Generating and decoding methylated DNA with a Human Epigenetic Foundation Model

Anonymous Authors[1]

Abstract

Gene expression in humans is regulated beyond the four-letter genetic code; cytosine methylation programs cell identity and regulates expression in response to environmental cues. We present Pleiades, a series of whole-epigenome foundation models (90M/600M/7B) trained on 1.9T tokens of methylated and unmethylated human DNA, establishing a new paradigm beyond the modeling of pure DNA sequences. Pleiades achieves state-of-the-art performance compared to leading DNA foundation models on human genomic annotation tasks, such as predicting histone modifications and gene regulatory elements; notably, we find that scaling model size yields consistent gains across all tasks, with the 7B model outperforming both smaller variants and DNA-only baselines. Finally, we show that Pleiades supports a number of cell-free DNA (cfDNA) tasks, opening the door to a new era of direct clinical application of biological foundation models via cfDNA.

1. Introduction

The application of artificial intelligence and advanced language modeling to the life sciences have in recent years established a new era for the discovery of biological knowledge, promising diagnostics and therapeutics across a range of complex human diseases (Jumper et al., 2021; Ji et al., 2021; Zhou, Zhihan et al., 2023; Žiga Avsec et al., 2021; Cui et al., 2024). Many such models effectively capture the underlying statistics of DNA patterns by focusing on the human genome, (Dalla-Torre et al., 2024; Ji et al., 2021; de Lima Camillo et al., 2024; Ying et al., 2024; Nguyen et al., 2024; Brixi et al., 2025), but have thus far neglected the epigenome; the set of dynamic chemical changes to the genetic code critical for organismal development, cellular fate, and both the onset and progression of multiple diseases (Holliday & Pugh, 1975; Ferguson-Smith et al., 1991; Consortium et al., 2015; Wang et al., 2015; Flavahan et al., 2017; Farh et al., 2015).

DNA methylation is a critical class of epigenetic alterations (Vanyushin et al., 1970; Schübeler, 2015) that has enabled cancer diagnostics and revealed mechanisms of age-related disease (Wan et al., 2017; Dai et al., 2024). In recent years, these advances have been extended through cell-free DNA (cfDNA), short fragments of DNA that circulate in biofluids including plasma and cerebrospinal fluid(Wan et al., 2025; 2017); crucially, the methylation status of cfDNA reflects cellular origin, as well as disease-associated changes(Loyfer et al., 2023; Wan et al., 2025) allowing the early detection of cancers, monitoring of treatment response, and the discovery of novel biology for therapeutic application (Wan et al., 2025; Baca et al., 2023; Alexandra Bartolomucci et al., 2025).

Here, we showcase Pleiades, a series of foundation models for the human epigenome trained on a unique data corpus integrating genomic sequence with cytosine methylation information(Loyfer et al., 2023; Caggiano et al., 2021; Byrska-Bishop et al., 2022). Our main results are as follows:

- We demonstrate state-of-the-art performance on human genomic annotation tasks, outperforming leading genomic foundation models including Nucleotide Transformer (Dalla-Torre et al., 2024)) and DNABERT-2 (Zhou, Zhihan et al., 2023).
- We show that Pleiades can accurately generate cfDNA fragments and predict cellular tissue of origin from plasma cfDNA, enabling the enrichment of patient samples with implications for diagnostics across novel disease categories (Pollard et al., 2023).

2. The Pleiades Model Series

2.1. Base model design and tokenization

Model design. Pleiades is an autoregressive language model based on the generative pretrained transformer architecture (Brown, 2020). We use squared ReLU activations (So et al., 2021) in the feed-forward (MLP) blocks. Relative token or-

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

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

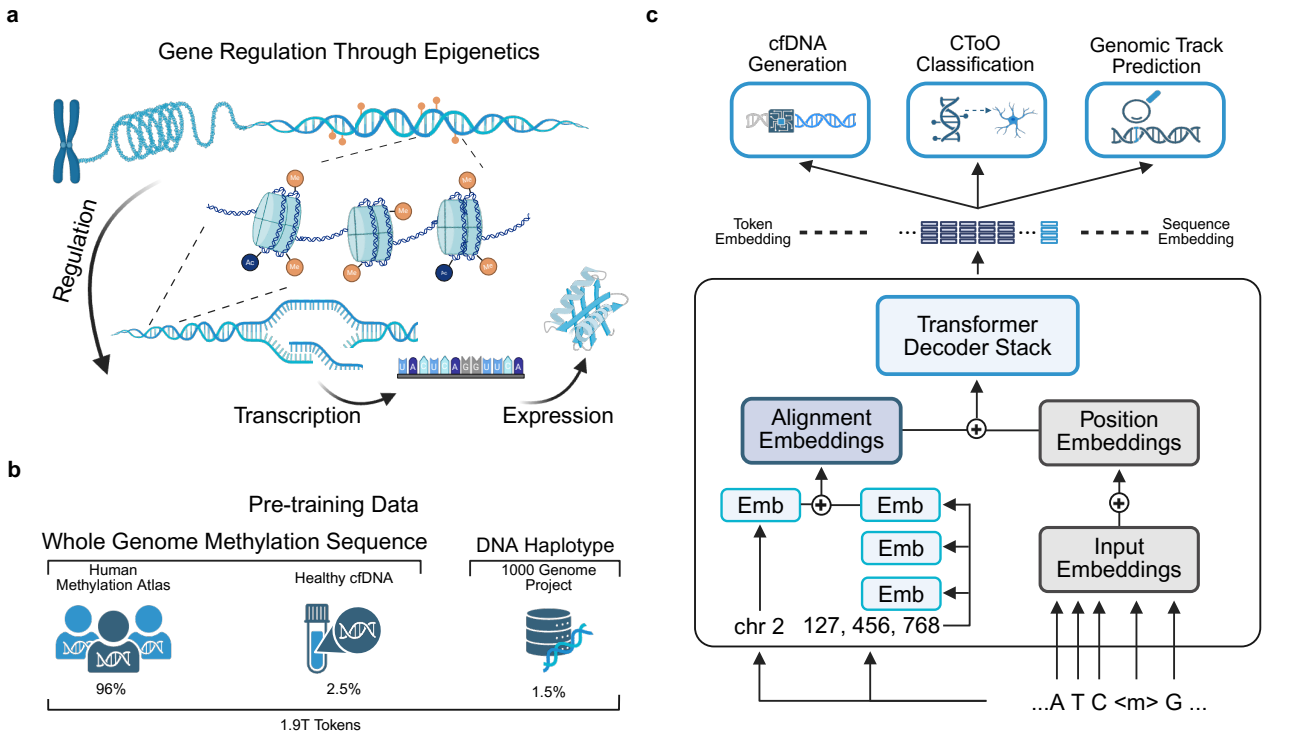

Figure 1. **Epigenomic Foundation modeling with Pleiades (a)** Epigenetic regulation. DNA and histone modifications modulate the accessibility, transcription, and downstream expression of DNA, without altering base sequence. Cell type–specific profiles encode high-resolution regulatory signals. **(b)** Pretraining. Pleiades is pretrained on 1.9T tokens, including whole-genome methylated DNA sequences from 39 different cell types, methylated cfDNA, and DNA haplotypes representative of human genome diversity. **(c)** Model architecture and tasks. Pleiades supports cfDNA generation, sequence classification such as Cell-Type-of-Origin and genomic track prediction. The model takes in token-level inputs (sequence, methylation, position, alignment), which are processed by a transformer decoder to produce per-token and sequence embeddings. Emb refers to embedding.

der is represented with rotary positional embeddings (RoPE) (Su et al., 2024). Full architecture details can be found in Appendix Section A.1.

Tokenization. We use character-level tokenization, representing each nucleotide as a single token (`A`, `C`, `T`, `G`) and methylation with a dedicated token (`<m>`). During pretraining, sequences from different sources are distinguished with special delimiter tokens: `<dna>` for nucleotide-only sequences, `<mdna>` for sequences containing methylation markers, and `<cfdna>` for cfDNA fragments. If the cell type of origin is known, a `<cell_type>` token is used followed by the name of the cell type and the closing token `</cell_type>`.

2.2. Alignment Embeddings

Accurate epigenomic modeling requires precise representation of genomic context, including long-range regulatory interactions (Chang, 2009). We introduce *Alignment Embeddings* (AEs), a novel architectural component to encode absolute genomic coordinates in Pleiades' token representations, enabling the model to distinguish biologically meaningful differences between reads without requiring prohibitively long transformer context windows.

Position encoding. For each nucleotide in a read, we encode its genomic coordinate based chromosome index c and within-chromosome position p. We encode chromosomes as $c \in \{1, \ldots, 25\}$, where chromosomes X, Y, and the mitochondrial chromosome are represented as 23, 24 and 25. We decompose p into three components capturing millions, thousands, and ones:

$$p_m = \left\lfloor \tfrac{p}{10^6} \right\rfloor, \quad p_k = \left\lfloor \tfrac{p}{10^3} \right\rfloor \bmod 10^3, \quad p_u = p \bmod 10^3. \tag{1}$$

This yields a 4-tuple (c, p_m, p_k, p_u) for every nucleotide.

Learned embeddings. Each coordinate component is embedded via its own learned lookup table, all producing vectors in $\mathbb{R}^d$. For a nucleotide with genomic coordinate tuple (c, p_m, p_k, p_u), we define the Alignment Embedding as the sum of the component embeddings:

$$e_{\mathrm{AE}}(c, p_m, p_k, p_u) = E_{\mathrm{chr}}(c) + E_m(p_m) + E_k(p_k) + E_u(p_u) \in \mathbb{R}^d, \quad (2)$$

where $E_{\mathrm{chr}} : \{1, \ldots, 25\} \to \mathbb{R}^d$, $E_m : \{0, \ldots, 249\} \to \mathbb{R}^d$, and $E_k, E_u : \{0, \ldots, 999\} \to \mathbb{R}^d$. All embedding tables are trained jointly with the model and produced for all bases in the input (Fig. 1d).

Comparison to prior work. In contrast to positional schemes such as CpGPT (de Lima Camillo et al., 2024) that provide location only at CpG sites ($\sim$ 30M loci in GRCh38), AEs encode absolute chromosome and single-base position for all nucleotides across the genome ($\sim$ 3.1B positions).

2.3. Pretraining

Pleiades is pretrained on a large corpus of high-quality sequence data from a variety of human cell types and samples. We curated a corpus of methylation and genomic data in order to maintain homogeneous sample quality composed of (see Appendix A.1 for processing details):

1. **The methylation atlas of normal human cell types**: whole-genome bisulfite sequencing (WGBS) of 39 cell-type groups obtained via fluorescence-activated cell sorting from 205 tissue samples across 137 healthy donors (Loyfer et al., 2023).
2. **Plasma-derived cfDNA**: WGBS and enzymatic methylation sequencing of plasma-derived cfDNA from 20 healthy individuals, combining in-house data with Caggiano et al. (2021).
3. **Human genome diversity**: a genome graph representative of the 1000 Genomes Project (Byrska-Bishop et al., 2022).

3. DNA Sequence Classification Benchmarks

3.1. Experimental Setup

We evaluate Pleiades on the Nucleotide Transformer (NT) genomic classification benchmarks (Dalla-Torre et al., 2024), which span promoter, enhancer, splice-site, and histone modification prediction tasks (Fig. 2a).

We identified a strong positional bias within the original dataset: negative sequences consistently started from genomic positions divisible by 1,000 (Appendix Fig. A.1). To address this, we introduced a randomized jitter in the range $[-500, 499]$ to the start positions of negative sequences, yielding what we refer to as the *Unbiased Nucleotide Transformer Benchmark*; Supplementary Section A.2).

We compare Pleiades (90M, 600M, 7B) to NT MS 2.5B (Dalla-Torre et al., 2024) and DNA-BERT2 (Zhou, Zhihan et al., 2023), fine-tuning each model for five epochs on the unbiased benchmark. To mitigate potential distribution shift effects between the predominantly methylomic pretraining data for Pleiades and this purely genomic fine-tuning dataset, the 90M and 600M models additionally underwent one epoch of fine-tuning on the DNA subset of pretraining data. Pleiades 7B did not undergo any additional fine-tuning.

When fine-tuning the Pleiades models, we append a `[CLS]` token to each input sequence and feed its final hidden state to a two-layer MLP with a ReLU nonlinearity. The first layer maps $d_{\mathrm{model}} \to d_{\mathrm{model}}$, and the second projects to the number of task labels. Full hyperparameters, compute details, and baseline settings are provided in Appendix A.2.

3.2. Results

Across tasks, Pleiades achieves the strongest performance as measured by Matthews correlation coefficient (MCC) (Fig. 2b). Pleiades 7B attains the highest MCC in $15/18$ tasks (macro-average MCC 0.98), while the 90M and 600M models outperform baselines on most tasks (macro-average MCC 0.76 and 0.77), compared with DNA-BERT2 (MCC 0.63) and NT MS 2.5B (MCC 0.67). Full results are provided in Table A.3. Gains are particularly pronounced on histone modification prediction (Fig. 2c): Pleiades 90M exceeds NT MS 2.5B despite having $27\times$ fewer parameters, consistent with known coupling between DNA methylation and histone modification (Cyrus Martin, 2005; Howard Cedar, 2009).

Finally, we assessed few-shot learning capabilities. On the H3K27ac histone modification task, Pleiades 7B reaches near-perfect MCC (0.9925) with only 152 examples. Other models achieve lower MCC even after training on the full dataset of approximately $30,000$ samples for 2 epochs (Fig. 2e).

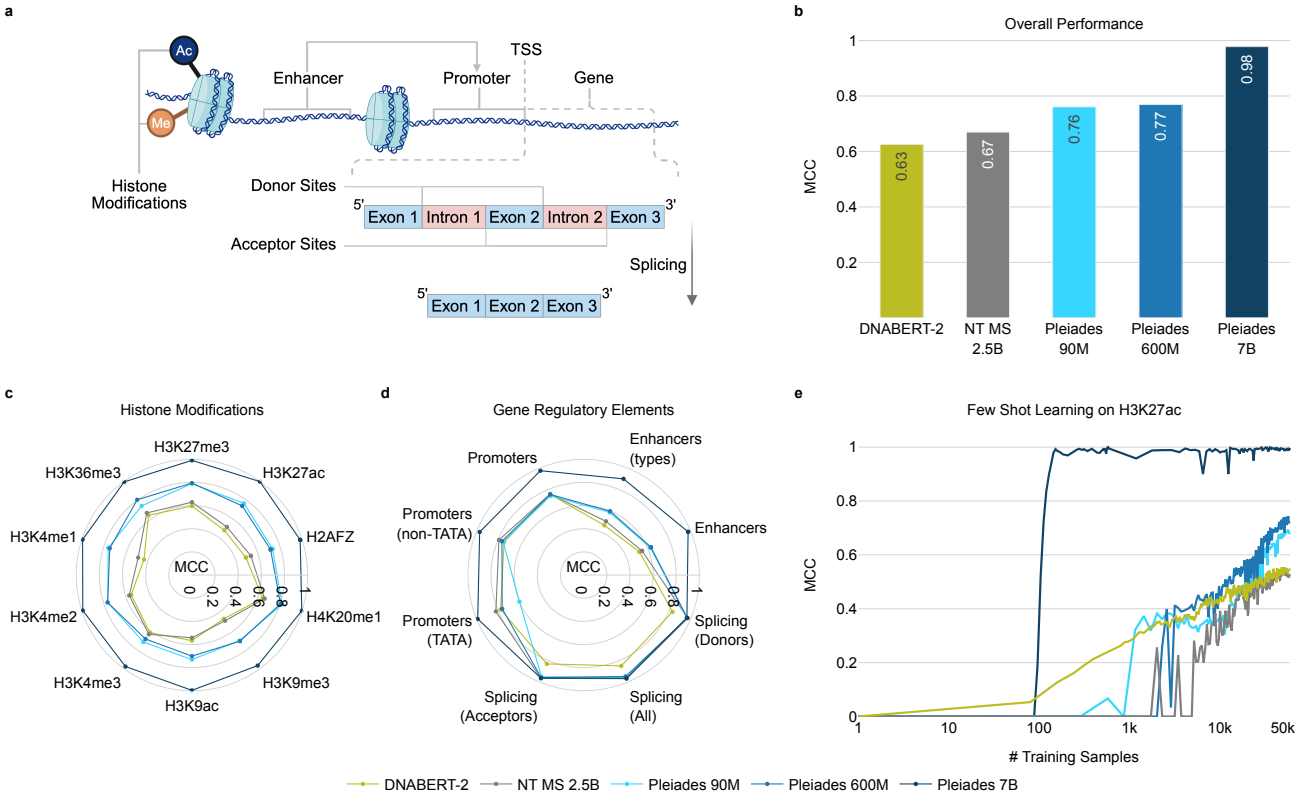

Figure 2. **Pleiades Performance Evaluation on Unbiased Nucleotide Transformer Benchmarks (a)** Schematic breakdown of Nucleotide Transformer benchmark tasks. **(b)** Overall performance of all models in terms of Matthews Correlation Coefficient (MCC). **(c)** MCC performance on histone modification tasks. **(d)** MCC performance on gene regulatory elements, including Enhancers, Promoters and Splice Sites. **(e)** Few-shot learning capability on an example NT Benchmark (H3K27ac). For Pleiades 7B, MCC of 0.9925 was achieved after training with 152 sequences.

4. Epigenomic Sequence Generation

4.1. Experimental Setup

Next, we explored the generative capabilities of Pleiades. We focused on cfDNA, and specifically assessed the feasibility of generating *in silico* biological data using held-out cfDNA samples *in silico*. We use biosamples from the pretraining test split (Caggiano et al., 2021), prepared with WGBS and sequenced to 30–50× depth. For each biosample, we designate 10% of fragments as a seed set for prompt construction and score generations against the remaining 90% fragments as ground truth. Analyses focus on 68 repeat-masked, high-coverage 1 kb regions.

Generation is conditioned on region-matched prompts: each prompt includes 5 observed seed fragments from a shared 1 kb window, followed by the cfDNA start token `<cfdna>` and a three-nucleotide prefix of a non-overlapping target fragment from the same window. We decode with top-k sampling ($k = 2$) at temperature $T = 0.7$, terminating upon emitting the cfDNA end token `</cfdna>` or reaching the remaining context budget $c - \text{length}(p)$, where $c = 1024$ is the model context length and p is the prompt.

We quantify generation quality at three resolutions: (i) *nucleotide fidelity* via per-position accuracy and relative longest common subsequence (LCS) (Fig. 3a,b); (ii) *methylome concordance* via context-specific (CpG/CHG/CHH) methylation ratios and binned methylation correlation (Fig. 3c,d); and (iii) *fragmentomics* via insert-length distributions and periodicity (Fig. 3e). For baseline comparison, we include Evo 2 7B, a DNA-only language model trained on multi-species genomic sequences (Brixi et al., 2025).

4.2. Results

At nucleotide resolution, scaling markedly improved fidelity. Pleiades 7B achieved 97% accuracy at the first nucleotide and 73% at base 150 (mean 83%); Pleiades 600M declined from 40% to 19% (mean 25%) and Evo 2 7B remained comparatively flat (mean 42%) (Fig. 3a). LCS analysis was consistent: Pleiades 7B reproduced on average 85% of each 150-nt window contiguously (median relative LCS 0.98), while Pleiades 600M and Evo 2 7B achieved only 0.07 and 0.08 on average (≈11–12 nt), respectively (Fig. 3b).

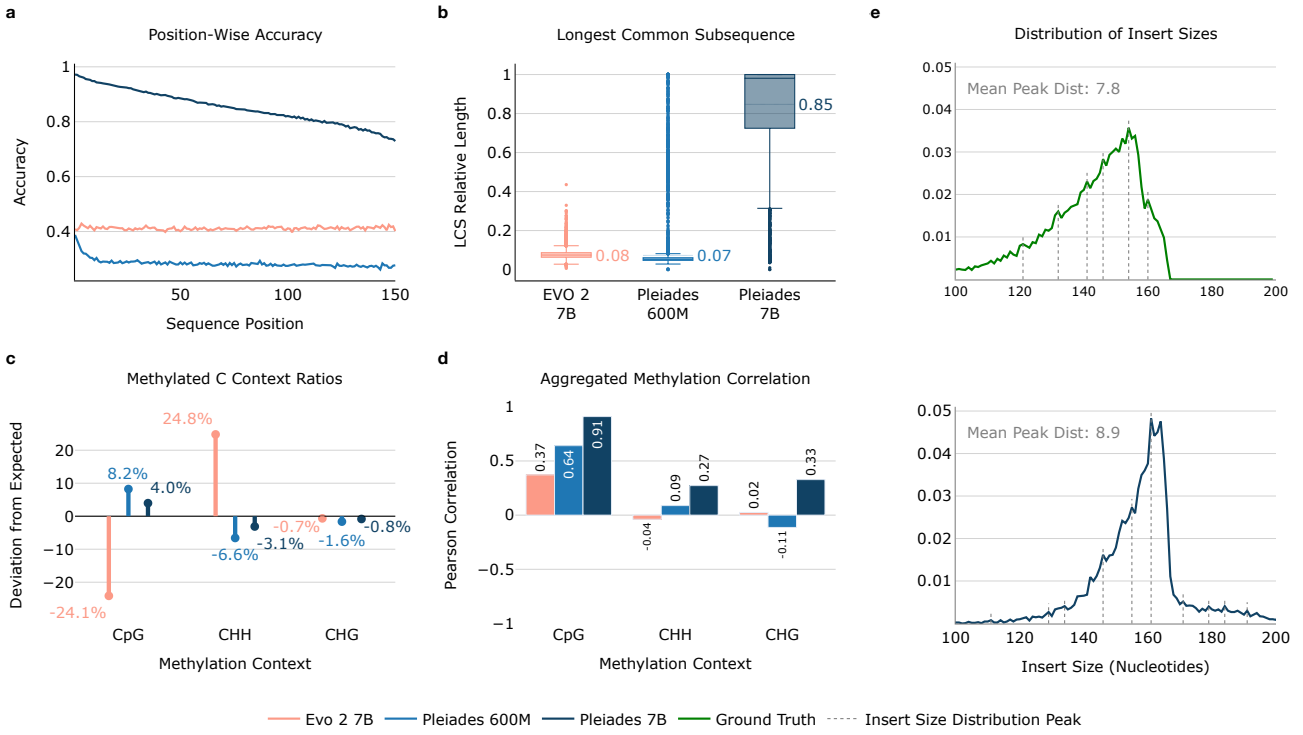

Figure 3. ***In silico*** **cfDNA Generation with Pleiades (a)** Position-wise nucleotide accuracy of the generated fragment. **(b)** Longest common *subsequence* length between generated and true fragments, expressed relative to the ground-truth length. **(c)** Cytosine-context distribution of methylated sites. **(d)** Pearson correlation of 1kb-binned methylation ratios between generated and true samples. **(e)** Insert size distribution of generated fragments for Pleiades 7B.

We next assessed the concordance of methylation across CpG, CHG, and CHH cytosine contexts[1] (Fig. 3c,d). In the ground truth set, methylation was predominantly CpG (91%), with smaller contributions from CHH (7%) and CHG (2%). Methylation–context ratio analysis shows that Evo 2 7B under-represented CpG contexts by 24.1% while over-generating CHH by 24.8%, indicating a bias towards non-canonical methylation sites. In contrast, Pleiades 600M reduced these errors to $+8.2\%$ (CpG) and -6.6% (CHH) and Pleiades 7B further to $+4.0\%$ (CpG) and -3.1% (CHH); CHG deviations remained below 1.6% for all models (Fig. 3d).

Pleiades 7B achieved the strongest correlation with ground truth methylation (Pearson $r = 0.91$ for CpG), compared with 0.64 for Pleiades 600M and 0.37 for Evo 2 7B (Fig. 3d). Non-CpG correlations were lower overall, but Pleiades 7B remained best for both CHH (0.27 vs. 0.09 and -0.04) and CHG (0.33 vs. -0.11 and 0.02). Together, these results indicate that scale and methylation-focused pretraining reduce context-distribution biases and improve methylation recall.

[1]Methylation can occur across three genomic contexts: CpG, CHG, and CHH. While the vast majority of methylation in the human genome occurs at CpG sites, non-CpG methylation is critical especially for brain biology and brain disease (Jeong et al., 2021; Shireby et al., 2022; Lister et al., 2013; Tian et al., 2023).

In silico generated fragments were plotted by insert size (Fig. 3e). The insert size distribution retained nucleosome-associated periodicity with a modest right-shift (mode 154 nt in ground truth vs. 163 nt generated; $\Delta = +9$ nt, $+5.8\%$). Rotational phasing, measured as the spacing between successive mini-peaks, was largely preserved (8.9 ± 0.3 nt *in silico* vs. 7.8 ± 0.2 nt empirically). Notably, these chromatin signatures emerged *de novo* without explicit length targets or nucleosome annotations, suggesting base-resolution pretraining can recover higher-order organization.

Overall, scaling from 600M to 7B increased mean per-nucleotide accuracy from 25% to 83%, relative LCS from 0.07 to 0.85, and CpG methylation correlation from 0.64 to 0.91, while preserving nucleosome-driven fragment length structure. Despite matching the 7B parameter budget, the DNA-only baseline underperformed across all metrics.

5. Cell Type-of-Origin (CToO)

5.1. Experimental Setup

Cell Type-of-Origin (CToO) refers to the cell type from which a cfDNA fragment originates. As plasma cfDNA is a mixture of fragments from many tissues, accurate CToO assignment is essential for tissue-specific assessment of

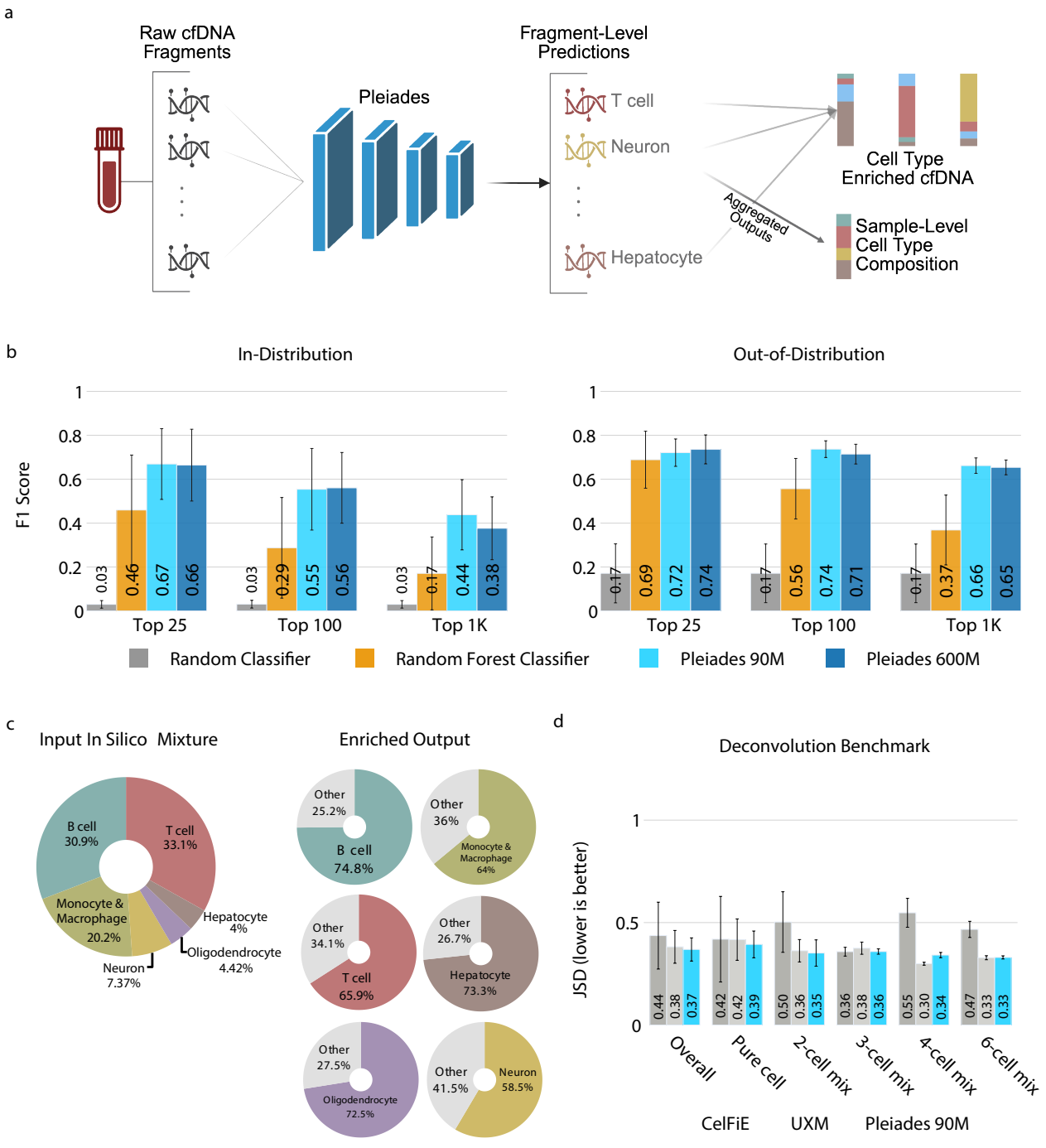

Figure 4. **Cell Type-of-Origin (CToO).** **(a)** An overview of CToO and downstream tasks. **(b)** Macro F1 scores on in-distribution and external out-of distribution (OOD) dataset. In-distribution contains 39 cell types and OOD contains 6 cell types. Bars represent means; error lines denote standard deviation. **(c)** OOD *in silico* composition before and after cell type enrichment. Enrichment was done with fine-tuned Pleiades 600M model over Top 1K cell type markers. **(d)** Deconvolution benchmark against 2 well known deconvolution tools. Bars present the mean Jensen-Shannon divergence score between true cell type ratio and estimated ratios; error bars indicate the standard deviation.

physiological state and enables early diagnosis, disease monitoring, and targeted intervention (Loyfer et al., 2023;

Liu et al., 2020). Most existing deconvolution methods infer only aggregate, sample-level proportions (Loyfer et al., 2023; Caggiano et al., 2021); here, we define the CToO task as predicting the source cell type of individual cfDNA fragments (Fig. 4a).

Training data are derived from the DNA methylation atlas (Loyfer et al., 2023), with DMRs computed on the training split to prevent information leakage. We evaluate a fragment-level CToO classifier using progressively larger panels of Differentially Methylated Regions (DMRs) (Hansen et al., 2012): (i) the published Top-25 marker regions from the tissue methylation atlas (Loyfer et al., 2023), and (ii) Top-100 and (iii) Top-1000 marker panels discovered from our training cohort (Fig. 4b). We additionally evaluate generalization on an independent out-of-distribution dataset (Do et al., 2020), restricting analysis to healthy samples and fragments overlapping the predefined marker panels.

We fine-tune Pleiades to classify fragments overlapping these marker panels. Because fragments within a marker region are typically dominated by non-target cell types, training is strongly class-imbalanced. To stabilize training and encourage separable representations, we combine standard supervised classification with a contrastive objective that clusters scarce positives and repels non-target fragments using a contrastive loss incorporating mean-based terms and hard negative mining (Robinson et al., 2020; Chen et al., 2020). Full training and dataset details are provided in Appendix A.3.

5.2. Results

Across panel sizes, Pleiades 90M and 600M achieve higher macro F1 than a tuned random forest baseline (e.g., Top-25: 0.67 vs. 0.46; Fig. 4b), with the performance gap widening as the panel expands to Top-100 and Top-1K. Stricter DMR filtering improves precision within narrowly defined regions. We further assessed generalization on an independent out-of-distribution (OOD) dataset containing 6 cell types (Do et al., 2020) (Fig. 4b).

We next explore the model's potential for cell type enrichment: selectively retaining fragments originating from a cell-of-interest. Using the OOD dataset from earlier we constructed an *in silico* admixture (Fig. 4c) and applied predictions from the Pleiades 600M model to select DNA fragments with the highest relative probabilities assigned to each target cell type.

This procedure produced large increases in target cell-type proportions within the top 1K DMR regions (Fig. 4c), including substantial enrichment of rare components: neuronal fractions increased from 7.37% to 58.5% and hepatocytes from 4.0% to 73.3%. Enrichment exhibited precision–recall tradeoff, with post-enrichment purity balanced against fragment retention (Fig. A.2).

Finally, we benchmarked Pleiades against two established deconvolution methods, UXM (Loyfer et al., 2023) and CelFiE (Caggiano et al., 2021), which infer sample-level cell-type proportions from aggregated cfDNA features.

Each trial randomly sampled 500,000 fragments from mixtures with known ground-truth ratios and was repeated 5 times with different sampling seeds. All methods require predefined marker regions, restricting the total number of fragments for evaluation. We quantified performance using Jensen–Shannon divergence (Sun et al., 2024); for full comparability, UXM and Pleiades both used the published Top-25 marker set (Loyfer et al., 2023). Pleiades achieved performance comparable to UXM and CelFiE (Fig. 4d). Pleiades performed best in $3/5$ mixture categories and tied with UXM in one, while UXM performed best for 4-cell-type mixtures (Fig. 4d).

6. Discussion

Our work demonstrates that modeling both DNA and methylation unlocks capabilities beyond DNA-only language models. Methylation represents a critical feature of the epigenome, the dynamic set of modifications that extensively influence cellular identity, function, and change throughout age and disease (Vanyushin et al., 1970; Schübeler, 2015; Dai et al., 2024). The Pleiades Series was created to capture these changes and showcase the utility of epigenetic pretraining across technical and biological applications.

Pleiades was trained on a unique corpus of 1.9T tokens comprising human DNA and methylation sequences, including a comprehensive atlas of the human methylome, spanning 39 cell-type groups (Loyfer et al., 2023).

The Pleiades series (90M, 600M, and 7B parameters) achieves state-of-the-art performance on the Nucleotide Transformer benchmark. Small Pleiades models outperform DNA-only models with many-fold higher in parameter count, and Pleiades 7B achieves MCC of 0.98. Pleiades 7B additionally demonstrated few-shot learning capabilities.

We also showcased several novel applications on cfDNA, including generative capabilities. Pleiades models generated *in silico* fragments with high accuracy across methylation context and fragment size, identified cellular origins of plasma-derived cfDNA matching state-of-the-art deconvolution methods, and enriched samples for fragments from specific cells of interest.

Future Work

Looking forward, multi-modal foundation modeling for biology offers promise to enable precision medicine and un-

lock novel insights for complex diseases. Pleiades is a first step to jointly modeling DNA and methylation in a unified, general-purpose foundation model. In future work, we believe expanding Pleiades to additional epigenomic modalities (ATAC–seq, ChIP–seq, etc.) and architectural improvements to accommodate longer context windows are promising directions towards a foundation model with deep mechanistic insight into genome regulation.

References

Alexandra Bartolomucci, M. N., Ferrier, T., Dickinson, K., Kaorey, N., Nadeau, A., Castillo, A., and Burnier, J. V. Circulating tumor dna to monitor treatment response in solid tumors and advance precision oncology. *npj Precision Oncology*, 9(1), Mar 2025. ISSN 2397-768X. doi: 10.1038/s41698-025-00876-y. URL https://doi.org/10.1038/s41698-025-00876-y.

Baca, S. C., Seo, J.-H., Davidsohn, M. P., Fortunato, B., Semaan, K., Sotudian, S., Lakshminarayanan, G., Diossy, M., Qiu, X., Zarif, T. E., Savignano, H., Canniff, J., Madueke, I., Saliby, R. M., Zhang, Z., Li, R., Jiang, Y., Taing, L., Awad, M., Chau, C. H., DeCaprio, J. A., Figg, W. D., Greten, T. F., Hata, A. N., Hodi, F. S., Hughes, M. E., Ligon, K. L., Lin, N., Ng, K., Oser, M. G., Meador, C., Parsons, H. A., Pomerantz, M. M., Rajan, A., Ritz, J., Thakuria, M., Tolaney, S. M., Wen, P. Y., Long, H., Berchuck, J. E., Szallasi, Z., Choueiri, T. K., and Freedman, M. L. Liquid biopsy epigenomic profiling for cancer subtyping. *Nature Medicine*, 29 (11):2737–2741, Nov 2023. ISSN 1078-8956. doi: 10.1038/s41591-023-02605-z. URL https://doi.org/10.1038/s41591-023-02605-z.

Brixi, Garyk, Durrant, G., M., Ku, Jerome, Poli, Michael, Brockman, Greg, Chang, Daniel, Gonzalez, A., G., King, H., S., Li, B., D., Merchant, T., A., Naghipourfar, Mohsen, Nguyen, Eric, Ricci-Tam, Chiara, Romero, W., D., Sun, Gwanggyu, Taghibakshi, Ali, Vorontsov, Anton, Yang, Brandon, Deng, Myra, Gorton, Liv, Nguyen, Nam, Wang, K., N., Adams, Etowah, Baccus, A., S., Dillmann, Steven, Ermon, Stefano, Guo, Daniel, Ilango, Rajesh, Janik, Ken, Lu, X., A., Mehta, Reshma, Mofrad, R.K., M., Ng, Y., M., Pannu, Jaspreet, Ré, Christopher, Schmok, C., J., John, St., J., Sullivan, Jeremy, Zhu, Kevin, Zynda, Greg, Balsam, Daniel, Collison, Patrick, Costa, B., A., Hernandez-Boussard, Tina, Ho, Eric, Liu, Ming-Yu, McGrath, Thomas, Powell, Kimberly, Burke, P., D., Goodarzi, Hani, Hsu, D., P., Hie, and L., B. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, 2025. doi: 10.1101/2025.02.18.638918. URL https://www.biorxiv.org/content/early/2025/02/21/2025.02.18.638918.

Brown, T. B. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell*, 185(18):3426–3440, 2022.

Caggiano, C., Celona, B., Garton, F., Mefford, J., Black, B. L., Henderson, R., Lomen-Hoerth, C., Dahl, A., and Zaitlen, N. Comprehensive cell type decomposition of circulating cell-free dna with celfie. *Nature communications*, 12(1):2717, 2021.

Chang, H. Y. Anatomic demarcation of cells: genes to patterns. *Science*, 326(5957):1206–1207, 2009.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Consortium, R. E., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K. H., Feizi, S., Karlic, R., Kim, A. R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L. H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015. doi: 10.1038/nature14248.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pp. 1–11, 2024.

Cyrus Martin, Y. Z. The diverse functions of histone lysine methylation. *Nature Reviews Molecular Cell Bi-*

ology, 6(11):838–849, Nov 2005. ISSN 1471-0072. doi: 10.1038/nrm1761. URL https://doi.org/10.1038/nrm1761.

Dai, W., Qiao, X., Fang, Y., Guo, R., Bai, P., Liu, S., Li, T., Jiang, Y., Wei, S., Na, Z., Xiao, X., and Li, D. Epigenetics-targeted drugs: current paradigms and future challenges. *Signal Transduction and Targeted Therapy*, 9(1), Nov 2024. ISSN 2059-3635. doi: 10.1038/s41392-024-02039-0. URL https://doi.org/10.1038/s41392-024-02039-0.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., Richard, G., Skwark, M., Beguir, K., Lopez, M., and Pierrot, T. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, November 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02523-z. URL http://dx.doi.org/10.1038/s41592-024-02523-z.

de Lima Camillo, L. P., Sehgal, R., Armstrong, J., Higgins-Chen, A. T., Horvath, S., and Wang, B. Cpgpt: a foundation model for dna methylation. *bioRxiv*, pp. 2024–10, 2024.

Do, C., Dumont, E. L., Salas, M., Castano, A., Mujahed, H., Maldonado, L., Singh, A., DaSilva-Arnold, S. C., Bhagat, G., Lehman, S., et al. Allele-specific dna methylation is increased in cancers and its dense mapping in normal plus neoplastic cells increases the yield of disease-associated regulatory snps. *Genome biology*, 21:1–39, 2020.

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., and Nahnsen, S. The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnology*, 38(3):276–278, 2020.

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., Jager, P. L. D., Kuchroo, V. K., Epstein, C. B., Daly, M. J., Hafler, D. A., and Bernstein, B. E. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, Feb 2015. ISSN 0028-0836. doi: 10.1038/nature13835. URL https://doi.org/10.1038/nature13835.

Ferguson-Smith, A. C., Cattanach, B. M., Barton, S. C., Beechey, C. V., and Surani, M. A. Embryological and molecular investigations of parental imprinting on mouse chromosome 7. *Nature*, 351(6328):667–670, Jun 1991. ISSN 0028-0836. doi: 10.1038/351667a0. URL https://doi.org/10.1038/351667a0.

Flavahan, W. A., Gaskell, E., and Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science*, 357(6348), Jul 2017. ISSN 0036-8075. doi: 10.1126/science.aal2380. URL https://doi.org/10.1126/science.aal2380.

Garrison, E., Sirén, J., Novak, A., et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36:875–879, 2018. doi: 10.1038/nbt.4227.

Hansen, K. D., Langmead, B., and Irizarry, R. A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10):R83, 2012. doi: 10.1186/gb-2012-13-10-r83.

Holliday, R. and Pugh, J. E. Dna modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, Jan 1975. ISSN 0036-8075. doi: 10.1126/science.187.4173.226. URL https://doi.org/10.1126/science.187.4173.226.

Howard Cedar, Y. B. Linking dna methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, 10(5):295–304, May 2009. ISSN 1471-0056. doi: 10.1038/nrg2540. URL https://doi.org/10.1038/nrg2540.

Jeong, H., Mendizabal, I., Berto, S., Chatterjee, P., Layman, T., Usui, N., Toriumi, K., Douglas, C., Singh, D., Huh, I., Preuss, T. M., Konopka, G., and Yi, S. V. Evolution of dna methylation in the human brain. *Nature Communications*, 12(1):2021, April 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21917-7. URL https://doi.org/10.1038/s41467-021-21917-7.

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

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021. ISSN 0028-0836. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Krueger, F. and Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *bioinformatics*, 27(11):1571–1572, 2011.

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., Lucero, J., Huang, Y., Dwork, A. J., Schultz, M. D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J. C., Rao, A., Esteller, M., He, C., Haghighi, F. G., Sejnowski, T. J., Behrens, M. M., and Ecker, J. R. Global epigenomic reconfiguration during mammalian brain development. *Science*, 341(6146), Aug 2013. ISSN 0036-8075. doi: 10.1126/science.1237905. URL https://doi.org/10.1126/science.1237905.

Liu, M. C., Oxnard, G. R., Klein, E. A., Swanton, C., Seiden, M. V., Liu, M. C., Oxnard, G. R., Klein, E. A., Smith, D., Richards, D., Yeatman, T. J., Cohn, A. L., Lapham, R., Clement, J., Parker, A. S., Tummala, M. K., McIntyre, K., Sekeres, M. A., Bryce, A. H., Siegel, R., Wang, X., Cosgrove, D. P., Abu-Rustum, N. R., Trent, J., Thiel, D. D., Becerra, C., Agrawal, M., Garbo, L. E., Giguere, J. K., Michels, R. M., Harris, R. P., Richey, S. L., McCarthy, T. A., Waterhouse, D. M., Couch, F. J., Wilks, S. T., Krie, A. K., Balaraman, R., Restrepo, A., Meshad, M. W., Rieger-Christ, K., Sullivan, T., Lee, C. M., Greenwald, D. R., Oh, W., Tsao, C.-K., Fleshner, N., Kennecke, H. F., Khalil, M. F., Spigel, D. R., Manhas, A. P., Ulrich, B. K., Kovoor, P. A., Stokoe, C., Courtright, J. G., Yimer, H. A., Larson, T. G., Swanton, C., Seiden, M. V., Cummings, S. R., Absalan, F., Alexander, G., Allen, B., Amini, H., Aravanis, A. M., Bagaria, S., Bazargan, L., Beausang, J. F., Berman, J., Betts, C., Blocker, A., Bredno, J., Calef, R., Cann, G., Carter, J., Chang, C., Chawla, H., Chen, X., Chien, T. C., Civello, D., Davydov, K., Demas, V., Desai, M., Dong, Z., Fayzullina, S., Fields, A. P., Filippova, D., Freese, P., Fung, E. T., Gnerre, S., Gross, S., Halks-Miller, M., Hall, M. P., Hartman, A.-R., Hou, C., Hubbell, E., Hunkapiller, N., Jagadeesh, K., Jamshidi, A., Jiang, R., Jung, B., Kim, T., Klausner, R. D., Kurtzman, K. N., Lee, M., Lin, W., Lipson, J., Liu, H., Liu, Q., Lopatin, M., Maddala, T., Maher, M. C., Melton, C., Mich, A., Nautiyal, S., Newman, J., Newman, J., Nicula, V., Nicolaou, C., Nikolic, O., Pan, W., Patel, S., Prins, S. A., Rava, R., Ronaghi, N., Sakarya, O., Satya, R. V., Schellenberger, J., Scott, E., Sehnert, A. J., Shaknovich, R., Shanmugam, A., Shashidhar, K. C., Shen, L., Shenoy, A., Shojaee, S., Singh, P., Steffen, K. K., Tang, S., Toung, J. M., Valouev, A., Venn, O., Williams, R. T., Wu, T., Xu, H. H., Yakym, C., Yang, X., Yecies, J., Yip, A. S., Youngren, J., Yue, J., Zhang, J., Zhang, L., Zhang, L. Q., Zhang, N., Curtis, C., and Berry, D. A. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of Oncology*, 31(6): 745–759, June 2020. ISSN 0923-7534, 1569-8041. doi: 10.1016/j.annonc.2020.02.011. Publisher: Elsevier.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Loyfer, N., Magenheim, J., Peretz, A., Cann, G., Bredno, J., Klochendler, A., Fox-Fisher, I., Shabi-Porat, S., Hecht, M., Pelet, T., et al. A dna methylation atlas of normal human cell types. *Nature*, 613(7943):355–364, 2023.

Loyfer, N., Rosenski, J., and Kaplan, T. wgbstools: A computational suite for dna methylation sequencing data representation, visualization, and analysis. *bioRxiv*, pp. 2024–05, 2024.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. Mixed precision training. In *International Conference on Learning Representations*, 2018.

Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D. B., Bartie, L. J., Thomas, A. W., King, S. H., Brixi, G., et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723): eado9336, 2024.

Pollard, C., Aston, K., Emery, B. R., Hill, J., and Jenkins, T. Detection of neuron-derived cfdna in blood plasma: a new diagnostic approach for neurodegenerative conditions. *Frontiers in Neurology*, 14:1272960, 2023. doi: 10.3389/fneur.2023.1272960.

Robinson, J. D., Chuang, C.-Y., Sra, S., and Jegelka, S. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2020.

Schübeler, D. Function and information content of dna methylation. *Nature*, 517(7534):321–326, Jan 2015. ISSN 0028-0836. doi: 10.1038/nature14192. URL https://doi.org/10.1038/nature14192.

Shireby, G., Dempster, E. L., Policicchio, S., Smith, R. G., Pishva, E., Chioza, B., Davies, J. P., Burrage, J., Lunnon, K., Vellame, D. S., Love, S., Thomas, A., Brookes, K., Morgan, K., Francis, P., Hannon, E., and Mill, J. Dna methylation signatures of alzheimer's disease neuropathology in the cortex are primarily driven by variation in non-neuronal cell-types. *Nature Communications*, 13(1), Sep 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-33394-7. URL https://doi.org/10.1038/s41467-022-33394-7.

So, D., Mańke, W., Liu, H., Dai, Z., Shazeer, N., and Le, Q. V. Searching for efficient transformers for language modeling. *Advances in neural information processing systems*, 34:6010–6022, 2021.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Sun, T., Yuan, J., Zhu, Y., Yang, S., Zhou, J., Ge, X., Qu, S., Li, W., Li, J. J., and Li, Y. Systematic evaluation of cell type deconvolution methods for plasma cell-free dna. *bioRxiv*, pp. 2024–03, 2024.

Tian, W., Zhou, J., Bartlett, A., Zeng, Q., Liu, H., Castanon, R. G., Kenworthy, M., Altshul, J., Valadon, C., Aldridge, A., et al. Single-cell dna methylation and 3d genome architecture in the human brain. *Science*, 382(6667): eadf5357, 2023.

Vanyushin, B. F., Tkacheva, S. G., and Belozersky, A. N. Rare bases in animal dna. *Nature*, 225(5236):948–949, Mar 1970. ISSN 0028-0836. doi: 10.1038/225948a0. URL https://doi.org/10.1038/225948a0.

Wan, J. C. M., Massie, C., Garcia-Corbacho, J., Mouliere, F., Brenton, J. D., Caldas, C., Pacey, S., Baird, R., and Rosenfeld, N. Liquid biopsies come of age: towards implementation of circulating tumour dna. *Nature Reviews Cancer*, 17(4):223–238, Apr 2017. ISSN 1474-175X. doi: 10.1038/nrc.2017.7. URL https://doi.org/10.1038/nrc.2017.7.

Wan, J. C. M., Sasieni, P., and Rosenfeld, N. Promises and pitfalls of multi-cancer early detection using liquid biopsy tests. *Nature Reviews Clinical Oncology*, Jun 2025. ISSN 1759-4774. doi: 10.1038/s41571-025-01033-x. URL https://doi.org/10.1038/s41571-025-01033-x.

Wang, A., Yue, F., Li, Y., Xie, R., Harper, T., Patel, N., Muth, K., Palmer, J., Qiu, Y., Wang, J., Lam, D., Raum, J., Stoffers, D., Ren, B., and Sander, M. Epigenetic priming of enhancers predicts developmental competence of hesc-derived endodermal lineage intermediates. *Cell Stem Cell*, 16(4):386–399, Apr 2015. ISSN 1934-5909. doi: 10.1016/j.stem.2015.02.013. URL https://doi.org/10.1016/j.stem.2015.02.013.

Ying, Kejun, Song, Jinyeop, Cui, Haotian, Zhang, Yikun, Li, Siyuan, Chen, Xingyu, Liu, Hanna, Eames, Alec, McCartney, L, D., Marioni, E., R., Poganik, R., J., Moqri, Mahdi, Wang, Bo, Gladyshev, and N., V. Methylgpt: a foundation model for the dna methylome. *bioRxiv*, 2024. doi: 10.1101/2024.10.30.621013. URL https://www.biorxiv.org/content/early/2024/11/04/2024.10.30.621013.

Zhou, Zhihan, Ji, Yanrong, Li, Weijian, Dutta, Pratik, Davuluri, Ramana V, and Liu, Han. Dnabert-2: Efficient foundation model and benchmark for multi-species genomes. In *The Twelfth International Conference on Learning Representations*, 2023.

Žiga Avsec, Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10): 1196–1203, Oct 2021. ISSN 1548-7091. doi: 10.1038/s41592-021-01252-x. URL https://doi.org/10.1038/s41592-021-01252-x.

A. Appendix

A.1. Details on the Pleiades Model Series

Table A.1. Pleiades Architecture Details

Property	Sequence Model		
	90M	**600M**	**7B**
Layers	12	32	42
Model Dimension	768	1280	4096
FFN Dimension	768	1280	4096
Attention Heads	12	20	32
Peak Learning Rate	10^{-4}	10^{-4}	10^{-5}
Warmup Steps	200	200	2000
Vocabulary Size	598	598	598
Max Context	1024	1024	1024

A.1.1. Model architecture

Pleiades uses a standard Transformer decoder architecture, scaled across three model sizes while keeping core design choices fixed. In particular, all variants share the same tokenization/vocabulary (598 tokens) and a maximum context length of 1024, with scaling achieved by increasing depth, hidden size, and the number of attention heads. Full architectural hyperparameters for the 90M, 600M, and 7B models are summarized in Table A.1.

A.1.2. Pretraining Dataset Processing

Our pretraining corpus contains both single-end and paired-end sequencing reads. We treat single-end reads and haplotype-derived reference reads as individual fragments. For paired-end data, overlapping mates are merged into a single fragment to increase effective context and reduce redundancy.

For WGBS and EM-Seq cfDNA, FASTQ files are processed with a modified MethylSeq pipeline (Ewels et al., 2020) and the Bismark aligner (Krueger & Andrews, 2011). We then apply the same preprocessing and filtering as the DNA methylation atlas (Loyfer et al., 2023).

For the 1000G diversity source, we build graphs with VG (v1.62.0) (Garrison et al., 2018) using the `construct` command and index each graph with the 1000G preset to retain haplotype-specific pathways. We then sample localized diversity using 1 Mbp sliding windows along reference coordinates, randomly sampling 40 haplotype paths per window from the retained pathways. From these sampled haplotypes we simulate sequencing reads at approximately $\sim 20\times$ coverage per haplotype, with random start positions within each window and read lengths restricted to 500–1000 nucleotides.

A.1.3. Pretraining procedure

Pretraining uses the standard autoregressive cross-entropy objective. Given a token sequence $x = (x_1, \ldots, x_T)$, we minimize the negative log-likelihood

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t}; \theta), \tag{3}$$

where $P(x_t \mid x_{<t}; \theta)$ denotes the model probability assigned to token x_t conditioned on the preceding context $x_{<t}$.

The models are optimized with AdamW (Loshchilov & Hutter, 2019) using a peak learning rate of 10^{-4} and a cosine annealing schedule. To reduce memory usage and accelerate training, we use bfloat16 mixed precision (Micikevicius et al., 2018). Full hyperparameters are reported in Table A.1. We pretrained Pleiades 7B for approximately 10 days on 256 H100 GPUs (32 nodes $\times$ 8).

Table A.2. Pleiades Hyperparameters for Unbiased Nucleotide Transformer Benchmarks

Property	90M	600M	7B	7B High LR[1]
Peak Learning Rate	10^{-4}	10^{-4}	10^{-6}	6×10^{-5}
Global Batch Size	48	288	288	96
Unfrozen Layers	*All*	*All*	*All*	Last 2 Layers

A.2. Nucleotide Transformer Benchmark

A.2.1. Unbiased Nucleotide Transformer Benchmark

The official Nucleotide Transformer (NT) benchmarks (Dalla-Torre et al., 2024) comprise 18 classification tasks. For each task, positive sequences are sampled from genomic regions with a target annotation (e.g., promoters, enhancers), while negative sequences are sampled from the remainder of the genome. However, the negative set is not fully random: negative sequences are drawn only from genomic start positions divisible by 1000. As shown in Fig. A.1a, label 0 sequences always start at loci with start position $\bmod 1000 = 0$, whereas label 1 (and label 2 where applicable) sequences are approximately uniform with respect to start position modulo 1000.

This systematic offset introduces a positional bias that inadvertently reduces the diversity of the negative set. Because Pleiades explicitly encodes genomic coordinates, it achieves near-perfect scores on the original NT benchmarks (Fig. A.1c) To remove this bias, we add a random jitter in the range $[-500, 499]$ to the start positions of negative sequences (Fig. A.1b).

A.2.2. Model fine-tuning

Both baseline models were fine-tuned using their published code and with default hyper-parameters. DNABERT-2 was fine-tuned using the fine-tune script in this official github page with learning rate 10^{-4} on V100 GPUs with the maximum batch size that would fit per device. NT 2.5B MS was fine-tuned using LORA and with a learning rate 5×10^{-4} on H200 GPUs.

Pleiades 90M and 600M were fine-tuned for an epoch on the DNA-only portion of our pretraining dataset, to bring their representations closer to pure DNA before fine-tuning for the benchmark tasks. This was not performed for Pleiades 7B. Table A.2 shows the exact hyperparameters used for fine-tuning Pleiades models on NT tasks. All Pleiades models were fine-tuned on H200 GPUs for between $\sim$4 minutes and $\sim$1 hour (depending on the task)

Fig. A.1d shows that our results after fine-tuning the baseline models DNABERT-2 and NT MS 2.5B for five epochs on our Unbiased NT Benchmarks.

A.3. Cell Type-of-Origin

A.3.1. Loss Function and Training Parameters

CToO fine-tuning is strongly class-imbalanced because, within a marker region, fragments from the target cell type are typically a minority. We therefore train with a combined objective that (i) supervises cell-type prediction and (ii) shapes the representation space with a contrastive loss that clusters scarce positives while repelling non-target fragments. For each anchor positive fragment, we sample a tractable set of negatives (re-sampled every epoch to improve generalization) and optionally augment positives. The contrastive term combines a mean-based component, which stabilizes training by comparing the anchor to an average negative, with hard negative mining to enforce fine-grained separation when cell types exhibit similar methylation patterns. In parallel, the model predicts cell type labels with a classification head operating on pooled sequence representations. The total loss is

$$\mathcal{L}_{\text{total}} = m_{\text{class}}\,\mathcal{L}_{\text{classification}} + m_{\text{contr}}\,\mathcal{L}_{\text{contrastive}}. \tag{4}$$

Contrastive loss. We define the contrastive loss as a convex combination of mean-based and hard-negative terms,

$$\mathcal{L}_{\text{contrastive}} = (1-\alpha)\mathcal{L}_{\text{mean}} + \alpha\mathcal{L}_{\text{hard}}, \tag{5}$$

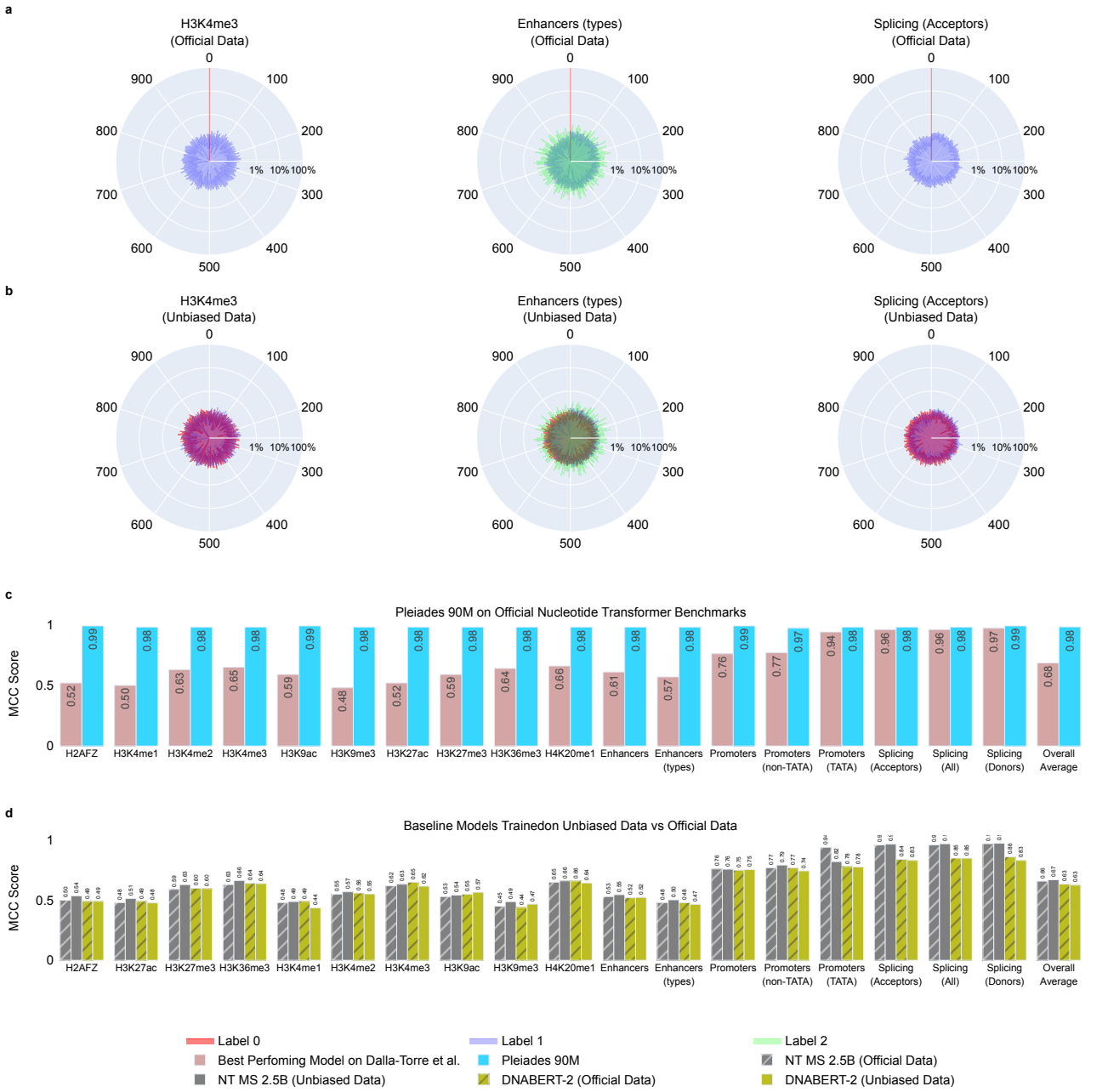

Figure A.1. **NT Benchmarks Positional Bias, Unbiased Dataset and Performance Comparisons (a)** Polar plot showing distribution of start position modulo 1000 of Official NT Benchmarks, separated by label. **(b)** Unbiased dataset where the bias is removed. **(c)** Comparison of MCC for Pleiades 90M vs Best Performing Baseline Models from Nucleotide Transformer (Dalla-Torre et al., 2024) on official NT Benchmarks Data. **(d)** MCC Comparison for Baseline Models on NT-Benchmarks Official Data vs our Unbiased Data.

where $\alpha \in [0, 1]$ controls the weight of hard negative mining. For an anchor embedding a, with positive set S_+ and negative set S_-, we compute

$$d_{\text{pos}} = \frac{1}{|S_+|} \sum_{p_i \in S_+} \delta(p_i, a) \qquad d_{\text{neg}} = \frac{1}{|S_-|} \sum_{n_j \in S_-} \delta(n_j, a)$$
$$d_{\text{hardest\_neg}} = \min_{n_j \in S_-} \delta(n_j, a) \tag{6}$$

and define margin-based losses with temperature scaling:

Table A.3. Performance (AuROC) of all models on the unbiased NT benchmark.

Task	DNABERT-2	NT MS 2.5B	Pleiades 90M	Pleiades 600M	Pleiades 7B
H2AFZ	0.4903	0.5358	0.7325	0.7147	**0.9837**
H3K27ac	0.4778	0.5141	0.7658	0.7390	**0.9988**
H3K27me3	0.5965	0.6294	0.7893	0.7948	**0.9912**
H3K36me3	0.6387	0.6636	0.7391	0.8043	**0.9953**
H3K4me1	0.4353	0.4886	0.7602	0.7466	**0.9933**
H3K4me2	0.5523	0.5695	0.7651	0.7661	**0.9915**
H3K4me3	0.6171	0.6328	0.7148	0.6827	**0.9768**
H3K9ac	0.5662	0.5423	0.7307	0.7001	**0.9960**
H3K9me3	0.4675	0.4876	0.6996	0.7063	**0.9670**
H4K20me1	0.6434	0.6608	0.8192	0.8051	**0.9982**
Enhancers	0.5231	0.5461	0.6230	0.6297	**0.9770**
Enhancers (types)	0.4658	0.5031	0.5860	0.5977	**0.9000**
Promoters	0.7546	0.7581	0.7416	0.7540	**0.9759**
Promoters (non-TATA)	0.7440	0.7921	0.7530	0.7643	**0.9725**
Promoters (TATA)	0.7767	0.8211	0.6002	0.7616	**0.9908**
Splicing (Acceptors)	0.8308	**0.9657**	0.9509	0.9593	0.9644
Splicing (All)	0.8496	**0.9693**	0.9597	0.9481	0.9659
Splicing (Donors)	0.8309	**0.9736**	0.9652	0.9694	0.9660

$$\begin{aligned} \mathcal{L}_{\text{mean}} &= \frac{\max(0,\, d_{\text{pos}} - d_{\text{neg}} + \text{margin})}{\text{temperature}}, \\ \mathcal{L}_{\text{hard}} &= \frac{\max(0,\, d_{\text{pos}} - d_{\text{hardest\_neg}} + \text{margin})}{\text{temperature}}. \end{aligned} \tag{7}$$

We use cosine distance,

$$\delta(v, w) = 1 - \frac{v}{\|v\|_2} \cdot \frac{w}{\|w\|_2}. \tag{8}$$

To improve stability and discourage representation collapse, we also experimented with adding a diversity penalty:

$$\mathcal{L}_{\text{contrastive+reg}} = \mathcal{L}_{\text{contrastive}} + \beta \sum_{i \neq j} \left(\tilde{\mathbf{E}}^{\top} \tilde{\mathbf{E}} \right)^2_{ij}, \tag{9}$$

where $\tilde{\mathbf{E}} \in \mathbb{R}^{N \times d}$ contains row-wise ℓ_2-normalized embeddings and β controls the penalty strength.

Classification loss. We supervise cell type prediction with cross-entropy:

$$\mathcal{L}_{\text{classification}} = \text{CrossEntropy}(\text{softmax}(\mathbf{y}_{\text{pred}}),\, y)\,, \tag{10}$$

where y denotes the one-hot cell type label. The classification head uses concatenated mean/max pooling of token embeddings h,

$$\begin{aligned} x &= [\,\text{mean}(h);\ \max(h)\,], \\ \mathbf{y}_{\text{pred}} &= W_2\, \sigma(W_1 x + b_1) + b_2. \end{aligned} \tag{11}$$

Training. Unless otherwise specified, we use $\alpha = 0.3$, $\text{margin} = 0.1$, and $\text{temperature} = 0.1$. Pleiades 600M was trained for ∼18h on 8 H200 GPUs for Top 25 regions, ∼11h on 64 H200 GPUs for Top 100 regions and ∼22h on 64 H200 GPUs for Top 1000 regions.

A.3.2. Cell-Type Differentially Methylated Regions

Samples from the DNA methylation atlas (Loyfer et al., 2023) were randomly split into training and test sets. We identified cell-type marker regions by calling differentially methylated regions (DMRs) in a 1-vs-all scheme for each cell type, using only the training split to prevent information leakage. DMR discovery was performed with `wgbstools` (Loyfer et al., 2024) via the `segment` and `find_markers` commands, requiring each candidate region to contain at least two CpGs, have length 50–2000 bp, satisfy an absolute delta-mean methylation difference threshold of ≥ 0.4, and achieve significance $p \leq 0.01$, using Bayesian pseudocounts of 15 for both C and T counts. From the resulting candidates, we selected the top 100 and top 1,000 markers per cell type, yielding 3,801 and 34,134 regions in total, respectively. We additionally report results on the published UXM Top-25 marker panel for reference (Loyfer et al., 2023).

Fragments intersecting a given DMR were labeled relative to that DMR's target cell type: fragments from the target cell type were assigned a positive label, and fragments from all other cell types were assigned a negative label. Because non-target fragments within a marker region exhibit broadly similar methylation patterns, we collapse non-target labels into a single generic class (e.g., `not_neuron` for neuron-targeted DMRs). To encourage separable representations within each region, we used a contrastive sampling scheme in which each positive fragment served as an anchor, paired with a fixed number of additional positive examples (3) and negative examples (36) drawn from other cell types within the same region. Negative examples were re-sampled each epoch to improve generalization.

Final classification metrics were computed over all fragments intersecting the evaluated marker panels.

A.3.3. Out-of-distribution data for evaluation

The out-of-distribution dataset (Do et al., 2020) consists of 478 FACS-processed samples from various cell types. Of the total samples, 96 failed quality control criteria: 92 had low sequencing coverage and 4 showed evidence of failed bisulfite conversion. We grouped related cell types into major groups defined by the methylation atlas (Loyfer et al., 2023) - six groups were selected. Data was processed similarly to our cfDNA data, but with an underlying aligner such as Bismark (Krueger & Andrews, 2011) due to its native support for per-base DNA methylation calling.

From the filtered cohort, we selected six representative cell types: B cell, monocyte/macrophage, T cell, liver hepatocyte, oligodendrocyte, and neuron. Using these samples, we constructed 14 equal-proportion mixtures: six pure (single-cell-type) mixtures, three two-way mixtures, three three-way mixtures, one four-way mixture, and one six-way mixture. The full mixture specification is provided in Table A.4.

For a direct comparison between fragment-level Pleiades and UXM, we used the same marker regions and corresponding reference atlas. CelFiE requires a different atlas format and marker-calling procedure and operates on coverage rather than methylation percentages; we therefore constructed a CelFiE-specific atlas to make the comparison as consistent as possible.

We fine-tuned Pleiades 90M on the published UXM Top-25 marker regions (Loyfer et al., 2023) and performed fragment-level classification. For each mixture sample, we aggregated fragment-level predictions to estimate the sample-level cell-type composition. UXM was run on the same Top-25 marker set using `--rlen 2`, corresponding to a minimum of two CpG sites per fragment.

For CelFiE, we used scripts provided by Caggiano et al. (2021) to identify DMRs for each of the 39 cell type groups, selecting the top 100 CpG sites per marker as recommended. We then estimated cell-type proportions using the authors' deconvolution script with default parameters, excluding any unknown cell types.

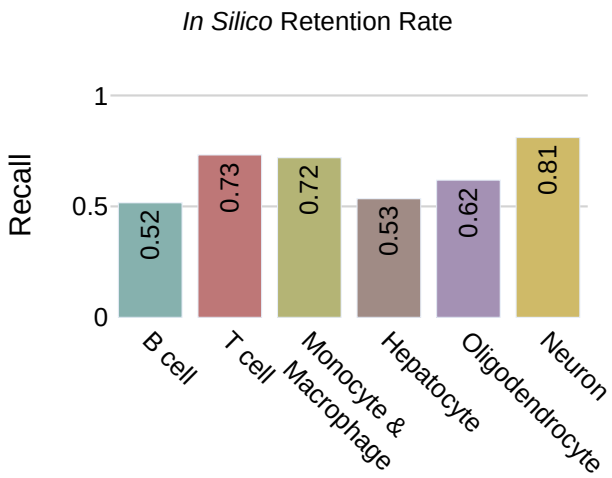

Figure A.2. Out-of-distribution CToO Recall over Top 1000 markers called from training data.

Table A.4. Cell-Type mixture proportions used for deconvolution benchmarking.

Mixture Name	B cell	T cell	Monocyte & Macrophage	Neuron	Oligodendrocyte	Hepatocyte
1-Cell Mix	1					
1-Cell Mix		1				
1-Cell Mix			1			
1-Cell Mix				1		
1-Cell Mix					1	
1-Cell Mix						1
2-Cell Mix	0.5	0.5				
2-Cell Mix			0.5	0.5		
2-Cell Mix		0.5				0.5
3-Cell Mix		0.33		0.34	0.33	
3-Cell Mix				0.34	0.33	0.33
3-Cell Mix	0.33			0.34		0.33
4-Cell Mix	0.25	0.25	0.25	0.25		
6-Cell Mix	0.17	0.17	0.17	0.16	0.16	0.17