PROJECT DESCRIPTION

INTRODUCTION

Viruses are the most abundant and diverse biological entities on the planet (Cobian-Güemes et al, 2016). They are found in all environments with life, and, besides causing devastating diseases in animals and plants, they play key roles in the ecology and evolution of cellular organisms (Flint et al, 2008; Weinbauer, 2004; Weitz, 2016; Forterre, 2006). Nevertheless, the understanding of viral evolution and function is hindered by the high mutation rates and genetic diversity of viruses (Krupovic et al, 2019). Alternatively, the proteins that form the shell (capsid) that protects the viral genome adopt a limited number of three-dimensional configurations (folds), providing a framework to study viral evolution (Bamford et al, 2005; Krupovic and Koonin, 2017). However, it is not known how capsid folds determine capsid architecture. Elucidating this relationship would shed light on how viruses have explored capsid structures through their evolutionary history. It would also enable the prediction of viral capsid architectures from genetic information, facilitating the study of viral evolution as well as the characterization and engineering of viral particles for nanotechnological and biomedical applications (Mateo et al, 2013).

About 60% of known viral taxa and 90% of viral particles observed in the environment adopt icosahedral symmetry (Krupovic and Koonin, 2017; Ackermann, 2007). These quasi-spherical capsids range in size from 20 to 800 nm and are built from 60 to thousands of copies of the coat proteins. The origin of icosahedral symmetry in capsids is due to their thermodynamic advantage over other architectures as well as the optimization of the viral genetic economy (Zandi et al, 2004; Luque et al, 2010). Icosahedral symmetry minimizes the cost of coding the capsid surface and maximizes the volume available to store the viral genome with respect the amount of proteins produced to form the capsid shell (Crick and Watson, 1956; Caspar and Klug, 1962). Viruses have devised different strategies to form icosahedral capsids. As reported in our most recent work in *Nature Communications*, the geometrical landscape of icosahedral capsid architectures can be generated using at least eight layouts unified under the umbrella of the local molecular equivalence principle (Twarock and Luque, 2019) (Figure 1).

Genetic and structural studies indicate that viral icosahedral capsids have emerged at least 11 times during evolution (Krupovic and Koonin, 2017). In each case, the major capsid protein adopted a different three dimensional-structure (or fold), ranging from alpha helix-rich to beta strand-rich (Krupovic and Koonin, 2017). These capsid protein folds have led to the definition of viral structural lineages, which group viruses that, for the most part, do not have DNA or RNA sequence similarity or infect the same organism. Nonetheless, these viruses share the capsid protein fold as well as replication and assembly strategies that indicate a common origin (Abrescia et al, 2012). The conservation of capsid protein folds, thus, has been fundamental in viral evolution, but the factors restraining these folds remain unknown (Cheng and Brooks, 2013). Here we propose that the limited number of regular and semi-regular geometric layouts available to build icosahedral capsids has constrained viral capsid protein folds. If confirmed, this would provide a quantitative framework to study viral evolution and would offer a structural target to develop generic antiviral strategies as well as a guide for capsid engineering.

BACKGROUND

The classification of viral structures in the last 60 years has been based on the quasi-equivalence principle framework of capsid proteins introduced by Caspar and Klug (Caspar and Klug, 1962; Carrillo-Tripp et al, 2008; Mannige and Brooks, 2010). This framework generates structures analogous to Goldberg polyhedra and their duals, geodesic polyhedra (Schein, 2014; Coxeter, 1973). These capsid layouts are constructed from a hexagonal grid (lattice) by replacing 12 symmetrically distributed hexagons by pentagons (Caspar and Klug, 1962) (Figure 1, left column). The introduction of these 12 pentagonal disclinations is required by Euler's Theorem to generate a closed polyhedral shape (Conway et al, 2008).

In this construction, each hexagon has associated six proteins (hexamer) and each pentagon five (pentamer); the organization of the capsid proteins exhibits the 5-fold, 3-fold, and 2-fold rotational symmetry axes of an icosahedron, and all proteins occupy the same local molecular environment. In the dual configuration, the proteins are organized in triangles made of three proteins (trimers). The structures are indexed by the triangulation number or *T*-number $T(h,k) = h^2 + hk + k^2$, where *h* and *k* are the coordinates in the hexagonal grid that take only integer values (non-negative by convention). This quadratic Diophantine equation generates an infinite series of *T*-numbers that correspond to the Löshchian numbers (Marshall, 1975; Arlinghaus and Arlinghaus, 1989). The *T*-number indicates the number of quasi-equivalent positions or proteins required in the asymmetric unit to recover the capsid structure when applying icosahedral symmetry. This construction only permits capsid layouts with $60T$ coat proteins organised into 12 pentamers and $10(T-1)$ hexamers, and the first capsids of the series contain 60 ($T = 1$), 180($T = 3$), 240 ($T = 4$), and 420 ($T = 7$) coat proteins.

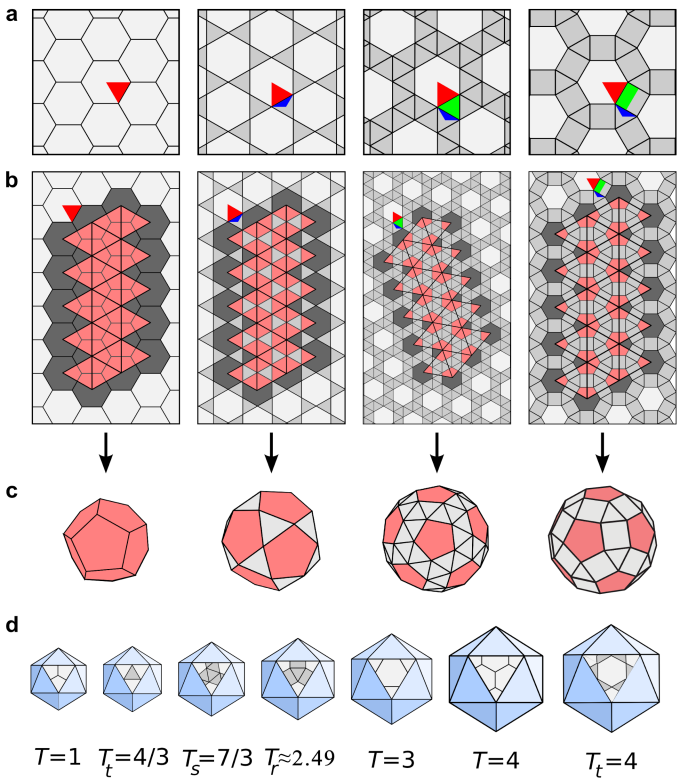


Figure 1. Twarock and Luque capsid icosahedral framework. **a**, The four Archimedean lattices permitting the Caspar-Klug construction: hexagonal (*6,6,6*), trihexagonal (*3,6,3,6*), snub hexagonal ($3^4,6$), and rhombitrihexagonal (*3,4,6,4*). The parenthesis indicate the regular polygons meeting at a vertex. The colors highlight the repeat unit of each lattice. **b,** Construction of Archimedean solids via replacement of 12 hexagons by pentagons in analogy to the CK construction. **c**, First elements of each icosahedral series. The process is analogous for the duals of the Archimedean lattices (Laves lattices**)**, which lead to triangular, rhomb, florets, and kite shaped tiles. **d**, Series combining icosahedral capsids from the different lattices. The structures are ordered based on capsid size (same hexagonal tile size).

Viral capsids reconstructed using high-resolution methods, such as X-ray crystallography and cryo-electron microscopy, have been assigned icosahedral *T*-numbers by comparing the number of coat proteins with the protein stoichiometry formula, $60T$, predicted in the CK framework (Baker et al, MMBR, 1999; Carrillo-Tripp et al, 2008). Over the years, however, an increasing number of viral icosahedral capsids have displayed unexpected protein stoichiometries, like 120 proteins (Duquerroy et al, 2009; Nemecek et al, 2013), 360 proteins organized in 72 pentamers (Salunke et al, 1989), or 180 proteins with a quasi-equivalence lower than the expected $T = 3$ (Kuhn et al, 2002; Sevvana et al, 2018). Other outliers display unexpected stoichiometries because they include minor capsid proteins, which are not contemplated in the CK framework (Pietilä et al, 2013; Grose et al, 2014; Yu et al, 2017; Dai et al, 2018; Bayfield et al, 2019). Additionally, icosahedral capsids that apparently conform to the same *T*-number structure, such as the abundant architecture $T = 3$ (180 proteins), display a variety of protein layouts other than the classical hexagonal and triangular patterns proposed in the CK framework (Johnson and Chandrasekar, 1998; Tang et al, 2001). Some of these outliers were explained case-by-case using Viral

Tiling Theory (Twarock, 2004; Elsawy et al 2008). But, until now, no geometrical framework predicted the existence of all these variety of icosahedral capsid architectures.

Such unifying framework for icosahedral capsids has been just published by the PI in *Nature Communications* (Twarock and Luque, 2019). This study shows that the CK quasi-equivalence approach is a particular case of the principle of regular vertex-transitive lattices (Archimedean lattices) and their face-transitive duals (Laves lattices). Among the 11 existing Archimedean lattices (Kepler, 1619; Conway et al, 2008), four lattices—hexagonal, trihexagonal, snub hexagonal, and rhombitrihexagonal—contain a hexagonal sublattice and, thus, are amenable to generate series of icosahedral capsids applying an analogous method to the CK theory (Figure 1). The same procedure can be applied to the associated Laves lattices, which generate uniform lattices with different tile shapes, respectively, triangles, rhombs, florets, and kites. The series associated with the hexagonal lattice and its dual recover the CK icosahedral capsids associated to Goldbergh and geodesic polyhedra, while the other six series generate icosahedral structures that had not been previously classified, except for the first element of each series, which is associated to either an Archimedean or a Catalan solid (Cromwell 1997; Conway et al, 2008). When compared with respect to the original hexagonal lattice, the six new icosahedral series present a rescaling of the capsid surface (that is, capsid size), which is captured in the generalized *T*-number $T_j(h,k) = \alpha_j(h^2 + hk + k^2) = \alpha_j T(h,k)$. Here $j = h,t,s,r$ indicates the lattice type used in the construction, respectively, hexagonal, trihexagonal, snub hexagonal, and rhombitrihexagonal. The factors are, respectively, $\alpha_h = 1$, $\alpha_t = 4/3 \approx 1.33$, $\alpha_s = 7/3 \approx 2.33$, and $\alpha_r = 4/3 + 2\sqrt{3} \approx 2.49$. These factors are directly related to the presence of minor tiles in the Archimedean based icosahedral capsids and the different size of the tiles in the Laves based icosahedral capsids.

These eight icosahedral layouts have provided an explanation under the same framework for the classical icosahedral capsids as well as the outliers discussed above (Twarock and Luque, 2019). The presence of hexagons and other polygons in the underlying Archimedean lattices rationalizes the existence of minor capsid proteins in viral capsids; the dual Laves lattices explain why icosahedral capsids can adopt different patterns with the same stoichiometry; and the clustering of different number of proteins respecting the local symmetry of a tile justifies the existence of non-quasi equivalent icosahedral capsids.

The icosahedral lattices described above are all built from individual or a combination of tiles with different shapes. It is thus reasonable to expect that the shape of the capsid protein (or, in particular, the capsid protein fold) might be related to each icosahedral lattice. In our recent publication, we confirmed this for a small number of viruses (Twarock and Luque, 2019). Viruses from the HK97 viral lineage, for example, display a trihexagonal lattice, while the virus from the beta-sheet unique lineage adopt a trihexagonal dual lattice (dimers forming rhombs). Alternatively, viruses from the PRD1 viral lineage adopt the classical hexagonal lattice in the CK framework (Benson et al, 1999). The lattice also seems sensitive to variations of a similar fold. Viruses with a capsid protein made of a single copy of the single jelly roll fold adopted the hexagonal dual lattice (forming triangular timers) while a virus with a capsid protein made of three similar repeats of the single jelly roll adopted a rhombihexagonal dual lattice (made of kites) (Twarock and Luque, 2019). This suggests that there could be a strong relationship between the capsid protein fold of viral lineages and the icosahedral lattice of the capsid.

The capsid protein fold has been used as the main proxy to define viral lineages, but these lineages have been subsequently refined based on additionally shared structural and genetic properties that suggest common ancestry (Bamford et al, 2005; Abrescia et al, 2012; Krupovic and Koonin, 2017). One element that is prevalent in well consolidated lineages is the packing and replication strategy of the genome. In particular, viruses in the same lineage follow the same route to synthesize mRNA from the viral genome and start the translation of viral proteins. This is known as the Baltimore classification: dsDNA (Class I), ssDNA (Class II), dsRNA (Class III), (+)ssRNA (Class IV), (–)ssRNA (Class V), ssRNA-RT (Class VI), and dsDNA-RT (Class VII) (Baltimore, 1971; 8 et al, 2014). Here, ds means double stranded, ss single stranded, (+) positive-sense, (–) negative sense, and RT reverse transcriptase. There are examples of

viruses that adopt icosahedral capsids in all the classes (Krupovic and Koonin, 2017). In fact, each class can contain several viral lineages that explore different specific genome packing and delivery steps. Class I (dsDNA genomes), for example, contains three lineages. Viruses in the HK97-lineage pre-assemble an empty capsid (procapsid) that is subsequently packed with dsDNA at high-densities, reaching a near quasi-crystalline state (Earnshaw and Casjens, 1980; Panja and Molineux, 2010). Viruses in the PRD1-lineage also pre-assemble an empty capsid but instead pack the dsDNA in combination with proteins that facilitate its condensation (Martín-González *et al.*, 2019). Finally, viruses in the single jelly roll vertical lineage co-assemble capsids made of 72 pentamers (360 capsid proteins) around a mini-chromosome made of dsDNA and histone proteins, like chromatin in the nucleus of eukaryotic cells (Salunke et al, 1989; Conway and Meyers, 2009). Most viral lineages, however, have just been recently identified bioinformatically, and it has not been confirmed the unique relationship between the capsid protein fold and the viral replication strategy (Koonin and Krupovic, 2017).

HYPOTHESIS

Here we hypothesize that each viral replication strategy is related to a specific capsid protein fold and icosahedral lattice (Figure 2). If confirmed, this would provide a unifying framework for genomic replication, molecular structure, and viral geometry, which will facilitate the classification, evolutionary studies, and structural prediction of viruses.

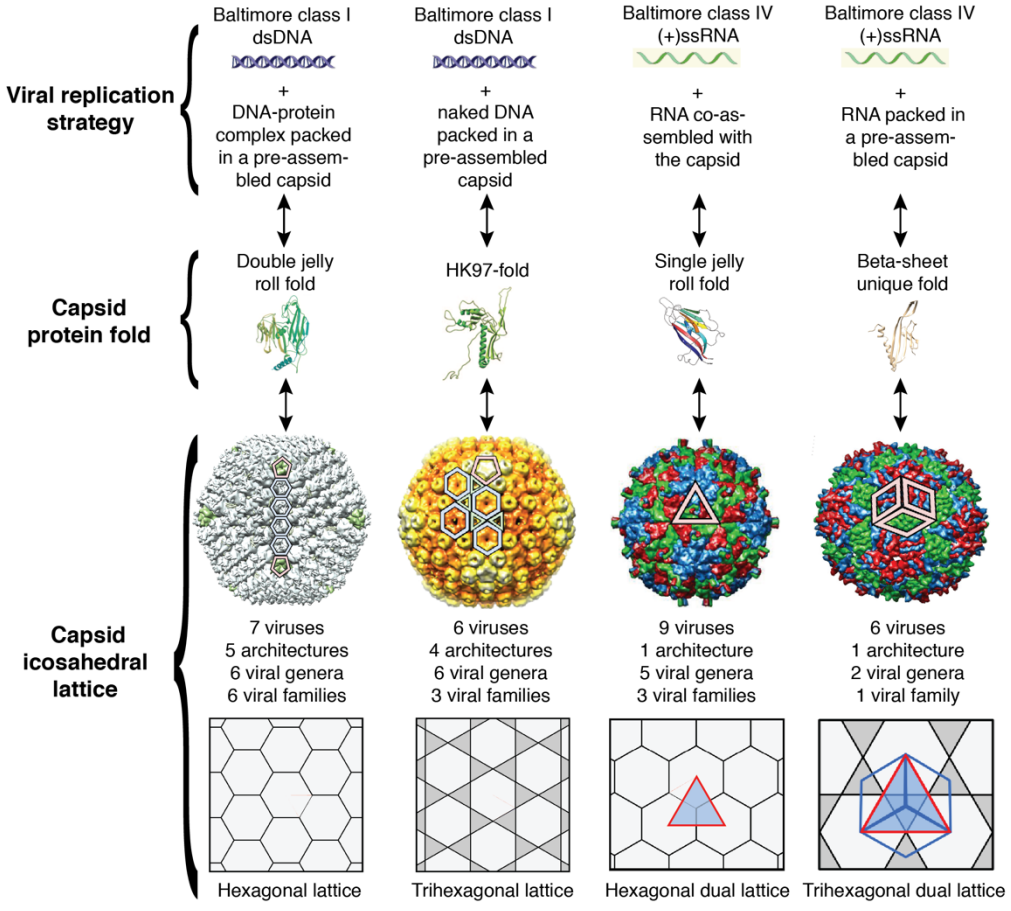


Figure 2. Preliminary data supporting our central hypothesis: one-to-one relation between viral replication strategy, capsid protein fold, and capsid icosahedral lattice. The preliminary data corresponds to four viral replication strategies that contained structural data for multiple viruses. The replication strategy includes the Baltimore class and genome packing mechanism. The protein fold and capsid examples were rendered with Chimera. The number of viruses, architectures, viral genera, and viral families as well as the associated capsid lattice is listed for each viral replication strategy.

PRELIMINARY DATA

We have assessed the validity of the central hypothesis for four icosahedral lattices that were identified multiple times with no ambiguity from 28 viruses (13 DNA and 15 RNA genomes) (Figure 2).

(i) The hexagonal lattice was observed among seven viruses that displayed 5 different architectures (*T*-numbers) and belonged to 6 viral genera and 6 viral families (Figure 5, left column). Viruses in this group infected bacteria, archaea, and vertebrates, including humans. The capsid protein displayed the double jelly roll fold characteristic of the PRD1-lineage. The hexagons of the lattice were formed by three capsid proteins each containing two similar copies of a jelly roll fold, that is, six protein folds per hexagon. These viruses belong to the Baltimore class I (dsDNA) and pre-assemble an empty capsid and pack the dsDNA in combination with proteins that facilitate its condensation (Martín-González *et al.*, 2019).

(ii) The trihexagonal lattice was observed among six viruses that displayed four different architectures and belonged to 6 viral genera and 3 viral families (Figure 5, mid left column). Viruses in this group infected bacteria, archaea, and vertebrates, including humans. The major capsid protein of all these viruses adopted the HK97-fold characteristic of the HK97-lineage. The hexagons were formed by the major capsid protein and each interstitial triangle was formed by three minor capsid proteins or three reinforcement proteins. These viruses belong to the Baltimore class I (dsDNA) and pre-assemble an empty capsid (procapsid) that is subsequently packed with dsDNA at high-densities, reaching a quasi-crystalline state and using the accumulated pressure to initiate the ejection of the genome in the host (Earnshaw and Casjens, 1980; Panja and Molineux, 2010).

(iii) The hexagonal dual lattice was observed among nine viruses that displayed architecture $T_h^D = 3$ and belonged to five viral genera and three viral families (Figure 5, mid right column). Viruses in this group infected hosts including vertebrates, invertebrates, and plants. The capsid protein of all these viruses adopted the single jelly roll fold and assemble in trimers to form the characteristic triangular tiles of the hexagonal dual lattice. These viruses belong to the Baltimore class IV and co-assemble their capsid around (+)ssRNA. The genomes contain either a single or two segments, encode subgenomic RNA, reach a similar density in the capsid, and display an analogous genetic architecture (with the open reading frame coding for the RNA-dependent RNA polymerase gene preceding the open reading frame of the capsid protein gene).

(iv) The trihexagonal dual lattice was observed among six viruses that displayed the same architecture ($T_t^D = 3$) and belong to the same viral genus and family (Figure 5, right column). Viruses in this group infected bacteria from two different classes in the same phylum. The capsid proteins of all these viruses adopted the beta-sheet unique fold. The capsid proteins formed dimers in the shape of the rhomb tiles of the lattice. These viruses belong to the Baltimore class IV and pre-assemble the capsid and subsequently pack (+)ssRNA. Their genetic architecture is opposite to case (iii). In this case the capsid protein gene precedes the RNA-dependent RNA polymerase gene. Additionally, these viruses use a maturation portal to deliver the genome into the host.

This preliminary analysis is consistent with our working hypothesis. Each viral replication strategy (genome packing and release) has associated a unique capsid protein fold and icosahedral lattice. The Baltimore class is consistent in each group, but is not sufficient to discriminate viral lineages. Additionally, as mentioned in the Background, we identified other icosahedral lattices with individual viral candidates, which displayed capsid protein folds and viral replication strategies consistent with our hypothesis (Twarock and Luque, 2019).

So far, the classification of viral capsids has been done by protein stoichiometry and visual inspection. This limits the analysis of a large number of capsids and can be ambiguous and arbitrary in many cases, as illustrated in Figure 8 for the procapsid of phage P22. This structure is made out of 420 capsid proteins (Figure 3a). Originally it was identified as classic $T(2,1) = 7$ icosahedral capsid ($60T$) adopting the hexagonal lattice (Figure 3b). In the new framework, however, it can be also characterized as a snub hexagonal $T_s(2,1) = \alpha_s T(1,1) = 7$ architecture, which contains 420 proteins and has the same surface area (Figure 3c). This layout seems to capture better the local 3-fold symmetry displayed by the

interactions with the underlying scaffolding proteins. Yet based just on the capsid proteins it is hard to determine which lattice would be more appropriate. Additionally, if one focuses on the different domains of the capsid protein, the organization of domains resembles a trihexagonal lattice with an architecture $T_t(2,1) = \alpha_t T(2,1) = 28/3$ (Figure 3d). This would imply a stoichiometry of 420 major domains in the pentamers and hexamers and 420 minor domains associated to trimers, analogous to the layout observed in other capsids in the viral lineage of this virus, HK97-fold (Twarock and Luque, 2019). One of the main tasks proposed in this research is to develop a computational quantitative method to assign capsid lattice.

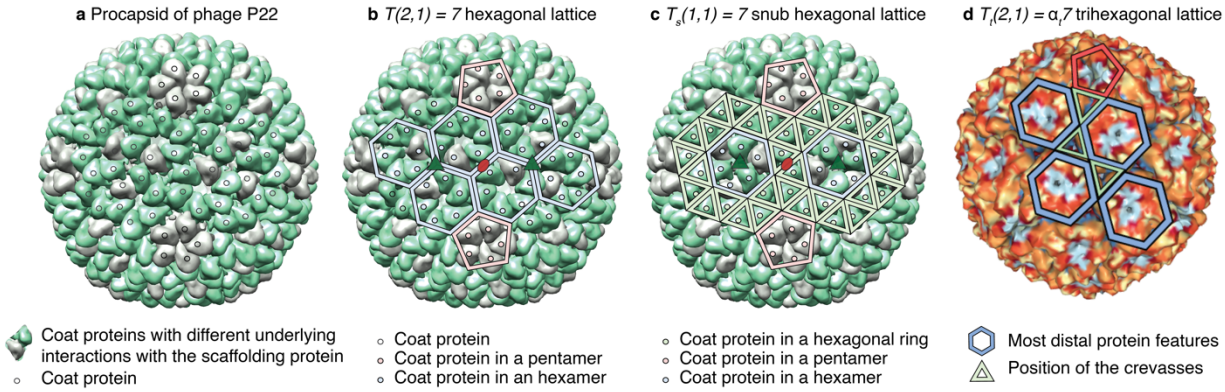


Figure 3. Alternative lattices for the phage P22 procapsid. All rendered structures are based on the Protein Data Bank structure 2XYY (Chen *et al.*, 2011).

PROPOSED RESEARCH

The research plan aims to develop rigorous methods to assess the one-to-one relationship between replication strategy, protein fold, and capsid lattice (Specific Aim 1), predict capsid architectures from genetic information (Specific Aim 2), and extend the principles used in the icosahedral lattices to generate new elongated and conical capsid geometries (Specific Aim 3). The main elements of each aim are summarized in Figure 4.

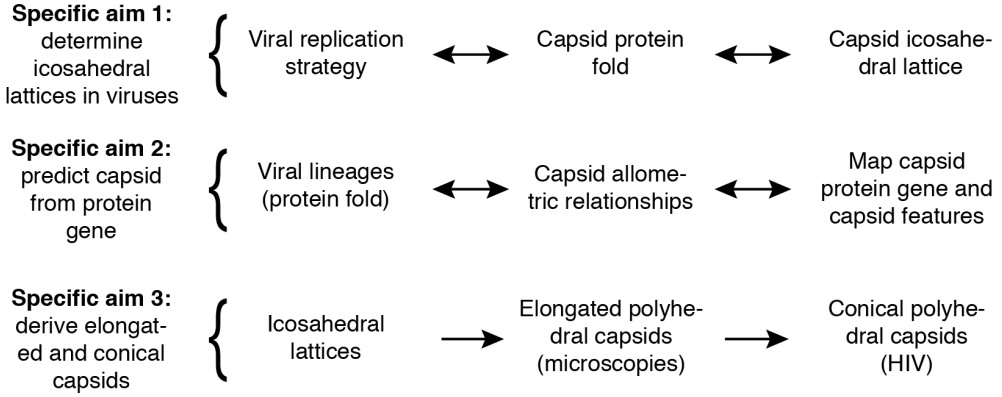


Figure 4. Proposed research. Each row focuses on one specific aim.

Specific Aim 1 (SA1): Determine the icosahedral lattices among viral lineages.
This aim will test if the relation between viral replication strategy, capsid protein fold, and icosahedral capsid lattice is one-to-one. This will be addressed in three sub aims. In SA1-1, we will develop a relational database including genomic, molecular, and icosahedral capsid data. In SA1-2, we will develop a computational method using Voronoi tessellations and protein-protein contact clusters to determine the icosahedral lattice of viral capsids quantitatively. In SA1-3, the computational method will be applied to the capsid structures in the relational database to test our central hypothesis.

SA1-1 Database of viral icosahedral capsids and viral lineages

The icosahedral virus database VIPERdb (Carrillo-Tripp et al, 2008) has currently deposited 482 capsid structures (August 21, 2019). These structures are classified based on the classical Caspar and Klug framework. Each entry has basic structural information as well as an associated high-resolution molecular model with an assigned Protein Data Bank identifier (PDB ID). The structures and information associated with each entry will be mined to generate an internal database. The viral replication strategy will be extracted from ViralZone, the Swiss Institute of Bioinformatics web-resource that integrates viral taxonomy, molecular genetics, assembly, and epidemiology (Hulo et al, 2010). The capsid protein fold associated with each virus will be extracted from the recent bioinformatic analysis that investigated the capsid proteins of icosahedral viruses (Krupovic and Kooning, 2017). The generated database will include the VIPERdb link, current T-number assigned, virus name, family, genus, resolution of the reconstruction, PDB ID, date, genome type (Baltimore class) viral lineage, host, experimental imaging method, accessible capsid surface, external radius, internal radius, number of major capsid proteins, capsid protein fold, a description for the genome packing and replication strategy, and the associated Viral Zone entry link. The information will be mined using python web scraping scripts combined with quality control procedures. This will provide a reliable and semi-automatic method to update the database as new structures are deposited.

SA1-2 Computational method to determine the icosahedral lattice of a viral capsid

As discussed in the Preliminary Data section, a key bottleneck in the analysis of capsid architectures is the lack of a computational method to determine the topological and geometrical properties of these structures. This also limits the opportunity to identify unexpected layouts and was the main cause that for more than 60 years, the CK framework was overextended. To circumvent this, we propose to develop a topological and geometrical computational method that will score and rank the icosahedral lattices associated to a capsid as illustrated in Figure 5 for the procapsid of phage P22.

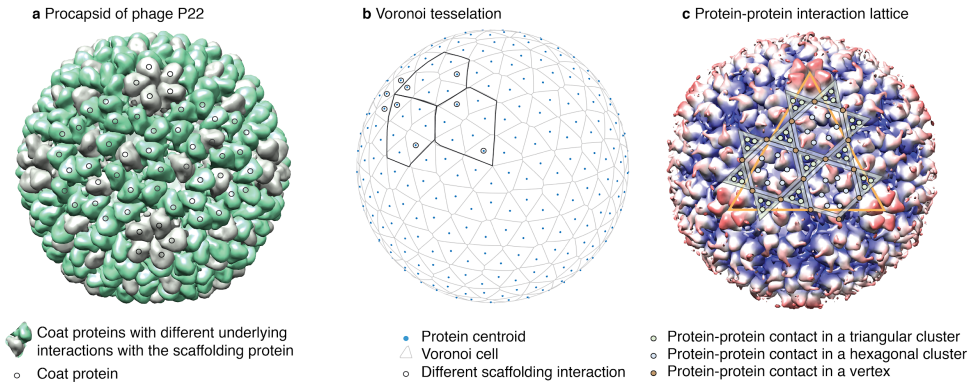


Figure 5. Computational methods to determine the icosahedral lattice of the capsid. **a**, Voronoi tessellation of the capsid protein centroids. **b**, Clustering of protein-protein contacts.

The topological method will generate the tiling of the capsid based on the centroids of the capsid proteins (Figures 5a and 5b). The centroids for each capsid protein will be obtained by averaging the position of the atoms of each protein chain in the PDB file of the associated viral capsid. The convex hull of the set of centroids will be obtained, and the Voronoi spherical tessellation will provide a capsid tiling (Augenbaum and Peskin, 1985; Na et al, 2002; Caroli et al, 2010; Luong and Bruno, 2018). We propose to transform the Voronoi spherical tessellation onto a flat surface with a 623 symmetry by applying the inverse transformation of the CK method, that is, introducing a facet on each pentagon (Caspar and Klug, 1962; Conway, 2008). This will conserve the planar angles and relative distances between vertices. Each icosahedral lattice will be superposed to the Voronoi lattice. The relative error of each lattice will be calculated by summing the distance between each lattice vertex and the nearest vertex in the projected Voronoi tessellation and dividing by the total number of vertices compared. The relative error of each lattice will be minimized by varying the lattice scale and orientation using Gradient Descent (Boyde and

Vandenberghe, 2004) and Monte Carlo algorithms (Metropolis, 1987). The minima obtained across lattices will be ranked. The smallest value will be used as the primary selected lattice. Our preliminary analysis for phage P22 procapsid led to a tessellation of seven different pentagonal tiles resembling triangles (Figure 5b). The best lattice identified corresponded to a snub hexagonal dual lattice T_s^D (*1,0*) with kites made of seven proteins each. This lattice was not expected from our initial visual analysis, but it actually accommodates better the distortions observed across multiple vertices in the tessellation as well as the interaction with the underlying scaffolding proteins. In addition to providing a quantitative way to assign the icosahedral lattice, the findings of this method can provide insights on the capsid assembly. In particular, the results for P22 indicate that in the assembly process proteins may interact with the scaffolding protein to form heptamer-kites, which would reduce kinetic traps in the capsid assembly (Luque et al, 2012). This exemplifies how our approach could be also used to generate new hypotheses about virus assembly.

Alternatively, the geometrical method will determine the tiling of the capsid proteins based on the protein-protein interaction network. Standard molecular cutoffs will be used to determine the interaction of amino-acids between different capsid proteins (Schlick, 2010). The interactions generated will be clustered by icosahedral symmetry. From this point on the process will be similar to the Voronoi tessellation method. In the case of phage procapsid P22, the protein-protein interaction approach led to a trihexagonal lattice $T_t(2,1) = \alpha_t T(2,1) = 28/3$. Thus, the capsid interactions seem compatible with the lattice observed among the other HK97-lineage members discussed above.

It is possible that the Voronoi tessellations and interaction clusters could adopt configurations that cannot be explained satisfactorily by the new icosahedral framework. In that case, it would be possible to test additional lattices by relaxing the requirement of having an edge-to-edge lattice (Twarock and Luque, 2019). We have identified nine non-edge-to-edge lattices (eighteen including the duals) that contain hexagonal sublattices (Grünbaum and Shepard, 1987). Here we propose that these lattices are amenable to generate additional icosahedral series.

SA1-3 Test of prediction 1: viral replication strategies and icosahedral lattices
The methodology developed in SA1-2 will be applied to the capsid structures in the database generated in SA1-1 to test our working hypothesis. The viral replication strategy and capsid protein fold will be compared for each of these viruses. If viruses with different replication strategies or capsid protein folds share the same icosahedral lattice, or if viruses with the same replication strategy display different lattice icosahedral groups, that will falsify our hypothesis. Otherwise, our hypothesis will be confirmed, at least for the set of data available to date. The null hypothesis in our case will be represented by icosahedral lattices that are independent of the viral replication strategy and capsid protein fold. Intermediate results could be also possible: different viruses in the same viral lineage could adopt a distinct icosahedral lattice. Any of these outcomes would provide valuable insight about virus classification and evolution.

Specific Aim 2: Predict icosahedral capsid architectures from genetic information
In each virus, the capsid protein has been selected to form a capsid with a well-defined size (Bruinsma et al, 2003; Mannige and Brooks, 2010). If, in addition, the capsid protein fold is related to the icosahedral capsid lattice, the information encoded in the capsid protein gene could be used to infer the capsid architecture of viruses. This would be feasible even if a viral lineage adopts multiple icosahedral lattices as far as variations in the capsid protein fold correlate with the different lattices. The challenge here is to obtain a large enough database relating capsid protein fold variations and capsid architecture to determine the validity of our rationale and the predictive power of our approach. This will be addressed in three sub-aims. In SA2-1, we will obtain the allometric relationships between capsid geometrical properties and genome size. In SA2-2, we will use these relationships to estimate the capsid architecture from the public library of sequenced viruses. In SA2-3, we will use statistical learning methods to test and assess the ability of the capsid protein gene to predict the capsid architecture. As described below, our preliminary analysis of the HK97-lineage suggests a ~80% accuracy in the prediction of capsid architecture form capsid protein genetic information.

SA2-1 Allometric relationships in viral lineages

The fact that viral lineages use different strategies to store the genome at different densities implies that there should be a distinct allometric relationship between genome size and capsid geometrical properties for each viral lineage. To confirm this conjecture, we will fit the icosahedral mesh in UCSF Chimera to measure the internal/external capsid radius, capsid thickness, internal/external surface, internal/external volume, internal/external sphericity, internal/external major capsid protein area, and packed genome density of viruses in the database generated in SA1-1 (Figure 6). We will also compile and measure cryo-EM reconstructed capsids that have no associated molecular models. These capsids will be obtained from the Electron Microscopy Data Bank (EMDB), which contains 1436 entries in the predefined category of virus particles (Aug 21, 2019). Redundant icosahedral capsids will be discarded.

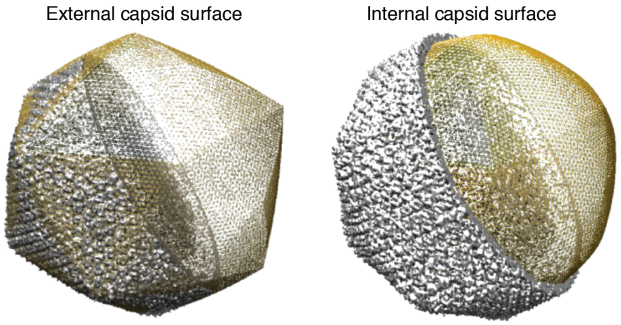


Figure 6. Capsid measurement using Chimera**.** The size and sphericity of a high-resolution icosahedral mesh (gold) is fitted to the external (left) and internal (right) capsid surfaces to measure the size, thickness, surface area, and volume of a capsid. The capsid structure rendered in grey using Chimera corresponds to PBCV-1 phage (EMDB 5378).

We performed a preliminary analysis of cryo-EM reconstructions of capsids in the HK97-lineage (n=35) with genome sizes ranging from 17 kbp to 289 kbp (kilo base pairs). The genome density (0.48 ± 0.06 bp/nm^3) and internal capsid protein surface (23 ± 2 nm^2) were constant. The external capsid diameter followed a power function relationship with the genome size with exponent 0.35 ± 0.3 ($R^2 = 0.97$), which is consistent with the genome occupying the full capsid volume (theoretical exponent $1/3 \approx 0.33$). As shown in the next sub-aim, a similar relationship was observed between the capsid architecture (*T*-number) and genome size. This supports our idea that capsids in a viral lineage follow a well-defined allometric relationship with respect to the genome size. We expect that even if the scaling exponent is similar in different lineages, the prefactor constant will differ due to variations in genome densities and capsid protein fold sizes among viral lineages.

SA2-2 Prediction of capsid architectures based on genome size

Most viruses that have been sequenced have not been investigated structurally due to the higher cost associated with high-resolution reconstructions compared to genome sequencing. The National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) collection contains currently 9222 complete viral genomes (Aug 21, 2019). Only a small fraction of those entries (10% or less) have associated a high-resolution reconstruction. In this sub-aim we propose to estimate the capsid architecture associated to all these genomes. This will require to determine the viral lineage associated to each genome and then apply the allometric relationships as a function of genome size obtained in SA2-1 to estimate the *T*-number, capsid diameters (interior and exterior), and sphericity of each virus.

For those genomes that have been already annotated, the major capsid protein fold will be identified using the homology markov model bioinformatic tool HHpred (Söding *et al.*, 2005; Krupovic and Koonin, 2017). For those cases were the major capsid protein has not been annotated, the capsid protein gene will be identified using the bioinformatic platforms viralPro (Galiez *et al.*, 2016) and viral ANN (Seguritan *et al.*, 2012). The viral genomes will be filtered based on the eleven capsid protein folds that have been identified among icosahedral viruses (Krupovic and Kooning, 2017).

We performed a preliminary analysis for the HK97-fold (n=35) to assess the accuracy of predicting the capsid architecture (*T*-number) using the architecture-genome allometric relationship. The power function model led to a scaling exponent $\alpha = 0.71 \pm 0.04$ with a coefficient of determination $R^2 = 0.98$ (Figure 7a). This is consistent with the *T*-number scaling as the surface of the capsid (theoretical exponent: $2/3 \approx$

0.67). The accuracy of the model was investigated by varying the fraction of data used for training and testing. The relative error decreased exponentially reaching a plateau of ~7% when using 20 data points or more (Figure 7b). Thus, our current method could predict the architecture of HK97-fold capsids from the genome size with 90-95% accuracy. We expect that other viral lineages could lead to a similar result.

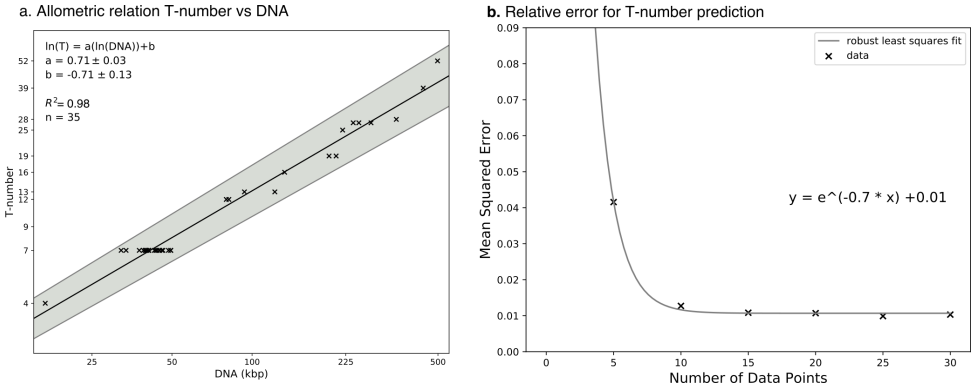


Figure 7. T-number prediction from genome size. **a**, Allometric relationship between T-number and genome size in the HK97-lineage. **b**, Relative error in the prediction of the T-number versus the number of data points used in the allometric model.

SA2-3 Prediction of capsid architectures based on the capsid protein gene

To determine if information from the capsid protein gene can predict the capsid architecture, we will apply a non-parametric statistical learning regression, such as random forest (Hastie et al, 2016), to obtain the relationship between capsid properties and input features from the capsid protein sequence (Figure 8a). The models will be trained using the frequency of amino-acids, the amino-acid sequence length, and the iso-electric point of the capsid protein as inputs. These properties were shown to discriminate the function of viral genes, in particular, the capsid protein function (Seguritan et al, 2012). This method could predict the capsid architecture even for capsid protein genes that have no apparent similarity with viral genes present in public databases. This is particularly important considering that about 70% of the genes in a viral genome do not usually show sequence similarity with known genes.

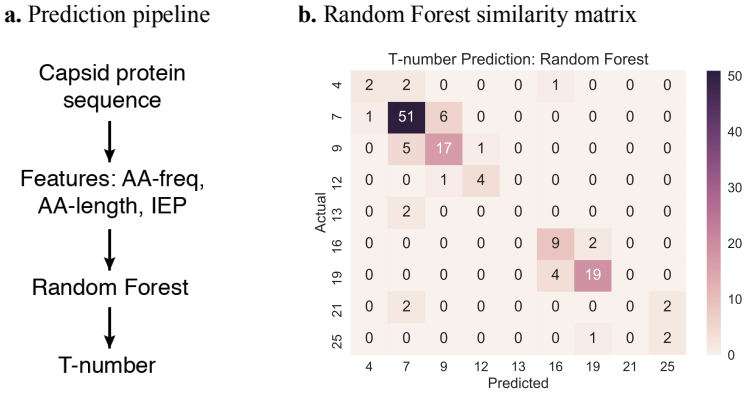


Figure 8. Prediction of T-number from capsid protein genetic information. **a**, Pipeline for the prediction of *T*-numbers from capsid protein genetic information. **b**, Random forest similarity matrix for 600 capsid proteins, using 80% of the proteins as a training set

To test the feasibility of this method, we implemented this approach for 600 viral genomes where the capsid protein was annotated and had assigned the HK97-fold. The capsid architecture was obtained using the *T*-number-genome allometric model discussed in SA2-2. The regression of *T*-number as a function of the capsid protein features mentioned above was obtained using random forest. Multiple sets were generated to investigate the accuracy of the method. In each round, 80% of the data was used as training and 20% for testing. Our initial analysis led to an accuracy of 85% (Figure 8b). This supports our hypothesis about the relationship between capsid protein fold variations and capsid architecture. The goal of this sub-aim is to refine this model for the HK97-lineage and extend the pipeline to the rest of viral lineages.

Specific Aim 3. Establish the landscape of icosahedral-based elongated and conical polyhedra

The mathematical principles used to obtain the sets of icosahedral architectures in the new viral capsid framework can be extended to generate also elongated and conical polyhedra. These two shapes are relevant in the virosphere and are of particular interest in biotechnology and biomedical applications (Luque and Reguera, 2010; Luque et al, 2010). The geometrical characterization and study of these capsid architectures is the most relevant mathematical innovation of this proposal.

SA3-1 New sets of elongated capsid structures

Elongated viral capsids (aka, bacilliform or prolate) are characterized by hemi-spherical-like caps and an elongated body (Luque and Reguera, 2010; Luque et al, 2010). They represent about ~20% among 5,500 microscopies of capsids from virus isolates and belong to viral families that also include icosahedral viruses (Ackermann, 2007; Krupovic and Koonin, 2017). These structures share important geometrical and physical properties with quasi-spherical capsids.

Elongated capsids can be obtained by extending the central part of an icosahedron while centering the caps on any of the icosahedral axes of symmetry, that is 5-fold, 3-fold, and 2-fold (Luque and Reguera, 2010) (Figure 9). Each cap contains six disclinations (pentamers) to concentrate the necessary Gaussian curvature to close the capsid due to Euler's Theorem. The symmetry fixed on the central part of the caps determines how the original 20 triangular faces of the icosahedron have to be distributed between the tubular body and the caps (Figure 9). The number of triangles in each case is $\Delta_{cap}^{5-fold} = 5$, $\Delta_{cap}^{3-fold} = 4$, $\Delta_{cap}^{2-fold} = 4$ as shown in Figures 9a, 9b, and 9c, respectively. The global symmetries determine the number of non-equivalent triangles in the body. This considers the global axial symmetry and the 2-fold symmetry at the center of the body, which guarantees that the underlying hexagonal sublattice is commensurate. The total symmetry product is $s_t = 2s$, where s is the rotational symmetry. The number of non-equivalent triangles in the body is given by $\Delta_{neq} = \Delta_{body}/s_t$, that is, 1 (5-fold), 2 (3-fold), and 3 (2-fold) as displayed in Figures 9a, 9b, and 9c, respectively. Two lattice vectors, one for the cap, $C_T(h,k)$, and one for the body, $C_Q(h',k')$, generate the catalog of possible elongated structures (Figure 9). For each rotational symmetry and fixed T-number in the cap, there is an infinite set of elongated architectures with different discrete body lengths (Luque and Reguera, 2010; Luque et al, 2010).

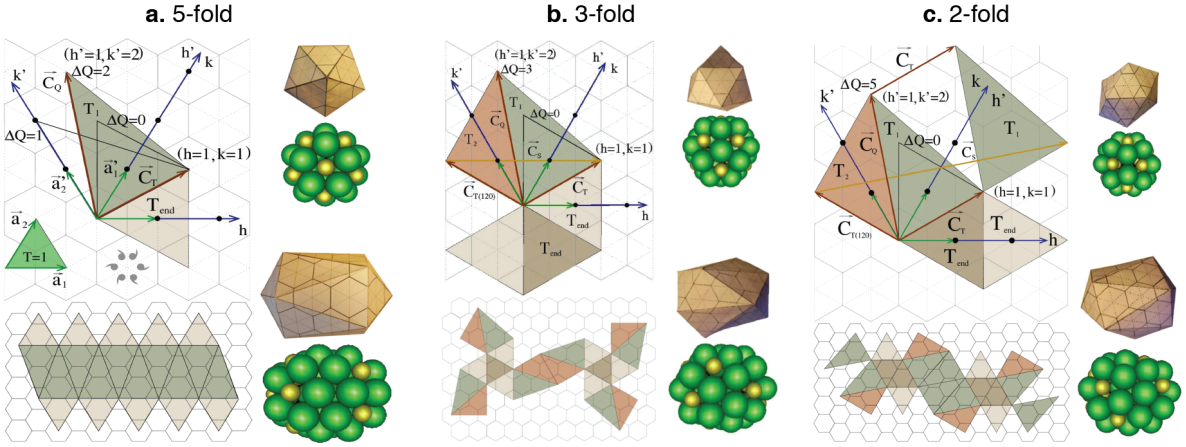


Figure 9. Templates for elongated viral capsids. Each panel illustrates the geometrical description to generate an elongated capsid for each icosahedral axis of symmetry: 5-fold (**a**), 3-fold (**b**), and 2-fold (**c**). In each panel, the top left lattice depicts the non-equivalent triangles in the cap (light and dark browns) and the non-equivalent triangles in the elongated body (green and orange). The bottom lattice contains the 20 triangles necessary to close the elongated capsid. To the right, the folded structure is displayed oriented through the axial symmetry axis (top) and from the 2-fold body symmetry axis (bottom). A spherocylindrical coarse-grained molecular model is included with spherical capsomers of different size: hexamers (green) and pentamers (yellow).

We previously applied this construction using the hexagonal lattice, obtaining specific geometrical properties and protein stoichiometry formulas for each symmetry (Luque and Reguera, 2010). We proved that the lengths are discretized: a specific number of proteins is required to increase the size (length) of

the capsid prolate, and this number depends on the symmetry and *T*-number in the cap. These properties facilitated the prediction of elongated capsids that were validated with structural information. We also developed a coarse-grained molecular model that assessed the thermodynamic feasibility of each symmetry, showing that the 5-fold case is favorable across *T*-numbers, the 3-fold case is favorable only among small *T*-numbers, and the 2-fold case is favorable only for $T = 1$ (Luque et al, 2010).

The goal of this sub-aim is to extend the construction of elongated capsids to the new set of icosahedral lattices and use the molecular coarse-grained model to determine those architectures that might be more feasible in the virosphere. The geometrical construction will provide the geometrical radius, length, protein positions, and protein stoichiometry for the extended catalog of elongated capsids with icosahedral caps. The allometry laws derived for each viral lineage in SA2-1 will be used to scale the geometrical radius and length to facilitate the comparison with electron microscopy images of elongated capsids (Ackermann, 2007; Actinobacteria db). The equilibration of the molecular coarse-grained model will be done using a Monte Carlo method scheme that fixes the number of tiles and optimizes the radius and length of the structure (Zandi et al, 2004; Luque et al, 2010).

SA3-2 New sets of conical capsid structures

Conical capsids are observed in some viral lineages that contain icosahedral viruses (Krupovic and Koonin, 2017). In particular, HIV adopts a capsid with this shape (Briggs et al, 2006). The conical structure is equivalent topologically to a sphere, and, based on Euler's theorem, 12 disclinations (pentamers) must be present to generate the curvature of the cone. Contrary to elongated capsids, in conical architectures the number of pentamers must be different in each cap, that is, 10+2, 9+3, 8+4, and 7+5 (large cap + small cap) (Figure 10). The case 11+1 reduces to the sphere and the case 6+6 reduces to an elongated structure. These configurations predict conical capsids with different angles that have been observed for HIV (Ganser-Pornillos and Sundquist, 2008). The precise reconstruction of the HIV capsid, and, in particular, the position of the pentamers, is still under debate (Zhao et al, 2013). Here we hypothesize that the icosahedral lattices can be used to generate conical capsids and provide the landscape of geometrical structures to characterize the architecture of HIV and other conical viruses.

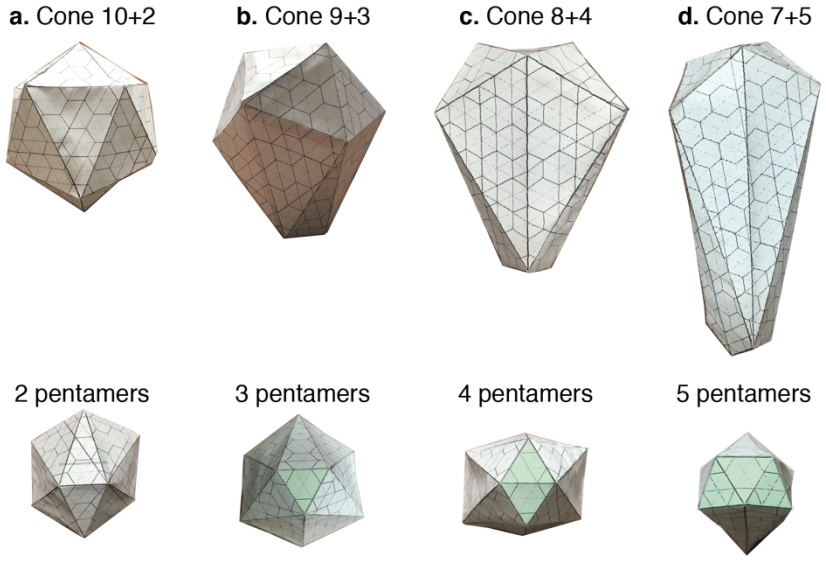


Figure 10. Geometrical models for conical capsids. The panels display folded structures for conical capsids with an increasing number of pentamers in the small quasi-spherical cap: 2 pentamers (**a**), 3 pentamers (**b**), 4 pentamers (**c**), and 5 pentamers (**d**). Top images: side views of the cones. Bottom images: axial view of the small spherical cap.

The extension to build conical architectures shares similarities with the construction of elongated capsids, but it is more challenging. A few previous attempts to generate conical structures used the CK framework, but the rules missed known examples (Nguyen et al, 2005; Sadre-Marandi et al, 2014). Below we illustrate our new approach, which is valid for all conical angles and *T*-numbers. The number of pentamers were fixed in the small and large cap. In each case, the possible rotational symmetry was evaluated. Since the caps in the conical structures are of different size, each cap must adopt a different *T*-

number, T_{small} and T_{large}. Our working hypothesis is that these two *T*-numbers must belong to the same *P*-class, so the lattice of the conical body can be commensurate with both caps. The *P*-class is defined by those *T*-numbers that generate self-similar capsid architectures. It is given by the equation $T(h,k) = f^2 P(h_0,k_0) = f^2(h_0{}^2 + h_0k_0 + k_0{}^2)$, where *f* is the common divisor of *h* and *k*, and the greatest common divisor of h_0 and k_0 is one (Caspar and Klug, 1962; Luque and Reguera, 2010; Aznar et al, 2012). The architectures $T = 1, 4$, and 9 are in class $P = 1$, and $T = 3$, 12, and 27 are in class $P = 3$. The examples provided in Figure 10 were generated by $T_{small} = 3$ and $T_{large} = 12$. Contrary to elongated capsids, once the *T*-number of the caps and conical angle (distribution of disclinations), the length of the body should be unique.

The goal of this sub-aim is to prove the theorem associated to this approach and generate the catalog of conical structures derived from the eight icosahedral lattices introduced in our prior work. As in the case of elongated structures, we will derive the radius of each cap, length, position of proteins, and protein stoichiometry rules as a function of the distribution of pentamers, the icosahedral class *P*, *T*-numbers, and icosahedral lattice. The outputs from these structures will be compared with the molecular models of HIV obtained from cryo-EM reconstructions (Zhao et al, 2013).

INTELLECTUAL MERIT

This work will test if icosahedral lattices are specific to each capsid protein fold and viral replication strategy. It will also test if the capsid protein gene alone can predict the architecture of viral capsids. This aims to unveil fundamental relationships between viral geometry, molecular structure, and genetics, and it aligns with one of the ten NSF's big ideas: Understanding the Rules of Life, that is, elucidating the sets of rules that predict an organism's phenotype. This process will require the development of mathematical and computational methods that will be valuable independently of the outcome of the hypothesis test. The project will provide a computational method that will assess the icosahedral lattice of viral capsids quantitatively, a process that is currently manual and ambiguous. It will also build a database relating geometrical, molecular, and genetic properties of viral capsids. The allometric relationships and conserved architectural properties will be established for each viral lineage, and a computational pipeline will be developed to predict the architecture of viral capsids using information from the capsid protein gene. Additionally, the mathematical principles behind the construction of icosahedral capsids will be extended to generate new infinite sets of elongated and conical capsids, establishing the geometrical landscape of icosahedral based capsids.

BROADER IMPACTS

The publications associated with the proposed research will provide an integrative framework for capsid geometry, capsid protein structure, and viral genetics. The characterization and prediction of capsid structures from genetic information will stimulate the environmental study of viral capsid architecture from sequence data. The computational methods developed to assess the icosahedral lattice of viral capsids will be integrated into the widely used UCSF Chimera software. This will promote the quantitative analysis of the geometry of viral capsids by the community, overcoming the manual—and often misplaced—classification of viral icosahedral capsids that has accompanied the field of structural virology for more than sixty years. The geometrical, molecular, and genetic relationships of viruses elucidated in our database will be shared with the online database for icosahedral virus capsid structures VIPERdb, the International Committee on Taxonomy of Viruses (ICTV), and the SIB Swiss Institute of Bioinformatics web-resource ViralZone, which integrates molecular, epidemiological, and virion information. This will facilitate the impact of our findings beyond the field of structural virology. Additionally, the mathematical approach proposed here could be further extended to generate new quasi-spherical, elongated, and conical polyhedra with other symmetries, e.g., octahedral, observed among Platonic, Archimedean, and Catalan solids.

Outreach workshops will be developed at the local and national levels combining geometrical virology, origami, 3D printing, and 3D photogrammetry-modeling. The goal will be to (i) promote the participation of underrepresented students doing research in mathematical and computational sciences at San Diego State University (SDSU), and (ii) train the new generation of structural virologist to assess and characterize the geometrical properties of viral capsids. Pilot versions of the workshops were offered to K-12, undergraduate, and graduate students by the PI during the academic year 2018–2019. In this proposal we aim to refine these workshops and expand them to train students in the field of structural virology at the National level.

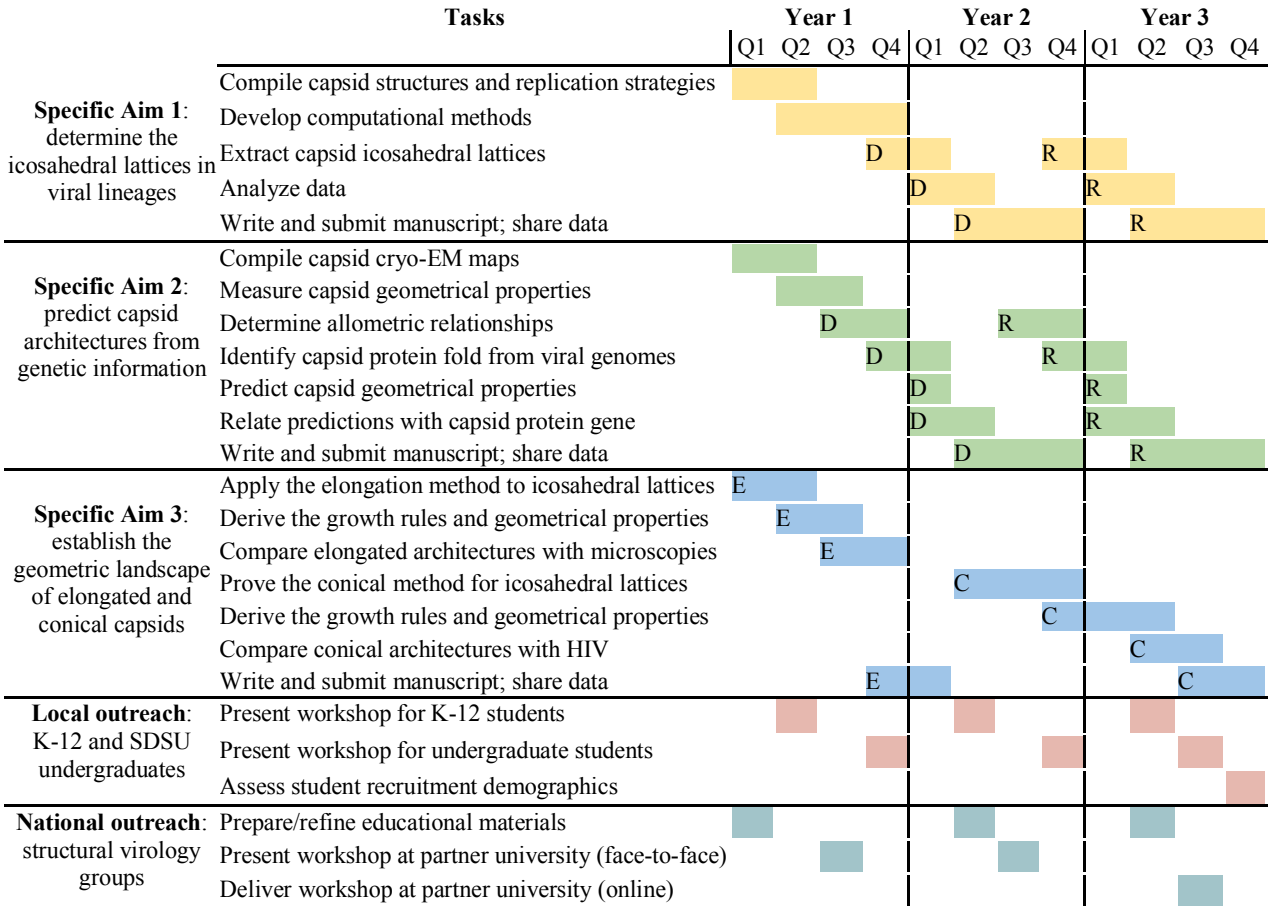
At the SDSU level, a day-long workshop will be offered every Spring to K-12 students during the SDSU Marine Science Day; students will build icosahedral capsids of marine viruses using origami. The second workshop, "The icosahedral capsid challenge", will be offered to undergraduate students every Fall at SDSU. We will coordinate with SDSU underrepresented undergraduate research programs, like MARC and IMSD, to increase the diversity of the student pool. In this one-day workshop, students will learn the geometrical constraints behind the architecture of viral capsids and design and fold newly established viral icosahedral structures (Twarock and Luque, 2019). The goal will be to promote interdisciplinary research and recruit undergraduate students to any of the twenty labs in the Viral Information Institute (VII) at SDSU. When concluding the grant, we will count the number of students that participated in the workshop and initiated research projects in the Viral Information Institute. The outreach will be considered a success if the demographics of the students initiating research reflects the diversity of SDSU as a Hispanic Service Institution.

At the National level, a two-day workshop will be developed to train graduate students and senior researchers in structural virology to use the new icosahedral framework of viral capsids. The workshop will be offered at the University of Connecticut, which has three interdisciplinary labs in structural virology (combining cryo-electron microscopy and molecular modeling) and will attract external groups. Dr. Simon White (Assistant Professor at the University of Connecticut) will be the host in the partner institution. The PI will offer the workshop face-to-face in years 1 and 2. The goal is to generate sufficient online materials that can be used in the third year by faculty in the partner institution, so the workshop can be offered regularly and shared with other institutions. The goal by the end of the grant is to obtain an agreement to disseminate the workshop with two additional institutions.

The research associated with this project will train two PhD students and six undergraduate students. SDSU is a designated Hispanic Serving Institution, which will facilitate the promotion of research among underrepresented students in mathematical sciences. Among the 15 students that have been trained or are doing research in the PI's lab, 47% belong to underrepresented minorities and 33 % are female.

TEAM AND TIMELINE

The project is based on the extensive experience of the PI in modeling the geometrical and physical properties of viral capsids (Luque et al, 2010; Luque and Reguera, 2010, Carrasco et al, 2011; Aznar et al, 2012; Luque et al, 2012; Luque and Reguera, 2013; Hernando-Perez et al, 2014; Twarock and Luque, 2019). All specific aims will be developed simultaneously. Specific Aims 1 and 2 will be led by two graduate students from the Joint-Doctoral program in Computational Science between University of California, Irvine and SDSU under the PI's mentorship. The PI will take the lead of Specific Aim 3 to develop the math required for the new sets of elongated and conical architectures. Undergraduate students from the Applied Mathematics program at SDSU will participate as research assistants in the three aims. The timeline for the project is shared below and is divided into quarterly periods. The letters D, R, E, and C designate research focusing on DNA viruses, RNA viruses, elongated architectures, and conical architectures, respectively.



RESULTS FROM PRIOR NSF SUPPORT

Luque: Collaborative Research: A National Consortium for Synergistic undergraduate Mathematics via Multi-institutional Interdisciplinary Teaching Partnerships (SUMMIT-P) (NSF DUE 1625166, 09/15/2016 – 08/31/2021). This grant is a consortium of eleven institutions collaborating to revise and improve the curriculum for lower division undergraduate mathematics courses in conjunction with the Committee on Curriculum Renewal Across the First Two Years (CRAFTY) of the Mathematical Association of America (MAA). New methodologies for teaching calculus were developed, implemented, and assessed. Additionally, new instruments have been developed to determine the needs of life sciences courses regarding mathematical sciences topics at SDSU. And the PI won x awards for the work he did improving the curriculum of math at SDSU.