BIOINVARIANT: EVALUATING APPROPRIATE BIOLOGICAL INVARIANCE AND SENSITIVITY IN GENOMIC FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Genomic foundation models are increasingly used as general-purpose DNA encoders, but they are still evaluated primarily through downstream accuracy. Predictive accuracy alone does not reveal whether representation geometry reflects known biological structure or is dominated by simpler sequence statistics such as GC content. We introduce **BioInvariant**, a counterfactual evaluation framework for frozen genomic representations. BioInvariant tests two complementary properties: appropriate invariance under transformations that preserve biological meaning, and appropriate sensitivity under perturbations that disrupt biological meaning, always relative to composition-matched controls. We evaluate DNABERT-2, Nucleotide Transformer, HyenaDNA, and Caduceus across four axes: reverse-complement consistency, coding consequence, clinical variant signal, and regulatory motif disruption. Raw embedding similarity can be misleading: Caduceus attains the highest raw reverse-complement similarity, but its reverse-complement-specific signal nearly disappears after GC-matched correction; DNABERT-2 and Nucleotide Transformer show stronger corrected specificity. Sensitivity results are also axis-dependent, and no model is consistent across all axes. Additional audits show that context length, GC shortcuts, background perturbations, pooling, and control position substantially affect interpretation; in particular, HyenaDNA's apparent motif sensitivity depends on where the control substitution is placed, and only DNABERT-2's motif effect is significant for both upstream and downstream controls. BioInvariant therefore complements downstream benchmarks with a representation-level criterion: whether frozen genomic encoders move for biologically appropriate reasons under controlled, matched counterfactuals.

1 INTRODUCTION

Genomic foundation models learn representations directly from DNA sequence and are increasingly used as reusable encoders for variant interpretation, regulatory annotation, sequence design, and other genomic tasks. Recent models differ substantially in architecture and inductive bias. DNABERT-2 uses a learned BPE tokenizer rather than fixed k-mers (Zhou et al., 2024); Nucleotide Transformer scales masked language modeling across large genomic corpora (Dalla-Torre et al., 2025); HyenaDNA uses long-convolution sequence modeling to extend context length at single-nucleotide resolution (Nguyen et al., 2023); and Caduceus is designed around reverse-complement symmetry, either through an RC-equivariant architecture (Caduceus-PS) or through RC data augmentation with post-hoc conjoining of predictions (Caduceus-Ph) (Schiff et al., 2024). These design choices suggest different hypotheses about what structure a representation should preserve and what changes should move it.

Most current evaluations answer a different question: whether a model supports downstream prediction. Benchmarks such as BEND and DART-Eval are valuable because they measure practical utility on biologically meaningful tasks (Marin et al., 2024; Patel et al., 2024). However, downstream accuracy does not directly test whether internal representation geometry is consistent with known biological structure. A model may score well by exploiting dataset-specific shortcuts such as GC composition, annotation bias, or variant-class imbalance. Conversely, useful geometric structure may

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.

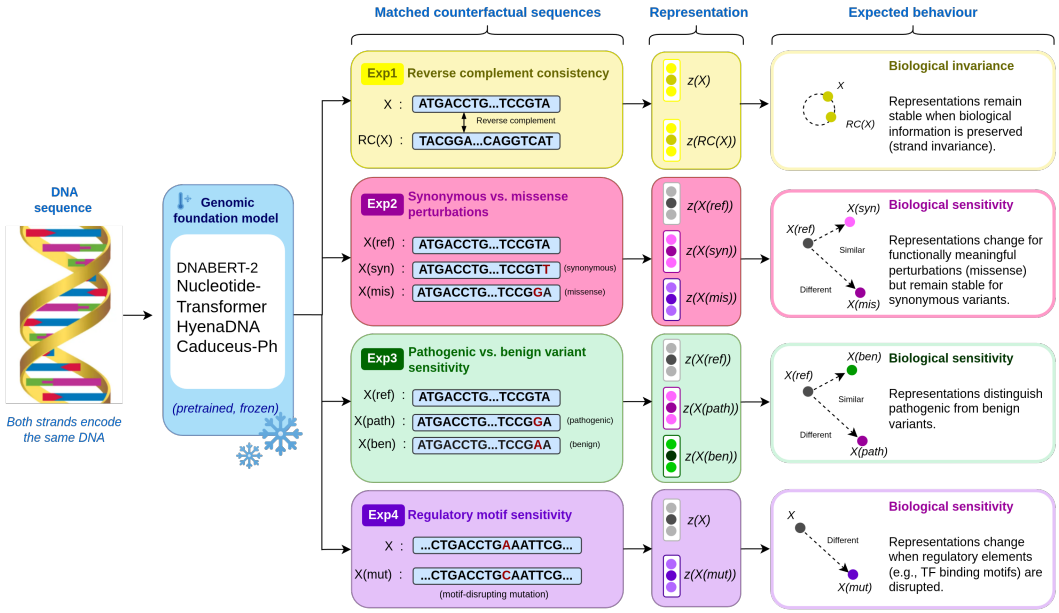

Figure 1: Overview of the BioInvariant evaluation protocol. Frozen genomic encoders are probed with controlled counterfactual sequence pairs. Experiment 1 tests reverse-complement invariance after correcting for composition-matched controls, whereas Experiments 2–4 test sensitivity to coding consequence, clinical variant signal, and regulatory motif disruption.

be obscured by a particular classifier or fine-tuning recipe. This motivates an evaluation protocol that probes frozen representations under controlled counterfactuals.

We introduce **BioInvariant**, a framework for testing whether genomic representations are invariant and sensitive in biologically appropriate ways. If a transformation preserves relevant biological meaning, the representation should remain stable beyond what is expected from generic sequence similarity. If a perturbation disrupts biological structure, the representation should change more than under a matched background perturbation. In both cases, raw cosine similarity or distance is insufficient: measurements must be compared against sequence controls matched for simple composition features.

BioInvariant evaluates four axes. Experiment 1 tests reverse-complement (RC) specificity: a sequence and its reverse complement describe the same double-stranded DNA locus read from opposite strands, so a strand-consistent representation should place them closer than composition-matched non-RC controls. Experiment 2 tests whether missense variants induce larger representation shifts than synonymous variants. Experiment 3 tests whether ClinVar pathogenic or likely pathogenic variants induce larger shifts than benign or likely benign variants. Experiment 4 tests whether disruption of a CTCF motif induces a larger shift than a matched outside-motif substitution. Together these axes cover invariance and sensitivity, in both coding and regulatory contexts.

Our main contribution is a unified counterfactual protocol for evaluating whether frozen genomic representations respond to sequence changes in ways consistent with specified biological hypotheses. Across four axes, we find that the behavior of four widely used encoders depends on both the biological question and the evaluation control. For example, Caduceus has the highest raw reverse-complement similarity, yet little reverse-complement-specific margin after comparison with GC-matched sequences. On the sensitivity axes, Nucleotide Transformer shows the clearest clinical variant signal, DNABERT-2 shows motif-disruption sensitivity with both upstream and downstream controls, and coding-consequence sensitivity is weak or reversed. Audits of context length, GC composition, background substitutions, pooling, and control position show how these choices can change the interpretation of raw similarity or distance. BioInvariant therefore offers a complementary evaluation of representation behavior alongside downstream predictive accuracy (Figure 1).

2 Related Work

Genomic foundation models. We evaluate DNABERT-2, Nucleotide Transformer, HyenaDNA, and Caduceus. These models span learned tokenization, large-scale masked language modeling, long-context sequence modeling, and RC-aware training (Zhou et al., 2024; Dalla-Torre et al., 2025; Nguyen et al., 2023; Schiff et al., 2024). We use the Caduceus-Ph checkpoint, which is trained with RC augmentation and is not RC-equivariant at the representation level. BioInvariant does not propose a new encoder; it evaluates the representation behavior of existing published encoders.

Benchmarks for genomic models. BEND and DART-Eval measure predictive performance on downstream genomic tasks, including regulatory annotation and variant effect prediction (Marin et al., 2024; Patel et al., 2024). These benchmarks are complementary to BioInvariant: they ask whether a model predicts external labels, while BioInvariant asks whether representation geometry reacts appropriately to controlled biological transformations and perturbations.

Zero-shot variant effect prediction. DNA language models are commonly used for zero-shot variant effect prediction through the log-likelihood ratio (LLR) of alternate versus reference alleles, as in GPN (Benegas et al., 2023) and Evo (Nguyen et al., 2024), and they are increasingly fine-tuned for pathogenicity prediction (Lin et al., 2025). BioInvariant measures representation distance rather than LLR. Benchmark studies also report that DNA foundation models rank differently across tasks (Feng et al., 2025), which motivates axis-wise evaluation.

Invariance and consistency tests. Recent work has studied reverse-complement consistency and mechanistic invariance in genomic language models (Ma, 2025; Cheng and Zhang, 2026). BioInvariant differs in two ways. First, it subtracts composition-matched background similarity rather than reporting only raw strand consistency. Second, it applies one statistical framework across multiple biological axes rather than a single transformation or motif family.

3 Experimental Setup

Models and representations. We evaluate frozen checkpoints of DNABERT-2, Nucleotide Transformer, HyenaDNA, and Caduceus. For an input DNA sequence X, we extract final-layer hidden states, remove special tokens, and mean-pool the remaining tokens to obtain $f_m(X)$. All experiments use cosine similarity or cosine distance. Inference runs in float32 without mixed precision, and distances are computed in float64 as $\frac{1}{2}\|\hat{a} - \hat{b}\|^2$ on unit-normalized vectors, which is equivalent to $1 - \cos(\hat{a}, \hat{b})$. Re-embedding the same windows in a different batch gives a maximum distance of 5.7×10^{-10} (DNABERT-2), 6×10^{-14} (Nucleotide Transformer), 10^{-14} (HyenaDNA), and 5×10^{-16} (Caduceus); every VES at $L = 3{,}000$ bp exceeds the 99th percentile of this noise. Raw distances differ by up to four orders of magnitude across models, so models are compared only through AUROC, win rate, Cohen's d, and RBS. Implementation details for tokenizers, checkpoints, filtering, and statistical tests are given in the Supplementary Material.

Data. All sequence windows are extracted from GRCh38. Experiments 1–3 use ClinVar SNVs (Landrum et al., 2020); Experiment 3 keeps only unambiguous pathogenic/likely-pathogenic and benign/likely-benign labels. Experiment 2 uses molecular consequence annotations to compare synonymous and missense variants independently of clinical labels. Experiment 4 uses ENCODE CTCF ChIP-seq peaks and a JASPAR CTCF PWM. We remove ambiguous bases, unwanted overlapping windows, and variants whose reference allele does not match GRCh38.

3.1 Reverse-Complement Invariance

3.1.1 Formal Definition

For a DNA sequence X, its reverse complement $\mathrm{RC}(X)$ represents the same double-stranded locus read from the opposite strand. We first define raw reverse-complement similarity as

$$\mathrm{RCI}_m(X_i) = \cos\left(f_m(X_i), f_m(\mathrm{RC}(X_i))\right). \tag{1}$$

Table 1: Experiment 1 on ClinVar held-out windows (chr17–22), $L^* = 3{,}000$ bp. $n = 500$ windows; $\Delta\mathrm{RC}$ and Retrieval@1 use the 498 windows whose GC bin contains at least one other window. $\Delta\mathrm{RC}$ uses $K = 20$ GC-matched controls; Retrieval@1 ranks $\mathrm{RC}(X_i)$ against all same-bin candidates (mean bin size 33.3, random baseline 3.8%). Raw RCI and corrected $\Delta\mathrm{RC}$ rank models differently.

Model	$\mathrm{RCI}_m \uparrow$ (95% CI)	$\Delta\mathrm{RC}_m \uparrow$ (95% CI)	Retrieval@1
DNABERT-2	0.9701 [0.9683, 0.9717]	$+0.0211$ [0.0195, 0.0228]	30.3%
Nucleotide Transformer	0.9898 [0.9889, 0.9906]	$+0.0094$ [0.0088, 0.0101]	36.9%
HyenaDNA	0.9959 [0.9953, 0.9964]	$+0.0021$ [0.0017, 0.0025]	14.5%
Caduceus	**0.9985** [0.9982, 0.9987]	-0.00059 [-0.00079, -0.00042]	7.4%

Table 2: Experiment 1 robustness on the same held-out windows. Top: Retrieval@1 with a fixed number K of same-bin distractors; K is capped at the bin size minus one (bins hold at most 53 windows), so $K = 50$ and "all" nearly coincide. The random baseline is the mean of $1/(k_{\text{eff}} + 1)$. Bottom: $\Delta\mathrm{RC}$ [95% CI] under alternative controls and representations: controls that are the $K = 20$ nearest same-bin windows in dinucleotide-frequency space; cosine after subtracting the mean embedding (anisotropy check); and mean pooling over only the central 20% of tokens.

	DNABERT-2	NT	HyenaDNA	Caduceus
Retrieval@1, $K = 10$ (random 9.3%)	43.4%	53.6%	25.9%	12.7%
Retrieval@1, $K = 20$ (random 5.4%)	35.3%	44.2%	16.9%	8.6%
Retrieval@1, $K = 50$ (random 3.8%)	30.3%	37.1%	14.5%	7.4%
Retrieval@1, all in bin (random 3.8%)	30.3%	36.9%	14.5%	7.4%
$\Delta\mathrm{RC}$, GC + dinucleotide-matched controls	$+0.0183$ [0.0167, 0.0199]	$+0.0086$ [0.0080, 0.0092]	$+0.0017$ [0.0013, 0.0021]	-0.00057 [-0.00076, -0.00041]
$\Delta\mathrm{RC}$, centered cosine	$+0.278$ [0.259, 0.298]	$+0.316$ [0.296, 0.336]	$+0.220$ [0.195, 0.245]	-0.009 [-0.036, 0.017]
$\Delta\mathrm{RC}$, central-20% pooling	$+0.0362$ [0.0324, 0.0404]	$+0.0168$ [0.0153, 0.0184]	$+0.0050$ [0.0043, 0.0057]	$+0.00046$ [0.00015, 0.00075]

Raw similarity alone is insufficient because reverse complementation preserves simple composition statistics such as GC content. Therefore, for each X_i we sample $K = 20$ controls $Y_{ik} = \mathrm{RC}(X_j)$, $j \neq i$, from held-out windows of the same length and the same 2%-wide GC bin (K is smaller when the bin has fewer than 21 windows), and compute

$$\mathrm{GCS}_m(X_i) = \frac{1}{K}\sum_{k=1}^{K} \cos\left(f_m(X_i), f_m(Y_{ik})\right), \qquad \Delta\mathrm{RC}_m(X_i) = \mathrm{RCI}_m(X_i) - \mathrm{GCS}_m(X_i). \quad (2)$$

A positive $\Delta\mathrm{RC}_m$ means that the true reverse complement is closer than composition-matched non-RC controls. We also report Retrieval@1, where $\mathrm{RC}(X_i)$ must be the nearest of the candidates $\{\mathrm{RC}(X_j)\}$ over all X_j in the same GC bin.

3.1.2 Results

We select $L^* = 3{,}000$ bp on development chromosomes 1–16 and evaluate $n = 500$ windows on held-out chromosomes 17–22. L^* is within the pretraining context of Nucleotide Transformer ($\sim$12 kb), HyenaDNA (160 kb), and Caduceus (131 kb), and longer than the $\sim$700 bp sequences DNABERT-2 was pretrained on (Zhou et al., 2024). Length and sample-size sweeps are in the Supplementary Material.

Table 1 shows that Caduceus has the highest raw RCI, consistent with RC-augmented training, but its corrected signal is close to zero. DNABERT-2 and Nucleotide Transformer have lower raw RCI yet stronger corrected RC specificity. Retrieval@1 gives the same interpretation at every candidate-set size (Table 2): Nucleotide Transformer and DNABERT-2 retrieve the true RC partner much more often than Caduceus among matched candidates. The ranking survives stricter dinucleotide-matched controls, centered cosine, and central pooling. The sign of the small Caduceus $\Delta\mathrm{RC}$ does not: it is slightly negative under whole-window pooling, indistinguishable from zero after centering, and slightly positive under central pooling. We therefore read Caduceus as having essentially no RC-specific cosine margin beyond composition, not as anti-RC; its Retrieval@1 (7.4% vs. 3.8% chance) shows a weak rank-order signal. As $\Delta\mathrm{RC}$ is bounded by $1 - \mathrm{GCS}$, we compare models via Retrieval@1 and centered $\Delta\mathrm{RC}$. Thus, raw RC similarity can overstate biologically specific invariance. Figure 2 illustrates how the model ranking changes after control subtraction.

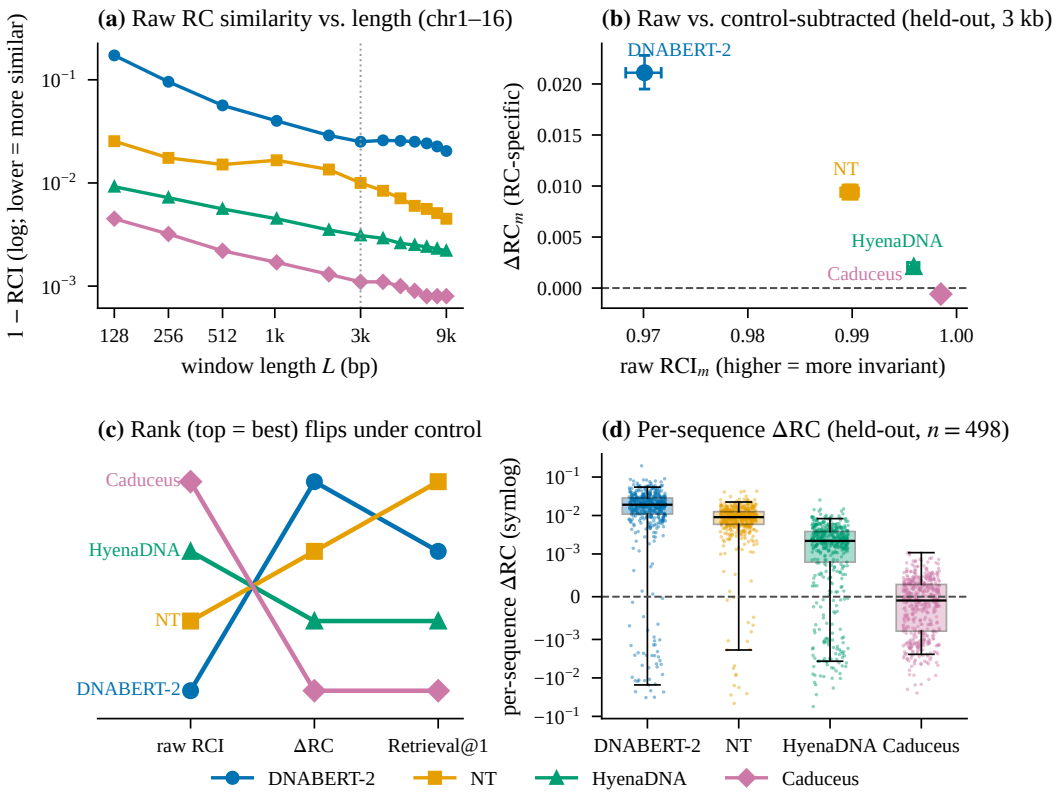

Figure 2: Experiment 1. Raw RCI increases with context length, but corrected ΔRC changes the model ranking on held-out data. Panel (d) shows the $n = 498$ held-out windows with at least one GC-matched control.

3.2 Synonymous versus Missense Representation Sensitivity

3.2.1 Formal Definition

In coding regions, synonymous variants preserve the encoded amino acid whereas missense variants alter it. For each SNV, we construct reference and alternate windows X_{ref} and X_{alt} and define the Variant Effect Score

$$\text{VES}_m = 1 - \cos\left(f_m(X_{\text{ref}}), f_m(X_{\text{alt}})\right). \tag{3}$$

If frozen representations capture coding consequence, missense variants should tend to have larger VES than synonymous variants. We evaluate this with AUROC and with 750 missense–synonymous pairs matched within gene. Throughout the paper, win rates are tested with an exact two-sided sign test, AUROC with a Mann–Whitney test, and paired distance differences with a sign-flip permutation test; confidence intervals are percentile bootstrap intervals (Efron and Tibshirani, 1993). Significance marks (*/†) use $\alpha = 0.05$ after Holm correction across the four models in each table block; raw p-values are shown.

3.2.2 Results

No model shows a robust positive missense-over-synonymous effect (Tables 3–4). DNABERT-2 and HyenaDNA are close to random, with AUROC 0.489 and 0.527. Nucleotide Transformer and Caduceus are below random, with AUROC 0.461 and 0.446. In the gene-matched analysis, only Caduceus remains significant after correction: missense variants move the representation less than synonymous variants in the same gene. The reversal of Nucleotide Transformer and Caduceus is not explained by simple mutational context. It keeps the same direction in every transition/transversion and CpG/non-CpG stratum, and remains significant after correction for Nucleotide Transformer among transitions and for Caduceus at non-CpG sites (stratified table in the Supplementary Material). It also persists, or strengthens, when pooling only a fixed 128 bp window around the variant (AUROC

Table 3: Experiment 2 paired analysis: 750 same-gene missense–synonymous pairs. $\bar{\Delta}_m = \mathrm{VES}_{\mathrm{missense}} - \mathrm{VES}_{\mathrm{synonymous}}$; raw $\bar{\Delta}_m$ is on each model's own scale and is not comparable across models. p_{sign}: exact sign test of the win rate. */†: win rate significantly above/below 0.5 after Holm correction.

L	Model	$\bar{\Delta}_m$ [95% CI]	p_{perm}	Win rate [95% CI]	p_{sign}
3 000 bp	DNABERT-2	6.81×10^{-5} $[-6.94 \times 10^{-8}, 1.44 \times 10^{-4}]$	0.062	0.496 [0.460, 0.532]	0.855
	Nucleotide Transformer	-2.02×10^{-6} $[-5.72 \times 10^{-6}, 9.14 \times 10^{-7}]$	0.272	0.459 [0.423, 0.495]	0.026
	HyenaDNA	1.13×10^{-8} $[-3.96 \times 10^{-8}, 6.29 \times 10^{-8}]$	0.669	0.497 [0.461, 0.533]	0.913
	Caduceus	-2.93×10^{-8} $[-3.67 \times 10^{-8}, -2.20 \times 10^{-8}]$	$< 10^{-4}$	0.364† [0.329, 0.397]	9×10^{-14}

Table 4: Experiment 2 missense-versus-synonymous VES and AUROC ($n = 750$ per class, $L = 3{,}000$ bp). */†: AUROC significantly above/below 0.5 (Mann–Whitney, Holm-corrected).

Model	VES synonymous (mean)	VES missense (mean)	AUROC ↑ (95% CI)
DNABERT-2	2.217×10^{-4}	2.334×10^{-4}	0.489 [0.461, 0.519]
Nucleotide Transformer	8.888×10^{-6}	8.094×10^{-6}	0.461† [0.432, 0.490]
HyenaDNA	6.971×10^{-7}	7.587×10^{-7}	0.527 [0.498, 0.556]
Caduceus	9.746×10^{-8}	8.626×10^{-8}	0.446† [0.418, 0.476]

0.452 and 0.428), whereas the weak HyenaDNA tendency disappears under that pooling (0.500). These results should be read as a test of representation sensitivity under this protocol, not as a claim about true functional severity.

3.3 Clinical Variant Signal

3.3.1 Formal Definition

Using ClinVar labels, Experiment 3 asks whether pathogenic/likely-pathogenic variants induce larger VES than benign/likely-benign variants. This is again a representation-level test rather than a clinical predictor: a positive association means that the embedding geometry is aligned with a curated clinical signal, not that the model has learned a causal mechanism of pathogenicity.

3.3.2 Results

Nucleotide Transformer shows the strongest clinical signal, with AUROC 0.630 and gene-matched win rate 0.602 (Tables 5–6). HyenaDNA has a smaller positive AUROC, DNABERT-2 is near random, and Caduceus is below random. The paired analysis is weaker: only Nucleotide Transformer is significant after correction. When we restrict to pairs that are both same-gene and missense, its win rate decreases from 0.602 to 0.545 but remains above chance. With pooling restricted to 128 bp around the variant (750 per class), the Nucleotide Transformer signal strengthens (AUROC 0.662), HyenaDNA is no longer significant (0.535), and the Caduceus reversal disappears (0.488, n.s.), so the Caduceus result on this axis depends on the pooling. Stratified analyses in the Supplementary Material show that Nucleotide Transformer's signal is strongest for loss-of-function consequences such as nonsense variants, indicating that part of the pathogenic-versus-benign contrast reflects molecular consequence class rather than pathogenicity alone.

3.4 Regulatory Motif Disruption

3.4.1 Formal Definition

For each CTCF motif instance, we create a motif-disrupting sequence X_{motif} by changing the highest-information position of the PWM hit to the base the PWM considers least likely and a matched control X_{control} by applying the same substitution (same original and new base) at a randomly chosen position outside the motif but within the central 128 bp (mean distance to the motif 37 bp). The Biological Sensitivity Index is

$$\mathrm{BSI}_m = [1 - \cos(f_m(X_{\mathrm{ref}}), f_m(X_{\mathrm{motif}}))] - [1 - \cos(f_m(X_{\mathrm{ref}}), f_m(X_{\mathrm{control}}))]. \quad (4)$$

A positive BSI means that the representation reacts more strongly to motif disruption than to a matched background substitution. Because raw BSI is on each model's own distance scale, we

Table 5: Experiment 3 matched analysis at $L = 3{,}000$ bp (1,000 pairs per block). p_{sign}: exact sign test of the win rate. */†: win rate significantly above/below 0.5 after Holm correction within each block.

Pair set	Model	$\overline{\Delta}_m$ [95% CI]	p_{perm}	Win rate [95% CI]	p_{sign}
Same gene	DNABERT-2	-4.35×10^{-6} [-7.03×10^{-5}, 6.17×10^{-5}]	0.899	0.498 [0.467, 0.529]	0.924
	Nucleotide Transformer	6.81×10^{-6} [3.20×10^{-6}, 1.24×10^{-5}]	$< 10^{-4}$	**0.602*** [0.572, 0.632]	1×10^{-10}
	HyenaDNA	3.59×10^{-8} [-5.23×10^{-9}, 7.47×10^{-8}]	0.080	0.514 [0.482, 0.544]	0.393
	Caduceus	-8.22×10^{-9} [-1.46×10^{-8}, -2.02×10^{-9}]	0.009	0.467 [0.435, 0.498]	0.040
Same gene, same missense	DNABERT-2	4.09×10^{-6} [-6.40×10^{-5}, 8.06×10^{-5}]	0.916	0.466 [0.435, 0.498]	0.034
	Nucleotide Transformer	-1.27×10^{-6} [-3.93×10^{-6}, 8.06×10^{-7}]	0.295	**0.545*** [0.515, 0.575]	0.005
	HyenaDNA	3.58×10^{-8} [-3.77×10^{-9}, 7.50×10^{-8}]	0.083	0.538 [0.507, 0.569]	0.018
	Caduceus	-7.71×10^{-9} [-1.43×10^{-8}, -1.33×10^{-9}]	0.014	0.462 [0.431, 0.492]	0.018

Table 6: Experiment 3 pathogenic-versus-benign VES and AUROC at $L = 3{,}000$ bp ($n = 1{,}000$ per class). */†: AUROC significantly above/below 0.5 (Mann–Whitney, Holm-corrected).

L	Model	VES pathogenic (mean)	VES benign (mean)	AUROC [95% CI]
3 000 bp	DNABERT-2	2.026×10^{-4}	2.453×10^{-4}	0.503 [0.476, 0.528]
	Nucleotide Transformer	1.599×10^{-5}	8.119×10^{-6}	**0.630*** [0.606, 0.655]
	HyenaDNA	7.038×10^{-7}	6.421×10^{-7}	0.538* [0.513, 0.563]
	Caduceus	8.191×10^{-8}	9.338×10^{-8}	0.448† [0.423, 0.473]

compare models through paired Cohen's d (Cohen, 1990) and the relative biological sensitivity $\text{RBS}_m = \text{median}\,[D_{\text{motif}}/D_{\text{control}}]$.

3.4.2 Results

DNABERT-2 and HyenaDNA have positive BSI at the default control seed and remain positive across the three control seeds reported in the Supplementary Material (default, `cs0`, and `cs1`). Nucleotide Transformer is null at the default seed but significant at both others (seed-unstable). Control position, however, matters. HyenaDNA is causal, so a substitution earlier in the input reaches more downstream tokens of the mean pool. Its positive BSI comes entirely from anchors whose control lies downstream of the motif ($d = 0.161$); when the control lies upstream, the effect vanishes ($d = 0.004$, $p = 0.87$; Table 7). We therefore do not count HyenaDNA's motif sensitivity as evidence. Nucleotide Transformer is null overall but positive when the control is upstream ($d = 0.062$) and null when it is downstream; we report this asymmetry without a mechanistic interpretation. Caduceus's negative effect comes only from downstream controls. We count motif sensitivity as position-robust only if the effect is positive and significant in *both* halves. Only DNABERT-2 meets this criterion ($d = 0.085$ upstream, 0.060 downstream). All effects are small ($|d| < 0.17$).

4 Audit Analyses for Robustness and Interpretation

We additionally audit five sources of interpretive instability: context length, GC shortcuts, sensitivity to background perturbations, pooling, and control position. Pooling and control position are reported with Experiments 1–4 (Tables 2 and 7); the remaining three are described below. These audits are not used to re-select the main protocol. Two of them change conclusions of the main experiments: control position removes HyenaDNA's motif effect, and variant-window pooling removes Caduceus's clinical reversal.

Context length. When repeating all four axes over multiple context lengths, the best length is not monotonic and depends on the pair of model and biological axis. RC specificity often benefits from longer contexts, whereas sensitivity axes can peak at intermediate lengths. Some main patterns hold across lengths: Caduceus has AUROC below 0.5 on coding consequence at all twelve lengths from 128 bp to 9 kb, and Nucleotide Transformer remains strongest on clinical signal at long context. Other patterns do not: Caduceus's whole-window clinical reversal disappears under 128 bp variant-window pooling, and its negative Experiment 4 effect at 3 kb turns positive beyond 4 kb.

GC shortcut. On the Experiment 4 CTCF anchors we compare two outside-motif perturbations that use the same anchors, positions, and number of substitutions: one that changes GC content by

Table 7: Experiment 4 at $L = 3{,}000$ bp ($n = 4{,}385$ CTCF anchors). p: paired permutation test on D_{motif} vs. D_{control}; */†: significant positive/negative effect after Holm correction. The last two columns split anchors by whether the control substitution is upstream ($n = 2{,}189$) or downstream ($n = 2{,}196$) of the motif in the model input.

Model	Mean BSI (95% CI)	Cohen's d (p)	RBS	d, control upstream (p)	d, control downstream (p)
DNABERT-2	7.10×10^{-5} [4.27×10^{-5}, 1.01×10^{-4}]	**0.073*** ($< 10^{-4}$)	1.60	**0.085*** ($< 10^{-4}$)	**0.060*** (0.004)
Nucleotide Transformer	5.02×10^{-7} [-3.49×10^{-6}, 3.00×10^{-6}]	0.004 (0.991)	1.14	**0.062*** (2×10^{-4})	-0.009 (1.00)
HyenaDNA	7.60×10^{-8} [4.98×10^{-8}, 1.03×10^{-7}]	0.083* ($< 10^{-4}$)	1.09	0.004 (0.865)	**0.161*** ($< 10^{-4}$)
Caduceus	-3.24×10^{-9} [-5.68×10^{-9}, -8.33×10^{-10}]	$-0.040^\dagger$ (0.008)	0.97	-0.009 (0.677)	$-0.066^\dagger$ (0.001)

Table 8: GC shortcut audit on CTCF anchors: geometric-mean ratio D_{gc}/D_{pres} of the embedding distance under a GC-changing versus a GC-preserving outside-motif perturbation with the same positions and number of substitutions. Values above 1 indicate sensitivity to GC content itself. k-mer rows are normalized k-mer count vectors.

	256 bp		3 000 bp	
Model	$\Delta GC = 2\%$	$\Delta GC = 5\%$	$\Delta GC = 2\%$	$\Delta GC = 5\%$
DNABERT-2	0.99	1.08	1.23	1.50
Nucleotide Transformer	1.09	1.20	1.50	1.82
HyenaDNA	1.94	3.41	3.95	6.21
Caduceus	5.18	11.4	17.2	23.6
k-mer, $k = 3$	1.17	1.44	2.50	3.81
k-mer, $k = 6$	1.01	1.01	1.05	1.11

ΔGC (X_{gc}) and one that preserves it (X_{pres}). Table 8 shows that Caduceus is the most GC-sensitive model: GC-changing perturbations move its embedding 5–24 times more than GC-preserving ones, more than a trinucleotide-count baseline. This is consistent with Caduceus having very high raw RCI but near-zero ΔRC once controls are GC-matched. DNABERT-2 is the least GC-sensitive.

Background sensitivity. On the CTCF anchors we also substitute $m \in \{1, 4, 16\}$ outside-footprint positions and compare the resulting distance with the single motif disruption. The background distance grows roughly linearly with m: four background substitutions usually move the representation more than one motif disruption, and sixteen move it several times more. The position analysis in Table 7 adds a second caveat: for three of the four models, *where* the control falls relative to the motif can create or remove an apparent effect. Raw sensitivity is therefore not enough; it must be compared against a background control matched in number, type, and position of substitutions.

5 Discussion

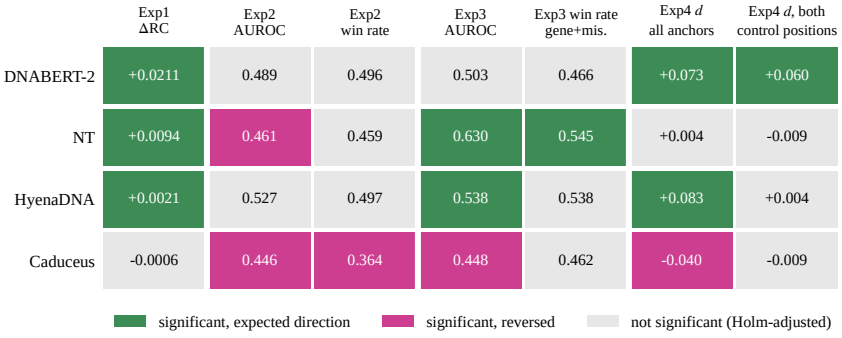

Figure 3: Summary across the four BioInvariant axes at $L = 3{,}000$ bp. A cell is colored when its test is significant at $\alpha = 0.05$ after Holm correction across the four models in that column (AUROC: Mann–Whitney; win rate: exact sign test; Exp4: paired permutation test). Exp1 is colored only if ΔRC has the same significant sign under all four variants in Table 2 (GC-matched, dinucleotide-matched, centered cosine, central pooling). The "all anchors" Exp4 column pools both control positions and is superseded by the last column, which is colored only if the Exp4 effect is significant in the same direction for both upstream and downstream controls; it shows the smaller of the two d values (Table 7). Green: expected direction; pink: opposite direction; gray: not significant.

From prediction accuracy to representation geometry. BioInvariant asks a complementary question to standard predictive benchmarks: beyond whether a model predicts downstream labels, it tests whether representation geometry is stable or sensitive in ways consistent with known biological structure Marin et al. (2024); Patel et al. (2024). The central result is that raw, uncorrected metrics are unreliable for model ranking Cheng and Zhang (2026). Caduceus has the highest raw RCI but the weakest corrected RC specificity and the lowest Retrieval@1 among the four models.

No single scalar "biological quality". Model rankings change across biological axes. Nucleotide Transformer has the strongest clinical signal and high RC retrieval, but it is below random on coding consequence and its motif effect depends on control position. DNABERT-2 has strong corrected RC specificity (second to Nucleotide Transformer on Retrieval@1 and centered $\Delta\mathrm{RC}$) and the only motif effect that is significant for both upstream and downstream controls, but is near null on coding and clinical axes at 3 kb. HyenaDNA shows a small positive clinical AUROC that is not significant under variant-window pooling; its apparent motif sensitivity is explained by control position under causal pooling. Caduceus often shows weak or opposite-direction behavior under the sensitivity designs, but several of its reversals are small and pooling- or length-dependent Ma (2025). The results summarized in Figure 3 fall into four behavior types: *Type I (desired invariance)*, RC consistency beyond composition, holds for DNABERT-2, Nucleotide Transformer, and HyenaDNA, whereas Caduceus's high raw RCI is largely explained by composition. *Type II (desired sensitivity)* appears for Nucleotide Transformer on clinical signal and for DNABERT-2 on motif disruption. *Type III (false sensitivity)* appears as Caduceus's strong response to GC-changing background edits and HyenaDNA's position-driven motif effect. *Type IV (false invariance)* appears on coding consequence, where no model moves more for missense than for synonymous variants, and for DNABERT-2 on clinical signal at 3 kb. BioInvariant should therefore be read as an axis-wise behavioral map, not a single leaderboard.

Limitations. The experiments measure associations between representation geometry and known biological structures; they do not prove internal causal mechanisms. Some Caduceus and HyenaDNA effects are small; they exceed the numerical noise floor, but their sign or significance depends on pooling or control position. Experiments 2–4 use $L = 3{,}000$ bp as the main operating point even though some axes peak at other lengths, and this length requires extrapolation for DNABERT-2. At $L \approx 1$ kb DNABERT-2 shows much stronger clinical and motif effects (Supplementary Material), so its near-null 3 kb results are not model-intrinsic. Whole-window mean pooling dilutes a single-SNV signal by roughly $1/L$; 128 bp variant-window pooling changes several Caduceus effects, and Experiment 4 has no local-pooling ablation. For Caduceus we use the Ph checkpoint and do not test the representation-level equivariance of Caduceus-PS. The matched designs control gene identity and, for Experiment 2, substitution type and CpG context, but not splice proximity, conservation, or transcript position. Sample sizes (500–4,385 per analysis) are modest, allele-frequency strata from gnomAD are not yet included, and we evaluate one checkpoint size per model family. We do not measure downstream predictive utility. The motif analysis considers CTCF and one PWM-based perturbation design. ClinVar labels inherit ascertainment bias and uneven gene coverage. Finally, BioInvariant evaluates frozen zero-shot representations and does not predict what happens after fine-tuning or calibration.

Implications. BioInvariant adds an evaluation axis complementary to downstream accuracy: whether representation geometry is biologically consistent under matched controls. The broad conclusion is not that one model is universally best, but that raw similarity can mislead about invariance, raw distance can mislead about biological sensitivity, and representation behavior depends on model, biological axis, and context length. Future genomic foundation model benchmarks should therefore report controlled invariance/sensitivity tests alongside downstream accuracy. The protocol itself is not specific to DNA. It needs only a frozen encoder, a transformation whose biological meaning is known, and a background perturbation matched on the nuisance statistics that could explain the effect. The same design applies to protein language models (e.g., synonymous-codon or conservative substitutions against matched controls) and to other scientific foundation models, where the same pitfalls we document are likely to recur: raw similarity inflated by composition, very small raw distances, pooling dilution, and position effects in causal encoders.

AI USE DISCLOSURE

The authors used an AI coding and editing assistant based on a large language model to help with manuscript editing, formatting, and cross-reference checks. Experimental design, data analysis, and scientific interpretation were performed by the authors.

REFERENCES

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., and Liu, H. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. ICLR, 2024. arXiv:2306.15006.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. Nature Methods, 22:287–297, 2025.

Nguyen, E., Poli, M., Faizi, M., et al. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. NeurIPS, 36:43177–43201, 2023.

Schiff, Y., Kao, C. H., Gokaslan, A., Dao, T., Gu, A., and Kuleshov, V. Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling. ICML, PMLR 235:43632–43648, 2024.

Marin, F. I., Teufel, F., Horlacher, M., Madsen, D., Pultz, D., Winther, O., and Boomsma, W. BEND: Benchmarking DNA Language Models on Biologically Meaningful Tasks. ICLR, 2024. arXiv:2311.12570.

Patel, A., Singhal, A., Wang, A., Pampari, A., Kasowski, M., and Kundaje, A. DART-Eval: A Comprehensive DNA Language Model Evaluation Benchmark on Regulatory DNA. NeurIPS Datasets and Benchmarks Track, 2024. arXiv:2412.05430.

Ma, M. Reverse-Complement Consistency for DNA Language Models. arXiv:2509.18529, 2025.

Cheng, B. and Zhang, J. The Mechanistic Invariance Test: Genomic Language Models Fail to Learn Positional Regulatory Logic. arXiv:2604.06549, 2026.

Feng, X. et al. Benchmarking DNA foundation models for genomic and genetic tasks. Nature Communications, 2025.

Cohen, J. Things I Have Learned (So Far). American Psychologist, 1990.

Efron, B. and Tibshirani, R. J. An Introduction to the Bootstrap. Chapman & Hall/CRC, 1993.

Benegas, G., Batra, S. S., and Song, Y. S. DNA language models are powerful predictors of genome-wide variant effects. PNAS, 120(44):e2311219120, 2023.

Nguyen, E., Poli, M., Durrant, M. G., et al. Sequence modeling and design from molecular to genome scale with Evo. Science, 386(6723):eado9336, 2024.

Landrum, M. J., Chitipiralla, S., Brown, G. R., et al. ClinVar: improvements to accessing data. Nucleic Acids Research, 48(D1):D835–D844, 2020.

Lin, Z., Sun, H., Guo, Y., et al. EnTao-GPM: DNA Foundation Model for Predicting the Germline Pathogenic Mutations. arXiv:2507.21706, 2025.