Alignment-Free Viral Metagenomics: External Validation of K-mer Based Ensemble Classifiers on an Independent Amazonian Dataset

Hiago D. Muniz[1], **João Paulo P. de Almeida**[2],
Carlos Augusto C. S. Júnior[1], **Ulisses B. Corrêa**[1]

[1] Centro de Desenvolvimento Tecnológico – Universidade Federal de Pelotas (UFPel)
Caixa Postal 354 – 96.010-610 – Pelotas – RS – Brazil

{hdmuniz, ubcorrea, cacsjunior}@inf.ufpel.edu.br

[2]University of Strasbourg
Strasbourg – Grand Est – France

jpdealmeida@unistra.fr

Abstract. *The evaluation of alignment-free viral classifiers has been conducted predominantly through internal cross-validation on curated databases, without exposure to novel environmental viromes. This work presents an out-of-distribution validation case study of a previously proposed alignment-free pipeline, in which ExtraTrees and XGBoost classifiers operate over k-mer embeddings generated by Word2Vec and fastText. Trained on the ZOVER repository, the models are applied without retraining to a fully independent metagenomic dataset derived from an RNA mosquito virome from the Amazon Basin. The validation set contains 7,624 sequences, of which 82 are homology-confirmed viruses from highly divergent families such as Phasmaviridae and Phenuiviridae, embedded in 7,542 host transcripts under strong class imbalance close to one virus per ninety-two host sequences. Without any retraining, ExtraTrees recovered 84.6% of the viral sequences and XGBoost 82.3%, with specificity around 88% and ROC-AUC near 0.95, classifying every input sequence in seconds on CPU and without gene annotation. In sensitivity, both models are comparable to three established tools, namely DeepVirFinder, VirSorter2 and geNomad, with no detectable difference given the small positive set, although with lower specificity. The results show that classical ensemble classifiers based on k-mer embeddings retain useful discriminative signal on divergent wild viromes, positioning them as a fast, annotation-free option for first-pass metagenomic screening.*

1. Introduction

The identification of viral sequences in metagenomic data is a central challenge in computational virology, with direct implications for public health surveillance and the early detection of emerging pathogens. Despite the potential of metagenomics to expose a vast viral diversity, most assembled contigs cannot be attributed to known viruses, a phenomenon known as viral dark matter [Krishnamurthy and Wang 2017]. Traditional alignment-based tools such as BLAST [Altschul et al. 1990] are inherently limited in this

context, which makes them inadequate for the discovery of novel or highly divergent pathogens that escape homology-based detection.

To circumvent this limitation, alignment-free machine learning methods have emerged as a prominent alternative. Although recent literature has favored deep learning architectures, these approaches usually benefit from dedicated hardware such as GPUs and demand substantial training time. By contrast, pipelines based on k-mer vector representations generated by Word2Vec and fastText, coupled to classical ensemble classifiers such as ExtraTrees and XGBoost, have achieved performance statistically equivalent to that of complex neural networks at a fraction of the computational cost [Silveira Júnior 2026]. A persistent gap in the evaluation of these models, however, is the reliance on internal cross-validation over previously curated databases. To assess the practical utility and the out-of-distribution generalization of an alignment-free model, it must be tested against an independent environmental dataset divergent from the training domain.

This paper presents an external validation case study of the ExtraTrees and XGBoost classifiers under domain shift and strong class imbalance. No adaptation or tuning to the target domain was performed: the models are trained on the ZOVER repository and evaluated on the Amazon data, which never enters training. For this validation we use a dataset originating from a recent metagenomic surveillance study in the Amazon Basin, which introduces 82 newly described RNA viral sequences from highly divergent families such as *Phasmaviridae* and *Phenuiviridae*. The models are challenged to recover these lineages, absent from the training domain, embedded in thousands of host transcripts, which reproduces the rarity of the viral class in a real sequencing scenario. Because the original annotation of the dataset was established by state-of-the-art homology tools, we use that curation as an external reference, independent of our pipeline. By submitting the dataset to the inference stage, we evaluate both the sensitivity in detecting divergent viruses and the specificity against eukaryotic background noise, conditions that determine the practical viability of alignment-free classifiers in real viral surveillance.

2. Related Work

The computational identification of viral sequences in metagenomes has driven the development of four major paradigms, each with distinct implications for the out-of-distribution generalization investigated in this work.

2.1. Tools based on deep learning over raw sequences

DeepVirFinder [Ren et al. 2020] introduced a CNN trained on one-hot encoding of nucleotide sequences, without dependence on gene annotation. ViraMiner [Tampuu et al. 2019] extended this approach with a dual-branch CNN architecture aimed at human metagenomes, while Seeker [Auslander et al. 2020] showed that a minimalist LSTM, with far fewer parameters than typical deep architectures, can match larger models for bacteriophage identification. A common denominator across these tools is their training distribution: all were calibrated predominantly on DNA viruses of prokaryotic origin or on viromes from well-characterized clinical and laboratory environments. Exposure to highly divergent negative-sense RNA viruses, belonging to families such as *Phasmaviridae* and *Phenuiviridae*, collected in a remote wild environment, represents a domain shift

scenario for which none of these tools has been explicitly validated, regardless of the length of the submitted sequences.

2.2. Tools based on functional annotation and hybrid approaches

VirSorter2 [Guo et al. 2021] uses multiple Random Forest classifiers fed by features derived from databases of viral marker genes, achieving robust performance when the target sequences show detectable homology to the reference profiles. VIBRANT [Kieft et al. 2020] combines Hidden Markov Model annotation with a multilayer perceptron, reporting near-perfect metrics on curated benchmarks. geNomad [Camargo et al. 2024], the current state-of-the-art reference, represents a hybrid approach, since it combines a neural sequence encoder, based on the IGLOO architecture, with a tree-based classifier trained on marker protein profiles. It thereby inherits the limitations of both paradigms, the dependence on the composition of the training corpus of neural architectures and the dependence on detectable homology of annotation-based tools. The structural limitation of these tools, in the scenario investigated here, is not the length of the sequences, which range from RdRp-bearing segments to nearly complete genomes and carry detectable marker genes [Fuques et al. 2026], but rather protein divergence. Viruses sufficiently divergent from the reference profile databases escape homology-based detection regardless of genome length, a pattern documented in independent evaluations of these tools on real environmental data [Wu et al. 2024]. In this scenario, the intrinsic precision of these tools does not translate into predictive coverage.

2.3. Alignment-free approaches based on k-mer embeddings

ProtVec [Asgari and Mofrad 2015] established the conceptual basis by applying Skip-gram Word2Vec to protein 3-mers, reaching 93% accuracy in protein family classification over 7,027 families with a support vector machine. dna2vec [Ng 2017] extended this approach to variable-length DNA k-mers, showing that vector operations over k-mer embeddings preserve biologically meaningful relationships. Virus2Vec [Ali et al. 2023] applied minimizer-based representations coupled to classical classifiers for viral host prediction, reinforcing the viability of lightweight sequence representation approaches.

Consolidating this line for the task of binary virus identification in metagenomes, a prior study [Silveira Júnior 2026] conducted a systematic comparison of 192 experimental configurations, covering nine classical classifiers, seven neural architectures, six values of k and two embedding models, over a dataset derived from the ZOVER repository [Zhou et al. 2022], evaluated under nested stratified cross-validation with paired significance tests. Under this protocol, the ensemble classifiers ExtraTrees, with an F1 of 0.894, and XGBoost, with an F1 of 0.891, achieved performance statistically equivalent to that of the best neural architecture evaluated, the CNN-Transformer, with an F1 of 0.880, with no McNemar test rejecting the null hypothesis of equivalence, and at a training cost roughly twenty times lower, without GPU dependency. That study acknowledges, among its limitations, that generalization to environmental metagenomes of distinct composition is not guaranteed, and points to validation on real samples from diverse environments as a priority direction for future work. The present article responds directly to that gap.

2.4. External validation and benchmarking on real data

Wu et al. [Wu et al. 2024] conducted the most comprehensive public benchmark of viral identification tools on real metagenomic data to date, evaluating multiple tools on samples

from different biomes and showing that the performance reported in controlled settings often does not hold on environmental data. That study establishes empirically that evaluation by internal cross-validation is insufficient to estimate the practical utility of viral classifiers, a systematic problem that remains poorly addressed in the literature. The analysis of Wu et al., however, did not include alignment-free approaches based on k-mer embeddings. To the best of our knowledge, no existing benchmark evaluates classical ensemble classifiers, applied without retraining, against a divergent wild RNA virome under this degree of class imbalance. This is the scenario investigated in this work.

3. Methodology

3.1. External validation dataset

The dataset used for validation is derived from the study of Fuques et al. [Fuques et al. 2026], conducted in the Juruá River region, Western Amazon Basin, Brazil. The 82 viral sequences that make up the positive class were identified and annotated by the original study through BLASTX search against the viral RefSeq database with a minimum bitscore threshold of 300 and 70% identity, completeness validation via CheckV [Nayfach et al. 2021], as applied by [Fuques et al. 2026], and phylogenetic analysis based on the RNA-dependent RNA polymerase (RdRp). These sequences span 10 viral families, with a predominance of *Phasmaviridae* and *Rhabdoviridae*, and range in completeness from partial segments bearing the RdRp to complete or nearly complete genomes, with the longer ones exceeding 10 Kb.

The divergence of these sequences relative to the training domain of the models is a structural feature of the evaluated scenario. The ZOVER repository [Zhou et al. 2022], used in training, focuses on zoonotic and arthropod-vector viruses with a predominant representation of known DNA and RNA lineages, whereas the most divergent families in the Amazonian set, in particular *Phasmaviridae* and *Phenuiviridae*, are composed of insect-specific negative-sense RNA viruses, with protein divergence often below 80% relative to the closest reference sequences [Fuques et al. 2026], which places them below the ICTV species demarcation thresholds and configures a verifiable domain shift scenario.

To rule out leakage from redundant sequences between training and validation, we compared the 82 Amazonian viral sequences against all ZOVER viral sequences used in training, with MMseqs2 [Steinegger and Söding 2017] in nucleotide search mode. No Amazonian sequence showed a high-identity match along its length to any training sequence. The highest similarity observed over a substantial fraction of the length was about 72% nucleotide identity, restricted to a small group of sequences, while most viruses showed no relevant alignment. The highest identity of any hit, measured at minimal query coverage, was 89.1%. Since this value is well below the threshold that would indicate identical or redundant sequences, we confirm the absence of duplicates between the sets. The recall measured below therefore reflects generalization to divergent viruses, and not the recovery of sequences previously seen in training.

The negative class represents the non-viral material that predominates in a mosquito metatranscriptomics experiment. Since the study of Fuques et al. [Fuques et al. 2026] sequenced RNA and recovered RNA viruses, the relevant biological background for screening is not raw genomic DNA, but rather the transcripts of the

host and its symbionts. For this reason the negative class was built from reference transcripts deposited in NCBI, and not from the raw reads of the original study, which could contain still unannotated viruses and contaminate the negative class with viral material.

We gathered 7,542 distinct transcripts, each used a single time, without fragmentation or overlap among them. The composition prioritizes the host of the study and reflects the diversity expected in a real sample. There are 2,950 mRNA transcripts of *Anopheles darlingi*, the host species of the original virome, 1,476 mRNA transcripts of other species of the genus *Anopheles*, 988 of *Culex* and 953 of *Aedes aegypti*, which are other relevant vectors, 960 gene sequences of *Wolbachia*, a symbiotic bacterium common in mosquitoes, and 215 ribosomal RNA sequences of Culicidae. The final set submitted to inference comprises 7,624 sequences, of which 82 are viral and 7,542 non-viral, which corresponds to an imbalance ratio of approximately 1 to 92, close to the viral prevalence observed in real metagenomic samples.

3.2. Preprocessing of input sequences

The sequences were submitted to inference in the original FASTA format, without assembly, trimming or additional normalization, so as to reproduce the scenario in which the models would receive contigs directly from an assembly pipeline. The only filtering applied to the negative class was the removal of transcripts outside the 500 to 25,000 bp range, which kept the host sequences in a length range comparable to that of the evaluated viral genomes. The k-mer tokenization ignores positions with ambiguous bases that do not form a valid k-mer, without any other character substitution.

3.3. Models and inference settings

The models used are the ExtraTrees classifier [Geurts et al. 2006], coupled to fastText embeddings, and XGBoost [Chen and Guestrin 2016], coupled to Word2Vec embeddings, both with $k = 6$. These two combinations were selected in [Silveira Júnior 2026] as the best classical configurations among 192 experimental configurations evaluated over the ZOVER repository [Zhou et al. 2022] under nested stratified cross-validation with paired statistical tests. Following that configuration, the models are trained on ZOVER and applied to the Amazon data without any retraining or tuning for the Amazonian domain, which is used only for evaluation.

Each sequence is tokenized into overlapping k-mers and represented by the arithmetic mean of its k-mer embeddings, generated by Word2Vec [Mikolov et al. 2013] for XGBoost and fastText [Bojanowski et al. 2017] for ExtraTrees, following the embedding configuration of [Silveira Júnior 2026]. This yields a fixed-length dense vector regardless of the sequence length, and the classifier issues a binary prediction for every input sequence, without length restriction or dependence on gene annotation.

The default decision threshold of 0.5 over the posterior probability was kept for both models, without any adjustment to the validation domain. Embedding training is not deterministic, as it depends on random initialization. So that the results do not depend on a single favorable run, the whole procedure of embedding generation and classifier training was repeated twenty times, and the results are reported as mean and standard deviation across the repetitions. Each repetition is trained on an 80% partition of the ZOVER data with a fixed random seed (42) for the partition and the classifiers; only the embedding training was left non-deterministic, as it is the dominant source of variability.

3.4. Computational environment

All experiments were run on a workstation with a 14-core processor providing 20 logical threads and 64 GB of RAM, without GPU use. Inference, including embedding generation and classification, runs entirely on CPU, which is one of the central arguments of this work. The pipeline was implemented in Python 3.13 with scikit-learn 1.7.2, XGBoost 3.1.3 and gensim 4.4.0. The independence from marker-gene and protein-profile databases distinguishes the approach from the annotation tools evaluated in the comparison.

The three reference tools were executed on CPU, on the same workstation and over the same sequences as the proposed classifiers, each with all 20 threads. geNomad 1.11.2 was run with `end-to-end --cleanup --enable-score-calibration` against database version 1.9, and a sequence was taken as viral when reported in the virus summary, which is the tool's own calibrated decision, with no additional cutoff. VirSorter2 2.2.4 was run against the database installed by `virsorter setup`, with `--include-groups dsDNAphage,ssDNA,NCLDV,RNA,lavidaviridae`, `--min-score 0.5` and `--min-length 0`, and a sequence was taken as viral when reported in `final-viral-score.tsv`. DeepVirFinder was run from commit 2635ec8, since the tool publishes no versioned release, with its bundled pretrained models and `-l 1`, and a sequence was taken as viral when the score was at least 0.5 with $p < 0.05$. The comparison is conservative with respect to parallelism: the reference tools were given all 20 threads, whereas the dominant cost of the proposed pipeline, the vectorization of each sequence into mean-pooled k-mer embeddings, runs on a single core.

To support reproducibility, the inference scripts, the exact command lines used for the reference tools, the GenBank accessions of the 82 validation viruses and the NCBI accessions of all host transcripts used in the negative class are available at `https://github.com/HiagoMuniz/alignment-free-amazon-validation`.

3.5. Evaluation metrics

Performance was evaluated with metrics suited to the strong class imbalance of the scenario. Sensitivity, or recall, quantifies the fraction of the 82 viral sequences correctly recovered and is the primary metric of the study, given the goal of detecting viral dark matter in a real environment. Specificity quantifies the fraction of host sequences correctly rejected, and is indicative of robustness against eukaryotic metagenomic noise. Precision measures the fraction of positive predictions corresponding to real viruses. The F1-score provides the harmonic mean of precision and sensitivity, and the Matthews correlation coefficient, or MCC, summarizes accuracy considering the four cells of the confusion matrix [Rainio et al. 2024]. To compare the methods with one another, we computed 95% Wilson confidence intervals for recall and specificity, and applied the paired McNemar test between pairs of methods over the same sequences, adopting a significance level of 0.05. For the proposed models, whose predictions vary across the 20 embedding runs, the McNemar test against each external tool was applied per run, and the resulting p-values are summarized by their median and range.

We also report two threshold-independent metrics. The area under the ROC curve measures the ability of the model to rank viral sequences above non-viral ones and does

not depend on the proportion between classes. The area under the precision-recall curve, PR-AUC, summarizes the trade-off between precision and recall under the real imbalance of the set and is particularly informative in scenarios with a rare positive class, as is ours. Since recall and specificity do not depend on the proportion between classes, while precision, F1 and MCC do, we report these last three at the natural imbalance of the set and discuss their behavior under different prevalences.

4. Results

4.1. Overall classifier performance

Table 1 presents the performance of the two classifiers at the default decision threshold of 0.5. Recall, specificity and the areas under the ROC and precision-recall curves are the central metrics, as they do not depend on class prevalence; precision and F1 are addressed below and in the Discussion.

Table 1. Classifier performance on the out-of-distribution Amazonian validation set, with 82 viral and 7,542 non-viral sequences. Values are mean ± standard deviation over 20 independent embedding repetitions.

Model	Recall	Specificity	ROC-AUC	PR-AUC
ExtraTrees	$84.6 \pm 1.1\%$	$88.7 \pm 1.2\%$	0.948 ± 0.005	0.507 ± 0.026
XGBoost	$82.3 \pm 1.3\%$	$87.8 \pm 0.9\%$	0.949 ± 0.005	0.547 ± 0.025

Both models recovered over 82% of the viral sequences and reached a ROC-AUC close to 0.95, which indicates a good ability to rank viral sequences above non-viral ones regardless of the adopted threshold. The PR-AUC was 0.507 for ExtraTrees and 0.547 for XGBoost. Under the positive prevalence of about 1.1% in this set, a random classifier would have a PR-AUC equal to that prevalence, so the obtained values are about 47 to 51 times higher than chance, which confirms the presence of discriminative signal even under strong imbalance. At the natural imbalance of the set, of approximately 1 virus per 92 hosts, the prevalence-dependent metrics are low for both models, with precision of 7.5% and 6.9%, F1 of 0.138 and 0.126, and MCC of 0.232 and 0.216 for ExtraTrees and XGBoost, respectively. These are values expected when the positive class is rare and are analyzed in Section 5. Inference over the 7,624 sequences, including embedding generation, was completed in about thirteen seconds on CPU for each model, without GPU use.

4.2. Performance by viral family

Table 2 stratifies the sensitivity of the classifiers by viral family, allowing the identification of error patterns associated with specific taxonomic groups.

Sensitivity varies markedly across families. The two most numerous, *Phasmaviridae* and *Rhabdoviridae*, were recovered stably and almost completely by both models, close to 97%. *Phenuiviridae*, *Xinmoviridae* and *Peribunyaviridae* were also fully recovered. At the opposite end, *Mesoniviridae* was not detected in any run by either model, and *Togaviridae* was recovered in about one fifth of the cases at most. *Iflaviridae* and *Flaviviridae* occupied an intermediate position. Eight of the ten families have $n \leq 6$, and *Peribunyaviridae* is represented by a single sequence. In these families a single sequence shifts the sensitivity by at least 16.7 percentage points, so the differences between

Table 2. Sensitivity by viral family ($n = 82$ sequences in total). Estimates for families with $n \leq 6$ are unstable and should be read as indicative; see the caveat in the text.

Family	n	ExtraTrees	XGBoost
Phasmaviridae	35	97.3%	98.3%
Rhabdoviridae	16	96.6%	97.5%
Iflaviridae	6	85.8%	59.2%
Togaviridae	5	2.0%	21.0%
Flaviviridae	4	66.2%	61.3%
Mesoniviridae	4	0.0%	0.0%
Totiviridae	4	100.0%	61.3%
Phenuiviridae	5	100.0%	100.0%
Xinmoviridae	2	100.0%	100.0%
Peribunyaviridae	1	100.0%	100.0%

the two classifiers and the values of 100% should be read as indicative rather than as reliable estimates. Only *Phasmaviridae* ($n = 35$) and *Rhabdoviridae* ($n = 16$) support a firmer reading. The central point is that the model error is not uniformly distributed, but concentrates in specific families that are absent or poorly represented in the training domain.

4.3. Comparison with reference tools

Table 3 compares the proposed classifiers with three widely used tools, run over the same 82 viral sequences and the same 7,542 host sequences. Recall is measured over the viral sequences and specificity over the host sequences.

Table 3. Comparison of the proposed classifiers with reference tools on the same Amazonian set. Recall measured on 82 viral sequences and specificity on 7,542 host sequences.

Method	Recall	Specificity
geNomad	89.0%	97.3%
DeepVirFinder	85.4%	79.6%
ExtraTrees (proposed)	84.6%	88.7%
XGBoost (proposed)	82.3%	87.8%
VirSorter2	81.7%	97.2%

In recall, the five methods occupy a narrow band, between 81.7% and 89.0%. For the three external tools, which are deterministic, the 95% Wilson confidence intervals are wide and overlap, geNomad at [80.4%, 94.1%], DeepVirFinder at [76.1%, 91.4%] and VirSorter2 at [72.0%, 88.6%], and the paired McNemar test among them points to a significant difference only for the VirSorter2 and geNomad pair ($p = 0.03$). For the proposed classifiers, whose predictions vary across embedding runs, we applied the paired McNemar test between each of the 20 runs of each model and each external tool, over the same 82 sequences. In none of these comparisons was the recall difference significant, with median p-values ranging from 0.24 to 1.00 and no individual run reaching $p < 0.05$. We therefore find no statistical evidence that the proposed classifiers differ in recall from the reference tools, although the small size of the positive class limits the power of

this comparison. In specificity, measured over 7,542 sequences, the difference is clear and the Wilson interval of each external tool does not reach the value of the proposed classifiers. geNomad and VirSorter2 reach about 97%, the proposed classifiers about 88%, and DeepVirFinder about 80%. This contrast outlines the central trade-off of the study, analyzed in Section 5: annotation tools reject host noise more strictly, while the proposed classifiers reach similar recall, classify every sequence without depending on gene annotation, and are much faster.

5. Discussion

The results should be interpreted in light of the adverse conditions of the validation scenario. The models were applied without any adaptation to a virome composed of insect-specific RNA viruses from families underrepresented in the ZOVER repository [Zhou et al. 2022], embedded in a large volume of host sequences. Even in this scenario, the classifiers recovered most of the viral sequences, with a ROC-AUC close to 0.95. The high ROC-AUC shows that the separation between virus and host remains sharp in terms of ranking, which indicates that the k-mer embeddings preserve useful discriminative signal even under domain shift. The drop in precision and F1 at the natural imbalance does not reflect a loss of that ranking ability, but rather the rarity of the positive class, as discussed below.

The per-family analysis shows that the error is not uniformly distributed. This contrast, between families recovered almost completely and families missed entirely, is consistent with the hypothesis that k-mer embeddings capture signals of global nucleotide composition, such as codon usage bias and dinucleotide patterns, which distinguish viral genomes from host genomes [Silveira Júnior 2026], but become insufficient for families whose compositional profile is absent from the training domain. The observed limit is therefore one of training set coverage, and not of the method itself, which points to taxonomic expansion of the base as a direct path to improvement. All three reference tools recovered the four *Mesoniviridae* sequences that both proposed models missed, which confirms that the lineage is detectable and the failure lies in training coverage.

The comparison with the reference tools reveals a clear trade-off. geNomad and VirSorter2 reject host noise more strictly. In recall, however, we find no evidence of a difference between the methods, and the proposed classifiers still classify every input sequence, without discarding those without detectable homology, and run on CPU in seconds. This pattern is consistent with that documented by Wu et al. [Wu et al. 2024], who showed systematic degradation of annotation-dependent tools on real environmental data, and extends the observation to a highly divergent wild RNA virome. The difference in cost is substantial. On the same workstation, each of the proposed classifiers processed the full set of 7,624 sequences in about thirteen seconds on CPU, and the two together in under thirty seconds, since they do not share an embedding. The reference tools, run over the same sequences in two batches, which counts each tool's fixed startup cost twice, took about 3 minutes (geNomad), 4 minutes (DeepVirFinder) and 49 minutes (VirSorter2). The alignment-free classifiers thus position themselves as a fast and comprehensive screening stage, complementary to confirmation by annotation tools.

The specificity of approximately 88% deserves attention. It means that between 11 and 12% of the host sequences are classified as viral, which, under the real prevalence

of about 1 virus per 92 hosts, generates a high absolute number of false positives and explains the low precision observed. In absolute terms, averaged over the 20 repetitions, ExtraTrees recovers 69 of the 82 viral sequences and misclassifies 854 of the 7,542 host transcripts, while XGBoost recovers 68 and misclassifies 918. Each run therefore returns fewer than a thousand candidate sequences for downstream inspection, rather than the thousands the percentages alone might suggest. In large-scale metagenomic screening, where each false positive implies an investigation cost, this behavior reinforces the role of the classifiers as an initial filter, and not as a standalone final decision. This same collapse of precision under high negative prevalence affects all evaluated tools, and is not specific to the proposed approach.

Some limitations should be considered. The validation set, although biologically diverse and methodologically independent, is restricted to mosquito RNA viruses from a single biome, which prevents generalizing the behavior of the models to viromes of other hosts or environments, and contains only 82 viral sequences, which makes the recall confidence intervals wide and limits the power of the comparisons between methods. In addition, both the viral and the host sequences are complete genomes and transcripts, whereas a real screening scenario operates over fragmented contigs of variable length, so the performance measured here tends to be optimistic relative to practical use. The negative class, in turn, was built from NCBI reference transcripts, and not from the host contigs of the original study, a choice that avoids contamination by unannotated viruses but makes specificity an estimate over a background curated by us, whose composition, in particular the ribosomal RNA fraction, does not exactly reproduce that of a real RNA library. It is also worth noting that we validated only the classical classifiers, and not the neural architectures reported as equivalent in a controlled setting [Silveira Júnior 2026], so the out-of-distribution generalization of those networks remains open. A further caveat concerns the ground truth itself. The positive set was defined by the original study through homology search [Fuques et al. 2026], so it is biased toward viruses that retain detectable homology and, by construction, under-represents the most divergent lineages that escape homology entirely. Our recall is therefore measured against a homology-defined reference, and the hardest cases for any method, viruses with no detectable similarity to known sequences, are absent from the evaluation set. One host transcript could not be scored by DeepVirFinder, since a third of its bases are ambiguous, and was counted as correctly rejected; scoring it as a false positive would change that tool's specificity by 0.01 percentage points.

Regarding the independence between training and validation, four *Flaviviridae* sequences have a divergent homolog, with about 72% nucleotide identity, in the training set. Removing these four sequences does not reduce the recall of the models beyond their run-to-run variation, which indicates that the reported performance is not driven by them. Finally, the non-deterministic training of the embeddings introduces variation of a few percentage points in the out-of-distribution recall between runs, handled by reporting mean and standard deviation, and the ZOVER repository [Zhou et al. 2022] underrepresents wild insect-specific families predominant in the Amazonian set, which helps explain the failures observed in *Mesoniviridae* and *Togaviridae*.

6. Conclusion

This work presented an out-of-distribution validation case study of alignment-free classifiers based on k-mer embeddings, submitting the ExtraTrees and XGBoost models, trained on the ZOVER repository without any reconfiguration, to a mosquito RNA virome from the Amazon Basin composed of highly divergent viral lineages absent from the training domain, embedded in host sequences under strong class imbalance.

ExtraTrees recovered on average 84.6% of the viral sequences and XGBoost 82.3%, with specificity around 88% and ROC-AUC close to 0.95, each processing all 7,624 sequences in about thirteen seconds on CPU, without gene annotation. In recall, the two models perform similarly to the three reference tools evaluated, with no detectable statistical difference, although with specificity lower than that of geNomad and VirSorter2. These results indicate that the competitiveness observed for classical classifiers in a controlled setting [Silveira Júnior 2026] holds, in recall, under real domain shift, positioning these models as a fast and comprehensive alternative for the initial screening of viruses in metagenomes, especially where computational resources or reference annotation are limited. This evidence comes from a single external dataset from one biome. It is therefore sufficient to show that the discriminative signal survives a real domain shift, but not to establish general practical utility across viromes.

The per-family analysis exposed the main limit of the approach. Performance is high and stable in families represented in training, but drops to zero in *Mesoniviridae* and stays very low in *Togaviridae*, both underrepresented in ZOVER. As future work, the expansion of the training base with NCBI GenBank sequences of broader taxonomic coverage is the natural step to reduce these blind spots. The replication of this protocol on viromes of other biomes [Wu et al. 2024] and the evaluation of the models on more than one external collection are direct extensions that would strengthen the conclusions presented here.

References

Ali, S. et al. (2023). Virus2Vec: viral sequence classification using machine learning. In *Proceedings of the Machine Learning for Healthcare Conference (MLHC)*, volume 209 of *PMLR*, pages 14–29.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Asgari, E. and Mofrad, M. R. K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLOS ONE*, 10(11):e0141287.

Auslander, N. et al. (2020). Seeker: alignment-free identification of bacteriophage genomes by deep learning. *Nucleic Acids Research*, 48(21):e121.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Camargo, A. P. et al. (2024). Identification of mobile genetic elements with geNomad. *Nature Biotechnology*, 42(8):1303–1312.

Chen, T. and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York. ACM.

Fuques, E., Massey, A. L., Qureshi, F., Campos-Silva, J. V., Ferreira da Silva, D. J., Peres, C. A., Levi, T., and Vega Thurber, R. L. (2026). Large-scale metagenomic surveillance study expands the known diversity of RNA viruses in mosquito populations from the Amazon Basin. *PeerJ*, 14:e20880.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Guo, J. et al. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*, 9(1):37.

Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, 8(1):90.

Krishnamurthy, S. R. and Wang, D. (2017). Origins and challenges of viral dark matter. *Virus Research*, 239:136–142.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

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

Ng, P. (2017). dna2vec: consistent vector representations of variable-length k-mers. arXiv:1701.06279.

Rainio, O., Teuho, J., and Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14:6086.

Ren, J. et al. (2020). Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*, 8(1):64–77.

Silveira Júnior, C. A. C. (2026). Identificação de sequências virais em dados metagenômicos: uma abordagem de alto desempenho utilizando embeddings. Master's thesis, Universidade Federal de Pelotas, Pelotas. Available at: `https://guaiaca.ufpel.edu.br/handle/prefix/20331`.

Steinegger, M. and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028.

Tampuu, A. et al. (2019). ViraMiner: deep learning on raw DNA sequences for identifying viral genomes in human samples. *PLOS ONE*, 14(9):e0222271.

Wu, L.-Y., Wijesekara, Y., Piedade, G. J., Pappas, N., Brussaard, C. P. D., and Dutilh, B. E. (2024). Benchmarking bioinformatic virus identification tools using real-world metagenomic data across biomes. *Genome Biology*, 25(1):97.

Zhou, S. et al. (2022). ZOVER: the database of zoonotic and vector-borne viruses. *Nucleic Acids Research*, 50(D1):D943–D949.