Small uncultured viral-like entities redraw the origin of viruses and cellular compartments

Abelardo Aguilar[1], Ingrid Miranda[2], Omer Nadel[1], Stephen Nayfach[3], Anca M. Segall[4], Forest L. Rohwer[4], Simon Roux[3], Simon White[5], and Antoni Luque[1,6,*]

[1]Department of Biology, University of Miami, Miami, FL 33146, USA.

[2]Department of Evolutionary Biology, Universidad Nacional Autónoma de México, Ciudad Universitaria, CDMX 04510, Mexico.

[3]Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

[4]Department of Biology, San Diego State University, San Diego, CA 92182, USA.

[5]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269, USA.

[6]Department of Physics, University of Miami, Miami, FL 33146, USA.

[*]Corresponding author: antoni.luque@miami.edu

ABSTRACT

The HK97-fold is a conserved three-dimensional protein structure specialized in forming protein shells. It is the primary building block of capsids that protect the genome of viruses in the *Duplodnaviria* realm, infecting organisms across all domains of life. It is also the primary building block of encapsulins, compartments that confine biochemical reactions in prokaryotes. Recent genomic studies have hypothesized that encapsulins evolved from HK97-fold viruses. However, this evolutionary pathway is challenging to justify biophysically because HK97-fold viruses form larger and more complex protein shells than encapsulins. We addressed this paradox by searching for smaller and simpler HK97-fold viral entities in metagenomes across ecosystems. The investigation yielded a well-defined group of viral entities, which encode HK97-fold proteins displaying significant molecular similarities with encapsulins and lacking tail genes. The structural phylogenetic analysis of HK97-fold entities revealed likely bidirectional evolutionary transitions between encapsulins and HK97-fold viruses. A potential evolutionary mechanism responsible for such transitions was identified based on the presence of lysogeny-associated genes in the viral entities and the structural and molecular parallels between encapsulins and procapsids, the immature state of viral capsids before genome packaging. Structural comparisons indicated that procapsids are akin to the common HK97-fold protein shell ancestor and might still facilitate transitions between modern viruses and encapsulins. This also implies that viral procapsids may retain the ability to carry out biochemical reactions during infections or when released in the environment, which, if confirmed, would require a fundamental revision of the role of viral capsids in the biosphere.

KEYWORDS

Viral capsids, encapsulins, HK97-fold, bioinformatics, metagenomics

INTRODUCTION

Viruses are Earth's most abundant biological entities, infecting hosts across all ecosystems (Cobián Güemes et al., 2016; Paez-Espino et al., 2016; Wigington et al., 2016; Wommack & Colwell, 2000). Most viruses protect their genetic material in protein shells called capsids, which assemble from multiple copies of the major capsid protein (MCP) encoded in the viral genome (Koonin et al., 2022; Twarock & Luque, 2019). Viruses in the ancient *Duplodnaviria* realm, which include herpes-like viruses in Eukarya and the abundant tailed bacteriophages in Bacteria and Archaea, protect their genome in icosahedral capsids built from MCPs adopting the HK97-fold (Figure 1A) (Duda & Teschke, 2019; Wikoff et al., 2000). This fold is conserved despite the lack of protein sequence similarity across *Duplodnavira* (Abrescia et al., 2012; Krupovic & Koonin, 2017). The HK97-fold is also found in the capsids of mobile elements related to tailed phages, such as gene transfer agents (GTAs) (Bárdy et al., 2020a; Duda & Teschke, 2019; Esterman et al., 2021) and phage-inducible chromosomal islands (PICIs) (Fillol-Salom et al., 2018; Penadés & Christie, 2015).

This fold is also found in the proteins that form encapsulins, which are icosahedral compartments in bacteria and archaea (Giessen & Silver, 2017; McHugh et al., 2014; Sutter et al., 2008) (Figure 1A). Encapsulin operons encode the HK97-fold protein and a cargo protein, which is stored in the icosahedral compartment and confers a biochemical function such as oxidative stress response or iron mineralization (Andreas & Giessen, 2021). The shared use of the HK97-fold protein shells has led to the hypothesis that these viral and cellular systems share a common evolutionary ancestor (Andreas & Giessen, 2021; Krupovic & Koonin, 2017). The fact that ancestral HK97-fold protein compartments evolved the ability to carry biochemical reactions and

gene delivery functions in prokaryotes implies that this fold played an important role in the early evolution of cellular life. However, the directionality and mechanisms underlying the evolution of the critical functionalities of HK97-fold compartments remain unresolved.

Structural and phylogenetic analyses of HK97-fold proteins initially grouped encapsulins in a monophyletic lineage nested within HK97-fold proteins of viral origin (Krupovic & Koonin, 2017). However, the presence of encapsulins in archaeal groups, like the phylum Crenarchaeota, which are not known to be infected by HK97-type viruses (Heinemann et al., 2011), complicated the picture, leaving the evolutionary directionality between encapsulins and HK97-fold viruses uncertain (Krupovic et al., 2019; Krupovic & Koonin, 2017). More recent studies combining new experimental data and bioinformatic analyses based on hidden Markov models (HMMs) of HK97-fold proteins discovered thousands of new encapsulin-like systems, including two novel families (Andreas & Giessen, 2021). The updated phylogeny and structural comparison of HK97-fold proteins of viral and encapsulin origin revealed three encapsulin families forming a monophyletic group, and one of the new encapsulin families (Family 4) was found nested within the HK97-fold viral proteins. This study comparing HK97-fold proteins reinforced the hypothesis that encapsulins are derived from HK97-type viruses.

However, encapsulins form protein shells that are far less complex than the capsids of HK97-type viruses (Figure 1B) (Chmelyuk et al., 2023a; Duda & Teschke, 2019; Jones et al., 2024; Luque et al., 2020). Most encapsulins form protein shells built from 60 HK97-fold proteins distributed in 12 pentamers (oligomers of five proteins) as the vertices of the simplest icosahedral architecture, or T=1, and a few form T=3 and T=4 icosahedral architectures combining pentamers at the vertices and hexamers on the faces (Andreas & Giessen, 2021; Chmelyuk et al., 2023a; Kashif-Khan et al., 2024). The T-number (or triangulation number) is the number of major capsid

proteins in the asymmetric, which follows a sequence (T=1, 3, 4, 7, ...) defined by the equation $T = h^2 + hk + k^2$, which determines the number of major capsid proteins, 60T, in an icosahedral shell (Caspar & Klug, 1962; Twarock & Luque, 2019). Most HK97-type viruses form instead T=7 structures followed by T=16 structures (Dai & Hong Zhou, 2018; Helgstrand et al., 2003; Lander et al., 2013; Luque et al., 2020; Suhanovsky & Teschke, 2015; Yap & Rossmann, 2014) (Figure 1B). The largest HK97-type viral capsids reported correspond to jumbo phages forming T=52 architectures that contain more than 3,000 major capsid proteins (Hua et al., 2017). On the lower limit, the smallest reconstructions correspond to tailed phages forming either elongated T=3 capsids or quasi-spherical T = 4 capsids, both containing the same number of proteins (Choi et al., 2006; Luque & Reguera, 2010; Tao et al., 1998)(Aksyuk et al., 2012; Suhanovsky & Teschke, 2015). The smallest HK97-type capsid characterized, associated with moving genetic material (gene transfer agent or GTA), forms a shortened T=3-like structure, forming an oblate head (Bárdy et al., 2020). The lack of the simplest icosahedral architectures, T=1, among HK97-type viruses or mobile elements is an intriguing structural puzzle if encapsulins evolved from HK97-type viruses.

The differences in complexity of assembly between the biochemical encapsulins, and genetic viruses and mobile elements, HK97-fold compartments, reinforce the structural evolutionary paradox. HK97-type viruses undergo a complex maturation process requiring scaffold proteins that guide the assembly of the correct capsid size, followed often by the assembly of minor proteins reinforcing the structure (Fokine & Rossmann, 2014; Hardy et al., 2020; Podgorski et al., 2025). The HK97-type viruses and mobile elements require also to replace one pentameric vertex by a portal that facilitates the packaging the genetic material as a double-stranded DNA molecule and the appendage of tail that facilitates its delivery to the host (Hawkins et al., 2023; Rao et al., 2021; Xu et al., 2019). The HK97-fold proteins guided by the scaffold

proteins assemble around the portal forming an empty quasi-spherical structure called a procapsid, which expands, releases the scaffold proteins, and becomes more faceted upon the packaging of the genetic material (Ionel et al., 2011; Oh et al., 2014; Roos et al., 2012; Veesler & Johnson, 2012; Wikoff et al., 2006). The HK97-fold proteins in encapsulins assemble their icosahedral shells without requiring a scaffold, auxiliary structural proteins, or a portal, although they can co-assemble around several copies of a protein called cargo, responsible for catalyzing biochemical reactions (Chmelyuk et al., 2023a; Jones et al., 2024).

The structural evolutionary paradox between biochemical and genetic HK97-type compartments could be resolved if simpler T=1-like HK97-type viruses exist. This possibility was first proposed in 2020 (Luque et al., 2020). That study was built on the observation that the genome length of tailed phages follows a well-established allometric relationship with their capsid size and architecture (Cui et al., 2014; Lee et al., 2022; Luque et al., 2020). The existence of canonical T=1 capsids was predicted to be associated with genomes ranging roughly from 1 to 5 kb (Luque et al, 2020). Genomes ranging from 5 kb to the isolated threshold ~11 kb were predicted to form expanded T=1-like capsids, possibly reinforced by minor proteins (Luque et al., 2020; Twarock & Luque, 2019). The application of this model predicted that the smallest isolated tailed phage, *Salmonella phage astrithr* with a 11.6 kb genome length, forms a quasi-spherical T = =3 caspid (Luque et al, 2020), a prediction in agreement with recent low-resolution imaging of similar Astrithrvirus (Pourcel et al., 2024). The investigation of uncultured viral genomes from human gut metagenomes found over twenty candidates below the threshold of isolated tailed phage genomes (< 11.6 kb) (Luque et al., 2020). The shortest candidate displayed a genome of about 5 kb, and the shortest containing a major capsid protein (HK97-fold) had a 7.4 kb genome. However, no

candidates were predicted to form canonical T=1 capsids (< 5 kb), and no evolutionary analysis was carried out comparing these small HK97-fold viral entities and encapsulins.

Here, this intriguing evolutionary gap of HK97-fold compartments was addressed by mining Earth's metagenomes, including aquatic, terrestrial, host-associated, and engineered biomes. The goal was to identify HK97-fold proteins associated with genomes in the twilight viral zone of tailed phages, that is, below their isolated threshold. The hypothesis was that proteins related to smaller genomes would display a stronger structural relationship with encapsulins and reveal a basal group in the evolution of HK97-fold compartments. However, the analysis revealed a more nuanced evolution of HK97-fold compartments, including lineages of encapsulins nested within HK97-fold phage groups and vice versa. The discovery of a well-defined temperate and tailless group in the twilight zone led to the formulation of a molecular evolutionary mechanism to resolve the bidirectional evolution of HK97-fold compartments: procapsids represent a nexus between encapsulins and viruses that could facilitate their bidirectional evolution. This model opens a controversial question regarding the possibility that procapsids may carry biochemical functions during viral infections, adding a metabolic function to viral capsids.

METHODS

Identifying environmental HK97-type small genomes

Viral sequences from world metagenomes were obtained from the IMG database (Paez-Espino et al., 2019) and the Global Ocean Virome (GOV) 2.0 (Gregory et al., 2019). IMG/VR was accessed in December 2020, containing 18,373 cultivated and 2,314,329 viral sequences (Figure 1A). GOV was accessed in December 2020, containing uncultivated sequences from 145 samples across the

world's oceans. The biome categories were assigned following the hierarchical metagenomic classification schema for three hierarchical levels (Ivanova et al., 2010) (Data File S1).

The bioinformatics pipeline CheckV estimated the completeness of viral genomes and removed non-viral regions from integrated proviruses (Nayfach et al., 2021). The viral contigs selected for analysis (Data File S2) satisfied three criteria. First, the contigs contained a direct terminal repeat (DTR) window of at least 20 base pairs (bp), which was used as a bioinformatic proxy for a complete viral genome (Benler et al., 2021; Casjens & Gilcrease, 2009; Nayfach et al., 2021). Second, each selected genome contained at least one marker gene associated with tailed phages, including major capsid protein, portal protein, terminase, baseplate, or tail genes (Koonin et al., 2020). Open reading frames were identified using Prodigal v2.6.3 (option '-p' for metagenome mode) (Hyatt et al., 2010). Third, the contigs analyzed were filtered to have a length of 20 kb or less. This criterion captured the *twilight zone* (5 kb to 11.6 kb) below the smallest isolated-tailed phage containing a major capsid protein, *Salmonella phage ashtrithr* (Luque et al., 2020; Olsen et al., 2020), and a control region above it (11.6 kb and 20 kb). The twilight zone was divided into micro contigs (< 5 kb), which had not been previously observed among HK97-fold entities, and mini contigs (5 to 11.6 kb).

Genome annotation

The open reading frames (Data File S4) were annotated in three complementary steps (Figure 1B). First, high-level functional categories were predicted using the VPF-PLM virus protein function protein language model (Data File S5), running the script faa_prediction.sh (https://github.com/kellylab/viral-protein-function-plm) (Flamholz et al., 2024), which was trained on the prokaryotic virus remote homologous groups (PHROGs) database (Terzian et al.,

2021). Functions were assigned when obtaining likelihood scores greater than 60%; the highest-ranked function was selected. The method achieves a weighted F1-score of 85% at this threshold for assigning correct high-level functional categories to viral proteins validated on ocean virome sequences.

Second, VIBRANT frame (Kieft et al., 2020), the Virus Identification By iteRative aNnoTation tool, which combines neural networks and hidden Markov models (HMMs), was applied to predict specific functions for each open reading frame (Data File S6). Version 1.2.1 of VIBRANT was executed with default parameters, which included curated HMM profiles from three major databases: KEGG (March 2019 release), Pfam (v32), and Virus Orthologous Groups (VOG; release 94). In cases where the specific function predicted by VIBRANT conflicted with the high-level functional annotation from VPF-PLM, annotations were resolved by manual inspection using HHpred, described below (Data File S7).

Third, Major Capsid Proteins (MCPs) of the HK97 fold were identified through an additional HMM-based annotation step. From the set of proteins annotated as "Head and packaging" by VPF-PLM, and those annotated as "Major Capsid Protein" by VIBRANT, MCPs were confirmed based on domain annotation against the Pfam clan CL0373 ("Phage-coat"), which encompasses all known HK97-fold protein domains. The clan comprised 17 HMM profiles, including viral capsid protein domains (13), encapsulin shell domains (3), and a domain of unknown function (1). The HMM profile (41798) from the efam database (Zayed et al., 2021) was also added, as it was recently proposed as a novel lineage of viral HK97-fold major capsid proteins identified using protein language models (Flamholz et al., 2024). Domain annotation was performed using hmmscan (HMMER v3) with a cutoff of E-value $< 10^{-3}$. The HMM profile for the efam database entry and the clan entries summary are provided in Data File S8. After domain-

based identification, a representative set of 60 viral contigs (encompassing 639 open reading frames) was selected for detailed analyses. These open reading frames were manually annotated using the HMM-based remote protein homology detection server HHpred (Söding et al., 2005). Queried annotation libraries included Pfam-A (v37), NCBI Conserved Domains (v3.18), and PHROGs (v4) with default server parameters and a MAC realignment threshold of 0.3. The Protein Data Bank (PDB) library was included to enhance structural evolutionary insights. Hits returned by HHpred with less than 20% probability were discarded. The top hit was selected unless the hit region was shorter than subsequent hits with comparable significance. The selected function for each open reading frame was labeled based on statistical confidence: very high (>99%, ***), high (95–99%, **), moderate (70–95%, *), and low (50–70%,?). Annotations below 50% probability were considered hypothetical. If the functional category identified by manual annotation differed from the automated annotation, the manual annotation was prioritized. The annotation files are provided in Data File S7.

Phylogeny of environmental major capsid proteins

The environmental viral contigs were clustered using a sequence-based phylogenetic inference of the HK97-fold major capsid proteins (Data File S9, S10). The multiple sequence alignment (Data File S11) was performed using MAFFT v6.240 (Katoh, 2002). The alignment of conserved cores was optimized by the L-INS-i algorithm, which adds a generalized gap cost to accommodate variable regions (Katoh & Toh, 2008). The alignment was refined, and artifacts were reduced by increasing iterative cycles up to 1000 using the *--maxiterate* parameter. Gap penalties were adjusted with gap thresholds of 0.05 and 0.15, which are standard when aligning highly variable multi-domain viral proteins (Kazlauskas et al., 2019). Phylogenetic inference was performed using

maximum likelihood as implemented in IQ-TREE 2 v2.3.6 (Minh et al., 2020). The best-fit substitution model was based on the Bayesian Information Criterion (BIC) and obtained with ModelFinder, integrated into IQ-TREE 2. The tree was inferred with ultrafast bootstrap approximation using 1000 replicates, and the default consensus tree was generated. The outputs, including the maximum-likelihood tree, log file, and likelihood distances, are provided in Data File S12. The tree was visualized and annotated using iTOL v6 (Letunic & Bork, 2024), overlaying the biome of origin and genome length. The groups derived from the phylogeny were explored by partitioning the tree into clusters using TreeCluster v1.0.3 (Balaban et al., 2019), which generates biologically meaningful clusters based on tree-based distances. Clustering was performed using the *avg_clade* method, which groups leaves into clusters so that the average pairwise distance between leaves does not exceed a specified threshold while ensuring each cluster forms a distinct clade. A threshold of 0.001 was applied to identify highly specific groupings.

Comparative analysis across HK97-fold proteins

Reference HK97-fold protein database. The HK97-fold proteins obtained from environmental samples were compared with a reference database composed of HK97-fold proteins from curated sources, including viral genomes from NCBI RefSeq, viral-related mobile elements (Gene Transfer Agents [GTAs] and Phage-Inducible Chromosomal Islands [PICIs]), and encapsulins. Viral MCP sequences belonging to the *Duplodnaviria* group were retrieved from the NCBI Viral Genomes Resource (Brister et al., 2015) in December 2023 using the Boolean query ("MCP" OR "Major Capsid Protein") AND "viral" NOT "human" AND "refseq", filtering to select only those belonging to complete genomes. These sequences were further confirmed to contain HK97-fold domains using hmmscan against the Pfam clan CL0373 ("Phage-coat") with an E-value cutoff <

10^{-3}. The resulting validated sequences constitute the final reference dataset of 921 sequences (Figure 3B).

The major capsid proteins of phage-inducible chromosomal islands (PICIs), which are entities found in bacteria that are related to tailed phages and often encode major capsid proteins (Penadés & Christie, 2015) were obtained using the NCBI identifiers reported in PICIs publications (Fillol-Salom et al., 2018; Alqurainy et al., 2023), yielding 66 MCPs. Finally, a set of encapsulin sequences was obtained from a recent study that expanded the encapsulin universe (Andreas & Giessen, 2021), yielding 58 proteins from four encapsulin families: family I (51 proteins), family II (30), family III (7), and family IV (13). The identifiers of the reference HK97-fold database are available in Data File S15.

Estimating packaging capacity of encapsulins. To carry comparative packaging capacities between viral entities and encapsulins, the internal volume of encapsulins of T=1 (PDBs: 3DKT, 6X8T), T= (4PT2, 2E0Z), and T=4 (6NJ8), molecular reconstructions were measured using the Blob and Surface tools in ChimeraX (Meng et al., 2023). The full encapsulin structure was built from the asymmetric unit by applying icosahedral symmetry operations, a density map was generated, and the contour level was adjusted to define the capsid boundaries. The Blob Picker tool selected the inner capsid surface and measured the enclosed volume (Figure S1). The encapsulin packaging capacity was projected using the average density of double-stranded DNA in tailed phage capsids, ~0.5 kb/nm^3 (Luque et al., 2020; Podgorski et al., 2023; Stone et al., 2019).

Pfam domain analysis. The HK97-fold hidden Markov model (HMM) profiles captured in the phage_coat clan (CL0373) from Pfam were mapped to the major capsid protein sequences for the environmental groups (micro, mini, and control) and the reference HK97-fold database.

The protein sequence regions associated with the detected domains were recorded. For each protein, the phage_coat coverage was calculated as the number of residues related to the non-overlapping Pfam domains identified divided by the protein's amino-acid sequence length. The encapsulin enrichment of a protein was obtained as the fraction of the protein region associated with a Pfam encapsulin domain with respect to the full region mapped to phage_coat domains. This domain analysis enabled the identification of viral MCPs exhibiting Pfam domains typically associated with encapsulins, interpreted here as evidence of quantitative phenotypic overlap between encapsulins and viral capsids. These encapsulin-enriched viral MCPs were mapped onto the phylogeny of MCPs (Figure 4A) to identify if they form robust monophyletic groups.

Structural phylogeny. To investigate the evolutionary relationship of HK97-fold compartments, including the novel encapsulin-enriched MCPs, a structural phylogeny of HK97-fold proteins was constructed in three steps: candidate selection, protein folding, and phylogenetic reconstruction. First, 220 major capsid proteins (MCPs) from environmental contigs ensured representation across the length-based groups of micro, mini, and control viral contigs. This included 113 MCPs enriched in encapsulin domains. 147 MCPs were selected from the reference group (NCBI Refseq and Uniprot), including the 16 proteins used in prior HK97-fold studies due to their reconstructed PDB structures (Krupovic & Koonin, 2017). 14 additional MCPs from RefSeq were included as representatives of proteins with encapsulin domain enrichment identified in this study. Additionally, 6 encapsulin shell proteins represented the structures available in the Protein Data Bank (PDB), encompassing families I, II, and IV. 4 MCPs from PICIs were included, as well as 1 MCP encoding Gene transfer agents (GTA), taxonomically restricted to *Rhodobacter* (Bárdy et al., 2020a). The proteins selected are provided in Data File S13. Second, protein folds were predicted using AlphaFold2 in ColabFold (Jumper et al., 2021). Structures were evaluated

using the pLDDT (predicted Local Distance Difference Test) confidence metric. Folded protein structures scoring above 40 were retained for phylogenetic analysis. For proteins with experimentally determined structures available in the PDB, the PDB coordinates were used in place of AlphaFold predictions. The protein structures and their respective pLDDT scores are provided in Data File S13. Third, the structural phylogenetic tree for the HK97-fold proteins was constructed using FoldTree, which outperforms traditional sequence-based phylogenetic methods for closely and distantly related proteins (Moi et al., 2023). FoldTree uses local structural alphabets that capture key genomic features of protein features, avoiding issues associated with global structural metrics. The calculation used the default parameters and relied on FoldSeek for the all-vs-all structural comparison (van Kempen et al., 2023). Structural tree files are available as Data File S14.

Protein-domain co-occurrence analysis. Proteins containing more than one Pfam domain were investigated using a co-occurrence network using the Python NetworkX library (Hagberg et al., 2008). Pfam domains were the nodes. Edges were drawn between nodes co-occurring in the same protein and were labeled as canonical (connecting viral-viral or encapsulin-encapsulin Pfam domains) and non-canonical (connecting viral with encapsulin domains). The network layout was visualized using the spring layout algorithm, which positions nodes with a larger degree of connectivity (number of connections or spring strength) closer to each other. The algorithm adjusts the position of the nodes iteratively until equilibrium is reached (minimizing the overall spring tensions in the network). Node size was proportional to that node's total degree of connectivity (number of connections), scaled between 100 and 500 units. The potential evolutionary trajectories across the domain network were obtained using a spanning path analysis. Edge weights were added to reflect biological assumptions: edges between domains of the same

group (viral-viral or encapsulin-encapsulin) were assigned the minimum cost factor, equal to the distance between the nodes, while transitions between viral and encapsulin domains (non-canonical) were penalized, with twice the distance between nodes as the cost. All possible spanning trees were enumerated, and their costs were computed. The best paths were selected to minimize the total edge cost and the number of non-canonical transitions to follow a parsimony assumption.

RESULTS

Twilight viral entities are relatively frequent but few contain an identifiable MCPs

There were 4,849 viral contigs in the twilight zone (< 11.6 kb) containing direct terminal repeats (DTRs) and predicted to be related to tailed-phage-like viruses (Figure 2C). These represented 65% of the total viral contigs, including the control region (11.6 – 20 kb) with the same characteristics (2,611 viral contigs). The viral contigs from both regions were frequent across biomes (Figure 2F-G). However, the fraction of viral contigs where HK97-fold-like major capsid proteins (MCPs) were identified was significantly smaller in the twilight zone (11% or 536 contigs) than in the control region (28.61% or 798). Most twilight zone MCPs were from human gut (39.8% or 531 MCP-containing contigs), followed by freshwater (17.47% or 233 MCP-containing contigs), and a few additional biomes contained more twilight zone MCPs than control ones, including non-marine saline, invertebrates, peat, and oil reservoir biomes.

MCPs from mini viral contigs are enriched in encapsulin domains

The main hypothesis in the study is that HK97-fold capsid proteins from the twilight zone will be more closely related to encapsulin proteins. This was tested by comparing the presence of encapsulin domains in viral HK97-fold proteins (Figure 3). The analysis included all protein

domains associated with hidden Markov model profiles in the HK97-fold phage_coat clan from Pfam (Figure 3A). This included the phage-coat coverage of the protein and the fraction associated with each domain identified. The phage-coat amino-acid coverage decreased for MCPs from smaller viral entities (Figure 3B). The MCPs in micro contigs displayed the lowest coverage with a median of 50%. The mini and control contigs displayed a median coverage of 70%. The RefSeq group displayed the largest coverage, with a median of 95%, as expected. The viral contigs displaying the most substantial enrichment in encapsulin domains were from the twilight zone's mini contigs (5 kb to 11.6 kb) (Figure 3C). Nearly 20% of these candidates (93 out of 491 or 19%) contained encapsulin domains, with 57 (61%) MCPs displaying encapsulin domains as their only phage-coat domain and 22 (24%) displaying an encapsulin domain fraction between 0.20 and 0.55. In contrast, none of the 45 MCPs from micro candidates (< 5 kb) contained encapsulin domains. The enrichment of the mini contigs was significantly higher than the environmental control group, where only 2.5% (20 out of 798) of the MCPs contained encapsulin domains, although 4 displayed an encapsulin domain fraction of 1. In the RefSeq group, only 1.9% (17 out of 910) MCPs contained a trace of encapsulins domains, with all of them below a 0.5 fraction except for a member in the fringe group (11.6 kb to 20 kb) with an encapsulin fraction of 1. No MCPs in the viral mobile elements, GTAs, or PICIs displayed encapsulin domains.

The prevalence of encapsulin domains across viral MCPs decreased with the associated genome length (Figure 3D). The encapsulin-enriched twilight mini contigs (5 kb - 11.6 kb) occupied the genome length region between the two largest encapsulin systems, which displayed a projected genome capacity of ~5 kb and 10 kb, respectively. Encapsulin-enriched MCPs in the control group were associated with genome lengths ranging between 13 kb and 19 kb, right above the maximum projected packaging capacity of known encapsulins. The viral genomes from RefSeq

containing MCPs with encapsulin domains had genome lengths between 42 kb and 58 kb, except for one virus with a 103 kb genome (Figure 3D) (Data File S16).

The dominant group of encapsulin-enriched viral entities is tailless and likely temperate

The phylogeny of major capsid proteins (MCPs) revealed that the encapsulin-enriched MCPs form 10 different groups and 10 singletons (Figure 4A). One dominant, well-defined phylogenetic group contained 57% (75) of the encapsulin-enriched MCPs. Their genomic functional analysis revealed common genomic patterns (Figures 4B and 4C). Their average genome length was 8.9 ± 0.6, and they came from host-associated samples, as illustrated with characteristic representatives in Figure 4A. Structural genes covered an average of 22% (16 to 28%) of the contig, representing 1.5 kb (1.0 to 2.0 kb), and contained the MCP and scaffold (Figures 4A and 4B). Transcription regulation genes covered 21% (15 to 27%) of the contig, representing 1.4 kb (1.0 to 1.8 kb), and included RNA polymerase sigma factors, transcriptional repressors, and often quorum sensing proteins. DNA and RNA metabolism genes comprised 17% (10 to 24%) of the contig length, covering 1.2 kb (0.8 to 1.6 kb), and typically included RepA, DNA helicase, or primosomal DNA I. Lysogenic genes represented 16% (12 to 20%) of the contig, encompassing 1.1 kb (0.9 to 1.3 kb) and included integrase and, in some cases, excisionase. Genes of unknown function were comparable in size, accounting for 24% (16 to 32 %) of the contig, representing 1.7 kb (1.0 to 2.4 kb). A minor fraction of the genome, 0.4% (0 to 2.4 %), spanning 0.026 kb (0 to 0.126 kb) included potential auxiliary metabolic genes, such as ribosomal proteins, membrane receptors, or formin-like proteins. No terminases or packaging genes were found. No tail connector genes were found, and the only traces of tail genes were related to tail fibers in a few genomes and a tail tape measure protein in one genome (ORF DTR_497289_3).

The RefSeq phages containing encapsulin domains in their MCPs were also temperate (Figure 4B). They primarily infect bacteria in the microbiome of mammals, including humans, and a few infect archaea from the order *Poseidonales* (Data File S16). Their genome composition strongly represented tail genes, comprising an average of 19 ± 6 (SD). The twilight viral entities that were not encapsulin-enriched did not form a homogenous group (Figure 4A). They displayed a diverse genome composition that included a wide diversity of tail genes (Data File S17). Even some micro contigs of ~2.5 kb included connectors and tail fibers (Figure S2).

Bidirectional evolution of HK97-fold viruses and encapsulins

Including encapsulin-enriched viral MCPs in the structural phylogeny revealed an entangled evolutionary relationship between HK97-fold proteins across encapsulins, environmental viral contigs, isolated viruses, and mobile elements (GTAs and PICIs) (Figure 5A). These MCPs were scattered across multiple branches of the tree, rather than forming a single clade. The most abundant group, derived from human gut mini contigs, was not recovered as a sister to any encapsulin family. However, some encapsulin-enriched viral MCPs did fall into positions closer to encapsulins than to other viral groups, as discussed below.

In contrast, a clearer evolutionary connection was observed between Gp7 viruses and Family II encapsulins, which are enriched in the Srpl-like domain, which is characteristic of these encapsulins. In the phylogeny, Gp7 MCPs are nested within the clade formed by the Srpl-like domain. The likely placement of the most recent common ancestor of Family II encapsulins lies deeper in the tree than the inferred emergence of the Gp7 lineage, suggesting that Family II encapsulins predate the origin of Gp7 viruses. This topology supports a possible encapsulin-to-virus transition, in which Family II encapsulins gave rise to a viral lineage.

Encapsulin families generally formed consistent monophyletic groups. A major clade corresponding to Family I encapsulins was structured into two sub-groups: one comprising archaeal representatives (Bathyarchaeota, Crenarchaeota, Korarchaeota, Methanobacteria) and another comprising bacterial representatives (Thermotoga, Quasibacillus, Myxococcus). Both groups include encapsulins forming T=1 and T=3 geometries, indicating that T-number variation has occurred independently within domain-specific lineages.

Family I encapsulins are sister to a subset of Family IV encapsulins, specifically those from Archaea (e.g., Pyrococcus, Paleococcus, Thermococcus). In contrast, bacterial Family IV encapsulins form a separate, more derived group, suggesting that Family IV encapsulins are polyphyletic. In this scenario, archaeal Family IV encapsulins may have evolved from Family I ancestors, while bacterial Family IV encapsulins may trace back to a different origin, possibly linked to viruses. The preservation of domain-driven grouping across these clades supports this hypothesis.

A clade comprising MCPs characterized by the Phage capsid E Pfam domain was identified near the root of the tree and included representatives from extremophilic environments such as hydrothermal vent contigs (e.g., DTR248616) and thermophilic viruses like Oshimavirus. A large region of the tree contains viral MCPs associated with diverse Pfam domains (e.g., P22, Gp13, Gp23, Gp10, L_lactis, efam_41798). Although the relationships among these groups are not fully resolved, their overall position suggests they represent diversification events within viral lineages, rather than early transitions of the HK97 fold.

Another clade of interest includes Family III encapsulins, viral MCPs with Family I-like domains, and bacterial Family IV encapsulins. The grouping of these lineages points to non-canonical evolutionary trajectories between viral and cellular compartments, although the

directionality remains uncertain. The abundance and characteristics of the viral representatives make a viral origin plausible, but further resolution is needed.

The co-occurrence analysis of phage_coat protein families revealed that the Linocin_M18 and Srpl-like encapsulin domains exhibited extensive non-canonical connections with viral domains, displaying 83 and 11 connections, respectively (Figure 5C). These non-canonical links shaped the global organization of the domain network, bringing these encapsulin domains into proximity with viral domains. Notably, *Linocin_M18* (Family I encapsulins) and *phage_capsid* domains, as well as *Srpl-like* (Family II encapsulins) and *Gp7* domains, co-occurred frequently, suggesting evolutionary transitions. We applied a parsimony-based analysis to explore the most plausible evolutionary trajectories within this network. We searched for spanning paths that connected all domains without forming cycles, ranking them according to a cost function that favored canonical connections and penalized non-canonical transitions. The best-scoring evolutionary paths connected domains under these conditions. Nevertheless, non-canonical transitions remained necessary to connect the network fully, supporting evolutionary transitions between viral and encapsulin entities.

Structure and network-based evolutionary analysis support the existence of bidirectional transitions between viruses and encapsulins. However, they do not resolve whether the universal HK97-fold ancestor was initially a biochemical compartment (encapsulin-like), a container of genetic material (capsid-like), or whether it simultaneously fulfilled both roles.

Discussion

Twilight zone viruses display encapsulin domains and reveal HK97-type tail-less virus clades

This study was motivated by the paradox that encapsulins are inferred to derive from HK97-type viruses based on molecular and structural phylogenetic analyses (Andreas & Giessen, 2021; Giessen & Silver, 2017; Krupovic & Koonin, 2017), yet HK97-type viruses form much more complex protein shells than encapsulins(Chmelyuk et al., 2023a; Duda & Teschke, 2019; Jones et al., 2024; Luque et al., 2020). It had been hypothesized that the existence of HK97-type viruses forming simpler capsids than those observed among isolated HK97-type viruses would resolve this paradox (Luque et al., 2020). These viruses were predicted to encode genomes smaller than the isolated tailed phage genome threshold (11.6 kb) (Berg & Roux, 2021; Luque et al., 2020). The mining of uncultured viruses across ecosystems, containing direct terminal repeats (DTRs) as a proxy for completion and HK97-fold major capsid proteins (MCPs), yielded 536 viral candidates encoding HK97-fold MCPs in the twilight zone (< 11.6 kb) (Figure 2), comprising 491 mini viruses (5-11.6kb), and the first 45 micro viral genomes (< 5 kb) reported to encode HK97-fold proteins, with the smallest displaying a 2.4 kb long genome (Figure S2). The assumption of using DTRs as a proxy for genome length is supported by the observation of similar genome length distributions between DTR-based and non-DTR-containing genomes among isolated tailed phages (Gulyaeva et al., 2022; Luque et al., 2020; Tisza & Buck, 2021). The assumption that genomes in the twilight zone assemble smaller capsids than those among isolated tailed phages was supported by a concurrent study (Nadel, Cioffi, Aguilar et al 2025).

The molecular analysis of HK97-fold MCPs using hidden Markov model (HMM) profiles confirmed the hypothesis that MCPs in the twilight zone were more similar to encapsulins than MCPs from genomes above the isolated tailed phage threshold (Figure 3). About 20% of the HK97-fold proteins in the twilight zone displayed encapsulin profiles compared to 2% in the

environmental control (Figure 3C). However, in the twilight zone, the encapsulin-enriched MCPs were found only among mini genomes (5 kb to 11.6 kb). No MCPs from micro genomes (5 kb to 11.6 kb) displayed encapsulin domains. This was a surprising finding because the working hypothesis predicted that MCPs from micro genomes would be the most similar to encapsulins because they are predicted to form T=1 capsids, the most prevalent phenotype among encapsulins. Nonetheless, mining HK97-fold proteins from micro genomes had a much lower yield, about 3.01% (45 out of 1495 genomes), compared to the mini genome region, where 14.62% (491 out of 3358 genomes) displayed HK97-fold genes. The yield was even higher for the environmental control group above the isolated threshold, 28.61% (798 out of 2798 genomes). The fact that the inference methods based on HMMs and PLMs are based on isolated genomes and that MCPs in the twilight zone are expected to be more divergent, may explain why the yield of HK97-fold proteins decreases for smaller genomes. Thus, to further confirm the absence of encapsulin-domain profiles in micro genomes, it would be necessary to increase the sensitivity of the methods used to identify HK97-fold proteins by building, for example, on combining metagenomic data with AI methods. Additionally, the absence of HK97-fold proteins as genomes become smaller could be also attributed to the likely need for helper viruses or host elements for the replication cycle of the viral-like entities in the twilight zone, as suggested by the observation of satellite life styles among isolated tailed phages with relatively short genomes (12 to 20 kb) and HK97-type mobile elements (11 to 18) (Agarwal et al., 1990; deCarvalho et al., 2023; Penadés et al., 2025; Pourcel et al., 2024). The need for helper viral elements or unusual host conditions would explain the difficulty isolating the HK97-fold twilight viruses.

The sequence and structural analyses of encapsulin-enriched HK97-type proteins suggested twilight zone viral entities are spread across multiple clades, with some associated with

more ancestral branches, while others are in more derived groups (Figures 4A and 5). The most well-defined clade of encapsulin-enriched MCPs was observed among mini genomes in the twilight zone (Figure 4). This group was primarily from human gut samples with ~8 kb genomes containing lysogenic genes, indicating a likely temperate lifestyle. The group lacked portal or tail apparatus genes (Figures 4B and 4C). Assuming that this is a tailed-like phage entity, it would imply that this group requires a helper tailed phage or a host-related tail-like secretion system to pack and translocate the genome. However, the presence of the replicase RepA-like genes invites an alternative scenario. This gene is commonly found in the replication of ssDNA viruses and plasmids (de la Higuera et al., 2020; Kazlauskas et al., 2019), including the plasmid phase of HK97-type phages P4 and N15 (Ravin, 2015). In ssDNA viruses, the activity of RepA is either directly or indirectly associated with the packaging of the genome in the empty capsid through 3-fold or 5-fold pores without requiring the typical portal related to the packaging of dsDNA (Mankertz & Hillenbrand, 2001; Sonntag et al., 2011). This opens the possibility that this well-defined encapsulin-enriched clade could represent the first case of an HK97-type ssDNA. The discovery of ssDNA viruses with small genomes (< 11 kb) in a structural viral lineage (double-jelly fold) traditionally associated with dsDNA viruses (Kejzar et al., 2022; Krupovic et al., 2024; Laanto et al., 2017; San Martín & Van Raaij, 2018) further reinforces the possibility that the portal-less and tail-less HK97-type twilight viruses reported here may store their genome as ssDNA.

Encapsulin enriched MCPs reveal and entangled evolution of HK97-type compartments

The structural phylogenetic analysis of HK97-fold proteins combined with the co-occurrence network analysis of encapsulin-viral molecular domains revealed encapsulin groups deriving from viruses as well as viruses deriving from encapsulin groups, corresponding to the viral clade with

MCPs displaying the GP7 domain (Figure 6A). This finding suggests that modern HK97-type viruses and encapsulins may have evolved from each other multiple times, contrary to the current paradigm. The most divergent group of viruses and encapsulins was primarily associated with extremophiles and archaea and did not include any apparent phylogenetic transitions between encapsulins and viruses (Figure 6A). However, this group also included the least amount of twilight zone viral entities, probably due to the difficulty of sampling these regions. Since the presence of encapsulin-enriched twilight zone entities was key to establishing the entanglement in the other viral-encapsulin groups, it is hypothesized that increasing the sampling of twilight zone viruses among extremophile ecosystems will reveal encapsulin-enriched candidates linking these divergent groups.

The evolutionary entanglement observed between encapsulins and HK97-type viruses led to the hypothesis of a procapsid-like common ancestor (Figure 6A), supported by three observations. First, the early assembly stages of viral capsids in the Duplodnaviria realm (procapsids or proheads) structurally resemble encapsulins, both containing the protein forming the capsid shell and a cargo protein, the so-called scaffold protein in the case of viruses (Agirrezabala et al., 2007; Gertsman et al., 2009; Heymann et al., 2003; Tuma et al., 1998). Second, the structural comparison of the only reconstruction of the same T-number (T=4) for encapsulins and viruses demonstrated striking similarities between the encapsulin from Bacillus thermotolerans (PDB 6NJ8) and the procapsid (PDB 7RWZ) and mature capsid (PDB 8VD4) from the Staphylococcus aureus Pathogenicity Island 1 (SaPI1). The encapsulin and procapsid enclosed comparable volumes (~2.12×10^4 nm^3 vs. ~2.18×10^4 nm^3), corresponding to similar packaging capacities (~10.6 kb for encapsulin, ~10.9 kb for procapsid). The structural phenotype of the T=4 procapsid is quasi-identical to the encapsulin, although they are unrelated phylogenetically: the

iron storage encapsulin from Bacillus thermotolerans (PDB 6NJ8) has just 24.6% sequence identity to the viral HK97-fold protein, and the encapsulin's host belongs to a different family than the viral host. Importantly, both procapsids and encapsulins form porous shells compared to mature capsids (Chen et al., 2011; Hryc et al., 2017; Selivanovitch et al., 2021a; Wu et al., 2016; Giessen & Silver, 2017). The pores enable encapsulins to transport reactants (ions or other molecules) necessary for biochemical functions (Chmelyuk et al., 2023b; Ross et al., 2022; Williams et al., 2018), while in viral procapsids, they likely facilitate ion transport and scaffold exit during maturation (Chen et al., 2011; Ignatiou et al., 2019; Selivanovitch et al., 2021). However, the mature capsid volume (~3.51×10^4 nm^3) is substantially larger, accommodating the full genome length (~15.2 kb) (Figure 6B).

The pattern of procapsid volume insufficiency relative to genome size extends to larger viral systems with T=7 and T=13 architectures (Figure S3). The T=7 procapsids of Enterobacteria phages HK97 and λ exhibited genome packaging capacities of approximately 19.7 kb and 23.9 kb, respectively, insufficient for their actual genome sizes of about 39.7 kb and 48.5 kb. Similarly, the T=13 procapsid of Enterobacteria phage T4, with a packaging capacity of approximately 105 kb, could not accommodate its genome size of about 168.9 kb. Thus, procapsids consistently show volumes incompatible with their genome lengths, supporting their functional analogy to encapsulins rather than mature viral capsids.

Third, the encapsulin-enriched twilight viral entities exhibit strong lysogenic signatures, indicating potential mechanisms for genomic integration and host genome coexistence. Together, these observations indicate plausible evolutionary pathways facilitating transitions between encapsulin and viral compartments. Specifically, upon viral genome integration into a host genome, the procapsid scaffold protein could evolve catalytic capabilities or be replaced by

cargo-like proteins, converting the procapsid into a biochemical reactor. If the scaffold interaction with viral portal proteins is disrupted, the resulting shell loses genome packaging capability but retains a porous structure suitable for reactant transport and biochemical functionality, potentially becoming an encapsulin-like compartment (Figure 6B).

Conversely, recombination events reintroducing cargo-like proteins capable of interacting with portal complexes could rapidly restore genome packaging capabilities, transforming encapsulin-like compartments back into functional procapsids. This molecular evolutionary model makes the evolutionary back-and-forth between viral procapsids and cellular encapsulins plausible, depending on selection pressures (Figure 6). Furthermore, this model combined with encapsulin domains identified among isolated tailed phages associated with relatively large genomes, invites speculation about the existence of large encapsulins forming T>4 architectures capable of carrying biochemical reactions in significantly larger enclosed volumes. However, no encapsulin-like compartments larger than T=4 have yet been identified experimentally.

Although this structural and genomic analysis strongly suggests potential biochemical functions of viral procapsids, we emphasize that this hypothesis remains speculative. Experimental validation is essential to establish whether procapsids indeed perform biochemical roles during viral infection or when released into the environment.

If procapsids represent the molecular link between viruses and encapsulins, then it is plausible that viral procapsids could carry biochemical functions in nature during virus infection. The similar size and porous structure between procapsids and encapsulins would be conducive to the transport of reactants. The fact that procapsid-like particles have been repurposed biotechnologically as biochemical reactors further supports this function in nature (Selivanovitch

et al., 2021b; Wang & Douglas, 2022). A fraction of viral-like particles produced during viral infections form empty capsids (Laurenceau et al., 2020; Subramanian et al., 2025), even if only a fraction of particles produced, ~1%, where to carry biochemical reactions, given the sheer number of viral-like particles formed on Earth, ~10^{31}, this will imply that about 10^{30} procapsid-like particles could be carrying biochemical functions, which could influence substantially processes at planetary scales. A key step to investigate the potential biochemical role of viral procapsids would be to determine if the scaffold can carry biochemical reactions, or alternatively, if homologous proteins to the scaffold or encapsulin-cargo are present in viral genomes and are packed in procapsids to carry this biochemical role.

Conclusions

The discovery of HK97-type viruses in the twilight zone has filled an evolutionary gap, potentially revealing entanglement between the evolution of encapsulin and viruses and multiple transitions back and forth. The observation of procapsids representing the potential link between cellular and viral compartments opens the doors to a likely extant role of viral procapsids in biochemical processes, which could lead to a potential metabolic phase of viral capsids during infection or after being released to the environment, which would require a fundamental revision of the role of viruses on the planet.

Acknowledgments

The authors acknowledge Sean Benler for fruitful discussions throughout different stages of the project. A.A. and A.L. acknowledge the support from the Gordon and Betty Moore Foundation award GBMF9871 (https://doi.org/10.37807/GBMF9871). The work conducted by the U.S.

Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

References

Abrescia, N. G. A., Bamford, D. H., Grimes, J. M., & Stuart, D. I. (2012). Structure unifies the viral universe. *Annual Review of Biochemistry*, *81*(Volume 81, 2012), 795–822. https://doi.org/10.1146/ANNUREV-BIOCHEM-060910-095130/CITE/REFWORKS

Agarwal, M., Arthur, M., Arbeit, R. D., & Goldstein, R. (1990). Regulation of icosahedral virion capsid size by the in vivo activity of a cloned gene product. *Proceedings of the National Academy of Sciences*, *87*(7), 2428–2432. https://doi.org/10.1073/PNAS.87.7.2428

Agirrezabala, X., Velázquez-Muriel, J. A., Gómez-Puertas, P., Scheres, S. H. W., Carazo, J. M., & Carrascosa, J. L. (2007). Quasi-Atomic Model of Bacteriophage T7 Procapsid Shell: Insights into the Structure and Evolution of a Basic Fold. *Structure*, *15*(4), 461–472. https://doi.org/10.1016/j.str.2007.03.004

Aksyuk, A. A., Bowman, V. D., Kaufmann, B., Fields, C., Klose, T., Holdaway, H. A., Fischetti, V. A., & Rossmann, M. G. (2012). Structural investigations of a *Podoviridae streptococcus* phage C1, implications for the mechanism of viral entry. *Proceedings of the National Academy of Sciences*, *109*(35), 14001–14006. https://doi.org/10.1073/pnas.1207730109

Andreas, M. P., & Giessen, T. W. (2021). Large-scale computational discovery and analysis of virus-derived microbial nanocompartments. *Nature Communications*, *12*(1), 4748. https://doi.org/10.1038/s41467-021-25071-y

Balaban, M., Moshiri, N., Mai, U., Jia, X., & Mirarab, S. (2019). TreeCluster: Clustering biological sequences using phylogenetic trees. *PLOS ONE*, *14*(8), e0221068. https://doi.org/10.1371/journal.pone.0221068

Bárdy, P., Füzik, T., Hrebík, D., Pantůček, R., Thomas Beatty, J., & Plevka, P. (2020a). Structure and mechanism of DNA delivery of a gene transfer agent. *Nature Communications*, *11*(1), 3034. https://doi.org/10.1038/s41467-020-16669-9

Bárdy, P., Füzik, T., Hrebík, D., Pantůček, R., Thomas Beatty, J., & Plevka, P. (2020b). Structure and mechanism of DNA delivery of a gene transfer agent. *Nature Communications 2020 11:1*, *11*(1), 1–13. https://doi.org/10.1038/s41467-020-16669-9

Benler, S., Yutin, N., Antipov, D., Rayko, M., Shmakov, S., Gussow, A. B., Pevzner, P., & Koonin, E. V. (2021). Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome*, *9*(1), 78. https://doi.org/10.1186/s40168-021-01017-w

Berg, M., & Roux, S. (2021). Extreme dimensions — how big (or small) can tailed phages be? *Nature Reviews Microbiology 2021 19:7*, *19*(7), 407–407. https://doi.org/10.1038/s41579-021-00574-z

Casjens, S. R., & Gilcrease, E. B. (2009). Determining DNA Packaging Strategy by Analysis of the Termini of the Chromosomes in Tailed-Bacteriophage Virions. In *Bacteriophages. Methods in Molecular Biology™* (Vol. 502, pp. 91–111). https://doi.org/10.1007/978-1-60327-565-1_7

Caspar, D. L. D., & Klug, A. (1962). Physical Principles in the Construction of Regular Viruses. *Cold Spring Harbor Symposia on Quantitative Biology*, *27*(0), 1–24. https://doi.org/10.1101/SQB.1962.027.001.005

Chen, D. H., Baker, M. L., Hryc, C. F., DiMaio, F., Jakana, J., Wu, W., Dougherty, M., Haase-Pettingell, C., Schmid, M. F., Jiang, W., Baker, D., King, J. A., & Chiu, W. (2011). Structural basis for scaffolding-mediated assembly and maturation of a dsDNA virus. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(4), 1355–1360. https://doi.org/10.1073/PNAS.1015739108/SUPPL_FILE/SM03.MP4

Chmelyuk, N. S., Oda, V. V., Gabashvili, A. N., & Abakumov, M. A. (2023a). Encapsulins: Structure, Properties, and Biotechnological Applications. *Biochemistry (Moscow)*, *88*(1), 35–49. https://doi.org/10.1134/S0006297923010042

Chmelyuk, N. S., Oda, V. V., Gabashvili, A. N., & Abakumov, M. A. (2023b). Encapsulins: Structure, Properties, and Biotechnological Applications. *Biochemistry (Moscow) 2023 88:1*, *88*(1), 35–49. https://doi.org/10.1134/S0006297923010042

Choi, K. H., Morais, M. C., Anderson, D. L., & Rossmann, M. G. (2006). Determinants of Bacteriophage ϕ29 Head Morphology. *Structure*, *14*(11), 1723–1727. https://doi.org/10.1016/j.str.2006.09.007

Cobián Güemes, A. G., Youle, M., Cantú, V. A., Felts, B., Nulton, J., & Rohwer, F. (2016). Viruses as Winners in the Game of Life. *Annual Review of Virology*, *3*(1), 197–214. https://doi.org/10.1146/annurev-virology-100114-054952

Cui, J., Schlub, T. E., & Holmes, E. C. (2014). An Allometric Relationship between the Genome Length and Virion Volume of Viruses. *Journal of Virology*, *88*(11), 6403–6410. https://doi.org/10.1128/JVI.00362-14

Dai, X., & Hong Zhou, Z. (2018). Structure of the herpes simplex virus 1 capsid with associated tegument protein complexes. *Science (New York, N.Y.)*, *360*(6384). https://doi.org/10.1126/SCIENCE.AAO7298

de la Higuera, I., Kasun, G. W., Torrance, E. L., Pratt, A. A., Maluenda, A., Colombet, J., Bisseux, M., Ravet, V., Dayaram, A., Stainton, D., Kraberger, S., Zawar-Reza, P., Goldstien, S., Briskie, J. V., White, R., Taylor, H., Gomez, C., Ainley, D. G., Harding, J. S., … Stedman, K. M. (2020). Unveiling crucivirus diversity by mining metagenomic data. *MBio*, *11*(5), 1–17. https://doi.org/10.1128/MBIO.01410-20/SUPPL_FILE/MBIO.01410-20-S0001.DOCX

deCarvalho, T., Mascolo, E., Caruso, S. M., López-Pérez, J., Weston-Hafer, K., Shaffer, C., & Erill, I. (2023). Simultaneous entry as an adaptation to virulence in a novel satellite-helper system infecting Streptomyces species. *The ISME Journal*, *17*(12), 2381–2388. https://doi.org/10.1038/S41396-023-01548-0

Duda, R. L., & Teschke, C. M. (2019). The amazing HK97 fold: versatile results of modest differences. *Current Opinion in Virology*, *36*, 9–16. https://doi.org/10.1016/j.coviro.2019.02.001

Esterman, E. S., Wolf, Y. I., Kogay, R., Koonin, E. V, & Zhaxybayeva, O. (2021). Evolution of DNA packaging in gene transfer agents. *Virus Evolution*, *7*(1). https://doi.org/10.1093/ve/veab015

Fillol-Salom, A., Martínez-Rubio, R., Abdulrahman, R. F., Chen, J., Davies, R., & Penadés, J. R. (2018). Phage-inducible chromosomal islands are ubiquitous within the bacterial universe. *The ISME Journal*, *12*(9), 2114–2128. https://doi.org/10.1038/s41396-018-0156-3

Flamholz, Z. N., Biller, S. J., & Kelly, L. (2024). Large language models improve annotation of prokaryotic viral proteins. *Nature Microbiology*, *9*(2), 537–549. https://doi.org/10.1038/s41564-023-01584-8

Fokine, A., & Rossmann, M. G. (2014). Molecular architecture of tailed double-stranded DNA phages. *Bacteriophage*, *4*(2), e28281. https://doi.org/10.4161/bact.28281

Gertsman, I., Gan, L., Guttman, M., Lee, K., Speir, J. A., Duda, R. L., Hendrix, R. W., Komives, E. A., & Johnson, J. E. (2009). An unexpected twist in viral capsid maturation. *Nature*, *458*(7238), 646–650. https://doi.org/10.1038/nature07686

Giessen, T. W., & Silver, P. A. (2017). Widespread distribution of encapsulin nanocompartments reveals functional diversity. *Nature Microbiology*, *2*(6), 17029. https://doi.org/10.1038/nmicrobiol.2017.29

Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., Dimier, C., Domínguez-Huerta, G., Ferland, J., Kandels, S., Liu, Y., Marec, C., Pesant, S., Picheral, M., Pisarev, S., … Wincker, P. (2019). Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*, *177*(5), 1109-1123.e14. https://doi.org/10.1016/j.cell.2019.03.040

Gulyaeva, A., Garmaeva, S., Kurilshikov, A., Vich Vila, A., Riksen, N. P., Netea, M. G., Weersma, R. K., Fu, J., & Zhernakova, A. (2022). Diversity and ecology of *Caudoviricetes* phages with genome terminal repeats in fecal metagenomes from four Dutch cohorts. In *bioRxiv* (p.

2022.09.02.506393). Cold Spring Harbor Laboratory. https://doi.org/10.1101/2022.09.02.506393

Hagberg, A., Swart, P. J., & Schult, D. A. (2008). *Exploring network structure, dynamics, and function using NetworkX*.

Hardy, J. M., Dunstan, R. A., Grinter, R., Belousoff, M. J., Wang, J., Pickard, D., Venugopal, H., Dougan, G., Lithgow, T., & Coulibaly, F. (2020). The architecture and stabilisation of flagellotropic tailed bacteriophages. *Nature Communications*, *11*(1), 3748. https://doi.org/10.1038/s41467-020-17505-w

Hawkins, D. E. D. P., Bayfield, O. W., Fung, H. K. H., Grba, D. N., Huet, A., Conway, J. F., & Antson, A. A. (2023). Insights into a viral motor: the structure of the HK97 packaging termination assembly. *Nucleic Acids Research*, *51*(13), 7025. https://doi.org/10.1093/NAR/GKAD480

Heinemann, J., Maaty, W. S., Gauss, G. H., Akkaladevi, N., Brumfield, S. K., Rayaprolu, V., Young, M. J., Lawrence, C. M., & Bothner, B. (2011). Fossil record of an archaeal HK97-like provirus. *Virology*, *417*(2), 362–368. https://doi.org/10.1016/j.virol.2011.06.019

Helgstrand, C., Wikoff, W. R., Duda, R. L., Hendrix, R. W., Johnson, J. E., & Liljas, L. (2003). The Refined Structure of a Protein Catenane: The HK97 Bacteriophage Capsid at 3.44 Å Resolution. *Journal of Molecular Biology*, *334*(5), 885–899. https://doi.org/10.1016/J.JMB.2003.09.035

Heymann, J. B., Cheng, N., Newcomb, W. W., Trus, B. L., Brown, J. C., & Steven, A. C. (2003). Dynamics of herpes simplex virus capsid maturation visualized by time-lapse cryo-electron

microscopy. *Nature Structural & Molecular Biology*, *10*(5), 334–341. https://doi.org/10.1038/nsb922

Hua, J., Huet, A., Lopez, C. A., Toropova, K., Pope, W. H., Duda, R. L., Hendrix, R. W., & Conway, J. F. (2017). Capsids and Genomes of Jumbo-Sized Bacteriophages Reveal the Evolutionary Reach of the HK97 Fold. *MBio*, *8*(5). https://doi.org/10.1128/mBio.01579-17

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. https://doi.org/10.1186/1471-2105-11-119

Ignatiou, A., Brasilès, S., El Sadek Fadel, M., Bürger, J., Mielke, T., Topf, M., Tavares, P., & Orlova, E. V. (2019). Structural transitions during the scaffolding-driven assembly of a viral capsid. *Nature Communications 2019 10:1*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-12790-6

Ionel, A., Velázquez-Muriel, J. A., Luque, D., Cuervo, A., Castón, J. R., Valpuesta, J. M., Martín-Benito, J., & Carrascosa, J. L. (2011). Molecular rearrangements involved in the capsid shell maturation of bacteriophage T7. *The Journal of Biological Chemistry*, *286*(1), 234–242. https://doi.org/10.1074/JBC.M110.187211

Ivanova, N., Tringe, S. G., Liolios, K., Liu, W., Morrison, N., Hugenholtz, P., & Kyrpides, N. C. (2010). A call for standardized classification of metagenome projects. *Environmental Microbiology*, *12*(7), 1803–1805. https://doi.org/10.1111/j.1462-2920.2010.02270.x

Jones, J. A., Andreas, M. P., & Giessen, T. W. (2024). Structural basis for peroxidase encapsulation inside the encapsulin from the Gram-negative pathogen Klebsiella pneumoniae. *Nature Communications*, *15*(1), 2558. https://doi.org/10.1038/s41467-024-46880-x

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kashif-Khan, N., Savva, R., & Frank, S. (2024). Mining metagenomics data for novel bacterial nanocompartments. *NAR Genomics and Bioinformatics*, *6*(1). https://doi.org/https://doi.org/10.1093/nargab/lqae025

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Katoh, K., & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, *9*(4), 286–298. https://doi.org/10.1093/bib/bbn013

Kazlauskas, D., Varsani, A., Koonin, E. V., & Krupovic, M. (2019). Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nature Communications*, *10*(1), 3425. https://doi.org/10.1038/s41467-019-11433-0

Kejzar, N., Laanto, E., Rissanen, I., Abrishami, V., Selvaraj, M., Moineau, S., Ravantti, J., Sundberg, L. R., & Huiskonen, J. T. (2022). Cryo-EM structure of ssDNA bacteriophage ΦCjT23 provides insight into early virus evolution. *Nature Communications 2022 13:1*, *13*(1), 1–13. https://doi.org/10.1038/s41467-022-35123-6

Kieft, K., Zhou, Z., & Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, *8*(1), 90. https://doi.org/10.1186/s40168-020-00867-0

Koonin, E. V., Dolja, V. V., & Krupovic, M. (2022). The logic of virus evolution. *Cell Host & Microbe*, *30*(7), 917–929. https://doi.org/10.1016/j.chom.2022.06.008

Krupovic, M., Dolja, V. V., & Koonin, E. V. (2019). Origin of viruses: primordial replicators recruiting capsids from hosts. *Nature Reviews Microbiology*, *17*(7), 449–458. https://doi.org/10.1038/s41579-019-0205-6

Krupovic, M., & Koonin, E. V. (2017). Multiple origins of viral capsid proteins from cellular ancestors. *Proceedings of the National Academy of Sciences*, *114*(12), E2401–E2410. https://doi.org/10.1073/pnas.1621061114

Krupovic, M., Kuhn, J. H., Fischer, M. G., & Koonin, E. V. (2024). Natural history of eukaryotic DNA viruses with double jelly-roll major capsid proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(23), e2405771121. https://doi.org/10.1073/PNAS.2405771121/SUPPL_FILE/PNAS.2405771121.SAPP.PDF

Laanto, E., Mäntynen, S., De Colibus, L., Marjakangas, J., Gillum, A., Stuart, D. I., Ravantti, J. J., Huiskonen, J. T., & Sundberg, L. R. (2017). Virus found in a boreal lake links ssDNA and dsDNA viruses. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(31), 8378–8383. https://doi.org/10.1073/PNAS.1703834114/SUPPL_FILE/PNAS.201703834SI.PDF

Lander, G. C., Johnson, J. E., Rau, D. C., Potter, C. S., Carragher, B., & Evilevitch, A. (2013). DNA bending-induced phase transition of encapsidated genome in phage λ. *Nucleic Acids Research*, *41*(8), 4518–4524. https://doi.org/10.1093/NAR/GKT137

Laurenceau, R., Raho, N., Forget, M., Arellano, A. A., & Chisholm, S. W. (2020). Frequency of mispackaging of Prochlorococcus DNA by cyanophage. *The ISME Journal 2020 15:1*, *15*(1), 129–140. https://doi.org/10.1038/s41396-020-00766-0

Lee, D. Y., Bartels, C., McNair, K., Edwards, R. A., Swairjo, M. A., & Luque, A. (2022). Predicting the capsid architecture of phages from metagenomic data. *Computational and Structural Biotechnology Journal*, *20*, 721–732. https://doi.org/10.1016/j.csbj.2021.12.032

Letunic, I., & Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research*, *52*(W1), W78–W82. https://doi.org/10.1093/nar/gkae268

Luque, A., Benler, S., Lee, D. Y., Brown, C., & White, S. (2020). The Missing Tailed Phages: Prediction of Small Capsid Candidates. *Microorganisms*, *8*(12), 1944. https://doi.org/10.3390/microorganisms8121944

Luque, A., & Reguera, D. (2010). The Structure of Elongated Viral Capsids. *Biophysical Journal*, *98*(12), 2993–3003. https://doi.org/10.1016/j.bpj.2010.02.051

Mankertz, A., & Hillenbrand, B. (2001). Replication of porcine circovirus type 1 requires two proteins encoded by the viral rep gene. *Virology*, *279*(2), 429–438. https://doi.org/10.1006/VIRO.2000.0730

McHugh, C. A., Fontana, J., Nemecek, D., Cheng, N., Aksyuk, A. A., Heymann, J. B., Winkler, D. C., Lam, A. S., Wall, J. S., Steven, A. C., & Hoiczyk, E. (2014). A virus capsid-like

nanocompartment that stores iron and protects bacteria from oxidative stress. *The EMBO Journal*, *33*(17), 1896–1911. https://doi.org/10.15252/embj.201488566

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., & Dessimoz, C. (2023). Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. In *bioRxiv*. https://doi.org/10.1101/2023.09.19.558401

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology*, *39*(5), 578–585. https://doi.org/10.1038/s41587-020-00774-7

Oh, B., Moyer, C. L., Hendrix, R. W., & Duda, R. L. (2014). The delta domain of the HK97 major capsid protein is essential for assembly. *Virology*, *456–457*(1), 171–178. https://doi.org/10.1016/J.VIROL.2014.03.022

Olsen, N. S., Hendriksen, N. B., Hansen, L. H., & Kot, W. (2020). A New High-Throughput Screening Method for Phages: Enabling Crude Isolation and Fast Identification of Diverse Phages with Therapeutic Potential. *PHAGE*, *1*(3), 137–148. https://doi.org/10.1089/phage.2020.0016

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., & Kyrpides, N. C. (2016). Uncovering Earth's virome. *Nature*, *536*(7617), 425–430. https://doi.org/10.1038/nature19094

Paez-Espino, D., Roux, S., Chen, I.-M. A., Palaniappan, K., Ratner, A., Chu, K., Huntemann, M., Reddy, T. B. K., Pons, J. C., Llabrés, M., Eloe-Fadrosh, E. A., Ivanova, N. N., & Kyrpides, N. C. (2019). IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Research*, *47*(D1), D678–D686. https://doi.org/10.1093/nar/gky1127

Penadés, J. R., & Christie, G. E. (2015). The Phage-Inducible Chromosomal Islands: A Family of Highly Evolved Molecular Parasites. *Annual Review of Virology*, *2*(1), 181–201. https://doi.org/10.1146/annurev-virology-031413-085446

Penadés, J. R., Seed, K. D., Chen, J., Bikard, D., & Rocha, E. P. C. (2025). Genetics, ecology and evolution of phage satellites. *Nature Reviews Microbiology 2025*, 1–13. https://doi.org/10.1038/s41579-025-01156-z

Podgorski, J. M., Podgorski, J., Abad, L., Jacobs-Sera, D., Freeman, K. G., Brown, C., Hatfull, G. F., Luque, A., & White, S. J. (2025). Stabilization mechanism accommodating genome length variation in evolutionarily related viral capsids. *Nature Communications 2025 16:1*, *16*(1), 1–10. https://doi.org/10.1038/s41467-025-58298-0

Podgorski, J. M., Podgorski, J., Abad, L., Jacobs-Sera, D., Freeman, K. G., Brown, C., Hatfull, G., Luque, A., & White, S. J. (2023). A novel stabilization mechanism accommodating genome length variation in evolutionarily related viral capsids. *BioRxiv*, 2023.11.03.565530. https://doi.org/10.1101/2023.11.03.565530

Pourcel, C., Essoh, C., Ouldali, M., & Tavares, P. (2024). *Acinetobacter baumannii* satellite phage Aci01-2-Phanie depends on a helper myophage for its multiplication. *Journal of Virology*, *98*(7). https://doi.org/10.1128/jvi.00667-24

Rao, V. B., Fokine, A., & Fang, Q. (2021). The Remarkable Vial Portal Vertex: structure and a plausible model for mechanism. *Current Opinion in Virology*, *51*, 65. https://doi.org/10.1016/J.COVIRO.2021.09.004

Ravin, N. V. (2015). Replication and Maintenance of Linear Phage-Plasmid N15. *Microbiology Spectrum*, *3*(1). https://doi.org/10.1128/MICROBIOLSPEC.PLAS-0032-2014

Roos, W. H., Gertsman, I., May, E. R., Brooks, C. L., Johnson, J. E., & Wuite, G. J. L. (2012). Mechanics of bacteriophage maturation. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(7), 2342–2347. https://doi.org/10.1073/PNAS.1109590109/SUPPL_FILE/PNAS.1109590109_SI.PDF

Ross, J., McIver, Z., Lambert, T., Piergentili, C., Bird, J. E., Gallagher, K. J., Cruickshank, F. L., James, P., Zarazúa-Arvizu, E., Horsfall, L. E., Waldron, K. J., Wilson, M. D., Logan Mackay, C., Baslé, A., Clarke, D. J., & Marles-Wright, J. (2022). Pore dynamics and asymmetric cargo loading in an encapsulin nanocompartment. *Science Advances*, *8*(4), 4461. https://doi.org/10.1126/SCIADV.ABJ4461/SUPPL_FILE/SCIADV.ABJ4461_MOVIES_S1_TO_S3.ZIP

San Martín, C., & Van Raaij, M. J. (2018). The so far farthest reaches of the double jelly roll capsid protein fold 06 Biological Sciences 0601 Biochemistry and Cell Biology. *Virology Journal*, *15*(1), 1–6. https://doi.org/10.1186/S12985-018-1097-1/TABLES/1

Selivanovitch, E., LaFrance, B., & Douglas, T. (2021). Molecular exclusion limits for diffusion across a porous capsid. *Nature Communications 2021 12:1*, *12*(1), 1–12. https://doi.org/10.1038/s41467-021-23200-1

Sonntag, F., Köther, K., Schmidt, K., Weghofer, M., Raupp, C., Nieto, K., Kuck, A., Gerlach, B., Böttcher, B., Müller, O. J., Lux, K., Hörer, M., & Kleinschmidt, J. A. (2011). The Assembly-Activating Protein Promotes Capsid Assembly of Different Adeno-Associated Virus Serotypes. *Journal of Virology*, *85*(23), 12686–12697. https://doi.org/10.1128/JVI.05359-11/ASSET/9822EA7F-6689-4755-9978-884B6872AD15/ASSETS/GRAPHIC/ZJV9990952780010.JPEG

Stone, N. P., Demo, G., Agnello, E., & Kelch, B. A. (2019). Principles for enhancing virus capsid capacity and stability from a thermophilic virus capsid structure. *Nature Communications*, *10*(1), 4471. https://doi.org/10.1038/s41467-019-12341-z

Subramanian, S., Kerns, H. R., Braverman, S. G., & Doore, S. M. (2025). The structure of Shigella virus Sf14 reveals the presence of two decoration proteins and two long tail fibers. *Communications Biology 2025 8:1*, *8*(1), 1–11. https://doi.org/10.1038/s42003-025-07668-x

Suhanovsky, M. M., & Teschke, C. M. (2015). Nature′s favorite building block: Deciphering folding and capsid assembly of proteins with the HK97-fold. *Virology*, *479–480*, 487–497. https://doi.org/10.1016/j.virol.2015.02.055

Sutter, M., Boehringer, D., Gutmann, S., Günther, S., Prangishvili, D., Loessner, M. J., Stetter, K. O., Weber-Ban, E., & Ban, N. (2008). Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nature Structural & Molecular Biology*, *15*(9), 939–947. https://doi.org/10.1038/nsmb.1473

Tao, Y., Olson, N. H., Xu, W., Anderson, D. L., Rossmann, M. G., & Baker, T. S. (1998). Assembly of a Tailed Bacterial Virus and Its Genome Release Studied in Three Dimensions. *Cell*, *95*(3), 431–437. https://doi.org/10.1016/S0092-8674(00)81773-0

Terzian, P., Olo Ndela, E., Galiez, C., Lossouarn, J., Pérez Bucio, R. E., Mom, R., Toussaint, A., Petit, M.-A., & Enault, F. (2021). PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genomics and Bioinformatics*, *3*(3). https://doi.org/10.1093/nargab/lqab067

Tisza, M. J., & Buck, C. B. (2021). A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proceedings of the National Academy of Sciences*, *118*(23), e2023202118. https://doi.org/10.1073/pnas.2023202118

Tuma, R., Prevelige, P. E., & Thomas, G. J. (1998). Mechanism of capsid maturation in a double-stranded DNA virus. *Proceedings of the National Academy of Sciences*, *95*(17), 9885–9890. https://doi.org/10.1073/pnas.95.17.9885

Twarock, R., & Luque, A. (2019). Structural puzzles in virology solved with an overarching icosahedral design principle. *Nature Communications*, *10*(1), 4414. https://doi.org/10.1038/s41467-019-12367-3

Veesler, D., & Johnson, J. E. (2012). Virus maturation. *Annual Review of Biophysics*, *41*(1), 473–496. https://doi.org/10.1146/ANNUREV-BIOPHYS-042910-155407/CITE/REFWORKS

Wigington, C. H., Sonderegger, D., Brussaard, C. P. D., Buchan, A., Finke, J. F., Fuhrman, J. A., Lennon, J. T., Middelboe, M., Suttle, C. A., Stock, C., Wilson, W. H., Wommack, K. E., Wilhelm, S. W., & Weitz, J. S. (2016). Re-examination of the relationship between marine virus and microbial cell abundances. *Nature Microbiology*, *1*(3), 15024. https://doi.org/10.1038/nmicrobiol.2015.24

Wikoff, W. R., Conway, J. F., Tang, J., Lee, K. K., Gan, L., Cheng, N., Duda, R. L., Hendrix, R. W., Steven, A. C., & Johnson, J. E. (2006). Time-resolved molecular dynamics of

bacteriophage HK97 capsid maturation interpreted by electron cryo-microscopy and X-ray crystallography. *Journal of Structural Biology*, *153*(3), 300–306. https://doi.org/10.1016/J.JSB.2005.11.009

Wikoff, W. R., Liljas, L., Duda, R. L., Tsuruta, H., Hendrix, R. W., & Johnson, J. E. (2000). Topologically Linked Protein Rings in the Bacteriophage HK97 Capsid. *Science*, *289*(5487), 2129–2133. https://doi.org/10.1126/SCIENCE.289.5487.2129

Williams, E. M., Jung, S. M., Coffman, J. L., & Lutz, S. (2018). Pore Engineering for Enhanced Mass Transport in Encapsulin Nanocompartments. *ACS Synthetic Biology*, *7*(11), 2514–2517. https://doi.org/10.1021/ACSSYNBIO.8B00295

Wommack, K. E., & Colwell, R. R. (2000). Virioplankton: Viruses in Aquatic Ecosystems. *Microbiology and Molecular Biology Reviews*, *64*(1), 69–114. https://doi.org/10.1128/MMBR.64.1.69-114.2000

Xu, J., Wang, D., Gui, M., & Xiang, Y. (2019). Structural assembly of the tailed bacteriophage ϕ29. *Nature Communications 2019 10:1*, *10*(1), 1–16. https://doi.org/10.1038/s41467-019-10272-3

Yap, M. L., & Rossmann, M. G. (2014). Structure and function of bacteriophage T4. *Future Microbiology*, *9*(12), 1319–1337. https://doi.org/10.2217/FMB.14.91

Zayed, A. A., Lücking, D., Mohssen, M., Cronin, D., Bolduc, B., Gregory, A. C., Hargreaves, K. R., Piehowski, P. D., White III, R. A., Huang, E. L., Adkins, J. N., Roux, S., Moraru, C., & Sullivan, M. B. (2021). efam: an *e* xpanded, metaproteome-supported HMM profile database of viral protein *fam* ilies. *Bioinformatics*, *37*(22), 4202–4208. https://doi.org/10.1093/bioinformatics/btab451

Figures

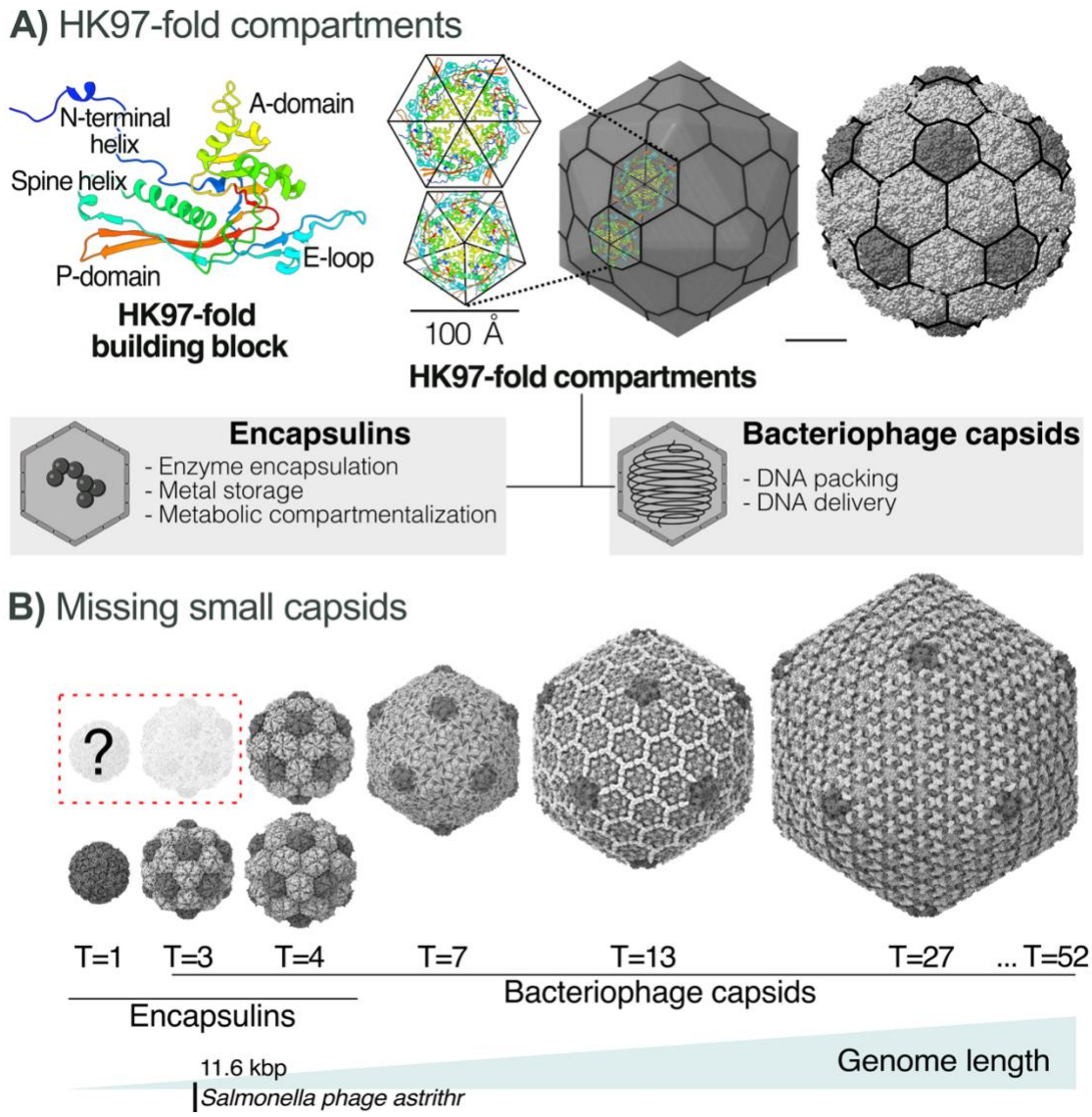

Figure 1. HK97-fold compartments composition and missing tailed phages with predicted small capsids. A) Schematic representation of the HK97-fold building block and its assembly into 3D compartments. The HK97 fold consists of the spine helix, P-domain, A-domain, and E-loop, which together form a unit for capsid assembly. Two major types of HK97-fold compartments are shown: Encapsulins, involved in enzyme encapsulation, metal storage, and metabolic compartmentalization; and bacteriophage capsids, specialized for DNA packaging and delivery. B) Size range of HK97-fold compartments showcasing geometric capsid architectures (T-numbers) ranging from encapsulins (T=1 to T=4) to larger bacteriophage capsids (T=4 to T=52). The dashed red box indicates missing small capsid intermediates (T=1 and T=3), which may exist but have not yet been characterized. The genome size gradient highlights the relationship between capsid size and genomic content, including the smallest phage known; *Salmonella phage astrithr* (11.6 kb) as a reference.

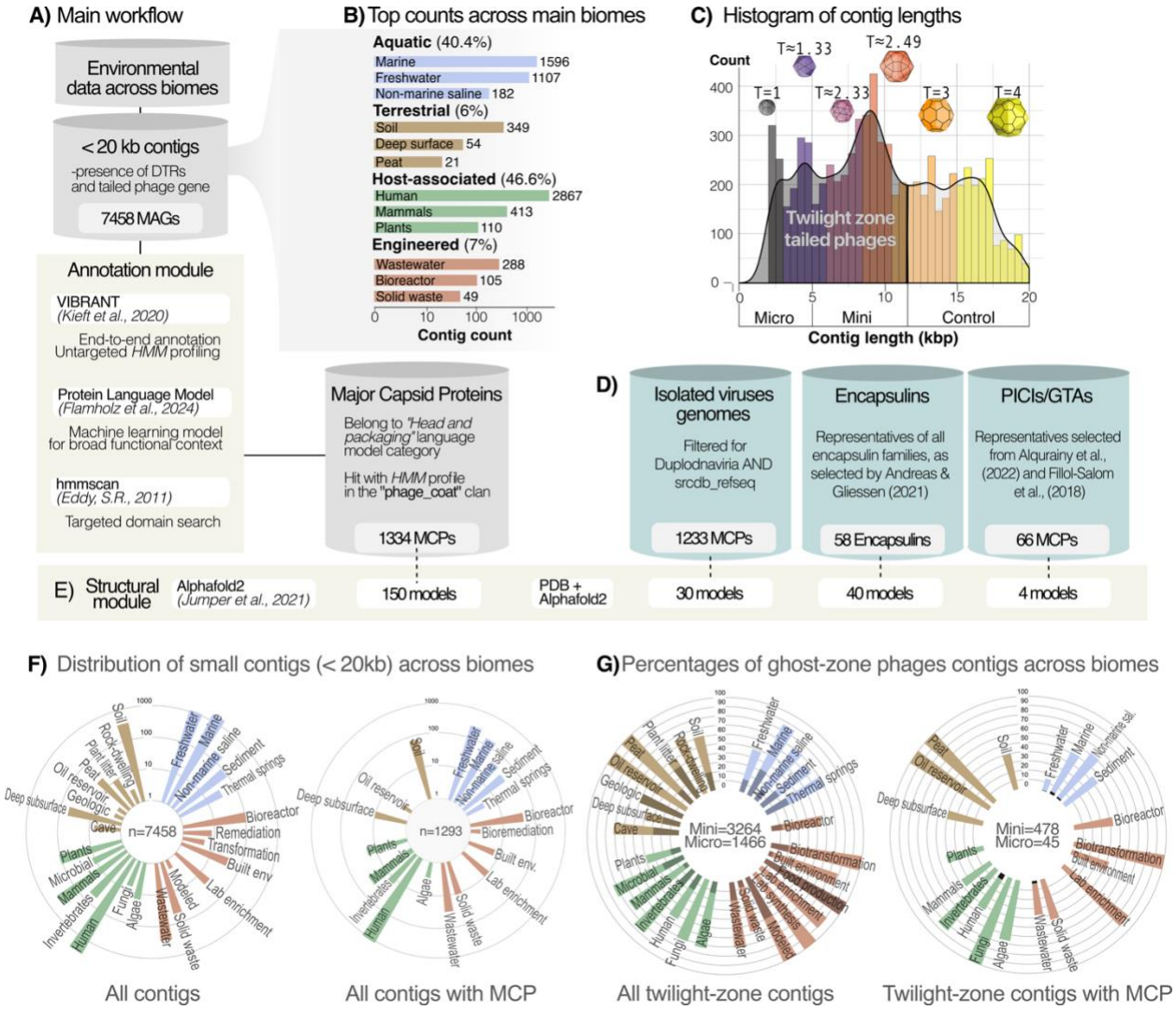

Figure 2. Discovery and comparison pipeline for twilight zone viruses. A) Flowchart illustrating the process from data acquisition to structural prediction. It is divided into three steps: filtering conditions, annotation, and major capsid protein identification. B) The bar plot displays the number of contigs (log scale) contributing from the top three biomes category (log scale) in each Level 1 category: Aquatic (blue), terrestrial (brown), host-associated (green), and engineered (red). C) Histogram of contig lengths. The solid line is a kernel density distribution with a bandwidth of 700. The colors are associated with different icosahedral capsids with the structures and T-numbers displayed on top. The vertical line represents the genome length threshold associated with the smallest isolated tailed phage genome, 11.6 kb, and the region below this value (the twilight zone) is shaded and includes mini (5 kb to 11.6 kb) and micro (< 5 kb) contigs. D) Overview of external datasets used for comparative analyses containing major capsid proteins (MCPs). E) Structural module indicating the number of structures predicted with AlphaFold2 or obtained from the Protein Data Bank (PDB). F) Circular bar plot representing the distribution of viral contigs across biomes, including all filtered contigs (left) and those with annotated MCP (right). The center includes the total number of contigs displayed (n). G) Similar plot to F but focuses on the Twilight Zone region. The center displays the number of mini and micro contigs.

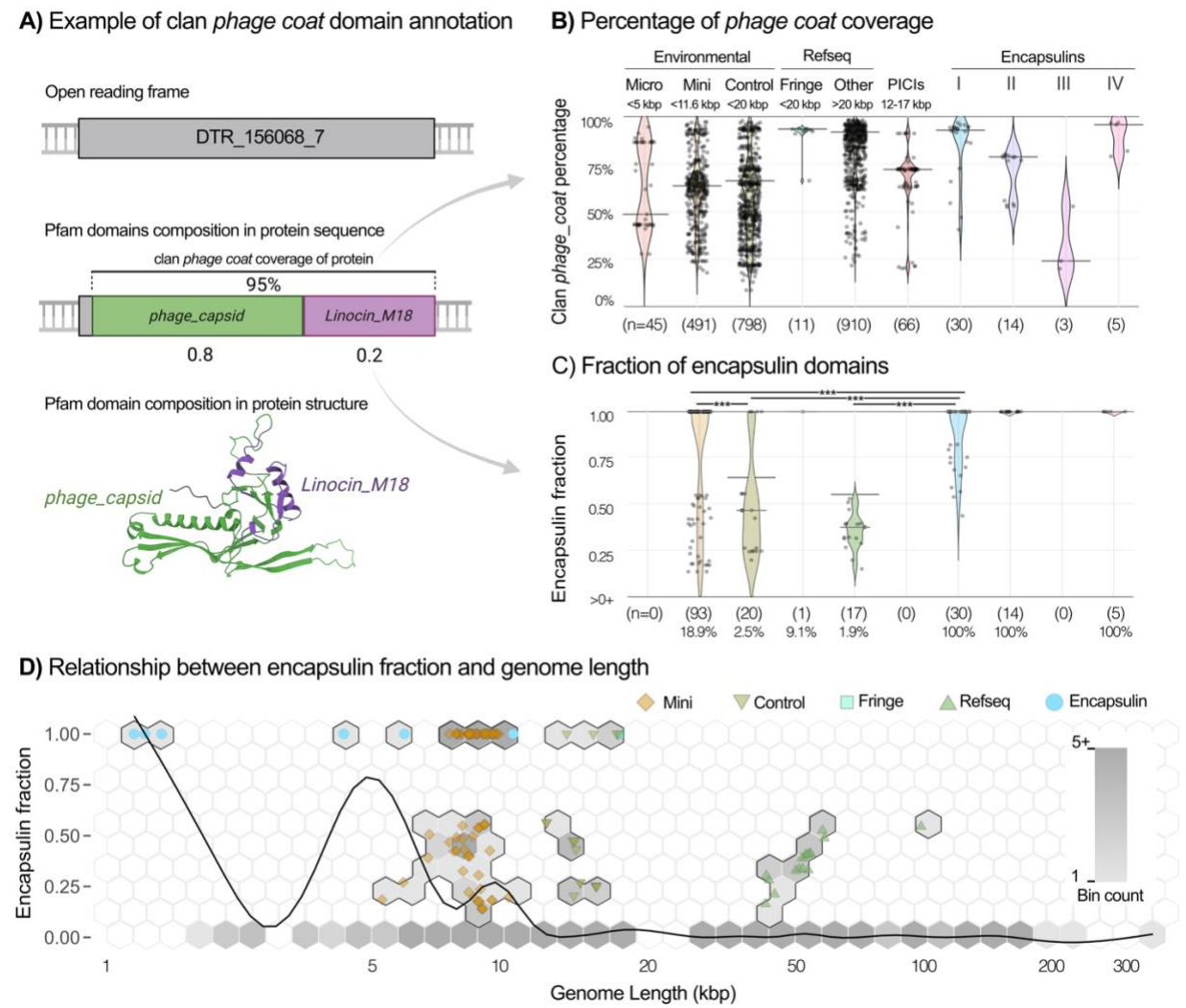

Figure 3. Encapsulin domain analysis across viral entities. A) Schematic representation of the major capsid protein domain analysis. The open reading of an annotated major capsid protein is displayed in grey. The protein domains identified are displayed in different colors, indicating the names of the specific family domains. The black horizontal bar covers the residues captured by the combined phage_coat clan domains, and the percentage of coverage is displayed. The numbers under the domains indicate the fraction of each domain. The predicted 3D protein structure is colored based on the sequence associated with each phage_coat domain identified. B) Percentage of phage_coat coverage plotted for the MCPs for viral and encapsulin groups. The black dots are the values for each representative. The horizontal bars represent the median. Each violin plot is displayed in a different color based on the group, with names on top. The number of members (n) in each group is displayed in parentheses at the bottom. C) Fraction of the encapsulin domains observed in the major capsid proteins for each group, following the same logic as in panel B). The symbol >0% in the y-axis indicates that only MCPs with an encapsulin domain fraction above zero are displayed. The horizontal bars indicate the groups that were compared statistically, displaying stars if they were different statistically. D) Scatterplot of encapsulin fraction versus genome length for those elements with an encapsulin signal (>0%). Each group is represented with a different symbol (see legend). The colors correspond to the same schema as in B) and C). For the encapsulins, the projected genome length corresponds to the calculated packaging capacity. The hexagonal net is a hexbin plot showing the density of data points on a grey scale, as indicated by the color bar. It includes the members that contain no encapsulin domain signal.

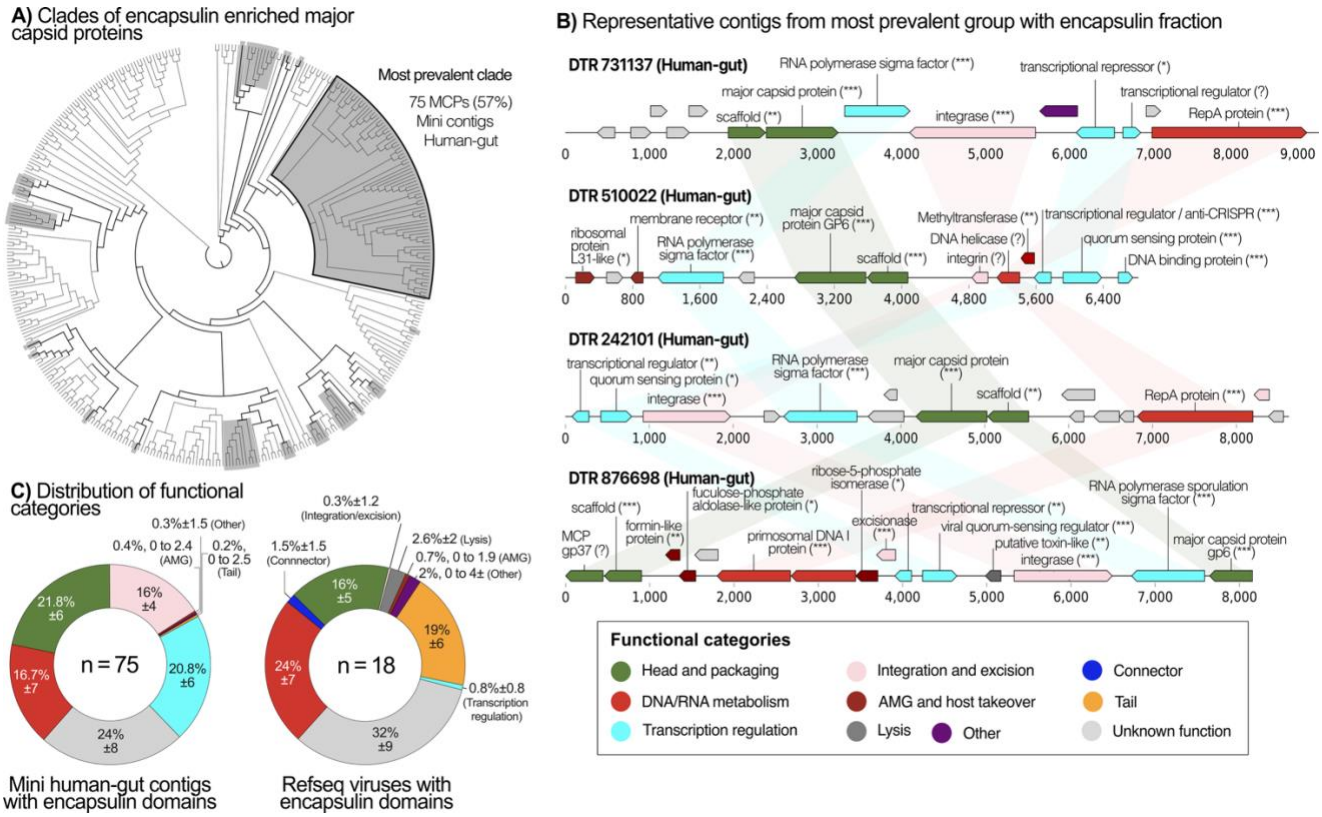

Figure 4. Genome composition of encapsulin-enriched mini viral entities. A) Gene maps of representative genomes from the encapsulin-enriched mini virus group. The horizontal line represents the putative genome, and the numbers are the length in bp. The open reading frames are displayed as thick-colored arrows. The colors correspond to high-level genome functions listed in the legend on the top right of the figure. The genes annotated successfully include the predicted function as a label with its statistical significance in parenthesis (see methods). Each genome also includes the identifier (DTR), all of them come from human-gut samples. B) Distribution of functional categories for the encapsulin-enriched mini viral group (left) and the encapsulin-enriched representatives from RefSeq (right)among viruses with encapsulin fractions. The colors are related to the functional categories as in panel A). The sections of the pies include the average percentages.

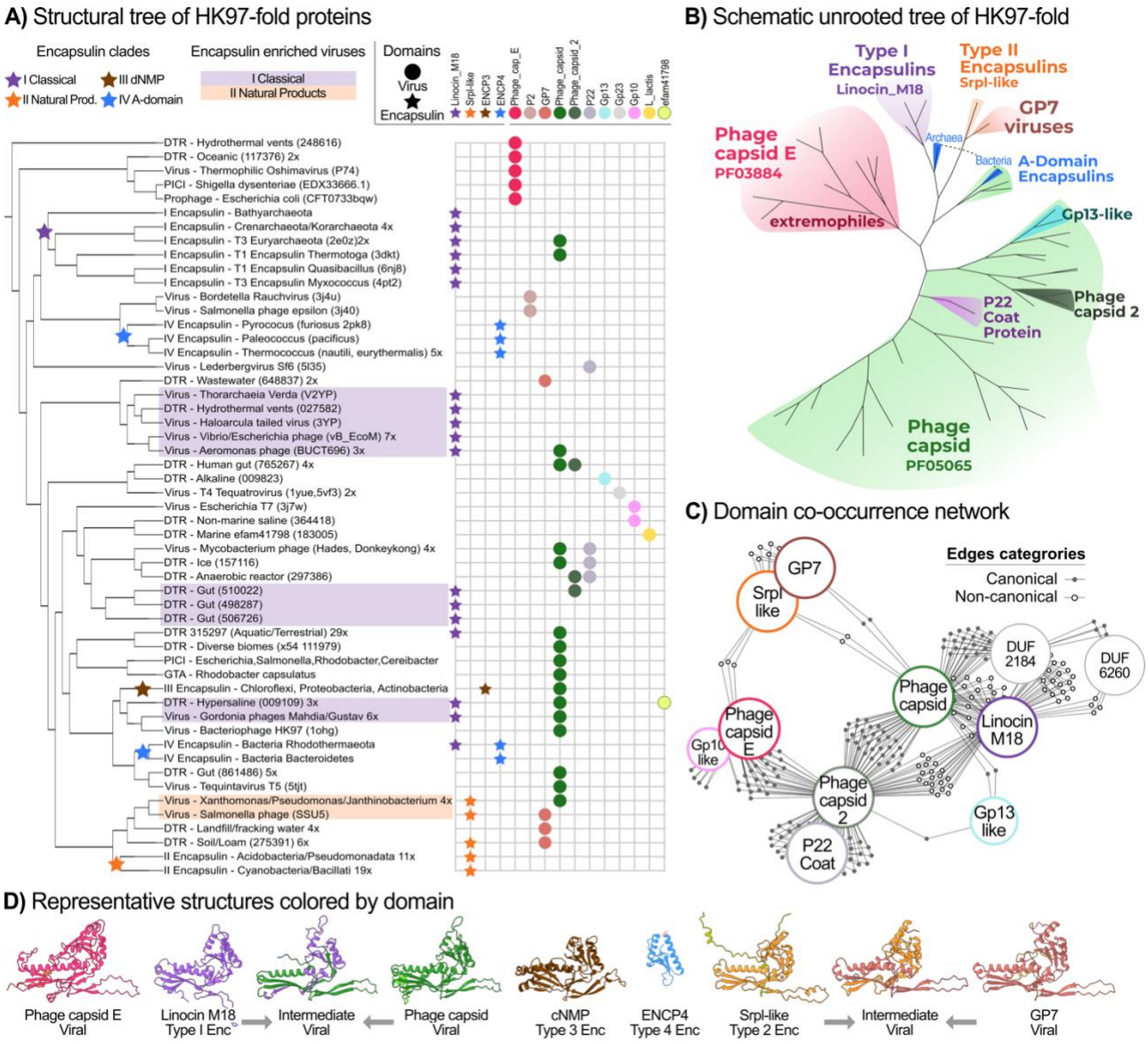

Figure 5. Structural phylogenetics and domain co-occurrence network. A) Rooted structural phylogeny. The tree compares HK97-fold structures, including encapsulins, twilight-zone viruses (G), environmental viral control (C), and RefSeq representatives (R). Each leave includes the associated Pfam domains from phage_coat to the right based on the color coding displayed in the legend on top. Stars correspond to encapsulin domains and circles to viral domains. A folded representative of each Pfam protein domain is included to the right. The shaded leaves highlight viral groups containing encapsulin domains. B) Unrooted structural phylogeny. The colors and labels are associated with the Pfam domains. C) Co-occurrence network of HK97-fold protein domains. The nodes correspond to the Pfam domains with the color coding based on the legend in A). The size of the circles corresponds to the number of edges of a node. Each edge joining two nodes implies an occurrence of those two protein domains found in the same major capsid protein. The symbol in the middle point of an edge indicates if the co-occurrence was canonical, that is, viral-and-viral or encapsulin-and-encapsulin domains (closed gray circles) or non-canonical, that is, viral-and-encapsulin domains (opened circles). D) Representative structures per lineage and intermediate viral MCPs enriched with encapsulin domains.

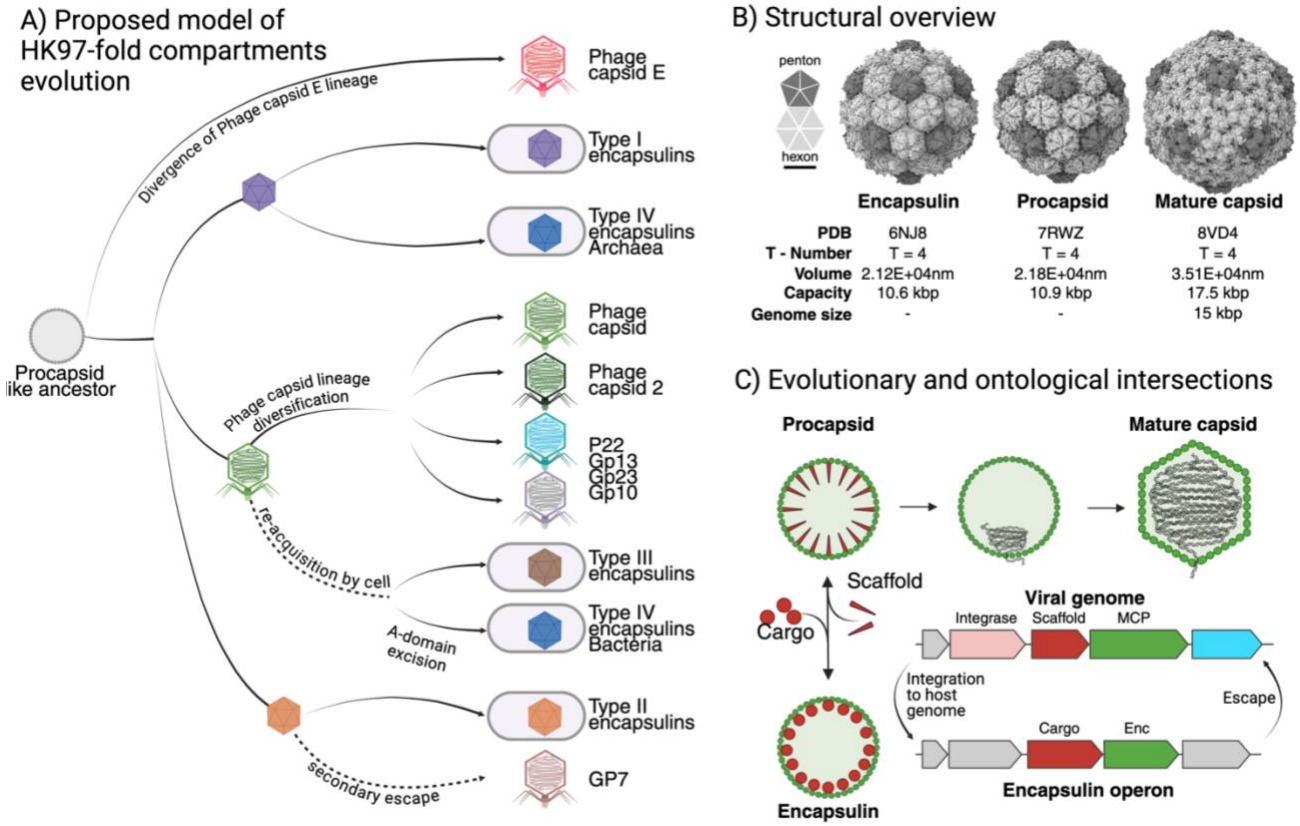

Figure 6. Schematic model of HK97 fold protein evolution. A) HK97 fold evolutionary model enhanced with twilight-zone phages. This model posits non-canonical transitions between viral MCPs and encapsulins. It suggests that encapsulins did not emerge as a single monophyletic group but rather through multiple independent transitions from viral ancestors, and vice versa. The model also assumes a procapsid-like ancestor, linking the biochemical and genetic delivery roles of HK97-fold proteins. B) Evolutionary and ontological intersections between viral MCPs and encapsulins. This panel explores the parallel between scaffolding proteins in viral capsids and cargo proteins in encapsulins, suggesting that both play a functionally equivalent role in determining the final assembly and purpose of the compartment. The procapsid, an immature and empty stage of viral capsids, could represent a bifurcation point in HK97-fold evolution, where modifications to its function led to the divergence of viral capsids and encapsulins. C) Structural comparison of HK97-fold compartments. The structural resemblance between encapsulins (PDB: 6NJ8) and viral procapsids (PDB: 7RWZ) is highlighted. Additionally, the comparison emphasizes a key constraint in viral evolution: at the procapsid stage, the internal volume is insufficient to accommodate the genome that the mature capsid will later encapsulate, necessitating further structural rearrangements.