scDNAm-GPT: A Foundation Model for Capturing Long-Range CpG Dependencies in Single-Cell Whole-Genome Bisulfite Sequencing to Enhance Epigenetic Analysis

Chaoqi Liang[1,9†], Peng Ye[2,3†], Hongliang Yan[9†], Peng Zheng[2†], Jianle Sun[2,4], Yanni Wang[9], Yu Li[9], Yuchen Ren[2,5], Yuanpei Jiang[9], Ran Wei[2], Junjia Xiang[1], Sizhe Zhang[1], Linle Jiang[9], Weiqiang Bai[2], Xinzhu Ma[2,3], Tao Chen[6], Wangmeng Zuo[1*], Lei Bai[2*], Wanli Ouyang[2*], Jia Li[7,8,9*]

[1]Harbin Institute of Technology.
[2]Shanghai Artificial Intelligence Laboratory.
[3]The Chinese University of Hong Kong.
[4]Carnegie Mellon University.
[5]The University of Sydney.
[6]Fudan University.
[7]State Key Laboratory of Respiratory Disease, Guangzhou Institute of Cancer Research, the Affiliated Cancer Hospital, Guangzhou Medical University.
[8]Department of Laboratory Medicine, National Clinical Research Center for Respiratory Disease, National Center for Respiratory Medicine, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou Medical University.
[9]Guangzhou National Laboratory.

*Corresponding author(s). E-mail(s): wmzuo@hit.edu.cn; bailei@pjlab.org.cn; ouyangwanli@pjlab.org.cn; jiali@gzhmu.edu.cn;
Contributing authors: lcqfacai@gmail.com; yepeng@pjlab.org.cn; yhldhit@gmail.com; zhengpeng0108@gmail.com; jianles@andrew.cmu.edu; wangyanni541@gmail.com; jade.li1788@gmail.com; renyuchen@pjlab.org.cn;

yuanpei130054@gmail.com; weiran@pjlab.org.cn;
2022110188@stu.hit.edu.cn; 2022110818@stu.hit.edu.cn;
jiang_linle@gzlab.ac.cn; baiweiqiang@pjlab.org.cn;
maxinzhu@pjlab.org.cn; eetchen@fudan.edu.cn;
†These authors contributed equally to this work.

Abstract

Accurately identifying development- and disease-associated DNA methylation features from single-cell DNA methylation data remains challenging due to the genome-wide scale and the sparse, stochastic nature of CpG coverage. We present scDNAm-GPT, a novel framework that integrates CpG token design, a Mamba backbone, and a cross-attention head to efficiently process ultra-long sequences while preserving both local CpG interactions and broader genomic context. Pretrained on over one million single cells from 28 human and mouse tissues, scDNAm-GPT effectively reconstructs sparse methylation landscapes, enhancing the resolution and accuracy of epigenetic analyses. It outperforms existing methods across key biomedical applications, including improved cell clustering, enhanced trajectory inference for precise mapping of differentiation pathways, identification of disease-relevant DNA methylation features, and robust, reference-free cell type deconvolution from cfDNA data. scDNAm-GPT learns regulatory features in a hierarchical manner and and its attention scores exhibit high biological interpretability by highlighting functionally relevant CpG regions. These advancements establish scDNAm-GPT as a scalable and generalizable solution for single-cell epigenomic analysis, paving the way for broader applications in single-cell DNA methylation profiling and uncovering novel insights into the epigenetic mechanisms underlying health and disease. The code for scDNAm-GPT is available at https://github.com/ChaoqiLiang/scDNAm-GPT.

Keywords: Single-Cell, Whole-Genome, DNA methylation, GPT

1 Introduction

The rapid advancement of single-cell data analysis technologies has revolutionized our understanding of cellular heterogeneity, lineage differentiation, and epigenetic regulation. At the forefront of this revolution are generative language models, particularly Transformer-based architectures, which have transformed single-cell omics research. Models like scBERT [1], scGPT [2] and scFoundation [3] have demonstrated state-of-the-art performance in single-cell RNA sequencing (scRNA-seq) analysis, excelling in tasks such as cell clustering, annotation and gene perturb prediection. These models leverage powerful deep learning techniques to extract complex interrelations within single-cell data, significantly improving our ability to interpret biological systems. However, Transformer-based language models face a critical limitation: the computational complexity of their self-attention mechanism scales quadratically with sequence length, restricting their applicability to datasets with sequences longer than a few

thousand words or elements. This limitation excludes many important types of biological data, such as single-cell whole-genome methylation data, from benefiting fully from these transformative approaches.

Single-Cell Whole-Genome Bisulfite Sequencing (scWGBS) [4, 5] is a state-of-the-art technique that profiles DNA methylation at single-nucleotide and single-cell resolution, providing unparalleled insights into whole-genome epigenetic regulation. scWGBS reveals cell-to-cell variability, enabling a deeper understanding of developmental processes, cellular heterogeneity, and disease-associated epigenetic changes. The technique involves bisulfite treatment of DNA to distinguish methylated from unmethylated cytosines, followed by high-throughput sequencing to generate methylation data spanning hundreds of thousands to millions of CpG sites per cell. This comprehensive dataset holds immense potential for understanding epigenetic regulation. But their vast length and sparse coverage make them computationally prohibitive for Transformer-based language models.

Despite its promise, the analysis of scWGBS data presents significant computational and analytical challenges. The vast length of the sequences—often spanning millions of CpG sites—and their sparse, variable coverage render traditional machine learning approaches inadequate. Current methods [6] typically divide the genome into fixed regions (e.g., 10-kb windows) and calculate regional methylation rates. While this approach is computationally manageable, it sacrifices single-nucleotide precision and fails to capture the intricate dependencies between CpG sites. Moreover, these methods overlook the broader genomic context, leading to substantial information loss and limiting their ability to uncover subtle but critical biological mechanisms.

In response to this challenge, we developed a generative foundational language model named scDNAm-GPT. Notably, scDNAm-GPT can process a single cell with 10 million CpG sites at single-CpG resolution, demonstrating its scalability and efficiency for high-throughput epigenomic analysis. scDNAm-GPT fuses three key components: the CpG special token design, Mamba's Selective State Space Models (SSMs) [7], and Cross-Attention head [8], specifically designed to analyze scWGBS data at single-CpG resolution. The Mamba architecture, grounded in SSMs, is specifically designed to address the challenges associated with processing extremely long input sequences. The SSMs framework allows the model to selectively focus on biologically relevant portions of the input sequence, effectively filtering out less informative regions to reduce computational complexity. Additionally, the CpG special token design centers on a 6-mer strategy around CG motifs to integrate the DNA background, while incorporating methylation information by multiplying the tokens with the methylation rate before inputting them into the model. The incorporation of Cross-Attention head in the final layer enables dynamic integration of nucleotide-level context, allowing the model to capture both local and long-range dependencies between CpG sites. This fusion is particularly well-suited to the biological significance of CpG methylation, where interactions between distant sites often play critical roles in gene regulation and epigenetic mechanisms.

By leveraging this hybrid architecture, scDNAm-GPT efficiently handles the ultra-long sequences characteristic of scWGBS data, where individual cells contain an average of 2 million CpG sites and the longest sequences reach up to 10

million. Pre-trained on data from over 1,000,000 single cells spanning 28 different tissues from humans and mice [9–12], the model benefits from a comprehensive and diverse representation of DNA methylation patterns across species and biological contexts. This extensive pre-training enables scDNAm-GPT to capture intricate patterns within sparse methylation data, preserving both local CpG interactions and broader genomic context. This innovative design ensures scDNAm-GPT not only overcomes the computational bottlenecks of traditional methods but also provides biologically meaningful representations of DNA methylation across the entire genome. With its ability to represent sparse, stochastic methylation patterns at unprecedented resolution, scDNAm-GPT enables precise insights into cellular heterogeneity and epigenetic regulation across a wide range of biological systems.

Our research highlights that scDNAm-GPT significantly addresses the limitations of existing scWGBS analysis methods by leveraging the power of a generative language model. In cell annotation tasks, scDNAm-GPT achieves over 94% accuracy, surpassing previous approaches by offering more distinct and biologically relevant groupings of cells, while uncovering subtle aspects of cellular heterogeneity. In trajectory inference, the model excels in accurately mapping the differentiation pathways of cells, enabling detailed and continuous representations of cell state transitions. Moreover, scDNAm-GPT automatically identifies key single-cell methylation features, providing researchers with an invaluable tool for uncovering critical epigenetic markers. Collectively, these advancements deepen our understanding of DNA methylation at the single-cell level and establish a new benchmark for analyzing ultra-long genomic sequences in epigenetic research.

2 Results

2.1 The overview of scDNAm-GPT

We introduce scDNAm-GPT (Fig. 1), the first CpG language model designed to analyze single-cell methylation data with unprecedented precision and scalability. This work aims to leverage language models to decode DNA CpG site epigenetic modification information at single-cell resolution. Each CpG site, along with its surrounding nucleotides, forms a CpG motif—"XXC(M)GXX", where "X" represents any nucleotide [13]. In specific, we treat the CpG motif as a "word" and, based on this concept, construct a CpG language. We then use this framework to develop a CpG language processing model—scDNAm-GPT.

scDNAm-GPT adopts the cutting-edge Mamba architecture [7], specifically designed to handle the immense scale and complexity of scWGBS data, including millions of CpG words across ultra-long genomic sequences. Leveraging Selective State Space Models (SSMs) and a modular, scalable design, Mamba excels in computational efficiency through dynamic computation graphs and advanced memory optimization, minimizing memory usage while maximizing performance even on massive and sparse datasets. A key innovation is the introduction of a novel "CpG fuzzy embedding" approach, where each CpG site's embedding is a weighted sum of methylation and unmethylation motifs, reflecting their respective methylation rates to preserve probabilistic epigenetic information. To further enhance sequence structure, special tokens

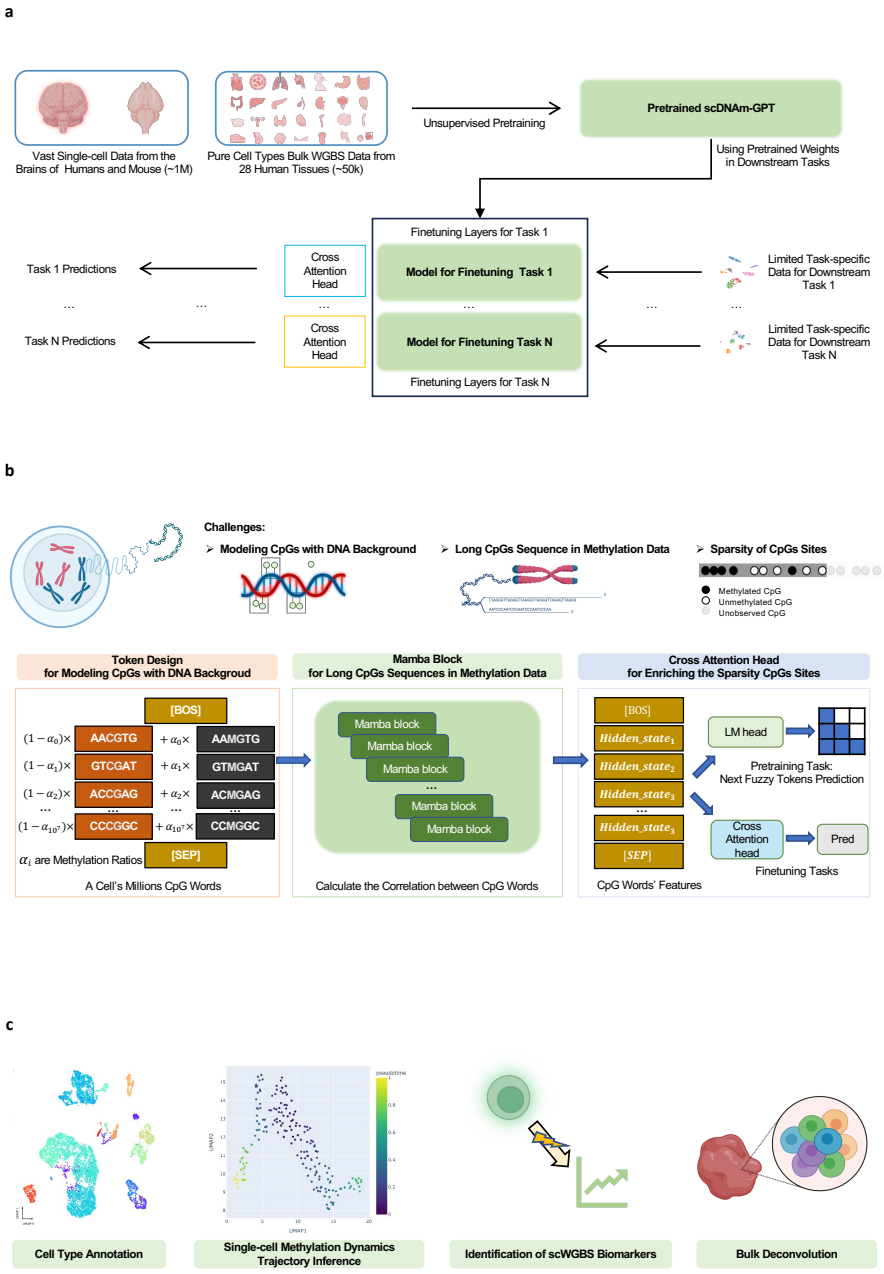

Fig. 1 Schematic of development and applications of scDNAm-GPT. a, The overall pipeline of establishing a foundation model, including an initial self-supervised pretraining on a large-scale single-cell dataset of multiple tissues and the fine-tuning process with limited downstream data. In the fine-tuning procedure, additional layers are added to the model with shared weights to be fine-tuned on task-specific datasets for various downstream tasks. Since the model has learned strong generalization capabilities through initial large-scale pretraining, it can easily adapt to various downstream tasks with a small amount of fine-tuning data. **b,** The architecture and pretraining process of the proposed scDNAm-GPT. We highlight the main components used to deal with the challenges in building a foundation model for scWGBS, including the Mamba blocks for long sequences in methylation data, cross attention mechanism for the important CpG sites, and token design to modeling CpGs with DNA background. **c,** Main applications of scDNAm-GPT: 1) cell type annotation, 2) dynamics trajectory inference, 3) Biological interpretability analysis, and 4) bulk deconvolution.

[BOS] and [SEP] are introduced to mark dataset boundaries, enabling effective management of variable-length sequences. The model's defining feature, Cross-Attention [8], allows the final hidden state of the [SEP] token to dynamically attend to the hidden states of all CpG sites, integrating both local and global dependencies to generate expressive single-cell representations. This innovative design not only captures the biological complexity of DNA methylation patterns but also scales seamlessly for diverse applications, from cell annotation and differentiation trajectory inference to cross-species and cross-tissue studies, making scDNAm-GPT a groundbreaking tool in single-cell epigenomics.

In the pretraining stage, scDNAm-GPT employs a novel **next fuzzy token prediction** strategy, an extension of traditional next token prediction tasks [14] tailored for DNA methylation data. At each genomic position, the model simultaneously predicts both the methylation and unmethylation tokens, with the loss weights dynamically adjusted based on the methylation rate and its complement (1-methylation rate). This design effectively captures the probabilistic nature of CpG methylation, enabling the model to learn nuanced epigenetic patterns from sparse and complex single-cell data. During fine-tuning, the final representation of the `[SEP]` token, obtained through a Cross-Attention head across all input tokens, serves as a comprehensive single-cell embedding. This embedding integrates both local and long-range dependencies between CpG sites, facilitating downstream tasks such as cell annotation, differentiation trajectory inference, and disease state prediction with high precision and biological relevance.

To leverage the scaling law of language models, we have curated a comprehensive cross-species dataset of one million single-cell methylation profiles from humans and mice, encompassing a diverse range of tissues. Our dataset includes methylation profiles from all 28 mouse tissue types and various human tissues, representing one of the most extensive collections for single-cell methylation studies. A total of 900,000 single-cell methylation profiles were collected from publicly available data repositories, including the NCBI Gene Expression Omnibus and Sequence Read Archive [9–11]. Among these, 400,000 profiles are from mouse brain regions, while 500,000 profiles are derived from human brain regions.

To further enhance tissue diversity, we incorporated bulk WGBS data from over 20 different tissues and organs, focusing on purified cell types [12]. This approach allows us to approximate these bulk datasets to scWGBS profiles, contributing approximately 50,000 samples. On average, each single-cell sample yields millions of "CpG words," providing a rich resource for training. These datasets, encompassing methylation profiles from diverse human and mouse tissues, offer a comprehensive foundation for training models to learn tissue-specific methylation patterns and their regulatory roles in gene expression. By incorporating single-cell resolution and cross-species data, our dataset enables the model to capture cellular heterogeneity and identify key regulatory marks across the genome, providing valuable insights into epigenetic regulation and disease mechanisms.

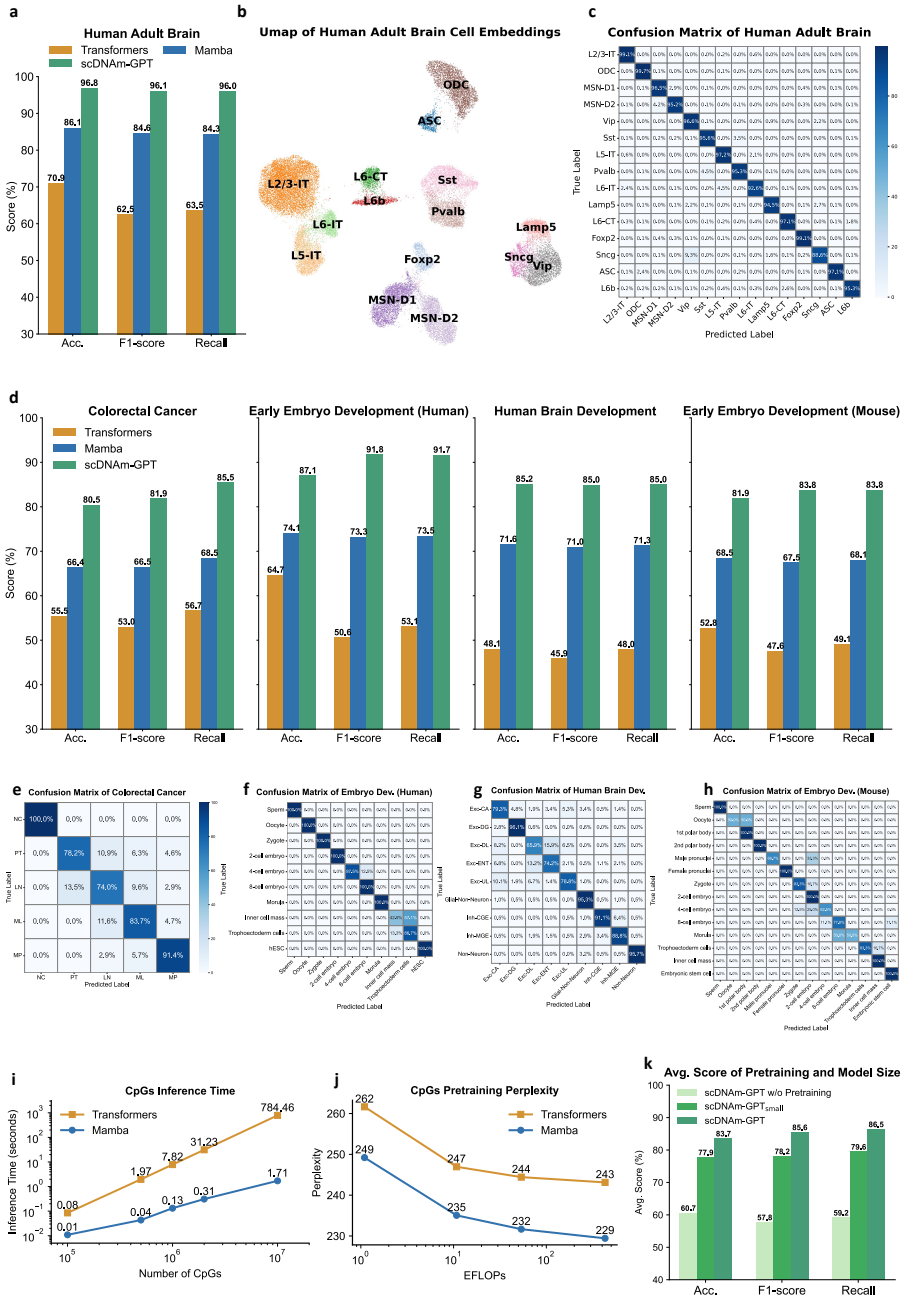

Fig. 2 Benchmarking scDNAm-GPT for efficient, scalable, and accurate modeling of single-cell DNA methylation. a, Performance comparison of Transformer, Mamba, and scDNAm-GPT on adult brain cell classification, evaluated by accuracy, F1-score, and recall. **b,** UMAP visualization of predicted adult brain cell types by scDNAm-GPT. **c,** Confusion matrix on the adult brain test set, showing prediction accuracy across 15 cell types. **d,** Performance comparison across four downstream tasks—colorectal cancer, early embryo development (human and mouse), and human brain development—evaluated by accuracy, precision, and F1-score. scDNAm-GPT consistently outperforms both baselines. **e–h,** Confusion matrices of scDNAm-GPT on the test sets of the four downstream tasks, illustrating classification accuracy and error patterns across cell or tissue types. **i,** Inference time of Transformer and Mamba models as a function of CpG input sequence length. Mamba scales linearly and maintains low latency even for 10^7 CpGs, while Transformer latency increases exponentially. **j,** Relationship between training compute (EFLOPs, 1×10^{18} floating-point operations) and model perplexity on fuzzy next-token prediction over CpG sequences. Mamba consistently achieves lower perplexity than Transformer. **k,** Effect of model depth and pretraining on generalization. Average performance across four unseen datasets for scDNAm-GPT w/o Pretraining, scDNAm-GPT$_{\text{small}}$ (4-layer with pretraining), and scDNAm-GPT (8-layer with pretraining). Both larger model size and pretraining enhance downstream accuracy.

2.2 Benchmarking the efficiency, scalability, and generalizability of scDNAm-GPT

To systematically evaluate the capabilities of scDNAm-GPT, we conducted comprehensive benchmarking across diverse biological contexts, focusing on inference speed, training efficiency, predictive accuracy, and generalization across tissue types. The evaluation utilized five datasets: human adult brain (used for pretraining) [11], and four unseen datasets—human developing brain cells [15], colorectal cancer (CRC) [16], and early embryonic development in both human [17] and mouse [18]. These unseen datasets enabled a rigorous assessment of the model's generalization ability. All models compared in this study, including Transformer, Mamba, and scDNAm-GPT, were pretrained on the same large-scale single-cell DNA methylation dataset (comprising 1 million cells) to ensure a fair comparison.

To benchmark the performance of our model on predicting cell types, our results demonstrate that across all four unseen datasets (Fig. 1d), Mamba consistently achieves about 70% in accuracy, F1-score, and recall—representing an improvement of approximately 15% over Transformers. Although Mamba provides an efficient backbone for modeling long genomic sequences, it lacks a mechanism to selectively focus on biologically informative regions, which are critical for epigenetic interpretation. To overcome this limitation, scDNAm-GPT incorporates a lightweight cross-attention module on top of the Mamba architecture, enabling selective enhancement of key CpG site representations. This architectural refinement achieves over 80% in accuracy, F1-score and recall, representing a substantial performance gain of about 15% over Mamba.

We also evaluate the impact of model depth on performance. Compared to the 4-layer version, the 8-layer model yields an additional 5.4% to 7.8% improvement in three metrics across the four unseen datasets (Fig. 2h and Table A3). These results demonstrate that the Mamba architecture, equipped with a cross-attention mechanism with an 8-layer depth, achieves superior performance (accuracy at 80%) in handling ultra-long sequences and capturing the most informative genomic regions from less than 1,000 cells.

To assess the impact of cell number on the performance of scDNAm-GPT, we systematically evaluated classification accuracy across five datasets: the human early embryo dataset (10 developmental stages; trained on 195 cells, tested on 85), mouse early embryo (14 cell types; trained on 494 cells, tested on 213), CRC (five categories; trained on 895 cells, tested on 384), developing human brain (nine major cell types; trained on 4,096 cells, tested on 1,759), and adult human brain (15 cell types; trained on 116,643 cells, tested on 38,881 cells). The confusion matrices reveal that scDNAm-GPT achieves accuracies of 87.1% (human early embryo), 81.9% (mouse early embryo), 80.47% (CRC), and 85.1% (developing human brain), 96.8% (adult human brain), demonstrating robust performance across datasets with training cell counts ranging from fewer than 300 to over 5,000.

To assess the impact of pretraining, we compared scDNAm-GPT to a randomly initialized variant trained from scratch. The pretrained model achieved substantially better performance, with gains of 23.0% in accuracy, 27.8% in F1-score, and 27.3% in recall, highlighting the benefits of large-scale methylation pretraining (Fig. 2h). We

also evaluated computational efficiency using the pretraining dataset. Mamba showed linear scalability in inference speed, completing inference on over 10 million CpGs in under two seconds, while Transformer latency increased exponentially with sequence length (Fig. 2f). In terms of training efficiency, Mamba consistently achieved lower perplexity than Transformer across all compute budgets (Fig. 2g), supporting its suitability for modeling ultra-long genomic sequences.

In summary, scDNAm-GPT demonstrates robust performance across diverse biological contexts, highlighting its ability to capture meaningful patterns from a wide range of biological systems while maintaining scalable computational efficiency and strong algorithmic performance.

2.3 Trajectory inference for single-cell DNA methylation data using scDNAm-GPT

The inherently asynchronous states of cells at a given time point enable the application of trajectory inference algorithms to reconstruct dynamic cellular transitions during development and disease progression using single-cell data. To evaluate the capability of scDNAm-GPT in inferring cellular trajectories from DNA methylation profiles, we utilized scWGBS datasets encompassing multiple stages of early embryonic development in both human [17] and mouse [18]. The human dataset includes 280 cells spanning developmental stages from the oocyte to the blastocyst. We aggregated the embeddings learned by scDNAm-GPT across all layers and applied diffusion pseudotime analysis to the resulting representations. The model accurately annotated cell stages and successfully reconstructed the developmental trajectory from the zygote to the inner cell mass (ICM) and trophectoderm (TE), in strong concordance with established embryonic progression (Fig. 3a). Similar performance was observed on the mouse dataset (Fig. 3b), further demonstrating the robustness and cross-species generalizability of scDNAm-GPT. Notably, the model was not provided with any temporal annotations; the inferred ordering emerged solely from static DNA methylation profiles following fine-tuning on cell type labels. This highlights the model's ability to capture temporal dynamics from static epigenomic snapshots.

To elucidate how scDNAm-GPT infers dynamic DNA methylation processes, we analyzed the cell embeddings extracted from each layer of the model fine-tuned on the human dataset. We employed UMAP to visualize the cell representations from all eight layers of scDNAm-GPT, thereby illustrating the progression of feature abstraction and the distinct information captured at each layer. In the shallow layers (e.g., layers 1 to 3), only the most pronounced differences were captured—such as the separation of sperm cells from all other cell types—while finer distinctions among more closely related developmental stages were largely omitted, resulting in the mixing of these stages in the embedding space. In contrast, the deeper layers (e.g., layers 4 to 7) captured more nuanced differences, enabling the separation of late-stages (e.g., ICM and TE stages) from early-stages (e.g., Oocyte, Zygote and Morula). At the eighth layer, each stage forms a compact and well-defined cluster, indicating that the deepest layer encodes highly stage-specific DNA methylation signatures. These observations suggest a progressive refinement of developmental trajectories across the model's hierarchical layers.

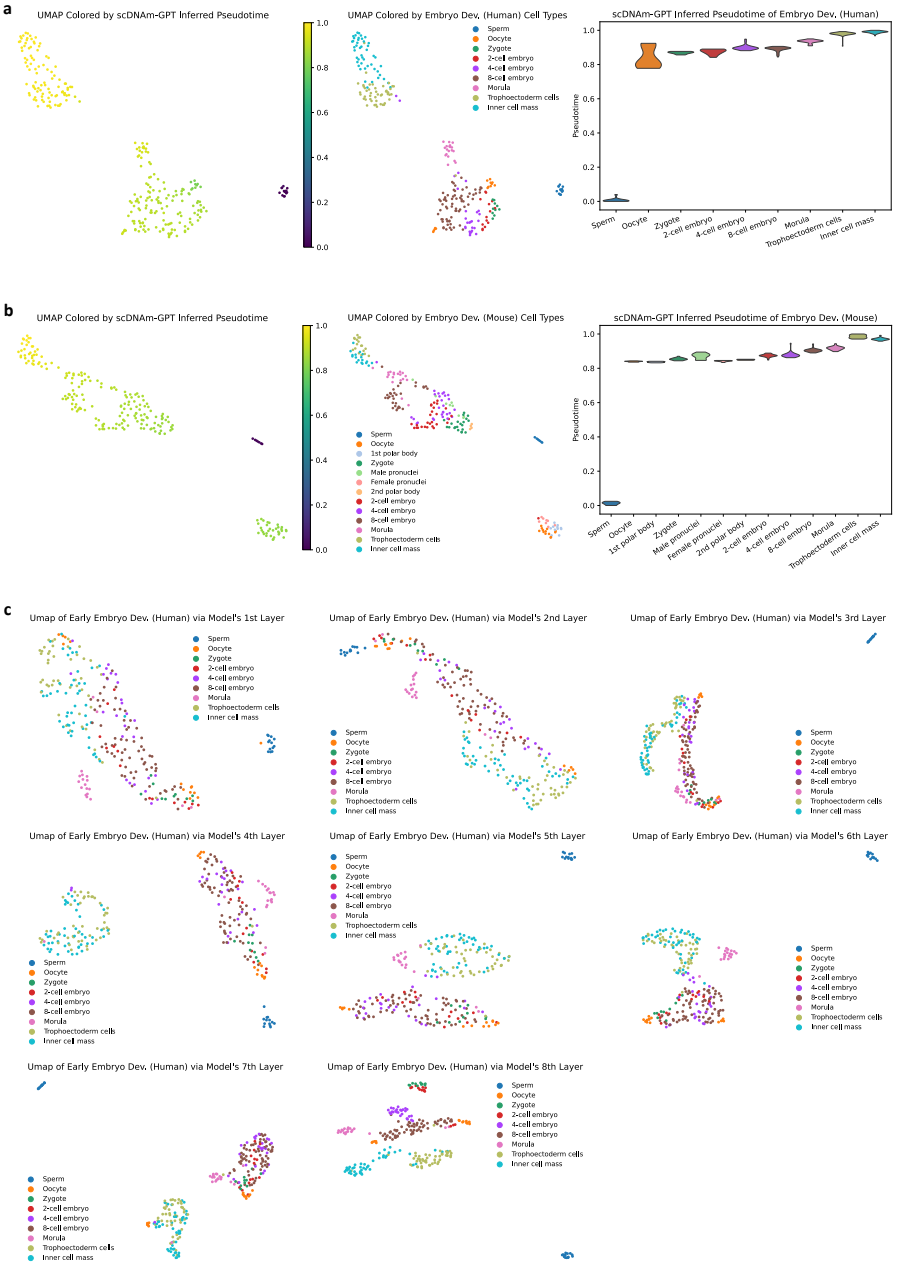

Fig. 3 Representation dynamics and pseudotime inference during early embryo development. a, Pseudotime inference for human early embryo development based on aggregated embeddings across all model layers, projected into 2D and colored by inferred pseudotime, revealing a clear developmental trajectory. **b,** Pseudotime inference for mouse early embryo development using the same embedding aggregation strategy, demonstrating the cross-species generalizability of scDNAm-GPT's representations. **c,** UMAP visualizations of scDNAm-GPT embeddings from different layers for human early embryo development. Different layers capture distinct types of information, with some layers emphasizing cell-type separation (classification) and others reflecting developmental continuity (pseudotime).

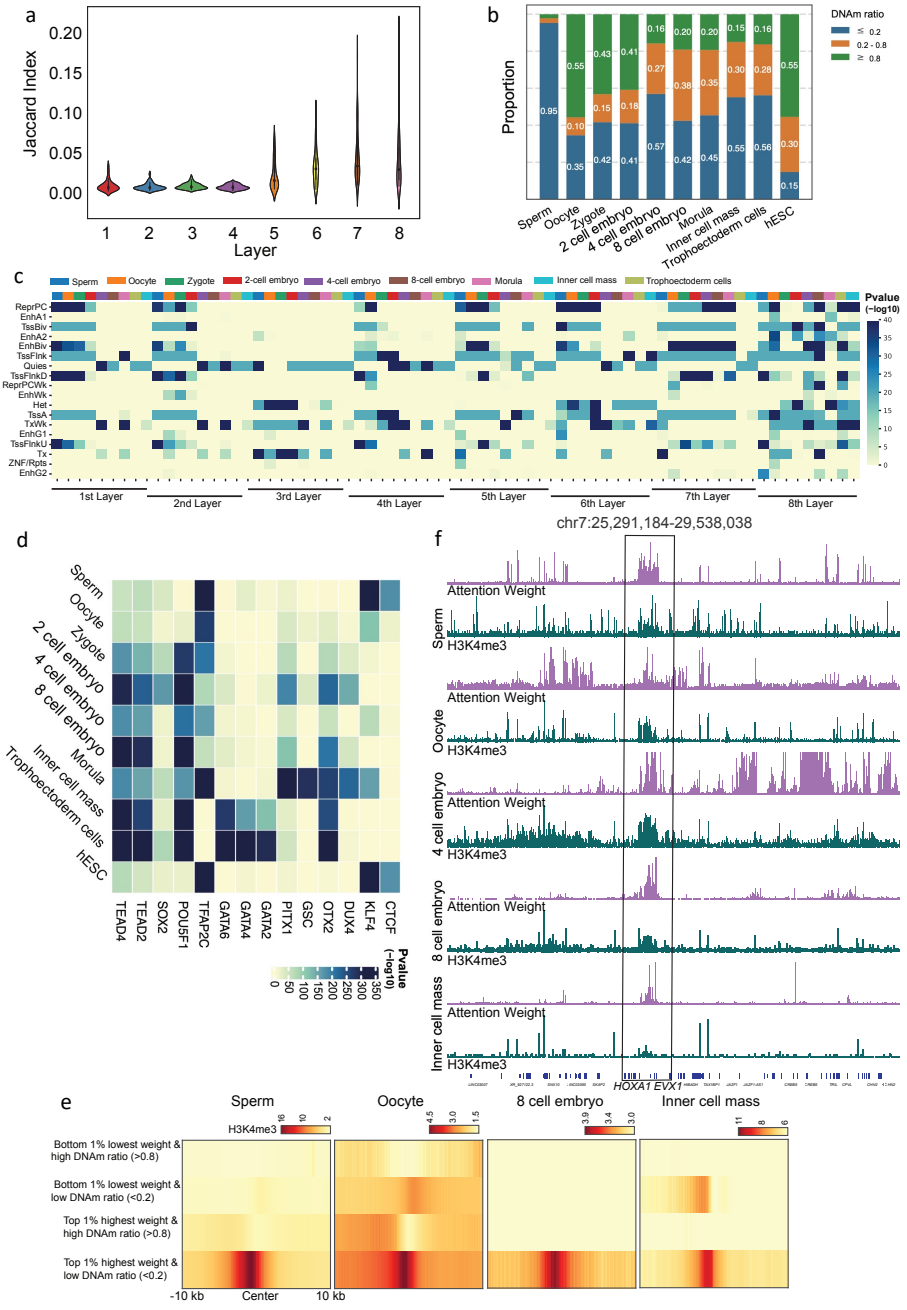

Fig. 4 Hierarchical token-level representations and regulatory relevance of scDNAm-GPT attention. a, Jaccard indices of the top 1% high-attention CpG sites across cell classes at each layer. **b,** Distribution of DNA methylation ratios for the top 1% high-attention CpGs at layer 8, stratified by cell class. **c,** Chromatin state enrichment analysis of the top 1% high-attention CpGs across layers and cell classes. **d,** Enrichment of transcription factor motifs within the top 1% high-attention regions at layer 8. **e,** Heatmap showing enrichment patterns around the top 1% high-attention and bottom 1% low-attention CpG sites, stratified by DNA methylation ratio at layer 8. **f,** IGV browser snapshots showing attention scores and H3K4me3 ChIP-seq signals at the HOXA1 locus.

2.4 The Biological Interpretability of the Context Learned by scDNAm-GPT

To investigate how scDNAm-GPT encodes information at the token level, we analyzed the contextual representations learned across its layers. For each layer, we examined the top 1% of genomic regions with the highest attention scores. Analysis of attention distributions revealed that CpG sites with the top 1% highest attention scores for each class exhibited minimal overlap in the early layers (layers 1–4), with Jaccard indices consistently below 0.05. In contrast, the Jaccard index gradually increased from layers 5 to 8, indicating a transition from class-specific to more generalized representations (Fig. 4a). This hierarchical trend suggests that scDNAm-GPT captures progressively complex features of CpG sites, evolving from low-level, localized patterns in shallow layers to high-level, abstract representations in deeper layers. Moreover, the distribution of DNA methylation ratios among the top 1% high-attention regions in layer 8 was notably diverse across classes (Fig. 4b), highlighting the model's capacity to integrate a broad spectrum of epigenetic information.

To further characterize the contextual learning of the model, we performed chromatin state enrichment analysis on the top 1% attention-scoring regions. Distinct enrichment patterns were observed across layers. Layers 1 to 4 showed significant enrichment in transcription start site (TSS) regions while being depleted for enhancer and repeat-associated regions. In contrast, layers 5 to 8 displayed broader enrichment profiles, including TSSs, enhancers, and repetitive elements. Notably, layer 8 exhibited comprehensive coverage of diverse regulatory elements across different cell types (Fig. 4c). This progressive expansion in chromatin context suggests that scDNAm-GPT learns regulatory features in a hierarchical manner, moving from narrow, localized signals to a diverse and integrated view of the epigenomic landscape.

We further conducted transcription factor (TF) binding motif enrichment analysis using the JASPAR database. This analysis revealed dynamic and stage-specific motif enrichment within the top 1% high-attention regions identified by layer 8 (Fig. 4d). For example, motifs corresponding to inner cell mass (ICM)-specific TFs such as GATA2, GATA4, and GATA6 were significantly enriched in ICM-associated regions. In embryonic stem cell (ES) stages, motifs for CTCF and TFAP2C were prominently enriched. At the 8-cell stage, motifs corresponding to OTX2 and PITX1 were significantly overrepresented. These findings are consistent with previously reported stage-specific TF activity profiles [19], supporting the biological relevance of the model's attention-based prioritization.

To further explore the epigenomic context of the model's high-attention regions, we integrated histone modification data. We observed that regions with top 1% attention scores and low DNA methylation levels (<0.2) were significantly enriched for the active histone mark H3K4me3. In contrast, high-attention regions with high DNA methylation ratios (>0.8) showed marked depletion of H3K4me3 signals. Notably, regions within the lowest 1% attention scores—regardless of methylation status—did not exhibit any significant H3K4me3 enrichment (Fig. 4e). These results suggest that scDNAm-GPT preferentially attends to regulatory regions marked by active chromatin features, particularly within hypomethylated contexts. For example, at the HOXA1 locus, regions with high attention scores aligned with areas of enriched H3K4me3 signal

across various early embryonic stages (Fig. 4f), in agreement with previous findings [20].

Collectively, these analyses demonstrate that scDNAm-GPT not only learns meaningful sequence-level and methylation-level representations but also captures biologically interpretable and hierarchically structured features of the epigenomic landscape.

2.5 Reference-Free Deconvolution of Heterogeneous Bulk DNA Methylation Profiles by scDNAm-GPT

Accurate inference of the cellular composition of bulk DNA methylation samples—such as those derived from cell-free DNA—is a critical task known as cell type deconvolution. Traditional deconvolution methods typically rely on predefined reference profiles for specific cell types, which are often limited by the difficulty of obtaining comprehensive and representative profiles, the lack of generalizability across tissues and conditions, and their inability to capture the extensive heterogeneity within individual cell types. To address these limitations, we propose a reference-free, end-to-end deconvolution strategy leveraging the pre-trained scDNAm-GPT model.

In silico mixtures with known cell type proportions are essential for evaluating deconvolution performance. Previously published methods generated methylation signals for each cell type by sampling from a single randomly selected methylome profile within that cell type [21]. However, this approach overlooks the inherent variability among samples of the same cell type. We simulated bulk DNA methylation data using single-cell DNA methylation data from five different neuronal cell types [11]. For each simulation, we randomly sampled between zero and ten cells from each cell type to generate mixtures that preserved both inter-cell-type and intra-cell-type heterogeneity. To better mimic real bulk sequencing conditions, we maintained variations in sequencing depth and coverage across the mixed single-cell samples. These variations reflect technical noise arising from differences in capture efficiency among cells and uneven sequencing depth across genomic loci.

The cell type proportions predicted by scDNAm-GPT showed high concordance with the true proportions across five distinct neuronal subtypes, achieving an average Pearson correlation coefficient of 0.84. In contrast, other reference-free deconvolution methods—RefFreeCellMix [22], EDec [23], and MeDeCom [24]—were unable to accurately infer cell type compositions (Fig. A3). This discrepancy may be attributed to the subtle DNA methylation differences among the five neuronal subtypes, which are not readily captured by previous approaches primarily designed to distinguish cell types across divergent lineages. These findings highlight the superior capability of scDNAm-GPT in resolving fine-grained cell type compositions from DNA methylation data, particularly within closely related cell subtypes, such as neuronal subtypes.

To further assess the model's ability to detect disease-associated DNA methylation patterns, we fine-tuned scDNAm-GPT on a binary classification task using single-cell whole-genome bisulfite sequencing (scWGBS) data from healthy colorectal cells and primary colorectal tumor cells. The resulting cell embeddings exhibited clearly separated clusters corresponding to normal and tumor cells, indicating that scDNAm-GPT effectively captures tumor-specific DNA methylation signatures. We

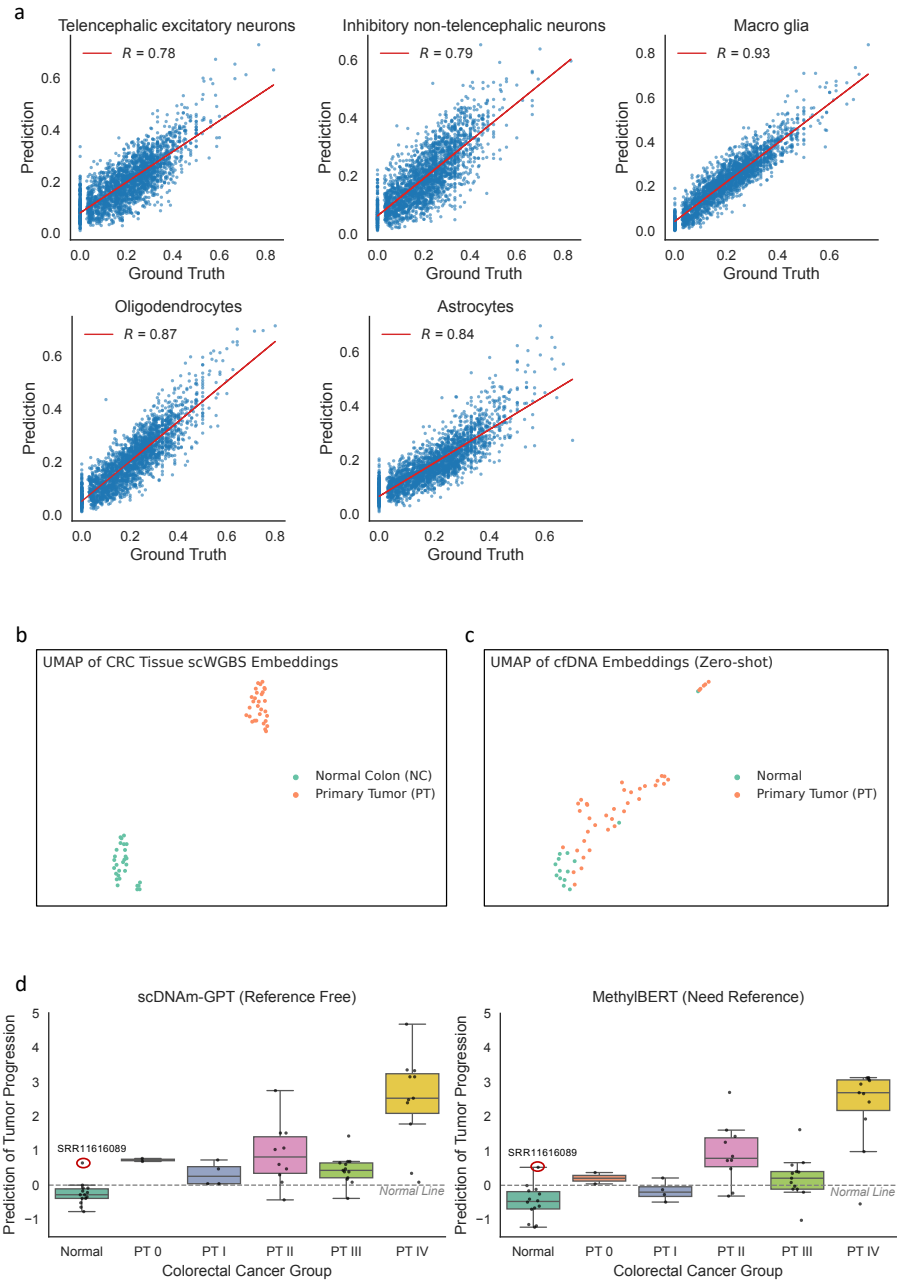

Fig. 5 Reference-free deconvolution and cfDNA-based cancer detection using scDNAm-GPT. a, Predicted vs. true average cell-type proportions in simulated bulk methylation samples across five neuronal types ($R = 0.78–0.93$). **b,** UMAP of cell embeddings after fine-tuning on healthy and tumor scWGBS data shows clear separation. **c,** UMAP of cfDNA samples reveals separation between normal and cancer groups. **d,** Tumor probability scores predicted by scDNAm-GPT better distinguish cancer from normal cfDNA samples than MethylBERT, without using reference profiles.

then conducted a zero-shot evaluation by directly applying the tissue-finetuned model to circulating cell-free DNA (cfDNA) samples derived from peripheral blood, without any additional training on cfDNA data. Remarkably, the model was still able to accurately distinguish cfDNA from healthy individuals and colorectal cancer patients (Fig. 5c), demonstrating strong generalization capability and the ability to robustly capture tumor-specific DNA methylation. features despite the high background noise inherent in cfDNA samples.

We next compared our model with MethylBERT, a state-of-the-art method for cfDNA-based tumor detection. To ensure a fair comparison, MethylBERT was fine-tuned on the same normal and tumor scWGBS data, using manually curated reference regions selected for their relevance to colorectal cancer. The tumor probabilities predicted by scDNAm-GPT effectively distinguished normal cfDNA samples from those at various tumor stages, with the exception of a single outlier sample (SRR11616089). In contrast, MethylBERT failed to clearly separate several cfDNA samples from patients with stage I and stage III tumors (Fig. 5d). These results suggest that scDNAm-GPT outperforms MethylBERT in distinguishing normal from cancer-derived cfDNA. Notably, we introduce a framework that fine-tunes scDNAm-GPT using a limited number of scWGBS samples, without requiring manual feature selection, and achieves improved deconvolution performance with enhanced clinical relevance.

3 Discussion

We introduce scDNAm-GPT, the first large language model specifically designed for single-cell DNA methylation analysis. By leveraging the innovative Mamba architecture and pretraining on over one million single-cell DNA methylation profiles with one million parameters, our model captures complex interactions among CpG sites and their sequence contexts at the single-cell level. This design enables scDNAm-GPTto generate biologically meaningful embeddings that robustly represent epigenetic states, ultimately providing a versatile tool for downstream analyses. A notable feature of our framework is the integration of cross attention mechanisms. Although our model is built upon the Mamba architecture—which is highly efficient in processing high-dimensional, sparse data—it still employs cross attention to effectively fuse information across different representations. This mechanism plays a critical role in aligning and integrating various aspects of the input data (e.g., local CpG site features and broader sequence context), thereby enhancing the model's capability to capture long-range dependencies and subtle epigenetic signals.

Due to the sparsity and high dimensionality of scWGBS data, clustering and cell type prediction remain significant challenges. Traditional methods mitigate sparsity by averaging DNA methylation ratios over 100,000 bp bins; however, this approach also obscures informative DNA methylation signals and leads to inconsistent clustering results. In contrast, scDNAm-GPT achieves a clustering accuracy of 94.4%, effectively generates meaningful clustering results that align with cell types, even for unseen data. Notably, regions with higher attention weights correspond to low DNA methylation ratios and are marked by active histone modifications, underscoring the interpretability of our model. The sizes of regions with higher attention scores range

from hundreds to tens of thousands of base pairs, demonstrating the model's superior resolution in identifying biologically meaningful DNA methylation regions.

Beyond clustering and cell type prediction, scDNAm-GPT demonstrates promising capabilities in reconstructing cell differentiation trajectories. This feature provides novel insights into epigenetic dynamics during development and lineage commitment, positioning our model as a powerful tool for investigating cell fate decisions and the underlying mechanisms of epigenetic regulation. Additionally, preliminary experiments suggest that scDNAm-GPTmay aid in deconvolving cell type compositions from cell-free DNA methylation data—a challenging task when relying on traditional reference-based methods.

Deconvolution using cfDNA methylation data to infer cell type composition, when based on reference data, poses significant limitations. scDNAm-GPT has certain limitations. First, its pretraining data primarily originate from brain samples, which may not sufficiently capture the features of all organs and diseases. Expanding the training dataset to include scWGBS data from diverse organs, diseases, and cell types will be essential for improving its generalizability in the future. Additionally, scDNAm-GPT currently does not integrate other omics data, such as scRNA-seq, scHi-C, or scCUT&Tag, which could enhance its applicability across various tasks. Furthermore, incorporating metadata such as individual patient information, clinical imaging, and physiological parameters (e.g., blood pressure) could significantly enhance scDNAm-GPT's ability to predict clinical status at an individual level.

We envision that incorporating more datasets, including scWGBS data from diverse tissues, disease conditions, and cell types, along with other omics data and metadata, will enhance scDNAm-GPT's performance across various tasks. Additionally, we plan to pretrain scDNAm-GPT on time-series and perturbation data, enabling it to learn causal relationships and infer cell behavior in response to DNA methylation alterations. As a foundation model, scDNAm-GPT has the potential to accelerate the discovery of epigenetic dynamics at the single-cell level, benefiting biomedical research.

4 Methods

4.1 Model

4.1.1 scWGBS Tokenizer

The scWGBS tokenizer is a critical component of scDNAm-GPT, as it transforms raw genomic data into a format suitable for processing by the model. Given that scWGBS data consists of methylation levels at each CpG site across large genomic regions, the tokenizer's primary role is to encode these methylation values into a tokenized sequence that preserves both the spatial and biological characteristics of the input data.

In contrast to traditional NLP tokenization, where words are split into subwords or characters, the scWGBS tokenizer is designed to handle the unique challenges posed by biological data, particularly the representation of DNA sequences and methylation patterns. Each CpG site is treated as a "token," and its methylation state is encoded in a way that retains the crucial epigenetic information.

The core idea behind the scWGBS tokenizer is to treat DNA sequences as strings of "CpG words," where each CpG site is a token. This tokenizer design incorporates both methylated and unmethylated states as part of the token representation, ensuring that methylation patterns are preserved for downstream processing.

To achieve this, each CpG site is encoded as a binary token that represents its methylation state: - A methylated CpG is assigned a token value of 1, - An unmethylated CpG is assigned a token value of 0.

In addition to binary methylation states, we incorporate flanking nucleotide context to account for the surrounding DNA sequence, which can have biological significance in regulating methylation patterns. This allows the tokenizer to generate more informative representations for each CpG site.

Thus, each token is a combination of the methylation state and its local DNA context, forming a sequence of tokens that represent the entire genomic region.

The tokenization process can be broken down into the following steps:

1. **DNA Sequence Splitting**: The genomic data is split into windows of CpG sites, where each window corresponds to a continuous region of the genome (e.g., a 10kb region). Each window contains a sequence of CpG sites, and the surrounding nucleotides are encoded as part of the context.

2. **Encoding Methylation States**: For each CpG site in the window, its methylation state is encoded as a binary value (0 or 1). Additionally, the surrounding nucleotides (e.g., $\pm$ 3 bases around each CpG site) are incorporated to form a context-aware token. This results in a k-mer token representation, where each CpG is accompanied by its local sequence context.

3. **Sequence Assembly**: The tokenized sequences are then assembled into a sequence of tokens, where each token represents a CpG site along with its methylation state and surrounding context. The final sequence represents the entire window of genomic data.

$$\text{Token}_i = (\text{MethylationState}(\mathbf{x}_i), \text{FlankingContext}(\mathbf{x}_i)), \tag{1}$$

where $\mathbf{x}_i$ is the i-th CpG site in the genomic window, and the MethylationState refers to the methylation status (0 or 1), while the FlankingContext represents the surrounding nucleotides.

4. **[BOS] and [SEP] Token Insertion**: A special [BOS] token is inserted at the beginning of each window to signal the start of a new sequence, as well as to serve as the query in the cross-attention mechanism, which aggregates global context from the entire sequence.

Once the tokenization process is complete, the resulting sequence of tokens can be fed into the Mamba Backbone for further processing. The [SEP] token plays a crucial role in generating the query vector for the cross-attention mechanism, as described in Section A.1.3. This integration of tokenized genomic data with the cross-attention mechanism allows scDNAm-GPT to capture both the methylation state of individual CpG sites and their relationships with distant sites across the genome.

For example, consider a window of CpG sites:

$$\text{Genomic Window} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n].$$

The tokenizer processes this window by encoding each CpG site $\mathbf{x}_i$ as a token that includes its methylation state and flanking nucleotide context. The resulting sequence is then represented as:

$$\text{Tokenized Sequence} = [[\text{BOS}], \mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n, [\text{SEP}]].$$

This tokenized sequence is then passed through the Mamba Backbone, where the [SEP] token generates the query vector for the cross-attention mechanism, which attends to each CpG site (represented by the tokens $\mathbf{t}_i$) based on their methylation states and contextual relationships.

The scWGBS tokenizer offers several advantages for modeling DNA methylation data:

1. **Preserving Methylation Information**: By encoding both methylation states and surrounding nucleotide contexts, the tokenizer ensures that key epigenetic information is preserved for downstream processing.

2. **Contextualized Tokenization**: The inclusion of flanking nucleotides in the tokenization process allows the model to learn not only the methylation state of each CpG site but also the surrounding sequence context that may influence methylation patterns.

3. **Scalability**: The tokenizer's ability to efficiently process large genomic windows and generate contextually rich tokens enables scDNAm-GPT to scale to millions of CpG sites, making it suitable for analyzing large-scale genomic data.

The scWGBS tokenizer is specifically designed to capture biologically relevant features of the genome. The methylation status of each CpG site is crucial for understanding gene regulation, and the flanking nucleotide context helps reveal potential regulatory motifs that influence methylation patterns. By integrating these features into the tokenization process, we ensure that the model can learn meaningful biological patterns from the data.

4.1.2 Mamba Backbone

Mamba is a state-of-the-art sequence modeling framework designed for efficient processing of ultra-long sequences. Unlike traditional transformer-based architectures, which suffer from quadratic complexity in self-attention, Mamba leverages Selective State Space Models (SSMs) [7] to achieve linear time complexity while maintaining strong representation capabilities.

The motivation for using Mamba in scDNAm-GPT stems from the inherent challenges of single-cell whole-genome bisulfite sequencing (scWGBS) data. Each cell's methylation profile consists of millions of CpG sites, forming an ultra-long sequence that requires efficient modeling of both local CpG interactions and long-range dependencies. Mamba's ability to capture global dependencies without the prohibitive computational costs of transformers makes it a very natural fit for this problem.

Long genomic sequences present unique challenges in computational biology [25, 26]. Methylation patterns at distant CpG sites often exhibit correlations due to chromatin organization, regulatory regions, and epigenetic inheritance. Traditional deep learning models, including recurrent neural networks (RNNs) and CNNs, struggle with these dependencies due to limited receptive fields or memory constraints.

Mamba addresses these limitations through a continuous-time selective state space model (SSM), which allows for efficient and memory-friendly long-sequence modeling. The core state update equation is:

$$\mathbf{h}_t = \mathbf{A}\mathbf{h}_{t-1} + \mathbf{B}\mathbf{x}_t, \tag{2}$$

where: $\mathbf{h}_t$ is the hidden state at time step t, $\mathbf{A}$ is a learned transition matrix controlling sequence memory, $\mathbf{B}$ is an input transformation matrix, $\mathbf{x}_t$ is the input at time step t.

Unlike transformers, which require explicit attention over all previous positions, Mamba's state-space model implicitly propagates long-range dependencies without explicit the token-by-token interactions, significantly reducing computational overhead.

The Mamba Backbone in scDNAm-GPT consists of a deep stack of **Mamba Blocks**, each designed to refine feature representations at increasing levels of abstraction. Each block operates as follows:

1. **Convolutional Input Projection**: The input sequence $\mathbf{X}$ is first processed by a depth-wise convolutional layer to extract local CpG patterns:

$$\mathbf{X}' = \text{Conv1D}(\mathbf{X}). \tag{3}$$

2. **State-Space Model (SSM) Evolution**: Each token is passed through a continuous-time state space layer, updating the internal sequence representation:

$$\mathbf{H}_t = \mathbf{A}\mathbf{H}_{t-1} + \mathbf{B}\mathbf{X}'_t. \tag{4}$$

3. **Selective Gating Mechanism**: A sigmoid gating function dynamically modulates the influence of new information:

$$\mathbf{G}_t = \sigma(\mathbf{W}_g\mathbf{X}_t), \tag{5}$$

where σ is the sigmoid activation and $\mathbf{W}_g$ is a learned weight matrix.

4. **Linear Projection and Normalization**: The output is projected to a new feature space and normalized:

$$\mathbf{Y}_t = \text{LayerNorm}(\mathbf{W}_y\mathbf{H}_t). \tag{6}$$

Each Mamba Block stacks these transformations, progressively refining long-sequence representations.

Network Configuration and Parameters

The Mamba Backbone in scDNAm-GPT consists of 8 stacked Mamba Blocks, leading to a total parameter count of approximately 1M parameters. This configuration balances computational efficiency with modeling capacity.

The architecture can be expressed as a sequential composition:

$$\mathbf{X} \xrightarrow{\text{Mamba Block 1}} \mathbf{Y}_1 \xrightarrow{\text{Mamba Block 2}} \mathbf{Y}_2 \ldots \xrightarrow{\text{Mamba Block 8}} \mathbf{Y}_8. \tag{7}$$

This deep stacking mechanism enables the model to capture both short-range CpG patterns and long-range regulatory dependencies in scWGBS data. The Mamba Block

computations are optimized using CUDA kernels for efficient parallelization, allowing scDNAm-GPT to process ultra-long sequences efficiently.

4.1.3 Cross Attention Head

Cross-attention is a powerful mechanism that allows a model to focus on relevant parts of the input sequence while processing a different sequence. In traditional attention mechanisms, such as in the Transformer model, a query vector interacts with all keys and values within the same sequence. However, in cross-attention, the query vector is derived from one sequence, and it attends to another sequence's keys and values [8]. This approach is especially useful when the relationships between two different data representations are essential, such as in multi-modal tasks or when integrating sparse information like DNA methylation patterns.

In our scDNAm-GPT framework, we leverage cross-attention to capture long-range dependencies and contextualize CpG site interactions across the entire genome. Specifically, we use a [SEP] token to generate query, key, and value representations (denoted as $\mathbf{Q}, \mathbf{K}, \mathbf{V}$) for the cross-attention mechanism. The [SEP] token's final hidden state serves as the query, while each CpG site's hidden state is used to compute the keys and values.

The mechanism of the employed cross-attention is clarified as follows: Given an input sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, where $\mathbf{x}_i$ represents the hidden state corresponding to the i-th token in the sequence, cross-attention computes the interactions between the query $\mathbf{Q}$ and the keys $\mathbf{K}$, along with their corresponding values $\mathbf{V}$.

The cross-attention operation can be described as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{8}$$

where: $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$ is the query matrix, $\mathbf{K} \in \mathbb{R}^{n \times d_k}$ is the key matrix, $\mathbf{V} \in \mathbb{R}^{n \times d_v}$ is the value matrix, d_k is the dimension of the key vectors.

The softmax function ensures that the attention weights sum to 1, providing a probabilistic distribution over the values $\mathbf{V}$ based on the relevance of the queries to the keys.

In scDNAm-GPT, we utilize cross-attention to capture the intricate dependencies between CpG sites across the entire genome, focusing on biologically meaningful interactions. To achieve this, we employ the [SEP] token in the following way:

1. The [SEP] token's final hidden state is used as the query $\mathbf{Q}$ for cross-attention. This token is specially designed to aggregate global contextual information from the entire sequence.

2. Each CpG site's hidden state, generated from the Mamba Backbone, serves as both the key $\mathbf{K}$ and the value $\mathbf{V}$.

Thus, for each position in the sequence (representing a CpG site), the attention weights are computed between the [SEP] token's query and the CpG site's key. This allows the [SEP] token to attend to the relevant CpG sites and capture their relationships in the final representation.

The cross-attention operation for a single CpG site can be described as:

$$\mathbf{H}_{\text{att}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{9}$$

where: $\mathbf{H}_{\text{att}}$ represents the attention-weighted hidden state for each CpG site after the cross-attention computation, $\mathbf{Q}$ is the hidden state of the [SEP] token, $\mathbf{K}$ and $\mathbf{V}$ are the hidden states of each CpG site.

This attention-weighted hidden state $\mathbf{H}_{\text{att}}$ is then used for further processing in downstream layers, where it contributes to the final genomic sequence representation.

The use of cross-attention in scDNAm-GPT offers several benefits for modeling genomic data:

1. **Capturing Long-Range Dependencies**: Cross-attention allows the [SEP] token to attend to distant CpG sites, which is essential for modeling the long-range interactions that occur in DNA methylation patterns. These long-range dependencies often play crucial roles in gene regulation and epigenetic modifications.

2. **Contextualizing CpG Interactions**: By attending to the entire genomic sequence, the cross-attention mechanism helps the model contextualize the interaction between CpG sites, ensuring that biologically relevant correlations are learned.

3. **Efficient Representation Learning**: The ability of the [SEP] token to serve as a global query ensures that even sparse methylation data can be processed effectively, enabling the model to capture meaningful patterns from incomplete or sparse data.

The cross-attention mechanism is particularly effective in processing large-scale genomic datasets, such as those in scDNAm-GPT, where each input sequence can span millions of CpG sites. By using the [SEP] token to generate the query vector, we avoid the need for expensive pairwise attention calculations across all tokens, enabling the model to efficiently capture global dependencies while maintaining computational feasibility.

In scDNAm-GPT, the cross-attention mechanism is integrated within the architecture to handle ultra-long sequences efficiently. By applying attention selectively through the [SEP] token query, the model can scale to handle large genomic sequences with millions of CpG sites while still extracting biologically meaningful relationships between distant CpG sites.

Code availability

Code used for model training and evaluation is available at https://github.com/yourname/scWGBS-GPT.

Appendix A

A.1 Exploratory Evaluation of CpG Tokenization and Embedding Strategies

To inform the model architecture and CpG token representation choices, we conducted a series of low-cost exploratory experiments using a compact scDNAm-GPT$_{\text{small}}$ version of our model, referred to as scDNAm-GPT$_{\text{small}}$, based on a Mamba backbone.

We first evaluated the impact of pretraining dataset size on cell type annotation accuracy across various CpG k-mer strategies. As shown in Fig. A1a, larger pretraining sets consistently improved accuracy, with the 6-mer configuration demonstrating the most robust and balanced performance among k-mer lengths ranging from 2 to 8. This finding is also consistent with prior work in DNA sequence modeling, such as DNABERT and Nucleotide Transformers, which both adopt 6-mers as the default tokenization unit. To assess the prediction quality of the 6-mer strategy, we visualized the confusion matrix of cell type annotations on the test set (Fig. A1b), which reveals strong concordance between predicted and true labels.

We further compared different strategies for generating cell-level embeddings from CpG-level representations. Heatmaps of cell embeddings produced using average-pooled hidden states (Fig. A1d) versus cross-attention–weighted embeddings (Fig. A1c) show that the latter yields markedly improved cell-type separation, highlighting the biological relevance captured by the attention mechanism. Fig. A1e presents a UMAP visualization of cross-attention–derived embeddings, in which cells cluster distinctly according to their biological identities—further validating the effectiveness of this embedding approach.

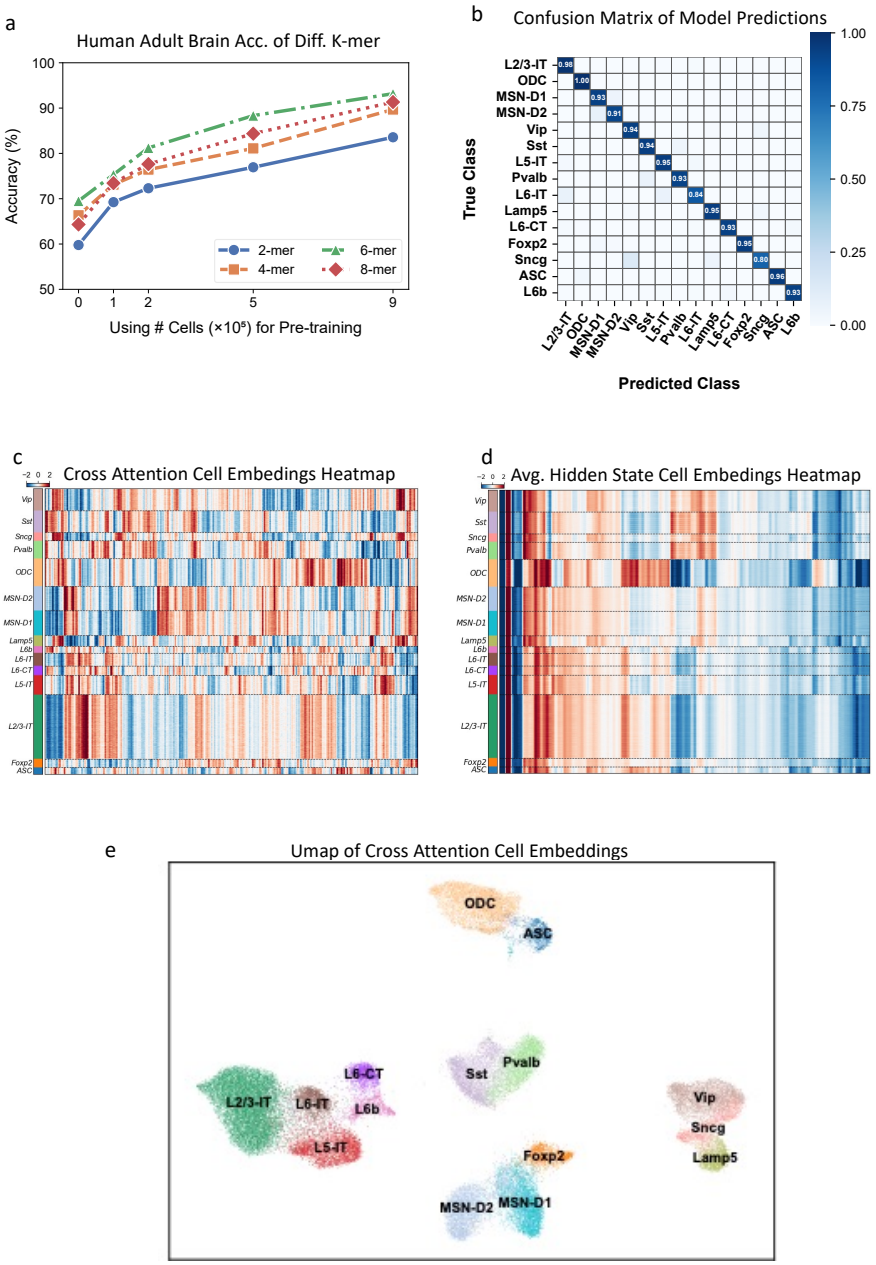

Fig. A1 Evaluation of k-mer strategies and embedding methods in scDNAm-GPT$_{\text{small}}$. **a,** Impact of pretraining dataset size on cell type annotation accuracy. Larger numbers of pretraining cells improve performance across different CpG k-mer strategies. **b,** Confusion matrix of cell type predictions on the test set, demonstrating high agreement between predicted and true labels. **c–d,** Heatmaps of cell embeddings extracted using different strategies. Cross-attention–based embeddings (**c**) yield clearer cell type separation compared to average-pooled hidden states (**d**). **e,** UMAP visualization of cross-attention–based embeddings. Cells form distinct clusters corresponding to their biological identities, highlighting the model's capacity to learn meaningful representations.

Table A1 Benchmark datasets used for model evaluation. Summary of classification datasets spanning diverse biological contexts, including cancer, development, and species differences. Each entry reports the number of annotated cell types, number of training and testing cells, and whether the dataset was included in the pretraining corpus.

Dataset	Cell Types	Training Cells	Testing Cells	In Pretraining
Human Adult Brain	15	116,643	38,881	Yes
Colorectal Cancer	5	895	384	No
Human Early Embryo	10	195	85	No
Mouse Early Embryo	14	494	213	No
Human Brain Development	9	4,096	1,759	No

Table A2 Performance on the Human Adult Brain (seen in the pretraining dataset). scDNAm-GPT is the full pretrained model; w/o PT indicates a randomly initialized version trained from scratch; scDNAm-GPT$_{\text{small}}$ denotes a 4-layer compact variant. This setting evaluates the model's memorization and training capacity on data seen during pretraining.

Metric	scDNAm-GPT	w/o PT	scDNAm-GPT$_{\text{small}}$	w/o PT
Accuracy (%)	**96.80**	68.07	**94.54**	69.89
F1-score (%)	**96.09**	60.92	**93.40**	70.20
Recall (%)	**95.97**	60.59	**93.20**	69.56

Table A3 Performance comparison of scDNAm-GPT and its variants on unseen datasets. We report accuracy, F1-score, and recall across four biological contexts. We compare the scDNAm-GPT (8-layer) and a compact scDNAm-GPT$_{\text{small}}$ (4-layer) with and without pretraining. These results assess generalization to datasets not seen during pretraining.

Dataset	scDNAm-GPT	w/o PT	scDNAm-GPT$_{\text{small}}$	w/o PT
Accuracy (%)				
Colorectal Cancer	80.47	68.49	75.00	61.20
Early Embryo Dev. (Human)	87.06	72.94	75.29	74.12
Human Brain Dev.	85.15	26.31	79.19	30.96
Early Embryo Dev. (Mouse)	81.94	75.00	81.94	75.00
Avg.	**83.66**	60.69	**77.86**	60.32
F1-score (%)				
Colorectal Cancer	81.86	69.56	75.66	62.81
Early Embryo Dev. (Human)	91.77	70.62	75.85	76.41
Human Brain Dev.	84.98	23.73	79.12	28.14
Early Embryo Dev. (Mouse)	83.77	67.35	82.11	65.74
Avg.	**85.59**	57.82	**78.19**	58.28
Recall (%)				
Colorectal Cancer	85.47	69.55	78.30	68.65
Early Embryo Dev. (Human)	91.70	71.69	78.81	77.07
Human Brain Dev.	85.04	26.30	79.09	30.91
Early Embryo Dev. (Mouse)	83.83	69.40	82.10	67.82
Avg.	**86.51**	59.24	**79.58**	61.11

A.2 Comparison of Full and Compact Models With and Without Pretraining

To evaluate the impact of model architecture and pretraining on predictive performance, we systematically compared four variants of scDNAm-GPT: the full model, a compact 4-layer version (scDNAm-GPT$_{\text{small}}$), and their respective counterparts trained from scratch without pretraining. Performance was assessed on a diverse panel of datasets spanning different biological contexts, including cancer, development, and cross-species generalization (Table A1).

As shown in Table A3, pretrained models consistently outperformed their non-pretrained counterparts across all four unseen datasets and three key evaluation metrics—accuracy, F1-score, and recall—underscoring the critical role of large-scale methylation pretraining. In particular, scDNAm-GPT achieved an average accuracy of 83.66% on the unseen datasets, compared to only 60.69% for the same architecture without pretraining. Similarly, the compact scDNAm-GPT$_{\text{small}}$ reached 77.86% accuracy with pretraining, far exceeding its randomly initialized counterpart (60.32%).

Importantly, scaling up model depth from 4 layers to the full scDNAm-GPT architecture led to notable performance gains, especially on complex datasets such as Human Brain Development and Colorectal Cancer. On the brain development dataset, for example, scDNAm-GPT achieved 85.15% accuracy, outperforming scDNAm-GPT$_{\text{small}}$ by over 6 percentage points, and the performance gap widened dramatically in the absence of pretraining (26.31% vs. 30.96%). These results demonstrate that pretraining and model capacity act synergistically to enable accurate and robust representation learning, particularly under the challenge of sparse single-cell methylation input.

To assess whether the model can retain information from the pretraining distribution, we further evaluated performance on the Human Adult Brain dataset, which was included in the pretraining corpus (Table A2). As expected, all models showed improved performance on this seen dataset. scDNAm-GPT achieved 96.8% accuracy, indicating strong memorization and representation retention capabilities, while its non-pretrained variant reached only 68.07%, suggesting that pretraining not only aids generalization but also preserves high-fidelity encoding of training distributions.

Together, these findings highlight two key conclusions: (i) pretraining is essential for learning biologically meaningful patterns from sparse, high-dimensional methylation data, and (ii) increasing model depth substantially enhances predictive performance, especially in biologically heterogeneous or epigenetically complex settings.

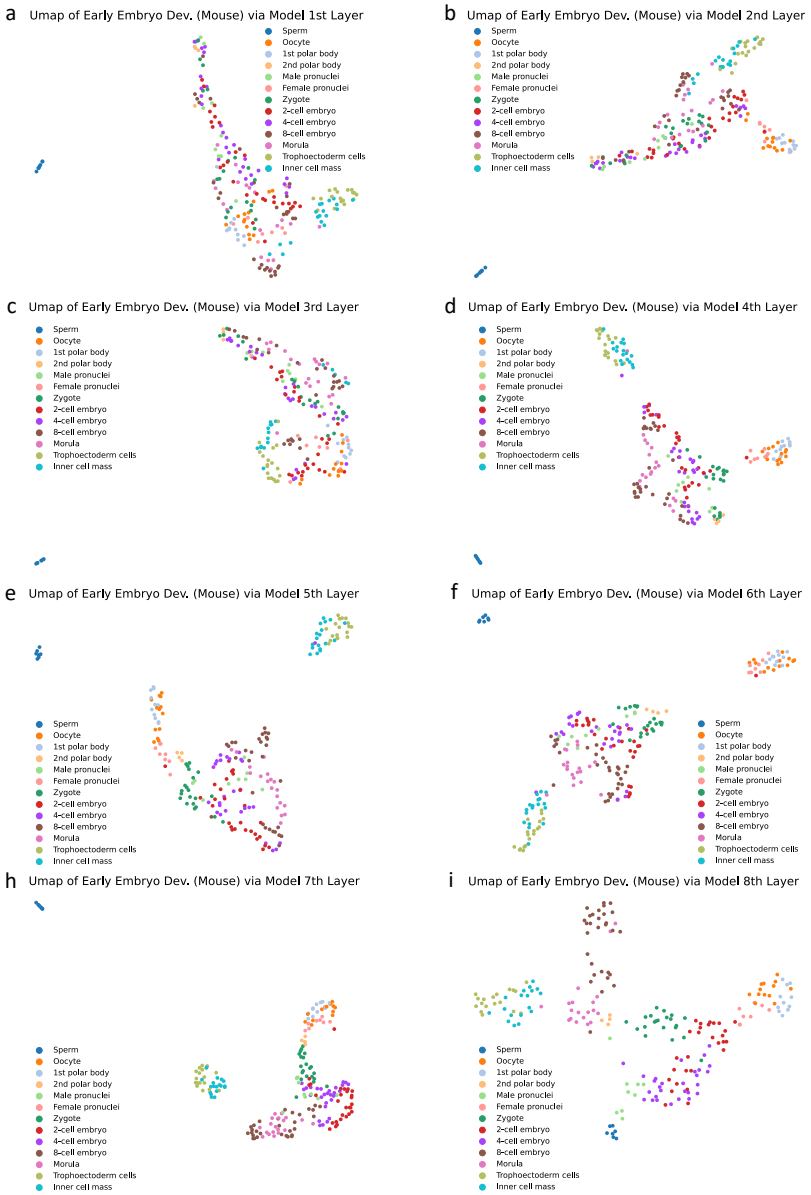

Fig. A2 Layer-wise representations reveal hierarchical encoding of developmental signals. a–i, UMAP visualizations of scDNAm-GPT embeddings from different layers for mouse early embryo development. Different layers capture distinct types of information, with some layers emphasizing cell-type separation (classification) and others reflecting developmental continuity (pseudotime).

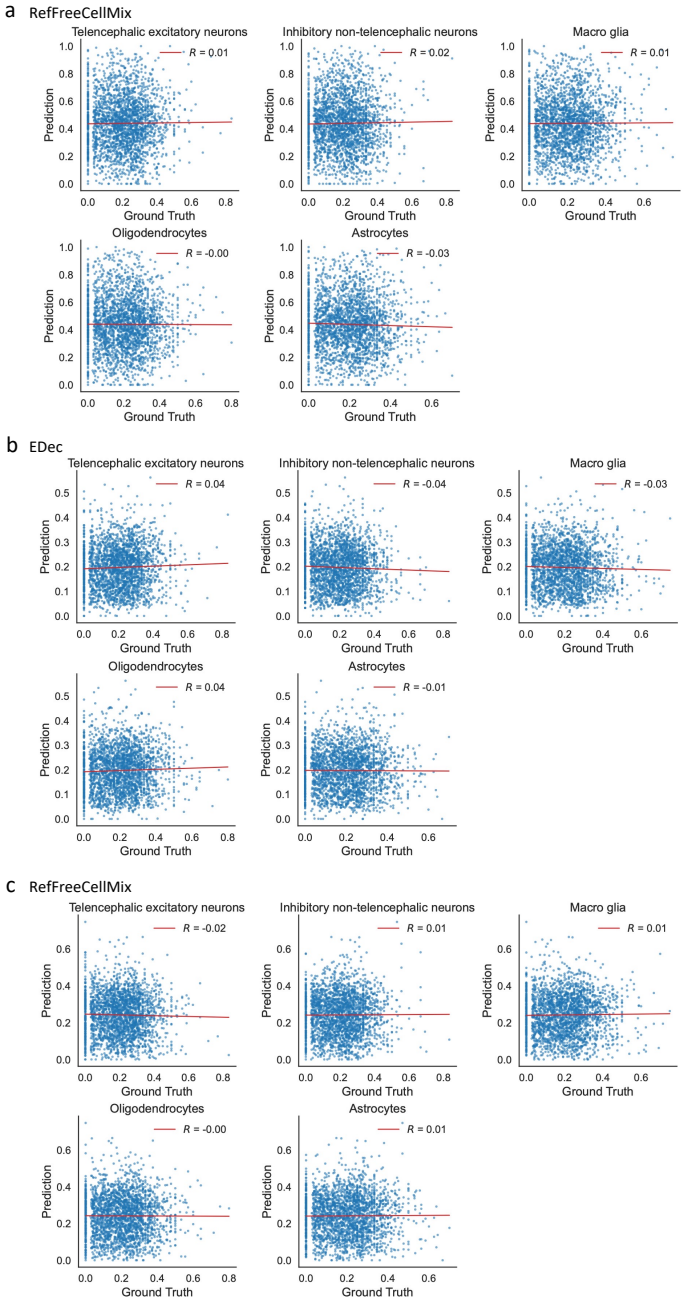

Fig. A3 Comparison of reference-free deconvolution methods on neuronal subtype composition inference. a–c, Estimated cell type proportions from RefFreeCellMix (a), EDec (b), and MeDeCom (c) show poor concordance with true proportions across five neuronal subtypes (average Pearson $-0.05 < r < 0.05$). These methods fail to resolve fine-grained methylation differences among closely related neuronal subtypes, in contrast to scDNAm-GPT, which achieves an average $r = 0.84$ (see Fig. 5d). The results highlight the superior resolution of scDNAm-GPT for fine-scale epigenetic deconvolution.

References

[1] Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., Yao, J.: scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. Nature Machine Intelligence **4**(10), 852–866 (2022)

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

[3] Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., Song, L.: Large-scale foundation model on single-cell transcriptomics. Nature Methods, 1–11 (2024)

[4] Farlik, M., Sheffield, N.C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., Bock, C.: Single-cell dna methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell reports **10**(8), 1386–1397 (2015)

[5] Mulqueen, R.M., Pokholok, D., Norberg, S.J., Torkenczy, K.A., Fields, A.J., Sun, D., Sinnamon, J.R., Shendure, J., Trapnell, C., O'Roak, B.J., *et al.*: Highly scalable generation of dna methylation profiles in single cells. Nature biotechnology **36**(5), 428–431 (2018)

[6] Kremer, L.P., Braun, M.M., Ovchinnikova, S., Küchenhoff, L., Cerrizuela, S., Martin-Villalba, A., Anders, S.: Analyzing single-cell bisulfite sequencing data with methscan. Nature Methods **21**(9), 1616–1623 (2024)

[7] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)

[8] Gheini, M., Ren, X., May, J.: Cross-attention is all you need: Adapting pretrained transformers for machine translation. arXiv preprint arXiv:2104.08771 (2021)

[9] Liu, H., Zhou, J., Tian, W., Luo, C., Bartlett, A., Aldridge, A., Lucero, J., Osteen, J.K., Nery, J.R., Chen, H., *et al.*: Dna methylation atlas of the mouse brain at single-cell resolution. Nature **598**(7879), 120–128 (2021)

[10] Liu, H., Zeng, Q., Zhou, J., Bartlett, A., Wang, B., Berube, P., Tian, W., Kenworthy, M., Altshul, J., Nery, J., et al.: Single-cell dna methylome and 3d multi-omic atlas of the adult mouse brain. biorχiv (2022)

[11] 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), 5357 (2023)

[12] 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)

[13] Zhou, W., Dinh, H.Q., Ramjan, Z., Weisenberger, D.J., Nicolet, C.M., Shen, H., Laird, P.W., Berman, B.P.: Dna methylation loss in late-replicating domains is linked to mitotic cell division. Nature genetics **50**(4), 591–602 (2018)

[14] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)

[15] Heffel, M.G., Zhou, J., Zhang, Y., Lee, D.-S., Hou, K., Pastor-Alonso, O., Abuhanna, K.D., Galasso, J., Kern, C., Tai, C.-Y., *et al.*: Temporally distinct 3d multi-omic dynamics in the developing human brain. Nature **635**(8038), 481–489 (2024)

[16] Bian, S., Hou, Y., Zhou, X., Li, X., Yong, J., Wang, Y., Wang, W., Yan, J., Hu, B., Guo, H., *et al.*: Single-cell multiomics sequencing and analyses of human colorectal cancer. Science **362**(6418), 1060–1063 (2018)

[17] Li, L., Guo, F., Gao, Y., Ren, Y., Yuan, P., Yan, L., Li, R., Lian, Y., Li, J., Hu, B., *et al.*: Single-cell multi-omics sequencing of human early embryos. Nature cell biology **20**(7), 847–858 (2018)

[18] Guo, F., Li, L., Li, J., Wu, X., Hu, B., Zhu, P., Wen, L., Tang, F.: Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. Cell research **27**(8), 967–988 (2017)

[19] Wu, J., Xu, J., Liu, B., Yao, G., Wang, P., Lin, Z., Huang, B., Wang, X., Li, T., Shi, S., *et al.*: Chromatin analysis in human early development reveals epigenetic transition during zga. Nature **557**(7704), 256–260 (2018)

[20] Xia, W., Xu, J., Yu, G., Yao, G., Xu, K., Ma, X., Zhang, N., Liu, B., Li, T., Lin, Z., *et al.*: Resetting histone modifications during human parental-to-zygotic transition. Science **365**(6451), 353–360 (2019)

[21] De Ridder, K., Che, H., Leroy, K., Thienpont, B.: Benchmarking of methods for dna methylome deconvolution. Nature Communications **15**(1), 4134 (2024)

[22] Houseman, E.A., Kile, M.L., Christiani, D.C., Ince, T.A., Kelsey, K.T., Marsit, C.J.: Reference-free deconvolution of dna methylation data and mediation by cell composition effects. BMC bioinformatics **17**, 1–15 (2016)

[23] Onuchic, V., Hartmaier, R.J., Boone, D.N., Samuels, M.L., Patel, R.Y., White, W.M., Garovic, V.D., Oesterreich, S., Roth, M.E., Lee, A.V., *et al.*: Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types. Cell reports **17**(8), 2075–2086 (2016)

[24] Lutsik, P., Slawski, M., Gasparoni, G., Vedeneev, N., Hein, M., Walter, J.:

Medecom: discovery and quantification of latent components of heterogeneous methylomes. Genome biology **18**, 1–20 (2017)

[25] Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., et al.: Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. Advances in neural information processing systems **36** (2024)

[26] Fishman, V., Kuratov, Y., Petrov, M., Shmelev, A., Shepelin, D., Chekanov, N., Kardymon, O., Burtsev, M.: Gena-lm: A family of open-source foundational models for long dna sequences. bioRxiv, 2023–06 (2023)