viruses

MDPI

Systematic Review

Reading the Host from the Virus: A Systematic Review of Sequence-Based Prediction of Host-Related Properties for Eukaryote-Infecting Viruses

Julia Galeeva [1,*], Polina Kuzmichenko [1], Nikolay Popov [1,2], Alexander Lukashev [1,2] and Elena Ilina [1]

[1] Research Institute for Systems Biology and Medicine (RISBM), Department of Mathematical Biology and Bioinformatics, 117246 Moscow, Russia
[2] Martsinovsky Institute of Medical Parasitology, Tropical and Vector Borne Diseases, Sechenov University, 119435 Moscow, Russia
* Correspondence: olgagaleeva546@gmail.com

Abstract

Viral metagenomics and metatranscriptomics have expanded our understanding of viral diversity and associations between viruses and hosts. For eukaryotic viruses, identifying host-related properties remains challenging because infection depends on receptor compatibility, tissue specificity and mechanisms of immune evasion. Interspecies transmission and zoonotic spillover pose major risks to global health, highlighting the need for computational methods that can prioritise plausible host associations and related risks for known and novel viruses. Following PRISMA 2020 guidelines, we systematically evaluated 18 software tools for predicting host-related properties of eukaryote-infecting viruses. The tools were classified into six prediction-target categories, ranging from host-taxon assignment to spillover-risk ranking. We also examined their input requirements, feature representations, modelling strategies and validation approaches. Citation analysis identified 735 tool-citation links, but only 28 documented applications in 27 publications across six tools. Functional testing showed that 16 of the 18 tools returned a host-related prediction for a user-provided sequence. Published performance could not be pooled or directly ranked because tools within the same prediction-target category differed in prediction unit, label space, construction of the held-out evaluation set and reported metric. Tool choice therefore follows from the biological question and the available input, and predicted associations remain hypotheses requiring independent biological validation.

Keywords: eukaryote-infecting viruses; viral host prediction; host range; metagenomics; computational tools; machine learning; predictive virology

Check for updates

Academic Editor: Jinmin Ma

Received: 1 October 2026
Revised: 10 November 2026
Accepted: 11 November 2026
Published: 18 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Much of the virosphere remains unexplored. Estimates of global viral diversity range from approximately 10^7 to 10^9 species, of which only about 17,000 have been formally recognized. Viral diversity may therefore be one or two orders of magnitude greater than that of cellular organisms, making the virosphere a major source of previously unknown genetic material [1]. For a long time, our knowledge of eukaryote-infecting viruses came mainly from studies of clinically or economically important viruses of humans, domestic animals, and crops. Metagenomic and metatranscriptomic studies have since revealed previously unrecognized viruses in vertebrates, invertebrates, plants, fungi, and other eukaryotic groups [2–4].

Identifying the actual hosts remains a major challenge, as the mere recovery of a viral sequence does not guarantee an active infection. The genetic material might instead originate from environmental exposure or ingested food. It could also pass transiently through the organism or be introduced via contamination [5]. Proving a biological association does not mean the organism supports productive infection or tissue-specific replication. For this reason, we separate basic host association from host range, tissue tropism, and reservoir competence. Actual infection requires the virus to breach susceptible cells, multiply, and survive the immune response. Transmission is a distinct step dependent on viral shedding and opportunities for further spread, and long-term maintenance in a host population relies on even broader ecological dynamics [6,7]. Cross-species transmission has occurred widely over evolutionary timescales, although the opportunities for a particular virus to change hosts remain constrained by viral and host phylogeny, ecology, and molecular compatibility [8,9].

For eukaryote-infecting viruses, the involvement of vectors and alternate hosts means that identifying a vector does not confirm the reservoir. Computational host predictors reflect this complexity. Some assign a virus to a broad taxonomic group. Others attempt to identify a specific host species, reservoir, or vector. Assessing human infection or cross-species transmission requires evidence beyond a predicted host association. These platforms often analyze similar sequence data but generate fundamentally different biological predictions.

The computational approaches underlying these predictions are heterogeneous. Some tools rely on sequence similarity or phylogenetic relationships, whereas others use engineered genomic features, learned nucleotide representations or protein-derived signals. Several approaches also combine viral molecular data with host or ecological information. Their prediction targets are equally diverse, ranging from host taxon and reservoir or vector identity to human infectivity, host adaptation and spillover risk. Despite these differences, the tools share the assumption that host-related properties can be inferred, at least in part, from molecular characteristics of the virus. Such outputs should be interpreted as model-based predictions rather than direct evidence of biological compatibility. A predicted host association does not by itself establish tissue tropism, reservoir competence, productive infection or onward transmission.

We conducted a PRISMA-compliant systematic review of 18 computational tools. By connecting theoretical modeling features with empirical data scope, this review establishes a practical guide for matching computational tools to appropriate biological questions, available sequence data, and their predictive limitations.

2. Biological Context for Virus–Host Prediction

2.1. Host Range and Tropism as Multilayered Biological Properties

Viruses are often studied as pathogens of a particular host, although this describes only part of their relationship with cellular organisms. The same organism can harbour a diverse viral community, including viruses involved in persistent, commensal or occasionally beneficial associations. Considering viruses only as pathogens of individual species therefore leaves much of their biology unexplained. Their relationships with cellular life also involve ecological and evolutionary processes, which underlie the description of the virosphere as an interconnected replicator space [1].

Host range and tropism are not simple taxonomic labels. They represent composite traits governed by ecological exposure, host phylogeny, receptor availability, tissue-specific permissiveness, intracellular compatibility, immune responses, and opportunities for sustained transmission. A virus may therefore be present in a host-associated sample without establishing replication, while productive infection in one tissue does not necessarily imply

systemic dissemination, onward transmission, or reservoir competence. Viral sequences recovered from wildlife samples may originate from active infection, transient passage, dietary material, environmental contamination, or organisms associated with the sampled host. For example, viral profiling of vampire bats revealed diverse viral assemblages but also highlighted the difficulty of distinguishing stable reservoir associations from transient or incidental detection [10].

An unusual illustration of tissue-restricted permissiveness is the reported association of Covert mortality nodavirus, previously linked primarily to aquatic animals, with a persistent ocular disease, hypertensive viral anterior uveitis, in humans exposed to raw or live aquatic species [11]. Although this case should not be generalized to imply frequent cross-kingdom transmission, it suggests that productive infection may sometimes depend on access to a specific permissive anatomical niche rather than on complete adaptation to the entire host organism.

Host range and tropism should therefore be interpreted across ecological, organismal, tissue, cellular, and molecular scales. These distinctions need to be retained when comparing host-prediction tools. We use them to separate the biological questions addressed by the predictions, from identifying a possible host to assessing its role in maintaining and transmitting the virus.

2.2. Molecular and Structural Barriers to Viral Entry

Viral entry relies on structural proteins engaging host receptors—mediated by fusion-protein architectures in enveloped viruses or capsid-driven genome delivery in non-enveloped pathogens like Enterovirus D68 [12,13]. Host specificity is partly dictated by the capacity of viral proteins to bind species-specific orthologues, such as spike-ACE2 interactions in coronavirus HKU5 or receptor preference shifts in influenza A haemagglutinin [14,15]. While these structural requirements provide the biological rationale for predictive computational models, entry is merely an initial prerequisite; productive infection additionally demands intracellular replication capability, immune evasion, and transmission competence.

Strains within a single viral species can exploit entirely distinct cellular entry receptors, whereas highly divergent viruses often convergently target the same host surface molecules. This biological complexity fundamentally frustrates attempts to model receptor interactions on a generalized, pan-viral scale. Specialized bioinformatic frameworks predict receptor specificity for well-characterized lineages, such as inferring HIV-1 coreceptor usage from V3 loop sequences [16,17].

2.3. Genomic Composition and Intracellular Immune Selection

After entry, viruses encounter intracellular environments that differ substantially among host lineages and cell types. Pattern-recognition receptors and restriction factors recognize features of non-self nucleic acids, including double-stranded RNA, uncapped RNA termini and unusual sequence composition. Nucleotide and dinucleotide frequencies, codon usage and codon-pair bias may therefore influence the resemblance of viral genomes to host transcripts and their recognition by antiviral pathways [18].

In mammals, CpG dinucleotides (cytosine–phosphate–guanine dinucleotide) occur relatively rarely in genomes and transcripts. Many RNA viruses that infect vertebrates share this characteristic [19,20]. In experiments, adding CpG sites to viral RNA can reduce viral replication. The ZAP (zinc-finger antiviral) protein drives part of this restriction [21]. It binds to CpG-rich RNA and acts alongside other proteins to inhibit viral protein synthesis or promote RNA degradation [18]. The strength of this restriction depends on more than just the total CpG count. The neighboring sequence, the distance between motifs, and the

RNA structure all matter, while proteins like TRIM25 and KHNYN further influence ZAP activity [22,23]. Elevated UpA (uridine–adenosine dinucleotide) levels restrict replication of the virus. ZAP, OAS3 (2′-5′-oligoadenylate synthetase 3), and RNase L contribute to the restriction of RNA viruses that have high CpG and UpA content [24].

When a virus changes hosts, it faces different antiviral defenses. Invertebrates rely mainly on RNA interference. Vertebrates, by contrast, deploy interferon-associated mechanisms like ZAP and the OAS/RNase L pathway. Reducing the dinucleotides associated with this restriction might improve replication for some viruses, but the actual benefit varies [24,25].

At the same time, changes in the composition of viral RNA can have a number of consequences. For example, a nucleotide substitution that removes a CpG site may not alter the protein but could affect the structure of the RNA or the efficiency of protein synthesis. Therefore, a reduction in the number of recognizable motifs does not always occur without consequences for viral replication [26].

RNA editing by host–cell enzymes is one possible explanation for the sequence changes observed in SARS-CoV-2. These processes include C-to-U editing by APOBEC (apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like) enzymes and A-to-I editing by ADAR (adenosine deaminase acting on RNA) enzymes. These processes can generate mutation patterns over short time intervals. However, to link changes in circulating viruses to the action of these enzymes, it is necessary to distinguish the effects of editing from other sources of mutations and to account for subsequent selection [27].

The sequence composition may provide clues about the virus's host, although the same pattern may have multiple explanations. The frequencies of nucleotides, dinucleotides, codons, codon pairs, and amino acids may reflect the history of interaction with the host. The resulting non-random distribution of these features creates a viral genomic syntax. This syntax carries signals of host-associated selection, but viral phylogeny, genome architecture, and replication mechanisms also shape it. Furthermore, historical sampling biases influence the patterns seen in available datasets.

2.4. Evolutionary Constraints and Host-Associated Predictive Signals

Entry barriers and intracellular selection drive a continuous host–virus evolutionary conflict [28]. Viral adaptation remains strictly constrained by fitness trade-offs across multiple scales: surface protein mutations may help evade antibody recognition, but they often impair structural stability, folding, or receptor binding within interconnected fitness landscapes [29–31]. Similar trade-offs operate at the genomic level, where modifying nucleotide composition—such as reducing immunostimulatory CpG or UpA motifs to evade intracellular restriction—simultaneously alters translation efficiency, RNA structure, and codon-pair composition, carrying fitness costs elsewhere in the life cycle [26].

These strict evolutionary constraints form both the biological foundation and the primary limitation of computational host prediction. Viral genomes encode compositional and phylogenetic signals associated with known reservoir hosts or vector groups, enabling sequence-based prediction under specific conditions [32]. Importantly, these identical sequence patterns also capture viral ancestry, taxonomic relatedness, genome architecture, and dataset sampling biases. Recent analyses demonstrate that train–test dataset composition heavily skews prediction performance, with taxonomic signals frequently overpowering genuine host-associated patterns when closely related viruses populate both datasets [33].

Consequently, computational models yield probabilistic hypotheses requiring experimental validation rather than definitive evidence of reservoir status. The mechanisms outlined above establish the biological framework evaluated in the following sections across various computational representations.

3. Materials and Methods

3.1. Eligibility Criteria and Study Selection

We included computational tools that infer a host-related property of eukaryote-infecting viruses from viral molecular information, alone or in combination with candidate-host, reference or ecological metadata. A tool was eligible if it was described in peer-reviewed original research as predicting a viral host, host taxon, host range, reservoir or vector category, host compatibility, host adaptation, human infectivity or spillover risk. The tool also had to be publicly available as source code, an installable package, a command-line pipeline, a notebook with trained models or a functional web server. In addition, it had to be documented as accepting user-provided viral nucleotide or protein sequences and returning a host-related prediction for sequences beyond those analysed in the original study, without retraining the complete model-development workflow.

Tools combining viral sequences with host or ecological metadata were retained. Methods whose primary input consisted of host-trait, ecological or network data were excluded even when their prediction target was host-related. Methods requiring a metagenomic cohort rather than a query sequence were also excluded.

Eligibility was restricted to reports published in English, with no restriction on publication year. Records indexed as reviews, systematic reviews, preprints, comments, editorials or letters were excluded as ineligible publication types. We also excluded reports presenting only a model or a fixed-dataset analysis without a released implementation, tools restricted to viruses of prokaryotes, and methods for viral detection, taxonomic classification or virus–host protein interaction prediction.

Tools developed specifically for influenza viruses or coronaviruses were excluded. General-purpose tools applicable to a broader range of viruses were retained even when evaluated on influenza or coronavirus data.

The review was registered retrospectively in OSF Registries on 28 September 2026 (https://doi.org/10.17605/OSF.IO/GV6TN), and no protocol was registered prospectively. The initial PubMed search, the first round of citation searching, screening and extraction had been completed before registration. The updated PubMed search and the two later rounds of citation searching were run after registration.

Studies were identified through a systematic PubMed search and backward and forward citation searching.

The initial PubMed search was run in August 2026 with the registered query, which restricted the host-related prediction terms to titles. In a second round of searching, these terms were extended to titles and abstracts, and the search was run on 21 October 2026 using the following query:

(“host prediction”[Title/Abstract] OR “predicting host”[Title/Abstract] OR “predicting hosts”[Title/Abstract] OR “host predictor”[Title/Abstract] OR “host taxon predictor”[Title/Abstract] OR (host[Title/Abstract] AND taxon[Title/Abstract] AND predictor[Title/Abstract]) OR “host range”[Title/Abstract] OR “host association”[Title/Abstract] OR “virus–host association”[Title/Abstract] OR “virus–host associations”[Title/Abstract] OR “host adaptation”[Title/Abstract] OR “virus adaptation”[Title/Abstract] OR “viral-host adaptation”[Title/Abstract] OR “reservoir host”[Title/Abstract] OR “reservoir hosts”[Title/Abstract] OR “vector prediction”[Title/Abstract] OR “human-infecting virus”[Title/Abstract] OR “human-infecting viruses”[Title/Abstract] OR “human infectivity”[Title/Abstract] OR “spillover risk”[Title/Abstract] OR “cross-species transmission”[Title/Abstract]) AND (virus[Title/Abstract] OR viral[Title/Abstract]) AND (computational[Title/Abstract] OR model[Title/Abstract] OR framework[Title/Abstract] OR tool[Title/Abstract] OR algorithm[Title/Abstract] OR “machine learning”[Title/Abstract] OR “deep learning”[Title/Abstract] OR “language model”[Title/Abstract] OR “maximum entropy”[Title/Abstract]).

Backward and forward citation searching was performed with the OpenAlex API [34] using included reports as seeds. The search was repeated whenever a tool was added to the review, giving three rounds: an initial round in August 2026, a round on 19 October 2026 seeded with VPF-Class, and a round on 21 October 2026 seeded with GiantHost and VirID. Records were deduplicated on normalized titles and cross-checked by DOI and PMID against records from the PubMed search and earlier citation-search rounds.

Before manual title and abstract screening, records from both routes were assessed using the same publication-type criterion and a reproducible topical filter. The filter retained records whose title or abstract contained at least one viral term, one host-related term and one computational term.

After application of the publication-type criterion, topical filtering and deduplication, records from both search routes underwent title and abstract screening. The study-selection process and record counts at each stage are shown in Figure 1. All search and screening records are provided in Supplementary Table S1.

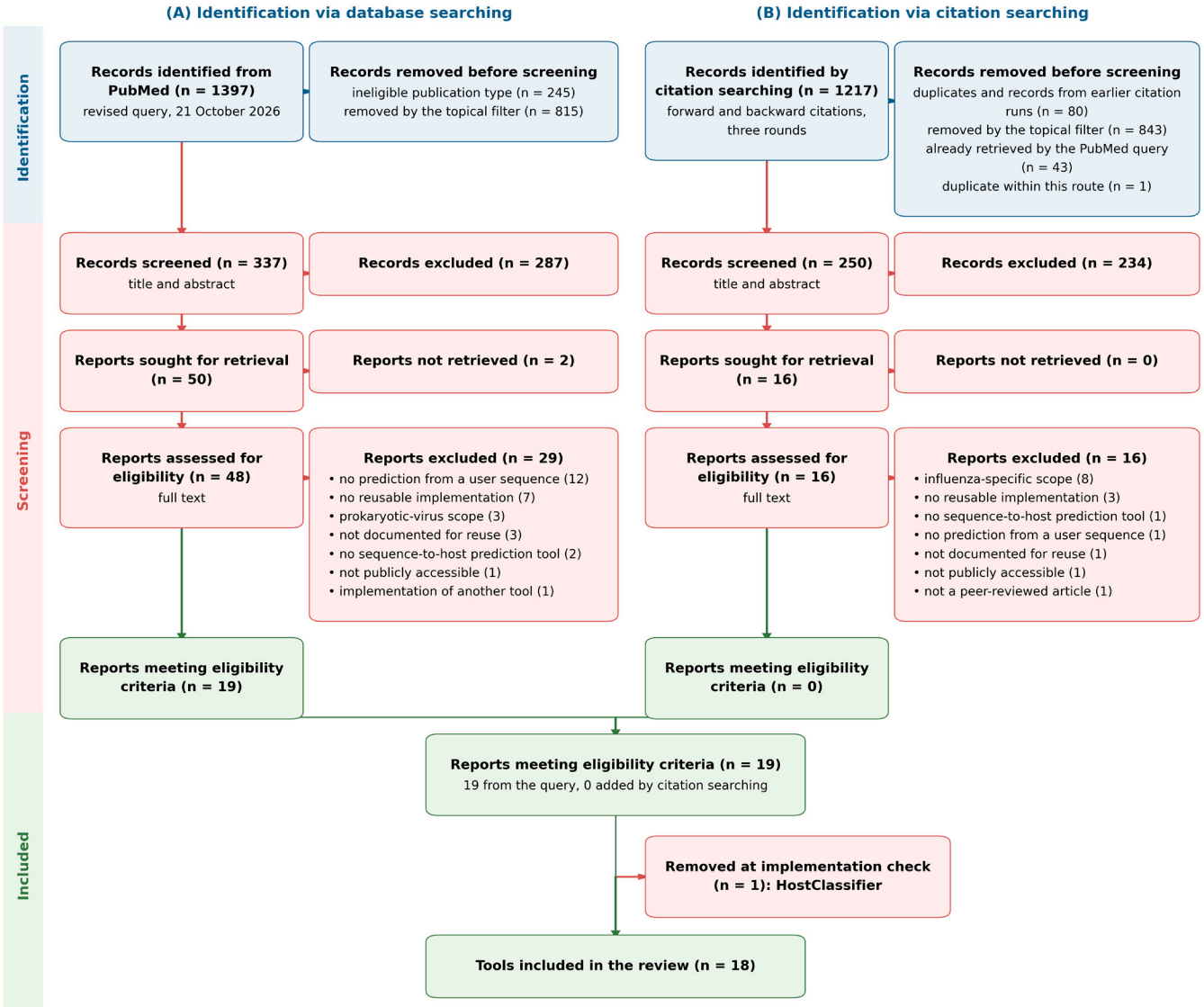


Figure 1. Study selection. PRISMA flow diagram showing the two study-identification routes. (**A**) PubMed search, including duplicate removal, title and abstract screening, full-text assessment, and exclusions by primary reason. (**B**) Backward and forward citation searching from studies included through PubMed, followed by the same selection procedure.

Every record retained for title and abstract screening was assessed by two authors. One author recorded the first decision and the second assessed the same record with that decision visible. Screening was therefore duplicated but not blinded or independent. Exclusion reasons were selected from a fixed list.

Potentially eligible reports were assessed in full text. Both authors assessed every full-text report from both identification routes and reached an agreed decision for each. Tool repositories, Supplementary Materials and documentation were inspected to confirm that the implementation was publicly accessible and documented. Each excluded report was assigned one primary exclusion reason.

No formal risk-of-bias or critical-appraisal instrument was applied because we did not identify a validated instrument designed specifically for reports of predictive software of this type.

3.2. Citation Dynamics and Documented Empirical Applications

Citation dynamics and empirical applications were analysed separately. Citation records for each primary tool publication were retrieved from OpenAlex using its DOI, cleaned and deduplicated. All 18 tools were assessed without a citation-count threshold. Citation counts were interpreted as measures of scholarly visibility rather than evidence of tool use.

Candidate applications were identified by semi-automated full-text screening in Europe PMC. Citing articles with accessible XML full text were screened after title-based exclusion of reviews, benchmarks and methodological comparisons. Methods and Experimental sections were searched for citations to the primary tool publication, tool names and predefined phrases indicating practical application. The complete phrase sets and screening rules are provided in Supplementary Table S2. Retrieved contexts were ranked and manually reviewed, and no application was included without full-text assessment.

Applications were classified into three hierarchical and mutually exclusive evidence levels. A record was classified as tool-applied when the tool or its original method was used on the study's own data, as computationally supported when the resulting conclusion was additionally corroborated by independent computational evidence, and as experimentally supported when laboratory evidence supported the predicted relationship. Every record was assessed by two authors, and disagreements were resolved against the quoted evidence.

Applications were classified as tool-applied, computationally supported, or experimentally supported. Citation data were processed in Python 3.9.6 using pandas 2.3.3, and the figure was generated using matplotlib 3.9.4 and seaborn 0.13.2.

3.3. Functional Testing of Released Implementations and Comparability of Reported Performance

Each included tool was installed from its released materials and run twice, once on the material shipped with it (where available) and once on a sequence not distributed with the tool. The test assessed whether the released implementation accepted a user-provided sequence and returned a host-related prediction. Accuracy was not evaluated, no model was retrained, and outputs were not compared with reference answers. One external sequence from GenBank or ENA was used per tool.

Runs were executed using a Python wrapper that captured the software environment, command, version, input and output files, and execution status for each test. The script uses only the Python standard library and is released together with the run log. Failed executions and any departures from the documented software environment were retained as part of the test record.

Each tool was assigned one of four levels of functional applicability according to whether obtaining a prediction required modification of the released code or its runtime

behaviour. Installation and packaging difficulties were recorded separately and did not affect this classification. Full run details are provided in Table S4 and the accompanying run log.

Separately, the comparability of published performance values was assessed before any quantitative synthesis. Reported values were compared across tools according to prediction target, unit of prediction, label set, construction of the held-out evaluation set and reported metric. They were not compared with the results of functional testing. Quantitative synthesis was considered appropriate only where tools within the same prediction target classified the same type of object, used comparable label categories, constructed the held-out set in a comparable manner and reported the same metric on a comparable basis.

4. Results

4.1. Included Tools and Prediction Targets

Following systematic screening, 18 computational tools met the eligibility criteria for eukaryotic host prediction and were included in this review. These selected tools differ substantially in their primary biological prediction targets, taxonomic scope, input requirements, and sequence representations. Rather than addressing a single task, they cover a diverse spectrum of objectives ranging from broad taxonomic assignment to highly specific epidemiological risk assessments.

We define the prediction target as the biological or epidemiological property that a tool is designed to infer from molecular data. The included tools fall into six prediction-target categories. These differ in what is predicted, which participant in the transmission cycle the prediction concerns, and the level of resolution. Tools are therefore compared within these categories rather than across them. Viral scope, host label spaces and output formats are described separately in Section 4.4.

Host-taxon classification. Seven tools assign viral sequences to broad host taxa or to small, predefined host panels: DeepHoF (2021) [35], RNAVirHost (2024) [36], Host Taxon Predictor (HTP) (2019) [37], MENB (2025) [38], VPF-Class (2021) [39], GiantHost (2026) [40] and VirID (2024) [41]. Three have qualifications related to the prediction target. The released HTP classifiers separate viruses of eukaryotes from bacteriophages, so host origin is resolved at the domain level only. GiantHost is restricted to giant viruses of the Nucleocytoviricota. VirID assigns host groups by transferring annotations from reference neighbours rather than using a trained host classifier.

Species-level host assignment. UniVH (2026) [42], EvoMIL (2024) [43], HostNet (2023) [44] and VIDHOP (2021) [45] resolve host relationships at higher resolution, addressing individual host species or specific virus–host pairs. In the HostNet model distributed for flaviviruses, the classes are vector species rather than hosts.

Human infectivity. VirHostPRED (2026) [46], zoonotic_rank (2021) [47] and DeePaC-vir (2021) [48] predict human infection from viral molecular information. The tools address the same general target, but their outputs are defined differently, so they are described individually in Section 4.4 and are not compared numerically here. zoonotic_rank characterises the virus rather than its relationship with any individual host.

Reservoir and vector prediction. ViralHostPredictor (2018) [32] accepts RNA viruses only and comprises three independent models for reservoir host, arthropod-borne transmission and vector taxon. Its reservoir assignment is based on labels transferred from phylogenetic neighbours rather than on a prediction made directly from the query sequence. MosViR (2024) [49] separates mosquito-associated viruses, arboviruses and other viruses from viral contigs, so its classes describe the vector rather than the host.

Host adaptation. GIVAL (2025) [50] predicts the adaptation of a viral lineage from any viral protein sequence. It maps the input onto a reference set and trains a classifier for

the matched group, so the labels depend on that group. Adaptation describes the fit of a viral lineage to a host and does not establish that a given host is susceptible to the virus.

Spillover risk. SPHAK (2026) [51] scores the risk that a virus crosses into a new host, treating animal viruses as candidates for spillover into humans and plant viruses as candidates for spillover into food grains. The score orders candidates for further investigation rather than predicting infection of a named host.

Not all included tools were built exclusively for eukaryotic hosts. Prokaryotic classes present in some tools are described with the host label spaces in Section 4.4.

The chronological distribution of these tools by publication year and their corresponding biological prediction targets is shown in Figure 2.

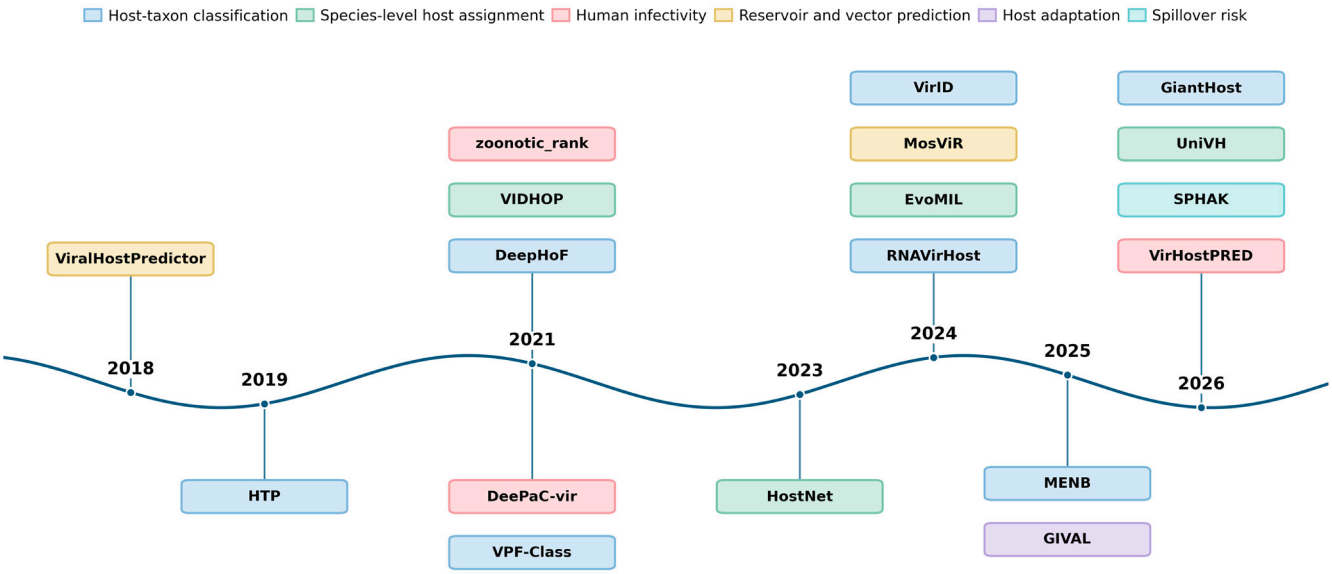


Figure 2. The timeline illustrates the development of tools for virus–host prediction. Each tool name is colored according to its prediction target.

4.2. Input Feature Classes

Computational methods differ both in the extracted viral information and how they transform this data into host-related predictions. To reduce comparative duplication, we distinguish feature classes, which serve as the input representations, from modeling approaches, which encompass the statistical or machine learning inference mechanisms.

Computational virus–host prediction fundamentally assumes that prolonged host interactions leave measurable signatures across viral genomes, proteins, and evolutionary histories. Algorithms translate these signatures into computational features to infer host association, reservoir identity, vector involvement, or zoonotic spillover potential. As newly discovered viruses are frequently highly divergent or recovered solely as fragmented metagenomic contigs, modern frameworks increasingly bypass classical alignment-based similarity, leveraging compositional, phylogenetic, embedding-based, and protein-level representations.

4.2.1. Alignment-Free Genomic Composition

Alignment-free approaches represent viral genomes through the frequencies or relative enrichment of short sequence motifs. Common features include mono-, di-, and trinucleotide frequencies, longer nucleotide k-mers, codon usage, codon-pair bias, and amino acid usage inferred from coding regions. These descriptors capture mutational biases, constraints imposed by genome architecture, and selective pressures associated

with replication in specific hosts. For instance, dinucleotide composition is highly relevant because vertebrate antiviral mechanisms can impose strict selection on motifs such as CpG and UpA [18,21,24].

HTP exemplifies an alignment-free classifier based entirely on explicit nucleotide composition. It constructs its input vectors by extracting absolute and relative mono-, di-, and trinucleotide frequencies directly from unaligned viral sequences [37].

Similarly, RNAVirHost utilizes a broader set of compositional descriptors, integrating nucleotide, dinucleotide, codon and amino acid usage preferences to represent the viral genome [36].

Other tools tailor these features for specific epidemiological prediction targets. ViralHostPredictor derives evolutionary signatures from codon, dinucleotide, and amino acid biases to represent ssRNA viruses from taxonomic groups that contain human pathogens [32].

The zoonotic_rank model calculates dinucleotide biases across entire genomes, coding regions, codon-bridge positions, and non-bridge positions, alongside measuring compositional similarity to human interferon-stimulated, housekeeping, and other functional genes [47].

MosViR represents each contig by (1,2)-mer frequencies, the conditional frequency of each dinucleotide following each single nucleotide, giving 64 values per sequence [49].

GiantHost summarises each giant-virus contig in a 137-dimensional vector of nucleotide and amino acid frequencies and codon usage [40]. A tetranucleotide vector was evaluated as an alternative but not retained, because trait-based features proved more stable across the shift from cultivated to environmental genomes.

Finally, MENB provides a more interpretable use of compositional signals by learning probability distributions constrained by observed motif frequencies, explicitly mapping the compositional changes associated with host adaptation [38].

4.2.2. Alignment-Dependent Estimation

Similarity-based features rely on the tendency of evolutionarily related viruses to share hosts. These approaches provide powerful input signals for viruses with well-characterized relatives, though their utility depends on the presence of homologous sequences in reference databases.

RNAVirHost incorporates this principle by aligning the query genome against reference sequences and converting the alignment scores into a sequence homology feature vector, allowing the model to use both taxonomic relatedness and alignment-free signals simultaneously [36].

ViralHostPredictor explicitly incorporates a phylogenetic-neighbourhood component that summarizes the reservoir or vector labels of closely related viruses [32]. This distinction is important because phylogenetic-neighbourhood features transfer information from previously characterized relatives, whereas sequence-composition features can provide signals that generalize across more distantly related viral taxa. VirID relies even more strongly on phylogenetic context. The polymerase region of a query sequence is placed on a precomputed reference tree whose tips carry host annotations curated from the Virus–Host Database, and the host assignment is then derived from the position of the query relative to these annotated reference sequences rather than from a separately trained host classifier. VPF-Class applies the same principle at the level of protein families [39]. Translated contigs are searched against classified viral protein family profiles, and the host annotation of the matched families is carried over, so reliability falls with distance from the reference families rather than from any one genome.

GIVAL utilizes sequence similarity for customized BLAST+ retrieval (BLAST+ 2.16.0 in the original implementation) to identify related proteins and construct a query-specific

dataset for downstream protein language model representation [50]. Here, similarity supports reference retrieval and semi-supervised model construction rather than acting as the final predictive signal.

4.2.3. Engineered Protein Features

Protein-level representations target the functional biomolecules executing receptor binding, cell entry, replication, immune evasion, and transmission. Operating exclusively in the amino acid space filters out synonymous nucleotide mutations and directly captures the physicochemical properties dictating host-virus interactions. Engineered features of this kind are computed by a formula fixed in advance, unlike the representations that models learn for themselves (Section 4.2.5), and they span multiple resolutions, from broad sequence composition metrics to short local motifs.

SPHAK extracts host-associated patterns by partitioning protein sequences into overlapping 6-amino acid k-mers to construct host-specific pattern matrices [51]. This methodology assumes that short linear motifs, frequently representing exposed surface loops, protease cleavage sites, or receptor-binding interfaces, carry sufficient evolutionary signal to distinguish host ranges and prioritize spillover risk.

VirHostPRED uses dipeptide composition and pseudo-amino acid composition only as baselines, comparing them with the ESM-2 embeddings on which its model is built [46].

GiantHost encodes the protein level as a sparse binary vector of roughly 8000 giant virus orthologous groups, so the unit is membership of a protein family rather than a residue frequency [40].

4.2.4. Learned Nucleotide Representations

Predefined composition features reduce genomes to global k-mer frequencies, discarding the linear order of nucleotides. The approaches in this section instead process raw sequences directly, converting nucleotides into numerical form so that the network itself, rather than the designer, determines which patterns are predictive.

One-hot encoding converts each nucleotide or codon into a rigid binary grid, functioning as a mathematical barcode. Frameworks such as DeepHoF [35] and VIDHOP [45] deploy this baseline approach, alongside tools like DeePaC-vir [48], which applies one-hot encoding directly to unaligned next-generation sequencing (NGS) reads using reverse-complement neural networks to guarantee strand-invariant host predictions. Lacking inherent semantic context, this initial encoding requires downstream neural networks (including convolutional or recurrent layers) to extract all predictive features by scanning the encoded fragments for local patterns and broader genomic context.

Advanced methods treat viral genomes as natural language by dividing the continuous sequence into discrete vocabulary units called tokens (typically single nucleotides, codons, or short k-mers) and mapping them to continuous numerical vectors. HostNet adapts the Word2Vec algorithm from natural language processing, pretraining it on a purpose-built corpus of viral genomes, to evaluate the genomic neighborhood of specific k-mers [44]. Distinct k-mers frequently sharing the same genomic context receive similar mathematical representations, a mechanism capturing the contextual grammar of the viral sequence prior to host prediction.

4.2.5. Pretrained Language-Model Representations

The approaches in the previous section learn from raw viral sequences during the host-prediction task itself. Protein language models (pLMs) employ a contrasting strategy and undergo initial pretraining on massive, unlabeled databases of known proteins. Through millions of fill-in-the-blank tasks (predicting masked amino acids based on surrounding

context), these models capture statistical regularities of protein sequences that reflect structure and evolutionary constraint, before encountering any host-prediction task.

The initial processing step requires dividing the protein into readable units, a process called tokenization. Frameworks such as VirHostPRED utilize standard architectures (including ESM-2) designed to read proteins one amino acid at a time [46]. Certain tools conceptualize this single-residue approach as biologically unnatural. GIVAL introduces vBERT (virus Bidirectional Encoder Representations from Transformers), which uses a hidden Markov model tokenizer to partition viral proteins into chunks of 1 to 5 amino acids. Instead of processing isolated residues, the model reads intact functional motifs to learn their biological context prior to classifying the virus [50].

A pLM returns one vector per residue, so a protein must be summarised before classification. The simplest option is to average these per-residue vectors into a single vector for the protein, as VirHostPRED does. Because a virus comprises multiple proteins, EvoMIL treats the entire viral proteome as a "bag" of individual proteins [43]. Through a technique called multiple-instance learning, this framework lets the model determine which specific viral proteins dictate host infection and weigh them heavily to construct a master profile for the entire virus.

UniVH extends this logic to the virus and the host at once [42]. It encodes the coding sequences of both organisms with LucaOne and groups them by gene family, so the representation spans both partners of the pair rather than the virus alone. By analysing these paired profiles alongside ecological data, the model represents the interaction itself rather than the virus alone (Section 4.2.6).

4.2.6. Host-Derived and Ecological Features

Most tools infer their prediction target from the viral sequence alone, but UniVH and zoonotic_rank also use host biology. UniVH represents both members of a candidate virus–host pair, and zoonotic_rank compares the viral genome with sets of human genes. UniVH embeds viral and host coding sequences with the LucaOne genomic language model and groups them into gene families.

These molecular representations are supplemented with ecological realm information and geographic overlap metadata as input features. For example, if a virus and a candidate host are both recorded in the Palearctic realm, UniVH additionally quantifies what proportion of the virus's observed spatial range overlaps the host distribution, and conversely, what proportion of the host range overlaps the virus distribution [42].

zoonotic_rank needs no candidate host at all. It measures how closely the viral genome resembles human interferon-stimulated, housekeeping and other functional gene sets, so the feature is anchored in host biology while the input remains the viral sequence alone [47].

4.3. Modeling Approaches

Based on their dominant inference mechanisms, we organize the 18 reviewed tools into six principal modeling paradigms: classical machine-learning models, de novo neural nucleotide classifiers, transfer learning from pretrained models, reference-augmented models, generative models, and multimodal models.

4.3.1. Classical Machine-Learning Models

Classical supervised classifiers map predefined input variables to fixed labeled outcomes using algorithms like logistic regression, support vector machines, and gradient boosting machines. While highly efficient for moderately sized datasets, they restrict predictions entirely to the labels represented during training.

For example, HTP evaluates logistic regression, quadratic discriminant analysis, k-nearest neighbors, and support vector classification to distinguish eukaryote-infecting viruses from bacteriophages [37].

zoonotic_rank uses gradient-boosted trees to rank viruses by their likelihood of infecting humans, and reports SHAP (SHapley Additive exPlanations) values showing which genome features raised or lowered each score [47].

RNAVirHost employs XGBoost within a two-level hierarchical framework, with a separate classifier trained for each virus order [36].

ViralHostPredictor applies gradient boosting machines to predict the ecological role of a virus rather than its host taxon [32].

MosViR builds two sequential soft-voting ensembles, each resampled from a pool of classical algorithms including support vector machines, k-nearest neighbors, logistic regression, random forests and discriminant analyses [49]. The first stage isolates mosquito-associated viruses, the second distinguishes dual-host arboviruses from strictly mosquito-specific lineages.

4.3.2. Neural Classifiers Trained on Viral Sequence Data

The models in this group learn task-specific representations directly from the input data rather than relying on a general-purpose pretrained protein language model. They capture complex nonlinear relationships using feed-forward, convolutional, recurrent and attention-based architectures, although they generally require larger training datasets and offer limited direct interpretability.

DeepHoF implements a BiPathCNN (bi-path convolutional neural network) architecture with separate convolutional paths for base and codon representations, and aggregates fragment-level predictions into a score for the whole sequence [35].

HostNet integrates Transformer, convolutional, and bidirectional gated recurrent unit layers to predict hosts or vectors for specific viral groups [44].

Similarly, VIDHOP aggregates short subsequence predictions using LSTM (long short-term memory) and CNN-LSTM architectures to build virus-specific classifiers [45].

DeePaC-vir applies reverse-complement convolutional and recurrent neural networks to identify human-infecting viruses, processing both strand orientations so that predictions are strand-invariant [48].

GiantHost uses a dual-tower multilayer perceptron, one branch for the trait vector and one for the orthologous-group profile, fused into a shared representation from which separate heads predict viral taxonomy and host assignment [40].

4.3.3. Transfer Learning and Pretrained Models

Transfer learning reduces the need to train massive sequence models on limited virus–host records by utilizing protein language models pretrained on extensive sequence collections.

EvoMIL applies attention-based multiple-instance learning to aggregate ESM-1b embeddings of individual viral proteins, treating the entire proteome as a bag of protein instances to predict virus-level host associations [43]. The resulting attention weights indicate model importance rather than direct experimental evidence of host specificity.

VirHostPRED evaluates individual viral proteins using fixed ESM-2 embeddings combined with a radial basis function support vector machine [46].

While both methods leverage transfer learning, EvoMIL integrates data for whole-virus predictions, whereas VirHostPRED classifies individual proteins.

4.3.4. Reference-Augmented Models

Reference-augmented approaches dynamically anchor queries against curated external databases to construct adaptive, context-dependent feature spaces. Rather than relying on static sequence properties or universal models, these methods generate localized analytical contexts by directly retrieving phylogenetic relatives, functional homologs, or reference profiles.

The architectural implementation of this dynamic retrieval varies significantly across predictive frameworks. GIVAL employs a customized BLAST+ search (BLAST+ 2.16.0) to retrieve query-related sequences, extracts vBERT embeddings, and trains a query-specific ResNet predictor to generate a context-dependent host-adaptation score [50].

SPHAK trains no model either. It derives its score from the overlap between query and reference k-mers using Bayes' theorem, after a family-level pre-assignment restricts the reference set [51].

VPF-Class is not a trained classifier at all. It matches query proteins against viral protein-family HMM (hidden Markov model) profiles and derives the host assignment deterministically from the profiles that match [39].

Ultimately, the predictive reliability of all such frameworks remains fundamentally constrained by the completeness and relevance of the underlying reference annotations.

4.3.5. Generative Models

Probabilistic generative models provide an interpretable alternative to discriminative classifiers. Maximum Entropy Nucleotide Bias (MENB) models learn probability distributions constrained by observed short motif frequencies associated with specific viral families and hosts.

MENB evaluates sequence compatibility against these learned distributions, and does so without requiring sequence alignment [38].

4.3.6. Multimodal Models

Pairwise models uniquely evaluate explicit candidate virus–host combinations, incorporating information from both organisms simultaneously. The modalities combined here are the virus and the candidate host, which distinguishes these models from architectures that fuse several descriptions of the virus alone.

UniVH represents viral and host genes using LucaOne embeddings and processes them through a Transformer encoder organised by functional category, with a binary classification head over the pair [42].

4.4. Viral Scope, Host Label Spaces and Output Types

The prediction target specifies the biological or epidemiological property inferred by a tool. Viral scope, host label space and output type define, respectively, the range of applicable viruses, the set of host categories considered and the form of the resulting prediction. RNAVirHost, UniVH, ViralHostPredictor and MENB accept RNA viruses only. MENB was trained on four single-stranded RNA virus families: Coronaviridae, Flaviviridae, Picornaviridae and Orthomyxoviridae. VPF-Class accepts DNA viruses only. VIDHOP and HostNet are limited to the viral groups represented by their distributed models. GIVAL depends on the viral groups represented in its reference set, from which training data are assembled at run time, while SPHAK is restricted to 18 viral families. DeepHoF is not restricted to a particular viral taxon, and HTP accepts both DNA and RNA viruses. GiantHost is restricted to giant viruses of the Nucleocytoviricota and excludes contigs shorter than 5 kb by default. VirID accepts RNA viruses for which an RNA-dependent RNA polymerase region can be recovered. Not all included tools were developed exclu-

sively for eukaryotic hosts. DeepHoF includes a prokaryotic germ class alongside four eukaryotic classes, although the distributed *p*-value library contains no positive example for the prokaryotic class. RNAVirHost includes Bacteria among its first-level labels, although this label is assigned through viral taxonomy rather than by the classifier. VPF-Class was developed primarily for viruses of prokaryotes but also reports Eukaryota as a host domain and can assign eukaryotic host families and genera. EvoMIL was trained on 22 prokaryotic and 36 eukaryotic hosts.

Across eukaryotic hosts, the tools range from broad host classes to species-level assignments. HTP, DeepHoF, RNAVirHost and MENB use fixed sets of broad host categories. VPF-Class reports host taxonomy, although family- and genus-level assignments mainly concern phages. VIDHOP, HostNet and EvoMIL use predefined host species, while the flavivirus model distributed with HostNet predicts vector species rather than hosts. UniVH scores user-specified virus–host pairs, with candidate hosts drawn from a fixed vocabulary of 525 species. ViralHostPredictor distinguishes reservoir groups and arthropod vector taxa, MosViR uses ecological classes, and GIVAL returns an adaptation label. VirHostPRED, DeePaC-vir, zoonotic_rank and SPHAK provide human-infectivity or spillover-related predictions. GiantHost assigns one of eight predefined host groups. VirID returns membership in four broad groups, human, vertebrate, plant and other, based on placement in a reference tree, and a sequence may receive more than one group label.

The outputs differ both in statistical meaning and in the unit to which the prediction applies. DeepHoF reports an empirical *p*-value for each host class. VPF-Class returns a membership ratio, a virus hit score and a confidence score derived from hit strength. MENB returns probabilities over three host classes that sum to one, including for viruses outside its stated scope. UniVH produces a score for each virus–host pair. SPHAK returns a classification and score for each submitted protein, while DeePaC-vir produces read-level probabilities. zoonotic_rank reports a calibrated probability, an ensemble interval, a binary classification and a priority tier. GIVAL returns an adaptation label. EvoMIL returns a probability for each candidate host species together with attention weights over viral proteins, although the distributed implementation does not expose an interface for user-supplied sequences. HostNet, as distributed, reports model-level summary metrics rather than per-sequence prediction scores. GiantHost returns a conformal set of candidate hosts with a user-defined marginal coverage, set to 90% by default, so the returned set may contain one host, several hosts or none. VirID returns binary group membership transferred from the reference tree, with the human label additionally requiring more than 80% amino acid consistency with a neighbouring human pathogen.

Eligibility was assessed from the published description and documentation of each tool. Differences between documented capabilities and the behaviour of the distributed implementation are reported separately in Section 4.7 and Table S4.

4.5. Validation Strategies and Evaluation Design

Validation design is central to interpreting reported performance because it determines how well an evaluation reflects a tool's ability to predict hosts for viruses that differ from those used in training. Validation strategies varied across the 18 tools, from internal data splits to evaluations using temporal, taxonomic or sequence-similarity separation from the training data.

VirHostPRED, EvoMIL, HostNet, VIDHOP and ViralHostPredictor relied primarily on internal evaluation through train–test splits, cross-validation or repeated resampling.

zoonotic_rank combines internal splitting with a separate holdout set of animal viruses.

Other tools assessed generalization using data separated by time or taxonomy. DeepHoF evaluated genomes from after 2018, while UniVH tested virus species from after 2020.

RNAVirHost excluded entire genera from training, and MENB evaluated performance on unseen viral families. HTP combined temporal testing on data collected after its reference snapshot with tests across fragment lengths. DeePaC-vir used species-aware splits and an external virome dataset.

Several evaluations reflected the specific tasks addressed by the tools. VPF-Class measured accuracy at different thresholds on prophage datasets, while MosViR used external mosquito metatranscriptomic contigs. GIVAL combined three-fold cross-validation on the training set retrieved for each query with independent validation sets for influenza HA and SARS-CoV-2 Spike. SPHAK used a held-out split of its reference families and an influenza case study.

GiantHost reports that isolating a large dedicated hold-out test set is impractical for giant viruses, and evaluates instead by taxonomy-aware cross-validation, leaving out one host species or one genus at a time, with standard five-fold cross-validation as a check and accuracy averaged per genome rather than per contig.

VirID validates only the phylogenetic placement step, by leave-one-out cross-validation measured as the recall of reference tree branches, and reports no validation of the host-group assignment itself, which transfers curated annotation from reference neighbours rather than applying a trained host model.

Studies also differed in how they defined negative or comparison examples. UniVH used computationally constructed negative virus–host pairs, and EvoMIL compared multiple negative-sampling strategies. During model development, zoonotic_rank treated viruses without documented human infection as non-human-infecting.

We summarized key information from Section 4.1 to Section 4.5 ($n = 20$ parameters) for all 18 tools based on their original publications in Table S3. The summary included details such as methodology, database source, input type, and output format.

The properties that decide how a tool is obtained and run, namely the unit it takes as input, the shortest input reported by its authors, the features and the model behind it, the availability of a web server, the hardware it needs and the licence of the released code, are summarised in Table 1.

Table 1. Profile of the 18 tools. The input unit, the shortest input, the feature class and the model family follow Supplementary Table S3, the web server and hardware entries follow the released code and Supplementary Table S4, and the licence was taken from the repository of each tool. Not stated means that these sources give no value.

Tool	Input Unit	Shortest Input	Feature Class	Model Family	Web Server	Hardware	Licence
DeepHoF	Genome or contig	400 bp, model fragment	learned nucleotide	neural	No	CPU, GPU optional	None stated
RNAVirHost	Genome or contig	Not stated	composition, similarity	classical, reference-augmented	No	CPU	MIT
HTP	Genome or contig, one record	500 nt, fair accuracy	composition	classical	No	CPU	GPL-3.0
MENB	Genome	Not stated	composition	generative	No	CPU	MPL-2.0
VPF-Class	Contigs	Not stated	similarity	reference-augmented	No	CPU	BSD-3-Clause
GiantHost	Contigs	5000 bp, default cutoff	composition, protein profiles	neural	No	CPU, GPU optional	None stated
VirID	Contigs, or reads	600 bp, workflow filter	similarity, phylogeny	reference-augmented	No	CPU	MIT

Table 1. *Cont.*

Tool	Input Unit	Shortest Input	Feature Class	Model Family	Web Server	Hardware	Licence
UniVH	CDS and proteins, host panel	Not stated	learned protein and gene, host traits	transfer learning, multimodal	No	GPU	None stated
EvoMIL	Proteins	25 aa, model section	learned protein	transfer learning	No	GPU	None stated
HostNet	Genome with a label table	Not stated	learned nucleotide	neural, transfer learning	No	CPU, GPU for training	Apache-2.0
VIDHOP	Genome or contig	100 to 400 nt, model fragment	learned nucleotide	neural	No	CPU	CC BY-NC-SA 4.0
VirHostPRED	Protein or proteome	Not stated	learned protein	transfer learning	Yes	Server only	None stated
zoonotic_rank	Genome, coding regions annotated	Not stated	composition, host traits	classical	No	CPU	GPL-3.0
DeePaC-vir	Reads, or a genome	250 bp, model read	learned nucleotide	neural	No	CPU, GPU optional	MIT
ViralHostPredictor	CDS, or a feature table	Not stated	composition, similarity	classical	Yes	CPU	GPL-3.0
MosViR	Contigs	500 bp, required	composition	classical	No	CPU	Artistic-2.0
GIVAL	Protein, a segment suffices	40 aa, stable accuracy	similarity, learned protein	transfer learning, generative	No	GPU	None stated
SPHAK	Proteins	Not stated	similarity, engineered	reference-augmented	No	CPU	MIT

4.6. Citation Dynamics and Applications

Scholarly citations and documented direct applications were analysed separately. After empty placeholder removal, tool-name normalization, and within-tool deduplication, the combined archived OpenAlex data contained 735 valid tool-citation links across the 18 tool publications shown in Figure 3A. All displayed tools were included without a citation-count threshold. ViralHostPredictor had the highest number of archived links ($n = 209$), followed by VPF-Class ($n = 124$), zoonotic_rank ($n = 124$), VIDHOP ($n = 78$), DeePaC-vir ($n = 65$), and HTP ($n = 39$). Each remaining tool had 19 or fewer links. No valid citing record was available for SPHAK or for GiantHost at the archived retrieval date. The counting unit was a tool-citation link, meaning that one citing publication could contribute separately to more than one tool. These counts describe scholarly visibility and were not treated as evidence of software execution or successful deployment.

The current audit retained 28 direct tool-publication applications across 27 unique publications and six tools (Figure 3B). Applications were identified for zoonotic_rank ($n = 18$), ViralHostPredictor ($n = 5$), RNAVirHost ($n = 2$), HTP ($n = 1$), VPF-Class ($n = 1$), and MosViR ($n = 1$). No direct application was documented for the remaining tools.

Within the 28 finalized records, 23 generated predictions without any secondary support. Four applications used orthogonal computational support, including cross-classifier concordance or phylogenetic contextualization, to corroborate the reported host or zoonotic-risk conclusion, and only one study was experimentally supported.

These applications occupied distinct positions within analytical workflows, representing risk ranking, host annotation, candidate selection, and sequence characterization.

zoonotic_rank functioned primarily post-discovery to rank viral sequences by predicted compatibility with human infection rather than identifying natural ecological hosts.

Its deployment spanned broad wildlife surveillance [10,52–56], environmental and host-specific viromics [57–59], and the targeted characterization of specific isolates [60–68]. *Bat ledantevirus 2* was the only experimentally supported case among all tools. After the virus received a very high zoonotic_rank score, neutralizing activity against a pseudovirus carrying its G protein was detected in 12 of 423 human serum samples (2.8%) [69].

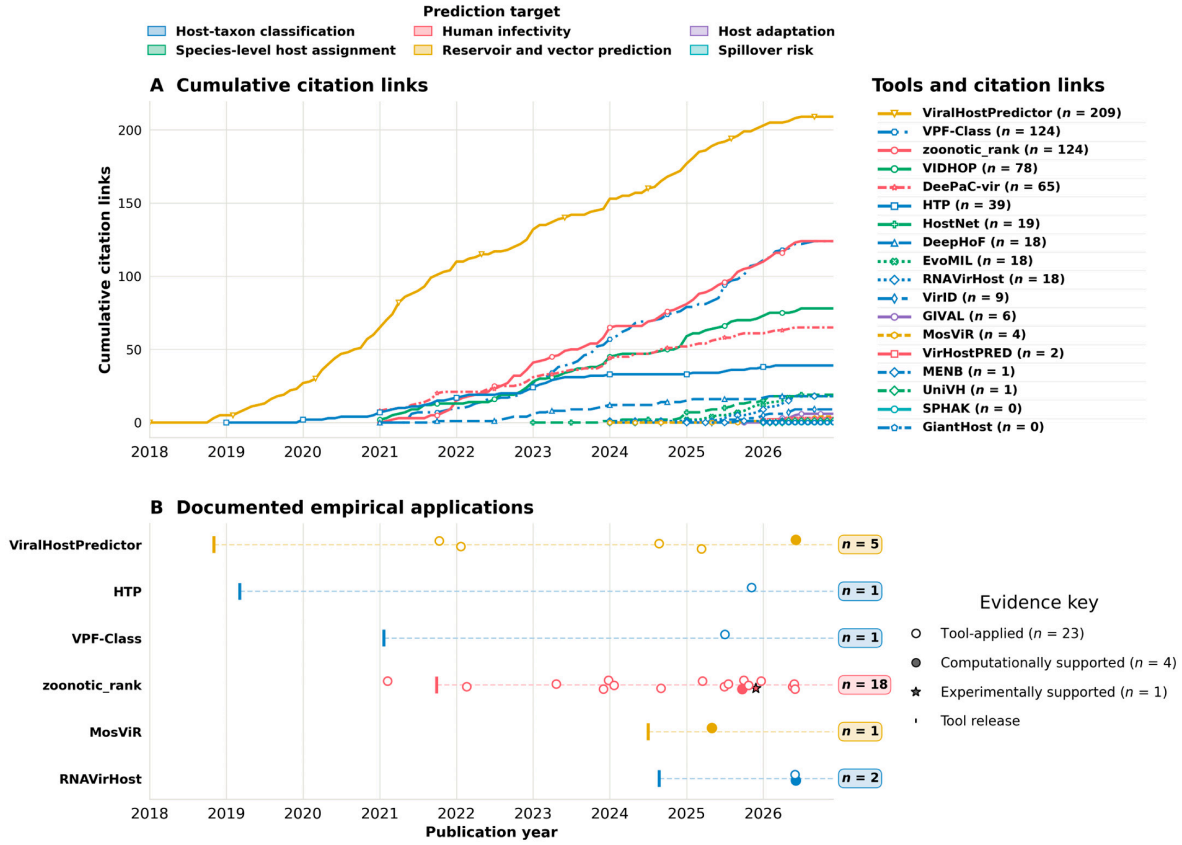


Figure 3. Scholarly visibility and documented applications of virus–host prediction tools. (**A**) Cumulative citation dynamics from OpenAlex records for all 18 tools. Colours denote prediction targets. Line styles and markers distinguish individual tools. The key displays final citation counts. (**B**) Publication timeline mapping direct tool applications to independent viral sequences or biological samples. Open circles denote isolated applications lacking secondary support. Filled circles indicate predictions corroborated by complementary evidence. Stars highlight experimental validation. Vertical bars mark tool release dates, and *n* gives the number of retained applications.

ViralHostPredictor was applied to wastewater viromes [70], a mosquito-associated iflavirus [71], the HCQD-2020 isolate (genus *Alphacoronavirus*) [72], the MUNV rhabdovirus [73], and understudied orthobunyaviruses [74].

RNAVirHost supported direct host annotation in wastewater [70] and molluscan transcriptomes [75].

HTP was used to classify rumen vOTUs (viral operational taxonomic units) as eukaryote-infecting viruses or bacteriophages [76].

VPF-Class was applied to estuarine vOTUs [77].

MosViR was applied to Atherstone and Deva viruses. Atherstone virus was predicted to be a putative arbovirus, whereas Deva virus was not classified as an arbovirus [78].

4.7. Functional Testing of the Released Implementations

Functional testing assessed whether each released implementation could process a user-provided sequence and return a host-related prediction. The assessment concerned

functional applicability rather than predictive performance. Sixteen of the 18 included tools produced an individual host-related prediction from user-supplied input.

The two exceptions were HostNet and EvoMIL. HostNet accepted user-supplied labelled data and completed execution in its rabies model, but the distributed entry point returned dataset-level summary metrics rather than a host assignment for an individual sequence. Its flavivirus model could not be reproduced from the released materials as provided because the distributed weights and label structure were incompatible. EvoMIL reproduced the evaluation workflow described in the publication but provided no input path for a user-supplied sequence, as the data path was fixed in the code and the available command-line options controlled training parameters. These cases illustrate that the capabilities described in a publication may differ from those directly accessible through the released implementation.

The intervention required to execute each released implementation was recorded on a four-level scale according to whether the tool's own code had to be modified. Ten tools produced a prediction through an author-provided interface without code changes. Two required an alternative route to the result without modification of the source code, and five required changes to the released code. EvoMIL provided no route for obtaining a prediction from a user-supplied sequence. HostNet was classified according to the intervention required to execute its distributed workflow, although the resulting output was a dataset-level metric rather than an individual host prediction. Installation and packaging difficulties were recorded separately and did not affect the classification.

Full run details, including commands, versions, dependencies, runtimes, failure modes, submitted sequences and returned outputs, are provided in Supplementary Table S4.

4.8. Comparability of Reported Outcomes

The tools fall into six prediction-target categories, four of which contain more than one tool. In none of these four groups were the reported outcomes directly comparable (Table 2).

Table 2. Comparability of reported outcomes across the six prediction targets. Tools are grouped by the target defined in Section 4.1. Quantitative synthesis was treated as possible only where the tools within a target classify the same type of object, use comparable label categories, build the held-out evaluation set in a comparable manner and report the same metric on a comparable basis. Per-tool values, their sources and the full assessment are given in Supplementary Table S5.

Prediction Target	Tools	Unit of Prediction	Label Set	Reported Metric	Synthesis
Host-taxon classification	DeepHoF, RNAVirHost, HTP, MENB, VPF-Class, GiantHost, VirID (7)	100 to 400 bp fragment to complete genome	2 to 8 categories, two hierarchical, one open below the domain level	AUC, accuracy at four different label cardinalities, two tools report none in the text	No
Species-level host assignment	UniVH, EvoMIL, HostNet, VIDHOP (4)	virus and host pair, all proteins of one virus, nucleotide subsequence	binary against a 525-species panel, or 6 to 36 species of a single virus	accuracy, AUC per host species, AUC per virus	No
Human infectivity	VirHostPRED, zoonotic_rank, DeePaC-vir (3)	one protein, one virus species, one 250 bp read pair	binary, with three different definitions of the positive class	accuracy and AUC, AUC, accuracy	No
Reservoir and vector prediction	ViralHostPredictor, MosViR (2)	one ssRNA genome, one contig of at least 500 bp	11 taxonomic reservoir groups with a binary vector-borne flag and 4 vector taxa, against 4 host-range labels in two stages	accuracy, AUC	No

Table 2. *Cont.*

Prediction Target	Tools	Unit of Prediction	Label Set	Reported Metric	Synthesis
Host adaptation	GIVAL (1)	one protein sequence	determined per query by clustering	none in the running text	Not applicable
Spillover risk	SPHAK (1)	one protein	binary, human for animal viruses and food grains for plant viruses	none for the host-related target	Not applicable

The three tools that predict human infectivity illustrate the problem. They classify, respectively, a protein, a virus species and a pair of 250-base reads, so the unit contributing to evaluation differs substantially across tools. Where accuracy-like proportions are reported, label sets range from two to eight categories, giving nominal chance levels under uniform guessing from 0.125 to 0.50. Four tools do not report a numerical performance estimate for the relevant prediction target in the main text, and only one of the eighteen publications reports an uncertainty estimate for its highlighted performance value. Variance-based pooling was therefore not possible for the remaining studies.

Evaluation-set construction also differed across tools. No two tools were assessed on identically defined held-out sets. The effect of this choice is visible even within a single tool, where the same architecture evaluated on the same underlying data yields 0.93 when the held-out set is defined by species and 0.68 when it is defined by genus.

A formal meta-analysis and direct performance ranking were therefore not performed. The synthesis is structured rather than quantitative. Per-tool values and their sources are provided in Supplementary Table S5.

4.9. Tool Selection Guide

Figure 4 summarises the findings as a task-matching guide rather than a ranking of tools. The roadmap combines prediction target, input requirements and functional applicability, with the underlying properties reported in Supplementary Tables S3 and S4. EvoMIL is not shown because its released implementation accepts no user-provided sequence. HostNet is shown with the dataset-level metrics its entry point returns.

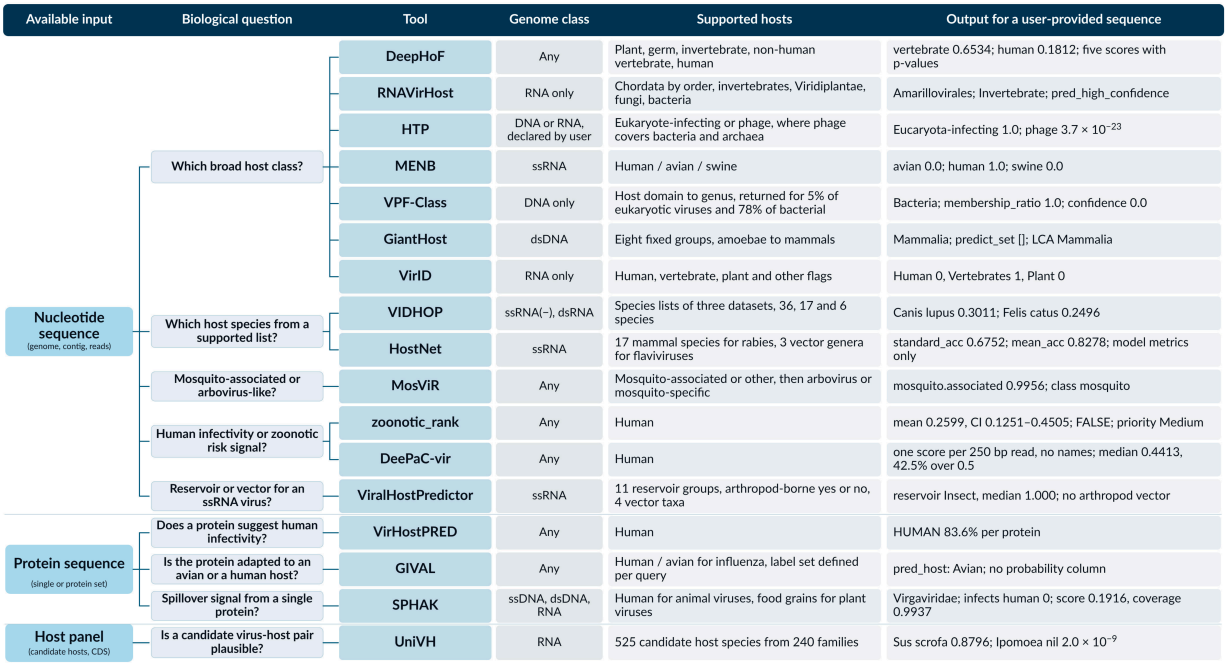


Figure 4. Tool selection scheme.

5. Discussion

This systematic review examines 18 tools for predicting potential eukaryotic virus hosts. Since the problem itself is broad, we classified these approaches, highlighting six separate biological tasks. Some methods are capable of determining exclusively the taxon of the carrier animal that the virus can infect. Other algorithms solve a different problem and assess the risk of pathogen transmission to humans or new animal species. From the perspective of utility for epidemiology and practical research, such tools can narrow the circle of working hypotheses, provided that their limitations are transparent to researchers.

The gap between a sequence-derived prediction and an actual biological outcome is crucial for risk-oriented tools. While a viral sequence contains information about host interaction, real-world infection also depends on exposure, cellular entry, replication, immune restriction, and transmission between individuals. Models rarely capture these barriers [7,79]. Therefore, a high infectivity score justifies closer investigation without automatically identifying a natural reservoir or predicting sustained spread. Likewise, evolutionary relatedness is informative for viruses with known relatives, but does not guarantee that a highly divergent virus shares their exact host range.

The way models use sequence composition highlights this issue. Some algorithms rely on basic nucleotide counts (e.g., HTP), while others analyze complex codon biases or protein data. These statistical patterns do capture real evolutionary boundaries. However, a single genetic signature can stem from entirely different biological mechanisms. For example, experiments on RNA viruses showed that a GC content mismatch reduced genome stability. However, this did not prove that GC content is a universal rule for predicting hosts [80]. Similarly, looking at codon usage in RNA viruses shows it is hard to separate nucleotide composition from translational pressure using only sequence data [20]. Therefore, finding a predictive signal does not automatically prove a specific host-adaptation mechanism. Adding complex architecture does not solve this interpretation problem. RNAVirHost and ViralHostPredictor compare data to known reference viruses, while EvoMIL and UniVH use pre-trained AI representations. Finding a match makes the prediction easier to track, and attention weights show which data influenced the model. Yet, neither proves the actual cause of host specificity. For poorly studied viruses, researchers must check if the model still works when close relatives are missing, instead of assuming the complex architecture will automatically generalize the data.

Standard performance metrics can give an impression of comparability even when the underlying evaluation settings differ substantially. We examined this directly by comparing the prediction target, unit of prediction, label set, construction of the held-out evaluation set and reported metric for each tool. No multi-tool prediction-target category was identified in which the reported outcomes shared a common evaluative basis (Table 2). Evaluation strategies and the definition of negative examples also differed substantially between studies. For example, VirHostPRED reports that bacteriophages constitute 73.2% of its negative class and that the test AUC decreases from 0.914 to 0.850 when they are removed, illustrating how strongly a reported value can depend on the composition of the comparison set. By contrast, taxonomically disjoint holdouts, such as those used by MENB, provide a stricter test of generalization to taxonomically distinct viruses. External analyses back this up. Model metrics naturally climb when training and test data share close evolutionary ties. Testing on entirely isolated families is a much harder task for these models [81]. Other work demonstrates that high overall scores can hide a model's failure on specific zoonotic lineages [82]. These findings support lineage-aware evaluation. They do not, however, prove that every tool universally fails on divergent viruses. The construction of negative examples also deserves careful attention. Given the limited data on virus–host biological incompatibility, researchers often use clearly mismatched pairs to train their models. The

fact that an algorithm easily spots these biologically impossible pairs does not mean a closely related candidate species is actually safe from the virus.

While citation metrics indicate considerable interest in these tools within the research community (735 archived links), their actual practical application remains fragmented. Documented uptake consists of only 28 applications in 27 publications across six tools, with over half of these instances attributed to zoonotic_rank. Although these tools are actively used to select target sequences from genomic and metagenomic datasets, only one retained application received experimental support for a prediction. *Bat ledantevirus 2* received a high zoonotic_rank score, after which neutralizing activity against a pseudovirus carrying its G protein was detected in 12 of 423 human serum samples (2.8%) [69]. This finding supports possible human exposure to *Bat ledantevirus 2* or an antigenically related *ledantevirus*, but it does not establish productive infection, disease, or transmission. For example, while the novel MUNV rhabdovirus was physically isolated from patient plasma, computational predictions identifying artiodactyls as the natural reservoir and midges as the vector lack empirical confirmation and remain strictly theoretical [69,73]. Testing computational predictions for wildlife viruses in the lab is difficult because cell lines for non-model species are rarely available. Field surveillance has similar limits. Finding a virus with RT-qPCR only proves it was present in the animal, not that the species is a competent reservoir. Data on host switches is also incomplete. Current databases lack records of semi-successful spillovers, where a novel pathogen causes a few isolated cases but fails to establish sustained circulation.

The discrepancies identified during functional testing should not be interpreted as shortcomings of the original studies. In this field, the boundary between a computational method, a trained predictive model and a reusable software tool is often not explicit. A publication may demonstrate predictive capability without making clear whether the released implementation can be applied directly to new user-provided sequences. The functional assessment therefore highlights a broader reporting issue: documented predictive performance and practical reusability are related but distinct properties.

Computational host prediction is currently used mainly to explore bioinformatics algorithms and patterns of virus evolution. At best, it serves as an exploratory triage tool for generating hypotheses, not as a standalone evidence framework. Because the same terms, method, framework, and tool, are applied both to released software and to code accompanying an analysis, the publication alone rarely indicates whether a usable predictor is available. To promote further advancement, prospective implementations should explicitly define the prediction unit, label space, and operational domain. Practical utility depends entirely on code, trained models, and dependency environments remaining uniformly accessible.

These challenges may be particularly relevant for protein language-model frameworks, which can require large pretrained weight files, GPU resources and specific versions of deep-learning libraries. Their continued usability may therefore depend on the availability of model weights, compatibility with upstream models and the reproducibility of the required software environment. Only when these methodological and translational gaps are addressed can computational frameworks reliably prioritize candidates without their outputs being misinterpreted as direct empirical evidence of infectious potential. Even so, the barriers described here do not diminish the scientific value of these models. Their function is not to deliver a verdict but to help researchers select sequences of interest for further study: faced with a continuous stream of metagenomic data, they narrow an enormous search space to a small set of testable hypotheses, acting as a first-pass triage that indicates to virologists and epidemiologists which taxa merit their limited laboratory and field resources.

6. Limitations

This systematic review focuses on reusable software and therefore does not provide a complete survey of all models for predicting host-related properties of viruses. Identification relied on a single bibliographic database, PubMed, supplemented by backward and forward citation searching. Although the revised query searched host-related prediction concepts in both titles and abstracts, reliance on one database means that relevant tools indexed elsewhere or embedded in broader analytical pipelines may still have been missed. Methods without a publicly accessible implementation were outside the scope of the review.

Eligibility decisions were duplicated but not independent. During title and abstract screening, the second author assessed each record with the first decision visible, and at full text the two authors reached a single agreed decision rather than assessing records independently. These procedures reduce the likelihood that an eligible report was missed by one assessor but do not provide a fully independent check, so a shared interpretation of the eligibility criteria could have affected study selection.

No formal critical-appraisal instrument was applied to the studies describing the included tools, and the review therefore does not grade the methodological quality of individual reports. Relevant methodological characteristics were extracted for each tool in Table S3 and are reported descriptively rather than converted into a quality score. Comparison with reviews using a standard appraisal instrument is therefore limited.

The structural differences among the reviewed tools prevent formal meta-analysis or direct performance ranking. This conclusion was based on a structured assessment of comparability, reported in Section 4.8 and Supplementary Table S5. The tools differ in prediction target, unit of prediction, label space, construction of the held-out evaluation set and reported metric, so their published values reflect study-specific evaluation designs rather than a common basis for comparison. Only one of the eighteen publications reports an uncertainty estimate for its highlighted performance value, so the information required for variance-based pooling is generally unavailable. Inconsistent reporting of training, validation and test datasets further limits comparison across studies.

Citation and application audits depend on how citing authors report their methods. Citation counts reflect scholarly visibility rather than software reliability or empirical use. Because the application audit required explicit evidence that a tool had been applied to independent data, the single identified case of experimental support reflects only what could be verified in the literature and does not establish that no further validation exists. Software availability was also assessed at a particular point in time and may change as repositories and web services are updated, migrated or discontinued.

Functional testing establishes only whether a prediction could be obtained and does not assess whether that prediction is correct. One sequence was submitted per tool, so the test demonstrates functional availability rather than performance across different inputs. All runs were performed by one assessor on one system between October and November 2026, with package versions recorded for each run, and results may differ in other software environments. For many tools, the documented dependency versions could not be resolved together, so testing used current compatible versions rather than the exact environments specified by the authors. This may affect the assigned applicability level because modifications required under newer dependencies might not be necessary in the original environment. Where testing was performed through a web service, its behaviour may change over time and cannot be fully reproduced from the run log. The applicability levels therefore describe the conditions encountered during this assessment and should not be interpreted as a ranking of tool quality.

7. Conclusions

This review covered 18 released tools for predicting host-related properties of eukaryote-infecting viruses. The tools address six prediction targets, from host-taxon assignment to spillover-risk ranking, and they differ in the input they accept, the label space they return and the resolution of their output. Tool choice therefore follows from the biological question and the available input rather than from reported accuracy.

Functional testing showed that 16 of the 18 released implementations returned a host-related prediction for a user-provided sequence, with the command, version, dependencies, running time and returned output recorded for every run. Published performance could not be pooled or ranked, because tools within the same prediction target differ in the unit of prediction, the label set, the construction of the held-out evaluation set and the reported metric.

Predicted host associations remain hypotheses. They narrow a large search space to a small set of candidates for laboratory and field work, and their usefulness depends on how clearly each tool states its prediction unit, its label space and the conditions under which its reported performance was obtained.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/v18111284/s1. Table S1. Search and screening records for all study selection stages. Table S2. Phrase sets and screening rules for the identification of documented applications, with the classification of every retained application. Table S3. Characteristics of the 18 virus–host prediction tools. Table S4. Functional testing of the released implementations, with the command, version, dependencies, running time, failure mode, submitted sequence and returned output of every run. Table S5. Assessment of the comparability of reported outcomes across the 18 tools. Supplementary Code. Scripts used for the PubMed search, the citation search, the logging of tool runs and the identification of documented applications.

Author Contributions: J.G. conducted the literature search, performed data analysis, generated visualizations and wrote the manuscript. P.K. contributed the literature search. N.P. performed citation analysis and wrote the manuscript. A.L. participated in the review process. E.I. supervised the group and managed funding. All authors have read and agreed to the published version of the manuscript.

Funding: This work was carried out with financial support from the Russian Ministry of Education and Science. Agreement No. 075-15-2025-530.

Data Availability Statement: All data supporting the findings are provided as Supplementary Material. Supplementary Tables S1–S5 contain the search and screening records, the application audit with its phrase set, the characteristics of the eighteen tools, the functional testing log and the comparability assessment. The analysis code, including the search, citation, application-audit and run-logging scripts, is provided as Supplementary Code.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

ACE2	angiotensin-converting enzyme 2
ADAR	adenosine deaminase acting on RNA
APOBEC	apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like
AUC	area under the receiver operating characteristic curve
CDS	coding sequence
CpG	cytosine–phosphate–guanine dinucleotide
dsDNA	double-stranded DNA
dsRNA	double-stranded RNA
ESM	Evolutionary Scale Modeling protein language model
HMM	hidden Markov model
ICTV	International Committee on Taxonomy of Viruses

MENB	maximum-entropy nucleotide bias
MIL	multiple-instance learning
MSL	Master Species List
NGS	next-generation sequencing
OAS	2′-5′-oligoadenylate synthetase
pLM	protein language model
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PseAAC	pseudo-amino acid composition
SHAP	SHapley Additive exPlanations
ssDNA	single-stranded DNA
ssRNA	single-stranded RNA
UpA	uridine-adenosine dinucleotide
VMR	Virus Metadata Resource
vOTU	viral operational taxonomic unit
VPF	viral protein family
ZAP	zinc-finger antiviral protein

References

1. Krupovic, M.; Koonin, E.V. Organization and Evolution of the Virosphere and the Replicator Space. *Comptes Rendus. Biol.* **2026**, *349*, 35–75. [CrossRef] [PubMed]
2. Shi, M.; Lin, X.-D.; Tian, J.-H.; Chen, L.-J.; Chen, X.; Li, C.-X.; Qin, X.-C.; Li, J.; Cao, J.-P.; Eden, J.-S.; et al. Redefining the Invertebrate RNA Virosphere. *Nature* **2016**, *540*, 539–543. [CrossRef] [PubMed]
3. Edgar, R.C.; Taylor, B.; Lin, V.; Altman, T.; Barbera, P.; Meleshko, D.; Lohr, D.; Novakovsky, G.; Buchfink, B.; Al-Shayeb, B.; et al. Petabase-Scale Sequence Alignment Catalyses Viral Discovery. *Nature* **2022**, *602*, 142–147. [CrossRef] [PubMed]
4. Shi, M.; Lin, X.-D.; Chen, X.; Tian, J.-H.; Chen, L.-J.; Li, K.; Wang, W.; Eden, J.-S.; Shen, J.-J.; Liu, L.; et al. The Evolutionary History of Vertebrate RNA Viruses. *Nature* **2018**, *556*, 197–202. Correction in *Nature* **2018**, *561*, E6. [CrossRef] [PubMed]
5. Porter, A.F.; Cobbin, J.; Li, C.-X.; Eden, J.-S.; Holmes, E.C. Metagenomic Identification of Viral Sequences in Laboratory Reagents. *Viruses* **2021**, *13*, 2122. [CrossRef] [PubMed]
6. Haydon, D.T.; Cleaveland, S.; Taylor, L.H.; Laurenson, M.K. Identifying Reservoirs of Infection: A Conceptual and Practical Challenge. *Emerg. Infect. Dis.* **2002**, *8*, 1468–1473. [CrossRef] [PubMed]
7. Plowright, R.K.; Parrish, C.R.; McCallum, H.; Hudson, P.J.; Ko, A.I.; Graham, A.L.; Lloyd-Smith, J.O. Pathways to Zoonotic Spillover. *Nat. Rev. Microbiol.* **2017**, *15*, 502–510. [CrossRef] [PubMed]
8. Geoghegan, J.L.; Duchêne, S.; Holmes, E.C. Comparative Analysis Estimates the Relative Frequencies of Co-Divergence and Cross-Species Transmission within Viral Families. *PLoS Pathog.* **2017**, *13*, e1006215. [CrossRef] [PubMed]
9. French, R.K.; Anderson, S.H.; Cain, K.E.; Greene, T.C.; Minor, M.; Miskelly, C.M.; Montoya, J.M.; Wille, M.; Muller, C.G.; Taylor, M.W.; et al. Host Phylogeny Shapes Viral Transmission Networks in an Island Ecosystem. *Nat. Ecol. Evol.* **2023**, *7*, 1834–1843. [CrossRef] [PubMed]
10. Bergner, L.M.; Mollentze, N.; Orton, R.J.; Tello, C.; Broos, A.; Biek, R.; Streicker, D.G. Characterizing and Evaluating the Zoonotic Potential of Novel Viruses Discovered in Vampire Bats. *Viruses* **2021**, *13*, 252. [CrossRef] [PubMed]
11. Liu, S.; Hu, D.; Xu, T.; Yin, J.; Shan, X.; Xia, J.; Wang, Z.; Xu, R.; Wang, C.; Anderson, D.E.; et al. An Emerging Human Eye Disease Is Associated with Aquatic Virus Zoonotic Infection. *Nat. Microbiol.* **2026**, *11*, 892–906. Correction in *Nat. Microbiol.* **2026**, *11*, 1756. [CrossRef] [PubMed]
12. White, J.M.; Delos, S.E.; Brecher, M.; Schornberg, K. Structures and Mechanisms of Viral Membrane Fusion Proteins: Multiple Variations on a Common Theme. *Crit. Rev. Biochem. Mol. Biol.* **2008**, *43*, 189–219. [CrossRef] [PubMed]
13. Liu, Y.; Sheng, J.; van Vliet, A.L.W.; Buda, G.; van Kuppeveld, F.J.M.; Rossmann, M.G. Molecular Basis for the Acid-Initiated Uncoating of Human Enterovirus D68. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E12209–E12217. [CrossRef] [PubMed]
14. Zhang, Y.; Li, Y.; Gao, W.; Li, D.; Xia, L.; Zhou, Q. Molecular Insights into Species-Specific ACE2 Recognition of Coronavirus HKU5. *Nat. Commun.* **2025**, *16*, 10889. [CrossRef] [PubMed]
15. Sun, Y.; Zhang, K.; Qi, H.; Zhang, H.; Zhang, S.; Bi, Y.; Wu, L.; Sun, L.; Qi, J.; Liu, D.; et al. Computational Predicting the Human Infectivity of H7N9 Influenza Viruses Isolated from Avian Hosts. *Transbound. Emerg. Dis.* **2021**, *68*, 846–856. [CrossRef] [PubMed]
16. Kumar, R.; Raghava, G.P.S. Hybrid Approach for Predicting Coreceptor Used by HIV-1 from Its V3 Loop Amino Acid Sequence. *PLoS ONE* **2013**, *8*, e61437. Correction in *PLoS ONE* **2013**, *8*. https://doi.org/10.1371/annotation/5c57dcdc-e5d9-4999-a7d0-32004427cba5. [CrossRef] [PubMed]
17. Cameron, K.; Rozano, L.; Falasca, M.; Mancera, R.L. Does the SARS-CoV-2 Spike Protein Receptor Binding Domain Interact Effectively with the DPP4 (CD26) Receptor? A Molecular Docking Study. *Int. J. Mol. Sci.* **2021**, *22*, 7001. [CrossRef] [PubMed]

18. Lo, R.; Gonçalves-Carneiro, D. Sensing Nucleotide Composition in Virus RNA. *Biosci. Rep.* **2023**, *43*, BSR20230372. [CrossRef] [PubMed]
19. Jenkins, G.M.; Holmes, E.C. The Extent of Codon Usage Bias in Human RNA Viruses and Its Evolutionary Origin. *Virus Res.* **2003**, *92*, 1–7. [CrossRef] [PubMed]
20. Belalov, I.S.; Lukashev, A.N. Causes and Implications of Codon Usage Bias in RNA Viruses. *PLoS ONE* **2013**, *8*, e56642. [CrossRef] [PubMed]
21. Takata, M.A.; Gonçalves-Carneiro, D.; Zang, T.M.; Soll, S.J.; York, A.; Blanco-Melo, D.; Bieniasz, P.D. CG Dinucleotide Suppression Enables Antiviral Defence Targeting Non-Self RNA. *Nature* **2017**, *550*, 124–127. [CrossRef] [PubMed]
22. Gonçalves-Carneiro, D.; Mastrocola, E.; Lei, X.; DaSilva, J.; Chan, Y.F.; Bieniasz, P.D. Rational Attenuation of RNA Viruses with Zinc Finger Antiviral Protein. *Nat. Microbiol.* **2022**, *7*, 1558–1567. [CrossRef] [PubMed]
23. Bohn, J.A.; Meagher, J.L.; Takata, M.A.; Gonçalves-Carneiro, D.; Yeoh, Z.C.; Ohi, M.D.; Smith, J.L.; Bieniasz, P.D. Functional Anatomy of Zinc Finger Antiviral Protein Complexes. *Nat. Commun.* **2024**, *15*, 10834. [CrossRef] [PubMed]
24. Odon, V.; Fros, J.J.; Goonawardane, N.; Dietrich, I.; Ibrahim, A.; Alshaikhahmed, K.; Nguyen, D.; Simmonds, P. The Role of ZAP and OAS3/RNAseL Pathways in the Attenuation of an RNA Virus with Elevated Frequencies of CpG and UpA Dinucleotides. *Nucleic Acids Res.* **2019**, *47*, 8061–8083. [CrossRef] [PubMed]
25. Simón, D.; Megrian, D.; Walt, H.K.; Moreno, P.; Musto, H.; Hoffmann, F.G.; Moratorio, G. Impact of the Zinc Antiviral Protein on the Genomic Composition of RNA Viruses Infecting Vertebrates. *Mol. Biol. Evol.* **2025**, *42*, msaf135. [CrossRef] [PubMed]
26. Gonçalves-Carneiro, D.; Bieniasz, P.D. Mechanisms of Attenuation by Genetic Recoding of Viruses. *mBio* **2021**, *12*, e02238-20. [CrossRef] [PubMed]
27. Schwemmle, M. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere* **2020**, *5*, e00408-20. [CrossRef] [PubMed]
28. Daugherty, M.D.; Malik, H.S. Rules of Engagement: Molecular Insights from Host-Virus Arms Races. *Annu. Rev. Genet.* **2012**, *46*, 677–700. [CrossRef] [PubMed]
29. Starr, T.N.; Greaney, A.J.; Hilton, S.K.; Ellis, D.; Crawford, K.H.D.; Dingens, A.S.; Navarro, M.J.; Bowen, J.E.; Tortorici, M.A.; Walls, A.C.; et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **2020**, *182*, 1295–1310.e20. [CrossRef] [PubMed]
30. Greaney, A.J.; Starr, T.N.; Gilchuk, P.; Zost, S.J.; Binshtein, E.; Loes, A.N.; Hilton, S.K.; Huddleston, J.; Eguia, R.; Crawford, K.H.D.; et al. Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain That Escape Antibody Recognition. *Cell Host Microbe* **2021**, *29*, 44–57.e9. [CrossRef] [PubMed]
31. Tenthorey, J.L.; Emerman, M.; Malik, H.S. Evolutionary Landscapes of Host-Virus Arms Races. *Annu. Rev. Immunol.* **2022**, *40*, 271–294. [CrossRef] [PubMed]
32. Babayan, S.A.; Orton, R.J.; Streicker, D.G. Predicting Reservoir Hosts and Arthropod Vectors from Evolutionary Signatures in RNA Virus Genomes. *Science* **2018**, *362*, 577–580. [CrossRef] [PubMed]
33. Perelygin, F.S.; Lukashev, A.N.; Aleshina, Y.A. The Effect of Taxonomic, Host-Dependent Features and Sample Bias on Virus Host Prediction Using Machine Learning and Short Sequence K-Mers. *Sci. Rep.* **2025**, *15*, 31592. [CrossRef] [PubMed]
34. Priem, J.; Piwowar, H.; Orr, R. OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts. *arXiv* **2022**, arXiv:2205.01833.
35. Guo, Q.; Li, M.; Wang, C.; Guo, J.; Jiang, X.; Tan, J.; Wu, S.; Wang, P.; Xiao, T.; Zhou, M.; et al. Predicting Hosts Based on Early SARS-CoV-2 Samples and Analyzing the 2020 Pandemic. *Sci. Rep.* **2021**, *11*, 17422. [CrossRef] [PubMed]
36. Chen, G.; Jiang, J.; Sun, Y. RNAVirHost: A Machine Learning-Based Method for Predicting Hosts of RNA Viruses through Viral Genomes. *Gigascience* **2024**, *13*, giae059. [CrossRef] [PubMed]
37. Gałan, W.; Bąk, M.; Jakubowska, M. Host Taxon Predictor—A Tool for Predicting Taxon of the Host of a Newly Discovered Virus. *Sci. Rep.* **2019**, *9*, 3436. [CrossRef] [PubMed]
38. Di Gioacchino, A.; Lecce, I.; Greenbaum, B.D.; Monasson, R.; Cocco, S. Deciphering the Code of Viral-Host Adaptation Through Maximum-Entropy Nucleotide Bias Models. *Mol. Biol. Evol.* **2025**, *42*, msaf127. [CrossRef] [PubMed]
39. Pons, J.C.; Paez-Espino, D.; Riera, G.; Ivanova, N.; Kyrpides, N.C.; Llabrés, M. VPF-Class: Taxonomic Assignment and Host Prediction of Uncultivated Viruses Based on Viral Protein Families. *Bioinformatics* **2021**, *37*, 1805–1813. [CrossRef] [PubMed]
40. Qu, F.; Chen, G.; Sun, Y. GiantHost: A Domain-Adaptive and Uncertainty-Aware Framework for Giant Virus Host Prediction. *Bioinformatics* **2026**, *42*, btag498. [CrossRef] [PubMed]
41. Yang, Z.; Shan, Y.; Liu, X.; Chen, G.; Pan, Y.; Gou, Q.; Zou, J.; Chang, Z.; Zeng, Q.; Yang, C.; et al. VirID: Beyond Virus Discovery—An Integrated Platform for Comprehensive RNA Virus Characterization. *Mol. Biol. Evol.* **2024**, *41*, msae202. [CrossRef] [PubMed]
42. Zhao, D.; Wang, Y.-F.; Yin, Z.-F.; Wu, Y.-F.; Yu, H.-J.; Xia, L.-Y.; Liu, X.-H.; Cui, X.-M.; Shi, X.-Y.; Zhu, D.-Y.; et al. A Unified Framework to Prioritize RNA Virus Cross-Species Transmission Risk Across an Expansive Host Landscape. *Viruses* **2026**, *18*, 211. [CrossRef] [PubMed]

43. Liu, D.; Young, F.; Lamb, K.D.; Robertson, D.L.; Yuan, K. Prediction of Virus-Host Associations Using Protein Language Models and Multiple Instance Learning. *PLoS Comput. Biol.* **2024**, *20*, e1012597. [CrossRef] [PubMed]
44. Ming, Z.; Chen, X.; Wang, S.; Liu, H.; Yuan, Z.; Wu, M.; Xia, H. HostNet: Improved Sequence Representation in Deep Neural Networks for Virus-Host Prediction. *BMC Bioinform.* **2023**, *24*, 455. [CrossRef] [PubMed]
45. Mock, F.; Viehweger, A.; Barth, E.; Marz, M. VIDHOP, Viral Host Prediction with Deep Learning. *Bioinformatics* **2021**, *37*, 318–325. [CrossRef] [PubMed]
46. Beltrán, J.F.; Belén, L.H.; Parraguez-Contreras, F.; Yañez, A.J. Protein Language Models Enable Accurate Viral Host Range Prediction. *Sci. Rep.* **2026**, *16*, 7606. [CrossRef] [PubMed]
47. Mollentze, N.; Babayan, S.A.; Streicker, D.G. Identifying and Prioritizing Potential Human-Infecting Viruses from Their Genome Sequences. *PLoS Biol.* **2021**, *19*, e3001390. [CrossRef] [PubMed]
48. Bartoszewicz, J.M.; Seidel, A.; Renard, B.Y. Interpretable Detection of Novel Human Viruses from Genome Sequencing Data. *NAR Genom. Bioinform.* **2021**, *3*, lqab004. [CrossRef] [PubMed]
49. de Andrade, A.A.S.; Brustolini, O.; Grivet, M.; Schrago, C.G.; Vasconcelos, A.T.R. Predicting Novel Mosquito-Associated Viruses from Metatranscriptomic Dark Matter. *NAR Genom. Bioinform.* **2024**, *6*, lqae077. [CrossRef] [PubMed]
50. Jiang, S.-Y.; Zhao, S.-S.; Wei, J.-Q.; Zhang, S.; Zhao, Z.; Tong, Y.; Liu, W.; Wang, J.; Jiang, T.; Li, J. General Intelligence Framework to Predict Virus Adaptation Based on a Genome Language Model. *Research* **2025**, *8*, 0871. [CrossRef] [PubMed]
51. N G, V.; Prakash, A.; Kavya, S.; Rajalakshmi, C.; Paul, P.M.; Thomas, V.I. Identifying Host-Specific Patterns in Viral Protein Sequences to Predict Host Spillover Risk in Animal and Plant Kingdoms. *Sci. Rep.* **2026**, *16*, 21894. [CrossRef] [PubMed]
52. Zhu, W.; Yang, J.; Lu, S.; Jin, D.; Pu, J.; Wu, S.; Luo, X.-L.; Liu, L.; Li, Z.; Xu, J. RNA Virus Diversity in Birds and Small Mammals From Qinghai-Tibet Plateau of China. *Front. Microbiol.* **2022**, *13*, 780651. [CrossRef] [PubMed]
53. Guo, Y.; Su, C.; Liang, H.; Jiang, X.; Yang, R.; Ye, J.; Gillespie, T.R.; Gao, Z.; Xu, L. Virome Diversity and Potential Sharing of Wild Mammals in a Biodiversity Hotspot, Yunnan, China. *Virol. J.* **2025**, *22*, 79. [CrossRef] [PubMed]
54. Dong, T.; Wang, Q.; Que, T.; Si, H.; Su, J.; Chen, Y.; Yang, K.; Li, C.; Qin, M.; Li, B.; et al. Healthy Pangolin Virome Reveals Mammalian Viral Diversity and Zoonotic Risk. *Microbiome* **2025**, *14*, 42. [CrossRef] [PubMed]
55. Lin, Y.; Li, F.; Liang, P.; Zhou, Y.; Zhang, L.; Zhou, W.; Liang, Y.; Yu, R.; Yang, W.; Zhou, Z.; et al. Infectome Landscape of Rodents and Shrews in Guangdong Province Reveals Diverse Pathogens with Zoonotic Potential in Wildlife. *Viruses* **2026**, *18*, 584. [CrossRef] [PubMed]
56. Wang, H.; Meng, Y.; Chen, X.; Cui, X.; Zuo, Q.; Han, N.; Liang, X.; Shen, X.; Li, C.; Li, D.; et al. Unveiling Novel Viral Diversity, Biogeography, and Host Networks in Wildlife Through High-Throughput Sequencing Data Mining. *Adv. Sci.* **2025**, *12*, e11920. [CrossRef] [PubMed]
57. Lin, Y.; Pascall, D.J. Characterisation of Putative Novel Tick Viruses and Zoonotic Risk Prediction. *Ecol. Evol.* **2024**, *14*, e10814. [CrossRef] [PubMed]
58. Yi, B.; Deng, Q.; Guo, C.; Li, X.; Wu, Q.; Zha, R.; Wang, X.; Lu, J. Evaluating the Zoonotic Potential of RNA Viromes of Rodents Provides New Insight into Rodent-Borne Zoonotic Pathogens in Guangdong, China. *One Health* **2023**, *17*, 100631. [CrossRef] [PubMed]
59. Liu, Y.; Zhang, Z.; Jiao, N.; Zhu, Y.-G.; Zhang, R.; Dominguez-Huerta, G.; Wang, H.; Feng, M.; Wen, R.; Ji, M.; et al. Diverse Novel RNA Viruses in the Cryosphere of the Tibetan Plateau. *Natl. Sci. Rev.* **2025**, *12*, nwaf361. [CrossRef] [PubMed]
60. Lamba, V.; Herlekar, I.; Chatterjee, D.; Patel, K.; Garg, K.M.; Chattopadhyay, B. Signature of Viral Fossils: A Comparative Genomics Approach to Understand the Diversity of Endogenous Retroviruses in Bats. *Virus Res.* **2024**, *350*, 199484. [CrossRef] [PubMed]
61. Claro, I.M.; Hua, X.; Viveros, A.; Moreira, F.R.R.; Proença-Módena, J.L.; Plante, K.S.; Weaver, S.C.; de Souza, W.M. Genomic Characterization of Arboviruses Discovered between the 1950s and 1980s. *J. Virol.* **2025**, *99*, e0121425. [CrossRef] [PubMed]
62. Kim, J.; No, J.S.; Park, K.; Lee, S.-H.; Natasha, A.; Pangestu, H.S.; Klein, T.A.; Kim, H.C.; Kim, W.-K.; Song, J.-W. Epidemiological Surveillance and Genomic Characterisation of Betacoronavirus in from Gyeonggi and Gangwon Provinces Republic of Korea, 2017–2019. *One Health* **2026**, *22*, 101381. [CrossRef] [PubMed]
63. Orf, G.S.; Olivo, A.; Harris, B.; Weiss, S.L.; Achari, A.; Yu, G.; Federman, S.; Mbanya, D.; James, L.; Mampunza, S.; et al. Metagenomic Detection of Divergent Insect- and Bat-Associated Viruses in Plasma from Two African Individuals Enrolled in Blood-Borne Surveillance. *Viruses* **2023**, *15*, 1022. [CrossRef] [PubMed]
64. Roev, G.V.; Borisova, N.I.; Chistyakova, N.V.; Vyhodtseva, A.V.; Akimkin, V.G.; Khafizov, K.F. Bastroviruses (): Genetic Diversity and Potential Impact on Human and Animal Health. *Vopr. Virusol.* **2023**, *68*, 505–512. [CrossRef] [PubMed]
65. Oh, Y.; Natasha, A.; Park, J.; Kim, W.-K.; Cho, H.-S. Lumpy Skin Disease Virus: Assessing Zoonotic Potential through Machine Learning. *Korean J. Vet. Serv.* **2025**, *48*, 173–180. [CrossRef]
66. Park, K.; Kim, S.-G.; Kim, J.-A.; Prayitno, S.P.; Kim, J.; Lee, S.-H.; Lee, J.; Seo, Y.-R.; Natasha, A.; Lee, H.I.; et al. Molecular Evolution and Zoonotic Potential of Muju Virus (Orthohantavirus Puumalaense) in Craseomys Regulus, Republic of Korea. *J. Med. Virol.* **2025**, *97*, e70503. [CrossRef] [PubMed]

67. Lee, S.; Lee, S.S.; Kim, B.; Pangestu, H.S.; Natasha, A.; Byun, J.H.; Park, J.; Prayitno, S.P.; Lee, W.; Oh, Y.; et al. Amplicon-Based MinION Sequencing, Genomic Characterization, and Zoonotic Potential of Patient-Derived Chikungunya Virus Imported from Thailand, Republic of Korea. *J. Bacteriol. Virol.* **2025**, *55*, 120–130. [CrossRef]
68. Pangestu, H.S.; Yang, I.; Natasha, A.; Rajoriya, S.; Hennisa, H.; Park, J.; Park, K.; Kim, J.; Kim, S.-G.; Klein, T.A.; et al. Molecular Prevalence, Genomic Characterization, and Zoonotic Potential of Novel Paramyxovirus and Hepacivirus in Alexandromys Fortis, Republic of Korea. *Vet. Res.* **2026**, *57*, 96. [CrossRef] [PubMed]
69. Zhang, X.-A.; Zhang, M.-Q.; Liu, Y.-W.; Lin, L.; Zhang, J.-T.; George, T.; Jalloh, M.B.; Sevalie, S.; Kargbo, K.B.; Jiang, B.-G.; et al. Virome Characterization of Wild Small Mammals Provides New Insight into Zoonotic Pathogens in West Africa. *Microbiome* **2025**, *13*, 242. [CrossRef] [PubMed]
70. Xu, F.; Qin, H.; Yu, F.; Qiu, Z.; Zhang, D.; Li, N.; Xiao, P. Comparative Virome Analysis of Lake and Domestic Wastewater Revealed the Unexpected Presence of Swine Acute Diarrhea Syndrome Coronavirus and -Associated Viruses. *Front. Microbiol.* **2026**, *17*, 1831710. [CrossRef] [PubMed]
71. Guimarães, L.d.O.; Bahia, S.L.; Ribeiro, G.d.O.; Ramos, E.d.S.F.; Villanova, F.; Dos Santos Morais, V.; Telles-de-Deus, J.; Helfstein, V.C.; Santos, J.M.D.; Pandey, R.P.; et al. New Iflavirus Species Characterized from Mosquitoes Captured in the Sao Paulo Zoological Facilities. *Microorganisms* **2024**, *12*, 1749. [CrossRef] [PubMed]
72. Do, H.-Q.; Nguyen, V.-G.; Chung, C.-U.; Jeon, Y.-S.; Shin, S.; Jang, K.-C.; Pham, L.B.H.; Kong, A.; Kim, C.-U.; Park, Y.-H.; et al. Genomic Characterization of a Novel Alphacoronavirus Isolated from Bats, Korea, 2020. *Viruses* **2021**, *13*, 2041. [CrossRef] [PubMed]
73. Edridge, A.W.D.; Abd-Elfarag, G.; Deijs, M.; Jebbink, M.F.; Boele van Hensbroek, M.; van der Hoek, L. Divergent Rhabdovirus Discovered in a Patient with New-Onset Nodding Syndrome. *Viruses* **2022**, *14*, 210. [CrossRef] [PubMed]
74. Sankhe, S.; Dieng, I.; Kane, M.; Diallo, A.; Ndiaye, N.A.; Top, N.M.; Dia, M.; Faye, O.; Sall, A.A.; Faye, O.; et al. Genomic Insights into Neglected Orthobunyaviruses: Molecular Characterization and Phylogenetic Analysis. *Viruses* **2025**, *17*, 406. [CrossRef] [PubMed]
75. Li, J.; Chen, K.; Chen, X.; Zheng, K.; Sun, S.; Yu, H.; Dong, Y.; Wang, Z.; Xu, Y.; McMinn, A.; et al. A Comprehensive RNA Virome from Molluscan Transcriptomes Reveals Extensive Diversity and Modular Genome Evolution. *Virol. Sin.* **2026**, *41*, 532–542. [CrossRef] [PubMed]
76. Navarro Marcos, C.; Gutiérrez-Rivas, M.; Goiri, I.; García-Rodríguez, A.; González-Recio, Ó. The Association of the Rumen Virome with Methane Emissions in Dairy Cattle. *Commun. Biol.* **2025**, *8*, 1534. [CrossRef] [PubMed]
77. Hussain, S.; Wang, X.; Pan, C.; Chen, S.; Xie, J.; Mahtab, N.; Hou, S.; Li, S. Salinity-Driven Shifts in Estuarine Viral Community Composition and Diversity near the Shenzhen Coast. *Appl. Environ. Microbiol.* **2025**, *91*, e0040725. [CrossRef] [PubMed]
78. Pilgrim, J.; Seechurn, N.; Cunningham-Oakes, E.; Dobbs, P.; Lopez, J.; Kohl, A.; Hughes, G.; Blagrove, M.S.C.; Baylis, M.; Darby, A. Emergence Risks from within? Metagenomic Analysis of Mosquito Viromes from Two Zoos Reveals Mosquito-Associated Orthobunyaviruses in the UK. *bioRxiv* **2025**. [CrossRef]
79. Mollentze, N.; Streicker, D.G. Predicting Zoonotic Potential of Viruses: Where Are We? *Curr. Opin. Virol.* **2023**, *61*, 101346. [CrossRef] [PubMed]
80. Kawahara, S.; Nao, N.; Tulakarnwong, S.; Suzuki, S.; Suzuki, R.; Chimnaronk, S.; Tamura, T.; Fukuhara, T. GC Content Mismatch of Transgene Destabilizes RNA Virus Genomes. *J. Virol.* **2026**, *100*, e00882-26. [CrossRef] [PubMed]
81. Reddy, T.; Schneider, A.; Hall, A.R.; Witmer, A.; Hengartner, N. An Improved Dataset for Predicting Mammal Infecting Viruses from Genetic Sequence Information. *PLoS Comput. Biol.* **2026**, *22*, e1014125. [CrossRef] [PubMed]
82. Kawasaki, J.; Suzuki, T.; Hamada, M. Hidden Challenges in Evaluating Spillover Risk of Zoonotic Viruses Using Machine Learning Models. *Commun. Med.* **2025**, *5*, 187. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.