METHYLATION-AWARE EMBEDDING GEOMETRY EMERGES FROM BISULFITE PRETRAINING IN DNA LANGUAGE MODELS

Jiajie Xiao*, **Salwan Butrus***, **Nathan Hunkapiller**
Curve Biosciences
{jiajie,salwan,nathan}@curvebio.com
*Equal contribution

ABSTRACT

DNA methylation encodes regulatory information beyond the DNA sequence, but most genomic language models (gLMs) miss this important modality because they are pretrained on native DNA only. We test whether a widely used DNA checkpoint can be *retrofitted* into a methylation-aware model by continual pretraining on bisulfite sequencing (BS-seq) reads, where methylation is implicitly encoded into token identities via C→T conversion. Rather than proposing a new architecture, we ask for compact, interpretable evidence that methylation is encoded in representation space. Using DNABERT2 continually pretrained on a multi-tissue BS-seq atlas, we show two simple geometric diagnostics: (i) per-read embedding norms become bimodal and align with hypo/hypermethylated contexts, and (ii) cosine distances between genomically matched tumor–normal read pairs increase substantially after BS-seq adaptation, relative to the native checkpoint. These results suggest that simple BS-seq retrofitting can endow a standard DNA gLM with biologically meaningful, increased, label-light epigenetic sensitivity.

1 INTRODUCTION

Language models are increasingly used to model biology and medicine Hayes et al. (2025); Tu et al. (2024), including genomic language models (gLMs) that treat DNA as a sequence of tokens. gLMs pretrained on native DNA have improved prediction of diverse sequence-linked phenomena, but biological state is not written in DNA alone.

The epigenome is a collection of chemical modifications to DNA and associated proteins that provides dynamic regulation of cellular function Dupont et al. (2009). Among these, DNA methylation of cytosines that are followed by guanine (CpG) is one of the most studied and biologically consequential mechanisms, with central roles in development and complex diseases Smith & Meissner (2013); Kulis & Esteller (2010). While specialized models have begun to look at methylation data de Lima Camillo et al. (2024); Jeong et al. (2025); Li et al. (2024); Niki et al. (2025), they offer limited mechanistic insight into how methylation signals are reflected in gLM representation geometry. Among these, CpGPT de Lima Camillo et al. (2024) additionally explores fusing DNA embeddings adapted from pretrained gLMs with methylation embeddings at $\sim$100k[1] CpGs, motivating complementary methods to retrofit widely used DNA-only checkpoints into methylation-aware gLMs without changing architectures or downstream tooling.

To address these gaps, we focus on DNA methylation captured by bisulfite sequencing (BS-seq), which "writes" epigenetic state into the sequence itself. In BS-seq data, unmethylated cytosines in CpG context are chemically converted and sequenced as thymines (henceforth C→T), while methylated cytosines remain unchanged; thus methylation status is encoded directly at the level of input tokens. Rather than proposing a new architecture or training objective, we ask three linked questions about *feasibility*, *interpretability*, and *diagnosis*: can continual pretraining on BS-seq reads confer methylation sensitivity to a standard gLM; what compact, measurable signatures in representation

[1]This accounts for only roughly 0.3% of all CpGs in the human genome.

space reflect that sensitivity; and can such signatures be computed cheaply and remain informative under distribution shift (e.g., tumor vs. normal) to decide whether retrofitting suffices or a more costly approach is needed?

Retrofitting is attractive only if we can verify methylation sensitivity without immediately resorting to task-specific supervised benchmarks or protocol-specific labels, which introduce additional design choices and can entangle methylation sensitivity with downstream objectives and evaluation details. We therefore focus on compact representation-level diagnostics that (i) can be computed from embeddings alone (or with minimal methylation annotations for interpretation), (ii) reflect epigenetic differences in a way that is consistent with epigenetic variation (rather than only global token-composition shifts), and (iii) remain informative under distribution shift (e.g., tumor vs. normal). These readouts are cheap to compute at scale and can be applied *before* committing to downstream task benchmarking or fine-tuning: if adaptation fails to move them in the expected direction, downstream task-specific evaluation may be premature; if adaptation succeeds, the same diagnostics help characterize where the model is sensitive (or brittle) under shift.

Contributions.

- We formalize gLMs as estimators of context-conditional probability distributions and present a phenomenological model that yields testable predictions for how BS-seq pretraining reshapes representation geometry under distribution shift.
- We propose two simple, label-light geometric diagnostics of methylation sensitivity after BS-seq adaptation: per-read embedding norms become bimodal and align with hypo/hypermethylated contexts, and cosine distances between genomically matched read pairs amplify.
- We validate these signatures on in-distribution (Hepatocytes vs. Granulocytes) and out-of-distribution (DLBCL vs. B-cells) read pairs, where BS-seq adaptation increases geometric separation between epigenetically distinct sequences relative to the native checkpoint.

2 METHODS: PREDICTIONS AND STUDY DESIGN

Theoretical perspective. We treat gLMs as estimators of context-conditional token distributions within their pretraining data distribution (see Theorem 1). Continual pretraining on BS-seq data induces a distribution shift from native DNA to bisulfite-converted reads, wherein epigenetic state directly modulates token identities through C→T conversion.

Geometric predictions. This perspective yields two testable predictions for embedding geometry post-adaptation (see more details and derivation in Appendix A.1):

- **Norm bifurcation.** After BS-seq adaptation, per-read pooled embedding norms become bimodal. Reads with more bisulfite conversion (hypomethylated CpG contexts, hence more C→T) shift toward a higher-norm mode, while reads that remain closer to native DNA (hypermethylated) stay near the original mode or slightly lower (see Theorem 2).
- **Cosine-distance amplification.** For reads at the *same genomic locus* across conditions, cosine distances between pooled embeddings should increase after BS-seq adaptation; the right tail strengthens when methylation divergence or distributional shift is greater (e.g., tumor vs. normal) (see Theorem 3).

Representative instantiation (model and data). We instantiate this study with DNABERT2-117M Zhou et al. (2023) and a large multi-tissue WGBS atlas spanning 39 tissues Loyfer et al. (2023), focusing on differentially methylated regions (DMRs) to enrich for biologically meaningful epigenetic variation. DNABERT2 and this atlas serve as representative examples: DNABERT2 is a widely used open DNA foundation model, and the atlas provides a diverse, whole-genome methylation reference of various healthy cell types. Our diagnostics (embedding norms and cosine distances) and predictions (geometric shift) are intended to be broadly applicable across gLM backbones that produce pooled sequence embeddings. More detailed discussion can be found in Appendix A.1.

Diagnostics and evaluation design. Starting from the native-DNA checkpoint, we continually pretrain on BS-seq reads (see Appendix A.2 for a detailed pretraining setup). For each BS-seq read we extract a pooled embedding and compute its ℓ_2 norm. For read pairs that overlap the *same genomic locus* across two conditions, we compute cosine distance. This matched-locus design

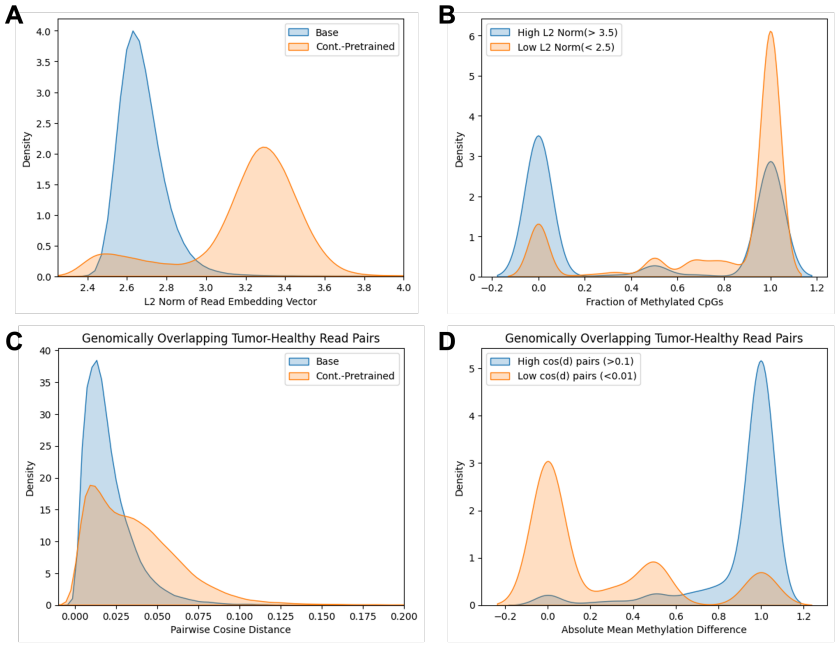

Figure 1: **Embedding geometry encodes methylation after BS-seq continual pretraining.** (A) Embedding-norm distributions shift from unimodal (base DNABERT2) to bimodal (BS-seq adapted). (B) High-norm reads are enriched for hypomethylated CpG contexts, while low-norm reads align with hypermethylated contexts. (C) For genomically overlapping tumor–normal read pairs, cosine distances increase after BS-seq pretraining, indicating greater epigenetic sensitivity. (D) Read pairs with larger cosine distances exhibit larger absolute methylation differences, linking angular separation to biochemical divergence.

controls for sequence differences and isolates epigenetic effects written into BS-seq tokens. We evaluate (a) *in-distribution* hepatocyte–granulocyte comparisons (both present in pretraining) and (b) an *out-of-distribution* DLBCL vs. normal B-cell comparison. See Appendix Figure S1 for study pipeline overview, Appendix A.2 for training details, and Appendix A.3 for embedding extraction methods.

3 Results

BS-seq adaptation induces methylation-aligned norm bimodality. After continual pretraining, the per-read embedding norm distribution becomes distinctly bimodal (Figure 1A). The mean per-read embedding norm increased by 0.51 ± 0.001 (Cramér–von Mises $T = 1056698$, $p = 3.5 \times 10^{-6}$, Cohen's $d = 1.59$), and a two-component fit exhibits clear separation (Ashman's $D = 3.07$) with modes centered at 2.58 ± 0.15 and 3.30 ± 0.18, consistent with a split between hyper and hypomethylated read populations.

Angular separation increases for genomically matched tumor–normal pairs. For genomically overlapping DLBCL–B-cell read pairs, cosine distances increase after BS-seq adaptation (Figure 1C), with a mean shift $\Delta d_{\cos} = 0.0160$ (95% CI [0.0157, 0.0163]; paired $t = 98.9$, $p < 10^{-16}$; Cohen's $d = 0.60$) and 76.3% of pairs exhibiting higher distance post-adaptation. In addition, right-tail enrichment for $d_{\cos} > 0.05$ rose from 5.0% [4.8–5.3%] to 26.9% [26.4–27.4%] ($\Delta = 21.9\%$, 95% CI [21.3–22.5%]; $z = 69.6$, $p < 10^{-16}$), suggesting that the adapted model amplifies separation in epigenetically divergent subpopulations.

Geometry quantitatively tracks methylation state and divergence. Reads with large embedding norms (> 3.5) were strongly enriched for hypomethylated CpGs, whereas those with small norms (< 2.5) clustered near fully methylated contexts (Cramér–von Mises $T = 1012.4$, $p =$

2.2×10^{-7}; $\Delta\mu = -0.33$ (95% CI $[-0.34, -0.33]$); Cohen's $d = -0.78$) (Figure 1B). Likewise, read pairs exhibiting large cosine distances ($d_{\cos} > 0.1$) showed substantially higher absolute mean methylation differences than low-distance pairs ($d_{\cos} < 0.01$) ($\Delta = 0.66$ (95% CI [0.63, 0.68]; Welch $t = 54.6$, $p < 10^{-16}$; Cohen's $d = 2.20$) (Figure 1D; see also Appendix Figure S3 for joint distributions). These associations demonstrate that both embedding magnitude and angular separation quantitatively track methylation state and inter-read divergence. Together, these results indicate that simple geometric statistics provide label-light readouts of methylation sensitivity.

Across in-distribution hepatocyte–granulocyte comparisons (both present in pretraining), we observe smaller but consistent geometry–methylation correspondence (Appendix Figure S2 and Appendix A.4), supporting that the diagnostics reflect methylation variation even when overall epigenetic programs are closer to the pretraining distribution.

4 Discussion and conclusion

Our framework reveals several insights about gLMs as epigenetic learners. Continual pretraining on BS-seq data induces a geometric reorganization of the embedding space that reflects underlying epigenetic states. The emergence of bimodal norm distributions and the amplification of cosine distances provide measurable signatures of methylation-aware representation learning, with distributional patterns (unimodal vs. bimodal) reflecting heterogeneity in epigenetic differences between tissue types.

Importantly, these signatures are lightweight and easy to compute: they can be computed without task-specific labels and can serve as representation-level diagnostics before committing to downstream benchmarks or more costly specialized training. The stronger right-tail separation observed in out-of-distribution DLBCL comparisons (Figure 1) is consistent with our theory that cosine geometry reflects both epigenetic divergence and distributional novelty, and suggests that the learned methylation-aware geometry remains useful as a diagnostic under this distribution shift.

Limitations and future directions. We use DNABERT2 and a focused set of WGBS-derived benchmarks as representative examples to test our theoretical predictions instead of an exhaustive survey of various architectures and biological contexts. Empirically, we primarily evaluate two normal tissue types (in-distribution) and DLBCL versus B-cell comparisons (out-of-distribution); broader tissue diversity, disease states, and assay types are left for future work. Most importantly, we emphasize understanding and diagnosing representation geometry rather than optimizing a particular downstream task: supervised benchmarks (e.g., read-level methylation calling or sample classification) are highly task- and protocol-dependent, and systematic evaluation is deferred to follow-up studies. We also do not include a head-to-head comparison against specialized methylation models; we view retrofitting as complementary when one wants a drop-in methylation-aware gLM starting from widely used DNA-only checkpoints. This setting also allows us to test the hypothesis that retrofitting preserves sequence interactions learned during DNA pretraining that are additive to methylation signals learned from BS-seq, relative to specialized methylation models trained from scratch.

Beyond this specific proof-of-concept, our framework is intended to be broadly applicable across architectures and scales. Applying the same representation-space diagnostics to diverse model classes (e.g., HyenaDNA Nguyen et al. (2023), Evo Nguyen et al. (2024)) and larger scales can help determine whether the observed signatures generalize beyond DNABERT2. Task-driven evaluations via supervised fine-tuning can clarify what task-specific signatures are needed beyond generic methylation-aware genomic embeddings learned through self-supervision. Finally, studying data scaling and sampling strategies for epigenetic pretraining may inform more efficient resource allocation relative to specialized alternatives.

We have presented a theoretical and empirical framework showing that continual pretraining on bisulfite sequencing data can induce clear, measurable geometric signatures in gLM embeddings that track methylation variation in our evaluated settings. Our phenomenological analysis provides a compact set of representation-level diagnostics for assessing BS-seq adaptation before investing in heavier downstream evaluations.

MEANINGFULNESS STATEMENT

A meaningful biological representation should align geometric structure in embedding space with real biochemical state and enable scientific interpretation beyond predictive accuracy. Our work demonstrates that continual pretraining on BS-seq data organizes DNA language model representations such that simple geometric quantities track methylation state and divergence, including tumor–healthy distribution shift. This interpretability-first design provides compact diagnostics to assess epigenetic learning before committing expensive computational resources to downstream tasks or specialized architectures. The streamlined multimodal approach—retrofitting existing DNA checkpoints via bisulfite tokens—establishes reusable diagnostic tools that reveal whether models capture regulatory programs beyond sequence alone, guiding future extensions and architecture choices.

AUTHOR CONTRIBUTIONS

J.X. and S.B. conceived and designed the study. J.X. and S.B. implemented model pretraining, finetuning, and analyses. S.B. and J.X. curated the data. J.X., S.B., and N.H. interpreted the results. All authors discussed the results, contributed to manuscript preparation and revision, and approved the final version.

ACKNOWLEDGMENTS

We thank Mahdi Baghbanzadeh for helpful discussions related to model selection and pretraining strategy during his internship at Curve Biosciences. We thank Dr. Ritish Patnaik and Prof. Shan X. Wang for helpful discussions throughout the project. This work was supported in great part by resources and services provided by Google Cloud Platform and Amazon Web Services.

REFERENCES

Lucas Paulo de Lima Camillo, Raghav Sehgal, Jenel Armstrong, Henry E Miller, Jessica A Lasky-Su, Albert T Higgins-Chen, Steve Horvath, and Bo Wang. Cpgpt: a foundation model for dna methylation. *bioRxiv*, pp. 2024–10, 2024.

Cathérine Dupont, D Randall Armant, and Carol A Brenner. Epigenetics: definition, mechanisms and clinical perspective. In *Seminars in reproductive medicine*, volume 27, pp. 351–357. © Thieme Medical Publishers, 2009.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Yunhee Jeong, Clarissa Gerhäuser, Guido Sauter, Thorsten Schlomm, Karl Rohr, and Pavlo Lutsik. Methylbert enables read-level dna methylation pattern identification and tumour deconvolution using a transformer-based model. *Nature Communications*, 16(1):788, 2025.

Marta Kulis and Manel Esteller. Dna methylation and cancer. *Advances in genetics*, 70:27–56, 2010.

Mingyang Li, Ruichu Gu, Shiyu Fan, Yu Fan, Bo He, Jinmin Yang, Yuting Chen, Mengling Xin, Han Wen, and Chengqi Yi. Methylqueen: A methylation encoded dna foundation model. *bioRxiv*, pp. 2024–12, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Netanel Loyfer, Judith Magenheim, Ayelet Peretz, Gordon Cann, Joerg Bredno, Agnes Klochendler, Ilana Fox-Fisher, Sapir Shabi-Porat, Merav Hecht, Tsuria Pelet, et al. A dna methylation atlas of normal human cell types. *Nature*, 613(7943):355–364, 2023.

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. *Advances in neural information processing systems*, 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.

Pouya Niki, Christoforos Nalmpantis, Javkhlan-Ochir Ganbat, Donal Byrne, Husam Babikir, Anjeet Jhutty, Will Rowe, Timing Liu, Netanel Loyfer, Sofia Toniolo, et al. Human whole epigenome modelling for clinical applications with pleiades. *bioRxiv*, pp. 2025–07, 2025.

Zachary D Smith and Alexander Meissner. Dna methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3):204–220, 2013.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

A Appendix

A.1 Theoretical framework and derivations

A.1.1 Foundational Definitions

Definition 1 (Sequence, Epigenetic, and Observation Spaces). *Let $\mathcal{X} = \{A, T, C, G\}^*$ denote the set of finite DNA strings over the canonical alphabet. Let $\mathcal{E}$ denote a measurable space of epigenetic states (e.g., CpG methylation patterns). A pair $(X, E) \in \mathcal{X} \times \mathcal{E}$ is drawn from some joint distribution P_{gen}.*

Bisulfite sequencing (BS-seq) defines an observation map

$$\pi : \mathcal{X} \times \mathcal{E} \to \mathcal{X}_{\text{BS}}, \qquad Y \;=\; \pi(X, E),$$

which (ideally) converts unmethylated cytosines to uracil (sequenced as thymine) while leaving methylated cytosines as cytosines; other bases are unchanged. Here $\mathcal{X}_{\text{BS}} \subseteq \{A, T, C, G\}^*$ is the induced post-conversion space.

Definition 2 (Language Models, Embeddings, and Predictive Heads). *Let $f : \mathcal{X} \to \mathbb{R}^d$ be a genomic language model (gLM) pretrained on native DNA, and let $g : \mathcal{X}_{\text{BS}} \to \mathbb{R}^d$ be the same architecture after continual pretraining on BS-seq observations. Here d is the embedding dimension. For an input $U \in \mathcal{X}$ or $\mathcal{X}_{\text{BS}}$, write $Z_f = f(U)$ and $Z_g = g(U)$ for (pooled) sequence embeddings.*

Let $p_\theta(\cdot \mid \text{ctx})$ denote the model's token predictive distribution (the softmax of a linear head applied to intermediate representations), with "ctx" the appropriate context such as a masked context for masked language modeling (MLM) or a prefix for next-token prediction (NTP).

A.1.2 The Statistical Nature of gLMs

Theorem 1. *Pretrained gLMs are estimators of the context-conditional probability distribution over tokens.*

Proof. Let P_{pretrain} be the data-generating distribution over token and context pairs in the pretraining dataset. The population objective during MLM pretraining is

$$\mathcal{L}_{\text{MLM}}(\theta) = \mathbb{E}_{X \sim P_{\text{pretrain}},\, M}\left[-\sum_{i \in M} \log p_\theta(x_i \mid X_{\setminus M}) \right], \tag{1}$$

where M is a masking pattern independent of the masked tokens given the unmasked context. More generally, the pretraining objective can be written as

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathrm{ctx}}\Big[H(P_{\mathrm{pretrain}}(\cdot \mid \mathrm{ctx})) + \mathrm{KL}\big(P_{\mathrm{pretrain}}(\cdot \mid \mathrm{ctx}) \,\big\|\, p_\theta(\cdot \mid \mathrm{ctx})\big)\Big], \tag{2}$$

where $H(P_{\mathrm{pretrain}}(\cdot \mid \mathrm{ctx}))$ denotes the Shannon entropy of the true pretraining distribution conditioned on context ctx, and $\mathrm{KL}(P_{\mathrm{pretrain}}(\cdot \mid \mathrm{ctx}) \,\|\, p_\theta(\cdot \mid \mathrm{ctx}))$ is the Kullback–Leibler (KL) divergence measuring how far the model distribution p_θ is from the true pretraining data distribution.

Minimizing $\mathcal{L}(\theta)$ over θ is equivalent to minimizing the conditional KL divergence. In the infinite-data, well-specified, and sufficient-capacity limit, any sequence of approximate minimizers satisfies $p_\theta(\cdot \mid \mathrm{ctx}) \to P_{\mathrm{pretrain}}(\cdot \mid \mathrm{ctx})$ almost surely. □

Proposition 1. *Continual pretraining on BS-seq data yields a model fine-tuning under distributional shift in P_{pretrain}. Directly following from Theorem 1, under sufficient model capacity and data, continual pretraining can lead to distributional convergence from $P_{\mathcal{X}}$ to $P_{\mathcal{X}_{\mathrm{BS}}}$.*

Proposition 2 (Methylation-Aware Attention)**.** *During continual pretraining on BS-seq data, attention-based models can develop attention patterns that preferentially weight methylation-informative positions. Let $A_g(Y)_{ij}$ denote the attention weight from position i to j in the BS-seq adapted model g for sequence Y. Then positions j in CpG contexts receive higher expected attention weights:*

$$\mathbb{E}[A_g(Y)_{ij} \mid j \in \textit{CpG}] > \mathbb{E}[A_g(Y)_{ij} \mid j \notin \textit{CpG}]. \tag{3}$$

Proof. The BS-seq transformation π creates systematic dependencies between token identities and methylation states: for positions in CpG contexts, C tokens indicate potential methylation (resistance to bisulfite conversion) while T tokens indicate unmethylated cytosines (successful conversion).

These patterns create reliable statistical signals for token prediction during MLM or NTP: in both settings, the objective is an expected negative log-likelihood of a target token given a context ("ctx"), i.e., $\mathbb{E}[-\log p_\theta(x \mid \mathrm{ctx})]$, where $\mathrm{ctx} = X_{\setminus M}$ under MLM and $\mathrm{ctx} = x_{<t}$ under NTP. Because CpG-context tokens in Y carry systematic information about the underlying (unobserved) methylation state, they reduce uncertainty about nearby targets under $P_{\mathcal{X}_{\mathrm{BS}}}(\cdot \mid \mathrm{ctx})$ and thus provide higher expected utility for prediction.

In attention-based architectures, improving $p_\theta(\cdot \mid \mathrm{ctx})$ requires routing information from informative positions into the hidden states used by the predictive head. Gradient-based training therefore increases capacity allocated to methylation-informative CpG positions—in transformers, this manifests as increased expected attention weight onto CpG-context indices (whether the target is a masked token under MLM or the next token under NTP). □

Remark (beyond attention). The same qualitative conclusion is expected for non-attention architectures. While such models do not expose an attention matrix A_g, the training objective still performs credit assignment: parameters connected (via the architecture's receptive field) to methylation-informative positions and their local context receive systematically larger gradient signal because CpG-context tokens are more predictive under BS-seq conversion. One can formalize this notion by an influence score at position j, e.g.,

$$I(j) := \mathbb{E}\big[\|\nabla_{h_j} \ell\|_2\big], \tag{4}$$

where h_j is the position-j hidden state and ℓ is the per-example MLM/NTP loss; then we expect $I(j)$ to be larger on CpG-context positions after BS-seq adaptation. In transformers, attention weights provide one observable proxy for this localization, but the underlying effect (allocation of representational capacity toward methylation-linked contexts) is not transformer-specific.

A.1.3 The Geometry of Epigenetic Learning

Definition 3 (Representation Geometry)**.** *For a sequence $Y \in \mathcal{X}_{\mathrm{BS}}$, let $Z_g(Y) \in \mathbb{R}^d$ denote its embedding after BS-seq adaptation. The* embedding norm $\|Z_g(Y)\|_2$ *measures the magnitude of activation. When the sequence context is fixed, the* epigenetic separation *between sequences Y_1 and Y_2 is usefully summarized by their cosine distance:*

$$d_{cos}(Z_g(Y_1), Z_g(Y_2)) = 1 - \frac{Z_g(Y_1) \cdot Z_g(Y_2)}{\|Z_g(Y_1)\|_2 \|Z_g(Y_2)\|_2} \tag{5}$$

Theorem 2 (Bimodal Norm Emergence under BS-seq Pretraining)**.** *Let g be obtained by continual pretraining of f on BS-seq data. Fix a conversion score $s(Y) \geq 0$ that measures CpG conversion. Concretely, for $Y = \pi(X, E)$ let $\mathcal{C}(X)$ be the set of cytosine positions in CpG context in the underlying locus X, and define $s(Y) := \frac{1}{|\mathcal{C}(X)|} \sum_{j \in \mathcal{C}(X)} \mathbf{1}\{Y_j = T\}$, which equals the fraction of CpG cytosines that are* unmethylated *under E (since BS-seq converts unmethylated C→T and leaves methylated C unchanged). Assume that there exist a function $U : \mathcal{X} \to \mathbb{R}^d$, a unit vector $v \in \mathbb{R}^d$, a scalar $\alpha > 0$, and an error term $\varepsilon(Y)$ such that for $Y = \pi(X, E)$,*

$$Z_g(Y) = U(X) + \alpha\, s(Y)\, v + \varepsilon(Y), \qquad \|\varepsilon(Y)\|_2 \leq \eta, \tag{6}$$

where $\eta \geq 0$ is a uniform bound capturing residual embedding variation not explained by the locus-dependent term $U(X)$ and the methylation-linked conversion score $s(Y)$. Assume further that $\langle U(X), v \rangle \geq 0$ for the loci considered. Here one can interpret $U(X)$ as a locus-dependent "baseline" embedding that captures sequence content shared across reads from the same genomic context, v as a (learned) representation direction aligned with methylation-linked conversion features, and α as the strength with which those features are expressed in the pooled embedding after BS-seq adaptation. Then for sequences $Y = \pi(X, E)$:

1. *If $s(Y)$ is larger (e.g., hypomethylated CpG contexts, hence more C→T), then $\|Z_g(Y)\|_2$ is (up to $O(\eta)$) larger.*
2. *If $s(Y)$ is near zero (e.g., hypermethylated contexts, hence few conversions), then $\|Z_g(Y)\|_2$ remains within $O(\eta)$ of $\|U(X)\|_2$.*
3. *If the marginal distribution of $s(Y)$ is a two-component mixture with well-separated components (hyper vs. hypo) and η is small relative to the induced separation, then the marginal distribution of $\|Z_g(Y)\|_2$ is bimodal.*

Intuition. BS-seq conversion writes epigenetic state into the token sequence: in CpG contexts, C vs. T is systematically linked to methylation status. Under MLM/NTP, reducing loss on BS-seq reads therefore incentivizes the model to represent (and use for prediction) conversion patterns and their surrounding context. A convenient mental model is that the pooled embedding decomposes into (i) a baseline component driven by the underlying genomic locus X and (ii) an additional component whose magnitude increases with the amount of conversion $s(Y)$. Importantly, because hypermethylated reads in CpG contexts undergo few conversions, they remain closer to native-DNA-like token statistics and can be represented using the model's existing "DNA" features (sometimes even more efficiently after adaptation); conversely, hypomethylated reads exhibit many C→T conversions and thus present more novel, systematic token patterns that the adapted model allocates representational capacity to, which can increase embedding magnitude. If the dataset contains two subpopulations (hyper vs. hypo) with clearly different conversion rates, this can naturally yield two norm modes.

Throughout, this intuition is intentionally coarse: it reduces BS-seq effects to the scalar CpG conversion score $s(Y)$ and does not attempt to disentangle CpG-context-specific pattern learning from simpler sensitivity to token-composition shifts.

Proof. By equation 6 and the reverse triangle inequality,

$$\big|\, \|Z_g(Y)\|_2 - \|U(X) + \alpha s(Y) v\|_2 \,\big| \leq \|\varepsilon(Y)\|_2 \leq \eta. \tag{7}$$

It therefore suffices to analyze $\|U(X) + \alpha s v\|_2$ as a function of $s \geq 0$.

Let $u := U(X)$. Since $\langle u, v \rangle \geq 0$ and $\|v\|_2 = 1$,

$$\|u + \alpha s v\|_2^2 = \|u\|_2^2 + 2\alpha s \langle u, v \rangle + \alpha^2 s^2 \tag{8}$$

is nondecreasing in s for $s \geq 0$, hence $\|u + \alpha s v\|_2$ is also nondecreasing in s. This proves that larger conversion score $s(Y)$ implies a larger norm up to the η perturbation in equation 7, giving (1). If $s(Y) \approx 0$, then equation 7 yields $\|Z_g(Y)\|_2 = \|u\|_2 + O(\eta)$, giving (2).

For (3), suppose $s(Y)$ is a mixture $s \sim \lambda P_{\text{hyper}} + (1 - \lambda) P_{\text{hypo}}$ whose components are well separated in the sense that their supports are separated by a gap $\Delta_s > 0$. Since $s \mapsto \|u + \alpha s v\|_2$ is nondecreasing and continuous, the pushforward of this mixture under the map $s \mapsto \|u + \alpha s v\|_2$ is itself a separated two-component mixture. Choosing η sufficiently small relative to the induced separation in norm space preserves two distinct modes after adding the perturbation $\varepsilon(Y)$. □

Theorem 3 (Cosine Distance Distribution and Epigenetic Heterogeneity). *Let $Y_A = \pi(X, E_A)$ and $Y_B = \pi(X, E_B)$ be BS-seq observations from the same genomic locus X but under different biological conditions with epigenetic states E_A and E_B respectively. After continual pretraining, the distribution of cosine distances between matched pairs (Y_A, Y_B) from the same genomic context is given by:*

$$P(d_{cos}) = \mathbb{E}_{X\sim\mathcal{X},\, E_A\sim P_A,\, E_B\sim P_B}\Big[\delta\Big(d_{cos} - d_{cos}\big(Z_g(Y_A), Z_g(Y_B)\big)\Big)\Big], \tag{9}$$

where d_{cos} is the cosine distance variable and $\delta(\cdot)$ is the Dirac delta function. Moreover, under the affine representation model equation 6 applied to both conditions,

$$Z_g(Y_A) = U(X) + \alpha s_A v + \varepsilon_A, \qquad Z_g(Y_B) = U(X) + \alpha s_B v + \varepsilon_B, \tag{10}$$

with $\|\varepsilon_A\|_2, \|\varepsilon_B\|_2 \le \eta$, and assuming in addition that the learned methylation direction v is orthogonal to the locus-dependent baseline component $U(X)$ (i.e., $\langle U(X), v\rangle = 0$), the cosine distance admits the clean quadratic approximation

$$d_{\cos}\big(Z_g(Y_A), Z_g(Y_B)\big) = \frac{\alpha^2}{2\,\|U(X)\|_2^2}\,(s_A - s_B)^2 + O\!\left(\frac{\alpha^4(|s_A| + |s_B|)^4}{\|U(X)\|_2^4}\right) + O\!\left(\frac{\eta}{\|U(X)\|_2}\right), \tag{11}$$

whenever $\alpha \max\{|s_A|, |s_B|\} \ll \|U(X)\|_2$. If $\langle U(X), v\rangle \neq 0$, the Taylor expansion generally includes an additional first-order (in α) term proportional to $(s_A - s_B)$; we impose orthogonality only to simplify the leading dependence on the conversion-score gap.

Consequently, we define the following decomposition of the expected cosine distance:

$$\mathbb{E}[d_{cos}] = \mathbb{E}[d_{cos}]_{naive} + \Delta_{epi}(E_A, E_B \mid X) + \Delta_{domain}(P_A, P_B, P_{pretrain} \mid X) \tag{12}$$

where $\mathbb{E}[d_{cos}]_{naive}$ is a baseline matched-locus expectation (e.g., the expected cosine distance when $E_A = E_B$ and there is no distribution shift relative to pretraining), and where we define *Δ_{epi} to be the expected contribution attributable to epigenetic/conversion-score mismatch at fixed locus and Δ_{domain} to collect any remaining systematic increase due to distribution shift. This decomposition is meant as an interpretive bookkeeping device (motivated by the leading dependence on $(s_A - s_B)^2$ in equation 11), rather than a uniquely identifiable separation of causal effects.*

Intuition. By matching reads to the same genomic locus X, we largely hold fixed the base sequence and isolate the effect of methylation-linked BS-seq token changes. When both conditions are *familiar* to the adapted model—i.e., their BS-seq read statistics are close to the distribution the model was optimized on during continual pretraining—the model can explain both conditions using a shared locus-dependent baseline representation, and the remaining epigenetic signal induces relatively modest angular separation. In contrast, under distribution shift (e.g., tumor vs. normal) or when one condition's methylation/conversion patterns are less represented during adaptation, the conditional density-estimation objective is satisfied by allocating more representational capacity to methylation-informative contexts; as a result, differences in conversion features (and associated local context) are amplified in embedding space, strengthening the right tail and potentially introducing mixture-like structure in the cosine-distance distribution.

Proof. By constraining comparisons to the same genomic locus X, we isolate epigenetic effects from sequence variation. The embedding difference arises solely from the epigenetic state difference E_A vs E_B through the BS-seq transformation π.

Equation equation 9 is the standard pushforward characterization of the distribution of a statistic of random variables.

For the approximation equation 11, plug the affine forms for $Z_g(Y_A)$ and $Z_g(Y_B)$ into the cosine distance definition equation 5. Under the orthogonality assumption $\langle U(X), v\rangle = 0$ and the small-perturbation regime $\alpha \max\{|s_A|, |s_B|\} \ll \|U(X)\|_2$, a Taylor expansion of the numerator and denominators around $U(X)$ yields the leading term proportional to $(s_A - s_B)^2$. Under $\langle U(X), v\rangle = 0$, the cosine similarity depends on α only through α^2, so the next correction is fourth-order in α (equivalently, $O((\alpha \max\{|s_A|, |s_B|\}/\|U(X)\|_2)^4)$); the $O(\eta/\|U(X)\|_2)$ term accounts for bounded embedding noise.

Finally, equation 12 is a *definition* of a two-term decomposition: Δ_{epi} isolates the portion attributable to epigenetic mismatch at fixed locus (suggested by the leading $(s_A - s_B)^2$ dependence in equation 11), while Δ_{domain} aggregates any remaining systematic contribution due to distribution shift (e.g., changes in read composition beyond what is summarized by s). □

A.2 CONTINUAL PRETRAINING

To empirically validate our theoretical predictions, we continually pretrained the DNABERT2 model with a 512 base-pair (bp) context window on whole-genome bisulfite sequencing (WGBS) data (Appendix Figure S1). Starting from the DNABERT2-117M checkpoint on HuggingFace, the model was trained for 100,000 steps using a biased masking scheme – 80 % of masked tokens drawn from CG (methylated CpG)/TG (unmethylated CpG) sites and 15 % selected at random – and optimized with a focal loss Lin et al. (2017) to enhance its ability to capture genomic-epigenetic interactions. Training was performed on four NVIDIA A100 GPUs using Distributed Data Parallel (DDP) with mixed-precision training (bfloat16), and a per-device batch size of 1024.

The WGBS data were derived from the DNA methylation atlas of normal human cell types spanning 39 tissues Loyfer et al. (2023). To capture biologically meaningful variation, we focused on approximately 50,000 differentially methylated regions (DMRs) identified by Loyfer et al., representing key inter-tissue methylation diversity. This subset comprised roughly one billion reads, about half of which were sampled for pretraining within our computational budget.

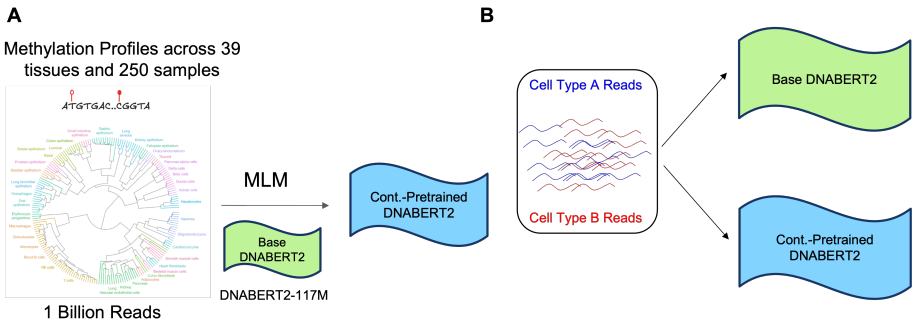

Figure S1: (A) DNABERT2 was continually pretrained on bisulfite-converted reads derived from a normal tissue methylation atlas. The model was trained using a masked-language-modeling (MLM) objective to adapt the base DNABERT2 model for methylation-aware sequence modeling. (B) The base and continually pretrained DNABERT2 models were then used to embed bisulfite-converted reads from distinct cell types. Comparison of read-level embeddings across cell types assessed whether continual pretraining enhanced the model's sensitivity to methylation-driven sequence variation.

A.3 EMBEDDING EXTRACTION

After continual pretraining, we extracted sequence embeddings from the DNABERT2 model by forwarding bisulfite-converted read sequences through the encoder and pooling the final hidden states. Bisulfite sequences were tokenized using the original DNABERT2 tokenizer using padding and truncation to a fixed length of 512 tokens. Tokenized inputs (`input_ids` and `attention_mask`) were passed to either the pretrained DNABERT2 encoder or the continually pretrained DNABERT2 encoder in inference mode, bypassing MLM head computation.

For each bisulfite-converted sequence $Y = (y_1, \ldots, y_L)$, hidden states of valid non-padding tokens were aggregated by padding-aware mean pooling:

$$Z_g(Y) = \frac{\sum_{t=1}^{L} m_t H_t}{\sum_{t=1}^{L} m_t}, \tag{13}$$

where $H_t \in \mathbb{R}^{768}$ is the final-layer representation of token t, and $m_t \in \{0, 1\}$ is the corresponding attention-mask indicator specifying valid (non-padded) positions. Embedding norms were computed directly as $\|Z_g(Y)\|_2$ and cosine distances between pairs of embeddings were evaluated as Equation 5.

In-distribution and out-of-distribution evaluation To assess both within- and out-of-distribution behavior, we analyzed hepatocyte–granulocyte (in-distribution during pretraining) and DLBCL–B–cell (out-of-distribution) read pairs using the same embedding extraction pipeline as follows. For each read overlapping the top differentially methylated regions (DMRs) between the two tissues (250 DMRs for in-distribution, 3000 for out-of-distribution), we computed the pooled embedding $Z_g(Y)$ and quantified the distribution of embedding norms $\|Z_g(Y)\|_2$ within each tissue. To assess inter-tissue geometric separation, we sampled genomically overlapping Tissue A–Tissue B read pairs and computed pairwise cosine distances. These metrics established empirical results for investigating the emergence of bimodal norms and distance amplification predicted by Theorems 2 and 3.

A.4 Embedding Comparison for In-Distribution Read Pairs

To assess whether the geometric signatures observed in tumor–normal comparisons also emerge in tissues represented during pretraining, we analyzed embedding pairs from hepatocytes and granulocytes (Appendix Figure S2). As expected for in-distribution tissues, cosine distances between genomically overlapping reads were narrowly concentrated near zero (Appendix Figure S2C), indicating that both cell types share highly similar representational geometry. Nonetheless, when we stratified these read pairs by cosine distance, we found that those exhibiting higher angular separation were still enriched for larger absolute methylation differences (Appendix Figure S2D). This observation confirms that even within stable, in-distribution epigenetic contexts, the model's embedding space retains sensitivity to subtle methylation variation: small geometric deviations in cosine space correspond to biologically meaningful methylation differences. Thus, while large-scale cosine divergence primarily emerges under out-of-distribution conditions, the same geometric–epigenetic correspondence persists at finer scales within normal tissue variation.

By contrast, tumor–normal comparisons display a pronounced right-shift and occasional bimodality in cosine distances, indicating greater heterogeneity in the learned representation space (Figure 1C). This pattern reflects both increased epigenetic variability within DLBCL samples and systematic deviations of their methylation landscapes from the sequence patterns captured during pretraining. Importantly, this geometric divergence persists even when analyses are restricted to reads overlapping known differentially methylated regions, suggesting that the model encodes broader methylation-driven representational structure beyond explicitly differential sites. Together, these findings show that cosine-distance amplification provides a measurable geometric signature of epigenetic divergence under distribution shift.

A.5 Cosine Distance–Methylation Relationship Across Models

To further quantify how geometric separation in embedding space reflects biological methylation differences, we examined the joint distribution between cosine distance and absolute mean methylation difference for genomically overlapping read pairs (Appendix Figure S3). The scatter and marginal distributions illustrate that, for both the base DNABERT2 model and the continually pretrained model, read pairs with small cosine distances correspond primarily to methylation-similar regions, while those with larger cosine separations are enriched for pairs exhibiting high methylation differences.

In the base model (Appendix Figure S3A), this relationship is weakly expressed: most read pairs cluster at low cosine distances regardless of methylation divergence (Pearson's $r = 0.17$, Spearman's $\rho = 0.19$, Kendall's $\tau = 0.15$). After continual pretraining (Appendix Figure S3B), the association is stronger, with higher cosine distances increasingly aligning with larger methylation differences, forming distinct density contours along the upper right of the joint distribution (Pearson's $r = 0.41$, Spearman's $\rho = 0.51$, Kendall's $\tau = 0.39$). A Fisher r-to-z comparison confirmed that the correlation enhancement was highly significant ($z = 28.2$, $p < 10^{-16}$). These patterns demonstrate that the methylation-aware model learns a mapping between angular distance in embedding space and biochemical divergence, even though the overall range of cosine distances remains

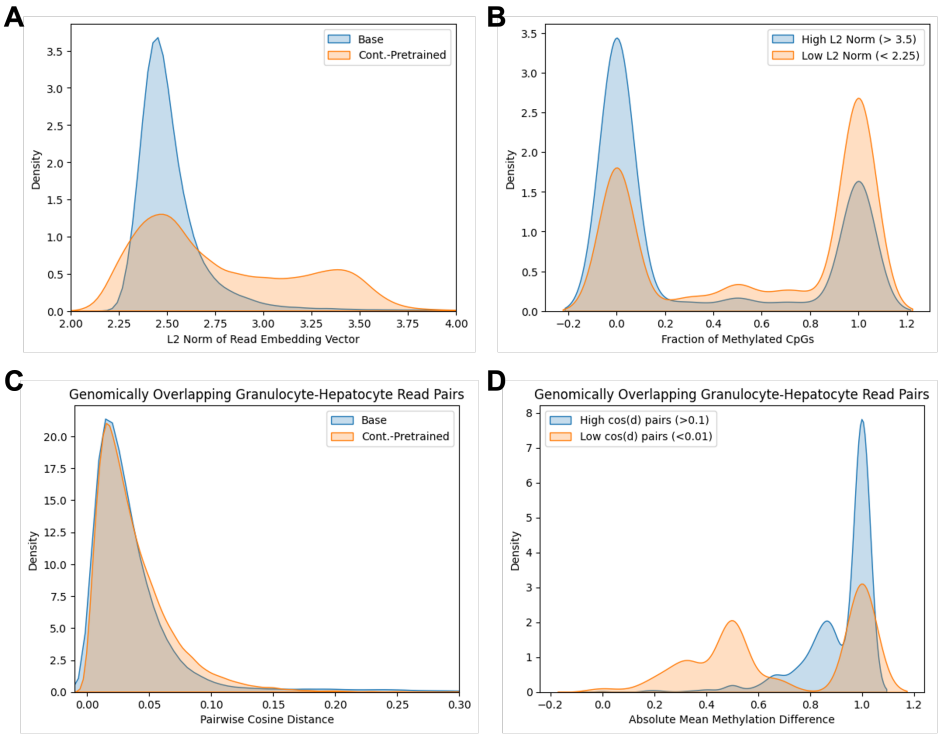

Figure S2: (A) Embedding-norm distributions for individual Hepatocyte and Granulocyte bisulfite-converted reads before (blue) and after (orange) continual pretraining. (B) Reads with high embedding norms correspond predominantly to hypomethylated CpG contexts, whereas those with low norms are enriched for hypermethylated regions, confirming the link between geometric magnitude and methylation state. (C) Cosine-distance distributions between embeddings of genomically overlapping granulocyte–hepatocyte read pairs. Continual pretraining only slightly increases pairwise angular separation for in-distribution cell types (Theorem 3). (D) Read pairs with large cosine distances show greater absolute mean methylation differences, establishing that embedding-space separation tracks biological methylation divergence

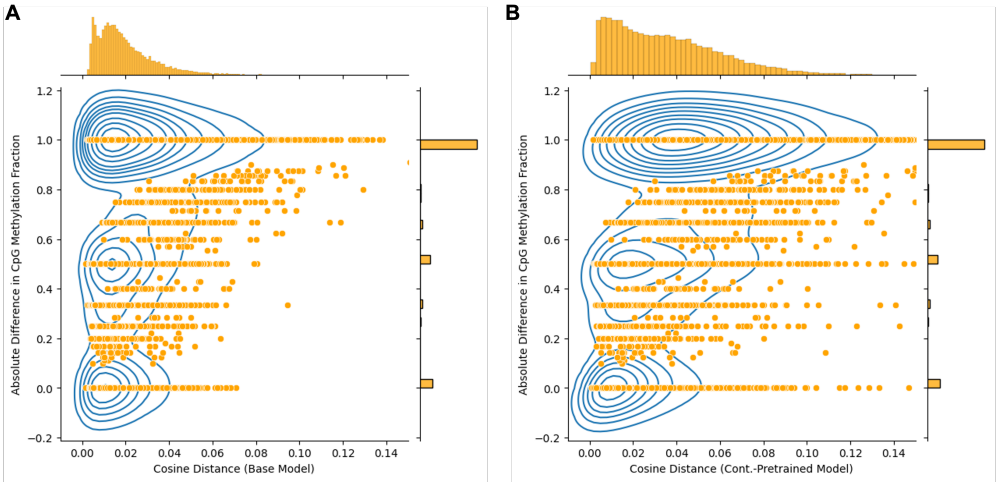

Figure S3: (A) Joint distribution between cosine distance and absolute mean methylation difference for genomically overlapping read pairs using the base DNABERT2 model. Cosine distances remain narrowly distributed and show weak correspondence with methylation differences, indicating limited methylation sensitivity in the pretrained DNA–only representation. (B) Same analysis for the methylation–aware model after bisulfite-sequence pretraining. A clearer positive trend emerges–read pairs with larger cosine distances are enriched for greater methylation divergence, demonstrating improved geometric alignment between embedding separation and biological methylation variability.

small. Together, these results reinforce that cosine geometry encodes methylation state information: pretraining on bisulfite data does not merely amplify distances globally but reshapes their alignment with underlying epigenetic variability.