MethylProphet: A Generalized Gene-Contextual Model for Inferring Whole-Genome DNA Methylation Landscape

Xiaoke Huang [1] **Qi Liu** [2] **Yifei Zhao** [2] **Xianfeng Tang** [3] **Yuyin Zhou** ✉ [1] **Wenpin Hou** ✉ [2]

Abstract

DNA methylation (DNAm), an epigenetic modification, regulates gene expression, influences phenotypes, and encodes inheritable information, making it critical for disease diagnosis, treatment, and prevention. While human genome contains approximately *28 million* CpG sites where DNAm can be measured, only 1–3% of these sites are typically available in most datasets due to complex experimental protocols and high costs, hindering insights from DNAm data. Leveraging the relationship between gene expression and DNAm offers promise for computational inference, but existing statistical, machine learning, and masking-based generative Transformers face critical limitations: they cannot infer DNAm at unmeasured CpGs or in new samples effectively. To overcome these challenges, we introduce MethylProphet, a gene-guided, context-aware Transformer model designed for DNAm inference. MethylProphet employs a Bottleneck MLP for efficient gene profile compression and a specialized DNA sequence tokenizer, integrating global gene expression patterns with local CpG context through a Transformer encoder architecture. Trained on whole-genome bisulfite sequencing data from ENCODE (1.6B training CpG-sample pairs; 322B tokens), MethylProphet demonstrates strong performance in hold-out evaluations, effectively inferring DNAm for unmeasured CpGs and new samples. In addition, its application to 10842 pairs of gene expression and DNAm samples at TCGA chromosome 1 (450M training CpG-sample pairs; 91B tokens) highlights its potential to facilitate pan-cancer DNAm landscape inference, offering a powerful tool for advancing epigenetic research and precision medicine. All codes, data, protocols, and models are publicly available via https://github.com/xk-huang/methylprophet/.

[1]UC Santa Cruz [2]Columbia University [3]Amazon. Correspondence to: Yuyin Zhou <yzhou284@ucsc.edu>, Wenpin Hou <wh2526@cumc.columbia.edu>.

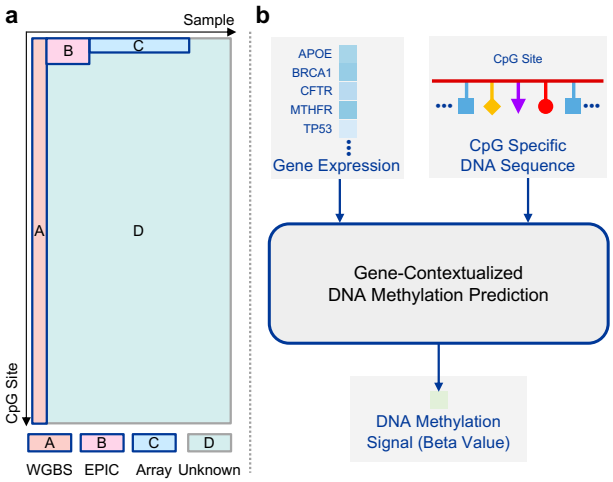

Figure 1: (a) Illustration of the scale of DNAm data. Part A, B, and C refers to existing DNAm samples. Part D refers to unmeasured CpG sites and new samples with gene expression measurements that users can apply MethylProphet to reliably predict their DNAm profiles. (b) Given gene context of a sample, MethylProphet aims to infer whole-genome DNAm at individual CpG resolution.

1. Introduction

DNA methylation (DNAm) is a key epigenetic modification that regulates gene expression, cell differentiation, and disease development (Feinberg, 2018; Loyfer et al., 2023). DNAm predominantly occurs at CpG (cytosine-phosphate-guanine) sites on the DNA sequence, whose tissue-specific and dynamic nature makes them valuable biomarkers (Hitz et al., 2023; The ENCODE Project Consortium, 2012; Luo et al., 2020; The Cancer Genome Atlas Research Network, 2008). However, DNAm profiling is limited by sparse coverage. Array-based technologies measure only 1–3% of CpG sites, while whole-genome bisulfite sequencing (WGBS) is too costly for large-scale studies (Shu et al., 2020). As a result, most CpG sites remain unmeasured in typical datasets (Figure 1 (a)). In contrast, large-scale gene expression profiling has enabled foundation models like Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2024a), scBERT (Yang et al., 2022), and scFoundation (Hao et al.,

Table 1: The scale of DNAm data included in this study. Columns include data source, sequencing techniques, CpG sites, tissue or cancer types, samples, CpG-sample pairs, and CpG-sample pairs with DNAm levels.

Data Source	Sequen.	# CpG Sites	# Tissues / Cancers	# Samples	# CpG-Sample Pairs	# Pairs w/ DNAm
ENCODE	WGBS	27,078,450	57	95	2,572,452,750	2,572,452,750
TCGA	Array	408,399	33	9,194	3,754,820,406	3,684,770,086
	EPIC	740,296	4	1,706	1,262,944,976	1,188,102,524
	WGBS	23,047,052	17	32	737,505,664	737,505,664

2024a). Given the link between gene expression and DNAm, leveraging this relationship offers a promising solution for genome-wide DNAm prediction and analysis.

Several models have been developed for DNAm prediction. DeepCpG (Angermueller et al., 2017) adopts an imputation-based approach, predicting binary DNAm states using local DNA sequences and neighboring methylation patterns. Other methods, including MuLan-Methyl (Zeng et al., 2022), StableDNAm (Zhuo et al., 2023), and MethylNet (Levy et al., 2020), focus on predicting DNAm levels but are limited to specific CpG subsets and sample distributions. Recent foundation models, such as CpGPT (De Lima Camillo et al., 2024) and MethylGPT (Ying et al., 2024), employ a self-supervised learning strategy similar to Masked Language Modeling (MLM) to learn global sample representations for downstream tasks. However, these models struggle to predict DNAm at unmeasured CpG sites and fail to generalize to new samples, restricting their applicability in reconstructing genome-wide DNAm landscapes.

To address this, we introduce MethylProphet, a gene-contextual Transformer model that integrates gene expression with CpG-specific DNA sequence through a novel encoding architecture (Figure 1(b)). Unlike prior DNAm models, which rely on sample-specific patterns, MethylProphet learns robust gene expression representations to infer comprehensive DNAm landscapes with high accuracy and generalizability through the following two key designs.

First, we enhance generalization to new samples by learning expressive representations of sample-specific gene expression profiles. To achieve this, we develop an efficient gene profile compression strategy that effectively captures gene expression levels. To enable gene-contextualized DNAm prediction, it is crucial to incorporate comprehensive gene profiles. However, existing Transformer-based gene encoding methods (Cui et al., 2024b; Bai et al., 2024; Hao et al., 2024b) typically include only a few thousand genes. In contrast, we found that approximately 25K genes, encompassing both protein-coding and non-protein-coding genes, contribute valuable information to this prediction task. To address this challenge, MethylProphet employs an efficient MLP that compresses long-range gene profiles into a compact embedding, enabling sample-specific guidance even for previously unseen samples.

Second, we represent each CpG site by encoding its surrounding DNA sequence context. By employing a DNA tokenizer, the model captures CpG-specific DNA sequence patterns, which not only reduces computational complexity but enables accurate predictions at novel CpG sites. This is achieved by learning from patterns observed across all training CpG sites, allowing the model to generalize effectively. Consequently, this approach addresses a critical limitation of existing models, which are constrained to making inferences only at trained CpG sites and fail to generalize to previously unseen sites.

By integrating global gene expression patterns and CpG-specific DNA sequence context within a Transformer encoder, MethylProphet achieves robust DNAm prediction at unmeasured CpG sites and generalizes effectively to new samples. Besides, we additionally encode CpG Island (CGI) index and region types, and chromosome locations to enhance CpG-specific genomic information.

To systematically evaluate MethylProphet and establish benchmarks for future research, we compile and process two *billion-scale* methylation datasets from ENCODE and TCGA (Table 1): (1) ENCODE whole-genome bisulfite sequencing (WGBS) data, comprising 1.6B training CpG-sample pairs (322B tokens), and (2) TCGA chromosome 1 DNAm data, containing 450M training CpG-sample pairs (91B tokens). Our model achieves strong performance across multiple evaluation scenarios, including prediction at unmeasured CpG sites and in trained samples, with Median of Across-Sample Pearson Correlation Coefficient (MAS-PCC) reaching 0.72 for ENCODE and maintaining robust performance across different prediction scenarios in TCGA.

Our contributions are summarized as follows:

• We develop a novel encoding scheme that uniquely integrates gene expression profiles with DNA sequence context to predict DNA methylation, overcoming the limitations of existing methods in inferring methylation at unmeasured CpG sites and generalizing to new samples.

• We introduce a encoding framework that combines an efficient Bottleneck MLP for gene profile compression with

a specialized DNA sequence tokenizer, enabling the processing of large-scale DNAm data (322B tokens from ENCODE, 91B tokens from TCGA) with efficiency.

• We demonstrate strong generalization performance across unmeasured CpG sites and unseen samples, achieving a PCC of 0.72 on ENCODE data while maintaining robust performance across diverse prediction scenarios in TCGA.

• We release comprehensive datasets, models, and protocols to support reproducibility and further advancements in DNA methylation research.

2. Related Works

Previous computational methods (Zeng et al., 2022; Angermueller et al., 2017; Zhuo et al., 2023; Levy et al., 2020; Wang et al., 2024; Levy-Jurgenson et al., 2019) have primarily focused on predicting DNAm levels at a limited subset of CpG sites and samples.

While recent Transformer-based models (De Lima Camillo et al., 2024; Ying et al., 2024) have demonstrated the ability to samples with approximately 10K CpGs, this represents only 0.03% of the 28M CpGs in whole genome. These models adopt masked modeling objectives, similar to those used in large language models (Achiam et al., 2023; Brown et al., 2020; Radford et al., 2019; Touvron et al., 2023), for pre-training. However, since their primary goal is to learn a global representation of the methylome for downstream tasks, this approach inherently limits their ability to predict for unmeasured CpG sites or for new samples.

Moreover, existing methods either fail to incorporate critical biological context, such as gene expression, or face computational constraints when processing genome-wide DNAm patterns. To address these limitations, we propose a comprehensive approach that leverages gene expression profiles to predict complete DNA methylomes, enabling generalization to unmeasured CpG sites and unseen samples.

3. Characteristics of DNAm Data

DNA methylation (DNAm) data is represented as a matrix where rows correspond to CpG sites, columns represent samples, and entries are β values. Two fundamental characteristics define the complexity of DNAm data analysis:

First, the human genome contains approximately 28 million CpG sites where methylation can occur. Current methylation profiling technologies offer different coverage-cost trade-offs (Shu et al., 2020). Array-based platforms (Illumina 450K and EPIC BeadChips) provide cost-effective but limited coverage (1-3% of whole-genome CpG sites), while Whole Genome Bisulfite Sequencing (WGBS) offers comprehensive coverage but at substantially higher cost. This high dimensionality poses significant challenges for computational modeling, particularly for approaches that attempt to directly represent each CpG site as an embedding. Such approaches not only require prohibitive memory resources (approximately 86GB for float32 embeddings with dimension 768) but also fail to generalize to unmeasured CpG sites. Figure 1 (a) and Table 1 illustrates the data scale of different for different sequencing techniques.

Second, effective methylation prediction requires integration of gene expression data as biological context. However, the high dimensionality of gene expression profiles (approximately 2×10^4 genes) presents its own challenges. Current Transformer-based gene encoding models are typically limited to processing around 1,000 genes due to the quadratic computational complexity of attention mechanisms. This limitation necessitates the development of computational approaches that are both memory-efficient and capable of handling long-range dependencies across the full gene expression profile.

These characteristics demand novel architectural solutions that can effectively (1) *represent and generalize across millions of CpG sites*, and (2) *efficiently process comprehensive gene expression profiles while maintaining computational tractability*.

4. MethylProphet

In this section, we introduce MethylProphet, a gene-contextual model that is capable of inferring the whole-genome DNAm landscape by integrating gene expression with CpG-specific sequence. To encode long-range gene profiles, we utilize a sample gene MLP Embedder (Bachmann et al., 2023) for efficient gene profile compression. Additionally, a DNA tokenizer (Zhou et al., 2024) is employed to capture CpG-specific DNA sequences by encoding their surrounding sequence context. These representations are then integrated with global gene expression patterns and supplementary genomic information, including CpG location, region type, and chromosomal context. A standard Transformer model (Vaswani, 2017) then fuses these embeddings into a unified representation, which is ultimately mapped to the dimension of N_{embed} before they are fed into the Transformer model.

4.1. Encoding

Gene expression To enable sample-specific DNAm prediction, we develop an efficient gene profile compression strategy that captures gene expression levels. Previous approaches that rely on large-scale single-cell datasets for Transformer-based gene encoding (Cui et al., 2024b; Bai et al., 2024; Hao et al., 2024b) is infeasible as only limited (around 1K) genes can be processed. In gene-contextualized

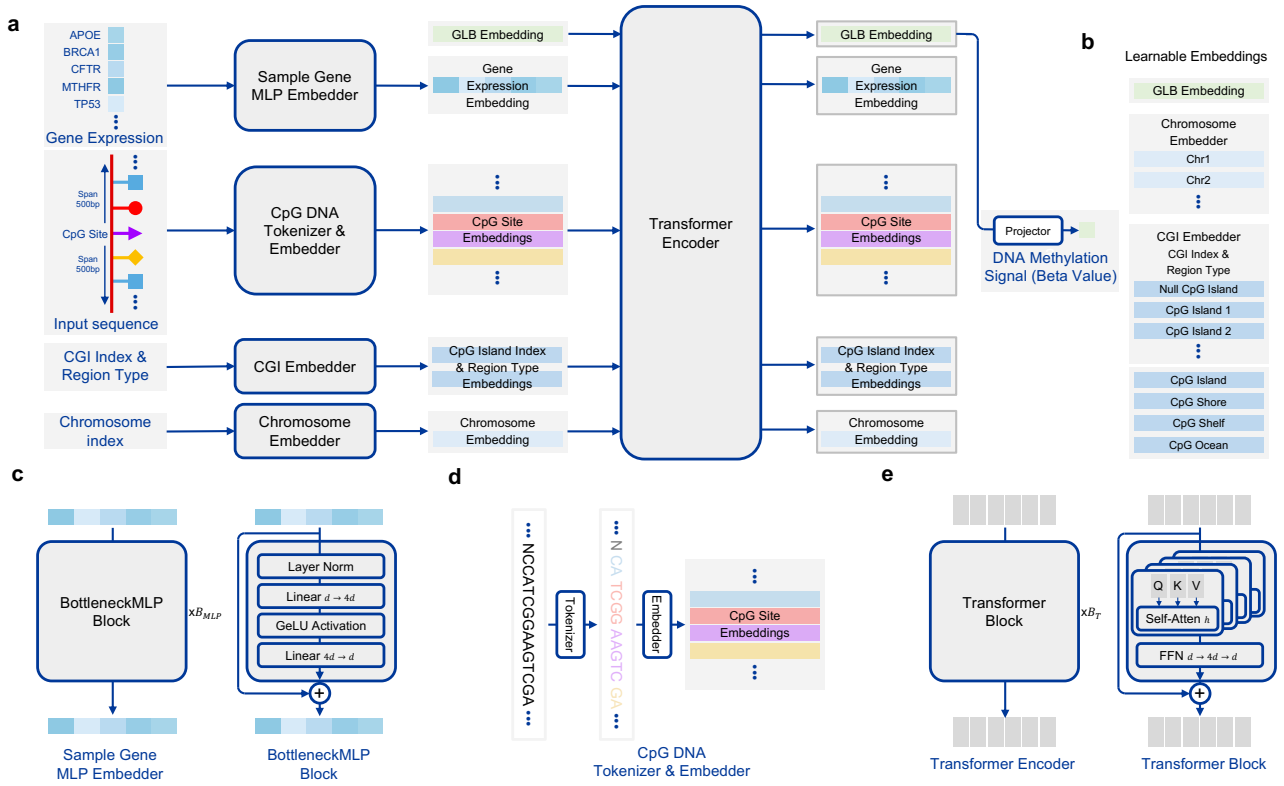

Figure 2: **Overview of our proposed pipeline.** (a) Model architecture of MethylProphet; (b) The learnable Global, chromosome, and CPG island-related embeddings; (c) Model architecture of efficient gene profile compression MLP; (d) DNA Tokenizer for CpG-specific DNA sequence; (e) Model architecture of the Transformer encoder that aggregates all the embeddings.

DNAm prediction, we would like to encode long-range of gene profiles (approximately 25K). Our method addresses the challenge of long-range through a Bottleneck MLP architecture (Bachmann et al., 2023). This design choice offers several advantages: 1) It efficiently compresses the high-dimensional gene expression vector $g_{\text{gene}} \in \mathbb{R}^{N_{\text{gene}}}$ into a compact embedding $x_{\text{gene}} \in \mathbb{R}^{N_{\text{embed}}}$; 2) It maintains model flexibility through minimal inductive bias; 3) It enables robust generalization despite limited training samples; 4) It accommodates long-range gene sets across different experimental contexts. The gene embedding pipeline is illustrated in Figure 2 (a,c).

CpG site To efficiently represent CpG sites, we use a DNA sequence tokenizer that captures the CpG-specific local genomic context. Our approach utilizes the DNABERT-2 (Zhou et al., 2024) variable-length tokenizer, which provides several advantages: 1) It employs Byte Pair Encoding (Gage, 1994) to identify and compress recurring DNA motifs; 2) It reduces sequence redundancy while maintaining biological relevance; 3) It achieves a 5× compression ratio (from 1000bp to approximately 200 tokens); 4) Similar DNA sequence patterns can be shared across CpG sites.

Given a 1 kilobase (1Kb) DNA sequence $S = \{s_i\}$, where $i = 1, 2, \ldots, 1000$ and $s_i \in \{\text{A, T, C, G}\}$, we generate approximately 200 tokens $T = \{t_j\}$ using a DNA tokenizer. These tokens are mapped to embeddings $X^{\text{nbase}} = \{x_j^{\text{nbase}}\}$, where $j = 1, 2, \ldots, 200$ and $x_j^{\text{nbase}} \in \mathbb{R}^{N_{\text{embed}}}$. This process is shown in Figure 2 (a,d).

CpG Island (CGI) and region type Although CpG-related DNA sequence can provide the majority of information to differentiate each CpG site, the repetivive or similar sequence may still introduce ambiguity. We further encodes the CpG island and region type information in the hope to alleviate the above problem. A set of optimizable CpG island index embeddings $\{x_i^{\text{Island_idx}}\}, i = 0, 1, 2, \ldots, N_{\text{island}}$ are initialized for each CpG island. Note that index $i = 0$ is used for represent CpG sites which are in CpG ocean region. Besides, we prepare additional CpG region type embeddings $\{x_{\text{region}}^{\text{island}}, x_{\text{shore}}^{\text{region}}, x_{\text{shelf}}^{\text{region}}, x_{\text{ocean}}^{\text{region}}\}$ to encode CpG island, shore, shelf, and ocean regions. For each CpG site, we add up its CpG island index embedding and CpG region type embedding to construct the CpG island embedding x_{island}. The illustration of GGI embedder is shown in Figure 2 (a,b).

Table 2: The data statistics among all the data source and splits in our experiments. The number of tokens is estimated by the average sequence length (*i.e.*, 200) of the input embeddings of the Transformer encoder.

Dataset	Chr.	Sequen.	Split	# CpG Sites	# Tissues	# Samples	# Pairs w/ Me.	# Tokens
ENCODE	1 - 22	WGBS	Train: Train CpG - Train Sample	24,363,170	57	66	1,607,969,220	321,593,844,000
			Val: Train CpG - Val Sample	24,363,170	22	29	706,531,930	141,306,386,000
			Val: Val CpG - Train Sample	2,707,033	57	66	178,664,178	35,732,835,600
			Val: Val CpG - Val Sample	2,707,033	22	29	78,503,957	15,700,791,400
TCGA	1	Array	Train: Train CpG - Train Sample	33,885	33	8,258	275,018,849	55,003,769,800
		EPIC		71,748	4	1,706	115,856,100	23,171,220,000
		WGBS		1,999,446	17	32	63,982,272	12,796,454,400
		Array	Val: Train CpG - Val Sample	33,885	33	920	30,638,464	6,127,692,800
			Val: Val CpG - Train Sample	6,742	33	8,258	55,141,308	11,028,261,600
			Val: Val CpG - Val Sample	6,742	33	920	6,143,360	1,228,672,000
TCGA	1 - 3	Array	Train: Train CpG - Train Sample	78,211	33	8,258	632,281,133	126,456,226,600
		EPIC		172,722	4	1,706	276,181,739	55,236,347,800
		WGBS		5,396,193	17	32	172,678,176	34,535,635,200
		Array	Val: Train CpG - Val Sample	78,211	33	920	70,443,801	14,088,760,200
			Val: Val CpG - Train Sample	14,893	33	8,258	121,617,682	24,323,536,400
			Val: Val CpG - Val Sample	14,893	33	920	13,550,097	2,710,019,400

Chromosome To further distinguish CpG sites, we encode this chromosome information as well. A set of learnable chromosome index embeddings $\{x_i^{\text{chr}}\}, i = 1, 2, \ldots, N_{\text{chr}}$ are used by our model. The length of each embedding is N_{embed}. Figure 2 (a,b) illustrate the chromosome embedder.

Global embedding Following the convention in the Transfomrer model (Devlin, 2018) and previous genome (Zhou et al., 2024) and gene fundation models (Cui et al., 2024b; Bai et al., 2024; Hao et al., 2024b), we use a learnable Global Embedding x_{GLB} to aggregate information among the embeddings across modalities. The encoded x_{GLB} embedding would be sent to a prediction head to output the final methylation level for the given CpG site and sample. Figure 2 (a,b) depicts the x_{GLB} embedding.

4.2. Model

Sample gene MLP embedder We use BottleneckMLP (Bachmann et al., 2023; 2024), a scalable state-of-the-art MLP encoder to process our gene (Figure 2 (c)). The model consists of several MLP block with skip connections (He et al., 2016).

Transformer encoder We use a Transformer encoder (Vaswani, 2017) to aggregate embeddings from all the above modalities. The attention mechanism in Transformer layers enables high modeling flexibility and capacity. The Transformer encoder leverage bi-directional attention where each embedding can attend to any others to collect information. We concatenate different types of embeddings in an order in Figure 2 (a), to form the input embedding sequence. The visual depiction of the Transformer model is illustrated in Figure 2 (a,d).

DNAm level projector We use a simple linear layer followed by a sigmoid function to project the encoded GLB embedding into the predicted DNAm level. The sigmoid function constrains the output in the range from 0 to 1 (Figure 2 (a)).

Training We use Mean Square Error (MSE) loss function to measure the error between the predicted DNAm levels with the gold standards. The derived loss can be used to update model weights end-to-end with back-propagation. For implementation details please refer to the Appendix.

5. Experiments

5.1. Data Source, Protocols, and Pre-processing

Our analysis uses two major datasets: ENCODE WGBS data (95 samples in 57 healthy tissue or cell types) (Hitz et al., 2023; The ENCODE Project Consortium, 2012; Luo et al., 2020) and TCGA methylation array data (9,194 samples in 33 cancer types) (The Cancer Genome Atlas Research Network, 2008), supplemented with TCGA EPIC and WGBS data to enhance CpG site coverage. The combined datasets contain billion-level CpG methylation signals, with varying coverage across datasets and samples. Each sample has corresponding RNA-seq data. We implemented a structured data partitioning strategy, creating four splits: training/validation CpG sites $\times$ training/validation samples with tissue-based partitioning for ENCODE (50% validation) and cancer-type-based partitioning for TCGA (10% validation), while CpG sites were randomly divided by chromosome (10% validation) (see Table 2). Data preprocessing included quality control, DNA sequence extraction (1Kb windows), CpG island and chromosome location annotations, and gene expression normalization (see Appendix).

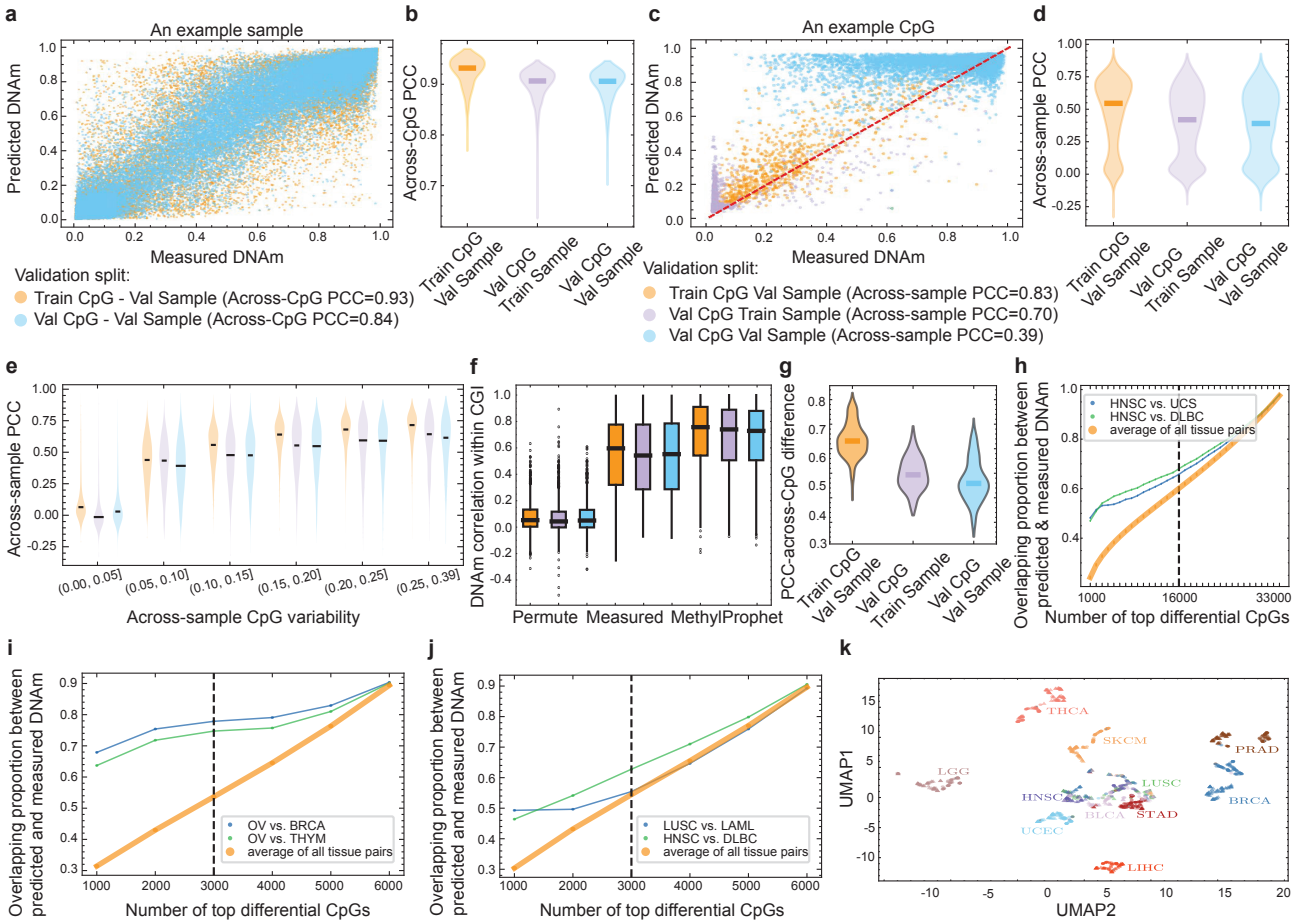

Figure 3: Cross-validation results on TCGA chromosome 1 data. (a) An example sample to demonstrate the calculation of across-CpG PCC. (b) Across-CpG PCC in three validation splits. (c) An example CpG to demonstrate the calculation of across-sample PCC. (d) Across-sample PCC in validation splits. (e) Across-sample PCC by DNAm variability in different train/validation splits, including Train CpG - Val Sample , Val CpG - Train Sample , Val CpG - Val Sample . (f) Predicted signal similarity within CGIs, with the same color scheme as (e). (g) The PCC of DNAm cell-type differences obtained from predicted and measured values. (h-j) DMR overlapping proportion between measured and predicted values. (k) UMAP of measured (triangles) and predicted (circles) samples.

5.2. Analysis Procedure

We evaluate the performance of MethylProphet on TCGA and ENCODE datasets using three validation splits. To quantify agreement between predicted and measured DNAm values, we employ six metrics:

(a) *Across-CpG Pearson correlation coefficient (PCC)* assesses how well the model preserves each sample's overall DNAm profile. (b) *Across-sample PCC* evaluates the model's ability to infer DNAm behavior at individual CpG sites. To further analyze across-sample PCC, we stratify CpG sites into bins based on their DNAm variability and examine PCC distributions within each bin. (c) *Tissue and cell-type preservation* is assessed by comparing pairwise tissue or cell-type differences between predicted and measured DNAm values. (d) *CpG island (CGI) coherence* is evaluated by measuring correlation between CpG pairs within the same CGI. As a baseline, we compute correlations after randomly permuting CGI indicators. (e) *Differentially methylated regions (DMRs)* are identified using `limma`, and we compare the top-ranked DMRs from predicted and measured DNAm by computing their overlap in descending significance cutoffs. (f) *Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)* verifies whether the inferred DNAm landscape captures tissue and cancer differences while preserving variation. For visualization, we project both predicted and measured DNAm onto the first two UMAP coordinates, where circles represent measured samples and triangles denote predicted samples.

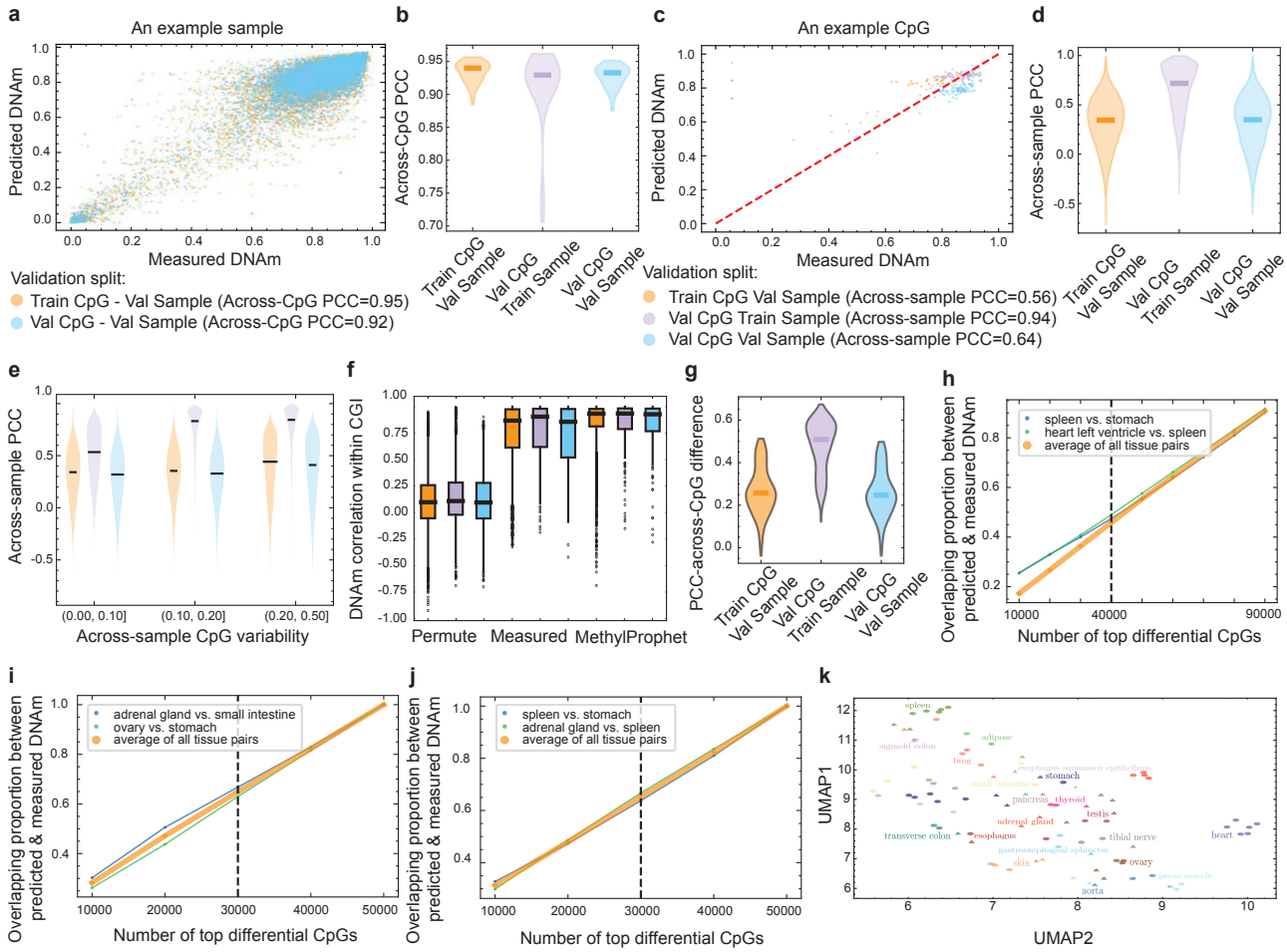

Figure 4: Cross-validation on ENCODE data. Similar to that of Figure 3, except that the results are based on the validation on ENCODE data. The sample differences (g) were calculated by comparing tissue/cell types rather than cancer types.

5.3. Results

MethylProphet performance on TCGA data Both across-CpG PCC (Figure 3 (a, b)) and across-sample PCC (Figure 3 (c, d)) reach the highest values in the *Train CpG - Val Sample* split, indicating that the model effectively captures site-wise DNAm patterns while generalizing well to new samples. Specifically, the predictions are consistently more accurate when generalizing to new samples rather than to new CpGs compared with splits of Val CpG - Train Sample and Val CpG - Val Sample (Figure 3 (b)). If a sample exhibits high across-CpG PCC, it suggests that the within-sample variability of CpGs is well captured (Figure 3 (a)). This result is expected, as the overall DNAm profile of a sample consists of a long vector of CpG elements, and global trends in DNAm are typically easier to learn and predict. For across-sample PCC (Figure 3 (d)), we observe a large variability, particularly when generalizing to both unseen CpGs and samples. The CpGs with high across-sample PCC indicate that the model can predict the CpG's variability across samples (Figure 3 (c)) well. This is very important because the ability to predict a CpG's behavior across individuals is highly related to its potential as a therapeutic target. We found that the across-sample PCC positively correlates with a CpG's variability across samples (Figure 3 (e)). Specifically, the highest median PCC values are observed for CpGs with a standard deviation (SD) in the range $(0.25, 0.36]$, reaching 0.70 for Train-CpG Val-Sample, 0.63 for Val-CpG Train-Sample, and 0.60 for Val-CpG Val-Sample.

MethylProphet successfully maintains intra-CGI correlation patterns across different validation splits (Figure 3 (f)), indicating regional epigenetic regulation.

In addition, MethylProphet is able to preserve cancer-specific DNAm differences (Figure 3 (g)). The Train CpG - Val Sample split exhibits the highest median PCC difference, indicating that the model effectively maintains cancer-specific DNAm patterns when predicting new samples using a fixed set of CpGs. However, the Val CpG - Train Sample

Table 3: Results of training models on different data sources. Datasets: E and T denotes ENCODE and TCGA, respectively; in brackets, A, E, and W denote Array, EPIC, and WGBS samples, respectively.

Train Data	Val Data	Train CpG - Val Sample				Val CpG - Train Sample				Val CpG - Val Sample			
		MAS-PCC	MAC-PCC	MSE	MAE	MAS-PCC	MAC-PCC	MSE	MAE	MAS-PCC	MAC-PCC	MSE	MAE
E (W)	E (W)	0.3436	0.9398	0.0079	0.0608	0.7165	0.9297	0.0108	0.0679	0.3411	0.9330	0.0086	0.0634
T (A)	T (A)	0.4000	0.8669	0.0363	0.1216	0.2769	0.7914	0.0555	0.1498	0.2597	0.7930	0.0557	0.1504
T (A+W)	T (A)	0.4705	0.3244	0.2981	0.9112	0.8674	0.8673	0.0252	0.0365	0.0369	0.1006	0.1205	0.1212
T (A+E)	T (A)	0.5226	0.9232	0.0222	0.0920	0.3727	0.8738	0.0350	0.1147	0.3451	0.8743	0.0355	0.1157
T (A+E+W)	T (A)	0.5455	0.9320	0.0199	0.0882	0.4194	0.9065	0.0266	0.1000	0.3904	0.9059	0.0271	0.1011

Table 4: The ablation of increasing training data scale by adding chromosomes for TCGA data.

Train Chr.	Val Chr.	Train CpG - Val Sample				Val CpG - Train Sample				Val CpG - Val Sample			
		MAS-PCC	MAC-PCC	MSE	MAE	MAS-PCC	MAC-PCC	MSE	MAE	MAS-PCC	MAC-PCC	MSE	MAE
1	1	0.5455	0.9320	0.0199	0.0882	0.4194	0.9065	0.0266	0.1000	0.3904	0.9059	0.0271	0.1011
1+2+3	1	0.4928	0.9249	0.0219	0.0915	0.3760	0.8961	0.0294	0.1047	0.3505	0.8960	0.0298	0.1057
1	1+2+3	0.3025	0.8012	0.0535	0.1473	0.2654	0.8216	0.0492	0.1362	0.2513	0.8230	0.0495	0.1368
1+2+3	1+2+3	0.4872	0.9246	0.0224	0.0919	0.3736	0.8993	0.0290	0.1027	0.3460	0.8992	0.0295	0.1037

and Val CpG - Val Sample splits show a decline in PCC differences, suggesting reduced performance in capturing cancer-type variation when generalizing to unseen CpGs.

The differential CpGs achieves the highest overlap between predicted and measured DNAm in the Train CpG - Val Sample split, followed by Val CpG - Train Sample and Val CpG - Val Sample splits (Figure 3 (h-j)). In addition, MethylProphet-predicted DNAm landscape successfully preserves cancer-specific differences, as samples from the same cancer type remain well-clustered (Figure 3 (k)).

MethylProphet performance on ENCODE data Unlike TCGA, where MethylProphet performs best in the Train CpG - Val Sample split, ENCODE shows a different trend across validation splits. For across-CpG PCC (Figure 4 (a, b)), the performance is similar across splits, while for across-sample PCC (Figure 4 (c, d)), MethylProphet performs best in the *Val CpG - Train Sample* split, possibly due to the limited testing samples in ENCODE data. Similar to that in TCGA, MethylProphet predicts methylation patterns more accurately for highly variable CpGs, where across-sample PCC increases with CpG variability (Figure 4 (e)).

In this cancer cohort, MethylProphet also effectively captures CpG co-methylation dynamics within CGIs (Figure 4 (f)). In the assessment of MethylProphet's ability to preserve tissue-specific DNAm differences, the Val CpG - Train Sample split exhibits the highest median PCC-across-CpG difference (Figure 4 (g)). This contrasts with TCGA, where the Train CpG - Val Sample split performed best.

The top-ranked DMRs obtained using predicted and measured DNAm achieve a relatively high overlap across all validation splits (Figure 4 (h-j)). However, unlike in TCGA, MethylProphet performs comparably across splits. This suggests that the DMR list is more stable, likely due to the significantly larger number of CpGs included in ENCODE data. Overall, MethylProphet successfully preserves tissue differences (Figure 4 (k)), with predicted and measured samples of the same cancer types cluster together.

Ablations We conduct ablations on data mixing strategies and scaling effects on TCGA data, with results presented in Table 3 and Table 4. Performance is evaluated using four metrics: 1) Median of Across-Sample Pearson Correlation Coefficient Median (MAS-PCC); 2) Median of Across-CpG Pearson Correlation Coefficient Across CpG Median (MAC-PCC); 3) Mean Square Error (MSE) and 4) Mean Absolute Error (MAE) between predicted and ground truth methylation values in validation sets.

In the data mixing ablation (Table 3), MethylProphet demonstrates strong CpG encoding capability on ENCODE WGBS data, achieving high MAS-PCC scores of 0.72 on the Val CpG - Train Sample split. However, performance on Val Sample is moderate, likely due to the limited sample size (Table 2) constraining the model's ability to learn generalized gene encodings for sample-specific features.

Training solely on Array data yields suboptimal performance, with MAS-PCC scores of 0.40, 0.28, and 0.26 on Train CpG - Val Sample, Val CpG - Train Sample, and Val CpG - Val Sample splits, respectively. This limitation stems from the restricted number of CpG sites in Array data (Table 2). Performance improves consistently when incorporating additional data sources, with optimal results achieved

by combining Array, EPIC, and WGBS data, yielding MASPCC scores of 0.54, 0.42, and 0.39 for the respective splits.

The scaling ablation (Table 4) reveals that models trained exclusively on chromosome 1 show limited generalization to additional chromosomes (2 and 3). While training on all three chromosomes slightly decreases validation performance on chromosome 1, it significantly improves the model's ability to generalize across chromosomes.

6. Conclusion

We present MethylProphet, a novel Transformer-based approach that enables comprehensive DNA methylation inference by integrating gene profile with genomic context. Trained on extensive datasets from ENCODE (1.6B training CpG-sample pairs; 322B tokens) and TCGA chromosome 1 data (450M training pairs; 91B tokens), Our model demonstrates robust performance in inferring genome-wide methylation patterns across diverse tissues and cancer types. We hope this capability to reconstruct complete methylomes from limited experimental data could advance both epigenetic research and precision medicine applications.

Impact Statement

While our primary objective is to enhance epigenetic research and precision medicine capabilities, we acknowledge that advances in genomic prediction technologies may have broader societal implications, including privacy considerations and ethical questions regarding genetic information accessibility. We have focused on developing methods that maintain scientific rigor while adhering to established ethical guidelines in computational biology and medical research. Our model, data source, data processing pipelines, and evaluation protocols are designed with transparency and reproducibility in mind, and we will release all code, data, protocols, and models to facilitate open scientific discourse and validation.

Acknowledgements

W. H. and Y.F. Z. were supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number K99/R00HG011468, through which W. H. received related training from Drs. Hongkai Ji, Stephanie Hicks, and Andrew Feinberg. Q. L. was supported by the General Fund at Columbia University Department of Biostatistics. We would like to thank Columbia HPC (including Insomnia and C2B2), TPU Research Cloud (TRC) program and Google Cloud Research Credits program for supporting our computing needs.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol*, 18(1): 67, December 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1189-z.

Bachmann, G., Anagnostidis, S., and Hofmann, T. Scaling MLPs: A Tale of Inductive Bias, October 2023.

Bachmann, R., Kar, O. F., Mizrahi, D., Garjani, A., Gao, M., Griffiths, D., Hu, J., Dehghan, A., and Zamir, A. 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities, June 2024.

Bai, D., Mo, S., Zhang, R., Luo, Y., Gao, J., Yang, J. P., Wu, Q., Singh, D., Rahmani, H., Amariuta, T., Grotjahn, D., Zhong, S., Lewis, N., Wang, W., Ideker, T., Xie, P., and Xing, E. scLong: A Billion-Parameter Foundation Model for Capturing Long-Range Gene Context in Single-Cell Transcriptomics. Preprint, Bioinformatics, November 2024.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

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

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods*, 21(8):1470–1480, August 2024b. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-024-02201-0.

De Lima Camillo, L. P., Sehgal, R., Armstrong, J., Higgins-Chen, A. T., Horvath, S., and Wang, B. CpGPT: A Foundation Model for DNA Methylation. Preprint, Systems Biology, October 2024.

Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Feinberg, A. P. The key role of epigenetics in human disease prevention and mitigation. *New England Journal of Medicine*, 378(14):1323–1334, 2018.

Gage, P. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.

Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., and Song, L. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, pp. 1–11, 2024a.

Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., and Song, L. Large-scale foundation model on single-cell transcriptomics. *Nat Methods*, 21(8):1481–1491, August 2024b. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-024-02305-7.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hitz, B. C., Jin-Wook, L., Jolanki, O., Kagda, M. S., Graham, K., Sud, P., Gabdank, I., Strattan, J. S., Sloan, C. A., Dreszer, T., et al. The ENCODE Uniform Analysis Pipelines. *Biorxiv*, 2023.

Levy, J. J., Titus, A. J., Petersen, C. L., Chen, Y., Salas, L. A., and Christensen, B. C. MethylNet: An automated and modular deep learning approach for DNA methylation analysis. *BMC Bioinformatics*, 21(1):108, December 2020. ISSN 1471-2105. doi: 10.1186/s12859-020-3443-8.

Levy-Jurgenson, A., Tekpli, X., Kristensen, V. N., and Yakhini, Z. Predicting methylation from sequence and gene expression using deep learning with attention. In *Algorithms for Computational Biology: 6th International Conference, AlCoB 2019, Berkeley, CA, USA, May 28–30, 2019, Proceedings 6*, pp. 179–190. Springer, 2019.

Loyfer, N., Magenheim, J., Peretz, A., Cann, G., Bredno, J., Klochendler, A., Fox-Fisher, I., Shabi-Porat, S., Hecht, M., Pelet, T., Moss, J., Drawshy, Z., Amini, H., Moradi, P., Nagaraju, S., Bauman, D., Shveiky, D., Porat, S., Dior, U., Rivkin, G., Or, O., Hirshoren, N., Carmon, E., Pikarsky, A., Khalaileh, A., Zamir, G., Grinbaum, R., Abu Gazala, M., Mizrahi, I., Shussman, N., Korach, A., Wald, O., Izhar, U., Erez, E., Yutkin, V., Samet, Y., Rotnemer Golinkin, D., Spalding, K. L., Druid, H., Arner, P., Shapiro, A. M. J., Grompe, M., Aravanis, A., Venn, O., Jamshidi, A., Shemer, R., Dor, Y., Glaser, B., and Kaplan, T. A DNA methylation atlas of normal human cell types. *Nature*, 613(7943):355–364, January 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-022-05580-6.

Luo, Y., Hitz, B. C., Gabdank, I., Hilton, J. A., Kagda, M. S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., Baymuradov, U. K., Graham, K., Litton, C., Miyasato, S. R., Strattan, J. S., Jolanki, O., Lee, J.-W., Tanaka, F. Y., Adenekan, P., O'Neill, E., and Cherry, J. M. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, 48(D1): D882–D889, January 2020. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkz1062.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Shu, C., Zhang, X., Aouizerat, B. E., and Xu, K. Comparison of methylation capture sequencing and infinium methylationepic array in peripheral blood mononuclear cells. *Epigenetics & chromatin*, 13:1–15, 2020.

The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, October 2008. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature07385.

The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature11247.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. LLaMA: Open and Efficient Foundation Language Models, February 2023.

Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Wang, H., Huang, T., Wang, D., Zeng, W., Sun, Y., and Zhang, L. MSCAN: Multi-scale self- and cross-attention network for RNA methylation site prediction. *BMC Bioinformatics*, 25(1):32, January 2024. ISSN 1471-2105. doi: 10.1186/s12859-024-05649-1.

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

Ying, K., Song, J., Cui, H., Zhang, Y., Li, S., Chen, X., Liu, H., Eames, A., McCartney, D. L., Marioni, R. E., Poganik, J. R., Moqri, M., Wang, B., and Gladyshev, V. N. MethylGPT: A foundation model for the DNA methylome. Preprint, Systems Biology, November 2024.

Zeng, W., Gautam, A., and Huson, D. H. MuLan-Methyl—multiple transformer-based language models for accurate DNA methylation prediction. *GigaScience*, 12: giad054, December 2022. ISSN 2047-217X. doi: 10.1093/gigascience/giad054.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., and Liu, H. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome, March 2024.

Zhuo, L., Wang, R., Fu, X., and Yao, X. StableDNAm: Towards a stable and efficient model for predicting DNA methylation based on adaptive feature correction learning. *BMC Genomics*, 24(1):742, December 2023. ISSN 1471-2164. doi: 10.1186/s12864-023-09802-7.

A. Implementation Details

The implementation details is shown in Table 5. For the experiments on ENCODE WGBS and TCGA chromosome 1, 2, and 3, we use 64 GPUs with 512 batch size per accelerator, taking about 1 GPU day for each experiment. While for those on TCGA chromosome 1, we use 32 GPUs with batch size 256, taking about half of GPU day for each experiment. We turn on gradient checkpointing to reduce memory usage and enable flah-attention 2 to speed up attention operator. The parameters specification and their computational cost are shown in Table 6.

Table 5: The implementation details.

Optimization	
Optimizer	AdamW (0.9, 0.95)
LR	1.00E-04
LR Decay Ratio	10x
LR Decay	cosine
Weight Decay	1.00E-03
LR Warmup	Linear
Warmup steps	2000
Gradient Clipping	1
Data Epoch	1
Batch Size*	256/512
Accelerator Type	NVIDIA L40s
# Accelerator	32/64
Training Precision	Mixed bf16

B. Data

B.1. Data Source

ENCODE data. Processed RNA-seq (TPM) and WGBS (β values) data were downloaded from The Encyclopedia of Elements (ENCODE) portal (https://www.encodeproject.org/). We identified wild-type samples with both RNA-seq and WGBS profiles, along with matched summary information including species, sex, age, tissue, and bioSample information. Technical replicates were combined by averaging their gene expression and their DNA methylation profiles. The averaged TPM values were $\log_2$-transformed after adding a pseudocount of 1. For WGBS data aligned to the hg19 genome, genome coordinates were converted to hg38 using liftover. Samples with WGBS data covering more than 80% of all CpG sites on autosomes and chromosome X were retained. Finally, all CpGs located on chromosomes X and Y were removed. A total of 95 samples covering 28,301,739 CpG sites and 55,503 genes were included in the final dataset.

TCGA data. Processed RNA-seq (TPM), 450K array and EPIC (β values) data were downloaded from the Cancer Genome Atlas Program (TCGA) data portal (https://portal.gdc.cancer.gov/). Processed whole-genome bisulfite sequencing (WGBS) data (β values) were downloaded from a static website provided by TCGA (https://zwdzwd.s3.amazonaws.com/directory_listing/trackHubs_TCGA_WGBS_hg38.html). For RNA-seq data, the TPM values were averaged for samples belonging to the same case. The averaged TPM values were $\log_2$-transformed after adding a pseudocount of 1. For 450K array and EPIC data, CpG sites with missing values across all samples were filtered out, and the β values were averaged for samples belonging to the same case. The WGBS data provided β values for each case. CpG sites with missing values across all cases were filtered out, and those located on chromosomes X and Y were removed. The final dataset included 9,194 450K array samples covering 408,399 CpG sites, 1,706 EPIC samples covering 740,296 CpG sites, and 32 WGBS samples covering 23,047,052 CpG sites. Additionally, gene expression profiles spanning 60,660 genes were included for each sample.

B.2. Data Partition and Protocols

Our model takes CpG-related information and gene expressions as inputs and predicts the methylation level for the given CpG site. Originally, there are three raw files to be processed, a raw methylation beta file, a sample gene profile, and a reference human DNA sequence template (hg38. The raw methylation beta profile consists of a matrix $M \times N$, where there are M CpG sites and N samples, while the gene expression profile includes the expression of L genes for all samples N. The data partition pipeline is shown in Figure 5.

Data sharding and sanity check. Since the raw methylation beta matrix is enormous, reaching an order of magnitude of billion (2.8 billion for ENCODE WGBS and 3 billion for TCGA Array), we first split the gigantic matrix into small shards. Sharding can leverage parallel computation and thus speed up the data pre-processing. We split the methylation beta matrix by rows (*i.e.*, by CpG sites) where every 10k rows assemble a shard file. During methylation matrix sharding, the corresponding DNA sequence for each CpG site in a shard is saved simultaneously using the reference human DNA sequence template (hg38). The window size of DNA sequence is 1Kb for the given CpG site. In addition, we filter out NaN entries and deduplicate genes and CpG sites.

Sample split. To split samples in to training and validation set, we first count the number of samples for each tissue / cancer types (ENCODE WGBS, Figure 6; TCGA Array, Figure 7). Then we split the samples based on the types.

Table 6: Model size and computation. ∗: Number of parameters includes the DNA tokenizer embeddings. †: FLOPs are estiamted with batch size equal 1.

Transformer Size	# of Hidden Layers	Hidden Size	# of Attention Heads	# of Params *	FLOPs †
Base	12	768	12	110M	104G
MLP Size	**# of Hidden Layers**	**Hidden Size**	**Bottleneck Factor**	**# of Params**	**FLOPs**
B_6-Wi_1024	6	1024	4	70M	70M

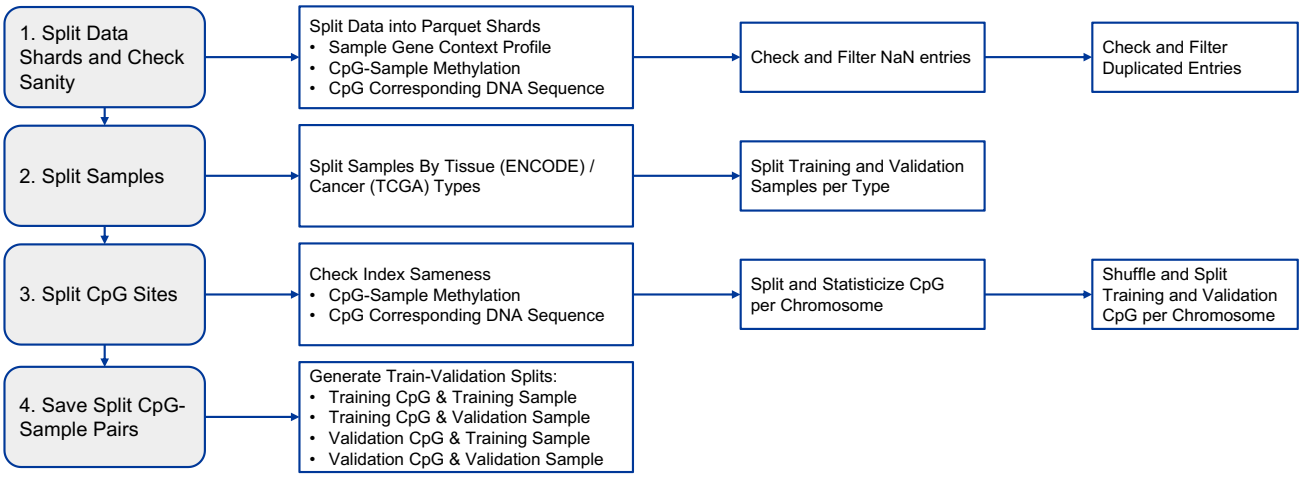

Figure 5: Data partition diagram.

There are 57 tissue types and 95 samples in total in ENCODE. For those tissues with more than one samples, We randomly sampled half of them as the validation samples. All the rest samples are used for the training set.

In TCGA, there are 33 cancer types with 9194 samples summed up. We randomly choose 10% of the samples for each tissue type as validation samples, and the rest are left for training. For those do not have cancer type assigned, we treat them as type "Unknown".

CpG split. We first check the methylaton matrix and the corresponding DNA sequence have the same CpG index. Then we statisticize CpG sites for each chromosome. We randomly pick 10% for CpG sites in each chromosome as training CpG sites. For ENCODE, we temporarly sample another 10% as training split. While for TCGA, we use the rest 90% as training. We supplement TCGA with addition EPIC and WGBS data which have no intersected with Array data.

CpG-sample split. The CpG sample splits are based on the previous sample and CpG splits. For ENCODE WGBS and TCGA Array, we would have four splits, where the first split is used for training, and the rest three splits are used for validation and performance report:

1. "Training CpG and Training Sample", for training;
2. "Training CpG and Validation Sample", for validation;
3. "Validation CpG and Training Sample", for validation;
4. "Validation CpG and Validation Sample", for validation.

To further synergy the limited CpG sites in TCGA array data, we additionally incorporate TCGA EPIC and TCGA WGBS data, which have no intersections with TCGA array data.

B.3. Data Pre-processing

CpG-specific DNA sequence. We extract the DNA sequence around the CpG site to represent the CpG site. The window size is 1Kb for each site. Besides, we record the CpG island index, as well as its region types (CpG island, CpG shore, CpG shelf, and CpG ocean). For those sites in CpG ocean, we assign -1 as their CpG island index. We embed the above information numerically, as shown in Figure 2 (a), (b), and (d).

Gene expression. The RNA counts are $\log_2$-transformed after adding a pseudocount of 1. Genes with mean and standard deviation below the specified cutoffs (ENCODE: mean

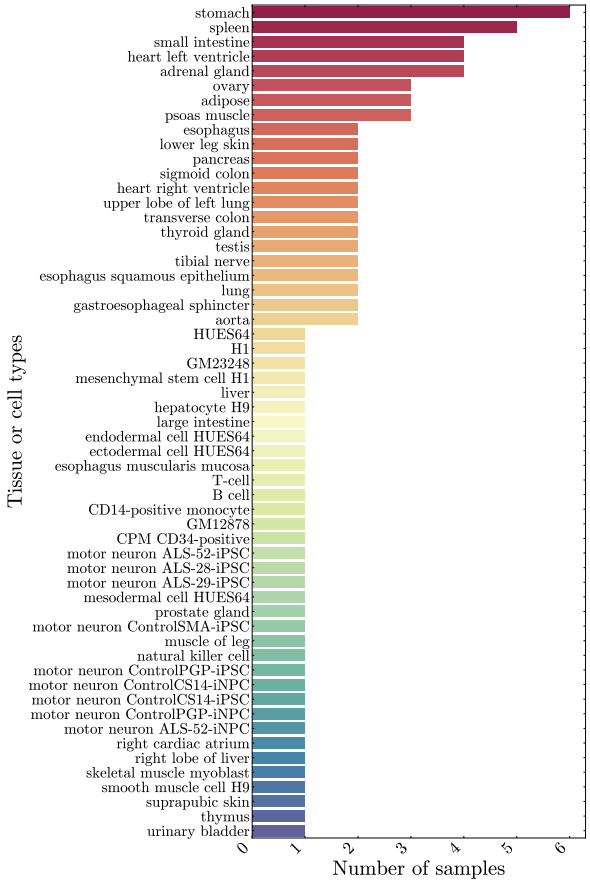

Figure 6: Samples counts by tissue types in ENCODE data.

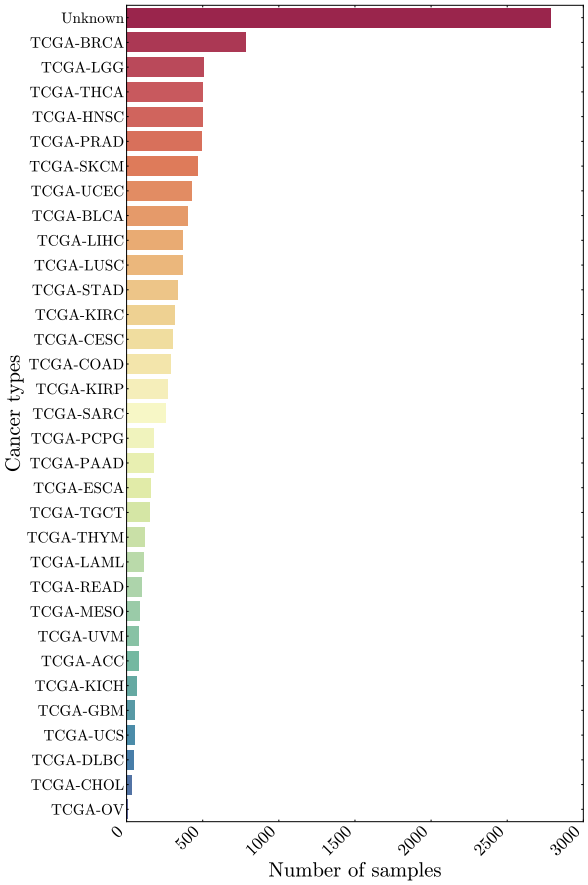

Figure 7: Samples counts by cancer types in TCGA data.

$= 0.1$, std $= 0.1$; TCGA: mean $= 0.5$, std $= 0.5$) are filtered out. Mitochondrial, proline-rich and ribosomal protein genes are removed. As a result, 24,337 genes are retained in the ENCODE dataset and 25,017 genes in the TCGA dataset. Note that both protein-coding and non-protein-coding genes are included prior to filtering. To mitigate batch effects, we apply the quantization technique(Cui et al., 2024b) where the $\log_2$-transformed RNA counts are quantized based on their probability densities. The quantized values are then linearly mapped to the range $[0, 1]$ to mitigate batch effects. The resulting gene expression vectors are subsequently encoded in the downstream model.

C. Addtional Experimental Results of Abaltion Studies

Figure 8 and Figure 9 illustrate the distribution of PCC for our ablation studies, which supplement Table 3 and Table 4: 1) the effect of mixing TCGA data with different sequencing techniques. 2) the effect of increasing data scale of TCGA.

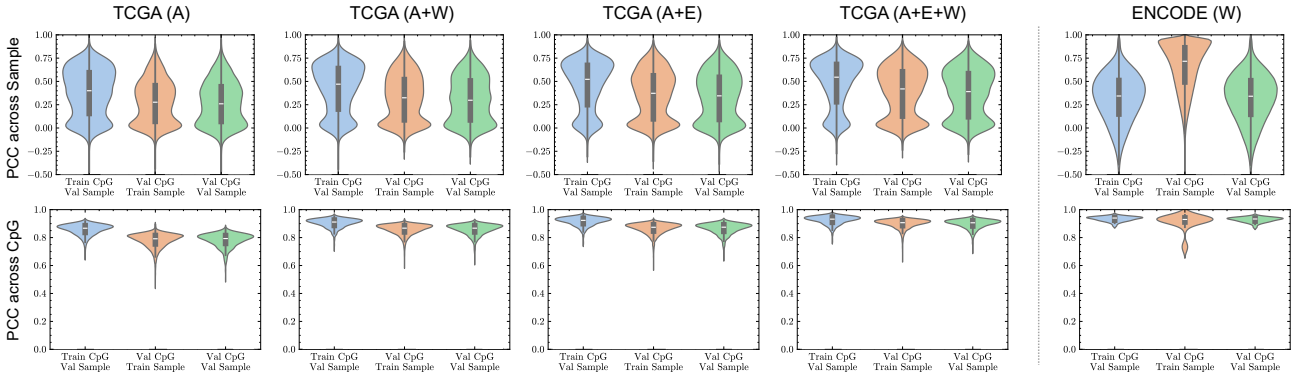

Figure 8: The distribution of PCC across Sample / CpG on validation sets for TCGA chromosome 1 data.

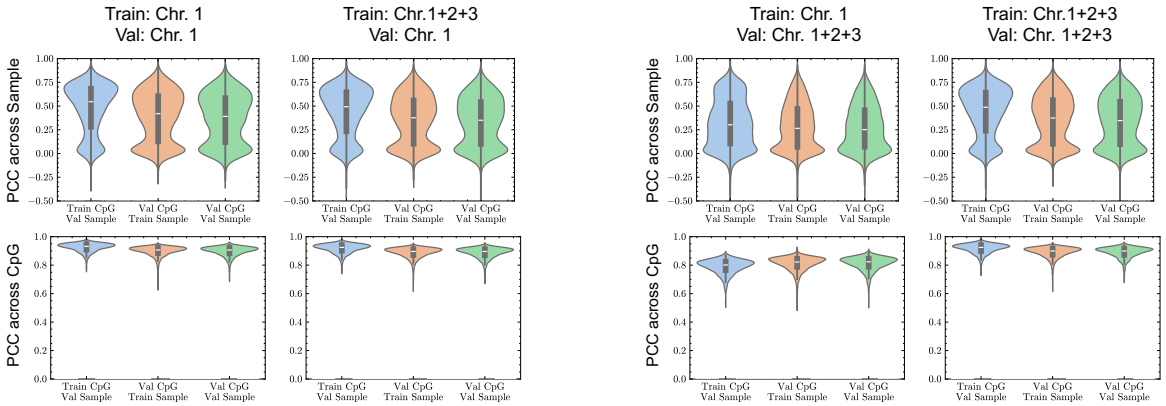

Figure 9: The distribution of PCC across Sample / CpG when increasing TCGA data scale by adding more chromosomes.