POSITIONAL ATTENTION SIGNATURES DISTINGUISH AI-GENERATED FROM NATURAL BIOLOGICAL SEQUENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Biological foundation models generate increasingly realistic DNA and proteins, raising two questions: how do generated sequence distributions differ from natural ones, and how can those sequences be screened efficiently? We study positional attention variance (pAV), the variation of attention-weighted distance across positions in a frozen biological language model. Across six DNA generators, mean pAV is 53–84% of a matched natural reference under a multispecies Nucleotide Transformer. Preserving dinucleotide counts while shuffling order lowers median pAV by 16.2% across 200 natural windows, establishing arrangement sensitivity. A protein panel shows analogous differences under ESM2. In matched detection tests, inexpensive features are stronger on clean data, while standalone pAV retains macro AUROC 0.867 at 20% substitution, compared with 0.822 for compression. Adding pAV to a clean-trained compression classifier does not improve this stress-test performance. We separately evaluate DNA screening with multiscale sequence features and frozen observer representations, obtaining macro AUROC 0.995 in five-fold sequence-level cross-validation. Distilling the observer contribution enables a compact service with measured throughput of 334.6 sequences per second on one GPU. Together, these studies establish an arrangement-sensitive statistical probe and a computationally practical screening approach, with distinct measurement, evaluation, and calibration boundaries.

1 INTRODUCTION

Foundation models now support biological sequence generation at scales ranging from proteins to long genomic contexts. GenomeOcean, Evo, Evo 2, and GENERator learn from large DNA corpora and produce sequences with increasingly realistic local statistics (Wu et al., 2025b; Nguyen et al., 2024; Brixi et al., 2026; Wu et al., 2025a). Protein language models and diffusion models similarly generate diverse amino-acid sequences (Ferruz et al., 2022; Nijkamp et al., 2023; Alamdari et al., 2026). These advances make sequence provenance a concrete data-management problem. A repository curator may need to distinguish naturally sampled records from unlabeled model outputs, while a model developer needs measurements that expose differences not apparent from a small set of marginal statistics.

Sequence similarity and provenance are different questions. Homology search identifies relationships to known sequences, but a generated sequence may have no strong database match, as may an uncharacterized natural sequence. Composition-based screening supplies another useful view. Prior work has identified compositional differences in generated bacteriophages, and Synsor detects signatures of engineered DNA (Ratcliff, 2024; Tay et al., 2024). These approaches motivate strong sequence-based controls rather than an assumption that neural observers are necessary for every detection setting.

We investigate a complementary measurement: how a frozen biological language model distributes its attention across sequence positions. A transformer attention head assigns each query position a distribution over keys (Vaswani et al., 2017). Its mean attention distance summarizes the span of that distribution; variation of this distance across queries describes positional heterogeneity inside the observer. We call this quantity positional attention variance (pAV). A difference in pAV between

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.

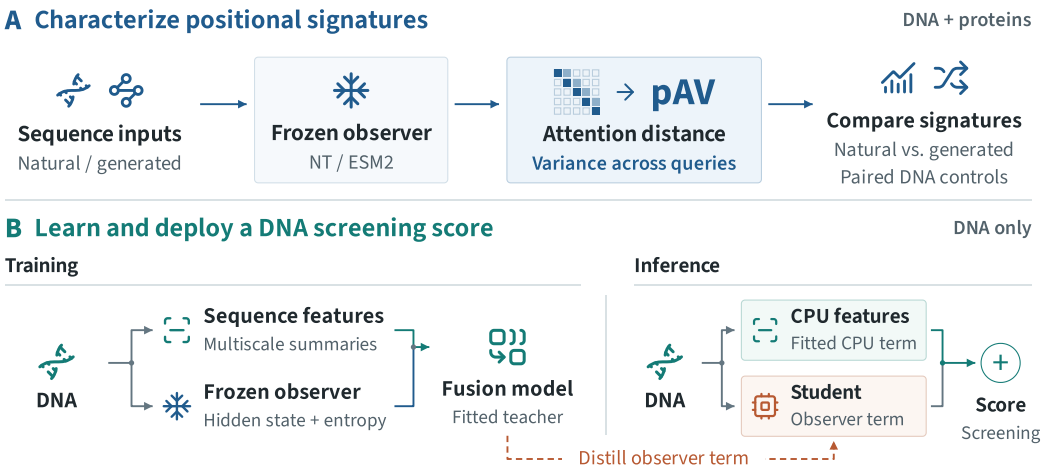

Figure 1: Two uses of frozen biological observers. (A) Attention-distance variance characterizes natural and generated DNA and proteins; paired DNA controls probe arrangement sensitivity. (B) Training combines multiscale sequence features and frozen observer representations in a fusion teacher. The dashed arrow denotes distillation of its observer contribution. During inference, DNA feeds the CPU and student branches; their logit contributions combine into a screening score without running the teacher. Icons are schematic, not measured attention maps or performance curves.

natural and generated populations is an empirical property of the sequence–observer pair, not a direct assay of regulatory interactions or biological function.

We organize the study around two uses of frozen observers (Figure 1). The first is measurement: we ask whether pAV differs across sequence populations, which transformations change it, and how those findings depend on the observer. The second is screening: we evaluate whether observer representations and inexpensive sequence features can support accurate, computationally practical DNA classification. The shared observer connects these questions, but they have different success criteria. A useful statistical probe need not be the strongest classifier, and a successful representation-based detector does not establish the marginal value of pAV.

The experiments establish three findings. Generated DNA populations show lower mean pAV under the multispecies observer, and dinucleotide-preserving controls lower median pAV by 16.2%, identifying sensitivity to sequence arrangement. Matched baseline and perturbation tests then show where this scalar signal is useful and where inexpensive features are stronger. Finally, a separate representation-fusion study evaluates DNA screening and distillation of the costly observer contribution. Together, these findings connect sequence characterization to practical screening without treating detection accuracy as evidence of biological function or individual authenticity.

2 Related Work

Biological sequence models. Genomic models differ in training distributions, tokenization, architecture, and context length. Nucleotide Transformer (NT) provides frozen bidirectional observers trained on human or multispecies genomes (Dalla-Torre et al., 2025). DNABERT-S supplies species-aware DNA embeddings (Zhou et al., 2025). Evo and HyenaDNA use long-context sequence architectures, while GenomeOcean and GENERator provide autoregressive genomic generators (Nguyen et al., 2024; 2023; Wu et al., 2025b;a). In proteins, ESM2 learns sequence representations relevant to structure prediction, and ProtGPT2, ProGen2, and EvoDiff provide distinct generation paradigms (Lin et al., 2023; Ferruz et al., 2022; Nijkamp et al., 2023; Alamdari et al., 2026). We use these models as experimental components, without equating their internal attention statistics with independent functional measurements.

Generated-sequence detection. DetectGPT and Fast-DetectGPT study statistical signatures of generated text through probability curvature and conditional probability curvature (Mitchell et al., 2023; Bao et al., 2024). They motivate the broader question of observer-based detection, but DNA tokenization and perturbations define a different measurement problem. Engineered-DNA detectors such as Synsor address a related, distinct provenance task (Tay et al., 2024). We therefore compare attention statistics with compression, short-word entropy, GC content, k-mer frequencies, and Markov likelihood ratios. The principal comparison uses the same training sequences and natural reference for every method.

3 METHODS

3.1 POSITIONAL ATTENTION VARIANCE

For an input sequence, let $A^{(\ell,h)}$ be the post-softmax attention matrix of layer ℓ and head h under a frozen observer. We restrict the matrix to its L active token positions using the attention mask. For query position j, define

$$\mathrm{MAD}_j^{(\ell,h)} = \sum_{k=0}^{L-1} A_{jk}^{(\ell,h)} |j-k|, \qquad \mathrm{pAV}^{(\ell,h)} = \frac{1}{|I|} \sum_{j \in I} \left(\mathrm{MAD}_j^{(\ell,h)} - \overline{\mathrm{MAD}}^{(\ell,h)} \right)^2, \tag{1}$$

where $I = \{1, \ldots, L-2\}$ and the bar denotes the mean over I. The DNA implementation removes padding before computing distances and excludes the first and last active query positions from the variance. The key set retains all active tokens, including special tokens when present, without renormalizing after query-boundary exclusion.

Distances are in token indices, not base pairs. This matters when comparing 6-mer and byte-pair encoding (BPE) observers, or sequences with different token counts. We compare natural and generated populations within a fixed observer and window protocol. Relative values divide by the mean of the corresponding natural reference; they are not observer-independent physical units.

For NT-2.5B, the per-sequence score averages a fixed set H of 20 heads among 32 layers and 20 heads per layer:

$$s_{\mathrm{pAV}}(x) = |H|^{-1} \sum_{(\ell,h) \in H} \mathrm{pAV}^{(\ell,h)}(x). \tag{2}$$

The head set was selected by discrimination and is a supervised component. It is held fixed throughout the DNA replication and perturbation analyses; we do not select heads separately for each target generator. Appendix B lists the indices and the available head-selection comparison. Attention extraction uses float32, a frozen evaluation-mode observer, and the same token-handling convention across experimental arms.

3.2 CONTROLLED TRANSFORMATIONS

To test what the measurement responds to, we transform the same 200 natural windows. Reverse complementation tests strand reversal under the observer. Mononucleotide shuffling preserves base composition; dinucleotide shuffling preserves exact 2-mer counts; codon-block shuffling preserves the multiset of nonoverlapping triplets. The latter is a triplet-order control, not a claim that every input window is an annotated coding sequence. These paired interventions separate retained local counts from altered sequence order.

A second experiment tests discrimination under distribution shift. We apply random point substitutions to generated DNA at rates of 0%, 5%, 10%, and 20%, keeping the natural reference unchanged and classifiers fitted on clean training sequences. The substitutions are not optimized against the detector and need not preserve biological function. Refitting at each rate would instead measure adaptation to that condition; the fixed-model design isolates performance under the specified shift.

3.3 DNA SCREENING MODELS

Screening uses observer information at two levels of detail. At the scalar level, standalone pAV uses $-s_{\mathrm{pAV}}(x)$ as the generated-sequence score without fitting a classifier. GenomeOcean-Sentinel

(GOS) combines mean pAV with four sequence features: gzip compression ratio, LZMA compression ratio, k-mer periodicity, and stop-codon density. Standardized features feed a regularized logistic regression. Appendix E.1 defines these sequence features.

At the representation level, we retain hidden features and attention entropy rather than reducing the observer to pAV. The 3,248-dimensional fusion vector contains a 2,560-dimensional mean hidden representation from layer 24 of NT-2.5B, 640 attention-entropy summaries, 47 multiscale CPU summaries, and an observer-availability indicator. A StandardScaler and balanced logistic regression are fitted within each training fold. This detector addresses the screening objective with a different feature set; its evaluation is separate from the scalar pAV ablations. Appendix F specifies feature order and fitting.

The multiscale branch uses 500-bp subwindows with stride 250. Each subwindow yields nine measurements: two compression-payload ratios, stop-codon density, four word entropies (1-, 2-, 3-, and 6-mer), the fraction of distinct 6-mers, and a linear periodicity statistic. The mean, maximum, standard deviation, minimum, and top-three mean of each measurement provide 45 features. Availability and window-count indicators complete the 47-dimensional branch. This describes within-window variability without requiring additional observer passes.

3.4 Distilling the Observer Contribution

Representation fusion requires an observer pass for every input. To reduce that inference cost, we decompose its logit into a CPU contribution and an observer contribution, then train a compact sequence model to approximate the latter. The deployed score combines the student's observer logit with the CPU contribution. We use a GenomeOcean-100M v1.2 student, evaluating detection with out-of-fold predictions and service cost with a resident model and bounded queues.

Let $b(x)$ denote the CPU-plus-intercept logit and $r(x)$ the teacher's observer contribution. If the student predicts a normalized scalar $q_\theta(x)$, the training objective is

$$\mathcal{L}(x, y) = \left(q_\theta(x) - \frac{r(x) - \mu_r}{\sigma_r} \right)^2 + 0.25 \ \mathrm{BCEWithLogits}(b(x) + \sigma_r q_\theta(x) + \mu_r, y) \,. \quad (3)$$

The target mean μ_r, standard deviation σ_r, and teacher coefficients are estimated on the corresponding training fold. This keeps the student target tied to the fitted detector while preserving a binary learning signal. At inference, the probability is the sigmoid of the combined CPU and student logit.

4 Experimental Setup

DNA and natural references. The six DNA sources are GenomeOcean-100M, GenomeOcean-500M, GenomeOcean-4B, GENERator-v2, Evo-1-131k, and Evo2-7B. The natural pool combines reference genomes and environmental collections, including GTDB, RefSeq, human, viral, and metagenomic sources (Parks et al., 2018; O'Leary et al., 2016; Alteio et al., 2020; Sunagawa et al., 2015; The Integrative HMP (iHMP) Research Network Consortium, 2019). We remove internal whitespace, uppercase sequences, split at non-ACGT symbols, and extract central 1,000-bp windows from qualifying fragments. Exact windows are deduplicated before seeded sampling. This treatment is particularly important for released BPE-tokenized sequence files and runs of ambiguous bases.

The characterization panel contains 200 windows per generated source and 200 natural windows. The matched baseline experiment trains on 1,500 natural and 1,500 generated windows and evaluates 500 windows per generated source against 500 held-out natural windows. All baselines use the same sequences. The selected generated training sample consists entirely of GenomeOcean-4B high-complexity outputs. GenomeOcean-100M, GenomeOcean-500M, GENERator-v2, and Evo-1 form the fixed four-model macro panel. GenomeOcean-4B and Evo2-7B are reported separately: the former shares the training generator, whereas Evo2 is also absent from the selected baseline training sample. Sampling uses seed 42 for training and seed 43 for evaluation. Exact training–evaluation overlap is checked; homologous or pretraining-corpus overlap is not thereby excluded.

Proteins and additional observers. The protein panel uses reviewed Swiss-Prot proteins and sequences from ProtGPT2, ProGen2-small, and EvoDiff-seq-640M, with a reported 1,000 analyzed

sequences per group and an 80–512-amino-acid natural reference (The UniProt Consortium, 2023; Ferruz et al., 2022; Nijkamp et al., 2023; Alamdari et al., 2026). ESM2-650M is the frozen observer. Protein-domain aggregate results are presented separately from the DNA replication; their numerical scale is not directly comparable. We also examine NT-500M and DNABERT-S to characterize observer-specific behavior.

Evaluation protocols. AUROC treats generated sequences as positive. Macro averages weight each target generator equally. In the matched pAV/baseline analysis, TPR at 5% FPR is an empirical ROC operating point, determined from the scored natural reference. It is not an independently calibrated deployment guarantee. The fusion experiment uses a distinct 5,000-row panel: 3,000 generated windows from six sources and 2,000 natural windows from four collections, all of length 1,000 bp. Five-fold cross-validation is stratified by source, so each source appears in both training and evaluation folds. We call this sequence-level cross-validation, not generator-held-out evaluation. Thresholds for its reported 1.0%, 1.3%, and 5.0% FPR points are chosen from pooled out-of-fold natural scores.

Paired controls use two-sided Wilcoxon signed-rank tests. Effect sizes distinguish the ratio of group means from the mean or median of paired relative changes. DNA distribution plots use per-sequence scores; the protein and additional-observer panels report group means. Appendix H specifies which analyses support record-level statistical reproduction.

5 Results

We first characterize the pAV signature and its dependence on sequence arrangement, domain, and observer. We then evaluate scalar detection against matched baselines before turning to representation-based screening and its serving cost.

5.1 A Reproducible DNA Signature

The frozen NT-2.5B observer assigns lower mean pAV to all six generated DNA populations (Figure 2). The natural reference has mean 1,264.6 and standard deviation 240.6. Relative means are 81.1% for GenomeOcean-100M, 83.6% for GenomeOcean-500M, 76.0% for GenomeOcean-4B, 52.7% for GENERator-v2, 73.5% for Evo-1, and 70.3% for Evo2-7B. Thus, the direction of the difference is shared across BPE, 6-mer, and character-level generator families under this observer.

The distributions also reveal meaningful overlap. The signature is a population comparison, and its magnitude differs across generators. A low score supplies statistical evidence under a reference distribution rather than proof of artificial origin for an individual sequence. Using the separately sampled 500-window evaluation panel, standalone pAV obtains macro AUROC 0.902 over the four model-level held-out generators and empirical TPR 74.0% at 5% FPR. Its AUROC ranges from 0.819 for GenomeOcean-500M to 0.989 for GENERator-v2. The natural reference changes between these two experiments, so their AUROCs should not be read as repeated estimates on identical samples.

5.2 Arrangement Sensitivity Under Matched Local Counts

Dinucleotide-preserving shuffling reduces mean pAV to 83.2% of the original natural mean. The median within-sequence reduction is 16.2%, and 159 of 200 windows decrease (79.5%; two-sided Wilcoxon $p = 6.71 \times 10^{-21}$). Because these windows retain their lengths, GC content, and exact 2-mer counts, those quantities alone cannot explain the paired change. The result supports sensitivity to sequence arrangement beyond these matched statistics; it does not isolate a specific biological interaction.

Other controls help localize the effect. Mononucleotide shuffling produces a median change of -15.9%, and codon-block shuffling produces -8.2%. Reverse complementation has a much smaller population shift: its mean is 101.2% of the original reference and its median paired change is $+1.8\%$ ($p = 0.50$). Individual strand differences remain appreciable. The median absolute relative difference is 14.1%, with a 90th percentile of 33.1%. Consequently, absence of a population shift is not per-sequence strand invariance, and this comparison does not establish equivalence within a prespecified margin.

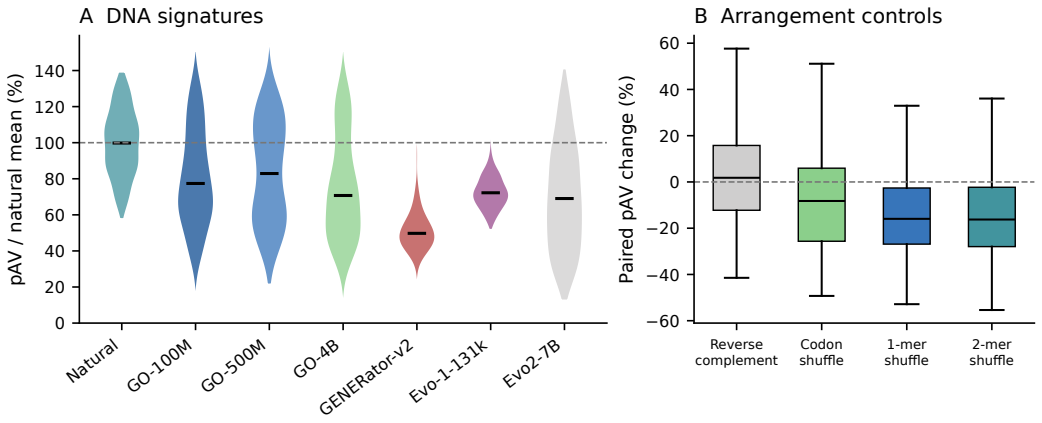

Figure 2: DNA signatures and paired controls. (A) Per-sequence pAV normalized by the natural mean, with 200 windows per population. Violin medians are shown in black. (B) Paired relative change for the same 200 natural windows. Boxes show the median and interquartile range; whiskers extend to 1.5 interquartile ranges and outliers are omitted from the display, not the analysis. Exact dinucleotide preservation with reordered positions changes pAV substantially.

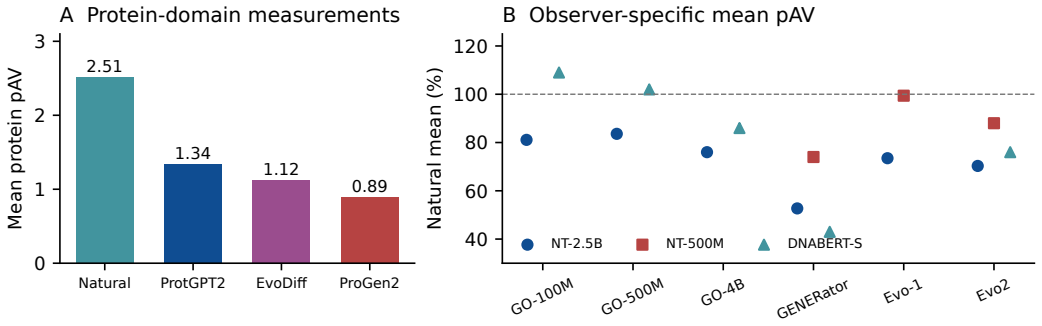

Figure 3: Protein and observer comparisons. (A) Mean pAV in the ESM2 protein panel, with 1,000 analyzed sequences per group. These are aggregate measurements; no uncertainty estimates are inferred from unavailable individual scores. (B) Selected observer-specific ratios. NT-2.5B uses the 200-window DNA panel; the other observers use separate panels. Missing points are unreported in this comparison, not measured zeros.

5.3 Domain and Observer Dependence

The protein panel provides evidence that the observation is not restricted to the DNA alphabet. Under the ESM2 observer, the recorded mean is 2.51 for Swiss-Prot, compared with 1.34 for ProtGPT2, 1.12 for EvoDiff, and 0.89 for ProGen2-small. The corresponding reported discrimination AUROCs are 0.830, 0.899, and 0.930. Both autoregressive and diffusion-generated proteins exhibit the directional difference. These aggregate results support a cross-domain empirical comparison; they do not establish that lower pAV implies reduced folding quality or function. Length- and composition-matched protein evaluation remains necessary for a stronger interpretation.

Observer choice changes what can be distinguished. The narrower NT-500M observer places Evo-1 near its natural baseline (99.4%, AUROC 0.495). DNABERT-S places GenomeOcean-100M at 109% and GenomeOcean-500M at 102% of its natural mean, while retaining a clearer GENERator-v2 difference. These measurements motivate normalization and calibration within each observer. Because training corpus, model capacity, and tokenizer vary together, the comparison does not identify tokenization as a sole causal factor. The stable conclusion is that pAV is a property of the

Table 1: Matched baseline comparison. AUROC on four model-level held-out generators, each with 500 windows, against the same 500 natural windows. Discriminative baselines share the 3,000-window training sample. pAV is a fixed, unfitted score. Macro averages exclude the separately reported GenomeOcean-4B and Evo2-7B batches.

Feature set	GO-100M	GO-500M	GENERator	Evo-1	Macro
Entropy (2-, 3-, 4-mer)	0.998	0.995	0.989	0.996	0.995
Compression (gzip, LZMA)	0.997	0.994	0.992	0.987	0.992
GC plus compression	0.996	0.993	0.987	0.979	0.989
pAV, fixed 20 heads	0.832	0.819	0.989	0.967	0.902
4-mer frequencies	0.807	0.708	0.277	0.750	0.636
Order-3 Markov LLR	0.898	0.795	0.450	0.316	0.615
GC content	0.756	0.625	0.071	0.164	0.404

chosen observation procedure as well as the input distribution. Additional aggregate panels and their coverage are in Appendix D.

Head selection is a second source of measurement dependence. In a 200-window-per-source comparison, selected-head pAV yields macro AUROC 0.908, versus 0.645 for averaging all 640 heads and 0.615 for the mean of 1,000 random 20-head subsets (Figure 4B). The supervised head set is fixed across target generators, but complete original discovery-set membership is unavailable. This result therefore describes the chosen measurement and its transfer, rather than an unsupervised whole-model invariant or demonstrated discovery/test independence.

5.4 Detection Baselines and Perturbation Stability

Does this population signature translate into a competitive detection score? On matched training and evaluation sequences, entropy over 2-, 3-, and 4-mers obtains macro AUROC 0.995 and the two-feature compression classifier obtains 0.992. The pAV-only result is 0.902, whereas k-mer frequencies, Markov likelihood ratios, and GC content yield lower discrimination in this panel (Table 1). Thus, inexpensive features provide the stronger clean-data baseline despite the clear attention signature.

The perturbation experiment identifies a setting where standalone pAV retains more discrimination. Across 0–20% substitutions in generated sequences, compression macro AUROC changes from 0.992 to 0.822, while pAV changes from 0.897 to 0.867 (Figure 4A). At 20%, the pAV advantage is 0.045 AUROC; for Evo-1 specifically, the values are 0.934 versus 0.630. This is a comparison on the same transformed inputs and unchanged natural reference. It supports stability under this stress test, not robustness to every adaptive generation strategy.

The incremental-feature test asks whether that stability benefits a fitted detector. A compression-plus-pAV logistic regression trained on clean inputs obtains macro AUROC 0.979 at 0% and 0.726 at 20%, compared with 0.992 and 0.822 for compression alone. The different degradation curves therefore do not establish useful complementarity under this training procedure. The evidence supports pAV as an arrangement-sensitive probe with distinct stress behavior, not as a uniformly beneficial classifier input.

5.5 Representation-Based DNA Screening

The application question is broader than whether a single statistic improves compression. Here we evaluate a detector that retains observer representations together with multiscale sequence features, then ask whether its observer contribution can be served inexpensively. This is a separate screening study, not a resolution of the scalar-feature result above.

The five-feature GOS detector provides a compact starting point for sequence screening. On the six-source 1,000-bp panel, its recorded evaluation yields macro AUROC 0.959 and macro recall 73.0%, with 12 of 2,000 natural windows called positive (0.6% FPR). The separate 2,500-bp benchmark yields macro AUROC 0.924 and recall 73.0%. These two window configurations use their respective fitted models and calibration sets. In a mosaic experiment, detection rises from 5.0% on natural

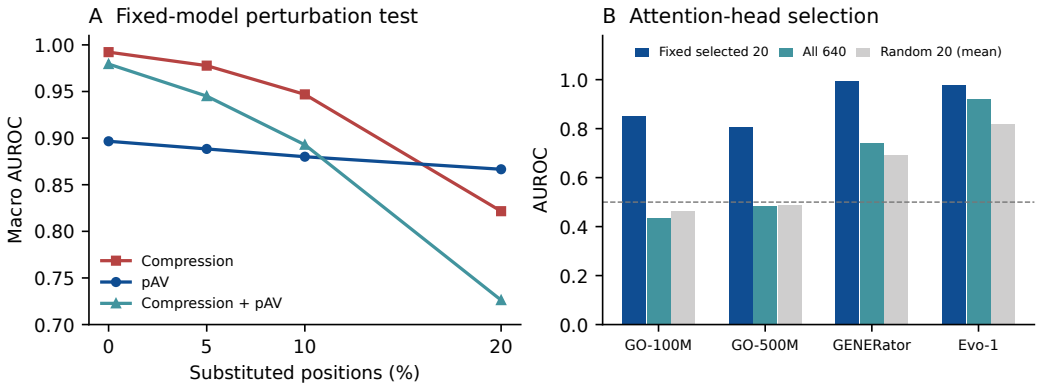

Figure 4: Perturbation and head-selection analyses. (A) Macro AUROC across four generated sources, with 250 sequences per source and substitution rate, against 500 clean natural references. Classifiers remain fixed after clean-data fitting. (B) Head-set comparison on 200 windows per source and natural reference. Random-20 bars average 1,000 subsets. Panels summarize separate experiments and are not interchangeable with the 500-window baseline table.

Table 2: Representation-fusion performance in five-fold sequence-level cross-validation. Each generated source has 500 windows. TPR values are empirical operating points on pooled out-of-fold scores with 2,000 natural references. All six source families occur in training folds; these are not unseen-generator results.

Generated source	AUROC	TPR at 1.3% FPR	TPR at 5% FPR
Evo-1	0.998	98.0%	99.4%
Evo2-7B	0.986	86.4%	93.0%
GENERator-v2	0.995	97.0%	98.6%
GenomeOcean-100M	0.998	97.8%	99.6%
GenomeOcean-4B	0.995	90.4%	97.8%
GenomeOcean-500M	0.997	95.8%	99.2%
Macro	0.995	94.2%	97.9%

windows to 91.0% on fully generated windows, showing a graded response to synthetic content. Appendix G reports the complete comparisons and natural-source error rates.

The representation-fusion detector achieves macro AUROC 0.995 in the 5,000-window sequence-level cross-validation experiment. At the empirical 1.3% natural-FPR point, it detects 2,827 of 3,000 generated windows (94.2%) with 26 of 2,000 natural windows called positive. Recall is 93.4% at 1.0% FPR and 97.9% at 5.0% FPR. Table 2 gives source-level results. These values establish high discrimination within the evaluated panel. Because every source family occurs in training folds, they do not measure unseen-generator transfer or provide a controlled improvement over the separately evaluated five-feature model.

5.6 Distillation and Serving Cost

The practical bottleneck of representation fusion is computing observer features. A six-epoch GenomeOcean-100M v1.2 student configuration approximates the observer contribution and reaches out-of-fold macro AUROC 0.996, with 97.3% macro recall at the 1.3% empirical FPR boundary. Its Evo2-7B recall is 89.2% at 1.3% FPR and 94.6% at 5% FPR. The configuration was selected from a one-to-six-epoch sweep on this panel, so these results describe the selected configuration rather than an untouched final holdout or established superiority over the teacher.

The service separates CPU feature preparation from a resident GPU student using bounded queues. On one NVIDIA GB10 device, processing 5,000 windows takes 14.941 seconds, or 334.6 sequences per second end to end. Peak allocated GPU memory is 0.573 GB. This is a measured finite-batch

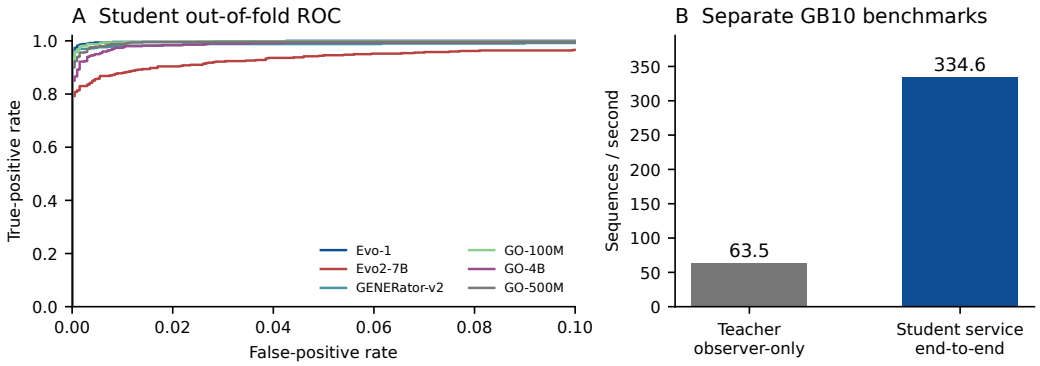

Figure 5: Compact screening. (A) Source-specific ROC curves recomputed from the six-epoch student's 5,000 out-of-fold scores, showing the low-FPR region. (B) Separate timing boundaries: teacher observer-only serving and full student-service throughput on NVIDIA GB10. The service benchmark uses a two-epoch checkpoint of the same student architecture, not the six-epoch checkpoint in panel A; bars are not a paired speedup measurement.

throughput after model initialization; the equivalent 1.205 million sequences per hour is a rate conversion, not an hour-long load test. Median end-to-end latency is 7.36 seconds and the 95th percentile is 10.97 seconds, including queuing in this workload. Service replay accuracy is not substituted for out-of-fold accuracy because the resident model is refitted on the panel.

6 Discussion and Conclusion

The central measurement finding is arrangement sensitivity: pAV changes when sequence order changes even with length and exact dinucleotide counts held fixed. Its lower mean across the DNA generators and the analogous protein comparison make it a useful descriptive probe. Observer and head-set dependence explain why its interpretation belongs to a defined measurement procedure rather than an observer-independent scale of biological quality.

The detection results sharpen that interpretation. Compression and entropy are strong clean-data baselines; pAV degrades differently under substitution, but its inclusion does not improve the tested clean-fitted scalar fusion. Representation-based screening addresses a different objective by retaining richer observer information, and distillation reduces the cost of that approach. These results connect measurement and application without making one a proxy for the other's success.

Several boundaries determine the next validation steps. Head selection is supervised, discovery and pretraining relationships are not fully resolved, and the protein analysis lacks a record-level matched reanalysis. Fusion and student results require generator-held-out testing and independent source-level calibration before supporting deployment error guarantees. The six-epoch accuracy study and two-epoch service benchmark must also be tested together before claiming a joint accuracy–throughput operating point. Within the evaluated settings, pAV characterizes sequence distributions and the screening models support provenance-aware triage, not proof of molecular function or an automatic exclusion rule.

AI USE STATEMENT

Generative AI tools were used for two categories of work. First, AI assistance supported manuscript organization, drafting, language editing, figure-generation code, reference retrieval, and checks of reported calculations. The authors are responsible for the final text, references, and interpretations. Second, biological generative models produced the synthetic DNA and protein datasets analyzed in this study; the model families and experimental settings are described in Section 4 and Appendix A. Generated sequences are treated as model outputs, not as experimentally validated biological functions.

ETHICS STATEMENT

This study concerns statistical characterization and provenance screening of biological sequences. A generated-sequence score is neither a functional assay nor a determination of harmfulness. False positives can affect legitimate synthetic biology and environmental sampling, so screening should use a representative natural reference and human review. We do not report functional enhancement or experimental synthesis of sequences. Data and model redistribution must follow the licenses of their respective providers.

REPRODUCIBILITY STATEMENT

Appendices A–H specify preprocessing, head indices, comparison protocols, and numerical reproduction coverage. The accompanying CPU script checks the supplied per-sequence attention scores, refitted-baseline scores, student out-of-fold logits, and aggregate arithmetic. Frozen-observer extraction and detector training are separate operations and are not rerun by this numerical package.

REFERENCES

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Neil Tenenholtz, Robert Strome, Alan M. Moses, Alex X. Lu, Nicolò Fusi, Ava P. Amini, and Kevin K. Yang. Protein generation with evolutionary diffusion: sequence is all you need. *eLife Reviewed Preprint*, 2026.

L. V. Alteio, F. Schulz, R. Seshadri, N. Varghese, W. Rodriguez-Reillo, E. Ryan, D. Goudeau, S. A. Eichorst, R. R. Malmstrom, R. M. Bowers, L. A. Katz, J. L. Blanchard, and T. Woyke. Complementary Metagenomic Approaches Improve Reconstruction of Microbial Diversity in a Forest Soil. *mSystems*, 2020.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. In *International Conference on Learning Representations*, 2024.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Jonathan C. Schmok, Ali Taghibakhshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Michael T. Pearce, Elana Simon, Etowah Adams, Zachary J. Amador, Euan A. Ashley, Stephen A. Baccus, Haoyu Dai, Steven Dillmann, Stefano Ermon, Daniel Guo, Michael H. Herschl, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R. K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, John St. John, Jeremy Sullivan, Joseph Tey, Ben Viggiano, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Sudarshan Pinglay, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modelling and design across all domains of life with Evo 2. *Nature*, 2026.

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, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 2025.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 2022.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 2023.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *International Conference on Machine Learning*, 2023.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. In *Advances in Neural Information Processing Systems*, 2023.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 2024.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 2023.

Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 2016.

Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, 2018.

Jeremy Ratcliff. Transformer model generated bacteriophage genomes are compositionally distinct from natural sequences. *NAR Genomics and Bioinformatics*, 2024.

Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R. Mende, Adriana Alberti, Francisco M. Cornejo-Castillo, Paul I. Costea, Corinne Cruaud, Francesco d'Ovidio, Stefan Engelen, Isabel Ferrera, Josep M. Gasol, Lionel Guidi, Falk Hildebrand, Florian Kokoszka, Cyrille Lepoivre, Gipsi Lima-Mendez, Julie Poulain, Bonnie T. Poulos, Marta Royo-Llonch, Hugo Sarmento, Sara Vieira-Silva, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans coordinators, Chris Bowler, Colomban de Vargas, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Fabrice Not, Hiroyuki Ogata, Stephane Pesant, Sabrina Speich, Lars Stemmann, Matthew B. Sullivan, Jean Weissenbach, Patrick Wincker, Eric Karsenti, Jeroen Raes, Silvia G. Acinas, and Peer Bork. Structure and function of the global ocean microbiome. *Science*, 2015.

Aidan P. Tay, Kieran Didi, Anuradha Wickramarachchi, Denis C. Bauer, Laurence O. W. Wilson, and Maciej Maselko. Synsor: a tool for alignment-free detection of engineered DNA sequences. *Frontiers in Bioengineering and Biotechnology*, 2024.

The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. *Nature*, 2019.

The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017.

Wei Wu, Qiuyi Li, Yuanyuan Zhang, Zhihao Zhan, Ruipu Chen, Mingyang Li, Kun Fu, Junyan Qi, Yongzhou Bao, Chao Wang, Yiheng Zhu, Zhiyun Zhang, Jian Tang, Fuli Feng, Jieping Ye, Yuwen Liu, Hui Xiong, and Zheng Wang. GENERator: A Long-Context Generative Genomic Foundation Model. *arXiv preprint arXiv:2502.07272*, 2025a.

Weimin Wu, Zhihan Zhou, Robert Riley, Mohammad Abdulqader, Xuefeng Song, Satria Kautsar, Rob Egan, Steven Hofmeyr, Gefei Liu, Shira Goldhaber-Gordon, Mutian Yu, Harrison Ho, Yaping Liu, Andrei Stecca Steindorff, Fengchen Liu, Feng Chen, Rachael Morgan-Kiss, Lizhen Shi, Han Liu, and Zhong Wang. Uncovering the Genomic Manifold via Scalable Learning from the Global Microbiome. *bioRxiv*, 2025b.

Zhihan Zhou, Weimin Wu, Harrison Ho, Jiayi Wang, Lizhen Shi, Ramana V Davuluri, Zhong Wang, and Han Liu. DNABERT-S: Pioneering Species Differentiation with Species-Aware DNA Embeddings. *Bioinformatics*, 2025.

A Data and Evaluation Protocols

A.1 Sources and Sample Construction

DNA training references combine GTDB representatives, chromosomal/mitochondrial/plasmid records, human and viral sequence, and Antarctic, GRE, HMP2, Mendota, and Tara metagenomic collections. Generated candidate training inputs comprise GenomeOcean-4B high- and low-complexity CSV files and an Evo2 training file. The sequence loader recognizes FASTA and one-sequence-per-line input, removes all internal whitespace, converts to uppercase, and splits at symbols outside ACGT. From every fragment of length at least 1,000, it takes the central 1,000 bases. It deduplicates exact windows and uses deterministic shuffling.

Natural sampling draws round-robin across source files after per-file shuffling. The generated-training loader collects qualifying distinct windows in file order, with an early collection limit, then shuffles the pooled windows. All 1,500 selected generated training windows map to the GenomeOcean-4B high-complexity input; the low-complexity and Evo2 candidate files contribute none. Recorded temperatures span 0.2–1.6 in increments of 0.2, top-k values include the unrestricted sentinel and 10–100 in increments of 10, top-p values span 0.05–0.95 in increments of 0.10, and recorded generation seeds are 42, 123, and 1024. These are observed settings, not a claim that their full Cartesian product is balanced. The four-source macro panel is kept fixed rather than enlarged after observing this composition.

A.2 Generation and Homology

The model identities are GenomeOcean-100M/500M/4B, GENERator-v2-Eukaryote-3B, Evo-1-131k, and Evo2-7B. These are source-specific generated corpora, not a factorial experiment that holds tokenizer, model size, training data, prompt, and sampling settings fixed. Claims concern the evaluated corpora rather than causal effects of architecture. Protein generators are ProtGPT2, ProGen2-small, and EvoDiff-seq-640M. Exact generation configurations and model revisions should accompany regenerated corpora; values not recoverable from the retained records are not imputed from library defaults.

A supplementary homology screen evaluated 50 GenomeOcean-4B and 50 Evo2-7B sequences against NCBI nt using BLASTn. The recorded no-hit rates are 66% and 90%, respectively. Significant-hit rates at $E < 10^{-5}$ are 10% and 2%; remaining weak-hit rates are 24% and 8%. This is a small,

Table 3: Separate evaluation panels. Counts denote analyzed windows or protein sequences, not independent genomes or model-training seeds.

Panel	Size	Purpose
DNA characterization	200 natural; 200 per six generators	Population pAV distributions
Paired controls	Same 200 natural windows	Controlled sequence transformations
Matched baselines	1,500 training per class; 500 test natural; 500 per six generators	Fixed four-model macro panel, two additional batches
Substitution stress	500 natural; 250 per four generators per rate	Fixed clean-trained classifiers
Head comparison	200 natural; 200 per four generators	Fixed, all-head, and random subsets
Protein extension	Reported 1,000 per four groups	Separate ESM2 aggregate comparison
Representation fusion	3,000 generated; 2,000 natural	Five-fold sequence-level cross-validation
Service timing	5,000 windows	Finite-batch throughput and latency

database-dependent novelty screen, not a direct classifier with a matched natural false-positive rate. Absence of a hit does not establish biological plausibility or synthetic origin.

B Attention Measurement and Selected Heads

NT-2.5B has 32 layers with 20 attention heads each. Flattening is layer-major, with zero-based index $20\ell + h$. The fixed set is

578	171	558	320	635	574	564	537	494	547
465	530	253	502	370	625	143	351	361	321

For each attention tensor, active rows and columns are sliced using the attention mask. The weighted token distance is summed across keys, and the population variance is computed across interior queries with an $L - 2$ divisor. The mean of the 20 resulting head scores is used without fitting a new pAV classifier. Float32 attention probabilities and accumulation are used. Sequence length, token count, special-token conventions, and boundary treatment are part of the measurement definition.

The head-comparison experiment applies this fixed set, all 640 heads, and 1,000 random 20-head subsets to the same sequences. Random subsets are sampled without replacement within each subset using NumPy seed 42. They are alternatives for a sensitivity analysis, not independent biological replicates. The selected set transfers without reselection across the four named targets, but original discovery-set membership is not completely available. Accordingly, this experiment measures fixed-head transfer, not a proof of discovery/test independence.

B.1 Token Geometry and Interpretation

Position-dependent mean distances need not require sequence-specific attention. For a uniform row distribution $A_{jk} = 1/L$ over zero-based positions, direct summation gives

$$\mathrm{MAD}_j = \frac{j(j+1) + (L-1-j)(L-j)}{2L}. \tag{4}$$

This quantity varies with query position even though every row assigns identical key probabilities. Thus pAV is not synonymous with attention entropy or a distance-free measure of biological complexity. Matched token handling, window length, and the paired controls are essential to interpreting a measured population difference. In particular, larger absolute pAV at longer windows is not on its own evidence of greater biological organization, because the distance scale itself changes.

Table 4: Head-set discrimination. Entries are AUROC; random subsets are summarized by their mean and 5th–95th percentile, not a confidence interval over new datasets.

Source	Fixed 20	All 640	Random mean	Random range
GO-100M	0.851	0.434	0.462	0.226–0.709
GO-500M	0.805	0.483	0.489	0.271–0.708
GENERator-v2	0.995	0.742	0.691	0.368–0.906
Evo-1	0.980	0.920	0.820	0.547–0.963
Macro	0.908	0.645	0.615	–

C Full Paired-Control Results

For each transformed score s'_i and original score s_i, the paired relative change is $100(s'_i - s_i)/s_i$. Its median is different from $100(\overline{s'}/\overline{s} - 1)$; both are useful but must be named explicitly. The 200 paired windows are the unit of the signed-rank test. Small p-values describe this sample and transformation, not independent generator-training replications.

Table 5: Natural-window controls, $n = 200$. The first numeric column is a ratio of group means. The next two summarize within-sequence changes. All p-values are two-sided Wilcoxon comparisons with identity.

Transformation	Mean ratio (%)	Median change (%)	Decreasing (%)	p
Reverse complement	101.2	+1.8	47.5	0.50
Codon-block shuffle	89.9	−8.2	65.0	3.93×10^{-9}
Mononucleotide shuffle	83.6	−15.9	77.5	4.79×10^{-19}
Dinucleotide shuffle	83.2	−16.2	79.5	6.71×10^{-21}
Order-1 Markov resampling	83.3	−16.2	79.5	8.54×10^{-20}
Order-3 Markov resampling	80.8	−19.2	85.0	9.26×10^{-26}
Order-5 Markov resampling	74.7	−25.8	91.0	7.46×10^{-32}

Exact dinucleotide shuffling is the cleanest count-matched control in this panel. Markov resampling preserves fitted transition statistics only in expectation. Fitting high-order chains on a single 1,000-bp window yields sparse contexts, particularly for order 5, and can introduce resampling artifacts. We therefore do not interpret the monotone Markov trend as a calibrated biological scale. Reverse-complement differences describe observer orientation sensitivity, not random measurement noise from repeated identical inference.

D Protein and Observer-Specific Aggregates

The protein comparison uses ESM2-650M, a 33-layer model with 20 heads per layer. The natural group contains reviewed Swiss-Prot proteins in the stated length range. The per-group means are 2.51, 1.34, 1.12, and 0.89 for natural, ProtGPT2, EvoDiff, and ProGen2, respectively, with generated-versus-natural AUROCs of 0.830, 0.899, and 0.930. Only group-level results are available for this panel. Figure 3A therefore reports means without uncertainty intervals, and length-matched reanalysis or independent AUROC recomputation is not supported.

For NT-500M, the retained values include GENERator-v2 at 74%, Evo-1 at 99.4%, and Evo2-7B at 88% of its natural mean. DNABERT-S values include GenomeOcean-100M at 109%, GenomeOcean-500M at 102%, GenomeOcean-4B at 86%, GENERator-v2 at 43%, and Evo2-7B at 76%. Within-observer normalization is essential. Comparisons across these separately sampled panels are descriptive and do not isolate training-corpus or tokenizer effects.

E BASELINES, PERTURBATIONS, AND INCREMENTAL CONTRIBUTION

E.1 FIVE-FEATURE GOS DEFINITIONS

For an ACGT sequence of length N, the two compression features divide the complete gzip or LZMA compressed byte length by the uncompressed byte length. The gzip implementation uses its default level 9; LZMA uses its default preset. Container overhead is included, unlike the payload-adjusted ratios in the multiscale branch. Lower ratios indicate greater compressibility, not greater complexity.

The original periodicity feature contrasts normalized circular autocorrelations at two fixed lag sets. For base a, define the centered indicator $u_{a,i} = \mathbf{1}\{x_i = a\} - N^{-1}\sum_{j=0}^{N-1}\mathbf{1}\{x_j = a\}$. Its autocorrelation is

$$c_a(d) = \frac{\sum_{i=0}^{N-1} u_{a,i}u_{a,(i+d) \bmod N}}{\sum_{i=0}^{N-1} u_{a,i}^2}, \qquad \overline{c}(d) = \frac{1}{4}\sum_{a\in\{A,C,G,T\}} c_a(d). \tag{5}$$

A constant channel contributes zero. The implementation uses an unpadded FFT and its inverse, giving circular rather than linear autocorrelation. With $D_1 = \{5, 6, 10, 12, 15, 18, 21, 24, 30\}$ and $D_0 = \{2, 3, 4, 7, 8, 9, 11, 13, 14\}$, the feature is

$$P(x) = \frac{1}{|D_1|}\sum_{d\in D_1}\overline{c}(d) - \frac{1}{|D_0|}\sum_{d\in D_0}\overline{c}(d). \tag{6}$$

It is set to zero for $N < 31$. This is a specified lag contrast, not a causal attribution of a sequence pattern to its tokenizer. The multiscale representation branch uses a separate linear-autocorrelation implementation.

Stop-codon density counts TAA, TAG, and TGA in each of the three reading frames of both the forward sequence and its reverse complement, then divides by the total number of complete triplets visited. For $N \geq 3$, the denominator is $2(N-2)$; shorter sequences receive zero. All sliding triplets are visited once per strand. The measurement does not require an annotated coding region or a predicted open reading frame.

E.2 MATCHED CPU BASELINES

Compression features are compressed byte length divided by input byte length, using gzip level 9 and default LZMA. Entropy features are empirical Shannon entropies over overlapping words of lengths 2, 3, and 4. K-mer classifiers use normalized frequencies for $k \in \{4, 5, 6\}$. The GC baseline uses the fraction of G and C bases. Discriminative baselines use StandardScaler and balanced logistic regression with a maximum of 2,000 iterations. Both transformations are fitted on training data. Markov models use context orders 3 or 5, unit pseudocounts, and length-normalized generated-versus-natural log-likelihood ratios.

All methods share the natural floor and four model-held-out targets in the main comparison. The six-source panel additionally contains GenomeOcean-4B and Evo2-7B batches, which are not included in that fixed macro average. GenomeOcean-4B shares the baseline training generator; Evo2-7B does not occur in the actual selected training sample. Standalone pAV AUROC on these batches is 0.870 and 0.913. Compression AUROC is 0.980 and 0.850. These values characterize source-specific trade-offs without redefining the macro denominator.

Table 6: Incremental scalar-feature analysis under substitution stress. The natural floor is fixed and classifiers are fitted on clean data. Each rate uses 250 windows per generated source. All values are macro AUROC over four targets. Differences are computed before rounding, so they can differ from subtraction of the displayed entries.

Rate (%)	Compression	+ pAV	Difference	Entropy + comp.	+ pAV	pAV only
0	0.992	0.979	−0.013	0.995	0.989	0.897
5	0.978	0.945	−0.033	0.985	0.967	0.888
10	0.947	0.893	−0.054	0.958	0.922	0.880
20	0.822	0.726	−0.095	0.812	0.727	0.867

The perturbation generator independently replaces selected bases with another base, using a fixed random stream. Because substitutions are not conditioned on preserving coding or function, the sweep is a statistical stress test. It does not demonstrate a function-preserving adversarial attack. Results from a classifier refitted at each rate would answer a different question and are not pooled with these fixed-model values.

F Representation Fusion and Distillation

The fusion feature order is an availability indicator, 2,560 layer-24 mean hidden features, 640 attention-entropy summaries, and 47 CPU multiscale summaries, totaling 3,248. In the complete-observer cross-validation panel, the availability indicator is constant. The fitted pipeline uses training-fold StandardScaler followed by balanced logistic regression with the liblinear solver and a 1,000-iteration limit. Five StratifiedKFold splits use random seed 7 and source stratification. The 5,000 input sequences are distinct under exact matching, and each receives one out-of-fold score. Thresholds are selected from pooled out-of-fold natural scores; they are empirical ROC summaries rather than external calibration estimates.

The fusion logit is decomposed algebraically into an observer component and a CPU-plus-intercept component. Fold-specific teacher coefficients and normalizers generate student targets using the corresponding training fold. A global teacher is also fitted for final service construction; its in-sample outputs must not replace fold-specific evaluation. A 116,411,905-parameter GenomeOcean-100M v1.2 model provides the compact student. The studied configuration uses batch size 16, learning rate 2×10^{-5}, weight decay 0.01, a binary-cross-entropy weight of 0.25, and up to six epochs. Its token budget is 200; the interpretation is the actual truncated/tokenized input, not unrestricted full-length coverage.

The six-epoch sweep reports observer-logit RMSE 3.544 and out-of-fold macro AUROC 0.996. At empirical 1.3% FPR it reports macro recall 97.3%, Evo2 recall 89.2%, and GenomeOcean-4B recall 98.0%. At 5% FPR the corresponding figures are 98.7%, 94.6%, and 99.4%. Epoch choice uses this same evaluation panel, so a separate final test would be needed for an unbiased assessment of the selection procedure.

F.1 Computational Measurement

The service uses six CPU feature workers, bounded work and feature queues, and one resident GPU student. It streams the 5,000-window panel in 128-sequence chunks, with GPU sub-batching. The recorded NVIDIA GB10 run uses BF16 inference. End-to-end wall time is measured from stream start through consumption of the final result and excludes model loading. Throughput is row count divided by that wall time. Latency includes the queueing induced by this finite workload; it is not single-request model latency.

Peak PyTorch allocation (0.573 GB) is distinct from total device memory reserved by the framework or operating system. The finite run does not establish steady-state capacity under arbitrary arrival processes, nor does its refitted model supply an independent accuracy test. Teacher serving, student forward-only capacity, and full-service throughput have different boundaries and are not placed in a single unqualified speedup ratio.

G Additional Detection and Sampling Analyses

G.1 Window Length and Natural Background

The five-feature detector was evaluated at 1,000 and 2,500 bp. The longer-window experiment uses a separate fitted configuration and natural-source panel. It therefore provides an application-level comparison rather than a controlled causal estimate of increasing input length. Its per-source natural false-positive rates range from 0% for GTDB to 11.5% for mitochondrial sequences; the overall recorded rate is 3.3%. The remaining rates are 0.4% for GRE, 0.4% for Harvard Forest, 1.3% for HMP2, 3.8% for Mendota, 0.8% for Tara, and 6.5% for Antarctic metagenomes. Domain-matched calibration is required when these sources dominate an application.

Table 7: Five-feature GOS and likelihood-based comparisons. AUROC values are from their respective recorded benchmark panels. The perturbation score is a DNA-specific approximation described below, not an exact implementation of published Fast-DetectGPT. GOS is fitted; likelihood scores require no labeled detector fitting.

	1,000 bp			2,500 bp		
Source	GOS	Perturbation	pLL	GOS	Perturbation	pLL
Evo-1	0.991	0.524	0.526	0.982	0.519	0.862
Evo2-7B	0.906	0.665	0.769	0.852	0.628	0.560
GENERator-v2	0.994	0.864	0.938	0.993	0.853	0.960
GO-100M	0.967	0.914	0.902	0.955	0.952	0.979
GO-4B	0.938	0.838	0.903	0.870	0.910	0.955
GO-500M	0.959	0.918	0.862	0.890	0.936	0.963

The DNA likelihood-perturbation score subtracts the mean approximate pseudo-log-likelihood of five 15%-substituted variants from the unperturbed approximate pseudo-log-likelihood. Its all-masked single-pass approximation and nucleotide perturbations differ from the conditional probability curvature estimator in Fast-DetectGPT (Bao et al., 2024). We report the implemented estimator under its own description. Raw pLL supplies an additional ablation. Recorded likelihood evaluations use 200 generated sequences per source and length; their thresholds must not be compared with GOS as if both had the same natural operating FPR. The table therefore reports threshold-free AUROC and is separate from the matched baseline experiment.

A separate HyenaDNA evaluation uses 1,000 generated sequences at 1,000 bp and obtains GOS AUROC 0.918 and TPR 52.2% at the specified 5% calibration target. This external architecture check applies to the five-feature model, not the representation-fusion or student experiments. It is retained as an aggregate result rather than pooled into either macro average.

G.2 Temperature, Local Noise, and Mixtures

The GenomeOcean-4B temperature sweep spans $T = 0.2$ to 2.0, with 200 sequences per temperature bin. Recorded mean pAV remains below the natural reference at high temperatures, approaching roughly 75% of the reference mean. From $T = 1$ to $T = 2$, five-feature GOS detection changes from 87.5% to 76.0%, while its AUROC changes from 0.963 to 0.911. These are sampling-sensitivity measurements for one generator, not a guarantee against adaptive selection. The natural point-substitution analysis is also distinct from generated-sequence stress: a paired 150-window experiment has a mean pAV change of $+8.0$ at 10% substitution ($p = 0.32$), whereas a separately normalized 500-window panel reports about 95% of its initial mean. Different samples and normalizations are not merged into one trajectory.

Mosaic windows are formed by inserting a central GenomeOcean-100M fragment into a natural 1,000-bp window, with 200 examples per generated fraction. GOS uses its 5% natural-reference operating threshold, while Synsor uses probability greater than 0.5. The resulting rates in Table 8 describe these two operating points, not a matched-FPR ranking of the methods.

Table 8: Mosaic sequence detection, $n = 200$ per fraction. The zero-fraction row is the achieved natural false-positive rate for this experiment.

Generated content (%)	GOS detection (%)	Synsor detection (%)
0	5.0	2.0
10	5.0	3.0
25	7.0	4.0
50	24.0	5.5
75	47.0	11.0
90	67.5	18.5
100	91.0	21.5

H CALIBRATION AND REPRODUCTION COVERAGE

The intended output is a screening score under a specified reference population. Empirical FPR measured on a mixed natural pool need not transfer to a new taxon or sequence source. Source-specific FPR should be measured before deployment, especially for compositionally unusual organellar or environmental sequences. A calibrated threshold also requires a calibration set independent of the final assessment if it is to support an out-of-sample error statement.

Table 9: Numerical reproduction coverage. Recomputing a statistic from retained scores is distinct from rerunning its generating model.

Analysis	Available verification scope
DNA distributions and controls	Per-sequence pAV; means, paired changes, tests, and plots can be recomputed
Standalone pAV detection	Per-sequence natural and generated scores; AUROC and empirical TPR can be recomputed
CPU baselines	Original sequences and deterministic fitting protocol; re-fitted scores and metrics retained
Perturbation and head comparisons	Per-source aggregate metrics; macro arithmetic checked; observer outputs are not fully available
Protein and additional observers	Aggregate measurements; per-sequence reanalysis is not supported by the retained numerical artifact
Representation fusion	Protocol code, fold summaries, source counts, and aggregate metrics; feature caches and checkpoints are not included
Compact student	Per-sequence out-of-fold logits for six epochs; AUROCs and operating-point counts can be recomputed; training is not rerun
Service timing	Hardware, row counts, timing and resource summaries; throughput arithmetic can be checked

The accompanying CPU script recomputes score-based metrics and checks aggregate arithmetic and file hashes. Biological validation, generator-held-out evaluation of the fusion model, model training, and independently calibrated deployment error rates are outside this numerical reproduction package. The figures use the corresponding numerical records; aggregate-only panels cannot support reconstruction of individual sequence scores.