DNA-RADIO: TOWARDS AN AGGLOMERATIVE DNA EMBEDDING MODEL

Anonymous authors
Paper under double-blind review

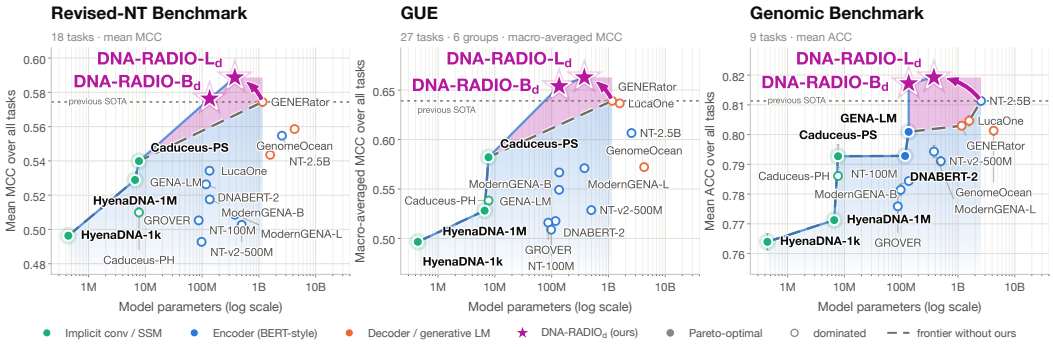


Figure 1: **DNA-RADIO sets a new Pareto-frontier of DNA embedding models.** We evaluate 16 models on 54 tasks across Revised-NT, GUE, and Genomic Benchmark via linear probing. By distilling multiple DNA language models, our 136M (DNA-RADIO-B_d) and 377M (DNA-RADIO-L_d) models achieve SOTA performance with up to 29.4× fewer parameters.

ABSTRACT

DNA language models (LMs) achieve strong genomic prediction performance, yet their strengths remain distributed across models, with no single model consistently dominating. Through layer-wise linear probing and centered kernel alignment, we find that downstream performance often saturates early and deeper layers contribute little to additional downstream-relevant signal, suggesting that these complementary capabilities can be consolidated into a compact representation. We introduce DNA-RADIO, an agglomerative embedding model that distills heterogeneous DNA LMs into a single student up to 29.4× smaller than its teachers. We then enrich its representations through post-training on large-scale labeled sequences while preserving inherited knowledge of distillation. Under a frozen-feature linear evaluation protocol, distillation using only unlabeled sequences improves average performance by up to 15.85% over its teachers and other sequence-only DNA LMs on held-out sequences, establishing state-of-the-art results despite its substantially smaller size. Post-training yields a further improvement on BEND, a benchmark held out in its entirety with its sequences excluded from training. These results establish multi-model distillation and post-training as an effective route to compact, unified, general-purpose DNA representations.

1 INTRODUCTION

Deoxyribonucleic acid (DNA) is the primary carrier of genetic information in most living organisms and is commonly represented as a long sequence over four nucleotide bases: A, T, G, and C. Motivated by structural analogies between natural language and DNA sequences (Mantegna et al., 1994; Searls, 1997), as well as the success of language models (LMs) (Vaswani et al., 2017; Brown et al., 2020), DNA LMs have adapted LM pre-training objectives, such as masked language modeling (Devlin et al., 2019) and autoregressive generation (Radford et al., 2018), to base sequences. Through pre-training on large-scale DNA corpora and subsequent transfer to downstream tasks, DNA LMs have achieved strong performance across a wide range of genomic prediction problems.

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.

However, recent studies (Ledneva et al., 2026) show that existing DNA LMs excel at different tasks rather than exhibiting uniformly strong performance: their strengths vary substantially across benchmarks, with no model consistently dominating across tasks. This is perhaps unsurprising, as DNA LMs differ in tokenization, architecture, training objective and corpora, each of which induces different inductive biases. As a consequence, no single model can currently serve as a general-purpose DNA representation, and the complementary knowledge spread across models remains unexploited.

To understand how existing DNA LMs use their capacity, we analyze the intermediate representations of diverse DNA LMs through layer-wise linear probing across multiple benchmarks and layer-wise centered kernel alignment (CKA) (Kornblith et al., 2019). We find that downstream performance often saturates at relatively early layers, while early-layer representations remain highly similar to the final layer. These results suggest that existing DNA LMs do not fully utilize their depth to progressively refine downstream-relevant representations; consistent with this view, reported correlations between model size and downstream performance in DNA LMs are mixed and often weak (Ledneva et al., 2026; Patel et al., 2024). We refer to this redundancy as unused capacity, in the sense that deeper layers contribute little additional downstream-relevant signal, which suggests that the knowledge held by a DNA LM can be represented more compactly.

These observations are complementary: the first motivates combining DNA LMs, while the second suggests their combined knowledge can fit in a compact model. Building on both, we propose DNA-RADIO, which consolidates knowledge from heterogeneous DNA LMs and additional training signals into one compact model. Inspired by agglomerative models (Ranzinger et al., 2024b; Heinrich et al., 2025; Ranzinger et al., 2026), DNA-RADIO distills multiple DNA LMs with different tokenizers, architectures, and objectives into a student up to $29.4\times$ smaller than its teachers, leveraging their underused capacity to combine capabilities unavailable in any individual model. We then introduce a post-training stage, inspired by general-purpose text embedding models (Wang et al., 2022; Lee et al., 2025; Zhang et al., 2025), that uses diverse labeled sequences to enrich the representation while preserving knowledge from pre-training and multi-teacher distillation.

Since DNA-RADIO targets general-purpose representations, we follow the standard self-supervised foundation model evaluation protocol (Caron et al., 2021; Oquab et al., 2024; Assran et al., 2023)-linear probing or zero-shot. Distillation uses only unlabeled sequences, holding out entire chromosomes and masking benchmark test loci. DNA-RADIO improves average performance up to 15.85% over sequence-only DNA LMs on Revised-NT Benchmark (Dalla-Torre et al., 2025), GUE (Zhou et al., 2024), and Genomic Benchmark (Grešová et al., 2023), surpassing its strongest teacher despite being far smaller (up to $29.4\times$ smaller). We further evaluate post-training on the fully held-out BEND (Marin et al., 2024), excluding its sequences from the training through coordinate-overlap, match, and reverse-complement checks. Post-training improves BEND performance, exceeding baselines, indicating transfer to held-out tasks rather than fitting to the post-training distribution. Together, these results show that multi-model distillation and large-scale post-training yield unified, general-purpose DNA embeddings. Our contributions can be summarized as follows:

- We found that existing DNA LMs underuse their depth with layer-wise probing and CKA, suggesting compact downstream-relevant knowledge.
- We introduce DNA-RADIO, which consolidates heterogeneous DNA LMs into a single model 3.24-$29.4\times$ smaller than its teacher by aligning heterogeneous tokenizations on nucleotide intervals, and refines it with information-constrained relational post-training.
- We show that each DNA-RADIO training stage consistently improves baselines across four eukaryotic benchmarks comprising 61 tasks: distillation achieves SOTA performance, while post-training yields further gains on a fully held-out benchmark.

2 Related Work

2.1 DNA Language Models

As language models (LMs) have advanced, successful LM pre-training strategies have been adapted to DNA LMs for diverse downstream tasks. DNABERT series (Ji et al., 2021; Zhou et al., 2024) adopted masked language modeling (MLM) (Devlin et al., 2019) and different tokenization (*e.g.,* k-mer and BPE) to learn DNA grammar and structural patterns, while the Nucleotide Transformer

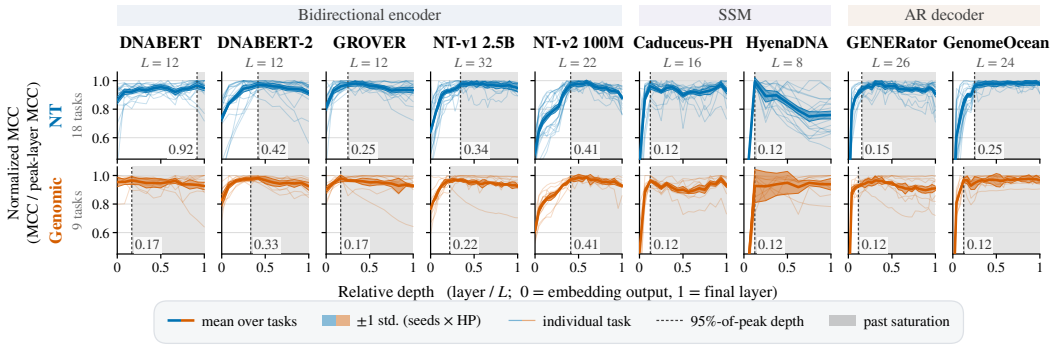


Figure 2: **Layer-wise linear probing of nine DNA language models on two benchmarks.** Each panel shows one model on NT Benchmark (18 tasks) and Genomic Benchmark (9 tasks). The x-axis is relative depth (layer/L); the y-axis is test MCC normalized by each task's best layer. Thin lines show tasks, bold lines their mean, and shading ± 1 s.d. over 3 seeds $\times$ the hyperparameter grid, pooled across tasks. The dashed line marks the first layer reaching 95% of peak mean performance; later layers (shaded) yield no gain. Probes use mean-pooling for encoders and last-token for others.

series (Dalla-Torre et al., 2025; Boshar et al., 2025) introduced nucleotide-level tokenization for fine-grained prediction. Other work has explored architectures for long-context modeling (Nguyen et al., 2023; Schiff et al., 2024), and generative auto-regressive training (Wu et al., 2025; Brixi et al., 2026). However, recent studies (Patel et al., 2024; Ledneva et al., 2026) show that downstream performance varies substantially with training objectives and architectures and, unlike natural-language LMs, does not consistently scale with model size or training data.

2.2 Embedding Models

Learning embeddings that faithfully represent modalities such as images and text has been extensively studied. For images, prior work has learned transferable representations through self-supervised learning (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025) and weak supervision from image-text pairs (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025). More recently, motivated by the complementary strengths of image embedding models across tasks, the RADIO series (Ranzinger et al., 2024b; Heinrich et al., 2025; Ranzinger et al., 2026) distills multiple models into a unified representation. Similarly, pre-trained LMs have been post-trained on large-scale labeled data to produce general-purpose text embeddings that transfer across tasks with little or no task-specific adaptation (Wang et al., 2022; Lee et al., 2025; Zhang et al., 2025). In contrast, DNA LMs are typically evaluated through full fine-tuning on large task-specific datasets, obscuring the transferability of their pretrained representations (Patel et al., 2024; Ledneva et al., 2026). Despite complementary strengths across task categories, no prior work consolidates multiple heterogeneous DNA LMs into a single embedding model.

3 Method

3.1 How Much Representational Capacity Do DNA Language Models Use?

Viewing DNA as language, prior work has adapted successful LLM architectures and pre-training objectives, yet model size translates weakly into downstream performance. Caduceus (Schiff et al., 2024) has approximately 2M parameters versus Evo-1 (Nguyen et al., 2024) with 7B—**3,627$\times$ larger**—but differs by less than **3%** on some tasks (Ledneva et al., 2026), suggesting that what these models actually use may be smaller than their size. We investigate this behavior via layer-wise linear probing of nine DNA LMs spanning diverse data formats, objectives, and architectures: bidirectional encoders (Ji et al., 2021; Zhou et al., 2024; Sanabria et al., 2024; Dalla-Torre et al., 2025), state-space models (SSMs) (Schiff et al., 2024; Nguyen et al., 2023), and autoregressive (AR) decoders (Wu et al., 2025; Zhou et al., 2025). On the Nucleotide Transformer Benchmark

(NT) (Dalla-Torre et al., 2025) and Genomic Benchmark (Grešová et al., 2023), each model saturates at its own characteristic depth, consistent across both benchmarks and typically in the early layers (Fig. 2).

To examine whether this saturation in predictive performance is accompanied by saturation in representational structure, we measure layer-wise Centered Kernel Alignment (CKA) (Kornblith et al., 2019) with radial basis function (RBF) kernel for four models spanning different architectures and training objectives, using samples from NT and Genomic Benchmark. As shown in Fig. 3, despite model-specific variation, representations after the saturation point remain highly similar to the final-layer embeddings, with CKA values > 0.7. Together with the linear-probing results, this suggests that early layers account for most downstream gains, while later layers add little to either predictive performance or representation structure. This saturation indicates that substantial model depth may be functionally under-utilized, providing a plausible explanation for weak scaling and implying that what these models use can be captured by a considerably smaller one.

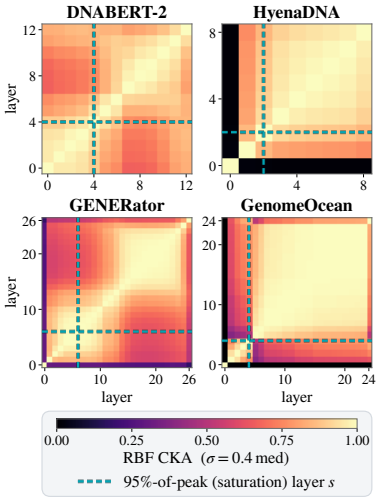


Figure 3: **CKA analysis on four DNA LMs.** Layer-vs.-layer CKA, averaged over two benchmarks.

3.2 Multi-Teacher Distillation Across Heterogeneous Tokenizers

The lack of a DNA LM that dominates all tasks in Sec. 2.1, together with the redundancy observed in Sec. 3.1, motivates distilling existing DNA LMs into a compact student. Inspired by agglomerative multi-teacher distillation in RADIO (Ranzinger et al., 2024b; Heinrich et al., 2025; Ranzinger et al., 2026), we integrate complementary representations from multiple DNA LMs. However, different tokenizers yield non-matching token counts, boundaries, and spans for the same sequence, making token-wise correspondence ill-defined. So we align student and teacher representations in coordinates according to their covered sequence intervals. See Appendix Alg. A1 for detailed algorithm.

Aligning different tokenizations. Different DNA tokenizers partition the same sequence at different boundaries, making token indices incomparable across models. We therefore align tokens by the nucleotide intervals they cover. For a single teacher, omitting teacher index k, let $S_u = [s_u, e_u)$ and $T_v = [s_v, e_v)$ denote the nucleotide intervals covered by student token u and teacher token v.

We then construct each student-token target by aggregating representations from overlapping teacher tokens. Let z_v be teacher token v's representation. Each representation is weighted by its nucleotide overlap with the student token and by a confidence $c_v \in [0, 1]$, which downweights representations affected by padding or crop boundaries. The final overlap weight a_{uv} and aligned target $\widetilde{z}_u$ are

$$a_{uv} = c_v \cdot |S_u \cap T_v|, \qquad \widetilde{z}_u = \frac{\sum_v a_{uv} z_v}{\sum_v a_{uv}}. \tag{1}$$

Applying this procedure to each teacher yields an aligned target $\widetilde{z}_u^{(k)}$ for every student token u.

Matching heterogeneous teacher representations. Aligned token representations vary across teachers in dimensionality and activation scale. We apply PHI-S (Ranzinger et al., 2024a) to normalize each teacher's dense features, balancing activation scales while preserving its native dimension d_k. The resulting targets remain in teacher-specific spaces and are not directly comparable with the student hidden state $h_u \in \mathbb{R}^{d_S}$. We therefore use a teacher-specific adaptor $A_k : \mathbb{R}^{d_S} \to \mathbb{R}^{d_k}$ to map each student representation into the corresponding teacher space.

We then compare $s_u^{(k)} := A_k(h_u)$ with the aligned target $\widetilde{z}_u^{(k)}$. Since padding and crop boundaries may support only part of student interval S_u, we exclude targets whose confidence-adjusted overlap covers less than half of S_u. For each remaining target, we set $w_u^{(k)} = c_u^{\mathrm{S}} \sum_v a_{uv}^{(k)}$, where $c_u^{\mathrm{S}} \in [0, 1]$ accounts for student-side crop-boundary effects as done in constructing c_v. Weighting by supported sequence span rather than token count prevents tokenizers producing more, shorter tokens from

exerting disproportionate influence. The pointwise multi-teacher distillation objective is then

$$\mathcal{L}_{\mathrm{MTD}} = \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_u w_u^{(k)} \left\| s_u^{(k)} - \widetilde{z}_u^{(k)} \right\|_2^2}{\left(\sum_u w_u^{(k)} + \epsilon\right) d_k}. \tag{2}$$

For each teacher, the numerator aggregates the squared matching errors over valid correspondences, while division by $\sum_u w_u^{(k)}$ produces their weighted average. Division by d_k further converts this error into a per-feature quantity, preventing teachers with wider representations from dominating the objective. The resulting teacher-specific losses are then averaged equally across the K teachers.

Auxiliary distillation objectives. Although $\mathcal{L}_{\mathrm{MTD}}$ aligns student and teacher features at individual token positions, it does not explicitly preserve relationships among positions or sequence-level information. We therefore complement it with relational and summary objectives.

The relational loss preserves the geometry of the teacher representation space. We match their cosine similarities using token pairs sampled across genomic-distance ranges:

$$\mathcal{L}_{\mathrm{rel}} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{(u,u')\sim\mathcal{P}_k} \left[\cos\left(s_u^{(k)}, s_{u'}^{(k)}\right) - \cos\left(\widetilde{z}_u^{(k)}, \widetilde{z}_{u'}^{(k)}\right) \right]^2, \tag{3}$$

where $\mathcal{P}_k$ contains valid token pairs sampled across training samples in the batch.

The summary loss transfers sequence-level information by aligning span-weighted pooled representations over support shared by the student and each teacher. As in C-RADIO (Ranzinger et al., 2026), teacher representations may differ substantially in angular dispersion, making unnormalized angular losses incomparable. We observe the same variation among DNA teachers and normalize each summary loss by its measured dispersion; statistics appear in Appendix Tab. A5. Let μ_k denote the mean direction of pooled teacher representations $\mathcal{Y}_k$. We define

$$\mathrm{Disp}_k = \mathbb{E}_{y\sim\mathcal{Y}_k}\left[\Theta(y,\mu_k)^2\right], \qquad \mathcal{L}_{\mathrm{sum}} = \frac{1}{K}\sum_{k=1}^{K} \frac{\Theta\left(B_k(\bar{h}^{(k)}), \bar{z}^{(k)}\right)^2}{\mathrm{Disp}_k + \epsilon}, \tag{4}$$

where $\bar{h}^{(k)}$ and $\bar{z}^{(k)}$ are the span-weighted pooled student and teacher representations, B_k is a teacher-specific summary adaptor, and Θ denotes angular distance. The total distillation objective is

$$\mathcal{L}_{\mathrm{distill}} = \mathcal{L}_{\mathrm{MTD}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}} + \lambda_{\mathrm{sum}}\mathcal{L}_{\mathrm{sum}}. \tag{5}$$

After distillation, we discard the teacher-specific adaptors and retain only the encoder.

3.3 Information Constrained Relational Post-Training

Multi-teacher distillation transfers complementary knowledge from DNA LMs but does not explicitly align sequence similarities with functional annotations. We introduce information constrained relational post-training on large-scale labeled data to improve this alignment. Since annotations capture only part of biological variation, they revise rather than define the model's existing neighborhood structure by shifting the relational distribution toward annotation by a controlled amount.

Pairwise relation scores. For a batch $\mathcal{B} = \{(x_i, y_i, D_i)\}_{i=1}^{B}$, where x_i, y_i, and D_i denote sequence, label, and source dataset, respectively, let $\mathcal{C}_i = \{1,\ldots,B\} \setminus \{i\}$ denote anchor i's candidates. We assign pairwise relationship scores using labels and source datasets:

$$r_{ij} = \begin{cases} +1, & D_i = D_j \text{ and } y_i = y_j, \\ -1, & D_i = D_j \text{ and } y_i \neq y_j, \\ 0, & D_i \neq D_j. \end{cases} \tag{6}$$

Cross-dataset pairs receive a score of zero because labels from different datasets do not necessarily refer to the same biological property.

Relational distributions. Pairwise scores indicate whether candidates should be closer or farther apart, but cannot distinguish candidates with equal scores. We therefore represent encoder similarities as a distribution over candidates. Let f_ϕ be a pooled sequence encoder with parameters ϕ. For

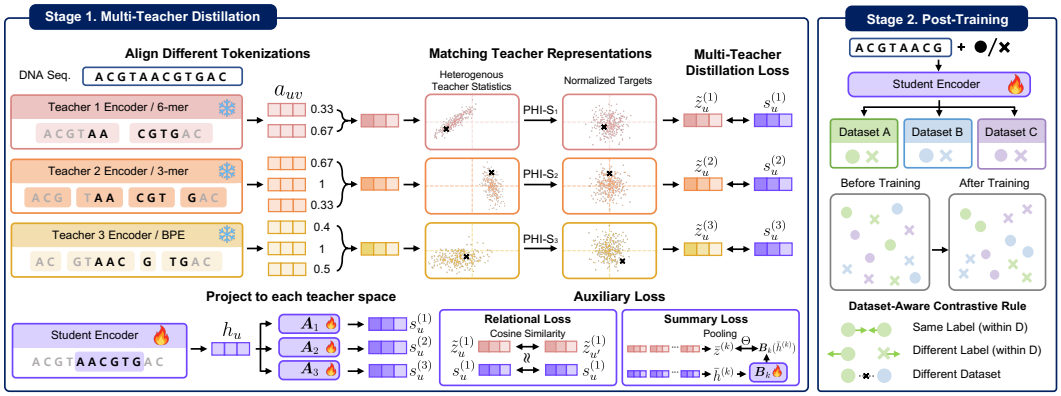


Figure 4: **Overview of DNA-RADIO.** To integrate complementary knowledge from heterogeneous DNA language models and align the resulting representations with annotations, DNA-RADIO employs a two-stage training framework. **(Stage 1)** We align heterogeneous teacher tokenizations and representation spaces to distill their knowledge into a shared student using pointwise, relational, and summary objectives. **(Stage 2)** We apply post-training, attracting same-label samples and repelling different-label samples within each dataset while leaving cross-dataset pairs unconstrained.

anchor i, its relational distribution over $\mathcal{C}_i$ is

$$p_{ij}^{\phi} = \frac{\exp\left(\cos(f_\phi(x_i), f_\phi(x_j))/\tau\right)}{\sum_{k\in\mathcal{C}_i} \exp\left(\cos(f_\phi(x_i), f_\phi(x_k))/\tau\right)}, \tag{7}$$

where $\tau > 0$ is the temperature. It preserves relative preferences among equally scored candidates. The temperature-scaled softmax converts pairwise similarities into a probability distribution that describes each candidate's relative affinity, a formulation widely used in contrastive representation learning (Oord et al., 2018; Chen et al., 2020). This probabilistic representation retains fine-grained neighborhood preferences and allows their revision to be controlled through KL divergence.

Information-constrained target construction. We train a student encoder with parameters θ using a baseline encoder with θ_{base}. The baseline relational distribution $p_i^{\theta_{\text{base}}}$ captures the neighborhood structure around each anchor sequence x_i to be revised. Both encoders are initialized from the multi-teacher-distilled parameters θ_0, such that $\theta_{\text{base}} = \theta_0$ and $\theta = \theta_0$ initially. Keeping the baseline separate fixes this structure within each iteration; its update across iterations is described below.

For each anchor x_i, we construct a target distribution q_i over its candidate sequences $\mathcal{C}_i$. The target should assign more probability to high-scoring candidates while preserving the baseline's finer-grained neighborhood structure. Maximizing expected score alone would place all probability on the highest-scoring candidates and leave their relative allocation unspecified. We therefore constrain the target's KL divergence from the baseline by an information budget. The target is the solution of

$$\max_{\{q_i\}} \frac{1}{B}\sum_{i=1}^{B}\sum_{j\in\mathcal{C}_i} q_{ij} r_{ij}, \quad \text{subject to } \frac{1}{B}\sum_{i=1}^{B} D_{\text{KL}}(q_i \| p_i^{\theta_{\text{base}}}) \le \delta, \tag{8}$$

where each q_i is a probability distribution over $\mathcal{C}_i$ and $\delta \ge 0$ is the information budget. The objective moves probability toward score agreement, while the constraint limits how much the scores can revise the existing neighborhoods.

When the constraint is active and admits a finite tilting strength, the solution is an exponential tilting of the baseline (see Appendix Sec. B.1 for detailed derivation),

$$q_{ij}(\alpha) = \frac{p_{ij}^{\theta_{\text{base}}} \exp(\alpha r_{ij})}{\sum_{k\in\mathcal{C}_i} p_{ik}^{\theta_{\text{base}}} \exp(\alpha r_{ik})}, \qquad \frac{1}{B}\sum_{i=1}^{B} D_{\text{KL}}(q_i(\alpha) \| p_i^{\theta_{\text{base}}}) = \delta, \tag{9}$$

where the tilting strength $\alpha \ge 0$ is shared across the batch and determined by the information budget δ (see Appendix Sec. B.2 for detail). For finite α, candidates with the same score retain their relative probabilities under the baseline.

Tilted target matching. The student produces its own relational distribution p_i^θ on the same candidates and at the same temperature, and we train it to match the target by minimizing

$$\mathcal{L}_{\text{tilt}} = \frac{1}{B}\sum_{i=1}^{B} D_{\text{KL}}\big(\text{sg}[q_i] \| p_i^\theta\big)\,, \tag{10}$$

where sg stops gradients through target construction, including the baseline distribution and the solved tilting strength. Through this objective, the student learns both the score-driven reallocation across groups and the baseline's distinctions within them.

EMA baseline updates. Fixing the baseline at θ_0 would make every target a tilt of the same initial distribution, preventing it from reflecting the student's progress. We instead update the baseline after each student optimization step using an exponential moving average (EMA),

$$\theta_{\text{base}} \leftarrow m \cdot \theta_{\text{base}} + (1-m)\cdot\theta, \tag{11}$$

where $m \in [0, 1)$ is the decay coefficient and θ denotes the updated student parameters. At the next iteration, this baseline defines the distribution tilted to construct the new target. Thus, the neighborhoods evolve with the student, while the moving average smooths update fluctuations without assuming that each update improves the representation.

4 Experiments

4.1 Experimental Setup

Datasets. We separate distillation from post-training. Distillation uses 327 eukaryotic reference assemblies (347.96 Gbp), with chromosome-level held-outs and benchmark-test-locus masking to limit test exposure. Post-training aggregates 69 datasets (4.15M labeled records) from NT (Dalla-Torre et al., 2025), GUE (Zhou et al., 2024), Genomic Benchmark (Grešová et al., 2023), and ENCODE cCRE. Deduplication, split reconciliation, and leakage filtering retain 3.29M records. We exclude all BEND datasets and BEND-overlapping records from the post-training corpus, using BEND as the primary end-to-end evaluation suite. For coordinate-bearing records, we remove overlaps with protected BEND intervals and discard 0.47 M records in training corpus. We also canonicalize reverse complements and enforce corpus-wide split disjointness to prevent exact sequence reuse. For the 43.6% of records without coordinates, we enforce exact sequence disjointness and audit possible containment in BEND windows (see Appendix Sec. C.2 for details).

Benchmarks. We perform evaluation on Revised-NT (Dalla-Torre et al., 2025), GUE (Zhou et al., 2024), Genomic Benchmark (Grešová et al., 2023), and BEND (Marin et al., 2024), especially on eukaryotic task. Following standard foundation-model practice, we use linear probing for Revised-NT, GUE, and Genomic Benchmark, and the released BEND protocol (CNN predictor; zero-shot for variant effect) for BEND, reporting five-seed results throughout. Revised-NT, GUE, and Genomic Benchmark span diverse genomic elements and species, but largely evaluate fixed-window classification, limiting assessment of nucleotide-resolved prediction, high-dimensional regulatory outputs and variant effects. BEND provides this broader coverage through seven heterogeneous tasks with varying sequence length (512 bp-0.1 Mbp) and genomic context. We therefore exclude BEND from post-training and use it as the primary held-out benchmark for generalization to unseen tasks. All the readouts are done from final layer of each model. See Appendix Sec. C.4 for details.

Implementation details. We use GENERator-1.2B (Wu et al., 2025), GenomeOcean-4B (Zhou et al., 2025), and NT-2.5B-multi (Dalla-Torre et al., 2025) as teachers and initialize the student weights from ModernGENA with modification on attention. Motivated by the observation in Sec. 3.1 and for efficiency, we use the earliest intermediate layer that reaches 95% of the peak linear-probe performance on held-out calibration dataset for each teacher (see Appendix Sec. D.1.1). Additional details such as architecture, optimization, computation and hyperparameters are provided in Appendix Sec. D.1.5.

4.2 Quantitative Results

Distilled DNA-RADIO comparison. In Fig. 1, we compare distilled DNA-RADIO with various models across multiple benchmarks, with detailed results presented in Tab. 1. On Revised-NT, although models such as GenomeOcean and GENERator achieve the best performance on some tasks,

Table 1: **Comparison for distilled DNA-RADIO.** We evaluate distilled DNA-RADIO against the teacher, the original student model, and several baselines on three benchmarks: **(Top)** the revised-NT benchmark across 18 tasks, **(Mid)** the GUE benchmark across 27 tasks, and **(Bottom)** the Genomic Benchmark across 9 tasks. For GUE, the reported Average is the unweighted mean of the six task-family means. We use MCC (Matthews Correlation Coefficient) as the evaluation metric for revised-NT and GUE, and accuracy for Genomic Benchmark. DNA-RADIO$_{\mathrm{d}}$ denotes the model trained with distillation only. All results are averaged over five random seeds.

Task	DNABERT-2	HyenaDNA-L	Caduceus-PS	NT-v2	ModernGENA-L	NT-multi	GenomeOcean	GENERator	**DNA-RADIO-B$_{\mathrm{d}}$**	**DNA-RADIO-L$_{\mathrm{d}}$**
Model size	117M	6.6M	7.73M	500M	377M	2.5B	4B	1.2B	136M	377M
Regulatory										
Enhancer	0.4892	0.4998	0.4790	0.4899	0.4892	0.4920	0.4899	0.4932	0.4973	0.5172
Enhancer type	0.3692	0.3598	0.3419	0.3566	0.3515	0.3879	0.3589	0.3952	0.3744	0.3794
Promoter all	0.6798	0.7167	0.7455	0.7113	0.7341	0.7354	0.7379	0.7744	0.7581	0.7554
Promoter non-TATA	0.7363	0.7380	0.7712	0.7405	0.7570	0.7567	0.7637	0.7631	0.7608	0.7826
Promoter TATA	0.7004	0.7550	0.7466	0.6999	0.6869	0.7670	0.7692	0.7925	0.8101	0.8455
Splice										
Splice acceptor	0.6219	0.4998	0.6255	0.4720	0.4731	0.6623	0.5875	0.6645	0.7072	0.7025
Splice site all	0.3464	0.3625	0.4254	0.3687	0.2906	0.4785	0.4205	0.5103	0.4969	0.5001
Splice donor	0.6469	0.4994	0.6581	0.5241	0.5363	0.6917	0.5962	0.6865	0.7218	0.7253
Histone										
H2AFZ	0.4037	0.4908	0.4174	0.3854	0.4054	0.4112	0.4734	0.4394	0.4419	0.4606
H3K27ac	0.4434	0.4520	0.4319	0.3916	0.4400	0.4615	0.4744	0.4725	0.4839	0.4733
H3K27me3	0.5559	0.5481	0.5587	0.5392	0.5447	0.5680	0.5888	0.5848	0.5790	0.5819
H3K36me3	0.5539	0.5442	0.5374	0.5065	0.5170	0.5766	0.5797	0.5791	0.5977	0.6087
H3K4me1	0.4349	0.4391	0.4131	0.3999	0.4168	0.4138	0.4607	0.4483	0.4575	0.4746
H3K4me2	0.4737	0.5237	0.5047	0.4726	0.4763	0.4792	0.5532	0.5374	0.5240	0.5408
H3K4me3	0.6138	0.6165	0.5887	0.6239	0.6187	0.6166	0.6254	0.6192	0.6233	0.6197
H3K9ac	0.5020	0.4858	0.5141	0.5188	0.5343	0.4924	0.5389	0.5312	0.5277	0.5405
H3K9me3	0.3019	0.3795	0.3900	0.3005	0.2980	0.3827	0.4212	0.4205	0.3964	0.4105
H4K20me1	0.6003	0.6090	0.5684	0.5466	0.5842	0.6113	0.6150	0.6273	0.6231	0.6360
Average	0.5263	0.5289	0.5399	0.5027	0.5086	0.5547	0.5586	0.5744	0.5767	0.5864

Task	DNABERT-2	HyenaDNA-L	Caduceus-PS	NT-v2	ModernGENA-L	NT-multi	GenomeOcean	GENERator	**DNA-RADIO-B$_{\mathrm{d}}$**	**DNA-RADIO-L$_{\mathrm{d}}$**
Model size	117M	6.6M	7.73M	500M	377M	2.5B	4B	1.2B	136M	377M
Yeast										
EMP	0.4557	0.4306	0.4750	0.4376	0.4448	0.5219	0.5512	0.5379	0.5358	0.5444
Mouse										
TF-M	0.5192	0.5005	0.5261	0.4939	0.6480	0.5478	0.4917	0.6149	0.6836	0.6880
Human										
TF-H	0.5720	0.5587	0.5957	0.5513	0.6359	0.5882	0.5958	0.6191	0.6570	0.6505
PD	0.7080	0.7456	0.7692	0.7323	0.7568	0.7829	0.7477	0.8204	0.8181	0.8217
CPD	0.5632	0.5693	0.5974	0.6214	0.6016	0.6099	0.5576	0.6312	0.6260	0.6390
SSP	0.2875	0.3640	0.5289	0.3352	0.3390	0.5888	0.4888	0.6113	0.6031	0.6330
Average	0.5176	0.5281	0.5821	0.5286	0.5710	0.6066	0.5721	0.6391	0.6539	0.6628

Task	DNABERT-2	HyenaDNA-L	Caduceus-PS	NT-v2	ModernGENA-L	NT-multi	GenomeOcean	GENERator	**DNA-RADIO-B$_{\mathrm{d}}$**	**DNA-RADIO-L$_{\mathrm{d}}$**
Model size	117M	6.6M	7.73M	500M	377M	2.5B	4B	1.2B	136M	377M
Mouse										
EE	0.8322	0.7504	0.8215	0.8554	0.8174	0.8479	0.8215	0.8322	0.8570	0.8702
Human										
CI	0.9113	0.8730	0.9355	0.8921	0.9251	0.9321	0.9423	0.9482	0.9404	0.9457
EC	0.7546	0.7263	0.7443	0.7245	0.7480	0.7554	0.7383	0.7554	0.7571	0.7593
EE	0.7567	0.7252	0.7492	0.7709	0.7674	0.7788	0.7713	0.7591	0.7779	0.7814
ER	0.6812	0.7358	0.6490	0.6860	0.6577	0.7299	0.6704	0.6541	0.7289	0.7141
non-TATA	0.8528	0.8383	0.8621	0.8414	0.8758	0.8631	0.8754	0.8787	0.8788	0.8834
OCR	0.6754	0.6742	0.6832	0.6816	0.6858	0.6971	0.6990	0.6943	0.7039	0.7127
Cross-species										
Human–Worm	0.9564	0.9309	0.9664	0.9473	0.9668	0.9628	0.9626	0.9672	0.9752	0.9772
Drosophila										
ES	0.7147	0.6871	0.7232	0.7202	0.7045	0.7347	0.7310	0.7376	0.7361	0.7306
Average	0.7928	0.7712	0.7927	0.7910	0.7943	0.8113	0.8013	0.8030	0.8173	0.8194

the base variant of DNA-RADIO already surpasses the previous state of the art overall, with the large variant yielding further performance gains. On GUE and Genomic Benchmark, the large and base variants of DNA-RADIO rank first and second, respectively, on most tasks. Notably, DNA-RADIO exceeds the best performance achieved by its teachers, suggesting that multi-teacher distillation enables generalization beyond the capabilities of individual teachers. All experiments use five random seeds. Standard deviation of each entry, comparisons with additional models, and evaluations on additional benchmarks are provided in Appendix Sec. E.2 and Sec. E.3.

Post-trained DNA-RADIO comparison. We apply post-training to the distilled DNA-RADIO-B$_{\mathrm{d}}$ model and compare it with various models on the BEND benchmark, with results presented in Tab. 2. Although distilled DNA-RADIO already performs well, post-training further improves its overall performance. Notably, we observe substantial gains on Var. effect expression and Var. effect disease, both of which are evaluated in a zero-shot setting. These improvements suggest that post-training promotes generalization beyond learning task-specific label information. We also show that our post-training also improves performance of other DNA LMs, which is shown in Appendix Sec. E.4.

Table 2: **Comparison for post-trained DNA-RADIO.** We perform information budget constrained post-training on top of distilled DNA-RADIO-B_d and evaluate the resulting model on the BEND benchmark across 7 tasks. Since the relevant metrics vary across tasks, we report the average rank.

Task	DNABERT-2	HyenaDNA-L	Caduceus-PS	NT-v2	ModernGENA-L	NT-multi	GenomeOcean	GENERator	**DNA-RADIO-B_d**	**DNA-RADIO**
Model size	117M	6.6M	7.73M	500M	377M	2.5B	4B	1.2B	136M	377M
Gene finding (MCC)	0.4376	0.3576	0.5685	0.6357	0.3182	0.6680	0.5230	0.5530	0.5798	0.6137
Enhancer annot. (AUPRC)	0.0250	0.0342	0.0302	0.0392	0.0273	0.0387	0.0296	0.0220	0.0372	0.0433
Chromatin access. (AUROC)	0.7625	0.7931	0.7663	0.7655	0.7675	0.7754	0.7313	0.6906	0.7568	0.7647
Histone mod. (AUROC)	0.7641	0.7509	0.7711	0.7516	0.7625	0.7671	0.7574	0.7319	0.7798	0.7826
CpG methyl. (AUROC)	0.8814	0.8675	0.8927	0.8780	0.8721	0.8978	0.8641	0.8533	0.9101	0.9303
Var. effect expr. (AUROC)	0.4838	0.4961	0.5402	0.4819	0.5296	0.5418	0.5108	0.3978	0.4994	0.5671
Var. effect disease (AUROC)	0.5092	0.5236	0.5450	0.4757	0.4944	0.7723	0.5227	0.6737	0.7368	0.7523
Average rank	7.1429	6.4286	4.2857	6.0000	6.7137	2.2857	7.2857	8.5714	4.1429	2.1429

4.3 Ablation Study

We ablated DNA-RADIO's distillation and post-training stages, evaluating each independently on the held-out BEND benchmark to quantify its contribution to performance.

On distillation components. (ID 1–3) Removing any of the three distillation objectives reduces average downstream performance relative to the full distillation baseline. Although the effects vary across tasks, each objective contributes to the performance of the distilled model.

Table 3: **Ablation study of DNA-RADIO components.**

ID	Study subject	CpG	Gene	Var. disease	Avg.
	DNA-RADIO-B_d	0.9101	0.5798	0.7368	0.7422
1	w/o $\mathcal{L}_{\text{MTD}}$	0.9119	0.5270	0.7369	0.7253
2	w/o $\mathcal{L}_{\text{rel}}$	0.9097	0.5708	0.7068	0.7291
3	w/o $\mathcal{L}_{\text{sum}}$	0.9077	0.5672	0.7076	0.7275
	DNA-RADIO	0.9303	0.6137	0.7523	0.7654
4	$\delta = 0$	0.9109	0.5710	0.5778	0.6866
5	w/o EMA	0.9194	0.6035	0.7308	0.7512

On post-training components. (ID 4–5) Setting $\delta = 0$ substantially reduces average downstream performance, highlighting the benefit of allowing neighborhood adaptation during post-training. Removing the EMA also reduces performance, suggesting that an evolving reference helps avoid constraining the model to a fixed baseline during post-training. These results support both components in the post-training.

5 Discussion

Through its distillation and post-training stages, DNA-RADIO achieves strong performance across a wide range of benchmarks and tasks. Our primary focus in this work is eukaryotic DNA sequences. Accordingly, the genomic corpus used for distillation consists only of eukaryotic genomes, and our evaluation is restricted to benchmarks targeting eukaryotic sequences. Extending this framework to prokaryotic and viral genomes is a promising direction for future work. In addition, DNA-RADIO outperforms all three of its teachers after the distillation stage alone; a similar effect has been reported for C-RADIO on some tasks. We hypothesize that this gain arises from synergy among the complementary capabilities of the individual DNA LMs, but we have not conducted experiments to verify this hypothesis. Identifying this mechanism and exploiting it to build stronger DNA models is another valuable avenue for future study.

6 Conclusion

We introduce DNA-RADIO, an agglomerative DNA embedding model trained in two stages. Our approach is motivated by two observations: existing DNA language models (LMs) excel on different benchmarks, and their deeper layers contribute limited changes to their representations, suggesting that their depth is underutilized. In the first stage, we distill three billion-scale teacher DNA LMs with different model sizes, tokenizers, and training objectives into compact 136M- and 377M-parameter students through multi-teacher distillation. In the second stage, we introduce information-budget constrained post-training to align DNA-RADIO's internal representations using over 3 million labeled sequence records. Across 54 tasks, even the distilled base variant outperforms SOTA models with over one billion parameters overall, while the large variant delivers further gains. Post-training also yields strong performance on the challenging, held-out BEND benchmark. Together, these results demonstrate the potential of combining multi-teacher distillation with information-budget constrained post-training to build compact, general-purpose DNA embedding models.

AI USE STATEMENT

In this work, we used generative AI tools for code generation, derivation verification, experiment managing and manuscript polishing. We have not used generative AI tools for idea generation. We have reviewed all AI-assisted work. All the generated code with generative AI tools was supervised and reviewed by authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, all reported performance results are averaged over five evaluation runs with different random seeds, and the corresponding standard deviations are provided in the Appendix. Our analyses are likewise conducted across multiple seeds and training configurations to assess the statistical reliability of our findings. The supplementary materials include the training and evaluation code for both our method and the baseline models, which enables reproduction of the results reported in Tab. 1 and Tab. 2.

REFERENCES

Alena Aspidova, Yuri Kuratov, Artem Shadskiy, Mikhail Burtsev, and Veniamin Fishman. Back to bert in 2026: Moderngena as a strong, efficient baseline for dna foundation models. *bioRxiv*, 2026. doi: 10.64898/2026.04.21.719816.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pp. 15619–15629. IEEE, 2023.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18 (10):1196–1203, 2021.

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.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A Gonzalez, et al. Genome modelling and design across all domains of life with evo 2. *Nature*, 652(8112):1349–1361, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9630–9640. IEEE, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607. PMLR, 2020.

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, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature methods*, 22(2):287–297, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pp. 4171–4186, 2019.

Veniamin Fishman, Yuri Kuratov, Maxim Petrov, Aleksei Shmelev, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. Gena-lm: A family of open-source foundational models for long dna sequences. biorxiv. *bioRxiv*, 2023.

Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25, 2023.

Yong He, Pan Fang, Yongtao Shan, Yuanfei Pan, Yanhong Wei, Yichang Chen, Yihao Chen, Yi Liu, Zhenyu Zeng, Zhan Zhou, et al. Generalized biological foundation model with unified nucleic acid and protein language. *Nature Machine Intelligence*, 7(6):942–953, 2025.

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *CVPR*, pp. 22487–22497, 2025.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, pp. 3519–3529. PMLR, 2019.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Daria Ledneva, Mikhail Nuridinov, and Denis Kuznetsov. Geneb: Why genomic models are hard to compare. *ICML*, 2026.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. In *ICLR*, volume 2025, pp. 79310–79333, 2025.

Rosario N Mantegna, Sergey V Buldyrev, Ary L Goldberger, Shlomo Havlin, Chung-Kang Peng, Michael Simons, and H Eugene Stanley. Linguistic features of noncoding dna sequences. *Physical review letters*, 73(23):3169, 1994.

Frederikke Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. Bend: Benchmarking dna language models on biologically meaningful tasks. In *ICLR*, volume 2024, pp. 15246–15281, 2024.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *NeurIPS*, 36:43177–43201, 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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024. ISSN 2835-8856.

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. *NeurIPS*, 37:62024–62061, 2024.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

Mike Ranzinger, Jon Barker, Greg Heinrich, Pavlo Molchanov, Bryan Catanzaro, and Andrew Tao. Phi-s: Distribution balancing for label-free multi-teacher distillation. *arXiv preprint arXiv:2410.01680*, 2024a.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, pp. 12490–12500. IEEE, 2024b.

Mike Ranzinger, Greg Heinrich, Collin McCarthy, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. C-radiov4 (tech report). *arXiv preprint arXiv:2601.17237*, 2026.

Melissa Sanabria, Jonas Hirsch, Pierre M Joubert, and Anna R Poetsch. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, 6(8):911–923, 2024.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. In *ICML*, volume 235 of *PMLR*, pp. 43632–43648, 2024.

David B Searls. Linguistic approaches to biological sequences. *Bioinformatics*, 13(4):333–344, 1997.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Wei Wu, Qiuyi Li, Yuanyuan Zhang, Zhihao Zhan, Ruipu Chen, Mingyang Li, Kun Fu, Junyan Qi, Yongzhou Bao, Chao Wang, et al. Generator: a long-context generative genomic foundation model. *arXiv preprint arXiv:2502.07272*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11941–11952. IEEE, 2023.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genomes. In *ICLR*, volume 2024, pp. 41642–41665, 2024.

Zhihan Zhou, Robert Riley, Satria Kautsar, Weimin Wu, Rob Egan, Steven Hofmeyr, Shira Goldhaber-Gordon, Mutian Yu, Harrison Ho, Fengchen Liu, et al. Genomeocean: An efficient genome foundation model trained on large-scale metagenomic assemblies. *bioRxiv*, 2025.

A CODE

We provide the code implementation of DNA-RADIO for reproducibility. Please refer to `README.md` in supplementary material for instructions on running the code for training and evaluation. We will release the model weights to the public upon acceptance.

B DERIVATION

B.1 DERIVATION OF EQ. 9

For notational convenience, let $p_{ij} = p_{ij}^{\theta_{\text{base}}}$. We assume that each candidate set $\mathcal{C}_i$ is nonempty and that $p_{ij} > 0$, as ensured by the softmax distribution with finite logits. Define

$$R(q) = \frac{1}{B}\sum_{i=1}^{B}\sum_{j\in\mathcal{C}_i} q_{ij} r_{ij}, \tag{A1}$$

$$K(q) = \frac{1}{B}\sum_{i=1}^{B} D_{\mathrm{KL}}(q_i \| p_i). \tag{A2}$$

The target construction problem is

$$\begin{aligned} \max_{\{q_i\}_{i=1}^{B}} \quad & R(q) \\ \text{subject to} \quad & K(q) \le \delta, \\ & \sum_{j\in\mathcal{C}_i} q_{ij} = 1, \qquad q_{ij} \ge 0 \quad \text{for all } i, j. \end{aligned} \tag{A3}$$

We use the convention $0\log(0/p_{ij}) = 0$. The feasible set is compact and nonempty, since $q = p$ is feasible for every $\delta \ge 0$, so an optimum exists. For $\delta > 0$, the baseline is strictly feasible. Since the objective is linear and the KL constraint is convex, Slater's condition ensures strong duality.

Lagrangian and exponential tilting. Introduce a multiplier $\eta \ge 0$ for the KL constraint and a multiplier λ_i for each normalization constraint. The Lagrangian is

$$\begin{aligned} \mathcal{L}(q, \eta, \lambda) = & \frac{1}{B}\sum_{i=1}^{B}\sum_{j\in\mathcal{C}_i} q_{ij} r_{ij} \\ & - \frac{\eta}{B}\sum_{i=1}^{B}\sum_{j\in\mathcal{C}_i} q_{ij}\log\frac{q_{ij}}{p_{ij}} + \eta\delta \\ & + \frac{1}{B}\sum_{i=1}^{B}\lambda_i\left(\sum_{j\in\mathcal{C}_i} q_{ij} - 1\right). \end{aligned} \tag{A4}$$

For $\eta > 0$, the KL-regularized objective is strictly concave in q and its maximizer has positive coordinates. The stationarity condition is therefore

$$\frac{\partial\mathcal{L}}{\partial q_{ij}} = \frac{1}{B}\left[r_{ij} - \eta\left(\log\frac{q_{ij}}{p_{ij}} + 1\right) + \lambda_i\right] = 0. \tag{A5}$$

Rearranging gives

$$q_{ij} = p_{ij}\exp\left(\frac{r_{ij}}{\eta}\right)\exp\left(\frac{\lambda_i}{\eta} - 1\right). \tag{A6}$$

Enforcing normalization yields

$$\exp\left(\frac{\lambda_i}{\eta} - 1\right) = \left[\sum_{k\in\mathcal{C}_i} p_{ik}\exp\left(\frac{r_{ik}}{\eta}\right)\right]^{-1}. \tag{A7}$$

Thus, defining $\alpha = 1/\eta$ and

$$Z_i(\alpha) = \sum_{k \in \mathcal{C}_i} p_{ik} \exp(\alpha r_{ik}), \tag{A8}$$

we obtain

$$q_{ij}(\alpha) = \frac{p_{ij} \exp(\alpha r_{ij})}{Z_i(\alpha)}. \tag{A9}$$

The tilting strength α is shared across anchors because the optimization problem has a single batch-averaged KL constraint and hence a single multiplier η. Only the normalization constant $Z_i(\alpha)$ depends on the anchor.

For the analysis below, define the log-partition function

$$\psi_i(\alpha) = \log Z_i(\alpha) \tag{A10}$$

and the batch-averaged KL divergence along the tilted family,

$$K(\alpha) := K(q(\alpha)). \tag{A11}$$

Global optimality. Let $\alpha > 0$ satisfy $K(\alpha) = \delta$. For any collection of candidate distributions q, Eq. A9 implies

$$\frac{1}{B} \sum_{i=1}^{B} D_{\mathrm{KL}}(q_i \| q_i(\alpha)) = K(q) - \alpha R(q) + \frac{1}{B} \sum_{i=1}^{B} \psi_i(\alpha). \tag{A12}$$

Applying the same identity to $q(\alpha)$ and subtracting yields

$$\begin{aligned} \alpha \left[R(q) - R(q(\alpha)) \right] = K(q) - \delta \\ - \frac{1}{B} \sum_{i=1}^{B} D_{\mathrm{KL}}(q_i \| q_i(\alpha)). \end{aligned} \tag{A13}$$

For every feasible q, the right-hand side is nonpositive. Thus $R(q) \leq R(q(\alpha))$, proving global optimality. Equality requires $q_i = q_i(\alpha)$ for every anchor, which also establishes uniqueness.

B.2 Determining the tilting strength α

When $\eta > 0$, complementary slackness gives

$$\eta\big(K(q) - \delta\big) = 0 \quad \Longrightarrow \quad K(q(\alpha)) = \delta. \tag{A14}$$

To characterize this equation, note that the first and second derivatives of $\psi_i(\alpha)$ satisfy

$$\psi_i'(\alpha) = \sum_{j \in \mathcal{C}_i} q_{ij}(\alpha) r_{ij}, \tag{A15}$$

$$\psi_i''(\alpha) = \sum_{j \in \mathcal{C}_i} q_{ij}(\alpha) r_{ij}^2 - \left(\sum_{j \in \mathcal{C}_i} q_{ij}(\alpha) r_{ij} \right)^2 \tag{A16}$$

$$= \mathrm{Var}_{j \sim q_i(\alpha)}[r_{ij}]. \tag{A17}$$

Taking the logarithm of Eq. A9,

$$\log \frac{q_{ij}(\alpha)}{p_{ij}} = \alpha r_{ij} - \psi_i(\alpha). \tag{A18}$$

Consequently,

$$D_{\mathrm{KL}}(q_i(\alpha) \| p_i) = \sum_{j \in \mathcal{C}_i} q_{ij}(\alpha) \left[\alpha r_{ij} - \psi_i(\alpha) \right] \tag{A19}$$

$$= \alpha \psi_i'(\alpha) - \psi_i(\alpha). \tag{A20}$$

The budget condition therefore reduces to the scalar equation

$$K(\alpha) = K(q(\alpha)) = \frac{1}{B}\sum_{i=1}^{B}[\alpha\psi_i'(\alpha) - \psi_i(\alpha)] = \delta. \tag{A21}$$

Differentiating gives

$$K'(\alpha) = \frac{\alpha}{B}\sum_{i=1}^{B}\psi_i''(\alpha) \tag{A22}$$

$$= \frac{\alpha}{B}\sum_{i=1}^{B}\mathrm{Var}_{j\sim q_i(\alpha)}[r_{ij}] \geq 0. \tag{A23}$$

If at least one anchor has candidates with different scores, then $K'(\alpha) > 0$ for every finite $\alpha > 0$, because $q_i(\alpha)$ has full support. Hence, a finite positive solution of Eq. A21, when it exists, is unique and can be found by a scalar root search.

B.3 Preservation of within-score structure

For any finite α and any $j, k \in \mathcal{C}_i$,

$$\frac{q_{ij}(\alpha)}{q_{ik}(\alpha)} = \frac{p_{ij}}{p_{ik}}\exp\left(\alpha(r_{ij} - r_{ik})\right). \tag{A24}$$

Therefore, candidates with the same score satisfy

$$r_{ij} = r_{ik} \quad \Longrightarrow \quad \frac{q_{ij}(\alpha)}{q_{ik}(\alpha)} = \frac{p_{ij}}{p_{ik}}. \tag{A25}$$

The target changes the probability mass assigned to different score groups while preserving the baseline conditional distribution within each group. For $r_{ij} \in \{-1, 0, +1\}$, the unnormalized probabilities are multiplied by $e^{-\alpha}$, 1, and e^{α}, respectively. Cross-dataset pairs therefore preserve their relative probabilities within the zero-score group, although their normalized probabilities may change through $Z_i(\alpha)$.

C Experimental Details

C.1 Constructing dataset for distillation

Provenance and assembly selection. The distillation stage uses unlabeled reference sequence only. We construct its corpus from two pools. First, we retain every non-plant eukaryotic genome in the NT-850 reconstruction (fungi, invertebrates, protozoa, mammalian vertebrates, and other vertebrates). Second, since NT-850 has no plant clade, we start with 212 RefSeq plant species and retain one representative per genus, preferring the higher assembly level and, when levels are equal, the smaller assembly. This rule avoids over-representing genera with several closely related assemblies. The resulting corpus contains 327 assemblies, 12,865 primary contigs, and 361.45 Gbp before the chromosome-level hold-out. Its composition after the split is reported in Tab. A1.

Assembly and contig filters. We apply the following filters before constructing the final manifest:

- **Core-assembly duplicates.** We remove species already represented among the six core assemblies, including GRCm39, *Danio rerio*, *Drosophila melanogaster*, and *Caenorhabditis elegans*.
- **Assemblies without a viable within-assembly split.** We remove nine assemblies with fewer than two contigs of at least 1 Mb, because they cannot provide both training sequence and a chromosome- or contig-level hold-out.
- **Assemblies dominated by benchmark masking.** We remove an assembly when the benchmark mask covers more than 50% of its usable sequence. This excludes *S. cerevisiae* S288C: 85.2% of its sequence is masked and all 25 sampled crops are rejected.

Table A1: **Composition of the unlabeled distillation corpus.** The table reports the number of retained assemblies and primary contigs, the sequence available for distillation, and the chromosome- or contig-level sequence held out before sampling for each clade. "Held out" denotes sequence on contigs that the distillation sampler never opens.

Clade	Assemblies	Primary contigs	Train sequence	Held out
Plant	151	4,993	140.11 Gbp	8,570 Mb
Non-mammalian vertebrate	54	2,326	104.41 Gbp	2,152 Mb
Mammalian vertebrate	31	3,464	77.21 Gbp	1,217 Mb
Invertebrate	38	1,457	24.75 Gbp	1,487 Mb
Fungi	43	502	1.17 Gbp	62 Mb
Protozoa	10	123	0.30 Gbp	11 Mb
Total	327	12,865	347.96 Gbp	13.50 Gbp

Table A2: **Final post-training corpus by source.** Each row reports the number of retained datasets and training records contributed by a source group after corpus harmonization and filtering. All retained examples consist of a DNA sequence paired with a class label.

Source	Datasets	Training records
ENCODE cCRE V4	1	1,608,133
Genomic Benchmark	9	713,087
GUE and GUE+ species tasks	16	370,574
Original NT	15	348,611
Revised NT	15	245,931
Total	56	3,286,336

For every retained assembly, we keep contigs of at least 1 Mb and exclude alternate, unplaced, organellar, and accession-scaffold sequences (patterns including `_random`, `chrUn`, `_alt`, `_fix`, `chrM`, and Arabidopsis `Mt`/`Pt`). The length floor retains 78% in fungi and protozoa while changing vertebrate sequence negligibly. The corpus uses 321 NCBI RefSeq assemblies, four UCSC assemblies, one GENCODE assembly, and one TAIR assembly.

Sampling. We use hierarchical sampling to balance domains without allowing genome size or the number of available assemblies to determine sampling frequency. Each crop is assigned to the human, mouse, or other domain with probabilities 0.27, 0.13, and 0.60, respectively. For an other-domain crop, we sample one of the six clades uniformly and then sample an assembly uniformly within that clade. Consequently, each clade contributes 10% of all draws, irrespective of the amount of sequence it contains or the number of assemblies it contains.

Labeled sources excluded by design. We do not use the labeled post-training corpus for distillation, because its role is to define and test the supervised stage rather than to supply distillation inputs

C.2 Constructing dataset for post-training

Corpus scope. We construct the post-training corpus from publicly released DNA classification resources. Every example consists of a DNA sequence paired with a class label. The final corpus contains 3,286,336 training records from 56 datasets after filtering, together with 95,049 validation and 90,098 test records. It draws from five source groups: ENCODE cCRE V4, Genomic Benchmark, GUE, and the original and revised Nucleotide Transformer (NT) suites. cCRE contributes an eight-class regulatory-element classification dataset, while the remaining sources cover sequence-level genomic classification tasks across human and non-human species. Tab. A2 reports the final composition by source, and Tab. A3 summarizes how the candidate records are transformed into the final corpus through deduplication, filtering, and BEND interval protection.

Table A3: **Construction ledger for the post-training corpus.** Starting from all candidate records, we report the records removed by cross-source canonical-DNA deduplication, class-collapse filtering, and BEND interval protection, followed by the final training, validation, and test splits. Removal counts are measured at their respective pipeline stages and therefore are not an additive partition.

Construction checkpoint	Records
Candidate records ingested for training	4,147,815
Removed by BEND interval protection	470,336
Removed by cross-source canonical-DNA deduplication	276,473
Removed by held-out interval collision	67,398
Removed after collapse below two usable classes	47,272
Final training split	3,286,336
Final validation split	95,049
Final test split	90,098

C.2.1 Corpus harmonization

Canonical sequence identity and split assignment. We normalize every sequence by uppercasing it and define its canonical DNA as the lexicographically smaller of the sequence and its reverse complement. Each canonical DNA receives exactly one split across the complete corpus; consequently, an identical sequence cannot occur in training for one source and evaluation for another. For coordinate-bearing records, the split is determined by genomic location, with any occurrence on GRCh38 chr20 or chr21 assigned to test. For sequence-only records, the split is determined by a stable hash of canonical DNA. We retain a source-provided official test partition when it agrees with this corpus-wide assignment; otherwise, we create validation data as a label-stratified 10% split of the source training partition. This replacement is necessary because related tasks can impose incompatible independent random splits on a shared sequence universe.

Cross-source duplicates and related task bundles. We remove cross-source duplicates using exact canonical-DNA equality and retain the instance from the highest-priority source: revised NT, cCRE, original NT, Genomic Benchmark, and GUE. The priority favors coordinate-bearing sources, which support interval-based BEND protection, and retains original NT over GUE because we identify GUE reuse of original-NT examples. We do not remove shared sequences within intentionally related tasks from the same source, such as parent and subtype promoter tasks; instead, we treat each such collection as a bundle with the single corpus-wide split. The resulting within-source shared-input counts are 255,071 for original NT, 70,850 for GUE, 69,818 for Genomic Benchmark, and 34,693 for revised NT. This distinction prevents duplicate samples across independent sources while preserving related tasks that assign different valid labels to the same input sequence.

Observed effects of harmonization. The procedure removes 276,473 cross-source duplicates. In particular, the ten GUE yeast `emp_*` tasks are duplicates of the ten original-NT yeast histone tasks. The GUE `prom_300_{all,notata,tata}` datasets lose their positive class after deduplication and are discarded, since an one-class dataset cannot support the post-training objective. After all filters, thirteen candidate datasets have no usable training records and do not contribute to any aggregate over adapted datasets. Full sequences remain in the manifest; truncation is solely a training-configuration effect. At the configured 512-token DNABERT-2 BPE limit, 0.8% of training records truncate (median 66, 90th percentile 199, 99th percentile 435, and maximum 2,157 tokens), predominantly in `gue/fungi_species_20`, `gue/virus_species_40`, and `genomic/dummy_mouse_enhancers_ensembl`. cCRE intervals are also retained at their annotated widths; class-median widths range from 232 bp (CA-CTCF) to 320 bp (PLS), so sequence length is a disclosed class cue for that source's own task.

C.2.2 BEND isolation and verification

BEND source exclusion and exact-sequence filtering. We exclude BEND entirely from the post-training sources: no BEND record, label, or split is used for post-training. In addition, we canonicalize every DNA sequence by uppercasing it and taking the lexicographically smaller of the sequence and its reverse complement. We remove any post-training record whose canonical DNA occurs in a

Table A4: **Measured scope, cost, and verification of BEND isolation.** We report the size of the merged GRCh38 BEND exclusion set, the post-training records removed by interval-overlap filtering, its source-specific cost for revised NT and cCRE, and verification on the final manifest. The final overlap count is evaluated over all retained coordinate-bearing post-training records.

Quantity	Value
Merged GRCh38 BEND exclusion union	334,527 intervals
Sequence spanned by the exclusion union	263.2 Mb (8.49% of GRCh38)
Post-training records removed by the interval-overlap gate	470,336 (11.3%)
Revised-NT records removed by BEND protection	27.3%
cCRE records removed by BEND protection	18.3%
Retained coordinate-bearing training records checked	1,854,064
Retained coordinate-bearing records overlapping a BEND exclusion	0

BEND evaluation split, ensuring that an identical sequence, including a strand-reversed copy, cannot be used for post-training and BEND evaluation.

Coordinate-based exclusion of BEND evaluation. Exact sequence matching alone does not detect distinct fragments drawn from the same genomic locus. We therefore apply an interval-overlap gate to every post-training record with auditable GRCh38 coordinates: a record is removed when its half-open genomic interval overlaps a BEND evaluation interval by at least one base. We construct one merged GRCh38 exclusion set using each task's evaluation protocol: validation and test intervals for histone modification, CpG methylation, and gene finding; all enhancer-annotation intervals because the benchmark uses ten-fold cross-validation; and all variant-effect loci expanded to BEND's 512-bp scoring window because the evaluation harness creates the held-out split. BEND chromatin accessibility is distributed in GRCh37 and is therefore not included in this coordinate-based exclusion set, as no validated lift-over was available.

Measured effect and manifest-level verification. The merged union spans 263.2 Mb, or 8.49% of GRCh38, and removes 470,336 post-training records. This is 11.3% of the coordinate-bearing candidate pool, including 27.3% of revised-NT records (25–35% by task) and 18.3% of cCRE records. We verified with care, checking 1,854,064 retained coordinate-bearing training records and find zero overlaps with a protected BEND interval; it also verifies canonical-DNA disjointness from every BEND evaluation split. Tab. A4 summarizes the measured scope and verification of this procedure.

C.3 Baselines

We restrict the scope of our comparison to models that are pre-trained on eukaryotic DNA sequences and are proposed as general-purpose DNA language models. Under this criterion, we evaluate a broad set of models, including those reported in the main text and additional models reported in the appendix: DNABERT2 (Zhou et al., 2024), GROVER (Sanabria et al., 2024), GENA-LM (Fishman et al., 2023) base, GENA-LM large, ModernGENA (Aspidova et al., 2026) base, ModernGENA large, Nucleotide Transformer-multi species (NT-multi) (Dalla-Torre et al., 2025), Nucleotide Transformer v2, HyenaDNA (Nguyen et al., 2023) tiny, HyenaDNA large, Caduceus (Schiff et al., 2024) PH, Caduceus PS, GENERator (Wu et al., 2025), GenomeOcean, and LucaOne v2 (He et al., 2025). We exclude models such as Enformer (Avsec et al., 2021), which rely on extensive supervised learning to perform well on a limited set of tasks and are therefore not intended to generalize across diverse downstream tasks.

C.4 Benchmark & Evaluation Protocol

C.4.1 Revised NT

Benchmark. The Revised Nucleotide Transformer (NT) benchmark (Dalla-Torre et al., 2025) is a downstream benchmark introduced for evaluating general-purpose genomic representations. It consists of 18 human genomic sequence classification tasks covering three broad categories: ten histone-mark prediction tasks, five regulatory-element annotation tasks for enhancers and promoters, and three splice-site prediction tasks. The revised benchmark replaces the original datasets with

more recent and higher-quality human annotations, uses genomic background sequences, and employs chromosome-held-out test sets to reduce sequence leakage between training and evaluation. Together, these tasks assess whether pretrained representations capture local sequence determinants associated with chromatin state, cis-regulatory elements, and RNA splicing.

Evaluation protocol. We evaluate frozen encoder representations using the reference linear-probing protocol. For each sequence, we apply attention-mask-weighted mean pooling to the final hidden states and standardize the resulting representation using the training-set mean and standard deviation. We then fit a class-balanced multinomial logistic-regression classifier using full-batch L-BFGS for 200 iterations. The inverse regularization strength is selected from $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$ according to development-set Matthews correlation coefficient (MCC). Since Revised NT does not provide a development split, we construct one by taking a stratified 10% split from the training set using a fixed seed and fit the classifier on the remaining 90%. We report the mean test MCC across all 18 tasks.

C.4.2 GUE

Benchmark. Genome Understanding Evaluation (GUE) (Zhou et al., 2024) was introduced as a standardized multi-species benchmark for comparing genomic foundation models across a broad range of sequence classification problems. The full benchmark comprises 27 datasets spanning diverse functional prediction tasks, including epigenetic-mark prediction, transcription-factor binding prediction in human and mouse, promoter and core-promoter detection, and splice-site prediction. As DNA-RADIO is designed for eukaryotic genomic sequences, we restrict our evaluation to the 27 eukaryotic datasets from the epigenetic-mark prediction, human and mouse transcription-factor binding, promoter detection, core-promoter detection, splice-site prediction task families. These tasks cover diverse sequence lengths and regulatory mechanisms across eukaryotic genomes, providing a heterogeneous test of the transferability of learned genomic representations.

Evaluation protocol. We use the same frozen-backbone linear-probing procedure as for Revised NT. Sequence representations are obtained by attention-mask-weighted mean pooling of the final hidden states and standardized using statistics computed from the training split. A class-balanced multinomial logistic-regression classifier is fitted with full-batch L-BFGS for each value of $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. Unlike Revised NT, GUE provides a development split, which we use directly to select C according to MCC before evaluating the selected classifier on the test split.

C.4.3 Genomic Benchmark

Benchmark. Genomic Benchmark (Grešová et al., 2023) was designed to provide a reproducible and easily accessible collection of datasets for genomic sequence classification, addressing inconsistencies in dataset construction and preprocessing that make comparisons between genomic models difficult. We evaluate 9 sequence-classification tasks drawn from multiple organisms and biological settings, including enhancer identification, promoter and open-chromatin-region detection, multiclass regulatory-element classification, coding-versus-intergenic sequence discrimination, cross-species human-versus-worm classification, and enhancer prediction in mouse and *Drosophila*. The benchmark therefore provides a complementary test of both functional-element recognition and more general discrimination of genomic sequence classes across species.

Evaluation protocol. We use the same frozen-representation probing pipeline as above: attention-mask-weighted mean pooling of the final hidden states, standardization using training-set statistics, and class-balanced multinomial logistic regression fitted by full-batch L-BFGS. We sweep $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$ and select it using accuracy on the released development split. The selected classifier is then evaluated on the test split. Following the benchmark's reporting convention, we report the mean test accuracy across all 9 tasks.

C.4.4 BEND

Benchmark. BEND (Marin et al., 2024) was developed to evaluate DNA language models on realistic and biologically meaningful human-genome annotation problems, with an emphasis on challenges that are not well represented by short sequence-level classification benchmarks. It contains

seven tasks spanning substantially different genomic scales and prediction granularities: nucleotide-wise gene finding, long-range enhancer annotation, chromatin-accessibility prediction, histone-modification prediction, CpG-methylation prediction, and two non-coding variant-effect prediction tasks measuring effects on gene expression and disease. The benchmark therefore ranges from local epigenetic signals to nucleotide-resolution annotation and long-range regulatory relationships, while the variant-effect tasks additionally test whether pretrained representations capture functional consequences of sequence variation without task-specific supervision.

Evaluation protocol. We follow the released BEND representation-based evaluation protocol using frozen DNA-RADIO representations. Since the BEND tasks require nucleotide-resolution features, we upsample each token representation according to the number of base pairs covered by the token, yielding one representation per nucleotide, and process long sequences in 4,096-bp chunks. For gene finding, chromatin accessibility, histone modification, CpG methylation, and enhancer annotation, we use the BEND downstream predictor consisting of two 1D convolutional layers with kernel size 3 and GELU activations, followed by task-dependent average pooling and a linear output layer. Following the released protocol, the predictor is trained for 100 epochs with AdamW, and the checkpoint with the highest validation metric is used for test evaluation. We report MCC for gene finding, AUROC for chromatin accessibility, histone modification, and CpG methylation, and AUPRC for enhancer annotation.

The two variant-effect tasks are evaluated using the released BEND zero-shot protocol without fitting a downstream predictor. For each variant, we embed 512-bp reference and alternate sequences centered on the variant, compute the cosine distance between their representations at the variant position, and report AUROC over 50,000 variants for each task.

D Implementation Details

D.1 Multi-teacher distillation

D.1.1 Models and representation extraction

Student models. We initialize DNA-RADIO-Base and DNA-RADIO-Large from the ModernGENA-base and ModernGENA-large checkpoints (Aspidova et al., 2026), respectively. Both students use the same 32k DNA BPE tokenizer and retain the backbone dimensions of their corresponding checkpoints. For the reported distillation runs, all sampled views fall within the full-attention regime, where biological tokens interact bidirectionally with one another and with 32 global slots. We replace token-index rotary positions with nucleotide-coordinate rotary positions computed from each token's covered sequence interval.

Teacher readout layers. For each teacher, we select the earliest layer whose aggregate linear-probe performance reaches at least 95% of the peak performance on a held-out calibration set. The calibration set is constructed from the NT Benchmark after excluding all sequences used in either multi-teacher distillation or post-training, ensuring that the calibration data are completely disjoint from both training stages. This criterion selects L10 for GENERator-Euk-1.2B (Wu et al., 2025), L8 for GenomeOcean-4B (Zhou et al., 2025), and L15 for Nucleotide Transformer-2.5B-multi (Dalla-Torre et al., 2025). All teachers remain frozen during distillation.

D.1.2 Teacher normalization and alignment

PHI-S normalization. For each teacher k, we fit the PHI-S transform (Ranzinger et al., 2024a) offline using $1{,}000{,}000$ teacher-token features sampled from the distillation training split. Let μ_k denote the feature mean, U_k the eigenvectors of the feature covariance, $\{\lambda_{ki}\}_{i=1}^{d_k}$ its eigenvalues, and H_k a fixed Hadamard mixing transform. The normalized feature is

$$\widehat{z} = \alpha_k H_k U_k^\top (z - \mu_k), \qquad \alpha_k = \left(d_k^{-1} \sum_{i=1}^{d_k} \lambda_{ki} \right)^{-1/2}, \tag{A26}$$

Table A5: Teacher-specific angular dispersion used to normalize the summary objective in Eq. 4.

Teacher	Disp_k
GENERator-Euk-1.2B	1.0542
GenomeOcean-4B	0.4765
Nucleotide Transformer-2.5B-multi	0.0546

where $z \in \mathbb{R}^{d_k}$ is a raw teacher feature. We retain the native teacher dimensions, with $d_k = 2048$, 3072, and 2560 for GENERator-Euk-1.2B, GenomeOcean-4B, and Nucleotide Transformer-2.5B-multi, respectively. The fitted transforms are fixed throughout distillation.

Boundary confidence and coverage. Tokens close to the boundary of a sampled view can be less reliable because they are observed with less surrounding context. We therefore assign a confidence score based on the token's distance to the nearest view boundary. Let $d(I, V)$ denote the distance from the midpoint of token interval I to the nearest boundary of view V. Given a boundary width r, we define

$$C(I, V; r) = \min\left\{1, \frac{d(I, V)}{r}\right\}. \tag{A27}$$

Thus, confidence increases linearly from zero at the boundary to one at distance r and remains one farther inside the view; non-biological tokens are assigned zero confidence. We apply the same rule to teacher and student tokens, giving $c_v^{(k)} = C(T_v^{(k)}, V_k; r_k)$ and $c_u^{\mathrm{S}} = C(S_u, V_{\mathrm{S}}; r_{\mathrm{S}})$. Using the confidence-weighted overlaps $a_{uv}^{(k)}$ from Eq. 1, we define

$$q_u^{(k)} = \min\left\{1, \frac{\sum_v a_{uv}^{(k)}}{|S_u|}\right\}, \qquad w_u^{(k)} = c_u^{\mathrm{S}} |S_u| q_u^{(k)}. \tag{A28}$$

We retain correspondences with $q_u^{(k)} \geq 0.5$. The cap on $q_u^{(k)}$ prevents overlapping teacher views from assigning more than full support to a student token.

D.1.3 Objective implementation

Teacher-specific adaptors. For dense matching, each teacher uses a training-only two-layer MLP adaptor A_k with a 512-dimensional hidden layer and GELU activation, preceded by RMSNorm. For the summary objective, we use a separate adaptor B_k consisting of RMSNorm followed by a linear projection to the teacher dimension d_k. All adaptors are optimized jointly with the student and discarded after distillation.

Relational pair sampling. The relational objective in Eq. 3 is evaluated using position pairs sampled within the local training views. Pairs are stratified by genomic distance, with up to 256 valid pairs sampled per teacher. The relational loss is evaluated every fourth distillation update.

Summary target construction. Summary targets are constructed from raw, uncentered teacher features. Teacher and student representations are span-weighted over their shared nucleotide support, and summary loss is applied only when the shared support covers at least 80% of the teacher summary support. The teacher-specific mean directions and angular dispersions used in Eq. 4 are estimated offline from the normalized pooled teacher representations, with the resulting dispersion values reported in Tab. A5.

D.1.4 Distillation algorithm

Algorithm A1 summarizes the complete distillation procedure using the notation introduced above. The three distillation paths are computed from the same local student view, with dense and relational loss using aligned teacher features and summary loss using pooled raw teacher features.

Algorithm A1 Multi-teacher distillation across heterogeneous tokenizers

Require: Corpus $\mathcal{D}$; $K = 3$ frozen teachers; student f_θ; adaptors $\{A_k, B_k\}_{k=1}^{K}$; $\lambda_{\mathrm{rel}}, \lambda_{\mathrm{sum}}$.
Ensure: Distilled encoder parameters θ_0.

Teacher preparation.

Offline: select the teacher readout layers and precompute the PHI-S transforms and summary statistics.
while training budget remains **do**

Student encoding.

Sample a local nucleotide view $x \sim \mathcal{D}$ with $|x| \le 4096$ bp.
Encode x with f_θ to obtain $\{h_u, S_u, c_u^{\mathrm{S}}\}_u$.
$\mathcal{L}_{\mathrm{MTD}}, \mathcal{L}_{\mathrm{rel}}, \mathcal{L}_{\mathrm{sum}} \leftarrow 0$.
for $k = 1, \ldots, K$ **do**

Teacher extraction and normalization.

Run frozen teacher k to obtain $\{z_v^{(k)}, T_v^{(k)}, c_v^{(k)}\}_v$ at the selected readout layer.
$\widetilde{z}_v^{(k)} \leftarrow \mathrm{PHIS}_k(z_v^{(k)})$.

Coordinate alignment and dense matching.

$a_{uv}^{(k)} \leftarrow |S_u \cap T_v^{(k)}| c_v^{(k)}$.
$q_u^{(k)} \leftarrow \min\{1, \sum_v a_{uv}^{(k)} / |S_u|\}$; retain u only if $q_u^{(k)} \ge 0.5$.
$\widetilde{z}_u^{(k)} \leftarrow \dfrac{\sum_v a_{uv}^{(k)} \widetilde{z}_v^{(k)}}{\sum_v a_{uv}^{(k)}}$.
$w_u^{(k)} \leftarrow c_u^{\mathrm{S}} |S_u| q_u^{(k)}$.
$\mathcal{L}_{\mathrm{MTD}} \mathrel{+}= \dfrac{1}{K} \dfrac{\sum_u w_u^{(k)} \left\| A_k(h_u) - \widetilde{z}_u^{(k)} \right\|_2^2}{d_k \left(\sum_u w_u^{(k)} + \epsilon \right)}$.
if relational loss is evaluated on this update **then**

Local relational matching.

Sample $\mathcal{P}_k$ from valid local pairs using the distance strata described above.
$\mathcal{L}_{\mathrm{rel}} \mathrel{+}= \dfrac{1}{K|\mathcal{P}_k|} \displaystyle\sum_{(u,u') \in \mathcal{P}_k} \left[\cos(A_k(h_u), A_k(h_{u'})) - \cos\left(\widetilde{z}_u^{(k)}, \widetilde{z}_{u'}^{(k)} \right) \right]^2$.
end if

Sequence-level summary matching.

Compute $\bar{h}^{(k)}$ and $\bar{z}^{(k)}$ by span-weighted pooling over the shared nucleotide support.
if the shared support covers at least 80% of the teacher summary support **then**
$\mathcal{L}_{\mathrm{sum}} \mathrel{+}= \dfrac{1}{K} \dfrac{\Theta\left(B_k(\bar{h}^{(k)}), \bar{z}^{(k)} \right)^2}{\mathrm{Disp}_k + \epsilon}$.
end if
end for

Student update.

$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{MTD}} + \lambda_{\mathrm{rel}} \mathcal{L}_{\mathrm{rel}} + \lambda_{\mathrm{sum}} \mathcal{L}_{\mathrm{sum}}$.
Update θ and $\{A_k, B_k\}_{k=1}^{K}$ using $\nabla \mathcal{L}$.
end while
return $\theta_0 \leftarrow \theta$ and discard all training-only adaptors.

D.1.5 Training configuration and compute

Loss weights and scheduling. For both reported models, we use $\lambda_{\mathrm{rel}} = 0.25$ and $\lambda_{\mathrm{sum}} = 0.014$, with unit weight on $\mathcal{L}_{\mathrm{MTD}}$. The dense and summary objectives are evaluated at every update. The relational objective is evaluated every fourth update and rescaled by four when active, preserving its nominal contribution across updates. All teacher-derived loss terms are linearly ramped over the first 5% of processed student tokens.

Sampling and optimization. Both models are trained using local views of 256, 512, $1{,}024$, $2{,}048$, and $4{,}096$ bp for a budget of 1.1B student tokens. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.10, and global gradient-norm clipping at 1.0. The peak learning rate is 10^{-4} for both models;

it is linearly warmed up over the first 2% of the token budget, held constant during the middle of training, and linearly decayed to 10% of the peak value over the final 10%. Training uses bfloat16 precision, a microbatch budget of $32{,}768$ student tokens, and four-step gradient accumulation.

Compute. Each model is distilled on 8 NVIDIA A100-80GB GPUs. The Base and Large models use the same data, tokenizer, teacher configuration, normalization artifacts, loss schedule, and optimization settings; only the initialized weights and backbone dimensions differ.

D.2 Information-Constrained Relational Post-Training

D.2.1 Post-training implementation

Backbone and representation extraction. We initialize post-training from the multi-teacher-distilled DNA-RADIO encoder and fully fine-tune it without introducing any additional projection or classification head. Sequence representations are obtained by attention-mask-weighted mean pooling of the final hidden states. We use a maximum token length of 256.

Structured global batches. Each data-parallel worker samples $D_b = 8$ datasets, $C_b = 2$ classes per dataset, and $M = 12$ sequences per class without replacement, giving a local batch of 192 sequences. The sampled dataset–class cells are shared across workers, while individual sequences are sampled independently. Representations are gathered across eight workers, yielding a global candidate pool of $B = 1{,}536$ sequences. For each sequence, all other distinct training examples in the global batch are used as candidates, excluding duplicate occurrences of the same example.

Fixed prior transform. Before constructing the relational prior, we fit a fixed soft-ZCA transform using 20,000 training sequences embedded by the distilled initialization θ_0. Let μ and Σ denote the corresponding feature mean and covariance. We use

$$T(z) = (z - \mu)W_{\mathrm{tf}}, \qquad W_{\mathrm{tf}} = \left(\Sigma + a\frac{\mathrm{tr}(\Sigma)}{d}I\right)^{-1/2}. \tag{A29}$$

The regularization strength a is selected by maximizing within-dataset retrieval mAP of the frozen initialization. The fitted transform remains fixed throughout post-training. The EMA-baseline distribution is constructed in the transformed space, whereas the student matching distribution is constructed from the raw pooled representation.

Information-budget solver. We use temperature $\tau = 0.07$ and information budget $\delta = 0.2$. The tilting strength α in Eq. 9 is shared across the global batch. Since $r_{ij} \in \{-1, 0, +1\}$, we aggregate the EMA-baseline probability mass within the three relation groups and solve the resulting batch-averaged KL constraint by scalar bisection with 28 iterations. The resulting target is treated as a no-gradient target.

EMA baseline. The EMA baseline uses momentum $m = 0.999$ and remains fixed at θ_0 for the first 500 optimization steps. Thereafter, it is updated after every student optimization step.

D.2.2 Post-training algorithm

Algorithm A2 summarizes the complete post-training procedure using the implementation details described above.

D.2.3 Optimization

Optimization. We optimize the student with AdamW using learning rate 2×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01, and global gradient-norm clipping at 1.0. The learning rate follows a cosine schedule parameterized over 48,000 optimization steps, with a 5% warm-up. For the reported model, we use the checkpoint at 4,000 optimization steps.

Algorithm A2 Information-constrained relational post-training

Require: Distilled encoder f_{θ_0}; labeled corpus $\mathcal{D}_{\mathrm{PT}}$; information budget δ; temperature τ; EMA momentum m; EMA warm-up W.
Ensure: Post-trained encoder parameters θ.

Initialization.

$\theta \leftarrow \theta_0$ and $\theta_{\mathrm{base}} \leftarrow \theta_0$; update θ_{base} only through EMA.
Fit and freeze the soft-ZCA transform T using representations from f_{θ_0}.
for each optimization step t **do**

Structured global batch.

Sample D_b datasets, C_b classes per dataset, and M sequences per class on each worker.
Encode the sampled sequences with f_θ and $f_{\theta_{\mathrm{base}}}$, and gather their representations across workers.
For each sequence x_i, use distinct training examples as candidates and construct relation scores r_{ij}.

Information-constrained target.

Construct the EMA-baseline distribution $p_i^{\theta_{\mathrm{base}}}$ in the fixed transformed space.
Solve a shared $\alpha \geq 0$ such that $\frac{1}{B}\sum_i D_{\mathrm{KL}}\left(q_i(\alpha) \| p_i^{\theta_{\mathrm{base}}}\right) = \delta$.
$q_{ij} \leftarrow \mathrm{sg}\left[\dfrac{p_{ij}^{\theta_{\mathrm{base}}} \exp(\alpha r_{ij})}{\sum_{k \in \mathcal{C}_i} p_{ik}^{\theta_{\mathrm{base}}} \exp(\alpha r_{ik})}\right]$.

Student matching.

Construct the student distribution p_i^θ from the raw pooled representations.
$\mathcal{L}_{\mathrm{tilt}} \leftarrow \dfrac{1}{B}\sum_i D_{\mathrm{KL}}\left(q_i \| p_i^\theta\right)$.
Update θ using $\nabla_\theta \mathcal{L}_{\mathrm{tilt}}$.

EMA update.

if $t > W$ **then**
$\theta_{\mathrm{base}} \leftarrow m\theta_{\mathrm{base}} + (1-m)\theta$.
end if
end for
return θ.

E Additional Results

E.1 Additional Analysis

E.1.1 Detailed linear probing results

We present the complete per-task results for every model and benchmark included in the layer-wise analysis in Fig. 2. For each model, task, and layer, performance is evaluated under three training configurations, each with three different random seeds. The results are shown in Fig. A1, Fig. A2, Fig. A3, Fig. A4, Fig. A5, Fig. A6, Fig. A7, Fig. A8, and Fig. A9.

E.1.2 Layer-wise linear probing of DNA-RADIO

To examine how the learned representations evolve across depth, we linearly probe every layer of ModernGENA-base, distilled DNA-RADIO-$\mathrm{B_d}$, and post-trained DNA-RADIO-B on the original NT Benchmark and Genomic Benchmark. As shown in Fig. A10, the advantage of DNA-RADIO becomes more pronounced in deeper layers: distillation improves over the pretrained backbone, and post-training further strengthens this gain. In contrast, the representations at the embedding layer remain similar across the three models.

ModernGENA-base tends to peak before the final layer and subsequently degrade, whereas DNA-RADIO maintains strong performance through the end of the network. This suggests that distillation and post-training primarily refine the deeper representations and make the final-layer features more useful for downstream tasks.

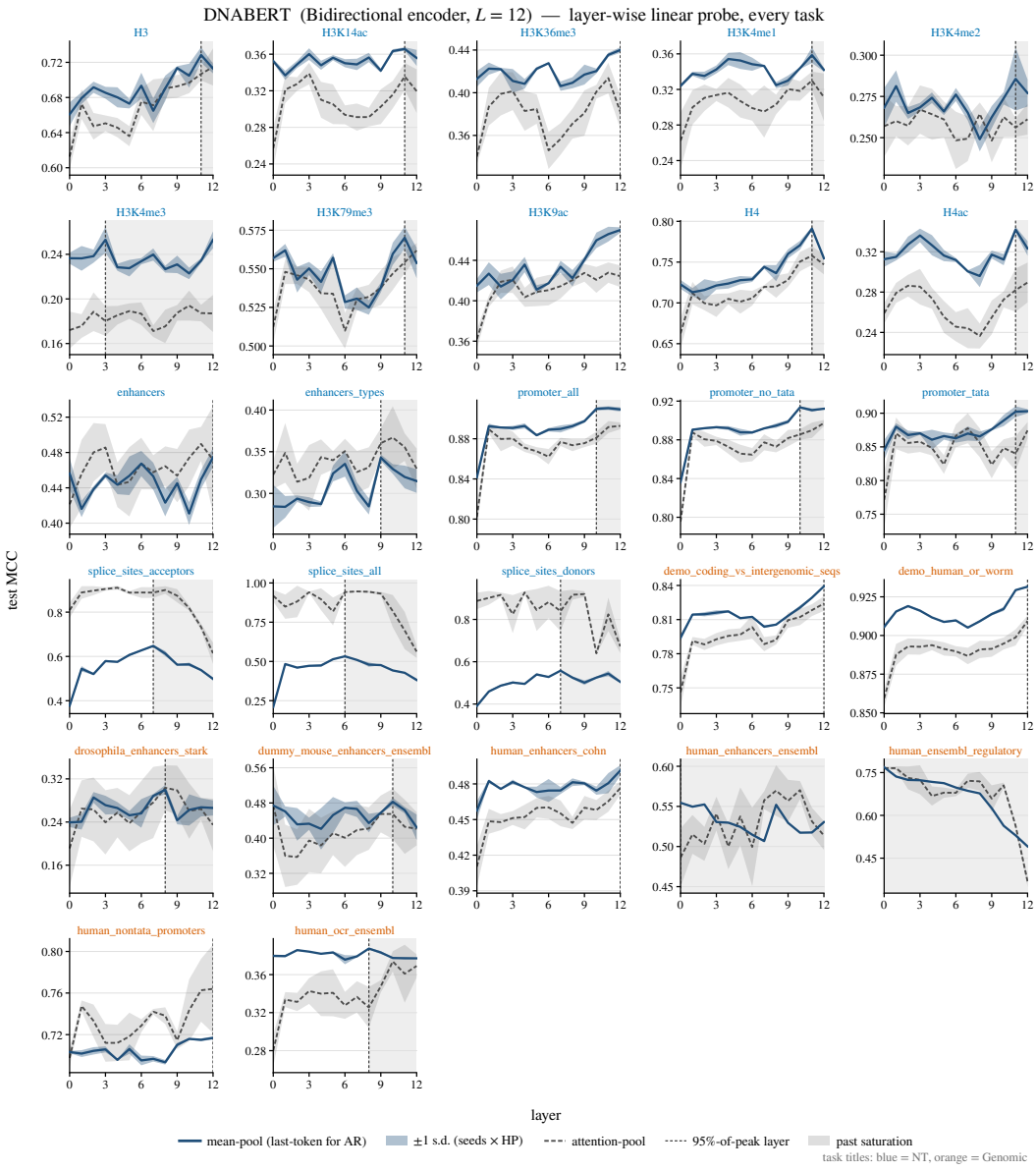


Figure A1: **Detailed layer-wise linear probing results for DNABERT.** We report layer-wise linear probing results for each task in the Nucleotide Transformer and Genomic Benchmark. We evaluate both mean-pool and attention-pool, with mean-pool generally yielding better performance. We also report standard deviations across three random seeds and three hyperparameter configurations.

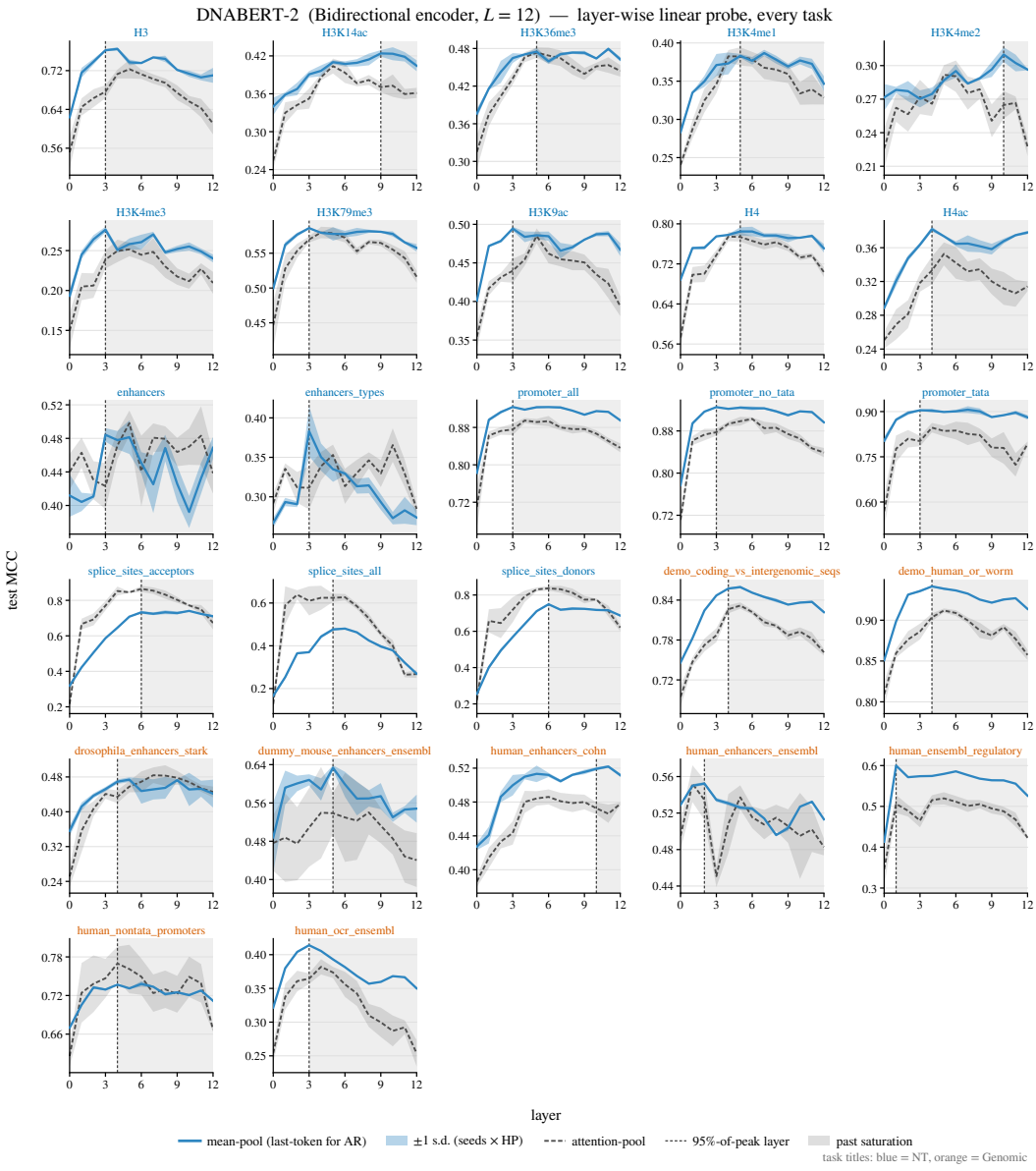


Figure A2: **Detailed layer-wise linear probing results for DNABERT2.** We report layer-wise linear probing results for each task in the Nucleotide Transformer and Genomic Benchmark. We evaluate both mean-pool and attention-pool, with mean-pool generally yielding better performance. We also report standard deviations across three random seeds and three hyperparameter configurations.

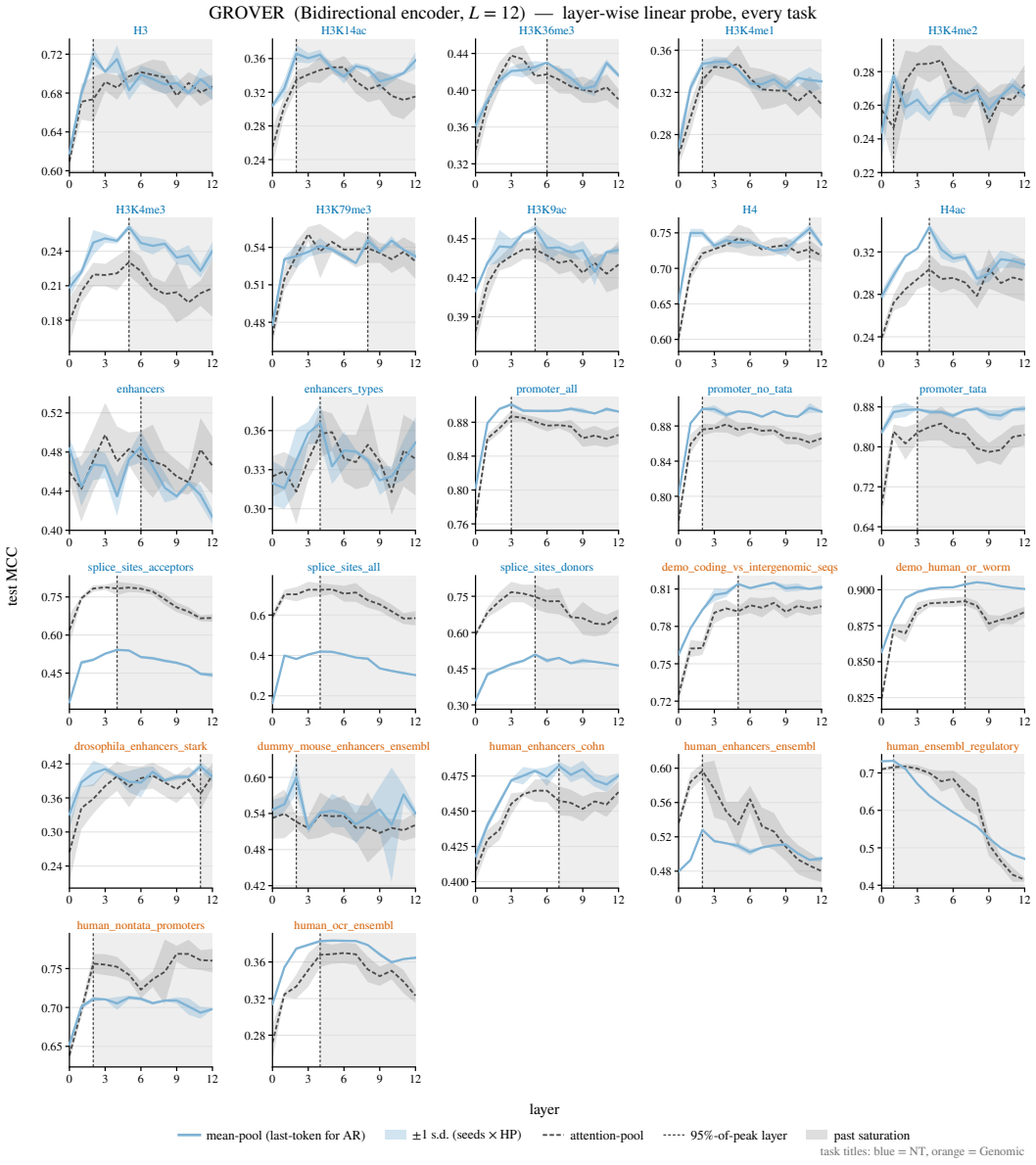


Figure A3: **Detailed layer-wise linear probing results for GROVER.** We report layer-wise linear probing results for each task in the Nucleotide Transformer and Genomic Benchmark. We evaluate both mean-pool and attention-pool, with mean-pool generally yielding better performance. We also report standard deviations across three random seeds and three hyperparameter configurations.

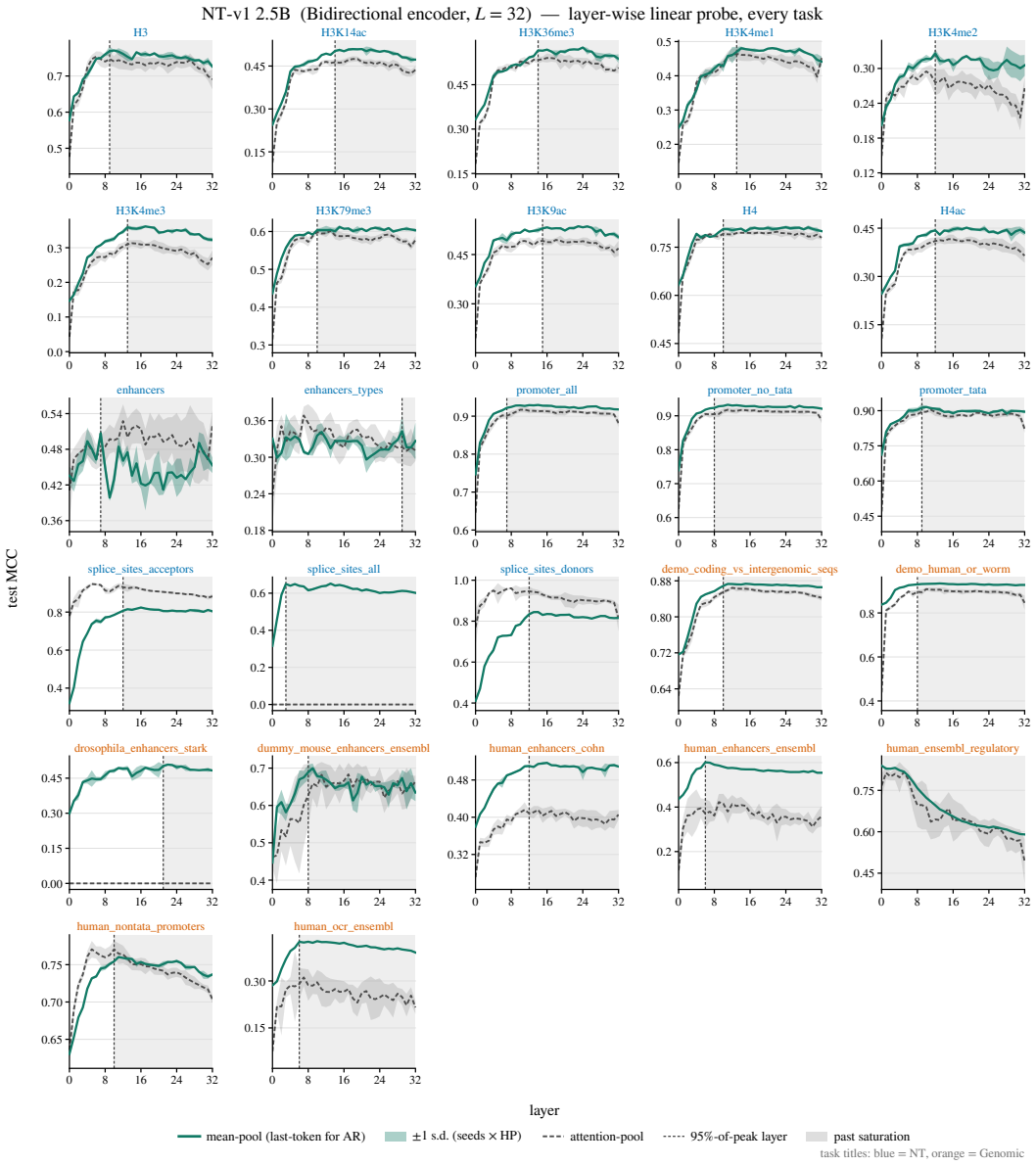


Figure A4: **Detailed layer-wise linear probing results for Nucleotide Transformer v1 Multi-Species 2.5B (NT-v1 2.5B).** We report layer-wise linear probing results for each task in the Nucleotide Transformer and Genomic Benchmark. We evaluate both mean-pool and attention-pool, with mean-pool generally yielding better performance. We also report standard deviations across three random seeds and three hyperparameter configurations.

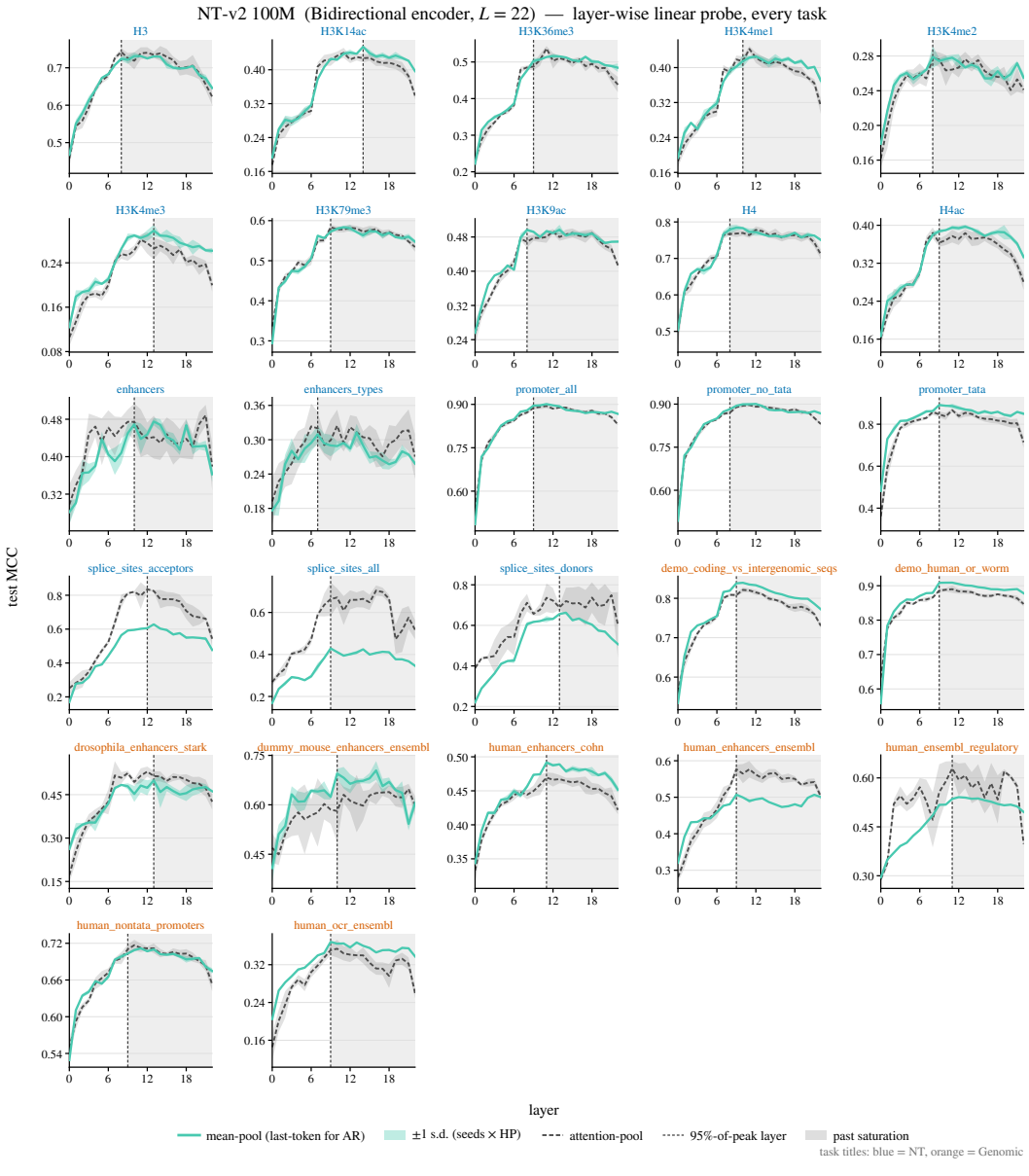


Figure A5: **Detailed layer-wise linear probing results for Nucleotide Transformer v2 100M (NT-v2 100M).** We report layer-wise linear probing results for each task in the Nucleotide Transformer and Genomic Benchmark. We evaluate both mean-pool and attention-pool, with mean-pool generally yielding better performance. We also report standard deviations across three random seeds and three hyperparameter configurations.

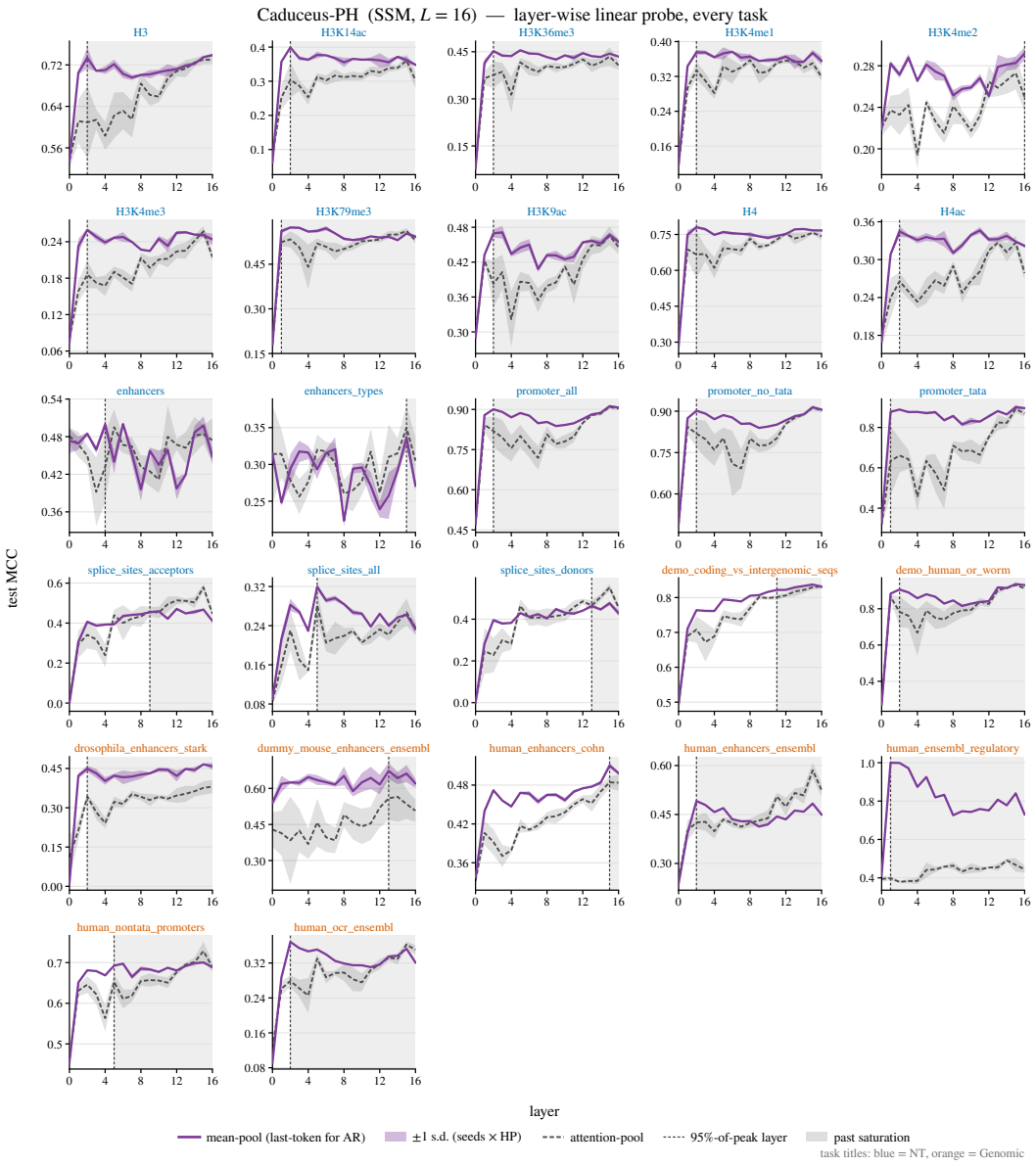


Figure A6: **Detailed layer-wise linear probing results for Caduceus-PH.** We report layer-wise linear probing results for each task in the Nucleotide Transformer and Genomic Benchmark. We evaluate both mean-pool and attention-pool, with mean-pool generally yielding better performance. We also report standard deviations across three random seeds and three hyperparameter configurations.

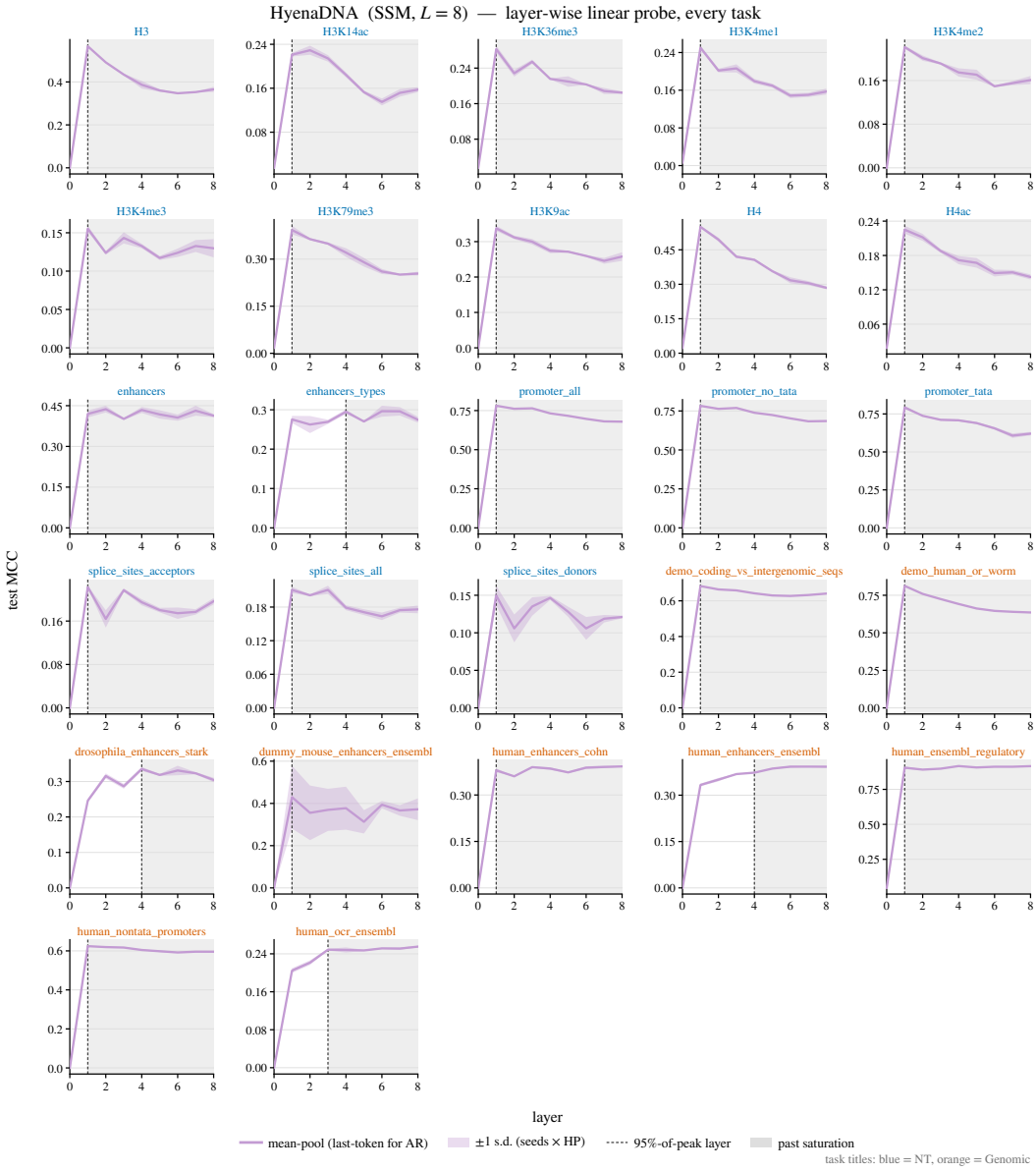


Figure A7: **Detailed layer-wise linear probing results for HyenaDNA.** We report layer-wise linear probing results for each task in the Nucleotide Transformer and Genomic Benchmark, using the representation of the end-of-sequence (EOS) token. We also report standard deviations across three random seeds and three hyperparameter configurations.

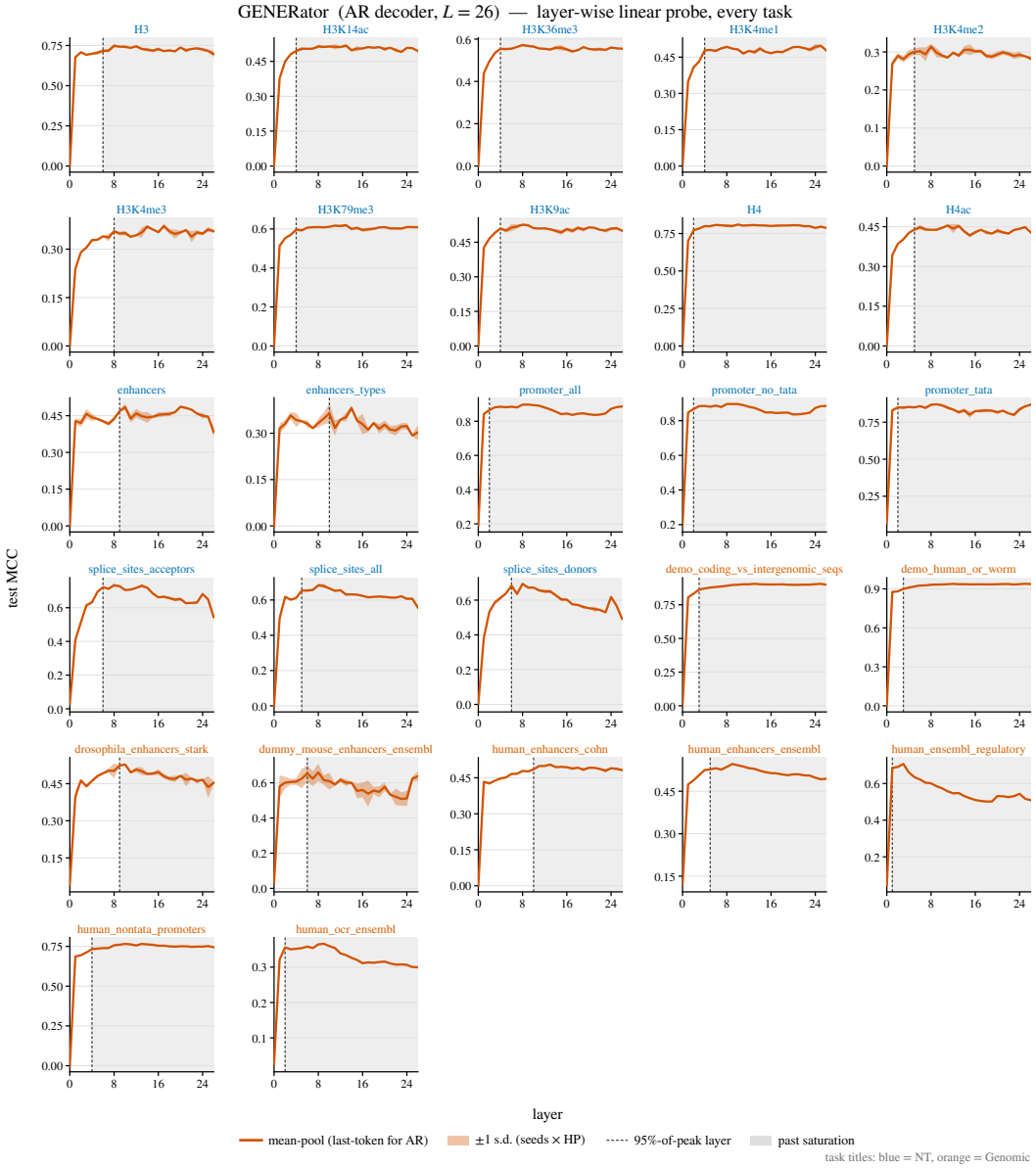


Figure A8: **Detailed layer-wise linear probing results for GENERator.** We report layer-wise linear probing results for each task in the Nucleotide Transformer and Genomic Benchmark, using the representation of the end-of-sequence (EOS) token. We also report standard deviations across three random seeds and three hyperparameter configurations.

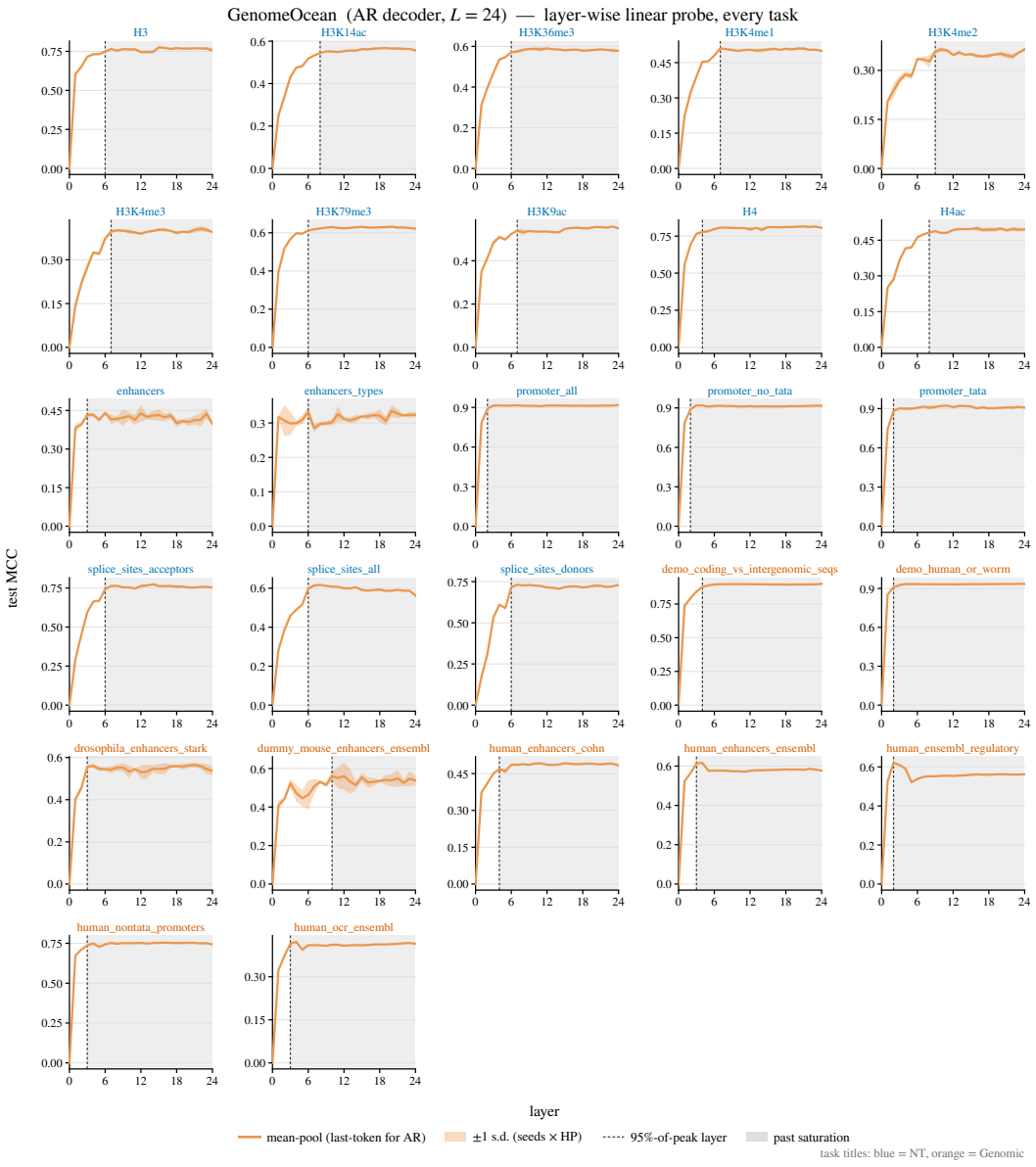


Figure A9: **Detailed layer-wise linear probing results for GenomeOcean 4B.** We report layer-wise linear probing results for each task in the Nucleotide Transformer and Genomic Benchmark, using the representation of the end-of-sequence (EOS) token. We also report standard deviations across three random seeds and three hyperparameter configurations.

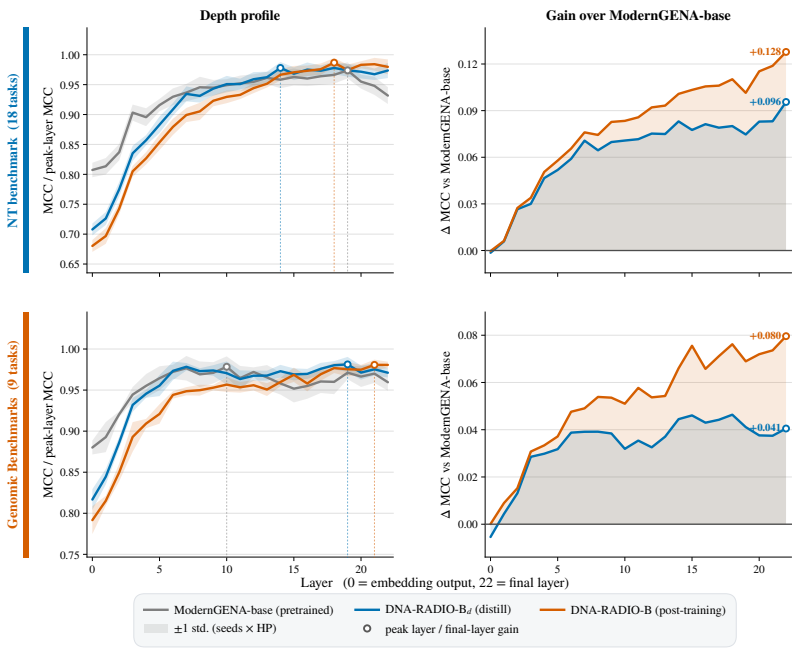


Figure A10: **Layer-wise linear probing of DNA-RADIO. Left:** task-wise MCC normalized by each task's peak-layer MCC and averaged across tasks. **Right:** absolute MCC gain over ModernGENA-base at the corresponding layer. Circles indicate the peak layer, and shading denotes ± 1 standard deviation over seeds.

E.1.3 t-SNE ANALYSIS BEFORE AND AFTER POST-TRAINING

We visualize the final-layer representations of ModernGENA-base, distilled DNA-RADIO-B_d, and post-trained DNA-RADIO-B on representative tasks from the original NT Benchmark. As shown in Fig. A11, the class structure becomes progressively more distinct after distillation and post-training. The separation is particularly pronounced for the promoter tasks, while the histone and splice-site tasks show more gradual changes. Overall, these visualizations provide a qualitative view of how distillation and subsequent post-training reshape the learned representation space.

E.2 ADDITIONAL BASELINE COMPARISON

Revised NT For the revised Nucleotide Transformer (NT) benchmark (Dalla-Torre et al., 2025), Tab. A6 reports the full results across histone-mark, regulatory-annotation, and splice-annotation tasks. DNA-RADIO achieves the best result on 10 of the 18 tasks and ranks within the top three on 15, with particularly strong results on promoter and splice-site prediction.

Genomic Benchmark Tab. A7 reports results on Genomic Benchmark (Grešová et al., 2023), which covers regulatory sequence classification across multiple species. DNA-RADIO ranks within the top three on eight of the nine tasks, including the best results on the mouse enhancer and human–worm classification tasks, indicating consistently strong performance across diverse sequence-classification settings.

GUE Tab. A8 presents the full Genome Understanding Evaluation (GUE) results (Zhou et al., 2024). On the aggregate task groups, DNA-RADIO achieves the best performance on TF-M, TF-H, PD, and CPD, while ranking second on EMP and SSP, showing strong performance across all major GUE task families.

BEND Tab. A9 presents the full BEND benchmark results (Marin et al., 2024). Across the seven BEND tasks, DNA-RADIO-B ranks second overall with an average score of 0.6363, narrowly trail-

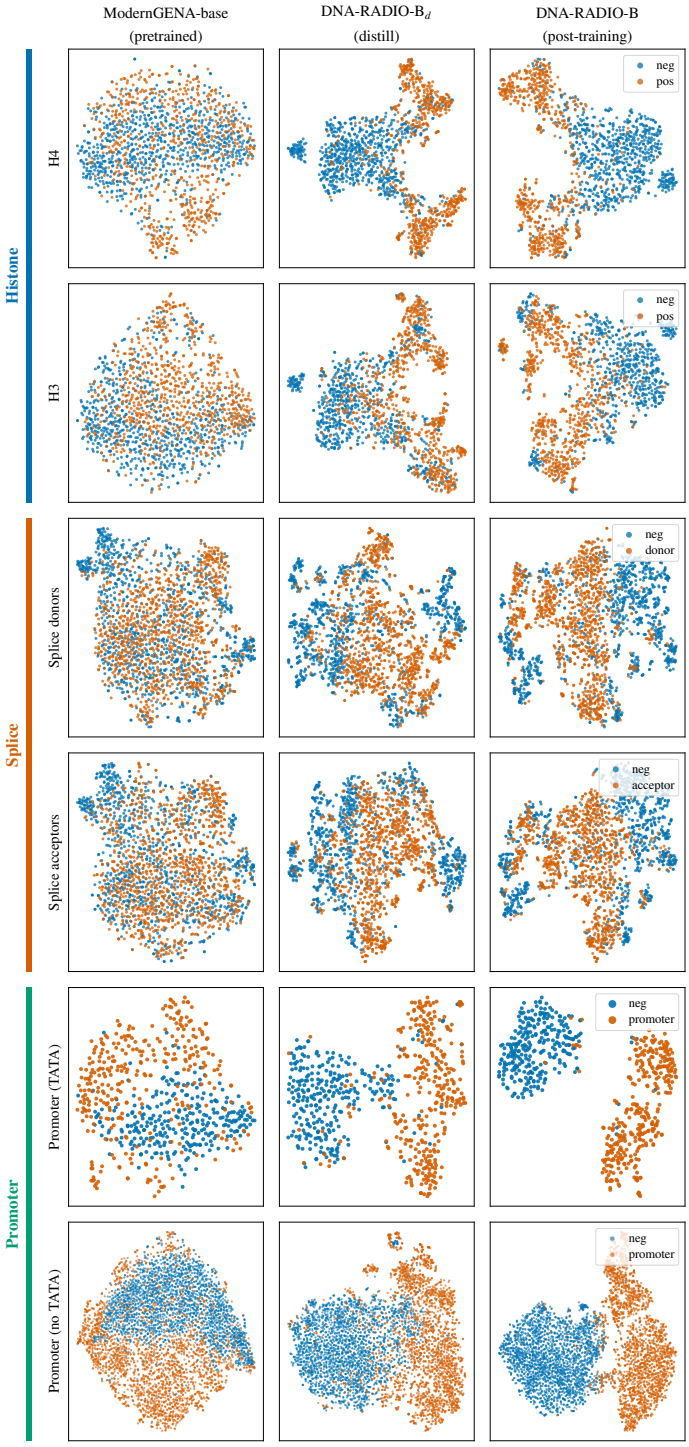


Figure A11: **t-SNE visualization before and after post-training.** Final-layer representations are shown for representative histone, splice-site, and promoter tasks from the original NT Benchmark. Columns correspond to ModernGENA-base, distilled DNA-RADIO-B$_{\mathrm{d}}$, and post-trained DNA-RADIO-B.

Table A6: Full benchmark results on the Revised NT benchmark downstream task suite. Each entry reports the mean with the standard deviation in subscript, over 5 random seeds. best, second, third within each row.

		Encoder-only (masked LM)						Nucleotide Transformer			Long-context (implicit conv. / SSM)				Generative / cross-modal			**Ours**	
		DNABERT-2 (117M)	GROVER (87M)	GENA-LM Base (110M)	GENA-LM-Base T2T-Multi (110M)	ModernGENA base (135M)	ModernGENA large (377M)	NT-multi (100M)	NT-multi (2.5B)	NT-v2 (500M)	HyenaDNA (tiny-1k) (0.4M)	HyenaDNA (large-1m) (6.6M)	Caduceus PH (7.7M)	Caduceus PS (7.7M)	GENERator (1.2B)	GenomeOcean (4B)	LucaOne v2.0 (1.8B)	**DNA-RADIO B_q**(136M)	**DNA-RADIO L_q**(377M)
Histone Markers	H2AFZ	$0.4037_{\pm.0022}$	$0.4138_{\pm.0016}$	$0.4127_{\pm.0048}$	$0.4043_{\pm.0007}$	$0.4113_{\pm.0087}$	$0.4054_{\pm.0093}$	$0.4016_{\pm.0035}$	$0.4112_{\pm.0044}$	$0.3854_{\pm.0084}$	$0.4115_{\pm.0030}$	$0.4908_{\pm.0037}$	$0.3978_{\pm.0026}$	$0.4174_{\pm.0073}$	$0.4394_{\pm.0089}$	$0.4734_{\pm.0057}$	$0.4650_{\pm.0008}$	$0.4419_{\pm.0070}$	$0.4606_{\pm.0087}$
	H3K27ac	$0.4434_{\pm.0048}$	$0.4484_{\pm.0061}$	$0.4496_{\pm.0068}$	$0.4836_{\pm.0068}$	$0.4236_{\pm.0019}$	$0.4400_{\pm.0061}$	$0.4089_{\pm.0091}$	$0.4615_{\pm.0031}$	$0.3916_{\pm.0055}$	$0.4156_{\pm.0066}$	$0.4520_{\pm.0065}$	$0.3954_{\pm.0057}$	$0.4319_{\pm.0037}$	$0.4725_{\pm.0144}$	$0.4744_{\pm.0024}$	$0.4284_{\pm.0063}$	$0.4839_{\pm.0080}$	$0.4733_{\pm.0151}$
	H3K27me3	$0.5559_{\pm.0091}$	$0.5558_{\pm.0037}$	$0.5667_{\pm.0021}$	$0.5812_{\pm.0033}$	$0.5494_{\pm.0059}$	$0.5447_{\pm.0066}$	$0.5211_{\pm.0038}$	$0.5680_{\pm.0011}$	$0.5392_{\pm.0049}$	$0.5070_{\pm.0027}$	$0.5481_{\pm.0046}$	$0.5677_{\pm.0028}$	$0.5587_{\pm.0051}$	$0.5848_{\pm.0050}$	$0.5888_{\pm.0055}$	$0.5515_{\pm.0054}$	$0.5790_{\pm.0054}$	$0.5819_{\pm.0038}$
	H3K36me3	$0.5539_{\pm.0079}$	$0.4919_{\pm.0134}$	$0.5600_{\pm.0041}$	$0.5376_{\pm.0030}$	$0.5113_{\pm.0056}$	$0.5170_{\pm.0027}$	$0.4798_{\pm.0048}$	$0.5766_{\pm.0021}$	$0.5065_{\pm.0065}$	$0.4948_{\pm.0031}$	$0.5442_{\pm.0090}$	$0.5235_{\pm.0040}$	$0.5374_{\pm.0084}$	$0.5791_{\pm.0040}$	$0.5797_{\pm.0045}$	$0.5561_{\pm.0058}$	$0.5977_{\pm.0040}$	$0.6087_{\pm.0041}$
	H3K4me1	$0.4349_{\pm.0014}$	$0.4191_{\pm.0068}$	$0.4279_{\pm.0022}$	$0.4296_{\pm.0046}$	$0.4002_{\pm.0046}$	$0.4168_{\pm.0064}$	$0.4121_{\pm.0074}$	$0.4138_{\pm.0026}$	$0.3999_{\pm.0030}$	$0.4125_{\pm.0031}$	$0.4391_{\pm.0061}$	$0.4031_{\pm.0069}$	$0.4131_{\pm.0064}$	$0.4483_{\pm.0074}$	$0.4607_{\pm.0034}$	$0.4398_{\pm.0036}$	$0.4575_{\pm.0033}$	$0.4746_{\pm.0049}$
	H3K4me2	$0.4737_{\pm.0054}$	$0.4808_{\pm.0031}$	$0.4944_{\pm.0061}$	$0.4859_{\pm.0067}$	$0.4806_{\pm.0099}$	$0.4763_{\pm.0085}$	$0.4678_{\pm.0071}$	$0.4792_{\pm.0051}$	$0.4726_{\pm.0098}$	$0.5192_{\pm.0074}$	$0.5237_{\pm.0068}$	$0.4637_{\pm.0042}$	$0.5047_{\pm.0078}$	$0.5374_{\pm.0040}$	$0.5532_{\pm.0057}$	$0.5113_{\pm.0126}$	$0.5240_{\pm.0029}$	$0.5408_{\pm.0028}$
	H3K4me3	$0.6138_{\pm.0089}$	$0.5917_{\pm.0070}$	$0.6439_{\pm.0099}$	$0.6094_{\pm.0014}$	$0.5836_{\pm.0058}$	$0.6187_{\pm.0019}$	$0.5743_{\pm.0077}$	$0.6166_{\pm.0034}$	$0.6239_{\pm.0046}$	$0.6104_{\pm.0086}$	$0.6165_{\pm.0083}$	$0.6016_{\pm.0074}$	$0.5887_{\pm.0079}$	$0.6192_{\pm.0160}$	$0.6254_{\pm.0110}$	$0.6105_{\pm.0053}$	$0.6233_{\pm.0071}$	$0.6197_{\pm.0046}$
	H3K9ac	$0.5020_{\pm.0109}$	$0.4814_{\pm.0068}$	$0.5217_{\pm.0102}$	$0.5060_{\pm.0119}$	$0.4939_{\pm.0075}$	$0.5343_{\pm.0112}$	$0.5131_{\pm.0080}$	$0.4924_{\pm.0073}$	$0.5188_{\pm.0114}$	$0.5121_{\pm.0037}$	$0.4858_{\pm.0064}$	$0.4819_{\pm.0061}$	$0.5141_{\pm.0208}$	$0.5312_{\pm.0134}$	$0.5389_{\pm.0074}$	$0.4876_{\pm.0045}$	$0.5277_{\pm.0048}$	$0.5405_{\pm.0084}$
	H3K9me3	$0.3019_{\pm.0061}$	$0.3465_{\pm.0193}$	$0.4038_{\pm.0081}$	$0.3852_{\pm.0120}$	$0.3521_{\pm.0094}$	$0.2980_{\pm.0203}$	$0.2895_{\pm.0073}$	$0.3827_{\pm.0089}$	$0.3005_{\pm.0168}$	$0.3478_{\pm.0089}$	$0.3795_{\pm.0114}$	$0.3873_{\pm.0186}$	$0.3900_{\pm.0071}$	$0.4205_{\pm.0139}$	$0.4212_{\pm.0060}$	$0.3846_{\pm.0096}$	$0.3964_{\pm.0075}$	$0.4105_{\pm.0247}$
	H4K20me1	$0.6003_{\pm.0062}$	$0.5607_{\pm.0057}$	$0.6174_{\pm.0048}$	$0.6140_{\pm.0046}$	$0.5617_{\pm.0043}$	$0.5842_{\pm.0089}$	$0.5588_{\pm.0050}$	$0.6113_{\pm.0070}$	$0.5466_{\pm.0011}$	$0.5687_{\pm.0028}$	$0.6090_{\pm.0043}$	$0.5726_{\pm.0035}$	$0.5684_{\pm.0071}$	$0.6273_{\pm.0047}$	$0.6150_{\pm.0055}$	$0.6163_{\pm.0037}$	$0.6231_{\pm.0048}$	$0.6360_{\pm.0057}$
Regulatory Annotation	Enhancer	$0.4892_{\pm.0088}$	$0.4888_{\pm.0047}$	$0.5054_{\pm.0060}$	$0.5000_{\pm.0093}$	$0.4864_{\pm.0055}$	$0.4892_{\pm.0053}$	$0.4646_{\pm.0036}$	$0.4920_{\pm.0036}$	$0.4899_{\pm.0074}$	$0.4504_{\pm.0021}$	$0.4998_{\pm.0032}$	$0.4770_{\pm.0066}$	$0.4790_{\pm.0067}$	$0.4932_{\pm.0071}$	$0.4899_{\pm.0039}$	$0.4886_{\pm.0038}$	$0.4973_{\pm.0079}$	$0.5172_{\pm.0033}$
	Enhancer type	$0.3692_{\pm.0082}$	$0.3551_{\pm.0041}$	$0.3706_{\pm.0057}$	$0.3658_{\pm.0044}$	$0.3524_{\pm.0051}$	$0.3515_{\pm.0041}$	$0.3389_{\pm.0044}$	$0.3879_{\pm.0035}$	$0.3566_{\pm.0049}$	$0.3212_{\pm.0109}$	$0.3598_{\pm.0025}$	$0.3453_{\pm.0040}$	$0.3419_{\pm.0044}$	$0.3952_{\pm.0043}$	$0.3589_{\pm.0055}$	$0.3608_{\pm.0048}$	$0.3744_{\pm.0039}$	$0.3794_{\pm.0080}$
	Promoter all	$0.6798_{\pm.0017}$	$0.7036_{\pm.0059}$	$0.7139_{\pm.0075}$	$0.7235_{\pm.0037}$	$0.7250_{\pm.0103}$	$0.7341_{\pm.0058}$	$0.7183_{\pm.0046}$	$0.7354_{\pm.0094}$	$0.7113_{\pm.0097}$	$0.7175_{\pm.0051}$	$0.7167_{\pm.0063}$	$0.7255_{\pm.0061}$	$0.7455_{\pm.0035}$	$0.7744_{\pm.0116}$	$0.7379_{\pm.0035}$	$0.7514_{\pm.0033}$	$0.7581_{\pm.0042}$	$0.7554_{\pm.0079}$
	Promoter non-TATA	$0.7363_{\pm.0030}$	$0.7197_{\pm.0016}$	$0.7476_{\pm.0033}$	$0.7538_{\pm.0041}$	$0.7364_{\pm.0048}$	$0.7570_{\pm.0035}$	$0.7280_{\pm.0057}$	$0.7567_{\pm.0038}$	$0.7405_{\pm.0048}$	$0.7428_{\pm.0046}$	$0.7380_{\pm.0070}$	$0.7516_{\pm.0049}$	$0.7712_{\pm.0012}$	$0.7631_{\pm.0072}$	$0.7637_{\pm.0028}$	$0.7562_{\pm.0061}$	$0.7608_{\pm.0029}$	$0.7826_{\pm.0065}$
	Promoter TATA	$0.7004_{\pm.0227}$	$0.7536_{\pm.0045}$	$0.7270_{\pm.0096}$	$0.6910_{\pm.0205}$	$0.7401_{\pm.0191}$	$0.6869_{\pm.0156}$	$0.6614_{\pm.0269}$	$0.7670_{\pm.0161}$	$0.6999_{\pm.0244}$	$0.7068_{\pm.0427}$	$0.7550_{\pm.0175}$	$0.7158_{\pm.0423}$	$0.7466_{\pm.0298}$	$0.7925_{\pm.0211}$	$0.7692_{\pm.0192}$	$0.7897_{\pm.0256}$	$0.8101_{\pm.0106}$	$0.8455_{\pm.0170}$
Splice Annotation	Splice acceptor	$0.6219_{\pm.0051}$	$0.4777_{\pm.0036}$	$0.5467_{\pm.0049}$	$0.5205_{\pm.0040}$	$0.5643_{\pm.0070}$	$0.4731_{\pm.0044}$	$0.4565_{\pm.0058}$	$0.6623_{\pm.0018}$	$0.4720_{\pm.0059}$	$0.4477_{\pm.0054}$	$0.4998_{\pm.0033}$	$0.5140_{\pm.0045}$	$0.6255_{\pm.0030}$	$0.6645_{\pm.0109}$	$0.5875_{\pm.0023}$	$0.5361_{\pm.0045}$	$0.7072_{\pm.0062}$	$0.7025_{\pm.0050}$
	Splice site all	$0.3464_{\pm.0072}$	$0.3012_{\pm.0102}$	$0.3231_{\pm.0088}$	$0.3192_{\pm.0040}$	$0.3570_{\pm.0076}$	$0.2906_{\pm.0054}$	$0.3403_{\pm.0133}$	$0.4785_{\pm.0058}$	$0.3687_{\pm.0053}$	$0.2932_{\pm.0038}$	$0.3625_{\pm.0063}$	$0.3196_{\pm.0047}$	$0.4254_{\pm.0030}$	$0.5103_{\pm.0027}$	$0.4205_{\pm.0040}$	$0.4666_{\pm.0072}$	$0.4969_{\pm.0056}$	$0.5001_{\pm.0042}$
	Splice donor	$0.6469_{\pm.0036}$	$0.5053_{\pm.0024}$	$0.5840_{\pm.0031}$	$0.5614_{\pm.0033}$	$0.5860_{\pm.0068}$	$0.5363_{\pm.0060}$	$0.5347_{\pm.0056}$	$0.6917_{\pm.0024}$	$0.5241_{\pm.0084}$	$0.4547_{\pm.0052}$	$0.4994_{\pm.0014}$	$0.5352_{\pm.0032}$	$0.6581_{\pm.0033}$	$0.6865_{\pm.0076}$	$0.5962_{\pm.0142}$	$0.5831_{\pm.0071}$	$0.7218_{\pm.0051}$	$0.7253_{\pm.0032}$

Table A7: Full benchmark results on the Genomic Benchmark downstream task suite. Each entry reports the mean with the standard deviation in subscript, over 5 random seeds. best, second, third within each row.

Task	Encoder-only (masked LM)						Nucleotide Transformer			Long-context (implicit conv. / SSM)				Generative / cross-modal			**Ours**	
	DNABERT-2 (117M)	GROVER (87M)	GENA-LM Base (110M)	GENA-LM-Base T2T-Multi (110M)	ModemGENA base (135M)	ModernGENA large (377M)	NT-multi (100M)	NT-multi (2.5B)	NT-v2 (500M)	HyenaDNA (tiny-1k) (0.4M)	HyenaDNA (large-1m) (6.6M)	Caduceus PH (7.7M)	Caduceus PS (7.7M)	GENERator (1.2B)	GenomeOcean (4B)	LucaOne v2.0 (1.8B)	**DNA-RADIO B_d**(136M)	**DNA-RADIO L_d**(377M)
Mouse																		
EE	$0.8322_{\pm.0047}$	$0.7736_{\pm.0279}$	$0.8438_{\pm.0183}$	$0.8355_{\pm.0054}$	$0.7959_{\pm.0063}$	$0.8174_{\pm.0171}$	$0.8455_{\pm.0129}$	$0.8479_{\pm.0103}$	$0.8554_{\pm.0117}$	$0.7215_{\pm.0238}$	$0.7504_{\pm.0133}$	$0.8231_{\pm.0079}$	$0.8215_{\pm.0332}$	$0.8322_{\pm.0047}$	$0.8215_{\pm.0107}$	$0.7512_{\pm.0054}$	$0.8570_{\pm.0069}$	$0.8702_{\pm.0086}$
Human																		
CI	$0.9113_{\pm.0008}$	$0.9048_{\pm.0010}$	$0.9271_{\pm.0005}$	$0.9269_{\pm.0007}$	$0.9227_{\pm.0006}$	$0.9251_{\pm.0004}$	$0.8861_{\pm.0002}$	$0.9321_{\pm.0021}$	$0.8921_{\pm.0004}$	$0.8778_{\pm.0004}$	$0.8730_{\pm.0006}$	$0.9226_{\pm.0002}$	$0.9355_{\pm.0004}$	$0.9482_{\pm.0004}$	$0.9423_{\pm.0017}$	$0.9530_{\pm.0003}$	$0.9404_{\pm.0006}$	$0.9457_{\pm.0008}$
EC	$0.7546_{\pm.0025}$	$0.7351_{\pm.0010}$	$0.7623_{\pm.0028}$	$0.7576_{\pm.0031}$	$0.7441_{\pm.0032}$	$0.7480_{\pm.0050}$	$0.7249_{\pm.0030}$	$0.7554_{\pm.0010}$	$0.7245_{\pm.0034}$	$0.7250_{\pm.0010}$	$0.7263_{\pm.0034}$	$0.7472_{\pm.0013}$	$0.7443_{\pm.0027}$	$0.7554_{\pm.0033}$	$0.7383_{\pm.0027}$	$0.7240_{\pm.0020}$	$0.7571_{\pm.0016}$	$0.7593_{\pm.0011}$
EE	$0.7567_{\pm.0008}$	$0.7471_{\pm.0006}$	$0.7538_{\pm.0008}$	$0.7550_{\pm.0004}$	$0.7466_{\pm.0008}$	$0.7674_{\pm.0006}$	$0.7488_{\pm.0003}$	$0.7788_{\pm.0006}$	$0.7709_{\pm.0005}$	$0.7368_{\pm.0005}$	$0.7252_{\pm.0003}$	$0.7423_{\pm.0004}$	$0.7492_{\pm.0002}$	$0.7591_{\pm.0005}$	$0.7713_{\pm.0013}$	$0.7971_{\pm.0004}$	$0.7779_{\pm.0009}$	$0.7814_{\pm.0005}$
ER	$0.6812_{\pm.0006}$	$0.6478_{\pm.0006}$	$0.6875_{\pm.0006}$	$0.6872_{\pm.0012}$	$0.6419_{\pm.0009}$	$0.6577_{\pm.0003}$	$0.6624_{\pm.0003}$	$0.7299_{\pm.0005}$	$0.6860_{\pm.0004}$	$0.7230_{\pm.0003}$	$0.7358_{\pm.0008}$	$0.6392_{\pm.0003}$	$0.6490_{\pm.0005}$	$0.6541_{\pm.0006}$	$0.6704_{\pm.0009}$	$0.7658_{\pm.0004}$	$0.7289_{\pm.0006}$	$0.7141_{\pm.0004}$
non-TATA	$0.8528_{\pm.0030}$	$0.8463_{\pm.0008}$	$0.8627_{\pm.0008}$	$0.8591_{\pm.0004}$	$0.8661_{\pm.0008}$	$0.8758_{\pm.0016}$	$0.8344_{\pm.0010}$	$0.8631_{\pm.0024}$	$0.8414_{\pm.0021}$	$0.8331_{\pm.0006}$	$0.8383_{\pm.0014}$	$0.8503_{\pm.0034}$	$0.8621_{\pm.0007}$	$0.8787_{\pm.0025}$	$0.8754_{\pm.0015}$	$0.8901_{\pm.0028}$	$0.8788_{\pm.0004}$	$0.8834_{\pm.0007}$
OCR	$0.6754_{\pm.0006}$	$0.6818_{\pm.0007}$	$0.6845_{\pm.0004}$	$0.6854_{\pm.0005}$	$0.6741_{\pm.0006}$	$0.6858_{\pm.0005}$	$0.6685_{\pm.0010}$	$0.6971_{\pm.0012}$	$0.6816_{\pm.0006}$	$0.6693_{\pm.0004}$	$0.6742_{\pm.0004}$	$0.6740_{\pm.0010}$	$0.6832_{\pm.0006}$	$0.6943_{\pm.0006}$	$0.6990_{\pm.0012}$	$0.7176_{\pm.0008}$	$0.7039_{\pm.0006}$	$0.7127_{\pm.0013}$
Cross-species																		
Human–Worm	$0.9564_{\pm.0005}$	$0.9506_{\pm.0004}$	$0.9653_{\pm.0003}$	$0.9674_{\pm.0009}$	$0.9607_{\pm.0007}$	$0.9668_{\pm.0004}$	$0.9407_{\pm.0004}$	$0.9628_{\pm.0010}$	$0.9473_{\pm.0004}$	$0.9366_{\pm.0002}$	$0.9309_{\pm.0003}$	$0.9627_{\pm.0004}$	$0.9664_{\pm.0003}$	$0.9672_{\pm.0005}$	$0.9626_{\pm.0007}$	$0.9690_{\pm.0011}$	$0.9752_{\pm.0006}$	$0.9772_{\pm.0004}$
Drosophila																		
ES	$0.7147_{\pm.0095}$	$0.6958_{\pm.0011}$	$0.7212_{\pm.0028}$	$0.7114_{\pm.0041}$	$0.7074_{\pm.0058}$	$0.7045_{\pm.0043}$	$0.7214_{\pm.0111}$	$0.7347_{\pm.0017}$	$0.7202_{\pm.0063}$	$0.6524_{\pm.0048}$	$0.6871_{\pm.0027}$	$0.7133_{\pm.0060}$	$0.7232_{\pm.0085}$	$0.7376_{\pm.0038}$	$0.7310_{\pm.0073}$	$0.6741_{\pm.0147}$	$0.7361_{\pm.0028}$	$0.7306_{\pm.0073}$

Table A8: Full benchmark results on the GUE benchmark downstream task suite. Each entry reports the mean with the standard deviation in subscript, over 5 random seeds. best, second, third within each row.

Task	Encoder-only (masked LM)						Nucleotide Transformer			Long-context (implicit conv. / SSM)				Generative / cross-modal			**Ours**	
	DNABERT-2 (117M)	GROVER (87M)	GENA-LM Base (110M)	GENA-LM-Base T2T-Multi (110M)	ModernGENA base (135M)	ModernGENA large (377M)	NT-multi (100M)	NT-multi (2.5B)	NT-v2 (500M)	HyenaDNA (tiny-1k) (0.4M)	HyenaDNA (large-1m) (6.6M)	Caduceus PH (7.7M)	Caduceus PS (7.7M)	GENERator (1.2B)	GenomeOcean (4B)	LucaOne v2.0 (1.8B)	**DNA-RADIO B_d**(136M)	**DNA-RADIO L_d**(377M)
Yeast																		
EMP-H3	$0.6865_{\pm.0126}$	$0.6581_{\pm.0110}$	$0.6882_{\pm.0020}$	$0.7042_{\pm.0067}$	$0.6663_{\pm.0042}$	$0.6799_{\pm.0202}$	$0.6377_{\pm.0096}$	$0.7382_{\pm.0070}$	$0.6473_{\pm.0059}$	$0.6494_{\pm.0057}$	$0.6812_{\pm.0067}$	$0.7122_{\pm.0077}$	$0.7307_{\pm.0028}$	$0.7374_{\pm.0083}$	$0.7645_{\pm.0032}$	$0.7360_{\pm.0031}$	$0.7757_{\pm.0047}$	$0.7773_{\pm.0126}$
EMP-H3K14ac	$0.3897_{\pm.0051}$	$0.3497_{\pm.0031}$	$0.3458_{\pm.0010}$	$0.3553_{\pm.0062}$	$0.3676_{\pm.0026}$	$0.3655_{\pm.0095}$	$0.3874_{\pm.0055}$	$0.4873_{\pm.0037}$	$0.3694_{\pm.0070}$	$0.3259_{\pm.0045}$	$0.3341_{\pm.0069}$	$0.3565_{\pm.0039}$	$0.4111_{\pm.0036}$	$0.5109_{\pm.0055}$	$0.5291_{\pm.0027}$	$0.4360_{\pm.0069}$	$0.4887_{\pm.0055}$	$0.4985_{\pm.0049}$
EMP-H3K36me3	$0.4399_{\pm.0059}$	$0.3807_{\pm.0072}$	$0.4014_{\pm.0060}$	$0.4096_{\pm.0021}$	$0.3957_{\pm.0027}$	$0.4242_{\pm.0109}$	$0.4606_{\pm.0043}$	$0.5243_{\pm.0053}$	$0.4362_{\pm.0052}$	$0.4047_{\pm.0015}$	$0.4006_{\pm.0019}$	$0.4134_{\pm.0024}$	$0.4543_{\pm.0026}$	$0.5481_{\pm.0043}$	$0.5567_{\pm.0027}$	$0.5073_{\pm.0063}$	$0.5334_{\pm.0040}$	$0.5514_{\pm.0019}$
EMP-H3K4me1	$0.3367_{\pm.0031}$	$0.3413_{\pm.0056}$	$0.3133_{\pm.0046}$	$0.3316_{\pm.0109}$	$0.3376_{\pm.0068}$	$0.3325_{\pm.0088}$	$0.3497_{\pm.0139}$	$0.4462_{\pm.0069}$	$0.3358_{\pm.0125}$	$0.3171_{\pm.0041}$	$0.3467_{\pm.0070}$	$0.3478_{\pm.0041}$	$0.3899_{\pm.0034}$	$0.4683_{\pm.0044}$	$0.4824_{\pm.0112}$	$0.4317_{\pm.0061}$	$0.4496_{\pm.0067}$	$0.4755_{\pm.0051}$
EMP-H3K4me2	$0.2772_{\pm.0084}$	$0.2586_{\pm.0109}$	$0.2605_{\pm.0058}$	$0.2623_{\pm.0075}$	$0.2593_{\pm.0118}$	$0.2736_{\pm.0030}$	$0.2528_{\pm.0112}$	$0.3057_{\pm.0042}$	$0.2472_{\pm.0074}$	$0.2702_{\pm.0046}$	$0.2781_{\pm.0040}$	$0.2691_{\pm.0027}$	$0.2840_{\pm.0034}$	$0.3264_{\pm.0097}$	$0.3569_{\pm.0059}$	$0.3258_{\pm.0083}$	$0.3313_{\pm.0036}$	$0.3403_{\pm.0056}$
EMP-H3K4me3	$0.2429_{\pm.0019}$	$0.2494_{\pm.0060}$	$0.2200_{\pm.0041}$	$0.2200_{\pm.0074}$	$0.2302_{\pm.0066}$	$0.2273_{\pm.0128}$	$0.2672_{\pm.0051}$	$0.3302_{\pm.0094}$	$0.2307_{\pm.0104}$	$0.2226_{\pm.0038}$	$0.2244_{\pm.0054}$	$0.2652_{\pm.0020}$	$0.2653_{\pm.0069}$	$0.3846_{\pm.0091}$	$0.3833_{\pm.0027}$	$0.3071_{\pm.0056}$	$0.3590_{\pm.0044}$	$0.3641_{\pm.0042}$
EMP-H3K79me3	$0.5801_{\pm.0012}$	$0.5447_{\pm.0032}$	$0.5415_{\pm.0029}$	$0.5439_{\pm.0033}$	$0.5623_{\pm.0026}$	$0.5681_{\pm.0047}$	$0.5631_{\pm.0027}$	$0.6189_{\pm.0054}$	$0.5734_{\pm.0027}$	$0.5419_{\pm.0124}$	$0.5496_{\pm.0045}$	$0.5524_{\pm.0019}$	$0.5730_{\pm.0020}$	$0.6375_{\pm.0040}$	$0.6395_{\pm.0048}$	$0.5910_{\pm.0072}$	$0.6315_{\pm.0066}$	$0.6228_{\pm.0042}$
EMP-H3K9ac	$0.4905_{\pm.0058}$	$0.4483_{\pm.0035}$	$0.4463_{\pm.0136}$	$0.4721_{\pm.0126}$	$0.4437_{\pm.0039}$	$0.4790_{\pm.0098}$	$0.4593_{\pm.0056}$	$0.5422_{\pm.0036}$	$0.4392_{\pm.0063}$	$0.4467_{\pm.0060}$	$0.4420_{\pm.0023}$	$0.4691_{\pm.0038}$	$0.5034_{\pm.0047}$	$0.5326_{\pm.0075}$	$0.5241_{\pm.0056}$	$0.5216_{\pm.0118}$	$0.5373_{\pm.0045}$	$0.5506_{\pm.0090}$
EMP-H4	$0.7533_{\pm.0047}$	$0.7380_{\pm.0117}$	$0.7387_{\pm.0055}$	$0.7576_{\pm.0081}$	$0.7281_{\pm.0081}$	$0.7642_{\pm.0067}$	$0.7683_{\pm.0041}$	$0.8071_{\pm.0057}$	$0.7498_{\pm.0077}$	$0.7305_{\pm.0029}$	$0.7375_{\pm.0089}$	$0.7650_{\pm.0019}$	$0.7746_{\pm.0031}$	$0.7935_{\pm.0036}$	$0.8046_{\pm.0057}$	$0.7762_{\pm.0039}$	$0.8065_{\pm.0099}$	$0.8076_{\pm.0038}$
EMP-H4ac	$0.3601_{\pm.0023}$	$0.3124_{\pm.0069}$	$0.3138_{\pm.0045}$	$0.3404_{\pm.0070}$	$0.3172_{\pm.0038}$	$0.3338_{\pm.0067}$	$0.3468_{\pm.0058}$	$0.4185_{\pm.0068}$	$0.3471_{\pm.0031}$	$0.2892_{\pm.0082}$	$0.3122_{\pm.0044}$	$0.3447_{\pm.0056}$	$0.3639_{\pm.0043}$	$0.4392_{\pm.0108}$	$0.4710_{\pm.0054}$	$0.4008_{\pm.0013}$	$0.4450_{\pm.0069}$	$0.4561_{\pm.0042}$
EMP	$0.4557_{\pm.0015}$	$0.4281_{\pm.0013}$	$0.4270_{\pm.0027}$	$0.4397_{\pm.0023}$	$0.4308_{\pm.0014}$	$0.4448_{\pm.0007}$	$0.4493_{\pm.0028}$	$0.5219_{\pm.0027}$	$0.4376_{\pm.0030}$	$0.4198_{\pm.0021}$	$0.4306_{\pm.0027}$	$0.4495_{\pm.0012}$	$0.4750_{\pm.0014}$	$0.5379_{\pm.0016}$	$0.5512_{\pm.0024}$	$0.5034_{\pm.0022}$	$0.5358_{\pm.0009}$	$0.5444_{\pm.0025}$
Mouse																		
TF-M-0	$0.4121_{\pm.0208}$	$0.3502_{\pm.0148}$	$0.4586_{\pm.0082}$	$0.4559_{\pm.0101}$	$0.5755_{\pm.0215}$	$0.6433_{\pm.0071}$	$0.3966_{\pm.0213}$	$0.4438_{\pm.0187}$	$0.3650_{\pm.0233}$	$0.3324_{\pm.0065}$	$0.3087_{\pm.0049}$	$0.3856_{\pm.0096}$	$0.3861_{\pm.0161}$	$0.5749_{\pm.0107}$	$0.3707_{\pm.0117}$	$0.4774_{\pm.0139}$	$0.6324_{\pm.0078}$	$0.6173_{\pm.0202}$
TF-M-1	$0.6986_{\pm.0019}$	$0.7006_{\pm.0084}$	$0.7544_{\pm.0022}$	$0.7647_{\pm.0020}$	$0.7578_{\pm.0025}$	$0.7885_{\pm.0012}$	$0.6733_{\pm.0026}$	$0.7820_{\pm.0038}$	$0.7184_{\pm.0020}$	$0.6927_{\pm.0027}$	$0.6759_{\pm.0041}$	$0.7595_{\pm.0024}$	$0.7659_{\pm.0014}$	$0.7834_{\pm.0028}$	$0.7314_{\pm.0023}$	$0.8166_{\pm.0017}$	$0.8315_{\pm.0005}$	$0.8385_{\pm.0017}$
TF-M-2	$0.6784_{\pm.0183}$	$0.6868_{\pm.0242}$	$0.6850_{\pm.0268}$	$0.7296_{\pm.0201}$	$0.7591_{\pm.0120}$	$0.7468_{\pm.0182}$	$0.6408_{\pm.0172}$	$0.6647_{\pm.0337}$	$0.6843_{\pm.0140}$	$0.5988_{\pm.0185}$	$0.6586_{\pm.0279}$	$0.6716_{\pm.0179}$	$0.7060_{\pm.0151}$	$0.7564_{\pm.0133}$	$0.6739_{\pm.0121}$	$0.7771_{\pm.0140}$	$0.7979_{\pm.0199}$	$0.8392_{\pm.0191}$
TF-M-3	$0.4482_{\pm.0167}$	$0.4780_{\pm.0326}$	$0.4764_{\pm.0625}$	$0.5396_{\pm.0270}$	$0.6056_{\pm.0143}$	$0.5869_{\pm.0234}$	$0.3063_{\pm.0158}$	$0.4599_{\pm.0602}$	$0.3774_{\pm.0287}$	$0.3618_{\pm.0341}$	$0.4838_{\pm.0263}$	$0.3996_{\pm.0338}$	$0.3562_{\pm.0313}$	$0.5257_{\pm.0334}$	$0.3409_{\pm.0484}$	$0.6031_{\pm.0425}$	$0.6601_{\pm.0108}$	$0.6171_{\pm.0385}$
TF-M-4	$0.3585_{\pm.0109}$	$0.3305_{\pm.0062}$	$0.4059_{\pm.0074}$	$0.4369_{\pm.0083}$	$0.4115_{\pm.0137}$	$0.4745_{\pm.0103}$	$0.3052_{\pm.0045}$	$0.3885_{\pm.0178}$	$0.3241_{\pm.0061}$	$0.3088_{\pm.0055}$	$0.3755_{\pm.0073}$	$0.3984_{\pm.0087}$	$0.4164_{\pm.0102}$	$0.4339_{\pm.0076}$	$0.3418_{\pm.0078}$	$0.4666_{\pm.0079}$	$0.4959_{\pm.0077}$	$0.5277_{\pm.0124}$
TF-M	$0.5192_{\pm.0081}$	$0.5092_{\pm.0129}$	$0.5561_{\pm.0155}$	$0.5853_{\pm.0082}$	$0.6219_{\pm.0031}$	$0.6480_{\pm.0077}$	$0.4644_{\pm.0084}$	$0.5478_{\pm.0138}$	$0.4939_{\pm.0056}$	$0.4589_{\pm.0084}$	$0.5005_{\pm.0056}$	$0.5229_{\pm.0065}$	$0.5261_{\pm.0069}$	$0.6149_{\pm.0.039}$	$0.4917_{\pm.0129}$	$0.6282_{\pm.0064}$	$0.6836_{\pm.0052}$	$0.6880_{\pm.0092}$
Human																		
TF-H-0	$0.6239_{\pm.0050}$	$0.6560_{\pm.0103}$	$0.6707_{\pm.0057}$	$0.6922_{\pm.0014}$	$0.6559_{\pm.0042}$	$0.6719_{\pm.0038}$	$0.5686_{\pm.0190}$	$0.6615_{\pm.0055}$	$0.6092_{\pm.0022}$	$0.6178_{\pm.0088}$	$0.6123_{\pm.0055}$	$0.6705_{\pm.0036}$	$0.6726_{\pm.0043}$	$0.6819_{\pm.0012}$	$0.6389_{\pm.0043}$	$0.6609_{\pm.0102}$	$0.6982_{\pm.0086}$	$0.6822_{\pm.0055}$
TF-H-1	$0.6845_{\pm.0063}$	$0.6783_{\pm.0140}$	$0.6965_{\pm.0021}$	$0.7037_{\pm.0036}$	$0.6693_{\pm.0115}$	$0.6993_{\pm.0044}$	$0.6322_{\pm.0065}$	$0.6647_{\pm.0073}$	$0.6564_{\pm.0061}$	$0.6800_{\pm.0039}$	$0.6849_{\pm.0058}$	$0.7071_{\pm.0210}$	$0.6880_{\pm.0079}$	$0.6968_{\pm.0082}$	$0.6929_{\pm.0127}$	$0.7229_{\pm.0062}$	$0.7474_{\pm.0028}$	$0.7431_{\pm.0054}$
TF-H-2	$0.5433_{\pm.0082}$	$0.5214_{\pm.0005}$	$0.5528_{\pm.0066}$	$0.5388_{\pm.0140}$	$0.5561_{\pm.0075}$	$0.6144_{\pm.0103}$	$0.5000_{\pm.0092}$	$0.5328_{\pm.0114}$	$0.5181_{\pm.0067}$	$0.4963_{\pm.0185}$	$0.5320_{\pm.0252}$	$0.5234_{\pm.0264}$	$0.5328_{\pm.0056}$	$0.5693_{\pm.0030}$	$0.5870_{\pm.0028}$	$0.6098_{\pm.0051}$	$0.6203_{\pm.0140}$	$0.6267_{\pm.0169}$
TF-H-3	$0.3696_{\pm.0077}$	$0.3775_{\pm.0155}$	$0.5012_{\pm.0075}$	$0.4736_{\pm.0106}$	$0.4471_{\pm.0079}$	$0.4570_{\pm.0087}$	$0.3407_{\pm.0079}$	$0.4438_{\pm.0091}$	$0.3635_{\pm.0152}$	$0.3677_{\pm.0054}$	$0.3649_{\pm.0337}$	$0.4567_{\pm.0153}$	$0.4389_{\pm.0124}$	$0.4680_{\pm.0099}$	$0.4066_{\pm.0074}$	$0.4539_{\pm.0105}$	$0.5195_{\pm.0227}$	$0.5076_{\pm.0048}$
TF-H-4	$0.6388_{\pm.0038}$	$0.6253_{\pm.0108}$	$0.6788_{\pm.0113}$	$0.6653_{\pm.0062}$	$0.6637_{\pm.0119}$	$0.7372_{\pm.0074}$	$0.5894_{\pm.0036}$	$0.6382_{\pm.0312}$	$0.6093_{\pm.0173}$	$0.6299_{\pm.0067}$	$0.5992_{\pm.0177}$	$0.6530_{\pm.0088}$	$0.6464_{\pm.0029}$	$0.6797_{\pm.0079}$	$0.6538_{\pm.0154}$	$0.6407_{\pm.0206}$	$0.6994_{\pm.0034}$	$0.6928_{\pm.0070}$
TF-H	$0.5720_{\pm.0034}$	$0.5717_{\pm.0054}$	$0.6200_{\pm.0017}$	$0.6147_{\pm.0045}$	$0.5984_{\pm.0052}$	$0.6359_{\pm.0036}$	$0.5262_{\pm.0026}$	$0.5882_{\pm.0064}$	$0.5513_{\pm.0057}$	$0.5583_{\pm.0045}$	$0.5587_{\pm.0107}$	$0.6021_{\pm.0047}$	$0.5957_{\pm.0040}$	$0.6191_{\pm.0034}$	$0.5958_{\pm.0043}$	$0.6176_{\pm.0028}$	$0.6570_{\pm.0054}$	$0.6505_{\pm.0047}$
PD-all	$0.7859_{\pm.0020}$	$0.7766_{\pm.0026}$	$0.7993_{\pm.0037}$	$0.7957_{\pm.0011}$	$0.8051_{\pm.0032}$	$0.8240_{\pm.0024}$	$0.7594_{\pm.0033}$	$0.8218_{\pm.0010}$	$0.7750_{\pm.0028}$	$0.7772_{\pm.0029}$	$0.7821_{\pm.0015}$	$0.7947_{\pm.0011}$	$0.8156_{\pm.0036}$	$0.8454_{\pm.0011}$	$0.8005_{\pm.0023}$	$0.8336_{\pm.0014}$	$0.8504_{\pm.0021}$	$0.8569_{\pm.0033}$
PD-notata	$0.8934_{\pm.0018}$	$0.8920_{\pm.0024}$	$0.9107_{\pm.0019}$	$0.9072_{\pm.0019}$	$0.9183_{\pm.0024}$	$0.9274_{\pm.0020}$	$0.8578_{\pm.0022}$	$0.9199_{\pm.0024}$	$0.8845_{\pm.0024}$	$0.8725_{\pm.0006}$	$0.8727_{\pm.0009}$	$0.9115_{\pm.0029}$	$0.9150_{\pm.0017}$	$0.9345_{\pm.0009}$	$0.8958_{\pm.0018}$	$0.9204_{\pm.0036}$	$0.9445_{\pm.0012}$	$0.9457_{\pm.0007}$
PD-tata	$0.4446_{\pm.0205}$	$0.4828_{\pm.0152}$	$0.5035_{\pm.0160}$	$0.5155_{\pm.0097}$	$0.5454_{\pm.0158}$	$0.5190_{\pm.0088}$	$0.5406_{\pm.0179}$	$0.6072_{\pm.0362}$	$0.5375_{\pm.0250}$	$0.5005_{\pm.0223}$	$0.5820_{\pm.0158}$	$0.5147_{\pm.0232}$	$0.5770_{\pm.0062}$	$0.6813_{\pm.0142}$	$0.5467_{\pm.0144}$	$0.6223_{\pm.0185}$	$0.6596_{\pm.0187}$	$0.6625_{\pm.0128}$
PD	$0.7080_{\pm.0067}$	$0.7172_{\pm.0053}$	$0.7378_{\pm.0056}$	$0.7395_{\pm.0031}$	$0.7563_{\pm.0050}$	$0.7568_{\pm.0032}$	$0.7193_{\pm.0068}$	$0.7829_{\pm.0117}$	$0.7323_{\pm.0081}$	$0.7167_{\pm.0071}$	$0.7456_{\pm.0052}$	$0.7403_{\pm.0073}$	$0.7692_{\pm.0029}$	$0.8204_{\pm.0045}$	$0.7477_{\pm.0050}$	$0.7921_{\pm.0068}$	$0.8181_{\pm.0066}$	$0.8217_{\pm.0051}$
CPD-all	$0.5722_{\pm.0026}$	$0.5789_{\pm.0058}$	$0.5890_{\pm.0023}$	$0.5927_{\pm.0030}$	$0.6066_{\pm.0024}$	$0.6197_{\pm.0015}$	$0.5715_{\pm.0024}$	$0.6055_{\pm.0024}$	$0.5903_{\pm.0058}$	$0.5568_{\pm.0017}$	$0.5818_{\pm.0028}$	$0.5939_{\pm.0021}$	$0.6145_{\pm.0054}$	$0.6385_{\pm.0017}$	$0.5726_{\pm.0016}$	$0.6237_{\pm.0051}$	$0.6452_{\pm.0041}$	$0.6468_{\pm.0028}$
CPD-notata	$0.6102_{\pm.0022}$	$0.6297_{\pm.0036}$	$0.6401_{\pm.0022}$	$0.6404_{\pm.0031}$	$0.6526_{\pm.0026}$	$0.6705_{\pm.0014}$	$0.5982_{\pm.0014}$	$0.6404_{\pm.0056}$	$0.6275_{\pm.0030}$	$0.5990_{\pm.0033}$	$0.6226_{\pm.0017}$	$0.6450_{\pm.0009}$	$0.6610_{\pm.0032}$	$0.6786_{\pm.0053}$	$0.6262_{\pm.0032}$	$0.6718_{\pm.0050}$	$0.6981_{\pm.0020}$	$0.6916_{\pm.0040}$
CPD-tata	$0.5073_{\pm.0095}$	$0.4822_{\pm.0079}$	$0.4764_{\pm.0056}$	$0.4974_{\pm.0127}$	$0.4494_{\pm.0070}$	$0.5147_{\pm.0238}$	$0.6053_{\pm.0106}$	$0.5839_{\pm.0084}$	$0.6464_{\pm.0091}$	$0.4591_{\pm.0115}$	$0.5035_{\pm.0096}$	$0.4636_{\pm.0175}$	$0.5169_{\pm.0111}$	$0.5766_{\pm.0218}$	$0.4741_{\pm.0213}$	$0.5800_{\pm.0111}$	$0.5348_{\pm.0156}$	$0.5786_{\pm.0090}$
CPD	$0.5632_{\pm.0034}$	$0.5636_{\pm.0028}$	$0.5685_{\pm.0020}$	$0.5768_{\pm.0057}$	$0.5695_{\pm.0031}$	$0.6016_{\pm.0077}$	$0.5917_{\pm.0035}$	$0.6099_{\pm.0041}$	$0.6214_{\pm.0029}$	$0.5383_{\pm.0044}$	$0.5693_{\pm.0035}$	$0.5675_{\pm.0062}$	$0.5974_{\pm.0056}$	$0.6312_{\pm.0082}$	$0.5576_{\pm.0072}$	$0.6252_{\pm.0046}$	$0.6260_{\pm.0059}$	$0.6390_{\pm.0021}$
SSP	$0.2875_{\pm.0048}$	$0.3058_{\pm.0051}$	$0.3856_{\pm.0054}$	$0.3487_{\pm.0049}$	$0.4233_{\pm.0024}$	$0.3390_{\pm.0024}$	$0.3021_{\pm.0028}$	$0.5888_{\pm.0035}$	$0.3352_{\pm.0031}$	$0.2872_{\pm.0042}$	$0.3640_{\pm.0036}$	$0.3475_{\pm.0029}$	$0.5289_{\pm.0018}$	$0.6113_{\pm.0043}$	$0.4888_{\pm.0049}$	$0.6533_{\pm.0024}$	$0.6031_{\pm.0032}$	$0.6330_{\pm.0025}$

ing NT-multi (2.5B; 0.6373). It achieves the best performance on 4 of 7 tasks, while ranking second on zero-shot variant-effect disease prediction and third on gene finding.

E.3 Additional Benchmark

Original NT For completeness, Tab. A10 reports results on the original NT downstream benchmark (Dalla-Torre et al., 2025). We primarily use the revised NT benchmark in Tab. A6, which updates the original benchmark with higher-quality human datasets, genomic negative samples, and chromosome-held-out test sets.

E.4 Post-training on Additional DNA LMs

Our information-constrained relational post-training is not specific to the DNA-RADIO architecture and can be applied to pretrained DNA encoders without modifying their backbone. To evaluate this generality, we apply the same post-training procedure to DNABERT-2, NT-v2-500M, and ModernGENA-base in addition to distilled DNA-RADIO, using the same post-training hyperparameters across all backbones, and evaluate all models using the same BEND protocol. As shown in Tab. A11, post-training increases the average BEND performance for each evaluated backbone. The gains across independently pretrained DNA LMs, obtained without backbone-specific hyperparameter tuning, indicate that the post-training objective is not restricted to DNA-RADIO.

E.5 Full Fine-Tuning Evaluation

Our main evaluation uses linear probing to assess the quality of the learned representations without allowing downstream supervision to reshape the encoder, following standard practice in representation learning (Chen et al., 2020; Caron et al., 2021). This choice is also consistent with GENEB (Ledneva et al., 2026), which evaluates frozen genomic representations under a unified probing protocol to isolate and enable controlled comparison of embedding quality across models. Full fine-tuning addresses a different question: because the pretrained encoder itself is updated, the resulting performance reflects not only the pretrained representation but also task-specific adaptation and model capacity. Indeed, prior work has shown that fine-tuning can substantially alter pretrained features (Kumar et al., 2022), while downstream transfer performance can also depend on model scale (Hernandez et al., 2021; Dalla-Torre et al., 2025). We therefore use linear probing as our primary protocol for comparing representation quality and report full fine-tuning as a complementary evaluation of downstream adaptability.

As shown in Tab. A12, DNA-RADIO-$\mathbf{L_d}$ achieves the highest overall average with only 377M parameters, substantially fewer than GENERator (1.2B) and NT-multi (2.5B). Compared with the linear-probing results in Tab. A6, the effect of end-to-end adaptation is particularly pronounced for splice annotation, where nearly all models show large performance gains and the relative rankings change substantially. This behavior illustrates why full fine-tuning should be interpreted as a measure of downstream adaptability rather than pretrained representation quality alone, as the entire model capacity can be exploited and reshaped by task-specific supervision. Notably, DNA-RADIO-$\mathbf{L_d}$ remains highly competitive in this setting and achieves the best overall average despite being substantially smaller than the largest baselines.

F Limitations

We introduce DNA-RADIO, which uses multi-teacher distillation to outperform substantially larger, state-of-the-art teacher models in aggregate performance. Nevertheless, several limitations remain. Although DNA-RADIO outperforms each teacher overall, it does not surpass the teachers on every individual task. Such task-specific gaps are not unexpected in distillation, but they motivate further investigation into when and how complementary capabilities across teachers translate into gains for the student. DNA-RADIO also does not outperform task-specific supervised models such as Enformer on their target tasks. Similar gaps have been observed for existing state-of-the-art DNA language models and for foundation models in other domains, where models trained specifically for a dataset can outperform general-purpose models. These results highlight the distinction between broad transfer performance and performance on tasks with dedicated supervision. Finally,

Table A9: Full benchmark results on the BEND downstream task suite. Entries for fine-tuned tasks report the mean with the standard deviation in subscript, over 5 random seeds. Variant-effect expression and disease prediction are evaluated zero-shot; therefore, their entries do not include standard deviations. best, second, third within each row.

Task	Encoder-only (masked LM)						Nucleotide Transformer			Long-context (implicit conv. / SSM)				Generative / cross-modal		**Ours**
	DNABERT-2 (117M)	GROVER (87M)	GENA-LM Base (110M)	GENA-LM-Base T2T-Multi (110M)	ModernGENA base (135M)	ModernGENA large (377M)	NT-multi (100M)	NT-multi (2.5B)	NT-v2 (500M)	HyenaDNA (tiny-1k) (0.4M)	HyenaDNA (large-1m) (6.6M)	Caduceus PH (7.7M)	Caduceus PS (7.7M)	GENERator (1.2B)	GenomeOcean (4B)	**DNA-RADIO B**(136M)
Gene finding (MCC)	$0.4376_{\pm.0028}$	$0.2975_{\pm.0035}$	$0.4160_{\pm.0047}$	$0.3988_{\pm.0071}$	$0.3833_{\pm.0024}$	$0.3182_{\pm.0029}$	$0.5150_{\pm.0047}$	$0.6680_{\pm.0020}$	$0.6357_{\pm.0016}$	$0.1387_{\pm.0022}$	$0.3576_{\pm.0038}$	$0.4410_{\pm.0024}$	$0.5685_{\pm.0012}$	$0.5530_{\pm.1894}$	$0.5230_{\pm.0056}$	$0.6137_{\pm.0023}$
Enhancer annot. (AUPRC)	$0.0250_{\pm.0020}$	$0.0259_{\pm.0032}$	$0.0279_{\pm.0024}$	$0.0248_{\pm.0021}$	$0.0243_{\pm.0008}$	$0.0273_{\pm.0017}$	$0.0300_{\pm.0024}$	$0.0387_{\pm.0016}$	$0.0392_{\pm.0020}$	$0.0274_{\pm.0058}$	$0.0342_{\pm.0062}$	$0.0308_{\pm.0068}$	$0.0302_{\pm.0019}$	$0.0220_{\pm.0116}$	$0.0296_{\pm.0031}$	$0.0433_{\pm.0031}$
Chromatin access. (AUROC)	$0.7625_{\pm.0021}$	$0.7682_{\pm.0032}$	$0.7499_{\pm.0043}$	$0.7449_{\pm.0089}$	$0.7593_{\pm.0037}$	$0.7675_{\pm.0021}$	$0.7486_{\pm.0019}$	$0.7754_{\pm.0030}$	$0.7655_{\pm.0032}$	$0.7342_{\pm.0050}$	$0.7931_{\pm.0018}$	$0.7719_{\pm.0023}$	$0.7663_{\pm.0026}$	$0.6906_{\pm.0015}$	$0.7313_{\pm.0018}$	$0.7647_{\pm.0023}$
Histone mod. (AUROC)	$0.7641_{\pm.0011}$	$0.7628_{\pm.0014}$	$0.7660_{\pm.0015}$	$0.7635_{\pm.0021}$	$0.7593_{\pm.0014}$	$0.7625_{\pm.0021}$	$0.7491_{\pm.0018}$	$0.7671_{\pm.0017}$	$0.7516_{\pm.0014}$	$0.7489_{\pm.0012}$	$0.7509_{\pm.0010}$	$0.7677_{\pm.0011}$	$0.7711_{\pm.0015}$	$0.7319_{\pm.0222}$	$0.7574_{\pm.0016}$	$0.7826_{\pm.0012}$
CpG methyl. (AUROC)	$0.8814_{\pm.0036}$	$0.8755_{\pm.0038}$	$0.8807_{\pm.0047}$	$0.8769_{\pm.0028}$	$0.8633_{\pm.0070}$	$0.8721_{\pm.0080}$	$0.8736_{\pm.0035}$	$0.8978_{\pm.0046}$	$0.8780_{\pm.0026}$	$0.8546_{\pm.0032}$	$0.8675_{\pm.0044}$	$0.8911_{\pm.0041}$	$0.8927_{\pm.0029}$	$0.8533_{\pm.0034}$	$0.8641_{\pm.0086}$	$0.9303_{\pm.0011}$
Var. effect expr. (AUROC)	0.4838	0.5610	0.4673	0.4768	0.4834	0.5296	0.4958	0.5418	0.4819	0.4586	0.4961	0.4467	0.5402	0.3978	0.5108	0.5671
Var. effect disease (AUROC)	0.5092	0.4995	0.4862	0.4729	0.4921	0.4944	0.5445	0.7723	0.4757	0.5133	0.5236	0.4586	0.5450	0.6737	0.5227	0.7523

Table A10: Full benchmark results on the original NT benchmark downstream task suite. Each entry reports the mean with the standard deviation in subscript, over 5 random seeds. best, second, third within each row.

		Encoder-only (masked LM)						Nucleotide Transformer			Long-context (implicit conv. / SSM)				Generative / cross-modal			**Ours**	
		DNABERT-2 (117M)	GROVER (87M)	GENA-LM Base (110M)	GENA-LM-Base T2T-Multi (110M)	ModernGENA base (135M)	ModernGENA large (377M)	NT-multi (100M)	NT-multi (2.5B)	NT-v2 (500M)	HyenaDNA (tiny-1k) (0.4M)	HyenaDNA (large-1m) (6.6M)	Caduceus PH (7.7M)	Caduceus PS (7.7M)	GENERator (1.2B)	GenomeOcean (4B)	LucaOne v2.0 (1.8B)	**DNA-RADIO B_{cl}**(136M)	**DNA-RADIO L_{cl}**(377M))
Histone Markers	H3	$0.7120_{\pm.0068}$	$0.6947_{\pm.0080}$	$0.7136_{\pm.0037}$	$0.7399_{\pm.0073}$	$0.6863_{\pm.0074}$	$0.7041_{\pm.0104}$	$0.6522_{\pm.0064}$	$0.7313_{\pm.0059}$	$0.6599_{\pm.0040}$	$0.6888_{\pm.0043}$	$0.6698_{\pm.0057}$	$0.7374_{\pm.0057}$	$0.7375_{\pm.0029}$	$0.7360_{\pm.0069}$	$0.7697_{\pm.0045}$	$0.7476_{\pm.0072}$	$0.7765_{\pm.0044}$	$0.7813_{\pm.0127}$
	H3K14ac	$0.4145_{\pm.0040}$	$0.3603_{\pm.0089}$	$0.3450_{\pm.0104}$	$0.3534_{\pm.0096}$	$0.3557_{\pm.0054}$	$0.3645_{\pm.0096}$	$0.3946_{\pm.0088}$	$0.4687_{\pm.0094}$	$0.3703_{\pm.0099}$	$0.3506_{\pm.0034}$	$0.3282_{\pm.0039}$	$0.3387_{\pm.0031}$	$0.4088_{\pm.0031}$	$0.4971_{\pm.0055}$	$0.5288_{\pm.0028}$	$0.4546_{\pm.0072}$	$0.4847_{\pm.0064}$	$0.4994_{\pm.0020}$
	H3K36me3	$0.4588_{\pm.0065}$	$0.4147_{\pm.0073}$	$0.4012_{\pm.0038}$	$0.4166_{\pm.0033}$	$0.4115_{\pm.0082}$	$0.4407_{\pm.0027}$	$0.4810_{\pm.0103}$	$0.5418_{\pm.0056}$	$0.4453_{\pm.0064}$	$0.4302_{\pm.0043}$	$0.4301_{\pm.0040}$	$0.4241_{\pm.0025}$	$0.4600_{\pm.0031}$	$0.5594_{\pm.0041}$	$0.5864_{\pm.0068}$	$0.5065_{\pm.0044}$	$0.5518_{\pm.0047}$	$0.5634_{\pm.0045}$
	H3K4me1	$0.3531_{\pm.0108}$	$0.3236_{\pm.0082}$	$0.3284_{\pm.0040}$	$0.3161_{\pm.0047}$	$0.3208_{\pm.0058}$	$0.3222_{\pm.0134}$	$0.3654_{\pm.0084}$	$0.4450_{\pm.0081}$	$0.3435_{\pm.0040}$	$0.3218_{\pm.0078}$	$0.3459_{\pm.0017}$	$0.3556_{\pm.0026}$	$0.3748_{\pm.0053}$	$0.4914_{\pm.0042}$	$0.4828_{\pm.0033}$	$0.4543_{\pm.0026}$	$0.4877_{\pm.0031}$	$0.4928_{\pm.0027}$
	H3K4me2	$0.2994_{\pm.0047}$	$0.2693_{\pm.0071}$	$0.2732_{\pm.0114}$	$0.2665_{\pm.0062}$	$0.2749_{\pm.0095}$	$0.2808_{\pm.0056}$	$0.2465_{\pm.0053}$	$0.3095_{\pm.0043}$	$0.2598_{\pm.0052}$	$0.2503_{\pm.0037}$	$0.2791_{\pm.0079}$	$0.2732_{\pm.0035}$	$0.2914_{\pm.0049}$	$0.3271_{\pm.0057}$	$0.3344_{\pm.0059}$	$0.3131_{\pm.0051}$	$0.3185_{\pm.0095}$	$0.3292_{\pm.0068}$
	H3K4me3	$0.2446_{\pm.0092}$	$0.2346_{\pm.0113}$	$0.2103_{\pm.0060}$	$0.2284_{\pm.0027}$	$0.2261_{\pm.0078}$	$0.2316_{\pm.0110}$	$0.2633_{\pm.0041}$	$0.3193_{\pm.0059}$	$0.2241_{\pm.0061}$	$0.2177_{\pm.0042}$	$0.2220_{\pm.0025}$	$0.2479_{\pm.0041}$	$0.2805_{\pm.0160}$	$0.3659_{\pm.0140}$	$0.3852_{\pm.0045}$	$0.3081_{\pm.0010}$	$0.3345_{\pm.0053}$	$0.3549_{\pm.0034}$
	H3K79me3	$0.5556_{\pm.0034}$	$0.5354_{\pm.0031}$	$0.5559_{\pm.0062}$	$0.5362_{\pm.0055}$	$0.5431_{\pm.0054}$	$0.5596_{\pm.0035}$	$0.5454_{\pm.0020}$	$0.5975_{\pm.0041}$	$0.5564_{\pm.0058}$	$0.5304_{\pm.0040}$	$0.5531_{\pm.0049}$	$0.5514_{\pm.0022}$	$0.5619_{\pm.0041}$	$0.6097_{\pm.0066}$	$0.6108_{\pm.0190}$	$0.5484_{\pm.0040}$	$0.6014_{\pm.0023}$	$0.6092_{\pm.0030}$
	H3K9ac	$0.4774_{\pm.0024}$	$0.4371_{\pm.0072}$	$0.4454_{\pm.0088}$	$0.4565_{\pm.0048}$	$0.4359_{\pm.0048}$	$0.4444_{\pm.0132}$	$0.4679_{\pm.0020}$	$0.5133_{\pm.0019}$	$0.4453_{\pm.0021}$	$0.4280_{\pm.0040}$	$0.4251_{\pm.0012}$	$0.4598_{\pm.0050}$	$0.4657_{\pm.0062}$	$0.5203_{\pm.0054}$	$0.5304_{\pm.0026}$	$0.4868_{\pm.0035}$	$0.5231_{\pm.0014}$	$0.5247_{\pm.0038}$
	H4	$0.7532_{\pm.0020}$	$0.7363_{\pm.0069}$	$0.7364_{\pm.0115}$	$0.7551_{\pm.0039}$	$0.7364_{\pm.0094}$	$0.7299_{\pm.0083}$	$0.7508_{\pm.0033}$	$0.7978_{\pm.0034}$	$0.7094_{\pm.0219}$	$0.7361_{\pm.0027}$	$0.7390_{\pm.0026}$	$0.7751_{\pm.0024}$	$0.7532_{\pm.0037}$	$0.7990_{\pm.0080}$	$0.8117_{\pm.0019}$	$0.7575_{\pm.0028}$	$0.7756_{\pm.0029}$	$0.7998_{\pm.0037}$
	H4ac	$0.3779_{\pm.0053}$	$0.3100_{\pm.0107}$	$0.3305_{\pm.0050}$	$0.3212_{\pm.0058}$	$0.3309_{\pm.0041}$	$0.3280_{\pm.0039}$	$0.3342_{\pm.0063}$	$0.4272_{\pm.0126}$	$0.3132_{\pm.0039}$	$0.2864_{\pm.0065}$	$0.3048_{\pm.0031}$	$0.3208_{\pm.0032}$	$0.3672_{\pm.0037}$	$0.4732_{\pm.0035}$	$0.4708_{\pm.0034}$	$0.3929_{\pm.0042}$	$0.4440_{\pm.0082}$	$0.4603_{\pm.0044}$
Regulatory Annotation	Enhancer	$0.4627_{\pm.0126}$	$0.4184_{\pm.0147}$	$0.4656_{\pm.0195}$	$0.4515_{\pm.0063}$	$0.4530_{\pm.0119}$	$0.4811_{\pm.0261}$	$0.3627_{\pm.0145}$	$0.4615_{\pm.0117}$	$0.3843_{\pm.0098}$	$0.4093_{\pm.0107}$	$0.4317_{\pm.0180}$	$0.4412_{\pm.0055}$	$0.4628_{\pm.0077}$	$0.4785_{\pm.0222}$	$0.4167_{\pm.0113}$	$0.5040_{\pm.0225}$	$0.4398_{\pm.0089}$	$0.4786_{\pm.0073}$
	Enhancer type	$0.2710_{\pm.0155}$	$0.3577_{\pm.0086}$	$0.3029_{\pm.0106}$	$0.2962_{\pm.0126}$	$0.3266_{\pm.0133}$	$0.3460_{\pm.0176}$	$0.2630_{\pm.0053}$	$0.3166_{\pm.0078}$	$0.2733_{\pm.0046}$	$0.3179_{\pm.0113}$	$0.2929_{\pm.0068}$	$0.3064_{\pm.0119}$	$0.3164_{\pm.0178}$	$0.3544_{\pm.0160}$	$0.3419_{\pm.0120}$	$0.3340_{\pm.0221}$	$0.3214_{\pm.0205}$	$0.3137_{\pm.0118}$
	Promoter all	$0.8945_{\pm.0018}$	$0.8930_{\pm.0021}$	$0.9181_{\pm.0018}$	$0.9098_{\pm.0019}$	$0.9241_{\pm.0020}$	$0.9362_{\pm.0007}$	$0.8670_{\pm.0017}$	$0.9177_{\pm.0009}$	$0.8812_{\pm.0023}$	$0.8693_{\pm.0007}$	$0.8605_{\pm.0018}$	$0.9037_{\pm.0013}$	$0.9144_{\pm.0012}$	$0.9314_{\pm.0015}$	$0.9064_{\pm.0018}$	$0.9181_{\pm.0047}$	$0.9432_{\pm.0028}$	$0.9513_{\pm.0011}$
	Promoter non-TATA	$0.8968_{\pm.0013}$	$0.8947_{\pm.0019}$	$0.9174_{\pm.0012}$	$0.9121_{\pm.0027}$	$0.9260_{\pm.0011}$	$0.9403_{\pm.0013}$	$0.8687_{\pm.0006}$	$0.9190_{\pm.0016}$	$0.8801_{\pm.0036}$	$0.8719_{\pm.0012}$	$0.8634_{\pm.0011}$	$0.9045_{\pm.0010}$	$0.9152_{\pm.0018}$	$0.9276_{\pm.0022}$	$0.9053_{\pm.0009}$	$0.9188_{\pm.0011}$	$0.9440_{\pm.0007}$	$0.9508_{\pm.0010}$
	Promoter TATA	$0.8749_{\pm.0081}$	$0.8799_{\pm.0025}$	$0.9017_{\pm.0063}$	$0.9014_{\pm.0096}$	$0.9018_{\pm.0054}$	$0.9167_{\pm.0041}$	$0.8407_{\pm.0105}$	$0.8884_{\pm.0116}$	$0.8587_{\pm.0124}$	$0.8655_{\pm.0103}$	$0.8694_{\pm.0037}$	$0.8940_{\pm.0098}$	$0.9010_{\pm.0054}$	$0.8988_{\pm.0065}$	$0.8729_{\pm.0129}$	$0.8891_{\pm.0112}$	$0.9258_{\pm.0096}$	$0.9313_{\pm.0084}$
Splice Annotation	Splice acceptor	$0.7031_{\pm.0056}$	$0.4527_{\pm.0067}$	$0.5361_{\pm.0076}$	$0.5270_{\pm.0070}$	$0.5610_{\pm.0050}$	$0.5182_{\pm.0053}$	$0.4725_{\pm.0089}$	$0.8051_{\pm.0074}$	$0.4644_{\pm.0024}$	$0.3892_{\pm.0019}$	$0.5116_{\pm.0044}$	$0.5110_{\pm.0033}$	$0.7317_{\pm.0060}$	$0.7810_{\pm.0024}$	$0.7173_{\pm.0066}$	$0.6189_{\pm.0055}$	$0.8021_{\pm.0048}$	$0.8354_{\pm.0039}$
	Splice site all	$0.2691_{\pm.0052}$	$0.3031_{\pm.0048}$	$0.3590_{\pm.0021}$	$0.3371_{\pm.0052}$	$0.3952_{\pm.0116}$	$0.2935_{\pm.0023}$	$0.3411_{\pm.0018}$	$0.5920_{\pm.0067}$	$0.3697_{\pm.0041}$	$0.3215_{\pm.0046}$	$0.4716_{\pm.0055}$	$0.3652_{\pm.0030}$	$0.5251_{\pm.0074}$	$0.6211_{\pm.0083}$	$0.5021_{\pm.0041}$	$0.7264_{\pm.0035}$	$0.6362_{\pm.0023}$	$0.6652_{\pm.0053}$
	Splice donor	$0.6807_{\pm.0130}$	$0.4582_{\pm.0042}$	$0.5086_{\pm.0022}$	$0.5243_{\pm.0071}$	$0.5844_{\pm.0056}$	$0.5100_{\pm.0049}$	$0.5079_{\pm.0048}$	$0.8142_{\pm.0061}$	$0.5223_{\pm.0115}$	$0.3678_{\pm.0066}$	$0.4352_{\pm.0078}$	$0.4976_{\pm.0027}$	$0.6871_{\pm.0130}$	$0.7895_{\pm.0055}$	$0.7136_{\pm.0036}$	$0.7295_{\pm.0040}$	$0.8094_{\pm.0049}$	$0.8206_{\pm.0049}$

Table A11: **Post-training across DNA language models on BEND.** We apply the same information-constrained relational post-training procedure to three existing DNA LMs and DNA-RADIO. Metrics are MCC for gene finding, AUPRC for enhancer annotation, and AUROC for the remaining tasks; **bold** indicates the better result within each backbone before and after post-training.

Task	DNABERT2	+post-train	NT-v2 500M	+post-train	ModernGENA-base	+post-train	DNA-RADIO-B_{cl} (Ours)	+post-train (Ours)
Gene finding	0.4376	**0.4840**	**0.6357**	0.6307	0.3833	**0.5216**	0.5798	**0.6137**
Enhancer annot.	0.0250	**0.0430**	0.0392	**0.0455**	0.0243	**0.0292**	0.0372	**0.0433**
Chromatin access.	0.7625	**0.7732**	0.7655	**0.7726**	0.7593	**0.7721**	0.7568	**0.7647**
Histone mod.	0.7641	**0.7775**	0.7516	**0.7654**	0.7593	**0.7780**	0.7798	**0.7826**
CpG methyl.	0.8814	**0.9074**	0.8780	**0.9170**	0.8633	**0.9169**	0.9101	**0.9303**
Var. effect expr.	0.4838	**0.5472**	0.4819	**0.5686**	0.4834	**0.5183**	0.4994	**0.5671**
Var. effect disease	0.5092	**0.5562**	0.4757	**0.6217**	0.4921	**0.6588**	0.7368	**0.7523**

Table A12: Full fine-tuning results on the Revised-NT benchmark. We report test MCC for each task. best, second, and third within each row.

Task	HyenaDNA-L	Caduceus-PS	GENA-LM-Base	NT-multi-100M	GENERator	NT-multi	ModernGENA-B	ModernGENA-L	**DNA-RADIO-B_{cl}**	**DNA-RADIO-L_{cl}**
Model size	6.6M	7.73M	110M	100M	1.2B	2.5B	136M	377M	136M	377M
Histone Markers										
H2AFZ	0.4920	0.4450	0.4806	0.4970	0.5052	0.5083	0.5264	0.5485	0.5294	0.5368
H3K27ac	0.4590	0.4660	0.4586	0.4850	0.5066	0.4879	0.5257	0.5436	0.5094	0.5323
H3K27me3	0.5770	0.5540	0.6005	0.5890	0.6187	0.6053	0.6359	0.5851	0.6215	0.6432
H3K36me3	0.5590	0.5510	0.6227	0.5970	0.6318	0.6234	0.6385	0.6310	0.6327	0.6439
H3K4me1	0.4630	0.4440	0.4627	0.4830	0.5189	0.4986	0.5071	0.5077	0.5223	0.4936
H3K4me2	0.5510	0.4910	0.5545	0.5320	0.5775	0.5815	0.5748	0.5792	0.5646	0.6142
H3K4me3	0.6330	0.6160	0.6359	0.6310	0.6653	0.6340	0.6599	0.6397	0.6436	0.6446
H3K9ac	0.5460	0.4930	0.5480	0.5300	0.5555	0.5691	0.5541	0.5611	0.5363	0.6175
H3K9me3	0.3890	0.4290	0.4456	0.4500	0.4750	0.4615	0.4658	0.5086	0.4400	0.5103
H4K20me1	0.6310	0.5950	0.6412	0.6580	0.6678	0.6495	0.6603	0.6641	0.6836	0.6681
Regulatory										
Enhancer	0.5240	0.4880	0.5126	0.5190	0.5305	0.5891	0.5539	0.5723	0.5501	0.5798
Enhancer type	0.4610	0.4620	0.4730	0.5170	0.5259	0.5521	0.5106	0.5482	0.5127	0.5045
Promoter all	0.7230	0.7340	0.7305	0.7770	0.7809	0.7822	0.7786	0.7834	0.7854	0.7677
Promoter non-TATA	0.7310	0.7320	0.7532	0.7930	0.7767	0.7539	0.7881	0.7875	0.7814	0.7782
Promoter TATA	0.7260	0.8780	0.7859	0.9060	0.8995	0.9532	0.9071	0.9155	0.8874	0.9058
Splice Annotation										
Splice acceptor	0.7790	0.8650	0.7872	0.9490	0.9607	0.9547	0.8477	0.8563	0.9288	0.9353
Splice site all	0.8420	0.8630	0.8415	0.9660	0.9695	0.9710	0.8848	0.8941	0.9365	0.9416
Splice donor	0.7740	0.8460	0.8267	0.9600	0.9734	0.9680	0.8640	0.8831	0.9587	0.9422
Average	0.6033	0.6084	0.6200	0.6577	0.6744	0.6746	0.6453	0.6672	0.6680	0.6811

our choices of teachers and student weight initialization rely on heuristics. Systematically identifying the best combination of teacher models and student initializations would require substantial computational resources. Developing principled criteria for these choices is therefore an important direction for future work.