GDTR: Layer-wise Settling Depth Reveals Biological Grammar in Genomic Foundation Models

Anonymous Authors[1]

Abstract

Genomic foundation models encode rich sequence regularities, yet existing tools rarely answer *where* in the layer stack a specific biological grammar becomes stable. We introduce GDTR (*Genomic Deep-Thinking Ratio*), a training-free residual-stream lens that assigns each nucleotide token a *settling depth* $c(t)$: the first layer at which its representation stabilises against the post-final-norm reference. On Evo 2 7B, splice donor/acceptor sites settle ~ 2 layers earlier than intronic and coding contexts (Cohen's $d = -0.43$), with ENCODE enhancer-like cCREs showing a smaller but measurable shift ($d = -0.19$); a single chr22 calibration transfers to held-out chr17 (94.6% effect retention). Crucially, $c(t)$ is informative in *both directions*: editing the central GT donor motif deepens settling, whereas shuffling the flanking grammar lets the isolated motif settle 3.18 layers earlier — so motif edits and flank shuffles dissociate motif detection from grammar integration. Six ClinVar molecular-consequence classes also differ in the layer at which variant-induced residual disruption peaks (Kruskal–Wallis $p = 3.0 \times 10^{-10}$), with synonymous substitutions peaking at the deepest layers. GDTR positions settling depth as a layer-wise interpretability axis for genomic foundation models, complementing existing variant scorers.

1. Introduction

Genomic foundation models — causal LMs (Evo 2 (Nguyen et al., 2026), HyenaDNA (Nguyen et al., 2023)), masked LMs (the Nucleotide Transformer (Dalla-Torre et al., 2024), DNABERT-2 (Zhou et al., 2024)), and bidirectional SSMs (Caduceus (Schiff et al., 2024)) — have catalysed a paradigm shift in functional genomics. Yet their internal layer-wise dynamics remain difficult to interpret. Existing tools ask *where* models attend (Abnar & Zuidema, 2020; Avsec et al., 2021), *what* they encode (Bricken et al., 2023; Cunningham et al., 2023), or *how* variants shift representations (Cheng et al., 2023; Jaganathan et al., 2019). A complementary question is largely unmeasured: *where in the layer hierarchy does a biological sequence element become stable?* We use *biological grammar* informally for the context-dependent integration of motif and flank.

We introduce GDTR, a layer-wise readout of residual-stream commitment depth. The Deep-Thinking Ratio of Chen et al. (2026) measures, in NLP transformers, the layer at which intermediate logit-lens distributions (nostalgebraist, 2020; Belrose et al., 2023; Pal et al., 2023) converge within ε of the final layer. GDTR adapts this idea to nucleotide sequences with three changes: a cosine residual-stream lens for small genomic vocabularies; a running-minimum envelope for non-monotone genomic-model trajectories; and an Evo 2-specific post-final-norm reference that handles representational saturation before the final block (§2). Fig. 1 summarises the readout.

This short paper makes three contributions. **(C1)** We define the per-token *settling depth* $c(t)$ for genomic causal LMs, including the architectural handling required for Evo 2's pre-final saturation, and validate that a single chr22 calibration transfers to held-out chr17. **(C2)** We show that splice sites and enhancer-like cCREs settle earlier than surrounding genomic contexts, that the splice signal is not explained by next-token entropy, and that targeted motif/flank perturbations cleanly separate motif detection from grammar integration. **(C3)** We show that ClinVar molecular-consequence classes differ in the layer at which a variant most disrupts the residual trajectory, providing a depth probe rather than a clinical scorer. The unifying claim: biological grammar changes not only *what* a genomic foundation model predicts, but also *where* its representation settles.

2. The GDTR Framework

For a nucleotide sequence of length T processed by a causal language model with L layers, we extract the residual-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

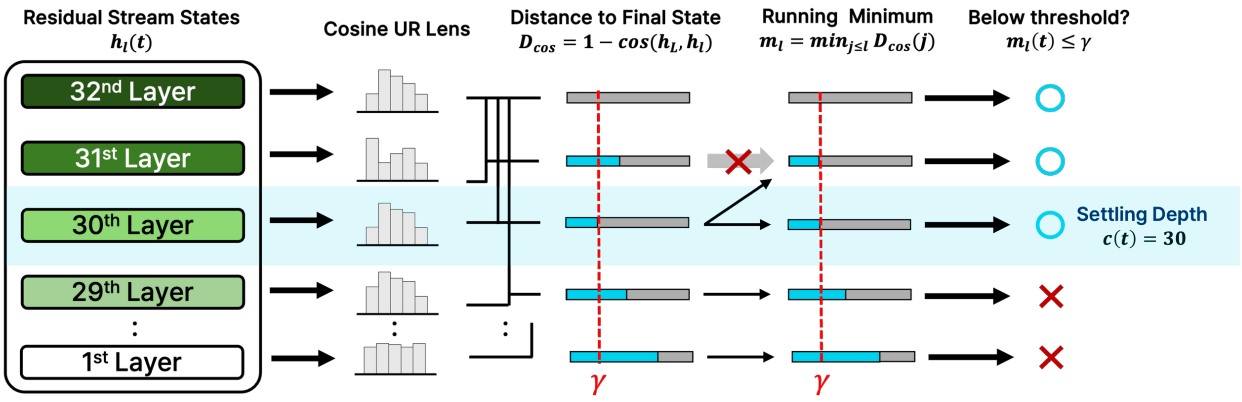

Figure 1. GDTR measures the layer at which a token's representation stabilises. Each Evo 2 layer emits a residual-stream tap $h_\ell(t)$; the *cosine lens* compares it to the final-layer state h_L via $D_{\cos}(\ell) = 1 - \cos(h_L, h_\ell)$, and the *running-minimum envelope* $m_\ell = \min_{j \le \ell} D_{\cos}(j)$ enforces monotone descent. The first layer with $m_\ell \le \gamma$ defines the settling depth $c(t)$ (Eq. 2). In the worked example, $c(t) = 30$: layer 31 alone has a larger $D_{\cos}$, but the envelope carries layer 30's smaller value forward. Lower $c(t)$ means earlier stabilisation.

stream activation $h_\ell(t)$ at every layer $\ell \in \{1, \dots, L\}$ and token position t. We replace the JSD lens of the parent DTR (Chen et al., 2026) with a cosine residual-stream lens that operates directly in hidden-state geometry,

$$D_{\cos}(\ell, t) \;=\; 1 - \cos\big(h_\ell(t),\, h_{\text{norm}}(t)\big), \qquad (1)$$

where $h_{\text{norm}}(t)$ is the post-final-norm output of the model after the canonical interpretively distinct tap at $L^\star = 29$ has been propagated through the final RMSNorm (blocks indexed $0, \dots, L-1$ throughout the Evo 2 analysis). We define the *settling depth* via a running-minimum envelope,

$$c(t) \;=\; \min\big\{\, \ell \;:\; \text{run-min}\; D_{\cos}(\ell, t) \le \gamma_{\cos} \big\}, \qquad (2)$$

with run-min $D_{\cos}(\ell, t) = \min_{k \le \ell} D_{\cos}(k, t)$. Lower $c(t)$ means earlier stabilisation; higher $c(t)$ means that the token remains unresolved until later layers. NLP transformers exhibit near-monotonic logit-lens convergence (nostalgebraist, 2020; Belrose et al., 2023); genomic CLMs do not (Hyena alignment rotations, Evo 2 pre-final saturation), so the running-min collapses non-monotonicity into a monotone-non-increasing trace. The threshold $\gamma_{\cos}$ is set by *regional* q_{70} *calibration*: the 70th percentile of the running-min at the penultimate layer over the analysis region. The operating point sits inside a flat 5×5 grid sweep over $(\gamma_{\cos}, \rho)$ (App. A.2).

Settling depth is two-sided. A lower $c(t)$ does not by itself imply a stronger biological signal. A token can settle early because (a) its representation is constrained by a biological grammar that the model commits to early, *or* (b) its surrounding context is simpler, so the running-min envelope reaches $\gamma_{\cos}$ with little integration. We treat this two-sidedness as a feature of the metric rather than a flaw: the perturbation experiments in §3.2 are designed to push $c(t)$ in opposite directions, separately stressing motif detection and flanking-grammar integration. Throughout the paper we therefore interpret depth signatures as *bidirectional* — both directions of movement carry information about how the model integrates biological grammar, and we apply this lens consistently to every per-context comparison reported below.

Representational saturation before the final block. Evo 2 reaches the post-norm direction before the last attention block executes: on chr22 sanity sequences, $\max_t |h_{30}(t) - h_{31}(t)| = 0$ exactly (block 31 is a residual passthrough), while $\cos(h_{29}, h_{\text{norm}}) = -0.013$ versus $\cos(h_{30}, h_{\text{norm}}) = \cos(h_{31}, h_{\text{norm}}) = +0.6855$ — block 30 rotates the representation into the post-norm direction, and block 29 is the deepest tap still interpretively distinct from h_{norm}. The canonical "tap the last block" convention therefore breaks; we lock the canonical tap at $L^\star = 29$ and use h_{norm} as the reference (full evidence in App. A.1). We treat $L^\star = 29$ as an empirical observation about Evo 2 representational saturation, not a claim about the model's architectural design intent.

Tuned-lens sanity check against frame mismatch. A per-layer tuned-lens affine fit (Belrose et al., 2023) reaches $\ge 98\%$ MSE recovery on $30/32$ Evo 2 layers (canonical $L^\star = 29$: 0.9996; App. A.4). This is a sanity check against gross norm-induced frame mismatch: the interior taps are linearly recoverable to the post-norm frame. The framework remains training-free because no affine weights are used at inference.

3. Layer-wise Biological Grammar

3.1. Splice and Enhancer Annotations Settle Early

We first test whether $c(t)$ delivers a biologically meaningful, transferable readout at the genome scale. Chromosome 22 serves as the calibration set ($\gamma_{\cos} = 0.397$ from q_{70} at the penultimate layer); chromosome 17 is held out for validation

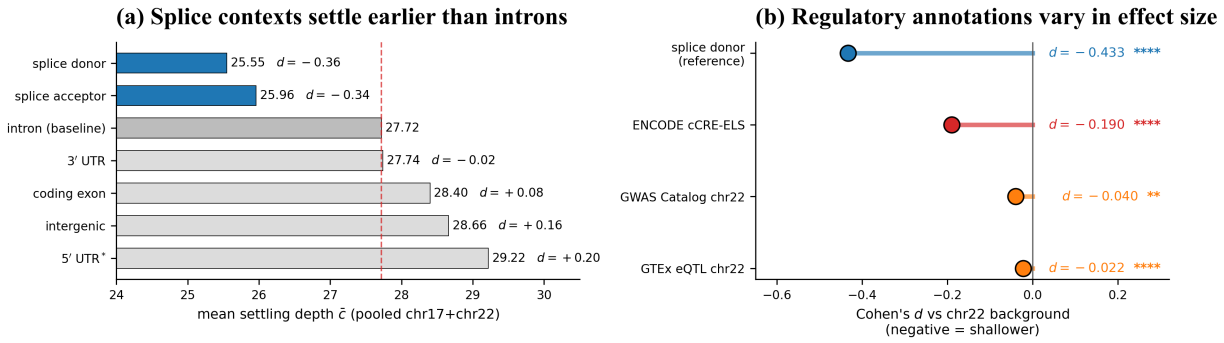

Figure 2. Splice and enhancer annotations shift toward shallower settling. **(a)** Per-context mean settling depth $\bar{c}$ pooled over chr17 (validation, frozen $\gamma_{\cos} = 0.397$) and chr22 (calibration). Splice donor and acceptor (blue) sit ~ 2 layers below the intron baseline ($\bar{c}=27.72$, red dashed; panel (b) d uses this same baseline). The asterisk on "5′ UTR*" marks an entropy-coupled context (§4) reported separately from the entropy-controlled hierarchy. Right of each bar: Cohen's d versus the intron baseline (Mann–Whitney significance: $*$, $**$, $***$, $***$ at 0.05, 0.01, 10^{-3}, 10^{-4}). **(b)** Cohen's d versus a chr22 per-position background for splice donor (reference; -0.433), ENCODE cCRE-ELS (-0.190), GTEx eQTL (-0.022) and GWAS Catalog (-0.040). cCRE-ELS gives a measurable enhancer signal; SNP-level eQTL/GWAS shifts are directionally consistent but biologically marginal.

with that γ frozen. The chr17 q_{70} recomputed independently equals 0.394 ($|\Delta|=0.0023$); the chr17 splice-donor effect reaches 94.6% of the chr22 magnitude (Cohen's $d=-0.349$ vs intron), and the relative ordering of the seven canonical contexts is preserved (Spearman $\rho=0.93$). The calibration transfers without re-tuning, and we use this single γ for all subsequent context- and variant-level comparisons.

Context-level ordering. On canonical genomic contexts (Fig. 2a), splice donor and acceptor sites settle ~ 2 layers earlier than the intronic baseline, and the remaining contexts cluster near or above it. Effect sizes are reported as Cohen's d throughout; Mann–Whitney p-values are dominated by the $\sim 10^7$ pooled positions and act only as significance sanity checks. Among the regulatory annotations (Fig. 2b), ENCODE cCRE-ELS enhancer-like elements (Moore et al., 2020) are the clearest non-splice signal at Cohen's $d = -0.190$ against a chr22 per-position background (vs. splice-donor $d = -0.433$): biologically modest but statistically robust and chr22→chr17 transferable. GTEx eQTLs (The GTEx Consortium, 2020) ($d=-0.022$) and GWAS Catalog SNPs (Sollis et al., 2023) ($d=-0.040$) shift in the same direction but with marginal magnitudes. The 5′ UTR shift is reported separately: at 5′ UTR, $c(t)$ couples strongly with next-token entropy ($\rho=+0.41$ vs. $|\rho|\leq 0.16$ elsewhere; §4), so it cannot be compared on the same axis as the entropy-controlled splice and cCRE-ELS signals.

Entropy does not explain the splice signal. A natural concern is that $c(t)$ merely tracks the per-position next-token Shannon entropy H_t of the model. On a control panel of 120 random chr22 windows (719,000 analysed positions) we compute H_t from the post-norm logits and find an overall Spearman $\rho(c, H_t) = -0.079$ (per-context $|\rho| \leq 0.16$ except 5′ UTR; §4; full per-context numbers in Tab. A3, App. A.3). On the same chr22 panel the splice-donor-versus-intron Cohen's d is -0.452; after regressing c on H_t and re-measuring on the residual it strengthens to -0.583. Next-token entropy therefore does not explain the splice depth ordering.

3.2. Motif Edits Separate Detection from Flanking Context

Settling depth is not simply a meter of motif intrinsic strength. We exploit its two-sided behaviour (§2) on 1,000 canonical GT-AG donors (chr22, ± 10 bp pad-averaged $\bar{c}$). Real donors settle at $\bar{c}=26.77$. (i) Replacing the central GT with AA (preserving the ± 100 bp flank) deepens $\bar{c}$ by 0.46 layers ($d = -0.086$, paired Wilcoxon $p = 2.3 \times 10^{-32}$): a small but reliable canonical-dinucleotide signal. (ii) Dinucleotide-shuffling the ± 100 bp flank while preserving the central GT *lifts* $\bar{c}$ to a shallower value by 3.18 layers ($d = +0.51$, $p = 4.1 \times 10^{-59}$): the isolated GT becomes easier to stabilise, whereas the real donor context requires deeper flanking-grammar integration. The two perturbations push $\bar{c}$ in opposite directions — the asymmetry expected if the depth signature reads out grammar integration rather than motif strength alone. From depth alone, we cannot decisively rule out an alternative reading — that on shuffled flanks the model *disengages* from uninformative noise and commits early to the lone GT ("early simplification" rather than reduced integration). The within-splice motif breakdown in App. D.2 mirrors this cautionary logic.

3.3. Variant Disruptions Peak at Consequence-Specific Layers

For variants, we use the layer at which $|\Delta D_{\cos}|$ peaks as a depth probe: it asks at which layer the residual trajectory changes most between reference and alternate sequence – a depth signature, not a pathogenicity score. We test this on a ClinVar (Landrum et al., 2020) 15-cancer-gene cohort, class-

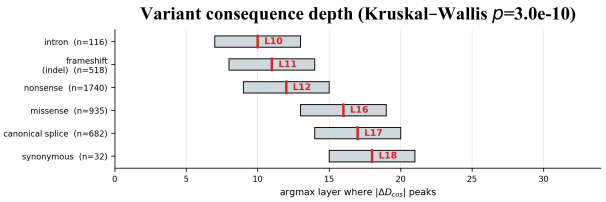

Figure 3. Variant consequences peak at different disruption layers. Argmax layer of $|\Delta D_{\cos}|$ across 4,023 ClinVar P/LP variants in 15 cancer-associated genes, grouped by molecular consequence; class-median layer in red. Kruskal–Wallis 6-way $p = 3.0 \times 10^{-10}$; the largest adjacent jump is nonsense → missense ($p_{\text{adj}} < 10^{-3}$). The broad overlap emphasises a population-level depth shift, not class separation.

stratified by ClinVar molecular-consequence (MC) priority, in decreasing order: canonical-splice, nonsense, frameshift, missense, synonymous, intron.

Argmax-layer class medians read *intron* < *frameshift* < *nonsense* < *missense* ≈ *canonical splice* < *synonymous* (Fig. 3), and the six classes differ globally at Kruskal–Wallis $p = 3.0 \times 10^{-10}$ (Kruskal & Wallis, 1952). The only adjacent gap individually resolved by Bonferroni-corrected Dunn tests (Dunn, 1964) is nonsense → missense ($p_{\text{adj}} < 10^{-3}$); the other adjacent contrasts are consistent with the global test but not individually resolved at this sample size ($n = 32$–1,740). The interpretable result is a population-level depth shift from early-truncating consequences (intron, frameshift, nonsense) toward missense and splice disruptions, and onward to synonymous substitutions, which peak at the deepest layers — consistent with protein-semantic information consolidating in late layers, where coding-frame–dependent disruptions register earlier and identity-preserving substitutions register only after codon-level meaning has been integrated.

The trajectory carries non-trivial information (a logistic regression on the full 32-dimensional $\Delta D_{\cos}$ trajectory reaches 10-fold stratified AUROC 0.844 vs. 0.729 for the best single-layer tap; per-gene LOO AUROC stays ≥ 0.77, and DeLong tests (DeLong et al., 1988) confirm $\Delta D_{\cos}$ adds information beyond Evo 2 log-likelihood with $p = 3.6 \times 10^{-15}$). We report these AUROC numbers as a sanity check that the trajectory is not noise; GDTR's contribution is the *layer at which disruption peaks*, not a competing pathogenicity classifier, and head-to-head comparison against task-specific variant scorers (Cheng et al., 2023; Jaganathan et al., 2019; Rentzsch et al., 2019) is orthogonal to that question. Full panels in App. C.

3.4. Robustness and Tokenisation Limits

Extending the chr22 windows to four genomic FMs (Evo 2 7B, HyenaDNA-large, NT-v2 500M, DNABERT-2) gives a bounded robustness result (Table 1; details in App. B). The donor < intron inequality replicates in the two per-bp causal LMs (Spearman $\rho = +0.516$). The two MLMs operate at k-mer/BPE token granularity, so their per-window readout cannot test a single-base splice junction — a replication within per-bp causal LMs and a diagnosis of tokenisation limits for MLMs.

Table 1. Cross-model summary. Full per-window numbers and MLM tokenisation caveats are in App. B.

Family (model)	Per-pos. readout	Donor < intron
Causal LM (Evo 2, HyenaDNA-large)	yes	yes
MLM (NT-v2, DNABERT-2)	no (per-window)	resolution-limited

4. Discussion and Limitations

The depth signature is *bidirectional*: motif edits deepen $\bar{c}$ while flank-shuffles lift it, and both directions are evidence that $c(t)$ reads out grammar integration rather than motif strength alone (§2, §3.2). Four scope conditions then frame the interpretation. *(i)* The genome-wide context-level findings (§3.1) are correlational readouts of representational dynamics, while the motif/flank perturbations (§3.2) are interventional but limited to a single locus class; establishing causal circuits at scale will require broader interventional follow-up (e.g. activation patching, sparse-dictionary edits). *(ii)* At 5′ UTR alone, $c(t)$ correlates strongly with prediction entropy ($\rho = +0.41$ vs. $|\rho| \leq 0.16$ elsewhere); the 5′ UTR depth shift therefore reflects a mixture of representational stabilisation and prediction confidence and is reported as a separate, entropy-coupled effect. *(iii)* The current entropy control isolates next-token uncertainty but does not control for sequence composition: k-mer rarity, GC content, and dinucleotide composition are the most plausible alternative explanation a reviewer would raise for the depth ordering, and we explicitly flag this as the leading composition confounder pending composition-matched negative controls. *(iv)* Single-base splice resolution requires a per-bp readout. Adding Caduceus (Schiff et al., 2024) as a second per-position model (bidirectional SSM rather than causal) would disentangle directional inductive bias from tokenisation in the cross-architecture comparison.

5. Conclusion

GDTR establishes a training-free, layer-resolved interpretability axis for genomic foundation models. Three findings define its message. *(1)* The depth signature is *bidirectional*: motif edits deepen $\bar{c}$ while flank-shuffles lift it, reading out grammar integration. *(2)* Splice sites and enhancer-like cCREs stabilise earlier than intronic/coding contexts, and the splice signal is not explained by next-token entropy — the readout reflects *detection plus context integration*. *(3)* Variant-induced $\Delta D_{\cos}$ peaks at consequence-specific layers, with synonymous substitutions peaking deepest. Whole-genome scaling, composition-matched controls, and sparse-

dictionary (Cunningham et al., 2023)/causal-edit (Meng et al., 2022) integration remain for future work.

References

Abnar, S. and Zuidema, W. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, 2021.

Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Burge, C. B., Padgett, R. A., and Sharp, P. A. Evolutionary fates and origins of U12-type introns. *Molecular Cell*, 2 (6):773–785, 1998.

Chen, W.-L. et al. Think deep, not just long: Measuring LLM reasoning effort via deep-thinking tokens. *arXiv preprint arXiv:2602.13517*, 2026.

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492, 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., et al. The Nucleotide Transformer: Building and evaluating robust foundation models for human genomics. *Nature Methods*, 2024.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

Dunn, O. J. Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252, 1964.

Jaganathan, K., Kyriazopoulou-Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., et al. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548, 2019.

Kruskal, W. H. and Wallis, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

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

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Moore, J. E., Purcaro, M. J., Pratt, H. E., et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583:699–710, 2020.

Nguyen, E., Poli, M., Faltings, B., et al. HyenaDNA: Long-range genomic sequence modelling at single-nucleotide resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Nguyen, E., Poli, M., Durrant, M. G., et al. Evo 2: Whole-genome modelling with context-length scaling. *Nature*, 2026.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong post, 2020.

Pal, K., Sun, J., Yuan, A., Wallace, B. C., and Bau, D. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, 2023.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894, 2019.

Schiff, Y., Kao, C.-H., Gokaslan, A., Dao, T., Gu, A., and Kuleshov, V. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *International Conference on Machine Learning (ICML)*, 2024.

Sollis, E., Mosaku, A., Abid, A., et al. The NHGRI-EBI GWAS Catalog: Knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1):D977–D985, 2023.

The GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

Wang, Z. and Burge, C. B. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5):802–813, 2008.

Zhou, Z., Ji, Y., Li, W., et al. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. In *International Conference on Learning Representations (ICLR)*, 2024.

Appendix

*Supplementary material for "*GDTR*: Layer-wise Settling Depth Reveals Biological Grammar in Genomic Foundation Models"*

A. Method Details

A.1. Architectural Quirk Handling: Evo 2's Idle Last Block

We discover that Evo 2 7B's last attention block is architecturally idle. Direct verification on chr22 sanity sequences (100 windows of 6 kb each) yields the values in Table A1. Two facts matter: (i) $\max_t |h_{30}(t) - h_{31}(t)| = 0$ exactly, so block 31 is a residual passthrough; (ii) the cosine alignment with the post-final-norm reference h_{norm} jumps from $\cos(h_{29}, h_{\text{norm}}) = -0.013$ (near-orthogonal) to $\cos(h_{30}, h_{\text{norm}}) = \cos(h_{31}, h_{\text{norm}}) = +0.6855$, so block 30 is the rotation that moves the representation into the post-norm direction. The deepest tap that is still *interpretively distinct* from h_{norm} is therefore $L^\star = 29$. The canonical NLP convention of "tap the last block" does not apply to Evo 2; we lock the canonical deep-thinking tap at $L^\star = 29$ and use h_{norm} as the convergence reference.

Table A1. Evidence that Evo 2's last attention block is architecturally idle. Block 31 is bit-identical to block 30 in value but is a distinct tensor; both differ from the post-final-norm reference, identifying block 29 as the deepest interpretively distinct tap.

Quantity	Measured value	Interpretation
$\max\lvert h_{30} - h_{31}\rvert$	0 (exact)	block 31 is residual passthrough
`data_ptr`(h_{30}) vs. h_{31}	distinct	physically separate copies
$\cos(h_{31}, h_{\text{norm}})$	0.6855	post-norm differs from raw h_{31}
$\cos(h_{30}, h_{\text{norm}})$	0.6855	identical to h_{31}
$\cos(h_{29}, h_{\text{norm}})$	-0.013	deepest distinct tap ($L^\star = 29$)

A.2. Hyperparameter Sensitivity

On chr22 the regional q_{70} calibration yields $\gamma_{\cos} = 0.397$. The 5×5 grid sweep (Table A2) produces a wide, flat plateau around the operating point ($\gamma_{\cos} = 0.40,\ \rho = 0.85$) with $\pm 0.10 \times \pm 0.05$ variation. Within this plateau the standard deviation across the 25 cells is 0.06, indicating low sensitivity to small perturbations in either hyperparameter. This robustness was first identified in Phase 0 (HyenaDNA-medium-160k), where the analogous q_{70} calibration produced a best operating point of ($\gamma_{\cos} = 0.50,\ \rho = 0.85$) with Cohen's $d = -1.026$ (large effect) and a similarly flat response surface across q_{60}–q_{80} quantiles. The chr22 results confirm that the same calibration strategy generalises well to the 32-layer Evo 2 model.

Table A2. Effect-size sweep over $(\gamma_{\cos}, \rho)$. Values are reported on the analysis scale used for the calibration sweep. The locked operating point $(0.40, 0.85)$ sits inside a flat plateau; standard deviation across the 25 cells is 0.06.

$\gamma_{\cos} \backslash \rho$	$\rho = 0.70$	$\rho = 0.75$	$\rho = 0.80$	$\rho = 0.85$	$\rho = 0.90$
0.30	5.04	5.18	5.21	5.20	5.15
0.35	5.10	5.21	5.24	5.23	5.18
0.40	5.16	5.25	**5.28**	**5.26**	5.21
0.45	5.13	5.22	5.25	5.24	5.19
0.50	5.08	5.17	5.20	5.19	5.14

A.3. Entropy Decoupling per Context

Table A3 reports per-context Spearman $\rho(c, H_t)$ on a 120-window chr22 control panel (719,000 analysed positions; the 1,000 shortfall versus $120 \times 6{,}000$ reflects truncated logits at window edges where the causal context is incomplete), where H_t is the per-position next-token Shannon entropy from the post-norm logits. The overall correlation is small and slightly negative ($\rho = -0.079$); per-context correlations stay $|\rho| \leq 0.16$ for every context except 5′ UTR ($\rho = +0.41$). The largest non-UTR couplings are splice acceptor at $|\rho| = 0.152$ and intron at $|\rho| = 0.108$ (Tab. A3). After regressing c on H_t on the same chr22 panel, the splice-donor-versus-intron Cohen's d *strengthens* from -0.452 to -0.583 on the residual, confirming that next-token entropy does not explain the splice depth ordering.

Table A3. Per-context Spearman correlation between settling depth $c(t)$ and next-token entropy H_t on the chr22 control panel (120 windows, 719,000 analysed positions; the 1,000 shortfall versus $120 \times 6{,}000$ reflects truncated logits at window edges). The next-largest non-UTR couplings are splice acceptor ($|\rho| = 0.152$) and intron ($|\rho| = 0.108$); 5′ UTR is the only entropy-coupled context.

Context	n	$\bar{c}$	$\bar{H}_t$	$\rho(c, H_t)$
intergenic	243,956	28.94	0.795	−0.024
intron	425,131	28.03	1.030	−0.108
coding exon	36,427	28.61	0.744	−0.080
3′ UTR	8,238	27.61	1.073	−0.028
splice donor	1,981	25.04	0.887	−0.083
splice acceptor	1,920	27.33	0.764	−0.152
5′ UTR	2,347	31.43	1.121	**+0.414**
overall	719,000	—	—	−0.079

A.4. Tuned-Lens Recovery Across All 32 Layers

Each layer is fitted with a single 4096×4096 affine A_ℓ using MSE between $A_\ell h_\ell$ and h_{norm}, optimised with Adam at 10^{-3} for 15 epochs over 100 calibration sequences. The post-norm tap is the prediction target. Recovery scores at five representative layers are reported in Table A4; $30/32$ layers reach $\geq 98\%$. The 98% threshold is descriptive of the empirical recovery distribution rather than a pre-registered cut-off: it summarises where the bulk of the per-layer recovery scores sit, not a criterion that the framework is required to clear at inference (the framework is training-free and uses no affine weights when computing $c(t)$).

Table A4. Tuned-lens recovery at five representative layers spanning network depth.

Layer / block type	Initial MSE	Final MSE (15 epochs)	Recovery
$\ell = 2$ (`hcl`)	1,259	5.6	0.9956
$\ell = 12$ (`hcm`)	742	13.6	0.9816 (worst)
$\ell = 22$ (`hcm`)	418	0.41	0.9990
$\ell = 28$ (`hcs`)	822	0.34	0.9996 (best below tap)
$\ell = 29$ (canonical tap, `hcm`)	510	0.20	0.9996

B. Cross-Architecture Replication: Scope, Granularity, Two-Tier Structure

This appendix expands the bounded robustness result of §3.4. We compare four publicly released genomic foundation models that span the per-bp causal-LM and the tokenised MLM design space: Evo 2 7B, HyenaDNA-large-1m, NT-v2 500M, and DNABERT-2 117M. Architectural specifications, the per-window token budget they impose on the same chr22 sequences, end-to-end runtime, and the per-model q_{70}-calibrated $\gamma_{\cos}$ are summarised in Table A5. All four models use the identical chr22 12,978-window panel.

Table A5. Architectural specifications, tokenisation, per-window context, runtime, and per-model q_{70}-calibrated $\gamma_{\cos}$ for the four genomic foundation models compared in §3.4.

Model	Architecture	Layers	Hidden	Tokens / 6 kb window	Wall-clock	$\gamma_{q_{70}}$
Evo 2 7B	Hybrid Transformer + StripedHyena 2	32	4096	6,000 (1 bp)	reused	0.396
HyenaDNA-large-1m	Pure Hyena	8	256	6,001 (1 bp + BOS)	~ 4 min	0.358
NT-v2 500M	Transformer MLM (k-mer)	29	1024	671 ($k = 6$, ~ 4 kb)	~ 7.5 min	0.533
DNABERT-2 117M	Transformer MLM (BPE)	12	768	~ 600 (BPE, ~ 3 kb)	~ 3 min	0.677

Per-context replication. Table A6 reports the per-context mean settling depth for each model. The two per-bp causal LMs (Evo 2, HyenaDNA-large) replicate donor $<$ intron and acceptor $<$ intron at depth-normalised scale (~ 3 layers below intron in 32-layer Evo 2; ~ 0.34 layers below intron in 8-layer HyenaDNA). The two MLMs tokenise at k-mer/BPE granularity: their per-window readout collapses splice-junction signal into the surrounding window mean and therefore matches the exon-dominant window mean to within rounding, consistent with a tokenisation-induced loss of single-base resolution rather than a model-class disagreement.

Pairwise rank concordance. Table A7 shows the pairwise Spearman ρ of per-window mean settling depth across the four models. The pattern is two-tier: within-family correlations are positive ($\rho_{\text{Evo2,Hyena}} = +0.516$; $\rho_{\text{NT,DNABERT}} = +0.663$),

Table A6. Per-context mean settling depth across the four models on the identical chr22 12,978-window set. Causal-LM models recover donor/acceptor < intron at depth-normalised scale; MLM models report per-window estimates with no per-position separation.

Model	Kind	L	$\gamma_{q_{70}}$	$\bar{c}$ (intron)	$\bar{c}$ (coding exon)	$\bar{c}$ (donor / acceptor)
Evo 2 7B	per_position	32	0.396	27.84	28.28	25.59 / 25.71
HyenaDNA-large	per_position	8	0.358	6.89	6.67	6.55 / 6.62
NT-v2 500M	per_window	29	0.533	n.a. (intron-dominant 0)	27.80 (exon-dom.)	27.85 (splice-cont.)
DNABERT-2 117M	per_window	12	0.677	n.a. (intron-dominant 0)	11.28 (exon-dom.)	11.27 (splice-cont.)

whereas every cross-family pair is weakly negative. This is the empirical basis for the "replication within per-bp causal LMs, tokenisation-bound for MLMs" claim of §3.4.

Table A7. Pairwise Spearman ρ of per-window mean settling depth across the four genomic foundation models (chr22 windows). Within-family correlations are positive; cross-family correlations are weakly negative.

	Evo 2 7B	HyenaDNA-large	NT-v2 500M	DNABERT-2
Evo 2 7B	1.000	+0.516	−0.119	−0.188
HyenaDNA-large	+0.516	1.000	−0.287	−0.166
NT-v2 500M	−0.119	−0.287	1.000	+0.663
DNABERT-2	−0.188	−0.166	+0.663	1.000

Visual summaries. The per-context bar chart (Fig. A1) shows the donor/acceptor < intron inequality directly for the per-bp causal LMs and the absence of separation for the tokenised MLMs. Fig. A2 consolidates the rank-concordance heatmap, the depth-normalised donor-vs-intron comparison, and a schematic of the two-tier structure into a single panel.

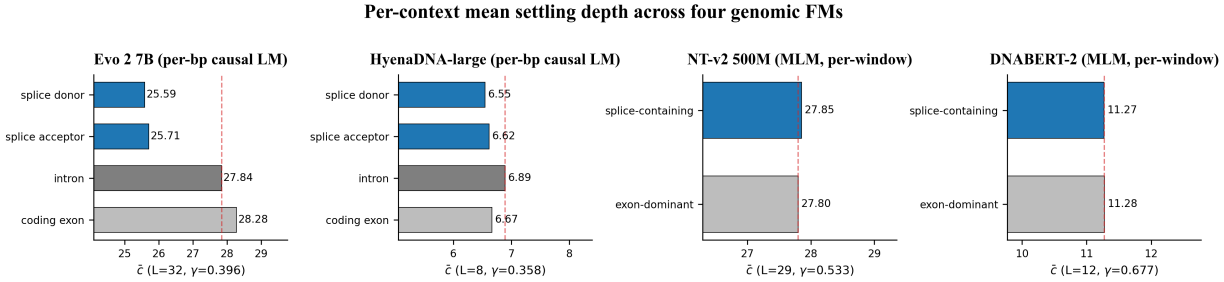

Figure A1. Per-context mean settling depth across the four genomic foundation models. Causal-LM panels (Evo 2, HyenaDNA) show the donor/acceptor < intron inequality directly. MLM panels (NT-v2, DNABERT-2) show no separation at tokenised window resolution.

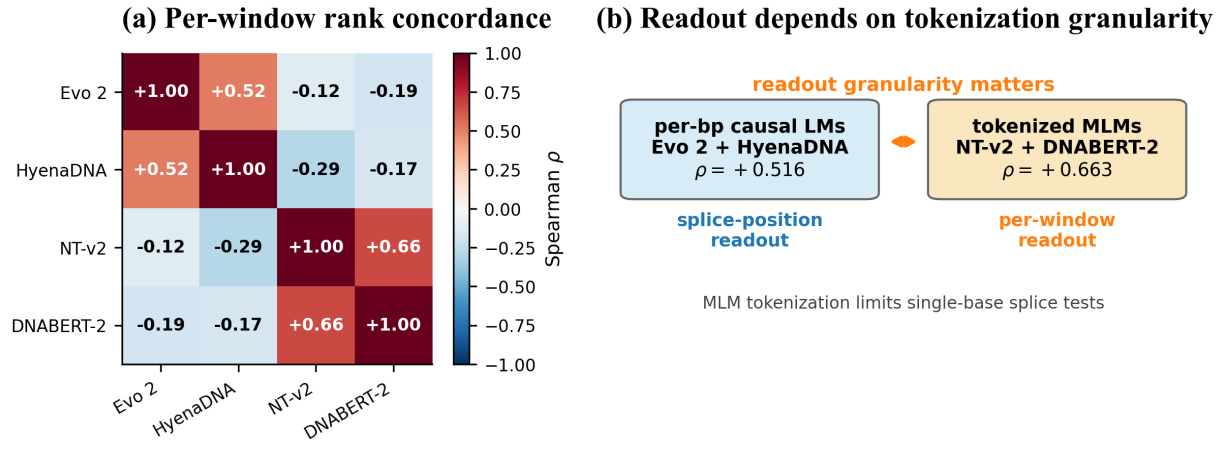

Figure A2. Cross-architecture two-tier structure (numerical detail in Tables A6, A7). **(a)** Pairwise Spearman ρ of per-window mean settling depth across four genomic foundation models on chr22, showing the within-family positive block-diagonal vs. cross-family weak-negative off-diagonal structure. **(b)** Schematic of the two-tier structure with the within/cross-family ρ values, summarising the "per-bp causal LMs replicate, tokenised MLMs are bound by readout granularity" interpretation.

C. Variant Pathogenicity: AUROC Summary, Per-Layer Ablation, and Panels

This appendix backs up the trajectory-information sanity check in §3.3. The cohort is the 8,008 ClinVar P/LP-vs-B/LB SNVs across 15 cancer-associated genes; the splitter is 10-fold stratified cross-validation with seed 42, and the LOGO-CV column reports leave-one-gene-out generalisation across the 14 evaluable genes. We treat AUROC strictly as a sanity check on whether the 32-d $\Delta D_{\cos}$ trajectory carries information beyond any single tap; we do not position GDTR as a clinical scorer (§3.3).

Table A8. Feature-level AUROC summary. The primary cosine lens reaches 0.844 with the full 32-d trajectory and only 0.729 at its best single tap – a +0.115 gap. Bracketed numbers are 1,000-bootstrap 95% confidence intervals.

Feature	dim.	Stratified 10-fold AUROC	LOGO-CV AUROC
Best single-layer $\Delta D_{\cos}$ ($\ell = 30$ tap)	1	0.729 [0.717, 0.741]	0.726 [0.694, 0.758]
Best single-layer ΔD_{jsd} ($\ell = 29$ canonical tap)	1	0.794 [0.781, 0.806]	0.787 [0.752, 0.821]
Evo 2 log-likelihood	1	0.751 [0.738, 0.764]	0.793 [0.740, 0.846]
32-d ΔD_{jsd} vector	32	0.823 [0.813, 0.832]	0.821 [0.790, 0.853]
32-d $\Delta D_{\cos}$ vector (primary)	32	**0.844** [0.831, 0.857]	**0.843** [0.811, 0.876]
Ensemble (ΔD+ Evo 2 LL)	33	0.861 [0.851, 0.871]	0.866 [0.832, 0.899]

Per-layer ablation. Table A9 reports single-layer logistic-regression AUROC at representative taps for both lenses. The two lenses agree on the U-shape of layer-wise discriminative mass but disagree on which tap is sharpest: JSD concentrates mass at the canonical tap $\ell = 29$, whereas cosine spreads mass across many taps and accordingly produces a much larger 32-d-vector-vs-best-single-tap gap.

Table A9. Selected per-layer AUROCs. JSD concentrates discriminative mass at the canonical tap $\ell = 29$; cosine spreads it across many taps, producing a much larger 32-d-vs-best-single-tap gap.

Layer	Block type	ΔD_{jsd} AUROC	$\Delta D_{\cos}$ AUROC
0	embed	0.605	0.519
7	`hcs` (shallow)	0.662	0.595
12	`hcm`	0.656	0.555
17	`attn`	0.723	0.646
24	`attn`	0.685	0.612
28	`hcs`	0.565	0.698
29 (canonical tap)	`hcm`	**0.794** (best JSD)	0.604
30 (post-norm−1)	—	0.512	**0.729** (best cos)
31 (idle, see App. A.1)	—	0.499 (degen.)	0.729 (= h_{30})
32-d vector	all	0.823	0.844

Why $L^\star = 29$, not $\ell = 30$, is the canonical tap. Although block 30 is the rotation step that aligns the representation with h_{norm} (App. A.1), it remains highly informative for downstream tasks: the best single-layer AUROC for the cosine lens occurs precisely at $\ell = 30$ (0.729). This is expected — once the residual stream has rotated into the final-norm frame, it carries maximal signal for linear classification. We therefore retain $L^\star = 29$ as the deepest *interpretively distinct* tap for the settling-depth metric (which compares h_ℓ to h_{norm} and would saturate at the rotation block), while acknowledging that the post-rotation representation ($\ell = 30$) is optimal for linear probing of variant pathogenicity. The two roles are consistent: the deepest tap that is non-trivially distinct from the reference defines the convergence target, and the rotated representation that matches the reference frame defines the maximally classifiable feature. The $\ell = 31$ AUROC inherits the same 0.729 because block 31 is a residual passthrough (Tab. A1); it is reported for completeness rather than as new information.

ROC, DeLong, and per-gene panels. Fig. A3 provides the four diagnostic panels referenced above. ROC curves (a) confirm the ranking from Table A8; the Paired DeLong tests in (b) show that $\Delta D_{\cos}$ adds significant information beyond Evo 2 ΔLL ($p = 3.6 \times 10^{-15}$); (c) is the per-layer ablation visualised across all 32 taps, and (d) is the leave-one-gene-out AUROC across the 14 evaluable genes (uniformly ≥ 0.77).

D. Splice Anatomy Beyond the Headline Contexts

This appendix supports the cautionary message in §3.2: the splice signal is real, but its internal structure is asymmetric and motif-class-dependent.

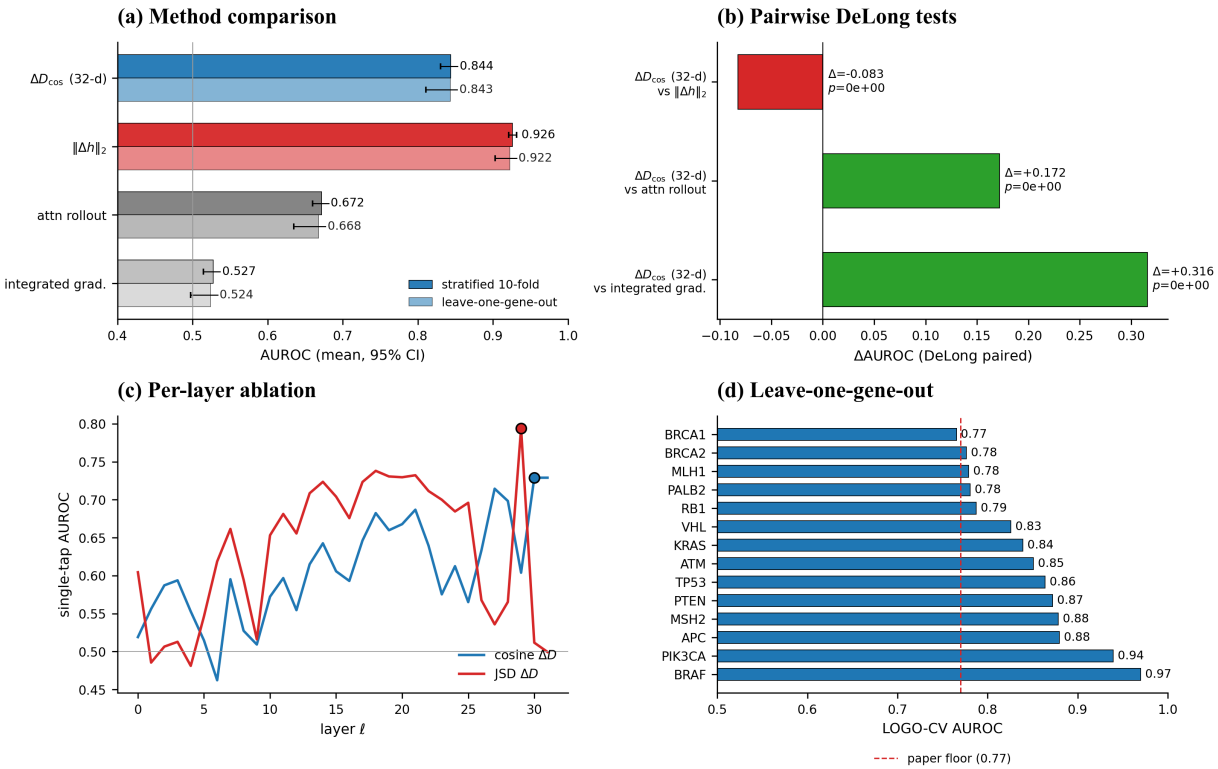

Figure A3. Variant pathogenicity discrimination on 8,008 ClinVar P/LP vs B/LB SNVs across 15 cancer-associated genes (numerical summary in Table A8). **(a)** ROC curves for the four scoring features. **(b)** Paired DeLong tests: ΔD adds significant information beyond Evo 2 ΔLL ($p = 3.6 \times 10^{-15}$). **(c)** Per-layer single-feature AUROC across all 32 layers – best single tap is $\ell = 29$ for JSD and $\ell = 30$ for cosine (the latter is the post-norm rotation block, not the canonical interpretive tap; see App. A.1); the 32-d vector beats either by $+0.05$ to $+0.12$. **(d)** Leave-one-gene-out AUROC across 14 evaluable genes is uniformly high (≥ 0.77).

D.1. Positional Fine-Profile Around Donor / Acceptor

On chr17 the donor profile reaches its mean settling-depth minimum at position $+20$ bp ($\bar{c} = 24.06$) and remains ≥ 1.5 layers below intronic baseline ($\bar{c} = 27.69$) out to ± 200 bp. On the same chromosome the acceptor profile reaches its minimum at $+50$ bp ($\bar{c} = 23.33$) and similarly remains depressed beyond ± 200 bp. On chr22 both donor and acceptor minima sit at coordinate 0 ($\bar{c} = 23.65$ and 23.64 respectively). The asymmetry — donor minimum exonic-side, acceptor minimum intronic-side — mirrors the asymmetric splice-grammar features (branch point, polypyrimidine tract) that lie predominantly on the intronic side of acceptors. The full positional profile is shown in Fig. A4; the per-side minima are summarised in Table A10.

Table A10. Per-side minima of the splice positional fine-profile. Pooled chr17 + chr22 analysis windows; positions measured from the splice junction. Donor exonic-side / acceptor intronic-side minima recapitulate the asymmetric splice-grammar context.

Side	arg-min position	$\bar{c}$ at min	$\bar{c}$ at -100 bp	$\bar{c}$ at $+100$ bp	n
splice donor	$+20$ bp	23.65	26.05	25.65	280,944
splice acceptor	$+50$ bp	23.64	26.01	24.50	278,955

D.2. Canonical vs. Non-Canonical Splice Motif Breakdown

A finer dissection by motif class – using the genomic dinucleotides immediately downstream of the donor and upstream of the acceptor — reveals an unexpected ordering (Table A11): *non-canonical* splice donors converge *earlier* than the dominant canonical GT-AG donors, while the minor canonical GC-AG class is the deepest of the three. The shallower-than-intron pattern ($\bar{c} = 27.72$) holds for every splice class, but the within-class ordering is the opposite of a naive "stronger motif $\Rightarrow$ shallower recognition" prior. The Cohen's d between canonical GT-AG and non-canonical donors is small (≈ 0.05), consistent with the constrained branch-point and polypyrimidine context that flanks non-canonical introns (Wang & Burge,

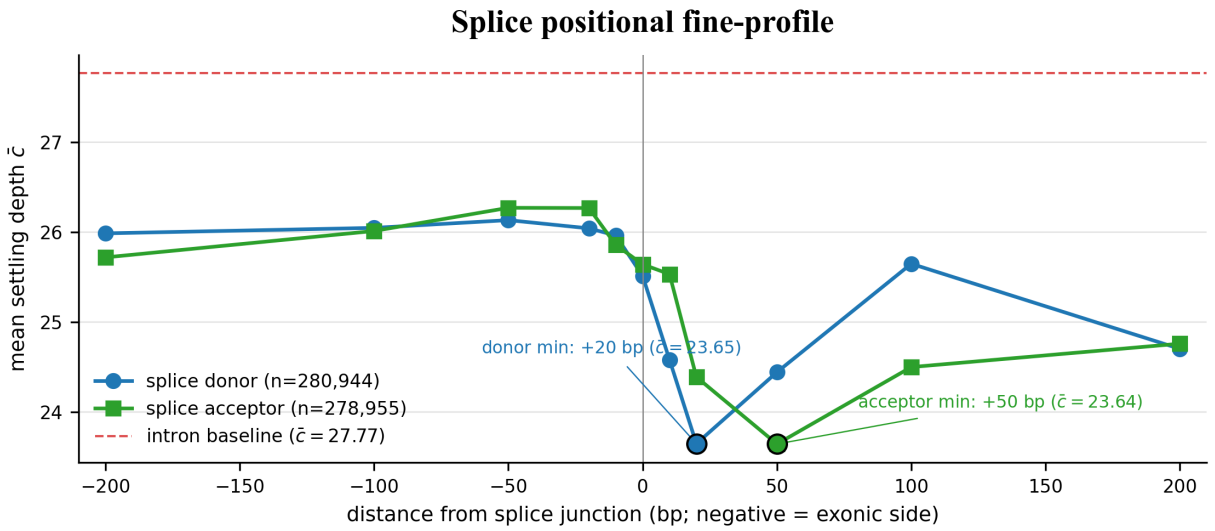

Figure A4. Mean settling depth $\bar{c}$ as a function of distance to the nearest splice donor / acceptor, computed on the pooled chr17 + chr22 analysis windows. Both profiles dip well below the intronic baseline ($\bar{c}=27.72$, red dashed) within ± 200 bp; donors minimise on the exonic side, acceptors on the intronic side, mirroring the asymmetric splice-grammar context (branch point and polypyrimidine tract on the intronic side of acceptors).

2008; Burge et al., 1998); we report the magnitudes honestly rather than over-interpret. The finding replicates qualitatively on the 8-layer HyenaDNA-large at depth-normalised scale.

Table A11. Within-splice motif breakdown by donor dinucleotide. The shallower-than-intron pattern ($\bar{c}=27.72$) holds for every splice class, but the within-class ordering is the opposite of a naive "stronger motif ⇒ shallower recognition" prior: non-canonical donors converge earliest, GC-AG latest.

Donor motif class	n	$\bar{c}$	vs. canonical GT-AG
non-canonical (AT-AC, other)	5,977	25.13	$d \approx -0.05$ (shallower)
canonical GT-AG (dominant)	350,552	25.79	— (reference)
canonical GC-AG (minor)	5,165	27.01	$d \approx +0.10$ (deeper)
intron baseline	$\sim 10^7$	27.72	— (deeper than every class)

D.3. Motif and Flank Perturbation Summary

Table A12 summarises the motif-edit and flank-shuffle controls of §3.2. On 1,000 canonical GT-AG donors (chr22, ± 10 bp pad-averaged c) the central GT motif contributes a small but reliable detection signal ($\bar{c}$ deepens by 0.46 layers when GT is replaced with AA), whereas dinucleotide-shuffling the ± 100 bp flank lifts $\bar{c}$ to a shallower value by 3.18 layers, isolating the GT for easier stabilisation. Real flanking grammar therefore demands deeper integration than the central motif on its own.

Table A12. Motif-edit and flank-shuffle controls on 1,000 canonical GT-AG donors. Real $\bar{c}=26.77$. Replacing the central GT with AA (while keeping the flank) deepens $\bar{c}$; shuffling the flank (while keeping GT) *lifts* $\bar{c}$ to a shallower value by 3.18 layers.

Condition	$\bar{c}$	$\Delta\bar{c}$ vs. real	Cohen's d	paired Wilcoxon p
real GT-AG donor (reference)	26.77	0 (ref.)	—	—
GT → AA, flank preserved	27.24	+0.46	−0.086	2.3×10^{-32}
GT preserved, ± 100 bp flank shuffled	23.59	**−3.18**	**+0.515**	4.1×10^{-59}

E. Reproducibility

All experiments use random seed 42 for cross-validation splits and bootstrap resamples. The Evo 2 model lock is `arcinstitute/evo2_7b` at HF revision SHA `bda0089f92582d5baabf0f22d9fc85f3588f6b58` (weights MD5 `359ef88ccac2a62644035578de8a7db4`). Data versions: GRCh38 primary assembly (UCSC, MD5 locked); GENCODE v44 GTF (per-chromosome filtered, persisted as `gffutils` SQLite); PhyloP 100-

way (UCSC); ENCODE SCREEN v3 cCRE catalog with ELS subset; RepeatMasker hg38; GTEx v8 cis-eQTL pairs (`Whole_Blood`, `Brain_Cortex`, `Liver`, `Lung` unioned); GWAS Catalog v1.0; ClinVar 2026-04-18. Software stack: `torch 2.4.1+cu124`, `evo2 0.3.0`, `vortex 1.0.8`, `transformer-engine 2.14.0`, `transformers 4.49.0`, `scipy 1.13`, `scikit-learn 1.4`. Hardware: a single NVIDIA H200 (141 GB). Total compute: $\sim$ 20 GPU-hours end-to-end. Code, dataset version locks, and figure-generation scripts will be released at an anonymous URL upon acceptance (an anonymized repository is provided as supplementary material for review).