Review

Not peer-reviewed version

Phage Bioinformatics Tools: A Review of Computational Approaches for Bacteriophage Research

Sean Jia Le Pang , Soon Keong Wee , Eric Peng Huat Yap *

Posted Date: 26 May 2026

doi: 10.20944/preprints202605.1740.v1

Keywords: bacteriophage; bioinformatics; foundation models; protein structure; host prediction; viral metagenomics

 doi:10.20944/preprints202605.1740.v1

Review

Phage Bioinformatics Tools: A Review of Computational Approaches for Bacteriophage Research

Sean Jia Le Pang [1], Soon Keong Wee [2] and Eric Peng Huat Yap [2,3,4,*]

[1] Institute for Digital Molecular Analytics and Science, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore
[2] Institute for Digital Molecular Analytics and Science, Nanyang Technological University, Singapore
[3] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
[4] National Centre for Infectious Diseases, Singapore
* Correspondence: ericyap@ntu.edu.sg

Abstract

Rising clinical interest in phage therapy and the exponential growth of metagenomic sequence catalogues have driven a rapid expansion of bacteriophage bioinformatics. More than 80 dedicated tools, mostly published since 2020, now span identification, assembly, annotation, taxonomy, lifestyle prediction, defence-system detection and host prediction. Aimed at experienced practitioners and developers, this review synthesizes the field through the lens of three successive computational paradigms: sequence homology, bounded by database completeness; machine learning, constrained by labelled training data; and foundation models, which now achieve Matthews correlation coefficients above 0.95 for identification via structure-informed prediction. Furthermore, we map the upstream components, namely gene callers, homology engines, protein language models, and structural search tools, that underpin most downstream pipelines, exposing shared infrastructure and ecosystem-level fragility when dependencies change. To translate this into practice, we propose web-based and command-line reference workflows calibrated to user expertise and sample types. Finally, we set an agenda for the next wave of tool development. Roughly half of phage genes still resist functional annotation despite structural methods; no broadly generalisable strain-level host predictor exists for phage therapy; varying true positive rates (0% to 97%) underscore the absence of standardised community benchmarks analogous to CASP or CAMI. As generative genome models begin designing synthetic phages, progress will depend less on producing standalone tools than on rigorous evaluation, interoperable infrastructure, and clinically meaningful prediction targets.

Keywords: bacteriophage; bioinformatics; foundation models; protein structure; host prediction; viral metagenomics

Introduction

Rising antimicrobial resistance has renewed clinical interest in bacteriophage therapy, while high-throughput sequencing has expanded phage diversity catalogues. IMG/VR has grown from approximately 268,000 sequences in its 2017 release to more than 15 million virus genomes and fragments in version 4[1], a trajectory mirrored by curated resources such as INPHARED[2]. This expansion supplies the training data on which modern computational tools depend. More than 80 dedicated phage bioinformatics tools now span eight functional categories: identification, assembly, annotation, taxonomy, lifestyle prediction, defense system detection, host prediction, and benchmarking, with the majority published between 2020 and 2025.

Tool development has followed three computational paradigms (Figure 1). The homology era (pre-2017) relied on BLAST alignments, HMM profiles, and curated reference databases; tools such

 doi:10.20944/preprints202605.1740.v1

as VirSorter[3] and vConTACT2[4] scanned sequences against known viral markers and built gene-sharing networks from shared protein clusters, but lost sensitivity on short metagenomic contigs and novel lineages lacking close database homologs. The machine-learning era (2017–2022) replaced explicit homology with statistical patterns learned from sequence composition; VirFinder[5] and PhaGCN [6] demonstrated reference-free identification and graph-based taxonomic classification, but remained bounded by labelled training-data availability and depended on short fragments. Foundation model approaches, accelerating since 2023 and built on pretrained transformer architectures and protein language models trained via self-supervised learning on massive unlabelled bodies of omic data have pushed performance ceilings higher. With geNomad[7] performing reference-independent identification and Phold[8] introducing structure-informed functional annotation, the dark matter gap of viral genes with unknown function that defines this period is starting to close.

Figure 1. Landscape of computational methods for phage analysis across major functional tasks from 2010 till today. Timeline showing the development of tools for phage identification, assembly, annotation, taxonomy, lifestyle prediction, defense detection, and host prediction, stratified by methodological paradigm: sequence homology (pre-2017), machine learning (2017–2022), and foundation models (2023+). Each point represents an individual tool positioned by publication year, with marker shape indicating algorithmic class (homology/alignment, machine learning, or foundation model) and colour denoting implementation complexity (web-based, local installation, or GPU/large-scale requirements). Dashed vertical line marks the transition from homology-based to learning-based approaches era. Key database milestones are indicated along the bottom, highlighting the co-evolution of data resources and analytical methods.

Recent guides such as Phage quest[9] have introduced phage bioinformatics to biological and medical researchers new to the field. This review takes a complementary but distinct position: it is aimed at experienced users applying tools in practice and at developers building the next generation of tools. We systematically survey more than 80 tools organised by computational paradigm (Table 1), synthesise benchmarking evidence and identify where foundation models are beginning to

 doi:10.20944/preprints202605.1740.v1

address persistent gaps (ranging from viral dark matter to strain-level host prediction) for phage therapy, and identify the key limitations of current approaches, and outline future directions to guide the next phase of tool development.

 doi:10.20944/preprints202605.1740.v1

Table 1. Bioinformatics tools for bacteriophage research.

Databases							
No.	Tool	Year	Description	Key feature	Citations (Apr 2026)	Availability	URL
T1	NCBI Viral Genomes Resource	2015	Curated complete viral genomes; RefSeq records	INSDC-linked reference standard	777	Web	https://www.ncbi.nlm.nih.gov/genome/viruses/
T2	PhagesDB	2017	Actinobacteriophage database (SEA-PHAGES)	>30,000 entries; >5,000 sequenced genomes as of 2026	540	Web	https://phagesdb.org/
T3	INPHARED	2021	Automated curation of complete phage genomes	GitHub-distributed; automated updates; revealed 75% sampling bias	371	GitHub	https://github.com/RyanCook94/inphared
T4	ICTV VMR	2022	Exemplar virus taxonomy reference	Official nomenclature alignment	1,504	Web	https://ictv.global/vmr
T5	IMG/VR v4	2023	Uncultivated virus genome repository	>15 million genomes; 6-fold increase over v3	446	Web	https://img.jgi.doe.gov/vr/
Identification and detection							
No.	Tool	Year	Approach	Key metric / feature	Citations (Apr 2026)	Availability	URL
T6	VirSorter	2015	Hallmark gene search + enrichment metrics	>95% recall on contigs ≥10 kb	1,182	GitHub	https://github.com/simroux/VirSorter
T7	MetaPhinder	2016	BLAST integration across reference genomes	Sequence-level classification	84	GitHub + Web	https://github.com/vanessajurtz/MetaPhinder

 doi:10.20944/preprints202605.1740.v1

T8	VirFinder	2017	Logistic regression on k-mer frequencies	78× higher TPR than VirSorter at 1 kb	701	GitHub	https://github.com/jessieren/VirFinder
T9	MARVEL	2018	Random Forest on genomic features	Extended RF-based detection	191	GitHub	https://github.com/LaboratorioBioinformatica/MARVEL
T10	PPR-Meta	2019	Deep learning	Three-way classification (phage/plasmid/chromosome)	200	GitHub	https://github.com/zhenchengfang/PPR-Meta
T11	DeepVirFinder	2020	Alignment-free CNN classifier	AUROC 0.93–0.98 for 300–3,000 bp	691	GitHub	https://github.com/jessieren/DeepVirFinder
T12	Seeker	2020	LSTM on raw DNA	Alignment-free; no feature engineering	137	GitHub	https://github.com/gussow/seeker
T13	VIBRANT	2020	Neural network + protein annotation	Automated recovery, annotation, curation	1,128	GitHub	https://github.com/AnantharamanLab/VIBRANT
T14	Kraken2	2019	k-mer exact matching	Precision 0.96 in mock community (Ho et al. benchmark)	7,218	GitHub	https://github.com/DerrickWood/kraken2
T15	VirSorter2	2021	Multi-classifier framework	F1 > 0.8 DNA and RNA virus detection	1,279	GitHub	https://github.com/jiarong/VirSorter2
T16	PhaMer	2022	Transformer on protein tokens	27% F1 improvement on real metagenomic data	65	GitHub + Web	https://github.com/KennthShang/PhaMer
T17	INHERIT	2022	DNABERT-style transformer	Representation learning for phage genomes	29	GitHub	https://github.com/Celestial-Bai/INHERIT
T18	geNomad	2023	IGLOO encoder + CRF for proviruses	MCC 95.3%; virus ID + taxonomy + annotation in one framework	884	GitHub	https://github.com/apcamargo/genomad

Prophage detection

No.	Tool	Year	Approach	Key metric / feature	Citations (Apr 2026)	Availability	URL
T19	PhiSpy	2012	Similarity + composition (7 genomic features)	94% prediction success; 0.66% FPR across 50 genomes	593	GitHub + Web	https://github.com/linsalrob/PhiSpy

 doi:10.20944/preprints202605.1740.v1

T20	PHASTER	2016	Curated database search (web server)	Widely used prophage web server	3,855	Web	https://phaster.ca/
T21	DEPhT	2022	Multimodal approach (3 run modes)	Improved prophage boundary determination	38	GitHub	https://github.com/chg60/DEPhT
T22	PHASTEST	2023	Updated PHASTER pipeline	31% faster; Higher sensitivity than phaster	554	Web	https://phastest.ca/
T23	PhageBoost	2021	Machine-learning-driven (evaluates viral genomic architecture)	Detects highly divergent prophages; tast, for high-throughput WGS	74	GitHub	https://github.com/ku-cbd/PhageBoost
Genome assembly and comparative genomics							
No.	Tool	Year	Approach	Key metric / feature	Citations (Apr 2026)	Availability	URL
T24	SPAdes	2012	de Bruijn graph assembly	Foundation for metaSPAdes/metaviralSPAdes	27,729	GitHub + Web	https://github.com/ablab/spades
T25	ViromeQC	2019	Sample-level contamination assessment	Quantifies non-viral contamination in VLP viromes	115	GitHub	https://github.com/SegataLab/viromeqc
T26	metaviralSPAdes	2020	Virus-specific SPAdes extension	Viral subgraph ID + completeness assessment; also used for contig verification	320	GitHub	https://github.com/ablab/spades
T27	VIRIDIC	2020	Nucleotide intergenomic similarity	ICTV-recommended algorithm for genus/species demarcation (also used in Taxonomy)	819	GitHub + Web	https://github.com/CristinaMoraru/VIRIDIC
T28	CheckV	2021	Genome quality assessment	5 quality tiers; 76,262 reference genomes; adopted by IMG/VR	1,790	Bitbucket	https://bitbucket.org/berkeleylab/checkv
T29	Clinker	2021	Gene cluster comparison visualisation	Automated comparison figures	1,393	GitHub	https://github.com/gamcil/clinker
T30	pyGenomeViz	2022	Genome comparison	Annotated comparative genomics figures	N/A	GitHub	https://github.com/moshi4/pyGenomeViz

 doi:10.20944/preprints202605.1740.v1

			visualisation (Python)				
T31	viralFlye	2022	Long-read viral assembly	Long-read metagenomics support	32	GitHub	https://github.com/Dmitry-Antipov/viralFlye
T32	vRhyme	2022	Viral contig binning	Coverage + nucleotide composition signals	113	GitHub	https://github.com/AnantharamanLab/vRhyme
T33	Phables	2023	Flow decomposition on assembly graphs	49% more high-quality genomes than existing tools	42	GitHub	https://github.com/Vini2/phables
T34	COBRA	2024	Paired-end extension of incomplete assemblies	Improves completeness and contiguity	46	GitHub	https://github.com/linxingchen/cobra
Gene annotation and functional prediction							
No.	Tool	Year	Approach	Key metric / feature	Citations (Apr 2026)	Availability	URL
T35	Prodigal	2010	Prokaryotic gene recognition	Fast; widely used general gene caller	13,001	GitHub	https://github.com/hyattpd/Prodigal
T36	pVOGs	2017	Prokaryotic virus orthologous groups	Viral marker gene detection; integrated into many pipelines	414	Web	https://ftp.ncbi.nlm.nih.gov/pub/kristensen/pVOGs/
T37	HHpred	2018	Profile-profile HMM comparison	Remote homology detection	2,719	Web	https://toolkit.tuebingen.mpg.de/tools/hhpred
T38	PHANOTATE	2019	Dynamic programming for phage ORFs	Finds genes missed by Prodigal/GeneMarkS/Glimmer; handles overlapping frames	319	GitHub	https://github.com/deprekate/PHANOTATE
T39	DRAM-v	2020	Metabolic pathway annotation	Identifies auxiliary metabolic genes (AMGs)	1,097	GitHub	https://github.com/WrightonLabCSU/DRAM
T40	PhANNs	2020	ANN structural protein classifier	F1 = 0.875 across 10 structural classes	113	GitHub + Web	https://github.com/Adrian-Cantu/PhANNs

 doi:10.20944/preprints202605.1740.v1

T41	PHROGs	2021	HMM protein clustering (38,880 clusters)	50.6% functional annotation across 17,473 reference viruses	425	Web	https://phrogs.lmge.uca.fr/
T42	MultiPhATE2	2021	Parallel multi-gene-finder pipeline	Runs multiple gene finders simultaneously	24	GitHub	https://github.com/carolzhou/multiPhATE2
T43	Pharokka	2023	Integrated pipeline (PHANOTATE + PHROGs)	Standard pipeline; <5 min for 50 kb genome	519	GitHub	https://github.com/gbouras13/pharokka
T44	VPF-PLM	2024	Protein language model annotation	+29% annotated ocean virome protein families	95	GitHub	https://github.com/kellylab/viral-protein-function-plm
T45	ProstT5	2024	Protein sequence → 3Di structural alphabet	Bilingual sequence-structure translation	364	GitHub	https://github.com/mheinzinger/ProstT5
T46	Foldseek	2024	Ultrafast structural search	4-5 orders of magnitude faster than Dali/TM-align at comparable sensitivity	2,365	GitHub + Web	https://github.com/steineggerlab/foldseek
T47	Empathi	2025	Hierarchical protein embeddings	2× on environment viromes (EnVhogDB); 3× on cultured phages	6	HuggingFace	https://huggingface.co/AlexandreBoulay/EmPATHi
T48	GOPhage	2025	Genomic context + transformer embeddings	+6.78% accuracy on divergent proteins	9	GitHub	https://github.com/jiaojiaoguan/GOPhage
T49	Phold	2026	Structure-informed (ProstT5 → Foldseek)	>50% gene annotation vs ~35% homology-only	30	GitHub	https://github.com/gbouras13/phold

Taxonomy and classification

No.	Tool	Year	Approach	Key metric / feature	Citations (Apr 2026)	Availability	URL
T50	VICTOR	2017	Genome-BLAST Distance Phylogeny	Automated species/genus demarcation	699	Web	https://victor.dsmz.de/
T51	ViPTree	2017	Genome-wide tBLASTx proteomic trees	Viral proteomic tree server	1,004	Web	https://www.genome.jp/viptree/
T52	vConTACT2	2019	Gene-sharing networks	96% genus-level ICTV agreement	990	Bitbucket	https://bitbucket.org/MAVERICLab/vcontact2

 doi:10.20944/preprints202605.1740.v1

				(pre-2022 framework)			
T53	PhaGCN	2021	Graph convolutional network + CNN	Semi-supervised classification	143	GitHub	https://github.com/KennthShang/PhaGCN
T54	PhaGCN2	2023	Extended PhaGCN (DNA + RNA viruses)	89.30% recall; 83.91% precision; applied to GPD and GOV2.0 datasets	112	GitHub + Web	https://github.com/KennthShang/PhaGCN2.0
T55	PhaBOX	2023	Integrated platform (PhaMer + PhaGCN + CHERRY + PhaTYP)	Unified ID + taxonomy from metagenomes (also used in Host prediction)	117	GitHub + Web	https://github.com/KennthShang/PhaBOX
T56	GRAViTy-V2	2024	Composite generalised Jaccard distances	Genome relationship analysis	10	GitHub	https://github.com/Mayne941/gravity2
T57	taxMyPhage	2025	Automated genus/species classification	Aligned with current ICTV revisions; dsDNA phages	76	GitHub	https://github.com/amillard/tax_myPHAGE
T58	vConTACT3	2025	ML-based hierarchical classification	>95% ICTV agreement (97.6% genus, 98.7% subfamily, 100% family/order)	6	Bitbucket	https://bitbucket.org/MAVERICLab/vcontact3

Lifestyle prediction

No.	Tool	Year	Approach	Key metric / feature	Citations (Apr 2026)	Availability	URL
T59	PHACTS	2012	Random Forest on protein similarity	99% precision (confident predictions); 88% overall sensitivity	305	GitHub	https://github.com/deprekate/PHACTS
T60	BACPHLIP	2021	HMM domains + Random Forest	98.3% accuracy on 423 independent test phages	294	GitHub	https://github.com/adamhockenberry/bacphlip
T61	DeePhage	2021	CNN on one-hot encoded DNA	89% accuracy; classifies contigs ≥100 bp	113	GitHub	https://github.com/shufangwu/DeePhage

 doi:10.20944/preprints202605.1740.v1

T62	PhaTYP	2023	BERT pre-trained + fine-tuned	Outperforms prior methods on short contigs	206	GitHub + Web	https://github.com/KennthShang/PhaTYP
T63	DeepPL	2024	NLP on nucleotide sequences	94.65% accuracy	10	GitHub	https://github.com/Wu-Microbiology/DeepPL
T64	ProkBERT PhaStyle	2025	Genomic language models (21–26M params)	BA 0.88–0.93; MCC 0.75–0.86 on 500 bp fragments	1	GitHub	https://github.com/nbrg-ppcu/PhaStyle

Anti-phage defense system detection

No.	Tool	Year	Approach	Key metric / feature	Citations (Apr 2026)	Availability	URL
T65	CRISPRDetect	2016	Array detection + boundary refinement	CRISPR array identification	393	GitHub + Web	https://github.com/davidchyou/CRISPRDetect_2.4
T66	CRISPRCasFinder	2018	Integrated array + Cas protein ID	Combined array and Cas detection	1,561	GitHub + Web	https://github.com/dcouvin/CRISPRCasFinder
T67	AcrFinder	2020	Homology + guilt-by-association + self-targeting spacers	Anti-CRISPR operon mining	80	GitHub + Web	https://github.com/HaidYi/acrfinder
T68	AcRanker	2020	XGBoost ranking	Identified AcrIIA20 and AcrIIA21	119	GitHub	https://github.com/amina01/AcRanker
T69	CRISPRCasTyper	2020	Automated subtype classification	98.6% accuracy	303	GitHub + Web	https://github.com/Russel88/CRISPRCasTyper
T70	PADLOC	2021/22	HMM + system completeness validation	Web server; customisable classifications	253	GitHub + Web	https://github.com/padlocbio/padloc
T71	DefenseFinder	2022	HMM profiles + MacSyFind	60 families; 151 subtypes across 21,000 genomes	775	GitHub + Web	https://github.com/mdmparis/defense-finder

 doi:10.20944/preprints202605.1740.v1

			er rule engine				
T72	AcrNET	2023	Deep learning anti-CRISPR prediction	Beyond homology-based methods	24	GitHub	https://github.com/banma12956/AcrNET
T73	AcaFinder	2023	Aca gene detection	Independent signal for novel anti-CRISPR loci	22	GitHub	https://github.com/boweny920/AcaFinder
T74	DefensePredictor	2025	Protein language model embeddings	45 novel defense systems validated across 69 *E. coli* strains	N/A	GitHub	https://github.com/Alextianyf/DefensePredictor

Host prediction

No.	Tool	Year	Approach	Key metric / feature	Citations (Apr 2026)	Availability	URL
T75	HostPhinder	2016	k-mer similarity to reference DB	First dedicated host prediction tool	193	GitHub + Web	https://github.com/julvi/HostPhinder
T76	WIsH	2017	Markov models on host genomes	Up to 63% genus accuracy; 100× faster than alignment	317	GitHub	https://github.com/soedinglab/WIsH
T77	RaFAH	2021	Random Forest on 43,644 protein clusters	Consistent across RefSeq/SAG/metagenomic benchmarks	145	GitHub	https://github.com/felipehcoutinho/RaFAH
T78	HostG	2021	Graph convolutional network (semi-supervised)	GCN-based host prediction	59	GitHub	https://github.com/KenthShang/HostG
T79	SpacePHARER	2021	Protein-level CRISPR spacer matching	1.4–4× sensitivity over BLASTN at metagenomic scale	109	GitHub	https://github.com/soedinglab/spacepharer

 doi:10.20944/preprints202605.1740.v1

T80	CHERRY	2022	Knowledge graph + graph convolutional encoder	Improved species-level accuracy	105	GitHub + Web	https://github.com/KenthShang/CHERRY
T81	PHIST	2022	k-mer-based alignment-free	+3–20 pp species accuracy; laptop-scale runtime	59	GitHub	https://github.com/refresh-bio/PHIST
T82	PHISDetector	2022	Unified CRISPR/prophage/similarity platform	Single platform for multiple interaction signals	52	GitHub + Web	https://github.com/HIT-ImmunologyLab/PHISDetector
T83	vHULK	2022	Neural network on viral protein family scores	Annotated genomic features input	36	GitHub	https://github.com/LaboratorioBioinformatica/vHULK
T84	iPHoP	2023	Integrated ensemble (homology + CRISPR + k-mer + ML)	1.5–13× more predictions at equivalent FDR; >300 GB database	413	Bitbucket	https://bitbucket.org/srouxjgi/iphop
T85	PhageHostLearn	2024	ESM-2 embeddings of RBP + receptor sequences	ROC AUC 81.8%; strain-level for *Klebsiella*	60	GitHub	https://github.com/dimiboeckaerts/PhageHostLearn
T86	PHPGAT	2025	Graph attention on heterogeneous knowledge graphs	Multimodal phage-host interaction prediction	14	GitHub	https://github.com/ZhaoZMer/PHPGAT
T87	PHIStruct	2025	SaProt protein structure embeddings	+7–9% over sequence-only for divergent phages	20	GitHub	https://github.com/bioinfodlsu/PHIStruct
T88	MoEPH	2025	Gated Mixture-of-Experts (transformer + statistical)	Combines PLM embeddings with statistical descriptors	0	N/A	N/A

 doi:10.20944/preprints202605.1740.v1

T89	RBPseg	2025	ESMFold + structural domain ID	First large-scale phage tail fibre structure atlas	3	GitHub	https://github.com/VKleinSousa/RBPseg

Integrated pipelines and structural resources

No.	Tool	Year	Description	Key feature	Citations (Apr 2026)	Availability	URL
T90	BFVD	2025	Big Fantastic Virus Database	351,242 predicted viral protein structures; >62% novel	78	GitHub + Web	https://bfvd.foldseek.com/
T91	Sphae	2025	Snakemake workflow wrapping 12 tools	End-to-end processing in <10 min	14	GitHub	https://github.com/linsalrob/sphae

 doi:10.20944/preprints202605.1740.v1

Search Strategy and Scope

To ensure a comprehensive and reproducible evaluation of the computational landscape, literature and repository searches were conducted across PubMed, Google Scholar, bioRxiv, and GitHub up to March 2026, with specific computational tasks (e.g., identification, assembly). Inclusion criteria: Dedicated phage bioinformatics tools (published from 2012 to 2026, including preprints) introducing novel algorithmic architectures or measurable performance advancements. Exclusion criteria: General-purpose computational utilities (e.g., BLAST, HMMER), which are discussed solely as infrastructural dependencies. Performance metrics were extracted directly from original publications or independent benchmarking studies; no novel benchmarking was conducted for this review.

Phage Sequence Data and Databases

Most computational phage tools depend on reference databases for training, validation, and comparison. The quality, completeness, and taxonomic framework of these databases can directly shape tool performance and set fundamental limits on what any algorithm can detect or classify. Five databases form the reference infrastructure for phage bioinformatics.

The NCBI Viral Genomes Resource curates complete viral genomes deposited in the International Nucleotide Sequence Database Collaboration (INSDC) databases and generates Reference Sequence (RefSeq) records for each recognized viral species [10]. Most phage identification and taxonomy tools use RefSeq entries as their training or reference database.

IMG/VR version 4 is the largest repository of uncultivated virus genomes, containing more than 15 million virus genomes and genome fragments derived from metagenomes, metatranscriptomes, and metaviromes a roughly six-fold increase over the previous version, driven by systematic application of geNomad for virus detection and CheckV for quality assessment [1]. IMG/VR applies current ICTV taxonomic standards and provides functional and ecological metadata within the Joint Genome Institute (JGI) platform.

The Actinobacteriophage Database (PhagesDB) catalogues Actinobacteriophages discovered through the SEA-PHAGES programme and affiliated groups [11]. It has grown to more than 30,000 entries by 2026, with over 5,000 fully sequenced genomes (accessed March 2026), the largest dedicated resource for Actinobacteriophage diversity.

INPHARED provides an automated pipeline for retrieving, curating, and analysing complete phage genomes from public repositories with structured metadata [2]. Beyond curation, INPHARED has also revealed systemic gaps: Cook et al. [2] identified severe sampling biases in the global collection, finding that 75% of all sequenced phages had been isolated on only 30 bacterial genera a bias that skews both database content and the training data available to computational tools. INPHARED is distributed through GitHub with automated updates.

The ICTV Virus Metadata Resource (VMR) provides the definitive list of exemplar viruses for every classified species, including accession numbers, genome composition, and host information [12]. Unlike the sequence-centric databases above, the VMR is a taxonomic reference used to align computational classification with official nomenclature. Following the 2022 ICTV reclassification, further discussed in the taxonomy and classification chapter, required databases and tools alike had to adopt the revised genomic nomenclature with INPHARED, IMG/VR, and vConTACT3 [13] now incorporating it natively.

Collectively, these resources are indispensable. They enable large-scale genomic and ecological landscape studies and serve as the foundation upon which downstream phage bioinformatics tools are developed. However, the pronounced sampling biases and taxonomic skew that currently characterize these databases, inevitably propagate into the training and evaluation of any tool built upon them. Expanding phage isolation and sequencing beyond the current 30 dominant host genera to uncover the true diversity is essential for improving tool generalisability.

 doi:10.20944/preprints202605.1740.v1

Phage Identification and Prophage Detection

Phage identification within metagenomic datasets is the starting point for most downstream analyses. A typical environmental metagenome contains millions of contigs, mostly of cellular origin, with viral sequences as a minority fraction that must be computationally separated[14]. Many viral lineages share no detectable homology with characterised references given the diversity of phage genomes, so purely database-dependent approaches may miss them. Over the past decade, tools have progressed from reference-dependent homology searches through supervised machine learning to transformer-based foundation models (Figure 1). Rather than relying on a single tool, analysts can combine multiple approaches tailored to their specific requirements to achieve better results. (Figure 2).

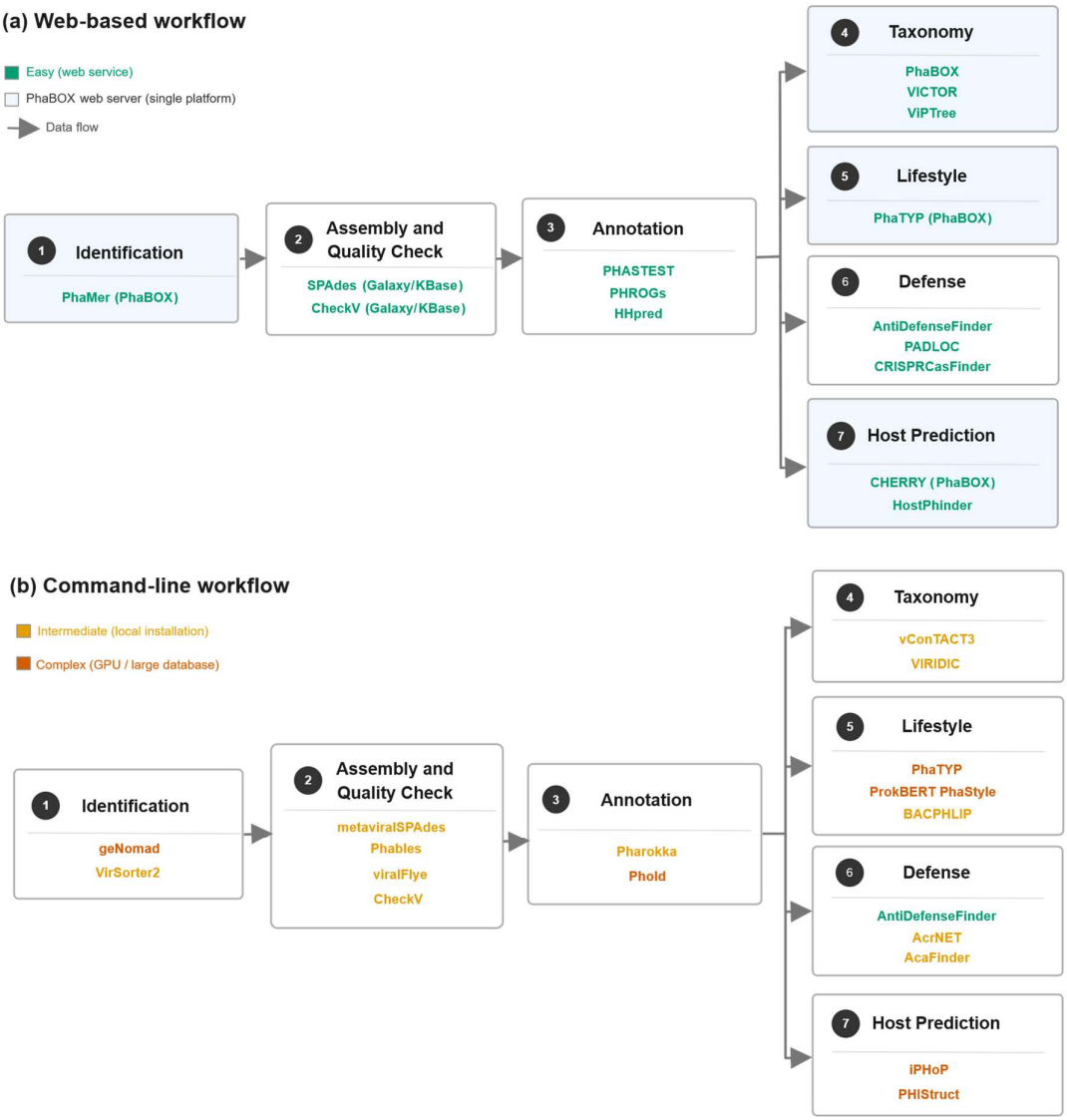

Figure 2. Proposed workflows for phage analysis of metagenomics data using (a) web-based or (b) command-line tools from identification through assembly/quality control and annotation to downstream analyses (taxonomy, lifestyle, defense systems, and host prediction), with data flow indicated between steps. Both workflows show analogous steps offering greater flexibility, scalability, and customisation based on the resources available to run the tools.

 doi:10.20944/preprints202605.1740.v1

VirSorter was the first widely adopted homology era tool for mining viral signals from microbial genomic data, combining reference-based hallmark gene searches with enrichment metrics to achieve greater than 95% recall on contigs of 10 kilobases or longer [3]. Its dependence on reference databases, however, meant sensitivity dropped on shorter contigs and novel phage lineages. Additional homology-era tools including MARVEL [15] extended detection using Random Forest classifiers trained on genomic features, resulting in better recall but remained constrained by database completeness.

VirFinder marked the shift from the homology era to the machine learning era, applying logistic regression to k-mer frequency signatures and achieving a 78-fold higher true positive rate than VirSorter on 1 kilobase contigs at equivalent false positive rates showing a considerable improvement on shorter contigs [5]. DeepVirFinder, its convolutional neural network extension, improved performance across all measured fragment lengths, reaching area under the receiver operating characteristic curve values of 0.93 to 0.98 for sequences of 300 to 3,000 bp [16] (Figure 3a). Seeker adopted a Long Short-Term Memory architecture operating directly on raw DNA sequences without feature engineering, enabling alignment-free identification of phages with little similarity to known families [17]. Other notable tools in this era, PPR-Meta [18], viralVerify [19] (see Assembly section), perform three-way classification of contigs as phage, plasmid, or chromosomal. VIBRANT [20] classifies contigs using HMM searches against viral protein families, and VirSorter2 [21] runs five random-forest classifiers in parallel, each specialised for a different viral group (dsDNA phages, NCLDVs, RNA viruses, ssDNA viruses, and virophages). Beyond these general-purpose identifiers, some tools target specific use cases: MetaPhinder [22] integrates BLAST hits across multiple reference phage genomes for sequence-level classification, while Kraken2, benchmarked by Ho et al. [14], prioritises false-positive minimisation (precision 0.96 in mock community evaluation), making it suited for studies where specificity matters more than sensitivity.

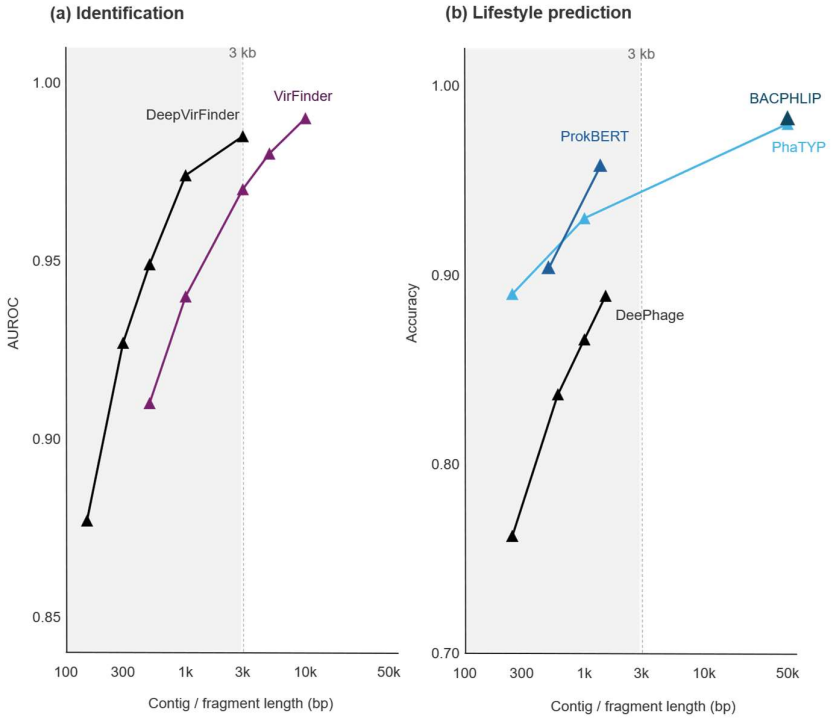

Figure 3. Performance of phage analysis models across contig lengths for (a) identification and (b) lifestyle prediction tasks. Area under the ROC curve (AUROC) for DeepVirFinder (black) and VirFinder (purple) increases with contig/fragment length, with both models approaching high accuracy (>0.98) for sequences ≥3 kb. Classification accuracy for DeePhage (black), PhaTYP (light blue), BACPHLIP (dark blue), and ProkBERT (blue) improves with fragment length, with BACPHLIP and PhaTYP achieving the highest accuracies at longer contigs.

 doi:10.20944/preprints202605.1740.v1

Dashed vertical line indicates the 3 kb contig length threshold used for reliable prediction. All values are drawn from the primary publications cited in the main text.

Transformer architectures entered phage identification in late 2022. PhaMer [23] and INHERIT [24] applied transformer and DNABERT-style [25] models respectively to phage contigs, with PhaMer achieving a 27% F1 score improvement over prior tools on real metagenomic data. For this field of virus classifications, geNomad [7] shows the strongest reported overall performance, MCC: 95.3%, F1-score: 97.3. geNomad uses a hybrid of alignment-free (IGLOO encoder) and more than 200,000 marker protein profiles to perform virus identification, plasmid detection, functional annotation, and taxonomic assignment within a single framework, achieving a Matthews correlation coefficient of 0.953 compared with 0.813 for VirSorter2.

Prophage detection requires precise boundary detection within bacterial chromosomes rather than whole-contig classification. Classical homology-based tools include PhiSpy[26], which combines similarity and composition strategies for 94% accuracy, and the widely used web server PHASTER [27] alongside its faster successor, PHASTEST[28]. For scalable analysis, DEPhT[29] offers multi-modal run settings (fast, normal and sensitive) for targeted screening, extraction, and annotation. To circumvent the computational bottlenecks and database dependencies of these homology methods, tools like PhageBoost[30] provide a purely machine-learning-driven alternative. By evaluating viral genomic architecture, such as gene density, sequence composition, and operon direction, rather than sequence similarity, PhageBoost detects highly divergent prophages at execution speeds explicitly tailored for high-throughput bacterial whole-genome sequencing (WGS) pipelines.

Despite these reported accuracy gains, independent benchmarks[14,31]reveal substantial tool-to-tool and dataset-to-dataset variability, with no single tool performing reliably across environments. Ho et al. found that 89.7% of true phage contigs were identified by more than half the tools tested, yet only 11.6% were detected by all ten tools [14], indicating that individual tools capture distinct subsets of viral diversity and thorough detection requires combining multiple approaches. Performance of various tools is highly dataset-dependent: Wu et al. documented a highly variable true positive rates ranging from 0% to 97% across nine tools evaluated on eight datasets from three biomes [31], underscoring that no single tool performs reliably across all environments; the right tool needs to be judiciously selected for the right purpose and that combining multiple complementary approaches is essential for thorough phage detection.

Genome Assembly, Quality Assessment and Comparative Genomics

Genome assembly is the checkpoint between identified viral contigs and biological interpretation. Incomplete or fragmented assemblies propagate errors into host prediction, lifestyle classification and taxonomy[32]. Phage-specific obstacles such as terminal repeats, high mutation rates, uneven coverage, and mosaic genome architectures make phage assembly more challenging than bacterial assembly[19,33] giving rise to a repertoire of quality check and assembly tools.

Tools from this category typically fall under the homology era starting with the generic assembler SPAdes [34] using multi-k-mer and a multiple de Bruijn graph framework for WGS. Its virus-specific extension metaviralSPAdes [19] is a homology-based tool that introduced modules for identifying viral subgraphs, distinguishing viral from bacterial contigs, and assessing genome completeness in metagenomic data. With the advent of long-read sequencing, new tools such as viralFlye [35] emerged for the assembly of long reads.

Post assembly refinement tools were also created to complement and boost genome assembly. COBRA [36] extends incomplete assemblies using paired-end information to generate longer and more complete assemblies. Phables [33] resolves tangled assembly graphs using flow decomposition to recover individual phage genomes through shared sequence regions. When assemblies remain fragmented, vRhyme[37] bins viral contigs into metagenome-assembled viral genomes (vMAGs) using coverage and nucleotide composition signals providing the genome-level resolution required for most downstream analyses in viral metagenomics.

 doi:10.20944/preprints202605.1740.v1

For viral genome quality assessment, CheckV [32] is the standard tool. Quality assessment is carried out through estimating completeness across five tiers (complete, high-quality, medium-quality, low-quality, and undetermined) and removing flanking host regions from provirus sequences against a database of 76,262 complete viral genomes. Its quality assessment has been adopted by IMG/VR. For sample level quality control, ViromeQC [38] quantifies non-viral contamination in VLP-enriched viromes.

For comparative genomics and visualisation, Clinker [39] and PyGenomeViz generate annotated genome comparison figures to visualize gene synteny analysis. Databases such as INPHARED [2] provide a continuously updated curated database supporting comparative analyses.

Early assembly workflows relied on general-purpose assemblers not optimised for viral genomes, often producing fragmented and incomplete contigs. The development of virus-specific modules within established frameworks (metaviralSPAdes), followed by phage-tailored approaches such as Phables, has progressively improved recovery rates. Quality assessment tools like CheckV now provide standardised completeness metrics that enable consistent evaluation across studies, though assembling complete genomes from short-read metagenomes remains difficult for phages with complex terminal structures.

Unlike identification and annotation, where foundation models have driven substantial performance gains, phage genome assembly and quality assessment remain rooted in classical graph algorithms and homology-based approaches. This persists because while language models excel at semantic pattern recognition, they currently struggle with the precise Eulerian path and De Bruijn graph resolutions required to untangle complex, fragmented assemblies caused by long terminal repeats and high mutation rates. Whether newer foundation architectures could improve genome recovery remains unexplored.

Gene Annotation and Functional Prediction

When relying solely on sequence homology, up to 80% of genes in a typical phage genome cannot be assigned a functional role [40,41], a phenomenon termed viral dark matter that reflects the deep evolutionary divergence of phage proteins from well characterised cellular homologs. Annotation tools have progressed through the same three computational eras: homology-based database searches, machine learning classifiers, and foundation model approaches including structure-informed prediction.

Gene calling, the first step in any annotation workflow, remains rooted in classical algorithms. PHANOTATE [42] addresses the distinctive features of phage gene organization, overlapping reading frames, non-standard start codons, and high gene density, using dynamic programming to evaluate every possible open reading frame. Across 2,133 complete genomes, PHANOTATE consistently identified genes missed by Prodigal [43], GeneMarkS[44], and Glimmer[45], though those four callers agreed on 82% of predictions[42]. Prodigal remains widely used for its speed and larger predicted coding sequences and forms the basis of many downstream foundational tools (Figure 4).

In the homology era, functional annotation relied on searching predicted genes against curated reference databases. PHROG [46] provides 38,880 protein clusters from 868,340 proteins across 17,473 reference prokaryotic viruses, with 5,108 functionally annotated clusters spanning nine categories via HMM profile-profile comparisons. These clusters serve as the primary annotation reference that tools such as Pharokka[47] query when assigning gene functions. Pharokka [47] has become the standard annotation pipeline, integrating gene calling (PHANOTATE or Prodigal), PHROGs assignment, and full functional annotation into a single tool that processes a 50 kb phage genome in under five minutes. Beyond Pharokka, alternative annotation pipelines include MultiPhATE2 [48], with the ability to run multiple gene finders in parallel for more comprehensive gene calling, and DRAM-v [49], which identifies auxiliary metabolic genes, phage-encoded genes that function within host metabolic pathways, during infection. Tools like HHpred[50] also extend homology-based annotation through profile-profile HMM comparisons, enabling remote homology detection for

 doi:10.20944/preprints202605.1740.v1

highly diverged proteins. Additionally, the availability of the pVOGs database [51] offers a complementary orthologous groups framework, widely used for viral marker gene detection and integrated into numerous downstream annotation and identification pipelines.

Machine learning classifiers brought supervised approaches to specific annotation tasks. PhANNs [52] classifies phage proteins into ten structural classes such as major capsid, tail fiber, and baseplate using an ensemble of artificial neural networks (weighted F1 = 0.875), enabling structural annotation of proteins that lack detectable homologs. Other tools such as PhageScanner[53] adopt similar machine learning strategies for phage protein classification.

Protein language models marked the shift to the foundation model era for annotation. VPF-PLM [54] uses mean-pooled embeddings from a pre-trained protein language model to classify viral proteins into nine PHROGs functional categories via a feedforward neural network, enabling alignment-free annotation of proteins beyond the reach of profile HMMs. Applied to the EFAM ocean virome database, VPF-PLM annotated 26,770 previously unannotated protein families, a 29.4% increase over existing annotations. Two 2025 tools extend this approach: Empathi [55] reorganises PHROGs into 44 hierarchical functional categories and trains binary classifiers on protein embeddings to double annotation rates relative to homology-only methods from 16% to 33%, and GOPhage [56] adds genomic context by using a Transformer to capture protein order along phage contigs, improving annotation of proteins lacking database homology.

A notable recent addition to the field, Phold [8], translates amino acid sequences into the 3Di structural alphabet via ProstT5 [57], and subsequently searches these representations against over 1.36 million predicted phage protein structures using Foldseek [58]. This pipeline annotates more than 50% of genes on a typical phage genome, compared with approximately 35% for homology-only methods such as Pharokka (Table 1). Despite these improvements, structure-informed annotation leaves roughly half of viral genes uncharacterized, a ceiling dictated by biophysics rather than computational limits. Many phage proteins feature intrinsically disordered regions (IDR)[59] or require physical interaction with host complexes to adopt functional folds. While monomeric prediction algorithms remain blind to host-dependent conformational dynamics, generalized models like AlphaFold 3[60] shift this paradigm. By accurately modeling multimolecular complexes, they provide a theoretical path to resolve viral folds dependent on host interactions. The primary bottleneck consequently shifts from algorithmic limitation to computational scalability, requiring the field to adapt these resource-intensive models for high-throughput viromics.

Taxonomy and Classification

After gene annotation, the next step is placing phages within a coherent taxonomic framework. The 2022 ICTV reclassification abolished the morphology-based families *Myoviridae*, *Siphoviridae* and *Podoviridae*, replacing them with genomically defined taxa under the class *Caudoviricetes* [61]. Morphological classification could not scale to metagenomic phages lacking electron micrographs, and genomic analyses had revealed all three families were polyphyletic. New computational tools were needed to assign phages within this revised framework (Table 1). These tools have progressed from sequence-based distance metrics through gene-sharing networks to deep learning classifiers capable of hierarchical classification across the full ICTV hierarchy.

Early approaches in this space relied on sequence-based distance metrics, which remain valuable for targeted analyses but face scalability limits. VICTOR[62] applies Genome-BLAST Distance Phylogeny to produce phylogenomic trees with automated species and genus demarcation. ViPTree[63] generates viral proteomic trees from genome-wide tBLASTx similarities. VIRIDIC[64] implements the ICTV-recommended algorithm for nucleotide-based intergenomic similarities used in formal taxonomic proposals. These alignment-based tools remain essential for targeted phylogenetic analyses but scale poorly to tens of thousands of genomes.

Several additional homology-based tools address specific taxonomic needs: taxMyPhage provides automated genus and species-level classification of dsDNA phages aligned with current ICTV revisions [65] while GRAViTy-V2 applies composite generalised Jaccard distances built from

 doi:10.20944/preprints202605.1740.v1

generalised Jaccard similarity indices for genome relationship analysis [66]. vConTACT2 addressed scalability by pioneering the gene-sharing network paradigm, connecting phage genomes by shared protein clusters and replicating 96% of existing genus-level ICTV assignments [4]. vConTACT2 was the methodological ancestor of the ML-era taxonomy tools, but the tool itself runs on classical homology and graph algorithms, classified only at the genus level, and left highly divergent phages in unresolved outlier clusters.

Deep learning integrated with network representations enabled classification under the new ICTV hierarchy. PhaGCN[67] combined CNN-learned DNA features with protein similarity in a semi-supervised graph convolutional network. PhaGCN2 extended this to DNA and RNA viruses, achieving 89.30% recall and 83.91% precision and classifying sequences across the Gut Phage Database (142,809 sequences) and GOV2.0 datasets. PhaBOX[68] consolidates PhaGCN with phage identification (PhaMer), lifestyle prediction (PhaTYP), and host prediction (CHERRY) into a single web server, delivering family-level taxonomic classification alongside the other analyses through one interface. An updated PhaBOX2 release extends coverage beyond bacteriophages to all viruses and adopts the ICTV 2024 taxonomy.

Among currently available tools, vConTACT3 provides the broadest hierarchical taxonomy coverage, employing machine-learning-based hierarchical classification from genus to order across four out of six officially recognised viral realms [13]. Evaluated on prokaryotic viruses across four realms, vConTACT3 achieved greater than 95% agreement with official ICTV taxonomy: 97.6% at genus, 98.7% at subfamily, and 100% at both family and order levels.

Together, these tools span the range from single-genome phylogenetics to large-scale automated classification, though all face the fundamental challenge of assigning consistent taxonomy to mosaic genomes where different genomic regions yield conflicting signals. Pervasive genome mosaicism remains a fundamental challenge. Smug et al. found extensive domain-level mosaicism across 133,574 representative phage proteins, with homologous domains shared across distinct functional classes in 45 of 101 categories [69]. Conflicting taxonomic signals across mosaic genomes expose a fundamental algorithmic flaw: enforcing strict hierarchical taxonomies onto reticulate, non-vertical evolutionary histories. Hierarchical models are mathematically misaligned with the pervasive horizontal gene transfer defining phage evolution. Future taxonomic algorithms might therefore shift from forced hierarchical clustering toward directed acyclic graphs (DAGs) or network-based models that natively resolve reticulate events.

Lifestyle Prediction

Phages follow several replication strategies, though two dominate current classification: the lytic cycle, in which the phage replicates and lyses the host cell, and the lysogenic cycle, in which it integrates as a prophage. Other strategies exist: chronic infection, pseudolysogeny, and carrier states[70], but computational tools have focused mainly on the lytic-lysogenic distinction. Distinguishing lytic from temperate phages is particularly important for clinical applications. Temperate phages carry inherent therapeutic risks, mediating horizontal gene transfer, mobilising toxin genes, and conferring antibiotic resistance, so phage therapy candidates should be confirmed as obligately lytic before clinical use. Lifestyle prediction tools have progressed through two computational eras: supervised machine learning on engineered features, followed by genomic language models fine-tuned from foundation model pretraining.

Early supervised learning approaches to lifestyle prediction were constrained by small training sets and complete-genome requirements. PHACTS[71], the first dedicated tool, combined protein similarity with a Random Forest classifier using a leave-one-out approach across 227 labelled phages, achieving 99% precision on 199 confident predictions from complete genomes (overall sensitivity 88%) with accuracy declining on partial proteomes. BACPHLIP[70] improved on this by feeding HMMER3-detected lysogeny-associated domain profiles, dominated by integrases and recombinases, to a Random Forest, reaching 98.3% accuracy on 423 independent test phages versus 79% for PHACTS; however it operates only on complete genomes and cannot evaluate fragmented

 doi:10.20944/preprints202605.1740.v1

contigs (Figure 3b). DeePhage[72] adopted a CNN architecture with one-hot encoded DNA, achieving accuracies from 76.2% on 100–400 bp fragments to 88.9% on 1,200–1,800 bp fragments, the first tool to classify contigs only a few hundred base pairs long.

Foundation models shifted the field to pretrained genomic language models fine-tuned for lifestyle classification. PhaTYP[73] adapts BERT through masked language model pre-training on RefSeq phage proteins before fine-tuning on the 1,867-phage DeePhage benchmark (1,290 virulent, 577 temperate), maintaining 0.89 accuracy on 100–400 bp fragments where DeePhage scored 0.87 and BACPHLIP could not run (Figure 3b) DeepPL[74] instead fine-tunes DNABERT on 6-mer tokenised nucleotide sequences, achieving 94.65% accuracy.

ProkBERT PhaStyle[75] benchmarked three genomic language models: ProkBERT, DNABERT-2[76], and Nucleotide Transformer on fragmented phage sequences; ProkBERT-mini (~20 million parameters) achieved balanced accuracy of 0.88–0.93 and MCC of 0.75–0.86 on 500 bp fragments, outperforming the substantially larger DNABERT-2 (117M parameters) and Nucleotide Transformer (50–500M), indicating that domain-specific pretraining matters more than model scale for this task. Figure 3b captures this generational shift: BACPHLIP's 98.3% complete-genome accuracy cannot extend to short contigs, DeePhage climbs gradually from 0.76 to 0.89 with increasing length, while PhaTYP and ProkBERT PhaStyle maintain 0.88–0.94 accuracy across the full-length range, a move from tools requiring complete genomes to foundation models built for metagenomic fragments.

Training data remain scarce: approximately 1,867 labelled examples constitute the primary benchmark dataset [73], and expansion requires labour-intensive experimental confirmation. The virulent-temperate binary inherently oversimplifies phage biology. To accurately model ecologically significant strategies such as pseudolysogeny, chronic infection, or carrier states, future foundation models must evolve from outputting discrete binary labels to generating continuous probability distributions. This 'soft classification' approach would mathematically capture the dynamic spectrum of phage lifecycles.

Defense and Counter-Defense System Prediction

Prokaryotic genomes encode at least 60 families of antiviral defense systems encompassing 151 subtypes identified by 2022 [77], well beyond the restriction-modification and CRISPR-Cas pathways identified through early molecular work. The current DefenseFinder catalogue spans these families, including abortive infection systems, retrons, cyclic oligonucleotide signalling pathways, and dozens of mechanistically uncharacterised systems [77]. These defense genes cluster within genomic defense islands, hotspots where multiple systems co-localise, often flanked by mobile genetic elements enabling horizontal transfer.

Defense-system detection relies on HMM profiles combined with architectural rules. CRISPR-specific tools established the earliest computational infrastructure: CRISPRDetect[78] for array detection and boundary refinement, CRISPRCasFinder for integrated array and Cas protein identification [79] CRISPRCasTyper[80] complements this by assigning subtypes directly from repeat sequences using a machine-learning classifier, useful for orphan arrays lacking adjacent Cas genes. DefenseFinder identifies antiviral systems using curated HMM profiles combined with MacSyFinder, a rule engine enforcing synteny and gene co-occurrence constraints [77]. Applied to 21,000 prokaryotic genomes, it provided the first systematic view of the antiviral arsenal, initially detecting 60 families with more than 150 subtypes, with regular updates expanding coverage. PADLOC takes a complementary approach, employing HMM-based searches followed by system completeness validation through gene presence-absence and synteny criteria [81]. Its web server allows customisable classifications incorporating newly described systems [82]. Running both tools on the same dataset provides broader coverage than either alone. Millman et al. [83] combined computational prioritisation with experimental phage-challenge plaque assays, confirming 21 new defense systems from 45 tested candidates and illustrating the pace at which the repertoire expands.

Counter-defense prediction has progressed through the three eras. Homology-based tools target anti-CRISPR (Acr) proteins through combined signal integration. AcrFinder[84] combines homology

 doi:10.20944/preprints202605.1740.v1

search, guilt-by-association via Aca-Aca operons, and self-targeting spacer detection, while AcaFinder[85] targets conserved Aca genes via HMM profiles or guilt-by-association with Acr homologs.

The machine-learning era brought AcRanker[86], which uses XGBoost on amino acid k-mer composition to rank anti-CRISPR candidates, successfully identifying AcrIIA20 and AcrIIA21 as inhibitors of Streptococcus iniae Cas9 and SpyCas9. Foundation-model approaches now extend beyond homology using protein language model embeddings. AcrNET[87] feeds ESM-1b embeddings alongside RaptorX structural and POSSUM evolutionary features into a CNN-based classifier to predict anti-CRISPR proteins. DefensePredictor[88] builds a gradient-boosting classifier on ESM2 embeddings to flag candidate defensive proteins including those with no detectable homology to known immune genes. It was applied across 69 E. coli strains and 45 previously unknown systems experimentally validated through phage-challenge plaque assays.

Foundation-model approaches are beginning to bypass the scaling limits of curated HMM catalogues by leveraging protein language model embeddings and defense-island context, extending discovery to systems beyond CRISPR. However, novel candidates still require phage-challenge validation, and the global defence landscape remains substantially under-sampled. Structure-informed anti-defence discovery via Foldseek could potentially be the next frontier.

Host Prediction and Phage-Bacteria Interaction

After identification, assembly, annotation, and taxonomic placement, the next question is which bacterial hosts a given phage can infect; this determination that is critical for phage ecology, therapeutic candidate selection, and interpretation of the millions of uncultured phage sequences catalogued in environmental and clinical metagenomes.

Host prediction draws on four signals from phage-host co-evolution: sequence homology, CRISPR spacer matching, codon usage bias, and metagenomic co-occurrence. Tools have progressed from single-signal methods through multi-feature machine learning to protein language model approaches.

In the homology era, early tools each targeted a single signal. HostPhinder [89] used k-mer similarity against a reference database of 2,196 phages with known hosts while WIsH [90] trained Markov models on host genomes, operating hundreds of times faster than alignment-based methods. SpacePHARER [91] enables protein-level CRISPR spacer matching with 1.4–4x greater sensitivity than BLASTN at metagenomic scale. These single-signal tools set accuracy ceilings that multi-feature integration subsequently raised.

The machine learning era enabled integration of multiple genomic features into unified prediction frameworks. RaFAH [92] applies a Random Forest trained on scores from 43,644 protein clusters, performing consistently across RefSeq, single-amplified genome, and metagenomic benchmarks. CHERRY [93] formulated host prediction as link prediction in a knowledge graph, using a graph convolutional encoder over protein and DNA features to substantially improve species-level accuracy. Several complementary tools extended coverage through alternative architectures: HostG [94] introduced graph convolutional networks for semi-supervised prediction; vHULK [95] feeds alignment scores against curated viral protein families into a neural network; PHIST [96] improved species-level accuracy by 3% over alignment based tools while running two orders of magnitude faster, suitable for routine use on a standard laptop. PHISDetector [97] unified detection of CRISPR spacers, prophage regions, and sequence similarity into a single ML-scored prediction pipeline.

iPHoP [98] addressed the fragmentation of these approaches by integrating BLAST homology, CRISPR spacer matching, three k-mer/composition tools (WIsH, VirHostMatcher, PHP), and RaFAH phage-protein content through a random-forest ensemble, providing 1.5–13× more host predictions. This gain confirms that combining orthogonal signal types yields greater coverage than refining any single approach. iPHoP's full reference database, however, requires more than 300 GB of storage, contrasting sharply with PHIST's laptop-scale runtime, a practical divide between thorough-but-heavy and fast-but-narrow tools that recurs across phage bioinformatics.

 doi:10.20944/preprints202605.1740.v1

Foundation model approaches now extend this frontier by replacing or supplementing engineered features with protein language model embeddings. PhaBOX [68] consolidates CHERRY, PhaTYP, PhaMer, and PhaGCN into a unified web server-based platform; PHPGAT [99] applies GATv2 graph attention networks over a multimodal heterogeneous phage-host knowledge graph. PHIStruct [100] applies SaProt protein structure embeddings (650M parameters) to achieve F1 gain over sequence-only machine learning methods, and 5–6% over BLASTp, specifically on phages where the training-test sequence similarity falls below 40%; and MoEPH [101] combines ProkBERT and ProT5 embeddings with statistical sequence descriptors through a gated Mixture-of-Experts fusion mechanism.

Genome-level prediction identifies likely host taxa but not the molecular determinants of specificity: the receptor-binding proteins (RBPs) on phage tail fibres. RBPseg [102] combines monomeric ESMFold predictions with structural domain identification and AF2M refinement to produce the first large-scale phage tail fibre structure atlas of 67 fibres grouped into 16 structural classes, validated by cryo-EM of five fibres from three phages.

PhageHostLearn [103] pushes toward strain-level resolution by predicting phage-host interactions directly from ESM-2 protein language model embeddings of RBP and bacterial receptor sequences, achieving a cross-validated ROC AUC of 81.8% in silico (79.3% in laboratory validation). Applied to Klebsiella phage-host specificity, it shows strain-level prediction is tractable, though extension across genera requires further work.

Most tools still operate at genus or species level. Strain-level prediction remains limited to single bacterial genera, all supervised methods depend on reference database completeness, and ensemble approaches carry computational costs that limit adoption in resource-constrained settings. While protein language models demonstrate sequence-level tractability for predicting interactions between receptor-binding proteins and host receptors, they abstract away the biophysics of viral infection. Strain-level specificity is a 3D thermodynamic process governed by dynamic binding affinities, not just sequence matching. To achieve clinical utility, next-generation predictors must integrate language-model embeddings with physics-based molecular docking and dynamics (MD) simulations to resolve precise receptor-ligand kinetics, as well as multi-omic models that compute the host's intracellular defense landscape.

Proposed Workflows for Phage Characterisation

The preceding sections reviewed more than 80 tools spanning eight categories. This breadth creates a practical selection problem: which tools should a researcher run, in what order, and with what alternatives? We propose two reference workflows calibrated to computational expertise using either web-based interfaces or command line interfaces (Figure 2). Additional tools are also included to augment and increase the sensitivity and output.

Browser-accessible platforms enable end-to-end phage analysis without local installation, lowering barriers for non-computational users (Figure 2a). Although no standalone phage-specific web assembler exists, general-purpose platforms such as Galaxy (https://usegalaxy.org) and KBase (https://kbase.us) provide integrated environments for assembly and quality check using SPAdes [34] and CheckV [32]. For downstream analysis, PhaBOX [68] acts as the central hub (shaded boxes in Figure 2a), integrating phage identification (PhaMer), taxonomy (PhaGCN), lifestyle prediction (PhaTYP), and host prediction (CHERRY) within a unified web interface. Specialised tools on other web browsers complement this workflow by providing additional analysis that may be required for more specific use cases. Prophage regions can be identified using PHASTEST [28] (phastest.ca). Functional annotation is supported by the PHROGs [46] database (using HMM-based clustering) with HHpred [50] for remote homology. For taxonomy analysis beyond PhaBOX, VICTOR [62] and ViPTree [63] provide phylogenomic and proteomic tree reconstruction via web servers. Defense system detection is supported by DefenseFinder [77] and PADLOC [81,82], which provide complementary HMM-based platforms for the detection of anti-defense systems, while CRISPRCasFinder [79] and CRISPRCasTyper [80] enable CRISPR array detection and subtype

 doi:10.20944/preprints202605.1740.v1

classification. Host prediction can be further refined using HostPhinder [89] and PHISDetector [97], both useful as independent checks on PhaBOX CHERRY's graph-based predictions . Overall, web-based workflows provide broad analytical coverage with minimal setup, but remain constrained in scalability and flexibility compared to command-line approaches.

Command-line implementations enable scalable, high-sensitivity analysis and integration of best-in-class tools across workflow stages (Figure 2b). Assembly is performed using metaviralSPAdes [19] for short reads or viralFlye [35] for long reads, while Phables [33] improve recovery of complete genomes from fragmented metagenome assemblies. For identification, geNomad [7] is recommended as a primary tool due to its strong performance, but combining it with VirSorter2 [21] or VIBRANT [20] improves recall, reflecting the low concordance observed across identification tools [14]. CheckV [32] provides standardized genome quality assessment. Functional annotation follows a two-stage process: Pharokka [47] enables rapid gene calling and homology-based annotation, followed by Phold [8], which incorporates structure information to improve functional assignment. DRAM-v [49] can be incorporated to identify auxiliary metabolic gene detection where required. Taxonomic classification is performed using vConTACT3 [13] for hierarchical clustering yielding greater than 95% ICTV agreement. VIRIDIC [64] can be applied for intergenomic similarity calculations in formal taxonomic assignments. Lifestyle prediction can be conducted using PhaTYP [73] or ProkBERT PhaStyle [75], both of which perform well on short contigs, while BACPHLIP [70] provides a simpler alternative for complete genomes. Defense system detection benefits from combining AntiDefenseFinder [77] and PADLOC [81,82] with CRISPRCasTyper [80] providing detailed CRISPR subtype annotation. Host prediction is performed using iPHoP [98], which integrates multiple signal types to improve coverage. Given the consistently low agreement between tools across categories [14,31], multi-tool strategies remain essential for robust phage genome analysis and tools should be selected based on input data type and required output regardless of user expertise. Furthermore, command-line implementations are stratified not merely by analytical intent, but by computational requirements, that could range from laptop-scale to High-Performance Computing (HPC) infrastructure.

Ultimately, the transition from fragmented, single-task execution to automated, end-to-end orchestration represents the most critical bottleneck in operationalizing these workflows. While the command-line ecosystem offers unparalleled analytical depth, it introduces substantial friction regarding data interoperability; the heterogeneous output formats of identification algorithms (e.g., geNomad) must be programmatically parsed and standardized to serve as inputs for downstream structural annotation or host prediction tools. To mitigate this fragility and ensure strict computational reproducibility, researchers could embed these individual tools within robust, containerized workflow management systems (such as Nextflow[104] and Snakemate[105]). Looking forward, the integration of agentic AI frameworks capable of autonomous workflow orchestration[106], that translate biological queries into executable multi-tool pipelines, holds the potential to seamlessly bridge the usability gap, bringing the accessibility of web platforms to the rigorous environment of high-performance computing.

Benchmarking, Current Limitations, and Future Directions

Across tool categories, successive computational paradigms have shifted performance baselines. In phage identification, geNomad increased reported Matthews correlation coefficient from 81.3% for VirSorter2 to 95.3% [7]. In functional annotation, Phold's structure-informed approach assigns functions to more than 50% of genes in a typical phage genome, compared with approximately 35% for homology-only workflows [8]. Lifestyle prediction has moved from complete-genome classifiers such as PHACTS to fragment-compatible foundation-model approaches, with ProkBERT PhaStyle reporting balanced accuracies of 0.88-0.93 on 500 bp fragments [75]. Host prediction has similarly benefited from ensemble learning, with iPHoP increasing host predictions rates by up to 13-fold at controlled false discovery rate [98], while vConTACT3 achieves greater than 95% agreement with ICTV taxonomy across multiple ranks [13].

 doi:10.20944/preprints202605.1740.v1

Despite these gains, performance remains strongly dependent on input length, dataset composition and evaluation design. Short contigs, particularly those below 3 kb, remain problematic across categories. DeepVirFinder's AUROC decreases from approximately 0.98 at 3 kb to 0.93 at 300 bp [16], and DeePhage accuracy drops from 88.9% on 1,200-1,800 bp fragments to 76.2% on 100-400 bp fragments (Figure 3). Host prediction benchmarks rarely report length-stratified performance, although tools such as WIsH show reduced genus-level accuracy as candidate-host space expands. As short contigs dominate many metagenomic assemblies, this length-dependence directly limits sensitivity and specificity.

Independent benchmarks further show that tool performance is highly dataset specific. Ho et al. found that only 89.7% of true phage contigs were detected by more than half of the tools tested, but only 11.6% were detected by all ten tools tested [14]. Wu et al. reported true positive rates ranging from 0% to 97% across tools and datasets from different biomes [31]. These results indicate that individual tools capture partially overlapping subsets of viral diversity rather than a common, stable signal. Multi-tool workflows are therefore currently more robust than any single classifier, though they increase computational burden and complicate interpretation when predictions disagree [107].

Three recurring data limitations explain much of this variability. First, supervised methods remain constrained by small datasets. For instance, lifestyle prediction still relies heavily on benchmark sets containing fewer than 2,000 experimentally labelled phages [73]. Second, reference databases are taxonomically and ecologically biased, with cultured phages dominated by a small number of bacterial host genera [2]. Thirdly, random-test splits can inflate performance when closely related genomes appear in both training and test sets. Further benchmarks should therefore use phylogenetically separated test sets, report performance by contig length and novelty of data inputs, and evaluate tools on independent environmental and clinical datasets [107].

Consequently, exceptionally high-performance metrics for recent foundation models must be interpreted cautiously. With 75% of sequenced phages originating from just 30 bacterial genera, these metrics likely reflect algorithmic overfitting to a hyper-characterized sequence space rather than true generalizability. Until evaluated against phylogenetically isolated holdout sets, these high accuracy ceilings merely represent the limits of deeply biased reference databases.

Viral dark matter remains a major unresolved challenge. Homology-based annotation leaves a large fraction of phage proteins uncharacterized [40,108]. Although structure-informed approaches such as Phold have improved annotation rates, roughly half of genes still lack confident functional assignments [8]. Furthermore, binary classification schemes (such as virulent versus temperate, or host versus non-host) oversimplify biological continua, omitting information such as pseudolysogeny, chronic infection, broad-host-range infection, and phage-plasmid lifestyles. As deep learning and foundation-model tools become more accurate, their reduced interpretability will need to be balanced against biologically grounded features such as domains, genomic neighbourhoods, structural similarity, and experimentally validated receptors[108].

Foundation models partially address the labelled-data bottleneck by learning representations from large unlabeled sequence and protein corpora [75]. Protein language models such as ESM-2 now support host prediction, protein family classification, and defense-system discovery [109,110], while structure-aware pipelines combining predicted protein structures, structural alphabets [103], and Foldseek enable functional inference beyond detectable sequence homology [8,47]. These dependencies create powerful shared infrastructure, but also ecosystem-level fragility: changes to core components such as gene callers, HMM databases, protein language models, or structural search tools can propagate through many downstream pipelines (Table 2). Future tool development should therefore prioritise versioned databases, containerised workflows, transparent model cards, and reproducible benchmark datasets.

Table 2. Key dependencies of machine learning and deep learning tools.

 doi:10.20944/preprints202605.1740.v1

Category	Tool	Year	Architecture	Gene calling (Prodigal, pyrodigal, PHANOTATE, GeneMark, six-frame translation)	Homology search (BLAST, DIAMOND, HMMER, HMMscan, hmmsearch, MMseqs2)	Genome LM (DNABERT, DNABERT-2, ProkBERT, Nucleotide Transformer)	Protein LM (ESM-1b, ESM-2, ProtBERT, ProT5, ProtTrans, SaProt)	Structure tools (AlphaFold, ESMFold, Foldseek, ProstT5)
Identification	VirFinder	2017	Logistic regression + k-mer	○	○	○	○	○
	MARVEL	2018	Random Forest	●	●	○	○	○
	PPR-Meta	2019	Bi-path CNN	○	○	○	○	○
	DeepVirFinder	2020	CNN on one-hot DNA	○	○	○	○	○
	Seeker	2020	LSTM on raw DNA	○	○	○	○	○
	VIBRANT	2020	NN + v-score metric	●	●	○	○	○
	VirSorter2	2021	Multi-classifier ML ensemble	●	●	○	○	○
	PhaMer	2022	Custom Transformer on protein tokens	●	●	○	○	○
	INHERIT	2022	DNABERT	○	○	●	○	○
	geNomad	2023	IGLOO encoder + CRF	●	●	○	○	○
Annotation	PhANNs	2020	ANN on amino-acid features	○	○	○	○	○
	VPF-PLM	2024	ESM-2 classifier	●	●	○	●	○
	Empathi	2025	Hierarchical PLM embeddings	●	●	○	●	○
	GOPhage	2025	ESM-2 + genomic-context Transformer	○	○	○	●	○
	Phold	2026	ProstT5 -> Foldseek structure-informed	●	●	○	○	●
Taxonomy	vConTACT2	2019	Gene-sharing network	●	●	○	○	○
	PhaGCN	2021	GCN + CNN	○	●	○	○	○
	PhaGCN2	2023	Extended GCN (DNA + RNA)	●	●	○	○	○
	PhaBOX	2023	Integrated DL platform	●	●	○	○	○
	vConTACT3	2025	Hierarchical ML classifier	●	●	○	○	○
Lifestyle	PHACTS	2012	Random Forest + FASTA similarity	○	●	○	○	○
	BACPHLIP	2021	RF on HMMER3 protein domains	○	●	○	○	○
	DeePhage	2021	CNN on one-hot DNA	○	○	○	○	○
	PhaTYP	2023	Custom BERT on protein tokens	●	●	○	○	○
	DeepPL	2024	DNABERT fine-tune	○	○	●	○	○
	ProkBERT PhaStyle	2025	Genomic language model	○	○	●	○	○
Defense	AcRanker	2020	XGBoost ranking	○	○	○	○	○
	AcrNET	2023	Deep NN + ESM-1b	○	○	○	●	○
	DefensePredictor	2025	PLM embeddings	○	●	○	●	○
Host prediction	HostG	2021	GCN	●	●	○	○	○
	RaFAH	2021	RF on 43,644 protein clusters	●	●	○	○	○
	CHERRY	2022	Graph encoder-decoder	●	●	○	○	○
	vHULK	2022	NN on viral protein families	○	●	○	○	○
	iPHoP	2023	Ensemble ML (5 methods)	●	●	○	○	○
	PhageHostLearn	2024	ESM-2 + XGBoost	●	●	○	●	○
	MoEPH	2025	Gated MoE of ProtBERT + ProT5	○	○	○	●	○
	PHIStruct	2025	SaProt (structure-aware PLM)	○	○	○	●	○
	PHPGAT	2025	Graph attention on multimodal KG	●	●	○	○	○
	RBPseg	2025	ESMFold + structural domain ID	○	○	○	○	●

The field urgently needs a community benchmarking framework analogous to Critical Assessment of Structure Prediction (CASP) or Critical Assessment of Metagenome Interpretation (CAMI) [111,112]. Such a benchmark should include phylogenetically separated train-test splits, standardised metrics, common reference datasets, contig-length strata, biome-specific test sets, negative controls, and clinically relevant tasks such as strain-level host prediction. Reporting should include not only accuracy, AUROC, F1 score, and Matthews correlation coefficient, but also calibration, false discovery rate, runtime, memory usage, and failure modes [107,113]. Without this standardisation, apparent performance gains will remain difficult to compare across studies.

Clinical translation imposes the most stringent requirements. Phage therapy would benefit greatly from strain-level and ideally receptor-level host prediction, yet most current tools operate at genus or species level. Recent studies in *Escherichia* and *Klebsiella* show that strain-level prediction is feasible in restricted settings, but no broadly generalisable predictor exists across major clinical pathogens[103]. Closing this gap will require larger experimentally validated phage-host interaction datasets, receptor-binding protein annotation, bacterial receptor modelling, and transfer-learning strategies that generalise across genera.

Overall, the next phase of phage bioinformatics should depend less on producing additional standalone tools and more on rigorous evaluation, interoperable infrastructure, and clinically meaningful prediction targets [31]. Structure-informed annotation, foundation-model representations, and generative genome models are already expanding what can be inferred from phage sequences [8,57,58]. Their impact will depend on whether the field can pair these advances with standardised benchmarks, experimentally grounded validation, and methods that resolve phage-host specificity at the strain and receptor levels.

Conclusions

Phage bioinformatics has moved through three computational paradigms in under a decade. Homology-based tools established the foundations of phage detection and annotation but remained bounded by reference database completeness. Machine learning classifiers bypassed that ceiling by learning statistical patterns from sequence composition, extending identification and classification to

 doi:10.20944/preprints202605.1740.v1

unculturable phages, yet introduced a new bottleneck in labelled training data. Foundation models pretrained on massive unlabelled corpora have now begun to lift both constraints simultaneously, enabling reference-independent classification, structure-informed functional annotation that narrows the viral dark matter gap, and the first generative models capable of producing viable synthetic phage genomes [114]. The field is therefore crossing a fundamental threshold: from computational analysis of natural phages to computational design of new ones.

Several substantive gaps persist. Roughly half of phage genes still resist functional annotation even with structural methods; no generalisable strain-level host predictor exists for phage therapy; and binary classification frameworks of phage lifecycles oversimplify biological phenomena. The most urgent community need, however, is standardized community benchmarking. Tool creation is outpacing evaluation, leaving researchers without a reliable basis for tool selection. These open problems define the next phase of phage bioinformatics and form the agenda for the tool developers this review is intended to equip.

Funding: This work was supported by Research Center for Excellence IDMxS. S.J.L.P. is funded by NTU Research Scholarship.

Data Availability Statement: No new data were generated or analysed in support of this review. All tools and databases discussed are publicly available and referenced in the text.

Conflicts of Interest: The authors declare no competing interests.

References

1. Camargo, A. P. et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.* **51**, D733–D743 (2022).
2. Cook, R. et al. INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. *PHAGE Ther. Appl. Res.* **2**, 214–223 (2021).
3. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
4. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
5. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, (2017).
6. Shang, J. PhaGCN: Phage taxonomic classification with graph convolutional networks. *Bioinformatics* https://github.com/KennthShang/PhaGCN (2021).
7. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2023).
8. Bouras, G. et al. Protein structure-informed bacteriophage genome annotation with Phold. *Nucleic Acids Res.* **54**, (2026).
9. Wendling, C. C., Vasse, M. & Wielgoss, S. Phage quest: a beginner's guide to explore viral diversity in the prokaryotic world. *Brief. Bioinform.* **26**, bbaf449 (2025).
10. Brister, J. R., Ako-adjei, D., Bao, Y. & Blinkova, O. NCBI Viral Genomes Resource. *Nucleic Acids Res.* **43**, D571–D577 (2014).
11. Russell, D. A. & Hatfull, G. F. PhagesDB: the actinobacteriophage database. *Bioinformatics* **33**, 784–786 (2016).
12. ICTV. ICTV Virus Metadata Resource (VMR). https://doi.org/10.5281/zenodo.18808244 (2022).
13. Bolduc, B. et al. Machine learning enables scalable and systematic hierarchical virus taxonomy. *Nat. Biotechnol.* 1–10 (2025) doi:10.1038/s41587-025-02946-9.
14. Ho, S. F. S., Wheeler, N. E., Millard, A. D. & van Schaik, W. Gauge your phage: benchmarking of bacteriophage identification tools in metagenomic sequencing data. *Microbiome* **11**, (2023).
15. Amgarten, D., Braga, L. P. P., da Silva, A. M. & Setubal, J. C. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Front. Genet.* **9**, (2018).

 doi:10.20944/preprints202605.1740.v1

16. Ren, J. et al. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* **8**, 64–77 (2020).
17. Auslander, N., Gussow, A. B., Benler, S., Wolf, Y. I. & Koonin, E. V. Seeker: alignment-free identification of bacteriophage genomes by deep learning. *Nucleic Acids Res.* **48**, e121–e121 (2020).
18. Fang, Z. et al. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience* **8**, (2019).
19. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics* **36**, 4126–4129 (2020).
20. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, (2020).
21. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, (2021).
22. Jurtz, V. I., Villarroel, J., Lund, O., Larsen, M. V. & Nielsen, M. MetaPhinder—Identifying Bacteriophage Sequences in Metagenomic Data Sets. *PLOS ONE* **11**, e0163111 (2016).
23. Shang, J., Tang, X., Guo, R. & Sun, Y. Accurate identification of bacteriophages from metagenomic data using Transformer. *Brief. Bioinform.* **23**, (2022).
24. Bai, Z. et al. Identification of bacteriophage genome sequences with representation learning. *Bioinformatics* **38**, 4264–4270 (2022).
25. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).
26. Akhter, S., Aziz, R. K. & Edwards, R. A. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* **40**, e126–e126 (2012).
27. Arndt, D. et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21 (2016).
28. Wishart, D. S. et al. PHASTEST: faster than PHASTER, better than PHAST. *Nucleic Acids Res.* **51**, W443–W450 (2023).
29. Gauthier, C. H. et al. DEPhT: a novel approach for efficient prophage discovery and precise extraction. *Nucleic Acids Res.* **50**, e75–e75 (2022).
30. Sirén, K. et al. Rapid discovery of novel prophages using biological feature engineering and machine learning. *NAR Genomics Bioinforma.* **3**, lqaa109 (2021).
31. Wu, S., Fang, Z., Tan, J., Li, M. & Wang, C. Benchmarking computational tools for virus identification in metagenomes across biomes. *Microbiome* **12**, 215 (2024).
32. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2020).
33. Mallawaarachchi, V. et al. Phables: from fragmented assemblies to high-quality bacteriophage genomes. *Bioinformatics* **39**, (2023).
34. Bankevich, A. et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
35. Antipov, D., Rayko, M., Kolmogorov, M. & Pevzner, P. A. viralFlye: assembling viruses and identifying their hosts from long-read metagenomics data. *Genome Biol.* **23**, (2022).
36. Chen, L. & Banfield, J. F. COBRA improves the completeness and contiguity of viral genomes assembled from metagenomes. *Nat. Microbiol.* **9**, 737–750 (2024).
37. Kieft, K., Adams, A., Salamzade, R., Kalan, L. & Anantharaman, K. vRhyme enables binning of viral genomes from metagenomes. *Nucleic Acids Res.* **50**, e83 (2022).
38. Zolfo, M. et al. Detecting contamination in viromes using ViromeQC. *Nat. Biotechnol.* **37**, 1408–1412 (2019).
39. Gilchrist, C. L. M. & Chooi, Y.-H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **37**, 2473–2475 (2021).
40. Hatfull, G. F. & Hendrix, R. W. Bacteriophages and their genomes. *Curr. Opin. Virol.* **1**, 298–303 (2011).
41. Hatfull, G. F. Dark Matter of the Biosphere: the Amazing World of Bacteriophage Diversity. *J. Virol.* **89**, 8107–8110 (2015).

 doi:10.20944/preprints202605.1740.v1

42. McNair, K., Zhou, C., Dinsdale, E. A., Souza, B. & Edwards, R. A. PHANOTATE: a novel approach to gene identification in phage genomes. *Bioinformatics* **35**, 4537–4542 (2019).
43. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, (2010).
44. Besemer, J., Lomsadze, A. & Borodovsky, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29**, 2607–2618 (2001).
45. Kelley, D. R., Liu, B., Delcher, A. L., Pop, M. & Salzberg, S. L. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res.* **40**, e9 (2012).
46. Terzian, P. et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genomics Bioinforma.* **3**, (2021).
47. Bouras, G. et al. Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics* **39**, (2022).
48. Ecale Zhou, C. L. et al. MultiPhATE2: code for functional annotation and comparison of phage genomes. *G3 GenesGenomesGenetics* **11**, (2021).
49. Shaffer, M. et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).
50. Zimmermann, L. et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
51. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* **45**, D491–D498 (2016).
52. Cantu, V. A. et al. PhANNs, a fast and accurate tool and web server to classify phage structural proteins. *PLOS Comput. Biol.* **16**, e1007845 (2020).
53. Frontiers | PhageScanner: a reconfigurable machine learning framework for bacteriophage genomic and metagenomic feature annotation. https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2024.1446097/full.
54. Flamholz, Z. N., Biller, S. J. & Kelly, L. Large language models improve annotation of prokaryotic viral proteins. *Nat. Microbiol.* **9**, 537–549 (2024).
55. Boulay, A., Leprince, A., Enault, F., Rousseau, E. & Galiez, C. Empathi: embedding-based phage protein annotation tool by hierarchical assignment. *Nat. Commun.* **16**, 9114 (2025).
56. Guan, J. et al. GOPhage: protein function annotation for bacteriophages by integrating the genomic context. *Brief. Bioinform.* **26**, bbaf014 (2025).
57. Heinzinger, M. et al. Bilingual language model for protein sequence and structure. *NAR Genomics Bioinforma.* **6**, (2024).
58. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2023).
59. Mishra, P. M., Verma, N. C., Rao, C., Uversky, V. N. & Nandi, C. K. Intrinsically disordered proteins of viruses: Involvement in the mechanism of cell regulation and pathogenesis. *Prog. Mol. Biol. Transl. Sci.* **174**, 1–78 (2020).
60. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
61. Turner, D. et al. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch. Virol.* **168**, (2023).
62. Meier-Kolthoff, J. P. & Göker, M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics* **33**, 3396–3404 (2017).
63. Nishimura, Y. et al. ViPTree: the viral proteomic tree server. *Bioinformatics* **33**, 2379–2380 (2017).
64. Moraru, C., Varsani, A. & Kropinski, A. M. VIRIDIC—A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses. *Viruses* **12**, 1268 (2020).
65. Millard, A. et al. taxMyPhage: Automated Taxonomy of dsDNA Phage Genomes at the Genus and Species Level. *PHAGE Ther. Appl. Res.* **6**, 5–11 (2025).
66. Mayne, R., Aiewsakun, P., Turner, D., Adriaenssens, E. M. & Simmonds, P. GRAViTy-V2: a grounded viral taxonomy application. *NAR Genomics Bioinforma.* **6**, lqae183 (2024).

 doi:10.20944/preprints202605.1740.v1

67. Shang, J., Jiang, J. & Sun, Y. Bacteriophage classification for assembled contigs using graph convolutional network. *Bioinformatics* **37**, i25–i33 (2021).
68. Shang, J., Peng, C., Liao, H., Tang, X. & Sun, Y. PhaBOX: a web server for identifying and characterizing phage contigs in metagenomic data. *Bioinforma. Adv.* **3**, (2023).
69. Smug, B. J., Szczepaniak, K., Rocha, E. P. C., Dunin-Horkawicz, S. & Mostowy, R. J. Ongoing shuffling of protein fragments diversifies core viral functions linked to interactions with bacterial hosts. *Nat. Commun.* **14**, (2023).
70. Hockenberry, A. J. & Wilke, C. O. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. *PeerJ* **9**, e11396 (2021).
71. McNair, K., Bailey, B. A. & Edwards, R. A. PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics* **28**, 614–618 (2012).
72. Wu, S. et al. DeePhage: distinguishing virulent and temperate phage-derived sequences in metavirome data with a deep learning approach. *GigaScience* **10**, (2021).
73. Shang, J., Tang, X. & Sun, Y. PhaTYP: predicting the lifestyle for bacteriophages using BERT. *Brief. Bioinform.* **24**, (2022).
74. Zhang, Y., Mao, M., Zhang, R., Liao, Y.-T. & Wu, V. C. H. DeepPL: A deep-learning-based tool for the prediction of bacteriophage lifecycle. *PLOS Comput. Biol.* **20**, e1012525 (2024).
75. Juhász, J. et al. ProkBERT PhaStyle: accurate phage lifestyle prediction with pretrained genomic language models. *Bioinforma. Adv.* **5**, (2024).
76. Zhou, Z. et al. DNABERT-2: Efficient Foundation Model and Benchmark for Multi-Species Genome. in (2024). doi:10.48550/arXiv.2306.15006.
77. Tesson, F. et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.* **13**, (2022).
78. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, (2016).
79. Couvin, D. et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* **46**, W246–W251 (2018).
80. Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A. & Sørensen, S. J. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *CRISPR J.* **3**, 462–469 (2020).
81. Payne, L. J. et al. Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Res.* **49**, 10868–10878 (2021).
82. Payne, L. J. et al. PADLOC: a web server for the identification of antiviral defence systems in microbial genomes. *Nucleic Acids Res.* **50**, W541–W550 (2022).
83. Millman, A. et al. An expanded arsenal of immune systems that protect bacteria from phages. *Cell Host Microbe* **30**, 1556-1569.e5 (2022).
84. Yi, H. et al. AcrFinder: genome mining anti-CRISPR operons in prokaryotes and their viruses. *Nucleic Acids Res.* **48**, W358–W365 (2020).
85. Wen, Y., Zhang, F. & Jiang, Y. AcaFinder: genome mining anti-CRISPR-associated genes. *mSystems* **8**, e00981-22 (2023).
86. Eitzinger, S. et al. Machine learning predicts new anti-CRISPR proteins. *Nucleic Acids Res.* **48**, 4698–4708 (2020).
87. Li, Y. et al. AcrNET: predicting anti-CRISPR with deep learning. *Bioinformatics* **39**, (2023).
88. DeWeirdt, P. C., Mahoney, E. M. & Laub, M. T. DefensePredictor: A Machine Learning Model to Discover Novel Prokaryotic Immune Systems. https://doi.org/10.1101/2025.01.08.631726 (2025) doi:10.1101/2025.01.08.631726.
89. Villarroel, J. et al. HostPhinder: A Phage Host Prediction Tool. *Viruses* **8**, 116 (2016).
90. Galiez, C., Siebert, M., Enault, F., Vincent, J. & Söding, J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**, 3113–3114 (2017).
91. Zhang, R. et al. SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts. *Bioinformatics* **37**, 3364–3366 (2021).

 doi:10.20944/preprints202605.1740.v1

92. Coutinho, F. H. et al. RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns* **2**, 100274 (2021).
93. Shang, J. & Sun, Y. CHERRY: a Computational metHod for accuratE pRediction of virus–pRokarYotic interactions using a graph encoder–decoder model. *Brief. Bioinform.* **23**, (2022).
94. Shang, J. & Sun, Y. Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning. *BMC Biol.* **19**, 250 (2021).
95. Amgarten, D., Iha, B. K. V., Piroupo, C. M., da Silva, A. M. & Setubal, J. C. vHULK, a New Tool for Bacteriophage Host Prediction Based on Annotated Genomic Features and Neural Networks. *PHAGE* **3**, 204–212 (2022).
96. Zielezinski, A., Deorowicz, S. & Gudyś, A. PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. *Bioinformatics* **38**, 1447–1449 (2021).
97. Zhou, F. et al. PHISDetector: A Tool to Detect Diverse In Silico Phage–Host Interaction Signals for Virome Studies. *Genomics Proteomics Bioinformatics* **20**, 508–523 (2022).
98. Roux, S. et al. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLOS Biol.* **21**, e3002083 (2023).
99. Liu, F., Zhao, Z. & Liu, Y. PHPGAT: predicting phage hosts based on multimodal heterogeneous knowledge graph with graph attention network. *Brief. Bioinform.* **26**, (2024).
100. Gonzales, M. E. M., Ureta, J. C. & Shrestha, A. M. S. PHIStruct: improving phage–host interaction prediction at low sequence similarity settings using structure-aware protein embeddings. *Bioinformatics* **41**, btaf016 (2025).
101. Chen, Q. et al. MoEPH: an adaptive fusion-based LLM for predicting phage-host interactions in health informatics. *Front. Microbiol.* **16**, (2025).
102. Klein-Sousa, V., Roa-Eguiara, A., Kielkopf, C. S., Sofos, N. & Taylor, N. M. I. RBPseg: Toward a complete phage tail fiber structure atlas. *Sci. Adv.* **11**, (2025).
103. Boeckaerts, D. et al. Prediction of Klebsiella phage-host specificity at the strain level. *Nat. Commun.* **15**, (2024).
104. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
105. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
106. Dip, S. A. et al. Large language model agents for biological intelligence across genomics, proteomics, spatial biology, and biomedicine. *Brief. Bioinform.* **27**, bbag110 (2026).
107. Shang, J. et al. From genomic signals to prediction tools: a critical feature analysis and rigorous benchmark for phage–host prediction. *Brief. Bioinform.* **26**, (2025).
108. Grigson, S. R., Bouras, G., Dutilh, B. E., Olson, R. D. & Edwards, R. A. Computational function prediction of bacteria and phage proteins. *Microbiol. Mol. Biol. Rev.* **89**, (2025).
109. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
110. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
111. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins Struct. Funct. Bioinforma.* **89**, 1607–1617 (2021).
112. Meyer, F. et al. Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
113. Gaborieau, B. et al. Prediction of strain level phage–host interactions across the Escherichia genus using only genomic information. *Nat. Microbiol.* **9**, 2847–2861 (2024).
114. King, S. H. et al. Generative design of novel bacteriophages with genome language models. https://doi.org/10.1101/2025.09.12.675911 (2025) doi:10.1101/2025.09.12.675911.

 doi:10.20944/preprints202605.1740.v1

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.