Acta Cryst A
FOUNDATIONS
ADVANCES

ISSN 2053-2733

Approximation of virus structure by icosahedral tilings

D. G. Salthouse,[a]* G. Indelicato,[a] P. Cermelli,[b] T. Keef[a] and R. Twarock[a]

[a]Departments of Mathematics and Biology, York Centre for Complex Systems Analysis, University of York, YO10 5DD, England, and [b]Department of Mathematics, University of Torino, Italy. *Correspondence e-mail: dsalthou@biologie.ens.fr

Received 7 February 2015
Accepted 2 April 2015

Edited by J.-G. Eon, Universidade Federal do Rio de Janeiro, Brazil

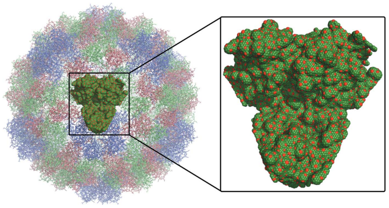
Keywords: virus capsid structure; icosahedral tilings; polygonal approximation.

Viruses are remarkable examples of order at the nanoscale, exhibiting protein containers that in the vast majority of cases are organized with icosahedral symmetry. Janner used lattice theory to provide blueprints for the organization of material in viruses. An alternative approach is provided here in terms of icosahedral tilings, motivated by the fact that icosahedral symmetry is non-crystallographic in three dimensions. In particular, a numerical procedure is developed to approximate the capsid of icosahedral viruses by icosahedral tiles *via* projection of high-dimensional tiles based on the cut-and-project scheme for the construction of three-dimensional quasicrystals. The goodness of fit of our approximation is assessed using techniques related to the theory of polygonal approximation of curves. The approach is applied to a number of viral capsids and it is shown that detailed features of the capsid surface can indeed be satisfactorily described by icosahedral tilings. This work complements previous studies in which the geometry of the capsid is described by point sets generated as orbits of extensions of the icosahedral group, as such point sets are by construction related to the vertex sets of icosahedral tilings. The approximations of virus geometry derived here can serve as coarse-grained models of viral capsids as a basis for the study of virus assembly and structural transitions of viral capsids, and also provide a new perspective on the design of protein containers for nanotechnology applications.

1. Introduction

Viruses are well known for their devastating impact on health and the economy, but the development of technology has enabled us to deepen our understanding of these entities, and the possible applications of these results in the medical field and nanotechnology are promising. The work presented here focuses solely on the study of protein shells (also called capsids) of icosahedral viruses. The capsid protects the genomic material of the virus and interacts with the host cell during the infection process. Understanding the structure of these capsids is therefore of utmost importance not only for the design of new antiviral drugs, but also for many other applications in nanotechnology. For example, the construction of protein cages from viral proteins (lacking viral genome), referred to as virus-like particles, can be used as gene vectors, *i.e.* to transport genetic material into cells for therapeutic purposes (Ma *et al.*, 2012). These (non-infectious) particles provide containers preventing premature degradation of drugs which, combined with the high host-cell specificity of viruses, can deliver these drugs to the specific targeted tissues. In cancer therapy, viruses carrying gold particles have been used to target cells for photothermal cancer treatment (Everts *et al.*, 2006).

The introduction of the concept of quasi-equivalence by Caspar & Klug (1962) was the stepping-stone for subsequent



mathematical models of the structural organization of viruses. Caspar and Klug demonstrated that if the proteins of the capsid are in *quasi-equivalent* local environments, *i.e.* if the local bonding environments of all proteins are similar, then their positions and orientations are encoded by a subtriangulation T of the icosahedral surface. Although applicable to many viruses, this theory provides no radial information on the structure of the capsid. Janner was the first to use three-dimensional lattices to study the mathematical principles underpinning the structural organization of viruses at different radial levels based on their crystal structures (Janner, 2006, 2010*a*,*b*). Although some of the essential viral features could be modelled in this context, the icosahedral nature of viral capsids suggests the use of tilings with non-crystallographic symmetry to better understand the geometrical constraints imposed on virus architecture. A technique for describing the structure of viral capsids was developed based on affine extensions of the symmetry group of the capsid (Keef *et al.*, 2008; Keef & Twarock, 2009). This approach uses subsets of the vertex sets of quasilattices to formulate constraints on virus architecture (Keef *et al.*, 2013), and it has been shown in Indelicato *et al.* (2012) that these constraint sets can be derived, *via* projection, from a higher-dimensional lattice using the cut-and-project scheme (Senechal, 1996). These constraint sets have been classified and matched to viral capsids using an automated best-fit algorithm. A study of a wide range of viruses has revealed that icosahedral symmetry and the surface tilings developed by Caspar and Klug in quasi-equivalence theory are part of a much wider set of constraints on virus architecture that are characteristic of the structural features of a virus at different radial levels.

Motivated by the fact that the constraint sets used earlier are by construction subsets of the vertex sets of quasilattices, we present here a study of how quasilattices and the three-dimensional associated tilings can help us understand the geometric principles underpinning the structures of viral capsids and provide coarse-grained models that capture essential features of a viral architecture. We use the cut-and-project method, as detailed in §2, to construct icosahedral tilings. In addition to the constraint sets derived in Keef *et al.* (2013), the tilings contain edges and tile surfaces. In §3 we present a new algorithm that evaluates the goodness of fit of such tilings to a capsid, evaluating which tile sets best approximate the capsid. In §4 we show an application of the procedure to a range of viral capsids.

Our work should be viewed as a further step towards an extended exploration of the use of quasicrystallographic techniques to describe assemblies of proteins with icosahedral symmetry, following on from Janner's seminal work of using lattice theory for the modelling of virus architecture. A description of viral capsids in terms of icosahedral tilings has a number of potential applications that are worth exploring. Firstly, icosahedral tilings obtained by the cut-and-project method are encoded by a number of parameters that are orders of magnitude smaller than the size of the original PDB (Protein Data Bank) file; such parameters are, for instance, the scaling and the actual tiles in the fundamental domain with occupancy larger than a fixed threshold. Hence, such encoding of capsid geometry is much more tractable, for instance, in order to identify surface features, such as local symmetry axes, the T number, and so on, by automatic computer programs designed for that purpose. Moreover, these tiling approximations of viral capsids can be used as coarse-grained models to study virus assembly and structural transitions in viral capsids. Finally, icosahedral tilings adapted to viral capsids can also be used as guidelines to design building blocks for synthetic viral nanoparticles. Indeed, since the tiles are obtained by projection and hence by construction they fit together to form an icosahedrally symmetric particle, building blocks designed according to the shapes of tiles would therefore have the correct interfacial structures to result in self-assembled particles.

2. Three-dimensional icosahedral tilings *via* the cut-and-project method

In this section we quickly review the cut-and-project method to construct three-dimensional tilings with icosahedral symmetry, using projections from six-dimensional space to a three-dimensional subspace invariant under the icosahedral group (Kramer & Schlottmann, 1989). We will refer to the corresponding tilings as 'gauge' tilings, that will be rescaled to fit the viral capsid.

A tiling of three-dimensional space is a countable collection of closed polyhedra, called tiles,

$$\mathcal{T} = \{t_1, t_2, \ldots\}$$

such that

$$\mathbb{R}^3 = \cup_{i=1}^{\infty} t_i, \qquad \mathrm{int}(t_i) \cap \mathrm{int}(t_j) = \emptyset, \quad i \neq j,$$

where $\mathrm{int}(t_i)$ is the interior of t_i. We assume that the tiles have positive volume and are the closures of their interiors. The 2-, 1- and 0-dimensional faces of each tile will be called facets, edges and vertices, respectively.

In order to enforce icosahedral symmetry, we construct tilings by projection from six-dimensional space. In fact, the lowest dimension in which a lattice has icosahedral symmetry as part of its point group, and a three-dimensional subspace invariant under icosahedral symmetry, is 6 (Levitov & Rhyner, 1988), and there are precisely three lattice types (Bravais lattices) with this property in six dimensions: the simple cubic (s.c.), the body-centred cubic (b.c.c.) and the face-centred cubic (f.c.c.) lattices (Levitov & Rhyner, 1988).

The algorithm for the construction of the three gauge tilings *via* the cut-and-project method is based on the following steps (Katz, 1989):

(*a*) Computation of the six-dimensional Voronoi cell at the origin. For a given six-dimensional lattice $\mathcal{L}$ in $\mathbb{R}^6$ (either the s.c., f.c.c. or b.c.c. lattice), we compute the Voronoi cell around the origin O of the lattice, *i.e.*

$$V(O) = \{u \in \mathbb{R}^6;\ |O - u| \leq |y - u|, \forall y \in \mathcal{L} \setminus \{O\}\}. \qquad (1)$$

The Voronoi cell is a convex polytope, and its k-dimensional faces can be computed iteratively using the *QHull* package

(Barber *et al.*, 1996). Actually, the collection of the Voronoi cells $V(q)$ at all lattice points $q \in \mathcal{L}$ is a cell complex and a periodic tiling of $\mathbb{R}^6$. Its dual complex is also a periodic tiling of $\mathbb{R}^6$, called the Delone tiling, constructed as follows: if B is a p-face of a Voronoi cell, *i.e.* a p-cell of the Voronoi complex, its dual cell B^* has dimension $n-p$ and is the convex hull of the centres of the Voronoi cells that intersect in B (Senechal, 1996).

(*b*) Projection of the tiles in three dimensions. There is a unique decomposition of six-dimensional space into the direct sum of two three-dimensional subspaces corresponding to the two inequivalent three-dimensional irreducible representations of the icosahedral group. The projections onto these subspaces will be denoted by $\Pi_\perp$ and $\Pi_\parallel$, and the subspaces themselves will be referred to as the perpendicular and parallel space [see Katz (1989) for their definition].

Tiles in three dimensions are constructed as projections onto the parallel space of the duals $B^*(q)$ of suitable 3-faces $B(q)$ of the Voronoi cells $V(q)$ at lattice points q. More precisely, the three-dimensional tiles are those $\Pi_\parallel(B^*(q))$ for which $\Pi_\perp(B(q))$ contains the origin in perpendicular space. Since the Voronoi cell $V(q)$ at a lattice point q is given by $V(q) = V(O) + q$, it is enough to compute the tiles as $\Pi_\parallel(B^*(O)) + \Pi_\parallel(q)$ for all lattice points q such that $\Pi_\perp(-q) \in \Pi_\perp(B(O))$, where $B(O)$ is a 3-face of the Voronoi cell at the origin.

(*c*) Computation of the glue tiles. If the projection of $-q$ onto the perpendicular space lies in the intersection of the projections of the boundaries of several 3-faces of $V(0)$, *i.e.* if

$$\Pi_\perp(-q) \in \left[\bigcap_j \Pi_\perp(\partial B_j(O))\right],$$

then inconsistencies arise in step (*b*). In this case it is necessary to define glue tiles as the projection onto the parallel space of the dual to the intersections of those boundaries, *i.e.*

$$\Pi_\parallel\left[\left(\bigcap_j \partial B_j(O)\right)^*\right] + \Pi_\parallel(q).$$

The union of the tiles determined in step (*b*), together with the glue tiles constructed in step (*c*), define the gauge tilings that we use to model virus structure. In this work we will denote the three gauge tilings either by

$$\mathcal{T}_{\rm s.c.}, \mathcal{T}_{\rm b.c.c.}, \mathcal{T}_{\rm f.c.c.},$$

or, when the underlying six-dimensional lattice is fixed, simply by $\mathcal{T}$. Examples of the three tiling types obtained by projection are given in Fig. 1.

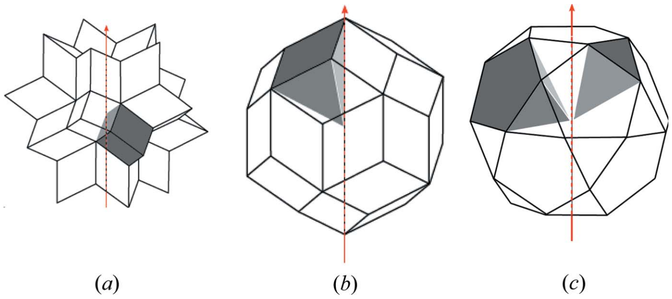


Figure 1
Vertex configurations at the origin for the (*a*) s.c., (*b*) b.c.c. and (*c*) f.c.c. tilings. For each case the distinct types of tiles are highlighted in grey.

3. The matching algorithm

We describe below the algorithm that matches the icosahedral tilings to the viral capsid. Experimental data of virus structure are derived from X-ray and cryo-electron microscopy experiments and are available as PDB files from the Viperdb website (Carrillo-Tripp *et al.*, 2009). Each PDB file contains the Cartesian coordinates of the atoms of the protein shell and, if available, of the genomic material present in the viral capsid.

3.1. Step 1, pre-processing of the PDB data: representation of the capsid surface by the surface mesh $\mathcal{M}$

Since our algorithm focuses only on the viral capsid, the information on the position of the genetic material of the virus must be removed from the PDB files prior to analysis. Then, after a centring that fits the data with the origin of the tilings, all atoms in the resulting PDB files are rotated such that their axes of symmetry align with those of the tilings, and the shape of the viral capsid is calculated as the solvent-excluded surface (SES) of the proteins using the *PyMol* software (The *PyMOL* Molecular Graphics System, Version 1.7.4, Schrödinger, LLC). The output is a mesh of points on the SES.

For simplicity, we assume that all atoms are represented by spheres centred at the atomic positions (given by the PDB file) with radius the average van der Waals radius computed over all species of atoms present in the protein shell. The atomic species (for example, carbon, nitrogen *etc.*) of each atom is available in the PDB file and the values of the van der Waals radii used to compute the average radius have been taken from Bondi (1964).

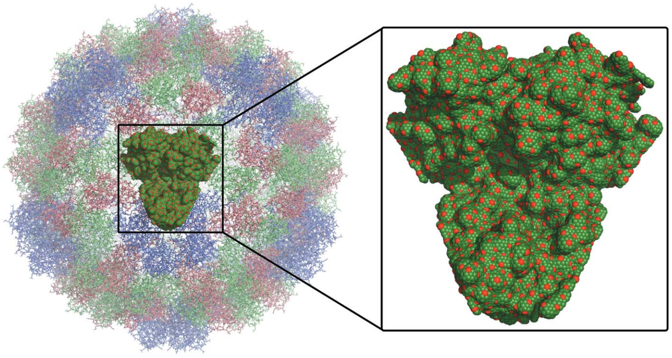


Figure 2
Illustration of the surface representation procedure described in Step 1, applied to the native form of Cowpea chlorotic mottle virus (PDB entry 1cwp). The portion of the SES surface output by *PyMol* that intersects the fundamental domain of the icosahedral group is represented in green; a close-up view representing the vertices of the original mesh (green spheres) as well as the reduced mesh $\mathcal{M}$ (red spheres) is shown in the box in magnification.

To reduce computation time the number of vertices representing the mesh is limited to one every 2 Å. We denote by $\mathcal{M}$ the resulting subset of vertices; the reduction enables faster computations while keeping the essential features of the capsid (see Fig. 2), and is also consistent with the fact that the finest resolution of experimental data is usually higher than 3 Å.

3.2. Step 2, scaling

Each of the gauge tilings constructed by the cut-and-project method is rescaled to fit the viral capsid by applying a scaling transformation, *i.e.* a linear transformation of $\mathbb{R}^3$ of the form $OP \mapsto sOP$ with O the origin in $\mathbb{R}^3$, $P \in \mathbb{R}^3$ and $s \in (0, +\infty)$. A rescaled tiling will be denoted by $\mathcal{T}(s)$.

3.3. Step 3, tile occupancy

In this section we fix a value of the scaling s and describe the rules which select the subset of tiles in $\mathcal{T}(s)$ that will be matched to the viral capsid. The process is iterated for different values of s in a given range (whose boundaries are explained in detail in §§3.4 and 3.5).

Intuitively, for a tile to be selected it must contain at least a minimum number of atoms of the viral capsid.

It is more convenient to work in terms of occupancy rather than in terms of the number of atoms within a tile, since tiles may have different sizes. For simplicity, we say that an atom is inside a tile if the centre of the atom is inside the tile, bearing in mind that in our model each atom is represented by a sphere of fixed radius. This can lead to cases where the volume occupied by atoms within a tile is larger than the volume of the tile (*i.e.* $\rho_i > 1$, see below), but these cases occur primarily for small values of s. Overall, this simplification remains good enough while reducing the computation time to reasonable levels.

Definition 3.1. We define the occupancy of a tile $t_i \in \mathcal{T}(s)$ as

$$\rho_i = \frac{N_i V}{\mathrm{vol}(t_i)} \tag{2}$$

where V is the volume of an atom, N_i is the number of atoms of the viral capsid whose centre lies in the tile t_i and $\mathrm{vol}(t_i)$ is the volume of t_i.

We fix a minimal threshold occupancy ρ by requiring that a tile is taken into consideration by the algorithm if its occupancy satisfies $\rho_i \geq \rho$.

Definition 3.2. We denote by $\mathcal{S}_{\mathcal{T}}(s, \rho)$ the subset of tiles in $\mathcal{T}(s)$ with occupancy larger than ρ, ordered in increasing order of tile occupancyρ_i, *i.e.*

$$\mathcal{S}_{\mathcal{T}}(s, \rho) = \{t_i \in \mathcal{T}(s) | \rho_i \geq \rho \text{ and } \rho_i \leq \rho_j \text{ for } i \leq j\}. \tag{3}$$

We denote by

$$\mathcal{A}_{\mathcal{T}}(s, \rho)$$

the boundary surface of $\mathcal{S}_{\mathcal{T}}(s, \rho)$, defined as the union of the facets of the tiles which are not shared by any two tiles in this set, and we denote by $|\mathcal{A}_{\mathcal{T}}(s, \rho)|$ the number of facets of $\mathcal{A}_{\mathcal{T}}(s, \rho)$. Further, we denote by

$$\mathcal{B}_{\mathcal{T}}(s, \rho)$$

the complex composed of the facets, edges and vertices of $\mathcal{A}_{\mathcal{T}}(s, \rho)$, and we write $|\mathcal{B}_{\mathcal{T}}(s, \rho)|$ for its cardinality.

Increasing the minimal threshold occupancy ρ results in a smaller set of tiles, *i.e.* $\mathcal{S}_{\mathcal{T}}(s, \rho') \subseteq \mathcal{S}_{\mathcal{T}}(s, \rho'')$ for $\rho' > \rho''$. On the other hand, the tiles in $\mathcal{S}_{\mathcal{T}}(s, \rho)$ can be the same for different values of ρ: the change in the set $\mathcal{S}_{\mathcal{T}}(s, \rho)$ occurs at discrete values ρ_α, $\alpha = 1, \ldots, N$, such that $\mathcal{S}_{\mathcal{T}}(s, \rho) = \mathcal{S}_{\mathcal{T}}(s, \rho_\alpha)$ for every $\rho \in (\rho_{\alpha-1}, \rho_\alpha]$. Indeed, ρ_α is the lowest value of ρ_i for tiles t_i in $\mathcal{S}_{\mathcal{T}}(s, \rho)$, where $\rho \in (\rho_{\alpha-1}, \rho_\alpha]$. Using this observation it is natural to define a non-increasing sequence of tilings using the occupancies of the tiles, *i.e.* by

$$\mathcal{S}_{\mathcal{T}}(s, \rho) \supseteq \cdots \supseteq \mathcal{S}_{\mathcal{T}}(s, \rho_i) \supseteq \mathcal{S}_{\mathcal{T}}(s, \rho_{i+1}) \supseteq \cdots \tag{4}$$

where ρ_i is the occupancy of tile $t_i \in \mathcal{T}(s)$.

Therefore, for each value of the scaling factor s, a finite non-increasing sequence of tilings is defined, obtained by increasing the minimal tile occupancy.

For example, suppose $\mathcal{S}_{\mathcal{T}}(s, 0.5)$ is composed of three tiles t_1, t_2 and t_3 with occupancy $\rho_1 < \rho_2 < \rho_3$. Then the sequence above becomes

$$\begin{aligned} \mathcal{S}_{\mathcal{T}}(s, 0.5) = \mathcal{S}_{\mathcal{T}}(s, \rho_1) = \{t_1, t_2, t_3\} \supset \mathcal{S}_{\mathcal{T}}(s, \rho_2) \\ = \{t_2, t_3\} \supset \mathcal{S}_{\mathcal{T}}(s, \rho_3) = \{t_3\}. \end{aligned}$$

3.4. Step 4, construction of the library of tilings

In order to construct a library of tilings that are good candidates to approximate the data, we restrict the algorithm to tilings $\mathcal{S}_{\mathcal{T}}(s, \rho)$ that fulfil the following properties:

(i) $\rho \geq 0.5$, *i.e.* each tile in $\mathcal{S}_{\mathcal{T}}(s, \rho)$ must have at least 50% occupancy.

(ii) The union of the tiles in $\mathcal{S}_{\mathcal{T}}(s, \rho)$ must contain at least 90% of the total main-chain atoms of the capsid (*i.e.* the chain of C, C_α and N atoms to which the side chains are attached) within experimental precision.

(iii) The scaling factor is restricted to a suitable interval $[s_{\min}, s_{\max}]$. To define the upper bound $s_{\max}$, notice that, after removing the genome, the capsid is a container whose interior is empty, and we require that *some* tiles within the capsid remain empty for any sampled value of s. For large scaling values, this is the case if at least the tiles in the vertex configuration at the origin O, *i.e.* the collection of tiles around the origin (see Fig. 1), do not intersect the protein shell, *i.e.* are contained in the empty region encompassed by the shell. Hence, we choose $s_{\max}$ as the maximal scaling such that no atoms of the capsid are within tiles with vertex O.

The lower bound $s_{\min}$ requires the introduction of the function σ_{sd} and will be defined later.

To construct the library, we restrict the algorithm to discrete values of the scaling factor:

$$s_k \in [s_{\min}, s_{\max}], \qquad k = 1, \ldots, n;$$

for each such value, as discussed above, there is an increasing sequence of tilings

$$\mathcal{S}_{\mathcal{T}}(s_k, \rho) \supseteq \cdots \supseteq \mathcal{S}_{\mathcal{T}}(s_k, \rho_i) \supseteq \mathcal{S}_{\mathcal{T}}(s_k, \rho_{i+1}) \supseteq \cdots \tag{5}$$

defined in terms of the tile occupancy. The set of tilings

$$\mathcal{S}_{\mathcal{T}}(s_k, \rho_i), \tag{6}$$

as k and i vary, is the library of tilings to which we shall restrict the algorithm in what follows.

3.5. Step 5, goodness of fit 1

The problem of finding the set of tiles that best approximates the surface of the protein shell can be formulated in a manner similar to polygonal approximation problems (Masood, 2008; Kolesnikov, 2012; Marji & Siy, 2003; Pikaz & Dinstein, 1995; Perez & Vidal, 1994; Ramer, 1972; Ray & Ray, 1993; Winzen & Niemann, 1994), in which polygonal curves are constructed to approximate a given one. In our case, the three-dimensional polyhedron to be approximated has vertices $\mathcal{M}$ on the SES of the viral capsid, and the approximating polyhedron is defined by the boundary surface $\mathcal{A}_{\mathcal{T}}(s, \rho)$ of $\mathcal{S}_{\mathcal{T}}(s, \rho)$. In this section we describe the scoring system that we use to quantify the approximation error.

We introduce some notation. For $v_i \in \mathcal{M}$, denote by $P_{\mathcal{A}}(v_i)$ the projection of v_i onto $\mathcal{A}_{\mathcal{T}}(s, \rho)$, *i.e.* the point of $\mathcal{A}_{\mathcal{T}}(s, \rho)$ having minimal distance from v_i (assuming this is unique), and by d_i the corresponding minimal distance between v_i and $\mathcal{A}_{\mathcal{T}}(s, \rho)$.

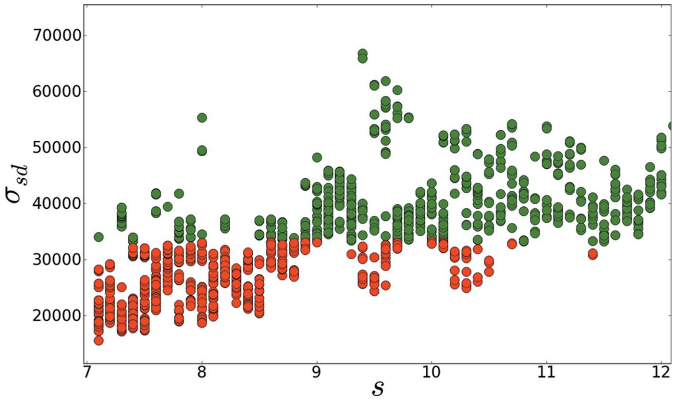


Figure 3
Determination of $s_{\min}$ for matching of the b.c.c. tiling to the viral capsid of the Minute mice virus (PDB entry 1mvm). The values of $\sigma_{sd}(s, \rho_i)$ are represented as a function of the scaling s for different values of ρ_i corresponding to the sequences of configurations $\mathcal{S}_{\mathcal{T}}(s, \rho_i)$ defined in equation (4). The values such that $\sigma_{sd}(s, \rho_i) \leq \Delta\sigma_{sd}$ are coloured in red. The minimal score $\min_{\rho_i} \sigma_{sd}(s, \rho_i)$ has local minima with respect to s for scalings s = 9.5, 10.3 and 11.4, which possibly correspond to good fits. For all $s \leq 9.0$, the sets of tiles satisfy $\min_{\rho_i} \sigma_{sd}(s, \rho_i) \leq \Delta\sigma_{sd}$. The measure of σ_{sd} is then limited by the uncertainty of the experimental measure, and therefore we choose $s_{\min} = 9.0$.

Definition 3.3. For a given tiling $\mathcal{S}_{\mathcal{T}}(s, \rho)$, the integrated square error $\sigma_{sd}(s, \rho)$ is defined by

$$\sigma_{sd}(s, \rho) = \sum_{\{i|v_i \in \mathcal{M}\}} d_i^2. \tag{7}$$

The quantity σ_{sd} is widely used in polygonal approximation problems (Masood, 2008; Kolesnikov, 2012; Marji & Siy, 2003; Perez & Vidal, 1994). This score penalizes each vertex in $\mathcal{M}$ with regard to its distance from $\mathcal{A}_{\mathcal{T}}(s, \rho)$. The bigger the value of σ_{sd}, the poorer the quality of the approximation.

In order to establish whether the score $\sigma_{sd}(s, \rho)$ is appropriate to assess the goodness of fit of our approximation, we have evaluated it for a sequence of tilings $\mathcal{S}_{\mathcal{T}}(s, \rho)$ for a test case in Fig. 3. Indeed, as the scaling s decreases, $\sigma_{sd}(s, \rho)$ tends to decrease with s. The local minima in Fig. 3 are due to the fact that the tiling $\mathcal{S}_{\mathcal{T}}(s, \rho)$ may be the same for a whole interval of the scaling factor s, in the sense that the number and type of tiles do not change up to rescaling. For each range in which $\mathcal{S}_{\mathcal{T}}(s, \rho)$ remains the same, there is a value of s for which the approximation, as measured by $\sigma_{sd}(s, \rho)$, is optimal: these correspond to the local minima in Fig. 3.

In our case, however, even for a simple example such as the one presented in Fig. 4, visual inspection of the tilings shows that σ_{sd} favours sets which do not match our intuitive notion of a good approximation of the surface of the shell by the tiles.

In order to better describe the fine features of the capsid surface, we decompose σ_{sd} into the sum of two scores σ_1 and σ_2 as follows. Denoting by $\mathcal{M}_j$ the set of vertices whose projection belongs to the jth element of $\mathcal{B}_{\mathcal{T}}(s, \rho)$ and by $D_j = \frac{1}{|\mathcal{M}_j|}\sum_{\{i|v_i \in \mathcal{M}_j\}} d_i$ the average distance of the jth element of $\mathcal{B}_{\mathcal{T}}(s, \rho)$ from the vertices in $\mathcal{M}_j$, then

$$\begin{aligned}
\sigma_{sd}(s, \rho) &= \sum_{j=1}^{|\mathcal{B}_{\mathcal{T}}(s,\rho)|} \sum_{\{i|v_i \in \mathcal{M}_j\}} d_i^2 = \sum_{j=1}^{|\mathcal{B}_{\mathcal{T}}(s,\rho)|} \sum_{\{i|v_i \in \mathcal{M}_j\}} (d_i - D_j + D_j)^2 \\
&= \sum_{j=1}^{|\mathcal{B}_{\mathcal{T}}(s,\rho)|} 2D_j \underbrace{\sum_{\{i|v_i \in \mathcal{M}_j\}} (d_i - D_j)}_{0} + \sum_{j=1}^{|\mathcal{B}_{\mathcal{T}}(s,\rho)|} \sum_{\{i|v_i \in \mathcal{M}_j\}} D_j^2 \\
&\quad + \sum_{j=1}^{|\mathcal{B}_{\mathcal{T}}(s,\rho)|} \sum_{\{i|v_i \in \mathcal{M}_j\}} (d_i - D_j)^2 \\
&= \sigma_1(s, \rho) + \sigma_2(s, \rho),
\end{aligned}$$

where

$$\sigma_1(s, \rho) = \sum_{j=1}^{|\mathcal{B}_{\mathcal{T}}(s,\rho)|} |\mathcal{M}_j| D_j^2, \quad \sigma_2(s, \rho) = \sum_{j=1}^{|\mathcal{B}_{\mathcal{T}}(s,\rho)|} \sum_{\{i|v_i \in \mathcal{M}_j\}} (d_i - D_j)^2. \tag{8}$$

The score $\sigma_1(s, \rho)$ penalizes sets for which the distance between the average position of the approximated points in $\mathcal{M}_j$ and the corresponding element in $\mathcal{B}_{\mathcal{T}}(s, \rho)$ is larger.

The score $\sigma_2(s, \rho)$ measures the difference between the distance of the vertices in $\mathcal{M}_j$ from the jth element in $\mathcal{B}_{\mathcal{T}}(s, \rho)$ and the average distance D_j. It can be viewed as a score that quantifies the alignment of the boundary of the tiling to the detailed features of the shape of the surface of the protein shell as modelled by $\mathcal{M}$. When all vertices in $\mathcal{M}_j$ are equidistant from the jth element in $\mathcal{B}_{\mathcal{T}}(s, \rho)$, the contribution of these vertices to σ_2 vanishes, see Fig. 5.

The reason why σ_{sd} is not a suitable score is due to the fact that in general σ_1 is larger than σ_2. In polygonal approximation studies, σ_{sd} is not governed by σ_1, as the vertices of the approximating polygon are chosen as a subset of the vertices of the approximated polygon. In our case, the approximating polyhedra are fixed by icosahedral symmetry *via* the cut-and-project method and a renormalization of the scores σ_1 and σ_2 is therefore required to make them comparable. It is important to recall that the scores σ_{sd}, σ_1 and σ_2 depend on the tiling $\mathcal{T}$ (either the b.c.c., f.c.c. or s.c. tiling). However, we will not make this dependence explicit in the notation unless strictly necessary, as in equation (11).

Lower bound for the scaling factor s. To devise a lower bound for s, we use the score σ_{sd} and the precision ε on the experimental data.

Definition 3.4. We define

$$\Delta\sigma_{sd} = \varepsilon^2 |\mathcal{M}|. \tag{9}$$

Notice that if $d_i < \varepsilon$ for every $v_i \in \mathcal{M}$, then $\sigma_{sd} < \Delta\sigma_{sd}$ and all points of $\mathcal{M}$ lie on the surface $\mathcal{A}_{\mathcal{T}}(s, \rho)$ within experimental precision, and decreasing σ_{sd} is meaningless. Hence, we define $s_{\min}$ as the largest scaling for which $\sigma_{sd}(t, \rho)$ is below the threshold $\Delta\sigma_{sd}$ for all $s \in [0, s_{\min}]$. Formally,

$$s_{\min} = \sup\{s : \min_{\rho \geq 0.5} \sigma_{sd}(t, \rho) \leq \Delta\sigma_{sd} \text{ for all } t < s\}.$$

An example is shown in Fig. 3.

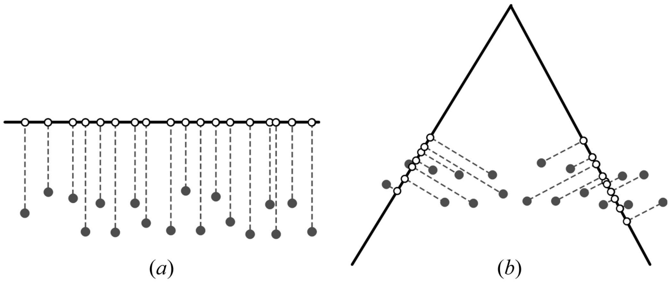


Figure 4
The vertices in $\mathcal{M}$ (grey dots) and their projections (empty circles) onto the facets of $\mathcal{A}(s, \rho)$ (black lines) are shown for two different tile selections in (*a*) and (*b*). The score σ_{sd} is higher in case (*a*) than in case (*b*), contradicting our intuitive notion of a good fit.

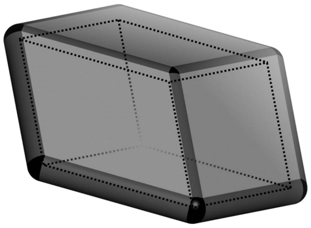


Figure 5
Here we assume that the approximating polyhedron is made of a single tile, whose edges are represented by dashed lines. In this case, the complex $\mathcal{B}_{\mathcal{T}}(s, \rho)$ is the union of the facets, the edges and the vertices of the polyhedron. The light-grey planar surfaces, the dark-grey cylinders and the black spheres are loci of points equidistant from the facets, the edges and the vertices of $\mathcal{B}_{\mathcal{T}}(s, \rho)$, respectively. The score $\sigma_2(s, \rho)$ is minimized if, for a given average distance from $\mathcal{B}_{\mathcal{T}}(s, \rho)$, the points of the set $\mathcal{M}$ to be approximated lie on the equidistant surfaces.

3.6. Step 6, goodness of fit II. Renormalized integrated square error

As discussed in the previous section, the contributions of the scores σ_1 and σ_2 to the total score σ_{sd} are not comparable, since typically σ_1 is larger than σ_2. Hence, it is better to define a new score

$$\sigma(s, \rho) := r\sigma_1(s, \rho) + \sigma_2(s, \rho), \tag{10}$$

in which σ_1 is weighted by a suitable renormalization factor r. In order to define it, let

$$\mathcal{R}(\gamma) := \sum_{\mathcal{T},k,i} (\gamma\sigma_{\mathcal{T},1}(s_k, \rho_i) - \sigma_{\mathcal{T},2}(s_k, \rho_i))^2 \tag{11}$$

where $\sigma_{\mathcal{T},1}$ and $\sigma_{\mathcal{T},2}$ are the scores σ_1 and σ_2 relative to the tiling $\mathcal{T}$, and the sequence (s_k, ρ_i) of all admissible values of the scale factors and occupancy is defined in equation (5). Notice that $\mathcal{R}(\gamma)$ contains contributions from the scores relative to all three tilings $\mathcal{T} = \mathcal{T}_{\text{s.c.}}, \mathcal{T}_{\text{f.c.c.}}, \mathcal{T}_{\text{b.c.c.}}$.

In order to minimize the difference between the contributions of $\gamma\sigma_1$ and σ_2 to σ we minimize $\mathcal{R}(\gamma)$ with respect to γ.

Definition 3.5. The renormalization factor r is the value of γ that minimizes $\mathcal{R}(\gamma)$, *i.e.*

$$r := \frac{\sum_{\mathcal{T},k,i} \sigma_{\mathcal{T},1}(s_k, \rho_i)\sigma_{\mathcal{T},2}(s_k, \rho_i)}{\sum_{\mathcal{T},k,i} \sigma^2_{\mathcal{T},1}(s_k, \rho_i)}. \tag{12}$$

3.7. Step 7, interpretation of the renormalized integrated square error

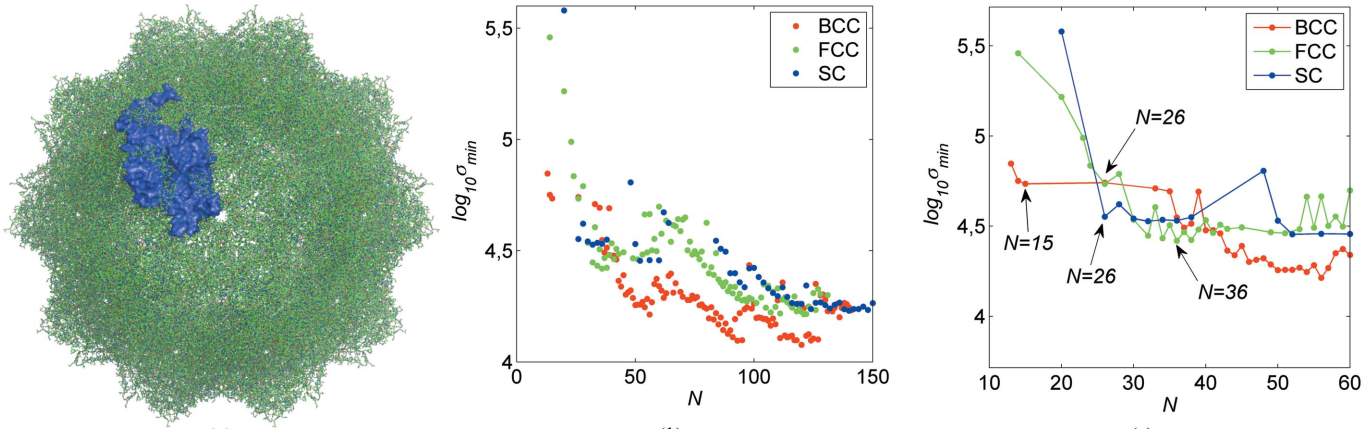
In two-dimensional polygonal approximation studies two approximating polygons with the same number of edges are compared. In the same spirit, in our three-dimensional case, we compare different approximations of $\mathcal{M}$ if they have the same number of facets. More precisely, we compute the score σ for every tiling and every discretized value of the scaling factor and occupancy and, among all the tilings that have the same number of facets, visualize only the minimum value of the score. Precisely, for each tiling

$$\mathcal{S}_{\mathcal{T}}(s_k, \rho_i),$$

we compute the set of boundary facets and the renormalized integrated square error

$$\mathcal{A}_{\mathcal{T}}(s_k, \rho_i), \quad \sigma(s_k, \rho_i).$$

Let $\mathcal{I}$ be the fundamental domain of the icosahedral group in three dimensions, which is a cone with vertex the origin in $\mathbb{R}^3$. Denoting by

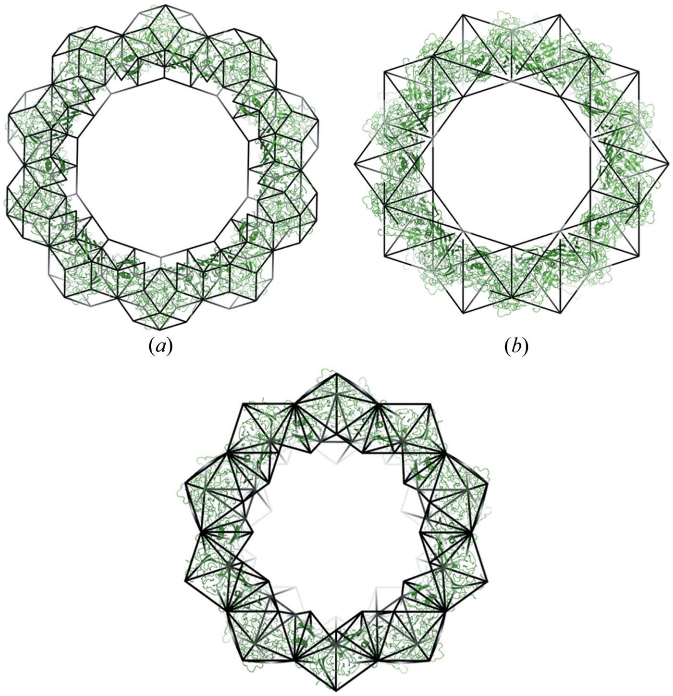


(*a*)

(*b*)

(*c*)

Figure 6
(*a*) The MVM capsid ($T = 1$) with a protein subunit represented in blue. (*b*) Plot of the score $\sigma_{\rm min}$ as a function of N for each of the three tilings with a renormalization factor $r_{\rm 1mvm} \sim 0.324$. (*c*) Magnification of the plot of the score $\sigma_{\rm min}$ showing some of the first minima ($N = 15, 26$), as well as $N = 36$, corresponding to the tiling shown in Figs. 7 and 8.



(*a*)

(*b*)

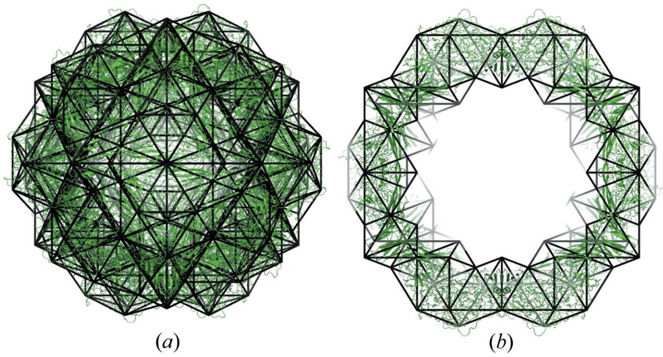
(*c*)

Figure 7
Comparison between some tilings fitting the Minute mice virus capsid: cross-sectional views along the fivefold axis. (*a*) s.c. tiling corresponding to $N = 26$; (*b*) b.c.c. tiling corresponding to $N = 15$; (*c*) f.c.c. tiling corresponding to $N = 36$. Visual inspection shows that the s.c. and b.c.c. tilings do not provide good approximations of the inner and outer surfaces of the MVM capsid, whereas the f.c.c. tiling provides a better approximation of the capsid layout.

$$|\mathcal{A}_{\mathcal{T}}(s_k, \rho_i) \cap \mathcal{I}|$$

the number of facets that intersect a fundamental domain, we define the minimum renormalized integrated square error corresponding to all tilings with the same number of facets as

$$\sigma_{\rm min}(N) = \min_{s_k, \rho_i}\{\sigma(s_k, \rho_i) : |\mathcal{A}_{\mathcal{T}}(s_k, \rho_i) \cap \mathcal{I}| = N\}.$$



(*a*)

(*b*)

Figure 8
f.c.c. tiling for $N = 36$ matching the Minute mice virus: (*a*) view along the twofold axis; (*b*) slice along the same axis.

Notice that the facets intersecting the boundary of the fundamental domain are counted at least twice in N. However, we assume that N is a reasonable proxy for the number of tiles. Typical plots of the minimum score $\sigma_{\rm min}(N)$ are shown in Figs. 6, 9, 11, 13, 16 and 18.

In these plots, we can distinguish between two regions:

(i) For low values of N (*i.e.* when the size of the tiles in $\mathcal{S}_{\mathcal{T}}$ is large) the values of $\sigma_{\rm min}$ undergo large oscillations, as the change of tiles corresponds to large changes in the shape of the approximating polyhedron. In some cases, tiles overlapping with the protein shell represent such a poor approximation that there is no tiling satisfying the restrictions on occupancy for some values of N.

(ii) For large values of N the size of the tiles compared to the capsid is small; as the number of facets increases, smaller features of the capsid, such as protrusions, can be approximated: when these features can be fitted, $\sigma_{\rm min}$ decreases to a local minimum or a plateau.

However, care must be taken when interpreting results for a large number of facets. In fact, given sufficiently many facets, the algorithm attempts to minimize the score by fitting smaller features, including those related to the tertiary structures of proteins. Small features of the capsid are highly mobile and it

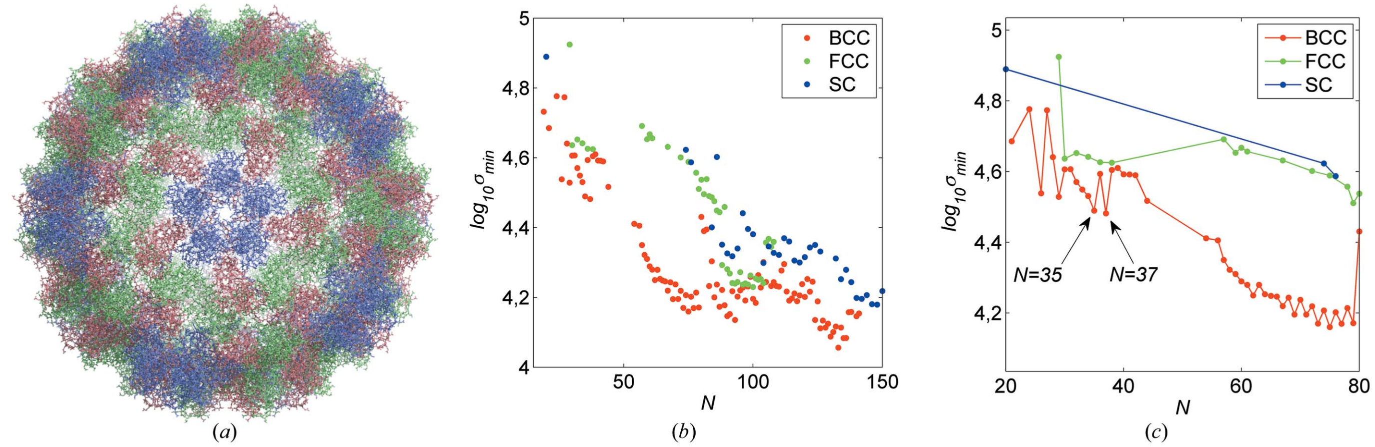


Figure 9
(*a*) The CCMV capsid with the '*A*', '*B*' and '*C*' chains coloured in blue, red and green, respectively. (*b*) Plot of the score $\sigma_{\rm min}$ as a function of *N* for each of the three tilings with a renormalization factor $r_{\rm 1cwp} \sim 0.392$. (*c*) Magnification of the plot in (*b*) showing two local minima at $N = 35$ and $N = 37$, the latter corresponding to the tiling shown in Fig. 10.

is not clear whether symmetry constraints apply here and, further, the reliability of the PDB data themselves is reduced. Therefore, also the reliability of our approximation is reduced for large *N*.

Since the renormalization factor *r* is independent of the type of tiling we show in each figure the scores corresponding to the three different tilings on the same plot for comparison (see for instance Figs. 6, 9, 11, 13, 16 and 18).

4. Application to viruses

In this section we apply the matching algorithm to some $T = 1$ and $T = 3$ viral capsids. The colour code is the same in all figures: the scores corresponding to s.c., b.c.c. and f.c.c. tilings are plotted as blue, red and green dots, respectively.

In the figures displaying the tilings, the edges of the tiles are represented as black lines and, for $T = 3$ viruses, the *A*, *B* and *C* chains have different colours.

When discussing the scores relative to each tiling, we keep in mind the issues discussed in the previous section:

(*a*) Global minimization of the score $\sigma_{\rm min}$ is not a viable criterion for the selection of the best approximation, since, apart from local fluctuations (see below), the score tends to decrease as the number of facets increases.

(*b*) Acceptable approximations should locally minimize the normalized integrated square error, in the sense that they should realize the minimum error among all those tilings with a comparable number of facets.

(*c*) Alternatively, approximations could be chosen as those that realize a plateau of the score.

(*d*) Acceptable approximations should be tractable and simple enough, in the sense that the number of facets should not be too large. As discussed above, tilings with too small tiles do not necessarily provide better approximations of the more delicate capsid features.

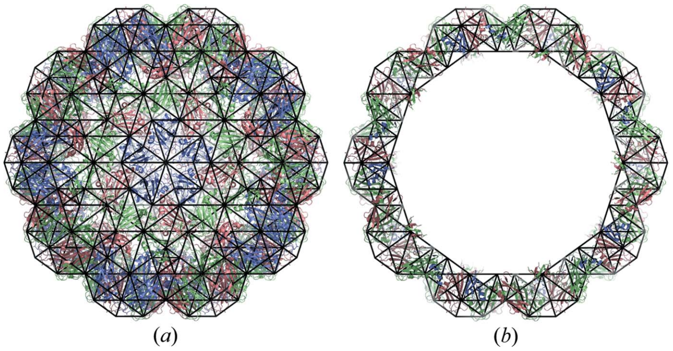


Figure 10
b.c.c. tiling for $N = 37$ matching the Cowpea chlorotic mottle virus capsid: (*a*) front view along a fivefold axis and (*b*) slice along the same axis.

4.1. *Minute mice virus* (MVM)

This $T = 1$ virus replicates in cells which are undergoing division and is responsible for the modifications of the response of biological systems where cell multiplication is important, such as in cancer research studies. The PDB file (PDB entry 1mvm, Llamas-Saiz *et al.*, 1997; available from Carrillo-Tripp *et al.*, 2009) includes information on the RNA of the virus with chain identifiers '*R*' and '*S*', that is removed prior to analysis (Fig. 6*a*).

Figs. 6(*b*), 6(*c*) show that a first set of local minima (as the number of facets increases) of the score $\sigma_{\rm min}(N)$ for the three tilings is reached at $N = 15$ for the b.c.c. tiling, and at $N = 26$ for the s.c. and f.c.c. tilings. In the range $30 \leq N \leq 39$, the f.c.c. tilings have a lower score than the other two. Indeed, inspection of Fig. 7 confirms that, as expected, the f.c.c. tilings provide a better approximation to the viral shell in this range. As an example, the f.c.c. tiling corresponding to $N = 36$ is depicted in Fig. 8. Notice that the contours of the inner and outer surfaces of the viral capsid, including the β-barrel involved in host-cell recognition (Llamas-Saiz *et al.*, 1997), are well approximated by this tiling.

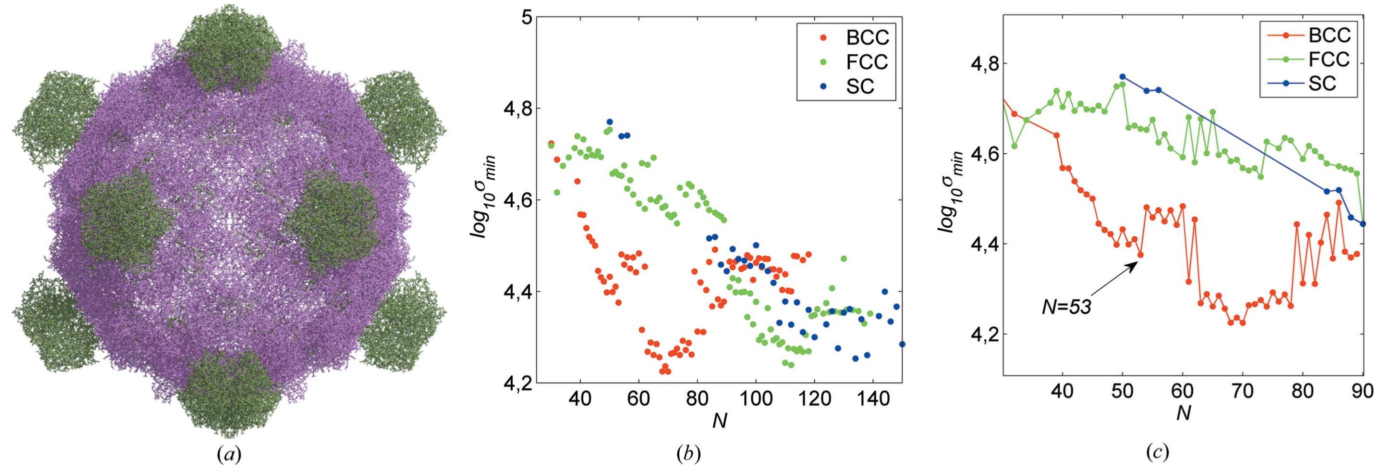


Figure 11
(*a*) The $T = 1$ bacteriophage $\alpha3$ shown along a twofold axis. (*b*) Plot of the score $\sigma_{\rm min}$ as a function of N for each of the three tilings with a renormalization factor $r_{\rm 1m06} \sim 0.360$. (*c*) Magnification of the plot in (*b*) showing the local minimum at $N = 53$ corresponding to the tiling shown in Fig. 12.

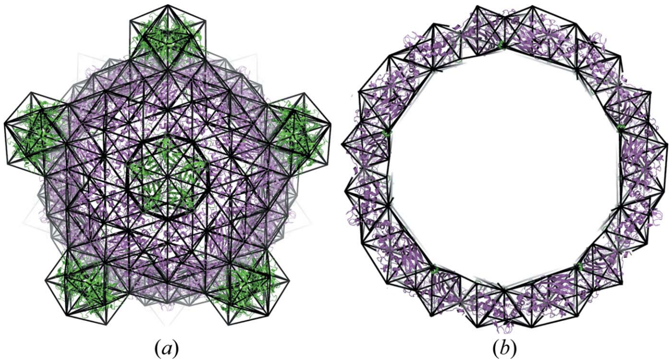


Figure 12
b.c.c. tiling for $N = 53$ matching the bacteriophage $\alpha3$ capsid: (*a*) outside view along a fivefold axis, and (*b*) a slice along the same axis.

4.2. *Cowpea chlorotic mottle virus* (CCMV)

CCMV is a $T = 3$ plant virus (PDB entry 1cwp, Speir *et al.*, 1995) which, thanks to reversible pH-dependent structural changes, can be used to package polymers (Douglas & Young, 1998) (Fig. 9*a*). Figs. 9(*b*), 9(*c*) show that the b.c.c. tiling yields lower scores than the f.c.c. and s.c. tilings when matched to the CCMV capsid. The bulk structure of the CCMV capsid is well approximated by fits corresponding to $N = 35$ and $N = 37$, which correspond to two local minima in the plot of the score. These two configurations differ very slightly in both s and ρ and they differ by just one single tile. The configuration for $N = 37$ is represented in Fig. 10. The thickness of the capsid shell and the structure of hexamers and pentamers are well approximated by this b.c.c. tiling, and hence no finer configuration is considered.

4.3. Bacteriophage $\alpha3$

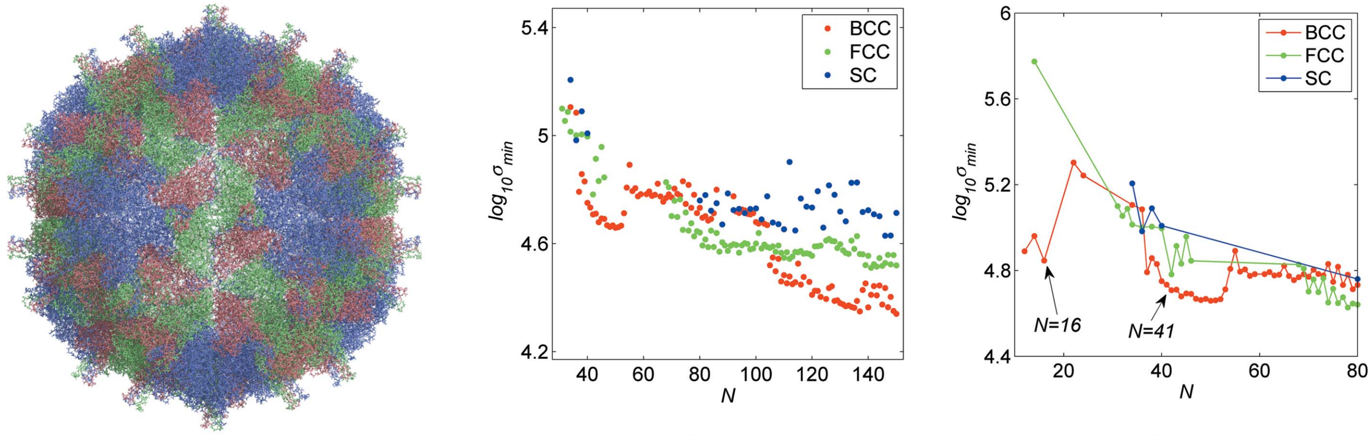
We now focus on the $T = 1$ bacteriophage $\alpha3$ capsid (PDB entry 1m06, Bernal *et al.*, 2003). The structure is mostly spherical, apart from the G-proteins positioned at the fivefold axes as shown in Fig. 11(*a*). In the intermediate range $35 \leq N \leq 90$, the b.c.c. tiling yields the lowest score (Figs. 11*b*, 11*c*). The configuration corresponding to the local minimum $N = 53$ is shown in Fig. 12.

4.4. *Pariacoto virus* (PaV)

PaV is a $T = 3$ insect virus; the 180 coat proteins cluster into 60 trimers with notable protrusions along the quasi-threefold axis (Tang *et al.*, 2001). These protrusions, see Fig. 13(*a*), are believed to be involved in host-cell recognition and are therefore important features which need to be described by our approximation. The best-match algorithm was run on the PDB file (PDB entry 1f8v, Tang *et al.*, 2001) from Carrillo-Tripp *et al.* (2009) after the removal of the '*R*' chain.

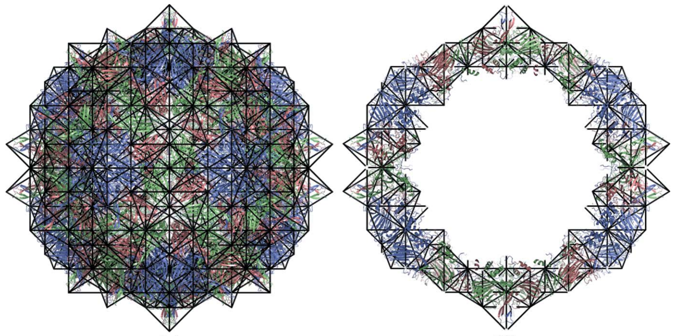
First note that the b.c.c. tiling provides the best match among all tilings in the range $37 < N < 70$ (Figs. 13*b*, 13*c*). For higher values of N, the algorithm lowers $\sigma_{\rm min}$ by fitting features of the tertiary structure of the proteins. The first local minimum of the b.c.c. tiling is located at $N = 16$, but the corresponding approximation by the tiles is too coarse, *i.e.* the outer and inner capsid surfaces are not well represented by the tile set. To match the protrusions, whose size is small when compared to the size of the virus (with an outer radius ~ 175 Å), a finer fit is needed. The score $\sigma_{\rm min}$ reaches a first plateau in the interval $N \in [40, 52]$, and matches corresponding to $N \in [40, 52]$, with the exception of $N = 40$ or 44 (which do not fit the protrusions), are visually very similar. We choose to display the tiling corresponding to $N = 41$ in Fig. 14 since it is the simplest representation in this class of b.c.c. matches. For this capsid we consider no finer fits as much higher values of N are required for f.c.c. or b.c.c. tilings to reach lower $\sigma_{\rm min}$ values.

We can compare our tile-matching approach with the one based on point arrays, developed in Keef *et al.* (2013), as applied to the PaV capsid. For clarity, we denote by $\mathcal{P}$ the point array obtained in Keef *et al.* (2013) that best matches PaV and by $\mathcal{P}_\mathcal{S}$ the set of vertices of $\mathcal{S}_{\rm b.c.c.}$ corresponding to $N = 41$, obtained by fitting the b.c.c. tiling to PaV. In Keef *et al.* (2013), a library of 569 point arrays has been created using



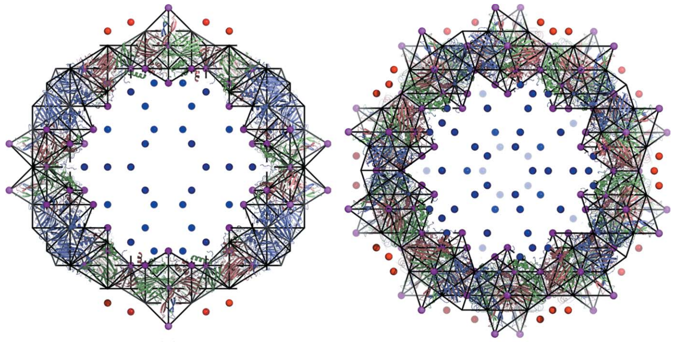
(*a*) (*b*) (*c*)

Figure 13
(*a*) The PaV capsid with the same colour coding as in Fig. 9(*a*). (*b*) Plot of the score $\sigma_{\rm min}$ as a function of N for each of the three tilings with a renormalization factor $r_{\rm 1f8v} \sim 0.369$. (*c*) Magnification of the plot in (*b*) showing the local minimum at $N = 16$ and the value $N = 41$ corresponding to the tiling shown in Fig. 14.



(*a*) (*b*)

Figure 14
b.c.c. tiling for $N = 41$ matching the Pariacoto virus: (*a*) outside view along a twofold axis, and (*b*) a slice along the same axis.



(*a*) (*b*)

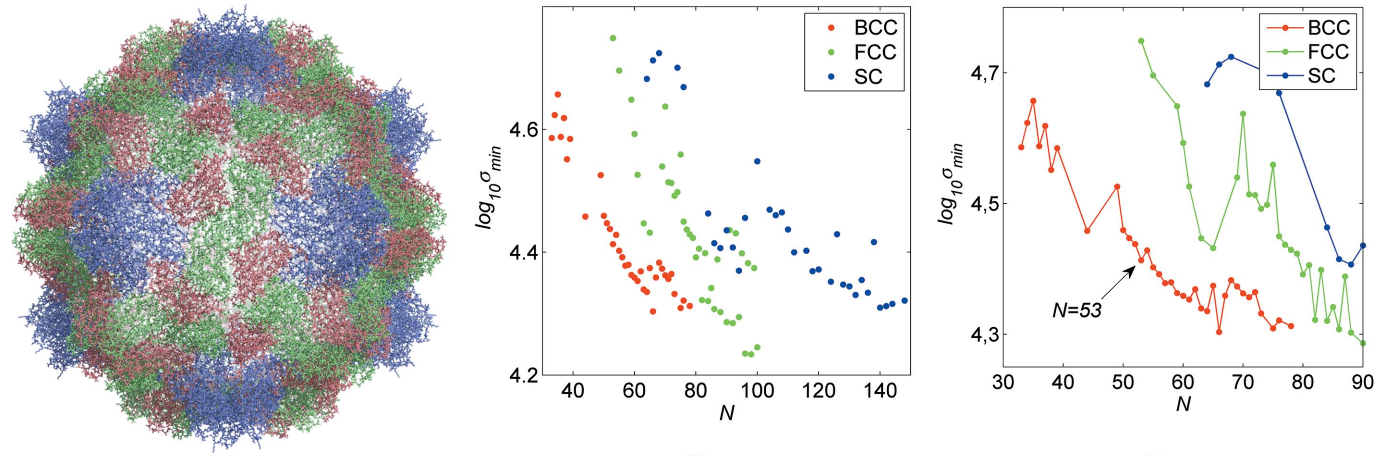
Figure 15
Comparison of the point array $\mathcal{P}$ derived in Keef *et al.* (2013) with the set of vertices $\mathcal{P}_{\mathcal{S}}$ of one of our tilings using two cross-sectional views: (*a*) slice along a twofold axis, and (*b*) a slice along a fivefold axis. For clarity the vertices of $\mathcal{P}$ are displayed with a 4 Å radius and are rescaled to match the vertices of the tile set $\mathcal{P}_{\mathcal{S}}$. Coloured in red are the vertices contributing to a poorer match. All vertices in $\mathcal{P}$ matching the capsid features (in magenta) are vertices of the tile set selected by our algorithm, while vertices with a higher approximation error (in red) and those (in blue) providing geometrical constraints on the genomic material (not shown), are not included in $\mathcal{P}_{\mathcal{S}}$.

affine extensions of the icosahedral group. After rescaling these vertex sets such that the outermost vertices match the protrusions of PaV, an RMSD-based score measuring the distance between vertices and the surface capsid has been used to rank these point arrays according to their proximity to atomic positions in the viral capsid [see Keef *et al.* (2013) for more details]. Notice that $\mathcal{P}$ and the set of points $\mathcal{P}_{\mathcal{S}}$ are projections of b.c.c. lattice points: hence, once scaled to the PaV capsid, $\mathcal{P}$ can be embedded into the gauge b.c.c. tiling (defined in §2 as the b.c.c. tiling with scaling factor $s = 1$) rescaled by a factor $s = 15.51$, whereas the scaling factor obtained by fitting $\mathcal{S}_{\rm b.c.c.}$ is $s = 16.05$. This small difference in the scalings corresponding to $\mathcal{P}$ and $\mathcal{P}_{\mathcal{S}}$ can easily be associated with the difference in the matching algorithms, *i.e.* minimizing the capsid-to-vertices distance (in the point array approach) as opposed to minimizing the capsid-to-tile subset distance (in the current study). As can be shown in Fig. 15 the two approaches correspond to similar approximations of the PaV capsid.

4.5. *Physalis mottle virus* (PhyMV)

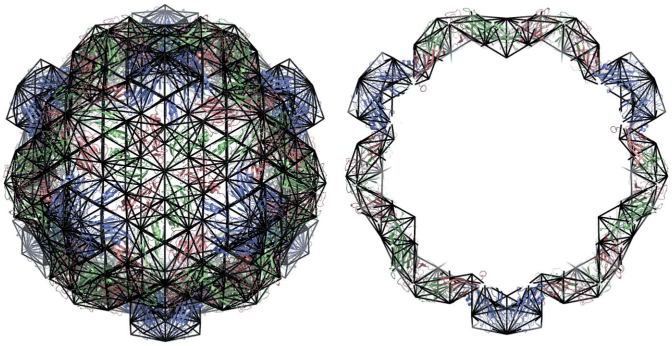
PhyMV is a $T = 3$ virus infecting plants. The PDB file (PDB entry 1e57, Krishna *et al.*, 2001) used in the tile-matching algorithm was created from X-ray crystallography with empty capsids (Fig. 16*a*). The score $\sigma_{\rm min}(N)$ is plotted in Figs. 16(*b*), 16(*c*).

For $N < 80$ the b.c.c. tilings give better fits with respect to the s.c. and f.c.c. tilings, but for $N < 50$ the matching of the hexamers' outer surface is poor. The tilings for $N \in [50, 66]$ share a common subset of tiles fitting the bulk of the hexamers, and the main differences between these tilings correspond to two different fittings of the pentamers' outer surfaces. As N increases and the approximation is finer, one of the two configurations occurs repeatedly. We choose to display the simplest representation (b.c.c. tiling corresponding to $N = 53$) of this class of approximations in Fig. 17.



(*a*) (*b*) (*c*)

Figure 16
(*a*) The PhyMV capsid seen from the outside along a twofold axis with the same colour coding as in Fig. 13(*a*). (*b*) Plot of the score $\sigma_{\rm min}$ as a function of N for each of the three tilings with a renormalization factor $r_{\rm 1e57} \sim 0.415$. (*c*) Magnification of the plot in (*b*) showing the value $N = 53$ corresponding to the tiling shown in Fig. 17.



(*a*) (*b*)

Figure 17
b.c.c. tiling corresponding to $N = 53$ for PhyMV: (*a*) outside view along a threefold axis, and (*b*) a slice along the same axis.

4.6. *Carnation mottle virus* (CarV)

CarV is a $T = 3$ virus responsible for mild mottling and chlorosis (*i.e.* whitening of green plant tissues due to deficiency of chlorophyll) in carnation crops (PDB entry 1opo, Morgunova *et al.*, 1994). Its capsid displays protrusions near the twofold axes, see Fig. 18(*a*). The results of the tile-matching algorithm are shown in Figs. 18(*b*), 18(*c*). As in the case of PaV, the b.c.c. tiling yields better fits than the s.c. or f.c.c. tilings for equal or lower values of N. Also, the first minimum at $N = 59$ provides only a poor description of the inner capsid surface and of the protrusion contour. We therefore look for a finer approximation to better match the capsid contour. A steep decrease in $\sigma_{\rm min}$ for $N \in [73, 80]$ suggests that smaller features of the viral capsid are matched. After this, $\sigma_{\rm min}(N)$ decreases almost linearly over $N \in [80, 114]$. Since $N = 80$ corresponds to the configuration for which the decrease of $\sigma_{\rm min}$ is the most significant, we choose this as the simplest representation of the essential features of the viral capsid (see Fig. 19).

5. Discussion

In this work we have explored the possibility of approximating viral capsids by icosahedral tilings, motivated by Janner's work on the approximation of virus structure by lattices. To do so, we have further developed ideas from polygonal approximation theory to define a scoring function that helps to rank the tilings according to how good they fit to the capsid surface. The question of the goodness of fit, of course, cannot have a conclusive answer, but our investigation will hopefully provide the basis for further research in this field.

Here we critically review our results and the open problems that they pose. First of all, we have constructed three gauge tilings (filling all space), by the well known cut-and-project scheme, that guarantees that the tilings have icosahedral symmetry. By rescaling each gauge tiling, we have obtained three families of tilings parametrized by the scaling parameter s. From each of these rescaled tilings we have selected finite subsets of tiles, by identifying those with occupancy greater than a given threshold ρ, *i.e.* by selecting those tiles such that a fraction at least ρ of their volume is occupied by atoms of the capsid. Indeed, for each fixed value of the scaling parameter, this allows one to define a sequence of such finite tilings, obtained by increasing the threshold occupancy.

Note that tilings with greater occupancy do not necessarily yield a better fit of the capsid. Tiles are discrete and their shapes are fixed, and the capsid may have protrusions or delicate external features such as protruding loops, and it may well be that, by omitting tiles with small occupancy, we miss some of these surface features. Hence, additional selection criteria are needed. We have chosen to explore criteria based on the goodness of fit of the surface of the tiling to the surface features of the capsid. In order to do so, we have defined a score σ as the weighted sum of two contributions, one measuring the average distance between the capsid surface and the surface of the tiling, and the second measuring how

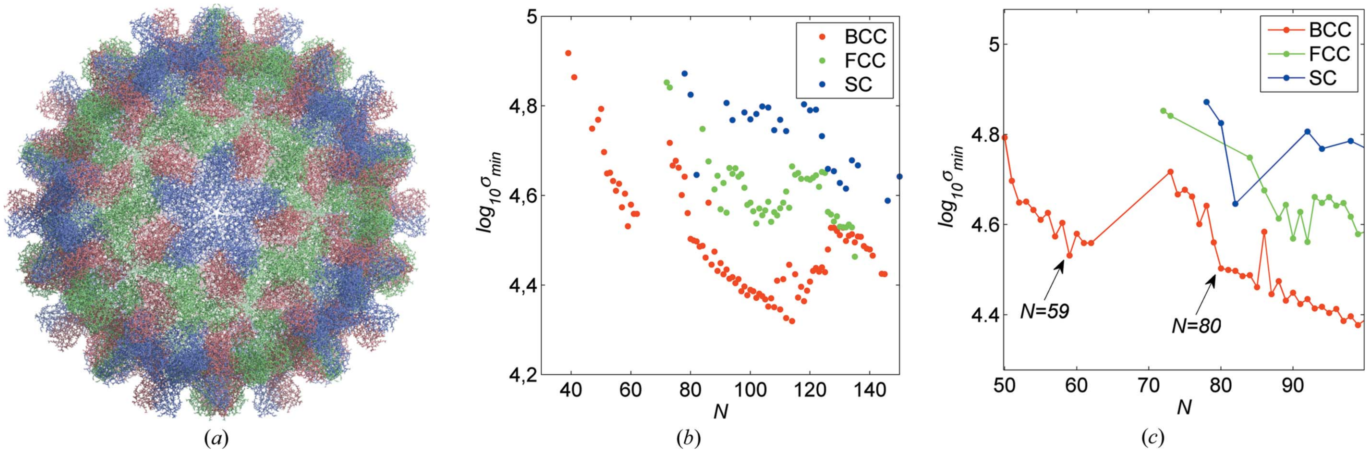


Figure 18
(*a*) The CarV capsid with the same colour coding as in Fig. 13(*a*). (*b*) Plot of the score $\sigma_{\rm min}$ as a function of N for each of the three tilings with a renormalization factor $r_{1\rm opo} \sim 0.362$. (*c*) Magnification of the plot in (*b*) showing the local minimum at $N = 59$ and the value $N = 80$ corresponding to the tiling shown in Fig. 19.

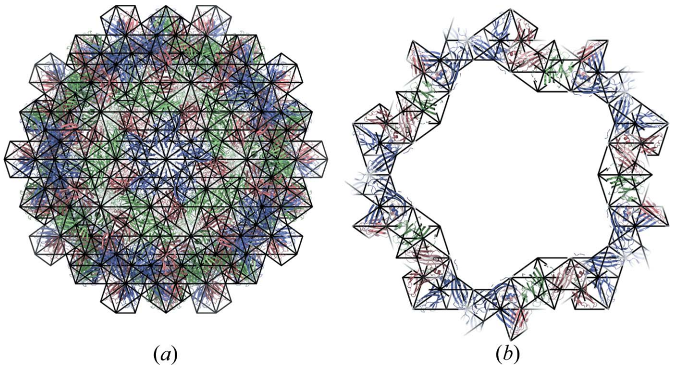


Figure 19
b.c.c. tiling for $N = 80$ matching the Carnation virus: (*a*) outside view along a fivefold axis, and (*b*) a slice along the same axis.

well each surface element of the tiling, *i.e.* its faces, edges and vertices, describe the fine features of the capsid surface [*cf.* equation (8) and Figs. 4 and 5].

As a basic parameter labelling each finite tiling, we have chosen a measure of its complexity, given by the number N of facets of the tiling surface that intersect the fundamental domain of the icosahedral group. Notice that this is a better measure than the scaling s, which may not be comparable among different viruses due to their different sizes. For each fixed number of facets N, we have selected the tiling with the lower score, and have plotted the minimum score as a function of N. Plots of the minimum score *versus* N are useful tools to understand goodness of fit and the viability of selection criteria for the best tilings. Inspection of our case studies, in fact, shows that there is no such unique criterion.

Consider our first two examples, MVM and CCMV (Figs. 6 and 9). These capsids do not have important protrusions and their surface is quite smooth. Hence, we expect that tilings with a relatively small number of facets give acceptable approximations both in terms of goodness of fit and low complexity. Hence, a viable selection criterion could be to choose tilings corresponding to the first local minima of the score as N increases. Inspection of Figs. 8 and 10 shows that this is indeed the case: the tilings with 36 (MVM) and 37 (CCMV) faces per fundamental domain give reasonable approximations of the capsid surfaces. The bacteriophage $\alpha 3$ capsid lies somehow at the boundary of the set of viruses for which coarse approximations are sufficient. It has large protrusions on the surface (Fig. 11), but no finer relevant features. Hence, the criterion of choosing the first local minimum, corresponding to $N = 53$, is satisfactory (Fig. 12). In a different class are the Pariacoto, Physalis mottle and Carnation mottle viruses (Figs. 13, 16 and 18). The capsids of these viruses have small, but tall protrusions that are important for their functionality and must be taken into account in the approximation. In this case, therefore, we expect that more complex tiles are needed to describe them, and the best selection rule turns out to be a choice of tiling which corresponds to a plateau of the score plot. Selecting the smaller complexity (*i.e.* N) tiling in these classes guarantees that a further increase of complexity does not change the score, and hence the goodness of fit, significantly. Figs. 14, 17 and 19 show the best fits to PAV, PhyMV and CarV, corresponding to $N = 41, 53$ and 80, respectively, which represent acceptable approximations of the capsids with relatively low complexity.

In conclusion, our work shows that approximations of viral capsids by icosahedral tilings are feasible, and provides the necessary basis for further work in this field. The approximations provide an alternative to Janner's blueprints for virus architecture in terms of lattice theory. They form a basis for the construction of coarse-grained models of viral capsids that can be used for the analysis of virus assembly and of the structural transitions in the surface lattices of viral capsids that in many viruses are important for infectivity.

Acknowledgements

R. Twarock and G. Indelicato thank the Leverhulme Trust for financial support *via* a Research Leadership Award (F/00224/AE). P. Cermelli and G. Indelicato acknowledge the Italian PRIN 2009 project 'Mathematics and Mechanics of Biological

Systems and Soft Tissues'. R. Twarock moreover acknowledges a Royal Society Leverhulme Trust Senior Research Fellowship (LT130088).

References

Barber, C., Dobkin, D. & Huhdanpaa, H. (1996). *ACM Trans. Math. Software*, **22**, 469–483.

Bernal, R., Hafenstein, S., Olson, N., Bowman, V., Chipman, P., Baker, T., Fane, B. & Rossmann, M. (2003). *J. Mol. Biol.* **325**, 11–24.

Bondi, A. (1964). *J. Phys. Chem.* **68**, 441–451.

Carrillo-Tripp, M., Shepherd, C. M., Borelli, I. A., Venkataraman, S., Lander, G., Natarajan, P., Johnson, J. E., Brooks, C. L. I. & Reddy, V. S. (2009). *Nucleic Acids Res.* **37**, D436–D442.

Caspar, D. L. D. & Klug, A. (1962). *Cold Spring Harbor Symposia on Quantitative Biology*, **27**, 1–24.

Douglas, T. & Young, M. (1998). *Nature (London)*, **393**, 152–155.

Everts, M., Saini, V., Leddon, J. L., Kok, R. J., Stoff-Khalili, M., Preuss, M. A., Millican, C. L., Perkins, G., Brown, J. M., Bagaria, H., Nikles, D. E., Johnson, D. T., Zharov, V. P. & Curiel, D. T. (2006). *Nano Lett.* **6**, 587–591.

Indelicato, G., Cermelli, P., Salthouse, D. G., Racca, S., Zanzotto, G. & Twarock, R. (2012). *J. Math. Biol.* **64**, 745–773.

Janner, A. (2006). *Acta Cryst.* A**62**, 270–286.

Janner, A. (2010*a*). *Acta Cryst.* A**66**, 301–311.

Janner, A. (2010*b*). *Acta Cryst.* A**66**, 312–326.

Katz, A. (1989). *An Introduction to the Mathematics of Quasicrystals*, pp. 147–183. San Diego and London: Academic Press.

Keef, T. & Twarock, R. (2009). *J. Math. Biol.* **59**, 287–313.

Keef, T., Twarock, R. & ElSawy, K. M. (2008). *J. Theor. Biol.* **253**, 808–816.

Keef, T., Wardman, J., Ranson, N., Stockley, P. G. & Twarock, R. (2013). *Acta Cryst.* A**69**, 140–150.

Kolesnikov, A. (2012). *Pattern Recognit. Lett.* **33**, 1329–1337.

Kramer, P. & Schlottmann, M. (1989). *J. Phys. A Math. Gen.* **22**, L1097–L1102.

Krishna, S. S., Sastri, M., Savithri, H. S. & Murthy, M. R. N. (2001). *J. Mol. Biol.* **307**, 1035–1047.

Levitov, L. & Rhyner, J. (1988). *J. Phys. Fr.* **49**, 1835–1849.

Llamas-Saiz, A., Agbandje-McKenna, M., Wikoff, W., Bratton, J., Tattersall, P. & Rossmann, M. (1997). *Acta Cryst.* D**53**, 93–102.

Ma, Y., Nolte, R. J. & Cornelissen, J. J. (2012). *Adv. Drug Deliv. Rev.* **64**, 811–825.

Marji, M. & Siy, P. (2003). *J. Pattern Recognit. Soc.* **36**, 2239–2251.

Masood, A. (2008). *J. Pattern Recognit. Soc.*, **41**, 227–239.

Morgunova, E., Dauter, Z., Fry, E., Stuart, D., Stel'mashchuk, V., Mikhailov, A. M., Wilson, K. S. & Vainshtein, B. K. (1994). *FEBS Lett.* **338**, 267–271.

Perez, J. & Vidal, E. (1994). *Pattern Recognit. Lett.* **15**, 743–750.

Pikaz, A. & Dinstein, I. (1995). *J. Pattern Recognit. Soc.* **28**, 373–379.

Ramer, U. (1972). *Comput. Graphics Image Processing*, **1**, 244–256.

Ray, B. K. & Ray, K. S. (1993). *J. Pattern Recognit. Soc.* **26**, 505–509.

Senechal, M. (1996). *Quasicrystals and Geometry*. Cambridge University Press.

Speir, J. A., Munshi, S., Wang, G., Baker, T. S. & Johnson, J. E. (1995). *Structure*, **3**, 63–77.

Tang, L., Johnson, K., Ball, L., Lin, T., Yeager, M. & Johnson, J. E. (2001). *Nat. Struct. Biol.* **8**, 77–83.

Winzen, A. & Niemann, H. (1994). *IEEE Int. Conf.* **1**, 228–232.