UNIVERSIDAD AUTÓNOMA DE MADRID

FACULTAD DE MEDICINA

TRABAJO FIN DE MÁSTER

Identification, Annotation and Structural Characterisation of Viral Terminal Proteins

Máster Universitario en Bioinformática y Biología Computacional

Autor: RAMÍREZ MONTÁNS, Juan Carlos

Director: REDREJO RODRÍGUEZ, Modesto
Departamento de BIOQUÍMICA

CURSO: 2024/2025
FECHA: mayo, 2025

Index

1. Introduction 2
1.1. Molecular insights of protein-primed DNA replication 2
1.2. Inherent hindrances for TP identification and annotation 4
1.3. Bioinformatic approaches on pPolB and DNA TP identification 5
1.4. Structure comparison versus sequence comparison 6
2. Motivation and objectives 7
3. Methods 7
3.1. Pre-dataset selection and structural modelling 7
3.1.1. Pre-dataset construction – GenBank and state-of-the-art references from PubMed 7
3.1.2. TP structural prediction/modelling 8
3.2. Ascertainment of standardizable discriminatory-definitory criteria for TP nature assessment 8
3.2.1. Surface charge density distribution asymmetry 8
3.2.2. Secondary structure 9
3.2.3. DNA binding probability 10
3.4. Homology search 10
3.4.1. Foldseek structure-based homology search 10
3.4.2. PSI-BLAST sequence-based homology search 11
3.4.3. BLASTp sequence-based homology search 12
3.4.4. Taxonomic in-depth scrutiny 12
3.5. Phylogenetic assessment 12
3.5.1. Sequence-based Clustal Omega (MSA) 12
3.5.2. Structure-based FoldMason (MSTA) 13
3.6. Clustering 13
3.6.1. Sequence-based – CD-Hit 13
3.6.2. Structure-based – qTMcluster 13
3.6.3. Sequence-based – BLAST2seq, k-Means and PCA 13
3.7. Genomic scrutiny 14
3.7.1. IPG-based genomic retrieval 14
3.7.2. PHROG-based proteome annotation 14
3.7.3. geNomad-based genome annotation 15
3.7.4. Viral-trace detection and non-viral set 16
3.8. eggNOG-mapper functional inference 16
3.9. TP:pPolB heterodimer inspection 16
3.9.1. Structural modelling 16
3.9.2. pPolB inspection 17
4. Results 17
4.1. Predataset selection and starter dataset shortlisting 17
4.2. Initial dataset phylogenetic reconstruction 18
4.3. The analysis of TP structures enables the validation of intrinsic features of TP 20
4.4. Sequence- and structure-based homology searches yield complementary results 25
4.5. Structure-based clustering (qTMcluster) 28
4.6. Functional inference (eggNOG-mapper) 31
4.7. Genome annotation and non-viral *contig* detection 32
4.8. Carbohydrate-metabolising and DNA-processing enzymes may be origins of TP 33
5. Discussion 36
6. Conclusions 41
7. Acknowledgements 42
8. Bibliography 42
9. Annex I. GitHub repository 47
10. Annex II. Supplementary figures 47

1. Introduction

1.1. Molecular insights of protein-primed DNA replication

According to the Central Dogma of Molecular Biology, five core molecular processes underpin life: DNA replication, RNA replication, transcription, reverse transcription, and translation. Among these, three –DNA replication, transcription, and translation– have traditionally been viewed as universal, fundamental, and largely invariant, with DNA replication historically considered the most conserved and unchanging. However, recent discoveries challenge this orthodoxy. Certain viruses –characterized by linear double-stranded DNA (dsDNA) genomes of 10-40kbp– along with related mobile genetic elements (MGE), exhibit non-canonical, largely uncharacterized mechanisms of genome replication.[1,2] Examples include dsDNA viruses such as bacteriophages (infecting both Gram-positive and Gram-negative bacteria), archaeal viruses, virophages (notably those in the order *Lavidavirales*), and members of the family *Adenoviridae*.[1,2] These viruses, which typically have icosahedral capsids and double jelly-roll major capsid proteins, replicate their genomes through atypical processes. Related non-viral MGE –such as polintons (*Polintoviricetes*), polinton-like viruses (PLV), casposons, transpovirons, and certain linear cytoplasmic plasmids found in fungi and eukaryotic organelles– employ similarly unconventional strategies.[1,2] According to the February 2025 update of the International Committee on Taxonomy of Viruses (ICTV), most of these replicons fall within the realms *Varidnaviria* (phylum Preplasmiviricota of the kingdom *Bamfordvirae*; *phylum Produgelaviricota* of the kingdom *Abadenavirae*) and *Duplodnaviria* (*phylum Uroviricota*, kingdom *Heunggongvirae*).[1–5] This exceptional phylogenetic and functional diversity presents both an opportunity to uncover novel molecular mechanisms and a significant challenge to the study of unorthodox DNA replication.

During DNA replication, DNA polymerases work on both DNA strands copying the template in 5'→3' direction, according to the semi-continuous canonical core cellular DNA replication, takes place in an asymmetrical manner, where (as both DNA strands are being replicated simultaneously) 5'→3' strand (or leading strand) is usually continuously polymerised without any interruption (and setting the direction of the replication fork), 3'→5' strand (antisense or lagging strand) replication is based in the generation of short fragments (Okazaki fragments) that allow replication to advance against fork's direction.[2,6–8] Remarkably, the 5'→3' polymerase activity is generally incompatible with a *sensu stricto* DNA priming or de novo DNA synthesis, with very few exceptions,[7,9] as most DNA polymerases are not able to start or initiate DNA polymerisation straightforwardly on the first position due to the need for a required pre-existing priming oligonucleotide (primer) responsible for enabling DNA polymerase-DNA binding by providing a free (3'-) hydroxyl group add new nucleotides to the growing chain. The primer is generally short ribooligonucleotide of around ten bases usually generated by DNA primases (DNA-dependent RNA polymerases) that will be ultimately replaced by DNA. However, as previously introduced, several viruses harbouring relatively small linear dsDNA genomes and related mobile genetic elements (MGE) can replicate their genomes by an alternative mechanism, called protein-primed DNA replication.[1,6] This way, those dsDNA replicons are replicated by an auxiliary protein that fulfil the primer role through the donation of free hydroxyl groups coming from donor Tyr, Ser or Thr moieties that undergo phosphoesterification with an initiating dNMP, directed by a 3'-internal template base.[1,2,7,9,10] These proteins, called DNA terminal proteins (TP), form a covalent bond with their encoding DNA by covalently binding to each strand's 5'-end, and may be, in addition, auxiliary in the genome encapsidation process.[1,2,9–12] Moreover, by permanently attaching TP to genome ends, this protein-primed alternative DNA replication mechanism can be regarded as a smart and efficient solution to the end replication problem of linear genomes, which entails a major relevance due to the utterly high replication rate frequently found in organisms or acellular elements bearing this replicative system.[6,13] Moreover, according to *in vitro* assays,

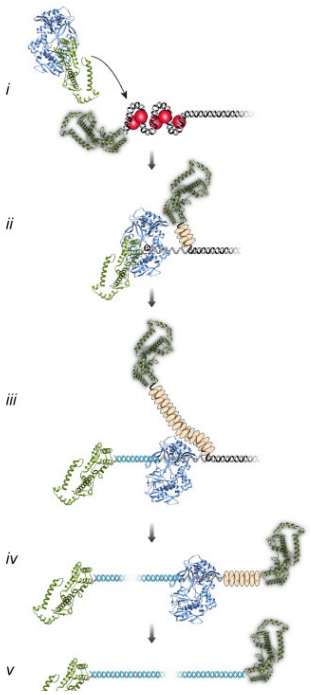

Figura 1. Schematic representation of TP-primed DNA replication from Redrejo-Rodríguez et al. (2012).[15]

only the two main elements, TP and their cognate DNA polymerases (pPolB), are required and sufficient for this replication mechanism,[6,12–16] and closing the knowledge gap regarding the first one can be considered the main purpose of this work.

Terminal proteins (TP), as key components of protein-primed DNA replication, were first characterized in the genome of *Salasvirus phi29* (Φ29) bacteriophage (class *Caudoviricetes*, *phylum Uroviricota*), a model system extensively studied by Margarita Salas for DNA replication.[6,12–17] This TP:pPolB system –where protein-primed DNA polymerase B (pPolB) uses a covalently attached TP to prime DNA synthesis– remains the most paradigmatic example. Since then, TP have been implicated in replication across a variety of linear double-stranded DNA (dsDNA) viruses (Baltimore Class I), particularly within the realms *Duplodnaviria* and *Varidnaviria*, and in phylogenetically related mobile genetic elements (MGE).[1,2,11,15,17] These include families such as *Tectiviridae* (e.g., PRD1, Bam35, GC1, forthebois),[1,18–27] *Adenoviridae* (e.g., *Mastadenovirus caesari*, also known as human adenovirus 2, hAd2),[1,25,26] and *Autolykiviridae* (e.g., *Paulavirus viph1080o*),[25] as well as rare archaeal viruses like *Bottigliavirus pozzuoliense* (ABV) and *Salterprovirus australiense* (His1).[1] Notably, although many of these genomes are predicted to use TP for protein-primed replication, only a small subset has been empirically validated. Moreover, despite the broad host range –including Gram-positive and Gram-negative bacteria, archaea, and eukaryotes– conserved gene synteny and phylogenetic patterns suggest that similar TP-dependent mechanisms may be widespread among related taxa.[1,2,7,12,25] Within the phylum *Preplasmiviricota*, the TP:pPolB system may be a definitory feature.[1,11,25] TP are often encoded directly upstream (5') of their cognate pPolB genes, with conserved synteny across polintons, PLV, virophages (*Virophaviricetes*), and linear plasmids.[1,11,25] Some exceptions show alternative arrangements (e.g., downstream TP or 5'-overlapping open reading frames), particularly in adenoviruses, where TP require proteolytic cleavage for activation.[1,25] Although adenoviruses share functional similarities with *Preplasmiviricota*-encoded TP, recent evidence suggests the adenoviral TP system likely evolved independently, perhaps via reverse evolution from *Polintoviricetes*.[1,25,29,30] Over the past decade, numerous dsDNA viruses –especially within *Caudoviricetes*– have been proposed to encode TP, including species from *Gueliniviridae* (e.g., *Brucesealvirus*), *Madridviridae* (*Cepunavirus*), and *Rountreeviridae* (*Andhravirus*), among others.[10,30–33] Bioinformatic predictions have expanded this list to include various genera related to Φ29, such as *Salasvirus*, *Huangshavirus*, *Gaunavirus*, *Claudivirus*, *Bundooravirus* or *Bahkauvirus*, *Andhravirus* or ophan *Microbacterium*-infecting species Ayka, Evcara, Curie or Voltaire.[1,10,28,31–42] An unusual case is *Bombyx mori densovirus 3* (BMD3), a linear single-stranded DNA (ssDNA) virus in *Bidnaviridae*, which appears to encode a pPolB-TP fusion acquired via horizontal gene transfer (HGT).[1,29] Despite increasing genomic evidence, experimental confirmation of TP remains limited, with Φ29,[17] PRD1, Bam35, and Cp1 among the few systems with structurally or mechanistically characterized TP.[1,6.12–15,32,43]

Thus, TP constitute a highly heterogeneous and phylogenetically dispersed group of proteins, sporadically encoded within the genomes of certain viruses and mobile genetic elements (MGE). Despite their apparent lack of evolutionary connectivity, TP often exhibit conserved interspecific genomic synteny. They are canonically located immediately upstream (5') of the open reading frame (ORF) encoding the cognate protein-primed DNA polymerase B (pPolB), with reported cases of genic overlap. However, recent studies have documented positional variations, including downstream (3') placements or insertion of intermediate ORF.

pPolB represent a specialized subfamily of DNA polymerases within the broader PolB family, and are structurally adapted to enable protein-priming via interaction with a TP, primarily through two conserved subdomains –TPR1 and TPR2– believed to mediate TP binding and facilitate primer-independent strand displacement, respectively.[1,6] While their interaction is assumed to be specific and essential in all TP-encoding MGE and viruses, this has not been experimentally validated in most cases.[1,2,44] The proposed replication mechanism involves the formation of a TP:pPolB heterodimer, stabilized by the asymmetrical surface charge density distribution of the TP, which promotes interaction with the negatively charged DNA backbone and positively charged pPolB regions.[1,6,16,17,22] This complex binds to replication origins typically located at genome termini.[6,16] This process often occurs in proximity to terminal repetitive sequences, slightly downstream of the 5' end, necessitating either a "sliding-back" (as seen in standalone TP systems

of *Caudoviricetes* and *Tectiliviricetes*)[1,6,10,12,16,17] or a "jumping-back" mechanism (common in *Adenoviridae* and related *Preplasmiviricota* systems featuring TP:pPolB polypeptides)[1,2,6,26,43]. The heterodimer remains stable during synthesis of the first ~10 nucleotides, after which conformational changes may occur.[1,6,16,17]

Notably, some TP (particularly those of adenoviruses) bear nuclear localization signals (NLS), such as lysine-rich motifs, potentially enabling eukaryotic nuclear import.[14,15] This suggests an evolutionary adaptation for host infection and may point to ancient prokaryote-eukaryote HGT events. Given the structural diversity of TP, the varied host range of TP-encoding replicons, and their potential in synthetic biology, deeper investigation into TP biology is both timely and necessary.[14,15] Their application in biotechnology is particularly compelling: TP and pPolB could underpin novel self-replicating, linear plasmid- or virus-based vectors capable of nuclear delivery and replication independent of host polymerases. Such vectors could support stable gene expression, transgenesis, targeted mutagenesis, and genome editing without integrating into host DNA, offering valuable tools for genetic engineering, functional genomics, and gene therapy.[14,15,26,45-48] Exploring the phylogenetic landscape and evolutionary history of TP would close key gaps in our understanding of DNA replication and open avenues for the design of advanced genetic tools. This requires a multidisciplinary strategy built on three core pillars: (i) large-scale *in silico* screening and annotation of TP in genomic databases; (ii) *in vitro* purification and biochemical characterization of TP:pPolB systems; and (iii) evolutionary analyses tracing the origins and functional divergence of TP and their associated pPolB. Here, we focus on the first and third of these strategies using a bioinformatic framework. Nonetheless, TP discovery and annotation remain challenging due to several computational and biological hindrances.

1.2. Inherent hindrances for TP identification and annotation

Unlike pPolB, TP are not conserved and lack distinctive catalytic motifs or common structural domains. Due to their unique nature and the atypical genomes that encode them, TP are often misannotated –or not annotated at all– in viral genome assemblies, especially those submitted without curation to public sequence databases. As a result, TP are frequently missing from closely related genomes, even within the same genus, complicating their detection through comparative genomics.[1,2] This issue is compounded by the widespread generation of low-quality draft genome assemblies –often consisting of fragmented contigs and scaffolds– which may include unrecognized viral sequences from virus, plasmids, or other MGE. Such contaminated or mixed-origin datasets are routinely deposited without adequate filtering or curation, introducing systemic annotation errors that propagate through downstream bioinformatic analyses. Similarly, in viral genomics, core host genes are sometimes misclassified as viral, further skewing annotations.[1,2,25]

Over time, this has led to a significant underrepresentation and misclassification of TP sequences. Many remain buried in unannotated or incorrectly annotated data, misidentified as host proteins, or lost due to insufficient sequence similarity with known TP.[1,16,25] Consequently, the number of validated and correctly annotated TP remains exceedingly small, necessitating exhaustive manual curation, a time-consuming and inefficient process. This challenge is further exacerbated by the extreme sequence divergence among known TP, the absence of reliable reference sequences, and the lack of universal or parameterizable structural criteria for their identification.[1,15,22]

Empirical structural data for TP is scarce and only one has been fully resolved by X-ray crystallography (Φ29 TP).[13,16] Structural models for a few others (e.g., PRD1, Bam35, Cp1) have been generated through predictive tools –especially after the introduction of AlphaFold2– but results vary widely in accuracy due to poor sequence quality and distant homology.[12,22] Many TP structures remain low-confidence (according to metrics like pLDDT, iPTM, and Q-score), despite their generally conserved cognate pPolB being more amenable to accurate prediction due to conservation.[1,49] To date, no structure-based phylogenetic framework exists for TP, leaving the evolutionary origins and molecular diversity of this protein group unresolved. The sporadic presence or absence of TP in closely related viral genomes and their high degree of sequence and structural divergence suggest a diffuse evolutionary origin, likely involving multiple, independent acquisition events.[1,25,49] This is consistent with the broader virological context, where viral genomes frequently incorporate additional genes through HGT. These acquired genes may not be

orthologous but can undergo evolutionary convergence, resulting in homoplasy with analogous functions, as seen in viral capsid proteins.[49] Gene shuffling among co-infecting viruses further contributes to this genomic plasticity, blurring phylogenetic relationships and complicating origin tracing.

Although the basic role of TP –donating a priming hydroxyl group– is mechanistically simple, the structural implications are highly complex. TP must simultaneously interact with template DNA, the nascent strand, and the initiating dNMP-bound pPolB, forming a covalent nucleoprotein complex of extreme topological complexity. These sterically impeded structures have yet to be empirically resolved and are only weakly predicted in silico, largely due to limitations in both sequence and model quality. Preliminary structural predictions suggest that TP:DNA and TP:pPolB interactions may be driven by exposed regions with asymmetric surface charge distributions: positively charged patches could interact with DNA's negatively charged phosphate backbone, while electronegative regions may align with basic regions of pPolB.[1,12,17] Despite their divergence, a few apparently shared features have emerged that might support standardizing TP identification.[10,12,15,22] These include: (i) asymmetric surface charge density distribution, facilitating simultaneous interaction with DNA and pPolB (Figure S1.), (ii) DNA-binding capacity, inferred from structural predictions and necessary for priming, and a (iii) predominantly α-helical secondary structure, common in predicted TP folds.[1,2,10,11,15] When combined with relative synteny and conserved genomic positioning adjacent to pPolB,[1,2,7,28] these features could offer a functional framework for validating TP candidates.

1.3. Bioinformatic approaches on pPolB and DNA TP identification

As discussed, the identification and validation of TP in viral genomes –especially those assumed to replicate via protein-primed mechanisms– has faced major challenges over the past two decades. In many cases, it remains unclear whether a TP is present at all in a given genome, due to incomplete assemblies, inconsistent annotations, and the absence of conserved sequence features or clear phylogenetic links among known TP. However, a breakthrough came in January 2024, when Mart Krupovic *et al*. proposed a novel, structure-based approach to bypass these limitations.[1] By leveraging cutting-edge protein modelling tools such as AlphaFold2 and structural homology search algorithms like DALI, they demonstrated that structure-based phylogenetics could effectively unblur viral evolutionary relationships that sequence-based methods had not been able to.[1,25] Applying this method to pPolB, their study suggested that some TP may have ended up as N-terminal structural domains fused to pPolB, with homologs possibly derived from bacterial proteins. Although they did not pinpoint a definitive ancestral TP or explore the full diversity of TP:pPolB systems, their work offered valuable insight into the evolutionary trajectory of TP within a subset of replicons, particularly those classified in the *phylum Preplasmiviricota*.[30] This built upon earlier suggestions that TP genes might overlap with the 5' end of the pPolB open reading frame. Recognizing that pPolB is consistently conserved across *Preplasmiviricota* –unlike capsid proteins, which evolve rapidly and unevenly– Krupovic *et al.* used pPolB as a phylogenetic anchor. Their structure-based approach enabled a clearer view of relationships among eukaryotic adenoviruses and related mobile genetic elements, setting the stage for broader applications in viral phylogenetics.[1,11,25]

Krupovic *et al.* propose that, in most *Preplasmiviricota* replicons (excluding *Tectiviridae* and *Adenoviridae*), as well as in *Bidnaviridae*, the TP is not separately encoded but rather fused as an N-terminal domain (D1) of a single TP:pPolB polypeptide (Figure 2).[1] This fusion complicates TP annotation, as the encoding gene may be cryptically embedded within the polypeptide pPolB ORF. The D1 domain resembles TP of PRD1 and adenoviruses, possibly by homology or convergent evolution.[1] It is often followed by a viral ovarian tumour-like cysteine deubiquitinlyase (vOTU or D2 domain), which may cleave and release the active TP in some contexts, though this cleavage appears inactive or absent in *Adenoviridae*, *Bidnaviridae*, and mitochondrial plasmids.[1] In PLV and transpovirons, the polypeptide pPolB may also include a vOTU domain, a canonical polymerase "palm" domain (reducing proofreading 5'→3' exonuclease activity)[1,4,50,51], and in some cases, a helicase domain from superfamily 1 (S1H) or 3 (S3H). Some PLV also harbour highly variable N-terminal extensions with no known structural homology, and in specific cases, a C-terminal GIY-YIG or HNH endonuclease domain. Interestingly, *Sputnikvirus* virophages (*Mividavirales*) appear to have replaced their ancestral pPolB with either a PolA or an archaeal-eukaryotic primase-polymerase (AEP), losing protein-primed replication capacity while maintaining overall phylogenetic proximity to other virophages.[1,52–58]

Structural analysis by Krupovic *et al*. suggests no detectable homology between TP of *Caudoviricetes* and *Preplasmiviricota*, pointing independent origins.[1] *Caudoviricetes* TP feature long α-helices, while *Preplasmiviricota* TP show mixed α/β folds.[1] Even within *Preplasmiviricota*, D1 domains show only modest structural conservation, suggesting deep evolutionary divergence. Despite this, structural patterns offer a promising route for identifying TP and studying protein-primed replication mechanisms. Comparative structural models from viruses infecting different hosts show differences in TP folding, potentially reflecting host-specific adaptation. Some structural resemblance among TP of *Bacillus*-infecting phages (across *Preplasmiviricota* and *Uroviricota*) hints at host-linked structural tropism. These differences, along with the lack of similarity between Bam35, PRD1, and Φ29 TP, support the hypothesis of independent HGT events at the origin of TP.[1,25] Krupovic *et al.* further hypothesize that *Tectivirus* and related MGE's TP or D1 domain may have originated from a "pinky finger" domain (LF/PAD) of Y-family DNA polymerases (e.g., Pol IV/V in bacteria, Pol ι/κ/η/Rev1 in eukaryotes), originally involved in translesion synthesis (TLS).[1,58] Over time, its function may have consistently shifted from dNTP binding to covalent dNMP attachment.[1]

The structural similarity between adenoviral and PRD1-like TP, alongside their consistent fusion in eukaryotic *Preplasmiviricota* genomes, suggests a shared ancestry from *Alphatectivirus*. This fusion may have occurred around the time of early eukaryogenesis, with evidence tracing co-evolution back to the Last Eukaryotic Common Ancestor (LECA).[1,59] The first major split within eukaryotic Preplasmiviricota pPolB likely separated mitochondrial linear plasmids (MLP) from other lineages –preceding the acquisition of the vOTU/D2 domain by *Polintoviridae* and cytoplasmic linear plasmids–, possibly during the migration of ancestral polintons from endosymbiotic protomitochondria into the cytoplasm.[1,15,25] Although adenoviruses encode a standalone TP resembling those of *Alphatectivirus*, Krupovic *et al.* argue they may have evolved directly from polintons, as their pPolB retains a now-inactive vOTU/D2 domain.[1] The split of the TP gene may have occurred later, paralleling similar events in some polintons.[1,2,25] In contrast, PLV, transpovirons, and *Lavidavirales* virophages likely diverged independently from polintons through loss of exonuclease activity and acquisition of helicase domains S1H or S3H.[1]

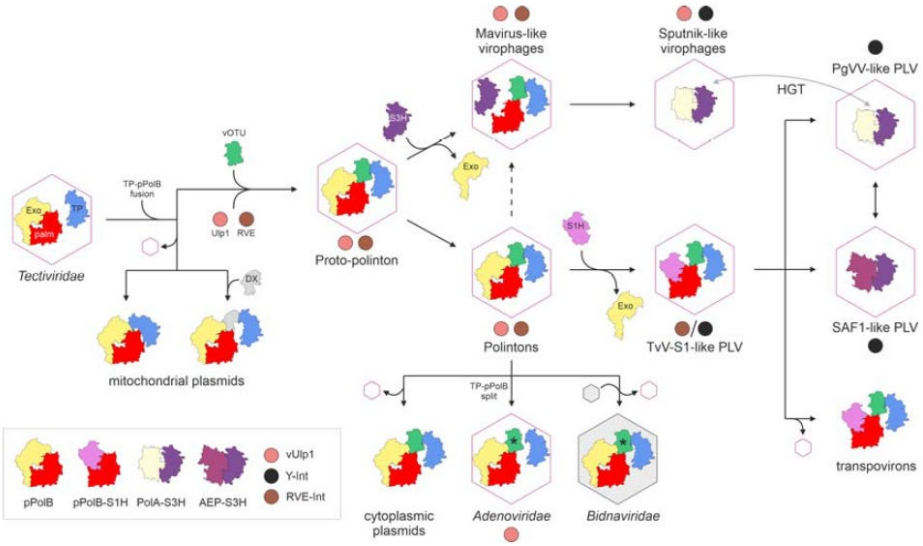

Figure 2. Proposed phylogenetic-evolutive relationships among members of *phylum Preplasmiviricota* according to structural homology of TP and polypeptide pPolB. Depicted in Kurpovic *et al*. (2024).[1]

Despite these advances, further structural-functional validation is needed to confirm their proposed evolutionary scenarios and to establish standardized criteria for TP identification, annotation, and re-annotation across viral lineages.

1.4. Structure comparison versus sequence comparison

Until recently, the lack of reliable structure prediction and structure-based phylogenetic tools made TP screening and identification highly challenging. Traditional sequence-based methods were hampered by TP

misannotation, low sequence conservation, cryptic gene locations, and inconsistent genome quality. However, Structural Bioinformatics has rapidly advanced over the past decade, culminating in a breakthrough with DeepMind's AlphaFold2 in 2021,[60] which dramatically outperformed coeval tools at CASP14.[60–63] This milestone spurred the development of new high-accuracy, deep-learning-based tools for protein structure modelling (e.g., AlphaFold3, Chai-1, Boltz-1/1x),[64–67] multimer prediction,[64,68] protein design using NLP (e.g., ProtGPT2, BioM3),[69–72] molecular docking,[65,73] and structure-based homology search (e.g., DALI, Foldseek)[63,74–77]. These tools have enabled exploration of previously inaccessible protein fold space and the functional annotation of vast amounts of uncharacterized genomic data. This revolution has led to the creation of new structure-based databases (e.g., AlphaFoldDB, BFVD, ESM Metagenomic Atlas, MGnify),[78–81] transforming structural bioinformatics into an essential toolbox for evolutionary, functional, and phylogenetic analysis. For example, Krupovic *et al.* applied AlphaFold via ColabFold to model TP:pPolB structures and used DALI to identify remote structural homologs.[1,60,61] Even this simple approach –without the use of newer phylogenetic algorithms– was sufficient to propose an evolutionary trajectory for TP-containing elements.

Protein sequence and structure comparisons offer complementary insights. Sequence alignment tools (e.g., BLAST, Clustal Omega, MMSeqs2) identify conserved motifs and infer function or ancestry but are limited in detecting distant homologs.[63,74] In contrast, structure-based tools (e.g., DALI, TM-align, Foldseek) detect conserved folds and topologies even when sequence similarity is minimal. Integrating both approaches enhance the robustness of evolutionary and functional inferences. Critically, structure-based homology searches have opened a new layer of evolutionary analysis by detecting deep relationships obscured by sequence divergence.[63,75,76] DALI and Foldseek represent two major paradigms in this space, where DALI compares inter-residue distance matrices to identify global fold similarities, offering high sensitivity but at a computational cost.[63] Foldseek, on the other hand, encodes structures into discrete alphabets for ultrafast comparison, sacrificing some sensitivity for speed and scalability.[74] On balance, DALI is ideal for deep yet small-scale evolutionary inference, while Foldseek excels in large-scale structural comparisons with reduced time consumption.[63,74]

2. Motivation and objectives

As discussed in the Introduction (see 1.), viral terminal proteins (TP) present an evolutionary and biochemical challenge. Very few have been functionally and biochemically characterized, and their greatly diverse sequences hinder accurate annotation and make it impossible to trace their evolutionary history. With the advent of deep-learning protein structure prediction methods and protein language models for improved structural comparison, we hypothesize that a systematic comparative analysis of TP sequences and structures could enhance TP classification and annotation. This main objective has been structured into three independent goals:

1. *In silico* prediction of three-dimensional structures of empirically validated, properly annotated and phylogenetically-diverse TP of known sequence.
2. Sequence and structure homolog search of selected TP, while suggesting and testing standardised definitory criteria.
3. Identification of putative non-viral homologs and possible structure-based phylogenetic history trace-back that unveil the primeval origin or origins of shortlisted TP clades.

3. Methods

3.1. Pre-dataset selection and structural modelling

3.1.1. Pre-dataset construction – GenBank and state-of-the-art references from PubMed

In order to perform an exhaustive screening for TP and putative cellular ancestors' identification, ensuring numeric and phylogenetic diversity representativity, we performed a manually-curated in-depth state-of-the-art bibliographic scrutiny of reliable viral species undergoing TP-primed genome replication on PubMed (https://pubmed.ncbi.nlm.nih.gov/), alongside a manual GenBank (https://www.ncbi.nlm.nih.gov/protein/) search on non-redundant database (nr) for protein entries annotated as TP, excluding perfectly identical sequences. Genomic contexts of putative TP were then inspected for relative synteny conservation and

DNA primase/helicase absence. Ambiguous *contigs* were further inspected on the Prokaryotic Virus Remote Homologous Groups Database (PHROGs, https://phrogs.lmge.uca.fr/) webserver.[82]

3.1.2. TP structural prediction/modelling

In order to generate structural prediction models for each starter TP they were recursively subjected to AlphaFold2 modelling through ColabFold (v1.5.5).[60,61] Several sets of hyperparameter values were pre-tested in order to ascertain how AlphaFold2 modelling worked for our TP dataset and which hyperparameter set fitted best for an optimal structural prediction. Thus, after testing high recycle values (namely 6 to 24, as more than 24 recycles are not time-affordable for non-clustered computing and for the number of sequences to be modelled), as equal or worse results (in terms of structure quality assessment based on scores pLDDT, pTM and iPTM) were obtained, such hyperparameter's value was eventually set to 3 recycles so that a substantial time consumption alleviation would be prioritised over minimal (or none) predictive accuracy gain. In addition, the recycle early stop tolerance was set to 0.0 in order to maximise stringency (by avoiding greatly divergent recycles), whilst reducing time consumption. Non-default hyperparameter set finally was defined by model_type = alphafold2_ptm, num_recylces = 3, and tol (recycle_early_stop_tolerance) = 0.0.

3.2. Ascertainment of standardizable discriminatory-definitory criteria for TP nature assessment

3.2.1. Surface charge density distribution asymmetry

For surface charge density distribution asymmetry assessment, protein structure PDB files were programmatically parsed by employing Bio.PDB.PDBParser module[a] from Biopython (v.1.76) library.[83] Each PDB's atoms were iterated over to extract their Cartesian coordinates and assigned partial charges via a custom function referencing an AMBER/CHARMM-based[84] atom-type-to-charge mapping. Extracted embedded atomic clusters, stored as coordinates and charges NumPy (v.2.2.0) arrays for efficient numerical computation, were then projected into a virtual three-dimensional Cartesian space generated using numpy.meshgrid,[b] for topology-aware surface charge density distribution asymmetry calculation. Three-dimensional grid was finetuned in size and density for proteins of interest in order to avoid over- or under-estimation of charge asymmetry, with a resolution of 50 points per dimension of a cubic mesh spanning from −40Å to +40Å along each axis, evaluating the local electrostatic potential for a total of 125,000 grid points. At each grid point, the electric potential in solution in the direction normal to a charged surface ($\nabla^2\psi$) was computed by summing the contributions from all atoms using a simplified screened Coulomb potential model (see Equations 1) derived from the linearized Debye-Hückel implementation of Poisson-Boltzmann equation, accounting for how local charges influence neighbouring charges. This was implemented by calculating the Euclidean distance between each grid point and all atomic positions, then computing the potential as the sum of each atomic charge divided by the dielectric-scaled distance with a screening constant (water dielectric constant ε_r = 80, and Debye-Hückel screening factor κ = 1). For the sake of computation, we presumed a uniform dielectric constant (ε_r) ignoring spatial variations in dielectric properties, treated atoms as point charges (q_i, located at $\mathbf{r_i}$), assumed a weak screening (κ instead of $\kappa^2\psi$) and neglected accounting for nonlinear ionic effects. A custom function then partitions the 3D grid into orthogonal hemispheres and sums the potential values in each. Fine octant-based partition was also tested. Asymmetry scores along each axis were calculated as the normalized difference in potential between paired hemispheres, and the final electrostatic asymmetry score was obtained by averaging these three directional scores. All data processing and visualization were conducted using Python NumPy and Matplotlib (v.3.10.1) matplotlib.pyplot[c] and mpl_toolkits.mplot3d modules.[d]

[a] https://biopython.org/docs/1.76/api/Bio.PDB.PDBParser.html; https://biopython.org/wiki/PDBParser
[b] https://numpy.org/doc/2.2/reference/generated/numpy.meshgrid.html
[c] https://matplotlib.org/stable/api/pyplot_summary.html#module-matplotlib.pyplot
[d] https://matplotlib.org/stable/api/toolkits/mplot3d.html

$$\begin{cases} \text{Poisson: } V = \nabla^2\psi = \dfrac{\partial^2\psi}{\partial x^2} + \dfrac{\partial^2\psi}{\partial y^2} + \dfrac{\partial^2\psi}{\partial z^2} = -\dfrac{\rho_e}{\epsilon_r\epsilon_0} \\ \text{Boltzmann: } c_i = c_{0_i} e^{-\dfrac{z_i e\psi}{k_B T}} \\ \text{Debye-Hückel: } \rho = \displaystyle\sum_{i=1}^{N} z_i q_i n_i = \sum_{i=1}^{N} N z_i q_i n_{0_i} e^{-\dfrac{z_i q_i \psi}{k_B T}} \end{cases} \longleftrightarrow V = \nabla^2\psi = -\sum_{i=1}^{N} \frac{z_i q_i n_{0_i}}{\epsilon_r\epsilon_0} e^{-\dfrac{z_i q_i \psi}{k_B T}}$$

$$\therefore -\sum_{i=1}^{N} \frac{z_i q_i n_{0_i}}{\epsilon_r\epsilon_0} e^{-\dfrac{z_i q_i \psi}{k_B T}} \approx -\sum_{i=1}^{N} \frac{z_i q_i n_{0_i}}{\epsilon_r\epsilon_0}\left(1 - \frac{z_i q_i \psi}{k_B T}\right) = -\left(\sum_{i=1}^{N} \frac{z_i q_i n_{0_i}}{\epsilon_r\epsilon_0} - \sum_{i=1}^{N} \frac{z_i^2 q_i^2 n_{0_i}^2 \psi}{\epsilon_r\epsilon_0 k_B T}\right)$$

$$\text{Gronwall: } V = \nabla^2\psi = \sum_{i=1}^{N} \frac{z_i^2 q_i^2 n_{0_i}^2 \psi}{\epsilon_r\epsilon_0 k_B T} = \kappa^2\psi : \text{Helmholtz}$$

$$\nabla^2\psi - \kappa^2\psi = 0\ ;\ \nabla^2\psi(\mathbf{r}) - \kappa^2\psi(\mathbf{r}) = \sum_{i=1}^{N} \frac{q_i}{\epsilon_r\epsilon_0}\delta(\mathbf{r} - \mathbf{r_i}) \rightarrow \psi(\mathbf{r}) = \sum_{i=1}^{N} \frac{q_i}{\epsilon_r|\mathbf{r} - \mathbf{r_i}|} e^{-\kappa|\mathbf{r}-\mathbf{r_i}|} \approx \sum_{i=1}^{N} \frac{q_i}{\epsilon_r|\mathbf{r} - \mathbf{r_i}| + \kappa}$$

$$V \cong \psi(\mathbf{r}) \approx \sum_{i=1}^{N} \frac{q_i}{\epsilon_r r + \kappa}$$

Equations 1. Simplified linearised Debye-Hückel implementation of Poisson-Boltzmann equation for point electrostatic potential calculation.

This simplified implementation was validated with external PDB2PQR and APBS toolkits in in order to validate this models' accuracy. A dataset comprising the protein structures of the 45 starter TP and other 190 non-TP control proteins (*N* = 235) from UniProtKB (https://www.uniprot.org/) was subjected to charge density asymmetry score calculation in order to assess the potential of charge density distribution asymmetry as discriminatory parameter for TP assessment (DATASET LOCATION). The control set included proteins of all taxonomic dominions and of diverse functionalities, length, folding and size so that a realistic representative set of protein diversity in nature could be employed as a contrast to TP structures. Charge asymmetry suitability for TP detection was then assessed by a statistical class-wise analysis of asymmetry scores consisting of non-parametric Wilcoxon rank-sum test followed by an effect-size evaluation using rank-biserial and Cohen's *d* as metric statistics. Statistic assessments were scripted in **R**; Durga (v.2.0)[e] package was employed for effect-size comparison and ggplot2 (v.3.5.2)[f] package was used for the generation of violin and *kernel* density plots.[85]

3.2.2. Secondary structure

Secondary structure content was computed for the TP- and control-containing dataset of target proteins (*N* = 235) using DSSP algorithm as implemented in Biopython (v.1.76) Bio.PDB.DSSP[g] module,[86,87] and protein structure PDB files were parsed using Bio.PDB.PDBParser. The DSSP algorithm intensively assigns secondary structure codes to each residue based on hydrogen bonding patterns and backbone dihedral angles; and these codes 8-element-based were then grouped into three major categories for the sake of simplicity: helices, β-sheets, and coils/turns. The number of residues in each category was tallied, and the total secondary structure content was calculated as a percentage of all DSSP-annotated residues. This scrutiny was iteratively applied over each protein structure, using resulting β-helix, sheet, and coil percentages for assessing secondary structure relative abundance suitability for TP discrimination. As in 3.2.1., a non-parametric Wilcoxon rank-sum test followed by an effect-size evaluation using Cohen's *d* as metric statistic was independently performed on helix and β-sheet distributions using Durga (v.2.0) and plotted employing ggplot2 (v.3.5.2) package.

In order to avoid PDB format version and source inconsistencies incompatible with DSSP's input requirements, all dataset structures were thoroughly and systematically re-formatted. DSSP implementation expects PDB files to conform to crystallography-based standards, specifically requiring the presence of a

[e] https://cran.r-project.org/web/packages/Durga/index.html
[f] https://ggplot2.tidyverse.org/
[g] https://biopython.org/docs/1.76/api/Bio.PDB.DSSP.html

'CRYST1' record in the metadata header typically replaced by 'MODEL 1' or 'MODEL1' stances in AlphaFold2-predicted structural models, or displaced by non-standard unsupported metadata headers in SWISS-MODEL predictions. In addition, DSSP prerequisites also includes an explicit description of the dimensions of a crystallographic unit cell, omitted in *in-silico*-predicted models' PDB files. Thus, inconsistencies in lexicon and syntax were recursively inspected, identified and re-formatted. All modifications were performed on a local Google Colab controlled environment to preserve original data integrity, while TP and control PDB files were processed separately for the sake of traceability.

3.2.3. DNA binding probability

In order to compute dual sequence-/structure-based DNA binding probability for its evaluation as possible TP discriminatory criterion, the aforementioned control-containing protein structure dataset was recursively uploaded to DNABIND (https://dnabind.szialab.org/) webserver (no RESTful API access) by Dr. András Szilágyi from the Hungarian HUN-REN.[88] Predicted DNA-binding probabilities (DATASET) were further used to assess the statistical validity of the binary classifier model and, hence, the model's confusion matrix and per-class accuracy, precision, recall/sensitivity, specificity and F1-score were calculated using Scikit-learn (v.1.6.0)[h] library's sklearn.metrics module. Thus, the assessment results were plotted on a heatmap and a barplot using Seaborn (v.0.13.2)[i] and Matplotlib (v.3.10.1) matplotlib.pyplot module, respectively.

3.4. Homology search

3.4.1. Foldseek structure-based homology search

To identify structural homologs, AlphaFold2-predicted protein models for starter TP were iteratively queried against the Foldseek online search server (https://search.foldseek.com) via its RESTful API access.[74] Metadata, including TP acronyms, and sequence lengths, were retrieved from a CSV-formatted dataset (DATASET) Python pandas (v.2.2.3)[j] library. Each PDB file was submitted to Foldseek via HTTP POST endpoint requests using a UNIX curl command (v.8.13.0)[k] executed through Python's subprocess module.[l] Queries were configured to use the TM-align structural alignment mode against six curated databases of interest covering over 220 million target protein structures (either empirically elucidated or AlphaFold2-predicted);[89] namely BFVD (v.2023_02),[79] AFDB50 (AlphFold/UniProt50, v4),[78,89] SWISS-PROT (AlphaFold/Swiss-Prot, v4),[90] AFDB-Proteome (v4),[89] RCSB PDB100 (20240101),[m] and GMGCL (v.1_2204)[n, 91]. Other available databases were omitted for containing predicted structures of metagenomic origin or isolated protein domains without taxonomic assignation, or protein complexes. Job status for each unique job ticket ID obtained upon submission, was monitored via periodic GET requests using the requests module, polling every 10 seconds until completion or failure. Upon job completion, each TP's tar-compressed result archives (.tar.gz format) was downloaded using streamed HTTP requests and written in chunks to accommodate large file sizes. Error handling was implemented for both submission and retrieval stages to ensure robust pipeline execution across the full TP set.

Upon decompression of recovered output files, headerless non-standard expanded 20-columned .m8 tabular files (employed by DIAMOND-like sequence-based alignment algorithms) gathering each protein's structural homology results from each queried database were parsed using pandas. Each file was checked for content validity before processing. Relevant alignment metrics –namely hit ID ('Hit'), percent identity ('%I'), query length ('len(Qry)'), E-value, Bit-Score, probability of homology ('P(H)'), taxonomic lineage ('taxonomy') and NCBI TaxonomyID ('TaxID')[92,o]– were extracted and standardized across databases. Alignment coverage ('%aln') was calculated as the ratio of query-referenced alignment aligned length ('len(aln)', computed as the subtraction of the last aligned position to the first aligned position plus one) to query length. Columns were renamed for clarity and restructured into a unified field schema. Database-

[h] https://scikit-learn.org/stable/
[i] https://seaborn.pydata.org/
[j] https://pandas.pydata.org/
[k] https://curl.se/
[l] https://docs.python.org/3/library/subprocess.html
[m] https://www.rcsb.org/
[n] https://gmgc.embl.de/
[o] https://www.ncbi.nlm.nih.gov/taxonomy

specific formatting variations were resolved via string manipulation. Retrieved entries from each database were annotated with their source database and TP and then concatenated into a TP-specific results dataframe that were further aggregated into a master dataframe containing results for the entire starter TP dataset.

Finally, all retrieved entries were queried against UniProtKB (https://www.uniprot.org/) database via its RESTful API, using Python's requests (v.2.32.3)[p] library, in order to retrieve GenBankID for each AFDB-linked entry (as BFVD and PDB do not cross-reference their entries) by targeting xref_embl field, so that ulterior deduplication with PSI-BLAST-derived hits could be carried out.[q] Entries lacking retrievable GenBank references –either due to API errors, deprecated UniProtKB entries, or entries exclusive to non-UniProtKB databases– were annotated accordingly with empty strings. The resulting dataframe was then exploded to normalize multiple GenBankID per hit, rearranged to follow a consistent column order, saved and duplicated to preserve the unfiltered state.

A filtering step was then applied to the raw dataframe to retain only high-confidence putative homologs, using a set ([Ξ]) of three conservative thresholds: E-value ≤ 0.01, homology probability (P(H)) ≥ 0.3, and aligned query coverage (%aln) ≥ 30%. Different confidence threshold set configurations were tested in order to empirically find the optimal compromise between a high number of diverse hits and confidence in the structural similarity. The filtered dataset was then sorted using a multi-criteria approach, prioritizing lowest E-values, highest P(H), and highest %aln. To address redundancy, rows with duplicated GenBankID were identified and collapsed, retaining only the highest quality instance of each unique GenBankID. Importantly, entries lacking GenBank identifiers were exempt from this deduplication step to avoid unintended data loss.

For analysing the top-tier taxonomic composition of Foldseek-yielded result dataset, those non-redundant entries containing a TaxID –those not referring to deleted, obsolete or rebranded UniProtKB entries, nor to BFVD-exclusive hits– were queried against NCBI Taxonomy database using Biopython's (v.1.76) Bio.Entrez package.[r] Taxonomy records were fetched in XML format via efetch, and the corresponding full lineage strings were extracted. A keyword-based matching strategy was applied to assign each hit to one of six NCBI Taxonomy broad domain-equivalent categories. Each query's results were appended to a lineage-containing dataframe , and summary counts for each category were compiled and then plotted down using dplyr (v.1.1.4),[s] tidyr (v.1.3.1)[t] and ggplot2 (v.3.5.2) **R** packages.

3.4.2. PSI-BLAST sequence-based homology search

In parallel to Foldseek, sequence-based homology searches were conducted using Position-Specific Iterated BLAST (PSI-BLAST) for far sequence homologs detection.[93,94] FASTA files for each TP in the starter dataset were retrieved beforehand by querying each GenBankID against NCBI Protein database (https://www.ncbi.nlm.nih.gov/protein/) via Biopython's (v.1.76) Bio.Entrez package. Fetched results were saved using Bio.SeqIO package.[u] FASTA files were batch-processed in parallel using GNU parallel (v.20120122)[v], where each sequence was locally queried against the clustered NCBI non-redundant sequence database (nr_cluster_seq, v.1.1., 2024-11-21)[w] using PSI-BLAST through *bash* scripting. After finetuning search parameters, PSI-BLAST was run with an E-value threshold of 0.001, and three iterations to enhance sensitivity through profile refinement. Results were formatted with tabular format and saved as TSV that were then parsed using custom Python scripts to extract relevant alignment statistics as done with Foldseek results files before (see 3.4.1.). Downstream processing was analogous to such carried out for structural putative homologs, including formatting, prefiltering, sorting and deduplication of redundant entries. Results pandas (v.2.2.3) dataframes for each starter TP were merged together into a single master

[p] https://pypi.org/project/requests/
[q] https://rest.uniprot.org/uniprotkb/{accession}
[r] https://biopython.org/docs/1.76/api/Bio.Entrez.html
[s] https://dplyr.tidyverse.org/
[t] https://tidyr.tidyverse.org/
[u] https://biopython.org/docs/1.76/api/Bio.SeqIO.html; https://biopython.org/wiki/SeqIO
[v] https://www.gnu.org/software/parallel/
[w] ftp://ftp.ncbi.nlm.nih.gov/blast/db/experimental/nr_clustered_seq.*.tar.gz

dataframe as done before, but default placeholder values were assigned for Foldseek-like fields unavailable from PSI-BLAST output (namely 'Taxonomy', 'TaxID' and homology probability 'P(H)'). Additional fields 'Method' (valued 'PSI-BLAST'), 'Database' (valued 'nr'), and TP acronym were added for further merger with Foldseek results. This time, as homology results were directly retrieved from nr, GenBankID for each entry was assigned by replicating its 'Hit' value.

To retrieve taxonomic context for each PSI-BLAST hit, protein GenBankID were queried against the NCBI Protein database using Biopython's (v.1.76) Bio.Entrez package. For each entry in the deduplicated results dataframe, the corresponding GenBank ID was used to fetch (using efetch) its GenBank flat file, from which the associated Taxonomy ID was parsed and recorded ('TaxID'). Species names ('Taxonomy') were also obtained by querying using Bio.Entrez against NCBI protein and parsing the 'ORGANISM' field of each full GenBank flat file. Subsequent downstream taxonomic lineages retrieval and descriptive statistics plotting were analogous to those carried out with Foldseek results (see 3.4.1.) .

Candidate homologous entries obtained via sequence- or structure-based homology were inspected for overlapping/coincidence both at entry and at TaxID levels via ggVennDiagram (v.1.4.4)[x] **R** package.[95,96] Format-compatible pre-processed and deduplicated sequence-based PSI-BLAST (Species_PSI-BLAST_results_updated.csv) and structure-based Foldseek (Dedup_FoldSeek_results_low_updated.csv) homology results datframes were merged into a single unified dataset (FINAL_results.csv) using pandas (v.2.2.3) and further deduplicated by 'GenBankID'.

3.4.3. BLASTp sequence-based homology search

For sequence-based BLASTp close homology search of selected UniProt-exclusive representative hits (select_fasta.sh). FASTA files for each representative candidate were retrieved beforehand by querying each UniProtKB identifier ('Hit') against UniProtKB database (https://www.uniprot.org/) by using requests (v.2.32.3) Python library for HTTP GET endpoint. They were batch-processed in parallel using GNU parallel (v.20120122), where each sequence was locally queried against nr_cluster_seq (v.1.1., 2024-11-21) database through *bash* scripting. After finetuning search parameters, BLASTp (v.2.16.0) was run with an E-value threshold of 10^{-5}, and a maximum of 5000 sequences per search.[97]

PROCESSING

3.4.4. Taxonomic in-depth scrutiny

In order to enhance taxonomic resolution across all candidate entries, the merged homology results were processed using the **R** taxonomizr (v.0.11.1) package.[y] The NCBI taxonomy database was prepared locally to allow efficient retrieval of full taxonomic hierarchies. Unique Taxonomy ID were extracted from the dataset and queried using getTaxonomy() against the local SQL database. The retrieved taxonomic ranks were merged back with the main dataset (FULL_results.csv).

For the completion of higher-rank taxa (namely realm and kingdom) systematically neglected by taxonomizr, an additional Python parsing post-processing was implemented. Full lineage strings previously obtained querying against NCBI Protein using Biopython (v.1.76) Bio.Entrez and Bio.SeqIO packages (FULL_taxonomy_classification_results.csv) were iteratively scanned using pandas (v.2.2.3) and taxonomic levels were inferred based on suffix patterns and known taxonomic naming conventions. Context-aware conditions were applied to distinguish taxa across domains/kingdoms. Inferred ranks were compiled into a lineage table, and realm and kingdom assignments were programmatically integrated into a full dataset (FULL_results+GenomeID+Tax.csv).

3.5. Phylogenetic assessment

3.5.1. Sequence-based Clustal Omega (MSA)

[x] https://cran.r-project.org/web/packages/ggVennDiagram/readme/README.html
[y] https://cran.r-project.org/web/packages/taxonomizr/index.html; https://github.com/sherrillmix/taxonomizr; https://cran.r-project.org/web/packages/taxonomizr/vignettes/usage.html

Both starter TP's and selected polypeptide pPolB's multi-FASTA formatted sequences were subjected to HMM profile-based multi-sequence alignment (MSA) and sequence-based phylogenetic reconstruction employing Clustal Omega (ClustalΩ, v.1.2.4)[z] webserver (https://www.ebi.ac.uk/jdispatcher/msa/clustalo) with default parameters. The resulting MSA were submitted to ESPript (v.3.0, https://espript.ibcp.fr/ESPript/cgi-bin/ESPript.cgi) for visual formatting and inspection.

3.5.2. Structure-based FoldMason (MSTA)

Both starter TP AlphaFold2-yielded models and selected polypeptide pPolB were subjected to structure-based phylogenetic reconstruction employing FoldMason (https://search.foldseek.com/foldmason) by uploading all PDB files to scrutinise to FoldSeek webserver (no RESTful API access, nor tuneable options were available).

3.6. Clustering

3.6.1. Sequence-based – CD-Hit

Either for pre-dataset or deduplicated homology results dataset sequence-based clustering, retrieved sequences were compiled into single multi-FASTA files (0_TP_Predataset.fasta or Main_Hit_Dataset.fasta) and subjected to clustering using CD-HIT algorithm via pycdhit (v.1.1.4)[aa] wrapper Python package.[98,99] Several identity threshold and word length configurations were tested in order to empirically find the optimal compromise among cluster/sequence number, sequence diversity and representativity. Eventually, clustering was performed with a sequence identity threshold of 60% (-c 0.6) and a word length (-n) of 4, while sequence comparison was carried out in a semi-global manner (-sc 1). The clustering process was executed in high-stringency accurate mode (-g 1), prioritizing alignment precision over speed. Description lengths were unrestricted (-d 0) for traceability.

3.6.2. Structure-based – qTMcluster

To assess Foldseek search accuracy and to group structurally similar Foldseek-yielded candidate hits and starter TP, we applied US-align's qTMcluster (v.0)[bb] module to the non-redundant dataset's and starter TP's structural models.[77] This method enables unsupervised structural clustering based on pairwise structural alignments, using qTM-score as a normalized similarity metric. Default parameter set was employed, setting sequence similarity threshold to 0.5.

qTMcluster structural clustering results (qTMcluster_results_queries.txt) were parsed to extract structure-wise cluster assignments. Each line was processed to isolate individual structure identifiers and identify the cluster representative structure, defined as the first listed. These data were compiled into a structured pandas (v.2.2.3) dataframe (qTMcluster_results.csv) containing 'Structure', 'Cluster', and 'Representative' fields. Subsequently, such dataframe was cross-referenced with Foldseek-derived candidate hits to assess whether starter TP and their associated hits belonged to the same structural cluster. For each Foldseek hit, the cluster membership of both itself and its associated starter TP was retrieved from the qTMcluster results dataset. A new annotated dataframe (qTMcluster_results_updated.csv) was generated with additional fields indicating the original TP (TP_Query), the TP's cluster (Query_Cluster), and a binary flag (Query_Coclustering) denoting whether the TP and hit co-clustered structurally.

In order to assess whether and how query coverage influenced co-clustering, the last cross-reference step above was added a filtering by '%aln' stance that was tested for a wide value set ranging 0.3-0.9. Results were plotted as barplots in **R** using ggplot2 (v.3.5.2).

3.6.3. Sequence-based – BLAST2seq, k-Means and PCA

As a theoretical contradistinction with sequence-based CD-Hit clustering, a pairwise-similarity unsupervised clustering was performed. All predataset TP FASTA-formatted sequence files were retrieved by querying each GenBankID against NCBI GenBank protein database (https://www.ncbi.nlm.nih.gov/protein/) as

[z] http://www.clustal.org/omega/
[aa] https://pypi.org/project/py-cdhit/
[bb] https://github.com/pylelab/USalign/blob/master/qTMclust.cpp

described before (see 3.4.2.), and were all-against-all pairwise-aligned using itertools (v.3.10.17)[cc] Python library and Biopython's (v.1.76) Bio.Blast.Applications module as a wrapper of NCBI's BLASTp (BLAST2seq) implementation.[dd,97] Each pairwise alignment percent identity score was stored in a 90×90 identity matrix which was then subjected to a 2-component principal component analysis (PCA) for dimensionality reduction through Scikit-learn's (v.1.6.0) sklearn.decomposition.PCA module.[ee] The reduced matrix was then employed for unsupervised k-Means clustering of TP through Scikit-learn's (v.1.6.1) sklearn.cluster.KMeans module,[ff] after empirically setting the optimal cluster number to 12. Clusters were plotted with Python's Seaborn (v.0.13.2) library and matplotlib.pyplot (v.3.10.1) package.

3.7. Genomic scrutiny

3.7.1. IPG-based genomic retrieval

Representative candidates obtained as a result of CD-Hit clustering (see 3.6.1.) –either original candidates or BLASTp synonyms of UniProtKB-exclusive hits– were queried by 'GenBankID' against NCBI's Identical Protein Groups (IPG) database (https://www.ncbi.nlm.nih.gov/ipg, 2025/01)[gg] to retrieve their genomic context (encoding genome/*contig*) for further investigating each hit's genomic nature and neighbourhood.[100] Genomic queries against IPG were performed using Biopython's (v.1.76) Bio.Entrez module either on a Python *ad hoc* script or via *bash* scripting, and protein GenBankID were employed using efetch utility, specifying ipg search (rettype) and XML results format (retmode). Retrieved identical clusters' relevant information (IPG_results_2.tsv) –namely 11 key attributes including IPG cluster ID, identical proteins' GenBankID, nucleotide (genome/*contig*) accession or 'GenomeID', identical proteins' genomic coordinates, strand orientation, organism metadata, and assembly data– was organised into a pandas (v.2.2.3) dataframe for further downstream processing and refinement comprising sieving, prefiltering and filtering. First, all entries originating from Swiss-Prot were excluded due to frequent lack of associated nucleotide accession numbers and incomplete positional information, which precludes reliable genomic localization. The remaining dataset was further prefiltered to remove entries lacking essential genomic coordinates or missing GenomeID. The genomic length of each remaining entry was then calculated (IPG_results_sieved.csv) and the dataset was further refined by selecting the longest representative entry –presumed to be the most complete or informative genomic representation– for each unique IPG group identifier (IPG_results_FILTERED.csv). This step enables the mapping of representative proteins to their most likely/informative genomic environments, facilitating the study of gene neighbourhood.

3.7.2. PHROG-based proteome annotation

In order to scrutinise genomic contexts of IPG-yielded representative proteins to determine its taxonomical nature and enable syntenic and functional neighbourhood analyses, automatically annotated proteomes were retrieved from NCBI GenBank nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore/) in GenBank (GBK) format from each genome's entry. Using the refined list of GenomeID from the filtered IPG results, *contigs* were fetched via Biopython's (v.1.76) Bio.Entrez package, setting retrieval type (rettype) as 'gbwithparts' and mode (retmode) to 'text' to ensure all annotated features were included, even for large genomes. Protein-coding sequences (CDS) located upstream and downstream of each IPG-yielded representative protein's *locus* from each GBK file were extracted by employing a Python script that parsed each file using Biopython's SeqIO package and filtered CDS features within a predefined symmetrical window of 50Mbp centred on the gene/protein of interest's midpoint, excluding the gene/protein itself. For each CDS in the defined window, protein sequences and associated protein_id tags were retrieved, along with a relative positional context ('upstream' or 'downstream') label; and were saved in FASTA format per genome/*contig* (Representative_Genomes_CONTEXT_NEW.tar.gz).

To functionally annotate CDS located near *loci* of genes/proteins of interest using PHROG, a custom HMM profile database was locally generated using MSA of curated PHROG families (viral of known function).

[cc] https://docs.python.org/es/3.10/library/itertools.html
[dd] https://biopython.org/docs/1.76/api/Bio.Blast.Applications.html
[ee] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
[ff] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
[gg] https://www.ncbi.nlm.nih.gov/ipg/docs/about/

HMM profiles were built for each PHROG alignment using hmmbuild form HMMER (v.3.4),[hh] concatenated, and indexed with hmmpress to create a searchable profile database.[101] Each genomic-context FASTA file was scanned against the database using hmmsearch with an E-value threshold of 10^{-3} and computation was distributed to 30 CPU. Domain-level results (--domtblout) were parsed to extract hits with E-value ≤ 0.01 and minimum query coverage of 50%. For each qualifying hit, the gene identifier, PHROG match, statistical metrics, and genomic relative position were recorded and saved in individual TSV files (PHROG_results_DEF.tar.gz). All files were then iteratively read and loaded as pandas (v.2.2.3) dataframes containing all PHROG-annotated CDS for each representative genome/*contig* excerpt. Each dataframe was then deduplicated to retain only the last occurrence of each unique 'GeneID', prioritizing the most confident (by E-value) annotation in order to resolve redundancy (PHROG_results_DEDUP.tar.gz).

To contextualize PHROG annotations with respect to genomic relative proximity, a cross-referencing step was performed between deduplicated PHROG results TSV files and their corresponding genomic context FASTA files. Each FASTA file was parsed to enumerate CDS, extract individual relative positional metadata ('upstream'/'downstream'), and assign sequential open reading frame (ORF) numbers. For each ORF, a relative distance to the IPG-yielded representative protein was computed and negative values were used for upstream (5') distances, whilst downstream (3') ones were marked as positive. All extracted data was merged with into the original TSV results files.

In order to obtain descriptive information regarding the absolute and relative genomic organization of representative genomes/*contigs*, all distance-annotated TSV files were first iteratively cross-referenced against a curated list of DNA-polymerase-associated PHROG identifiers (PHROG_polymerase_profiles.tsv) so that putative TP-pPolB synteny was assessed. A specific pPolB profile (phrog_1907) was individually searched amongst annotations for this purpose. For each genome, distance values for all ORF were analysed using numpy (v.2.2.0) and descriptive statistics were then computed into a dataframe (ORF_metrics_DEF+pol.tsv) –namely total ORF number, number of upstream/downstream ORF, an upstream/downstream count asymmetry , mean distance, standard deviation, median distance, maximum upstream/downstream distances, absolute minimum distance to representative protein, DNA polymerase and pPolB presence and distance– and visually inspected using Seaborn (v.0.13.2), matplotlib.pyplot (v.3.10.1) and Scikit-learn's (v.1.6.0) sklearn.preprocessing package.

3.7.3. geNomad-based genome annotation

In parallel to PHROG proteome annotation, all representative genomes were also annotated locally using geNomad (v.1.11.0) tool for MGE identification and provirus detection.[ii,102] Full-length nucleotide FASTA-formatted genome/*contig* sequences were retrieved beforehand by programmatically querying GenBankID against NCBI GenBank protein database (https://www.ncbi.nlm.nih.gov/protein/) (see 3.4.2.). For each genome, the corresponding FASTA file was symmetrically trimmed to a maximum length of 60kbp. Genomic coordinates for each IPG-yielded representative protein were retrieved from the IPG metadata table (see 3.7.1) and, for each genome, if the total length was ≤60kbp plus the protein's, length the entire sequence was retained; otherwise, a genomic window spanning ≤30kbp upstream and downstream protein ending positions was retained as context. New genomic contexts (Representative_Genomes_nt_CONTEXT.tar.gz) were first concatenated into a single multi-FASTA file (TP_genomes.fasta) with uniform line wrapping (60bp/line) using SeqKit (v.2.10.0).[jj,103] Briefly, individual FASTA files were padded to ensure trailing newlines, merged, reformatted with SeqKit, and collapsed of excess blank lines (tr -s '\n'). geNomad was then executed on 32 threads in "end-to-end" mode with score calibration enabled against geNomad database.

For viral nature or contamination identification amongst representative genome, the output classification summary file TP_genomes_virus_summary.tsv was parsed to identify geNomad-predicted genome contexts associated with viral nature or provirus contamination signatures. A list of genomes/*contigs* flagged as viral

[hh] http://hmmer.org/, http://eddylab.org/software/hmmer/Userguide.pdf, https://github.com/EddyRivasLab/hmmer
[ii] https://github.com/apcamargo/genomad/, https://portal.nersc.gov/genomad/
[jj] https://bioinf.shenwei.me/seqkit/, https://github.com/shenwei356/seqkit

was extracted and those genomes/*contigs* not included were classified as 'non-viral' into a list (geNomad_non_viral_genomes.txt) for downstream processing.

3.7.4. Viral-trace detection and non-viral set

To refine the classification of non-viral genomes, we cross-referenced geNomad-derived predictions with the PHROG-based viral ORF metrics dataset (ORF_metrics_DEF+pol.tsv) using pandas (v.2.2.3). Genomes previously flagged as non-viral by geNomad were compared against the list of genomes indeed annotated by PHROG; and, those not present or containing ≤3 predicted viral ORF were deemed likely to be confidently non-viral (geNomad_safe_non_viral_genomes_expanded.txt).

To investigate the functional classification of IPG-yielded representative proteins encoded in each allegedly non-viral genome/*contig*, we extracted COG (Clusters of Orthologous Groups) categories corresponding to each representative protein by cross-referencing each genome/*contig* identifier (see 3.8.) from the master results dataset. Dual-category annotations were split into individual COG letters to ensure accurate representation of multifunctional annotations. All resulting COG categories were compiled into a list (Non_viral_COG.txt) and where subsequently plotted using dplyr (v.1.1.4), tidyr (v.1.3.1) and ggplot2 (v.3.5.2) **R** packages.

To associate each starter TP with functional COG categories linked to related IPG-yielded representative proteins encoded in allegedly virus-free genomes/*contigs*, we parsed the master results dataset, and for each non-viral genome/*contig* identifier, we extracted the corresponding starter TP acronym and its IPG-yielded representative protein's COG category annotations. Dual-category annotations were also split into individual COG letters. The resulting associations were compiled into a dataframe, filtered for duplicates (Query_TP_eggNOG_COG.csv), and plotted for downstream functional enrichment and distribution employing both ggplot2 (v.3.5.2) and ggalluvial (v.0.12.5)[kk] **R** packages for flow/Sankey/alluvial plot creation.

3.8. eggNOG-mapper functional inference

Double-synonymous representative protein sequences were gathered in FASTA format by querying each GenBankID against NCBI GenBank protein database (https://www.ncbi.nlm.nih.gov/protein/) as described before (see 3.4.2.); and were submitted to EMBL eggNOG-mapper (v.2.1.12) webserver (http://eggnog-mapper.embl.de/) for functional inference and prediction.[104,105] E-value threshold was set to 0.001, whereas employed percentage identity was 40%, as a trade-off between accuracy and diversity retainment. Sequence-based functional prediction was performed by querying against both eggNOG 5 and novel ORFan-like protein families described by Álvarez del Río *et al*. (2022, 2024)[106,107] to ensure maximal functional diversity coverage. Results tabular TSV files containing functional annotations spanning several databases were parsed using Python pandas (v.2.2.3) to extract both an explicit functional description and COG (https://www.ncbi.nlm.nih.gov/research/cog/) categories for each protein for which inference had been possible.[ll] Extracted annotations were merged with the main results dataframe and their descriptive statistics plotted in R using ggplot2 (v.3.5.2).

3.9. TP:pPolB heterodimer inspection

3.9.1. Structural modelling

Some pPolB and TP:pPolB were structurally modelled using Google DeepMind's AlphaFold3 through its webserver (https://alphafoldserver.com/) access.[64] Some TP:pPolB heterodimeric models were also added an ATP molecule as computational substitute for *in vivo* dAMP initial nucleotide. In order to assess method-wise suitability for pPolB and TP, predictions were also carried out employing Chai-1, both programmatically and,[mm] before its algorithm release, through Chai Discovery webserver (https://lab.chaidiscovery.com/).[66] Chai-1 predictions were carried out either incorporating an MSA step to Chai-1 pipeline or not, in order to determine how accuracy loss might affect our specific protein set.

[kk] https://cran.r-project.org/web/packages/ggalluvial/vignettes/ggalluvial.html, https://cran.r-project.org/web/packages/ggalluvial/index.html
[ll] https://www.sbg.bio.ic.ac.uk/~phunkee/html/old/COG_classes.html
[mm] https://github.com/chaidiscovery/chai-lab

3.9.2. pPolB inspection

Structural predictions for pPolB and TP:pPolB heterodimers were both visually and computationally assessed for similarity. For this purpose, they were loaded into PyMOL (v.2.5.5) for all-against-all comparison through RMSD calculation employing align command. phi29 pPolB's crystallographically empirically-elucidated model (PDB: 2EX3) was set as gold standard template. Domains and/or peptide chains *ad hoc* were coloured for the sake of interpretability, taking 2EX3 as stencil.

4. Results

4.1. Predataset selection and starter dataset shortlisting

To enable a comprehensive exhaustive TP screening, including potential cellular ancestors, it was essential to assemble a robust and representative starting dataset of empirically validated or strongly inferred TP protein sequences. This dataset needed to capture maximal phylogenetic diversity to allow detection of distant orthologs during downstream analyses. Overrepresentation of closely related sequences would constrain the search space, resulting in redundant, phylogenetically narrow, and ultimately inconclusive outputs. Conversely, sufficiently diverse input sequences broaden the scope and depth of homology detection. However, the current pool of reliably annotated TP remains utterly limited, with many entries being either redundant or nearly identical. Therefore, our initial dataset was carefully tailored and curated to balance two key criteria: (i) maximize phylogenetic diversity by retrieving all available, reliable TP annotations across taxa, and (ii) minimize redundancy by filtering out highly similar sequences, retaining only those that contribute unique evolutionary information.

To address both redundancy and diversity, all non-identical entries –either empirically validated through state-of-the-art literature or annotated as TP in GenBank– were assessed for conserved genomic synteny with their putative cognate pPolB within their contigs or genomes. A so-called 'predataset' containing over 90 proteins (ANNEX TABLE) was hence tailored (see 3.1.1.), excluding known identical or highly similar TP –namely those from PRD1-like *Alphatectivirus PR4*, *A. PR5*, *A. L17*, *A. PR772*, PRDvermilion or PKJ.Ry.20.2,[18,19] *Betatectivirus GIL16*, *B. GIL01*, *B. Wip1* or *B. sole*,[12,21–23] *Deltatectivirus weeheim*,[24] *Cepunavirus Cp7* and *C. Cp5*,[108] *Claudivirus thornton* or *C. juan*,[109] *Tsarbombavirus tsarbomba*,[109] or *Harambevirus beachbum*–[109]. However, to minimise redundance, protein sequences were clustered using high-stringency CD-Hit algorithm at a 60% identity cut-off (see 3.6.1), yielding a total of 45 clusters, 34 of which were monotypic. Several identity thresholds were tested, but a trade-off between diversity preservation and over-fragmentation was sought, and choosing a conservative yet permissive for diversity cut-off was prioritised. For the generation of the starter dataset only one representative sequence was shortlisted from each cluster, reducing the dataset by 50% whilst retaining broad phylogenetic coverage.

Interestingly, even among non-monotypic clusters, only limited taxonomic clustering was observed, with related taxa often dispersed across separate clusters. For example, TP from the *Adenoviridae* family did not co-cluster with any other *Preplasmiviricota* members, nor was genus-level clustering achieved, highlighting the high phylogenetic divergence within this group. Some *Mastadenovirus* TP were distributed across two distinct clusters (#2 and #3) with low intra-cluster similarity. Only three *Aviadenovirus* TP (GAdV-4, DAdV-3, FAdV-4) clustered together (#5), and TP from *Sauropsida*-infecting *Siadenovirus* (CPAdV-2, PsAdV-2, RAdV-1) conformed a separate group. However, other *Mastadenovirus*, *Aviadenovirus*, and *Barthadenovirus* sequences remained as monotypic clusters, having failed to meet the similarity threshold. Overall, this points to poor taxonomy-based clustering, though a minor host-related (trophic/ethologic) pattern can be slightly discerned. A similar trend was seen in Tectiviridae: apart from PRD1-like *Alphatectivirus* TP co-clustering in cluster #10, most were individually clustered, lacking consistent family- or genus-level grouping. *Caudoviricetes* TP also exhibited scattered clustering. The largest cluster (#0) included *Salasvirus* and closely related *Beecentumtrevirus* (Nf, Goe1, hmny2), while other *Salasmaviridae* TP from *Claudivirus*, *Hemphillvirus*, *Klosterneuburgvirus*, *Layangcvirus*, and *Huangshavirus* were scattered across clusters #1, #6, and several monotypic clusters. TP from *Anjalivirus* and *Bundooravirus* formed micro-clustering groups, and the most taxonomically coherent cluster (#4) included all three *Guelinviridae Clostridiales*-infecting *Brucesealvirus* TP (ΦZP2, ΦCPR7, ΦCPV4). Even closely related TP from *Cyanobacteriota*-infecting *Kyanoviridae* (S-CREM2, S-SCSM1, DSL-LC02) failed to co-cluster. These

results underscore the extensive sequence-level diversity among TP, even within closely related taxa, which entails a major hindrance for TP screening and scrutiny.

All 45 shortlisted TP were subjected to AlphaFold2 high-accuracy modelling (see 3.1.2.), yielding high-quality confidence scores for the majority of proteins. As shown in Figure 3, most predicted models achieved pLDDT scores above 80 and pTM scores between 0.7 and 0.8, reflecting unexpectedly high reliability despite the minimal sequence conservation among TP.

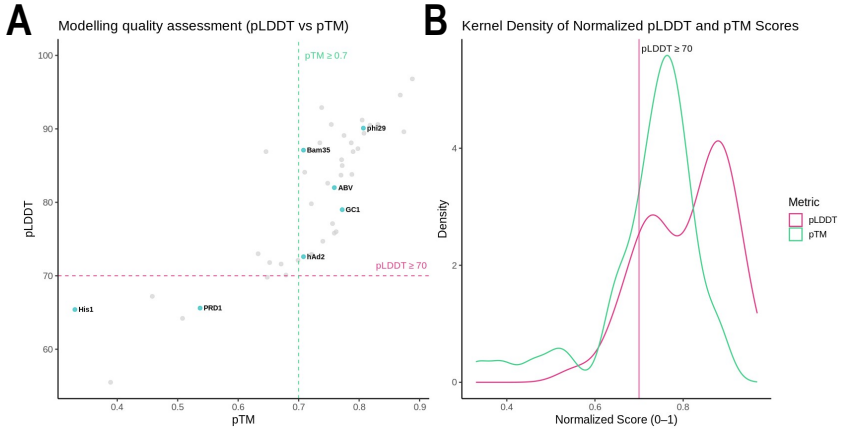

Figure 3. a) Quality assessment pLDDT-pTM score distribution of the 45 AlphaFold2 models for the shortlisted TP. Cyan dots depict representative empirically validated TP. Upper confidence thresholds for pLDDT (in magenta) and pTM (in teal) were represented with dashed lines, and the upper-right section (delimited by pLDDT ≥ 70 and pTM ≥ 0.7) gridles optimal models. b) Normalised pLDDT (magenta) and pTM (teal) score *kernel* density distributions for initial TP's AlphaFold2 models. Normalised quality threshold, set at 0.7, is represented with a magenta horizontal line.

4.2. Initial dataset phylogenetic reconstruction

Although sequence-based CD-Hit clustering highlighted profound divergence among the initial set of representative TP sequences at a >60% identity threshold (see 4.1.), we aimed to further investigate whether any minimal sequence or structural conservation was still preserved, and how such features are phylogenetically distributed –particularly in relation to established viral whole-genome phylogenies and current ICTV taxonomic classifications–.[5,25] Specifically, our goal was twofold: first, to evaluate the extent and directionality of intra-dataset divergence despite the acknowledged high variability, and second, to generate comparative sequence- and structure-based phylogenetic reconstructions of TP to determine how well TP-based phylogenies align with broader genome-wide evolutionary trajectories and taxonomic consistency. Additionally, we sought to compare phylogenetic assessments derived from sequence versus structure, given that protein fold and function are often more conserved than primary sequence. Notably, to date, structural alignment methods have not been systematically applied to phylogenetic reconstruction in viruses, and no existing studies have examined the TP:pPolB system using standardised structural-alignment-based phylogenetics.

To further address the putative evolutionary relationships and structural conservation of TP, initial TP sequences were first aligned using Clustal Omega (ClustalΩ), and a sequence-based phylogram was generated through neighbour-joining (NJ) with a distance matrix, while a structure-based guide tree was secondly inferred using FoldMason by comparing three-dimensional models of the corresponding proteins (see 3.5.).[76] This dual approach enabled the resolution of both evolutionary divergence at the sequence level and potential convergences or conservation at the structural level (Figure 4).

ClustalΩ-yielded sequence-based phylogram revealed distinct, well-supported clades that in some cases correspond to established taxonomic classifications. Some *Tectiviridae* family members formed a coherent clade, with clear subdivisions corresponding to recognized genera: *Betatectivirus* (AP50, Bam35, Sato) and *Epsilontectivirus* (Toil). This phylogenetic coherence supports a putative monophyletic origin for these tectiviral TP and their genus-level evolutionary divergence. However, at the sequence level, deviations from expected taxonomic proximity were observed. For instance, the TP from the remaining *Tectiviridae* genera *Alphatectivirus* (PRD1), *Gammatectivirus* (GC1) and *Deltatectivirus* (Forthebois) were externally clustered

into a completely different top-level clade, alongside *Adenoviridae* and –unexpectedly– some *Kyanoviridae* (S-CREM2 and S-SCSM1). On the other hand, beta/epsilontectiviral clade included *Madridviridae* (class *Caudoviricetes*) member *Cepunavirus Cp1*, that despite being taxonomically closer to *Salasmaviridae* and other Φ29-related phages, was unexpectedly placed alongside *Epsilontectivirus toil*. *Guleliniviridae Brucesealvirus* ΦZP2, which exhibits closer evolutionary ties to Φ29-like *Salasmaviridae*, was positioned as the outgroup to the entire beta/epsilontectiviral clade. These discrepancies may highlight the limitations of sequence-level resolution in capturing deeper evolutionary trajectories. Additional phylogenetic incongruities include the clustering of *Abadenavirae* kingdom *Vibrio*-infecting phages –*Paulavirus viph1080o* and *Livvievirus viph1249b*– more closely with the presumed beta/epsilontectiviral clade than its sister family *Adenoviridae*. Furthermore, the TP from the archaeal virus ABV was grouped alongside beta/epsilon and *Autolykiviridae* TP, a result that echoes earlier classifications placing these taxa together within the now-defunct ~~*Siphoviridae*~~ family infecting marine or halophilic bacteria. Together, these observations underscore the challenges of reconstructing accurate evolutionary relationships based solely on primary sequence data.

While at the sequence level *Adenoviridae* family conformed a single top-rank coherent clade (Figure 4.a.), it is clearly bifurcated, separating *Mastadenovirus* (hAd2, BAdV-3, PAdV-4, EAdV-1) and *Aviadenovirus* (FAdV-8, GAdV-4, DAdV-1, PiAdV-1, CrAdV-1) lineages. *Aviadenovirus* TP clustered with shorter inter-branch distances, reflecting lower divergence in *Adenoviridae* with avian hosts. Conversely, the *Mastadenovirus* members were more dispersed, consistent with their greater host range and the putative adaptive evolution of TP in mammalian systems. *Siadenovirus* CPAdV-2, while taxonomically related to *Aviadenovirus*, branched separately, suggesting moderate divergence from canonical TP sequences. Furthermore, alpha/gamma/deltatectiviral TP were externally co-clustered with adenoviral clade, pointing out a possible common ancestral origin for TP acquisition as pointed by Krupovic *et al*. based on structure-based pPolB phylogenetic reconstructions.

For *Caudoviricetes*, as for Adenoviridae family, the *Salasmaviridae* family formed a tightly grouped and taxonomy-aware clade, including most Φ29-taxonomically-related genera: Salasvirus (Φ29), Huangshavirus (DLc1), *Harambevirus* (Harambe), *Bundooravirus* (WhyPhy), *Karezivirus* (Karezi), *Bahkauvirus* (Chedec11) and *Layangcvirus* (LY3); indicating high sequence conservation of TP among *Bacillati*-infecting Φ29-like phages. On the other hand, TP belonging to order-orphan *Caudoviricetes* were grouped into a single clade branched into two markedly different subclades. Firstly, TP from phages infecting *Micrococcales* (Evcara, Curie, Ayka, PineapplePizza, Tillums and Voltaire) grouped together, potentially reflecting a shared evolutionary lineage or HGT of replication modules through common hosts. While a few of them have been recently taxonomically categorised, most of them remain unclassified, but according to their TP sequences, Ayka seems to be related to phages from genus *Anjalivirus*, whereas Evcara and Curie might be encompassed within genus *Amherstviurs*. The second subclade comprises TP from highly divergent viral genera infecting a broad range of hosts, including Gram-positive bacteria or *Archaea*.

The phylogram also highlighted several deeply branching or anomalously placed taxa. TP of *Epsilontectivirus toil* appeared as one of the most basal lineages (branch distance around −0.16), consistent with its unique TP architecture and divergent genome features. Similarly, ΦYS61, with the longest branch length in the phylogram (around 0.44), emerged as the most divergent TP, potentially reflecting either ancient divergence, a homoplastic origin or adaptive sequence-noticeable structural innovations distinct from other viruses. Interestingly, TP from *Kyanoviridae* phages are distantly grouped and *Synechococcus*-infecting DSL-LC02, exhibited higher divergence (branch length >0.4), suggesting either a more ancient split or unique functional adaptations.

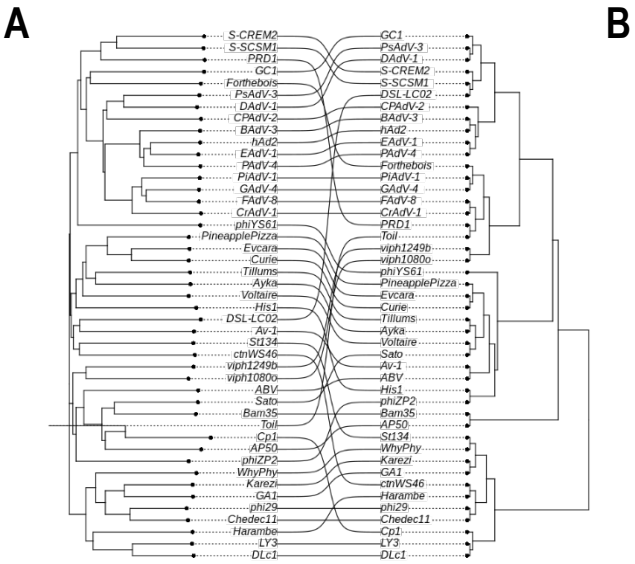

Figure 4. Tanglegram comparing both **a)** ClustalΩ sequence-based phylogram for initial TP dataset, and **b)** FoldMason structure-based cladogram for initial TP dataset. Connecting lines link each TP's placement in each phylogenetic tree.

The structure-based cladogram output by FoldMason largely mirrored the topology of the phylogram, yet exhibited key differences attributable to structural convergence (Figure 4.b.). Moreover, FoldMason's enhanced sensitivity relies on structure-driven domain-based alignments that are not usually captured when performing sequence-based approaches. Structural alignment of the initial TP dataset yielded an exceptionally low MSA-LDDT score of 0.285, reflecting the profound structural heterogeneity within the dataset and further substantiating the high divergence already evidenced at the sequence level.

Two major distinct clades were revealed and, while the first one mostly recapitulated the sequence-based grouping of TP from *Salasmaviridae* and other Φ29-related genera, it relocated TP belonging to ctnWS46 and *Andhravirus St134* adjacent to those of Harambe and WhyPhy, respectively. This rearrangement likely reflects common host-driven structural particularities, as the newly co-clustered phages all infect *Bacillales* hosts. The former sequence-based beta/epsilontectiviral cluster was absent in the structure-based reconstruction: TP from *Tectiviridae* are sparsely distributed among those of either *Adenoviridae* (non-betatectiviral TP) or non-*Salasmaviridae Caudoviricetes* (Sato), with the exception of *Betatectivirus AP50* and *Bam35*, which are consistently co-clustered as outgroup of the second major clade. Notably, at the structure level, TP from PRD1 and Toil co-cluster, alongside those of *Aviadenovirus*. Furthermore, GC1 was also positioned among *Aviadenovirus*, suggesting a closer yet unexpected evolutionary proximity. Despite consistent *Mastadenovirus* TP clustering, those from *Aviadenovirus* and *Siadenovirus* are diffusely distributed and are interspersed with *Tectiviridae*, *Autolykiviridae* and *Kyanoviridae* TP. These interleaving likely reflects structural idiosyncrasies that may correspond to host-specific functional constraints. Strikingly, unlike the structure-based approach, the structural analysis revealed a tightly grouped micro-cluster encompassing all Kyanoviridae TP, suggesting a conserved fold-level organization despite broader sequence divergence. Finally, TP structural assessments support Ayka relatedness to *Anjalivirus* phages and Curie and Evcara inclusion within genus *Amherstvirus*.

Structural analysis via FoldMason proved particularly useful for resolving ambiguous placements and for suggesting homology among poorly characterized proteins, especially in cases where sequence divergence was high. The combined data underscore the utility of TP as molecular markers for viral systematics and evolution and provide a robust framework for classifying emerging or uncharacterized viral lineages based on core replication protein architecture.

4.3. The analysis of TP structures enables the validation of intrinsic features of TP

Once high-accuracy structural models for shortlisted starter TP were obtained, we wanted to further ascertain whether TP can be systematically screened and imputed. Although TP lack conserved catalytic or structural motifs that would ease their computational identification and imputation, as described above,

function-related structural elements have arisen, when in combination with relative synteny conservation, as possible key signature elements. Structure-inferred features have been suggested relying on the few available empirical structural data, yet they have not been evaluated to a wider extent to this date. Firstly, it has been largely suggested that, mechanistically, TP must bear specific regions of high positive surface charge density that enable interaction and binding with densely negatively charged DNA's pentose-phosphate backbone, whilst others might be highly electropositive for facilitating interaction with basic regions of cognate pPolB.[15] Despite TP are suggested to be asymmetrically charged, surface charge density distribution symmetry elucidation in proteins entails major hindrances as they are topologically and intrinsically asymmetrical. To address this, all atoms of each TP were projected into a Cartesian three-dimensional space which was then partitioned (see 3.2.1.) and overall electrostatic potential ($\Delta\psi$) difference between spatial divisions was averaged ($\Delta\Delta\overline{\psi}$). For our surface charge density asymmetry estimation model, a trade-off between computation costs and accuracy was sought. Therefore, we applied a simplified yet accurate Debye-Hückel linearisation of Poisson-Boltzmann equation for point electrostatic potential and divided the Cartesian grid into hemispheres or octants (see 3.2.1.). Although electrostatic potential distribution modelling accuracy might not be high, the relative assessments of charge asymmetry can be effectively captured by applying such approximations. Both hemisphere- and octant-based models produced consistent results, with the coarser hemisphere-based method still providing reliable asymmetry estimates.

Statistical analysis (Figure 5) of $\Delta\Delta\overline{\psi}$ distributions between TP and control groups using the Wilcoxon test revealed a significant difference in median asymmetry scores. TP values followed a bimodal kernel density distribution (modes at -0.02 eV and -0.045 eV; median and mean ≈ -0.02 eV), while the control distribution showed only a subtle bimodality. As control sample size increases, its distribution tends toward a Poisson-like negative curve centred near -0.01 eV, reflecting the near-symmetrical charge profiles of most natural proteins. Additionally, although Cohen's *d* suggests a small effect size –indicating minimal average difference–, "the rank-biserial correlation reveals a small but meaningful effect in ranking: TP values tend to rank higher, while the control group contains a few extreme outliers, keeping group means close.

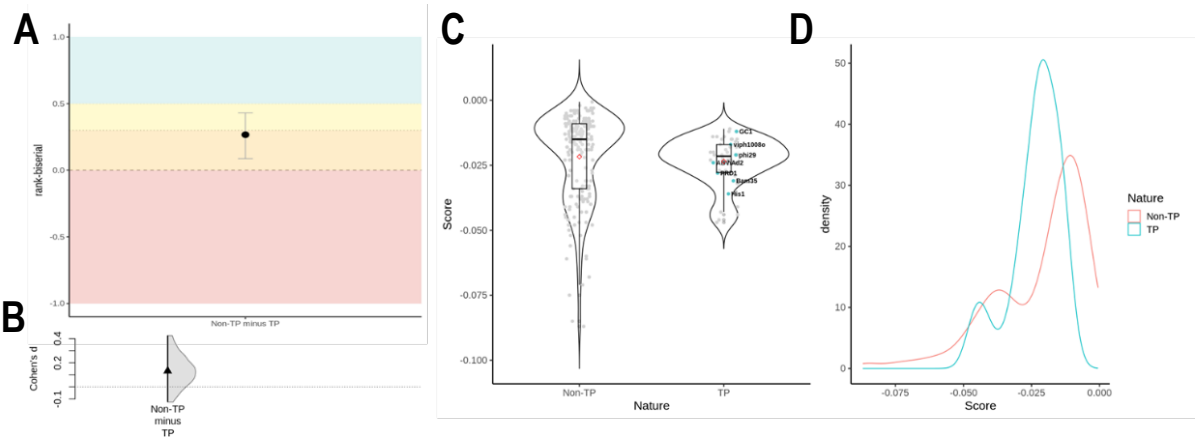

Figure 5. a) Rank-biserial correlation in black, including confidence interval. Background colours indicate effect size ranges according for rank-biserial correlation (red – no overall effect, $r_{rb} \leq 0$; orange – small effect, $0 < r_{rb} < 0.3$; yellow – moderate effect, $0.3 \leq r_{rb} \leq 0.5$; green – large effect, $r_{rb} > 0.5$). b) Cohen's *d* and confidence interval. c) Violin plots for TP and control asymmetry score distributions. Each dot jittered in light grey corresponds with a protein, whereas cyan dots depict representative empirically validated TP. Distribution means are represented as red rhombus, while the median is presented within each inner boxplot. d) Kernel density distributions for TP (cyan) and control (red) sets. Asymmetry score may be measured in terms of electrostatic potential units (eV).

While our charge density asymmetry approach provides valuable insight into TP identification, it faces a key limitation: TP are a subset of the broader set of all natural proteins (represented by controls). This subset relationship complicates significance interpretation, as the true centre of protein asymmetry distribution is unknown and lacks a defined standard. Moreover, although our control dataset spanned diverse genomes across all dominions of life and included proteins of varying sizes and functions –rather than TP-like proteins–statistical significance was still achieved, strengthening the robustness of our findings. Overall, we demonstrate that surface charge asymmetry, particularly scores between -0.045 eV and -0.02 eV (median ≈ -0.02 eV), can serve as a useful discriminative parameter for TP prediction. In conclusion, we have implemented a straightforward and quick method for analysing protein charge asymmetry that, although useful, it is not definitive by itself and may be needed to be combined with additional criteria.

As previously introduced, TP are thought to be enriched in α-helical content, as is common among DNA-binding proteins. However, this hypothesis has not been systematically validated across a broad TP dataset and remains based on a few empirical models. To address this, we quantified the proportion of each TP's atoms associated with specific secondary structure elements, aiming to identify whether any topological features are particularly enriched and potentially characteristic of TP. As described above, we applied the DSSP algorithm to assign secondary structure categories based on local topology, thus embedding each model with secondary structure annotations suitable for compositional analysis. While DSSP can differentiate eight structural elements –including α-, π-, and 3_{10}-helices; β-strands; β-bridges; β-/γ-turns; polyproline helices; and loops– we grouped all helical and all β-sheet-related structures into broader categories for simplified yet accurate comparison (see 3.2.2. for details). Although we did not evaluate specific folds or supersecondary motifs, this dictionary-based structural profiling effectively revealed secondary structure biases across the TP dataset.

As for surface charge density distribution asymmetry calculation validation, secondary structure α and β scores were computed for both TP (N = 45) and control subsets (N = 190). Statistical analysis of both distributions (see 3.2.2.) revealed marked differences between both scores and subsets (Figure 6). While Wilcoxon rank-sum test found a significant statistical difference in median α score of the two subsets, median β contents did not achieved statistical significance. Upon effect size estimation, both Cohen's d and rank-biserial correlation pointed out that differences between TP and control subsets, while statistically valid, range from small to moderate in magnitude. Most biochemically characterised TP display α contents above 80% (TP distribution median at 75%), whereas just a few of them (including GC1, whose folding contains an unusually high β-content) falling below 50%. Interestingly TP's kernel density distributions for α and β scores are almost symmetrical Poisson-like distributions comprising α contents spanning 5%-90%. On the other hand, although control ones are almost symmetrical, α score distribution's shape is almost bimodal (50% and 90%) gaussian, particularly depressed around TP's median/mode. This effect can also be appreciated in β content distributions, although subset mode opposition is subtler due to both exhibiting closer means and similar Poisson-like shapes.

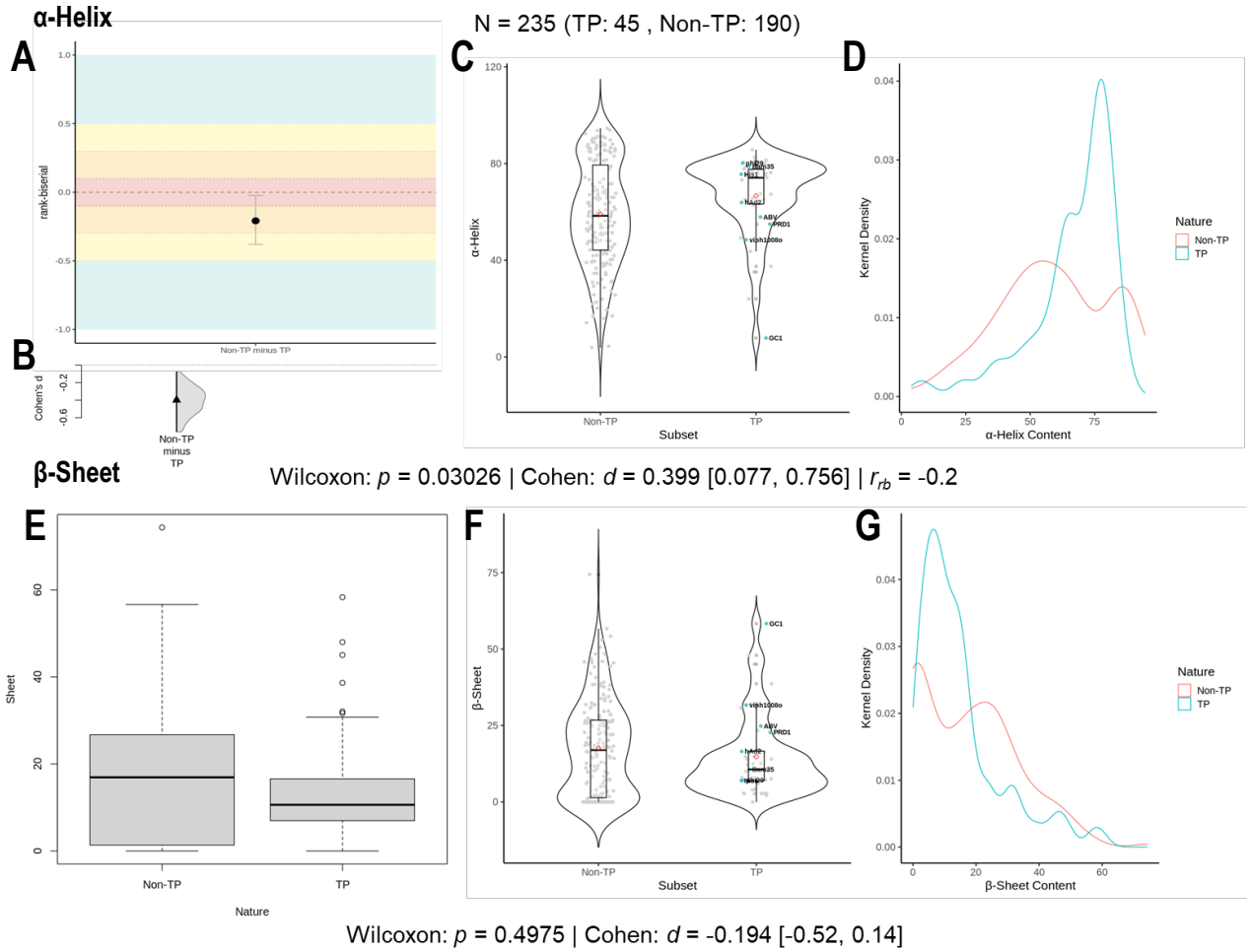

Figure 6. Statistical scrutiny for α (upper half panels) and β (lower half panels) content distributions for TP and control subsets (see Figure 4 for reference). a) Rank-biserial correlation for α content in black, including confidence interval. b) Cohen's *d* with confidence interval for α content. c) Violin plots for TP and control α score distributions. Metrics are depicted as in Figure 4. d) Kernel density distributions for TP (cyan) and control (red) subsets. α content should be measured as a percentage. e) Boxplots for TP and control β score distributions. f) Violin plots for TP and control α score distributions. Metrics are depicted as in Figure 4. g) Kernel density distributions for TP (cyan) and control (red) subsets. β content should be measured as a percentage.

While α-helix enrichment provides meaningful insight into TP identification, its standalone discriminative power remains limited. Despite α-helix enrichment is a strong and recurrent feature among TP, not all fully-helical or α-enriched proteins can be considered as TP. Some other proteins –e.g., DNA-binding proteins such as classical transcription factors with characteristic HLH, bHLH or bHLH-zipper motifs– might also bear highly α-helical structures. Moreover, not all TP are uniformly enriched in α-helices; despite a central tendency around 80%, some TP deviate from this pattern. Likewise, although we have hereby demonstrated that α content serves as a useful indicator of TP-like nature it lacks the specificity required for definitive classification. As for charge distribution asymmetry, it cannot be employed alone as a standalone parametric criterion for TP imputation, requiring additional complementary parameters to achieve robust and unambiguous TP annotation.

The third structurally-inferred feature for TP imputation hereby evaluated was DNA binding capacity. Nevertheless, it is not as straightforward to assess as the previous two criteria, as there are no common universal structural low-complexity determinants of DNA binding and it has not been yet suggested any sequence or structural pattern for TP:DNA binding. Consequently, as described in state-of-the-art reports on this subject, the most feasible strategy for evaluating this feature necessarily relies on structure- and sequence-informed Bayesian classification models.[110,111] Given that the TP:DNA binding mode remains unresolved and –due to utterly high TP divergence– may not be conserved across all TP taxa, any classifier trained on limited TP data risks introducing confounding biases or misclassification artifacts due to the

underrepresentation. Furthermore, the limited number of available TP structures compared with other DNA binding protein families further limits representativeness during training and would severely constrain the statistical robustness and generalizability of any novel classifier. To circumvent these limitations, we opted to test DNA-binding capacity using an existing pre-trained classifier rather than developing a new model. As previously introduced (see 3.2.3.), we employed the web-based DNABIND classifier by András Szilágyi (Hungarian HUN-REN), which integrates sequence and structural features and remained accessible.[88] Therefore, we tested Szilágyi's classifier against our TP-control dataset (as both TP's and each controls' DNA-binding capacity was known beforehand) and evaluated its performance using standard classification metrics –false negative count, false positive count, accuracy, precision, recall, specificity and F1-score–.

As shown by the classification metrics (Figure 7), DNABIND displays a marked class-dependent performance bias. For Class 0 ("Non-binding"), the model achieves high precision (0.95) but low recall (0.46), indicating that while most predictions labelled as 'non-binding' are correct, a substantial proportion of actual non-binding proteins are misclassified. In contrast, for Class 1 ("DNA-binding"), the model shows high recall (0.90) but low precision (0.30), meaning it successfully captures most actual DNA-binding proteins but also misclassifies a considerable number of non-binding proteins as binders. In essence, the model is more sensitive to detecting DNA-binding proteins, but this heightened sensitivity comes at the cost of a high false positive rate. Additionally, the classifier demonstrates a prediction imbalance, with a 61% frequency of Class 1 predictions versus 39% for Class 0. This skew toward the DNA-binding class may introduce further bias and should be taken into account. While the majority TP were properly classified as Class 1, most divergent ones (GC1, ABV, and *Kyanoviridae* S-CREM2 and S-SCSM1) were misclassified, either pointing to concerns regarding their nature or showcasing the classifier's bias against uncommon DNA binding proteins.

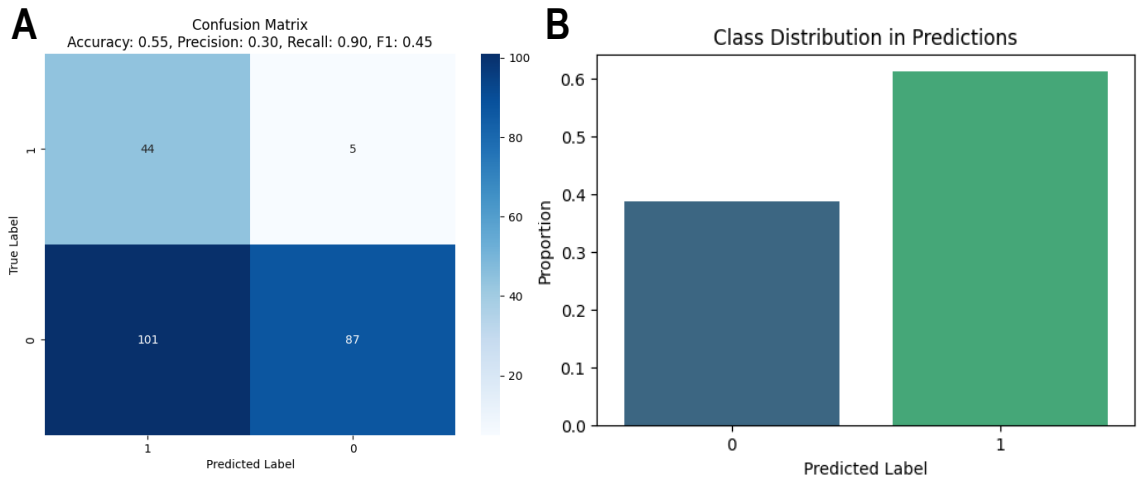

Figura 7. a) Confusion matrix for Szylági's DNABIND classifier. Heatmap shows classification counts and biases. Class 0 corresponds to 'DNA-binding' label, whereas Class 1 refers to 'Non-binding' label. As for supervised learning, each evaluated protein structure had a precomputed known label (true label) and was predicted a DNA-binding label by DNABIND (predicted label). True and predicted labels' counts were plotted once against the other. Each entry could be either properly (0 or 1) classified, matching true and predicted labels, or can be misclassified (type I error: 0-1; type II error: 1-0). **b)** Percent predicted label ratios for DNABIND classifier against joint TP-control dataset. Class 0 corresponds to 'DNA-binding' label, whereas Class 1 refers to 'Non-binding' label.

In line with the results presented above, despite the absence of any biochemically defined sequence- or structure-based DNA-binding motif, TP can still be reliably classified using generalist DNA-binding prediction models. An *ad hoc* trained classifier should be implemented for an appropriate classification. Overall, our data show that, despite sequence-level diversity, TP share a set of standardized, quantifiable structural features that can be leveraged for TP imputation. While none of the three criteria individually (surface charge asymmetry, α-helix content, and predicted DNA-binding capacity) suffice for definitive classification, their combination –particularly when coupled with sequence-derived context such as synteny conservation– offers a robust and comprehensive framework for TP identification. These parameters, taken together, provide a foundation upon which a dedicated Bayesian classifier may be effectively developed and trained.

4.4. Sequence- and structure-based homology searches yield complementary results

As previously detailed (see 3.4.1. and 3.4.2.), both sequence- and structure-based homology searches were performed using initial dataset 45 shortlisted TP included in order to (i) comprehensively capture distant homologs by inspecting both layers of diversity, and (ii) qualitatively and quantitatively compare methodological performance across divergent input sequences. Structure-based homology search was conducted through Foldseek's RESTful API querying against major non-metagenomic structure databases. Metagenomic-data-fed databases were excluded due to lacking relevant taxonomic information for TP phylogenetic trajectory and evolutionary history reconstruction. This search yielded 24,522 raw candidate homologous hits spanning all domains of life. After cross-referencing for GenBankID and accounting for entries with multiple mappings, the dataset expanded by 34.33% to 32,722 entries. To enhance data quality and reduce redundancy, a three-stage prefiltering curation pipeline was implemented (see 3.4.1.): (i) Sieving out low quality entries based on a compound threshold encompassing E-value, Foldseek homology probability and query alignment coverage; (ii) hit sorting based on combined quality metrics; and (iii) deduplication by GenBank ID, retaining only the entry with lowest E-value per ID. After testing several stringency parameter combinations, optimal filtering thresholds were set at E-value (ξ_E) ≤ 0.01, homology probability (ξ_H) ≥ 0.5, and query coverage (ξ_L) ≥ 0.3, overwhelmingly reducing the dataset by 92.14% to 2,572 candidates (Figure 8). We sought a trade-off between candidate load and diversity preservation. Variations in E-value cut-off produced minor fluctuations (±10^3 hits), while changes in homology probability had negligible impact if the other criteria were fixed. Final deduplication reduced the set by an additional 25.58%, yielding 1914 shortlisted candidate homologs –a cumulative 94.15% reduction–. Taxonomic assignation and inspection yielded 1211 unique TaxID –a cumulative 96.3% reduction–.

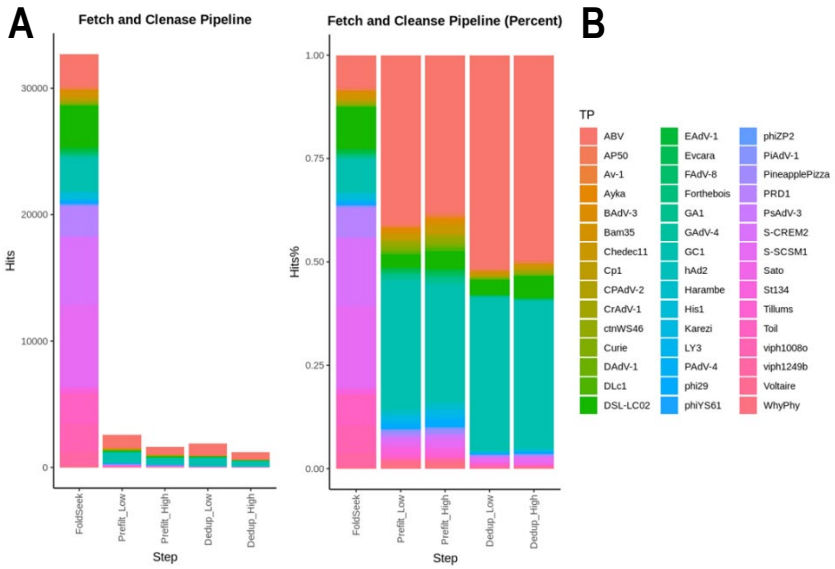

Figure 8. Prefiltering pipeline performance on Foldseek results. **a)** Depicts raw counts per curation stage, whereas **b)** represents percent per-TP contribution to each cleansing stage dataset. Stacked bars were employed per TP, and each TP was represented with a different colour. Prefiltering curation steps included initial raw GenBankID-including dataset ('FoldSeek'), quality-based sieving out ('Prefilt') and deduplication ('Dedup') datasets. For the sake of acknowledging how different quality parameter combinations influenced prefiltering and deduplication performance, tow compound threshold sets were included: low stringency cut-off ('Low') comprising [ξ_E ≤ 0.01, ξ_H ≥ 0.5, ξ_L ≥ 0.3] and selected for further analyses, and high stringency threshold ('High') comprising [ξ_E ≤ 10^{-3}, ξ_H ≥ 0.5, ξ_L ≥ 0.3].

Beyond its quantitative impact, the prefiltering step substantially altered the qualitative composition of the hit dataset. Initially, TP from *Kyanoviridae* S-CREM2 and S-SCSM1 were the major contributors to raw hit dataset, distantly followed by TP from DSL-LC02 (*Kyanoviridae*); GC1, PRD1 and Toil (*Tectiviridae*), viph1080o (*Autolykiviridae*), and ABV (*Ampullaviridae*) (Figure 8.b.). However, after applying the compound prefiltering curation step, dataset architecture shifted markedly and remained consistent across both prefiltered and deduplicated derived datasets. Notably, S-CREM2 and S-SCSM1 experienced prefiltering-driven drastic reductions in associated hit counts –by 99.96% and 99.69%, respectively– dropping their representation from ~15-20% to just ~1-2%. Conversely, TP from ABV and GC1 were minimally affected

during sieving, retaining most of their associated hits. Consequently, their relative contribution increased substantially, representing ~35-45% of the prefiltered dataset and ~42-52% post-deduplication. While the curation process sharply –yet heterogeneously– reduced hit counts for all TP, Ayka's TP lost all associated candidates, likely due to poor structural homology resulting from extreme divergence. Several other TP –Tillums, PsAdV-3, Cp1, BAdV-3, Forthebois, His1, Karezi, ΦYS61, and ΦZP2– retained only a single hit each after deduplication. This widespread reduction underscores the abundance of initially captured low-confidence or marginally similar hits that failed to meet the compound threshold criteria, highlighting the importance of rigorous quality control in structural homology analyses.

In parallel to Foldseek-based homology search, we performed a PSI-BLAST-driven sequence-based homology search against the nr-clustered (90%) database to capture distant homologs across both sequence and structural combined layers of diversity (see 3.4.2.). The clustered version of nr was selected to reduce computational burden while preserving sequence diversity and representativity. After testing multiple parameter sets, we optimized the search by setting the number of iterations to 3 and the per-iteration E-value threshold (ξ_E) to $\leq 10^{-3}$, balancing profile sensitivity and diversity explored with the need to prevent profile drift and false positive bias towards distant homologs. This search yielded 12,855 raw candidate hits. As for Foldseek candidate homologs, an analogous three-step curation pipeline –quality-based *ad hoc* sieving ($\xi_E \leq 10^{-3}$ and query coverage $\xi_L \geq 0.3$), hit sorting, and GenBankID-based deduplication– was applied. Sieving out low quality entries reduced the dataset marginally by 4.51% (12,272 entries), while deduplication further reduced it by 88.64%, resulting in 1,395 high-confidence hits –cumulative 89.15% reduction–. Although nr-clustered was employed to optimize PSI-BLAST efficiency, resulting clusters were not further expanded and inspected since the retained hits already preserved sufficient sequence diversity and the number of candidates was large enough. In fact, to facilitate downstream phylogenetic reconstruction and avoid redundancy, candidate hits from both Foldseek and PSI-BLAST were subsequently merged and clustered by sequence similarity (threshold set at 0.6), prioritizing strong-homology candidates for robust evolutionary inference. Taxonomic assignation and inspection of PSI-BLAST curated dataset yielded 763 unique TaxID –a cumulative 94.06% reduction–.

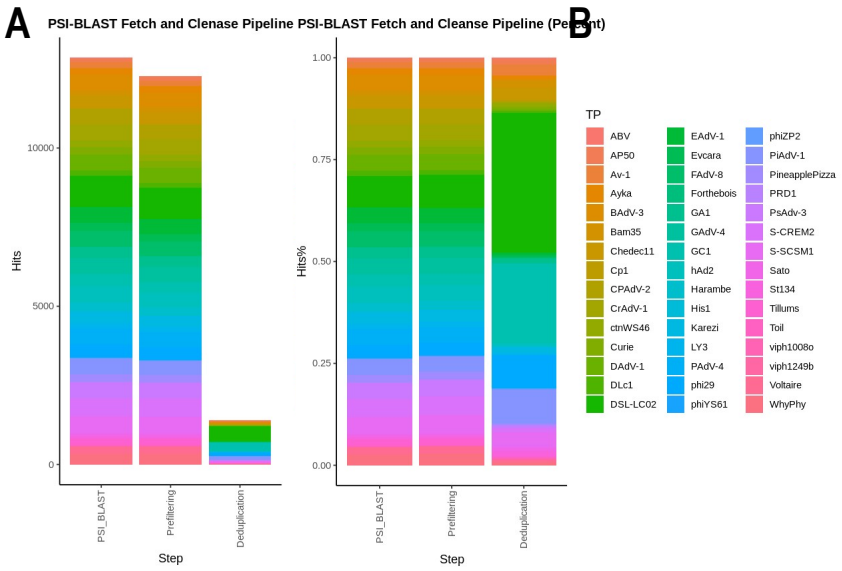

Figure 9. Prefiltering pipeline performance on PSI-BLAST results. **a)** Depicts raw counts per curation stage, whereas **b)** represents percent per-TP contribution to each cleansing stage dataset. Stacked bars were employed per TP, and each TP was represented with a different colour. Prefiltering curation steps included initial raw PSI-BLAST-yielded hit dataset ('PSI_BLAST'), quality-based sieving out ('Prefiltering') and deduplication ('Deduplication') datasets. Quality compound threshold for prefiltering was set at [$\xi_E \leq 0.01$, $\xi_L \geq 0.3$].

As for Foldseek- derived dataset, the curation pipeline reshaped the qualitative composition of PSI-BLAST's hit dataset (Figure 9.a.). While prefiltering itself had minimal impact on overall dataset composition, deduplication substantially altered its architecture and parental TP relative contributions (Figure 9.b.). Initial TP contributed relatively evenly to the raw dataset, except for *Kyanoviridae* DSL-LC02 TP, which accounted for ~10% of hits. However, deduplication stage had asymmetric effects: CrAdV-1 and PAdV-4 adenoviral TP lost all associated hits, while DSL-LC02 and GC1 TP underwent moderate hit count reductions (~29-50%), increasing their relative representation in the final deduplicated dataset to ~30% and ~20%, respectively. TP from PiAdV-1 and Φ29 also became prominent, each contributing ~15% of the deduplicated dataset. Conversely, several TP –BAdV-3, Forthebois, His1, ΦYS61, and ΦZP2– retained only a single hit each. Notably, TP of Forthebois, His1 and BAdV-3 each preserved just one hit per method, likely indicating high divergence. Additionally, ABV did not outstand as major contributor in sequence-based search, reinforcing methodological complementarity. The sharp reduction observed post-deduplication –but not during prefiltering– underscores sequence-based search tendency to identify many reliable yet closely related putative homologs, which become redundant in the context of large or divergent datasets.

To assess methodological differences between structure- and sequence-based homology searches cross-referenced both curated and non-curated Foldseek- and PSI-BLAST-derived datasets to identify shared candidates and overlapping taxonomic assignments. GenBankID comparison (Figure 10; see 3.4.2.) revealed that only a single hit was common between the raw, non-deduplicated datasets, and none remained shared post-deduplication. Taxonomic overlap was slightly higher but still minimal, with only 8% of TaxID shared across methods after curation.

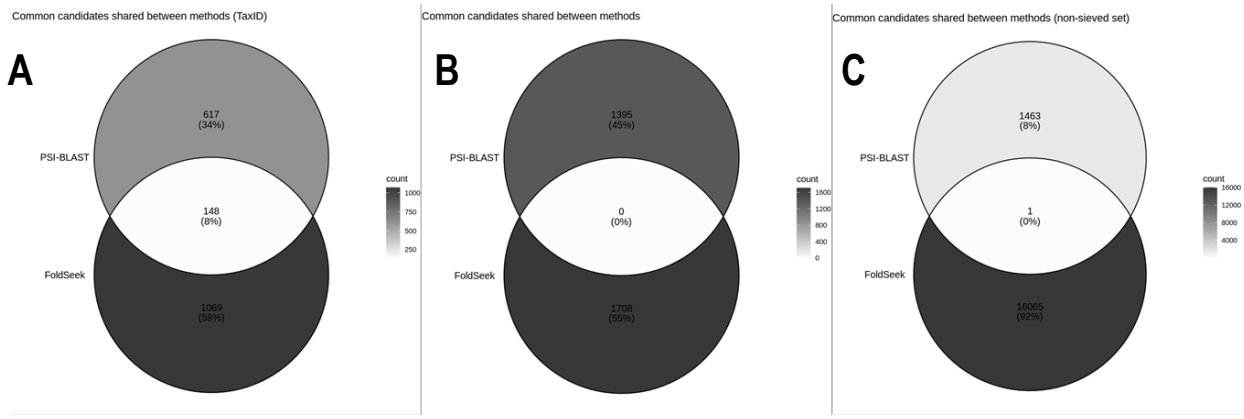

Figure 10. Venn diagrams representing Foldseek/PSI-BLAST datasets cross-method overlapping: **a)** TaxID of curated datasets, **b)** GenBankID of candidate hits after sieving, and **c)** GenBankID of candidate homologs before curation. Gray scale depict the magnitude of elements belonging to each or both method-wise datasets (the darker, the larger).

A detailed taxonomic analysis of curated Foldseek- and PSI-BLAST-derived datasets (Figure 11.b.) revealed clear method-dependent biases. PSI-BLAST retrieved a higher proportion of viral hits (49.15% vs. 33.22% of unique TaxID), whereas Foldseek was more sensitive to non-viral candidates, particularly distant homologs. Foldseek's taxonomic profile was dominated by bacterial entries (57.87%), with lower but notable contributions from eukaryotic (5.36%) and archaeal (2.97%) taxa –proportions consistently higher than PSI-BLAST's (bacteria: 47.84%; eukaryotic/archaeal: 1.3-1.7%)–. This divergence reinforces method and database complementarity for an exhaustive homology screening. Among viral hits, while PSI-BLAST captured hits exclusively from *Varidnaviria* and *Duplodnaviria* (majority) realms, while Foldseek additionally captured candidates from *Adnaviria* –archaeal dsDNA virus infecting ABV's host *Acidianus spp.*– and a single *Riboviria* hit (ssRNA *Orthobunyavirus bataiense*). Despite Foldseek's broader taxonomic sensitivity, Duplodnaviria dominance was even more pronounced than in PSI-BLAST, suggesting higher structural conservation within this realm.

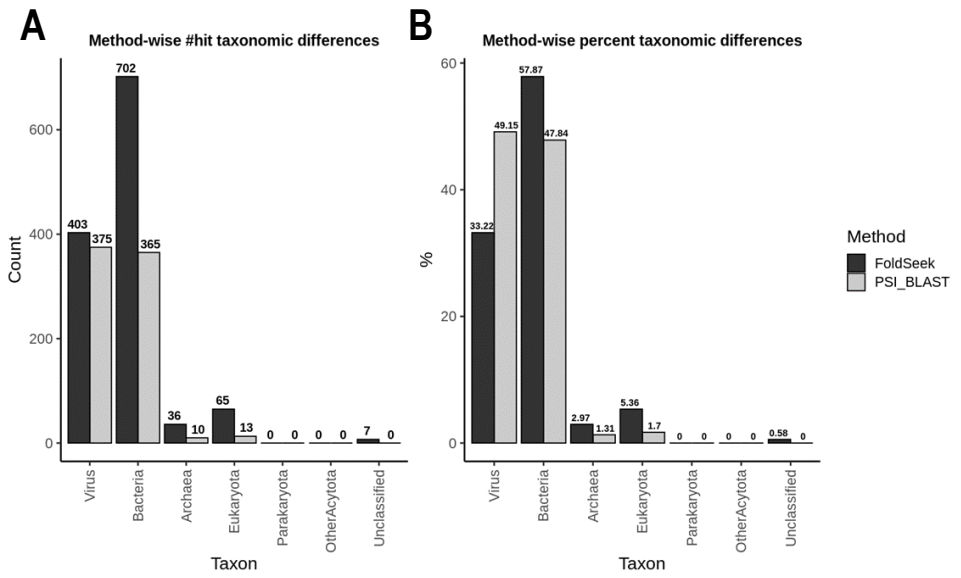

Figure 11. Absolute (**a**) and relative (**b**) taxonomic distributions for both Foldseek- and PSI-BLAST-derived curated datasets. Unique TaxID were counted as metric for taxonomic distributions. Distribution classification and stratification was performed by main paraphyletic top-rank taxonomic divisions of life. 'OtherAcytota' group gathered non-viral acellular replicons including plasmids, transposons, integrons, caposons, pipolins, viroids, virusoids, some satellite viruses and adeno-associated viruses, and obelisks; whereas 'Unclassified' category encompasses all NCBI GenBank metagenomic samples with no taxonomic assignation.

Finally, all curated candidates were merged (3309 hits) and deduplicated. A total of 2964 unique entries (based on GenBank and UniProtKB ID) were clustered using CD-HIT at a 0.6 identity threshold (see 3.6.1) in order to reduce sequence volume while retaining strong-homology hits. This resulted in 2153 clusters (-27.36%), most of which were monotypic (85.28%) or bitypic (7.8%). The largest cluster (#0) contained 47 adenoviral sequences. Overall, co-clustered Foldseek and PSI-BLAST candidates showed moderate taxonomic consistency and shared functional features or parental TP. Cluster representatives were extracted yielded, and 1433 entries linked to UniProtKB and cross-referenced with one or more contig-associated GenBank IDs were further subjected to sequence-based BLASTp to retrieve 'synonymous' GenBank entries. Each representative was then queried against NCBI Identical Protein Groups (IPG) database to assign a 'synonymous identical' protein (hereafter, IPG-representative) and its associated genome/*contig* for downstream classification of putative viral or cellular proteins. IPG mapping yielded 4255 entries, and, after collapsing to the longest *contig* per reference GenBankID, 2136 non-redundant GenomeID were retained (-0.008%), indicating minimal presence of representative candidates within shared IPG groups.

4.5. Structure-based clustering (qTMcluster)

Following Foldseek-driven structure-based homology search (see 4.4.), we employed US-align's qTMcluster (see 3.4.1.) to evaluate the structural co-clustering of initial TP queries and their respective candidate hits across the 1914 deduplicated entries. First, we observed that two main –namely ABV and GC1– so-called highly-prolific TP queries were the main contributors to the structure-based candidate dataset, collectively gauging ~95% of candidate volume and diversity (Figure 12.a.). However, upon accounting for co-clustering (Figure 12.b.), almost 95% of hits non-co-clustered with their parental query TP were precisely associated to either ABV (~65%) or GC1 (~30%), highlighting their poor clustering fidelity. Consequently, most putative homologs were not co-clustered (Figure 12.a.), and only ~5% of candidates co-clustered with their parental TP. Notably, within the latter group, the main contributor was *Kyanoviridae*-belonging DSL-LC02 TP, which alone accounted for ~35% of co-clustered hits, despite its limited overall contribution. Moderate co-clustering rates (~10-15%) were observed for TP from ABV, S-SCSM1 (*Kyanoviridae*), and GC1 (Figure 12.b.). Interestingly, although hits yielded by ABV and GC1 TP were mainly clustered independently to their respective TP, they were also among the most co-clustered in number, suggesting that their parental TP may exhibit broadly conserved folds with widespread fold similarity but moderate-to-low global homology.

On a per-TP basis, 100% of the putative homologs of biochemically validated TP from Φ29 and Sato, as well as those from under-characterized Harambe (genus *Harambevirus*) and St134 (genus *Andhravirus*) TP, exhibited complete co-clustered with their parental TP (Figure 12.c.). In contrast, all of the candidates related to TP from adenoviral/tectiviral CPAdV-2, hAd2, PRD1, or viph1080o or *Kyanoviridae* S-CREM2 failed to co-cluster with their respective TP queries, suggesting a lack of detectable global structural homology. Around 40-50% of the hits yielded by other empirically supported *Varidnaviria* TP –Toil, viph1249b, AP50 and Bam35– were co-clustered, while TP with limited prior characterization –Chedec11, ctnWS46, DSL-LC02, and S-SCSM1– showed moderate-to-high co-clustering ratios of 45-75%. Despite ABV and GC1 contributing the highest absolute number of co-clustered candidates (Figure 12.b.), their per-TP co-clustering ratios remained low (~2-5%), reinforcing the notion of widespread fold similarity with limited global structural homology.

This analysis revealed a clear dichotomy in the structural homology landscape of TP candidates. On one side, certain TP demonstrate strong structural conservation, suggesting highly conserved global structural features, despite potential sequence divergence, and may reflect stable evolutionary trajectories or functional constraints. Conversely, highly-prolific TP showed broad fold similarity but insufficient global structural alignment for clustering under TM-score thresholds. This pattern implies the presence of conserved structural motifs or domains embedded within otherwise divergent scaffolds, likely the result of convergent evolution, modularity, or domain shuffling. In sum, while some TP display clear structural homology indicative of conserved evolutionary origin, others suggest a mosaic nature with localized structural conservation insufficient for global clustering, underscoring the complexity and potential multi-origin nature of TP evolution.[1,25]

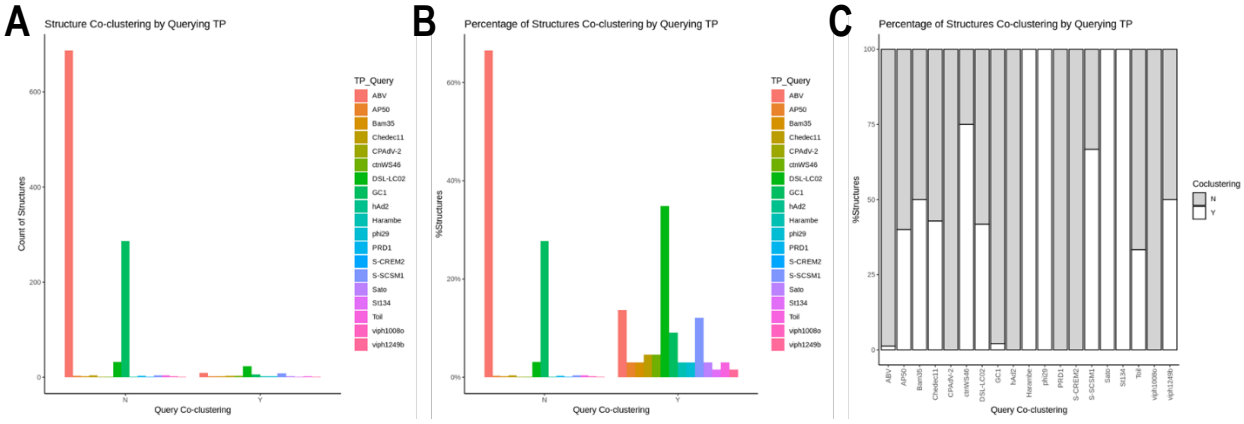

Figure 12. qTMcluster non-hierarchical structural clustering of initial TP queries and Foldseek-yielded putative homologs. **a)** Overall TP-hit co-clustered number of proteins. 'N' label represents non-co-clustered homologs, whereas 'Y' was employed to designate candidates co-clustered with their parental TP. **b)** Percentual overall TP-candidate co-clustering. The percentage of each parental TP's hits was assessed for each subset (Y/N), so that the sum of all co-clustered contributions makes up 100%, as well as for all non-co-clustering entries. Each bar and colour represent an initial TP's, either overall (a) or percentual (b), contribution. **c)** Per-TP break down of co-clustering rates. Each bar reflects the percentual share of homologs co-clustering (white) and not co-clustered (grey) with their parental TP. Alignment query coverage threshold ($\lfloor \xi_L \rfloor$) set at 30%.

At 30% query coverage threshold –lower coverage cut-off value previously applied during raw candidate dataset prefiltering (see 3.4.1. and 4.4.)– approximately 94% of the curated Foldseek-yielded candidate homologs failed to co-cluster alongside their respective initial querying TP, suggesting low overall accuracy of Foldseek's homology detection and limited reliability of the resulting predictions (Figure 13.c.). However, this trend may be partially influenced by methodological factors. First, as introduced before, highly-prolific initial TP queries such as ABV or GC1 –characterized by moderate global homology and widespread fold similarity– yielded a plethora of associated hits, most of which did not -co-cluster and may have indeed masked co-clustering among the remaining hit-scarce initial TP. Secondly, even if co-clustering scrutiny is assessed on a per-TP basis, while improved, performance remained modest (Figure 13.a.). Notably, TP such as phi29, Sato, Harambe, and St134 exhibited 100% co-clustering of their hits, despite their limited quantitative contribution to the overall dataset relative to ABV or GC1. However, third, this discrepancy likely reflects intrinsic algorithmic differences: unlike Foldseek, which is sensitive to local domain-level homology,

US-align's qTMcluster uses global RMSD values normalized by alignment length referred to query (see 3.4.1.) for non-hierarchical independent clustering. Consequently, qTMcluster may overlook localized structural conservation that originally triggered Foldseek hits, as it operates on full-length alignment rather than domain-level similarity. Therefore, qTMcluster clustering of potential polypeptide candidates may mask conserved local homology that led to Foldseek detection.

To address the apparent discrepancy between Foldseek hit detection and qTMcluster-based co-clustering, we tested a wide range of alignment coverage thresholds (30-90%) for filtering structurally clustered candidates to evaluate the impact of local homology on clustering accuracy. Raising the threshold from 30% to 70% filtering stringency had minimal effect on overall co-clustering rates, as it implied a drop-off of just a few candidates –e.g., increasing the cut-off to 50% sieved only 3 candidates out–. In contrast, applying a 90% coverage filter reduced the number of clustered candidates by 89.55%, primarily affecting hits associated with highly-prolific query TP such as ABV and GC1. Even at overall nete level, such increase of coverage stringency yielded a subtle yet significant increase in co-clustering ratio –rising by 123.3% to reach overall 13.4% Figure 13.d.–. On a per-TP basis, most initial TP exhibited enhanced clustering at higher coverage, with TP such as Bam35 and viph1249b achieving complete hit co-clustering (Figure 13.b.). Conversely, some TP, including those from CPAdV-2, hAdV2, PRD1, Toil, and *Kyanoviridae*-belonging S-CREM2, lost all associated hits. Interestingly, as suggested by Krupovic *et al*., TP and pPolB previously linked to adenoviral and PRD1-like tectiviral lineages (see 4.2.) are structurally related to polypeptide *Preplasmiviricota* pPolB encompassing both proteins.[1] Coincidentally, those TP were amongst those with the lowest co-clustering ratios. Thus, dropping their TP's candidate homologs by increasing coverage threshold may represent polypeptide-level homologs producing short, partial alignments. In summary, although qTMcluster initially appeared to contradict Foldseek homology predictions, stratifying clustering by alignment coverage reveals it as a viable validation strategy that has even hinted out possible evolutionary links. Nonetheless, its insensitivity to local domain-level similarity underscores the necessity to account for alignment coverage according to the specific protein dataset handled.

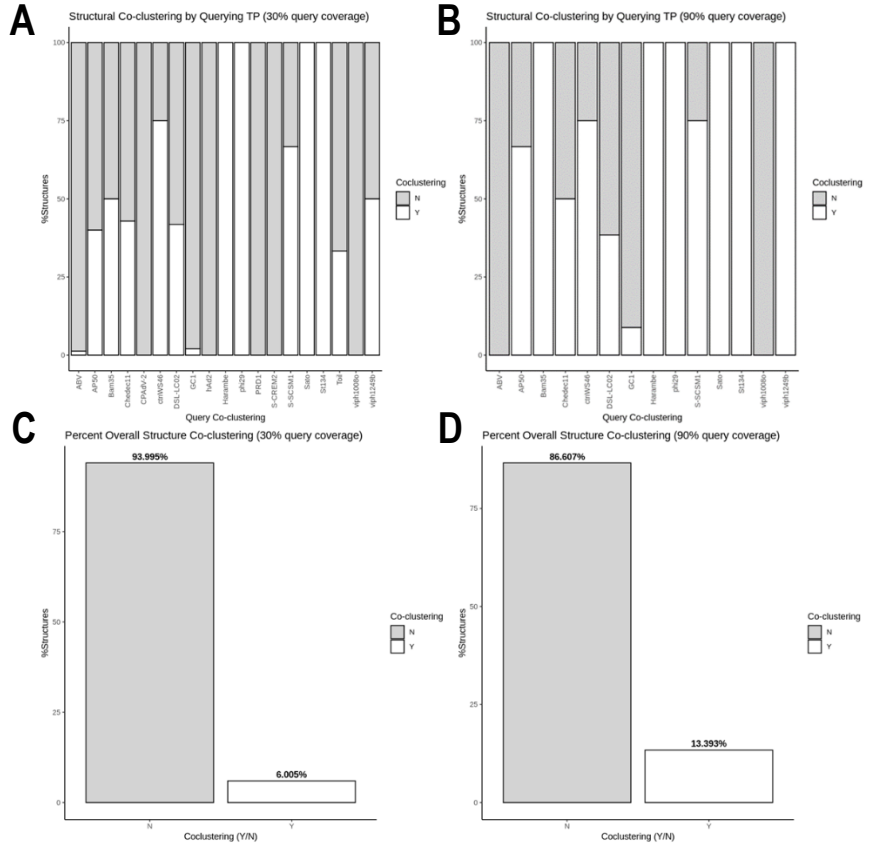

Figure 13. Impact of alignment query coverage threshold ($\lfloor \xi_L \rfloor$) on per-TP (a-b) overall (c-d) co-clustering percentual rates. In a-c $\lfloor \xi_L \rfloor$ was set at default 30%, whereas a $\lfloor \xi_L \rfloor$ = 90% was applied for b-d.

4.6. Functional inference (eggNOG-mapper)

To characterize the functional landscape of IPG-yielded representative proteins, functional annotations were inferred for all 2136 non-redundant TP-related protein entries employing eggNOG-mapper algorithm (see 3.8.). Both known functional descriptions and Cluster of Orthologous Groups (COG) categories were retrieved and further inspected. We decided to scrutinise representative proteins so that functional imputation could be extrapolated to the broader hit dataset. Regarding eggNOG novel ORFan-like families proposed by Rodríguez del Río *et al*. (2022, 2024),[106,107] 41 proteins (1.92% of input entries) were classified unknown-function categories according to eggNOG annotations (see 3.8; Figure 14.b.). Among these, NOVRBAUE and NOVTOAE9 were the most prevalent, accounting for 26.83% and 24.39% of valid assignments, respectively, while the remaining 46.78% were distributed across eight additional families. Notably, all IPG-derived representatives assigned to the NOVVKA6S category yielded top structural PDB matches annotated as 'primer terminal protein (TP)', strongly supporting the existence of distant homology linking canonical TP, candidate homologs, representative sequences, and inferred functions.

The resulting classification revealed a wide distribution across 20 distinct COG categories for 889 proteins (41.62% of input representatives), underscoring the functional heterogeneity associated with TP candidate homologs in representative genomes (Figure 14.a.). Despite this breadth, most IPG-yielded representative proteins were predicted no functional annotation using eggNOG-mapper, two out of the three most frequent COG categories were 'unknown function' ('-' or 'S'), comprising the 48.8% of the valid annotations and underlying the profound divergence of TP sequences and a putative vague functional origin or origins. Among not-unknown COG categories, the most prevalent functional class was COG category L ('Replication, recombination, and repair'), comprising 29% of the annotations, consistent with the established role of TP in protein-primed DNA replication and supports a likely evolutionary origin related to DNA-binding and -processing. Strikingly, a substantial number of proteins, though markedly less represented, were mapped to COG G ('Carbohydrate transport and metabolism', 7%) and COG M ('Cell wall/membrane/envelope biogenesis', 6.7%).The presence of COG G is particularly intriguing, as it suggests a potential mechanistic link: DNA-binding interactions often occur via (i) electrostatic interaction with the DNA's pentose-phosphate backbone or (ii) through direct base binding. Thus, carbohydrate metabolism entails oligosaccharide recognition and interaction which can, in turn, easily derive into DNA-binding through backbone interaction.

Additional annotations related to nucleic acid synthesis and metabolism were less common, with COG categories K ('Transcription') and F ('Nucleotide transport and metabolism') comprising 2.5% and 0.6% of the annotations, respectively. Testimonial (>0.5%) yet diverse representation was also observed for several categories –including COG V, E, P, C, Q, U, and singletons in T, W or O–. The presence of such diverse functional categories may reflect annotation artifacts, or multiple origin, convergent evolution and homoplasy of TP ancestors. However, multi-domain nature of candidate hits or representative proteins cannot be ruled out.

At this stage, we did not quantify the relative contribution of each initial TP to the IPG-derived representative dataset and its associated functional annotations. Consequently, it remains unevaluated whether specific COG categories were predominantly driven by homologs of a single initial TP or were collectively represented across multiple TP candidates (see 4.8.).

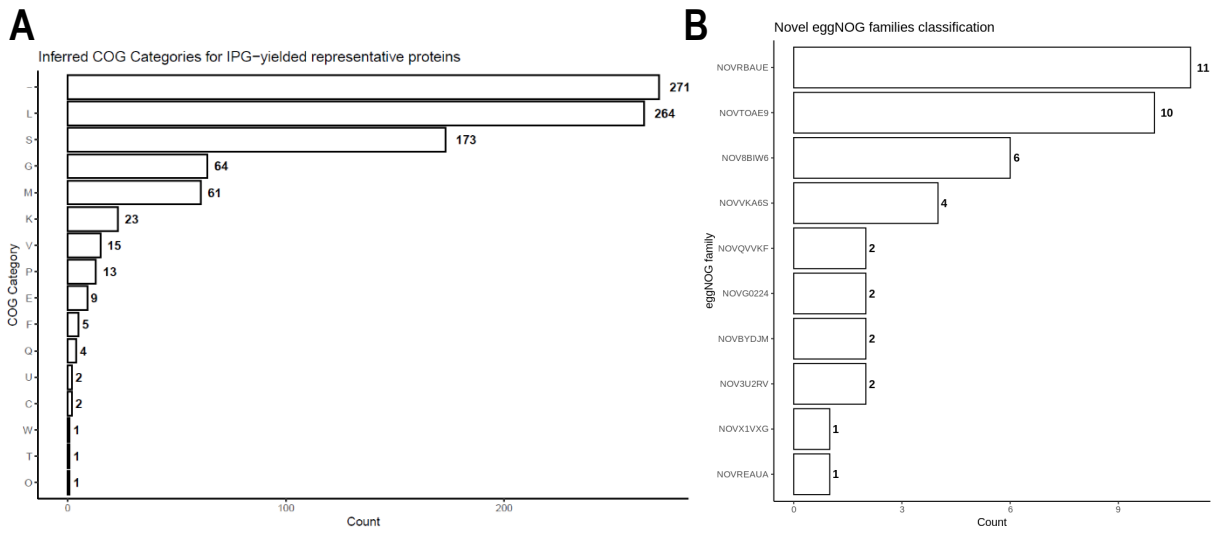

Figure 14. Distribution of (**a**) COG categories and (**b**) novel eggNOG ORFan-like families inferred by eggNOG-mapper for IPG-representative proteins. For detailed COG category label and description correlation see reference sheet from London Imperial College.[nn]

4.7. Genome annotation and non-viral *contig* detection

To ascertain the viral or non-viral nature of representative genomes/contigs for reliable TP origin inference –given the risk of taxonomic misannotation or viral contamination (see 1.2.)– we employed MMseqs2-based PHROG (for proteome annotation of viral ORF of known function) and geNomad (for comprehensive genome classification, annotation and viral detection), followed by cross-method validation (see 4.4., 3.7.2. and 3.7.3). As a result, PHROG identified only 105 (0.05%) genomes/*contigs* as lacking viral ORFs, classifying them as putatively non-viral. In contrast, geNomad detected 1327 viral signatures –and 23 plasmid sequences– within 1160 genomes (54.31%). To balance dataset representativity and minimize false positives, we tested multiple PHROG-based viral ORF count thresholds, ultimately setting a conservative yet permissive cut-off at ≤3 viral ORF ($\xi_{\mathrm{vORF}} \leq 3$). Cross-referencing both methods yielded a final set of 318 genomes/*contigs* (14.89%) considered ostensibly non-viral, representing 431 original curated hits (13% of the curated hit dataset).

PHROG-inferred viral genomes/*contigs* were further examined to confirm the presence of putative pPolB or other DNA polymerases supporting TP identity, and to assess synteny and viral ORF organization. Viral ORF counts across genomes followed a right-skewed distribution centred around 30 viral ORF per genome. However, among contigs encoding both a putative TP and a pPolB, matching PHROG pPolB-specific profile phrog_1907, the number of viral ORF sharply decreased to 1-3, forming a trimodal Gaussian mixture with a dominant peak at +1 ORF beyond the TP, suggesting low *contig* completeness but conserved gene relative synteny. Viral ORF proximity to TP was highly localized, with distance distribution showing a sharply leptokurtic Gaussian centred at 0, indicating that most ORFs were directly adjacent (±1) to the TP. For *contigs* encoding pPolB, the distance distribution shifted slightly upstream, with a mode at -1 and mean/median ~0, confirming tight genomic linkage. Rare pPolB placements occurred at +1, -2, and -5 positions.

[nn] https://www.sbg.bio.ic.ac.uk/~phunkee/html/old/COG_classes.html

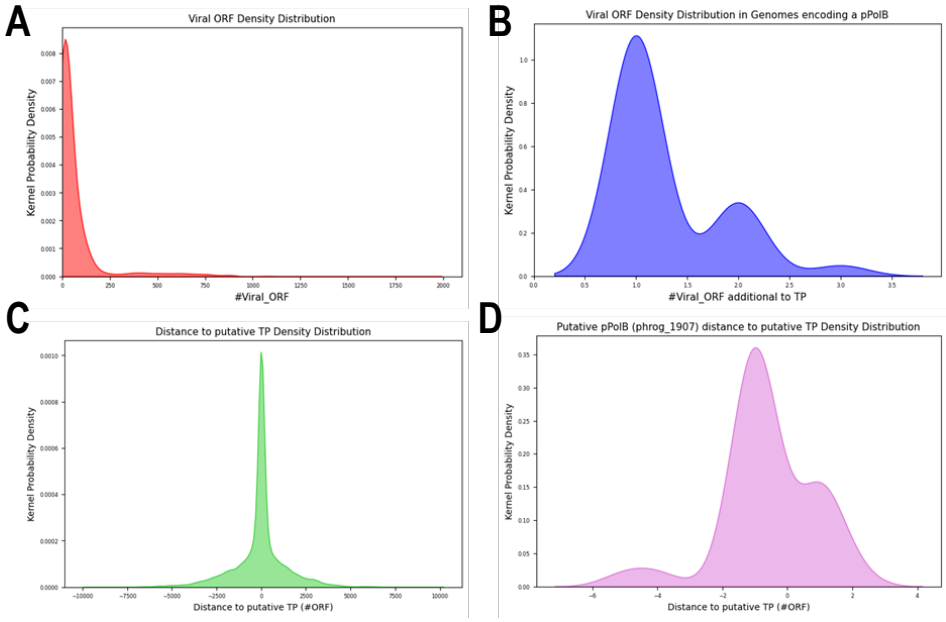

Figure 15. a) *Kernel* density distribution of viral ORF number among representative genomes/contig. **b)** *Kernel* density distribution additional to TP viral ORF number in *contigs* harbouring an ORF matching pPolB-specific phrog_1907 profile. **c)** ORF distance to TP *kernel* density distribution. Upstream positions are represented as negative values, whereas downstream ORF are represented as positive values. **d)** pPolB (phrog_1907) relative distance to TP *kernel* distribution.

4.8. Carbohydrate-metabolising and DNA-processing enzymes may be origins of TP

Following the exclusion of viral genomes/*contigs* and those containing proviruses or viral contamination, we decided to evaluate the remaining 318 genomes/*contig* and their corresponding original candidate hits as end-of-workflow scrutiny. To minimize redundancy, instead of evaluating all putative homologs left, we focused on the encoded 318 IPG-representative proteins. Before analysing any individual homolog or tracing evolutionary paths, we first assessed the functional diversity within this filtered dataset using eggNOG-mapper-inferred COG categories, as previously described (see 3.8 and 4.6.; Figure 16.a.). Some low-frequency categories –W ('Extracellular structures'), E ('Amino acid metabolism and transport') and Q ('Secondary structure')– of the non-filtered representative dataset were removed after filtering, likely reflecting viral features (W/Q) or host-virus interaction (E). Notably, while the complete dataset previously showed a high prevalence of unknown-function categories ('-' and 'S') among proteins with functional predictions (~41% of the original representative set) (see 4.6), their relative and absolute frequencies dropped significantly after removing representatives encoded in viral genomes. This reduction strongly supports validity for non-viral genome selection stage, as viral proteins often lack well-defined functional annotations or their roles remain unresolved. Importantly, aside from the sharp decrease in unknown-function categories, mainly '-', the overall COG category relative distribution remained largely unchanged after sieving viral representatives out, indicating proportional and homogeneous representation of viral proteins across functional categories. Interestingly, COG category G ('carbohydrate transport and metabolism') was overall the relatively least affected by viral protein exclusion, further supporting a potential evolutionary link between carbohydrate metabolism and TP's functional affinity for the DNA pentose-phosphate backbone.

To further investigate functional correlations per initial TP, all unique TP-hit-representative-function relationships were extracted (see 3.7.4) and visualized using an alluvial/Sankey plot (Figure 16.b.) to ease interpretability. While not directly accounting for absolute contribution, and despite the high diversity of inferred functions and their markedly heterogeneous distribution across initial TP, three main patterns emerged. (i) several COG categories were exclusively inferred for representative proteins associated with highly-prolific TP, either ABV –F, L and V– or GC1 –C, O, P, T and U–, while the remaining functional categories were shared by, at least, two initial non-highly-prolific TP. Notably, no functional overlap was observed between ABV- and GC1-associated representatives' inferred annotations. (ii) All representatives derived from adenoviral TP –CPAdV-2, DAdV-1, EAdV-1, GAdV-4, PiAdV-1, PsAdV-3– were assigned

solely to COG category S ('function unknown') and to no additional functional classes. (iii) Remaining representatives linked to *Kyanoviridae* TP family –namely S-CREM2, S-SCSM1 and DSL-LC02–, while sharing COG categories with representative proteins of at least another initial TP, they were never associated with COG S. Specifically, IPG-representative proteins linked to S-SCSM1 were exclusively predicted with COG K, whereas all S-CREM2-related representatives were annotated as '-'. Additionally, COG G ('carbohydrate transport and metabolism') was contributed only by GC1- and DSL-LC02-associated proteins, the latter contributing a single representative annotated as 'epimerase/dehydratase'. These results support a putative evolutionary link between TP and carbohydrate metabolism. Notably, all proteins inferred under COG L (DNA 'replication, recombination, and repair') originated exclusively from ABV-derived candidates.

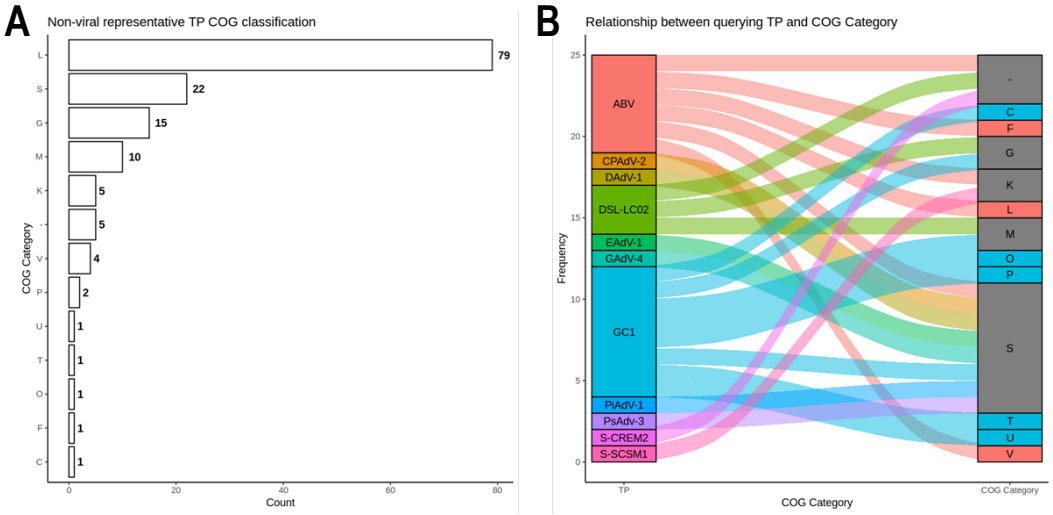

Figure 16. a) COG category absolute count distribution among IPG-yielded representative proteins encoded in ostensible virus-free genomes/*contig*. For detailed COG category label and description correlation see reference sheet from London Imperial College.[oo] **b)** Alluvial plot depicting initial TP relative contribution to each COG category. COG categories (right stacked bar) in dark grey were predicted to representative proteins yielded by more than one TP, whereas coloured (either red or process cyan) categories are only influenced by a single parental TP likewise coloured. COG C, O, P, T and U were only attributable to GC1-related representative proteins, whereas COG F, L and V were linked to ABV-exclusive candidates. Connector flows/alluvia are coloured coincidingly with each parental TP (left stacked bar). Alluvia breadth is homogeneous and equals the height of a stack slice. As TP-inferred function correlations are based on unique relationships, alluvia are not informative of the absolute abundance of either candidate hits nor representative proteins binding parental TP and COG categories. The height of TP stacked-bar slices is only proportional to the number of different COG categories linked. Conversely, the height of COG category stacked-bar slices is proportional to the number of different parental TP with at least one representative protein inferred each category.

After removing viral representative proteins, only one novel eggNOG ORFan-like family proposed by suggested by Rodríguez del Río *et al*. (2022, 2024) –NOVQVVKF– remained,[106,107] with two associated representative proteins. To further functionally characterise the definitive filtered dataset, non-standard eggNOG-mapper-inferred functional descriptions were examined for recurring patterns and initial TP. Remarkably, all IPG-representative proteins yielded by adenoviral were consistently annotated as 'Adenoviral DNA terminal protein' (data not shown). Upon taxonomic inspection, all of them were found to belong to either *Aviadenovirus* or *Mastadenovirus*, thus underscoring PHROG and geNomad's limited sensitivity to non-bacteriophage viral contigs and likely inoperancy for adenoviral ORF detection. Consequently, homologs and IPG-representative proteins related to adenoviral TP were manually removed from further analyses.

Subsequently, we focused on non-standard descriptions linked to COG-G: all but one were annotated as Glycosyl hydrolase family 28 (GHF28) or functionally related terms, collectively constituting the most abundant annotation group. GHF28 comprises polysaccharide-hydrolysing enzymes implicated in plant cell wall degradation or remodelling, including fungal and bacterial pectin-active enzymes such as exopolygalacturonases I (EC 3.2.1.67) and II (EC 3.2.1.82), endopolygalacturonases (EC 3.2.1.15), and

[oo] https://www.sbg.bio.ic.ac.uk/~phunkee/html/old/COG_classes.html

rhamnogalacturonases (EC 3.2.1.171).[112–115] Additionally, GHF28-alike annotations included pectate lyase superfamily members (EC 4.2.2.2 and EC 4.2.2.9) and analogous fold descriptions –e.g., 'Periplasmic copper-binding protein (NosD)', InterPro IPR007742. Notably, all GHF28-like annotations were exclusively linked to GC1-associated representative proteins, reinforcing polysaccharide metabolism –particularly pectin degradation– as an ostensible functional origin for GC1, consistent with the necessity of terminal proteins to bind DNA backbone-linked oligosaccharides.

Focusing on ABV-associated COG-L-related descriptions, we observed greater functional diversity compared to COG-G, though UvrD helicase and related terms predominated. UvrD helicases (EC 5.6.2.4) belong to the SF1A superfamily, comprising ATP-dependent, ssDNA-binding DNA helicases with RecA-like core folds and 3'→5' translocation polarity.[116] This molecular role provides a plausible, previously unreported likely origin for ABV TP, as helicases and other DNA-interacting proteins may easily evolve toward covalent DNA binding, akin to TP. The second most prevalent COG-L annotation, PD–(D/E)XK nuclease superfamily (EC 3.1.21.3),[117–121] comprises a vast and heterogenous group of restriction-like or homing DNA endonucleases I with an overall αββββαβ core fold and a conserved signature basic catalytic motif. Although its characteristic catalytic Lys residue is opposite in nature to TP's Thr, Ser or Tyr priming residue, their DNA-binding and specificity-determining residues may represent evolutionary precursors. Additionally, the high modularity of PD–(D/E)XK nucleases facilitates domain shuffling and functional divergence.[122] Notably, RecB-family nucleases, which interact with RecA, share PD–(D/E)XK motifs I–III,[123] thus linking the two dominant ABV-related COG L annotations –UvrD and PD–(D/E)XK– as putative components of bacterial or archaeal DNA repair and proto-immune systems,[124–126] which, in turn, may be at the evolutive origin of archaeal phage ABV.

Lastly, we evaluated original candidate hits yielded by non-highly-prolific TP, and applied conservative ORF number lower threshold of 3 for representative *contigs*, to enhance confidence in their non-viral origin. This filtering yielded 9 original candidates linked to only 5 parental TP –St134, Chedec11, WhyPhy, DSL-LC02 and S-SCSM1– across 6 IPG-representative proteins or *contigs* –encoding 4-7 ORF–. Despite undergoing multi-stage validation, short representative *contigs* lengths limited confidence in definitive taxonomic classification. All final candidates (Table 1) were encoded in either bacterial or archaeal genomes, predominantly Gram-positive *Bacillota*. However, species-level resolution remained largely unresolved. All final candidate homologs were further subjected to eggNOG-mapper functional inference.

Table 1. Final filtered candidate hits for non-ABV and non-GC1 initial TP. Asterisk-marked (*) candidate ID correspond to Foldseek-yielded AlphaFoldDB/UniProtKB entries, while the remaining ones are PSI-BLAST-yielded GenBank entries. Taxa marked with a double-dagger (‡) belong to Gram-negative bacteria –*phylum Bacteroidota* or *Pseudomonadota*–, those marked with a dagger (†) belong to *Archaea* –*phylum Thermoproteota*–, and the remaining unmarked species are Gram-positive bacteria –*phylum Bacillota*–.

TP	Candidate	Organism	Representative	Genome	Organism	COG
St134	MDN6162083.1 WP_141712621.1	*Atopostipes sp.* *Staphylococcus xylosus*	MDN6162083.1	JAKDRT010000259.1	*Atopostipes sp.*	?
Chedec11 **WhyPhy**	A0A355UWK0 * HBN10498.1	*Ruminococcus sp.* *Ruminococcus sp.*	HBN10498.1	DOBJ01000020.1	*Ruminococcus sp.*	?
DSL-LC02	NCV28323.1 WP_300102907.1 MFA6198908.1 MBQ5474055.1	*Nitrosomonadales sp.* ‡ *Flavobacterium sp.* ‡ *Bacteroidales sp.* ‡ *Lachnospiraceae sp.*	NCV28323.1 WP_300102907.1 MBQ5474055.1	RGKA01000088.1 NZ_JAQTPQ010000267.1 JAFNQR010000283.1	*Nitrosomonadales sp.* ‡ *Flavobacterium sp.* ‡ *Lachnospiraceae sp.*	G ? ?
viph1080o	A0A842PB43 *	*Nitrosopumilus sp.* †	MBC8501637.1	JACNFV010000031.1	*Nitrosopumilus sp.* †	?

Strikingly, all candidates yielded by *Caudoviricetes* TP are encoded in genomes of potential hosts for their respective associated TP –*Bacillus*-infecting Chedec11 or WhyPhy might be susceptible to infect related *Bacillota* members as *Ruminococcus sp*., and *Staphylococcus*-infecting St134 may be able to infect related *Bacilli* as *Atopostipes sp*.–. While such virus-host associations are plausible sources of direct viral gene acquisition, candidate TP cellular homologs may also reflect orthologs encoded in closely related taxa to an actually susceptible host. Notably, St134 yielded a hit from *Staphylococcus xylosus* (WP_141712621.1), a validated coagulase-negative staphylococcal host for *Andhravirus* phages including St134, Pike, Pontiff, and

SeAlphi,[28,127–129] supporting it as a strong candidate for andhraviral TP's origin. Another St134's hit from *Atopostipes sp.* (MDN6162083.1), co-clustered with the staphylococcal homolog, likely reflects true orthology potentially acquired via HGT. Such HGT events may constitute the underlaying mechanism behind candidate taxonomical diversity and candidate and TP potential hosts taxonomic assignation mismatch. That is, although putative candidate hits may be assigned to apparently incompatible hosting species –e.g., viph1080o is not able to infect archaeal *Nitrosopumilus spp*. or DSL-LC02 may not be able to infect Gram-negative *Nitrosomonadales spp.*, *Flavobacterium spp.* or *Bacteroidales spp.*–, assigned species may have acquired candidate hits through HGT phenomena from truly compatible host taxa –e.g., *Ruminococcus spp.* may have acquisitioned their associated candidates from *Bundooravirus*' unique known host *Bacillus pumilus*–[42] or vice versa.

Among all ultimate candidates, only DSL-LC02's hit NCV28323.1 stands out as functionally informative, being the sole validated candidate with a reliable functional annotation. Predicted as an 'epimerase/dehydratase', it is encoded in *Nitrosomonadales sp.* genome, where this such molecular role, as in most Gram-negative bacteria, is carried out by WcaG. Also known as GDP-L-fucose synthase –or GDP-4-keto-6-deoxy-D-mannose-3,5-epimerase-4-reductase, EC 1.1.1.271–, WcaG is a widespread conserved one-domain bifunctional NADPH-dependent 3,5-epimerase/4-reductase involved in fucose exopolysaccharide (FcEPS) biosynthesis.[130,131] In FcEPS-producing bacteria, such as especially *Nitrosomonadales spp.* or DSL-LC02' host *Synechococcus spp.*,[132–134] WcaG catalyses the carbonyl reduction of GDP-4-dehydro-6-deoxy-D-mannose (GDP-4k6d-Man) to GDP-L-fucose (GDP-Fuc) via intermediate dual epimerisation to GDP-4-dehydro-6-deoxy-L-glucose (GDP-4k6d-Glc).[130,135,136] This reaction is essential for synthesizing diverse GDP-Fuc-derived deoxysaccharides (e.g., colanic acid), critical for biofilm formation, community homeostasis, and environmental resilience.[130,133] Given WcaG's involvement in oligosaccharide binding and metabolism, as previously introduced (see 4.6.), may plausibly diverge toward DNA interaction via deoxyribose moieties. Thus, NCV28323.1/WcaG represents a compelling DSL-LC02 TP homolog, potentially found at the origin of this TP acquisition. Furthermore, *Nitrosomonadales sp.* may have acquisitioned WcaG from *Cyanobacteriota* DSL-LC02's *Synechococcus* hosts.

Finally, although Foldseek initially contribution the majority of hits to the unfiltered dataset, the final curated set shows a clear dominance of PSI-BLAST-derived candidates (7 out of 9). Remarkably, TP yielding the highest number of Foldseek hits were exclusively highly-prolific, whereas for non-highly-prolific TP the most informative candidates were PSI-BLAST-yielded. These results challenge the recent proposition of structure-based homology searches as the new gold standard for remote homolog detection as suggested by Krupovic *et al*. (2024)[1] (see 1.2.).

5. Discussion

5.1. Combined standardised structure- and genome-based features enable TP screening

Terminal proteins are highly divergent proteins, but poorly characterized structural and biochemically, thus lacking validated parametric feature for its programmatic screening and imputation. Here, we showed that TP can be structurally characterised be specific patterns regarding surface charge density asymmetry, α-helical content and DNA-binding probability, alongside previously suggested features related to synteny conservation. Particularly the presence of pronounced asymmetric surface charge density distributions facilitating DNA-backbone interaction was a recurrent property among known and newly predicted TP. Similarly, high α-helical content and DNA-binding capacity further supported the structural compatibility of candidate proteins with a TP molecular role. Nonetheless, these parameters are neither exclusive nor universally present in *bona fide* TP. In parallel, genome-informed features such as conserved viral synteny, genome ORF density and taxonomic origin added important contextual support. Notably, pPolB co-occurrence is highly informative, particularly when spatial proximity to TP-like ORF is observed. Taken together, these attributes constituted a multivariate, standardisable criteria set allowing TP identification and functional imputation with enhanced confidence.

Although out simplified Debye-Hückel linearisation for Poisson-Boltzmann equation for point electrostatic potential calculation accuracy might not be high, it was precise enough to effectively capture electrostatic potential variations and charge density asymmetry. To further support TP functional prediction, we employed Szilágyi's DNABIND classifier as the only publicly available online tool enabling dual sequence/structure probabilistic DNA-binding scoring.[88] Despite being originally trained on general DNA-binding proteins, its reported performance (~70–75% sensitivity and specificity) remains comparable to that of more recent, structure-informed methods. However, its use in the TP-specific context revealed important limitations. Unlike canonical DNA-binding domains, TP bind DNA via priming residues and generally lack structured DNA-recognition motifs, probably often falling outside traditional classifier boundaries. The results underscore that generic DNA-binding predictors fail to capture the nuanced biochemical and structural constraints of TP-DNA interactions. Hence, while Szilágyi's classifier provided supportive evidence in some instances, its predictive value is limited and may result in both false positives and negatives, especially in divergent or minimal TP variants, underscoring the necessity of implementing an *ad hoc* classifier trained on TP data.

The clear resolution afforded by combining structural, biochemical, and genomic evidence suggests the potential to develop a dedicated TP classifier, tailored to the unique sequence and structural landscape of TP. A supervised machine learning model –preferably a Bayesian classifier or a probabilistic graphical model due to their interpretability and robustness to small training sets– could be trained on a curated dataset of verified TP and non-TP sequences. Crucially, such a classifier could integrate both primary sequence descriptors (e.g., priming residue context, amino acid composition) and secondary structure-derived parameters (e.g., electrostatic asymmetry, α-helical ratio, DNA-binding probability), along with genome-informed variables (e.g., ORF context, synteny patterns). When coupled with existing virus prediction tools such as geNomad, this TP-specific module could extend current viral annotation capabilities to include functional inference and TP detection. This integrative framework could not only systematise TP discovery but also aid in de novo functional annotation of poorly understood small viral proteins at the interface of replication initiation and genome packaging.

5.2. Sequence-based vs. structure-based homology search methodologies

All along this work, we successfully addressed structure- and sequence-based homology search method-wise critical differences that have yielded a broad homology diversity exploration. The comparative analysis of PSI-BLAST and Foldseek results revealed a striking absence of overlap between the candidate homologs retrieved by each method, highlighting their methodological divergence and suggesting strong complementarity in distant homology detection. This discrepancy may primarily arise from the extreme diversity and sequence divergence among terminal proteins, which significantly challenge both structure- and sequence-based homology inference. For TP, characterized by small size, rapid evolutionary rates and putative homoplasy, sequence similarity is often eroded beyond the detection thresholds of traditional alignment tools, while structural convergence or modularity may also limit reliable fold-level matching. Importantly, the absence of shared hits between PSI-BLAST and Foldseek does not necessarily indicate a methodological failure but rather reflects the high divergence of the TP landscape and the differing evolutionary signals each method targets. As such, the non-overlapping results underscore a high degree of complementarity, collectively expanding the explored homologous space and capturing alternative aspects of evolutionary conservation as primary sequence versus global fold geometry.

However, this lack of overlap between PSI-BLAST- and Foldseek-yielded candidates may not be generalized beyond this TP use case. The degree of complementarity between structure- and sequence-based homology methods is highly dependent on the evolutionary divergence, domain architecture, and sequence complexity of the protein family being interrogated. Proteins with moderate divergence and well-conserved domains may exhibit greater overlap between methods, while highly novel or fast-evolving sequences like TP tend to diverge in the features detected by structure- versus sequence-alignment algorithms. Additionally, it is critical to account for confounding methodological factors such as database composition and indexing strategies. While, PSI-BLAST searches were conducted against nr-clustered

protein databases, Foldseek queries mainly relied on the AlphaFold structural database.[63,74] Differences in curation, redundancy filtering, and sequence-to-structure mapping across these repositories inherently impact hit composition, introducing potential biases unrelated to methodological power or accuracy *per se*.[74] Furthermore, the internal methodological variation within structure-based approaches also contributes to result divergence. Foldseek and DALI, for instance, operate under fundamentally distinct paradigms, as Foldseek uses fast, rotation-invariant embeddings (PLM) and vector quantization to capture structural similarities alignment-like comparisons of protein folds, while DALI applies C_α-C_α distance matrix comparisons across full atomic coordinates to detect global fold similarity.[63,74] These algorithmic differences affect sensitivity, runtime, and false-positive rates, and can produce discordant outputs for the same query. Importantly, while structure-based homology searches are increasingly adopted, their pipelines are still under active development, lack full benchmarking standardisation, and cannot yet be considered a universal gold standard for remote homolog detection. Nonetheless, they enable exploration of sequence-dark evolutionary space inaccessible to purely alignment-based tools. Our findings thus reinforce the view that structure- and sequence-based methods should not be treated as interchangeable or competing, but as synergistic tools whose combined application is crucial for exhaustive phylogenetical-functional inference, particularly in deeply divergent and under-characterised protein families such as TP.

5.3. Terminal proteins constitute a highly divergent and heterogeneous protein family

Despite their historical relevance and extensive experimental characterization,[1,12–17] neither *Salasmaviridae* (e.g., Φ29) nor *Tectiviridae* (e.g., PRD1, Bam35, Toil) TP yielded high-confidence hits after our structure- and sequence-based homology search and rigorous curation pipeline. This lack of retention contrasts sharply with their canonical roles in TP research and likely stems from intrinsic properties of structure-based algorithms such as Foldseek, which favour compact, ordered folds over multidomain flexibility or fragmented matches.[74] Notably, PRD1-like TP –foundational to Krupovic *et al.* (2024) structural reconstruction of pPolB/TP evolutionary history–[1] were identified using DALI, a tool with more permissive structural alignment scoring compared to Foldseek's highly optimized, high-throughput architecture.[63,74] Our inability to recover PRD1-like ancestral hits using Foldseek, despite their prominence in the DALI-based phylogeny of Krupovic et al. (2024),[1] highlights methodological differences rather than contradiction. Foldseek prioritizes global structural similarity and high-throughput scalability, offering a more conservative but robust alternative for homology detection.[74] Our findings reinforce the validity of structure-based screening pipelines for TP classification, providing a complementary framework that captures divergent yet functionally coherent homologs. Rather than undermining previous proposals, our results support a refined, architecture-driven model for TP evolutionary analysis and large-scale annotation. In particular, Krupovic et al. postulated, according to straightforward RMSD calculations, that adenoviral TP evolved from PRD1-like tectiviral precursors via polintons, basing their inference on structural similarity of pPolB D1 domains across preplasmiviricots.[1,25] However, our structural reconstructions, powered by FoldMason and Foldseek, suggest that some *Aviadenovirus* TP clusters more closely related to ancestral Tectiviridae e.g., GC1, while others to PRD1, rather than with other adenoviral clades such as *Mastadenovirus*. This refined architectural clustering points to a possible polyphyletic origin within *Adenoviridae*, challenging the assumption of a monolithic evolutionary path and encouraging a re-examination of TP phylogenies using alternative structural pipelines. Moreover, both our sequence and structural data were consistent showing a clear division between alpha/gamma/delta- and betatectiviral TP. Furthermore, the inability of structural searches to recover ostensibly ancestral PRD1 hits, in spite of high-confidence architecture-based evolutionary expectations, highlights a current limitation in TP structural homology resolution; one that ongoing development of structure alignment tools may eventually overcome.[138] These results strongly support the view that structure-based homology methods, while expanding discovery potential, are not yet standardized enough to be considered a definitive gold standard.

Phylogenetic trees built from our dataset consistently revealed that TP evolution may have homoplastic origin and polyphyletic and does exhibit host-driven structural similarities, with TP clustering patterns frequently reflecting host phylogenies more than viral taxonomy. For instance, TP from viruses infecting *Bacillota* (e.g., Chedec11, WhyPhy, St134) or *Cyanobacteriota* (e.g., DSL-LC02, S-SCSM1) tend to co-

cluster, despite diverging substantially at the genomic or taxonomic level. Notably, our structure-informed phylogenies enabled the tentative reassignment or refinement of several clades: Ayka seems to be related to phages from genus *Anjalivirus* –probably as independent sister genus– whereas Evcara and Curie might be encompassed within genus *Amherstviurs*. Toil TP, while sequence-divergent, retained architectural cores that linked it to *Tectiviridae*, a relationship undetectable by sequence comparison alone. Structural analysis via FoldMason proved particularly useful for resolving ambiguous placements and for suggesting homology among poorly characterized proteins, especially in cases where sequence divergence was high. The combined data underscore the utility of TP as molecular markers for viral systematics and evolution and provide a robust framework for classifying emerging or uncharacterized viral lineages based on core replication protein architecture. This implementation enhances confidence on our structure-yielded results that other structure-based phylogenies lack. On balance, structure- and sequence-based combined data underscore the utility of TP as molecular markers for viral systematics and evolution and provide a robust framework for classifying emerging or uncharacterized viral lineages based on core replication protein architecture.

5.4. Putative functional origins of ABV, GC1 and DSL-LC02 terminal proteins

Although terminal TP are mechanistically linked to DNA replication, we observed that among the final curated non-viral homologs, only those associated with *Bottigliavirus pozzuoliense* (ABV) retained annotations related to DNA-binding or processing functions. This absence is unexpected, as TP homologs would reasonably be anticipated to exhibit molecular roles involving DNA interaction. One explanation may lie in the curation pipeline itself: stringent thresholds applied for quality control (e.g., E-value, alignment coverage, probability scores) may have sieved out candidates exhibiting marginal similarity to known DNA-binding proteins. Additionally, the limited availability of homologs and well-annotated structural databases with comprehensive taxonomic coverage may have further constrained hit retention. Despite this, the robust association of DNA-processing annotations to ABV-related candidates strongly validates our approach and suggests that the employed filtering pipeline, although strict, retained biologically meaningful hits.

Among the most notable results were three high-confidence, functionally plausible cellular homologs for TP from ABV, GC1, and DSL-LC02. ABV-associated hits revealed enrichment in two functionally coherent DNA-processing enzyme families: UvrD helicases and PD–(D/E)XK nucleases. UvrD (EC 5.6.2.4), a RecA-associated 3'→5' SF1A DNA helicase, provides a credible evolutionary substrate for TP divergence, given its ssDNA-binding properties and directional translocation. Similarly, PD–(D/E)XK nucleases (EC 3.1.21.3), known for their structural modularity and DNA cleavage specificity, could serve as ancestral templates for TP evolution, especially given their conserved αβββαβ core fold and domain promiscuity. In addition, RadA and other related PD–(D/E)XK asgardarchaeal recombinases are present in *Promethearchaeati* archaea, sister kingdom of ABV's *Acidianus spp.*'s host (*Thermoproteati*), and are thought to be widespread among archaea, constituting a plausible origin for ABV TP. These annotations, exclusive to ABV, point to a bacterial or archaeal RecAB DNA repair and proto-immunity context at the evolutionary origin of this archaeal TP. Conversely, hits linked to GC1 and DSL-LC02 were exclusively annotated with carbohydrate-hydrolysing enzymes. GC1 candidates were dominated by Glycosyl Hydrolase Family 28 (GHF28) enzymes, implicated in bacterial and fungal pectin degradation, including exo-/endopolygalacturonases (EC 3.2.1.67, EC 3.2.1.82, EC 3.2.1.15), rhamnogalacturonases (EC 3.2.1.171) and pectate lyases (EC 4.2.2.2 and EC 4.2.2.9).[112–115] DSL-LC02's most informative candidate, NCV28323.1, was annotated as WcaG 'epimerase/dehydratase' (EC 1.1.1.271), a bifunctional GDP-L-fucose synthase involved in oligosaccharide processing and biofilm biosynthesis, crucial for DSL-LC02's *Synechococcus spp.* hosts.[129–135] Both GHF28 and WcaG share oligosaccharide-binding properties that plausibly represent ancestral functions facilitating the evolution of DNA backbone recognition and interaction via deoxyribose moieties, thus supporting a carbohydrate-processing origin for these TP. Moreover, GC1 showed the least surface charge asymmetry, which downplays interaction with DNA pentose-phosphate backbone but might favour carbohydrate interaction and metabolism, likely pointing alternative deoxyribose binding mechanism interaction with DNA, and reinforcing GHF28 as promising ancestral origins of this TP.

Beyond these high-confidence origins, we also identified strong host-related candidates for *Caudoviricetes* TP from *Bahkauvirus chedec11*, *Bundooravirus whyphy*, and *Andhravirus st134*. Remarkably, St134 yielded a homolog (WP_141712621.1) from *Staphylococcus xylosus*, a validated host for Andhravirus phages,[28,126–128] strongly suggesting this protein as a putative TP ancestor acquired through virus-host HGT. An additional St134 hit from *Atopostipes sp.* co-clustered with the former, likely representing a HGT ortholog. Similar patterns were observed for hits linked to Chedec11 and WhyPhy, which were associated with *Bacillus* hosts and candidates such as *Ruminococcus sp.*,[42] pointing to a host-taxa-related origin. HGT not only explains candidate taxonomic mismatches but also supports the view that TP evolution is shaped by sporadic assimilation from functionally versatile, host-derived proteins rather than a single monophyletic ancestor, likely pointing to multiple acquisition events followed by evolutive convergence.

5.5. Concluding remarks

While the primary goal of this study was not to pinpoint a definitive cellular ancestor for each TP or TP group, our multi-faceted investigation nonetheless succeeded in shedding light on key aspects of TP phylogeny and evolutionary trajectories. By integrating structural, functional, and genomic data, we uncovered several highly promising candidate ancestors, most notably the WcaG GDP-L-fucose synthase associated with DSL-LC02 TP. This constitutes one of the first 'direct' links between TP and a putative cellular homolog, substantially expanding the conceptual landscape of TP origin and proving host genomes as pools for TP ancestral acquisition. Moreover, we broadened the evolutionary scope of TP studies by moving beyond traditional TP-TP comparisons to consider potential host-derived and functionally analogous ancestors, which has historically been neglected or impossible to deepen in.[1,11,25] In parallel, our phylogenetic analyses also revealed novel patterns of co-evolution driven by host specificity, with terminal proteins from viruses infecting related bacterial or archaeal taxa often clustering together structurally despite taxonomic disparity at the genome level.

Second, we addressed a major knowledge gap in TP research by proposing and validating a robust set of structure-derived and genome-inferred parameters –such as surface charge asymmetry, α-helical content, DNA-binding likelihood, and relative synteny– to enable the systematic screening and imputation of terminal proteins. While none of these features are individually diagnostic, in combination they form a reproducible framework for TP identification. Although existing tools proved serviceable for preliminary classification, our findings highlight the pressing need for an *ad hoc* classifier tailored to TP-specific characteristics. Such a model, trained on curated TP and non-TP examples, could be readily integrated into viral annotation pipelines such as geNomad, enabling scalable and accurate TP detection in metagenomic contexts.

Third, by coupling sequence-based (PSI-BLAST) and structure-based (Foldseek) homology searches, we maximized the phylogenetic and taxonomic range of our screening effort. Despite their methodological divergence and non-overlapping results, both approaches contributed uniquely to the dataset, emphasizing their complementarity. This dual-pronged strategy revealed that well-characterized TP like those from Φ29 or PRD1 yielded few or no high-confidence homologs, while lesser-known proteins –including those from ABV, DSL-LC02, GC1, WhyPhy, and especially St134– facilitated deeper evolutionary insights and more extensive candidate retrieval. This paradox underscores the utility of underexplored and structurally divergent TP for evolutionary reconstruction, as well as the limitations of over-reliance on canonical models. Looking ahead, ongoing advancements in structural bioinformatics and the emergence of next-generation homology detection algorithms and specialized databases are likely to further standardize and refine TP screening protocols,[138] expanding their utility for virology and molecular evolution. Low redundancy of structure-based and sequence-based methods confirms the complementarity of both approaches and validates the use of combined methodologies to maximize diversity capture. Our findings position structure-based homology as a reliable, high-resolution tool for exploring deep evolutionary relationships and justify its use as a complementary approach in phylogenetic inference.[74,76] On the other hand, perhaps, we should have included more filtering steps or selecting only high-quality genomes for a finer candidate selection.

6. Conclusions

1. The analysis of an initial TP dataset from literature and databases confirmed the high diversity of TP sequences, even within the same viral family.
2. We implemented a simplified Debye-Hückel linearisation of the Poisson-Boltzmann equation to analyse protein charge distribution, confirming charge asymmetry as a common feature of TP.
3. Other robust TP features include secondary structure contents, DNA binding capacity prediction, and gene synteny.
4. Sequence- and structure-based searches of TP-related proteins yielded divergent datasets, highlighting the complementary utility of these approaches.
5. Despite numerous hits, many main TP groups lack clear cellular homologs after data curation and filtering, including some identified in previous studies.
6. TP from archaea *Bottigliavirus pozzuoliense* (Acidianus bottle-shaped virus, ABV) is highly similar to host cellular DNA-binding proteins, consistent with prior reports.
7. TP from *Gammatectivirus GC1* and *Kyanoviridae* phage DSL-LC02 share similarities with hosts' cellular pectin-hydrolysing enzymes and a cellular epimerase, suggesting alternative TP origin and DNA interaction mechanisms.
8. TP were likely acquired through multiple independent virus-host horizontal gene transfer events prior their convergent evolution.
9. TP from *Andhravirus St134*, *Paulavirus viph1080o*, *Bundooravirus whyphy* and *Bahkauvirus chedec11*, while poorly studied, found host core cellular homologs of unknown function that might be at their origins.
10. We implemented both TP-driven structure-based and sequence-based phylogenetic reconstructions that pointed to *Aviadenovirus* (and *Adenoviridae*) as ostensibly polyphyletic groups needing further consideration.
11. TP could be employed as molecular markers for viral taxonomy and systematics on uncharacterised viral lineages such as Ayka (proposed *Anjalivirus*) or Evcara and Curie (proposed *Amherstvirus*).

7. Acknowledgements

I want to specially thank Víctor Mateo Cáceres, MSc., and Eduardo Diego Lozano Escobar, MSc., for their worthy contributions on code scripting and their priceless points of view and advices about how to proceed on the development of this work's pipeline.

8. Bibliography

1. Krupovic, M., Kuhn, J. H., Fischer, M. G. & Koonin, E. V. Natural history of eukaryotic DNA viruses with double jelly-roll major capsid proteins. (2024) doi:10.1101/2024.03.18.585575.

2. Krupovic, M. & Koonin, E. V. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat. Rev. Microbiol.* **13**, 105–115 (2015).

3. Krupovic, M., Dolja, V. V. & Koonin, E. V. The LUCA and its complex virome. *Nat. Rev. Microbiol.* **18**, 661–670 (2020).

4. Koonin, E. V & Krupovic, M. Polintons, virophages and transpovirons: a tangled web linking viruses, transposons and immunity. *Curr. Opin. Virol.* **25**, 7–15 (2017).

5. Lefkowitz, E. J. *et al.* Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* **46**, D708–D717 (2018).

6. Pérez-Arnaiz, P. *et al.* Involvement of phage ϕ29 DNA polymerase and terminal protein subdomains in conferring specificity during initiation of protein-primed DNA replication. *Nucleic Acids Res.* **35**, 7061–7073 (2007).

7. Redrejo-Rodríguez, M. *et al.* Primer-Independent DNA Synthesis by a Family B DNA Polymerase from Self-Replicating Mobile Genetic Elements. *Cell Rep.* **21**, 1574–1587 (2017).

8. Meselson, M. & Stahl, F. The replication of DNA in Escherichia coli. *Proc. Natl. Acad. Sci.* **44**, 671–682 (1958).

9. Mateo-Cáceres, V. & Redrejo-Rodríguez, M. Pipolins are bimodular platforms that maintain a reservoir of defense systems exchangeable with various bacterial genetic mobile elements. *Nucleic Acids Res.* **52**, 12498–12516 (2024).

10. Redrejo-Rodríguez, M. & Salas, M. Multiple roles of genome-attached bacteriophage terminal proteins. *Virology* **468–470**, 322–329 (2014).

11. Yutin, N., Shevchenko, S., Kapitonov, V., Krupovic, M. & Koonin, E. V. A novel group of diverse Polinton-like viruses discovered by metagenome analysis. *BMC Biol.* **13**, 1–14 (2015).

12. Berjón-Otero, M., Villar, L., Salas, M. & Redrejo-Rodríguez, M. Disclosing early steps of protein-primed genome replication of the Gram-positive tectivirus Bam35. *Nucleic Acids Res.* gkw673 (2016) doi:10.1093/nar/gkw673.

13. Salas, M. PROTEIN-PRIMING OF DNA REPLICATION. *Annu. Rev. Biochem.* **60**, 39–71 (1991).

14. Redrejo-Rodríguez, M., Muñoz-Espín, D., Holguera, I., Mencía, M. & Salas, M. Nuclear localization signals in phage terminal proteins provide a novel gene delivery tool in mammalian cells. *Commun. Integr. Biol.* **6**, e22829 (2013).

15. Redrejo-Rodríguez, M., Muñoz-Espín, D., Holguera, I., Mencía, M. & Salas, M. Functional eukaryotic nuclear localization signals are widespread in terminal proteins of bacteriophages. *Proc. Natl. Acad. Sci.* **109**, 18482–18487 (2012).

16. Méndez, J., Blanco, L., Esteban, J. A., Bernad, A. & Salas, M. Initiation of phi 29 DNA replication occurs at the second 3' nucleotide of the linear template: a sliding-back mechanism for protein-primed DNA replication. *Proc. Natl. Acad. Sci.* **89**, 9579–9583 (1992).

17. Salas, M. & de Vega, M. Protein-Primed Replication of Bacteriophage Φ29 DNA. in 137–167 (Elsevier Inc., 2016). doi:10.1016/bs.enz.2016.03.005.

18. Quinones-Olvera, N. New Methods to Explore Mobile Genetic Elements in Bacteria. *Harvard Univ. Grad. Sch. Arts Sci.* (2024).

19. Parra, B. *et al.* Characterization and Abundance of Plasmid-Dependent Alphatectivirus Bacteriophages. *Microb. Ecol.* **87**, 85 (2024).

20. Xi, H., Fu, B., Sheng, Q., Luo, M. & Sun, L. Isolation and Characterization of a Lytic Bacteriophage RH-42-1 of Erwinia amylovora from Orchard Soil in China. *Viruses* **16**, 509 (2024).

21. Gillis, A., Hock, L. & Mahillon, J. Comparative Genomics of Prophages Sato and Sole Expands the Genetic Diversity Found in the Genus Betatectivirus. *Microorganisms* **9**, 1335 (2021).

22. Berjón-Otero, M. *et al.* Bam35 Tectivirus Intraviral Interaction Map Unveils New Function and Localization of Phage ORFan Proteins. *J. Virol.* **91**, (2017).

23. Jalasvuori, M. & Koskinen, K. Extending the hosts of Tectiviridae into four additional genera of Gram-positive bacteria and more diverse Bacillus species. *Virology* **518**, 136–142 (2018).

24. Caruso, S. M. *et al.* A Novel Genus of Actinobacterial Tectiviridae. *Viruses* **11**, 1134 (2019).

25. Koonin, E. V., Fischer, M. G., Kuhn, J. H. & Krupovic, M. The polinton-like supergroup of viruses: evolution, molecular biology, and taxonomy. *Microbiol. Mol. Biol. Rev.* **88**, (2024).

26. Al-Wassiti, H. A. *et al.* Adenovirus Terminal Protein Contains a Bipartite Nuclear Localisation Signal Essential for Its Import into the Nucleus. *Int.*

J. Mol. Sci. **22**, 3310 (2021).

27. Washington, J. M. *et al.* Expanding the Diversity of Actinobacterial Tectiviridae: A Novel Genus from Microbacterium. *Viruses* **17**, 113 (2025).

28. Hawkins, N. C., Kizziah, J. L., Hatoum-Aslan, A. & Dokland, T. Structure and host specificity of Staphylococcus epidermidis bacteriophage Andhra. *Sci. Adv.* **8**, (2022).

29. Krupovic, M. & Koonin, E. V. Evolution of eukaryotic single-stranded DNA viruses of the Bidnaviridae family from genes of four other groups of widely different viruses. *Sci. Rep.* **4**, 5347 (2014).

30. Klassen, R. & Meinhardt, F. Linear Protein-Primed Replicating Plasmids in Eukaryotic Microbes. in *Microbial Linear Plasmids* 187–226 (Springer Berlin Heidelberg). doi:10.1007/7171_2007_095.

31. Volozhantsev, N. V. *et al.* Molecular Characterization of Podoviral Bacteriophages Virulent for Clostridium perfringens and Their Comparison with Members of the Picovirinae. *PLoS One* **7**, e38283 (2012).

32. García, E., Gómez, A., Ronda, C., Escarmis, C. & López, R. Pneumococcal bacteriophage Cp-1 contains a protein bound to the 5' termini of its DNA. *Virology* **128**, 92–104 (1983).

33. Martín, A. C., Blanco, L., García, P., Salas, M. & Méndez, J. In VitroProtein-primed Initiation of Pneumococcal Phage Cp-1 DNA Replication Occurs at the Third 3' Nucleotide of the Linear Template: A Stepwise Sliding-back Mechanism. *J. Mol. Biol.* **260**, 369–377 (1996).

34. Cater, K. *et al.* A Novel Staphylococcus Podophage Encodes a Unique Lysin with Unusual Modular Design. *mSphere* **2**, (2017).

35. Li, C. *et al.* Isolation and Characterization of Bacillus cereus Phage vB_BceP-DLc1 Reveals the Largest Member of the Φ29-Like Phages. *Microorganisms* **8**, 1750 (2020).

36. Evseev, P., Gutnik, D., Evpak, A., Kasimova, A. & Miroshnikov, K. Origin, Evolution and Diversity of φ29-like Phages–Review and Bioinformatic Analysis. *Int. J. Mol. Sci.* **25**, 10838 (2024).

37. Xing, S. *et al.* Complete genome sequence of a novel, virulent Ahjdlikevirus bacteriophage that infects Enterococcus faecium. *Arch. Virol.* **162**, 3843–3847 (2017).

38. Skowron, P. M. *et al.* Sequence, genome organization, annotation and proteomics of the thermophilic, 47.7-kb Geobacillus stearothermophilus bacteriophage TP-84 and its classification in the new Tp84virus genus. *PLoS One* **13**, e0195449 (2018).

39. Pourcel, C., Essoh, C., Ouldali, M. & Tavares, P. Acinetobacter baumannii satellite phage Aci01-2-Phanie depends on a helper myophage for its multiplication. *J. Virol.* **98**, (2024).

40. Lossouarn, J. *et al.* Enterococcus faecalis Countermeasures Defeat a Virulent Picovirinae Bacteriophage. *Viruses* **11**, 48 (2019).

41. Peng, W. *et al.* Isolation and genomic analysis of temperate phage 5W targeting multidrug-resistant Acinetobacter baumannii. *Arch. Microbiol.* **204**, 58 (2022).

42. Stanton, C. R., Rice, D. T. F., Beer, M., Batinovic, S. & Petrovski, S. Isolation and Characterisation of the Bundooravirus Genus and Phylogenetic Investigation of the Salasmaviridae Bacteriophages. *Viruses* **13**, 1557 (2021).

43. King, A. J. & van der Vliet, P. C. A precursor terminal protein-trinucleotide intermediate during initiation of adenovirus DNA replication: regeneration of molecular ends in vitro by a jumping back mechanism. *EMBO J.* **13**, 5786–5792 (1994).

44. Kazlauskas, D., Krupovic, M., Guglielmini, J., Forterre, P. & Venclovas, Č. Diversity and evolution of B-family DNA polymerases. *Nucleic Acids Res.* **48**, 10142–10156 (2020).

45. Tian, R. *et al.* Establishing a synthetic orthogonal replication system enables accelerated evolution in E. coli. *Science (80-.).* **383**, 421–426 (2024).

46. Tian, R. *et al.* Engineered bacterial orthogonal DNA replication system for continuous evolution. *Nat. Chem. Biol.* **19**, 1504–1512 (2023).

47. Williams, R. L. & Liu, C. C. Accelerated evolution of chosen genes. *Science (80-.).* **383**, 372–373 (2024).

48. Rix, G. *et al.* Continuous evolution of user-defined genes at 1 million times the genomic mutation rate. *Science (80-.).* **386**, (2024).

49. Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular ancestors. *Proc. Natl. Acad. Sci.* **114**, (2017).

50. Jeong, D.-E. *et al.* DNA Polymerase Diversity Reveals Multiple Incursions of Polintons During Nematode Evolution. *Mol. Biol. Evol.* **40**, (2023).

51. Wuitschick, J. D. A novel family of mobile genetic elements is limited to the germline genome in Tetrahymena thermophila. *Nucleic Acids Res.* **30**, 2524–2537 (2002).

52. Mougari, S. *et al.* Guarani virophage, a new sputnik-like isolate from a Brazilian lake. *Front. Microbiol.* **10**, (2019).

53. Mougari, S., Sahmi-Bounsiar, D., Levasseur, A., Colson, P. & Scola, B. La. Virophages of giant viruses: An update at eleven. *Viruses* **11**, 1–28 (2019).

54. Gaia, M. *et al.* Zamilon, a novel virophage with Mimiviridae host specificity. *PLoS One* **9**, 1–8 (2014).

55. La Scola, B. *et al.* The virophage as a unique parasite of the giant mimivirus. *Nature* **455**, 100–104 (2008).

56. Desnues, C. *et al.* Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 18078–18083 (2012).

57. Iyer, L. M., Abhiman, S. & Aravind, L. A new family of polymerases related to superfamily A DNA polymerases and T7-like DNA-dependent RNA polymerases. *Biol. Direct* **3**, 39 (2008).

58. Kuznetsova, A. A., Fedorova, O. S. & Kuznetsov, N. A. Structural and Molecular Kinetic Features of Activities of DNA Polymerases. *Int. J. Mol. Sci.* **23**, 6373 (2022).

59. Krupovic, M., Dolja, V. V. & Koonin, E. V. The virome of the last eukaryotic common ancestor and eukaryogenesis. *Nat. Microbiol.* **8**, 1008–1017 (2023).

60. Cramer, P. AlphaFold2 and the future of structural biology. *Nat. Struct. Mol. Biol.* **28**, 704–705 (2021).

61. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

62. Lee, J.-W. *et al.* DeepFold: enhancing protein structure prediction through optimized loss functions, improved template features, and re-optimized energy function. *Bioinformatics* **39**, (2023).

63. Holm, L., Laiho, A., Törönen, P. & Salgado, M. DALI shines a light on remote homologs: One hundred discoveries. *Protein Sci.* **32**, (2023).

64. Graham, F. Daily briefing: AlphaFold3 is now open source. *Nature* (2024) doi:10.1038/d41586-024-03728-0.

65. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

66. Boitreaud, J. *et al.* Chai-1: Decoding the molecular interactions of life. (2024) doi:10.1101/2024.10.10.615955.

67. Wohlwend, J. *et al.* Boltz-1 Democratizing Biomolecular Interaction Modeling. (2024) doi:10.1101/2024.11.19.624167.

68. Álvarez-Salmoral, D. *et al.* AlphaBridge: tools for the analysis of predicted macromolecular complexes. (2024) doi:10.1101/2024.10.23.619601.

69. Ofer, D., Brandes, N. & Linial, M. The language of proteins: NLP, machine learning & protein sequences. *Comput. Struct. Biotechnol. J.* **19**, 1750–1758 (2021).

70. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).

71. Praljak, N. *et al.* Natural Language Prompts Guide the Design of Novel Functional Protein Sequences. (2024) doi:10.1101/2024.11.11.622734.

72. Liu, S. *et al.* A text-guided protein design framework. *Nat. Mach. Intell.* **7**, 580–591 (2025).

73. Masters, M. R., Mahmoud, A. H., Wei, Y. & Lill, M. A. Deep Learning Model for Efficient Protein–Ligand Docking with Implicit Side-Chain Flexibility. *J. Chem. Inf. Model.* **63**, 1695–1707 (2023).

74. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).

75. Moi, D. *et al.* Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. (2023) doi:10.1101/2023.09.19.558401.

76. Gilchrist, C. L. M., Mirdita, M. & Steinegger, M. Multiple Protein Structure Alignment at Scale with FoldMason. (2024) doi:10.1101/2024.08.01.606130.

77. Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods* **19**, 1109–1115 (2022).

78. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

79. Kim, R. S., Levy Karin, E., Mirdita, M., Chikhi, R. & Steinegger, M. BFVD–a large repository of predicted viral protein structures. *Nucleic Acids Res.* **53**, D340–D347 (2025).

80. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science (80-.).* **379**, 1123–1130 (2023).

81. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).

82. Terzian, P. *et al.* PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genomics Bioinforma.* **3**, (2021).

83. Goldman, G., Chati, P. & Ntranos, V. Uncovering differential tolerance to deletions versus substitutions with a protein language model. (2024) doi:10.1101/2024.06.27.601077.

84. Małolepsza, E. *et al.* Symmetrization of the AMBER and CHARMM force fields. *J. Comput. Chem.* **31**, 1402–1409 (2010).

85. Khan, M. K. & McLean, D. J. Durga: an R package for effect size estimation and visualization. *J. Evol. Biol.* **37**, 986–993 (2024).

86. Hekkelman, M. L., Álvarez Salmoral, D., Perrakis, A. & Joosten, R. P. DSSP 4: FAIR annotation of protein secondary structure. (2025) doi:10.1101/2025.04.11.648460.

87. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS One* **8**, e80635 (2013).

88. Szilágyi, A. & Skolnick, J. Efficient Prediction of Nucleic Acid Binding Function from Low-resolution Protein Structures. *J. Mol. Biol.* **358**, 922–933 (2006).

89. Varadi, M. *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).

90. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. in *Plant Bioinformatics* 89–112 (Humana Press, 2007). doi:10.1007/978-1-59745-535-0_4.

91. Coelho, L. P. *et al.* Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256 (2022).

92. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020**, (2020).

93. Altschul, S. F. & Koonin, E. V. Iterated profile searches with PSI-BLAST–a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**, 444–447 (1998).

94. Bhagwat, M. & Aravind, L. PSI-BLAST Tutorial. in *Comparative Genomics. Methods in Molecular Biology™* (ed. Bergman, N. H.) 177–186 (Humana Press, 2007). doi:10.1007/978-1-59745-514-5_10.

95. Gao, C.-H., Yu, G. & Cai, P. ggVennDiagram: An Intuitive, Easy-to-Use, and Highly Customizable R Package to Generate Venn Diagram. *Front. Genet.* **12**, (2021).

96. Gao, C. *et al.* ggVennDiagram: Intuitive Venn diagram software extended. *iMeta* **3**, (2024).

97. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

98. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).

99. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

100. Agarwala, R. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8–D13 (2018).

101. Eddy, S. R. A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation. *PLoS Comput. Biol.* **4**, e1000069 (2008).

102. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2024).

103. Shen, W., Sipos, B. & Zhao, L. SeqKit2: A Swiss army knife for sequence and alignment processing. *iMeta* **3**, (2024).

104. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

105. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).

106. Rodríguez del Río, Á. *et al.* Functional and evolutionary significance of unknown genes from uncultivated taxa. (2022) doi:10.1101/2022.01.26.477801.

107. Rodríguez del Río, Á. *et al.* Functional and evolutionary significance of unknown genes from uncultivated taxa. *Nature* **626**, 377–384 (2024).

108. Escarmis, C. *et al.* Inverted terminal repeats and terminal proteins of the genomes of pneumococcal phages. *Gene* **36**, 341–348 (1985).

109. Ismail, A. *et al.* Evidence of a Set of Core-Function Genes in 16 Bacillus Podoviral Genomes with Considerable Genomic Diversity. *Viruses* **15**, 276 (2023).

110. Si, J., Zhao, R. & Wu, R. An Overview of the Prediction of Protein DNA-Binding Sites. *Int. J. Mol. Sci.* **16**, 5194–5215 (2015).

111. Lou, W. *et al.* Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naïve Bayes. *PLoS One* **9**, e86703 (2014).

112. He, S. Y. & Collmer, A. Molecular cloning, nucleotide sequence, and marker exchange mutagenesis of the exo-poly-alpha-D-galacturonosidase-encoding pehX gene of Erwinia chrysanthemi EC16. *J. Bacteriol.* **172**, 4988–4995 (1990).

113. Sprockett, D. D., Piontkivska, H. & Blackwood, C. B. Evolutionary analysis of glycosyl hydrolase family 28 (GH28) suggests lineage-specific expansions in necrotrophic fungal pathogens. *Gene* **479**, 29–36 (2011).

114. Markovič, O. & Janeček, Š. Pectin degrading glycoside hydrolases of family 28: sequence-structural features, specificities and evolution. *Protein Eng. Des. Sel.* **14**, 615–631 (2001).

115. Villarreal, F., Stocchi, N. & ten Have, A. Functional Classification and Characterization of the Fungal Glycoside Hydrolase 28 Protein Family. *J. Fungi* **8**, 217 (2022).

116. Singleton, M. R., Dillingham, M. S. & Wigley, D. B. Structure and Mechanism of Helicases and Nucleic Acid Translocases. *Annu. Rev. Biochem.* **76**, 23–50 (2007).

117. Knizewski, L., Kinch, L. N., Grishin, N. V, Rychlewski, L. & Ginalski, K. Realm of PD-(D/E)XK nuclease superfamily revisited: detection of novel families with modified transitive meta

profile searches. *BMC Struct. Biol.* **7**, 40 (2007).

118. Steczkiewicz, K., Muszewska, A., Knizewski, L., Rychlewski, L. & Ginalski, K. Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily. *Nucleic Acids Res.* **40**, 7016–7045 (2012).

119. Laganeckas, M., Margelevičius, M. & Venclovas, Č. Identification of new homologs of PD-(D/E)XK nucleases by support vector machines trained on data derived from profile–profile alignments. *Nucleic Acids Res.* **39**, 1187–1196 (2011).

120. Knizewski, Ł., Kinch, L., Grishin, N. V., Rychlewski, L. & Ginalski, K. Human Herpesvirus 1 UL24 Gene Encodes a Potential PD-(D/E)XK Endonuclease. *J. Virol.* **80**, 2575–2577 (2006).

121. Kinch, L. N., Ginalski, K., Laszek, R. & Grishin, N. V. Identification of novel restriction endonuclease-like fold families among hypothetical proteins. *Nucleic Acids Res.* **33**, 3598–3605 (2005).

122. Sirias, D., Utter, D. R. & Gibbs, K. A. A family of contact-dependent nuclease effectors contain an exchangeable, species-identifying domain. (2020) doi:10.1101/2020.02.20.956912.

123. Šišáková, E., Stanley, L. K., Weiserová, M. & Szczelkun, M. D. A RecB-family nuclease motif in the Type I restriction endonuclease EcoR124I. *Nucleic Acids Res.* **36**, 3939–3949 (2008).

124. Haldenby, S., White, M. F. & Allers, T. RecA family proteins in archaea: RadA and its cousins. *Biochem. Soc. Trans.* **37**, 102–107 (2009).

125. Seitz, E. M., Brockman, J. P., Sandler, S. J., Clark, A. J. & Kowalczykowski, S. C. RadA protein is an archaeal RecA protein homolog that catalyzes DNA strand exchange. *Genes Dev.* **12**, 1248–1253 (1998).

126. Ferlez, B., Huang, P. T., Hu, J. & Brooks Crickard, J. Evolution of Eukaryotic Specific DNA Binding Sites in Asgard Archaeal RecA Recombinases. (2025) doi:10.1101/2025.02.19.639135.

127. Golosova, N. N. *et al.* Bacteriophage vB_SepP_134 and Endolysin LysSte_134_1 as Potential Staphylococcus-Biofilm-Removing Biological Agents. *Viruses* **16**, 385 (2024).

128. Culbertson, E. K. *et al.* Draft Genome Sequences of Staphylococcus Podophages JBug18, Pike, Pontiff, and Pabna. *Microbiol. Resour. Announc.* **8**, (2019).

129. Valente, L. G. *et al.* Isolation and characterization of bacteriophages from the human skin microbiome that infect Staphylococcus epidermidis. *FEMS Microbes* **2**, (2021).

130. Islam, R., Brown, S., Taheri, A. & Dumenyo, C. K. The Gene Encoding NAD-Dependent Epimerase/Dehydratase, wcaG, Affects Cell Surface Properties, Virulence, and Extracellular Enzyme Production in the Soft Rot Phytopathogen, Pectobacterium carotovorum. *Microorganisms* **7**, 172 (2019).

131. Xiao, M. *et al.* Fucose-containing bacterial exopolysaccharides: Sources, biological activities, and food applications. *Food Chem. X* **13**, 100233 (2022).

132. Verstraete, W. & Philips, S. Nitrification–denitrification processes and technologies in new contexts. in *Nitrogen, the Confer-N-s* (eds. Hoek, K. W. Van der, Erisman, J. W., Smeulders, S., Wisniewski, J. R. & Wisniewski, J.) 717–726 (Elsevier, 1998). doi:10.1016/B978-0-08-043201-4.50102-7.

133. Cruz, J. D. *et al.* Bioprospecting for industrially relevant exopolysaccharide-producing cyanobacteria under Portuguese simulated climate. *Sci. Rep.* **13**, 13561 (2023).

134. Cai, H. *et al.* Genomic Analysis and Taxonomic Characterization of Seven Bacteriophage Genomes Metagenomic-Assembled from the Dishui Lake. *Viruses* **15**, 2038 (2023).

135. Vogel, U., Beerens, K. & Desmet, T. Nucleotide sugar dehydratases: Structure, mechanism, substrate specificity, and application potential. *J. Biol. Chem.* **298**, 101809 (2022).

136. Beerens, K., Gevaert, O. & Desmet, T. GDP-Mannose 3,5-Epimerase: A View on Structure, Mechanism, and Industrial Potential. *Front. Mol. Biosci.* **8**, (2022).

137. Ahmed, S. H., Bose, D. B., Khandoker, R. & Rahman, M. S. StackDPP: a stacking ensemble based DNA-binding protein prediction model. *BMC Bioinformatics* **25**, 111 (2024).

138. Trinquier, J. *et al.* SoftAlign: End-to-end protein structures alignment. (2025) doi:10.1101/2025.05.09.653096.

9. Annex I. GitHub repository

All input and output data files, scripts and Google Colaboratory notebooks can be found in our *ad hoc* GitHub repository (https://github.com/rnrlab/TP).

10. Annex II. Supplementary figures

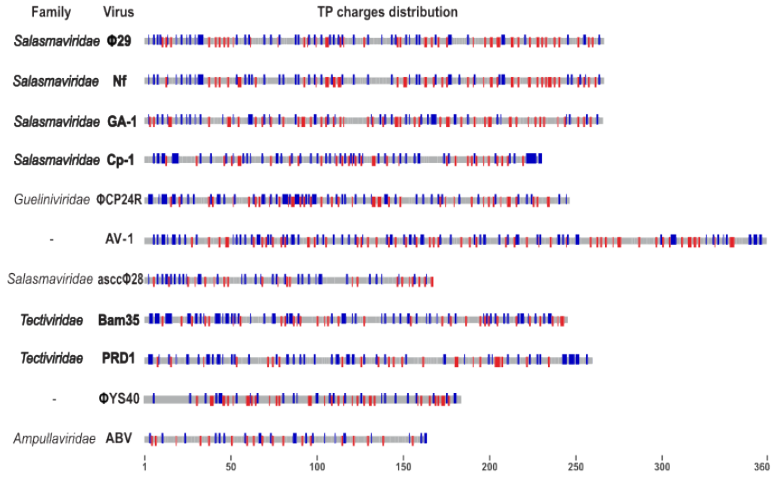

Figure S17. Empirically isolated and scrutinised TP residue charge distribution. Modified from Redrejo-Rodríguez *et al.* (2012)[15] with author's approval.