Title: vir2vec: A Viral Genome-Wide Viral Embedding

Author: Simone Rancati[1, 2, 3, †], Pablo Arozarena Donelli[1, †], Giovanna Nicora[1], Laura Bergomi[1], Tommaso Mario Buonocore[1], Micheal Aaron Sy[2], Sakshi Pandey[4], Mattia Prosperi[2], Marco Salemi[3, 5], Riccardo Bellazzi[1], Christina Boucher[4], Enea Parimbelli[1, *], Simone Marini[2,3, *]

[1]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Adolfo Ferrata 5, Pavia, 27100, Italy
[2]Department of Epidemiology, College of Public Health and Health Professions, University of Florida, 2004 Mowry Road, Gainesville, FL 32610, United States
[3]Emerging Pathogens Institute, University of Florida, 2055 Mowry Road, Gainesville, FL 32610, United States
[4]Department of Computer and Information Science and Engineering, University of Florida, 432 Newell Dr, Gainesville, FL 32611, United States
[5]Department of Pathology, Immunology and Laboratory Medicine, College of Medicine, University of Florida, 1600 SW Archer Road, Gainesville, FL 32610, United States
[†]**Equal Contribution**
[*]**Corresponding Authors**

Abstract: Genomic language models (gLMs) have recently emerged as powerful numerical surrogates for DNA, but existing architectures are largely focused on human DNA or trained on limited viral references, and no dedicated benchmark currently exists for viral genome understanding. Here we introduce vir2vec, a 422-million-parameter, decoder-only gLM obtained by continual pretraining of Mistral-DNA on a rigorously curated pan-viral corpus of 565,747 complete genomes spanning 295 viral species. vir2vec operates on byte-pair–encoded DNA subwords and exposes fixed-length genome-level embeddings that are reused across tasks. Additionally, we present vGUE, a unified benchmark for viral representation learning. In vGUE, we precompute vir2vec embeddings and feed them to simple classifiers (logistic regression, support vector machines, random forests) trained under nested cross-validation, to quantify how well they capture biologically motivated axes of viral variation. Using this framework, vGUE assesses genomic viral prediction tasks across: (i) organism-level discrimination (virus vs non-virus genomes and reads), (ii) genome-wide evolutionary fingerprints (DNA vs RNA viruses, host-range prediction), (iii) intra-genus species separation (HIV-1 vs HIV-2), (iv) fine-grained variant and subtype typing (SARS-CoV-2 lineages), and (v) phenotypic context signal detection (HIV-1 brain vs plasma Tropism). vir2vec attains the highest balanced accuracy across seven out of eight heterogenous classification tasks consistently outperforming both a human-trained genomic foundation model and a viral-specific one. By coupling a pan-viral gLM with a standardized evaluation suite, vir2vec and vGUE provide an open foundation for future viral genomic models, surveillance tools, and discovery pipelines.

BACKGROUND

Viruses remain among the most pervasive threats to global health. For example, seasonal influenza causes around one billion infections and 3–5 million severe cases every year, leading to an estimated 290,000–650,000 respiratory deaths worldwide [1]; and dengue is endemic in more than 100 countries, with roughly half of the world's population now living in at-risk areas, and an estimated 100–400 million infections annually [2]. Beyond headline pathogens, metagenomic surveys of environmental and human-associated microbiomes continue to uncover vast swath of viral diversity that has no close counterpart in current reference databases. Studies of marine and

gut viromes, for instance, find that only a minority of viral reads can be confidently assigned to known taxa, highlighting how incomplete our picture of the global virosphere still is [3, 4, 5].

The COVID-19 pandemic dramatically accelerated genomic surveillance, turning high-throughput sequencing into a routine instrument of public health. However, this success has also exposed the limitations of the computational infrastructure sitting downstream of the sequencers. Traditional bioinformatics pipeline relies heavily on alignment to reference genomes or profile-based homology search, which are computationally intensive and intrinsically constrained by database coverage. Benchmarking efforts show that alignment-based methods and machine-learning methods perform well on near-reference sequences but degrade as genomes diverge or when working with short, noisy contigs in complex metagenomes [6, 7]. This is particularly problematic for viruses: reference collections are taxonomically biased, many clades remain sparsely sampled or entirely unrepresented, and viral taxonomy itself is in rapid flux. As a result, a substantial fraction of viral signal in clinical and environmental samples remains unclassified, limiting timely pathogen discovery and hampering quantitative ecological and epidemiological analyses [8, 9].

In parallel, advances in deep learning have begun to reshape computational genomics. Inspired by successes in natural language processing, several groups have proposed treating DNA as a sequence-based language and trained transformer-based foundation models that learn generic representations directly from nucleotide strings. A variety of architectures have emerged, ranging from decoder-only models capable of generating synthetic sequences [10, 11] to encoder-only models that produce numerical representations while preserving biologically meaningful signal. Among the encoder-based approaches, the Nucleotide Transformer (NT) family scales encoder-only models up to 2.5 billion parameters, pre-trained on human genomes supplemented with hundreds of additional species and achieves state-of-the-art performance across diverse genomics benchmarks [12]. DNABERT-2 similarly introduces a more efficient tokenizer and architecture than earlier genomic language models (gLMs), replacing fixed *k*-mer vocabularies with byte-pair encoding and enabling multi-species training at scale [13]. More recently, encoder–decoder architectures such as ENBED have demonstrated that long-context, byte-level transformers can be adapted to sequence-to-sequence tasks in genomics, further expanding the design space for genomic foundation models [14].

Despite these advances, existing gLMs are still only partially aligned with the needs of computational virology: the main architectures are not optimized for viral genomes, as training data are typically multi-species across kingdoms or primarily human-centric. For example, Mistral-DNA-v1-422M-hg38 Genomic Language Model (gLM) is a 422M-parameter generative DNA model derived from the Mixtral-8x7B-v0.1 architecture [15], simplified for genomic applications by reducing the number of layers and the hidden size, and pretrained exclusively on 10 kb windows from the hg38 human genome assembly [16]. In addition, most published genomic foundation models are encoder-only architectures fine-tuned for specific classification or regulatory prediction tasks. As an example of such viral-focused efforts, the ModernBert-DNA-v1-37M-virus LLM is a 37M-parameter DNA model derived from ModernBERT [18], similarly simplified by reducing depth and hidden dimensionality, and pretrained on approximately 15,071 viral genomes longer than 1 kb, split into 1 kb segments [17]. While valuable, these models fall short of providing a foundational viral genomic model that serves as a robust base for fine-tuning across real-world applications, as they remain limited in size and in their ability to capture natural genomic variation, evolutionary fingerprints, and genome-wide features along full viral genomes. Recently, a handful of new models begun to develop the field of viral gLMs. ViraLM, for example, fine-tunes a DNABERT-2 backbone [9, 13, 16] on ~50,000 viral genomes and shows improved virus-versus-non-virus classification in metagenomic contigs compared with classical approaches [19]. Other recent work has explored the use of large language models for hierarchical viral taxonomy

prediction [20]. While promising, these approaches are typically optimized for a single downstream task or a narrow taxonomic slice and do not yet provide a general, reusable embedding space that is explicitly designed to be multispecies, multi-host, genome-wide, and pan-viral. Therefore, a single, scalable embedding space that faithfully encodes many viral genomes across bacteria, archaea, plants, invertebrates and vertebrate hosts would provide a common numerical "language" on top of which heterogeneous tasks can be layered: virus vs non-virus detection in metagenomes; DNA vs RNA virus discrimination; host- or tissue-range prediction; lineage-level classification within high-priority pathogens such as SARS-CoV-2; and even more exploratory applications such as zoonotic risk estimation or vaccine antigen design. To be practically useful, such an embedding engine must (i) be trained on a carefully curated, taxonomically broad, quality-controlled corpus; (ii) remain robust to unbalanced species distributions and heterogeneous genome lengths; and (iii) expose representations that are both geometrically well-structured (for clustering and visualization) and predictive in downstream supervised tasks.

To fill this research gap, here we introduce vir2vec, a viral language model and embedding space developed to address the challenges detailed above (Figure 1). vir2vec is obtained by continual pre-training of a 422-million-parameter, Mistral-style decoder-only transformer on a curated pan-viral corpus of more than 500,000 complete genomes spanning 295 species with hundreds to thousands of genomes represented per species. Building on the DNABERT-2 tokenizer, vir2vec learns context-rich nucleotide representations over this diverse corpus and exposes fixed-length, genome-level vectors via max-pooling over token embeddings.

Despite the growing interest in gLMs, the field has lacked a standardized benchmark for evaluating their understanding of DNA sequences. Prior work spans a wide range of tasks and datasets, resulting in performance comparisons that are not always direct. With the Genome Understanding Evaluation (GUE), DNABERT-2 introduced the first effort to develop a standard benchmark for gLMs [13]. However, this benchmark covered only two virus-focused tasks, underscoring the lack of a standard evaluation framework for current and future viral gLMs. For this reason, we not only propose a new model, but also a unified benchmark for viral representation learning: the viral Genomic Understanding Evaluation (vGUE) [13]. We systematically evaluate vir2vec embeddings across a suite of downstream tasks that probe viral biology at multiple resolutions: (i) broad organism-level distinctions; (ii) genome-wide signatures and evolutionary fingerprint recognition; (iii) variant/subtype typing; (iv) intra-genus species distinction; and (v) phenotypic context signal identification.

By positioning vir2vec as a pan-viral, multispecies embedding engine—rather than a model tailored to a single dataset or endpoint—we aim to bridge the current gap between general genomic foundation models and bespoke viral classifiers. Our results show that a single decoder-only transformer, when exposed to a sufficiently broad and carefully curated viral corpus, can learn a coherent representation of viral "language" that supports both coarse-grained and lineage-level tasks, and that generalizes beyond the training distribution. In doing so, vir2vec provides a versatile foundation on which future computational virology pipelines, surveillance systems, and discovery tools can be built.

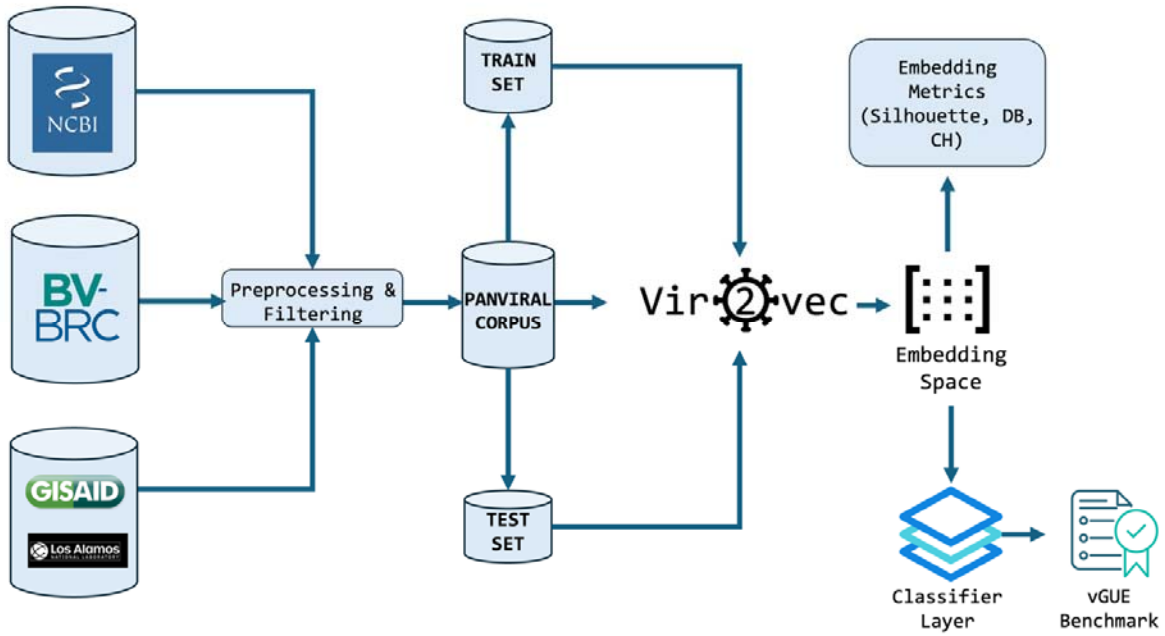

Figure 1. vir2vec. *Viral genomes collected from public repositories (NCBI, BV-BRC, and others) undergo preprocessing and are then split into a training set, used to build the vir2vec model, and a held-out test sets for evaluation. Primary evaluation relies on clustering metrics computed in the latent embedding space (Silhouette, Davies–Bouldin, and Calinski–Harabasz indices). These embeddings are subsequently fed into a classification layer, implemented as a carefully designed suite of machine-learning classifiers tailored to the downstream tasks defined by the vGUE benchmark.*

RESULTS

Overview of vir2vec architecture

vir2vec is architecturally a decoder, autoregressive model that can be used as a frozen encoder by extracting the hidden representations at each nucleotide position and aggregating them into a fixed-length vector. In practice, the transformer layers learn rich contextual embeddings as they predict the next token; max-pooling over these token-level states yields stable, genome-level representations that can be fed into shallow classifiers for diverse downstream tasks. We adopted this decoder-only design to make vir2vec compatible with both discriminative and generative use case. In its current form, it offers a versatile embedding space suitable for both classification and regression tasks. Looking ahead, the same backbone can be extended to support sequence generation, editing, and simulation of viral genomes—capabilities that address a notable gap in current viral gLMs, which predominantly rely on encoder-only architectures.

The architecture of vir2vec follows the Mistral/Mixtral family with mixture-of-experts feed-forward layers and sliding-window attention [15], and operates on DNA subword tokens learned via a BPE-style tokenizer adapted to genomic sequences. Each sequence is encoded into a fixed 4,096-dimensional embedding by max-pooling token representations along the genome, providing a common feature space for all downstream tasks. To specialize the model to panviral genomics, we performed continual pretraining on a rigorously curated pan-viral corpus of 565,747 complete genomes from 295 viral species. Genomes were aggregated from NCBI Virus [21] and BV-BRC

[22] and enriched with species-focused resources (GISAID [23], LANL-HIVdb [24], HBVdb [25]), filtered to retain only complete assemblies and high-quality genomes. The resulting dataset was split 70/30 at the species level into training and held-out sets, and the model was further trained with a standard causal language modeling objective, keeping the architecture fixed and updating only the weights to capture viral genome-wide constraints and natural variation.

PCA evaluation of the embedding space

To obtain a global, task-agnostic assessment of how vir2vec organizes viral diversity, we first analyzed its full 4,096-dimensional embeddings on the held-out pan-viral test set, comprising 30% of the corpus and stratified to preserve both the species composition and the relative proportion of genomes across 295 viral species. Using species labels only as a reference (i.e. treating each species as a putative cluster), we computed standard clustering quality metrics (**Figure 2**) in the original embedding space. The silhouette [26] coefficient was slightly negative (−0.0115), whereas the Davies–Bouldin [27] and Calinski–Harabasz [28] indices were 3.0241 and 814.3, respectively. This combination of a near-zero silhouette with a relatively low Davies–Bouldin and high Calinski–Harabasz score indicates that the model induces compact, partially separated species-level groups that slightly overlap in high-dimensional space. This is consistent with biological expectations: viruses frequently share conserved genes, replication strategies, and genome-wide constraints, so species are expected to lie along a continuum of similarity rather than forming perfectly isolated clusters.

To further characterize the geometry of the vir2vec embedding space, we visualized the 4,096-dimensional test-set embeddings with a two-dimensional principal component analysis (PCA) projection **(Figure 2A**). Each point corresponds to a single held-out genome, colored uniformly to avoid injecting label information into the visualization. As expected from the high intrinsic dimensionality, the first two principal components capture a modest fraction of the total variance (PC1 ≈ 44.2% and PC2 ≈ 8.0%), which is desirable because it indicates a nuanced embedding space where information is spread across many directions and likely reflects non-linear structure that cannot be reduced to a weighted sum of one or two components. The plot shows a clear structure with compact subregions and a pronounced elongation along the PC1 axis, consistent with broad evolutionary and genomic-composition differences across families. This is in line with the clustering indices computed in the original 4,096-dimensional space, where species form compact but partially overlapping groups rather than perfectly separated clusters.

To obtain a more interpretable view, we examined the same PCA space restricted to three clinically important respiratory viruses: Alphainfluenzavirus influenzae, Orthopneumovirus hominis and SARS-related coronavirus **(Figure 2B**). Even in two dimensions, these species occupy largely distinct, well-localized regions that can be separated with linear decision boundaries. This indicates that vir2vec already encodes species-discriminative information in its raw embeddings, without any task-specific fine-tuning, and suggests that simple linear classifiers on top of the frozen representation should be sufficient to support accurate downstream discrimination for at least a subset of virologically relevant organisms.

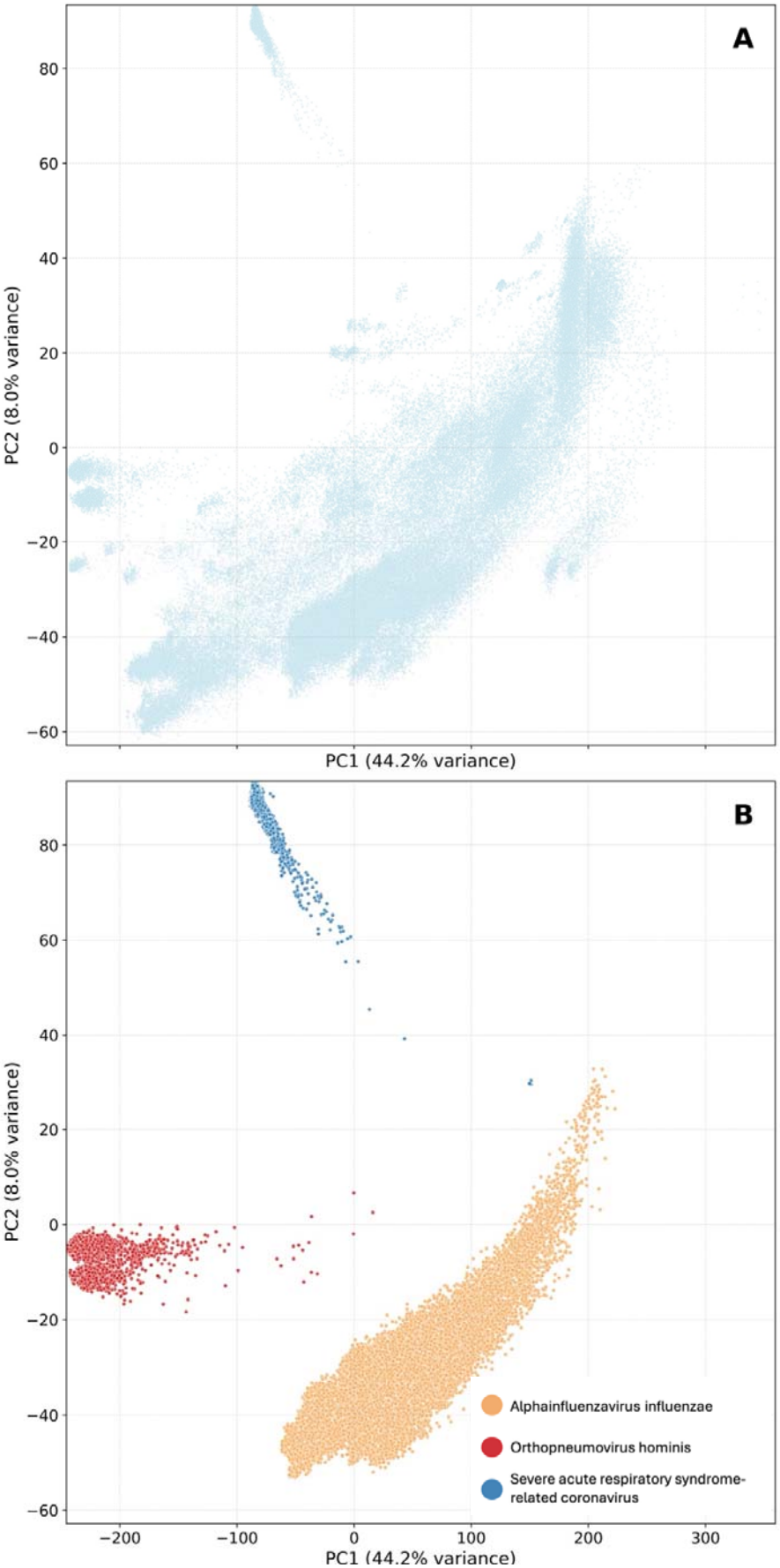
A
B
PC2 (8.0% variance)
PC1 (44.2% variance)
80
60
40
20
0
−20
−40
−60
−200
−100
0
100
200
300
Alphainfluenzavirus influenzae
Orthopneumovirus hominis
Severe acute respiratory syndrome-related coronavirus

Figure 2. Global and species-specific geometry of the vir2vec embedding space. *(A) Two-dimensional PCA projection of the 4,096-dimensional vir2vec embeddings for all 152,750 genomes in the held-out pan-viral test set. Each point is a single genome, shown in a uniform color; The distribution exhibits partially compact substructures and a broad gradient along PC1, consistent with a continuum of viral diversity rather than perfectly isolated species clusters. (B) Same PCA space, highlighting three clinically relevant respiratory viruses—Alphainfluenzavirus influenzae, Orthopneumovirus hominis and Severe acute respiratory syndrome-related coronavirus—in distinct colors. Even in this low-dimensional projection, the three species occupy largely non-overlapping regions that are approximately linearly separable, illustrating that the embeddings already encode useful species-level signal for downstream classification.*

Description of the vGUE Benchmark

To evaluate how well vir2vec embeddings support diverse viral inference problems, we applied the newly proposed vGUE benchmark introduced above. vGUE groups downstream tasks into biologically motivated categories—broad organism-level distinctions, genome-wide signature and evolutionary fingerprint recognition, intra-genus species distinction, variant/subtype typing, and phenotypic context signal identification—and instantiates each category with one or more representative use cases (**Table 1**). Across these five categories, vGUE comprises seven representative tasks which span full genomes and short reads, DNA and RNA viruses, diverse host taxa, and both taxonomic and phenotypic endpoints, providing a broad and comprehensive representation learning benchmark specific for viral gLMs. For every task, we used vir2vec strictly as a frozen encoder, extracted one 4,096-dimensional embedding per genome or read, and trained shallow classifiers under nested cross-validation, as described in the Methods.

Category	**Task**	**Purpose**	**Data**
Broad organism-level distinctions	Virus vs non-virus classification	Evaluate whether the model's embeddings capture the fundamental features across full genomes that differentiate viral from cellular life.	Viral full genomes from corpus test set and 9,600 bacterial complete genomes from RefSeq
	Virus vs non-virus reads classification	Tests the model's ability to recognize viral sequence signatures in a very limited context.	Synthetic NovaSeq reads generated with InSilicoSeq v1.9 [29]. 100,000 reads per class from 15,000 RefSeq viral genomes, 1,000 RefSeq bacterial genomes, and hrg38 primary assembly
Genome-wide signatures and evolutionary fingerprint recognition	DNA vs RNA classification	Probes whether the model has learned the deep evolutionary and genome-wide signatures that distinguish DNA-based viruses from RNA-based viruses.	5,000 DNA and 5,000 RNA complete genomes from corpus test set
	Host Prediction	Evaluates whether the model captures host-associated genomic-wide signatures	2,000 complete genomes from different families grouped in 4 classes based

		that allow assigning viruses to broad host categories.	on the host: vertebrates, invertebrates, bacteria, and plants
Intra-genus species distinction	HIV-1 vs HIV-2 classification	Evaluates the model's ability to detect subtle genomic differences within a genus	7,248 HIV-1 and 94 HIV-2 complete genomes
Variant/Subtype typing	SARS-CoV-2 Lineage Identification	Measures how well the model resolves very fine-grained sequence differences between closely related SARS-CoV-2 lineages that differ by only a small number of mutations across otherwise highly similar genomes.	1,000 complete genomes across 7 different SARS-CoV-2 lineages (Alpha, Beta, Gamma, Delta, Lambda, Mu, Omicron)
Phenotypic context signal identification	HIV-1 Tropism Prediction	Tests if the model can capture subtle adaptation of a same subtype virus to different environments within the host	HIV-1 subtype B env CDS: 764 brain seqs and 1,325 plasma seqs

Table 1. Overview of the vGUE benchmark. *Tasks are grouped by biological level (broad organism-level distinctions, genome-wide signatures, intra-genus species discrimination, variant/subtype typing and phenotypic context) and described in terms of their objective ("Purpose") and the corresponding dataset ("Data"), including full genomes and simulated reads drawn from our curated viral corpus and from public repositories. The specific tasks and datasets listed here are illustrative examples; vGUE is designed to be flexible, as long as evaluations cover the same high-level categories of viral genome understanding.*

vir2vec outperforms state-of-the art approaches

To contextualize vir2vec, we compared it with (i) the human-trained model Mistral-DNA-v1-138M-hg38, pretrained using 10kb DNA sequences from the hg38 human genome assembly [31], and (ii) ModernBERT-DNA-v1-37M-virus [17], pretrained on 15071 viral reference genomes split into 1 kb segments. All three models were used strictly as frozen encoders, and evaluated under the same vGUE pipeline and shallow classifiers. For each task and each embedding technique we developed a Logistic Regression (LogReg), a Radial Basis Function Support Vector Machine (RBF-SVM), and a Random Forest (RF) classifier. The best performances for each task by the best model based on each competing encoder are shown in **Figure 3**, while **Table 2** describe the performance in detail across classifiers. Importantly, vir2vec outperforms the other approaches in terms of Balanced Accuracy (BA) for all but one task.

Task	Best Classifier	vir2vec	Mistral-DNA	ModernBERT -virus
Virus vs Non-Virus (genomes)	**LogReg**	**0.98 ± 0.006**	0.95 ± 0.003	0.96 ± 0.003
Virus vs Bacteria (Reads)	**LogReg**	**0.64 ± 0.003**	0.60 ± 0.005	0.63 ± 0.005
Virus vs Human (Reads)	**RF**	**0.97 ± 0.001**	0.66 ± 0.004	0.96 ± 0.001
DNA vs RNA	**RBF-SVM**	**0.96 ± 0.006**	0.89 ± 0.006	0.92 ± 0.009
Host Prediction	**RF**	**0.84 ± 0.01**	0.70 ± 0.007	0.78 ± 0.014
HIV-1 vs HIV-2	**LogReg**	**1.00 ± 0.000**	0.96 ± 0.055	0.99 ± 0.018
SARS-Cov2 Variant	**LogReg**	0.99 ± 0.005	0.98 ± 0.004	**1.00 ± 0.001**
HIV-1 Tropism	**RBF-SVM**	**0.67 ± 0.17**	0.61 ± 0.14	0.64 ± 0.17

Table 2. Performance comparison of three genomic language models on the vGUE benchmark. For each task and model (vir2vec, Mistral-DNA and ModernBERT-virus), the table reports the mean balanced accuracy (mean ± standard deviation over 10 nested cross-validation folds) obtained by the best performing classifier trained on frozen embeddings (LogReg, SVM or RF). The SVM uses either a linear or RBF kernel depending on embedding dimensionality and task complexity, chosen for computational efficiency. Overall, viral-pretrained models consistently outperform the human-trained baseline on most genome-wide and organism-level tasks. vir2vec attains high balanced accuracy across the benchmark—reaching near-ceiling performance on variant/subtype typing and HIV-1 vs HIV-2, strong scores on DNA vs RNA and host prediction, robust separation in organism-level tasks, and more modest but consistently above-chance performance on the harder HIV-1 Tropism task—typically achieving the highest scores on DNA vs RNA discrimination, host prediction and intra-genus separation. Within each task, the top-performing classifier is consistent across embedding models, i.e., the same classifier is providing the best results for a single task for all the embedding models,

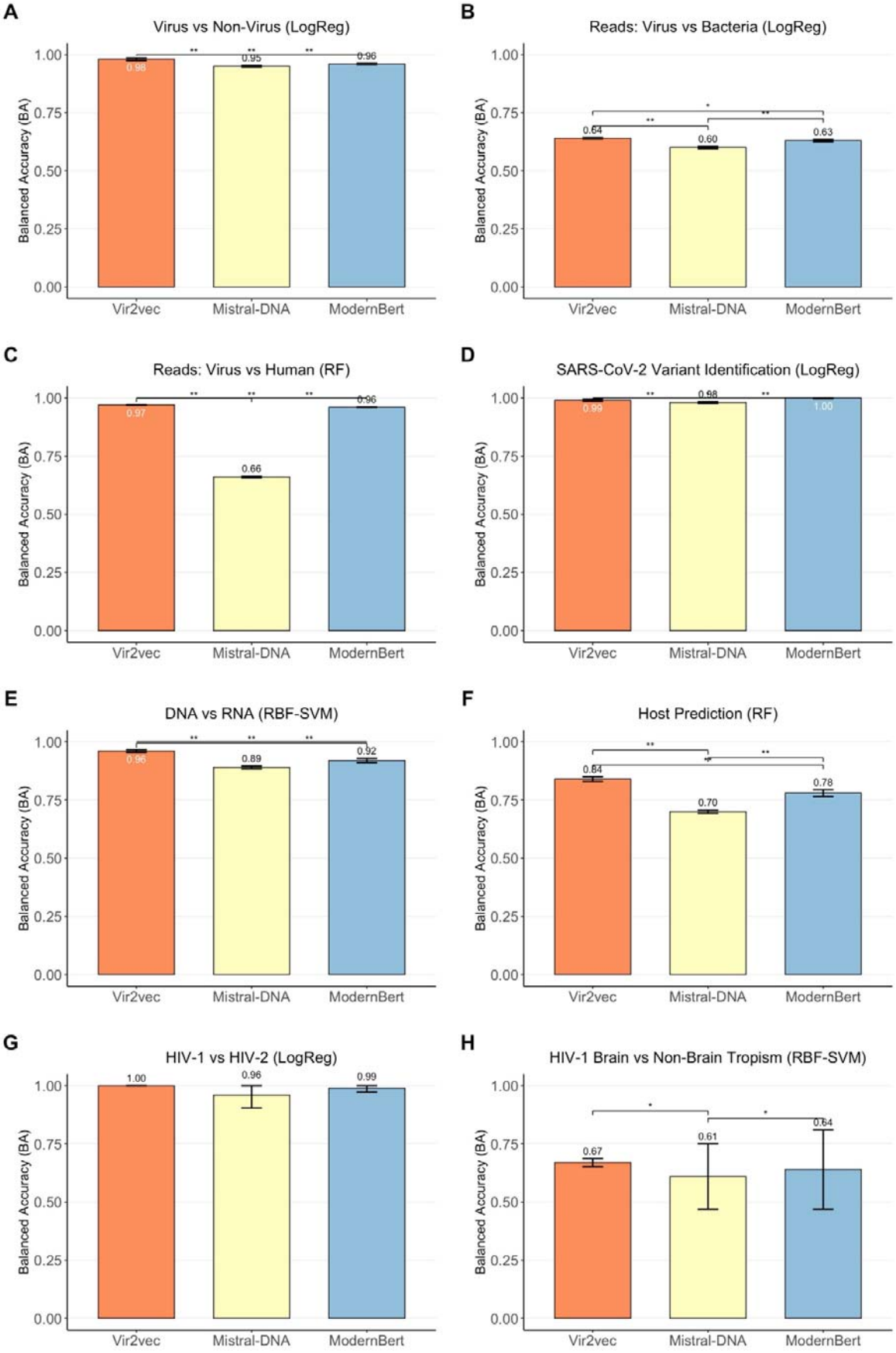
A
Virus vs Non-Virus (LogReg)
Balanced Accuracy (BA)
1.00
0.75
0.50
0.25
0.00
0.98
0.95
0.96
Vir2vec
Mistral-DNA
ModernBert
B
Reads: Virus vs Bacteria (LogReg)
0.64
0.60
0.63
C
Reads: Virus vs Human (RF)
0.97
0.66
0.96
D
SARS-CoV-2 Variant Identification (LogReg)
0.99
0.98
1.00
E
DNA vs RNA (RBF-SVM)
0.96
0.89
0.92
F
Host Prediction (RF)
0.84
0.70
0.78
G
HIV-1 vs HIV-2 (LogReg)
1.00
0.96
0.99
H
HIV-1 Brain vs Non-Brain Tropism (RBF-SVM)
0.67
0.61
0.64

Figure 3. Pairwise differences in balanced accuracy (ΔBA) between vir2vec, ModernBERT-virus, and the human-trained Mistral-DNA baseline across vGUE tasks. Bars show the mean ΔBA over cross-validation folds for the best-performing classifier per task (Logistic Regression, SVM, or Random Forest). Error bars represent the variability across folds (e.g. standard error or 95% confidence interval), and asterisks indicate statistically significant differences after Holm correction ($p < 0.05$, $p < 0.01$, $p < 0.001$; paired Wilcoxon signed-rank test).

To further compare models statistically, we computed pairwise differences in BA across cross-validation folds within each classifier (LogReg, SVM, RF), assessed significance with a paired Wilcoxon signed-rank test, and controlled for multiple comparisons using Holm correction. We report Holm-adjusted p-values alongside Δ (mean BA difference). vir2vec confirms its superior performance with statistical significance. Additional metrics and full performance by the three shallow classifiers are provided in the Supplementary Tables.

DISCUSSION

Our results reinforce the idea that genomic language models (gLMs) are highly effective numerical surrogates for genomes. Across the vGUE benchmark, a single frozen embedding space supports tasks ranging from broad organism-level discrimination to lineage typing and phenotype-associated signal detection, with balanced accuracies often matching or exceeding those reported for state-of-the-art genomic foundation models on more general benchmarks [13]. This is consistent with recent work on DNABERT-2 and the Nucleotide Transformer, which demonstrated that large, multi-species sequence models learn transferable representations that generalize across diverse downstream problems with minimal task-specific tuning [12, 13]. In this context, vir2vec can be viewed as a focused instantiation of the same paradigm: a genome-scale foundation model specialized on panviral genomics, rather than a single pathogen or host genome.

A central observation is that vir2vec consistently outperforms both the human-trained Mistral-DNA-v1-138M-hg38 and ModernBERT-virus models on tasks that depend on genome-wide or evolutionary fingerprints. These include DNA vs RNA viruses, host prediction, and intra-genus separation between HIV-1 and HIV-2, where vir2vec achieves the highest balanced accuracies among all compared models. These are precisely the settings in which training on hundreds of thousands of complete genomes from 295 viral species, with rich within-species diversity, should be most beneficial: the classifier must integrate weak, distributed signals (e.g. codon usage, dinucleotide biases, genome organization, host-adaptation footprints) rather than relying on a single marker region. The gains over a human-only baseline echo observations from broader genomic benchmarks, where moving from human-centric to multi-species pretraining substantially improves cross-organism generalization [12, 34].

Notably, the identity of the top-performing classifier is task-specific yet remarkably stable across embedding models (Table 2). For instance, RF consistently outperforms SVM and LogReg on Host Prediction (for vir2vec) and on Virus vs Human, whereas SVM is consistently the strongest choice for HIV-1 tropism across all embedding approaches. In contrast, LogReg emerges as the best-performing classifier in four tasks (Virus vs Non-virus, Virus vs Bacteria, HIV-1 vs HIV-2, and SARS-CoV-2 variant identification). Overall, this pattern suggests that the embedding spaces already encode highly discriminative signals, such that a simple linear decision rule can match or exceed more flexible non-linear models, which may be more prone to overfitting in these settings.

The read-level organism classification tasks highlight a different kind of challenge. Discriminating viral from bacterial reads at 150 bp is intrinsically hard: short fragments provide a narrow window

on the genome, compositional patterns are partially shared across domains of life, and many classic discriminants (e.g. genome segmentation, gene order) are simply not visible. Not surprisingly, virus-vs-bacteria reads yield some of the lowest balanced accuracies in vGUE, with all models clustering around modest but clearly above-chance performance. Yet despite this limited context, vir2vec still outperforms both the human-trained Mistral-DNA and the virus-reference ModernBERT across classifiers, suggesting that its 4096-dimensional embeddings capture more discriminative short-range sequence features. For virus-vs-human reads, the task is comparatively easier—human sequence is strongly distinct from viral content even in small windows—and here a non-linear classifier on vir2vec embeddings approaches near-perfect separation. These findings broadly align with recent metagenomic studies such as ViraLM, where foundation-model embeddings improve virus–non-virus discrimination on contigs of varying length, particularly when reference databases are incomplete [19].

Interestingly, the tasks where vir2vec obtains the lowest absolute balanced accuracies are HIV-1 Tropism prediction (brain vs plasma) and virus-vs-bacteria read classification, despite still outperforming the comparison models in both settings. In the Tropism task, the model must recover subtle phenotypic adaptation signal from env coding sequences, in the presence of limited sample size and noisy, heterogeneous metadata. These results suggest that phenotype-level predictions such as tissue Tropism may require larger, more carefully curated cohorts and explicit fine-tuning strategies—potentially multi-task or contrastive training that directly incorporates host and tissue context—building on top of the generic viral representations provided by vir2vec.

More broadly, our comparison confirms that viral-specific pretraining matters even in short-read settings. For virus-vs-bacteria reads, vir2vec is consistently better than both Mistral-DNA and ModernBERT-virus, though the performance gap is understandably smaller than for genome-wide tasks. This echoes observations from ViraLM and related work, where gLM-based virus detection gains are largest on challenging metagenomic contigs that are short, noisy, or distant from known references [19]. Taken together, our results support a layered view of viral representation learning: (i) large, pan-viral gLMs provide a strong, task-agnostic embedding space; (ii) shallow classifiers already unlock much of this potential across a wide task spectrum; and (iii) for particularly difficult or label-limited problems (e.g. Tropism, zoonotic risk estimation), task-aware fine-tuning on top of these embeddings is a promising direction.

Lastly, we report also that we probed a ninth task: Influenza HA/NA segment and H/N subtype prediction. In Influenza use-cases, all models – from lightweight baselines to vir2vec – achieved nearly perfect balanced accuracy, ranging from 0.99 to 1.00 (data not shown). In these problems, genomes within a class are almost identical except for sparse point mutations, often concentrated in well-known regions. Once a model has a reasonably faithful local sequence representation, distinguishing such variants becomes comparatively easy, and additional pretraining data or higher-dimensional embeddings offer limited headroom. For this reason, we do not further dissect these results and instead focus on vGUE tasks where genome-wide representation learning is more discriminative.

This study has several limitations: vGUE focuses on complete genomes and simulated short reads, leaving out fragmented, low coverage contigs from real metagenomic datasets, where assembly errors, chimeras, and host contamination are common. The viral corpus, while broad, is still biased toward intensively sequenced pathogens such as SARS-CoV-2 and influenza, and we explicitly down-sampled these taxa to control over-representation. As a result, extremely rare species and emerging viruses remain underexplored. In addition, our evaluation uses only shallow classifiers on frozen embeddings; deeper task-specific heads or end-to-end fine-tuning may yield further gains, particularly for phenotype-rich tasks. Finally, we benchmark vir2vec against two specific open-

source baselines; integrating more recent foundation models such as updated Nucleotide Transformer variants, HyenaDNA, or DNABERT-S for long reads would provide a broader landscape of comparison [12, 34].

Despite these caveats, the overall picture is clear: gLMs are not only viable but highly competitive tools for the numerical modeling of viral genomes, and domain-specialized pretraining on a large, carefully curated pan-viral corpus yields tangible benefits. vir2vec demonstrates that a single decoder-only transformer, used purely as a frozen encoder, can support a wide range of biologically meaningful tasks, from organism-level discrimination to fine-grained typing, with particularly strong gains whenever the relevant information is distributed along the entire genome rather than confined to a single region. By coupling this embedding engine with the vGUE benchmark, we provide both a model and an evaluation framework that can serve as a foundation for future viral gLMs, downstream applications in surveillance and discovery, and systematic comparisons across generations of genomic foundation models.

CONCLUSIONS

In this work, we introduced vir2vec, a pan-viral genomic language model trained on complete genomes from hundreds of viral species, and vGUE, a unified benchmark that spans tasks from broad taxonomic discrimination to fine-grained lineage typing and phenotype prediction We showed that a single embedding space can support a wide range of downstream classifiers with strong and often state-of-the-art performance, reinforcing the view that genomic language models are effective numerical surrogates for viral genomes.

vir2vec consistently outperformed both a human-trained genomic foundation model and a viral-specific baseline. The gains were particularly pronounced in genome-wide tasks, where the relevant signal is distributed along the entire sequence—such as DNA vs RNA virus discrimination, host prediction, and HIV-1 vs HIV-2 classification. In these settings, the combination of a long receptive field and training on diverse, complete viral genomes appears critical: the model captures evolutionary and compositional fingerprints that are not confined to a single locus, and these are readily exploited by shallow downstream classifiers.

Our results support a layered perspective on viral representation learning. First, domain-specialized pretraining on a large, carefully filtered pan-viral corpus yields clear benefits over more generic genomic models. Second, frozen embeddings plus shallow classifiers already unlock much of this potential, enabling efficient reuse across many tasks without per-task fine-tuning. Third, for particularly difficult or label-limited problems—such as Tropism, zoonotic risk estimation, or subtle phenotypes—vir2vec should serve as a strong initialization for more advanced strategies, including multi-task, contrastive, or retrieval-augmented fine-tuning.

Finally, vGUE itself provides a reusable evaluation framework for viral genomic foundation models. By spanning whole-genome and read-level tasks, diverse hosts, and both taxonomic and phenotypic labels, it enables systematic, apples-to-apples comparisons between successive generations of models. We release vir2vec and vGUE as open resources with the goal of catalyzing a new wave of pan-viral gLMs and downstream tools for surveillance, diagnostics, and discovery. Future work will extend this paradigm to real metagenomic data, longer and noisier reads, and richer clinical and ecological annotations, with the broader objective of turning viral genomic sequences into a structured, queryable latent space for virology and public health.

METHODS

vir2vec architecture

vir2vec builds on the Mistral-DNA-v1-422M-hg38 model, a 422-million-parameter large language model originally developed to model the human genome (**Figure 1**) [16]. The base architecture is a decoder-only transformer: given a DNA sequence, the model reads it from left to right and learns to predict each nucleotide based on all preceding nucleotides, in the same way that language models in natural language processing predict the next word in a sentence. Structurally, vir2vec follows the Mistral/Mixtral family of transformers in a reduced and DNA-specialized configuration. The network consists of a stack of transformer layers, and each layer combines a self-attention mechanism with a small feed-forward neural network [15]. In the feed-forward part, the model uses a mixture-of-experts design: instead of a single feed-forward block, several expert subnetworks are available, and the model learns which experts to activate for each input. This design increases the expressive power of the model without a proportional increase in computational cost. In the attention component, vir2vec employs efficient mechanisms such as grouped-query attention and sliding-window attention, which allow it to handle relatively long DNA sequences while keeping memory usage manageable [15].
Each nucleotide sequence is first converted into discrete tokens using the tokenizer distributed with Mistral-DNA [16]. This tokenizer is based on byte-pair encoding (BPE) [35]adapted to DNA: rather than representing the sequence only as individual bases (A, C, G, T, N), it learns a vocabulary of short subwords that correspond to frequently occurring motifs and oligonucleotides in genomic data. This tokenization scheme allows the model to capture recurring sequence patterns in a compact way and to generalize more effectively across different genomes and host species [12, 36, 35].

Internally, every token is mapped to a 4096-dimensional vector, and the series of transformer layers progressively refines these vectors to encode both local sequence motifs and long-range dependencies along the genome. To obtain a single embedding for an entire viral genome or sequencing read, we aggregate the token representations using max pooling over the sequence dimension: for each of the 4096 features, we retain the maximum value observed across all positions in the sequence. The result is a fixed-length, 4096-dimensional representation that summarizes the most salient patterns present in that viral sequence and can be used as input for downstream analyses.

Continual pretraining on viral genomes

To transform the original human-focused Mistral-DNA model into a pan-viral embedding engine, we performed continual pretraining on a curated corpus of viral genomes [37]. Starting from the pre-trained Mistral-DNA weights, we further trained the model on hundreds of thousands of viral sequences spanning multiple host groups and including both DNA and RNA viruses. During this phase, the learning objective remained unchanged: given a partially observed sequence, the model is trained to predict the next nucleotide, without ever seeing future positions.

For training, viral genomes were organized into a tabular dataset in which a dedicated column contained the raw nucleotide sequence (named Sequence). Each sequence was tokenized with the Mistral-DNA tokenizer and either truncated or padded to match the maximum input length supported by the model. We used standard causal language modeling without masking, so that at each position the model only has access to previous tokens.

The optimization protocol was deliberately kept simple and robust. We used the AdamW optimizer, a widely adopted variant of stochastic gradient descent with weight decay, with an initial learning rate of 5×10^{-4} and a weight decay coefficient of 0.01 to limit overfitting [38]. The batch size was

set to two sequences per device, and gradient accumulation over four steps was used to obtain an effective batch size of eight sequences per parameter update. Training was run for up to ten epochs over the viral corpus. A warmup phase of 1,000 gradient steps gradually increased the learning rate from zero to its target value, helping to stabilize the early stages of training. To further improve stability, gradients were clipped to a maximum norm of 1.0. All training was conducted in full 32-bit floating-point precision (fp32), avoiding complications related to mixed-precision arithmetic.
To monitor learning and reduce the risk of overfitting, we set aside a validation split of the viral dataset and evaluated the model at the end of each epoch. After every epoch, the current state of the model was checkpointed, and the corresponding validation loss was recorded. An early-stopping criterion with a patience of three epochs was applied [11]: if the validation loss did not improve for three consecutive epochs, training was stopped. The checkpoint with the lowest validation loss across all epochs was selected as the final version of vir2vec.

No architectural changes were introduced during continual pretraining: the number of layers, attention heads, and hidden dimensions remained identical to those of the original Mistral-DNA-v1-422M-hg38 configuration. Conceptually, the transition from Mistral-DNA to vir2vec consists solely of adapting the model's weights so that the transformer becomes specialized in viral genomics as the language of viruses, rather than being tuned mainly to the human genome. This strategy allows vir2vec to retain the general sequence-modeling capabilities acquired from human genomic data while reshaping its internal representations to better capture the diverse patterns of viral genomes. In the resulting configuration, vir2vec is used as a fixed feature extractor: given any viral genome or read, it produces a 4096-dimensional embedding that serves as the common representation for all downstream tasks considered in this study.

Data

Public viral repositories such as NCBI Virus and BV-BRC are intrinsically heterogeneous: high- and low-quality assemblies, as well as richly and sparsely annotated genomes, co-exist within the same resources. Building a training corpus for a viral genomic language model therefore requires rigorous data retrieval and systematic quality control, with particular attention to ambiguous bases and taxonomic imbalance.

Only complete viral genomes are selected for training vir2vec, rather than fragmented genomes, to expose the model to natural genome-wide constraints (e.g., gene order, codon usage, long-range dependencies), while still enabling downstream tasks at read or segment level by extracting embeddings from local windows of the same model. All complete genomes in NCBI Virus and BV-BRC are included in the corpus if they satisfiy two criteria: <1% ambiguous bases ("N") and no stretches longer than 20 consecutive Ns. Ns represent unresolved nucleotides and long N-runs typically indicate low-quality assemblies; at the same time, fully excluding sequences with Ns would substantially reduce genomic diversity. Our thresholds ensure that the model sees at least 99% resolved bases per genome, limiting the impact of missing data on self-attention, and are consistent with GISAID and Nextclade guidelines for flagging low-quality genomes [39].

The corpus is further enriched with sequences and metadata from species-specific resources (GISAID, LANL-HIVdb, HBVdb). Harmonizing taxonomic fields across repositories (e.g. "HIV-1" vs "Human immunodeficiency virus 1") required manual normalization but greatly facilitated the construction of downstream evaluation datasets; after this harmonization step, we removed exact duplicate sequences.

To address under-representation, we exclude taxa with fewer than 100 genomes, removing ~80,000 genomes. Although this reduced the nominal number of species from 28,229 to 295, many of the

discarded entries correspond to individual isolates that are effectively treated as distinct species in public databases; as noted in literature, GenBank-style entries often reflect viral isolates rather than ICTV-recognized species, and many are mislabeled as unique species [40]. For applications focused on genuinely rare or emerging viruses, vir2vec can be further adapted by fine-tuning on new data as it becomes available.

Over-representation is controlled by random down-sampling of SARS-CoV-2 and Alphainfluenzavirus, the dominant classes in both NCBI and BV-BRC, to 100,000 genomes each. For influenza, down-sampling is stratified by segment so that the segment distribution in the subsample mirrors the full collection.

The final corpus includes 565,747 genomes from 295 viral species. The dataset is split 70/30 into training and test preserving species-level proportions in both subsets (**Figure 1**); a small validation set is drawn from the test split for hyperparameter tuning and comparison of alternative training configurations. The reason of creating a test set is to avoid data leakage; only genomes in the test set are used to create the datasets for the downstream evaluations.

The vGUE Benchmark

The first attempt to introduce a benchmark for gLMs was made by the authors of DNABERT-2 (Zhou et al., 2024) who developed the Genome Understanding Evaluation (GUE) benchmark [13]. It compiles 36 distinct datasets spanning 9 different task types across multiple species; the tasks include genomic sequences across 4 different organisms (human, mouse, yeast/fungi, and virus). For the virus category specifically, GUE included a variant classification task focusing on SARS-CoV-2 and a species classification. As these were the only virus-specific tasks in the benchmark, important viral genomic challenges were not represented in the original GUE. This highlights a need for a more expansive benchmark dedicated to viral genomics, especially when evaluating gLMs trained solely with viruses.

For this reason, we propose the vGUE as a specialized benchmark tailored to viral gLMs. The rationale for vGUE is that evaluating a foundation model on viral sequences requires tasks attuned to the biology of viruses. Many standard genomics tasks (e.g. epigenetic mark prediction, promoter detection) are not applicable to viruses – as they have quite simple genomes also due to their unique way to replicate (they depend on hijacking cells). vGUE is designed to fill this gap by covering a diverse set of evaluation tasks that reflect the unique characteristics and challenges of viral genomes. These tasks assess a model's ability to capture biological and chemical features of viruses at multiple levels of granularity. It's worth emphasizing that the specific tasks in vGUE are illustrative – the exact dataset choices could evolve – but the core idea is to set a standard suite of evaluations covering key aspects of viral genome analysis: broad organism-level distinctions, short-read classification, genome-wide signatures and evolutionary fingerprint recognition, intra-genus species distinction, variant/subtype typing, and phenotypic context signals identification.

Embedding Generation

vir2vec is used strictly as a frozen encoder: given a raw nucleotide sequence, the model produces a single fixed-length embedding that summarizes the full genome or read. Architecturally, each input sequence is tokenized with the DNABERT-2–style tokenizer, every token is mapped to a 4096-dimensional vector, and the transformer stack refines these token representations. To obtain a sequence-level representation, we apply max-pooling across the sequence dimension, retaining for each of the 4096 features the maximum value observed over all positions. This yields a 4096-dimensional embedding that captures the most salient patterns in the sequence and serves as the common input to all downstream analyses.

For each task, we construct a tabular dataset in which one column contains the raw nucleotide string (genome, contig, or read) and additional columns store task-specific labels (e.g. host category, lineage, DNA vs RNA). Datasets are stored on disk as CSV files and processed in a streaming fashion to handle up to hundreds of thousands of sequences without exhausting system memory. During embedding extraction, sequences are read in chunks of N rows (adjusted based on dataset size) and passed to the tokenizer. The tokenizer operates in left-padded, truncated mode with a maximum input length of 4,096 tokens: sequences shorter than this limit are padded, whereas longer sequences are split into multiple windows by the tokenizer. Left padding is used to respect the causal, decoder-only nature of vir2vec (the model should see real bases before padding tokens). Genomes that exceed the 4,096-token context window are encoded as a small number of contiguous windows. For each window, we take the last hidden state from the model and apply max-pooling across the token dimension, obtaining a single 4096-dimensional vector per window. To recover a single embedding for the entire genome or read, we then compute the mean of these window-level vectors, effectively averaging over local max-pooled summaries. This two-step pooling (max across tokens, mean across windows) makes the representation invariant to genome length while still highlighting the strongest motif-level activations anywhere in the sequence. We deliberately chose max-pooling over token embeddings rather than simple mean-pooling because we are primarily interested in the presence of informative motifs (e.g. host- or lineage-associated patterns), which can be washed out by averaging over long genomes. The subsequent averaging across windows ensures that very long genomes do not dominate simply by contributing more windows.

Batched inference and memory management

Within each chunk, sequences are further divided into mini-batches for GPU inference. For each batch, the tokenized inputs are moved to the GPU and passed through the model. All inference is performed in full 32-bit floating-point precision (fp32), as in training, to avoid numerical issues associated with mixed precision.

To keep the pipeline scalable across tasks with very different sizes, chunk size and batch size are treated purely as computational knobs. This tuning affects throughput and memory usage but not the definition of the embedding itself.

Embedding storage and alignment with downstream tasks

The final embeddings for a given model–task pair are stored as an HDF5 matrix of shape

$$(n\ seq, 4096),$$

with one row per sequence in the original dataset. Labels and metadata are maintained in parallel structures and are always aligned by row index, ensuring that each embedding can be unambiguously paired with its corresponding sequence and task label. The same extraction code and pooling strategy is used for all vGUE tasks; only the input files and the chunk/batch sizes differ. This guarantees that differences in downstream performance reflect the information content of the embeddings rather than idiosyncrasies of the extraction procedure.

Classifier Layer

For all downstream tasks in vGUE, vir2vec (and all comparison models) is used strictly as a frozen feature extractor. For each viral genome or read, we compute a single fixed-length embedding as described above, obtaining one 4096-dimensional vector per sequence. These embeddings are then

used as input features to shallow classifiers trained separately for each task and each model, so that differences in performance can be attributed to the quality of the representations rather than to differences in classifier architecture or tuning.

We deliberately adopt a small, standard set of supervised classifiers that cover complementary inductive biases while remaining easy to interpret and compare: LogReg provides a linear decision boundary with probabilistic outputs. It is a natural baseline for testing whether the embedding space is already close to linearly separable for a given task; SVM is used either with a linear kernel or a RBF kernel, depending on the size of the dataset. Linear SVM behaves similarly to LogReg but optimizes a margin-based objective and is well suited for high-dimensional embeddings. The RBF kernel allows for non-linear decision boundaries and can exploit residual structure not captured by a purely linear separator; RF represents a non-parametric, tree-based ensemble that can capture non-linear interactions between embedding dimensions and is robust to outliers and heterogeneous feature scales.

All downstream datasets are constructed exclusively from genomes in the held-out test partition of the viral corpus, to avoid any data leakage between continual pretraining and evaluation. Within each task, we perform stratified 10-fold cross-validation, ensuring that class proportions are preserved in every fold. In each outer fold, approximately 90% of the data are used for training and 10% for testing. To tune hyperparameters without biasing the outer test estimates, we adopt a nested cross-validation scheme. For a given classifier and outer training split, we run a grid search over a small, task-agnostic hyperparameter grid using an inner 3-fold stratified cross-validation, with balanced accuracy as the optimization criterion. The best hyperparameters from the inner search are then used to refit the classifier on the full outer training fold, and the resulting model is evaluated on the held-out test fold. This procedure is repeated for all 10 folds, and we report the mean and standard deviation of the performance metrics across folds. Hyperparameter grids are deliberately compact and centered around commonly used values, to ease comparison across tasks while avoiding excessive overfitting to individual datasets. For example, for LogReg we vary the inverse regularization strength C over a small range around 1 and consider standard $L2$-penalized solvers with sufficient maximum iterations to guarantee convergence. For SVMs, we tune C and (for RBF kernels) the kernel width γ, exploring a logarithmic grid around default values. For RF, we vary the number of trees, maximum depth, minimum number of samples per leaf, and feature subsampling strategy. In tasks with pronounced class imbalance (e.g. HIV-1 vs HIV-2) we additionally enable class-balanced loss functions by setting class weights to be inversely proportional to class frequencies in the training data, or by using 'balanced_subsample' in the RF.

For each task in the benchmark, we used mean balanced accuracy over cross-validation folds as the primary performance metric. Balanced accuracy is preferable to plain accuracy in our setting to avoid issues with tasks that are affected by class imbalance; BA gives equal weight to each class and is therefore more informative than overall accuracy when some classes are rarer than others. For binary problems, balanced accuracy is defined as the average of sensitivity and specificity:

$$BA = \frac{1}{2}(TPR + TNR) = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right),$$

Where TP, TN, FP, FN, are true positives, true negatives, false positives and false negatives.

For multiclass problems, we compute balanced accuracy as the macro-average of per-class recall. For each class c, we calculate

$$Recall_c = \frac{TP_C}{TP_C + FN_c},$$

and then average across all C classes:

$$BA = \frac{1}{C}\sum_{c=1}^{C} Recall_c$$

All reported BA values are the mean over cross-validation folds, and we also report the corresponding standard deviation. In the Supplementary Tables, we additionally report other standard metrics: precision, recall, ROC-AUC, and F1 Score, also computed separately on each fold and then averaged across folds.

AKNWOLEDGEMENTS

We gratefully acknowledge funding from the NIH NIAID (R01 AI170187) and the University of Florida Artificial Intelligence and Complex Computational Research Award.

P.Arozarena Donelli is a PhD student enrolled in the National PhD program in Artificial Intelligence, XLI cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma.

DATA AVAILABILITY

The data and code used to generate the results reported in this paper are available at https://github.com/simoRancati/Vir2vec. The pretrained model can be accessed at https://huggingface.co/pabloarozarenad/Vir2vec upon request, subject to providing an institutional email address, a brief description of the intended use, and the associated IRB protocol number.

REFERENCES

[1] W. H. O. (WHO), "Influenza (seasonal)," 28 February 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/influenza-%28seasonal%29. [Accessed 18 November 2025].

[2] W. H. O. (WHO), "Dengue," 21 August 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue?. [Accessed 18 November 2025].

[3] F. S. C. G. G. e. a. Coutinho, "Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans," *Nat Commun*, vol. 8, no. 15955, 2017.

[4] V. L. G. &. B. F. Aggarwala, "Viral communities of the human gut: metagenomic analysis of composition and dynamics.," *Mobile DNA*, vol. 8, no. 12, 2017.

[5] S. Y. L. H. L. W. L. G. W. G. L. C. P. Y. Z. Q. D. Y. S. J. Y. J. Zhao L, "Uncovering 1058 Novel Human Enteric DNA Viruses Through Deep Long-Read Third-Generation Sequencing and Their Clinical Impact.," *Gastroenterology*, vol. 163, no. 8, pp. 699-711, 2022.

[6] S. M. H. A. Khot V, "Computational approaches in viral ecology.," *Comput Struct Biotechnol J.*, vol. 9, no. 18, pp. 1605-1612, 2020.

[7] R. V. J. B. E. L. K. L. M. S. J. W. A. Z. J. D. M. Hegarty B, "Benchmarking informatics

approaches for virus discovery: caution is needed when combining in silico identification methods," *mSystems,* Vols. 9:e01105-23., 2024.

[8] L. S. H. Y. C. S. Bai GH, "The Human Virome: Viral Metagenomics, Relations with Human Diseases, and Therapeutic Applications.," *Viruses.,* vol. 14, no. 2, p. 278, 2022 Jan 28.

[9] A. S. A. B.-M. e. a. Moreno Zolfo, "Discovering and exploring the hidden diversity of human gut viruses using highly enriched virome samples," *bioRxiv,* 2024.02.19.580813..

[10] M. G. D. J. K. e. a. Garyk Brixi, "Genome modeling and design across all domains of life with Evo 2," *bioRxiv,* vol. 638918, 202502.18.

[11] N. G. B. L. B. T. C. D. P. E. B. R. S. M. P. M. M. S. Rancati S, "SARITA: a large language model for generating the S1 subunit of the SARS-CoV-2 spike protein.," *Brief Bioinform.,* vol. 26(4):bbaf384, 2025 Jul 2.

[12] H. G. L. M.-R. J. e. a. Dalla-Torre, "Nucleotide Transformer: building and evaluating robust foundation models for human genomics.," *Nat Methods,* vol. 22, p. 287–297, 2025.

[13] Y. J. W. L. P. D. R. D. H. L. Zhihan Zhou, "DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome," *arXiv,* no. 2306.15006, 2024.

[14] K. H. T. D. L. N. A. V. Malusare A, "Understanding the Natural Language of DNA using Encoder-Decoder Foundation Models with Byte-level Precision.," *Bioinform Adv.,* vol. 4(1):vbae117, 2024 Aug 12.

[15] A. S. A. R. e. a. Albert Q. Jiang, "Mixtral of Experts," *arxiv,* vol. 2401.04088, 2024.

[16] R. Mourad, "HuggingFace," [Online]. Available: https://huggingface.co/RaphaelMourad/Mistral-DNA-v1-422M-hg38.

[17] R. Mourad, "HuggingFace," [Online]. Available: https://huggingface.co/RaphaelMourad/ModernBert-DNA-v1-37M-virus.

[18] A. C. B. C. e. a. Benjamin Warner, "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference," *arxiv,* vol. 2412.13663, 2024.

[19] J. S. J. G. D. W. Y. S. Cheng Peng, "ViraLM: empowering virus discovery through the genome foundation mode," *Bioinformatics,* vol. 40, no. 12, December 2024.

[20] H. Gunasekaran, N. Wilfred Blessing, U. Sathic and M. Husain, " Enhanced Viral Genome Classification Using Large Language Models.," *Algorithms* , vol. 18, no. 302, 2025.

[21] N. C. f. B. Information, "NCBI Virus," 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/. [Accessed November 2024].

[22] R. A. T. B. e. a. Robert D Olson, *Nucleic Acids Research,* vol. 51, no. D1, p. D678–D689, 2023.

[23] S. e. a. Khare, "GISAID’s Role in Pandemic Response," *China CDC Weekly,* vol. 3, no. 49, pp. 1049-1051, 2021.

[24] L. A. N. Laboratory. [Online]. Available: https://www.hiv.lanl.gov/. [Accessed November 2024].

[25] J. H. e. al., "HBVdb: a knowledge database for Hepatitis B Virus," *Nucleic Acids Research,* vol. 41, no. D1, p. D566–D570, 2013.

[26] K. R. S. a. C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA),* Sydney, 2020.

[27] D. L. D. a. D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vols. PAMI-1, no. 2, pp. 224-227, 1979.

[28] U. M. a. S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 24, no.

12, p. 1650–1654, 2002.

[29] H. G. e. al., "Simulating Illumina metagenomic data with InSilicoSeq," *Bioinformatics*, vol. 35, no. 3, pp. 521-522, 2019.

[30] A. F. M. e. a. Laura Heath, "Evidence for Limited Genetic Compartmentalization of HIV-1 between Lung and Blood," *PLOS ONE* , 2009.

[31] M. Raphael. [Online]. Available: huggingface.co/RaphaelMourad/Mistral-DNA-v1-138M-hg38.

[32] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80-83, 1945.

[33] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, p. 65–70, 1979.

[34] W. L. Z. B. H. C. Z. J. W. J. L. L. W. P. W. C. Feng H, "Benchmarking DNA Foundation Models for Genomic Sequence Classification.," *bioRxiv*, vol. 2024.08.16.608288, 2024 Aug 18.

[35] B. C. Testagrose C, "Tokenization and deep learning architectures in genomics: A comprehensive review," *Comput Struct Biotechnol J.*, vol. 27, pp. 3547-3555, 2025 Jul 28.

[36] H. A. P. M. Cherednichenko O, "Benchmarking DNA large language models on quadruplexes.," *Comput Struct Biotechnol J.*, vol. 27, pp. 992-1000, 2025 Mar 7.

[37] M. A. S. S. L. K. B. I. D. D. S. N. Gururangan S, "Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks.," *arXiv preprint*, vol. arXiv:2004.10964, 2020.

[38] H. F. Loshchilov I, "Decoupled Weight Decay Regularization.," *arXiv preprint*, vol. arXiv:1711.05101 , 2019.

[39] I. e. a. Aksamentov, "Nextclade: clade assignment, mutation calling and quality control for viral genomes," *Journal of Open Source Software*, 2021.

[40] M. H. V. Van Regenmortel, "Virus species and virus identification: Past and current controversies," *Infect. Genet. Evol.*, vol. 7, pp. 133-144, 2007.

[41] G. A. A. K. Martin C, "Protein Set Transformer: A protein-based genome language model to power high diversity viromics," *Res Sq [Preprint]*, Vols. rs.3.rs-4844047., 2024 Sep 23.

[42] L. A. N. Laboratory. [Online]. Available: https://www.hiv.lanl.gov/.