4(b) corresponds to when the distance between the charges is 1.4 nm; see Fig. 5(a) for a schematic presentation of the distribution of charges in both cases. Note that the charged amino acids are yellow and neutral ones are

blue in Fig. 5(a). The optimal number of monomers enclosed in the shell for Fig. 4(a) is $N = 658$ and for Fig. 4(b) is $N = 680$. The figure clearly shows that the polymer concentration is higher at the positions where the positive charges are located along the tails. When the distance between two charges is less than the Debye length $\lambda_D = 0.438$ nm, there is only one maximum in the profile. As the distance between the charges increases and goes beyond two Debye lengths, the genome density profile between the two charges goes almost to zero.

It is important to note that we have previously studied the impact of the number of branched points, which is closely connected to the f_b value, on the length of the encapsidated genome and found that the length of the genome increases with f_b [26]. Since our focus in this paper is only on the effect
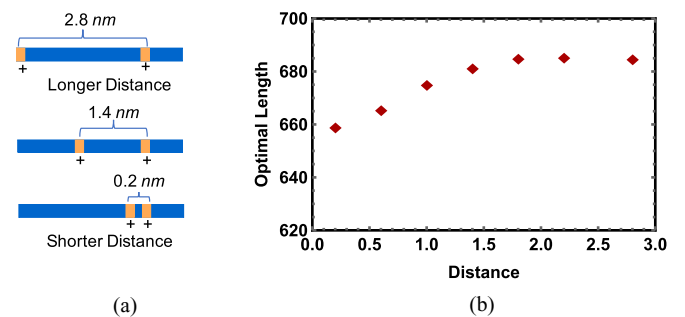

FIG. 5. (a) Schematic of an N-terminal tail. The distance between two positive charges along the N-terminal domain increases from bottom to top. Each yellow rectangle is 0.2 nm and denotes one positive charged amino acid. The smallest distance between the two charges is 0.2 nm. From the shortest to the longest distance, we examine seven different cases. The largest distance between the two charges is 2.8 nm. The charge on the right side is next to the wall and its position is fixed. (b) Optimal length of RNA encapsulated as a function of the distance between two charges for a capsid with radius $R = 9$ nm, the tail length 4 nm, and salt concentration $\mu = 500$ mM. RNA is modeled as an annealed branched polymer.
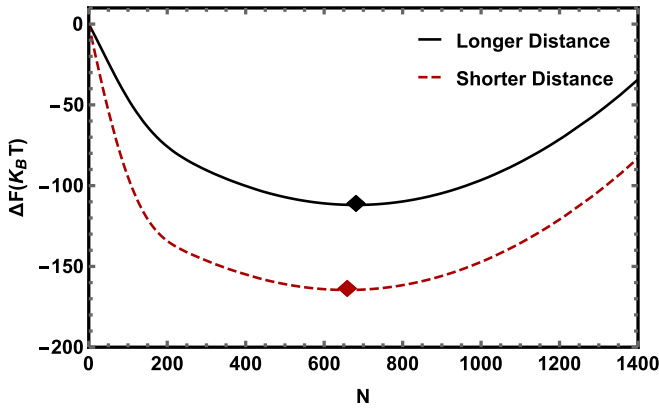
FIG. 6. Encapsulation free energy as a function of monomer number, N. The dashed line corresponds to the case in which the distance between the two charges is short ($d = 0.2$ nm) and the solid curve to when the distance between the charges is a little bit longer ($d = 1.4$ nm); see Fig. 5(a) for a schematic of two charge distributions. The other parameters are the capsid radius $R = 9$ nm, the tail length 4 nm, and salt concentration $\mu = 500$ mM, and the total positive charge on the capsid is $Qc = 120$. RNA is modeled as an annealed branched polymer and its fugacity is $f_b = 3.86$. The optimal number of packaged monomers for $d = 0.2$ nm is 658 while for $d = 1.4$ nm it is 680.

of charge distribution along the N-terminals, we set $f_b = 3.86$ for all the calculations presented here. In a previous paper, we found that this value of f_b would create a similar number of branch points to the case in the wild-type BMV genome [26]. The value of f_b does not play an important role in our findings of the effect of N-terminal charge distribution.

Figure 6 shows the encapsulation free energy as a function of N, the number of monomers, for a $T = 1$ structure. The dashed line in the figure corresponds to the case in which the distance between the charges is 0.2 nm and solid lines to when the distance between the charges is 1.4 nm; see Fig. 5(a) for a schematic of two charge distributions. As illustrated in Fig. 6, when the charges are closer to each other, the free energy of the system is lower; however, the minimum of the free energy moves toward longer chains as the distance between the charges increases.

Figure 5(b) shows the optimal length of encapsidated RNA as a function of the distance between two charges along the N-terminal domains. One charge is placed at the end of the N-terminal tail next to the capsid wall, but the location of the other varies from the wall all the way to the tip. The figure clearly shows that as the distance between charges increases, the optimal length of the genome increases too. Thus, the location of charges along the N-terminal domains has an impact on the amount of the polymer packaged. It appears as the distance between the charges goes up, at some point the optimal length of the packaged genome saturates and does not keep increasing. A careful examination of the first term in Eq. (1) shows that for this size of capsid and charge distribution, the optimal genome density is too small and the impact of entropy is not strong enough to have a significant role in the optimal length of the genome. As the distance between the charges increases and becomes more than two Debye lengths ($\lambda_D = 0.438$ nm for $\mu = 500$ mM), the electrostatic interaction becomes very

Charge distribution	Virions	Effective Free Energy	Optimal Length	Percent Change (Theory)	Percent Change (Experiment)
	WT	-1935.67	3478.35	N/A	N/A
	2HA$_7$	-1935.67	3478.35	0	+0.2
	2HA$_{15}$	-1963.96	3479.95	+0.06	+2.4
	2H$_7$	-3215.13	4547.57	+30.74	+19.6
	2H$_{15}$	-3170.14	4536.67	+30.43	+21.2
	4S	-1935.67	3478.35	0	+1.7
	4R	-3720.75	4436.86	+27.56	+9.0

FIG. 7. Table of seven charge distributions along N-terminals where each yellow rectangle represents a positively charged amino acid and blue triangles neutral ones. The table includes the optimal encapsulation free energy of the RNA confined into a spherical shell, the optimal length of encapsulated RNA, percent change (theory) of optimal length compared to the wild-type BMV, and the percent change (experiment) from Fig. 2. The salt concentration is 500 mM. The radius of the capsid is 12 nm. For wild type, 2HA$_7$, 2HA$_{15}$, and 4S, the total charge on capsid is $Qc = 1440$ but for 2H$_7$, 2H$_{15}$, and 4R it is $Qc = 2160$. The tail length for wild type, 4S, and 4R is 5 nm while for 2HA$_7$, 2HA$_{15}$, 2H$_7$, and 2H$_{15}$ it is 6.5 nm. The Debye length λ_D for 500 mM is 0.438 nm.

weak between the two charges. Thus, the genome will be mostly adsorbed in the close proximity of each positive charge along the peptide. Note that even though entropy prefers a uniform genome density, the electrostatic interaction is much stronger and thus the optimal length of encapsidated genome first increases with the distance between the charges and then it remains more or less constant.

B. A capsid with 180 tails ($T = 3$)

We now examine the impact of charge distribution along the N-terminal domain for a $T = 3$ capsid with 180 N-terminal tails. More specifically, we focus on the self-assembly studies of Ni *et al.* in which the impact on the length of packaged RNA of the location and distribution of positive charges along the N-terminal domains of BMV capsid proteins were studied [14]. Figures 2(a) and 2(b) show the distribution of charges along N-terminal domains and the length of encapsulated RNA for different mutants, respectively. The schematic of the charge distribution along the N-terminals for various mutants and wild-type capsid proteins based on our model is illustrated in the left column of Fig. 7. The length of N-terminal is set equal to 5 nm for the wild type and 6.5 nm for the mutants with eight extra amino acids. We assume all amino acids have the same size, which is set equal to 0.2 nm. The charged amino acids are yellow and neutral ones are blue as before.

Following the same procedures as described above for a $T = 1$ structure, we first obtain the genome profile for a given number of nucleotides and then use it to calculate the free energy of the system. Figure 8 shows the genome profiles for the wild type, 2HA$_{15}$, and two other mutant proteins. The schematic of charge distribution for each case is illustrated in Figs. 7 and 9. The total number of monomers in each plot in Fig. 8 is $N = 1390$, and the total number of charges in all capsids is $Qc = 1440$. There are eight positive charges on each N-terminal tail, whose length is 6.5 nm long for mutants
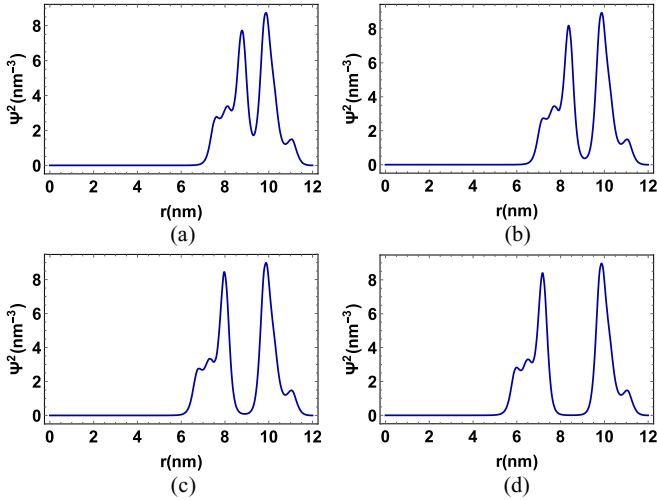
FIG. 8. The genome density profile vs r, the distance from the center of the capsid for four different charge distributions along the N-terminal domains: (a) wild type (WT), (b) 2HA$_{15}$(M1), (c) 2HA$_{15}$(M2), and (d) 2HA$_{15}$. The first column in Figs. 7 and 9 shows the schematics of N-terminal tails for each case. The peaks in the RNA profiles correspond to the position of positive charges along the N-terminal tails. As the distance between the charges located in the middle of N-terminal tails increases, the density of genome between the two peaks goes lower. However, the amount of RNA between two peaks due to the entropic contribution and the range of electrostatic interaction do not drop to zero in the case of 2HA$_{15}$(M1) (b). The genome density between the two peaks becomes smaller for 2HA$_{15}$(M2) (c) and becomes almost zero for 2HA$_{15}$.

and 5 nm for wild-type proteins. The genome is considered to be a branched polymer ($f_b = 3.86$).

Figure 10 shows the free energy of a branched polymer packaged by the wild-type and mutant proteins of Fig. 7. The symbols in the figure correspond to the optimal genome length for each case. The figure reveals that the encapsulation free energy of the wild type, 2HA$_7$, 4S, and 2HA$_{15}$ are almost the same. Note that all these mutants have the same number of charges on their capsids. The values of the minimum free energy, the corresponding optimal genome length, and the percent change (theory and experiment) of encapsulated genome compared to the wild-type case are presented in Fig. 7.

FIG. 10. The encapsidation free energy as a function of monomer numbers for the mutants presented in Fig. 7. 2HA$_7$ and 4S have the same free energy as wild type. The additional length inserted in 2HA$_7$ does not have a huge impact on the optimal encapsidated genome because it does not modify the distance between the charges along the N-terminal tails; see Fig. 7. The capsid radius is 12 nm and the tail length is 6.5 nm with total charges on the capsid $Qc = 1440$. The salt concentration is 500 mM.

Consistent with the experimental data presented in Fig. 2 and the last column of Fig. 7, our theoretical calculations show that as the number of positive charges on the N-terminal tails increases, the optimal length of the genome increases too. The mutants 4R, 2H$_7$, and 2H$_{15}$ have four extra positive charges compared to wild-type proteins and they all encapsidate longer genomes. Both mutants 2H$_7$ and 2H$_{15}$ have longer tails compared to 4R, and our results show that they encapsidate longer genomes, consistent with the experimental findings. Thus the length of N-terminal tails influences the amount of packaged RNA.

While there are many similarities between the experiments presented in Fig. 2 and our theoretical results shown in Fig. 7, there are also some differences. The comparison of the experiment and theory reveals that more genome is encapsidated by 2HA$_{15}$ proteins compared to wild-type or 2HA$_7$ proteins, which is not observed in our calculations. Note that to perform the numerical calculations, we consider that all amino acids have the same effective size (0.2 nm), and the Debye length in our system is $\lambda_D = 0.438$ nm. Since the parameter landscape is quite vast and there are several unknowns, instead of changing the size of each amino acid, we modify the distance between the fourth and fifth charged amino acids in the N-terminal tail of the mutant 2HA$_{15}$. More specifically, we systematically increase the distance between the fourth and fifth positive charges from 0.2 nm to 2.8 nm where the 8 amino acids were inserted for the case of the mutant 2HA$_{15}$ and then calculate the optimal length of encapsidated genome for three different salt concentrations of $\mu = 100$, 300, and 500 mM. As illustrated in Fig. 9, the optimal length of encapsidated genome depends on both the distance between the fourth and

Charge Distribution	Virions	Optimal Length (500mM)	Percent Change (500mM)	Optimal Length (300mM)	Percent Change (300mM)	Optimal Length (100mM)	Percent Change (100mM)
	2HA$_7$(M)	3431.4	N/A	2501.00	N/A	1831.2	N/A
	2HA$_7$	3478.35	+1.37	2528.29	+1.09	1840.28	+0.50
	2HA$_{15}$(M1)	3522.73	+2.67	2544.39	+1.73	1847.94	+0.91
	2HA$_{15}$(M2)	3501.42	+2.04	2548.31	+1.89	1852.57	+1.17
	2HA$_{15}$(M3)	3492.47	+1.78	2541.79	+1.63	1853.44	+1.21
	2HA$_{15}$	3479.95	+1.41	2529.88	+1.15	1850.17	+1.04

FIG. 9. Table of six different charge distributions along N-terminals. As before each yellow rectangle represents an amino acid with a positive charge and blue rectangles represent neutral amino acids. The table includes the optimal length of encapsulated RNA for three different salt concentrations, 500 mM, 300 mM, and 100 mM. The distance between the fourth positive charge (the fourth yellow rectangle) and the fifth positive charge from top to bottom is 0.2 nm, 0.6 nm, 1.0 nm, 1.4 nm, 1.8 nm, and 2.2 nm. The percent change (theory) of the optimal length of encapsidated RNA for each mutant relative to the RNA encapsidated by mutant 2HA$_7$(M) is also presented in the table. The capsid radius is 12 nm and the tail length is 6.5 nm with total charges on the capsid $Qc = 1440$. Debye length is $\lambda_D = 0.979$ nm for $\mu = 100$ mM, $\lambda_D = 0.565$ nm for $\mu = 300$ mM, and $\lambda_D = 0.438$ nm for 500 mM.
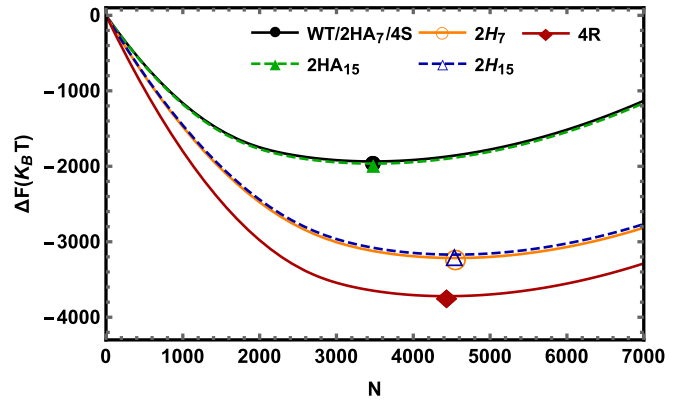
Charge Distribution	Virions	Effective Free energy	Optimal Length	Percent Change
	Tight	-2719.89	3387.44	N/A
	Loose	-2086.17	3471.32	+2.48%

FIG. 11. Schematic of two N-terminal tails with different charge distributions. As before each yellow rectangle represents an amino acid with a positive charge and effective size of $d = 0.2$ nm and blue rectangles represent neutral amino acids but with the same size. The table includes the effective encapsulation free energy of the RNA confined into a spherical shell, the optimal length of encapsulated RNA, and the percent change (theory) of optimal length of packaged RNA with respect to the first charge distribution. The salt concentration is $\mu = 500$ mM, and tail length is 4.5 nm for both cases. The total charge on the capsid is $Qc = 1440$. When the charges are distributed more evenly, the optimal length of the encapsidated genome increases. For the first line of the table, the distance between yellow rectangles is either zero or 0.2 nm, while for the second one it is either 0.2 nm or 0.4 nm.

fifth positively charged amino acids and the salt concentration. The figure reveals that as the distance increases from 0.2 to 2.2 nm, the optimal length of the encapsidated genome first increases and then later decreases.

To gain more insights into the experimental results, we also examined the impact on the optimal polymer length of a uniform charge distribution along the N-terminals versus a tight one as presented in Fig. 11. As shown in the figure, for a given tail length and number of positive charges, when the charges are distributed more uniformly along the N-terminals, the optimal length of encapsidated genome becomes longer.

IV. DISCUSSION AND SUMMARY

Despite the fact that many experiments have shown that the number of nucleotides packaged by capsid proteins increases with the number of charges on N-terminal tails, how the amount of encapsidated RNA depends on the distribution of the charges along and the length of the N-terminal domain of capsid proteins is not well understood. Our results presented in Fig. 5(b) for $T = 1$ capsid and Fig. 7 and Fig. 9 for $T = 3$ viruses show that the electrostatic interaction alone is not sufficient to explain the dependence of the amount of packaged RNA on the amino sequence of N-terminal tails in the BMV experiments [14]. For example, the amount of packaged RNA is different for the mutant $2HA_{15}$ and $2HA_7$ as illustrated in Fig. 2 whereas both have the same number of charges, very similar charge distribution, and the same peptide length. This reveals the importance of specific interactions that depend on the exact type of amino acids, RNA secondary or tertiary structures, and packaging sequences or signals, which involve the highly specific, nonelectrostatic interactions between sections of RNA and capsid proteins [8,9,42]. Our mean-field theory does not include this effect and thus cannot explain the experimental observation due to specific interactions; nevertheless, our theory can describe how the length of N-terminal tails and distribution of charges along the peptide control the amount of RNA packaged by BMV capsid proteins, consistent with the experimental data.

The simple case of two charges on the N-terminal tails of the $T = 1$ capsid (Fig. 5) shows clearly that when the distance between two positive charges increases, the optimal length of RNA encapsulated into the capsid also increases. A careful examination of Eq. (1) shows that the length of the encapsulated polyelectrolyte increases with the distance in order for the chain to be uniformly distributed between the two charges, lowering the entropy contribution [the first term in Eq. (1)] as much as possible. Figure 5 shows that the optimal length of the genome saturates and remains more or less constant beyond a certain distance between the two charges. This is mainly due to the fact that the optimal length of the genome for $T = 1$ is such that the density of the genome is low. When the distance between the charges is more than two Debye lengths ($\lambda_D = 0.438$ nm for $\mu = 500$ mM), the electrostatic interaction becomes very weak between the distant charges. It appears that the chain then prefers to reside only in the immediate vicinity of each positive charge along the peptide. More specifically, as the distance between the charges increases, the electrostatic does not promote encapsidation of longer genomes. Thus, the optimal length of the genome first increases and then it remains constant even if the distance between the charges increases further.

Figure 7 shows that the case for the $T = 3$ structure relevant to BMV experiments is more complex. As seen in the figure, the mutant 4R whose charge is increased by substitution instead of insertion (keeping the length constant) has less encapsidated RNAs than do $2H_7$ and $2H_{15}$, while all three mutants have the same number of charges on their tails. Our calculations reveal that since $2H_7$ and $2H_{15}$ have longer N-terminal tails, a longer genome is necessary for the chain to *uniformly* wrap around the tail keeping the entropic contribution in Eq. (1) low. However, the difference between the length of the genome encapsidated by wild-type proteins and the mutant $2HA_{15}$ proteins whose N-terminal length is increased by insertion of eight neutral amino acids is not as pronounced in our theory as in the experiments. This could be explained at least in part by the distance between the charged amino acids in the peptide. To understand the impact of the distance between the charges, we systematically examined the effect of the distance between the charges in the middle of the N-terminal tail as illustrated in Fig. 9. The results presented in the figure are quite intriguing as the optimal length of the genome first increases and then decreases for the three different salt concentrations presented in the figure.

The large distance between the two charges along the N-terminal tail provides more space for the genome to reside. The careful examination of Eq. (1) shows that due to the entropic consideration, the genome will be distributed more or less uniformly along the N-terminal leading to the packaging of longer genomes. However, as the distance between the charges increases and goes beyond two Debye length ($\lambda_D = 0.438$ nm for $\mu = 500$ mM, $\lambda_D = 0.565$ nm for $\mu = 300$ mM, and $\lambda_D = 0.979$ nm for $\mu = 100$ mM), the optimal length of RNA becomes shorter. This effect can be well understood by investigating the genome profiles presented in Fig. 8. When the distance between the fourth and fifth charges is very large, there will be two distinct peaks in the genome profile with almost no nucleotides between the charges indicating that the negatively charged RNA prefers to be localized mainly around

the positive charges. Figure 9 indicates that as the distance between the charges increases, at some point the optimal length of encapsidated RNA decreases resulting in the lower polymer density, which also reduces the entropy cost of formation of two completely separate peaks. Since the Debye length is longer for lower salt concentrations, the optimal length of the genome starts decreasing at $d = 1.8$ nm for $\mu = 100$ mM [2HA$_{15}$(M3)], $d = 1.4$ nm for $\mu = 300$ mM [2HA$_{15}$(M2)], and $d = 1.0$ nm for $\mu = 500$ mM [2HA$_{15}$(M1)]. While the behavior is the same for all three salt concentrations, the effect is less pronounced as the salt concentration decreases. Figure 11 further supports that for a given length and number of positive charges, the more uniformly charges are dispersed along the N-terminals, the longer the optimal length of the encapsidated genome becomes.

We emphasize that the goal of this paper has been to qualitatively explain the experimental results and to explore the impact of entropy and electrostatic interaction that depend on the distance between the charges and not the details of protein structures. A better quantitative comparison between the experiments and theory can be obtained if the theory includes many other effects such as counter-ion condensation, the presence of divalent ions, the structure of proteins, and the packaging signals discussed above.

In summary, in this paper we explore whether the variation in RNA packaging by BMV mutants observed in the experiments of Ni *et al.* and presented in Fig. 2 [14] can be understood by the mean-field theory incorporating electrostatics, excluded volume interaction, and RNA conformational entropy. In particular, we have calculated, as a function of the number and location of charges in the peptide tails, the free energy of an RNA confined in a spherical shell interacting with the N-terminal tails and ions. We find that the combined effect of the electrostatic interaction and the genome entropy considerations can shed light on many experimental data relevant to BMV assembly. While our mean-field theory cannot explain all the experimental data, we have been able to show that the location and the distance between charges along the N-terminal tails significantly influence the amount of packaged RNA. Understanding the factors contributing to the virus assembly and RNA packaging will pave the path for interfering with the different stages of the virus life cycle.

ACKNOWLEDGMENTS

[1] J. B. Bancroft, Adv. Virus Res. **16**, 99 (1970).

[2] M. Comas-Garcia, R. D. Cadena-Nava, A. L. N. Rao, C. M. Knobler, and W. M. Gelbart, J. Virol. **86**, 12271 (2012).

[3] C. Beren, L. L. Dreesens, K. N. Liu, C. M. Knobler, and W. M. Gelbart, Biophys. J. **113**, 339 (2017).

[4] A. Borodavka, S. Singaram, P. Stockley, W. Gelbart, A. Ben-Shaul, and R. Tuma, Biophys. J. **111**, 2077 (2016).

[5] M. F. Hagan and R. Zandi, Curr. Opin. Virol. **18**, 36 (2016).

[6] J. Ning, G. Erdemci-Tandogan, E. L. Yufenyuy, J. Wagner, B. A. Himes, G. Zhao, C. Aiken, R. Zandi, and P. Zhang, Nat. Commun. **7**, 13689 (2016).

[7] R. D. Cadena-Nava, Y. F. Hu, R. F. Garmann, B. Ng, A. N. Zelikin, C. M. Knobler, and W. M. Gelbart, J. Phys. Chem. B **115**, 2386 (2011).

[8] J. D. Perlmutter and M. F. Hagan, J. Mol. Biol. **427**, 2451 (2015).

[9] 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, and R. Tuma, J. Biol. Phys. **39**, 277 (2013).

[10] J. Sun, C. DuFort, M.-C. Daniel, A. Murali, C. Chen, K. Gopinath, B. Stein, M. De, V. M. Rotello, A. Holzenburg, C. C. Kao, and B. Dragnea, Proc. Natl. Acad. Sci. USA **104**, 1354 (2007).

[11] S. Li, G. Erdemci-Tandogan, J. Wagner, P. van der Schoot, and R. Zandi, Phys. Rev. E **96**, 022401 (2017).

[12] R. Zandi, B. Dragnea, A. Travesset, and R. Podgornik, Phys. Rep. **847**, 1 (2020).

[13] V. Sivanandam, D. Mathews, R. Garmann, G. Erdemci-Tandogan, R. Zandi, and A. L. N. Rao, Sci. Rep. **6**, 26328 (2016).

[14] P. Ni, Z. Wang, X. Ma, N. C. Das, P. Sokol, W. Chiu, B. Dragnea, M. Hagan, and C. C. Kao, J. Mol. Biol. **419**, 284 (2012).

[15] V. A. Belyi and M. Muthukumar, Proc. Natl. Acad. Sci. USA **103**, 17174 (2006).

[16] H. Tao, Z. Rui, and B. I. Shklovskii, Physica A **387**, 3059 (2008).

[17] C. Zeng, M. Hernando-Pérez, B. Dragnea, X. Ma, P. van der Schoot, and R. Zandi, Phys. Rev. Lett. **119**, 038102 (2017).

[18] D. L. Caspar and A. Klug, Cold Spring Harbor Symp. Quant. Biol. **27**, 1 (1962).

[19] A. Gopal, D. E. Egecioglu, A. M. Yoffe, A. Ben-Shaul, A. L. N. Rao, C. M. Knobler, and W. M. Gelbart, PLoS ONE **9**, e105875 (2014).

[20] J. D. Perlmutter, C. Qiao, and M. F. Hagan, eLife **2**, e00632 (2013).

[21] P. van der Schoot and R. Zandi, J. Biol. Phys. **39**, 289 (2013).

[22] P.-G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, New York, 1979).

[23] S. Li, H. Orland, and R. Zandi, J. Phys.: Condens. Matter **30**, 144002 (2018).

[24] I. Borukhov, D. Andelman, and H. Orland, Eur. Phys. J. B **5**, 869 (1998).

[25] A. Siber and R. Podgornik, Phys. Rev. E **78**, 051915 (2008).

[26] G. Erdemci-Tandogan, J. Wagner, P. van der Schoot, R. Podgornik, and R. Zandi, Phys. Rev. E **89**, 032707 (2014).

[27] G. Erdemci-Tandogan, J. Wagner, P. van der Schoot, R. Podgornik, and R. Zandi, Phys. Rev. E **94**, 022408 (2016).

[28] M. Janssen, A. Härtel, and R. van Roij, Phys. Rev. Lett. **113**, 268501 (2014).

[29] J. Wagner, G. Erdemci-Tandogan, and R. Zandi, J. Phys.: Condens. Matter **27**, 495101 (2015).

[30] I. Borukhov, D. Andelman, and H. Orland, Europhys. Lett. **32**, 499 (1995).

[31] A. Shafir, D. Andelman, and R. R. Netz, J. Chem. Phys. **119**, 2355 (2003).

[32] A. Siber, A. L. Bozic, and R. Podgornik, Phys. Chem. Chem. Phys. **14**, 3746 (2012).

[33] T. C. Lubensky and J. Isaacson, Phys. Rev. A **20**, 2130 (1979).

[34] S. I. Lee and T. T. Nguyen, Phys. Rev. Lett. **100**, 198102 (2008).

[35] K. Elleuch, F. Lequeux, and P. Pfeuty, J. Phys. I (France) **5**, 465 (1995).

[36] P.-G. de Gennes, Macromolecules **15**, 492 (1982).

[37] H. Ji and D. Hone, Macromolecules **21**, 2600 (1988).

[38] W. Bangerth, R. Hartmann, and G. Kanschat, ACM Trans. Math. Softw. **33**, 24/1 (2007).

[39] K. Bathe, *Finite Element Procedures* (Prentice Hall, New Jersey, 1996).

[40] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. (Springer, New York, 2006).

[41] G. Erdemci-Tandogan, H. Orland, and R. Zandi, Phys. Rev. Lett. **119**, 188102 (2017).

[42] 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, Proc. Natl. Acad. Sci. USA **112**, 2227 (2015).

[43] See http://www.rbvi.ucsf.edu/chimera.