# Representative vs. Load-bearing Layers: A Dissociation in Genomic Foundation Models

**Anonymous Authors**[1]

## Abstract

Genomic foundation models are typically evaluated by aggregating representations across layers (Pearce et al., 2026) or defaulting to the last layer (Dalla-Torre et al., 2024); neither asks which layer the joint model actually relies on. We probe this with a minimal training-free scalar — the L2 norm of the per-layer hidden-state shift at the variant token, $\|\Delta h_\ell\|_2$ — on 8,008 ClinVar (Landrum et al., 2018) single-nucleotide variants in NT-v2 500M (Dalla-Torre et al., 2024) (a masked language model, MLM) and Evo 2 7B (Brixi et al., 2026) (a causal language model, CLM, with a hyena/attention hybrid). The layer with peak single-feature AUROC (*representative*) is not the layer a joint multi-layer classifier most depends on (*load-bearing*, identified by leave-one-layer-out ablation drop, concordant with |SHAP| (Lundberg & Lee, 2017)). Representative layers sit mid-network in both models, while load-bearing depth lies at opposite ends of the depth axis — mid-shallow in the MLM, deep in the CLM hybrid. The dissociation has direct downstream consequences: in NT-v2 a 1-dim mid-layer scalar exceeds the canonical 1024-dim last-layer mean-pool baseline by $+0.049$ AUROC, whereas in Evo 2 the 4096-dim mean-pool is competitive with the joint $\|\Delta h_\ell\|_2$ feature. Standard last-layer pooling therefore leaves variant-relevant signal untapped specifically in MLM-based pipelines.

## 1. Introduction

Layer-wise analysis of genomic foundation models (FMs) has so far split between two camps. Aggregation methods pool features across all layers (Pearce et al., 2026), treating layer identity as a nuisance dimension; mechanistic analyses pick a single intermediate layer based on the density of biologically meaningful features (Brixi et al., 2026; Templeton et al., 2024); standard downstream pipelines, meanwhile, default to the last layer (Dalla-Torre et al., 2024). None of these asks the question that separates them: *which layer does the joint model actually rely on, and is it the same layer a univariate probe identifies as carrying the strongest signal?*

In this paper we show that, in two genomic FMs, it is not. The layer maximising single-feature variant discriminability is not the layer the joint multi-layer classifier most depends on, in both NT-v2 (a masked language model, MLM) and Evo 2 (a causal language model, CLM, with a hyena/attention hybrid backbone). We measure this with a minimal training-free probe and two operational metrics introduced in §2: the per-layer single-feature AUROC (*representative* — where does the signal live?) and the leave-one-layer-out ablation drop on a joint multi-layer classifier (*load-bearing* — what does the joint model rely on?), with |SHAP| (Lundberg & Lee, 2017) as a concordant attribution.

**Contributions.**

1. **Representative–load-bearing dissociation** in two genomic FMs: the per-layer AUROC peak and the joint-classifier load-bearing layer are operationally distinct (§3–§4).
2. **Cross-architectural depth shift, observational**: load-bearing depth differs by $\sim 20$ layers between the MLM (mid-shallow) and the CLM hybrid (deep), while representative depth stays mid-network in both — although objective, capacity, and architecture covary in our setup.
3. **Standard last-layer protocol is suboptimal in MLMs**: a 1-dim mid-layer scalar exceeds the canonical 1024-dim mean-pool baseline by $+0.049$ AUROC in NT-v2, whereas in Evo 2 the canonical mean-pool itself is competitive with the joint $\|\Delta h_\ell\|_2$ feature (App. H).

Our scope is narrow — one task, two main models with one weak control, linear readouts — but the dissociation replicates across four scalar reductions of $\Delta h_\ell$ (App. J).

[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.

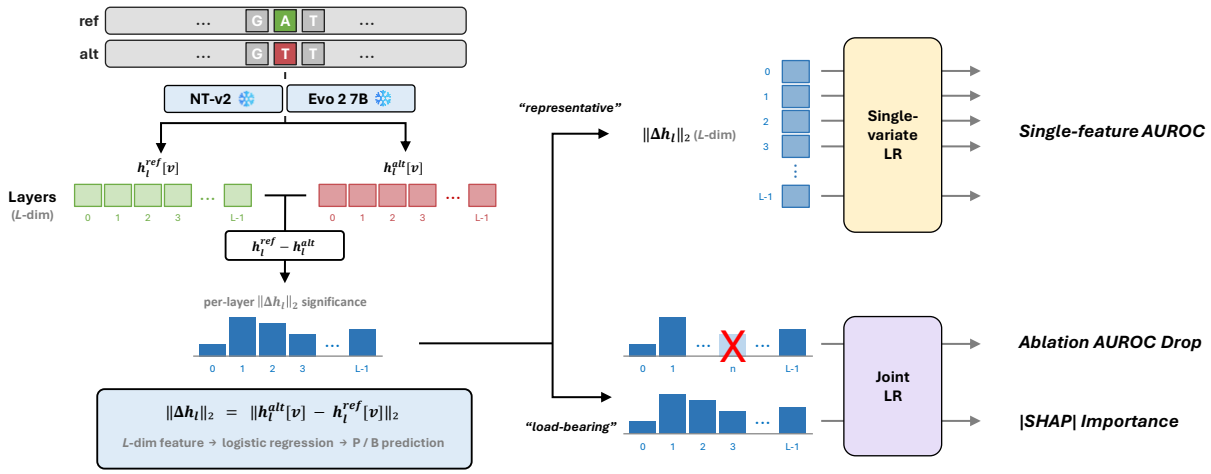


*Figure 1.* **Method overview and the two metric axes. Left:** matched ref/alt forward passes through a frozen genomic FM (NT-v2 or Evo 2). At the variant-token position $v$ we compute $\|\Delta h_\ell\|_2$, the L2 norm of the per-layer hidden-state difference; stacking across layers gives an $L$-dim feature ($L$=30 for NT-v2, $L$=32 for Evo 2; index $\ell$=0, . . . , $L$−1). **Right:** the two operational axes. The *representative* axis (top) reports per-layer single-feature AUROC; the *load-bearing* axis (bottom) reports leave-one-layer-out ablation drop (X marks the dropped layer) and |SHAP| on the joint classifier (§3–§4). The two axes need not select the same layer.

**Related work.** Layer-wise probing has a long tradition in NLP — logit lens (nostalgebraist, 2020), TunedLens (Belrose et al., 2023), BERT-depth analysis (Tenney et al., 2019) — with recent evidence that last-layer features are systematically suboptimal for downstream tasks (Skean et al., 2025). In genomic FMs, prior work either pools across layers (Pearce et al., 2026) or selects single intermediate layers via mechanistic-interpretability tooling (Brixi et al., 2026; Templeton et al., 2024). We instead ask which layer is most *relied upon*, and characterise it as operationally distinct from the most *decodable* layer.

## 2. Probe and setup

For each variant we run two forward passes through a frozen genomic foundation model — one on the reference sequence, one on the alternate (Fig. 1). With $h_\ell[v] \in \mathbb{R}^{d_\ell}$ the hidden state at layer $\ell$ and the variant-token position $v$, the probe is

$$\|\Delta h_\ell\|_2 \;\coloneqq\; \left\|h_\ell^{\mathrm{alt}}[v] - h_\ell^{\mathrm{ref}}[v]\right\|_2, \quad \ell = 0, \ldots, L-1, \;\; (1)$$

a minimal, training-free, rotation-invariant, head-decoupled scalar. Three alternative reductions (cosine, L1, directional projection) yield the same within-model dissociation (App. J); construction details are in App. A.

**Two operational definitions.** We define the **representative** layer as the per-layer single-feature AUROC peak — *"where does the signal live?"* — and the **load-bearing** layer as the largest leave-one-layer-out ablation drop on the joint multi-layer classifier, concordant with |SHAP| (Lundberg & Lee, 2017) — *"what does the joint model rely on?"* A single layer cannot answer both questions, and we show in §3–§4 that the two definitions identify operationally distinct layers in both architectures.

We use 8,008 ClinVar (Landrum et al., 2018) single-nucleotide variants across 15 high-penetrance cancer-associated genes (Pathogenic / Likely Pathogenic, P/LP, $n$=3,514; Benign / Likely Benign, B/LB, $n$=4,494; App. F). Main models: **NT-v2 500M** (Dalla-Torre et al., 2024) (MLM, 29 blocks) and **Evo 2 7B** (Brixi et al., 2026) (CLM, 32 hybrid hyena/attention blocks); a weak-CLM control, HyenaDNA-large (Nguyen et al., 2023), is in App. M. All models are frozen. Per-layer and joint AUROCs use stratified 10-fold logistic regression with no tuning; bootstrap CIs, DeLong tests, and three confound controls (substitution-stratified AUROC, within-gene label permutation, gene-balanced subsampling) all pass (App. C–D).

## 3. Decodability does not equal functional reliance

**Per-layer AUROC profile.** For each layer we fit a stratified 10-fold logistic regression on the single feature $\|\Delta h_\ell\|_2$ and report out-of-fold AUROC. The profile (Fig. 2) is consistent across architectures: a sharp rise out of the input embedding, a broad mid-network plateau, and decay or collapse at the final layer. NT-v2 plateaus over L9–L26 ($\geq$ 0.90), peaks at $\|\Delta h_{15}\|_2$=0.930, and collapses near chance level at the final block ($\|\Delta h_{29}\|_2$=0.550; 5 of 14 genes below chance under leave-one-gene-out; App. F). Evo 2 peaks at $\|\Delta h_8\|_2$=0.855 and decays more gently to $\|\Delta h_{31}\|_2$=0.750 ($\Delta$=+0.105, no chance-level collapse).

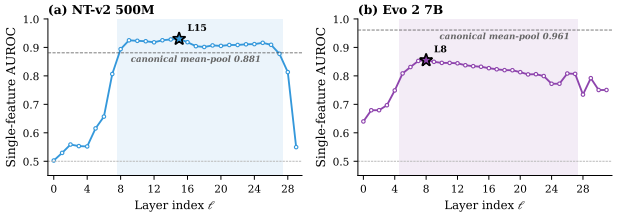


*Figure 2.* **Per-layer single-feature AUROC.** Out-of-fold AUROC of a 10-fold logistic regression on the single feature $\|\Delta h_\ell\|_2$. The filled star marks the representative layer (univariate peak); the grey dashed line marks the canonical last-layer mean-pool baseline (App. G–H); the lower grey dotted line at 0.5 marks chance. **(a)** NT-v2 500M (MLM): peak L15 (0.930) exceeds the 1024-dim mean-pool baseline (0.881); the L9–L26 plateau ($\geq$ 0.90) begins at the load-bearing L9 (Fig. 3). **(b)** Evo 2 7B (CLM hybrid): peak L8 (0.855) sits below the 4096-dim mean-pool baseline (0.961); the panel shows the within-architecture peak-vs-last gap, not a mean-pool deficiency.

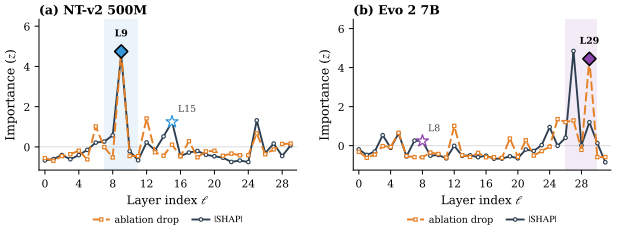


*Figure 3.* **Per-layer joint-classifier importance.** Tree-SHAP mean |SHAP| (navy solid) and leave-one-layer-out ablation drop (orange dashed), $z$-scored within architecture. The open star marks the representative layer (Fig. 2); the filled diamond marks the load-bearing layer (largest ablation drop); the pink band highlights the load-bearing region. **(a)** NT-v2 500M: load-bearing L9 (mid-shallow, L7–L11) vs. representative L15. **(b)** Evo 2 7B: load-bearing L29 (deep, L26–L30) vs. representative L8; |SHAP| peaks at L27 in the same band.

We interpret the NT-v2 collapse (§5) as a combination of final-layer basis rotation and MLM commitment to the reconstruction subspace. Both peak-vs-last gaps are highly significant ($p<10^{-50}$ paired DeLong; 100% of paired bootstrap resamples; App. C).

**Comparison with the canonical last-layer mean-pool baseline.** Genomic-FM downstream evaluations typically read out the last hidden state via window mean-pool. We compute this baseline on the same split for both architectures (full Evo 2 numbers in App. H, Tab. A3); the two models behave qualitatively differently:

**NT-v2 (MLM).** The canonical 1024-d mean-pool last-layer baseline reaches 0.881 AUROC (Fig. 2a, grey dashed). The 1-d mid-layer scalar $\|\Delta h_{15}\|_2$=0.930 exceeds this 1024-d protocol by $+\mathbf{0.049}$ at a 1024$\times$ feature-dimensionality reduction.

**Evo 2 (CLM hybrid).** The canonical 4096-d mean-pool last-layer baseline reaches 0.961 AUROC — comparable to (and slightly above) the joint 32-d $\|\Delta h_\ell\|_2$ feature (0.926). For Evo 2 the canonical mean-pool readout is therefore not impaired; the corresponding fine-tune-free single-scalar comparison ($\|\Delta h_8\|_2$=0.855 vs. $\|\Delta h_{31}\|_2$=0.750, $\Delta$=+0.105) shows the within-architecture peak-vs-last gap, not a mean-pool deficiency.

The asymmetry — the +0.049 win is MLM-specific — is consistent with NT-v2's near-chance final-layer collapse driving the canonical baseline down (§5). The representative-layer scalar already exceeds the canonical baseline in NT-v2 with a single dimension, whereas no single-feature scalar matches the canonical mean-pool baseline in Evo 2; §4 shows that this representative layer is, in both architectures, *not* the layer the joint multi-layer classifier most relies on. Stacking $\|\Delta h_\ell\|_2$ across all $L$ layers raises the joint AUROC to 0.962/0.926 (NT-v2 / Evo 2), within $\sim$ 0.04 of the cross-layer covariance aggregator EVEE (Pearce et al., 2026) (0.997 on a different split) at the cost of just one scalar per layer.

## 4. Joint models depend on different layers than probes suggest

When the joint multi-layer $\|\Delta h_\ell\|_2$ feature (30-dim for NT-v2, 32-dim for Evo 2) is fit by a logistic-regression classifier, which layer does the classifier most rely on? We measure this with two operationally independent attributions, both defined in §2: leave-one-layer-out ablation drop — the direct counterfactual on out-of-fold AUROC when $\|\Delta h_\ell\|_2$ is removed and the classifier is refit on the remaining $L-1$ layers — and tree-SHAP (Lundberg & Lee, 2017), a feature-attribution approximation computed on a LightGBM surrogate of the joint classifier (App. A). The load-bearing layer is the largest ablation drop; |SHAP| serves as a concordant attribution.

For NT-v2 the load-bearing layer is L9: both metrics rank L9 first (ablation drop +0.0027, |SHAP| 2.81, joint LR coefficient +4.21). The representative layer L15, although the strongest single-feature predictor at 0.930, is not the layer the classifier most depends on (ablation drop only +0.0001). For Evo 2 the load-bearing layer is L29: leave-one-out ablation drops AUROC by +0.0069 ($\sim$ 3$\times$ the runner-up). |SHAP| peaks at L27 (2.64 vs. 0.95 at L29); both layers lie in the deep last-five-layer band, where the ablation drop is also concentrated — consistent across the two attribution schemes that load-bearing in this CLM hybrid lies near the output, far from the representative layer L8 (Fig. 3b). Pairwise Spearman between the two metrics across all layers is 0.78 for NT-v2 and 0.65 for Evo 2 — moderate-to-strong agreement. Several layers enter with negative joint-LR coefficients despite high single-feature AUROC (App. K), a

suppressor pattern we leave for future analysis.

### 4.1. Cross-architectural pattern

Representative depth is similar across architectures (mid-network: relative depth $0.52$ NT-v2, $0.25$ Evo 2). Load-bearing depth, by contrast, shifts by $\sim 20$ layers in opposite directions: mid-shallow in NT-v2 (MLM; $0.31$, ahead of the representative layer) and deep in Evo 2 (CLM hybrid; $0.91$, near the last block). The two architectures differ in many factors (objective, capacity, sequence-mixing primitive, data); the within-model dissociation holds in both, and the direction of the cross-architectural shift is consistent with training-objective-dependent organisation of functional variance across depth, although confounded with the other factors (§5).

## 5. Discussion

**Why decodability and reliance dissociate.** A layer is *representative* when its signal alone discriminates the task; *load-bearing* when removing it from the joint feature drops classifier AUROC the most. The representative layers (Evo 2 L8, NT-v2 L15) are strong individually yet dispensable under leave-one-layer-out ablation once the joint classifier is refit on the remaining layers. The load-bearing layers (Evo 2 L29, NT-v2 L9) carry contributions the joint classifier weights heavily and that ablation cannot replace — the operational signature of *reliance*.

**Cross-architectural shift consistent with training-objective effects.** Representative depths fall within the mid-network band in both architectures (relative depths $0.52$ for NT-v2, $0.25$ for Evo 2); load-bearing depths, in contrast, lie at opposite ends of the depth axis (relative depths $0.31$ and $0.91$). A capacity-limited weak-CLM control (HyenaDNA-large, $\sim$ 6.6M params) collapses load-bearing to the input embedding (Table 1; App. M). The dissociation is therefore consistent across all three points, but load-bearing *depth* is shaped jointly by training objective and capacity. The shift direction (MLM mid-shallow, CLM deep) is consistent with each objective's gradient-pressure profile: MLM applies loss only at masked positions via a final-layer vocab head, committing the final layer to a reconstruction subspace and shielding task-relevant variance at earlier layers; CLM applies loss at every position, leaving deep layers under task-relevant pressure. *We treat this only as an observation: the observed shift in load-bearing depth is consistent with differences in training objective (MLM vs. CLM), though objective, capacity, architecture, and data are confounded in our setup.* A controlled same-architecture comparison (varying only the objective) would be needed for causal attribution; predicted directions of movement and the full mechanism sketch are in App. L.

*Table 1.* Three-architecture cross-comparison. Within-model dissociation holds in all three; load-bearing depth is shaped jointly by objective and capacity. HyenaDNA's L0 load-bearing reflects a capacity-limited regime where the model relies on substitution-bias carried at the input embedding (App. M).

| Model | Obj. | Cap. | Repr./LB | Joint |
|---|---|---|---|---|
| NT-v2 500M | MLM | 500M | L15/L9 | 0.962 |
| Evo 2 7B | CLM | 7B | L8/L29 | 0.926 |
| HyenaDNA | CLM | $\sim$6.6M | L4/L0 | 0.673 |

**Implications.**

- **Last-layer pooling under-uses MLM representations.** A 1-dim mid-layer scalar exceeds the canonical 1024-dim mean-pool in NT-v2 ($+0.049$ AUROC); for Evo 2 the canonical mean-pool already matches the joint $\|\Delta h_\ell\|_2$ feature.
- **Single-feature AUROC alone mischaracterises layer importance.** It is the easiest to compute and the least informative about joint-model reliance; load-bearing requires the joint classifier in the loop.
- **Pretraining shapes functional depth.** The cross-architectural shift is consistent with training-objective effects on load-bearing-layer location, with capacity as a second factor (HyenaDNA control, App. M).

**Limitations.** Pathogenicity labels are a damaging-vs-tolerant proxy in high-penetrance genes; cross-architectural inference rests on $n$=2 strong models plus one weak control; readouts are linear; objective and capacity are not isolated here. None of these affects the within-model dissociation, which is identified by leave-one-layer-out ablation and concordant with $|\mathrm{SHAP}|$ in both strong architectures and replicated under four scalar reductions of $\Delta h_\ell$ (App. J). A controlled same-architecture, same-data MLM-vs-CLM comparison would be needed to attribute the depth shift to objective alone; existing public checkpoints do not provide such a setup, and we therefore treat the cross-architectural shift as observational evidence rather than a causal claim (App. L).

**Closing.** The result is small in scope — one probe, two main models, one task — but the observation is structural: in genomic FMs, what is most decodable is not what is most used. Treating the two as identical, as standard last-layer probing protocols implicitly do, consistently mischaracterises which layers a model relies on, and — in MLMs at least — leaves variant-relevant signal at mid layers that downstream pipelines never read.

## References


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.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Brixi, G., Durrant, M. G., Ku, J., et al. Genome modelling and design across all domains of life with Evo 2. *Nature*, 2026. doi: 10.1038/s41586-026-10176-5.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Henryk Grzywaczewski, A., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *Nature Methods*, 2024. doi: 10.1038/s41592-024-02523-z.

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.

Ethayarajh, K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 55–65, 2019.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 2018. doi: 10.1093/nar/gkx1153.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

Mann, H. B. and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1): 50–60, 1947.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Birch-Sykes, C., Wornow, M., Patel, A., Rabideau, C., Massaroli, S., Bengio, Y., Ermon, S., Ré, C., and Baccus, S. HyenaDNA: Long-range genomic sequence modelling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*, 2023.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Pearce, M., Dooms, T., Yamamoto, R., Meehl, J., Molnar, C., Bissell, M., Hazra, D., Fang, C., Nguyen, N., Anderson, M., Osborne, C., Duffy, P., Toomey, B., Klee, E., Myasoedova, E., Ryu, A., Ayanian, S., Korfiatis, P., Redlon, M., Jain, A., Balsam, D., and Wang, N. EVEE: Interpretable variant effect prediction from genomic foundation model embeddings. *bioRxiv*, 2026. doi: 10.64898/2026.04.10.717844. Posted 2026-04-14; concurrent work.

Skean, O. et al. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint*, 2025.

Templeton, A. et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Anthropic Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Tenney, I., Das, D., and Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019.

Wu, J. et al. Anisotropy and last-layer geometry in masked language models. *arXiv preprint*, 2023.

# Supplementary material

## A. Probe construction details

**Window and variant-token offset.** We use a 6 kb window centred on the variant. For Evo 2 (single-base byte tokenization), the variant token index is fixed at $v = 3000$ (window midpoint). For NT-v2 (6-mer tokenization), $v$ is the first token where the reference and alternate token sequences differ; this is the 6-mer that contains the variant. We restrict the analysis to single-nucleotide variants so that the variant token is unambiguous in both architectures.

**Layer indexing.** Evo 2 7B has 32 transformer blocks, exposed by the HuggingFace wrapper as 33 hidden states (input embedding + 32 block outputs); we index $\ell \in \{0, \ldots, 31\}$ over the 32 post-block hidden states (matching the convention in Brixi et al. (2026)). NT-v2 has 29 transformer blocks; the wrapper exposes 30 hidden states; we index $\ell \in \{0, \ldots, 29\}$ inclusive of the input embedding (so $\ell = 29$ is the final block output). HyenaDNA-large has 8 transformer blocks (10 hidden states with input embedding); $\ell \in \{0, \ldots, 9\}$.

**Numerical precision.** Evo 2 forwards run at `bfloat16`; NT-v2 forwards run at `float32` (a known dtype-mismatch issue in the NT-v2 wrapper at `bfloat16`). Hidden states are converted to fp32 before computing $\|\Delta h_\ell\|_2$. We verify on a sample of 200 variants that the relative precision error in $\|\Delta h_\ell\|_2$ is below $10^{-3}$ per layer, well below the cross-fold AUROC noise floor.

**Implementation.** All extractions and analyses use a fixed `Pipeline([StandardScaler, LogisticRegression(C=1.0, solver=lbfgs, max_iter=1000)])` with seed 42; we do not tune. Tree models (used for SHAP) are LightGBM with 500 estimators, max-depth 6, seed 42. Reproducibility scripts are listed in App. N.

## B. Per-layer univariate signal: Mann–Whitney $U$ and Cohen's $d$

To map per-layer signal at the univariate level we test $\|\Delta h_\ell\|_2$ between P/LP and B/LB variants per layer using the Mann–Whitney $U$ test (Mann & Whitney, 1947) with Benjamini–Hochberg FDR correction (Benjamini & Hochberg, 1995) across the 32 (Evo 2) or 30 (NT-v2) layers (Fig. A1). All Evo 2 layers reject the null at FDR $p < 0.05$; $|d|$ ranges from 0.42 (L0) to 1.50 (L8). For NT-v2 all layers except the input embedding L0 reject ($|d|$ at L0 $= 0.03$, $p = 0.29$): unlike Evo 2's single-byte tokenization, the NT-v2 6-mer embedding-norm difference at the variant-token position is balanced across P/LP and B/LB at the input layer. Across the remaining 29 NT-v2 layers $|d|$ ranges from 0.71 (L19) to 2.03 (L10), with the maximum sitting inside the L9–L11 plateau also identified by single-feature AUROC; the small offset between the AUROC peak (L15) and the Cohen's $d$ peak (L10) reflects that $d$ depends on the marginal mean shift while AUROC additionally rewards monotone separability of the score. The final layer L29 rejects with a small, sign-flipped effect ($d = -0.15$, $p = 1.6 \times 10^{-14}$) consistent with the near-chance single-feature AUROC 0.55 reported in §3: a small non-task-aligned residual that is statistically detectable at $n = 8{,}008$ but not classifier-useful (and reverses sign in the basis-rotated subspace).

## C. Bootstrap and DeLong tests

We run paired bootstrap (B=1000) on the variant axis, holding the cross-validation fold structure fixed; for each draw we resample variants with replacement and recompute the AUROC of the representative- and last-layer single features and their difference. We also run paired DeLong tests (DeLong et al., 1988) on the out-of-fold scores.

*Table A1.* Bootstrap CI (B=1000, paired) and paired DeLong test for the representative-vs-last single-feature AUROC gap.

| Model | Repr. AUROC | Last AUROC | $\Delta$ pt. | 95% CI of $\Delta$ | DeLong $z$ |
|---|---|---|---|---|---|
| Evo 2 7B | 0.855 | 0.750 | +0.105 | [+0.093, +0.117] | 17.1 |
| NT-v2 500M | 0.930 | 0.550 | +0.380 | [+0.366, +0.394] | 52.6 |

In both architectures, 100% of bootstrap resamples show the representative AUROC strictly above the last AUROC.

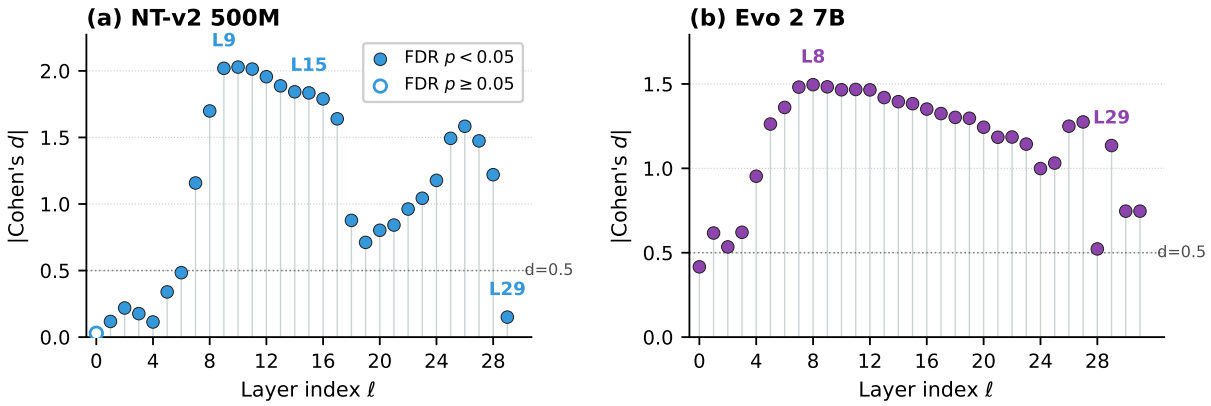


*Figure A1.* Per-layer |Cohen's $d$| for $\|\Delta h_\ell\|_2$ between P/LP and B/LB. Filled markers reject the null at FDR $p < 0.05$; open markers fail to reject. **(a)** NT-v2: L0 fails; the L9–L11 plateau peaks at $|d| \approx 2.0$; L29 has $|d| = 0.15$ with sign flip (still rejects at this $n$). **(b)** Evo 2: all 32 layers reject; $|d|$ peaks at L8.

## D. Confound controls

We pre-register three confound checks on the cached $\|\Delta h_\ell\|_2$ features and run them on both Evo 2 and NT-v2.

**Substitution-stratified.** For each of the 12 ordered (ref, alt) substitutions, we restrict the dataset to that substitution and recompute the joint $L$-dim AUROC. All 12 substitutions retain joint AUROC $> 0.85$ for Evo 2 (range 0.852–0.951) and $> 0.93$ for NT-v2 (range 0.933–0.971). Thus mutational-signature differences between P/LP and B/LB do not explain the headline AUROC.

**Within-gene label permutation.** We permute P/LP–B/LB labels within each of the 15 cancer genes (preserving per-gene base rates) and recompute peak-layer single AUROC and joint AUROC; 1,000 permutations for the single-feature null and 200 for the joint null (joint requires full 10-fold CV). Real performance vastly exceeds the 99th percentile of the permutation null in both architectures: e.g., NT-v2 real joint 0.962 vs. permutation 99th percentile 0.544.

**Gene-balanced sub-sampling.** For each gene with $\min(n_{\mathrm{P/LP}}, n_{\mathrm{B/LB}}) \geq 30$, we down-sample to a class-balanced subset and re-run the joint pipeline (50 trials per architecture). Both peak-layer and joint AUROCs drop by less than 0.01 in all three architectures (Evo 2 $\Delta = -0.006$; NT-v2 $\Delta = -0.002$; HyenaDNA $\Delta = -0.002$), ruling out per-gene class imbalance as the dominant signal source.

## E. L0 substitution-bias decomposition (Evo 2)

For Evo 2's byte-level tokenization, the L0 perturbation $\|\Delta h_0\|_2$ at the variant token depends only on the embedding-vector difference between the two single-base tokens; it reflects substitution-frequency biases by construction, with no biological signal beyond it. We decompose the L0 single-feature AUROC (0.640) into the part attributable to substitution-frequency differences alone (a 12-class lookup of P/LP rate per substitution: 0.704 AUROC) and the within-substitution residual (weighted-mean AUROC across the 12 substitutions: 0.541, near chance). Thus the L0 signal is fully explained by mutational-signature bias, with no biological signal beyond it. The mid-layer signal (L8 onwards), in contrast, is preserved within every substitution category (Appendix D), establishing that the layer-level findings are not artefacts of substitution distribution.

## F. Dataset construction and per-gene LOGO drill-down (NT-v2)

**Gene list and counts.** The 15 genes are *BRCA1*, *BRCA2*, *TP53*, *EGFR*, *KRAS*, *BRAF*, *PIK3CA*, *APC*, *MLH1*, *MSH2*, *PTEN*, *RB1*, *VHL*, *ATM*, *PALB2* — standard hereditary-cancer-panel genes with $\geq 92$ variants per gene with both classes represented. No gene-level filtering was performed based on AUROC.

**Why high-penetrance.** Condition-agnostic and condition-aware ClinVar labels closely agree on high-penetrance cancer-gene variants; the restriction therefore holds label noise low so that observed layer-level differences reflect representational structure rather than label-aggregation artefacts. AUROC on this benchmark is a damaging-vs-tolerant proxy and not a clinical-pathogenicity number at population scale (cf. Limitations, §5).

*Table A2.* Per-gene leave-one-gene-out AUROC for NT-v2: representative L15 single feature, last L29 single feature, and joint 30-d feature. The last-layer scalar is below chance for several genes (**bold**); the representative-layer scalar is at least 0.77 everywhere.

| Gene | Repr. L15 | Last L29 | Joint 30-d |
|---|---|---|---|
| APC | 0.952 | 0.624 | 0.974 |
| ATM | 0.972 | 0.443 | 0.984 |
| BRAF | 0.775 | 0.724 | 0.927 |
| BRCA1 | 0.890 | 0.604 | 0.952 |
| BRCA2 | 0.930 | 0.739 | 0.961 |
| KRAS | 0.789 | 0.715 | 0.924 |
| MLH1 | 0.923 | 0.490 | 0.941 |
| MSH2 | 0.974 | 0.465 | 0.992 |
| PALB2 | 0.968 | 0.645 | 0.955 |
| PIK3CA | 0.836 | 0.416 | 0.881 |
| PTEN | 0.902 | 0.558 | 0.951 |
| RB1 | 0.930 | 0.406 | 0.951 |
| TP53 | 0.903 | 0.557 | 0.923 |
| VHL | 0.882 | 0.509 | 0.953 |

The last layer of NT-v2 is below chance for 5 of 14 evaluable genes (EGFR has only B/LB variants and is excluded from LOGO), reinforcing the basis-rotation interpretation in §5.

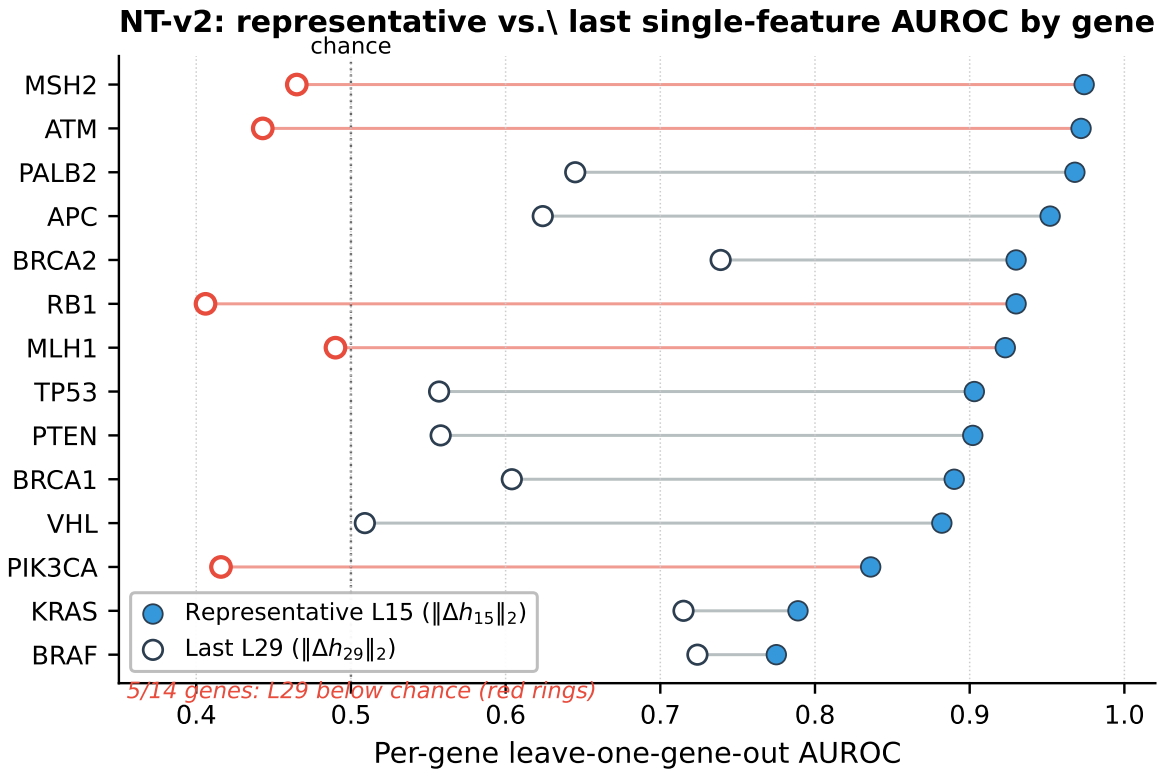


*Figure A2.* Per-gene LOGO single-feature AUROC for NT-v2: representative L15 (filled blue) vs. last L29 (open). Genes are sorted by L15 AUROC. Red rings mark the 5 of 14 genes where L29 falls below chance (0.5, dotted). The L29 collapse is gene-pervasive rather than driven by one or two outlier genes, supporting the basis-rotation account of §5.

## G. NT-v2 mean-pool baseline (controlled comparison)

We extract the last-layer ($\ell = 29$) hidden states for the reference and alternate sequences for all $8{,}008$ NT-v2 variants and pool by mean over the $\sim 1000$ token positions, yielding a 1024-d vector per side. The difference vector (alt $-$ ref) is fit by the same logistic-regression pipeline as the rest of the paper. Stratified 10-fold AUROC: $0.881$ (LOGO $0.833$). This is the canonical "last hidden state mean-pool" baseline standard in genomic-FM downstream evaluations (Dalla-Torre et al., 2024); the 1-d mid-layer scalar baseline ($\|\Delta h_{15}\|_2 = 0.930$) exceeds it by $+0.049$ at a $1024\times$ reduction in feature dimensionality.

## H. Evo 2 mean-pool baseline

We ran the same mean-pool extraction for Evo 2: last-layer ($\ell$=31) hidden states for ref and alt over the same window, yielding a 4096-d difference vector per variant (10,910 variants extracted; 8,008 P/LP $\cup$ B/LB used). Table A3 shows the full set of single-scalar and 4096-d baselines under the same logistic-regression pipeline as the rest of the paper.

*Table A3.* Evo 2 7B baseline AUROCs on the 8,008-variant split (stratified 10-fold pooled AUROC; LOGO in parentheses).

| Baseline | Dim | AUROC (10-f \| LOGO) |
|---|---|---|
| Joint $\|\Delta h_\ell\|_2$ (paper) | 32 | $0.926$ $(0.919)$ |
| Single-scalar $\|\Delta h_8\|_2$ (peak) | 1 | $0.855$ $(0.848)$ |
| Single-scalar $\|\Delta h_{31}\|_2$ (last) | 1 | $0.750$ $(0.739)$ |
| Mean-pool last (canonical) | 4096 | $\mathbf{0.961}$ $(\mathbf{0.956})$ |
| Mean-pool peak ($\ell$=8) | 4096 | $0.947$ $(0.856)$ |
| Variant-token last (vector) | 4096 | $0.824$ $(0.807)$ |
| Variant-token peak (vector) | 4096 | $0.770$ $(0.708)$ |

**Implication for the +0.049 NT-v2 result in the main text.** The canonical Evo 2 mean-pool baseline ($0.961$) is comparable to (slightly above) the joint 32-d $\|\Delta h_\ell\|_2$ feature ($0.926$), in clear contrast to NT-v2 where mean-pool ($0.881$) is exceeded by the 1-d mid-layer scalar ($0.930$). The +0.049 win is therefore MLM-specific. This is consistent with the architecture-dependent final-layer behaviour developed in §5: NT-v2's final layer collapses near chance (basis rotation + reconstruction-subspace commitment), degrading the mean-pool readout, whereas Evo 2's final layer remains task-informative and the mean-pool readout works as expected. The dissociation between representative and load-bearing layers, by contrast, is observed in both architectures regardless of whether the canonical mean-pool readout is competitive.

Numerical results are saved at `results/tier1_last_layer/evo2/baselines_auroc.csv` and `baselines_auroc.json` for reproducibility; reproduction recipe is in App. N.

## I. Multi-layer feature: $\|\Delta h_\ell\|_2$ stacked across all layers

Stacking $\|\Delta h_\ell\|_2$ across all layers and fitting a single logistic regression on the resulting $L$-dim feature vector gives:

- Evo 2 (32-d): stratified 10-fold AUROC $0.926$ (gain over single-feature peak $0.855$: $+0.071$).
- NT-v2 (30-d): stratified 10-fold AUROC $0.962$ (gain over $0.930$: $+0.032$).

The gain over the representative-layer scalar is precisely the contribution of the layers carrying task-relevant unique variance — principally the load-bearing layers (Evo 2 L29, NT-v2 L9) and a small set of supporting layers identified by the joint LR coefficients (Evo 2 L12 / L19 / L26; NT-v2 L9 / L15 / L25).

## J. Why the L2 norm? Comparison with other scalar reductions of $\Delta h_\ell$

**Why L2.** $\|\Delta h_\ell\|_2$ is one of several possible scalar reductions of the per-layer (ref, alt) contrast. We chose the L2 norm because it (i) is rotation-invariant in the model's representation geometry — a property cosine similarity lacks when ref/alt vectors are not unit-normalised in the trained space; (ii) requires no anchor distribution and no LM-head decoding (unlike logit-lens probes, which read off vocab logits and are therefore vocab-coupled); and (iii) is the simplest scalar consistent with the Jacobian-norm tradition in attribution methods (e.g. gradient $\times$ input).

**Empirical comparison.** We compare $\|\Delta h_\ell\|_2$ against three alternatives: cosine dissimilarity ($1 - \cos(h_\ell^{\text{alt}}, h_\ell^{\text{ref}})$ at the variant token); L1 norm ($\|\Delta h_\ell\|_1$); and the directional projection $\langle \Delta h_\ell, \hat{w}_\ell \rangle$ where $\hat{w}_\ell$ is the per-layer LR weight vector. The qualitative dissociation between representative and load-bearing layers — the same within-architecture layer rankings — holds across all four reductions; numerical differences and full per-layer tables are reported in the GitHub repository's `results/tier1_other_aggregators/`.

## K. Suppressor layers (cross-architectural)

The joint-LR coefficients on standardised $\|\Delta h_\ell\|_2$ features include negative entries among the top-5 by |coefficient| in both architectures. In all such cases the layer has a high single-feature AUROC yet enters the joint model with a negative coefficient.

*Table A4.* Top-5 layers by |LR coef| in each architecture. Negative-coefficient "suppressor" layers in **bold**.

| Model | Layer | coef | Single-feat. AUROC |
|---|---|---|---|
| Evo 2 7B | L12 | $+5.244$ | 0.844 |
| | L19 | $+4.210$ | 0.820 |
| | **L21** | $\mathbf{-3.376}$ | 0.805 |
| | L5 | $+2.402$ | 0.808 |
| | L26 | $+2.279$ | 0.808 |
| NT-v2 500M | L9 | $+4.209$ | 0.925 |
| | L25 | $+2.837$ | 0.916 |
| | **L12** | $\mathbf{-2.695}$ | 0.918 |
| | L15 | $+1.990$ | 0.930 |
| | **L20** | $\mathbf{-1.695}$ | 0.905 |

The pattern is consistent with subtraction of a generic perturbation-magnitude confound from the positive contributions of the load-bearing layer and its supporting layers. A full mechanistic account would require activation-patching experiments beyond the scope of this paper.

## L. Extended discussion: training-objective-dependent depth allocation

**Caveat on causal isolation.** The cross-architectural shift in load-bearing depth is consistent with an objective-driven account but does not isolate it. Evo 2 and NT-v2 differ in training objective *and* parameter count, sequence-mixing primitive, training data, and tokenization. Properly attributing the shift to objective alone would require controlled comparisons (same architecture, same data, different objective) which existing public checkpoints do not provide. The HyenaDNA control point (§M) suggests capacity is also a load-bearing factor in the weak-model regime, further complicating a clean objective-only interpretation. We treat the shift direction (deep in CLM, mid-shallow in MLM) as suggestive evidence and an invitation to such controlled experiments.

**Mechanism extension.** The intuition that MLM final-layer capacity is committed to the reconstruction subspace is well aligned with the documented anisotropy of MLM final-layer representations (Ethayarajh, 2019; Wu et al., 2023): the directions an LM head reads off are necessarily a low-dimensional subspace of the layer's hidden geometry, and any task-relevant variance not aligned with that subspace is preferentially preserved at earlier layers, where it is shielded from being overwritten. This is consistent with our observation that NT-v2's L29 collapses to near-chance task discriminability (AUROC $0.55$) while load-bearing weight concentrates at the mid-shallow L9. In CLM, by contrast, every position contributes to the loss at every step and deep-layer representations stay under task-relevant gradient pressure throughout training, allowing load-bearing variance to accumulate at depth (Evo 2 L29).

**What would falsify the hypothesis.** A controlled MLM-vs-CLM comparison at fixed architecture, data, and capacity would directly test the prediction; the directions in which the hypothesis predicts movement are: (i) the load-bearing layer should shift earlier under MLM and later under CLM; (ii) the representative layer should remain mid-network under both, since both objectives produce broadly distributed token-level features at intermediate depth.

## M. HyenaDNA: weak-CLM control

HyenaDNA-large (Nguyen et al., 2023) is a CLM trained on long-range genomic context with a hyena sequence-mixing primitive (no attention). On our benchmark its single-feature peak is L4 (AUROC $0.654$, single-feature) and the joint 10-d feature reaches $0.673$. The load-bearing layer is L0 (the input embedding); both leave-one-out ablation drop ($+0.0155$) and mean $|\mathrm{SHAP}|$ ($0.470$) rank L0 first. Bootstrap 95 % CI on the representative-vs-last gap is $[+0.086, +0.110]$ ($\Delta = +0.098$ point estimate, $100\%$ of resamples show gap $> 0$).

We interpret this as a capacity-limited regime: the model's representations carry only modest task-relevant signal, and the most non-redundant layer is the input embedding itself, on which the (ref, alt) substitution identity directly imprints. This pattern is consistent with the hypothesis (§5) that training objective and capacity together — not objective alone — shape the depth allocation pattern.

## N. Reproducibility

The complete analysis pipeline, cached per-layer $\|\Delta h_\ell\|_2$ feature CSVs, and figure-generation scripts are released at `[anonymized for review]`. All main-text and appendix numbers, tables, and figures — per-layer single-feature AUROC, joint logistic-regression coefficients, leave-one-layer-out ablation drops, tree-SHAP attributions, the cross-architectural unified analysis (Tab. 1), paired bootstrap and DeLong tests (App. C), the three confound controls (App. D), the L0 substitution-bias decomposition (App. E), and the per-gene LOGO drill-down (App. F) — are reproducible from the cached features alone, without re-extracting hidden states from the underlying foundation models. The mean-pool and full-hidden-state baselines (App. G–H) additionally require the cached-extraction scripts to be run on the public NT-v2 500M, Evo 2 7B, and HyenaDNA-large checkpoints. Implementation details (logistic-regression pipeline, tree-SHAP settings, random seed) are given in App. A.