Methylation-Aware Embedding Geometry Emerges from Bisulfite Pretraining in DNA Language Models

Jiajie Xiao*, Salwan Butrus*, Nathan Hunkapiller
[1]Curve Biosciences (San Mateo, CA; Dallas, TX). *Equal contribution.

2026 Workshop on Learning Meaningful Representations of Life

Introduction

Background

- DNA methylation (5mC at CpG sites) regulates gene expression, cell identity, development, and disease.
- In bisulfite sequencing (BS-seq): unmethylated cytosines (C) are converted to thymines (T) while methylated C are unchanged. This encodes epigenetic state directly into token identities.
- Most genomic language models (gLMs) are pretrained on native DNA only, without methylation context.
- Retrofitting *via* continual pretraining starting from well-developed DNA checkpoints could endow gLMs with methylation knowledge without new architectures and pretraining efforts.

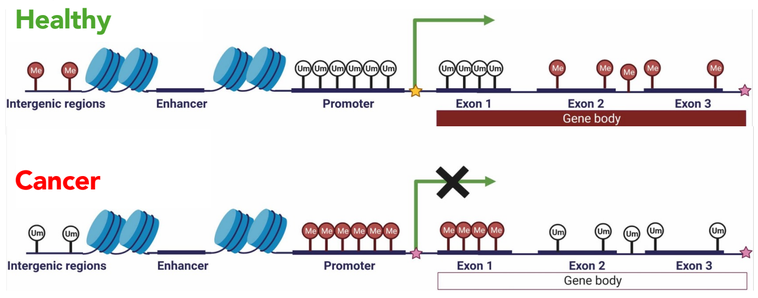

Figure 1. DNA methylation dysregulation in a tumor suppressor gene. Hypermethylation at the promoter silences the tumor suppressor gene, amplifying disease progression. Bisulfite sequencing encodes this state directly into token identities *via* C→T conversion.

Approach & Key Questions

- Can continual pretraining on BS-seq reads retrofit DNABERT2 with methylation sensitivity? (Feasibility)
- What compact, measurable signatures appear in representation space? (Interpretability)
- Can signatures be computed cheaply and remain informative under distribution shift? (Diagnosis)

Theoretical Framework

Theorem 1: gLMs as Distribution Estimators

- Pretrained gLMs approximate context-conditional token distributions:

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

- Continual pretraining on BS-seq data induces a distributional shift: $P_X \xrightarrow{\pi} P_{X_{BS}}$(Definition 1, Proposition 1), where $X = \{A, C, T, G\}$ denotes the set of finite DNA strings over the canonical alphabet and X_{BS} is the induced post-bisulfite conversion space.

Proposition: Methylation-Aware Attention

- Let $A_g(Y)_{ij}$ denote the attention weight from position i to j in the BS-seq adapted model g for sequence Y. After BS-seq pretraining:

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

- CpG tokens carry systematic methylation information → higher gradient signal → preferentially weighted by attention.

Definition:

For a sequence $Y \in X_{BS}$, let $Z_g(Y) \in \mathbb{R}^d$ denote its embedding after BS-seq adaptation. The embedding norm measures the magnitude of activation. The epigenetic separation between sequences Y_1 and Y_2 can be summarized by their cosine distance:

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

Theorem 2: Bimodal Norm Emergence

- After BS-seq adaptation, each read embedding decomposes into a sequence component and a methylation component

$$Z_g(Y) = U(X) + \alpha\, s(Y)\, v + \varepsilon(Y), \qquad \|\varepsilon(Y)\|_2 \le \eta,$$

- $U(X)$: sequence baseline shared across all reads from locus X regardless of methylation state.
- $s(Y)$: fraction of CpG cytosines unmethylated in read Y. Bimodally distributed across the genome: high = hypomethylated, low = hypermethylated. Two-state methylation shift.
- αv: learned methylation direction in embedding space; α scales with BS-seq adaptation strength.
- $\varepsilon(Y)$: bounded residual.

Theorem 3: Cosine Distance Amplification

- 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. The expected cosine distance can be derived as:

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

- $\Delta_{epi} \propto (s_A - s_B)^2$: epigenetic contribution grows with CpG conversion mismatch between conditions.
- Δ_{domain} captures additional distributional novelty (e.g., tumor vs. healthy out-of-distribution shift).

Testable Geometric Predictions

Prediction 1: Norm Bifurcation

- Per-read embedding norms $\|Z_g(Y)\|_2$ become bimodal after BS-seq adaptation.
- High-norm reads: hypomethylated contexts (more C→T, greater representational capacity allocated).
- Low-norm reads: hypermethylated contexts (fewer conversions → closer to native DNA distribution).

Prediction 2: Cosine-Distance Amplification

- Cosine distances between reads at the same locus increase after adaptation.
- Amplification is stronger when methylation divergence between conditions is larger.
- Right tail of cosine distribution enriched under large distribution shift (e.g., tumor vs. healthy).

Study Design & Data

Base Model

- DNABERT2-117M: continually pretrained from HuggingFace checkpoint; no architecture changes.

Data Used for Continuous Pretraining

- WGBS atlas: 39 tissues, 250 samples (Loyfer et al., 2023)
- Focused on ~50,000 differentially methylated regions (DMRs) representing inter-tissue variation.
- ~1 billion reads total; ~500 million sampled for pretraining.

Continuous Pretraining Setup

- 512 bp context window; Epigenetics-biased MLM: 80% CG/TG + 15% random masking, with 5% identity tokens to anchor genomic context.
- Focal loss: enhances learning of rare CpG-context patterns.
- 100,000 steps | Batch size: 1,024 | 4× NVIDIA A100 | bfloat16 / Distributed Data Parallel (DDP).

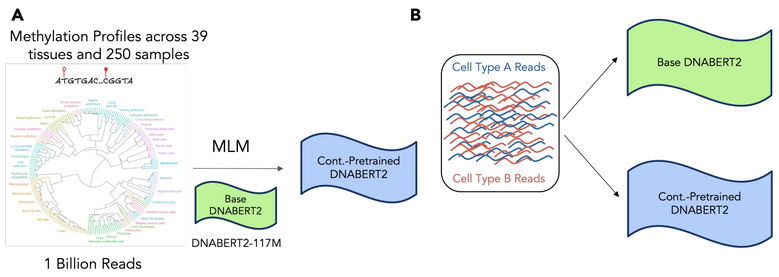

Figure 2. Study Design Overview. (A) DNABERT2 was continually pretrained on bisulfite-converted reads derived from a healthy tissue methylation atlas. The model was trained using a masked-language-modeling (MLM) objective to adapt the base DNABERT2 model for methylation-aware sequence modeling. (B) The 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.

Embedding Diagnostics

Two Label-Free Metrics from Embeddings:

- Embedding Norm $\|Z_g(Y)\|_2$: per-read magnitude; proxy for CpG conversion rate and methylation level.
- Cosine Distance: angular separation between paired embeddings; proxy for epigenetic divergence.

Matched-Locus Design:

- Read pairs drawn from the same genomic locus across conditions, with sequence differences controlled.
- Isolates epigenetic effects encoded via BS-seq; no protocol-specific labels needed.
- Applicable before any downstream fine-tuning or supervised benchmarks.

In-Distribution Cell Types (Hepatocyte vs. Granulocyte)

- Test whether the geometric signatures reflect genuine methylation sensitivity rather than distributional novelty.
- Analyzed embedding pairs from hepatocytes and granulocytes: two cell types present in the pretraining data.
- Top 250 DMRs; ~1M read pairs analyzed.

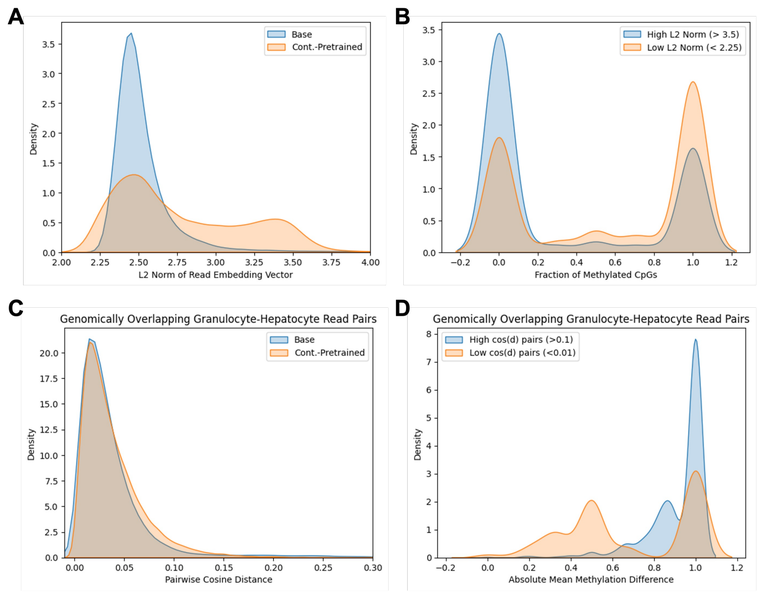

Figure 3. In-distribution read embedding evaluation. (A) Embedding-norm distributions for individual Hepatocyte and Granulocyte bisulfite-converted reads before (blue) and after (orange) continual pretraining. (B) Reads with high embedding norms correspond predominantly to hypomethylated CpG contexts, 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.

- Norm bifurcation and cosine–methylation correspondence are observed for cell types seen during pretraining, suggesting that the geometry reflects true epigenetic variation rather than distributional novelty.
- Cosine-distance amplification is minuscule, consistent with Theorem 3, which predicts weaker separation when both conditions are within the training distribution.
- The model is sensitive but not overactive: angular separation tracks methylation divergence at fine scales without inflating distances where epigenetic differences are small.

Out-of-Distribution Cell Types (Tumor vs. Healthy)

- Top 3,000 DMRs; testing geometric generalization.
- Bisulfite-adapted DNABERT2 produces bimodal embedding norm distributions across DLBCL reads at DMRs: the two modes correspond to hypo- and hypermethylated read populations, emerging without label supervision.

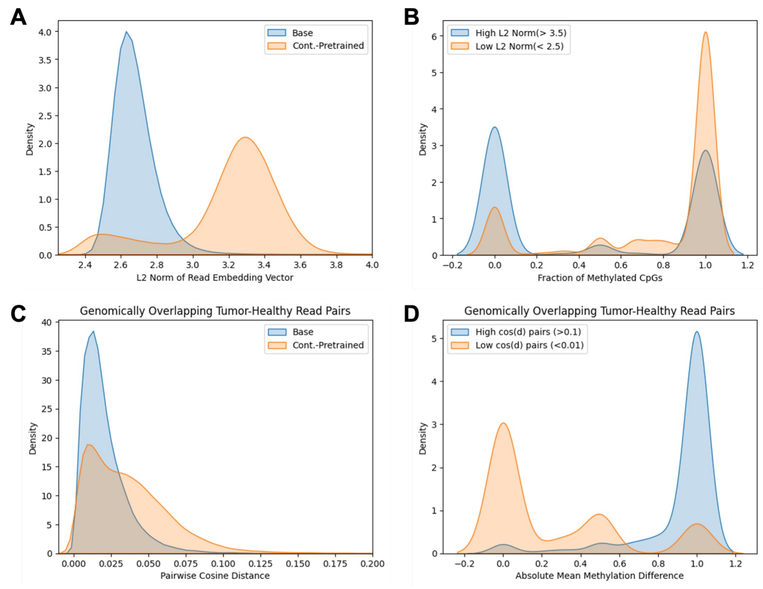

Figure 4: Embedding geometry encodes methylation after BS-seq pretraining (DLBCL vs. B-cells). (A) Norm distributions shift from unimodal (base) to bimodal (adapted). (B) High-norm reads: hypomethylated CpGs; low-norm: hypermethylated. (C) Cosine distances increase for tumor–healthy overlapping read pairs. (D) High cosine-distance pairs exhibit larger absolute methylation differences.

Norm Bimodality

- Mean norm increase: $\Delta = +0.51 \pm 0.001$ (Cramér–von Mises $T = 1{,}056{,}698$; $p < 10^{-5}$; Cohen's $d = 1.59$).
- Bimodal modes: 2.58 ± 0.15 (hypermethylated) and 3.30 ± 0.18 (hypomethylated); Ashman's $D = 3.07$.

Cosine-Distance Amplification

- Mean shift: $\Delta d_{cos} = 0.016$ (95% CI [0.0157, 0.0163]; paired t = 98.9; $p < 10^{-5}$; Cohen's $d = 0.60$).
- 76.3% of read pairs show higher cosine distance post-adaptation
- Right-tail enrichment ($d_{cos} > 0.05$): 5.0% → 26.9% ($\Delta = +21.9\%$; $z = 69.6$; $p < 10^{-5}$).

Methylation Correspondence

- Embedding norm magnitude tracks methylation state: high-norm reads are depleted for methylated CpGs: Cramér–von Mises $T = 1012$, $p < 10^{-5}$; $\Delta\mu = -0.33$ (95% CI [−0.34, −0.33]); Cohen's $d = -0.78$.
- Read pairs with high cosine distance show greater methylation divergence ($\Delta = 0.66$, Welch $t = 54.6$, $p < 10^{-5}$, Cohen's $d = 2.20$).
- Cosine–methylation correlation increases after continuous pretraining: Pearson's $r = 0.17 \rightarrow 0.41$, Spearman's $\rho = 0.19 \rightarrow 0.51$, Kendall's $\tau = 0.15 \rightarrow 0.39$).
- Fisher r-to-z: $z = 28.2$, $p < 10^{-5}$: cosine geometry increasingly tracks epigenetic divergence.

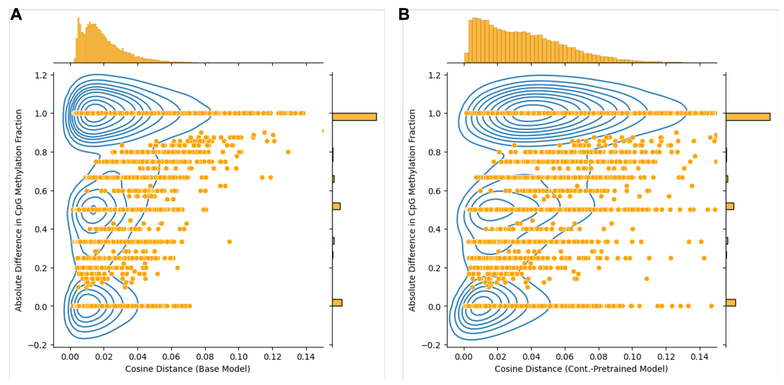

Figure 5. Joint distribution of cosine distance and absolute CpG methylation difference for tumor vs. healthy read pairs. (A) Base DNABERT2: weak correspondence between cosine distance and absolute mean methylation difference for genomically overlapping read pairs. (B) BS-seq adapted model: correspondence strengthens substantially.

Conclusion & Future Directions

Conclusions

- Continual pretraining on BS-seq data induces clear methylation-aware geometry: norm bimodality and cosine amplification emerge as theoretically predicted.
- Diagnostics are label-free, interpretable, and computationally cheap, serving as effective assessments before committing to costly downstream benchmarks.
- Learned geometry generalizes out-of-distribution (DLBCL tumor), demonstrating robust methylation encoding beyond training data.
- Simple retrofitting via BS-seq pretraining endows a standard DNA gLM with label-light epigenetic sensitivity.

Future Directions

- Extend to other gLM backbones at larger parameter scales.
- Broader tissue types, disease states (cancers, immune disorders), and assay types beyond WGBS.
- Task-driven evaluations to clarify when task-specific fine-tuning adds value beyond geometry-level diagnostics (e.g., demonstrating performance improvement using benchmarks for disease classification tasks).
- Data scaling and sampling strategies for more efficient epigenetic pretraining within computational budgets.
- Studying the extent to which BS-seq pretraining preserves sequence-interaction information learned during native DNA pretraining.

References

1. Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., & Liu, H. (2023). DNABERT-2: Efficient foundation model for multi-species genome. arXiv:2306.15006.
2. Loyfer, N., Magenheim, J., Peretz, A., Cann, G., Bredno, J., Klochendler, A., et al. (2023). A DNA methylation atlas of human cell types. Nature, 613, 355–364.
3. Do, C., Dumont, E. L. P., Salas, M., Castano, A., Mujahed, H., Maldonado, L., et al. (2020). 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, 153.
4. Smith, Z. D., & Meissner, A. (2013). DNA methylation: Roles in mammalian development. Nature Reviews Genetics, 14, 204–220.
5. Kulis, M., & Esteller, M. (2010). DNA methylation and cancer. Advances in Genetics, 70, 27–56.
6. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2980–2988).
7. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., et al. (2024). Towards generalist biomedical AI. NEJM AI, 1(3): AIoa2300138.

Acknowledgements

We thank Dr. Mahdi Baghbanzadeh for helpful discussions on 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 part by resources and services provided by Google Cloud Platform and Amazon Web Services.