


**Local rules for the self-assembly of a non-quasi-equivalent viral capsid**Giuliana Indelicato *Dipartimento di Matematica, Politecnico di Torino, 10129 Torino TO, Italy*Paolo Cermelli *Dipartimento di Matematica, Università di Torino, 10123 Torino TO, Italy*Reidun Twarock *Department of Mathematics and Department of Biology, University of York, York, YO10 5DD, United Kingdom*

(Received 17 December 2021; accepted 5 May 2022; published 3 June 2022)

The structures of many large bacteriophages, such as the P23-77 capsids, do not adhere strictly to the quasi-equivalence principle of viral architecture. Although the general architecture of the P23-77 capsids is classed as  $T = 28d$ , it self-assembles from multiple copies of two types of coat protein subunits, and the resulting hexameric capsomers do not conform to the Caspar-Klug paradigm. There are two types of hexamers with distinct internal organization, that are located at specific positions in the capsid. It is an open problem which assembly mechanism can lead to such a complex capsid organization. Here we propose a simple set of local rules that can explain how such non-quasi-equivalent capsid structures can arise as a result of self-assembly.

DOI: [10.1103/PhysRevE.105.064403](https://doi.org/10.1103/PhysRevE.105.064403)**I. INTRODUCTION**

Most viral capsids are made of multiple copies of a single protein subunit, referred to as the coat (or capsid) protein (CP). In general, the overall structure of an icosahedral viral capsid is well described in terms of a single CP and an architecture that conforms to either Caspar-Klug theory [1] or its extensions [2,3]. However, there are exceptions to this general rule, one of which is manifest in the capsids of some viruses that infect bacteria living in extreme environments, either with high salinity or high temperatures. It is thought that these phages have evolved relatively slowly and, therefore, the study of their structures may shed light on the properties of the progenitors of many of the more-recently emerging viruses.

Specifically, in this work we focus on bacteriophage P23-77, a dsDNA virus that infects *Thermus thermophilus* bacteria<sup>1</sup> [4–9]. Its capsid contains a lipidic membrane and is classed as a  $T = 28d$  architecture. However, its building blocks are two (distinct) CPs, VP16 and VP17. In the Caspar-Klug paradigm the CPs are organized in hexamers and pentamers, with the hexamers being made of six chemically identical proteins: Hence, every protein has approximately the same environment within the capsid, from which the name “quasi-equivalence” originates. On the other hand, in P23-77, the organization of the two CPs in the hexamers varies, breaking the quasi-equivalence principle.

Therefore, the following two questions arise: What is the evolutionary advantage of the specific layout of VP16 and

VP17 found in this virus, and which assembly mechanism can lead to this arrangement. The observation that there is a correlation between sites of stress concentration in the capsid and the structure of the hexamers at those sites suggests that the non-quasi-equivalent arrangement of the capsid proteins helps reinforce the shell against internal stresses [10]. Here we address the second question, proposing a local-rules model for self-assembly of the two major CPs that generates the observed capsid structure as a guaranteed outcome. A combinatorial approach akin to that developed in Ref. [11] shows that there are just three possible capsid layouts with the ratio of CP determined in the experimentally resolved capsid.

We will here explore the hypothesis that the specific arrangement of the two major capsid proteins seen in experiments is a consequence of a simple set of local assembly rules that arise from the rates of the attachment reactions between the two major CPs, i.e., they are a consequence of rules that determine the likelihood an attachment reaction occurs. Assembly is viewed here as a multistep process, by which each protein attaches according to a temporal sequence dictated by its affinity for the reactive sites, and the complexity of the capsid layout is an emergent feature of the timing of the individual assembly reactions. In this work we restrict ourselves to a limited set of local rules, motivated by suggestions in the experimental literature, that involve only two building blocks, and study the resulting assembly paths: This allows us to identify a unique and simple subset of local rules leading to the correct capsid layout.

As the literature on virus assembly is extensive, we only recall some of the main approaches based on local rules. One of the first are the geometric local rules by Berger *et al.* [12], which are formulated for a  $T = 7$  shell, such as polyoma

<sup>1</sup>Thermus virus P23-77 has been recently renamed as Hukuchivirus P23-77.

virus, in terms of angles between subunits, but lacking any information on affinities. The works of Zlotnick [13,14] are based on a thermodynamic model that allows us to show, among other results, how weak protein interactions can lead to stable viral capsids and how reversibility of bond formation allows for error correction [15,16]. Further contributions relevant to our work, in particular to the free energy of bond formation based on the intersubunit affinities, are due to Morozov *et al.* [17]. Our work has also been inspired by the approaches of Zandi *et al.* [18–21] and Hagan *et al.* [22,23] and their joint review [24]. Finally, we recall the work by Valbuena *et al.* [25,26], who exhibit explicit examples of self-assembly occurring by nucleation and subsequent growth and coalescence of patches, similarly to our model here.

## II. CAPSID STRUCTURE AND ASSEMBLY

Here we briefly describe the structure of the P23-77 capsid [4–9]. As mentioned above, it has a  $T = 28d$  structure, made of 60 penton proteins located at the fivefold axes, and 1620 major capsid proteins, of which 1080 are VP16 and 540 are VP17. The minor capsid protein VP11 is located peripheral to the lipidic membrane (without transmembrane domains) below the capsid. The basic structural units of the P23-77 capsid are the major coat proteins VP16 and VP17. There are substantial differences between these two proteins: While VP16 is small and occurs as a dimer, VP17 only occurs as a monomer and is larger than VP16 as it bears a turret protruding to the exterior of the capsid [6]. The presence of these turrets gives the capsid a typical crenellated appearance [cf. Fig. 1(a)].

Even though the structural units of the capsid are VP16 dimers and VP17 monomers, its overall layout can be described in terms of hexameric and pentameric capsomers [Fig. 1(b)]. Doing so, when differences between VP16 and VP17 are taken into account, it is clear that the capsid organization does not follow the Caspar-Klug scheme: There are hexamers near the twofold axes which have a different internal organization [Fig. 1(b), inset]. This violates quasi-equivalence, according to which all hexamers would be expected to be equal.

As to the minor capsid protein VP11, this occurs as a homodimer, with an estimated copy number of 147 in the capsid [7]. VP11 has a nonspecific strong Coulomb interaction with the DNA *in vitro*. However, due to its aspecific electrostatic nature, it is not known whether VP11 is actually associated with the DNA *in vivo*. What is important, though, is that VP11 dimers, in the presence of the lipidic membrane, tend to strongly attach to VP17. Hence, it has been conjectured that this protein acts as an assembly factor (scaffold protein), and its role is to help localize VP17 at specific sites on the membrane to facilitate assembly. Further, since VP11 occurs as a dimer in the virion, it is also likely that it plays a role in the formation of VP17 homodimers, that would not spontaneously form otherwise.

Regarding the assembly of P23-77, little is known, but it has been suggested that it might proceed as follows (for similar conjectures regarding other viruses of the PRD1-lineage see [8,27–30]): The lipidic membrane is formed around the circular dsDNA by recruiting some of the lipidic molecules

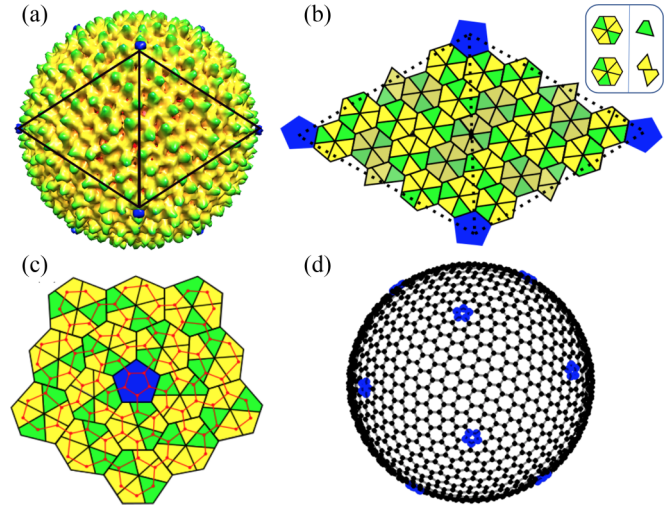


FIG. 1. The capsid layout of P23-77. (a) Cryo-EM reconstruction of the capsid (EMDB accession code EMD1525), showing the positions of VP16 dimers (yellow/light gray) and VP17 monomers (green/dark gray); (b) a tiling representation of two icosahedral faces in terms of tiles corresponding to the capsid building blocks, a VP17 monomer (green/dark gray) and VP16 homodimer (yellow/light gray) (inset, right); the shaded hexamers around the 2-fold axes have different internal organizations (see inset left for the internal structure of the two types of hexamer); (c) construction of the graph  $\mathcal{G}$ ; each protein corresponds to a node of the graph, and adjacent proteins are associated to nodes linked by an edge. Hence, edges of the graph correspond to interfaces between capsid proteins, but, as segments, are transverse to them. (d) the graph  $\mathcal{G}$  whose nodes represent the major capsid proteins.

of the cell; the membrane subsequently incorporates penton proteins as well as the minor capsid proteins VP11. In other viruses of the PRD1 lineage, for instance PRD1 itself, the lipidic membrane is formed before the DNA is internalized, and portal proteins are involved in the internalization process [31].

The general accepted view is that the penton proteins act as nucleation sites, around which assembly proceeds by incorporating VP16 and VP17 into the growing capsid shell around the lipidic membrane, with VP11 dimers having the role of facilitating the attachment of VP17 to the membrane in some of the positions.

## III. THE ASSEMBLY MODEL

The purpose of this paper is to explore whether there exist simple local assembly rules that can explain how the P23-77 capsid self-assembles without further regulation. Our model is based on a number of biologically motivated assumptions.

Our first main assumption is that assembly occurs by the incorporation of VP17 monomers and VP16 dimers at the lipidic membrane, which are the basic and only building blocks in this model. This is partly justified by the experimental evidence that VP17 are recruited as monomers by the minor capsid protein VP11 bound to the membrane, which has the consequence that VP17s, after the incorporation, attract the VP16 dimers to the membrane [7]. However, a comparison of the number of VP11 (147) and that of VP17 (540, i.e., 60 around the particle fivefold axes and 480 in other parts of

the capsid) suggests that VP11 cannot be solely responsible for all VP17 attachment events. Nevertheless, it is possible that assembly proceeds without incorporation of higher-order oligomers, and we investigate here the local rules under which this process results in the unique outcome seen experimentally. In the Discussion section, we compare this with other scenarios based on the incorporation of higher-order species, which would mostly require more complicated local rules.

Our second main assumption is regarding the basic biochemical mechanism underlying self-assembly. Namely, we postulate that the protein interfaces involved in the assembly process are endowed with different chemical affinities that, in turn, determine the speed at which each protein binds to its neighbors. Indeed, experimental evidence (yeast two-hybrid experiment reported in Ref. [6]) shows that the binding at inequivalent interfaces is characterized by dramatically different reaction rates. Attachment is here viewed as a sequential process, in which, at each step, all available reaction sites are saturated by those capsid proteins with greatest affinity among those available. Since binding reactions with higher affinities are faster, the peculiar layout of the capsid is an emergent feature of the timing of the attachment reactions as encoded by the interfacial energies.

Third, we further simplify our model by assuming that the detachment of already bound capsid proteins is not allowed, and the whole process is deterministic. It is well known that stochasticity is inherent in every chemical reaction, more so when weak bonds such as those involved in capsid assembly are considered. Also, detachment is one of the basic error-correction mechanisms in self-assembly and cannot be reasonably neglected. However, as stated above, our main focus here is on the basic combinatorics of local rules: The rule sets identified here can then be incorporated in future work in assembly models based on reaction kinetics using a Gillespie algorithm [32,33].

The last simplifying assumption we make here is that, from a mathematical point of view, we view assembly as the sequential coloring of the vertices of a predetermined  $T = 28d$  graph, which means that a  $T = 28d$  geometry is the defined outcome. In turn, this assumption has the consequence that, since nucleation occurs at the fivefold sites, which we assume to be occupied by the penton proteins, these sites are predetermined, and nucleation occurs precisely at the particle fivefold axes. There could be different biological mechanisms to ensure this. One possible scenario could be based on membrane-mediated interactions: Indeed, the pentons are the only capsid proteins in P23-77 located across the lipidic membrane and reaching into the interior of the capsid where they are in contact with the packaged genome [4]. It is therefore possible that interactions with the genome, and their strong anchoring in the membrane, define the penton positions. Hence, it seems reasonable to assume that pentons organize such that their mutual distance is maximal, possibly mediated also by membrane-mediated repulsive interactions, known to be relevant for proteins embedded in lipidic membranes [34,35]. The same physical mechanism could also keep the growing patches apart from each other, so that they might grow until merging. However, we do recognize this to be a fundamental issue to be further clarified by future research.

We now describe our mathematical model in some detail. Starting from the experimentally determined configuration of the P23-77 capsid [6] and the associated tiling representation [Fig. 1(b)], we first construct a spherical graph  $\mathcal{G}$  whose nodes correspond to the relative positions of the major capsid proteins VP16 and VP17, and whose edges correspond to the interfaces between adjacent proteins [Figs. 1(c) and 1(d)]. In what follows, we shall often employ the terms “site” and “interface” when referring to a node and an edge, respectively, of  $\mathcal{G}$  and denote by  $\mathcal{V}$  the set of nodes of  $\mathcal{G}$ . Note that the edges of  $\mathcal{G}$ , encoding adjacency, are transverse to the actual interfaces between the proteins. Note also that, were penton proteins be included, the spherical graph  $\mathcal{G}$  would be the skeleton of a typical Caspar-Klug  $T = 28d$  icosahedral capsid with 1680 proteins. However, to study assembly, we exclude the 60 penton proteins at the fivefold vertices, so that  $\mathcal{G}$  has 1620 nodes.

We model assembly as the sequential occupation of the graph  $\mathcal{G}$  by VP16 dimers and VP17 monomers. An *intermediate configuration* of the capsid is a mathematical representation of an assembly intermediate and is defined as a subset of the nodes of the graph together with a coloration of its nodes by two colors (or binary numbers), corresponding to VP16 and VP17. We choose here yellow (code 0) for VP16 and green (code 1) for VP17, and say that a site is *occupied* if it belongs to an intermediate configuration.<sup>2</sup>

As VP16 always occurs as a dimer, we restrict our considerations to intermediate configurations which are such that if a site is occupied by VP16, then also the adjacent site belonging to a different hexamer must be occupied by VP16 [Fig. 1(b)]. This implies that also VP17, with the exception of those adjacent to pentons, must occupy both adjacent sites belonging to different hexamers. Hence, two adjacent occupied sites in different hexamers must have the same color.

We denote by  $\mathcal{C} = \{(i, n_i)\}_{i \in \mathcal{I}}$  an intermediate configuration, where  $\mathcal{I}$  is the subset of occupied sites and  $n_i \in \{0, 1\}$  is the coloration of site  $i$ . Formally, then, self-assembly is described by a sequence  $\{\mathcal{C}_t\}_{t=0, \dots, T}$  such that  $\emptyset = \mathcal{I}_0 \subset \mathcal{I}_1 \subset \dots \subset \mathcal{I}_T = \mathcal{V}$  and if  $i \in \mathcal{I}_t$  then its color  $n_i$  is the same for all  $t' > t$ .

If two adjacent sites  $i$  and  $j$  are both occupied, we call the corresponding interface  $ij$  the *bond* between  $i$  and  $j$ . Each bond is characterized by its *affinity*, which only depends on the colors of the occupied sites. The affinity of a bond determines the ratio of the forward and backward rate of the reaction involved in its formation, a higher affinity meaning a faster reaction. We denote the affinities by  $\alpha_{10}$ ,  $\alpha_{01}$ ,  $\alpha_{00}$ , and  $\alpha_{11}$  (Fig. 2), where  $\alpha_{ij}$  correspond to an interaction of units of color  $i$  and  $j$ , and the order  $ij$  refers to their relative position within a hexamer. In particular,  $\alpha_{10}$  is the affinity for a bond between a VP17 and a VP16 belonging to the same hexamer, with the *VP17 preceding the VP16 in counterclockwise order* (Fig. 2). We distinguish between type I and type II depending on whether the incoming unit is a VP16 dimer or a VP17 monomer.  $\alpha_{01}$  denotes the bond between a VP16 and VP17 belonging to the same hexamer with the *VP16 preceding*

<sup>2</sup>Note that our choice here is different from the usual convention in lattice models, where the index 0 denotes unoccupied sites.

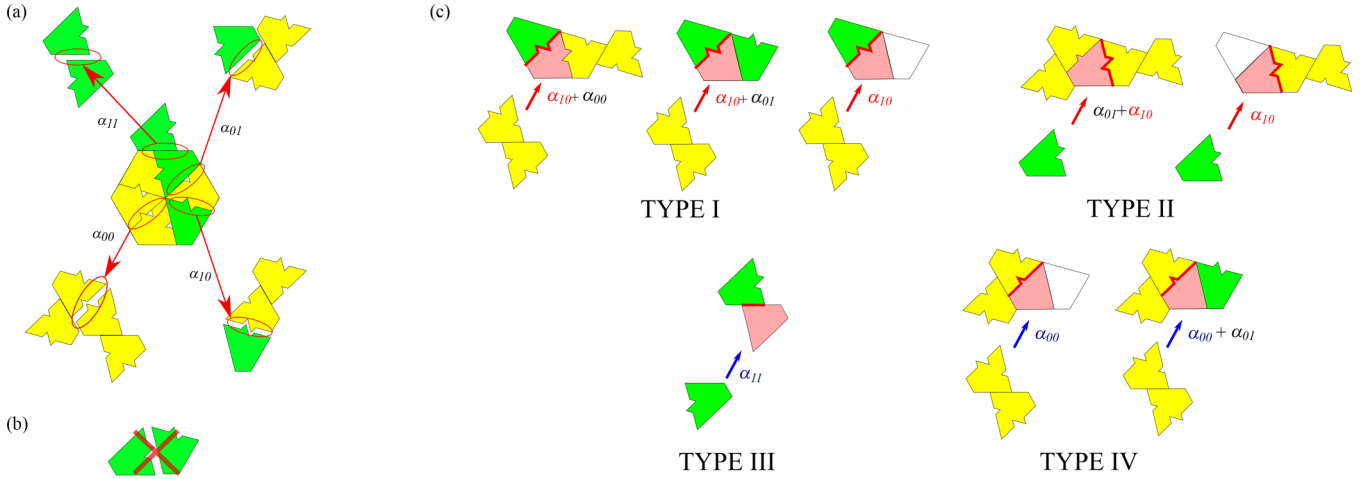


FIG. 2. A graphical depiction of the local rules. (a) The energies of different types of interfaces are shown; jagged lines at the protein interfaces encode allowed interactions and orientations of the protein units as jagged markings must match; e.g. (b) intra-hexamer bonds between VP17 are forbidden. (c) classification of the reactive sites, highlighted in pink (dark gray), with the relevant reactive interfaces in red (black), and the most probable reactions indicated by either red (light gray) (fast reaction) or blue (dark gray) (slow reaction) arrows. Type I and II, the highlighted reactive interfaces have affinity  $\alpha_{10}$  for VP16 and VP17, respectively; type III, the reactive interface has affinity  $\alpha_{11}$  for VP17 (this is the VP17 dimerisation reaction, possibly mediated by VP11); type IV, the reactive interface has affinity  $\alpha_{00}$  for VP16.

the VP17 in counterclockwise order.  $\alpha_{00}$  represents a bond between two adjacent VP16 in the same hexamer (Fig. 2, type IV), and  $\alpha_{11}$  a bond between two VP17 in different hexamers (Fig. 2, type III). This bond is involved in the formation of the putative VP17 dimers.

Note that we have not assigned an affinity to a bond between two VP17 belonging to the same hexamer; since this bond does not occur in the capsid we view it as forbidden in our assembly model [Fig. 2(b)].

Given an intermediate configuration  $C$ , a *reactive site* is an unoccupied node that is adjacent to an occupied site. Also, we define a *reactive interface* to be an interface between a reactive site and an adjacent occupied site. By the affinity of a reactive interface to some CP, or of some CP to a reactive interface, we mean the affinity of the bond formed when that CP occupies the reactive site to which the interface belongs.

### A. Local rules

Our goal is to determine whether there are local assembly rules, to be formulated in terms of binding affinities between the major capsid proteins VP16 and VP17, that guarantee the outcome of the P23-77 capsid architecture. We have examined a number of possible combinations, and the basic result is shown in Fig. 3: The observed capsid layout can only be obtained in the sequential attachment process if the affinities are ordered in the specific manner described in Fig. 2.

This gives rise to the following model assumptions:

(1) VP17 and VP16 preferentially bind when VP17 precedes VP16 in the cyclic counterclockwise order of the hexamer, i.e.,  $\alpha_{10} > \alpha_{01}$ .

(2) VP16 dimers bind to each other with an affinity lower to that for the counterclockwise binding of VP17 to VP16, i.e.,  $\alpha_{00} < \alpha_{10}$ , but higher than that for the counterclockwise binding of VP16 to VP17, i.e.,  $\alpha_{00} > \alpha_{01}$ .

(3) VP17 monomers belonging to different hexamers form dimers with weak affinity  $\alpha_{11}$ .

Note that assumption 3 is justified as follows: VP11 occurs as a homodimer and, when associated to the lipidic membrane, has high affinity for VP17, suggesting that this minor capsid protein plays a role in promoting the formation of VP17 dimers at these sites.<sup>3</sup>

The local rules above correspond to the following order relations between the affinities:

$$\alpha_{10} > \alpha_{00} > \alpha_{01}. \quad (1)$$

In particular, the main local constraint is that there is a preferential attachment of VP17 monomers at a specific reactive interface of the VP16 dimers, i.e., that  $\alpha_{10}$  dominates all other affinities. Also, recall that two VP17 monomers cannot bind to each other within the same hexamer.

These local assembly rules imply an order of sequential attachment for the incoming protein units: denoting by  $\tau_{ij}$  the expected time of the reaction with affinity  $\alpha_{ij}$ , and recalling that the reaction time is inversely proportional to the affinity, Eq. (1) yields

$$\tau_{10} < \tau_{00} < \tau_{01}. \quad (2)$$

In order to implement these rules, we categorize the reactive sites as follows [cf. Fig. 2(c)]:

Type I: These sites involve an intrahexamer VP17 reactive interface and can only be occupied by a VP16; the attachment reaction has affinity  $\alpha_{10}$ .

<sup>3</sup>The experimentally determined number of VP11 is 147 [7], which is consistent with either one or two dimers per fundamental domain. Indeed, it is possible to show (proof omitted here) that only one VP17 dimerisation event per fundamental domain is necessary to assemble the correct capsid as all other VP17 placements are consequences of rules 1 and 2.



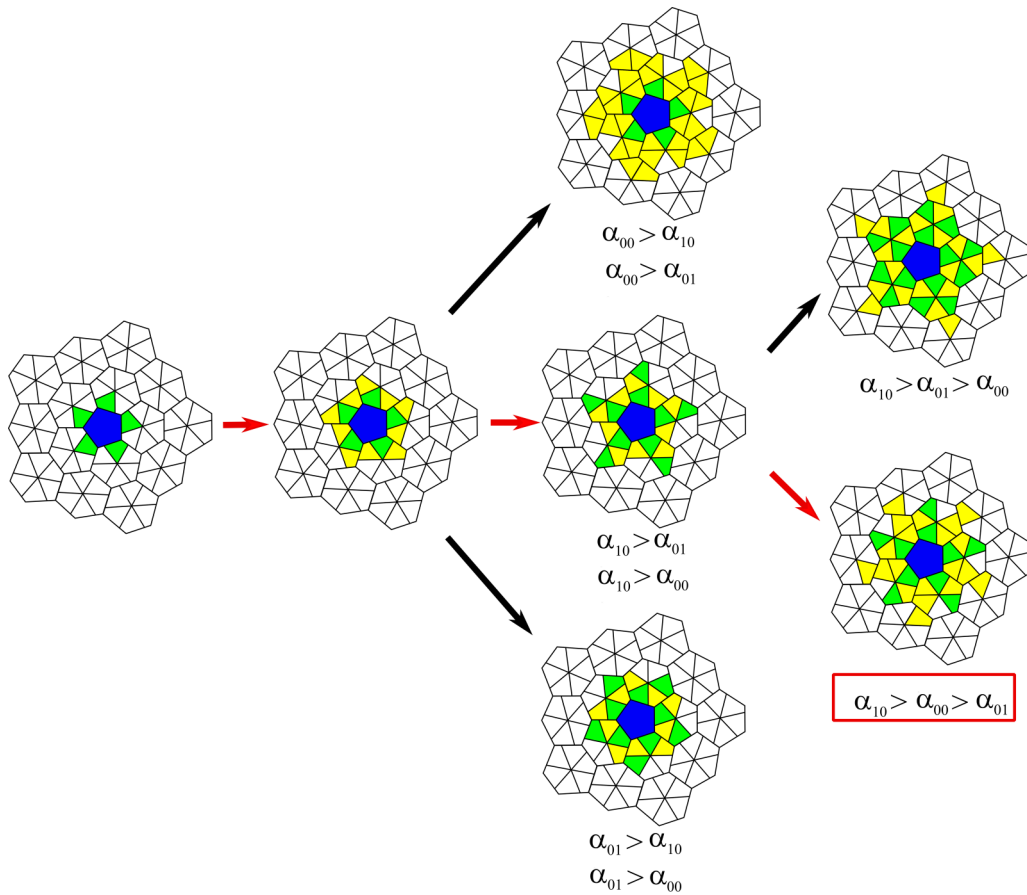


FIG. 3. Local rules and assembly paths in the sequential attachment model. The first steps of the growth process corresponding to different orderings of the affinities are shown. The unique assembly path leading to the observed capsid layout is highlighted in red (light gray).

Type II: These sites involve an intrahexamer VP16 reactive interface and can be occupied by either a VP16 or by a VP17. The attachment of a VP17 has affinity  $\alpha_{10}$ , while the attachment of a VP16 has lower affinity  $\alpha_{00}$ .

Type III: This is a VP17-dimerization site: It can only be occupied by a VP17 with affinity  $\alpha_{11}$ .

Type IV: These sites involve an intrahexamer VP16 reactive interface and can be occupied by either a VP16 or a VP17. Contrary to type II sites, the attachment of VP16 has affinity  $\alpha_{00}$ , while the attachment of a VP17 has affinity  $\alpha_{01}$ , which we assume to be negligible.

The above local rules are not sufficient to recover the correct capsid layout: As seen in the next section, it is necessary to assume also that the VP17 dimerization reaction is slower than the fast reaction, and its rate is comparable to that of the slow VP16-to-VP16 attachment, i.e.,  $\tau_{11} \sim \tau_{00}$ . Also, in what follows, we have assumed for simplicity that  $\alpha_{01} = 0$ , so that the counterclockwise attachment of a VP16 to a VP17 and the clockwise attachment of a VP17 to a VP16 does not spontaneously occur in our model. Since the buried surface area of the involved interfaces are nonnegligible, the corresponding affinity cannot strictly vanish: Our assumption just means that the reaction is much slower, and therefore dominated, by the other three reactions.

## B. Results

Using the local rules described above, we model assembly as a sequence of attachment reactions as follows.

After being triggered by nucleation at the pentons, occurring either at random, or at fixed times and sites as described below, assembly then proceeds by steps. The time interval between these steps is set by the fast reaction times: At each step, available type I and II reactive sites are occupied, while type III and IV reactive sites, involving slow reactions, are occupied every two steps. At each step, new reactive sites are created, and the procedure is iterated recursively. The assumption that slow reactions occur approximately every two steps (see Fig. 4, bottom) requires that the reaction times must satisfy a certain set of bounds, as discussed below. The procedure, following nucleation at a single penton site, is depicted in Fig. 4 and can be schematized as follows:

Step 1: Add one or more pentons at the capsid five-fold axes, that serve as nucleation sites, and attach VP17 monomers at the reactive sites around the pentons, so that there is now a single type of reactive site: type I.

Step 2: The fast reaction is the attachment of VP16 dimers at reactive sites of type I, with affinity  $\alpha_{10}$ . This generates type II and type IV reactive sites.

Step 3: Now the fastest reaction is the attachment of VP17 monomers at reactive sites of type II, with

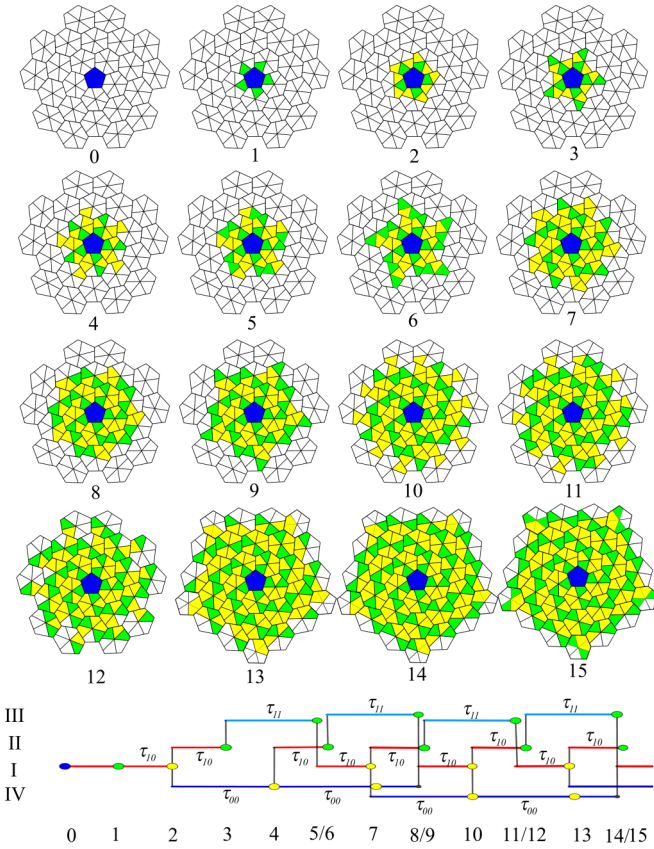


FIG. 4. Results of the sequential attachment model. Top, the first 15 steps of assembly following nucleation at a penton. Bottom, the assembly path; red (light gray) and blue (dark gray) lines are fast and slow reactions, respectively. The circles depicted at the end of each line indicate the type of CPs that attach in each reaction, while the Roman numerals to the left are the types of the corresponding reactive sites.

affinity  $\alpha_{10}$ . There are now two types of reactive sites: type III and type IV.

Step 4: Reactive sites of type IV, that were formed at step 2, are now available for the attachment of VP16. Since this is a slow reaction, with affinity  $\alpha_{00} < \alpha_{10}$ , it takes two iterative steps to be completed (from step 2 to step 4). The reactive sites are now types II–IV.

Step 5: The slow dimerisation reaction takes place now and type III reaction sites that were formed at step 3 are occupied by VP17 monomers: The reaction involves affinity  $\alpha_{11}$ . Now the reactive sites are types I, II, and IV.

Step 6: The type II reactive sites formed at step 4 are now occupied by VP17 monomers. Again here the reaction is fast with affinity  $\alpha_{10}$ . The reactive sites are type I, III, and IV.

Step 7: Reactive sites of type I, that have been formed at step 5, are now available for the attachment of VP16 (fast reaction). Also, reactive sites of type IV formed at step 4 are now available for the attachment of VP16 (slow reaction). These two reactions are not necessarily simultaneous, but their relative order is irrelevant: What is important is that they both occur between steps 6 and 8 (see below for bounds on the

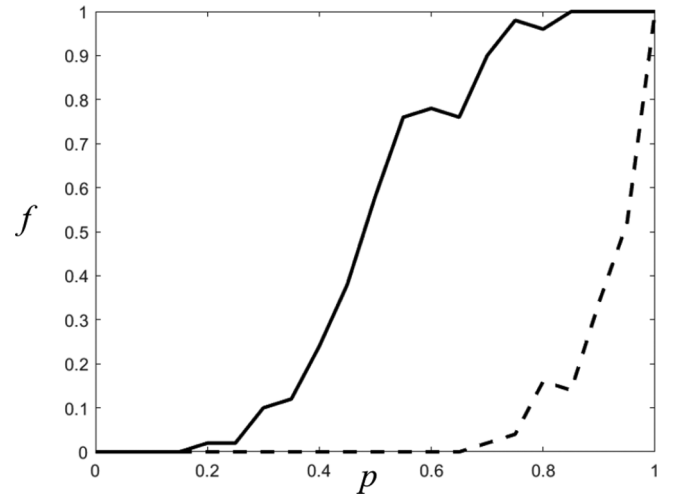


FIG. 5. Capsid yield for different nucleation probabilities and rules for the dimerisation of VP17. The proportion  $f$  of complete capsids is shown as a function of the nucleation probability  $p$  (50 replicas for each value of  $p$ ). The solid line corresponds to the dimerisation of VP17 at type III sites preceding the attachment of VP17 at type II sites [inequality (3)], and the dashed line to the reverse ordering.

reaction times that enforce this requirement). The reactive sites are now types II–IV.

Steps 5–7 are then iterated (steps 8 to 15 in Fig. 4). Since each penton generates a growing patch, the procedure is terminated if, and when, all patches have merged and the capsid is complete.

In our model the local rules determine the ordering of the attachment reactions. However, it is clear that if each reaction occurs with a given rate, the above ordering cannot be conserved after a finite number of steps, unless the slow reaction times are exactly twice the fast reaction times. But the ordering only needs to be conserved for the number of steps necessary to reach the stage at which different growth patches, nucleated at different penton sites, merge. Inspection of Fig. 4 shows that this occurs approximately after 13 steps (depending on when the other patches have nucleated, see Discussion below). Hence, we may use the diagram at the bottom of Fig. 4, in which generic reaction times are indicated, to show that the ordering chosen in the algorithm above is conserved up to the 14th step at least if the following inequalities hold for the reaction times:

$$\tau_{10} < \tau_{00} < \tau_{10} + \tau_{11} < 2\tau_{00} < 3\tau_{10} + \tau_{11}. \quad (3)$$

Also, we have assumed that the slow VP17 dimerization reaction always occurs after the occupation of type I sites and before the occupation of type II sites. This requirement can be removed, but, as shown in Fig. 5, this yields suboptimal capsid growth. Inspection of Fig. 4 shows that the above assumption requires that

$$\tau_{10} < \tau_{11} < \tau_{00} < 2\tau_{10}, \quad (4)$$

which, by the way, also implies the validity of Eq. (3). The model is valid as long as these inequalities are fulfilled, and there is therefore a large degree of robustness in the model against variation in the reaction times.

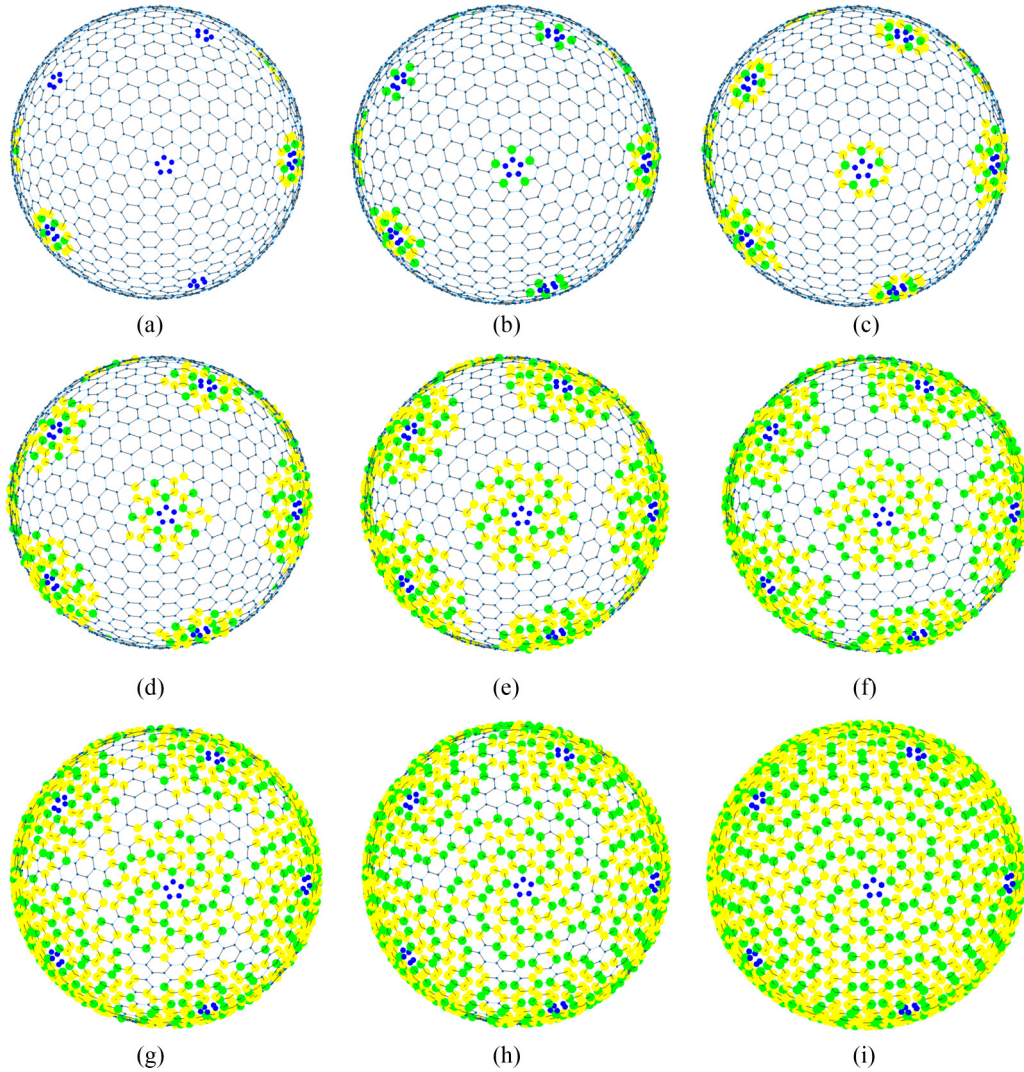


FIG. 6. Snapshots of a complete assembly path, with nucleation probability  $p = 0.85$

For a single nucleation site, our procedure yields the correct capsid layout up to the nearest twofold axes, but it starts to deviate from the correct layout a couple of steps after reaching the twofold axes (Fig. 4, step 15). We have therefore explored numerically in Matlab [36] different scenarios, in which a variable number of nucleation sites were activated at random steps, with constant nucleation probability  $p$  at each time step. The results are reported in Fig. 5, which shows that the proportion  $f$  of complete capsids is greater than 95% under the assumption Eq. (3) and for a nucleation probability  $p$  greater than 0.8. Also, Fig. 5 shows that the inequalities Eq. (3) are necessary for the robustness of the model as otherwise the yield is dramatically decreased. Hence, multiple nucleation sites are required to achieve the correct geometric outcome: Part of one of the numerically computed complete assembly paths is shown in Fig. 6, and the algorithm is sketched in Fig. 7.

#### IV. DISCUSSION

Cryo-electron microscopy has revolutionized our understanding of virus structure, providing unprecedented insights into complex viral architectures. It also has opened up questions regarding the processes that lead to the observed outcomes. The mechanisms underpinning the assembly of quasi-equivalent capsids are by now fairly well understood. In this case, the structural proteins abide to the same type of local interactions across the capsid, resulting in the latticelike architectures modelled in Caspar-Klug theory. It is much less clear, however, how the formation of architectures violating the quasi-equivalence principle is regulated. In particular, it has been an open question whether the distinct local environments seen in such structures could arise from a universal set of local rules. Otherwise, how would the system “sense” which of the distinct environments it needs to implement at a given site, especially since there does not appear to be a structured scaffold to guide the attachment?



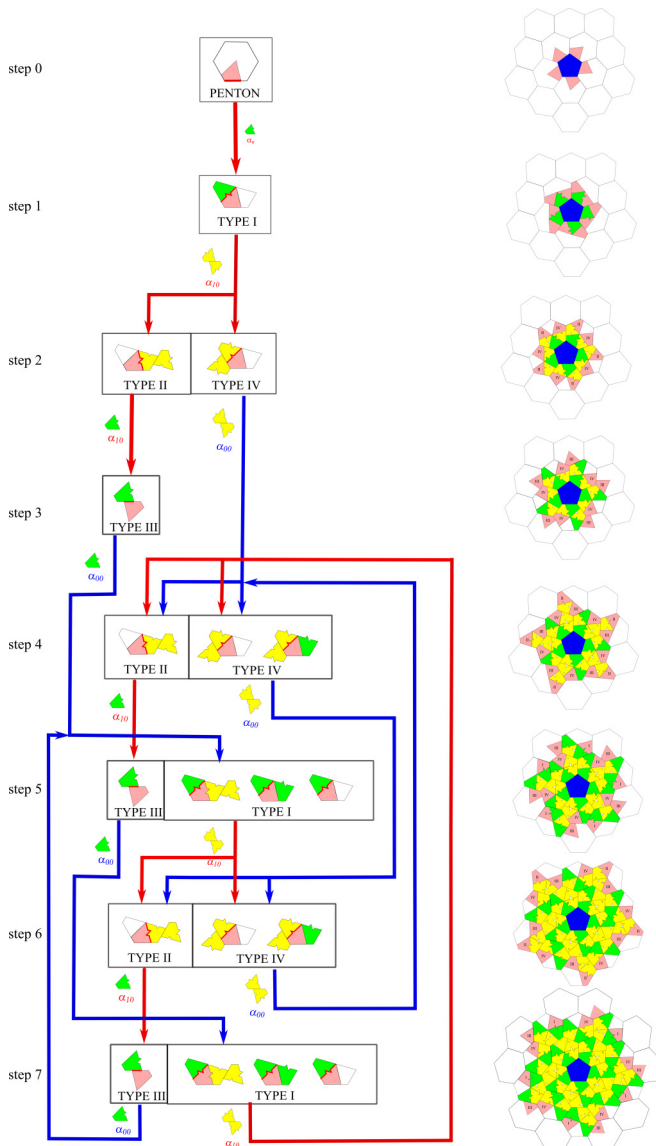


FIG. 7. A flow diagram for steps 1 to 7 of the assembly process. Fast and slow reactions are denoted by red (light gray) and blue (dark gray) arrows, respectively. The corresponding assembly intermediates are shown on the right, with reactive sites indicated in salmon (dark gray).

Using the P23-77 capsid as a model system, we demonstrate here that such complex behavior can indeed arise from a simple set of local rules that is applied universally across the capsid. We show that local rules formulated in terms of the binding affinities between the structural proteins at different interfaces can account for formation of the observed structures without any external regulation. The local rules we propose

here are based on two simple assumptions. First, there is a reactive interface whose affinity dominates all others. Second, since the affinities determine the reaction rates, assembly is fundamentally regulated by the interplay of fast and slow reactions. We note that the rules on the relative affinities at different interaction sites are required to guarantee the correct outcome during the early stages of assembly. In order to achieve correct assembly of the entire capsid using these rules, two further assumptions are required: (1) Assembly must (randomly) nucleate at all 12 pentamers within a relatively short period of time; and (2) there is one site per asymmetric unit (fundamental domain) of the capsid where a VP17 dimer is required. We suggest that the minor capsid protein VP11 that is connected to the lipidic membrane may play a role in this. This would be consistent with the experimental evidence that VP11 is required for correct assembly, as well as with the experimentally estimated copy number of VP11. It seems likely that the role of VP11 is to recruit VP17 to the lipidic membrane, in addition to possibly promoting its dimerization.

We note that our model uses VP17 monomers and VP16 dimers as capsomers, i.e., as the units of assembly. It is possible that assembly could also involve higher-order species such as the heterotrimer and heterotetramer observed by X-ray crystallography [6]. However, the local rules involving such building blocks would have to be more complex. In particular, given the fact that the heterotrimer is part of the heterotetramer, it is not *a priori* clear from the point of view of local rules why it is the heterotrimers, rather than the heterotetramers, that surround the pentons at the particle fivefold axes. We leave an in-depth investigation of which higher-order species might be consistent with different local assembly rule sets to future work.

It is, however, remarkable that simple local rules do exist that would guarantee the complex outcome of a large, non-quasi-equivalent capsid geometry from the most basic building blocks, the VP17 monomer and the VP16 dimer. Our model also suggests roles for other structural proteins, such as the minor capsid protein VP11, and informs on the importance of the nucleation step at multiple sites to achieve the correct outcome.

## ACKNOWLEDGMENTS

The authors acknowledge the support of the following grants; R.T. and G.I.: EPSRC Established Career Fellowship (EPR0232041); R.T.: Royal Society Wolfson Fellowship (RSWFR1180009); Wellcome Trust Joint Investigator Award with Professor Peter Stockley (University of Leeds) (110145 and 110146). P.C.: “Stochastic and statistical models and methods for the applications” (University of Torino, SACLRILO-18-01); P.C. and G.I.: PRIN 2017 project “Mathematics of active materials: from mechanobiology to smart devices.”

[1] D. Caspar and A. Klug, Physical principles in the construction of regular viruses, *Cold Spring Harb. Symp. Quant. Biol.* **27**, 1 (1962).

[2] R. Twarock and A. Luque, Structural puzzles in virology solved with an overarching design principle, *Nat. Commun.* **10**, 4414 (2019).



- [3] R. Twarock, A tiling approach to virus capsid assembly explaining a structural puzzle in virology, *J. Theor. Biol.* **226**, 477 (2004).
- [4] S. T. Jaatinen, L. J. Happonen, P. Laurinmäki, S. J. Butcher, and D. H. Bamford, Biochemical and structural characterisation of membrane-containing icosahedral dsDNA bacteriophages infecting thermophilic *Thermus thermophilus*, *Virology* **379**, 10 (2008).
- [5] I. Rissanen, A. Pawlowski, K. Harlos, J. M. Grimes, D. I. Stuart, and J. K. H. Bamford, Crystallization and preliminary crystallographic analysis of the major capsid proteins VP16 and VP17 of bacteriophage P23-77, *Acta Cryst. F* **68**, 580 (2012).
- [6] I. Rissanen, J. M. Grimes, A. Pawlowski, S. Mantynen, K. Harlos, J. K. H. Bamford, and D. I. Stuart, Bacteriophage P23-77 capsid protein structures reveal the archetype of an ancient branch from a major virus lineage, *Structure* **21**, 718 (2013).
- [7] A. Pawlowski, A. M. Moilanen, I. A. Rissanen, J. A. E. Määttä, V. P. Hytönen, J. A. Ihalainen, and J. K. H. Bamford, The minor capsid protein VP11 of Thermophilic Bacteriophage P23-77 facilitates virus assembly by using lipid-protein interactions, *J. Virol.* **89**, 7593 (2013).
- [8] A. Pawlowski, I. Rissanen, J. K. H. Bamford, M. Krupovic, and M. Jalasvuori, Gammasphaerolipovirus, a newly proposed bacteriophage genus, unifies viruses of halophilic archaea and thermophilic bacteria within the novel family Sphaerolipoviridae, *Arch. Virol.* **159**, 1541 (2014).
- [9] A. Pawlowski, *Thermus Bacteriophage P23-77: Key Member of a Novel, but Ancient Family of Viruses from Extreme Environments*, Academic Dissertation, University of Jyväskylä (Jyväskylä University Printing House, Jyväskylä, 2015).
- [10] G. Indelicato, P. Cermelli, and R. Twarock, Surface stresses in complex viral capsids and non-quasi-equivalent viral architectures, *J. R. Soc. Interface* **17**, 20200455 (2020).
- [11] G. Indelicato, P. Burkhard, and R. Twarock, Classification of self assembling protein nanoparticle architectures for applications in vaccine design, *Royal Soc. Open Sci.* **4**, 161092 (2017).
- [12] B. Berger, P. W. Shor, L. Tucker-Kellogg, and J. King, Local rule-based theory of virus shell assembly, *Proc. Natl. Acad. Sci. USA* **91**, 7732 (1994).
- [13] P. Ceres and A. Zlotnick, Weak protein-protein interactions are sufficient to drive assembly of hepatitis B virus capsids, *Biochemistry* **41**, 11525 (2002).
- [14] D. Endres and A. Zlotnick, Model-based analysis of assembly kinetics for virus capsids or other spherical polymers, *Biophys. J.* **83**, 1217 (2002).
- [15] A. Zlotnick, Are weak protein-protein interactions the general rule in capsid assembly? *Virology* **315**, 269 (2003).
- [16] A. Zlotnick, Distinguishing reversible from irreversible virus capsid assembly, *J. Mol. Biol.* **366**, 14 (2007).
- [17] 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).
- [18] R. F. Bruinsma, W. M. Gelbart, D. Reguera, J. Rudnick, and R. Zandi, Viral Self-Assembly as a Thermodynamic Process, *Phys. Rev. Lett.* **90**, 248101 (2003).
- [19] R. Zandi, B. Dragnea, A. Travasset, and R. Podgornik, On virus growth and form, *Phys. Rep.* **847**, 1 (2020).
- [20] S. Panahandeh, S. Li, and R. Zandi, The equilibrium structure of self-assembled protein nano-cages, *Nanoscale* **10**, 22802 (2018).
- [21] J. Ning, G. Erdemci-Tandogan, E. L. Yufenyuy, J. Wagner, B. A. Himes, G. Zhao, C. Aiken, R. Zandi, and P. Zhang, In vitro protease cleavage and computer simulations reveal the HIV-1 capsid maturation pathway, *Nat. Commun.* **7**, 13689 (2016).
- [22] M. F. Hagan, O. M. Elrad, and R. L. Jack, Mechanisms of kinetic trapping in self-assembly and phase transformation, *J. Chem. Phys.* **135**, 104115 (2011).
- [23] T. C. T. Michaels, M. M. J. Bellaiche, M. F. Hagan, and T. P. J. Knowles, Kinetic constraints on self-assembly into closed supramolecular structures, *Sci. Rep.* **7**, 12295 (2017).
- [24] M. Hagan and R. Zandi, Recent advances in coarse-grained modeling of virus assembly, *Curr. Opin. Virol.* **18**, 36 (2016).
- [25] A. Valbuena and M. G. Mateu, Kinetics of surface-driven self-assembly and fatigue-induced disassembly of a virus-based nanocoating, *Biophys. J.* **112**, 663 (2017).
- [26] A. Valbuena, S. Maity, M. G. Mateu, and W. H. Roos, Visualization of single molecules building a viral capsid protein lattice through stochastic pathways, *ACS Nano* **14**, 8724 (2020).
- [27] L. De Colibus, E. Roine, T. S. Walter, S. L. Ilca, X. Wang, N. Wang, A. M. Roseman, D. Bamford, J. T. Huisken, and D. I. Stuart, Assembly of complex viruses exemplified by a halophilic euryarchaeal virus, *Nat. Commun.* **10**, 1456 (2019).
- [28] S. Mäntynen, L.-R. Sundberg, H. M. Oksanen, and M. M. Poranen, Half a century of research on membrane-containing bacteriophages: bringing new concepts to modern virology, *Viruses* **11**, 76 (2019).
- [29] D. Gil-Carton, S. T. Jaakkola, D. Charro, B. Peralta, D. Castañó-Díez, H. M. Oksanen, D. H. Bamford, and N. G. A. Abrescia, Insight into the assembly of viruses with vertical single  $\beta$ -barrel major capsid proteins, *Structure* **23**, 1866 (2015).
- [30] I. Santos-Pérez, D. Charro, D. Gil-Carton, M. Azkargorta, F. Elortza, D. H. Bamford, H. M. Oksanen, and N. G. Abrescia, Structural basis for assembly of vertical single  $\beta$ -barrel viruses, *Nat. Commun.* **10**, 1184 (2019).
- [31] C. Hong, H. M. Oksanen, X. Liu, J. Jakana, D. H. Bamford, and W. Chiu, A structural model of the genome packaging process in a membrane-containing double stranded DNA virus, *PLoS Biol.* **12**, e1002024 (2014).
- [32] S. R. Hill, R. Twarock, and E. C. Dykeman, The impact of local assembly rules on RNA packaging in a  $T = 1$  satellite plant virus, *PLoS Comput. Biol.* **17**, e1009306 (2021).
- [33] E. C. Dykeman, P. G. Stockley, and R. Twarock, Solving a Levinthal's Paradox for Virus Assembly suggests a novel antiviral therapy, *Proc. Natl. Acad. Sci. USA* **111**, 5361 (2014).
- [34] J. Gao, R. Hou, L. Li, and J. Hu, Membrane-mediated interactions between protein inclusions, *Front. Mol. Biosci.* **8**, 811711 (2021).
- [35] T. Idema and D. J. Kraft, Interactions between Mmodel inclusions on closed lipid bilayer membranes, *Curr. Opin. Colloid Interface Sci.* **40**, 58 (2019).
- [36] MATLAB, (R2021b). The MathWorks Inc. Natick, MA (2010).