# WAVEBENCH: ASSESSING GENERALIZATION OF GENOME LANGUAGE MODELS IN WASTEWATER VIRUS DETECTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Genome language models (gLMs) are trained on DNA sequences to illuminate the relationship between genotype and phenotype. This goal is only useful if models are able to extend their reasoning beyond annotated data, or generalize out-of-distribution. However, this evaluation is confounded, since all biological sequences descend from the same origin. We introduce WAVEBench, a benchmark that withholds whole taxa, at ranks from species to class, from the labelled folds used to train a downstream probe. This way, each held-out taxon is novel to the probe. To test generalization, we devised a fold-assignment algorithm that, given any hierarchy of relatedness, produces a cross-validation partition where each fold's held-out data is stratified into progressively more distant novelty ranks from the remaining folds. We apply it to novel virus identification in wastewater metagenomics, a timely public-health surveillance task where the evolutionary distribution shift is intrinsic. We evaluated a logistic probe on 51 models from 11 gLM families and compared them to a 5-mer baseline. Median AUROC falls from species to class, but individual models do not decline uniformly; different ranks contain different clades, so these comparisons do not isolate an effect of evolutionary distance. Also, performance varies substantially relative to a k-mer composition baseline, with a subset of models achieving higher leaderboard AUROC. Overall, we contribute a fold-construction algorithm, a benchmark on a pressing public-health problem, and an assessment of gLMs' ability to generalize to taxa withheld from downstream training.

## 1 INTRODUCTION

Genome language models (gLMs) learn general-purpose representations of biological sequences intended to support downstream prediction tasks across diverse organisms and sequence contexts, ranging from regulatory element prediction (Boshar et al., 2025) to genome design (King et al., 2026). This ability is particularly promising as genomic sequences, and especially experimentally labeled sequences, represent only a fraction of biological diversity (Nayfach et al., 2021; West-Roberts et al., 2024). However, current gLMs can perform poorly on out-of-distribution sequences (Jiang et al., 2026; Hearne et al., 2026). It therefore remains unclear to what extent their learned representations support prediction across evolutionarily distinct taxa.

Because all biological sequences share common descent (Felsenstein, 1985), held-out sequences and even whole clades are never fully independent of training data. Common splitting protocols do not reliably eliminate leakage, which must be controlled explicitly (Bernett et al., 2024; Nagai et al., 2026). Instead, train–test similarity can be made an explicit, tunable axis of a benchmark (Petti & Eddy, 2022), and model performance measured as a function of it (Rafi et al., 2026). Detecting viral sequences from taxa withheld from training in raw, diverse wastewater metagenomics provides such a test at the far end of that axis, with critical applications for biosurveillance (The Nucleic Acid Observatory Consortium, 2021; Levy et al., 2023).

We introduce WAVEBench (Figure 1), a benchmark applied to wastewater metagenomics data to assess downstream out-of-taxon generalization of gLMs and, reciprocally, evaluate these models' potential to support wastewater surveillance through viral read detection. Our key contributions are:



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

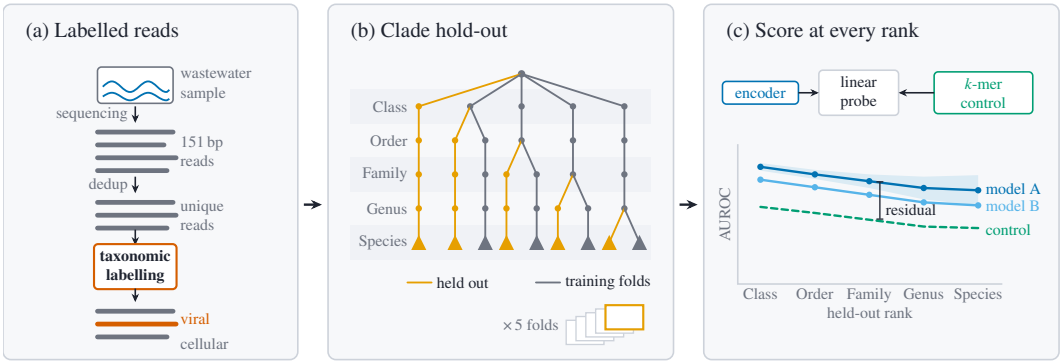



Figure 1: **WAVEBench overview.** (a) Wastewater reads are processed and taxonomically labelled. (b) Each fold holds out whole clades at ranks from class to species. A probe trained on the other four folds is evaluated separately at each rank. Orange marks held-out clades; novelty is relative to probe training, not encoder pretraining. (c) Frozen encoders, such as METAGENE-1 and Evo, and the $k$-mer composition control use the same probe protocol.

- **Graded taxonomic cross-validation.** We develop a fold-construction algorithm that enables a classifier to be evaluated at increasingly coarse holdout ranks within a single cross-validation partition.
- **A wastewater metagenomic data-based benchmark.** We apply this algorithm to a viral-versus-cellular classification task in wastewater sequences, providing a generalization assessment on taxa absent from downstream training.
- **Empirical maps of gLM out-of-taxon generalization.** We evaluate 51 models from 11 model families on the benchmark using a shared frozen-embedding linear-probe protocol. The results characterize how performance changes across held-out ranks.

Only 19 of 51 models beat a simple 5-mer composition baseline. As the withheld taxa become more distant, the best models far exceed restricted BLAST and DeepVirFinder trained on the same dataset. However, even the best models miss most viral reads at strict false-positive rates, and at realistic viral shares their calls need confirmation. WAVEBench thus measures how well gLMs recognize viruses from taxa withheld from downstream training, and what that ability is worth for wastewater surveillance (Section 5).

## 2 Related Work

**Evaluation and generalization of gLMs.** gLMs are pretrained through self-supervision on DNA sequences from human genomes, multispecies references, and metagenomic corpora, including environmental samples (Ji et al., 2021; Dalla-Torre et al., 2025; Liu et al., 2025). Existing benchmarks evaluate regulatory prediction, variant effects, gene properties, and tasks requiring long-range sequence context (Marin et al., 2024; West-Roberts et al., 2024; Kan-Tor et al., 2024; Kao et al., 2024; Cheng et al., 2025; Feng et al., 2025). However, high performance on downstream tasks does not guarantee generalization at evolutionary scale, since these benchmarks do not evaluate a graded series of taxonomic holdouts (Grešová et al., 2023; Liu et al., 2024; Patel et al., 2024). Jiang et al. (2026) linked the accuracy of loss-of-function prediction in synthetic expression cassettes to sequence likelihood, and suggested that sequence familiarity influences functional prediction. Separately, Hadad et al. (2026) demonstrated that performance on a promoter prediction task decreased the further the training genome was from the test genome. Extending to metagenomics, Hearne et al. (2026) reported elevated classification error and miscalibration on simulated reads from genera and families held out of training. WAVEBench measures these generalization failures across taxonomic ranks within a single cross-validation partition.

**Controlling evolutionary relatedness in biological benchmarks.** Robust evaluation of generalization requires evidence that is robust across many sequences, taxonomic levels, and partitions. Broad evidence for learning biological rules is subject to two sides of the same coin: memorization and

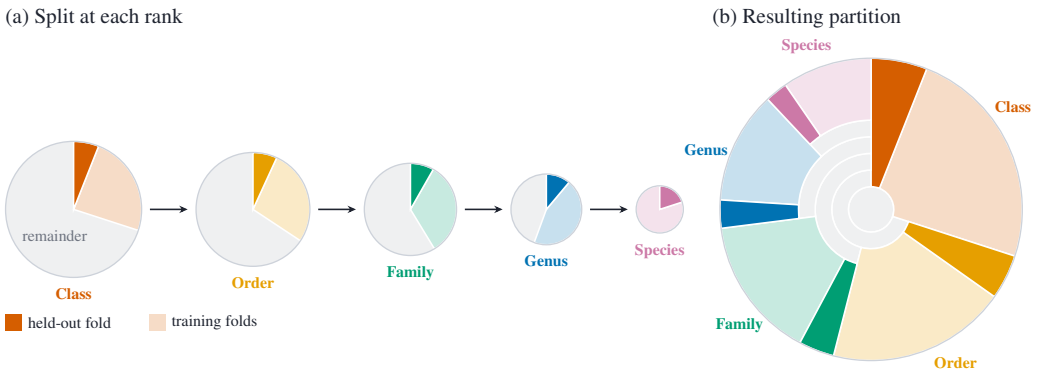



Figure 2: **Five holdout ranks in a single partition.** (a) Units are assigned to their coarsest held-out rank, from class to species; grey shows the remainder passed to finer ranks. A held-out clade lies entirely in one fold, with a same-rank sister clade represented outside that fold. (b) The resulting rank groups are disjoint. Within each group, saturated colour marks the held-out fold and lighter shades mark the four training folds.

generalization. Generalization is a model's ability to apply learned rules to make accurate predictions on new, unseen data. Memorization is the opposite: when a model only reproduces specific training examples. To distinguish memorization from generalization in biology, multiple deleakage methods have been developed to minimize the similarity between the train and test set, including GraphPart, SpanSeq, and DataSAIL (Teufel et al., 2023; Ferrer Florensa et al., 2024; Joeres et al., 2025). All these methods rely on minimizing similarity, as measured by homology, $k$-mer, or other metrics. However, due to the nature of biological sequences, minimizing similarity for a chosen partition does not necessarily fully deleak the sequence. An alternative to this approach is to quantify generalization across taxonomy as a gradient rather than an absolute number.

**Metagenomic biosurveillance of novel viral threats.** Generalization across taxonomy is critical for novel virus detection. Pandemic early-warning systems necessitate the detection of never-before-seen sequences. Current efforts pursue this goal via wastewater metagenomics (Rushford et al., 2025; Justen et al., 2026). However, read-based viral identification remains dependent on similarity to reference sequences (Kantor et al., 2026). Learned representations may therefore enable detection when similar labeled examples are unavailable. Some supervised tools, like DeepVirFinder can be used to identify viral sequences from nucleotide sequence (Ren et al., 2020). There are also a number of benchmarks and tools for virus detection. LAMBDA compares genomic models on detecting phages in bacterial genomes (Lindsey et al., 2026), ViroBench assesses viral genomic tasks using genus-disjoint and temporal splits (Ye et al., 2026), and vGUE includes a short-read viral classification task with frozen embeddings and nested class-stratified cross-validation, without reported taxonomic separation between downstream training and test folds (Rancati et al., 2025). These benchmarks do not systematically evaluate short-read viral detection across progressively broader taxonomic exclusions.

# 3 WAVEBENCH

## 3.1 CORPUS AND TASK

We use 18 wastewater metagenomic samples from the CASPER dataset (Justen et al., 2026), collected at 14 sites in nine US states over 275 days and sequenced as paired-end RNA-seq libraries (cDNA), amounting to 23.1 billion 151-bp-long sequenced reads (BioProject PRJNA1247874; Figure S2). Upstream nucleotide-MinHash deduplication, run separately within each of the 2,248 shards into which the runs were randomly divided, removed 72% of the reads. Because it never compared reads across shards, an estimated 45% of the 6.42 billion reads that remain are still exact copies of another read, mostly duplicates of one sequenced molecule within a run; the evaluation set therefore admits each sequence once and at most one read per sequenced fragment (Section 3.4). Then, we labeled the reads using Kraken2 PlusPFP dataset (`k2_pluspfp_20251015`; Langmead, 2025). We used

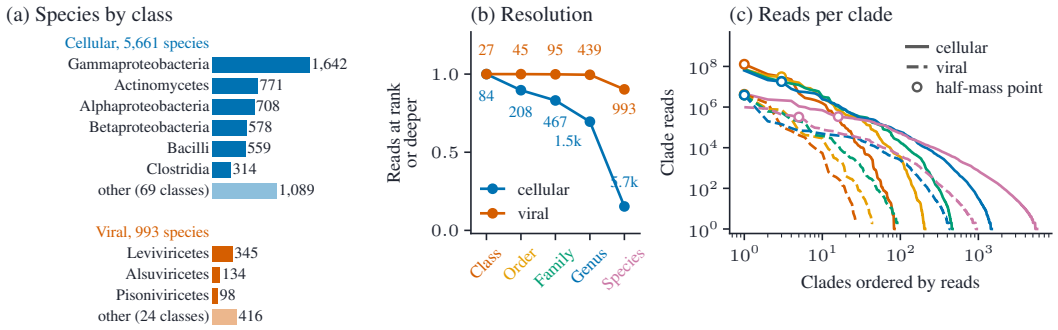



Figure 3: **Read counts are concentrated in a few clades.** The plots summarize cellular and viral assignment units at class or deeper (Appendix B). (a) Species counts by class. Classes containing at least 5% of a subset's species are named; the rest are grouped as "other". (b) Fraction of each subset's unit reads assigned at the indicated rank or deeper, with clade counts beside the points. (c) Reads per clade, ordered from largest to smallest within each rank. Rank colours match the labels in (b); solid lines are cellular and dashed lines viral. Each open circle marks where the cumulative count reaches half of that rank's reads.

a strict confidence threshold of 0.8 to favor label precision, to get 534.8 million annotated reads (Wood et al., 2019); restricting to the NCBI Viruses and cellular organisms roots (taxids 10239 and 131567) leaves 520.3 million reads. Restricting these to ranked assignment units leaves 305.8 million reads. In order to have sufficient data to split into five folds, we decided to stratify the partitions over five ranks: class, order, family, genus, and species. So, after the sister-class exclusion, we retained the 249.5 million reads annotated at these ranks or lower (Figure 3). Reads that Kraken2 leaves unclassified are excluded, so the task does not measure detection among unclassified sequences.

### 3.2 Hierarchical holdouts

We seek one five-fold partition in which every taxonomic rank at each held-out fold has novel clades relative to the rest of the partition (Figure 2). Novelty is simulated by withholding entire taxa from training folds, measuring transfer to unseen taxa as a prerequisite for flagging uncharacterized viruses. This tells apart a model able to recognize novel viruses from a model that can recognize any novel data, and measures how well a model reasons over data not seen in the downstream training set. The metagenomic reads assigned directly to the same taxon are quantized into *assignment units*, which are never split across folds. A rank-$r$ clade is *confined* to fold $f$ when all its descendant units lie in $f$ but a sibling under its immediate parent appears in one of the other folds (Figure 4). Each unit is assigned to the coarsest rank in which it is confined, so the five held-out ranks are disjoint. A *stratum* is one rank within one held-out fold, counted separately for viral and cellular reads.

### 3.3 Fold construction algorithm

**Objective.** We want to maximize the informative statistical power at each stratum. Each stratum contains clades, which are composed of reads. We assume reads from each clade are correlated, so we value a confined clade of $n_c$ reads by its effective sample size

$$u(n_c, \rho) = \frac{n_c}{1 + (n_c - 1)\rho}, \qquad \rho \in \{0, 0.001, 0.01, 0.05, 0.1, 1\}, \tag{1}$$

where $\rho$ is the within-clade correlation. As $\rho$ is unknown, we score assignments over a grid from $\rho = 0$ (counting reads) to $\rho = 1$ (counting clades). To compare ranks, each stratum's value is divided by its even-spread share, one fifth of its rank's total candidate value. The objective is the harmonic mean of these 150 ratios (5 ranks $\times$ 5 folds $\times$ 6 values of $\rho$), which is dominated by its smallest terms and so rewards strengthening the thinnest stratum over enriching well-covered ones.

**Fold construction.** We build the folds in two steps: a feasibility step and an optimization step. In the first stage, the objective is to ensure the minimum number of clades per strata are all satisfied.

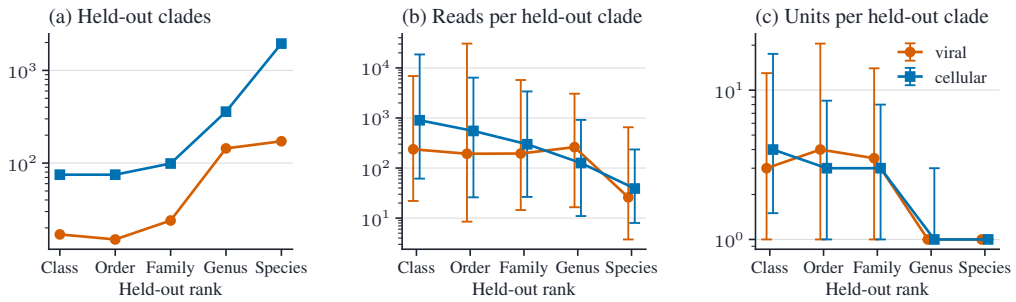



Figure 4: **Clade counts and sizes in the constructed folds.** (a) Total confined clades across the five folds at each rank. (b) Reads per clade and (c) assignment units per clade, shown as medians with interquartile ranges across clades. Orange denotes viral clades and blue cellular clades. All vertical axes use a log scale.

Every held-out rank in every fold is guaranteed at least 1,500 corpus reads and at least 3 viral or 15 cellular confined clades. Under the sister-class rule only 84 cellular classes can be withheld across the 5 folds, which bounds the cellular floor; 15 is the largest value tested that the corpus supports. The second step improves that assignment through simulated annealing. Each step moves one unit, moves one confined clade, or swaps two clades between folds; steps that would break a floor are rejected. Additionally for the second step, we use simulated annealing to occasionally accept worse steps and escape local optima. We run 50 seeded trajectories of 200,000 steps each and keep the assignment with the best objective value.

**Split validity.** A separate auditor re-derives every confined clade, rank–fold floor and objective value from the Kraken2 report and the released assignment. The released split passes the audit, and a copy with one of its four files removed is rejected (Appendix A.2).

### 3.4 Evaluation dataset

We evaluate on 200,000 reads sampled from the partition: 20,000 per label and fold, split equally across the strata. Within each stratum, reads are drawn round-robin over the confined clades, and within a clade by cycling over its assignment units in a seeded order, one read per unit per turn. At most one read is taken from each sequenced fragment, never both mates, and no sequence or its reverse complement occurs twice in the set; a clade whose pool holds fewer eligible reads than its share contributes all of them, and the remainder goes to the other clades of the stratum. No evaluation read is assigned to a metazoan taxon. The evaluation reads represent all 2,927 confined clades (Table 1), although their read counts differ; Appendix B (Figure S1) gives the read and clade inventory of every rank-by-fold cell before sampling. Because reads within a clade are correlated to a certain extent, we resampled at the clade level for confidence intervals. Nonetheless, the higher the rank the smaller the number of clades: for viruses, each fold contains only 3–4 held-out classes and 3 held-out orders.

Table 1: **Evaluated clades by rank.** Minimum–maximum across the five folds.

| Rank | Viral clades | Cellular clades |
|---|---|---|
| Class | 3–4 | 15 |
| Order | 3 | 15 |
| Family | 4–6 | 18–22 |
| Genus | 25–34 | 65–75 |
| Species | 26–43 | 348–466 |

## 4 Evaluation protocol

**Probing and evaluation.** The benchmark asks how well a gLM's embeddings can be used to separate viral from cellular reads when all related reads at the same taxonomic rank are missing from the labelled training set. We embed the same 151-bp reads across all 51 models and fit logistic probes on frozen embeddings using five-fold cross-validation. For fold $k$, each probe trains on the other four folds' evaluation reads, with every training read counted once, and scores all five held-out ranks

| Model | Leaderboard AUROC [95% CI] | AP |
|---|---|---|
| evo2_1b_base | 0.928 [0.911, 0.941] | 0.730 |
| METAGENE-1 | 0.916 [0.899, 0.932] | 0.727 |
| evo2_7b_base | 0.908 [0.891, 0.921] | 0.633 |
| evo2_7b_262k | 0.907 [0.889, 0.922] | 0.644 |
| evo2_7b | 0.906 [0.890, 0.921] | 0.631 |
| NTv3_650M_pre | 0.850 [0.825, 0.872] | 0.580 |
| *Reference baselines* | | |
| 5-mer control | 0.801 [0.777, 0.824] | 0.467 |
| DeepVirFinder | 0.827 [–, –] | – |

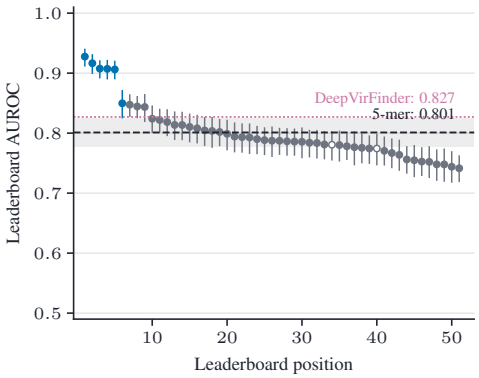



Table 2 and Figure 5: **Model performance relative to composition and DeepVirFinder.** Left: the six leading models and reference baselines, with leaderboard AUROC and average precision (AP). Right: all 51 models ordered by leaderboard AUROC, with 95% clade-bootstrap intervals. Horizontal lines mark the baselines; the grey band is the 5-mer interval. Blue highlights the six models listed in the table, not statistical significance. Open circles mark checkpoints flagged for numerical instability. Dashes denote unavailable estimates.

in fold $k$.[1] The leaderboard averages the five per-rank clade-balanced AUROCs, weighting ranks equally and, within each label, fold and rank, weighting clades equally and units equally within each clade. All models' pretraining corpora were compiled before these samples were collected, but clades withheld from the probe's folds may still appear in pretraining. A clade's AUROC is computed from that clade's reads against every read of the opposite label in the same rank and fold, and checkpoints are grouped into families by pretraining lineage as listed in Table S6. Discrimination is always reported clade-balanced; detection requires one threshold per fold and is therefore reported on pooled scores, so the two families of numbers are not interchangeable (Appendix I). Model-minus-control differences use paired clade bootstrapping (Appendix C.2); alternative estimators are compared in Appendix I.

**Comparison methods.** We use a 5-mer probe as the composition control; on the leaderboard statistic, it is within 0.001 AUROC of the strongest composition feature tested, the 6-mer (Table S3). We retrain DeepVirFinder (Ren et al., 2020), a CNN on one-hot read sequences, on each fold's training reads and score that fold's held-out reads, following its authors' published training procedure and using their released code unchanged. For alignment (restricted BLAST), we run `blastn` (Camacho et al., 2009) on every evaluation read against a reference rebuilt from the assemblies behind the Kraken2 database, excluding each (rank, fold) stratum's held-out clades. A read's score is its best viral minus best cellular bit score, with a no-hit side contributing 0 (Appendix G).

**A representative sample.** The evaluation reads balance clades, so they do not show how a probe behaves on reads in their natural proportions. For each fold we therefore draw a representative sample of the labelled corpus: 50,000 reads, uniformly at random with a fixed seed, from all Kraken2-classified deduplicated corpus reads whose taxon is an assignment unit of that fold, viral and cellular together; taxa then appear in proportion to their abundance in the corpus, with no clade balancing. A read is dropped if it shares a read identifier, a sequence or a sequencing fragment with the other folds' evaluation reads, on which that fold's probe was trained, and the sample keeps one read per sequence and per fragment. The same fitted probes score these reads. For each fold and each FPR target ($10^{-2}$, $10^{-3}$ and $10^{-4}$), the threshold is set on the fold's benchmark cellular reads and applied unchanged to its representative sample, and sensitivity and false-positive rate are averaged over the five folds with equal weight. At a viral share $\pi$, a detector with sensitivity $t$ and false-positive rate $f$ has precision $\pi t/[\pi t + (1-\pi)f]$. The sample holds only reads classified by Kraken 2. Figure 8 assumes unclassified reads are cellular and assumes the probe flags them at the rate it flags classified cellular reads.

[1] Appendix C gives the probe and evaluator details.

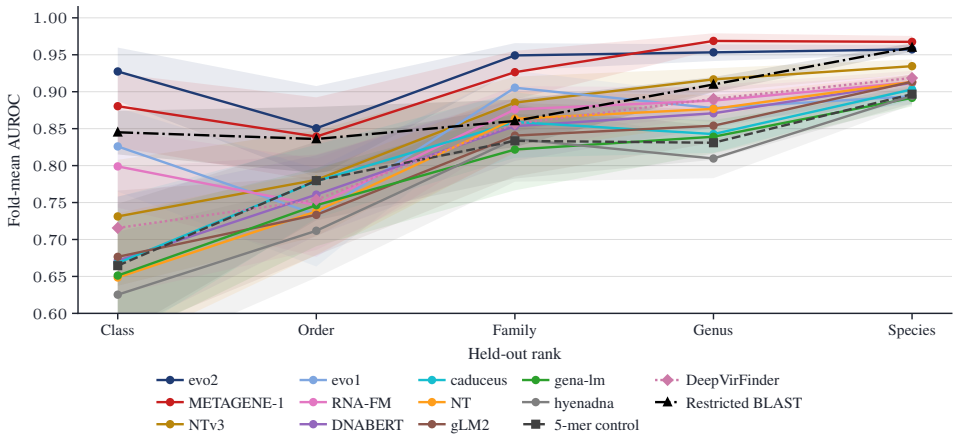



Figure 6: **Discrimination across held-out ranks.** Clade-balanced fold-mean AUROC for one representative per model family (200,000 reads); bands are 95% clade-bootstrap intervals. Dashed grey: 5-mer control; pink: DeepVirFinder; black: restricted BLAST. Ranks contain different clades, so the curves do not isolate a distance effect.

## 5 RESULTS

**Generalization.** AUROC is higher at species than at class, and varies more across models at broader ranks (Figure 6). Median AUROC decreases monotonically from lower to higher ranks (0.900, 0.842, 0.841, 0.729, 0.649 for species, genus, family, order and class, respectively). Across the 51 models, clade-balanced AUROC ranges from 0.83–0.97 at species (median 0.90), compared with 0.59–0.93 at class (median 0.65). Only 19 of 51 models outperform the baseline's leaderboard AUROC of 0.801 (Table 2 and Figure 5). Only 9 of 51 models outperform the fold-retrained DeepVirFinder CNN (leaderboard AUROC 0.827); for this dataset and labelled-data budget, most frozen representations do not generalize better than this CNN trained from scratch. Evo 2 1B and METAGENE-1 are the best performing models. Evo 2 1B with a leaderboard AUROC of 0.928 (0.957 at species, 0.927 at class) and METAGENE-1 with a leaderboard AUROC of 0.916 (0.967 at species, 0.880 at class), demonstrating that their frozen representations support discrimination even when broad taxonomic groups are absent from probe training. Evo 2 1B exceeds METAGENE-1 by $+0.011$ on the leaderboard statistic (95% paired clade-bootstrap interval $[+0.001, +0.021]$). Models tend to struggle with the same cellular clades: median pairwise Spearman correlations in per-clade AUROC range from 0.83 to 0.88 across ranks, after averaging each pair's correlations across folds. The corresponding median pairwise Spearman correlations for viral clades range from 0.40 to 0.76. This difficulty is also strongly correlated with the baseline probe, with median model–5-mer Spearman correlations in per-clade AUROC of 0.739 to 0.825. Model rankings varied depending on the taxonomic level evaluated. When doing paired clade resampling, models agree most strongly among species, genus and family holdouts (median Spearman $\rho = 0.90$) and least between class and the finer ranks (0.57), a gap of 0.34 $[0.08, 0.57]$. So models vary in performance depending on the clades' distance from the training set. Generalization to held-out clades is therefore uneven, and dependent on the clades. Most models' performance did not extend beyond a simple what a simple 5-mer baseline provided.

**Model family performance.** Generalization correlated most with model family. The 11 families account for 90% of the variance across models, and their best checkpoints span 0.15 AUROC, while no within-family difference between models moved performance by more than 0.05 (Figure S4a). Within the same family larger models were not necessarily better: Evo 2 1B (0.928) exceeds the Evo 2 7B base model (0.908), while NTv3 gains 0.032 from 8M to 650M parameters. Longer pretraining context also did not improve transfer: Evo 1 falls from 0.847 at 8k tokens to 0.822 at 131k, and Evo 2 7B scores 0.908 at 8k and 0.906 at 1M tokens. Changing the pretraining genomes moves a model in either direction: the 2.5B Nucleotide Transformer rises from 0.778 on human genomes to 0.808 on multiple species, and GENA-LM falls from 0.784 on the human genome to 0.741 on yeast. Variation between model families makes it difficult to isolate performance to a single knob.

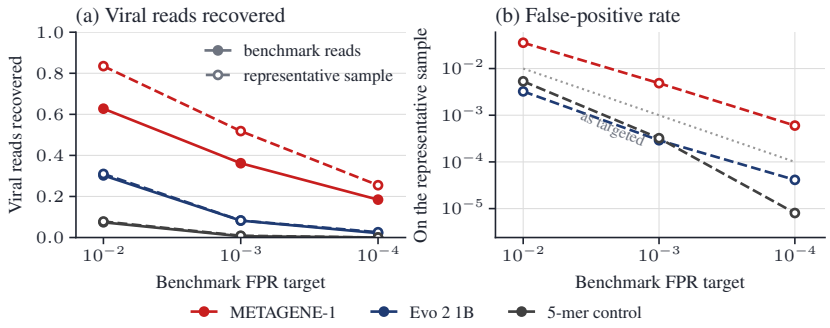



Figure 7: **Benchmark thresholds transfer to the representative sample.** Each fold's thresholds are set on its benchmark cellular reads at the stated FPR target and applied unchanged to 50,000 classified reads drawn from that fold's taxa in the representative sample (Section 4); values average the five folds with equal weight. (a) Share of viral reads recovered on benchmark (solid) and representative-sample (dashed) reads. (b) False-positive rate on representative-sample cellular reads; the dotted line marks the target.

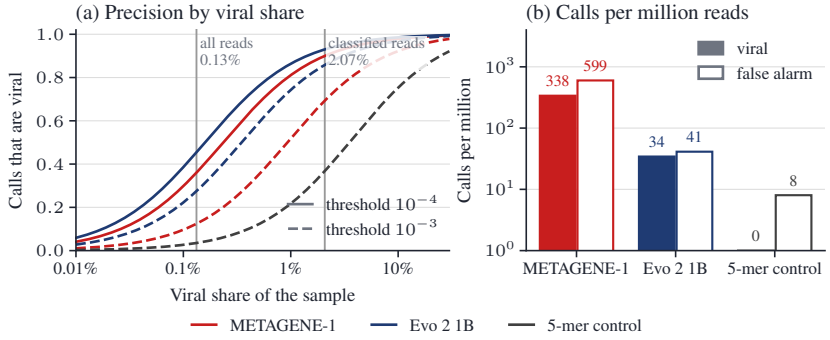



Figure 8: **What the detectors would report in wastewater.** Both panels apply the representative-sample sensitivity and false-positive rate of Figure 7, held fixed across viral shares. (a) Share of calls that are viral against the viral share of a sample, at the $10^{-4}$ (solid) and $10^{-3}$ (dashed) thresholds. Vertical lines mark the viral share of all deduplicated reads, counting unclassified reads as non-viral, and of classified reads in the evaluated taxa. The 5-mer control recovers no viral read at $10^{-4}$ and has no curve there. (b) Viral calls (filled) and false alarms (open) per million deduplicated reads at the all-read viral share and the $10^{-4}$ threshold.

**Comparison to baselines.** As novelty increases, restricted BLAST loses almost all its sensitivity, while both learned methods degrade far less (Figure S4b). At 1% FPR, with reads pooled within each rank, restricted BLAST recovers 53.58% of viral reads when only the species is withheld but 2.92% when the class is withheld, because 61% of class reads find no hit. From species to class, DeepVirFinder keeps 40% of its sensitivity (26.1% to 10.4%) and METAGENE-1, the model with the highest detection AUROC pooled across folds (0.960, against 0.949 for Evo 2 1B), keeps 48% (84.3% to 40.0%), against 5% for restricted BLAST. Learning from labelled reads thus avoids the collapse (diagnostics in Figure S6), and pretraining adds level: at class, METAGENE-1 recovers about 4 times as many viral reads as DeepVirFinder. At 1% FPR, only 1 of the 51 models beats restricted BLAST at species, but 48 beat it at class. Pooled over all ranks and folds, METAGENE-1 recovers more viral reads than restricted BLAST at every FPR target, by 44.3 $[36.6, 51.2]$, 25.9 $[16.9, 35.9]$ and 15.6 $[8.2, 24.3]$ percentage points at $10^{-2}$, $10^{-3}$ and $10^{-4}$ (paired 95% clade-bootstrap intervals).

**Application to biosurveillance.** The benchmark balances clades; wastewater does not. We therefore scored each fold's probe on a second sample of classified reads drawn in the representative sample (Section 4). Viral reads are rare and concentrated: they are 2.07% of the classified reads in these taxa and 0.13% of all deduplicated reads, and a single clade holds half of them at every rank from class to genus (Figure 3c). Detection transfers without re-tuning (Figure 7). At thresholds set on benchmark

reads for FPR targets of $10^{-2}$, $10^{-3}$ and $10^{-4}$, Evo 2 1B recovers 31.0, 8.3 and 2.6% of viral the representative sample, against 30.2, 8.2 and 2.1% on the benchmark. METAGENE-1 recovers more in the representative sample than on the benchmark (83.5, 51.8 and 25.5%, against 62.8, 36.2 and 18.5%), at a false-positive rate above target: 0.060% at the $10^{-4}$ target, 6 times the target. To express these rates as a surveillance program would see them, we apply them at a given viral share (Figure 8). At the viral share of all deduplicated reads, 36% of METAGENE-1's calls at the $10^{-4}$ threshold are viral: per million reads it makes 338 viral calls and 599 false alarms. Evo 2 1B is somewhat more precise (45%) but makes 34 viral calls, about 10 times fewer, and the composition control makes no viral call and 8 false alarms. Among classified reads alone, METAGENE-1's precision at the same threshold rises to 90%. METAGENE-1 thus finds about ten times more viral reads than Evo 2 1B, and composition finds none, but at the viral share of the full read stream its calls still need confirmation. These projections assume sensitivity and false-positive rate do not change with viral share.

## 6 Discussion and limitations

**Discussion.** gLMs are meant to yield general-purpose representations. Nonetheless, when separating viral from cellular reads across increasingly distant held-out taxa, less than half of representations outperform a 5-mer composition baseline ($19/51$) or a previously published model retrained on the same folds ($9/51$). However, among the best-performing models, it appears some useful information is embedded: given the 200,000-read dataset, probes trained on them outperformed a robust supervised baseline. Generalization is jagged and clade dependent: certain clades were difficult to generalize to across all models . Models strongly correlated with the composition baseline, hinting probes likely mostly capture composition. It is challenging to tease apart the effect of pretraining on performance. The best-performing models (METAGENE-1, Evo 2) were both pretrained on microbe-rich corpora. Models' predictions agree more on cellular than viral clades ($\rho = 0.83$–$0.88$ vs $0.40$–$0.76$), possibly because they saw more cellular genomes. These could hint that models are still forming connections that carry over to related taxa. An intreresting future direction would be to apply the same algorithm to pretraining data, and measure transfer to withheld taxa.

Overall, the benchmark demonstrates that AUROC falls as ranks get coarser. Fine-rank rankings don't predict class rankings ($\rho = 0.90$ vs $0.57$), so necessitating the need for holding out higher ranks than only species and genera, the current standard. However, coarse-rank resolution is set by the corpus: the dataset contained only $3 - -4$ viral classes per fold. A richer dataset gives more clades and smoother curves. Most models don't generalize, though some had biosurveillance value. BLAST search aline collapses when the class is withheld ($48/51$ models beat it at 1% FPR); learned models don't. Probes transfer to the representative sample without re-tuning. METAGENE-1 was able to have 25.5% recall on the representative sample at the $10^{-4}$ threshold, but at the viral share of the full read stream, METAGENE-1's calls are mostly false (36% viral at $10^{-4}$), providing a potentially useful stage in a screening pipeline.

**Limitations.** We use Kraken2 to define the labels and holdouts, without independent biological validation. We probe frozen mean-pooled embeddings linearly, which could miss more complex signals the representations may hold. Clade balancing limits abundance dominance but does not make the held-out taxa representative of all unseen viruses. The benchmark alone does not demonstrate what the model has learned.

## 7 Conclusion

We propose an algorithm for a 5-fold cross-validation split for graded evaluation of genome language models at multiple taxonomic ranks at once, constructed directly from metagenomic samples. The fold construction algorithm works in any hierarchy. We use it to determine that how well a representation transfers depends strongly on which clades are withheld: most of the 51 models fall below a 5-mer composition baseline, while METAGENE-1 and Evo 2 retain a margin even when whole classes are withheld from probe training. As detectors they exceed a BLAST search or DeepVirFinder, a supervised viral detector retrained on the same data. It demonstrates that generalization to distant taxa is jagged, but models still provide applications in metagenomic viral read detection.

## REPRODUCIBILITY STATEMENT

Both dataset and code are anonymous and downloadable. The dataset (`https://huggingface.co/datasets/WavebenchICLR/wavebench-ICLR`) holds the released split, the 200,000 evaluation reads with their sequences, the frozen embeddings of every model on those reads, the probe training weights, and the held-out per-read scores and probe records of every model, control and baseline. The code (`https://anonymous.4open.science/r/wavebench-ICLR-81EF`) contains the fold-construction algorithm, its auditor, the embedding and probe code, and a script that recomputes the evaluation statistics in this paper from the released scores, refits the training weights and the composition controls from the released reads and any model's probe from its released embeddings, and compares every recomputed value with the released one. In the paper, the fold-construction algorithm is specified in Section 3.3 and Appendix A and its auditor in Appendix A.2; the evaluation reads in Section 3.4; the probe, composition controls and estimators in Section 4 and Appendix C; the restricted BLAST baseline in Appendix G; and the corpus processing in Appendix B. Appendix J records the licences and the limits of reproduction.

## ETHICS STATEMENT

WAVEBench is built from 18 wastewater samples deposited in the NCBI Sequence Read Archive (BioProject PRJNA1247874) by the CASPER program, without use restrictions. We redistribute only the selected evaluation subset. The data carries no individual-level metadata, and no human-subjects research was conducted. Labels are Kraken2 assignments, not clinical or diagnostic findings, and the prevalence scenarios are neither prevalence estimates for the contributing communities nor operational surveillance recommendations. Embeddings follow the licence of the model that produced them (Appendix J).

## AI USE STATEMENT

We used generative AI to assist with software development, data processing, and writing the first draft of the manuscript. These uses included implementing code for fold construction, auditing, read collection and filtering, read selection, embedding, and evaluation; cleaning and reformatting data; interpreting results; preparing figures and tables from saved outputs; drafting and restructuring manuscript sections; and checking for consistency. We did not use generative AI to generate synthetic datasets, or develop conceptual frameworks. AI was only used to consult and implement, and did not independently come up with hypotheses and methods or interpretations. The authors reviewed all AI-assisted work, and take full responsibility for the final manuscript and all associated claims, code, data, and other artifacts.

## REFERENCES


Judith Bernett, David B Blumenthal, Dominik G Grimm, Florian Haselbeck, Roman Joeres, Olga V Kalinina, and Markus List. Guiding questions to avoid data leakage in biological machine learning applications. *Nature Methods*, 21(8):1444–1453, 2024. doi: 10.1038/s41592-024-02362-y.

Sam Boshar, Benjamin Evans, Ziqi Tang, Armand Picard, Yanis Adel, Franziska K. Lorbeer, Chandana Rajesh, Tristan Karch, Shawn Sidbon, David Emms, Javier Mendoza-Revilla, Fatimah Al-Ani, Evan Seitz, Yair Schiff, Yohan Bornachot, Ariana Hernandez, Marie Lopez, Alexandre Laterre, Karim Beguir, Peter Koo, Volodymyr Kuleshov, Alexander Stark, Bernardo P. de Almeida, and Thomas Pierrot. A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*, 2025. doi: 10.64898/2025.12.22.695963.

Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10: 421, 2009. doi: 10.1186/1471-2105-10-421.

Wenduo Cheng, Zhenqiao Song, Yang Zhang, Shike Wang, Danqing Wang, Muyu Yang, Lei Li, and Jian Ma. DNALongBench: A benchmark suite for long-range DNA prediction tasks. *Nature Communications*, 16:10108, 2025. doi: 10.1038/s41467-025-65077-4.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025. doi: 10.1038/s41592-024-02523-z.

Joseph Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15, 1985. doi: 10.1086/284325.

Haonan Feng, Lang Wu, Bingxin Zhao, Chad Huff, Jianjun Zhang, Jia Wu, Lifeng Lin, Peng Wei, and Chong Wu. Benchmarking DNA foundation models for genomic and genetic tasks. *Nature Communications*, 16(1):10780, 2025. doi: 10.1038/s41467-025-65823-8.

Alfred Ferrer Florensa, Jose Juan Almagro Armenteros, Henrik Nielsen, Frank Møller Aarestrup, and Philip Thomas Lanken Conradsen Clausen. SpanSeq: similarity-based sequence data splitting method for improved development and assessment of deep learning projects. *NAR Genomics and Bioinformatics*, 6(3):lqae106, 2024. doi: 10.1093/nargab/lqae106.

Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25, 2023. doi: 10.1186/s12863-023-01123-8.

Eyal Hadad, Noia Kogman, Lina Golan, Anva Avraham, Reut Ben-Hamo, Zhi Wei, Lior Rokach, and Isana Veksler-Lublinsky. Optimizing genomic language models for promoter prediction: a comparative study of tokenization and cross-species learning. *NAR Genomics and Bioinformatics*, 8(1):lqag025, 2026. doi: 10.1093/nargab/lqag025.

Gavin Hearne, Mohammad Saleh Refahi, Robi Polikar, and Gail L. Rosen. Improving genomic language model reliability under distribution shift. *bioRxiv*, 2026. doi: 10.64898/2026.03.19.712858. Preprint.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37 (15):2112–2120, 2021. doi: 10.1093/bioinformatics/btab083.

Shiyu Jiang, Xuyin Liu, and Zitong Jerry Wang. Evaluating DNA function understanding in genomic language models using evolutionarily implausible sequences. *ACS Synthetic Biology*, 15(6): 2256–2263, 2026. doi: 10.1021/acssynbio.6c00024.

Roman Joeres, David B. Blumenthal, and Olga V. Kalinina. Data splitting to avoid information leakage with DataSAIL. *Nature Communications*, 16:3337, 2025. doi: 10.1038/s41467-025-58606-8.

Lennart J. Justen, Clayton Rushford, Olivia S. Hershey, Róisín Floyd-O'Sullivan, Simon L. Grimm, William J. Bradshaw, Harmon Bhasin, Daniel P. Rice, Katherine Stansifer, Jo D. Faraguna, Michael R. McLaren, Alessandro Zulli, Alejandro Tovar-Mendez, Emma E. Copen, Kristen K. Shelton, Ayaaz Amirali, Sherin Kannoly, Sofia Pesantez, Aiden Stanciu, Iñigo Caballero Quiroga, Leopolda Silvera, Nicole Greenwood, Barbra Bongiovi, Austin Walkins, Ryan Love, Scott Lening, Kaylyn Patterson, Theresa Johnston, Sandra Hernandez, Aymara Benitez, Billie Jo McCarley, Samantha Engelage, Suguna Pillay, Cindy Calender, Brent Herring, Carey Robinson, Monett Wastewater Treatment Plant, Columbia Missouri Wastewater Treatment Plant, Daniel Cunningham-Bryant, Gordon Adams, Jillian Paull, Jamie Devlin, Vamsi Thiriveedhi, Sarah E. Turbett, Jacob E. Lemieux, Rose S. Kantor, David H. O'Connor, John J. Dennehy, Rachel Poretsky, Jason A. Rothman, Helena M. Solo-Gabriele, Jason R. Vogel, Pardis C. Sabeti, Jeff Kaufman, and Marc C. Johnson. Deep untargeted wastewater metagenomic sequencing from sewersheds across the United States, 2026. medRxiv preprint.

Yoav Kan-Tor, Michael Morris Danziger, Eden Zohar, Matan Ninio, and Yishai Shimoni. Does your model understand genes? a benchmark of gene properties for biological and text models, 2024.

Rose S. Kantor, Migun Shakya, Nelson Ruth, Jason A. Rothman, Clayton Rushford, Devon A. Gregory, Aidan Epstein, Jeff T. Kaufman, Jonathan E. Allen, Patrick S. G. Chain, David H. O'Connor, and Marc C. Johnson. A genome-resolved view of the wastewater RNA virome, 2026. medRxiv preprint.

Chia-Hsiang Kao, Evan Trop, McKinley Polen, Yair Schiff, Bernardo P. de Almeida, Aaron Gokaslan, Thomas Pierrot, and Volodymyr Kuleshov. Advancing DNA language models: The genomics long-range benchmark. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024. URL https://iclr.cc/virtual/2024/23032.

Samuel H. King, Claudia L. Driscoll, David B. Li, Daniel Guo, Aditi T. Merchant, Garyk Brixi, Max E. Wilkinson, and Brian L. Hie. Generative design of bacteriophages with genome language models. *Science*, 393(6811):eaec2657, 2026. doi: 10.1126/science.aec2657.

Ben Langmead. Kraken 2, KrakenUniq and Bracken indexes. https://benlangmead.github.io/aws-indexes/k2, 2025. PlusPFP index, version 2025-10-15.

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2 edition, 2014. doi: 10.1017/CBO9781139924801.

Joshua I Levy, Kristian G Andersen, Rob Knight, and Smruthi Karthikeyan. Wastewater surveillance for public health. *Science*, 379(6627):26–27, 2023. doi: 10.1126/science.ade2503.

LeAnn M. Lindsey, Nicole L. Pershing, Keith Dufault-Thompson, Ho-jin Gwak, Anisa Habib, Aaron Schindler, Arjun Rakheja, June L. Round, W. Zac Stephens, Anne J. Blaschke, Hari Sundar, and Xiaofang Jiang. LAMBDA: A prophage detection benchmark for genomic language models. *NAR Genomics and Bioinformatics*, 8(3):lqag103, 2026. doi: 10.1093/nargab/lqag103.

Ollie Liu, Sami Jaghouar, Johannes Hagemann, Shangshang Wang, Jason Wiemels, Jeff Kaufman, and Willie Neiswanger. METAGENE-1: Metagenomic foundation model for pandemic monitoring. *arXiv preprint arXiv:2501.02045*, 2025.

Zicheng Liu, Jiahui Li, Siyuan Li, Zelin Zang, Cheng Tan, Yufei Huang, Yajing Bai, and Stan Z. Li. GenBench: A benchmarking suite for systematic evaluation of genomic foundation models, 2024.

Frederikke Isa Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. BEND: Benchmarking DNA language models on biologically meaningful tasks. In *The Twelfth International Conference on Learning Representations*, 2024.

Masayuki Nagai, Alan E Murphy, Kaeli Rizzo, and Peter K Koo. Toward generalizable and interpretable AI in regulatory genomics. *Nature Genetics*, 58:1787–1802, 2026. doi: 10.1038/s41588-026-02670-3.

Stephen Nayfach, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, David Paez-Espino, I-Min Chen, Marcel Huntemann, et al. A genomic catalog of earth's microbiomes. *Nature Biotechnology*, 39:499–509, 2021. doi: 10.1038/s41587-020-0718-6.

Aman Patel, Arpita Singhal, Austin Wang, Anusri Pampari, Maya Kasowski, and Anshul Kundaje. DART-Eval: A comprehensive DNA language model evaluation benchmark on regulatory DNA. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 37, pp. 62024–62061, 2024. doi: 10.52202/079017-1981.

Samantha Petti and Sean R Eddy. Constructing benchmark test sets for biological sequence analysis using independent set algorithms. *PLoS Computational Biology*, 18(3):e1009492, 2022. doi: 10.1371/journal.pcbi.1009492.

Abdul Muntakim Rafi, Brett Kiyota, Nozomu Yachie, and Carl G de Boer. Characterizing homology-induced data leakage and memorization in genome-trained sequence models, 2026. bioRxiv preprint.

Simone Rancati, Pablo Arozarena Donelli, Giovanna Nicora, Laura Bergomi, Tommaso Buonocore, Michael Aaron Sy, Sakshi Pandey, Mattia Prosperi, Marco Salemi, Riccardo Bellazzi, Christina Boucher, Enea Parimbelli, and Simone Marini. Vir2vec: A genome-wide viral embedding, 2025. bioRxiv preprint.

Jie Ren, Kai Song, Chao Deng, Nathan A. Ahlgren, Jed A. Fuhrman, Yi Li, Xiaohui Xie, Ryan Poplin, and Fengzhu Sun. Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*, 8(1):64–77, 2020. doi: 10.1007/s40484-019-0187-4.

Clayton Rushford, Devon Gregory, Emma Copen, Jonathan Naydenov, Alejandro Tovar-Mendez, Po-E Li, Rose S. Kantor, Migun Shakya, Nelson Ruth, Patrick S. G. Chain, Hsinyeh Hsieh, Torin Hunter, Josh Kome, Matt Frank, Terri Lyddon, Jeff Kaufman, David H. O'Connor, and Marc C. Johnson. Untargeted longitudinal ultra deep metagenomic sequencing of wastewater provides a comprehensive readout of expected and unexpected viral pathogens, 2025. medRxiv preprint.

Felix Teufel, Magnús Halldór Gíslason, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Ole Winther, and Henrik Nielsen. GraphPart: homology partitioning for biological sequence analysis. *NAR Genomics and Bioinformatics*, 5(4):lqad088, 2023. doi: 10.1093/nargab/lqad088.

The Nucleic Acid Observatory Consortium. A global nucleic acid observatory for biodefense and planetary health. *arXiv preprint arXiv:2108.02678*, 2021.

Jacob West-Roberts, Joshua Kravitz, Nishant Jha, Andre Cornman, and Yunha Hwang. Diverse genomic embedding benchmark for functional evaluation across the tree of life. In *NeurIPS 2024 Workshops: AIDrugX*, 2024. URL https://mlanthology.org/neuripsw/2024/westroberts2024neuripsw-diverse/.

Derrick E Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome biology*, 20:257, 2019. doi: 10.1186/s13059-019-1891-0.

Dongxin Ye, Fang Hu, Han Hu, Shu Hu, Yang Tan, Wanli Ouyang, Stan Z. Li, Jie Cui, and Nanqing Dong. ViroBench: Benchmarking Nucleotide Foundation Models on Viral Genomics Tasks. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 12679–12690, 2026. doi: 10.1145/3770855.3819057.

# A CONSTRUCTION DETAILS AND AUDIT BOUNDARIES

## A.1 UNITS, CANDIDATES AND THE OBJECTIVE

The hierarchical Kraken reports define direct read counts and taxonomic ancestry. Read-bearing units are the indivisible objects placed in folds; candidate clades group the assignable descendants of canonical class, order, family, genus or species taxa. A candidate can be held out only if another candidate of the same rank under the same parent, a sister clade, has at least one unit in a training fold; reads assigned to the parent itself do not count. Confined clades are selected in coarsest-to-finest order so that their evaluated read sets do not overlap across novelty ranks.

For subset $s$, let $\mathcal{C}_{srf}$ be its confined clades at rank $r$ in fold $f$, and $\mathcal{A}_{sr}$ its candidate clades before assignment. The normalized cell value at correlation $\rho$ is

$$z_{srf\rho} = \frac{\sum_{c\in\mathcal{C}_{srf}} u(n_c,\rho)}{\frac{1}{5}\sum_{a\in\mathcal{A}_{sr}} u(n_a,\rho)}, \qquad H_s = \left[\frac{1}{150}\sum_{r,f,\rho} z_{srf\rho}^{-1}\right]^{-1}. \tag{2}$$

The denominators are fixed by the corpus and candidates, not by the candidate assignment under comparison. The $\rho$ grid is a design device spanning depth and breadth; it is not an estimate of empirical within-clade correlation. The implementation uses fixed-point contributions for deterministic scoring.

The feasibility step uses a mixed-integer feasibility solver; the optimization step improves feasible assignments through seeded, operation-limited search. Every rank–fold cell is guaranteed at least three viral or fifteen cellular clades, as reflected in Tables S1 and 1. The released method configuration lists the $50$ trajectory seeds and step budgets, which both subsets use. Seeded search does not establish uniqueness or convergence to a global optimum. Initial temperature $0.002$ and cooling rate $0.99999$ are implementation defaults rather than complete provenance by themselves. Reproduction therefore requires the matching implementation as well as the release configuration.

## A.2 SPLIT AUDIT

A separate auditor re-derives the validity of a split from the Kraken2 report and the released assignment. It re-applies the corpus filter and checks that the filter moves whole units, with their read counts, only from placed to dropped; checks that each unit has one valid fold; re-derives every confined clade, checks that each is absent from training at its labelled rank and has same-rank sister support in another fold, checks that every rank–fold cell meets its floors, and recomputes the objective values and the release checksums. It does not call the feasibility constructor or the local search, but it shares parsing and candidate data structures with them, so it is not independent of every upstream assumption, and taxonomic consistency does not validate the Kraken2 labels. The released split passes the audit, and a copy with one of its four files removed is rejected.

# B CORPUS PROVENANCE AND DISTINCT INVENTORIES

The full split inventory precedes the sampling of evaluation reads (Table S1); Table 1 in the main text summarizes the clades represented in the evaluated reads.

Table S1: **Full split inventories before evaluation sampling.** Confined-clade and read counts are minimum–maximum across all five folds of the accepted release.

| Rank | Viral clades | Viral reads | Cellular clades | Cellular reads |
|---|---|---|---|---|
| Class | 3–4 | 11,629–66,980 | 15 | 1,667,319–2,641,433 |
| Order | 3 | 29,873–265,234 | 15 | 2,547,542–18,490,429 |
| Family | 4–6 | 41,807–3,895,742 | 18–22 | 3,377,767–26,425,715 |
| Genus | 25–34 | 81,478–141,036 | 65–75 | 3,187,366–8,340,525 |
| Species | 26–43 | 86,509–159,857 | 348–466 | 927,771–2,906,070 |

The report inventory and the split assignment describe different stages. After the sister-class filter, the split assignment contains $1{,}127$ viral and $7{,}204$ cellular units, totalling $8{,}331$; the filter removes $1{,}432$ viral and $470$ cellular units.

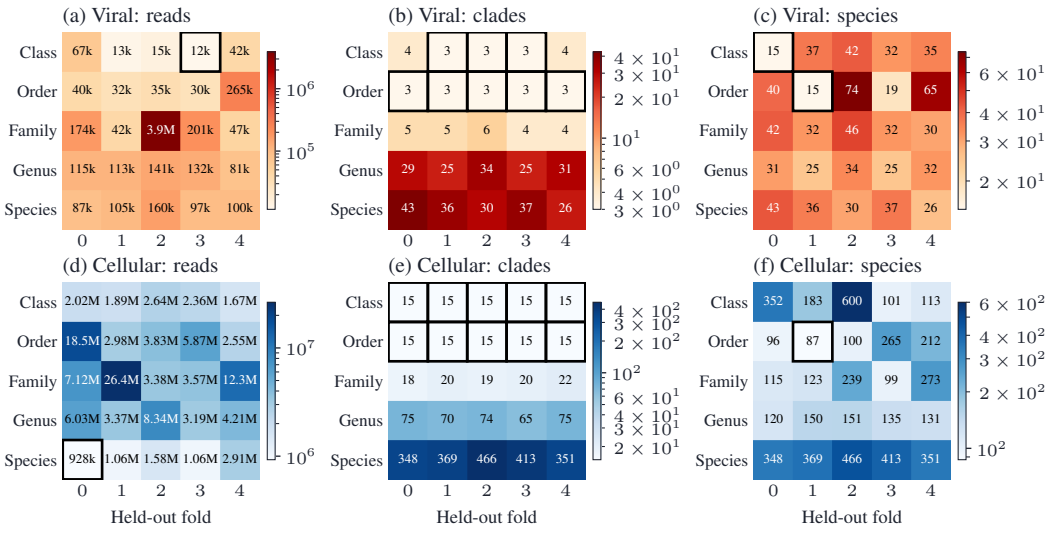



Figure S1: **Read and clade inventories by rank and fold.** Columns show reads (a, d), confined clades (b, e), and represented species (c, f), for viral (top) and cellular (bottom) subsets. Rows run from class to species within each panel. Colour scales are logarithmic; boxes mark the smallest cells, including ties. Counts precede evaluation-read sampling. All cells meet the construction floors, but their sizes remain uneven.

Figure S2 displays the source geography and collection dates. Its vector map reproduces the recovered 0.7-degree single-linkage construction: 16 mapped cohorts form 11 markers, with arithmetic-mean marker locations and areas proportional to summed deduplicated reads; Southern California and central Oklahoma lack coordinates and appear in the date panel only. The date panel is drawn from the cohort inventory. This inventory identifies eighteen cohorts across nine US states, with sixteen collected between 2025-11-19 and 2026-01-20. The dates outside this window are 2025-04-20 and 2025-09-28. Table S2 retains the upstream cohort inventory as contextual provenance, not as a denominator for the frozen benchmark: its rounded base-pair totals describe a larger pool, and the mapping between those totals, pooled portions and the released report snapshot has not been completely reconciled. The 2,248 manifest source identifiers refer to shards, not 2,248 independent biological samples. Geographic coverage and sampling dates do not establish representative coverage of wastewater communities.




Figure S2: **Source locations and collection dates.** (a) 16 of 18 cohorts, grouped into 11 markers by single linkage at 0.7°. Marker area scales with summed deduplicated reads. Ontario/Riverside, Chicago and New York combine two, three and three cohorts, respectively; six cohorts are located at city centroids. (b) Collection dates for all cohorts (2025–2026), including Southern California and central Oklahoma, which lack coordinates. Axis breaks omit unsampled intervals. Shards are not independent biological samples.

Table S2: **Upstream cohort inventory.** Rounded base-pair and deduplication figures for the larger source pool, retained as provenance context. These are not counts of the released benchmark snapshot; the mapping is not fully reconciled.

| Cohort | Dedup rate | bp before | bp after |
|---|---|---|---|
| CHI-D1_20250420 | 87% | 222B | 29B |
| CHI-D2_20250928 | 65% | 328B | 114B |
| MA_Boston_DITPN_20251119 | 68% | 133B | 42B |
| CA_Sacramento_20251120 | 63% | 249B | 92B |
| CA_PaloAlto_20251124 | 81% | 205B | 39B |
| MO_Columbia_20251125 | 80% | 176B | 36B |
| OK-B_20251202 | 35% | 180B | 117B |
| IA_Ottumwa_20251203 | 56% | 226B | 99B |
| NYC-Hospital-A_20251204 | 95% | 60B | 3B |
| NYC-Hospital-C_20251204 | 87% | 77B | 10B |
| NYC-Hospital-B_20251210 | 71% | 63B | 18B |
| CA_Riverside_20251218 | 94% | 190B | 12B |
| ID_Boise_20251229 | 59% | 173B | 71B |
| Southern_California_20251231 | 88% | 215B | 26B |
| CHI-C_20260105 | 92% | 260B | 21B |
| MO_KC_Westside_20260105 | 61% | 302B | 117B |
| FL_Miami_20260112 | 83% | 192B | 33B |
| CA_Ontario_20260120 | 63% | 237B | 87B |
| **Total** | **72%** | **3.5T** | **966B** |

**Deduplication.** The recorded upstream workflow uses nucleotide 5-gram MinHash signatures, eight bands of twenty hashes, and 64-bit hashing, with one representative per union–find component of candidate near-duplicates. For shingle-set Jaccard similarity $s$, the standard banding candidate probability is $1 - (1 - s^{20})^8$ (Leskovec et al., 2014, Section 3.4). This is a probabilistic candidate rule, not a guaranteed 90% identity threshold. A 151 bp read has 147 overlapping 5-gram positions, which need not all be distinct. The full raw/deduplicated corpora and preprocessing implementation are outside the benchmark release.

**Taxonomic labelling.** The documented workflow classifies sequence text with Kraken2 PlusPFP `k2_pluspfp_20251015`, confidence 0.8 and minimum base quality zero. Taxids 10239 and

131567 define the benchmark's viral and cellular subsets. Source row $i$ is named `read_i`; because malformed input rows can be skipped, classification records must be joined to sequences by identifier rather than output-line position. The historical classifier executable version and complete database-build provenance remain incompletely pinned. The supplied taxonomy contains `nodes.dmp` and `names.dmp`, but not `merged.dmp`; retired identifiers can therefore reduce resolved coverage.

Main-text Figure 3 shows report-level composition before the evaluation reads are drawn. Units above class (16% of cellular unit reads) are omitted from it and are never held out, and its "other" group includes species without a class in the taxonomy (33 cellular and 125 viral). From class to species, half of each rank's reads lie in $1, 3, 3, 3, 16$ cellular clades and $1, 1, 1, 1, 5$ viral clades. Viral reads are 4.4% of the reads in held-out clades. Figure S3 follows the read counts from sequencing to the evaluation set.

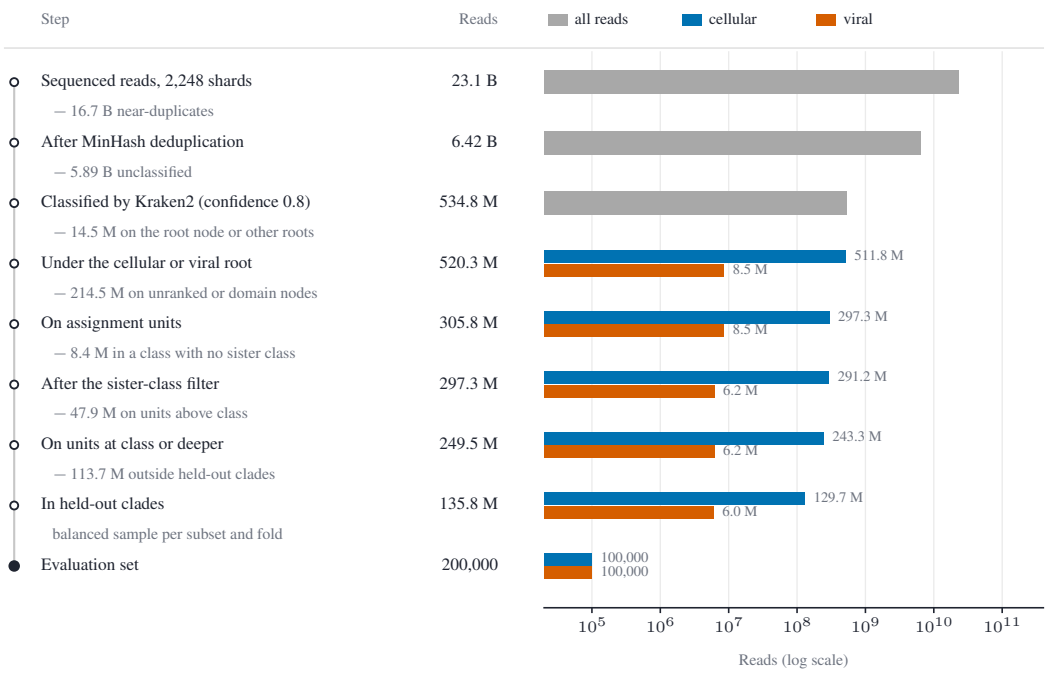



Figure S3: **Reads from sequencing to evaluation.** Rows give the reads remaining after each processing step; grey annotations between rows give the numbers removed. Counts are sequencing records: deduplication ran within each shard of a run, so an estimated 45% of the deduplicated records are exact copies of another record, whereas the evaluation set admits each sequence once. Bars show the totals on a log scale, with cellular and viral reads shown separately once the roots separate. The final row contains the 200,000 evaluation reads.

## C PROBE, CONTROLS AND STATISTICAL DEFINITIONS

Training folds alone determine each feature's mean and standard deviation; standard deviations below $10^{-8}$ are replaced with one. The standardized training data are shuffled with a seeded generator, and averaged SGD fits logistic loss for twenty epochs in batches of 8,192. Decision scores, rather than calibrated probabilities, are retained for held-out reads. The regularization coefficient is $1/n_{\text{train}}$, and the optimizer uses the `optimal` schedule, `average=True` and random state zero. Each probe trains on 160,000 reads: two labels in four folds of 20,000 reads each. The mean across models of the five-fold AUROC standard deviation is 0.0184.

**Training reads.** For fold $k$ and one label, the training reads are the other four folds' evaluation reads, and every read counts once in the probe's loss; no read is reweighted or capped. Feature standardization is unweighted.

**Operating points.** Sensitivity at a target FPR is the highest sensitivity at an attainable FPR no greater than the target; all reads tied at a score cross the threshold together, and a target below the empirical FPR resolution is undefined. At assumed prevalence $\pi$, sensitivity $t$ and FPR $f$, precision is $\pi t/[\pi t + (1-\pi)f]$, assuming both rates transfer to that population.

**Evaluator.** Embeddings use the full upstream model commits listed in the released code; embedding runs took place on RunPod, with each model's GPU and runtime recorded in its timing receipt, and hardware was not standardized across models. Model adapters produce frozen embeddings using their documented pooling paths. DNABERT-2 and DNABERT-S are loaded in native fp32 and select the PyTorch attention implementation, avoiding the custom Triton kernel that downcasts fp32 attention tensors. Evo 2 models are admitted only when TransformerEngine runs on a Hopper-class GPU; substituting a conventional linear layer is not the evaluated model. On an A40, DNABERT-2, DNABERT-S and DNA-BERT-4 were each embedded twice for the same $5{,}000$ reads; every tested read has cosine similarity $1.0000$ (including the minimum). Evo2_1b_base was checked separately on one NVIDIA H100 80GB HBM3 with native FP8: two passes over $5{,}000$ reads from the corpus give bit-identical vectors; a mismatched-pair negative control fails the same check. These checks establish agreement on the checked reads and hardware, not exhaustive cross-hardware reproducibility.

Sample-size dependence remains in the regularization and the released `optimal` schedule. The convergence diagnostic is the absolute AUROC difference between epochs ten and twenty; a small value is a practical stability check, not a proof of optimization to the global optimum. Matching the probe protocol controls the evaluation procedure, but does not ensure that each representation is equally well served by the chosen regularization or optimization budget.

GC content is a one-dimensional control; normalized $k$-mer counts have $4^k$ dimensions for $k = 3, 4, 5, 6$. Each control uses the same probe and folds as the models. All five controls are retained, because the AUROC-best control need not maximize sensitivity at a particular FPR.

Table S3: All composition controls. Mean and s.d. are computed across the five folds. Pooled AUROC is computed from the concatenated held-out decision scores.

| Control | Mean AUROC | Mean AP | Pooled AUROC |
|---|---|---|---|
| GC | $0.717 \pm 0.027$ | 0.636 | 0.716 |
| 3-mer | $0.799 \pm 0.019$ | 0.752 | 0.798 |
| 4-mer | $0.809 \pm 0.018$ | 0.765 | 0.809 |
| 5-mer | $0.822 \pm 0.020$ | 0.787 | 0.822 |
| 6-mer | $0.829 \pm 0.018$ | 0.795 | 0.829 |

### C.1 Fold averaging, pooled scores and weighting

Let $A_f$ denote AUROC on fold $f$ using only scores from its held-out probe. The fold-mean estimator is $\overline{A} = \frac{1}{5}\sum_f A_f$ with the sample standard deviation of the five values. The pooled estimator instead computes one AUROC after concatenating the five held-out score arrays. It includes cross-fold positive–negative pairs; monotone score transformations specific to a fold preserve $A_f$ but can change pooled AUROC. Per-rank AUROCs are clade-balanced fold means. The leaderboard averages these five per-rank AUROCs, giving each rank equal weight; detection uses pooled scores. Table S4 shows small positive mean fold-minus-pooled gaps, although individual gaps can be negative. Numerical comparisons must name the estimator.

Table S4: Fold-mean minus pooled-score AUROC across 51 models and five controls.

| Mean gap | s.d. of gap | Min gap | Max gap |
|---|---|---|---|
| +0.0003 | 0.0003 | -0.0001 | +0.0012 |

The per-rank estimator used by the leaderboard and novelty curves gives each held-out clade equal mass within each label, fold and rank, divides that mass equally among the clade's represented assignment units, then divides unit mass equally among its reads. Per-rank 95% confidence intervals use $2{,}000$ bootstrap draws of held-out clades with replacement within each label–fold–rank cell. Each draw recomputes the same fold-mean statistic. Random streams are independent across cells and

shared across models. The released split and fitted-probe scores remain fixed, so these intervals do not capture refitting or split-construction uncertainty.

For taxon $t$ with $n_{tfry}$ evaluated reads in fold $f$, rank $r$ and class $y$, each read receives weight $1/n_{tfry}$. Weighted AUROC averages positive–negative comparisons with the product of their weights, counting ties as one half. Each represented taxon has unit mass within its group, but groups are not additionally normalized to equal mass. This changes the estimand toward represented taxa, not all taxa in the source corpus, and does not add unseen clades or fit a new classifier.

### C.2 Clade intervals and their evaluated populations

For the paired residual analysis, confined clades are resampled with replacement within each (label, fold, rank). The model and control use the same sampled clades; fold AUROCs are averaged, and percentile intervals are computed over $2{,}000$ replicates. These intervals are conditional on the sampled clades, the released split and the fitted scores. They do not measure variation across newly constructed splits or all stages of training. The per-rank intervals have no multiple-comparison adjustment and do not provide simultaneous coverage across models and ranks. Table 1 documents coverage; viral cells at coarse ranks still contain few clades despite the clade-balanced selection.

The detection report uses a separate 2000-draw clade bootstrap, seed zero, over reads whose clade can be resolved from the provided taxonomy and exclusion specification. Unresolved reads are omitted from this bootstrap only. The full-set ROC point and bootstrap-subset point can therefore differ; Table S5 reports both alongside the conditional intervals. Resampling clades does not compensate for uncertainty in Kraken2 labels, missing taxa or unmeasured deployment shifts.

Table S5: 20k detection sensitivity (%). Intervals use 2000 clade-bootstrap draws on resolved reads; full-set and bootstrap-subset points are explicitly separated.

| Entry | FPR | Full set | Resolved subset | 95% interval |
|---|---|---|---|---|
| evo2_1b_base | 0.01 | 30.30 | 30.30 | [26.40, 34.76] |
| evo2_1b_base | 0.001 | 8.15 | 8.15 | [5.65, 10.20] |
| evo2_1b_base | 0.0001 | 1.89 | 1.89 | [1.05, 2.68] |
| 5-mer | 0.01 | 7.08 | 7.08 | [5.17, 9.29] |
| 5-mer | 0.001 | 0.10 | 0.10 | [0.00, 1.04] |
| 5-mer | 0.0001 | 0.00 | 0.00 | [0.00, 0.01] |
| Restricted BLAST | 0.01 | 18.54 | 18.54 | [14.54, 23.96] |
| Restricted BLAST | 0.001 | 9.52 | 9.52 | [5.08, 14.97] |
| Restricted BLAST | 0.0001 | 0.26 | 0.26 | [0.05, 2.25] |

## D Model inventory and full leaderboard

Main-text Figure 5 summarizes the full roster; Tables S6 and S7 list the models and their representation settings, and Table S8 gives the full leaderboard. In the registry and leaderboard, identifiers M01–M51 are reused consistently across tables. The secondary-structure variant of RNA-FM in the `multimolecule` collection is not listed separately: our adapter loads only its encoder, which is identical to RNA-FM's, so it yields the same representation. Model revisions are pinned in the released code. The per-rank, residual and weighting tables use the same identifiers.

Table S6: Complete model identifiers. Driver names describe adapter routing, not a biological taxonomy of model families. IDs are reused in every subsequent table. Licence: the model's licence as its Hugging Face card declares it, or, where the card declares none, the licence file of its repository.

| ID | Model identifier | Driver | Licence |
|---|---|---|---|
| M01 | evo2_1b_base | evo | Apache-2.0 |
| M02 | metagene-ai/METAGENE-1 | metagene | Apache-2.0 |
| M03 | evo2_7b_base | evo | Apache-2.0 |
| M04 | evo2_7b_262k | evo | Apache-2.0 |
| M05 | evo2_7b | evo | Apache-2.0 |
| M06 | InstaDeepAI/NTv3_650M_pre | nt_v2_v3 | InstaDeep OML v1.0 |
| M07 | togethercomputer/evo-1-8k-base | evo | Apache-2.0 |
| M08 | multimolecule/rnafm | rna_fm | AGPL-3.0 |
| M09 | InstaDeepAI/NTv3_650M_post | nt_v2_v3 | InstaDeep OML v1.0 |
| M10 | InstaDeepAI/NTv3_100M_pre | nt_v2_v3 | InstaDeep OML v1.0 |
| M11 | togethercomputer/evo-1-131k-base | evo | Apache-2.0 |
| M12 | InstaDeepAI/NTv3_8M_pre | nt_v2_v3 | InstaDeep OML v1.0 |
| M13 | zhihan1996/DNA_bert_3 | dnabert | Apache-2.0 |
| M14 | zhihan1996/DNA_bert_4 | dnabert | Apache-2.0 |
| M15 | kuleshov-group/caduceus-ps_seqlen-1k_d_model-256_n_layer-4_lr-8e-3 | caduceus | Apache-2.0 |
| M16 | InstaDeepAI/nucleotide-transformer-2.5b-multi-species | nt_v2_v3 | CC BY-NC-SA 4.0 |
| M17 | zhihan1996/DNA_bert_6 | dnabert | Apache-2.0 |
| M18 | tattabio/gLM2_650M | glm | Apache-2.0 |
| M19 | zhihan1996/DNA_bert_5 | dnabert | Apache-2.0 |
| M20 | tattabio/gLM2_650M_embed | glm | Apache-2.0 |
| M21 | kuleshov-group/caduceus-ps_seqlen-131k_d_model-256_n_layer-16 | caduceus | Apache-2.0 |
| M22 | tattabio/gLM2_150M | glm | Apache-2.0 |
| M23 | InstaDeepAI/nucleotide-transformer-v2-500m-multi-species | nt_v2_v3 | CC BY-NC-SA 4.0 |
| M24 | AIRI-Institute/gena-lm-bert-base-athaliana | gena_lm | MIT |
| M25 | AIRI-Institute/gena-lm-bert-base-lastln-t2t | gena_lm | MIT |
| M26 | kuleshov-group/caduceus-ph_seqlen-131k_d_model-256_n_layer-16 | caduceus | Apache-2.0 |
| M27 | AIRI-Institute/gena-lm-bert-base-fly | gena_lm | MIT |
| M28 | AIRI-Institute/gena-lm-bert-base-t2t-multi | gena_lm | MIT |
| M29 | InstaDeepAI/NTv3_100M_post | nt_v2_v3 | InstaDeep OML v1.0 |
| M30 | kuleshov-group/caduceus-ph_seqlen-1k_d_model-256_n_layer-4_lr-8e-3 | caduceus | Apache-2.0 |
| M31 | AIRI-Institute/gena-lm-bert-large-t2t | gena_lm | MIT |
| M32 | AIRI-Institute/gena-lm-bert-base-t2t | gena_lm | MIT |
| M33 | InstaDeepAI/nucleotide-transformer-v2-250m-multi-species | nt_v2_v3 | CC BY-NC-SA 4.0 |
| M34 | zhihan1996/DNABERT-2-117M | dnabert | Apache-2.0 |
| M35 | InstaDeepAI/nucleotide-transformer-v2-100m-multi-species | nt_v2_v3 | CC BY-NC-SA 4.0 |
| M36 | InstaDeepAI/nucleotide-transformer-2.5b-1000g | nt_v2_v3 | CC BY-NC-SA 4.0 |
| M37 | InstaDeepAI/nucleotide-transformer-v2-50m-multi-species | nt_v2_v3 | CC BY-NC-SA 4.0 |
| M38 | LongSafari/hyenadna-medium-160k-seqlen | hyena_dna | BSD-3-Clause |
| M39 | AIRI-Institute/gena-lm-bigbird-base-t2t | gena_lm | MIT |
| M40 | zhihan1996/DNABERT-S | dnabert | Apache-2.0 |

| ID | Model identifier | Driver | Licence |
|---|---|---|---|
| M41 | `kuleshov-group/caduceus-ps_seqlen-1k_d_model-118_n_layer-4_lr-8e-3` | caduceus | Apache-2.0 |
| M42 | `kuleshov-group/caduceus-ph_seqlen-1k_d_model-118_n_layer-4_lr-8e-3` | caduceus | Apache-2.0 |
| M43 | `LongSafari/hyenadna-tiny-1k-seqlen-d256` | hyena_dna | BSD-3-Clause |
| M44 | `InstaDeepAI/nucleotide-transformer-500m-human-ref` | nt_v2_v3 | CC BY-NC-SA 4.0 |
| M45 | `LongSafari/hyenadna-tiny-1k-seqlen` | hyena_dna | BSD-3-Clause |
| M46 | `InstaDeepAI/nucleotide-transformer-500m-1000g` | nt_v2_v3 | CC BY-NC-SA 4.0 |
| M47 | `LongSafari/hyenadna-small-32k-seqlen` | hyena_dna | BSD-3-Clause |
| M48 | `LongSafari/hyenadna-tiny-16k-seqlen-d128` | hyena_dna | BSD-3-Clause |
| M49 | `LongSafari/hyenadna-medium-450k-seqlen` | hyena_dna | BSD-3-Clause |
| M50 | `LongSafari/hyenadna-large-1m-seqlen` | hyena_dna | BSD-3-Clause |
| M51 | `AIRI-Institute/gena-lm-bert-base-yeast` | gena_lm | MIT |

Table S7: Representation metadata for the same model IDs. Dimension is observed in the final embedding receipt. Layer selection and pooling are policies from the pinned adapter source; stated fallbacks describe adapter selection policy and do not claim that an unrecorded fallback branch was observed at runtime.

| ID | Dimension | Layer selection | Pooling |
|---|---|---|---|
| M01 | 1920 | blocks.21.mlp.l3 | attention-mask mean of token representations |
| M02 | 4096 | final hidden state | attention-mask mean of token representations |
| M03 | 4096 | blocks.28.mlp.l3 | attention-mask mean of token representations |
| M04 | 4096 | blocks.28.mlp.l3 | attention-mask mean of token representations |
| M05 | 4096 | blocks.28.mlp.l3 | attention-mask mean of token representations |
| M06 | 1536 | final hidden state | attention-mask mean if token-valued; unchanged if already pooled |
| M07 | 4096 | final hidden state after model.norm (native Evo-1) | attention-mask mean of token representations |
| M08 | 640 | final hidden state | attention-mask mean of token representations |
| M09 | 1536 | post-deconvolution embedding; final-hidden fallback | attention-mask mean if token-valued; unchanged if already pooled |
| M10 | 768 | final hidden state | attention-mask mean if token-valued; unchanged if already pooled |
| M11 | 4096 | final hidden state after model.norm (native Evo-1) | attention-mask mean of token representations |
| M12 | 256 | final hidden state | attention-mask mean if token-valued; unchanged if already pooled |
| M13 | 768 | final hidden state | attention-mask mean of token representations |
| M14 | 768 | final hidden state | attention-mask mean of token representations |
| M15 | 512 | final hidden state | attention-mask mean of token representations |
| M16 | 2560 | final hidden state | attention-mask mean if token-valued; unchanged if already pooled |

| ID | Dimension | Layer selection | Pooling |
|---|---|---|---|
| M17 | 768 | final hidden state | attention-mask mean of token representations |
| M18 | 1280 | final hidden state; pooler-output fallback | attention-mask mean of final hidden state; native pooler fallback |
| M19 | 768 | final hidden state | attention-mask mean of token representations |
| M20 | 512 | final hidden state; pooler-output fallback | attention-mask mean of final hidden state; native pooler fallback |
| M21 | 512 | final hidden state | attention-mask mean of token representations |
| M22 | 640 | final hidden state; pooler-output fallback | attention-mask mean of final hidden state; native pooler fallback |
| M23 | 1024 | final hidden state | attention-mask mean if token-valued; unchanged if already pooled |
| M24 | 768 | final hidden state | attention-mask mean of token representations |
| M25 | 768 | final hidden state | attention-mask mean of token representations |
| M26 | 256 | final hidden state | attention-mask mean of token representations |
| M27 | 768 | final hidden state | attention-mask mean of token representations |
| M28 | 768 | final hidden state | attention-mask mean of token representations |
| M29 | 768 | post-deconvolution embedding; final-hidden fallback | attention-mask mean if token-valued; unchanged if already pooled |
| M30 | 256 | final hidden state | attention-mask mean of token representations |
| M31 | 1024 | final hidden state | attention-mask mean of token representations |
| M32 | 768 | final hidden state | attention-mask mean of token representations |
| M33 | 768 | final hidden state | attention-mask mean if token-valued; unchanged if already pooled |
| M34 | 768 | final hidden state | attention-mask mean of token representations |
| M35 | 512 | final hidden state | attention-mask mean if token-valued; unchanged if already pooled |
| M36 | 2560 | final hidden state | attention-mask mean if token-valued; unchanged if already pooled |
| M37 | 512 | final hidden state | attention-mask mean if token-valued; unchanged if already pooled |
| M38 | 256 | backbone hidden output (use_head=False) | attention-mask mean of token representations |
| M39 | 768 | final hidden state | attention-mask mean of token representations |
| M40 | 768 | final hidden state | attention-mask mean of token representations |
| M41 | 236 | final hidden state | attention-mask mean of token representations |
| M42 | 118 | final hidden state | attention-mask mean of token representations |
| M43 | 256 | backbone hidden output (use_head=False) | attention-mask mean of token representations |
| M44 | 1280 | final hidden state | attention-mask mean if token-valued; unchanged if already pooled |

| ID | Dimension | Layer selection | Pooling |
|---|---|---|---|
| M45 | 128 | backbone hidden output (use_head=False) | attention-mask mean of token representations |
| M46 | 1280 | final hidden state | attention-mask mean if token-valued; unchanged if already pooled |
| M47 | 256 | backbone hidden output (use_head=False) | attention-mask mean of token representations |
| M48 | 128 | backbone hidden output (use_head=False) | attention-mask mean of token representations |
| M49 | 256 | backbone hidden output (use_head=False) | attention-mask mean of token representations |
| M50 | 256 | backbone hidden output (use_head=False) | attention-mask mean of token representations |
| M51 | 768 | final hidden state | attention-mask mean of token representations |

Table S8: Full leaderboard: leaderboard AUROC (the mean over the five holdout ranks of the clade-balanced fold-mean AUROC), then the read-counted fold-mean AUROC $\pm$ s.d. and fold-mean AP, which count every held-out read once. Model IDs follow Table S6; row order is fixed by 20k leaderboard AUROC.

| ID | 20k leaderboard | 20k AUROC | 20k AP |
|---|---|---|---|
| M01 | 0.928 | $0.949 \pm 0.005$ | 0.937 |
| M02 | 0.916 | $0.961 \pm 0.015$ | 0.962 |
| M03 | 0.908 | $0.928 \pm 0.006$ | 0.899 |
| M04 | 0.907 | $0.931 \pm 0.006$ | 0.906 |
| M05 | 0.906 | $0.928 \pm 0.006$ | 0.899 |
| M06 | 0.850 | $0.891 \pm 0.015$ | 0.871 |
| M07 | 0.847 | $0.870 \pm 0.012$ | 0.803 |
| M08 | 0.844 | $0.864 \pm 0.014$ | 0.829 |
| M09 | 0.844 | $0.877 \pm 0.019$ | 0.856 |
| M10 | 0.824 | $0.857 \pm 0.014$ | 0.832 |
| M11 | 0.822 | $0.842 \pm 0.011$ | 0.772 |
| M12 | 0.818 | $0.861 \pm 0.015$ | 0.836 |
| M13 | 0.814 | $0.838 \pm 0.020$ | 0.799 |
| M14 | 0.814 | $0.838 \pm 0.021$ | 0.800 |
| M15 | 0.810 | $0.818 \pm 0.020$ | 0.782 |
| M16 | 0.808 | $0.851 \pm 0.015$ | 0.828 |
| M17 | 0.804 | $0.835 \pm 0.022$ | 0.796 |
| M18 | 0.804 | $0.831 \pm 0.020$ | 0.796 |
| M19 | 0.802 | $0.829 \pm 0.017$ | 0.791 |
| M20 | 0.799 | $0.827 \pm 0.021$ | 0.792 |
| M21 | 0.794 | $0.830 \pm 0.020$ | 0.798 |
| M22 | 0.793 | $0.827 \pm 0.022$ | 0.792 |
| M23 | 0.793 | $0.819 \pm 0.018$ | 0.784 |
| M24 | 0.790 | $0.811 \pm 0.017$ | 0.778 |
| M25 | 0.788 | $0.811 \pm 0.017$ | 0.767 |
| M26 | 0.787 | $0.823 \pm 0.022$ | 0.787 |
| M27 | 0.787 | $0.824 \pm 0.016$ | 0.789 |
| M28 | 0.786 | $0.819 \pm 0.016$ | 0.783 |
| M29 | 0.786 | $0.818 \pm 0.019$ | 0.783 |
| M30 | 0.786 | $0.805 \pm 0.023$ | 0.765 |
| M31 | 0.784 | $0.818 \pm 0.019$ | 0.778 |
| M32 | 0.784 | $0.818 \pm 0.018$ | 0.778 |
| M33 | 0.781 | $0.820 \pm 0.016$ | 0.787 |
| M34 | 0.781 | $0.811 \pm 0.018$ | 0.771 |
| M35 | 0.780 | $0.802 \pm 0.016$ | 0.762 |
| M36 | 0.778 | $0.813 \pm 0.020$ | 0.781 |
| M37 | 0.777 | $0.801 \pm 0.016$ | 0.759 |
| M38 | 0.775 | $0.799 \pm 0.025$ | 0.755 |

| ID | 20k leaderboard | 20k AUROC | 20k AP |
|---|---|---|---|
| M39 | 0.774 | $0.804 \pm 0.019$ | 0.765 |
| M40 | 0.774 | $0.803 \pm 0.021$ | 0.760 |
| M41 | 0.771 | $0.791 \pm 0.022$ | 0.750 |
| M42 | 0.767 | $0.787 \pm 0.026$ | 0.745 |
| M43 | 0.764 | $0.784 \pm 0.023$ | 0.734 |
| M44 | 0.756 | $0.765 \pm 0.022$ | 0.725 |
| M45 | 0.755 | $0.769 \pm 0.026$ | 0.717 |
| M46 | 0.753 | $0.788 \pm 0.022$ | 0.752 |
| M47 | 0.752 | $0.763 \pm 0.026$ | 0.704 |
| M48 | 0.748 | $0.754 \pm 0.027$ | 0.697 |
| M49 | 0.748 | $0.746 \pm 0.025$ | 0.677 |
| M50 | 0.744 | $0.751 \pm 0.031$ | 0.689 |
| M51 | 0.741 | $0.756 \pm 0.015$ | 0.706 |

Table S9: Per-rank fold-mean AUROC at 20k. All models and composition controls use the same reads and folds.

| ID / control | Class | Order | Family | Genus | Species |
|---|---|---|---|---|---|
| M01 | 0.927 | 0.851 | 0.949 | 0.953 | 0.957 |
| M02 | 0.880 | 0.839 | 0.926 | 0.969 | 0.967 |
| M03 | 0.917 | 0.832 | 0.937 | 0.922 | 0.930 |
| M04 | 0.921 | 0.819 | 0.937 | 0.925 | 0.933 |
| M05 | 0.918 | 0.818 | 0.943 | 0.920 | 0.932 |
| M06 | 0.731 | 0.781 | 0.885 | 0.917 | 0.935 |
| M07 | 0.826 | 0.734 | 0.905 | 0.878 | 0.894 |
| M08 | 0.799 | 0.746 | 0.876 | 0.888 | 0.913 |
| M09 | 0.699 | 0.804 | 0.868 | 0.913 | 0.934 |
| M10 | 0.677 | 0.779 | 0.875 | 0.883 | 0.907 |
| M11 | 0.775 | 0.719 | 0.890 | 0.846 | 0.878 |
| M12 | 0.683 | 0.764 | 0.853 | 0.881 | 0.912 |
| M13 | 0.671 | 0.761 | 0.853 | 0.871 | 0.912 |
| M14 | 0.667 | 0.757 | 0.861 | 0.869 | 0.914 |
| M15 | 0.668 | 0.779 | 0.859 | 0.843 | 0.903 |
| M16 | 0.649 | 0.738 | 0.863 | 0.877 | 0.913 |
| M17 | 0.655 | 0.747 | 0.847 | 0.863 | 0.911 |
| M18 | 0.677 | 0.733 | 0.841 | 0.854 | 0.914 |
| M19 | 0.626 | 0.760 | 0.850 | 0.861 | 0.912 |
| M20 | 0.671 | 0.710 | 0.845 | 0.857 | 0.910 |
| M21 | 0.633 | 0.729 | 0.848 | 0.857 | 0.905 |
| M22 | 0.641 | 0.714 | 0.849 | 0.846 | 0.917 |
| M23 | 0.649 | 0.724 | 0.850 | 0.838 | 0.903 |
| M24 | 0.651 | 0.746 | 0.822 | 0.839 | 0.892 |
| M25 | 0.634 | 0.720 | 0.846 | 0.842 | 0.900 |
| M26 | 0.629 | 0.718 | 0.837 | 0.848 | 0.905 |
| M27 | 0.634 | 0.738 | 0.818 | 0.851 | 0.896 |
| M28 | 0.622 | 0.734 | 0.835 | 0.840 | 0.901 |
| M29 | 0.630 | 0.738 | 0.833 | 0.842 | 0.886 |
| M30 | 0.644 | 0.733 | 0.820 | 0.829 | 0.902 |
| M31 | 0.625 | 0.699 | 0.849 | 0.845 | 0.902 |
| M32 | 0.642 | 0.713 | 0.832 | 0.836 | 0.895 |
| M33 | 0.618 | 0.703 | 0.848 | 0.841 | 0.895 |
| M34 | 0.650 | 0.707 | 0.837 | 0.826 | 0.883 |
| M35 | 0.628 | 0.731 | 0.834 | 0.821 | 0.887 |
| M36 | 0.650 | 0.715 | 0.802 | 0.832 | 0.892 |
| M37 | 0.638 | 0.709 | 0.829 | 0.820 | 0.887 |
| M38 | 0.625 | 0.712 | 0.836 | 0.810 | 0.895 |
| M39 | 0.628 | 0.685 | 0.833 | 0.835 | 0.892 |
| M40 | 0.645 | 0.664 | 0.836 | 0.833 | 0.894 |
| M41 | 0.630 | 0.693 | 0.821 | 0.816 | 0.893 |
| M42 | 0.622 | 0.713 | 0.809 | 0.804 | 0.887 |

| ID / control | Class | Order | Family | Genus | Species |
|---|---|---|---|---|---|
| M43 | 0.638 | 0.692 | 0.818 | 0.789 | 0.881 |
| M44 | 0.615 | 0.738 | 0.795 | 0.773 | 0.859 |
| M45 | 0.660 | 0.683 | 0.810 | 0.759 | 0.862 |
| M46 | 0.589 | 0.700 | 0.786 | 0.809 | 0.878 |
| M47 | 0.666 | 0.695 | 0.800 | 0.749 | 0.851 |
| M48 | 0.671 | 0.704 | 0.778 | 0.737 | 0.849 |
| M49 | 0.689 | 0.699 | 0.787 | 0.730 | 0.835 |
| M50 | 0.646 | 0.702 | 0.796 | 0.735 | 0.841 |
| M51 | 0.619 | 0.688 | 0.787 | 0.763 | 0.849 |
| GC | 0.708 | 0.708 | 0.765 | 0.668 | 0.806 |
| 3-mer | 0.652 | 0.750 | 0.828 | 0.802 | 0.889 |
| 4-mer | 0.656 | 0.759 | 0.831 | 0.818 | 0.891 |
| 5-mer | 0.665 | 0.780 | 0.833 | 0.831 | 0.897 |
| 6-mer | 0.673 | 0.785 | 0.823 | 0.842 | 0.886 |

Table S10: Per-rank composition-control comparison at 20k. For each rank, the strongest composition control is selected by its raw clade-balanced fold-mean AUROC; Above counts reported checkpoints with a strictly greater raw AUROC. This appendix comparison does not change the fixed 5-mer reference control used in the main results.

| Rank | Strongest composition control | AUROC | Above |
|---|---|---|---|
| Class | GC content | 0.708 | 9 |
| Order | 6-mer freq | 0.785 | 6 |
| Family | 5-mer freq | 0.833 | 32 |
| Genus | 6-mer freq | 0.842 | 26 |
| Species | 5-mer freq | 0.897 | 26 |

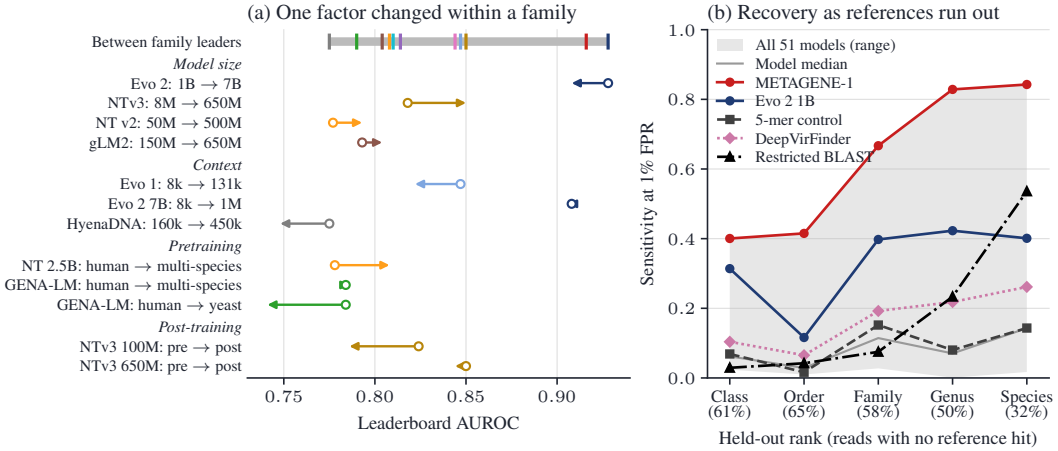



Figure S4: **What drives transfer, and where references run out.** (a) Leaderboard AUROC before (open circle) and after (arrowhead) changing one factor within a model family; the grey bar spans the best checkpoint of each family, one tick per family. (b) Sensitivity at 1% FPR by held-out rank, pooling reads within each rank rather than balancing clades; the grey band spans all 51 models. Percentages under each rank give the share of its reads with no hit in either BLAST reference.

## E NOVELTY CURVES AND PAIRED COMPOSITION RESIDUALS

Main-text Figure 6 shows the clade-balanced curves for the evaluation set. Paired residuals compare each model with the composition control using the same clade-balanced fold-mean estimator (Appendix C.2). These intervals condition on the fitted scores and released split and have no multiple-comparison adjustment. Comparing ranks compares different clades, not a controlled manipulation of evolutionary distance. Table S9 reports the per-rank estimates using consistent model identifiers, and Table S10 compares each rank with its strongest composition control.

**Plant-order subgroup.** METAGENE-1 separates held-out viral reads less accurately from flowering-plant reads than from other cellular reads at the order rank (pooled AUROC 0.824 versus 0.949).

## F DETECTION OPERATING POINTS

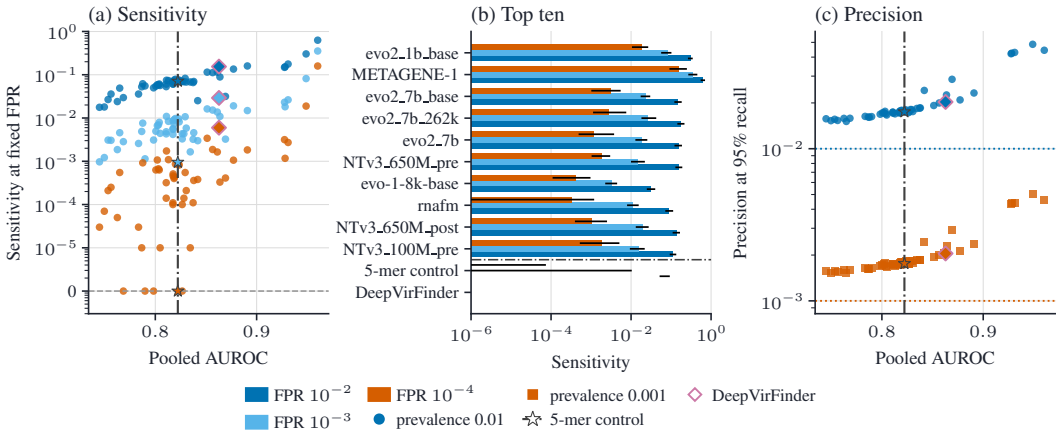



Figure S5: **Detection at low false-positive rates.** All panels use pooled held-out scores. (a) Sensitivity at three FPR targets against AUROC for all 51 models; points at "0" recover no viral reads. (b) The ten leading models by leaderboard AUROC, the 5-mer control and DeepVirFinder at the same targets. Error bars, where available, are 95% clade-bootstrap intervals. (c) Precision at 95% recall, calculated for assumed viral prevalences of 1% and 0.1%; dotted lines show a random classifier. Stars and the vertical dash-dotted line mark the 5-mer control; pink diamonds mark DeepVirFinder.

Table S11: **Pooled held-out sensitivity (%) at fixed FPR.** The model is the leaderboard leader; the control is the reference composition control. Thresholds describe test ROC curves; none was validation-selected for deployment. Restricted BLAST is the alignment baseline: no probe, each read searched with blastn against references with its cell's held-out clades removed.

| Entry | AUROC | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ |
|---|---|---|---|---|
| evo2_1b_base | 0.949 | 30.30 | 8.15 | 1.89 |
| 5-mer control | 0.822 | 7.08 | 0.10 | 0.00 |
| Restricted BLAST | 0.877 | 18.54 | 9.52 | 0.26 |

Figure S5 shows the operating points of all models. Sensitivity at target FPR $\alpha$ is the largest empirical TPR attainable at FPR no greater than $\alpha$, using thresholds that respect score ties. There are 100,000 negative reads, giving an FPR resolution of $10^{-5}$. Thus FPR $10^{-4}$ permits only ten false positives; tail estimates need particular care. No threshold is selected on an independent validation partition. Precision scenarios additionally assume that sensitivity and FPR transfer unchanged to the stated prevalence.

The standardized partial AUROC over FPR $[0, \alpha]$ is

$$\frac{1}{2}\left[1 + \frac{\int_0^\alpha \mathrm{TPR}(f)\,df - \alpha^2/2}{\alpha - \alpha^2/2}\right], \tag{3}$$

with linear interpolation at the cutoff. It equals 0.5 for the chance ROC and 1 for perfect separation. The detection report contains these metrics, precision at 95% and 99% recall, and operating points for all models and controls. A sensitivity ratio is undefined where its control denominator is zero. Tables S11, S5 and S13 report the operating points, conditional intervals and precision scenarios; Table S12 lists pooled AUROC and sensitivity for every model.

Table S12: Detection at 20k. Pooled AUROC, sensitivity (%) at each FPR, and hypothetical precision (%) at 95% recall and prevalence 0.1%. These are test-ROC summaries, not prospectively chosen thresholds.

| ID / control | AUROC | TPR@$10^{-2}$ | TPR@$10^{-3}$ | TPR@$10^{-4}$ | Precision |
|---|---|---|---|---|---|
| M01 | 0.949 | 30.30 | 8.15 | 1.89 | 0.503 |
| M02 | 0.960 | 62.87 | 35.38 | 15.90 | 0.458 |
| M03 | 0.928 | 14.82 | 2.31 | 0.32 | 0.434 |
| M04 | 0.931 | 17.68 | 2.63 | 0.27 | 0.443 |
| M05 | 0.928 | 15.24 | 1.82 | 0.12 | 0.437 |
| M06 | 0.891 | 15.96 | 1.49 | 0.18 | 0.235 |
| M07 | 0.869 | 3.16 | 0.33 | 0.04 | 0.291 |
| M08 | 0.864 | 8.89 | 1.09 | 0.03 | 0.228 |
| M09 | 0.877 | 13.52 | 1.91 | 0.11 | 0.211 |
| M10 | 0.857 | 11.21 | 1.52 | 0.19 | 0.202 |
| M11 | 0.842 | 2.52 | 0.30 | 0.04 | 0.245 |
| M12 | 0.861 | 11.46 | 1.37 | 0.16 | 0.205 |
| M13 | 0.837 | 6.60 | 0.32 | 0.01 | 0.186 |
| M14 | 0.838 | 6.99 | 0.46 | 0.04 | 0.187 |
| M15 | 0.817 | 7.05 | 0.56 | 0.01 | 0.173 |
| M16 | 0.851 | 11.48 | 1.62 | 0.34 | 0.194 |
| M17 | 0.835 | 6.81 | 0.17 | 0.00 | 0.185 |
| M18 | 0.831 | 6.78 | 0.58 | 0.02 | 0.177 |
| M19 | 0.828 | 6.90 | 0.43 | 0.01 | 0.182 |
| M20 | 0.827 | 6.98 | 0.43 | 0.01 | 0.174 |
| M21 | 0.829 | 8.27 | 1.07 | 0.06 | 0.182 |
| M22 | 0.827 | 7.07 | 0.33 | 0.00 | 0.176 |
| M23 | 0.819 | 7.56 | 0.93 | 0.04 | 0.180 |
| M24 | 0.811 | 7.46 | 1.09 | 0.17 | 0.172 |
| M25 | 0.811 | 5.31 | 0.48 | 0.01 | 0.175 |
| M26 | 0.823 | 7.47 | 0.97 | 0.09 | 0.179 |
| M27 | 0.824 | 7.49 | 0.81 | 0.08 | 0.181 |
| M28 | 0.818 | 6.82 | 0.73 | 0.04 | 0.175 |
| M29 | 0.818 | 7.01 | 0.86 | 0.05 | 0.174 |
| M30 | 0.805 | 5.65 | 0.44 | 0.00 | 0.168 |
| M31 | 0.818 | 6.05 | 0.55 | 0.05 | 0.178 |
| M32 | 0.817 | 6.13 | 0.42 | 0.01 | 0.176 |
| M33 | 0.820 | 7.85 | 0.97 | 0.09 | 0.180 |
| M34 | 0.811 | 6.07 | 0.43 | 0.03 | 0.175 |
| M35 | 0.802 | 5.88 | 0.67 | 0.11 | 0.174 |
| M36 | 0.813 | 7.34 | 0.84 | 0.11 | 0.171 |
| M37 | 0.801 | 5.44 | 0.50 | 0.06 | 0.173 |
| M38 | 0.798 | 4.98 | 0.31 | 0.00 | 0.170 |
| M39 | 0.804 | 6.11 | 0.79 | 0.06 | 0.171 |
| M40 | 0.803 | 5.83 | 0.69 | 0.07 | 0.177 |
| M41 | 0.790 | 5.51 | 0.11 | 0.00 | 0.164 |
| M42 | 0.787 | 5.27 | 0.49 | 0.00 | 0.162 |
| M43 | 0.783 | 3.58 | 0.26 | 0.00 | 0.164 |
| M44 | 0.765 | 4.90 | 0.46 | 0.06 | 0.152 |
| M45 | 0.768 | 2.96 | 0.28 | 0.00 | 0.159 |
| M46 | 0.788 | 6.20 | 0.76 | 0.09 | 0.162 |
| M47 | 0.763 | 2.40 | 0.19 | 0.01 | 0.159 |
| M48 | 0.753 | 2.67 | 0.31 | 0.02 | 0.157 |
| M49 | 0.745 | 1.77 | 0.10 | 0.00 | 0.158 |
| M50 | 0.750 | 1.80 | 0.12 | 0.01 | 0.154 |
| M51 | 0.756 | 3.58 | 0.30 | 0.03 | 0.154 |
| GC | 0.716 | 0.66 | 0.01 | 0.00 | 0.158 |
| 3-mer | 0.798 | 4.52 | 0.16 | 0.00 | 0.174 |
| 4-mer | 0.809 | 5.14 | 0.01 | 0.00 | 0.177 |
| 5-mer | 0.822 | 7.08 | 0.10 | 0.00 | 0.176 |
| 6-mer | 0.829 | 7.47 | 0.03 | 0.00 | 0.178 |

Table S13: Partial AUROC and precision scenarios for METAGENE-1 (MG1), all controls and restricted BLAST. pAUC columns use FPR cutoffs $10^{-2}, 10^{-3}, 10^{-4}$. Precision columns (%) give recall / assumed prevalence (%).

| Entry | pAUC .01 | pAUC .001 | pAUC .0001 | P:95/1 | P:95/.1 | P:99/.1 |
|---|---|---|---|---|---|---|
| MG1 | 0.757 | 0.630 | 0.551 | 4.438 | 0.458 | 0.252 |
| GC | 0.499 | 0.500 | 0.500 | 1.573 | 0.158 | 0.137 |
| 3-mer | 0.508 | 0.500 | 0.500 | 1.725 | 0.174 | 0.142 |
| 4-mer | 0.509 | 0.500 | 0.500 | 1.759 | 0.177 | 0.139 |
| 5-mer | 0.514 | 0.500 | 0.500 | 1.747 | 0.176 | 0.136 |
| 6-mer | 0.515 | 0.500 | 0.500 | 1.764 | 0.178 | 0.135 |
| Restricted BLAST | 0.575 | 0.522 | 0.501 | 3.480 | 0.356 | 0.255 |

## G RESTRICTED BLAST BASELINE

The Kraken2 PlusPFP database retains only its minimizer index, so the reference was rebuilt from the accessions in its `seqid2taxid.map`. Of 75,945 assemblies, 75,927 were retrievable from NCBI; 18 have since been suppressed. Sequences were split into a viral database (18,677 sequences) and a cellular database (1,003,726 sequences, 455.8 Gb); 18,286 plasmid records added after the database build were excluded. For each of the 25 (rank, fold) strata, the sequences whose taxa fall in that stratum's held-out clades were listed by the repository's exclusion code, which passed its audit and negative-control checks, and passed to `blastn` as a negative sequence-identifier list, so each read is searched only against a reference without its own held-out clades. The probe is held to the same exclusion, since its labelled training reads come only from the other folds; the unrestricted reference, from which the labels derive, is circular and is not used as a baseline. The comparison does not control the encoder's pretraining exposure or independently validate the Kraken2 labels.

Search used BLAST+ 2.14.1 with `-task blastn -word_size 11 -evalue 1e-3 -max_target_seqs 5`. The default megablast task was about five times faster but found 10 instead of 745 viral-database hits on one test stratum, and was not used. A read's score is its best viral bit score minus its best cellular bit score, with a missing side counted as 0. All 200,000 evaluation reads were searched. Section 5 reports the comparison and intervals from 2,000 clade-bootstrap draws over 2,927 represented clades. The no-hit share of evaluation reads is 52.9%. No-hit reads receive score zero, producing a tied block; the reported operating points respect whole ties rather than interpolating fractional reads. No-hit proportions count all reads in a rank, not only viral reads. The diagnostic figure's within-rank sensitivity pools the five folds and is distinct from the primary clade-balanced fold means. It cannot establish a conditional advantage restricted to no-hit reads.

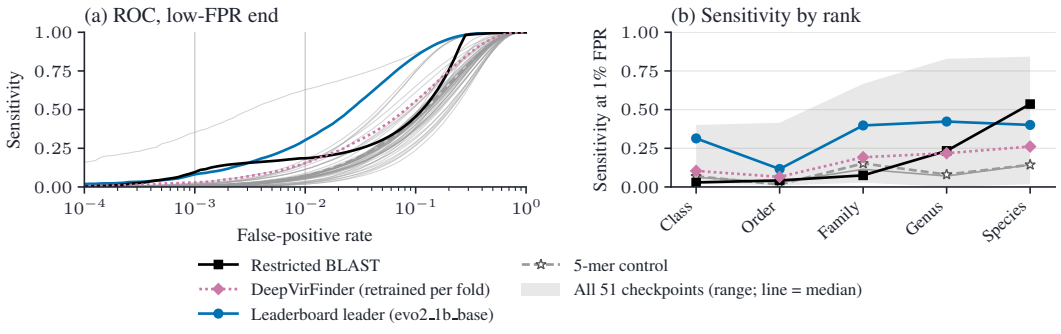



Figure S6: **Restricted BLAST compared with learned detectors.** Black denotes restricted BLAST, blue evo2_1b_base (the leaderboard leader), pink DeepVirFinder, and dashed grey the 5-mer control. (a) ROC pooled over all held-out reads, with the low-FPR region expanded on a log axis. Grey curves show all 51 gLMs. Lines connect empirical vertices, including ties; intermediate points may not be attainable. (b) Conservative sensitivity at 1% FPR, pooling read-weighted scores within each rank. The grey band spans the gLM minimum and maximum, with their median shown as a thin line; it is not a confidence interval. No confidence intervals are shown in either panel.

## H DEEPVIRFINDER WITH MORE TRAINING DATA

DeepVirFinder-large uses the same folds and held-out reads as the matched DeepVirFinder arm. Its leaderboard AUROC is 0.824 (Table S15), compared with 0.827 for the matched arm and 0.928 for evo2_1b_base, the leading model. The larger arm uses the TensorFlow GPU backend, whereas the matched table arm uses the Theano CPU backend.

Table S14: Reference baselines at 20k. Leaderboard AUROC is the mean over the five holdout ranks of the clade-balanced fold-mean AUROC, as for the checkpoints; fold-mean AUROC counts every held-out read once. DeepVirFinder is retrained on each fold's training reads; restricted BLAST searches a reference with each cell's held-out clades removed. Above: reported checkpoints whose leaderboard AUROC exceeds the baseline's. DeepVirFinder over 3 seeds: $0.830 \pm 0.004$ (mean $\pm$ s.d.).

| Baseline | Leaderboard AUROC | Fold-mean AUROC | Class | Order | Family | Genus | Species | Above |
|---|---|---|---|---|---|---|---|---|
| 5-mer control | 0.801 | 0.822 | 0.665 | 0.780 | 0.833 | 0.831 | 0.897 | 19 of 51 |
| DeepVirFinder (seed 1) | 0.827 | 0.865 | 0.716 | 0.754 | 0.857 | 0.890 | 0.919 | 9 of 51 |
| Restricted BLAST | 0.882 | 0.877 | 0.845 | 0.836 | 0.860 | 0.910 | 0.960 | 5 of 51 |

Table S15: DeepVirFinder trained on up to about ten times more reads of the training folds' taxa (large), on the same folds and holdout, against the matched arm and the 3 leading checkpoints at 20k. Leaderboard AUROC as in Table S14; per-rank values are clade-balanced fold means.

| Model | Leaderboard AUROC | Class | Order | Family | Genus | Species |
|---|---|---|---|---|---|---|
| DeepVirFinder-large | 0.824 | 0.708 | 0.734 | 0.864 | 0.901 | 0.914 |
| DeepVirFinder (matched) | 0.827 | 0.716 | 0.754 | 0.857 | 0.890 | 0.919 |
| evo2_1b_base | 0.928 | 0.927 | 0.851 | 0.949 | 0.953 | 0.957 |
| METAGENE-1 | 0.916 | 0.880 | 0.839 | 0.926 | 0.969 | 0.967 |
| evo2_7b_base | 0.908 | 0.917 | 0.832 | 0.937 | 0.922 | 0.930 |

## I ESTIMATOR ROBUSTNESS

We compare the primary estimator with the five-fold mean AUROC counting each read once, and with read- and taxon-weighted pooled AUROCs, all computed from the same fitted scores. The first comparison changes within-rank weighting and aggregation across ranks; the pooled comparisons additionally combine scores across folds.

The evaluation set represents every confined clade, but the observed read counts per clade are unequal (Section 3.4). Read weighting follows the representation of units and clades in the selected evaluation reads; taxon weighting gives every represented unit equal mass within its label–fold–rank group, which gives more total mass to clades containing more units. Neither reconstructs corpus-abundance weighting, and neither is the primary estimand, which weights each clade equally and then each unit within it. Figure S7 compares the leaderboard statistic with the five-fold mean that counts every held-out read once; Tables S16–S17 report the weighting analyses on the same fitted scores. Appendix C.1 defines the estimators and distinguishes pooled AUROC from fold means. Across 51 models, the two orderings have Spearman $\rho = 0.974$; 15 models change rank by at least 3 positions.

Tables S16 and S17 give the rank-level and model-level results for the evaluation set. Figure S8 visualizes the read- and taxon-weighted pooled-score sensitivity.

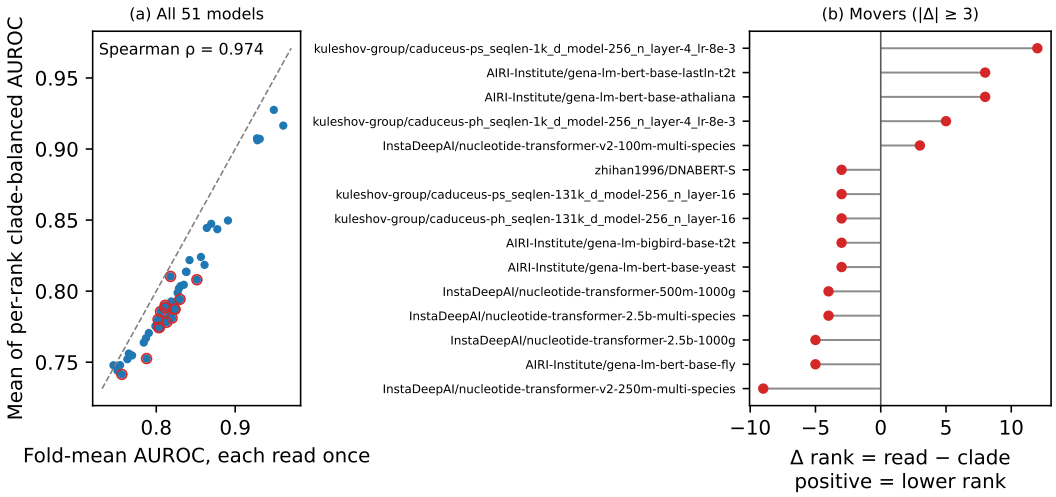



Figure S7: **Estimator robustness.** (a) Each point is one model: five-fold mean AUROC counting every held-out read once on the horizontal axis, and the primary rank-averaged clade-balanced AUROC on the vertical axis. Red outlines mark models whose leaderboard positions differ by at least 3. (b) The signed position changes for those models, calculated as read-counted position minus clade-balanced position; positive values indicate a worse position under read counting. Both statistics use the same fitted scores.

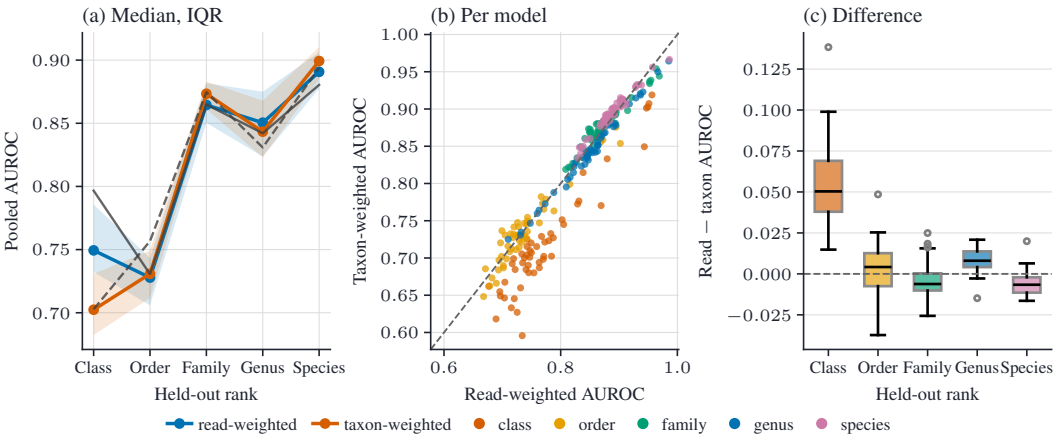



Figure S8: **How evaluation weighting changes AUROC.** (a) Median read- and taxon-weighted pooled AUROC across 51 models, with interquartile bands. Grey lines show the 5-mer control (solid: read-weighted; dashed: taxon-weighted). (b) One point per model and rank; points above the diagonal have higher taxon-weighted AUROC. (c) Read-weighted minus taxon-weighted AUROC across models. Boxes span the interquartile range, centre lines mark medians, whiskers extend to the most extreme values within 1.5 interquartile ranges of the box, and circles mark values beyond them. Positive differences indicate higher AUROC under read weighting.

Table S16: Pooled-score AUROC by novelty rank: median across models and the reference composition control (5-mer), under read and taxon weighting.

| Rank | Models/read | Models/taxon | Control/read | Control/taxon |
|---|---|---|---|---|
| Class | 0.749 | 0.702 | 0.797 | 0.703 |
| Order | 0.728 | 0.731 | 0.731 | 0.757 |
| Family | 0.864 | 0.873 | 0.865 | 0.875 |
| Genus | 0.850 | 0.843 | 0.842 | 0.831 |
| Species | 0.891 | 0.899 | 0.880 | 0.894 |

Table S17: Every model under read and taxon weighting. All values are pooled-score AUROC, not fold means.

| ID | 20k/read | 20k/taxon |
|---|---|---|
| M01 | 0.949 | 0.929 |
| M02 | 0.960 | 0.919 |
| M03 | 0.928 | 0.908 |
| M04 | 0.931 | 0.911 |
| M05 | 0.928 | 0.908 |
| M06 | 0.891 | 0.875 |
| M07 | 0.869 | 0.855 |
| M08 | 0.864 | 0.840 |
| M09 | 0.877 | 0.854 |
| M10 | 0.857 | 0.840 |
| M11 | 0.842 | 0.824 |
| M12 | 0.861 | 0.847 |
| M13 | 0.837 | 0.829 |
| M14 | 0.838 | 0.830 |
| M15 | 0.817 | 0.815 |
| M16 | 0.851 | 0.836 |
| M17 | 0.835 | 0.826 |
| M18 | 0.831 | 0.821 |
| M19 | 0.828 | 0.827 |
| M20 | 0.827 | 0.815 |
| M21 | 0.829 | 0.814 |
| M22 | 0.827 | 0.816 |
| M23 | 0.819 | 0.812 |
| M24 | 0.811 | 0.799 |
| M25 | 0.811 | 0.808 |
| M26 | 0.823 | 0.811 |
| M27 | 0.824 | 0.813 |
| M28 | 0.818 | 0.810 |
| M29 | 0.818 | 0.801 |
| M30 | 0.805 | 0.799 |
| M31 | 0.818 | 0.810 |
| M32 | 0.817 | 0.808 |
| M33 | 0.820 | 0.811 |
| M34 | 0.811 | 0.802 |
| M35 | 0.802 | 0.798 |
| M36 | 0.813 | 0.797 |
| M37 | 0.801 | 0.798 |
| M38 | 0.798 | 0.794 |
| M39 | 0.804 | 0.797 |
| M40 | 0.803 | 0.804 |
| M41 | 0.790 | 0.785 |
| M42 | 0.787 | 0.778 |
| M43 | 0.783 | 0.777 |
| M44 | 0.765 | 0.761 |
| M45 | 0.768 | 0.757 |
| M46 | 0.788 | 0.776 |
| M47 | 0.763 | 0.741 |
| M48 | 0.753 | 0.739 |
| M49 | 0.745 | 0.719 |
| M50 | 0.750 | 0.732 |
| M51 | 0.756 | 0.754 |

## J ARTIFACT SCOPE AND LICENCES

The anonymous code repository (`https://anonymous.4open.science/r/wavebench-ICLR-81EF`) contains the fold-construction, audit, read-selection, embedding and evaluation code, with the method configuration that lists every search seed and step budget. The anonymous dataset repository (`https://huggingface.co/datasets/WavebenchICLR/wave`

bench-ICLR) contains the released split: the viral and cellular unit-to-fold assignments with their checksums. It also contains the 200,000 evaluation reads with their sequences, Kraken2 taxa, assignment units, folds, labels, held-out ranks and clades (selection seed 20260925), and the run accession of every sample; the frozen embeddings of every model on those reads, with model revisions and per-run GPU and runtime provenance; the probe receipts and training weights; and the held-out decision scores of all 51 models, the five composition controls and the reference baselines. A script in the code repository recomputes the evaluation statistics from those scores without re-running an encoder, refits the training weights and the composition controls from the reads and any model's probe from its embeddings, and compares every recomputed value with the released one. Regenerating the embeddings themselves requires GPUs and the model revisions and adapters recorded in the code, where the licence of each model is listed alongside.

Not included: the Kraken2 PlusPFP index, which is distributed by its maintainers; the restricted-BLAST reference sequences and databases, which are rebuilt from the accessions in that index's sequence-to-taxon map; the raw reads, which are public in the Sequence Read Archive under BioProject PRJNA1247874; the deduplicated corpus and the per-read Kraken2 classifications; and the deduplication implementation, whose parameters are recorded in Appendix B. The Kraken2 executable version is not fully pinned, and the database build is recorded by its identifier only. No fine-tuned counterpart or independent mock-community or spike-in validation is reported, and Kraken2 defines both labels and holdouts, so agreement with it is not independent biological accuracy.

Code and documentation are released under the Apache License 2.0, and the split, labels, scores and receipts under CC BY 4.0. The reads remain under the Sequence Read Archive's terms and should be cited with Justen et al. (2026). Each model's embeddings follow that model's licence: the Nucleotide Transformer v1 and v2 models are under CC BY-NC-SA 4.0, the NTv3 models under InstaDeep's non-commercial model licence, whose outputs may not be re-licensed, and the RNA-FM port under AGPL-3.0; the repository lists the licence of every model.