WHICH LAYER TO READ? THE CONSEQUENCES OF LAYER CHOICE IN DNA FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen DNA foundation models are conventionally used through final-layer representations, under the implicit assumption that the final hidden state provides the most informative features. We investigate how this convention affects performance and model comparisons, and how useful layers should be selected. Across ten models from five architectural families and two task suites (280 model–task cells), selecting a layer via held-out validation raises mean test AUROC from 0.726 to 0.770 on the primary suite, with similar gains on the external suite. The validation-based layer selection also changes model rankings in both suites. We then examine how to select layers without probing every stored state. Selecting layers directly with representation and transferability metrics yields lower mean test AUROC than full-layer validation. Useful depth varies across models and tasks, yet every model's median validation-selected depth lies in the first half of the stack in both suites. Searching only the first half of stored states nearly matches full-layer validation in mean test AUROC across both suites while requiring approximately half as many layer-specific probe fits. When screening only 10% of stored states, LogME-guided nomination achieves higher mean test AUROC than the evaluated metric-free rules. These findings establish layer choice as a consequential component of frozen model evaluation and support explicit reporting of the extracted state and selection protocol.

1 INTRODUCTION

DNA foundation models are pretrained to provide reusable representations of genomic sequence. When a foundation model is used as a frozen feature extractor, it offers a sequence of representations across the depths. Thus, a downstream probe must choose where in the backbone to extract its input representation. However, treating the final state as the default does not remove this choice.

In practice, frozen-embedding evaluations of DNA foundation models commonly extract representations from the final layer (Feng et al., 2025; Patel et al., 2024). This convention persists despite evidence that intermediate layers often yield superior representations. For example, probing multiple depths in the Nucleotide Transformer checkpoints revealed that the best layer is model- and task-dependent and often not the final one (Dalla-Torre et al., 2025). In particular, a recent cross-architecture study of language models found that intermediate layers give better frozen embeddings than the final layer (Skean et al., 2025). Taking on this layer dependence, we investigate its implications for model evaluation and selection. Does the final layer convention alter measured downstream utility or relative model comparisons? Can simple endpoint redefinitions eliminate this gap? How many layers must held-out validation probe, and do layerwise criteria improve the selection efficiency?

DNA foundation models with diverse architectures have been trained on the same modality and evaluated on the same tasks. Thus, a shared extraction convention can be tested across architectures directly. We probe every stored state of ten models from five architectural families on 12 binary primary tasks and 16 external tasks (Feng et al., 2025). Layer-selection rules use training or held-out validation data, and test scores are reserved for evaluation (Section 3).

Relative to the final-layer convention, validation-based layer selection improves measured performance and changes model rankings (Section 4). Layerwise profiles further show that useful depth varies across models and tasks (Section 5). Despite this variation, the validation-selected depth

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.

concentrates in the first half of the stack. Validating only the earlier half of stored states closely approximates full validation on both suites (Section 6).

This paper makes three contributions. (i) We quantify how layer choice alters measured performance and relative model rankings across heterogeneous DNA foundation models. (ii) We show the possibilities and limitations of the examined endpoint corrections and characterize model- and task-dependent variations in representation utility. (iii) We show that early-layer screening closely approximates full validation on both suites, while metric-guided nomination improves selection when only a small fraction of states is probed. Together, these results establish layer choice as an explicit, reportable component of frozen-model evaluation.

2 Related Work

Depth-dependent utility of representations. The utility of learned representations varies with depth. Linear probing of every layer was introduced by Alain & Bengio (2016). In NLP, linguistic abstractions are stratified across BERT's layers (Tenney et al., 2019), and the best probing layer differs by task (Liu et al., 2019). Across language-model architectures, intermediate layers often give better frozen embeddings than the final layer (Skean et al., 2025). Moreover, some genomic studies also exhibit similar layer-dependent differences. The Nucleotide Transformer probed ten layers per model across its NT-v1/v2 checkpoints and found the best-performing layers are model- and task-dependent (Dalla-Torre et al., 2025). Evo 2 reports a per-block embedding sweep for variant classification (Brixi et al., 2026). These studies show that intermediate layers can be better. We ask instead what leaving the layer implicit does to model comparisons, and how a good layer can be found. While layer weighting (Peters et al., 2018) and feature concatenation offer alternative ways to construct downstream representations, DNA foundation model evaluations in practice rely on single layer representations. We take this single layer convention and evaluate the systematic consequences of leaving layer choice implicit.

Representation and transferability criteria. Several works have explored label-free or lightweight criteria to estimate representation quality without full fine-tuning. H-score (Bao et al., 2019) and LogME (You et al., 2021) score the alignment between features and target labels to assess transferability of the pre-trained models. RankMe evaluates unsupervised representations via feature-space rank (Garrido et al., 2023). Spectral criteria like α-ReQ assess representation quality through eigenspectrum decay (Agrawal et al., 2022), and TransRate evaluates candidate layer-truncation depths for fine-tuning transfer (Huang et al., 2022). For protein language models and image transformers, Valeriani et al. (2023) identify an expansion followed by contraction in intrinsic dimension and propose a profile-based rule for selecting useful layers. Whereas these criteria are often evaluated by their association with downstream accuracy across models or checkpoints (Li et al., 2021; Agostinelli et al., 2022), we use these criteria within individual models across layers, measuring the selection regret of their chosen layers and their value for nominating candidates.

Sensitivity to evaluation design. Model comparison is increasingly recognized as sensitive to evaluation choices. In speech, altering the probing head reorders the model rankings (Zaiem et al., 2023). In genomics, pooling strategies also affect the relative model performance (Feng et al., 2025). Although the extraction depth also affects the performances, DNA foundation model benchmarks commonly use the final layer by convention (Patel et al., 2024; Marin et al., 2024; Feng et al., 2025). We quantify the performance loss and ranking changes associated with this convention and examine their sensitivity to alternative endpoint definitions.

3 Setup

3.1 Models, Tasks, and Representations

We evaluate ten DNA foundation models spanning five architectural families (Appendix Table A1): four Transformer encoders (GENA-LM and GENA-LM-L, Fishman et al., 2025; DNABERT-2, Zhou et al., 2024; NT-v2, Dalla-Torre et al., 2025), one Transformer-based U-Net (NT-v3, Boshar et al., 2025), two bidirectional state-space models (Caduceus-PS and Caduceus-Ph, Schiff et al., 2024), one Hyena model (HyenaDNA, Nguyen et al., 2023), and two StripedHyena models (Evo,

Nguyen et al., 2024; Evo 2, Brixi et al., 2026). For each model, we extract its execution-ordered hidden states (10 to 34 states per model, Appendix Table A2). These comprise the initial token-embedding output (L_0) and a predefined sequence of stored hidden states. For GENA-LM-L, NT-v2, HyenaDNA, Caduceus, Evo, and Evo 2, the last stored state corresponds to the final layer normalization output. NT-v3 exposes 26 hidden states through its conv–Transformer–deconv interface, where the first state is extracted after the first convolutional block. We express relative depth as $\ell/(n-1)$ for stored state L_ℓ of a model with n stored states. Exact state mappings, including final normalization, are specified in Appendix A.1.

The primary suite contains 15 tasks. Headline analyses use the 12 binary classification tasks evaluated by AUROC (ten transcription-factor-motif detection tasks labeled by PWM (position weight matrix) scans, splice-site detection, and CpG-island detection; see Appendix A.2), yielding 10 models $\times$ 12 tasks $=$ 120 model–task cells. Because ten tasks detect local motifs, the primary suite emphasizes short-range sequence patterns. The remaining regression and multiclass tasks appear in Appendix A.2. For broader task coverage, an external suite adapted from Feng et al. (2025) adds 16 tasks with complete ten-model coverage, yielding 160 cells (15 binary AUROC tasks and one three-class macro-AUROC task). The external suite spans seven task families and multiple species and sequence granularities. Detailed task summaries appear in Appendix Tables A3 – A6.

For every model–sequence pair, all stored states are extracted via a single frozen forward pass per sequence and cached before probe fitting. Thus, backbone feature extraction is fixed. Token representations at each stored state are mean-pooled across non-padding, non-special-token positions (Appendix A.1). Each task uses fixed, disjoint train, validation, and test splits shared across all models and layers (Appendix Table A4). For classification tasks, at every stored state, ℓ_2-regularized logistic regression probes are fitted with regularization parameter $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ on the training split. The validation split determines both C and the selected layer, and test sets evaluate the final selection (Appendix A.3).

3.2 Baselines and Regret

All stored states, including L_0, form the candidate set. The primary baseline uses the final stored hidden state (Section 3.1). Alternative endpoint definitions are examined in the sensitivity analyses (Section 4.3; Appendix A.4). The *validation-selected layer* is the layer maximizing validation performance among the candidate states. The *diagnostic test optimum* is the layer maximizing test performance per cell.

For a model–task cell c and layer ℓ, simple regret is defined as follows:

$$r_c(\ell) = \max_{\ell'} \text{test}_c(\ell') - \text{test}_c(\ell) \geq 0,$$

where the maximum is taken over all stored states. In contrast, the *observed held-out gain* is the test-score difference between the validation-selected layer and final stored state. This gain measures the improvement achieved by validation-based selection, whereas regret measures the gap to the diagnostic test optimum. Because the diagnostic test optimum is the maximum of observed test scores across stored states, it is subject to selection optimism (Appendix A.6).

3.3 Layer Selection Methods and Statistical Analysis

The layer selection methods are categorized by their information requirements. *Label-free profiles* are computed from unlabeled training-split embeddings without accessing validation or test data. They comprise four spectral metrics (effective rank, RankMe, α-ReQ, and PR-var), an intrinsic-dimension (ID) profile rule, and prompt entropy (Skean et al., 2025) (Appendix A.5). These label-free profiles are task-agnostic and therefore yield the same standalone layer choice across tasks for a given model. *Training-label scores (LogME, H-score, TransRate)* use training labels to assess layer quality without fitting downstream probes. *Full validation* fits probes at every stored state and selects the layer with the highest validation performance. *Layer screening* fits probes only at candidate states nominated by a metric or a metric-free rule, then selects among them using validation scores (Section 6). For evaluation, two metric-free depth rules are used. *Early Layers* takes the first k stored states, while *Depth Prior* takes the k states nearest the model's median validation-selected relative depth from other primary tasks (Appendix A.5).

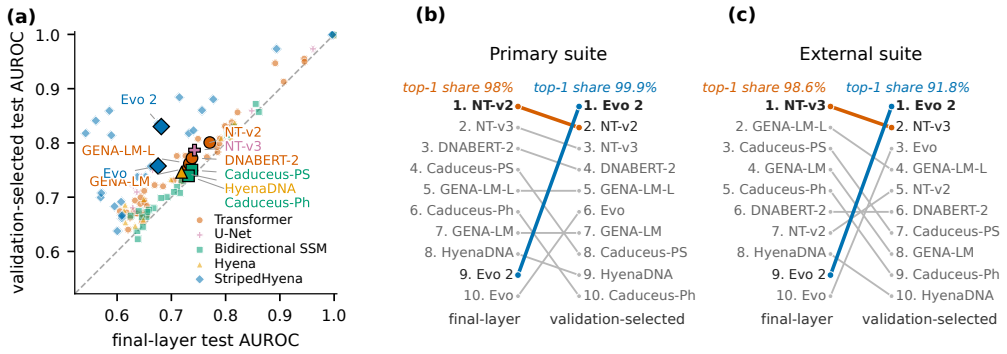

Figure 1: **Validation-based layer selection improves test performance and changes model rankings. (a)** Test AUROC at the validation-selected layer versus the final stored state for 120 primary-suite model–task cells. Colors and marker shapes indicate architectural families. Larger labeled markers denote per-model means. The dashed line denotes equal performance. **(b, c)** Model ranking under the final layer convention (left) and validation-selected layers (right) on the (b) primary suite and (c) external suite. Task-bootstrap top-1 shares are shown for the leading models.

The analysis unit is the individual model–task cell, and aggregate model comparisons use macro-averages over tasks. Screening comparisons against full validation use paired per-cell test-AUROC differences, defined as full validation minus screening, and positive values indicate performance loss. Uncertainty is quantified using a crossed bootstrap, independently resampling models and tasks (10,000 iterations; 95% percentile confidence intervals). Model-ranking statistics instead resample tasks only (Appendix A.6). The *bootstrap top-1 share* denotes the fraction of task-resampled replicates in which a given model achieves top rank.

4 The Final Layer Convention Changes Downstream Conclusions

We compare the final layer convention with full validation, which probes all stored states and selects the layer via held-out validation. Figure 1 summarizes the performance comparison on the primary suites and external suites.

4.1 The Final Layer Convention Incurs Performance Loss

Selecting layers via validation raises mean test AUROC from 0.726 to 0.770 across the 120 primary cells (mean gain +0.0437, crossed-bootstrap 95% CI $[0.0207, 0.0779]$). Test AUROC improves in 92.5% of cells, ties in 0.8%, and decreases in 6.7% (Figure 1a). The gains are positive across all five architectural families (mean AUROC gain: Transformer encoders +0.029, U-Net +0.044, bidirectional state-space models +0.010, Hyena +0.027, and StripedHyena +0.115). On the external suite, mean test AUROC rises from 0.818 to 0.859 (gain +0.0417), with positive mean gains in all seven task families, from +0.017 on promoter tasks to +0.067 on DNA methylation tasks.

Against the diagnostic test optimum, the final layer convention incurs a mean regret of 0.0471 on the primary suite and 0.0457 on the external suite. Validation selection leaves 0.0034 (primary) and 0.0039 (external), recovering 93% and 91%, respectively, of the oracle-referenced performance gap. It is noted that the final layer attains the exact diagnostic test optimum in only 1 of the 120 primary cells and lies within 0.005 AUROC of it in 11.7% of cells. The regression and multiclass results appear in Appendix Table B1.

4.2 Layer Choice Alters Model Rankings

The performance penalty of the final layer convention is not uniform across models, and the layer selection changes their relative rankings. On the primary suite, the top-ranked model shifts from

NT-v2 to Evo 2 (Figure 1b). Specifically, NT-v2 leads in 98.0% of task-bootstrap resamples under the final layer convention, whereas Evo 2 leads in 99.9% under the validation-based selection. On the external suite, the leading model shifts from NT-v3 (top-1 share 98.6%) to Evo 2 (91.8%) (Figure 1c).

The changes extend beyond the leading model. Kendall rank correlation between the two conventions is $\tau = 0.289$ on the primary suite and $\tau = 0.022$ on the external suite. Among 45 model pairs, 16 pairs (primary) and 22 pairs (external) reverse order. Of these reversals, 10 (primary) and 8 (external) are statistically supported by confidence intervals excluding zero under both conventions. The reversals concentrate in the StripedHyena models. Twelve of the 16 primary reversals and 15 of the 22 external reversals involve Evo or Evo 2. The leader change persists under macro, group-balanced, and Borda aggregation (Appendix Tables B2 and B3).

4.3 Sensitivity to the Final Layer Definition

To assess sensitivity to the extracted endpoint, we test two alternative final layer definitions. Under the block-back-off definition (stored block-output endpoint; Definition (b) in Appendix A.4), mean regret slightly increases from 0.0471 to 0.0494 on the primary suite and from 0.0457 to 0.0483 on the external suite. A norm-based endpoint definition (Definition (c) in Appendix A.4) changes the baseline state for NT-v3, Evo, and Evo 2 while retaining all stored states as candidates for validation selection and the diagnostic test optimum. It reduces mean regret to 0.0363 on the primary suite and 0.0228 on the external suite, but does not remove the performance loss. Kendall rank correlations increase to $\tau = 0.600$ on the primary suite and $\tau = 0.689$ on the external suite, while CI-supported pairwise reversals decrease to two and zero, respectively. The primary-suite leader still changes from NT-v2 to Evo 2, whereas Evo 2 leads under both layer conventions on the external suite. Thus, the external leader reversal depends on the endpoint definition.

The performance recovery comes predominantly from Evo and Evo 2, whose endpoints move to L_{10} and L_{28}, respectively. For the eight models other than Evo and Evo 2, mean regret stays at approximately 0.029 on the primary suite and decreases only from 0.025 to 0.024 on the external suite. Endpoint corrections therefore explain part of the observed ranking changes, while leaving performance gaps that extend beyond the StripedHyena models.

5 Useful Depth Is Model- and Task-Dependent

Layerwise profiles reveal three patterns that limit fixed endpoint rules: abrupt performance degradation accompanied by structural changes, gradual decline without comparable diagnostic changes, and task-dependent selected depths (Figure 2). All 120 primary-suite profiles appear in Appendix Figure B1.

5.1 Performance Degradation Can Accompany Structural Changes

We look for structural signatures with two label-free diagnostics, the mean token norm and the RankMe effective rank (Appendix Figure B4). Evo 2 on splice-site prediction illustrates performance degradation coinciding with a structural shift (Figure 2a). Test AUROC remains high near 0.973 across layers L_{24}–L_{28} (with validation selecting L_{27}), before dropping to 0.829 at L_{31}/L_{32}, and partially recovers at the final-normalization state. Evo exhibits an abrupt increase in token norms around L_{11}, followed by a sustained high-norm regime, whereas Evo 2 shows a sharp increase near the end of the stack (Appendix Figure B4). This growth persists with fp32 residual-stream computation (Appendix C). Evo 2 also shows a terminal drop in RankMe effective rank (Appendix Figure B4).

5.2 Gradual Decline Can Lack Clear Diagnostic Signatures

In many model–task cells, test AUROC peaks early and declines gradually toward the final state. Across the 120 primary cells, 89% exhibit no adjacent-layer AUROC drop exceeding 0.02 in the final third of the stack. Nevertheless, 75% have final layer AUROC more than 0.01 below their diagnostic test optimum.

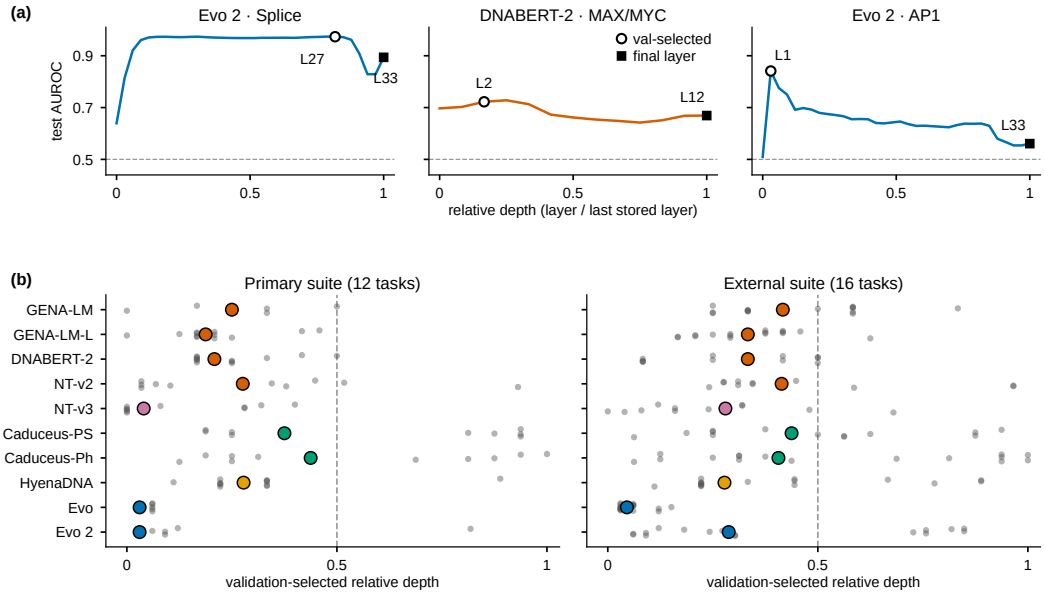

Figure 2: **Useful representation depth varies across models and tasks. (a)** Test AUROC over relative depth across three representative model–task cells: terminal degradation (Evo 2 on splice site prediction), gradual decline without a clear diagnostic signature (DNABERT-2 on MAX/MYC motif detection), and a task-dependent early optimum (Evo 2 on AP1 motif detection). Open circles mark validation-selected layers, filled squares are the final stored states, and the dashed line denotes chance performance (0.5). **(b)** Validation-selected relative depth per model across the primary (left) and the external (right) suites under identical model ordering and axis scales. Gray dots represent individual tasks, colored markers denote the per-model median across all tasks in each suite. Colors indicate architectural families. The dashed line is relative depth 0.5.

DNABERT-2 on MAX/MYC motif detection illustrates gradual decline after an early peak (Figure 2a, middle). Validation selects L_2 (test AUROC 0.722), while the final layer performs 0.053 AUROC lower. No corresponding abrupt transition appears in the model's token-norm or RankMe effective-rank profile (Figure B4).

5.3 Selected Depth Varies across Tasks

Useful depth varies across tasks within the same backbone. For Evo 2, validation selects L_{27} for splice-site prediction but L_1 for AP1 motif detection. On AP1, the selected layer achieves test AUROC 0.841, compared with 0.561 at the final state (Figure 2a). Thus, an endpoint suitable for one task can miss useful representations for another.

The primary suite also shows systematic differences between task types (Figure 2b). Across models, the median validation-selected relative depth is 0.20 for the ten PWM-motif tasks and 0.43 for splice-site prediction. CpG-island detection is excluded from this contrast because its layerwise profiles are nearly flat and saturated (Appendix A.2). Within individual models, the median selected depth for motif tasks is no deeper than the splice-site depth in eight of ten cases. In addition, the size of the gap differs across models, ranging from small for Evo (0.030 vs. 0.061) to large for Evo 2 (0.030 vs. 0.818).

Despite this variation, the selected depths concentrate in the first half of the stack. Across all tasks, the pooled median relative depth is 0.22 on the primary suite and 0.32 on the external suite. Every model's median selected depth lies in the first half on both suites, as do 87% (104/120) of primary cells and 76% (121/160) of external cells (Figure 2b).

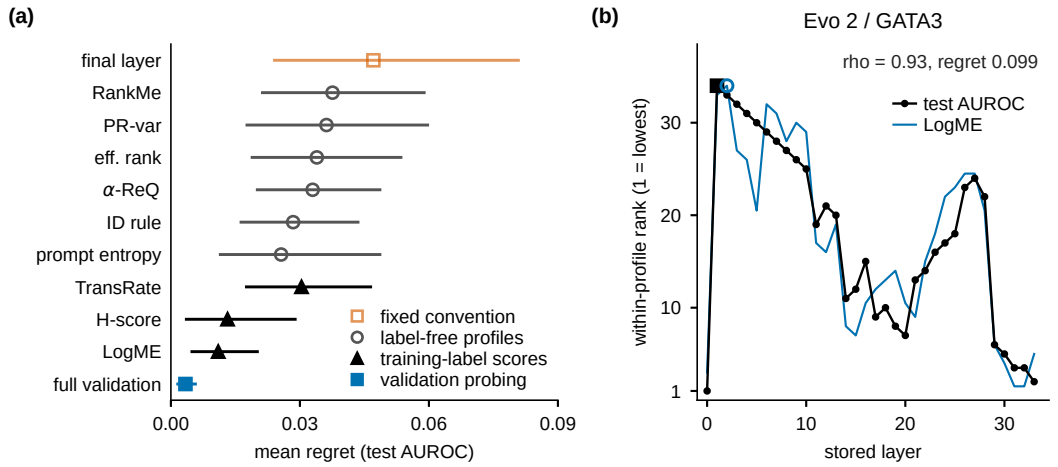

Figure 3: **Standalone metrics narrow but do not close the regret gap. (a)** Mean regret with crossed model×task bootstrap 95% confidence intervals ($n = 120$ primary suite cells). Marker style indicates information access levels. The final layer convention and full validation serve as baselines. The association-versus-regret summary and per-metric scatter plots appear in Appendix Figure B2. **(b)** Within-profile ranks of test AUROC (black) and LogME (blue) for Evo 2 on GATA3 across 34 stored states (1 = lowest). The filled square marks the diagnostic test optimum (L_1), and the open circle marks the LogME-selected layer (L_2). Spearman ρ between the two series and the regret of the LogME argmax are given in the panel.

6 From Layerwise Metrics to Layer Screening

The task dependence of the useful depth motivates task-specific layer selection. A layerwise metric can either select a layer directly or nominate a shortlist for which validation probes make the final choice. We evaluate both uses by comparing their selection regret with that of full validation.

6.1 Standalone Metrics Narrow but Do Not Close the Regret Gap

We first use each criterion of Section 3.3 as a standalone selector. Score-based criteria select their oriented argmax, while the ID rule selects the first local minimum after the first peak of the TwoNN profile (Appendix A.5). Against final-state regret of 0.0471, the spectral geometry profiles (effective rank, RankMe, α-ReQ, and PR-var) leave mean regrets of 0.033–0.038, TransRate 0.0304, the ID rule 0.0284, prompt entropy 0.0257, H-score 0.0132, and LogME 0.0110. Full validation achieves lower mean regret than every standalone selector, at 0.0034 (Figure 3a). Although all standalone selectors reduce mean regret, only LogME, H-score, and ID rule improve on the final layer with paired 95% intervals that exclude zero (Appendix Table B4). For the ID rule, the lower bound is only $+0.0005$. Even LogME leaves a gap of about 0.008 to full validation.

Profile correlation does not guarantee accurate layer selection. For Evo 2 on GATA3, LogME shows close rank-correlation with layerwise test AUROC ($\rho = 0.93$), but selects L_2 instead of the diagnostic test optimum at L_1, incurring regret of 0.099 (Figure 3b). Across the 120 model–task cells, high profile correlation does not guarantee low selection regret (Appendix B.4). These results lead us to evaluate metrics by the selections they induce and to test validation over metric-nominated candidate states.

6.2 Layer Screening Reduces Probe-Fitting Requirements

To reduce the number of layer-specific probe fits, we evaluate a subset of stored states. A nomination rule selects the top $k = \max(1, \mathrm{round}(bn))$ candidate states, where b is the fraction of n stored states to be probed. Probes are fitted only at these states, and validation scores determine the final selection. Across the ten models, probing 10% of states averages 2.1 candidates and probing 50% averages 10.7, compared with 21.9 under full validation.

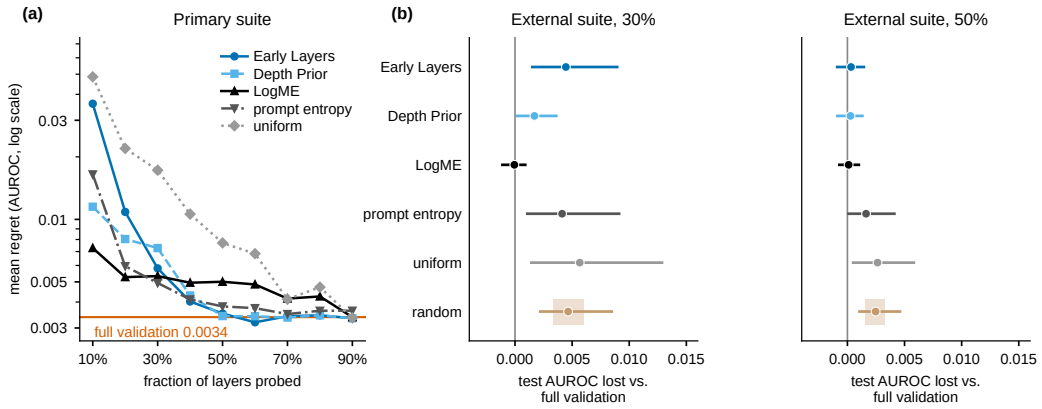

Figure 4: **Layer screening can approach full-validation performance.** **(a)** Mean regret versus the fraction of stored layers probed (primary suite, $n = 120$; layer screening from 10% to 90% in 10% steps; log scale). The curves show the two metric-free depth rules (Early Layers, Depth Prior), LogME, prompt entropy, and uniform spacing, with full validation as the labeled horizontal reference. The comprehensive comparison across all guides and baselines appears in Appendix Figure B3 and Appendix Tables B5 and B6. **(b)** Mean paired test-AUROC differences between full validation and screening on the external suite ($n = 160$) at 30% and 50% screening fractions. Positive values indicate lower performance under screening. Intervals are 95% crossed model–task bootstrap confidence intervals, except for random selection, which shows the mean and 5th–95th percentiles across 100 random-seed results.

We compare metric-guided nomination, metric-free depth rules, and generic screening baselines (Figure 4a). Metric guides rank states by their scores. The ID guide instead nominates the k states nearest its selected layer. Early Layers takes the first k stored states. Depth Prior takes the k states nearest the same model's median validation-selected relative depth on the remaining primary tasks, using leave-one-task-out evaluation. The resulting primary-suite depth prior is transferred to the external suite without retuning. Both depth rules were formulated after the depth pattern in Section 5 had been observed on both suites. The external results therefore test their consistency across suites rather than provide an independent validation. Generic baselines use uniform spacing, random subsets, adaptive coarse-to-fine search, or a fixed relative-depth band (Appendix A.5).

Screening fractions range from 10% to 90% in increments of 10 percentage points, with full validation as the reference. At a 10% fraction, LogME-guided validation yields mean regret of 0.0073, compared with 0.0115 for Depth Prior, 0.0165 for prompt entropy, and 0.028–0.054 for random, uniform, and coarse-to-fine screening. At 30%, prompt entropy (0.0049), LogME (0.0053), and Early Layers (0.0058) yield similar mean regrets.

At a 50% screening fraction, both depth rules have small mean test-AUROC losses relative to full validation, with 0.0001 for Early Layers (95% CI $[-0.0008, +0.0012]$) and 0.00004 for Depth Prior ($[-0.0007, +0.0009]$). Cell-level losses exceed 0.005 AUROC in 4% and 3% of cases, respectively. Thus, validating only the first half of stored states achieves nearly the same mean test AUROC as full validation while requiring approximately half as many layer-specific probe fits.

6.3 Screening Results Are Consistent on the External Suite

We apply the same screening rules to the 160 external suite cells without retuning (Figure 4b). Depth Prior uses each model's median validation-selected depth from the primary suite. The four spectral geometry guides, the ID rule and prompt entropy reuse the model-specific, task-agnostic profiles computed from the source input pools described in Appendix A.5. These profiles are not recomputed on external-suite inputs. Full validation yields mean regret of 0.0039. When probing 30% of stored states, Depth Prior and LogME stay close to full validation, whereas Early Layers, prompt entropy, and uniform spacing lose 0.004–0.006 AUROC (Figure 4b).

At a 50% screening fraction, both depth rules yield mean regret of 0.0042, corresponding to a mean test-AUROC loss of 0.0003 relative to full validation. The crossed-bootstrap 95% CIs are $[-0.0009, 0.0014]$ for Early Layers and $[-0.0009, 0.0013]$ for the transferred Depth Prior. Cell-level losses exceed 0.005 AUROC in 6.9% and 7.5% of cases, respectively, and no task family loses more than 0.002 on average.

LogME-guided screening uses fewer candidate states while maintaining high selection quality. Probing 10% of stored states yields a paired test-AUROC difference of -0.0008 relative to full validation (95% CI $[-0.0036, 0.0008]$; Appendix Table B6). On the primary suite, the same screening fraction yields a mean loss of 0.0039 (95% CI $[0.0005, 0.0091]$). Thus, LogME supports accurate selection from a small candidate set, although the remaining performance gap differs between suites.

7 Discussion

Layer choice is usually left implicit when frozen DNA foundation models are evaluated. Our results show that this choice changes the conclusions. The measured performance and model rankings depend on the extracted state. An evaluation that reads only the final layer therefore cannot separate the quality of a model's representations from the depth at which they are read. Performance gaps persist across the endpoint definitions examined, although the extent of ranking changes depends on the endpoint. The norm-based definition reduces rank reversals and eliminates the change in the top-ranked model on the external suite.

The concentration of useful representations in earlier layers also offers a practical way to reduce probe fitting. In the evaluated setting, validating the first half of stored states achieves nearly the same mean test AUROC as full validation while requiring approximately half as many layer-specific probe fits. LogME-guided nomination can narrow the candidate set further when only a few layers can be probed. These reductions concern probe fitting, that is, backbone feature extraction is held fixed, and metric-guided nomination additionally requires scoring the stored states.

Our evidence is limited to frozen linear probes on mean-pooled embeddings of sequences up to 1 kb. Ten of the twelve primary tasks concern PWM-motif detection. The external suite provides broader task coverage. The observed depth patterns and screening results may differ for longer sequences and tasks requiring broader context. We do not study fine-tuning or multi-layer aggregation, such as learned layer weighting or concatenation. Our diagnostics characterize limitations of the examined endpoint rules without ruling out more effective criteria.

Based on these findings, we recommend three practices for evaluations of frozen embeddings. (i) Report the exact hidden state that is read, including whether a final normalization is applied. (ii) Report the candidate layers, the data split, and the criterion used to select among them. (iii) Report how the headline conclusions change under alternative endpoint definitions. More broadly, the final layer of a frozen model is a hypothesis to be tested, not a default. Making these choices explicit improves the interpretability and reproducibility of downstream evaluations.

AI USE STATEMENT

In this work, we used generative AI tools as research and writing assistants. They provided feedback on the research methodology, experimental design, and interpretation of results, assisted with implementing source code, searching the literature, and drafting and editing parts of the manuscript in English. We did not use generative AI to generate synthetic data, to propose hypotheses, or to formulate or prove theoretical or mathematical claims. The authors reviewed and revised all AI-assisted outputs, including code, experimental results, and text, and made all final decisions on the study design and analyses. We take full responsibility for the final content of this work.

REFERENCES

Andrea Agostinelli, Michal Pándy, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. How stable are transferability metrics evaluations? In *Computer Vision – ECCV 2022*, volume 13694 of *Lecture Notes in Computer Science*, pp. 303–321. Springer, 2022. doi: 10.1007/978-3-031-19830-4_18.

Kumar Krishna Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-ReQ: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17626–17638, 2022.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2309–2313, 2019. doi: 10.1109/ICIP.2019.8803726.

Sam Boshar, Benjamin Evans, Ziqi Tang, Armand Picard, Yanis Adel, Franziska K Lorbeer, Chandana Rajesh, Tristan Karch, Shawn Sidbon, David Emms, et al. A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*, 2025. doi: 10.64898/2025.12.22.695963. URL https://www.biorxiv.org/content/10.64898/2025.12.22.695963.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Jonathan C. Schmok, Ali Taghibakhshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Michael T. Pearce, Elana Simon, Etowah Adams, Zachary J. Amador, Euan A. Ashley, Stephen A. Baccus, Haoyu Dai, Steven Dillmann, Stefano Ermon, Daniel Guo, Michael H. Herschl, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R. K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, John St. John, Jeremy Sullivan, Joseph Tey, Ben Viggiano, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Sudarshan Pinglay, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modelling and design across all domains of life with Evo 2. *Nature*, 652(8112):1349–1361, 2026. doi: 10.1038/s41586-026-10176-5.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025. doi: 10.1038/s41592-024-02523-z.

Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12(1):20005, 2022.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, 2017.

Haonan Feng, Lang Wu, Bingxin Zhao, Chad Huff, Jianjun Zhang, Jia Wu, Lifeng Lin, Peng Wei, and Chong Wu. Benchmarking DNA foundation models for genomic and genetic tasks. *Nature Communications*, 16(1):10780, 2025. doi: 10.1038/s41467-025-65823-8.

Veniamin Fishman, Yuri Kuratov, Aleksei Shmelev, Maxim Petrov, Dmitry Penzar, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. Gena-lm: a family of open-source foundational dna language models for long sequences. *Nucleic Acids Research*, 53(2):gkae1310, 2025. doi: 10.1093/nar/gkae1310.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 2023.

Aldo Glielmo, Iuri Macocco, Diego Doimo, Matteo Carli, Claudio Zeni, Romina Wild, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. DADApy: Distance-based analysis of data-manifolds in Python. *Patterns*, 3(10):100589, 2022. doi: 10.1016/j.patter.2022.100589.

Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. Frustratingly easy transferability estimation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9201–9225. PMLR, 2022.

Yandong Li, Xuhui Jia, Ruoxin Sang, Yukun Zhu, Bradley Green, Liqiang Wang, and Boqing Gong. Ranking neural checkpoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2663–2673, 2021.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 1073–1094, 2019. doi: 10.18653/v1/N19-1112.

Frederikke Isa Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. BEND: Benchmarking DNA language models on biologically meaningful tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uKB4cFNQFg.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*, volume 36, pp. 43177–43201, 2023.

Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336.

Aman Patel, Arpita Singhal, Austin Wang, Anusri Pampari, Maya Kasowski, and Anshul Kundaje. DART-Eval: A comprehensive DNA language model evaluation benchmark on regulatory DNA. In *Advances in Neural Information Processing Systems*, volume 37, pp. 62024–62061, 2024. doi: 10.52202/079017-1981. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/71998bfc3217ffe1cca1ee084dfadadd-Abstract-Datasets_and_Benchmarks_Track.html.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202/.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43632–43648. PMLR, 2024. URL https://proceedings.mlr.press/v235/schiff24a.html.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55854–55875. PMLR, 2025.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1452.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51234–51252, 2023.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143. PMLR, 2021.

Salah Zaiem, Youcef Kemiche, Titouan Parcollet, Slim Essid, and Mirco Ravanelli. Speech self-supervised representation benchmarking: Are we doing it right? In *Proc. Interspeech 2023*, pp. 2873–2877, 2023. doi: 10.21437/Interspeech.2023-1087.

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

APPENDIX

A EXPERIMENTAL DETAILS

A.1 MODELS AND STORED STATES

We evaluate ten DNA foundation models spanning five architectural families. All backbones are frozen. Exact Hugging Face checkpoint identifiers, resolved snapshot revisions, model loader classes, layer hook paths, and compute/storage precisions are detailed in Table A1.

For each model, we extract a predefined set of stored hidden states, ranging from 10 states for HyenaDNA to 34 for Evo and Evo 2, with a mean of 21.9 states per model (Table A2). For nine of the ten models, L_0 is the initial token-embedding output, while intermediate states correspond to residual-stream block outputs. The final stored state corresponds to a final layer-normalization state for NT-v2 (post `emb_layer_norm_after`), HyenaDNA (`ln_f`), Caduceus-PS/Ph (`norm_f`; raw layer-16 output is not separately saved), GENA-LM-L (`last_layer_ln`), Evo, and Evo 2 (backbone norm, L_{33}), and the raw final block output for DNABERT-2 and GENA-LM (whose pre-LayerNorm checkpoints omit final normalization).

NT-v3 constitutes an exception. As an official U-Net-like architecture (convolutional encoder, Transformer bottleneck, deconvolutional decoder with skip connections), its 26 stored states expose the official conv–Transformer–deconv hidden-state interface across heterogeneous spatial and sequence resolutions. Its L_0 state corresponds to the output of the first convolutional stage, and its final stored state represents the full-resolution decoder output prior to the language modeling head. These states are therefore not architecturally interchangeable with standard Transformer layers. For NT-v3, relative depth denotes position within this ordered sequence of stored states, rather than depth through a stack of architecturally uniform blocks.

GENA-LM checkpoints are pre-LayerNorm BERT variants requiring the official custom model loader class. Initializing them with generic Hugging Face `BertModel` classes silently reinitializes all LayerNorm parameters. For Evo 2, stored L_{31} and L_{32} representations are bitwise identical under the native `bf16` extraction setting. This duplication disappears with an `fp32` residual stream (Appendix C). Both states are retained in the primary candidate set.

Stored token representations are mean-pooled over non-padding, non-special-token positions to obtain fixed-dimensional embeddings for downstream probing. Padding positions are excluded using the attention mask, and special-token positions are additionally excluded by token ID.

For input processing, NT-v3 sequences are right-extended with artificial `N` characters to the nearest multiple of 128 (no `[PAD]` token is used; natural `N` bases remain valid sequence content), and mean pooling applies stage-aware fractional coverage excluding only artificially appended positions. For maskless models (HyenaDNA, Caduceus-PS/Ph) evaluated on variable-length external tasks, sequences were batched by exact tokenized length to prevent artificial padding from entering forward passes (unique-length sequences run at batch size 1). Attention-masked models (GENA-LM, GENA-LM-L, DNABERT-2, NT-v2) use standard padding masks; causal models (Evo, Evo 2) are right-padded, which does not affect preceding valid token states. DNABERT-2 maps natural `N` bases to `[UNK]` in 18 affected sequences project-wide ($\leq 0.05\%$ of total bases). These token positions are excluded from mean pooling by the special-token mask. Stored feature dimensions per model are listed in Table A2.

Table A1: **Checkpoints, loaders, stored-state hooks, and precisions.** Rev. = Hugging Face snapshot revision resolved at extraction time. Hooks specify modules from which stored states are extracted (L_0 = embedding output, except for NT-v3, whose L_0 follows the first convolutional stage, followed by block outputs). Compute = forward-pass precision; Storage = cached embedding precision. GENA-LM/-L use official custom pre-LayerNorm model classes to prevent parameters from being silently reinitialized. HS denotes the model's returned hidden-state sequence.

Model	Checkpoint (Hugging Face id)	Rev.	Loader / class	Stored-state hooks (L0; blocks; final)	Compute / storage
HyenaDNA	`LongSafari/hyenadna-medium-160k-seqlen-hf`	`7ebf717`	`AutoModel` (remote code)	HS; last = `ln_f`	fp16 / fp16
Caduceus-PS	`kuleshov-group/caduceus-ps_seqlen-131k_d_model-256_n_layer-16`	`d89eeb8`	`AutoModel` (remote code)	HS; last = `norm_f`	fp16 / fp16
Caduceus-Ph	`kuleshov-group/caduceus-ph_seqlen-131k_d_model-256_n_layer-16`	`b047752`	`AutoModel` (remote code)	HS; last = `norm_f`	fp16 / fp16
GENA-LM	`AIRI-Institute/gena-lm-bert-base-t2t`	`4f1352b`	official pre-LN class (remote code)	HS; last = raw block L12 (no final LN in checkpoint)	fp32 / fp16
GENA-LM-L	`AIRI-Institute/gena-lm-bert-large-t2t`	`f997a8a`	official pre-LN class (remote code)	HS; last = `encoder.last_layer_ln`	fp32 / fp16
DNABERT-2	`zhihan1996/DNABERT-2-117M`	`7bce263`	`AutoModel` (custom encoder)	`model.embeddings`; `model.encoder` (`output_all_encoded_layers`)	fp32 / fp16
NT-v2	`InstaDeepAI/nucleotide-transformer-v2-500m-multi-species`	`06615c1`	`AutoModelForMaskedLM` → `.esm`	HS; last = post `emb_layer_norm_after`	fp32 / fp16
NT-v3	`InstaDeepAI/NTv3_650M_pre`	`5e6050b`	`AutoModelForMaskedLM` (remote code)	HS (official conv–Transformer–deconv interface)	fp32 / fp16
Evo	`togethercomputer/evo-1-131k-base`	`c206aab`	`AutoModelForCausalLM` (remote code)	pre-hook `backbone.blocks[0]`; `backbone.blocks[`i`]`; `backbone.norm`	bf16 / fp32
Evo 2	`arcinstitute/evo2_7b_base`	`074097e`	official `evo2` package	pre-hook `model.blocks[0]`; `model.blocks[`i`]`; `model.norm`	bf16 / fp32

Table A2: **Evaluated models and stored states.** L_0 denotes the token-embedding output for nine models and the first convolutional-stage output for NT-v3. Stored states follow each model's extraction interface and may include a final normalization or decoder output. For Evo and Evo 2, the final layer-normalization output is stored as the final state (L_{33}) and serves as the primary final layer. Dim = representation dimension after mean pooling. Evo 2 stored L_{32} duplicates L_{31} (bfloat16 artifact; retained in reported data).

Model	Family	States	Primary final layer	Dim
HyenaDNA	Hyena	10	L9 (norm)	256
Caduceus-PS	Bidir. SSM	17	L16 (norm)	512
Caduceus-Ph	Bidir. SSM	17	L16 (norm)	256
GENA-LM	Transformer	13	L12	768
GENA-LM-L	Transformer	25	L24 (norm)	1024
DNABERT-2	Transformer	13	L12	768
NT-v2	Transformer	30	L29 (norm)	1024
NT-v3	U-Net (attn. core)	26	L25 (decoder out)	1536
Evo	StripedHyena	34	L33 (norm)	4096
Evo 2	StripedHyena	34	L33 (norm)	4096
Total / mean		219 / 21.9		

A.2 Task Accounting and Data Description

Primary suite. The primary benchmark suite comprises 15 tasks evaluated across all ten models: 10 binary PWM-based transcription-factor motif tasks (AP1, CEBPB, CTCF, ETS1, GATA3, MAX/MYC, NRF1, REST, SP1, and YY1), splice-site detection, CpG-island detection, GC-content regression, and two multiclass classification tasks (functional element classification, 5 classes; transposable element classification, 6 classes). Headline analyses aggregate the 12 binary AUROC tasks ($10 \times 12 = 120$ cells) to maintain metric commensurability; regression and multiclass results appear in Table B1.

Reference and genomic splits. All primary-suite tasks are constructed from the human GRCh38/hg38 reference genome. Chromosome-level partitioning ensures that windows from the same genomic locus cannot occur in different splits.

- **Training Split:** Odd autosomes (chr1, chr3, ..., chr21).
- **Validation Split:** Chromosomes chr2, chr4, and chr6.
- **Test Split:** Remaining even autosomes (chr8, chr10, ..., chr22) and chromosome chrX.

Primary input sequences comprise $1{,}000$ bp windows unless noted otherwise. Annotation sources include GENCODE v44 (splice sites, exons, introns, promoter contexts), UCSC `cpgIslandExt` (CpG islands), UCSC RepeatMasker `hg38.fa.out` (transposable elements), EPDnew (promoter transcription start sites), VISTA (experimentally verified enhancers), and JASPAR2024 CORE vertebrate non-redundant position frequency matrices (PFMs) scanned via FIMO at $p < 10^{-4}$ (motif binding tasks). The 15 tasks contain 241,673 sequence records in total. Nine windows are each shared by two tasks, always within the same split. Deduplication for embedding extraction therefore yields a unified pool of 241,664 unique sequences, with no cross-split sharing among these windows. All tasks utilize fixed, disjoint train, validation, and test splits. Primary task split sizes range over $4{,}435$–$23{,}811$ (train), $1{,}586$–$10{,}328$ (validation), and $2{,}967$–$15{,}860$ (test) sequences.

Task construction details.

- **GC Content (Regression):** $49{,}999$ windows ($1{,}000$ bp); the target continuous label is the window's GC sequence fraction.
- **Motif Binding Tasks (10 Binary Tasks):** Positive instances comprise $1{,}000$ bp windows centered on a significant FIMO hit ($p < 10^{-4}$) of the target JASPAR motif. Each positive 1 kb window is centered exactly on the motif midpoint, so the (short) motif occupies the central position. Negative instances are windows lacking motif hits, GC-matched to the

positive distribution within $\pm 2\%$. Tasks contain approximately $5{,}000$ positive and $5{,}000$ negative windows prior to deduplication.

- **Splice Site Detection (Binary):** Positives are centered on annotated GENCODE v44 splice junctions. Negatives contain central GT or AG dinucleotides without annotated splicing activity, GC-matched to the positive set within $\pm 2\%$.
- **CpG Island Detection (Binary):** Positives center on UCSC CpG islands. Negatives are random genomic windows situated at least 10 kb away from any annotated CpG island. Negatives are intentionally *not* GC-matched (mean GC fraction 0.399 vs. 0.638 for positives). This task therefore functions largely as a sequence-composition benchmark. A logistic probe on GC fraction alone reaches 0.991 test AUROC, and no stored layer of any model falls below 0.988 validation AUROC. Layerwise profiles are therefore nearly flat, and the selected layer carries little information for this task.
- **Functional Element Classification (5 Classes):** $\sim 5{,}000$ sequences per class spanning promoters (EPDnew TSS ± 500 bp), enhancers (VISTA), exons (GENCODE), introns (GENCODE), and intergenic regions.
- **Transposable Element Classification (6 Classes):** $5{,}000$ sequences per class centered on RepeatMasker annotation midpoints. Five repeat classes map RepeatMasker families to LINE, SINE/Alu, LTR, DNA transposon, and Simple Repeat (including low-complexity repeats). All unlisted repeat families are excluded. A non-repeat background class samples chromosome-length-weighted random $1{,}000$ bp windows excluding any RepeatMasker overlap (without GC matching).

Windows containing ambiguous `N` characters are discarded across all primary tasks.

External suite. From the 57 classification datasets in Feng et al. (2025), we constructed a coverage-oriented external suite of 16 single-dataset tasks with full ten-model coverage ($10 \times 16 = 160$ cells): 15 binary AUROC tasks and one three-class macro-AUROC task spanning seven task families and multiple species (Table A5). The GM12878 promoter task was replaced by an EPD-style non-TATA human promoter dataset because its input sequences exceeded GENA-LM's 512-token limit. Split sizes and task metadata appear in Tables A3, A5, and A6.

Auxiliary length analysis. Six promoter variants (three promoter types at 70 and 300 bp) are used for length-sensitivity analyses and are not counted as additional tasks in the 16-task external aggregate. The human non-TATA 300-bp variant shares 4,999 sequences with the external human promoter task. Among these shared sequences, 4,089 retain the same split assignment and 910 have different assignments across the two datasets (908 training–validation and two training–test). Probes are fitted separately for the two datasets, and the auxiliary length analysis is not used to construct or tune the screening rules.

Input length and truncation. Primary sequences are 1,000 bp long. Their tokenized lengths remain within the respective model context limits, so no primary input is truncated. External promoter variants used in the length-sensitivity analysis are 70 bp and 300 bp long.

Diagnostic analysis pools. Prompt entropy is computed on 500 sequences sampled from the training split of the unified pool with seed 0, each truncated to 256 bp. Spectral geometry profiles use 8,000 training-split sequences sampled with `RandomState(42)`. Profiles computed from 8,000 test-split sequences are used only for the input-source sensitivity analysis.

Table A3: **Task accounting.** Primary benchmark evaluates 15 tasks. Headline metrics use 12 binary AUROC tasks for commensurability. External suite comprises 16 classification datasets selected via coverage-oriented design (Table A5). Six promoter variants (three types at 70 and 300 bp) are used for auxiliary length-sensitivity analyses and are not counted as additional tasks in the external aggregate. The human non-TATA 300-bp variant shares 4,999 sequences with the external human promoter task (Appendix A.2).

Suite	Category	Tasks	Metric(s)	Cells
Primary	all evaluated	15	AUROC / macro-F1 / Pearson r	150
Primary	headline aggregate	12	test AUROC	120
Primary	appendix-only	3	macro-F1 (2), Pearson r (1)	30
External	evaluation suite	16	AUROC (15), macro-AUROC (1)	160
Auxiliary	promoter-length variants	6	AUROC	60

Table A4: **Primary-suite split sizes.** Counts are reported per task after quality filtering; windows shared across tasks are counted in each task. The unified embedding-extraction pool contains 241,664 unique sequences.

Task	Train	Val	Test
GC content (regression)	23,811	10,328	15,860
Motif: AP1	4,744	2,096	3,148
Motif: CEBPB	4,678	2,059	3,208
Motif: CTCF	4,595	1,686	3,027
Motif: ETS1	4,791	1,957	3,252
Motif: GATA3	4,776	2,152	3,072
Motif: MAX/MYC	4,829	1,879	3,276
Motif: NRF1	4,569	1,586	2,967
Motif: REST	4,620	1,743	3,133
Motif: SP1	4,435	1,600	3,044
Motif: YY1	4,805	1,993	3,202
Splice site	9,991	2,828	6,951
CpG island	10,139	3,492	6,369
Functional element (5-class)	12,511	4,655	7,816
Transposable element (6-class)	14,391	5,970	9,639

Table A5: **External-suite task list.** 16 single-dataset classification tasks adapted from the source benchmark catalog (Feng et al., 2025). Lengths reflect median sequence lengths (bp) of prepared datasets. The task list distinguishes seven families. For family-balanced aggregation, methylation and histone tasks are combined into an epigenetic group, yielding six analysis groups.

Task	Source dataset	Family	Metric	Len.	Species
Human TFBS 1	`tf/Human_TFBS_1`	TFBS	AUROC	101	human
Human TFBS 2	`tf/Human_TFBS_2`	TFBS	AUROC	101	human
Mouse TFBS 1	`mouse/mouse_TFBS_1`	TFBS	AUROC	101	mouse
6mA methylation	`iDNA_ABF/6mA`	methylation	AUROC	41	human
5mC methylation	`iDNA_ABF/5mC`	methylation	AUROC	41	human
Human promoter (EPD-style, non-TATA)	`prom/promoter_notata_300bps`	promoter	AUROC	300	human
Arabidopsis promoter (non-TATA)	`iPro-WAEL/Promoter_Arabidopsis_NonTATA`	promoter	AUROC	251	A. thaliana
Bacterial promoter (*B. amyloliquefaciens*)	`iPro-WAEL/Promoter_B_amyloliquefaciens`	promoter	AUROC	40	B. amyloliquefaciens
Enhancer	`enhancers/enhancer`	enhancer	AUROC	199	human
Enhancer (Ensembl)	`genomic_benchmark/enhancer_ensembl`	enhancer	AUROC	268	human
Splice acceptor	`splice/acceptors`	splice	AUROC	600	human
Splice donor	`splice/donors`	splice	AUROC	600	human
Splice-site type (3-class)	`splice/splice_site_type_DNABERT`	splice	macro-AUROC	400	human
Open chromatin region	`genomic_benchmark/open_chromatin_region`	open-chromatin	AUROC	317	human
DNase I hypersensitivity	`iDHS-EL/DNase_I`	open-chromatin	AUROC	243	human
Yeast H3K4me3	`EMP/Yeast_H3K4me3`	histone	AUROC	500	S. cerevisiae

A.3 Probe Details

Binary classification. Per-layer probes are ℓ_2-penalized logistic regression models (`lbfgs` solver, maximum 1,000 iterations, convergence tolerance 10^{-3}) fitted on training splits. The inverse regularization parameter is selected on validation splits across five log-spaced values, $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. Edge grid selections are recorded but the grid is not extended in primary evaluations. Binary tasks are evaluated via test AUROC.

Multiclass classification. The two primary multiclass tasks use the same logistic regression probe family, solver configuration, and five-value C grid. Regularization and layer selection use validation macro one-vs-rest AUROC, while primary test results are reported using macro-F1. The external suite's three-class task is evaluated using macro one-vs-rest AUROC.

Regression. The single primary regression task uses ridge regression, selecting regularization parameter α on validation splits from $\alpha \in \{10^{-2}, 10^{-1}, 1, 10, 100\}$, evaluated via Pearson r (mean-squared error and Spearman correlation are also stored).

Split roles and tie handling. All selection methods operate on the same underlying pool of stored states within each cell, including L_0. Full validation evaluates every state, whereas screening restricts validation to the nominated subset. Validation-score ties are resolved by selecting the first maximizer in candidate order: ascending stored-state index for full validation, metric-ranking or-

Table A6: **External-suite split sizes** (sequences per split after minimum-length filtering and stratified caps; validation splits are 15% per-class stratified carves from capped training splits, seed 42; source test splits are used up to stratified caps).

Task	Train	Val	Test
Human TFBS 1	12,750	2,250	8,000
Human TFBS 2	12,750	2,250	8,000
Mouse TFBS 1	4,818	850	2,430
6mA methylation	12,750	2,250	8,000
5mC methylation	2,789	492	1,407
Human promoter (EPD-style, non-TATA)	12,750	2,250	8,000
Arabidopsis promoter (non-TATA)	7,027	1,240	3,543
Bacterial promoter (*B. amyloliquefaciens*)	1,260	223	636
Enhancer	12,722	2,246	400
Enhancer (Ensembl)	12,750	2,250	8,000
Splice acceptor	12,750	2,250	6,654
Splice donor	12,750	2,250	6,592
Splice-site type (3-class)	12,750	2,250	7,999
Open chromatin region	12,750	2,250	8,000
DNase I hypersensitivity	605	106	306
Yeast H3K4me3	12,750	2,250	8,000
Total	156,721	27,657	85,967

der for metric-guided screening, and the order produced by each nomination rule for metric-free screening.

LogME, TransRate, and the spectral metrics are ranked at full precision, with exact ties resolved toward the shallower stored state. For primary-suite H-score and prompt entropy, ties in the rounded scores retain the ordering produced by the original sorting implementation. This does not guarantee preference for shallower states. Sensitivity checks over alternative resolutions of these rounded-score ties leave the reported aggregate screening results unchanged at the displayed precision.

Feature conditioning and convergence. A `StandardScaler` is fitted on training embeddings and applied to validation and test embeddings. Among the 2,628 probes fitted with the validation-selected C for each stored-state–task pair, 10 (0.4%) reach the 1,000-iteration `lbfgs` cap (Appendix C). Among the 2,628 stored-state–task configurations, 39% select the lower grid boundary ($C = 10^{-3}$) and 16% select the upper boundary ($C = 10$). Selected C values lie on the search-grid boundary in 54% of these configurations.

We expand the grid to $C \in \{10^{-4}, 10^{-3}, \ldots, 10^{3}\}$ for nine cells covering Caduceus-Ph, NT-v2, and Evo 2 on AP1, splice site, and CpG island. The expanded grid changes the validation-selected layer in one cell: Caduceus-Ph on CpG island shifts from L_{16} to L_{15}. Test AUROC does not decrease in any of the nine cells, and layerwise performance profiles retain Spearman correlations of $\rho \geq 0.98$ with those from the original grid. These checks indicate limited sensitivity to the regularization range in the evaluated cells.

A.4 Final Layer Definitions

- **Definition (a)** [Primary Convention]: Reads the final stored hidden state. For GENA-LM-L, NT-v2, HyenaDNA, Caduceus, Evo, and Evo 2, this represents a final layer-normalization state; for NT-v3, the final decoder output; and for GENA-LM and DNABERT-2, the raw final block output (Appendix A.1).
- **Definition (b)** [Block Back-Off]: Reads the deepest stored raw block output, before final normalization. This differs from Definition (a) for four models: HyenaDNA (L_8), Evo and Evo 2 (L_{32}), and GENA-LM-L (L_{23}). For GENA-LM-L the raw output of block 24 is not stored, so L_{23} is the penultimate block output. For Caduceus-PS/Ph and NT-v2, no raw final block output is stored, and Definition (b) falls back to (a). Definition (b) is therefore a stored-state approximation to a pre-normalization endpoint, rather than a uniform removal of final normalization across all models.

- **Definition (c)** [Norm-Based]: Tracks mean token norms across stored block outputs ($L \geq 1$) and identifies the last contiguous run of adjacent-layer transitions for which the later-to-earlier norm ratio exceeds 10. The endpoint is the stored state immediately preceding the first transition of that run. Models without such a run retain their final stored state. This rule yields L_{10} for Evo, L_{28} for Evo 2, and L_{24} for NT-v3, leaving the other seven models unchanged. For NT-v3, the rule detects a scale change at the deconvolution output rather than numerical explosion.

These definitions change only the baseline endpoint. The candidate states for validation selection and the diagnostic test optimum remain unchanged. Definitions (b) and (c) are used only in sensitivity analyses (Section 4.3).

A.5 Metric Definitions and Layer Screening Rules

Geometry profiles. Four spectral geometry metrics are calculated per layer from $N = 8{,}000$ unlabeled training-split sequence embeddings (mean-pooled, task-agnostic, sampled with a fixed seed from a pooled sequence set). For NT-v3, the same training-split sequences are processed using the model's official input format, with artificial `N`-padding to multiples of 128 and stage-aware weighted pooling that excludes the appended positions. This changes input processing, not the source sequences or their split assignments.

1. **Effective Rank:** For the centered embedding matrix with singular values σ_i, we compute $\exp(-\sum_i p_i \log p_i)$, where $p_i = \sigma_i / \sum_j \sigma_j$.
2. **RankMe:** Following Garrido et al. (2023), we compute $\exp(-\sum_i q_i \log q_i)$ from the singular values of the uncentered embedding matrix Z, where $q_i = \sigma_i(Z)/\sum_j \sigma_j(Z) + \varepsilon$ and $\varepsilon = 10^{-7}$. No renormalization is applied after adding ε. Effective rank and RankMe therefore differ in centering and in the numerical ε term. Setting ε to zero leaves all RankMe standalone selections and screening shortlists unchanged across the evaluated screening fractions on both suites.
3. **α-ReQ:** Spectral-decay exponent (Agrawal et al., 2022), computed from the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$ of the centered embedding covariance. Eigenvalues below or equal to $10^{-12}\lambda_1$ are discarded. For the fit, we further retain the m eigenvalues satisfying $\lambda_i > 10^{-4}\lambda_1$ and require $m \geq 8$. Excluding the leading eigenvalue, we fit $\log \lambda_i = a - \alpha \log i$ by least squares over $i = 2, \ldots, m$. Lower α is preferred.
4. **PR-var:** Participation ratio of the centered-covariance eigenvalue spectrum,

$$\text{PR-var} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2},$$

   using eigenvalues satisfying $\lambda_i > 10^{-12}\lambda_1$. Higher values are preferred.

These geometry metrics require no labels and no validation/test sequence inputs. As a sensitivity check, profiles were recomputed using $N = 8{,}000$ unlabeled *test-split* inputs. Standalone argmax layers are identical between input sources in 39 of 40 model–metric pairs, the exception being effective rank for Evo. Across 360 screening configurations (ten models $\times$ four geometry metrics $\times$ nine screening fractions from 10% to 90% in 10-percentage-point increments), 45 yield different candidate shortlists between training- and test-input profiles. The corresponding differences in mean guided regret (test-input minus training-input) range from -0.0017 to $+0.0014$ AUROC. Orientations are fixed a priori: higher is preferred for effective rank, RankMe, PR-var, LogME, H-score, and TransRate, and lower is preferred for α-ReQ. The ID rule has no orientation and is used only through its selected layer.

Prompt entropy. Matrix-based (von Neumann) entropy of sequence token embeddings at a given layer (Skean et al., 2025): for token matrix Z of a single sequence, entropy is defined as $-\sum_i w_i \log w_i$ over eigenvalues w of the trace-normalized Gram matrix $ZZ^\top$. Per-sequence values are averaged over a fixed, task-agnostic pool of 500 training-split sequences (256 bp) to produce one profile per model and layer. Prompt entropy consumes unlabeled training task inputs without accessing labels, validation, or test data (higher preferred).

Intrinsic-dimension profile rule. TwoNN intrinsic dimension (Facco et al., 2017) is computed per layer with DADApy (Glielmo et al., 2022) on the same $N = 8{,}000$ unlabeled training-split embeddings as the geometry profiles, discarding the largest 10% of neighbor-distance ratios. Embeddings are cast to float64, and duplicate vectors within a layer are removed. The rule selects the first local minimum after the first local maximum of the profile. This rule was fixed and committed before any regret was computed. If the profile has no local maximum or its first peak occurs at the last stored state, the rule selects the global ID minimum excluding L_0. This fallback is invoked only for GENA-LM, selecting L_9. As a sensitivity check, Gride estimates (Denti et al., 2022) at several scales change the selected layer for four models (GENA-LM-L, NT-v2, NT-v3, Evo). For layer screening, the guide nominates the k layers nearest the selected layer, with ties broken toward shallower layers.

Training-label scores. LogME (You et al., 2021), H-score (Bao et al., 2019), and TransRate (Huang et al., 2022) are computed from training embeddings and their labels without fitting downstream probes. Higher scores are preferred for all three methods.

TransRate (Huang et al., 2022) is computed with the official implementation ($\epsilon = 10^{-4}$) on the same training inputs as LogME and H-score. The official code adds normalization steps absent from the published formula, including division by a layer-dependent rank. The two versions select the same layer in only 10.6%–11.7% of cells, and we report the official version. The selected layer is also sensitive to ϵ. Layers selected at 10^{-4} and 10^{-2} agree in only 33–38% of cells.

Selection and nomination. Score-based metrics select a state by maximizing their oriented score and nominate screening candidates by ranking states under the same orientation. The ID rule instead selects a state from the profile shape and nominates nearby states. Both uses require metric profiles across all stored states. Missing metric values are excluded before candidate ranking.

Layer-screening protocol. For a screening fraction b and n_ℓ stored states, the candidate count is $k = \max(1, \mathrm{round}(bn_\ell))$, with half-integer ties rounded to the nearest even integer. Score-based guides nominate the top k states under their fixed orientation. Probes are fitted on training data for the nominated states, and validation performance determines the final selection. We evaluate fractions from 10% to 90% in 10-percentage-point increments. At 10%, 30%, 50%, and 70%, the mean numbers of states probed are 2.1, 6.6, 10.7, and 15.4, respectively, compared with 21.9 for full validation.

Metric-free nomination rules. Early Layers nominates the first k stored states. Depth Prior nominates the k states nearest the model's median validation-selected relative depth on the remaining primary tasks, excluding the target task. For external tasks, the prior is computed from the primary suite and transferred without modification. We additionally evaluate uniform spacing, random subsets, coarse-to-fine screening, and a fixed depth band.

- **Uniform:** Selects $\mathrm{round}(\mathrm{linspace}(0, n_\ell - 1, k))$ unique indices.
- **Random:** Draws k layers without replacement across 100 random seeds (master seed 42), reported via 5th–95th percentile seed bands.
- **Coarse-to-Fine:** Places $\min(\max(2, \lceil k/2 \rceil), k)$ uniform coarse probes, then expands iteratively around the coarse validation optimum until k probes are trained. For $k = 1$ this reduces to a single coarse probe (L_0).
- **Fixed Band:** Let n_{band} be the number of stored states with relative depth in $[0.2, 0.5]$. If $n_{\mathrm{band}} \leq k$, all band states are probed. Otherwise, k states are linspace-subsampled within the band.

A.6 Statistical Procedures and Test Independence

Analysis unit and uncertainty. The model–task cell is the primary analysis unit (120 primary, 160 external cells); aggregate comparisons macro-average over tasks. Confidence intervals use a crossed cluster bootstrap resampling models and tasks independently (10,000 iterations, seed 42; 95% percentile intervals). Rankings, top-1 shares, Kendall τ, and pairwise model differences instead

use a task bootstrap, because resampling models would change the set being ranked. Bootstrap top-1 share represents the fraction of task-resampled replicates in which a model ranks first.

Paired comparisons. For each nomination rule and screening fraction, we compute the within-cell test-AUROC difference $d_c = \mathrm{test}_c(\text{full validation}) - \mathrm{test}_c(\text{screening})$. Positive values indicate lower performance under screening. Mean differences and 95% confidence intervals are estimated using the crossed model–task bootstrap, with the same resampled cells used for both methods.

Selection inflation in the diagnostic test optimum. Maximizing over observed test scores introduces selection optimism into the diagnostic test optimum. Within each primary-suite cell, a half-split analysis selects a state on one half of the test set and evaluates it on the other. The estimated half-sample optimism is divided by $\sqrt{2}$ to approximate inflation at the full test-set size, yielding 0.0023 AUROC (95% CI [0.0010, 0.0037]). A bootstrap estimate is 0.0020 AUROC. A separate parametric bound based on per-cell bootstrap standard errors is 0.0052 AUROC. Because all selection rules share the same diagnostic test optimum within each cell, this reference cancels in pairwise regret differences.

Test independence. The reported selection rules use training and validation data to determine the selected state; test scores do not enter their selection criteria. Test scores are used to evaluate the selected states and for retrospective analyses, including layerwise profiles and diagnostic test optima. As an implementation check, permuting stored test scores leaves validation- and metric-based selections unchanged.

B Additional Results

B.1 Complete Layerwise Profiles

Figure B1 presents test-AUROC profiles across all stored states for the 120 primary-suite model–task cells. These profiles place the examples in Section 5 in the context of the complete evaluation. Open circles mark validation-selected states, and filled squares mark the final stored states under Definition (a).

The profiles show early peaks, broad performance plateaus, and both gradual and abrupt declines toward the final states. Useful depth also varies within a model across tasks: Evo 2 selects L_{27} for splice-site detection but L_1 for AP1. The CpG-island profiles are nearly flat, so differences in selected depth on this task should not be interpreted as strong evidence of depth preference.

B.2 Regression and Multiclass Task Results

Table B1 reports results for the three primary-suite tasks excluded from the binary-AUROC aggregate: GC-content regression, functional-element classification, and transposable-element classification. Regression uses Pearson r for validation selection and test evaluation. Multiclass tasks use validation macro one-vs-rest AUROC for regularization and layer selection, while test results are reported using macro-F1 and macro-AUROC. These tasks are reported separately without pooling their different evaluation metrics.

On GC-content regression, all ten models achieve test Pearson $r \geq 0.998$ at validation-selected states. Nine models select L_0 or L_1; NT-v2 selects L_{27}, with near-ceiling performance at both the selected and final states. The largest gains occur for Evo and Evo 2, whose final-state correlations are 0.9885 and 0.9897, respectively.

On multiclass tasks, the gains vary substantially across models. The largest macro-F1 improvements occur on transposable-element classification for Evo 2 (+0.2195) and Evo (+0.1616). For NT-v3 on functional-element classification, validation selection increases test macro-AUROC by 0.0061 but decreases macro-F1 by 0.0013. Thus, improvements under the selection metric need not translate into improvements under a different reporting metric.

B.3 Sensitivity of Model Rankings to Endpoint Definitions and Aggregation

Tables B2 and B3 report sensitivity to baseline endpoint definitions, aggregation methods, and task weighting. Endpoint definitions change only the baseline state; validation selection and the diagnostic test optimum continue to use all stored states.

Macro aggregation assigns equal weight to each task. Group-balanced aggregation first averages scores within each task group and then averages equally across groups. On the primary suite, the three groups are PWM-motif detection (ten tasks), splice-site detection, and CpG-island detection, each receiving weight $1/3$. The separate $0.5/0.5$ weighting assigns half the weight to the PWM-task mean and half to the mean of splice-site and CpG-island detection, giving the latter two tasks weight $1/4$ each. These are therefore distinct weighting schemes. Borda aggregation uses the negative mean of each model's within-task ranks, assigning average ranks to ties; higher aggregate scores are preferred.

On the primary suite, the baseline and validation-selected leaders differ under every examined condition. Under Definition (a), NT-v2 leads at the final stored state, whereas Evo 2 leads after validation selection, with top-1 bootstrap shares of 0.980 and 0.999. Group-balanced aggregation changes the baseline leader to NT-v3, while Evo 2 remains the validation-selected leader. Under the norm-based Definition (c), Kendall's τ increases from 0.289 to 0.600, and the number of CI-supported pairwise reversals decreases from ten to two. The baseline and validation-selected leaders still differ in 99.8% of task-bootstrap resamples.

On the external suite, leadership changes depend on the endpoint definition. Under Definition (a), NT-v3 leads at the baseline endpoint and Evo 2 leads after validation selection ($\tau = 0.022$; eight CI-supported reversals). Under Definition (c), Evo 2 leads under both conditions, τ increases to 0.689, and no pairwise reversals are CI-supported. The fraction of bootstrap resamples with differ-

ent leaders decreases from 0.918 to 0.082. Nevertheless, mean regret at the norm-based endpoint remains 0.0363 on the primary suite and 0.0228 on the external suite.

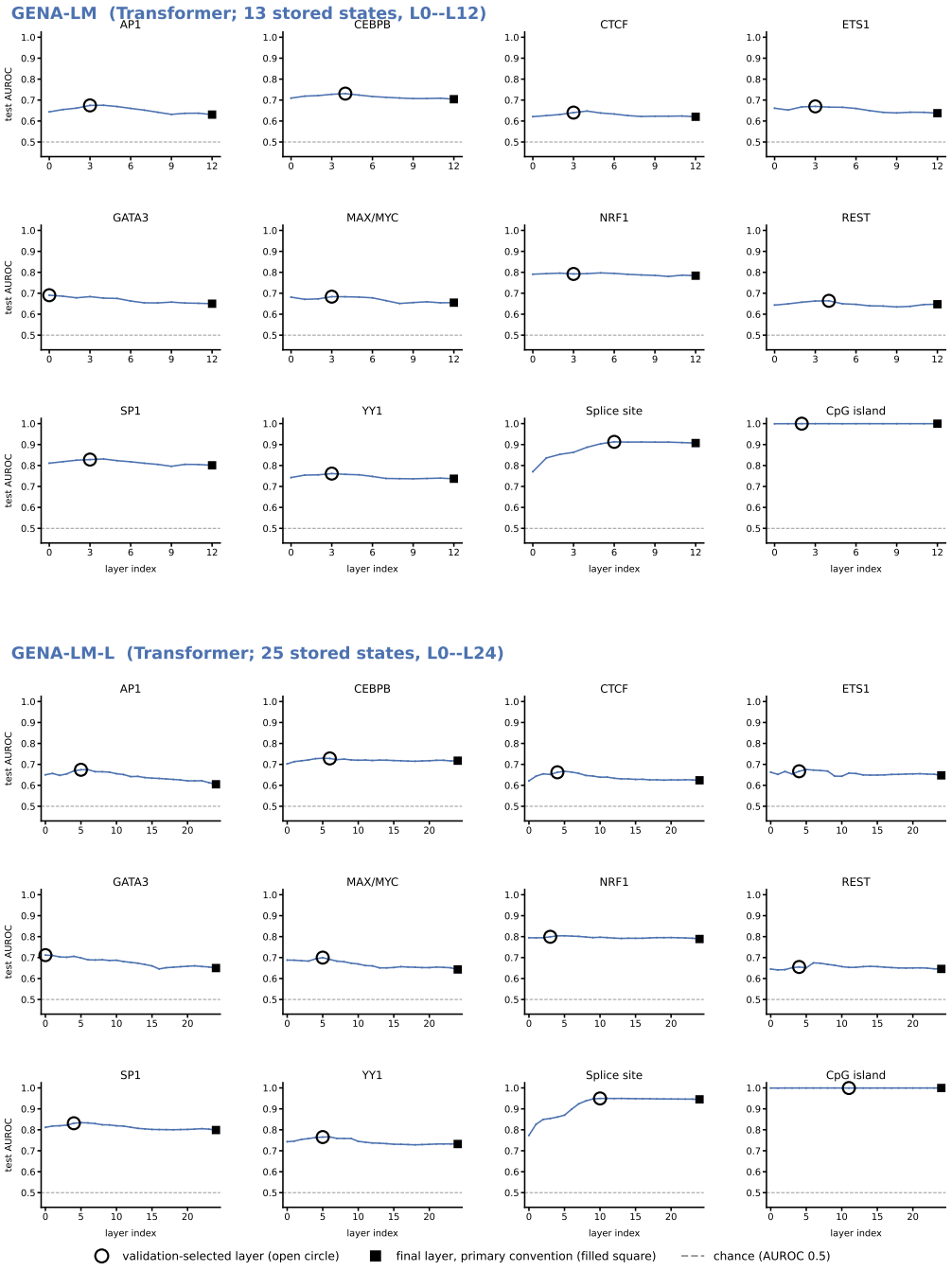

Figure B1: **Complete layerwise test-AUROC profiles for the 120 primary-suite model–task cells.** Page 1 of 5, with two models per page. The horizontal axis denotes the stored-state index. Open circles mark validation-selected states; filled squares mark final stored states under Definition (a). Dashed lines indicate chance performance (AUROC = 0.5).

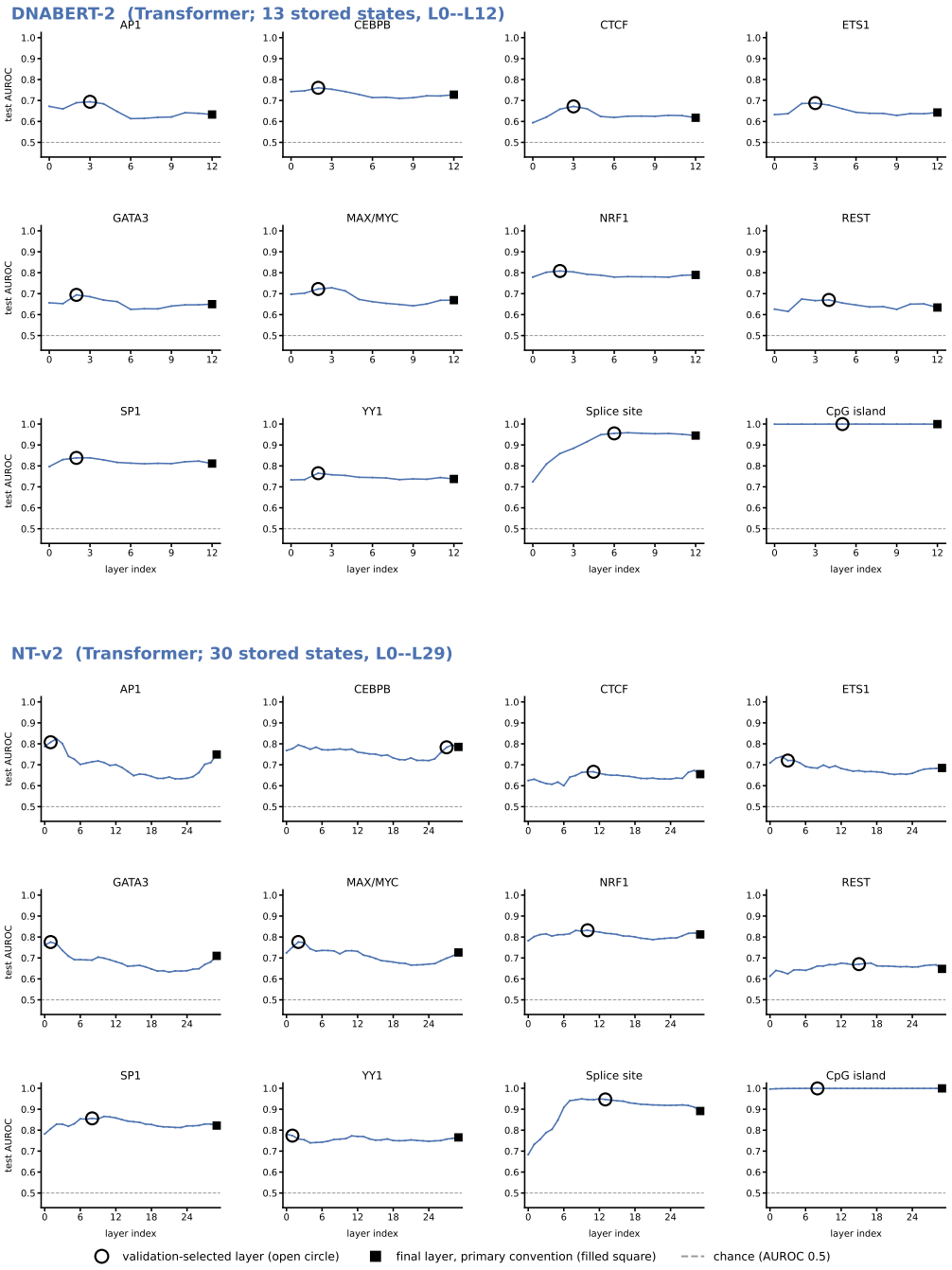

Figure B1 (continued): page 2 of 5.

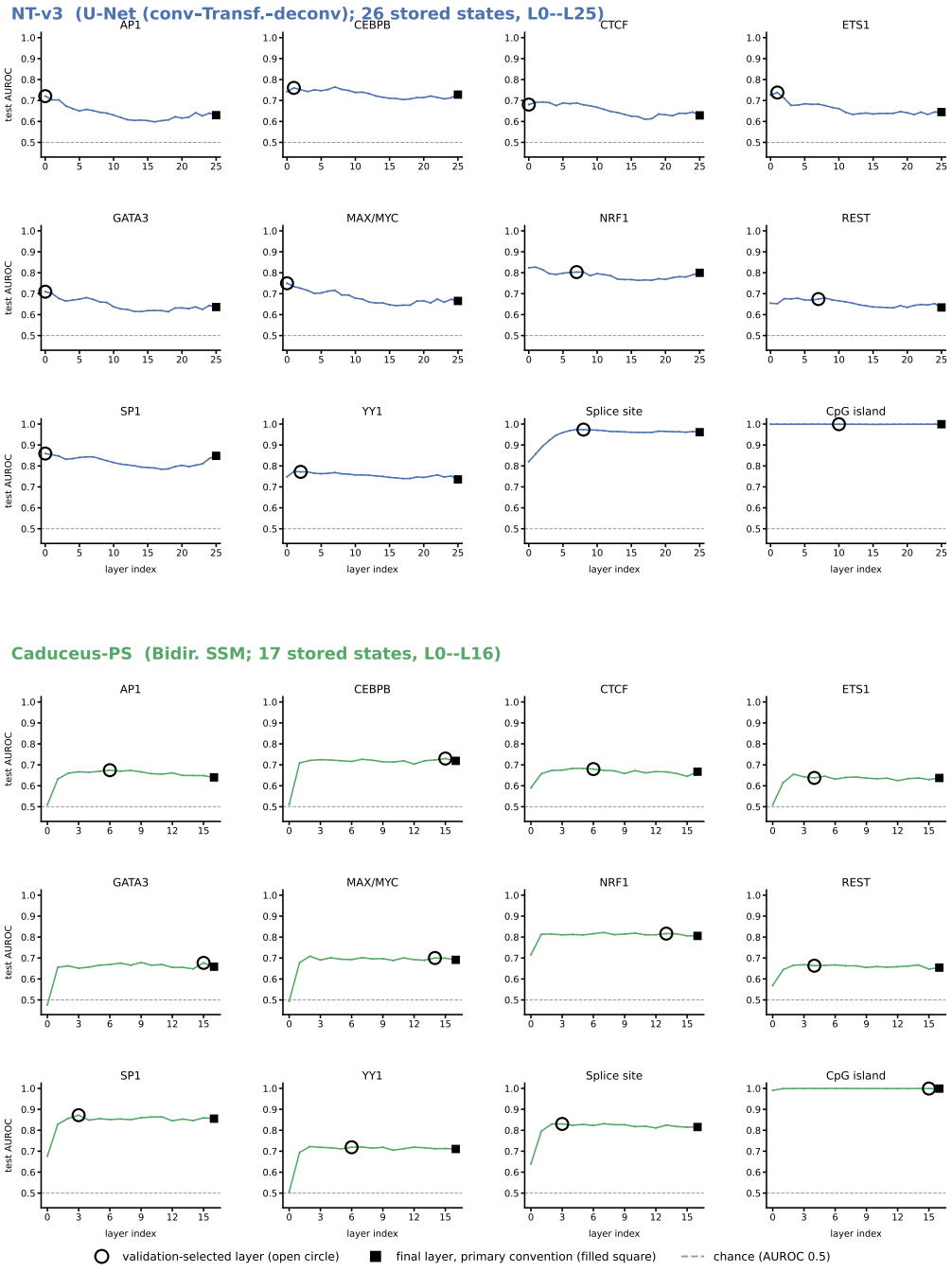

Figure B1 (continued): page 3 of 5.

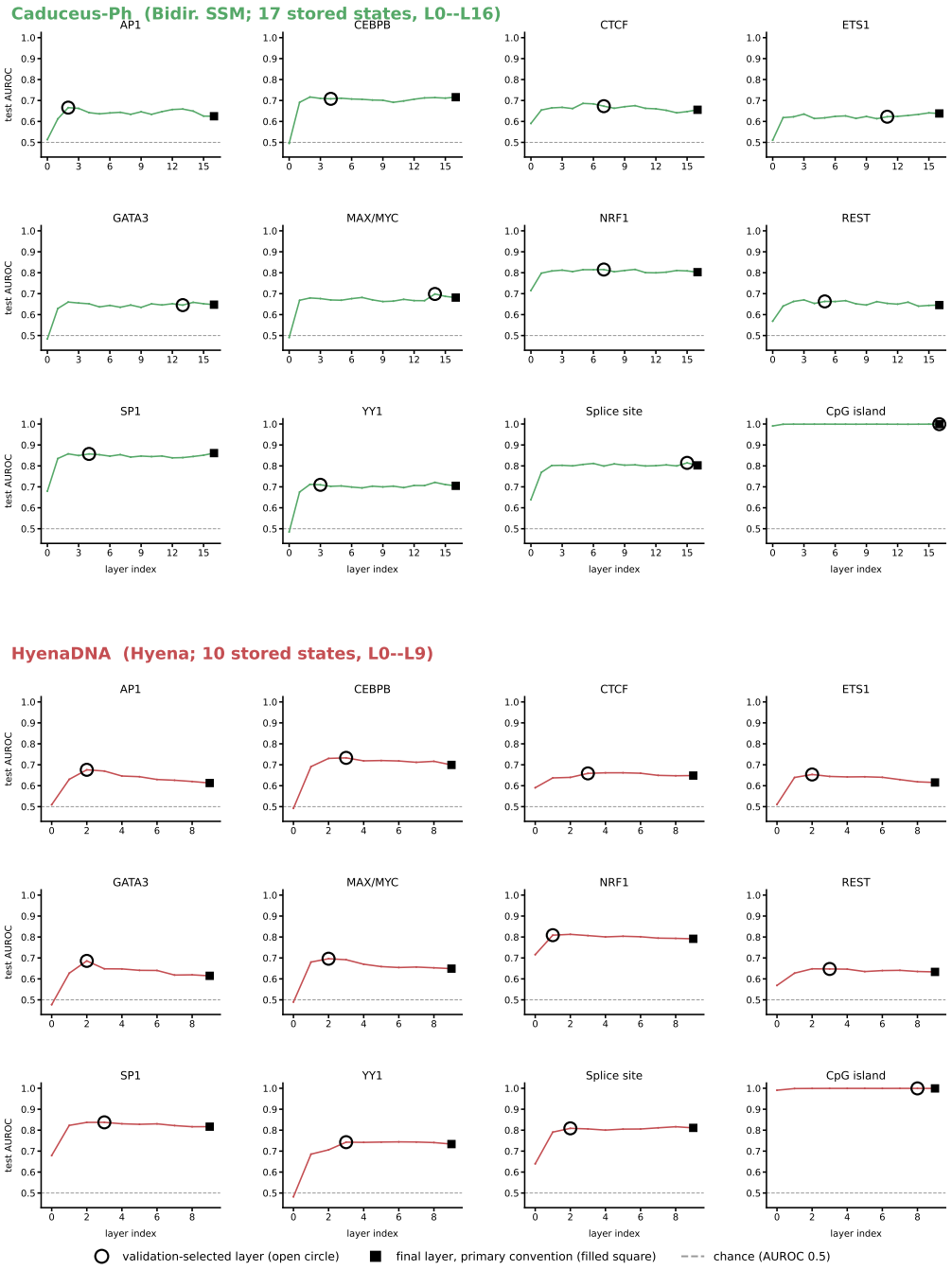

Figure B1 (continued): page 4 of 5.

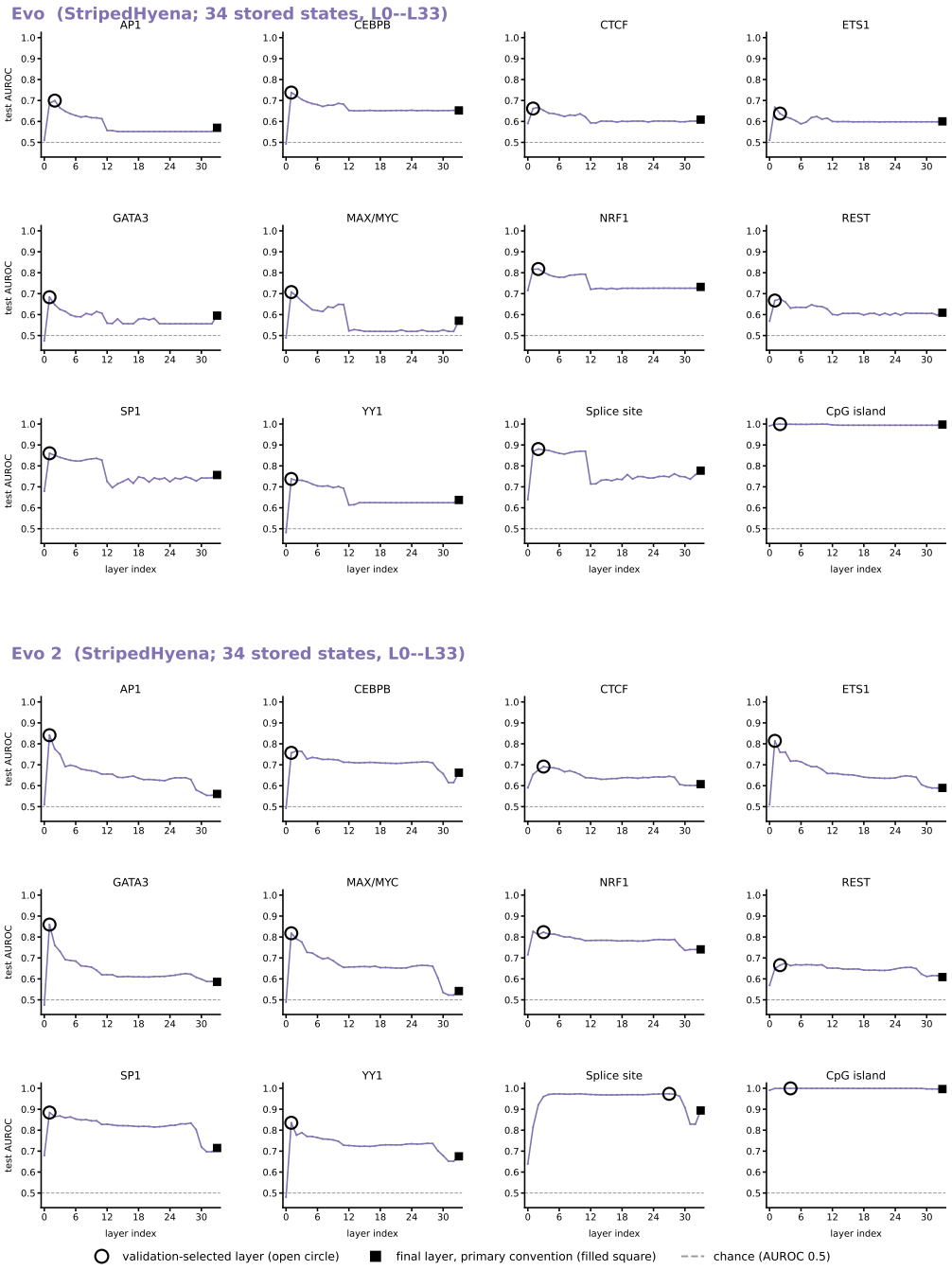

Figure B1 (continued): page 5 of 5.

Table B1: **Final-state and validation-selected performance on regression and multiclass tasks.** All reported scores are measured on the test split. Regression selection uses validation Pearson r; multiclass selection uses validation macro one-vs-rest AUROC. The score column reports Pearson r for regression and macro-F1 for classification. Δ is the validation-selected score minus the final-state score, computed before rounding.

	Final			Validation-selected			
Model	state	score	AUROC	state	score	AUROC	Δ
GC-content regression (score: Pearson r; ridge, $\alpha \in \{10^{-2}, \ldots, 10^{2}\}$)							
GENA-LM	L12	0.9931	—	L1	0.9980	—	+0.0049
GENA-LM-L	L24	0.9931	—	L1	0.9983	—	+0.0052
DNABERT-2	L12	0.9959	—	L0	0.9993	—	+0.0034
NT-v2	L29	0.9985	—	L27	0.9996	—	+0.0011
NT-v3	L25	0.9999	—	L0	1.0000	—	+0.0001
Caduceus-PS	L16	0.9999	—	L0	1.0000	—	+0.0001
Caduceus-Ph	L16	1.0000	—	L0	1.0000	—	+0.0000
HyenaDNA	L9	0.9975	—	L0	1.0000	—	+0.0025
Evo	L33	0.9885	—	L0	1.0000	—	+0.0115
Evo 2	L33	0.9897	—	L0	1.0000	—	+0.0103
Functional-element classification (5 classes; score: macro-F1; selection: validation macro-AUROC)							
GENA-LM	L12	0.6245	0.8744	L8	0.6267	0.8757	+0.0021
GENA-LM-L	L24	0.6612	0.8929	L10	0.6663	0.8956	+0.0051
DNABERT-2	L12	0.6404	0.8831	L10	0.6605	0.8918	+0.0201
NT-v2	L29	0.5752	0.8433	L14	0.6376	0.8827	+0.0624
NT-v3	L25	0.6731	0.8933	L9	0.6718	0.8994	-0.0013
Caduceus-PS	L16	0.5349	0.8148	L5	0.5411	0.8217	+0.0061
Caduceus-Ph	L16	0.5355	0.8149	L6	0.5358	0.8162	+0.0002
HyenaDNA	L9	0.5306	0.8088	L8	0.5312	0.8115	+0.0006
Evo	L33	0.4862	0.7810	L2	0.5691	0.8402	+0.0828
Evo 2	L33	0.5908	0.8480	L28	0.6981	0.9093	+0.1073
Transposable-element classification (6 classes; score: macro-F1; selection: validation macro-AUROC)							
GENA-LM	L12	0.6784	0.9041	L5	0.7141	0.9204	+0.0357
GENA-LM-L	L24	0.6411	0.9003	L8	0.7150	0.9248	+0.0739
DNABERT-2	L12	0.6488	0.8952	L5	0.7200	0.9254	+0.0712
NT-v2	L29	0.5668	0.8548	L14	0.6632	0.9055	+0.0964
NT-v3	L25	0.6332	0.8936	L7	0.6955	0.9219	+0.0623
Caduceus-PS	L16	0.5134	0.8210	L7	0.5565	0.8524	+0.0432
Caduceus-Ph	L16	0.5013	0.8175	L5	0.5505	0.8478	+0.0493
HyenaDNA	L9	0.5607	0.8524	L8	0.5668	0.8555	+0.0061
Evo	L33	0.4106	0.7571	L2	0.5721	0.8561	+0.1616
Evo 2	L33	0.4641	0.7933	L27	0.6836	0.9187	+0.2195

Table B2: **Ranking and final layer definition robustness: leadership.** Shares report top-1 bootstrap resample frequencies for leaders under final vs. validation-selected layers. $P(\text{flip})$ denotes the fraction of resamples where top-1 model identity changes. For the primary suite, three-group balancing assigns weights $(1/3, 1/3, 1/3)$ to PWM, splice, and CpG, whereas two-group reweighting assigns $(1/2, 1/4, 1/4)$; PWM tasks are equally weighted within their group.

Class	Condition	Baseline leader	Share (baseline)	Val. leader	Share (val.)	P(flip)
Primary suite (12 binary tasks)						
Primary analysis	definition (a): last stored state; macro aggregation	NT-v2	0.98	Evo 2	0.999	0.999
Definition sensitivity	definition (b): stored block-output endpoint	NT-v2	0.98	Evo 2	0.999	0.999
Definition sensitivity	definition (c): norm-based endpoint	NT-v2	0.97	Evo 2	1.00	0.998
Aggregation variant	three-group balanced aggregation	NT-v3	0.64	Evo 2	1.00	0.9996
Aggregation variant	Borda aggregation	NT-v2	0.98	Evo 2	0.96	0.965
Task-weighting sensitivity	two-group reweighting (0.5 PWM / 0.5 splice–CpG)	NT-v2	0.62	Evo 2	1.00	0.9996
External suite (16 tasks) — top-1 identity is definition-sensitive						
External-suite analysis	definition (a): last stored state; macro aggregation	NT-v3	0.986	Evo 2	0.918	0.918
Definition sensitivity	definition (b): stored block-output endpoint	NT-v3	0.99	Evo 2	0.92	0.918
Definition sensitivity	definition (c): norm-based endpoint	Evo 2	1.00	Evo 2	0.92	0.082
Aggregation variant	group-balanced aggregation	NT-v3	0.99	Evo 2	0.98	0.983
Aggregation variant	Borda aggregation	NT-v3	0.88	Evo 2	0.97	0.965

Table B3: **Ranking and final layer definition robustness: agreement, pairwise order changes, and regret.** Kendall's τ_b compares model rankings under baseline and validation-selected states. Pairwise order changes count changes in the sign of the between-model score difference, including transitions into or out of a tie. For Borda aggregation, the primary-suite count comprises 19 strict reversals and one tie transition; the external-suite count comprises 20 strict reversals and one tie transition. All other reported order changes are strict reversals. — means not computed for that condition. For the primary suite, three-group balancing assigns weights $(1/3, 1/3, 1/3)$ to PWM, splice, and CpG, whereas two-group reweighting assigns $(1/2, 1/4, 1/4)$; PWM tasks are equally weighted within their group.

Class	Condition	Kendall τ_b	Order Changes	CI-supported	Baseline regret
Primary suite (12 binary tasks)					
Primary analysis	definition (a): last stored state; macro aggregation	0.289	16/45	10	0.0471
Definition sensitivity	definition (b): stored block-output endpoint	0.289	16/45	10	0.0494
Definition sensitivity	definition (c): norm-based endpoint	0.600	9/45	2	0.0363
Aggregation variant	three-group balanced aggregation	0.511	11/45	7	—
Aggregation variant	Borda aggregation	0.135	20/45	9	—
Task-weighting sensitivity	two-group reweighting (0.5 PWM / 0.5 splice–CpG)	0.467	12/45	7	0.0346
External suite (16 tasks)					
External-suite analysis	definition (a): last stored state; macro aggregation	0.022	22/45	8	0.0457
Definition sensitivity	definition (b): stored block-output endpoint	0.022	22/45	10	0.0483
Definition sensitivity	definition (c): norm-based endpoint	0.689	7/45	0	0.0228
Aggregation variant	group-balanced aggregation	0.022	22/45	11	—
Aggregation variant	Borda aggregation	0.090	21/45	7	—

B.4 Association Versus Selection Regret

Figure B2 presents the association–selection comparison separately for each scored metric across the 120 primary-suite cells. For each cell, we compute Spearman correlation between the oriented metric scores and test AUROC across stored states, and the regret of the state maximizing the oriented score. The ID rule is excluded from this comparison because it selects from profile shape rather than by maximizing an oriented score.

Across 960 metric–cell evaluations, 18 cases combine high within-cell association ($\rho \geq 0.8$) with substantial selection regret (≥ 0.02 AUROC), spanning seven of the eight scored metrics. High profile correlation can coexist with substantial selection regret. Correlation summarizes agreement across the full ordering of states, whereas regret depends on the test performance of the single selected state. The two quantities therefore assess different aspects of a metric's usefulness for layer selection.

B.5 Near-Optimal Sets and Performance Tolerance

Exact agreement with the diagnostic test-optimal state does not account for performance plateaus. Table B4 therefore reports both exact-hit rates and the fraction of selections whose test AUROC lies within 0.005 of the diagnostic maximum. This tolerance is measured relative to the test optimum, rather than to full validation.

For full validation, the hit rate increases from 0.450 under exact matching to 0.750 under the 0.005 tolerance. For the final stored state, it increases from 0.008 to 0.117. The 0.005 near-optimal sets contain a mean of 3.94 states and a median of two, with a mean within-cell fraction of 20.2% of stored states. Using one bootstrap standard error of the diagnostic maximum as the tolerance yields a mean of 3.52 states and a median of three. The two set definitions have a mean within-cell Jaccard overlap of 0.698.

B.6 Complete Screening Results

Figure B3 reports all screening fractions from 10% to 90% in increments of 10 percentage points. Tables B5 and B6 provide numerical results at all nine screening fractions for the primary and external suites, respectively. Mean regret generally decreases as more states are probed, although test performance need not improve monotonically because the final selection is based on validation scores.

On the primary suite, LogME yields the lowest mean regret among the evaluated rules when probing 10% of stored states (0.0073). Depth Prior (0.0115) performs similarly to H-score (0.0119) and better than prompt entropy (0.0165). At 30%, prompt entropy (0.0049), LogME (0.0053), and Early Layers (0.0058) all yield lower mean regret than H-score (0.0070). These comparisons show that training-label guidance does not uniformly outperform metric-free depth rules.

Uniform spacing and coarse-to-fine screening yield mean regrets of 0.0486 and 0.0539 at 10%, decreasing to 0.0077 and 0.0063 at 50%. Random subsets average 0.0281 regret at 10% and 0.0082 at 50% across 100 seeds. The fixed depth band remains at 0.0188 regret from 30% onward because its candidate pool is exhausted. In comparison, Early Layers and Depth Prior reach mean regrets of 0.0035 and 0.0034 at 50%, respectively, against 0.0034 for full validation.

TransRate and the ID guide retain larger performance gaps. On the primary suite at 50%, their mean regrets are 0.0145 and 0.0120, compared with 0.0050 for LogME and 0.0071 for H-score. On the external suite, the ID guide at the same fraction has a mean test-AUROC loss of 0.0023 relative to full validation (95% CI $[0.0003, 0.0052]$).

External depth-rule results. When probing 50% of stored states, both depth rules yield mean regret of 0.0042, compared with 0.0039 for full validation. Their mean paired test-AUROC losses are both 0.0003, with crossed-bootstrap 95% CIs of $[-0.0009, 0.0014]$ for Early Layers and $[-0.0009, 0.0013]$ for the transferred Depth Prior. Losses exceed 0.005 AUROC in 6.9% and 7.5% of cells, respectively. The largest task-family mean losses are 0.0018 for Early Layers on enhancer tasks and 0.0013 for Depth Prior on TFBS tasks. The two rules select the full-validation layer itself in 76% and 75% of cells, respectively, while probing 10.7 states on average.

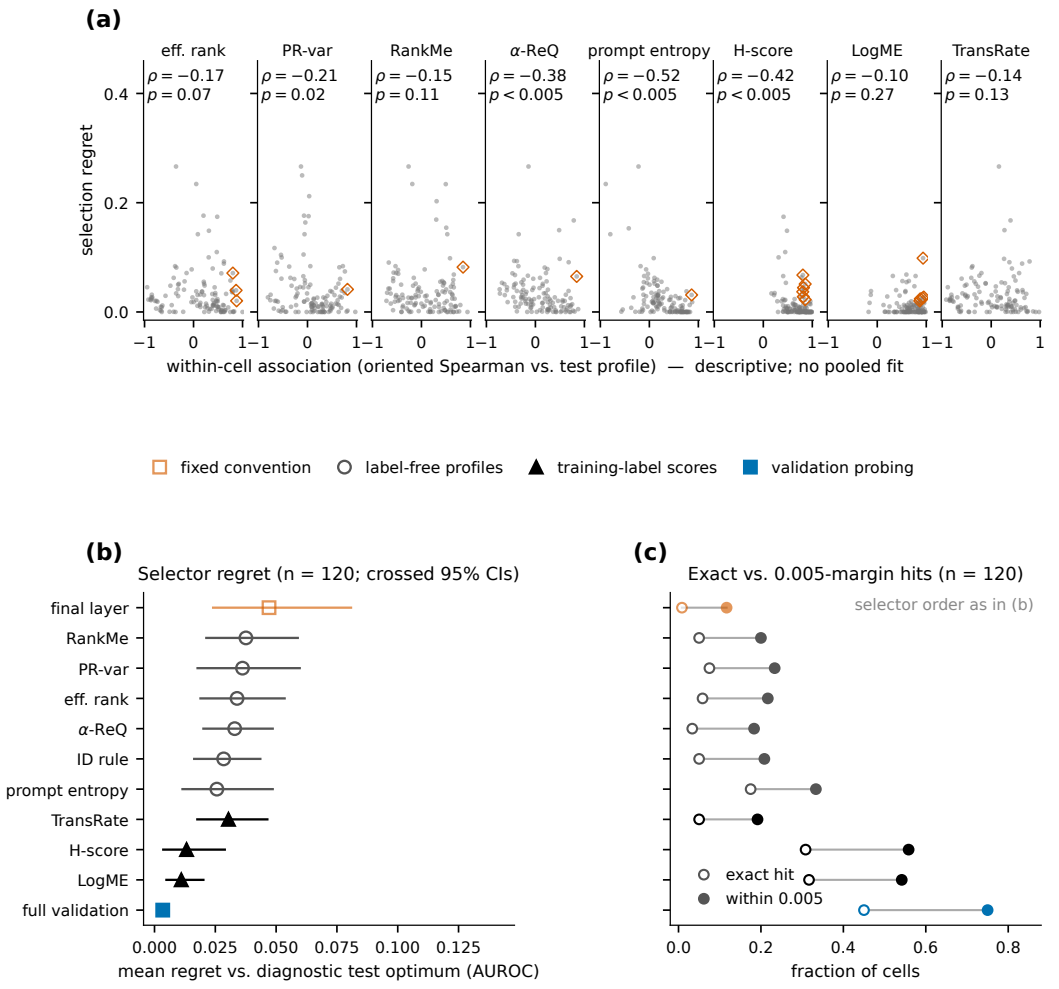

Figure B2: **Profile association, selection regret, and hit rates.** **(a)** Within-cell Spearman correlation between oriented metric scores and test AUROC versus the regret of the metric-selected state. Panel annotations summarize cross-cell association between the two plotted quantities. Diamonds mark cases with $\rho \geq 0.8$ and regret ≥ 0.02 AUROC. **(b)** Mean selection regret with 95% crossed model–task bootstrap confidence intervals. **(c)** Exact-match and within-0.005-AUROC hit rates relative to the diagnostic test optimum. Results cover 120 primary-suite model–task cells.

Table B4: **Standalone selector regret, exact hits, margin hits, and near-optimal set structure** (120 primary model–task cells). **(A)** Mean regret with crossed model×task cluster-bootstrap 95% intervals, the exact hit rate (selected layer = diagnostic optimum), and the within-0.005 AUROC margin hit rate by selector. Exact hit measures agreement with the first test-score maximizer in the stored-state ordering. Margin hit counts any selected state within 0.005 AUROC of the maximum. Gain denotes the selected-state test AUROC minus the final-state test AUROC, with paired crossed-bootstrap 95% confidence intervals. **(B)** Average size and stored layer fraction of 0.005-margin and 1-SE near-optimal sets.

(A) Regret and hit rates by selector ($n = 120$)

Selector	Mean regret	95% CI	Exact hit	Within 0.005	Gain vs. final (95% CI)
final layer	0.0471	[0.0241, 0.0808]	0.008	0.117	—
eff. rank	0.0339	[0.0189, 0.0535]	0.058	0.217	+0.0131 [−0.0125, +0.0370]
PR-var	0.0362	[0.0177, 0.0597]	0.075	0.233	+0.0109 [−0.0086, +0.0276]
RankMe	0.0376	[0.0213, 0.0589]	0.050	0.200	+0.0095 [−0.0112, +0.0277]
α-ReQ	0.0330	[0.0201, 0.0486]	0.033	0.183	+0.0141 [−0.0047, +0.0371]
ID rule	0.0284	[0.0163, 0.0435]	0.050	0.208	+0.0187 [+0.0005, +0.0455]
prompt entropy	0.0257	[0.0126, 0.0457]	0.175	0.333	+0.0214 [−0.0098, +0.0551]
TransRate	0.0304	[0.0176, 0.0464]	0.050	0.192	+0.0167 [−0.0003, +0.0378]
H-score	0.0132	[0.0036, 0.0289]	0.308	0.558	+0.0339 [+0.0188, +0.0530]
LogME	0.0110	[0.0049, 0.0201]	0.317	0.542	+0.0361 [+0.0171, +0.0631]
full validation	0.0034	[0.0016, 0.0057]	0.450	0.750	+0.0437 [+0.0207, +0.0779]

(B) Near-optimal-set structure ($n = 120$)

	0.005 margin set	1-SE set
mean layers in set	3.942	3.517
mean fraction of stored layers	0.202	0.194
median layers in set	2	3

Jaccard agreement between the two set definitions: mean 0.698, median 1.000.
Mean optimum-layer bootstrap SE: 0.0074.

Additional external metric-guide analysis. At a 30% screening fraction, mean excess regret relative to full validation is 0.0049 for effective rank, 0.0041 for prompt entropy, and 0.0046 for RankMe. For this analysis, model-clustered bootstrap 90% CIs are $[0.0029, 0.0071]$, $[0.0019, 0.0069]$, and $[0.0025, 0.0069]$, respectively. These intervals use a different resampling scheme from the crossed-bootstrap 95% intervals reported above. Secondary cell-level TOST analyses with an equivalence margin of ± 0.005 AUROC yield $p = 0.462$, $p = 0.177$, and $p = 0.291$, respectively. None establishes equivalence at this margin.

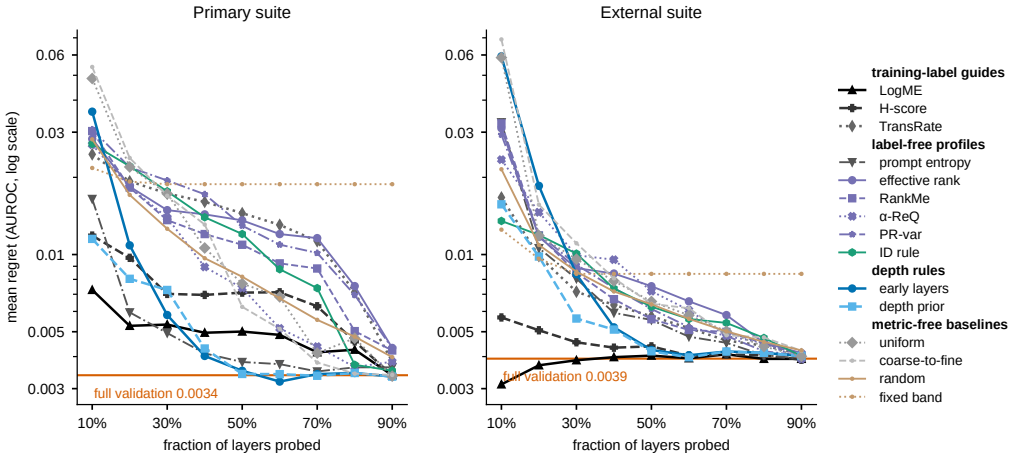

Figure B3: **Layer screening across all nomination rules.** Mean regret versus the fraction of stored states probed on the primary suite (left, $n = 120$) and external suite (right, $n = 160$). Screening fractions range from 10% to 90% in increments of 10 percentage points. Colors distinguish groups of nomination rules, and line styles and markers identify individual rules. Horizontal lines indicate full validation. Random-subset results are averaged across 100 seeds.

B.7 Sensitivity of Diagnostic Optimal Depth to Input Length

We compare diagnostic test-optimal depths between 70-bp and 300-bp promoter task variants across ten models and three promoter task types, yielding 30 paired model–task comparisons (Table B7). Because depths are selected using test performance, this analysis characterizes observed optima rather than evaluating a deployable layer-selection rule.

The median signed shift in relative depth (300 bp minus 70 bp) is -0.032, while the mean is -0.081. Agreement within one stored layer occurs in 11 of 30 pairs; agreement within 0.10 relative depth also occurs in 11 of 30 pairs. Mean absolute depth shifts are 0.143 for the combined promoter task, 0.303 for non-TATA promoters, and 0.286 for TATA promoters. Thus, the small aggregate signed shift coexists with substantial variation across individual comparisons. These results caution against assuming that the selected depths transfer unchanged between the two input-length variants.

Table B5: **Layer screening on the primary suite.** Entries report mean regret across 120 model–task cells at screening fractions from 10% to 90% in 10-percentage-point increments. **(A)** Metric-guided screening. **(B)** Metric-free baselines and depth rules. Full-layer validation yields mean regret of 0.0034. The ID guide nominates states nearest its selected layer. Random-subset results are averaged over 100 seeds.

(A) Metric-guided screening (primary suite)

Guide	10%	20%	30%	40%	50%	60%	70%	80%	90%
LogME	0.0073	0.0053	0.0053	0.0050	0.0050	0.0049	0.0042	0.0043	0.0034
H-score	0.0119	0.0097	0.0070	0.0070	0.0071	0.0071	0.0063	0.0046	0.0034
TransRate	0.0246	0.0193	0.0173	0.0160	0.0145	0.0131	0.0113	0.0071	0.0038
prompt entropy	0.0165	0.0060	0.0049	0.0041	0.0038	0.0037	0.0035	0.0036	0.0036
effective rank	0.0283	0.0184	0.0149	0.0143	0.0136	0.0120	0.0116	0.0075	0.0043
RankMe	0.0303	0.0182	0.0136	0.0120	0.0109	0.0092	0.0088	0.0050	0.0042
α-ReQ	0.0268	0.0183	0.0139	0.0089	0.0074	0.0052	0.0044	0.0036	0.0038
PR-var	0.0308	0.0219	0.0194	0.0172	0.0130	0.0109	0.0101	0.0070	0.0043
ID rule	0.0270	0.0221	0.0176	0.0140	0.0120	0.0088	0.0074	0.0037	0.0035

(B) Metric-free baselines and depth rules (primary suite)

Rule	10%	20%	30%	40%	50%	60%	70%	80%	90%
uniform spacing	0.0486	0.0220	0.0173	0.0106	0.0077	0.0068	0.0041	0.0047	0.0034
coarse-to-fine	0.0539	0.0238	0.0170	0.0131	0.0063	0.0051	0.0038	0.0034	0.0034
random subsets	0.0281	0.0171	0.0126	0.0097	0.0082	0.0067	0.0056	0.0048	0.0040
fixed depth band [0.2,0.5]	0.0218	0.0193	0.0188	0.0188	0.0188	0.0188	0.0188	0.0188	0.0188
Depth rules									
early layers	0.0361	0.0109	0.0058	0.0040	0.0035	0.0032	0.0034	0.0035	0.0033
depth prior	0.0115	0.0080	0.0073	0.0043	0.0034	0.0034	0.0034	0.0035	0.0033

Table B6: **Layer screening on the external suite.** Entries report mean regret across 160 model–task cells at screening fractions from 10% to 90% in 10-percentage-point increments. **(A)** Metric-guided screening. **(B)** Metric-free baselines and depth rules. Full-layer validation yields mean regret of 0.0039. The Depth Prior is transferred from the primary suite; training-label guides use external-suite training labels. The ID guide nominates states nearest its selected layer. Random-subset results are averaged over 100 seeds.

(A) Metric-guided screening (external suite)

Guide	10%	20%	30%	40%	50%	60%	70%	80%	90%
LogME	0.0031	0.0037	0.0039	0.0040	0.0040	0.0039	0.0041	0.0039	0.0039
H-score	0.0057	0.0051	0.0045	0.0043	0.0044	0.0040	0.0041	0.0040	0.0039
TransRate	0.0167	0.0099	0.0072	0.0064	0.0057	0.0052	0.0046	0.0044	0.0041
prompt entropy	0.0328	0.0106	0.0080	0.0059	0.0056	0.0048	0.0045	0.0040	0.0039
effective rank	0.0311	0.0117	0.0089	0.0084	0.0075	0.0066	0.0058	0.0045	0.0041
RankMe	0.0325	0.0119	0.0085	0.0067	0.0056	0.0051	0.0050	0.0045	0.0041
α-ReQ	0.0234	0.0146	0.0099	0.0095	0.0072	0.0061	0.0048	0.0045	0.0039
PR-var	0.0293	0.0120	0.0091	0.0074	0.0061	0.0052	0.0048	0.0042	0.0040
ID rule	0.0135	0.0119	0.0101	0.0074	0.0063	0.0056	0.0054	0.0047	0.0040

(B) Metric-free baselines and depth rules (external suite)

Rule	10%	20%	30%	40%	50%	60%	70%	80%	90%
uniform spacing	0.0586	0.0119	0.0096	0.0079	0.0066	0.0058	0.0050	0.0044	0.0040
coarse-to-fine	0.0690	0.0156	0.0111	0.0081	0.0064	0.0062	0.0048	0.0048	0.0042
random subsets	0.0215	0.0109	0.0086	0.0072	0.0064	0.0056	0.0050	0.0046	0.0042
fixed depth band [0.2,0.5]	0.0125	0.0096	0.0084	0.0084	0.0084	0.0084	0.0084	0.0084	0.0084
Depth rules									
early layers	0.0593	0.0185	0.0084	0.0052	0.0042	0.0040	0.0042	0.0041	0.0040
depth prior	0.0157	0.0098	0.0056	0.0051	0.0042	0.0040	0.0042	0.0041	0.0040

Table B7: **Diagnostic optimal depths across input-length variants.** Results cover 30 paired comparisons across ten models and three promoter task types at 70 bp and 300 bp. Signed shifts are computed as the diagnostic test-optimal relative depth at 300 bp minus that at 70 bp. Near-optimal set: within 0.005 AUROC of the cell maximum; all bootstrap intervals are 95% percentile (10,000 resamples).

Quantity (30 matched 70/300-bp promoter pairs; diagnostic test-optimal depths)	Value
Signed depth shift (300 bp − 70 bp), mean	-0.081 (TOST 90% CI of the mean $[-0.184, +0.022]$)
Signed depth shift, median	-0.032 (bootstrap 95% CI of the median $[-0.139, +0.020]$)
Selected-depth agreement within ± 1 layer	11/30
Selected-depth agreement within $\pm 10\%$ depth	11/30
$\lvert\Delta\text{layer}\rvert$, mean / median	5.57 / 2.0
Near-optimal-set overlap (Jaccard), mean / median	0.330 / 0.310
TOST at predeclared ± 0.10 depth margin	t-based $p = 0.38$; Wilcoxon-based $p = 0.21$
Per-type $\lvert\Delta\text{depth}\rvert$ means (combined / non-TATA / TATA)	0.143 / 0.303 / 0.286

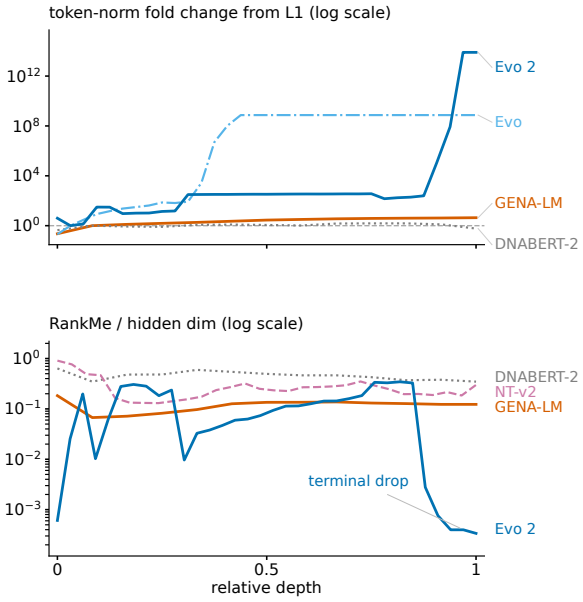

Figure B4: **Layerwise token norms and RankMe profiles. Top:** Mean token-norm fold change relative to L_1 (log scale). Evo shows an abrupt increase around relative depth 0.3, followed by a plateau, whereas Evo 2 shows a sharp increase near the end of the stack. GENA-LM exhibits a comparatively modest, gradual increase. **Bottom:** RankMe normalized by hidden dimension (log scale; Evo 2 exhibits a terminal drop).

C PIPELINE VALIDATION

We examine probe conditioning, tokenization, and numerical precision to assess the reliability of the evaluation pipeline.

Probe conditioning and regularization. Features are standardized using training-split statistics before probe fitting (Appendix A.3). On the GC-content regression task, test correlations from ridge probes fitted with and without standardization differ by less than 0.005 at every stored state. This check indicates limited sensitivity to standardization for this task. For each stored-state–task pair, iteration counts are recorded only for the probe fitted with the validation-selected C. Of these 2,628 probe fits, 10 (0.4%) reach the 1,000-iteration `lbfgs` cap, with five cases each in Evo and Evo 2. Iteration counts for the remaining regularization-grid fits are not retained. Boundary-selection frequencies and the expanded-grid sensitivity check are reported in Appendix A.3.

Tokenization checks. We compare tokenization outputs against official model implementations using twenty diagnostic sequences, including synthetic sequences enriched in `N`. Mean pooling excludes padding positions using the attention mask and additionally excludes special-token positions using an explicit token-ID mask before averaging. Special-token IDs are collected from the tokenizer's declared special tokens, with a model-specific padding ID for Evo 2. For GENA-LM, batched tokenization reproduces the official token sequences, including `[CLS]` and `[SEP]` framing and BPE vocabulary mapping. The checked inputs are not truncated at primary-suite sequence lengths.

Precision sensitivity of Evo and Evo 2. We repeat extraction on a diagnostic pool of 500 sequences of length 256 bp, using an `fp32` residual stream while retaining 16-bit attention computation (Table C1). Peak token-norm fold changes relative to L_1 remain similar between native `bf16` and `fp32` residual execution: 7.25×10^8 versus 7.30×10^8 for Evo, and 6.04×10^{13} versus 5.99×10^{13} for Evo 2. Thus, the observed norm growth persists with an `fp32` residual stream.

In a separate check at 1,000 bp, the bitwise identity between Evo 2's stored L_{31} and L_{32} states occurs in `bf16` but disappears with an `fp32` residual stream. Across the 256-bp diagnostic pool, the lowest layerwise mean cosine similarities between representations from the two precision settings are 0.994 for Evo and 0.987 for Evo 2. These checks support the precision robustness of the norm-growth observation, but do not directly establish unchanged downstream probe performance.

Table C1: **Precision sensitivity of Evo and Evo 2 representations** (500 diagnostic sequences, 256 bp; all stored layers). `fp32` denotes full-precision residual stream execution with 16-bit attention core computation. Stored $L_{31} \equiv L_{32}$ identity is evaluated at 1,000 bp sequence length.

	Evo		Evo 2	
	bf16	fp32	bf16	fp32
peak token-norm fold vs. L1	7.25×10^8	7.30×10^8	6.04×10^{13}	5.99×10^{13}
final-norm state fold vs. L1	23.3	23.4	216.7	215.2
L31 ≡ L32 (bitwise, 1,000 bp)	—	—	yes	no
lowest layer mean / lowest sequence cosine (bf16 vs. fp32)	0.994 (L_1) / 0.611 (L_5)		0.987 (L_{25}) / 0.949 (L_{25})	