DO GENOMIC LANGUAGE MODELS INHERIT THE RIGHT ASSUMPTIONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Genomic language models largely inherit three choices from natural-language models: subword tokenization, learned input embeddings, and likelihood-based pretraining. We test whether these choices align with DNA using controlled comparisons that hold architecture, data, and nucleotide budget fixed, together with model-free diagnostics.

Under DNABERT-2's published merge rules, BPE is mutation-stable: a substitution moves boundaries by a mean radius of 2.97 bp, less than one mean token length. The larger mismatch is placement. At coding-sequence edges, boundaries occur at $0.636\times$ a length-matched null; the depletion replicates across eleven windows spanning six chromosomes and persists against a compositional control with the opposite predicted sign. Landmark stratification shows depletion at splice sites but enrichment at stop codons.

Reverse-complement symmetry separates into segmentation, embedding, and contextual-representation levels. Five released embedding tables give $V/\mathrm{null} \in [0.85, 1.08]$; only two have enough reverse-complement pairs for formal inference, and an architecturally equivariant model still has an unconstrained embedding table.

Across four tokenizer arms, repeats receive 63–83% of total loss improvement versus 0.3–0.8% for the CDS-assigned class. In a controlled dose intervention, increasing repeat loss weight lowers retained splice-site signal in mean-pooled representations from 0.916 to 0.517 relative to a randomly initialised backbone across two seeds (blocked slope -0.352, 95% CI $[-0.499, -0.204]$). Placebo arms make a batch-only explanation less consistent with the observed pattern. These results motivate evaluating inherited NLP design choices against genomic invariances and functional structure rather than assuming direct transfer.

1 INTRODUCTION

Genomic function depends on sequence content and positional context, yet only a fraction of genomic positions have experimentally resolved annotations. DNA encoders therefore aim to make functional information recoverable from sequence without an assay for every query.

Genomic language models (gLMs) have expanded from DNABERT (Ji et al., 2021) through Nucleotide Transformer (Dalla-Torre et al., 2023), DNABERT-2 (Zhou et al., 2024), HyenaDNA (Nguyen et al., 2023), Caduceus (Schiff et al., 2024), and Evo (Nguyen et al., 2024), reaching megabase contexts and sequence generation. Many design choices still follow NLP conventions: frequency-driven subword tokenization (Sennrich et al., 2016), unconstrained learned input embeddings, and likelihood objectives whose improvement is commonly read as better sequence modelling.

DNA makes these assumptions testable. Coding and regulatory functions depend on position; a sequence and its reverse complement describe the same double-stranded sequence; and repeats occupy roughly half of the human genome while only 1–2% is protein-coding. A tokenizer can therefore be efficient yet place boundaries poorly, downstream equivariance cannot repair orientation-dependent input segmentation, and likelihood can be dominated by abundant sequence classes.

We ask which inherited choices are supported by genomic evidence. BPE boundaries are depleted at coding landmarks, released embedding tables show little reverse-complement correspondence

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.

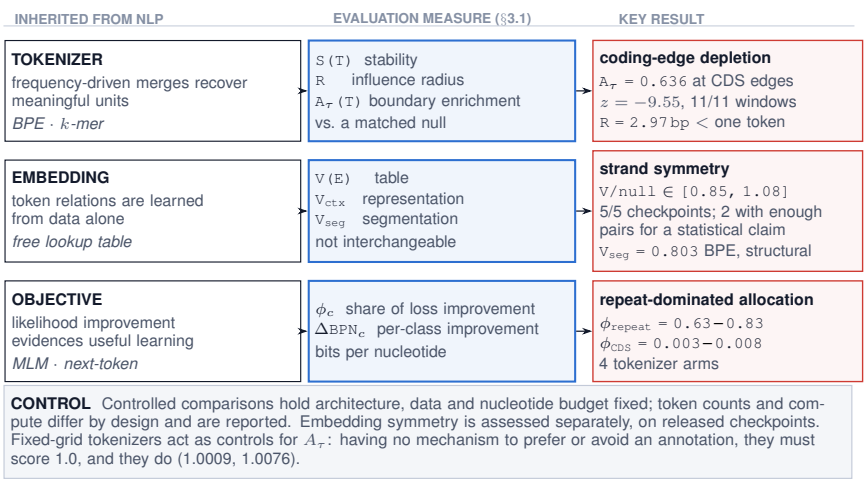


Figure 1: Three inherited assumptions, evaluation measures, and main findings. Controlled tokenizer and objective arms match architecture, data, and nucleotide budget (Appendix C).

beyond matched random baselines, and most loss improvement accrues to repeats rather than the CDS-assigned class. Because adjacent gLMs also differ in architecture, data, context, and objective, we isolate individual choices where possible while matching architecture, data, and nucleotide budget. The claim is not that these choices are universally wrong, but that their assumptions require genomic calibration. Figure 1 summarises the evaluation.

Contributions.

- Calibrated tokenizer measures showing mutation-stable BPE but 36% coding-edge depletion, replicated across eleven windows on six chromosomes (§4).
- A three-level reverse-complement decomposition separating segmentation, embedding, and contextual representation (§5).
- Annotation-wise loss accounting: repeats receive 63–83% of improvement and CDS-assigned sequence 0.3–0.8% (§6).
- A controlled dose intervention raising mean-pooled retained splice signal from 0.517 to 0.916, with placebo and similarity controls (§8).
- Downstream validity checks against simple baselines (§7) and seven silent implementation failures (Appendix P).

2 Related Work

Genomic sequence encoders. DNABERT's overlapping k-mers can leak masked content, whereas non-overlapping 6-mers introduce phase-dependent segmentation (Dalla-Torre et al., 2023). DNABERT-2 uses BPE (Sennrich et al., 2016); HyenaDNA uses single-base long contexts; Caduceus adds architectural RC equivariance; and Evo scales autoregressive generation. Adjacent models change multiple axes (Appendix A), limiting causal attribution.

Assumptions already questioned. DNABERT-2 addressed k-mer masking leakage while redesigning tokenization, and Caduceus imposed strand equivariance architecturally. We evaluate these assumptions directly and add objective allocation across sequence classes.

Evaluation. Compression does not reveal biological boundary placement (Appendix P), while zero-shot variant likelihood (Benegas et al., 2023; Frazer et al., 2021) and linear probes require simple controls. In §7, composition and 4-mer baselines expose both limitations.

3 EXPERIMENTS

3.1 EVALUATION MEASURES

Let a tokenizer T induce a boundary set $B(x)$ on $x \in \Sigma^L$, with compression $\rho = L/|T(x)|$. The two tokenizer measures below operate on boundary placement rather than token identities; Appendix I gives the full definitions for all evaluation measures.

Segmentation stability $S(T)$ and influence radius R. $S(T)$ is one minus the expected Sørensen–Dice disagreement between $B(x)$ and $B(x^{(p\to a)})$ under a single substitution, while R is the distance from the edited position p to the farthest boundary that moves. Defining stability on boundary sets rather than token edit distance matters because a substitution necessarily changes the *content* of the covering token, giving token edit distance a floor of $1/|T(x)|$ and a single-nucleotide tokenizer a score of 0.996 rather than 1. Fixed-grid tokenizers have $S(T) = 1$ and $R = 0$ by construction, so R carries the magnitude comparison for mutation-sensitive segmentation; we report it with 95% bootstrap intervals rather than against a threshold.

Annotation boundary alignment $A_\tau(T)$. $A_\tau(T)$ is the fraction of annotated landmarks within τ of a token boundary divided by the same quantity under a matched null. The matched null is essential because a raw hit rate rewards boundary density: a single-nucleotide tokenizer would achieve a raw hit rate of 1.0 trivially. *Length permutation* shuffles the tokenizer's own segment lengths, preserving their number and distribution while destroying only placement; for fixed-width tokenizers this is vacuous, so those arms use a *circular phase shift*. Under either null a single-nucleotide tokenizer gives $A_\tau = 1$ exactly, providing a fixed point for the scale. We report $\tau = 0$, where power is highest and type-I error is 0.022–0.033 against a nominal 0.05 (Appendices L and M).

Reverse-complement violation at three levels. We distinguish $V(E)$ at the embedding table, V_{ctx} at the pooled representation, and V_{seg} at the segmentation. These levels are measurable separately and are not interchangeable; this paper reports $V(E)$ and V_{seg}, leaving the contextual level to a layer-wise audit. An asymmetric embedding table can be tied and an asymmetric model constrained, but a tokenizer with $V_{\mathrm{seg}} > 0$ cuts the two strands in different places, and no downstream constraint can make the input segmentation itself equivariant. $V(E)$ requires a matched null because cosine distance between arbitrary rows of a trained table is not 1. By construction, $0 \le V_{\mathrm{seg}} \le 1$, with 0 denoting identical mirrored boundary sets and larger values indicating greater segmentation disagreement. Because fixed grids are phase-sensitive, V_{seg} must still be interpreted as a function of window length: a 6-mer scores 0.99 on 512 bp and exactly 0 on 510 bp (Appendix N).

Objective decomposition ϕ_c. Each token's negative log-likelihood is attributed to annotation classes in proportion to the bases it covers and reported in bits per nucleotide rather than token perplexity, since token units differ across arms. The accounting identity $\sum_c |I_c| \cdot \mathrm{BPN}_c = \text{corpus NLL}$ holds under any tokenization and is checked at every checkpoint, providing a direct check for token–label misalignment. ϕ_c is the share of total loss improvement attributable to class c. It identifies where optimisation gains accrue; whether those gains preserve downstream functional signal is tested in §8.

3.2 EXPERIMENTAL SETUP

In controlled comparisons, an *arm* denotes one experimental condition: arms share architecture, data, and nucleotide budget and differ only in the component under study.

Budgets are matched in nucleotides. Nucleotide exposure, token exposure, and compute cannot all be equalised. Matching token counts would give the 6-mer arm six times more genomic sequence, so we instead match nucleotides and base-pair context. Coarse-token arms therefore use 0.17–$0.33\times$ the single-base compute at the same nucleotide budget; coarse-token advantages are conservative, whereas single-base advantages are partly compute-confounded (Appendix C).

Data and annotation. We use hg38 with GENCODE v44 CDS (Frankish et al., 2023), ENCODE cCRE V3 (ENCODE Project Consortium et al., 2020), and RepeatMasker annotations. Overlaps are

resolved by fixed precedence (`exon > cCRE > repeat`); reversing it changes the CDS-assigned share from 1.33% to 0.96%. The precedence and labels are hashed into each mask fingerprint. Training uses 21 chromosomes (2.75 Gbp), with chr8 and chr9 held out (Appendix D).

Models. We use a HyenaDNA backbone with $21{,}750{,}528$ parameters (`d_model` 384, 12 layers), an $8{,}192$ bp context, and a token-based learning-rate schedule. The schedule is indexed by tokens rather than steps because the dose arms in §8 have different step counts by design.

Tokenizers. The BPE arm uses both DNABERT-2's published vocabulary and its published merge rules. This distinction is material: replacing the merge-rule encoder with a greedy longest-match encoder over the same vocabulary changes compression by only 3% but inflates the measured influence radius by $1.9\times$ (Appendix P).

4 TOKENIZATION: THE COST IS PLACEMENT, NOT FRAGILITY

Byte-pair segmentation is stable under substitution. At ClinVar positions with observed alternate alleles, BPE yields $\bar{R} = 2.97$ bp (95% bootstrap interval $[2.88, 3.05]$), compared with 2.50 when alternate alleles are sampled uniformly (Appendix G). BPE's mean token length over the ClinVar evaluation windows is 5.17 bp, so the mean influence radius is 0.57 times the mean token length. All three fixed-grid arms give $R = 0$ exactly in every stratum.

A secondary result is that pathogenic variants have a smaller radius than benign variants (2.52 vs. 3.03, with non-overlapping intervals), and the difference persists within the CDS-assigned class. This argues against annotation composition as the sole explanation.

Table 1 summarises the tokenizer-level measurements. The fixed-grid controls are invariant to substitution and remain calibrated at the matched boundary-alignment null, whereas BPE is slightly less stable and shows a boundary-placement deviation that the landmark analysis below localises to coding boundaries.

Table 1: Tokenizer results on chr20; BPE uses DNABERT-2 merge rules. $\bar{R}$ uses stratified positions with uniform alternate alleles; ClinVar BPE gives 2.97 bp $[2.88, 3.05]$. Here $\rho = 5.10$ is descriptive, versus the separately calibrated training value 5.21 (Table 6); A_τ pools CDS and cCRE boundaries.

Tokenizer	ρ	S	$\bar{R}$ (bp, unif.)	A_τ
3-mer	2.99	1.0000	0.0	1.0006
6-mer	5.99	1.0000	0.0	1.0014
single-base	1.00	1.0000	0.0	1.0000[†]
BPE (DNABERT-2)	5.10	0.9959	2.5 $[2.4, 2.6]$	0.9378

[†]Every position is a boundary, so $A_\tau = 1$ is saturated for the single-base tokenizer.

Boundaries are depleted at coding-sequence edges. Pooled across all annotation boundaries, BPE gives $A_\tau = 0.938$ ($z = -4.16$), below the matched-null expectation. Stratifying by annotation type shows that the depletion is concentrated at coding-sequence edges (0.636, $z = -9.55$, $n = 2{,}917$) and is absent at cCRE edges (0.994, n.s., $n = 15{,}685$). Because cCRE edges comprise 84% of the pooled boundary set, aggregation dilutes a 36% depletion at coding-sequence edges to about 6% overall.

The annotation types also differ in positional precision: splice sites and stop codons are exact genomic landmarks, whereas cCRE edges are computationally defined element boundaries.

The coding figure is itself two effects with opposite signs. "CDS edge" pools four landmarks that a frequency-driven merge process need not treat similarly. Stratifying them (Figure 2) shows that the pooled value of 0.636 combines opposing effects: BPE boundaries are depleted at splice sites (0.487 over donors and acceptors, $z = -15.0$; 0.444 at acceptors alone) but *enriched* at stop codons (1.802, $z = +10.0$). The fixed-grid controls remain within 0.05 of 1.0 at every landmark, providing a calibrated reference where the BPE effects are largest.

A simple sequence-composition account is consistent with both directions. Stop codons are three high-frequency triplets, so frequency-driven merges can align with them; by contrast, GT and AG are common dinucleotides whose splice identity depends on positional context that the merge objective does not encode. The pattern is therefore consistent with BPE aligning more readily with landmarks whose identity is locally compositional than with those defined by positional context. Splice is also the task that survived our search for a downstream axis (§7): both the tokenizer effect here and the later repeat-directed loss effect concern positionally defined splice structure, although we do not establish a shared mechanism.

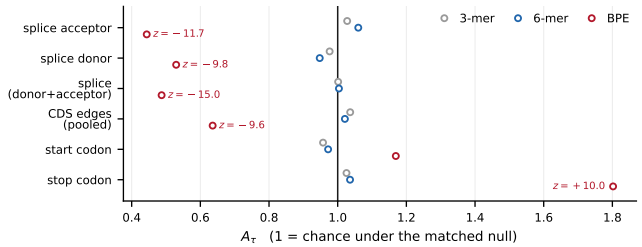


Figure 2: A_τ by landmark type on chr20, $\tau = 0$, with 2,000 null draws. Values below 1 indicate depletion relative to the matched null. z is annotated for BPE where $p_{\text{emp}} \leq 0.05$. The fixed-grid tokenizers remain close to the matched-null expectation, providing a calibrated reference for the BPE departures.

Replication. Because boundary enrichment could reflect regional sequence composition as well as tokenizer behaviour, we repeated the measurement in eleven independent 20 Mb windows spanning six chromosomes. BPE is below 1.0 and significant in $11/11$ windows, whereas only one of twenty-two grid-control tests is significant, consistent with null calibration (Appendix G).

A compositional explanation predicts the opposite sign. A simple GC-driven confound could produce boundary depletion without any relation to function: GC-rich sequence is more predictable, BPE merges it more aggressively, and longer tokens imply fewer boundaries per base. If this mechanism explained the observed depletion, tokens near annotations should be longer. They are shorter, which should increase A_τ; the observed depletion therefore persists despite a compositional tendency acting in the opposite direction (Appendix G).

Two auxiliary observations provide context: token–annotation NMI is low relative to an English reference, while the 3-mer grid shows a small CDS-edge enrichment consistent with reading-frame alignment (Appendix K).

5 Embeddings: Equivariance Does Not Induce Correspondence

Across the five released checkpoints whose embedding matrices we could read directly, none shows strong reverse-complement correspondence at the embedding-table level (Table 2). Formal statistical interpretation is limited to DNABERT-2 (796 reverse-complement pairs) and NT-50M (4,036 pairs). Their V/null values are 0.846 and 0.992, respectively: NT-50M lies close to the matched-random baseline, while DNABERT-2 shows greater correspondence but remains far from the $V/\text{null} \to 0$ expected under exact embedding correspondence. The character-level models contain only four DNA tokens each and therefore provide descriptive rather than inferential evidence.

The Caduceus pair is particularly informative because it separates data augmentation from architectural equivariance. The released `caduceus-ph` configuration sets `rcps = False` and relies on reverse-complement data augmentation rather than the architectural RCPS path. By contrast, `caduceus-ps` sets `rcps = True` and activates that path. Its embedding table has $V/\text{null} = 1.065$; with only four DNA-token pairs this value is descriptive, but it illustrates that architectural equivariance does not require the underlying embedding rows to correspond. RCPS concatenates a forward embedding with a channel-flipped complement embedding, leaving the lookup table itself unconstrained. Embedding-table correspondence and architectural equivariance are therefore distinct properties: enforcing the latter does not imply the former.

Table 2: Reverse-complement violation in five released checkpoints. $V/\text{null} = 1$ indicates no greater correspondence than matched random pairs; values below 1 indicate greater correspondence.

Model	dim	RC pairs	$V(E)$	matched null	V/null
DNABERT-2	768	796	0.5359	0.6336 ± 0.0061	0.846
NT-50M	512	4,036	0.9921	1.0000 ± 0.0007	0.992
HyenaDNA-1k	128	4	1.3044	1.2052 ± 0.0943	1.082
Caduceus-PH (`rcps=False`)	256	4	0.8521	0.8178 ± 0.0453	1.042
Caduceus-PS (`rcps=True`)	256	4	1.1055	1.0382 ± 0.0478	1.065

Merge-based tokenization introduces a separate segmentation-level constraint. Only 20.4% of DNABERT-2's DNA vocabulary is closed under reverse complement, so pairwise embedding tying cannot be applied uniformly across the vocabulary. Over the 20 Mb evaluation span, $V_{\text{seg}} = 0.803$. On this scale, 0 denotes identical mirrored boundary sets, and all three fixed-grid controls score 0.000 on the same span (Appendix G); the BPE value therefore reflects substantial segmentation disagreement. Once the boundaries differ, downstream embedding or representation constraints cannot make the input segmentation itself equivariant. Reverse-complement handling therefore has to be evaluated separately at the segmentation, embedding, and contextual representation levels.

6 Objective: loss improvement follows genome composition

We train four tokenizer arms under the matched budgets of Appendix C and log per-class bits per nucleotide from step 0, because the classwise decomposition cannot be reconstructed from the final loss alone.

Across all four arms, repetitive elements account for 63–83% of total loss improvement, whereas the CDS-assigned class accounts for only 0.3–0.8%; intergenic sequence receives 11–29% and cCREs 5–8% (Table 3). These shares partly reflect class size, but class mass does not explain the pattern by itself. After normalising by the number of nucleotides in each class, repeat BPN still improves by 2.0–$7.5\times$ more than CDS BPN, compared with an 80–$296\times$ gap in aggregate contribution. Because class sizes are identical across arms, the 0.63–0.83 spread in ϕ_{repeat} cannot be attributed to genome composition alone. It is also not aligned with total compute: the single-base arm receives the most compute ($1.00\times$) yet has the lowest repeat share (0.63), whereas the grid arms receive only 0.17–$0.33\times$ compute and have the highest shares (0.81–0.83). This does not isolate tokenizer as the sole cause, but it rules out a simple explanation based on class prevalence or total compute alone. Final CDS BPN is 1.88–1.91, only about 0.1 bit below the roughly 2.0-bit zero-order composition baseline.

Longer-context checkpoints improve most on repeats. If longer context primarily benefited long-range functional dependencies, larger BPN reductions would be expected in annotated functional classes than in repeats. Scoring three released HyenaDNA checkpoints on identical held-out nucleotides yields the opposite descriptive ordering: from 1 kb to 131 kb, repeat BPN decreases by 0.183, compared with 0.042 for cCRE, 0.013 for coding, and 0.009 for intergenic sequence—more than the other three combined. The checkpoints also differ in depth and parameter count, so this comparison does not identify a causal effect of context length (Appendix B). Within these released checkpoints, however, the largest observed reduction is in repetitive sequence. One possible explanation is that a longer window makes additional copies of repetitive elements available; the present comparison cannot separate that mechanism from capacity or memorisation effects.

Why the share varies across arms. Because ϕ_c is a normalised share, its cross-arm ordering need not match absolute BPN reduction (Appendix I). We therefore interpret the robust result within arms: repeats receive the majority of improvement and the CDS-assigned class less than 1% in every tokenizer arm, including the single-base arm where the repeat share is lowest at 0.63.

Table 3: Share of total loss improvement by annotation class, ϕ_c, after 1.0×10^{10} nucleotides. Repetitive elements account for the majority of improvement in every tokenizer arm, whereas the CDS-assigned class contributes well under one percent. Class size is included to distinguish corpus prevalence from the allocation of improvement.

Class	$\lvert I_c \rvert$ (Gb)	ϕ (3-mer)	ϕ (6-mer)	ϕ (BPE)	ϕ (single-base)
repeat	1.343	**0.8065**	**0.8291**	**0.6837**	**0.6274**
intergenic	1.103	0.1330	0.1144	0.2317	0.2891
cCRE	0.267	0.0569	0.0536	0.0767	0.0757
CDS-assigned	0.034	**0.0036**	**0.0028**	**0.0079**	**0.0078**

7 What can serve as a downstream axis?

Two candidate downstream evaluations do not separate the trained representations from simple baselines. On held-out ClinVar variants, zero-shot likelihood scoring (Landrum et al., 2018; Benegas et al., 2023) gives AUROC 0.4657–0.4883 across the four trained tokenizer arms, with every confidence interval containing 0.5. A zero-order composition baseline scores 0.4676 and lies inside the confidence interval of every trained arm. All five point estimates—the four trained arms and the zero-order baseline—are below 0.5. Because the trained-arm intervals include 0.5, we treat this common direction descriptively; importantly, the same direction is already present in a scorer with no trainable parameters. A replication on an independent variant set with a different architecture shows the same limitation (Appendix E).

On GUE linear probes, a 4-mer count baseline scores higher than the trained representation on two of three tasks (0.546 vs 0.510; 0.611 vs 0.561). Splice is the only task that survives this composition control, but under mean pooling a probe on a *randomly initialised* backbone reaches 0.7014, compared with 0.4334 after pretraining. Within this mean-pooled readout, pretraining therefore leaves less linearly decodable splice signal than random initialisation. Together with the repeat-dominated loss allocation in §6, this gap motivates the intervention in §8: we ask whether reducing repeat-directed loss recovers part of the signal relative to the random-initialisation reference. Thus neither zero-shot variant scoring nor an uncalibrated linear-probe score, by itself, provides a defensible axis for the intervention. Appendix E gives the independent replication, shows robustness to scoring span and evaluation context, and provides evidence against CpG hypermutability as the sole explanation (Cooper et al., 2005).

The estimand we can defend. Within the dose series, tokenizer, architecture, context and pooling are fixed. We therefore normalise each trained splice probe between a composition baseline and a shared random-initialisation reference:

$$\text{retained} = \frac{\text{probe}_{\text{trained}} - \text{probe}_{\text{composition}}}{\text{probe}_{\text{random-init}} - \text{probe}_{\text{composition}}}. \tag{1}$$

On this scale, $\text{retained} = 0$ matches the composition baseline and $\text{retained} = 1$ matches the random-initialisation reference. All probe scores are Matthews correlation coefficients; the corresponding references are 0.0695 for composition and 0.6785 ± 0.0232 for random initialisation over five backbone seeds.

This quantity is not a general measure of biological information retained by the network. It measures linearly decodable splice signal in a *mean-pooled* representation. Because mean pooling discards position, the high random-init reference makes the estimand pooling-dependent; a per-position probe could change both the reference and the measured dose effect, which we have not tested (Appendix E). We therefore scope §8 to this mean-pooled estimand. Splice probing is the measurement axis for the intervention, not a claim that it is the preferred biological endpoint.

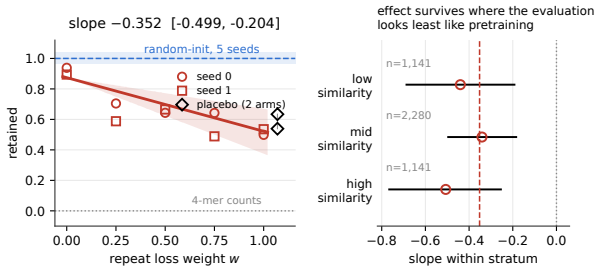


Figure 3: **Left:** retained splice signal against repeat loss weight for both seeds; the band shows the fitted trend and 95% uncertainty, with random-init and 4-mer references. Placebos use $w = 1$ with the $w = 0$ batch and step configuration. **Right:** slopes within similarity strata; every interval excludes zero.

8 Repeat-directed likelihood reduces positional information recoverable from pooled representations

We directly test whether the repeat-dominated loss allocation in §6 is inert for the mean-pooled splice estimand defined in §7. Holding tokenizer and scale fixed, we vary the repeat-class loss weight and ask whether reducing repeat-directed loss changes retained splice-site signal relative to the shared random-initialisation reference.

Design. An observational association between $\Delta\text{BPN}_{\text{repeat}}$ and downstream signal is not identified across tokenizer arms: tokenizer determines both the representation and the loss allocation, leaving essentially no independent variation after tokenizer and scale are accounted for (Appendix J). We therefore fix tokenizer and scale and manipulate the repeat-class loss weight directly at $w \in \{1, .75, .5, .25, 0\}$; $w = 0$ sets the repeat-class contribution to the training loss to zero.

Three controls keep this intervention interpretable as a change in relative loss allocation. The weighted loss is normalised by realised weight mass so that changing w does not mechanically change the overall gradient scale. Batch size is increased from 32 to 63 across the series to hold Kish effective sample size approximately constant. Because this adjustment changes the number of updates at fixed nucleotide exposure, the learning-rate schedule is indexed by tokens rather than steps (Appendix J). Two $w = 1$ placebo runs use the $w = 0$ arm's batch and step configuration to assess whether those implementation changes account for the observed dose pattern. The dose series uses two seeds per dose. This limited replication constrains precision, so we report a small-sample t interval and an equivalence-power sensitivity below rather than treating the blocked fit as a high-precision estimate.

Result. The fitted association between repeat weight and retained signal is negative in both seeds (Figure 3). Across the ten dose runs, a blocked linear fit with seed as a fixed effect gives a slope of -0.352 per unit w, with a normal-approximation 95% CI of $[-0.499, -0.204]$ on 7 residual degrees of freedom; the seed-specific slopes are -0.377 and -0.326. The fitted change across the full $w = 0$ to $w = 1$ range is therefore 0.352 retained units, about 4.2 times the blocked residual standard deviation of 0.084. With only seven residual degrees of freedom, a t interval is wider, $[-0.530, -0.173]$, but still excludes zero and all three candidate equivalence margins. Appendix E discusses how a per-position probe could alter this estimand and, under the stated reference-only alternative, increase the slope magnitude.

The linear slope summarises overall direction rather than a linear dose law: seed 1 rises at $w = 0.5$, and ten runs do not resolve curvature under a quadratic sensitivity analysis (Appendix F).

The conclusion is insensitive to the margin. The normal-approximation blocked interval, $[-0.499, -0.204]$, and the wider t interval, $[-0.530, -0.173]$, both lie outside all three candidate equivalence margins. At the observed residual variation, achieving 0.80 power for the tightest candidate margin would require about thirteen seeds per dose, whereas this study uses two (Appendix O). Within the margins considered, the data therefore favour a non-zero dose association over equivalence.

The placebo argues against a batch-only explanation. Matching effective sample size raises the batch size from 32 at $w = 1$ to 63 at $w = 0$, so batch and step changes are potential alternative explanations. The two $w = 1$ placebo runs use the $w = 0$ batch and step configuration and retain 0.634 and 0.539. Their mean (0.587) is much closer to the ordinary $w = 1$ mean (0.517) than to the $w = 0$ mean (0.916), although the two placebo runs themselves differ by 0.095. This control therefore argues against a batch-only explanation while leaving run-to-run noise visible.

The dose association persists after similarity stratification. If evaluation–training similarity were driving the result, the association should weaken in the least pretraining-like sequences. Re-estimating the slope within strata of 5-mer profile similarity to pretraining windows (Figure 3, right) gives intervals excluding zero in every stratum; the least-similar quartile has slope -0.440 with 95% CI $[-0.691, -0.188]$, and the slope ordering is not monotone in similarity. Seed 1 reproduces the negative slopes across strata (Appendix F). The association therefore remains present in the least-similar stratum tested. Its level does shift there: the $w = 0$ arm reaches $\mathrm{retained} = 1.18$, meaning that its probe exceeds the shared random-initialisation reference on that subset. All of these sequences remain human genomic sequence, so this analysis does not establish out-of-distribution generality.

Most separation is already present at the earliest common point ($\sim 26\%$ of each schedule) in both seeds, with no systematic accumulation thereafter (Appendix F).

The observed dose response is inconsistent with repeat-directed loss being inert for this estimand. Its scope is one probe, one task, two seeds, and a mean-pooled representation referenced to random features: within these bounds, changing repeat-class loss allocation changes retained splice signal under intervention, not biological information in general, a task-universal effect, or out-of-distribution generality.

9 Discussion

The mismatches are distinct, but each requires genomic calibration. At 21.75M parameters and 3.6 corpus passes, claims are relative and non-dose arms are single-seed. The results do not imply one mechanism: BPE is locally stable yet places boundaries selectively, embedding correspondence is not implied by architectural equivariance, and loss improvement concentrates in repeats; the dose intervention links that allocation to the mean-pooled splice estimand in §7.

Implications for evaluation and design. Tokenizer evaluation should report boundary placement alongside compression; RC handling should be audited at segmentation, embedding, and contextual levels; and downstream evaluations should include composition and random-initialisation baselines. Objective allocation should likewise be reported by annotation class. Across the four arms, ϕ_{repeat} spans 0.63–0.83 despite identical class prevalence and does not track total compute: the single-base arm receives the most compute ($1.00\times$) yet has the lowest repeat share (0.63). The intervention in §8 shows that this allocation is not inert for the mean-pooled splice estimand.

Scope and open questions. The controlled training experiments use one 21.75M-parameter backbone; the non-dose arms are single-seed and the dose series uses two seeds per dose. The released long-context checkpoints differ in capacity as well as context, so that comparison is descriptive rather than a causal estimate of context length. The intervention is evaluated on one splice task with mean pooling and entirely within human genomic sequence, and we do not provide a layerwise audit of V_{ctx}. These limits leave several questions open: which repeat families drive the allocation, whether the dose effect transfers to other tasks, architectures, pooling schemes, or species, and whether deliberately increasing A_τ improves downstream representations. The present results therefore motivate genomic calibration of inherited design choices rather than a universal ranking of tokenizers, embedding schemes, or objectives.

Use of large language models. An LLM assistant supported prose editing, analysis code, and methodological discussion, including experimental-prediction framing and the identification argument in §8. All results were author-generated and inspected; independently detected LLM-assisted failures are documented in Appendix P. **Reproducibility.** The supplement records annotation precedence, mask fingerprint, dose schedule, equivalence margins, and released code with 107 tests.

REFERENCES

Gonzalo Benegas, Sanjit Singh Batra, and Yun S Song. DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120 (44):e2311219120, 2023.

Gregory M. Cooper, Eric A. Stone, George Asimenos, NISC Comparative Sequencing Program, Eric D. Green, Serafim Batzoglou, and Arend Sidow. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, 15(7):901–913, 2005. doi: 10.1101/gr.3577405.

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. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023. doi: 10.1101/2023.01.11.523679.

ENCODE Project Consortium, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A. Davis, Alexander Dobin, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020. doi: 10.1038/s41586-020-2493-4.

Adam Frankish, Sílvia Carbonell-Sala, Mark Diekhans, Irwin Jungreis, Jane E. Loveland, Jonathan M. Mudge, Cristina Sisu, James C. Wright, Carme Arnan, If Barnes, et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Research*, 51(D1):D942–D949, 2023. doi: 10.1093/nar/gkac1071.

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

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.

Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J. Bradley Holmes, Brandi L. Kattman, and Donna R. Maglott. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 2018. doi: 10.1093/nar/gkx1153.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin W. Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton M. Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Christopher Ré. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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*, 386(6723), 2024. doi: 10.1126/science.ado9336.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *International Conference on Machine Learning (ICML)*, 2024.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1715–1725, 2016. doi: 10.18653/v1/P16-1162.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. In *International Conference on Learning Representations (ICLR)*, 2024.

A Design axes across genomic language models

Table 4: What changes between adjacent genomic language models. At least three design axes move at every step, limiting clean causal attribution from cross-model comparisons.

Model	Architecture	Tokenizer	Data	Objective	Axes
DNABERT	Transformer	overlap k-mer	human	MLM	—
DNABERT-2	Transformer+ALiBi	BPE-4096	135 species	MLM	4
NT	Transformer	non-overlap 6-mer	850 species	MLM	3
HyenaDNA	Hyena	single-base	human	AR	4
Caduceus	Mamba, RC-equiv.	single-base	human	MLM	3
Evo	StripedHyena	single-base	prokaryote+	AR	3
This work	*fixed*	varied, 4 levels	*fixed*	*fixed*	1

B Context gain by annotation class

Table 5: Bits per nucleotide by annotation class for three released HyenaDNA checkpoints, scored on identical held-out nucleotides. Across these checkpoints, the largest BPN reduction is observed for the repeat class.

Checkpoint	context	CDS	cCRE	repeat	intergenic
HyenaDNA-1k	1,024	1.9275	1.8489	1.4585	1.8925
HyenaDNA-32k	32,768	1.9071	1.8003	**1.2614**	1.8814
HyenaDNA-1M†	131,072	1.9142	1.8074	1.2755	1.8840
Δ (1k $\rightarrow$ 1M)		$+0.013$	$+0.042$	$+\mathbf{0.183}$	$+0.009$

†The released `hyenadna-large-1m` checkpoint, scored on a 131,072 bp window rather than its full 1 Mb context, so that all three checkpoints are read on the same held-out nucleotides.

The three checkpoints differ in depth and parameter count as well as context, and were trained on the whole human genome, so chr8 is not unseen. Absolute BPN is therefore optimistic, only differences between context lengths are interpretable, and the disproportionate repeat gain could in part reflect the larger checkpoints' greater capacity to memorise composition rather than context length alone.

C Resolved budgets

Table 6: Nucleotide-matched budgets. Genomic context is 8,192 bp for every arm; token context and compute differ and are reported rather than equalised.

Tokenizer	ρ	context (tok)	steps	total bp	rel. compute
3-mer	3.00	2,732	38,147	1.0×10^{10}	$0.33\times$
6-mer	6.00	1,366	38,147	1.0×10^{10}	$0.17\times$
single-base	1.00	8,192	38,147	1.0×10^{10}	$1.00\times$
BPE	5.21	1,573	38,147	1.0×10^{10}	$0.19\times$

D Data, annotation and verification

Reference hg38; GENCODE v44 CDS features; ENCODE cCRE V3; RepeatMasker, all on hg38 coordinates. Class shares on chr20 under the fixed precedence are 49.3% repeat, 37.1% intergenic, 12.3% cCRE and 1.3% coding. The reference assembly used in these experiments lacks chrY and chrM.

The failure we most wanted to exclude is a mixed genome build: a single GRCh37 file would shift every coordinate while the code terminates normally. We therefore checked the reference base of all 7,246 two-star ClinVar SNVs on chr20 against the FASTA and found no mismatches.

Precedence is a CLI argument, hashed into the mask fingerprint together with the labels, and a sensitivity table over both orders is emitted with every run. The whole-genome masks were also validated against the chr20 fingerprint.

E WHY TWO CANDIDATE DOWNSTREAM AXES FAILED

What a per-position probe could change. The estimand of §8 reads mean-pooled representations. One reference-only alternative is that random projections preserve information in proportion to the positions entering the pooling; under that alternative, a per-position probe would lower the random-init reference. If the trained-arm scores were unchanged, lowering this denominator would increase `retained` and magnify the fitted slope, making -0.352 conservative under this reference-only scenario. How the trained arms themselves would move under a per-position probe is untested, so a measured sign change would fall outside this argument rather than contradict it.

The failure replicates on a different architecture. We rescored an independent set of $2{,}000$ ClinVar variants ($1{,}000$ pathogenic, $1{,}000$ benign; $8{,}192$ bp window) with a separate mixture-of-experts architecture developed alongside this work. The sixteen configurations are not sixteen independent models: eight are architecture or ablation variants including the released checkpoint, and eight are expert-routing controls applied to one of them. Read that way, the eight variants span AUROC 0.5178–0.5241 and the eight routing controls 0.4948–0.5044; *no* configuration has a confidence interval excluding 0.5, the widest being $[0.4676, 0.5486]$. The same limitation therefore appears under a different architecture and variant set.

The two groups differ by 0.0216 in mean AUROC. We treat this as an exploratory routing-control comparison rather than a tested effect: the separation is $0.54\times$ the minimum effect this sample resolves, and the paired comparison was not confirmatory. In Table 7, a composition model with no trainable parameters also lies inside the confidence interval of every trained arm. These two descriptive comparisons are of the same order and both indicate limited discrimination on this evaluation axis; neither provides a precise estimate of learned-model contribution.

At $n = 2{,}000$ that design resolves AUROC differences of about 0.04, so the replication bounds the measure's discriminative power rather than proving it exactly zero; the balanced $1{,}000/1{,}000$ split is more efficient for AUROC estimation than the $3{,}655/22{,}670$ imbalance of the primary set, which is why a smaller sample was used. The architecture is not named here for anonymity and will be identified in the camera-ready version. The replication covers the evaluation axis, not the intervention: whether the dose slope holds at scale is untested.

One feature of that replication is consistent with the compositional account in Table 7. Scored separately, non-coding variants give higher AUROC than coding ones in all sixteen configurations, and in $8/8$ of the architecture variants considered alone (0.4964–0.5635 against 0.4853–0.5096 over the full set). In our own arms, all point estimates are below 0.5 (0.4657–0.4883), while the replication is near 0.5 overall; because the trained-arm intervals include 0.5, we treat this common direction descriptively rather than as evidence of a systematic below-chance effect. The below-chance direction is more apparent in the coding stratum than overall. This stratified pattern is consistent with, but does not establish, the compositional account developed below.

ClinVar. AUROC changes little across scoring spans (0.4869–0.4940 from ± 0 to ± 256 bp and over the whole window) and across evaluation context lengths (0.431–0.437 from 1 to 4 kb). On the label set, measured without any model, CpG transitions are enriched $1.09\times$ against $1.28\times$ for transversions, which does not support CpG hypermutability as the sole explanation (Cooper et al., 2005). A compositional account is consistent with the remaining pattern: pathogenic variants occur at C and G reference bases in 41.5% and 40.2% of cases against 38.3% and 38.4% for benign, in locally GC-poor context. A zero-shot ClinVar AUROC reported without a composition baseline therefore should not by itself be read as evidence of learned-model contribution, since a parameter-free composition model occupies the same range.

GUE probes. On promoter and transcription-factor prediction, the 4-mer count baseline scores higher than the trained representation. Splice-site prediction is the only task that exceeds this composition control, but under the single-base tokenizer a randomly initialised backbone scores higher than the pretrained representation under mean pooling. The tokenizer changes the number of tokens per 400 bp window (400 for single-base versus 78 for BPE), and the 6-mer pilot has a much lower random-init baseline (0.171). This dependence is why we use splice probing only through the calibrated estimand defined in §7, rather than treating the raw probe score as a general representation-quality measure.

Table 7: Zero-shot ClinVar AUROC on held-out chr8 and chr9 (26,325 variants, 3,655 pathogenic). The non-neural composition baseline lies inside the confidence interval of every trained 21.8M-parameter arm, and all point estimates are below 0.5. The Markov scorers have no trainable parameters; their transition probabilities are estimated from the reference, not fitted to the labels.

Scorer	trainable params	AUROC	95% CI
Markov order 0 (composition)	0	0.4676	[0.4484, 0.4854]
Markov order 2	0	0.4860	[0.4656, 0.5058]
trained, 4 tokenizer arms	21.8 M	0.4657–0.4883	all containing 0.5

F DOSE EFFECT OVER TRAINING

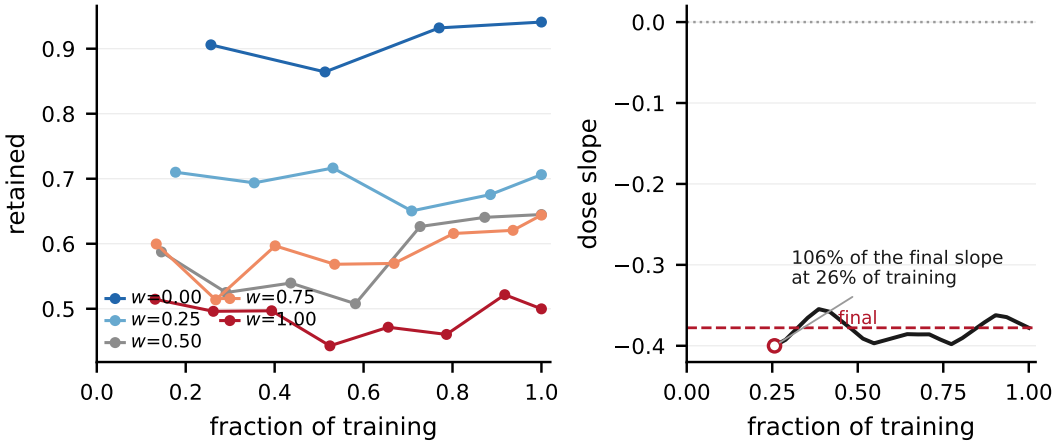


Figure 4: **Left:** retained signal over training for the five seed-0 dose arms. **Right:** the dose slope refitted at each point on a shared fraction grid. Most of the final separation is already present at the earliest point where all five arms can be compared, and the subsequent trajectory shows no systematic accumulation of the dose effect.

Seed 1 reproduces the similarity stratification as well as the trajectory: stratum slopes -0.400, -0.304 and -0.421 with the least-similar quartile at $[-0.731, -0.069]$, so the shallow middle stratum is a property of the stratification rather than one seed's noise.

At the earliest common point the dose slope is -0.400 against a seed-0 final of -0.378; the second seed gives -0.358 against a final -0.326.

A quadratic sensitivity term is $+0.396$ (95% CI $[-0.042, +0.834]$, $F(1, 6) = 3.14$, $p = 0.13$), with AIC favouring the quadratic fit by 2.2. With ten dose runs these statistics do not distinguish a line from a curve. Because w is an objective-allocation weight, curvature in w would not by itself identify nonlinearity in an underlying biological mechanism.

Progress is a fraction of each arm's own schedule rather than a step count: the arms have different step counts by design (38,147 at $w = 1$, 19,481 at $w = 0$), so a comparison at a common step would report the schedule instead of the dose. Arms are linearly interpolated onto a shared fraction grid, which on a curve this flat does no meaningful work.

G TOKENIZER DIAGNOSTICS ON CHR20

Table 8: Tokenizer diagnostics on chr20, DNABERT-2 merge rules, 20 Mb alignment span. S saturates at 1 for every grid-based arm; R carries the comparison. R **here is measured at positions sampled stratified by annotation class with a uniformly drawn alternate allele**; the ClinVar-conditioned value in §4, which uses real variant positions and real alternate alleles, is 2.97 bp. A_τ here pools two annotation types and understates the effect by more than half, see §4. ρ here is measured over the whole of chr20 and is descriptive. Replication: across eleven independent 20 Mb windows on six chromosomes BPE averages 0.9379 (sd 0.0190), significant in $11/11$, against grid means of 1.0009 and 1.0076 with one of twenty-two tests significant. Composition control: tokens average 4.995 bp at CDS edges against 5.152 at random positions, so they are shorter where the depletion is, not longer. The budget table (Table 6) quotes 5.21, computed on the separate budget-calibration span; the schedule was derived from this value, which is the reference value used throughout.

Tokenizer	ρ	S	$\bar{R}$ (bp, unif.)	A_τ	RC closure	$V_{\text{seg}}^{\ddagger}$
3-mer	2.99	1.0000	0.0	1.0006	1.000	0.000
6-mer	5.99	1.0000	0.0	1.0014	1.000	0.000
single-base	1.00	1.0000	0.0	1.0000†	1.000	0.000
BPE (DNABERT-2)	5.10	0.9959	2.5 [2.4, 2.6]	0.9378	0.204	0.803

† saturated: every position is a boundary, so $A_\tau = 1$ carries no information for this arm. ‡ measured over the 20 Mb evaluation span; V_{seg} is window-dependent for grid tokenizers, see Appendix N.

H WHERE EACH NUMBER COMES FROM

Two kinds of measurement appear in this paper and they are not comparable, so each appendix states which it reports. *Model-free diagnostics* (S, R, A_τ, $V(E)$, V_{seg}, ρ) are measured on hg38 chr20 unless another chromosome is named, using published vocabularies and released checkpoints. *Trained-arm results* (ϕ_c, BPN, retained, the dose slope) come from the four tokenizer arms and the dose series, trained on 21 chromosomes with chr8 and chr9 held out and evaluated on the held-out pair. A third, smaller set of numbers comes from the *synthetic fixture* used for power and coverage checks; those are labelled where they appear (Appendices L and Q), and their levels do not transfer to real sequence because the fixture's repeats are near-periodic.

I DIAGNOSTIC DEFINITIONS

Let $x \in \Sigma^L$ with $\Sigma = \{A, C, G, T\}$, and let a tokenizer T induce a boundary set $B(x) = \{b_1 < \cdots < b_n = L\}$, where token i spans $x_{b_{i-1}+1:b_i}$. Compression is $\rho = L/|T(x)|$. The two tokenizer diagnostics below operate on $B(x)$ rather than on the tokens themselves.

I.1 SEGMENTATION STABILITY

For a substitution of x_p by a, write $x^{(p\to a)}$. Define

$$\Delta_{\text{seg}}(x, p, a) = \frac{\big|B(x) \triangle B(x^{(p\to a)})\big|}{|B(x)| + |B(x^{(p\to a)})|} \quad \text{(Sørensen–Dice)}, \qquad S(T) = 1 - \mathbb{E}_{x,p,a}\big[\Delta_{\text{seg}}\big], \quad (2)$$

and the *influence radius*

$$R(x, p, a) = \max\big\{\, |q - p| \;:\; q \in B(x) \triangle B(x^{(p\to a)}) \,\big\}, \qquad R = 0 \text{ when the set is empty.} \quad (3)$$

Defining Δ on boundaries rather than on token edit distance matters: substituting a base necessarily changes the *content* of whichever token covers it, so token edit distance has a floor of $1/|T(x)|$ and a single-nucleotide tokenizer would score 0.996 rather than 1. For fixed-grid tokenizers, a substitution cannot move a cut, so $S = 1$ by construction and R carries the magnitude comparison, reported with 95% bootstrap intervals rather than against a threshold. Positions are sampled stratified by annotation class, since uniform sampling would inherit the genome's repeat composition.

I.2 Annotation boundary alignment

For an annotation boundary set $A(x)$ and tolerance τ,

$$\mathrm{hit}_\tau(B,x) = \frac{1}{|A(x)|}\sum_{a\in A(x)} \mathbf{1}\Big[\min_{b\in B}|a-b| \le \tau\Big], \qquad A_\tau(T) = \frac{\mathbb{E}_x[\mathrm{hit}_\tau(B(x),x)]}{\mathbb{E}_{x,B_0}[\mathrm{hit}_\tau(B_0,x)]}. \tag{4}$$

The null B_0 is the whole metric. A raw hit rate rewards having many boundaries, so a single-nucleotide tokenizer scores 1.0 trivially and BPE would score low for reasons unrelated to biology. We use two nulls and select between them automatically. *Length permutation* shuffles the tokenizer's own segment lengths, preserving both their number and their distribution and destroying only their placement. For fixed-width tokenizers this is vacuous — permuting identical lengths reproduces the observed boundaries — so those arms use a *circular phase shift*, which randomises the only degree of freedom a grid has. Under either null, a single-nucleotide tokenizer gives $A_\tau = 1$ exactly, providing a fixed point for the scale.

We report $\tau = 0$: power to detect planted alignment collapses immediately above it, and type-I error under a true null is 0.022–0.033 against a nominal 0.05 at $N = 600$ (Appendices L and M). The reported p is rank-based, since the null is skewed.

I.3 Reverse-complement violation at three levels

Writing $\bar{t}$ for the reverse complement of t and $E(\cdot)$ for the embedding table,

$$V(E) = \mathbb{E}_t\big[1-\cos(E(t),E(\bar{t}))\big], \qquad V_{\mathrm{ctx}} = \mathbb{E}_x\big[1-\cos(h(x),h(\bar{x}))\big], \tag{5}$$

$$V_{\mathrm{seg}} = \mathbb{E}_x\Big[\big|\mathrm{mirror}(B(x)) \triangle B(\bar{x})\big| \big/ \big(|B(x)| + |B(\bar{x})|\big)\Big], \tag{6}$$

Here $h(x)$ is the mean-pooled final-layer representation, and $\mathrm{mirror}(B) = \{L - b : b \in B \setminus \{L\}\} \cup \{L\}$, which keeps the terminal boundary rather than mapping it to 0.

These levels are measurable separately and are not interchangeable. An asymmetric table can be tied; an asymmetric model can be constrained architecturally or by a consistency loss; but a tokenizer with $V_{\mathrm{seg}} > 0$ *cuts the two strands in different places*, and no downstream constraint can restore a correspondence that the input representation does not have.

$V(E)$ requires a matched null. Cosine distance between arbitrary rows of a trained embedding table is not 1, so "$V(E)$ is far from zero" says nothing on its own; we compare against the same statistic computed on randomly paired rows of the same matrix.

V_{seg} must be reported as a function of window length: a fixed-width tokenizer anchors its grid at the left edge, so a 6-mer scores 0.99 on a 512 bp window and exactly 0 on 510 bp. Quoting one window length would report a padding artifact as a property (Appendix N).

I.4 Objective decomposition

Assign each nucleotide i a class $c(i)$ and attribute each token's negative log-likelihood to classes in proportion to the bases it covers. Reporting is in bits per nucleotide, never token perplexity, since token units differ across arms; $\sum_c |I_c| \cdot \mathrm{BPN}_c$ equals the corpus NLL under any tokenization, an identity checked at every checkpoint, providing a direct check for token–label misalignment. Then

$$\phi_c = \frac{|I_c| \cdot \Delta\mathrm{BPN}_c}{\sum_{c'} |I_{c'}| \cdot \Delta\mathrm{BPN}_{c'}}, \qquad \sum_c \phi_c = 1. \tag{7}$$

ϕ identifies where optimisation gains accrue. Its cross-arm ordering need not match absolute improvement: repeat BPN decreases by 1.78 bits for single-base, 1.19 for BPE, and 0.93–0.98 for the grid arms, while the single-base arm begins at higher BPN across all classes (2.76–2.83 versus 2.02–2.07). Whether these gains preserve downstream functional signal is a separate question, addressed in §8.

J DOSE-SERIES IMPLEMENTATION

Why the observational version is not identified. Regressing $\Delta\mathrm{BPN}_{\text{repeat}}$ on tokenizer (four levels) and scale gives $R^2 = 0.9997$, the coefficient of determination from an OLS regression with tokenizer indicators and scale as predictors. Almost none of its variance is free of the design variables, so a regression of downstream gain on $\Delta\mathrm{BPN}_{\text{repeat}}$ while controlling for tokenizer has essentially no residual variation to work with (variance inflation factor $\sim 3 \times 10^3$), and it removes the signal entirely. That is what the dose intervention is for: it varies the loss allocation while holding the design variables fixed.

The weighted loss is normalised by realised weight mass. Without it, repeats being roughly half the genome, the sum of weights falls from 1.00 to 0.50 across the dose range and gradient magnitude with it, which is arithmetically identical to halving the learning rate on the $w = 0$ arm and perfectly confounded with the dose. Normalisation holds gradient scale at 1.000 and makes the manipulation one of relative allocation: the repeat share of the objective falls from 0.50 to 0 while every other class rises proportionally.

Normalisation does not fix gradient noise. Kish effective sample size falls to 0.50 at $w = 0$, a $1.42\times$ increase in gradient standard deviation that also tracks the dose. A weight floor does not help — effective sample size is 0.55 at $w = 0.05$ — because unequal weights inflate variance whether or not any are zero. Batch size is therefore scaled by $\mathrm{ESS}(w{=}1)/\mathrm{ESS}(w)$, giving 32, 33, 36, 43, 63 across the series.

That fix creates a third constraint. Batch and steps are two degrees of freedom against three quantities: effective sample size per step, total nucleotides, and optimiser updates. Holding the first two and letting updates vary ($38{,}147 \rightarrow 19{,}481$) requires a token-based learning-rate schedule, since a step-based one would give the arms different schedules. Holding effective sample size and updates instead would give the $w = 0$ arm twice the training data, directly confounding dose with genomic exposure; we therefore hold nucleotide exposure fixed and allow update counts to vary.

K TOKEN IDENTITY AND FUNCTION

Bias-corrected normalised mutual information between token identity and annotation class on chr20, with a shuffle floor subtracted: 0.018–0.043 across the three DNA tokenizers, against 0.236–0.255 for English dependency-treebank text under the same estimator and the same correction. Without the shuffle floor the DNA figures are several times larger and the comparison is meaningless, since the floor is where most of the raw mutual information sits when the label alphabet is small. At CDS edges, the 3-mer arm lies slightly above its matched-null expectation ($A_\tau = 1.036$, $p = 0.020$), consistent with partial alignment of a 3 bp grid to reading frame; one of nine tests at this p value is not sufficient as standalone evidence.

L CHOICE OF τ

Power to detect planted boundary alignment, with 35% of annotations placed on token boundaries. This is a synthetic power experiment, so ρ below is the fixture's compression (6.7 for BPE) rather than the 5.10–5.21 measured on chr20. $\tau = 0$ is optimal for every arm and the loss above it is immediate. For grid arms the saturation flag fires; for BPE it does not, yet power still halves by $\tau = 2$, so saturation does not capture the whole loss.

Tokenizer	ρ	τ=0	τ=1	τ=2	τ=5	τ=12
3-mer	3.0	**0.49**	0.00	0.00	0.00	0.00
6-mer	6.0	**0.81**	0.14	0.00	0.00	0.00
BPE	6.7	**1.00**	0.76	0.48	0.45	0.38

M NULL CALIBRATION

Annotation boundaries drawn independently of the tokenizer on chr20, so true enrichment is 1.0 by construction; the question is how often the test rejects. At $N = 600$ (standard error 0.9 points on a 5%

rate): BPE with length permutation 0.032 $[0.018, 0.046]$; BPE with phase shift 0.033 $[0.019, 0.048]$; 6-mer with phase shift 0.022 $[0.010, 0.033]$. The calibration is slightly conservative. An earlier reading of 0.083 at $N = 120$ was Monte-Carlo noise — the standard error at that N is 2.0 points.

N V_{seg} PHASE CURVE

V_{seg} measured on chr20 against window length. The k-mer rows return to zero at every multiple of k; BPE has no zero phase at any tested width. These are short windows, so the BPE values sit below the 0.803 reported in §5 for a 20 Mb span: the longer the span, the more opportunities for the two strands to be cut differently. The phase structure, not the level, is what this table is for.

Window (bp)	510	511	512	513	514	515	516
3-mer	**0.000**	0.994	0.994	**0.000**	0.994	0.994	**0.000**
6-mer	**0.000**	0.988	0.988	0.988	0.988	0.988	**0.000**
BPE	0.525	0.585	0.520	0.672	0.657	0.546	0.688

O EQUIVALENCE MARGIN

Residual sensitivity and equivalence margins. The equivalence analysis considers three candidate margins, $\delta \in \{0.01, 0.02, 0.03\}$. Using the seed-0 residual standard deviation (0.069) gives an interval of $[-0.548, -0.206]$; using the blocked residual from both seeds (0.084) gives $[-0.499, -0.204]$. Neither interval approaches zero or any of the three candidate margins. Run-to-run variation is substantial: two arms retrained on a different accelerator under an identical seed and schedule differ by 0.068 and 0.085 in retained, supporting the blocked estimate as the more representative residual measure.

The table below reports equivalence power by seeds per dose at residual standard deviation 0.069. At the observed blocked residual of 0.084, reaching 0.80 power at $\delta = 0.01$ would require about thirteen seeds per dose. The value quoted in §8 is the same calculation evaluated at the observed residual. Blocking on seed — reusing the same seeds across doses — narrows the mean interval from 0.0161 to 0.0100 at no additional compute.

seeds/dose	runs	δ=0.01	δ=0.02	δ=0.03
2	10	0.14	**0.90**	1.00
3	15	0.30	0.99	1.00
4	20	0.50	1.00	1.00
6	30	0.79	1.00	1.00

P THE SEVEN FAILURES IN DETAIL

These experiments also exposed a broader measurement problem: several plausible implementations fail without raising an error. In the course of making the measurements we found seven of the same kind in our own instruments (Table 9). We report them because the failure modes are not specific to us, and because the first deserves stating on its own: a published vocabulary is not a published tokenizer. Loading DNABERT-2's vocabulary while segmenting with our own greedy longest-match encoder changed compression by 3% — and compression is the statistic most often reported in tokenizer comparisons — while every boundary-derived diagnostic, which is all of them, was measuring our encoder rather than the published one.

The common structure is that each produced plausible numbers. Three were caught by consistency assertions we had written for other reasons — bit conservation, a division guard, a scale assertion — which is an argument for writing more of them than seem necessary.

A published vocabulary is not a published tokenizer. Our BPE arm loaded DNABERT-2's vocabulary but segmented with a greedy longest-match encoder. Zero of 298 windows tokenized identically to the real merge rules; boundary Jaccard was 0.52. Compression differed by 3% — and

Table 9: Seven failures encountered in making these measurements. None raised an error and none appeared on synthetic fixtures; three were caught by consistency assertions written for other purposes. Three — the first, the last, and the budget resolved twice — share one cause: the same quantity computed independently in two places. Each failure is described in full below.

Failure	How it surfaced
Published vocabulary, own greedy encoder	compression off by 3%; boundary Jaccard 0.52, R inflated $1.9\times$
Configured "20M" instantiated 6.5M	parameter-count assertion; curves looked normal
Loss weights normalised per window, not per batch	an all-repeat window zeroed the weight sum and a division guard refused; the dose treatment had been silently deleted
Sampling from the chromosome start	30% N; k-mer vocabulary collapsed to two tokens and reported closure 1.000 correctly by accident
RC closure from observed k-mers	0.605 at 20 kb, 0.931 at 200 kb; true value 1.000 by construction
Append-mode log merged a retried run	min/max step summary: right when the first run finished, wrong when it did not
Budget resolved in two places	empirical BPE compression calibrated on different windows in the budget-calibration and analysis scripts: $\rho = 5.21$ vs 5.20, token context $1{,}573$ vs $1{,}574$; two of the paper's tables disagreed

compression is the statistic most often reported in tokenizer comparisons, so nothing looked wrong. Every other diagnostic is a function of the boundary set, so all of them were measuring our encoder: the influence radius was inflated $1.9\times$ (5.62 vs. 2.97 bp).

A configured model size is not a model size. The configuration labelled "20M" instantiated $6{,}543{,}872$ parameters. That label propagates into every run record, figure axis, and scaling claim.

Per-window loss normalisation erases the treatment. The dose experiment weights the loss by annotation class. Normalising each window to mean weight 1 — rather than the batch — makes a window that is 90% repeat and one that is 10% repeat contribute equally at $w = 0$, deleting exactly the quantity the dose manipulates. The loss magnitude looks normal. It surfaced only because an all-repeat window drove the weight sum to zero and a division guard refused; without that guard, eleven runs would have completed quietly wrong.

Sampling from the start of a chromosome. chr20 opens with 60 kb of telomeric N. Taking the first 200 kb put 30% N in the sample, emptied the reverse-complement window set, and reduced the k-mer vocabulary to two tokens — which then reported RC closure 1.000, correct by accident.

A vocabulary statistic that depends on sample size. k-mer RC closure computed from observed k-mers reads 0.605 at 20 kb, 0.931 at 200 kb, and 1.000 over the full 4^k space. The true value is 1.000 by construction.

An append-mode log merged two runs. A remote command timed out but had in fact executed; the retry produced a second series appended to the first. Summarisation used min and max step, so the corruption was invisible when the first run had completed and wrong when it had not.

The nucleotide budget was resolved twice. BPE compression is empirical — it is higher inside repeats than inside coding sequence — so the schedule is derived from a measurement rather than from the vocabulary size. The budget-calibration script used the first 400 kb of the chromosome; the analysis script used the first 200 kb of the analysis span and resolved the schedule again. The two agreed to within 0.2% ($\rho = 5.21$ vs. 5.20, token context $1{,}573$ vs. $1{,}574$), which is precisely why it survived: at that size the difference reads as a rounding choice rather than as two different numbers for one quantity. It reached the paper, where the budget appendix and the released `protocol.json` disagreed. The appendix table is now generated from the reference configuration, and the analysis script reads the schedule from it rather than measuring again.

This is the third instance in this table of the same structural fault — the first (two encoders over one vocabulary) and the fourth (a vocabulary statistic recomputed from a sample) are the others. Duplicated computation of a single quantity does not announce itself, because both copies are individually plausible; only comparing them reveals anything, and nothing prompts the comparison.

Q BOOTSTRAP INTERVAL COVERAGE

This check runs on the synthetic fixture, not on chr20, so its values are not comparable with the measurements in §4: the fixture's repeats are near-periodic, which inflates re-segmentation. Ground truth from an exhaustive $n = 3000$ evaluation there (BPE, 512 bp window) is $S = 0.8455$ and $R = 92.41$ bp, against $S = 0.9959$ and $\bar{R} = 2.97$ bp on real sequence. What transfers is the coverage, not the level. Coverage of a nominal 95% interval computed from a small sample: at $n = 100$, 0.98 and 0.97 for S and R; at $n = 300$, 0.97 and 0.95. The resulting intervals are slightly conservative at small n.