Appendix for "MethylProphet"

Contents

A Potential biological insights **13**

B Data **14**

B.1 Data Source 14

B.2 Data Partition and Protocols 15

B.3 Data Pre-processing 16

C Implementation Details **17**

C.1 Configurations of MethylProphet 17

C.2 Baselines 17

C.2.1 Levy-Jurgenson et al. (2019b) 17

C.2.2 CpGPT (De Lima Camillo et al., 2024) 18

D Additional evaluation metrics **18**

D.1 Across-sample PCC by DNAm variability 18

D.2 PCC of DNAm cell-type and tissue differences 18

D.3 DMR overlapping proportion between measured and predicted values 18

E Evaluation results **18**

E.1 Robustness to missing context DNAm 18

E.2 MethylProphet performance on TCGA data 19

E.3 MethylProphet performance on ENCODE data 21

F Discussion **22**

F.1 Limitation and future work 22

F.2 Broad impact 22

A Potential biological insights

MethylProphet enables genome-wide DNA methylation (DNAm) reconstruction from gene expression and sequence data alone, providing unique opportunities for advancing biological interpretation, methodological development, and genomic applications. This cross-modality prediction framework offers several important insights and use cases in real-world biomedical research.

First, MethylProphet facilitates low-cost methylome reconstruction in settings where whole-genome bisulfite sequencing (WGBS) or array-based profiling is infeasible. Many large-scale transcriptomic datasets lack matching methylome profiles, and thus cannot be directly leveraged for epigenetic discovery. For example, the ENCODE consortium has generated 1,699 RNA-seq samples but only 211 WGBS samples; the TCGA program includes more than 10,426 RNA-seq

samples but only 32 WGBS samples; and GEO hosts 241,014 RNA-seq samples but just 6,318 WGBS samples. By computationally inferring DNAm in these cohorts, METHYLPROPHET enables downstream epigenetic analyses without the need for additional profiling.

Second, METHYLPROPHET enhances public and disease biobank resources such as GTEx, ENCODE, TCGA, and PCAWG by providing whole-genome methylome predictions. This allows for deeper epigenetic insights, cancer subtype stratification, and biomarker discovery. Prior work, such as (Yang et al., 2024), predicted DNAm from GTEx and multi-omics TCGA data, but their scope was limited to Illumina EPIC array CpGs, covering only $\sim$3% of the genome. By contrast, METHYLPROPHET enables whole-genome prediction at more than 100$\times$ the sample scale, thereby extending coverage from 3% to 100% of the genome and broadening the landscape of epigenetic discovery.

Third, METHYLPROPHET supports sample-level methylation estimation in multi-omic and single-cell studies, where DNAm data are often sparse or missing. This ability to reconstruct complete sample-level methylomes from transcriptomic profiles enables downstream tasks such as DNAm regulation inference, cell-fate trajectory analysis, and multi-omic clustering, all without requiring methylation-specific assays.

In addition, METHYLPROPHET contributes to predictive biomarker development. For instance, 850K array-based methylation profiles have been used to predict brain metastases (Zuccato et al., 2025). By extending methylation reconstruction to the full genome, METHYLPROPHET opens new possibilities for noninvasive biomarker discovery and risk stratification in cohorts that lack direct methylation assays.

Another important application is in the development of DNA methylation clocks for aging and disease phenotyping. Epigenetic clocks such as Horvath and GrimAge estimate biological age based on a small number of CpGs, but their accuracy is limited by array coverage (1–3% of the genome). METHYLPROPHET provides genome-wide methylation inference, improving both the resolution and accuracy of aging models. Furthermore, it enables biological age estimation in transcriptome-only cohorts, thereby expanding the reach of age-related biomarkers in large-scale population and longitudinal studies.

Beyond these applications, METHYLPROPHET establishes cross-modality prediction as a powerful paradigm in multi-omics. Cross-modal inference is increasingly central to computational biology: studies have predicted DNAm from expression (Yang et al., 2024; Liu et al., 2024), chromatin accessibility from expression and DNA (Zhou et al., 2017), and gene expression from sequence (Avsec et al., 2021). More recently, ALPHAGENOME leveraged such predictions for virtual perturbation analyses (Avsec et al., 2025). These efforts collectively reduce experimental cost, enable retrospective analyses on existing data, and broaden the scope of multi-omic investigations, especially in disease contexts such as cancer, heart failure, and leukemia. Within this broader landscape, METHYLPROPHET demonstrates that accurate genome-wide DNAm prediction from transcriptome and sequence data is both feasible and biologically meaningful, thereby opening new directions for integrative epigenomic discovery.

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 A1.

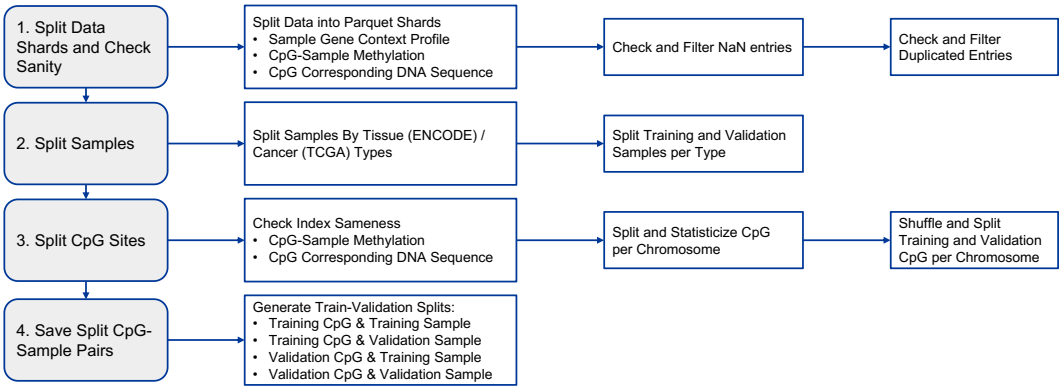

Figure A1: Data partition diagram.

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 A2; TCGA Array, Figure A3). Then we split the samples based on the types.

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

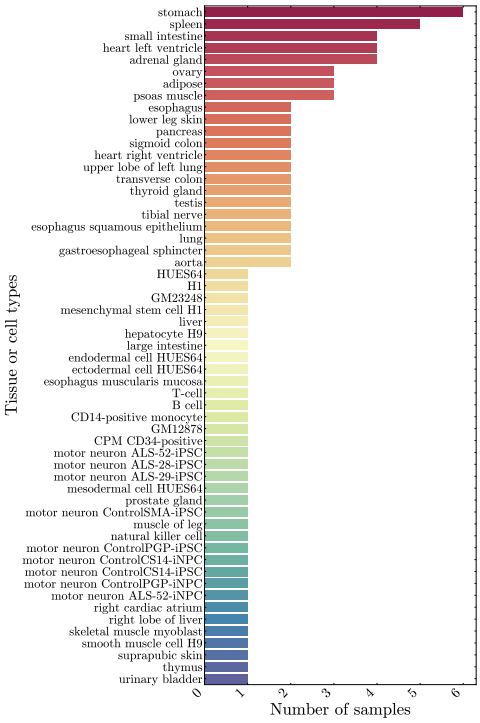

Figure A2: Samples counts by tissue types in ENCODE data.

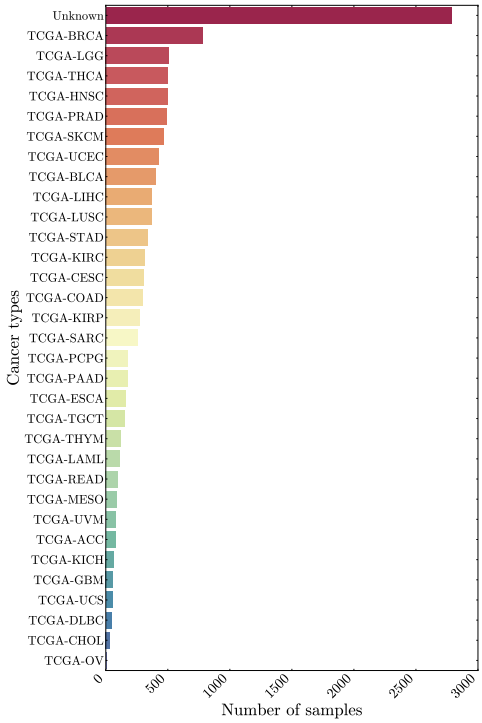

Figure A3: Samples counts by cancer types in TCGA data.

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.

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 $= 0.1$, 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 Implementation Details

C.1 Configurations of MethylProphet

The implementation details is shown in Table A1. 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 A2.

Table A1: 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

Table A2: 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

C.2 Baselines

C.2.1 Levy-Jurgenson et al. (2019b)

We implement the model described in Levy-Jurgenson et al. (2019b), which uses a multi-branch architecture with four subnetworks: two convolutional neural network (CNN) branches that process DNA sequences around CpG sites, and two attention-based MLP branches that incorporate gene expression and CpG-gene distance, respectively. The outputs of all branches are concatenated and passed through a final regression head to predict DNAm levels. We use the original model structure as described in the paper. To ensure fairness, we apply the same input preprocessing and trained on the same data splits as MethylProphet. Our reimplementation is based on the open-source code available at: https://github.com/YakhiniGroup/Methylation.

C.2.2 CpGPT (De Lima Camillo et al., 2024)

CpGPT is an imputation-based Transformer model trained via masked modeling on large-scale CpG methylation data. It learns context-aware representations of CpG sites by predicting masked methylation values based on the surrounding sequence. In our evaluation, we use the trained CpGPT-100M model to extract sample-level embeddings for 20 randomly selected samples from the Train Sample set. These embeddings are then used to predict DNAm levels at the corresponding Val CpG sites for each selected sample, following the Val CpG – Train Sample evaluation split. We use the publicly released trained model and inference code from: https://github.com/lcamillo/CpGPT.

D Additional evaluation metrics

To complement the main performance metrics, we provide more evaluations to better understand model behavior, particularly in capturing biologically meaningful DNA methylation (DNAm) signals.

D.1 Across-sample PCC by DNAm variability

We stratify CpG sites into bins according to their inter-sample DNAm variability, computed as the standard deviation of beta values across samples. For each bin, we compute the distribution of across-sample PCCs between predicted and measured methylation levels.

D.2 PCC of DNAm cell-type and tissue differences

To assess the preservation of biological variation, we compare pairwise differences in average methylation levels between tissues or cell types, calculated for predicted and measured data. For each tissue or cell-type pair, we compute the PCC between predicted and measured methylation differences across CpG sites. High correlations indicate that the model captures inter-tissue and inter-cell-type epigenetic distinctions.

D.3 DMR overlapping proportion between measured and predicted values

We identify Differentially Methylated Regions (DMRs) from both measured and predicted methylation matrices using the `limma` R package. We rank DMRs by statistical significance and compute the overlap proportion between top-ranked regions from the predicted and measured DNAm matrices, across varying thresholds (e.g., top 1000, 2000 DMRs).

E Evaluation results

E.1 Robustness to missing context DNAm

To assess the reliance on surrounding DNAm context, we conducted an ablation study by progressively reducing the percentage of available context CpG values for CpGPT. Table A3 and Table A4 report the performance across 200 held-out test samples.

Table A3: MAS-PCC (median across samples) under different levels of available context CpGs.

% surrounding DNAm	CpGPT	MethylGPT	MethylProphet
100%	0.19	0.23	0.31
80%	0.21	0.18	0.31
60%	0.13	0.15	0.31
40%	0.09	0.12	0.31
20%	0.06	0.08	0.31

When no surrounding context DNAm is available, CpGPT and MethylGPT degenerate (their output variance collapses), and the PCC metric becomes undefined. In contrast, *MethylProphet* remains

Table A4: MAC-PCC (median across CpGs) under different levels of available context CpGs.

% surrounding DNAm	CpGPT	MethylGPT	MethylProphet
100%	0.84	0.78	0.88
80%	0.88	0.69	0.88
60%	0.79	0.63	0.88
40%	0.69	0.54	0.88
20%	0.60	0.49	0.88

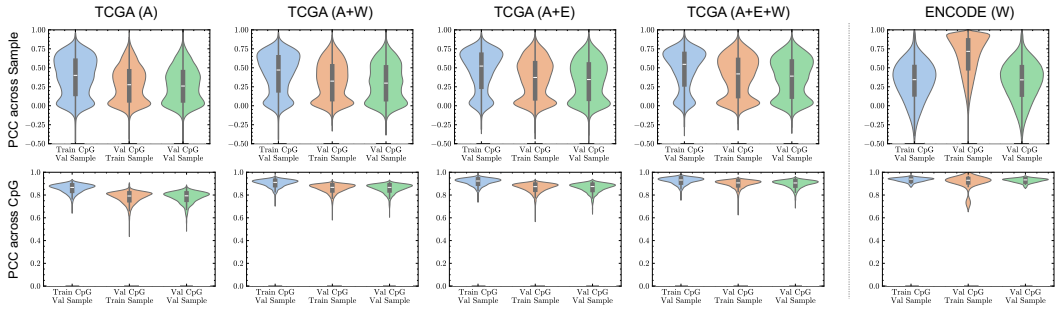

Figure A4: The distribution of PCC across Sample / CpG on validation sets for TCGA chromosome 1 data.

stable across all levels of context sparsity due to its reliance on gene expression and DNA sequence inputs, which are independent of neighboring CpG methylation measurements.

These results highlight that *MethylProphet* is not only competitive in predictive accuracy but also substantially more robust and generalizable in low-data or missing-data settings. This robustness is especially valuable for real-world applications where measured DNAm data may be sparse or unavailable.

E.2 MethylProphet performance on TCGA data

Figure A4 and Figure A5 illustrate the distribution of PCC for our ablation studies : 1) the effect of mixing TCGA data with different sequencing techniques. 2) the effect of increasing data scale of TCGA.

Both across-CpG PCC (Figure A6 (a, b)) and across-sample PCC (Figure A6 (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 A6 (b)). If a sample exhibits high across-CpG PCC, it suggests that the within-sample variability of CpGs is well captured (Figure A6 (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 A6 (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 A6 (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 A6 (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 A6 (f)), indicating regional epigenetic regulation.

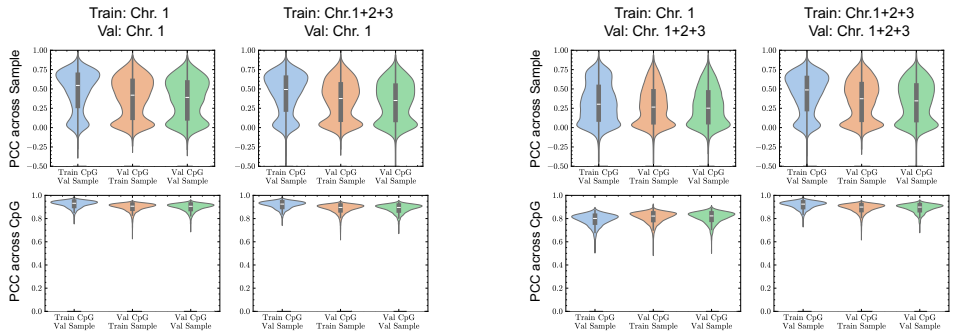

Figure A5: The distribution of PCC across Sample / CpG when increasing TCGA data scale by adding more chromosomes.

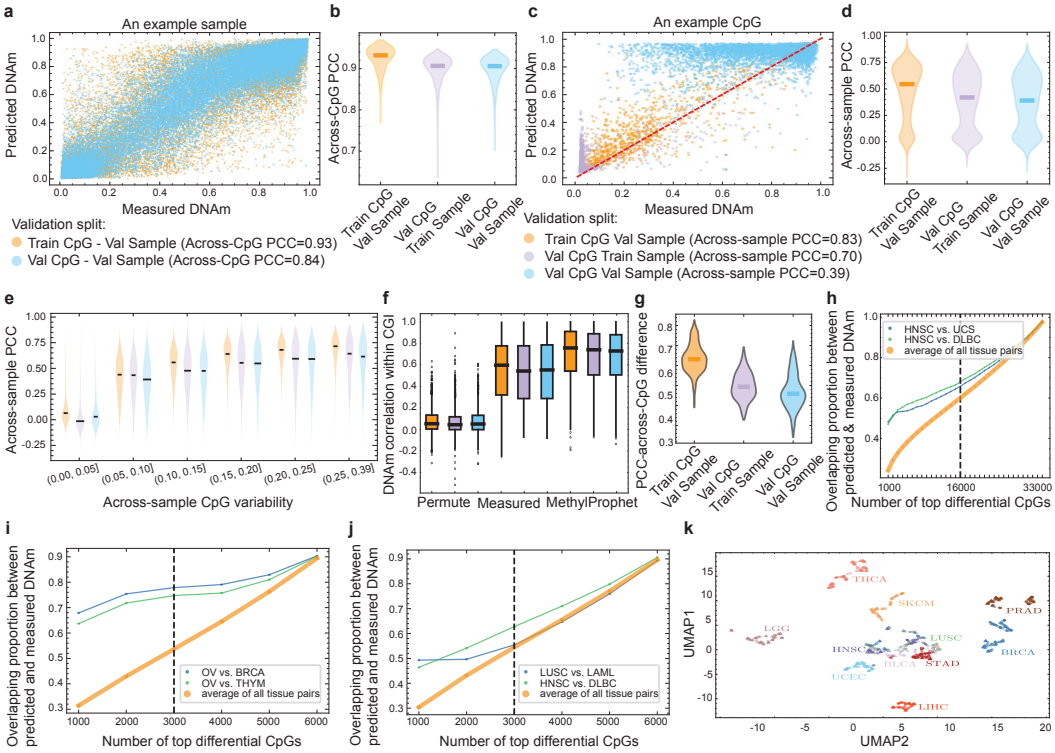

Figure A6: Cross-validation results on TCGA chromosome 1 data. (a) An example sample to demonstrate the calculation of across-CpG PCC. (b) Across-CpG PCC in three validation splits. (c) An example CpG to demonstrate the calculation of across-sample PCC. (d) Across-sample PCC in validation splits. (e) 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.

In addition, MethylProphet is able to preserve cancer-specific DNAm differences (Figure A6 (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 and Val CpG - Val Sample splits show a

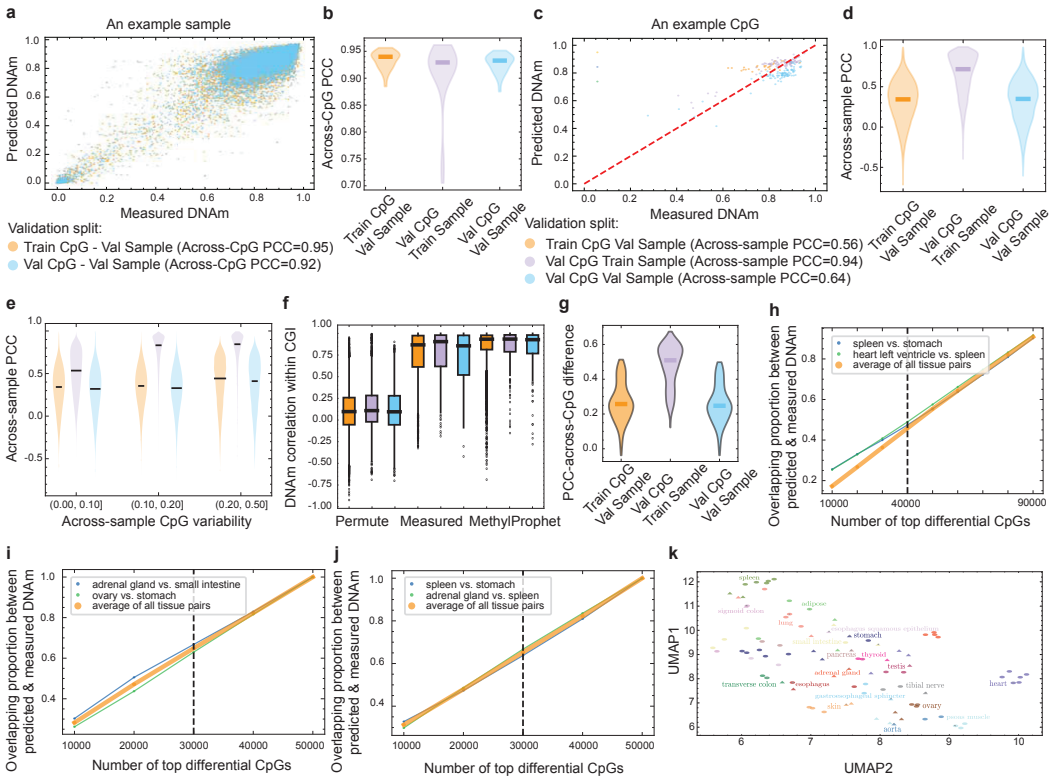

Figure A7: Cross-validation on ENCODE data. Similar to that of Figure A6, except that the results are based on the validation on ENCODE data. The sample differences (g) were calculated by comparing tissue/cell types rather than cancer types.

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 A6 (h-j)). In addition, MethylProphet-predicted DNAm landscape successfully preserves cancer-specific differences, as samples from the same cancer type remain well-clustered (Figure A6 (k)).

E.3 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 A7 (a, b)), the performance is similar across splits, while for across-sample PCC (Figure A7 (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 A7 (e)).

In this normal tissue cohort, MethylProphet also effectively captures CpG co-methylation dynamics within CGIs (Figure A7 (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 A7 (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 A7 (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 A7 (k)), with predicted and measured samples of the same cancer types cluster together.

F Discussion

F.1 Limitation and future work

This work should be regarded as a proof-of-concept study that demonstrates the feasibility of leveraging gene expression and genomic context for whole-genome DNA methylation inference. While MethylProphet introduces a new paradigm and achieves promising results, we do not propose fundamentally new model architectures nor do we systematically explore more efficient or specialized designs. Instead, our focus is on establishing baseline feasibility and potential, rather than optimizing for computational efficiency or architectural innovation. Future research could address these aspects by adopting alternative architectures or scaling strategies to further improve performance and resource efficiency.

F.2 Broad impact

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.