Computational and Structural Biotechnology Journal 27 (2025) 631–638

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Deep5mC: Predicting 5-methylcytosine (5mC) methylation status using a deep learning transformer approach

Evan Kinnear [a,1], Houssemeddine Derbel [a,1], Zhongming Zhao [b], Qian Liu [a,c,*]

[a] Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, 4505 S Maryland Pkwy, Las Vegas, NV 89154, USA
[b] Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[c] School of Life Sciences, College of Sciences, University of Nevada, Las Vegas, 4505 S Maryland Pkwy, Las Vegas, NV 89154, USA

ARTICLE INFO

Keywords:
DNA methylation prediction
Deep learning
Association of 5mC and genomic sequences

ABSTRACT

DNA methylations, such as 5-methylcytosine (5mC), are crucial in biological processes, and aberrant methylations are strongly linked to various human diseases. Genomic 5mC is not randomly distributed but exhibits a strong association with genomic sequences. Thus, various computational methods were developed to predict 5mC status based on DNA sequences. These methods generated promising achievements and overcome the limitations of experimental approaches. However, few studies have comprehensively investigated the dependency of 5mC on genomic sequences, and most existing methods focus on specific genomic regions. In this work, we introduce Deep5mC, a deep learning transformer-based method designed to predict 5mC methylations. Deep5mC leverages long-range dependencies within genomic sequences to estimate the probability of cytosine methylations. Through cross-chromosome evaluation, Deep5mC achieves Matthew's correlation coefficient over 0.86 and F1-score over 0.93, substantially outperforming state-of-the-art methods. Deep5mC not only confirms the influence of long-range sequence context on 5mC prediction but also paves the way for further studying 5mC-sequence dependency across species and in human diseases.

1. Introduction

DNA methylations play an indispensable role in various biological functions [1]. The most common form of DNA methylations in mammals is 5-methylcytosine (5mC). Other types of methylations, such as N6-methyladenine (6 mA) and 4-methylcytosine (4mC), are found in various organisms, while 5-carboxylcytosine (5caC) and 5-hydroxymethylcytosine (5hmC) are less prevalent in humans [2]. These epigenetic methylations are paramount for silencing genes [3], safeguarding against the activity of repetitive elements [4], maintaining genomic stability during mitosis [5], and imprinting genes based on their parental origin [6]. Abnormal alterations in DNA methylations have been found to precipitate a myriad of diseases, such as autoimmune rheumatic diseases and various forms of cancers [7,8], often exhibiting irregularities at gene promoters and regulatory regions [9,10]. Consequently, the identification of methylations is critical for a comprehensive understanding of the multifaceted roles of DNA methylations in human disorders.

Several high-throughput sequencing techniques have been developed to detect genome-wide 5mC, including bisulfite sequencing [11], oxidative bisulfite sequencing [12], PacBio single-molecule real-time (SMRT) sequencing [13], and Oxford nanopore sequencing [14]. However, these methods are often expensive and time-consuming. Thus, there is a pressing need for developing efficient computational methods to identify 5mC sites. These methods can offer a more cost-effective and rapid alternative for comprehensive analysis of DNA methylations, thereby advancing insights into epigenetic regulation and its health implications.

These computational methods can be categorized into three groups. The first group includes Methylator [15], MethCGI [16], and iDNA-Methyl [17], which used classical machine learning algorithms, such as support vector machines (SVM) [18], to predict methylations. These methods typically used short DNA sequences as input and were trained on small datasets. For example, iDNA-Methyl used hand-crafted features with 20 bp sequences, while Methylator was tested with DNA sequences of varying lengths (ranging from 9 bp to 89 bp) and trained using 39 bp sequences. MethCGI used longer DNA sequences (400 bp) as input but still relied on manually designed features. The second group of

* Corresponding author at: Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, 4505 S Maryland Pkwy, Las Vegas, NV 89154, USA.
E-mail address: qian.liu@unlv.edu (Q. Liu).
[1] Equal contribution

https://doi.org/10.1016/j.csbj.2025.02.007
Received 17 December 2024; Received in revised form 12 February 2025; Accepted 13 February 2025
Available online 14 February 2025

prediction methods were designed to identify cell-type specific methylations, such as iPromoter-5mC [19] and BiLSTM-5mC [20]. These methods used hand-crafted features extracted from 41 bp DNA sequences to train deep learning models on single-cell-type datasets to improve specificity. The third group, including 5mC_Pred [21], iDNA-ABF [22], BERT6mA [23] and DeepCpG [24], leveraged language models to extract high-level features for classification. While these studies have produced promising results, a comprehensive investigation of 5mC dependency on genomic sequences remains limited, and most methods focus on specific regions, such as promoters or cancer-related methylated sites, and their trained models are generally unavailable (except for iPromoter-5mC and 5mC_Pred).

In this study, we introduce Deep5mC, a novel deep learning method designed to identify 5mC methylation sites in DNA sequences using a transformer framework [25]. Deep5mC predicts 5mC status based on DNA sequences. Through cross-chromosome assessment, Deep5mC significantly outperforms existing methods, demonstrating that the occurrence of 5mC is closely linked to genomic sequences. Deep5mC can further be extended to comprehensively study the 5mC-sequence dependency across different species and conditions, including human diseases. Deep5mC is publicly available at https://github.com/qgenlab/Deep5mC.

2. Materials and methods

2.1. Datasets

The 5mC datasets for model training and testing were downloaded from the NIH roadmap epigenomics consortium [26]. These datasets are comprised of 5mC methylations across 37 distinct human epigenomes, covering chromosomes 1–22, X, Y, and M. As one of the most comprehensive collections of human epigenomes for primary cells and tissues, this dataset provides a robust foundation for studying DNA methylation patterns.

This methylation data was generated using whole genome bisulfite sequencing (WGBS) and consists of methylation percentages (ranging from 0 to 1) for CpG sites within these epigenomes. We preprocessed this data using the steps elucidated in Fig. 1. First, methylation sites were filtered out if they had low coverage (≤ 3) or were from chromosomes X, Y, or M, because these chromosomes have vastly different methylation characteristics from chromosomes 1–22 [27–29]. Second, the arithmetic mean of methylation percentages was computed for each methylation site across epigenomes, after excluding the highest and lowest values to mitigate the influence of outliers. The standard deviation of methylation percentages was calculated, and sites with a standard deviation > 0.1 were discarded to ensure methylation consistency across epigenomes. Following this filtering, 34,577,332 positive methylation sites and 4987,695 negative sites were retained. Fourth, we extracted a subsequence of 2561 bp centered each filtered site in the human reference genome, i.e, genome reference consortium human build 37 (GRCh37 or hg19). These subsequences served as input for transformers to learn sequence-based context information.

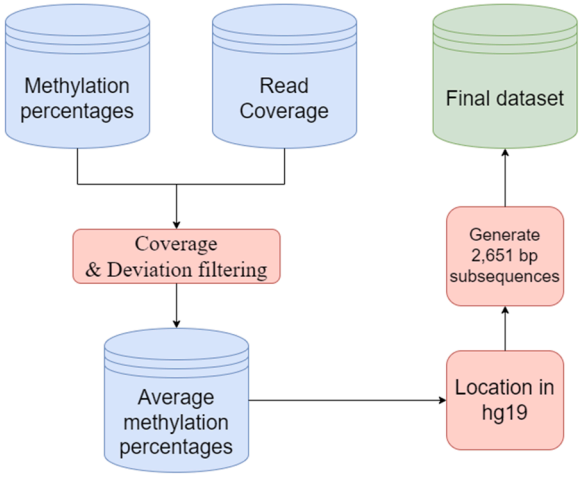

Fig. 1. Data preprocessing. Hg19: Genome Reference Consortium Human Build 37.

2.2. Deep learning framework

Our deep learning algorithm, named Deep5mC, is a transformer framework as illustrated in Fig. 2. Given an input DNA sequence, Deep5mC uses several steps listed below to predict methylation status: (1) a feature extractor transforms input sequences into one-hot encoding vectors incorporating position embedding and token embedding, (2) a transformer component generates representation vectors for each position using a transformer framework, and (3) two fully connected layers output prediction of a target position.

2.2.1. Feature generation

An input sequence contains low-level features of nucleotides, which were progressively combined to generate high-level features using a transformer model. To do so, an input sequence was converted to a vector through token embedding and position embedding.

Token embedding: In an input sequence, each nucleotide A, C, T, G and N (for unknown nucleotides) were denoted as a one-hot vector. In addition, [CLS], [SEP] and [PAD] tokens were included to mark the start and end of sequences as well as sequence padding, respectively. [CLS] and [SEP] are necessary for position embedding in transformers, while [PAD] ensures DNA sequences with varying lengths have the same number of embeddings: [PAD] is usually added to the encoding of shorter sequences so that all sequences have the same length of input in transformers. An example for sequence encoding is shown in Fig. 2.

Position embedding: To capture long-range dependencies in DNA sequences, Deep5mC employed position embedding. There are two types of position embedding: absolute position embedding and relative position embedding [30]. Deep5mC uses relative position embedding [30] because the distance between two bases is more critical than their absolute positions, and relative position embedding efficiently captures relative position representations or distances between nucleotides within sequences. Huang et al. [31] introduced a memory-efficient method of computing relative positional encoding representations, enabling the model to be informed by how far two positions are apart in a sequence. The relative attention used in Deep5mC is presented as.

$$\text{Relative Attention} = \text{Softmax}\left(\frac{QK^T + S_{rel}}{\sqrt{D_h}}\right)V$$

Where Q, V, and K are the query, value and key in a self-attention model, D_h is the dimension of one attention head h, $S_{rel} = QR^T$ and $R_{ij} = a_{ij}^K$, $QR^T = Skew(QE_r)$, E_r is the relative position embedding matrix. In Deep5mC, we utilized 16 attention heads with an embedding size D_h of 1024.

CNN Encoding: There are four main types of tokens in DNA sequences, and this limited-token property makes it challenging to capture high-level contextual information. To address this challenge, we incorporated convolutional neural networks (CNN) to capture local dependencies in DNA sequences by combining vectors of the current token with its adjacent neighboring bases. We applied a CNN layer with 1024 output channels and a kernel size of (3, 1024) to combine every three consecutive bases.

2.2.2. Transformer framework

DNA sequences share structural similarities with natural language sentences, making transformer methods a natural fit for sequence-based predictions. Transformer algorithms [25] rely on self-attention

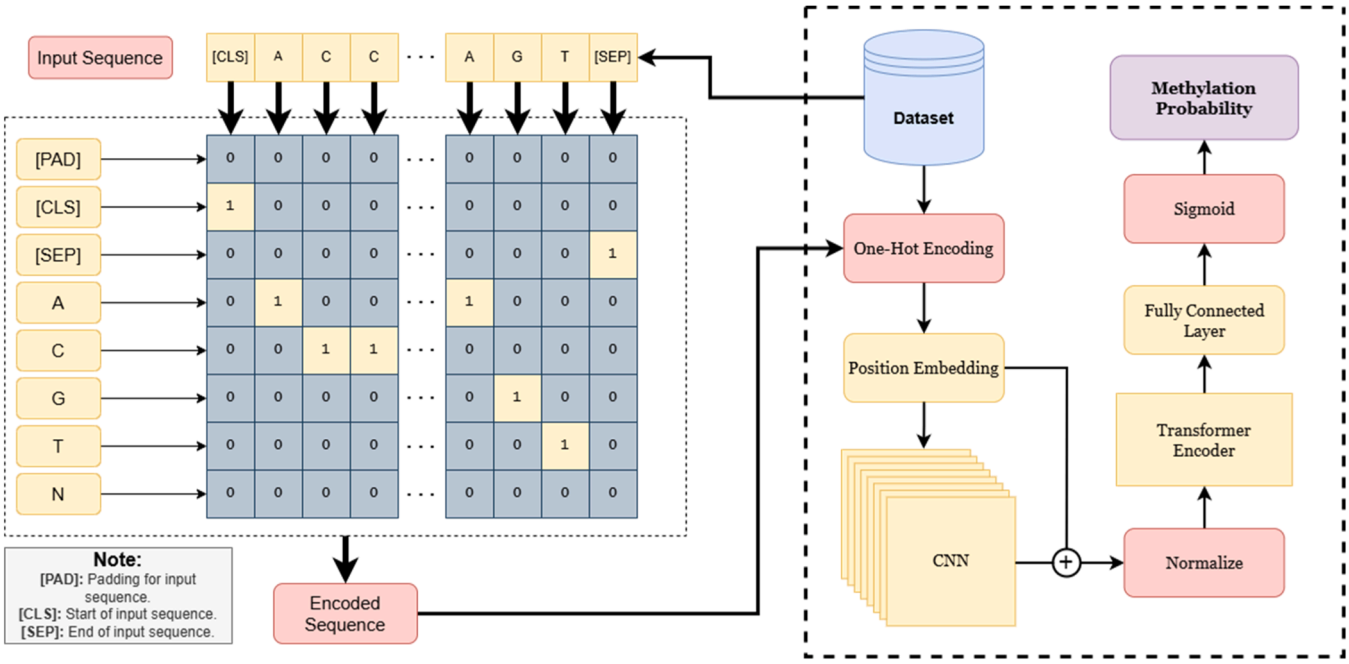

Fig. 2. Deep5mC architecture. The one-hot encoding, position embedding, and token embedding are for feature extractor.

mechanisms to process input data in parallel, making it highly efficient to process sequential data. Self-attention mechanisms enable transformers to weigh the importance of different elements in an input sequence and dynamically adjust their influence on the output. The self-attention is computed as follows:

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

Where Q, K, V are the concatenation of query, key, and value vectors, respectively. Given an input sequence $S = \{s_1, s_2, \ldots, s_n\}$ with its input matrix $M \in R^{n\times m}$, where n is the sequence length and m is the embedding size, Q, K, and V are calculated by multiplying M with three different learnable matrices W(K), W(Q), and W(V), with d_k is the dimension of the key vector. To enhance learning, Vaswani et al. [25] proposed multi-head attention which concatenates the output of different attention heads. In Deep5mC, we used a transformer model of 16 layers with 16 multi-heads and a hidden size of 1024.

These representation vectors learned through a transformer model were then combined into a single 1,024-element vector by a fully connected neural network, and the combined vector was then passed through a second fully connected layer to predict methylation percentage with a sigmoid activation function.

2.3. Training Deep5mC models

To evaluate 5mC predictions, we adopted a cross-chromosome testing strategy where the whole-genome data were split to training chromosomes and testing chromosomes. For each chromosome, we extracted subsequences of 2561 bp centered on CpG sites. We trained our model on training chromosomes, and generated evaluation performance on testing chromosomes. During the training process, we optimized the model using an Adam optimizer with a learning rate of 5e-6, a batch size of 512 and the loss function of mean absolute error (MAE) which is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \widehat{Y}_i|$$

Where n is the batch size, Y_i is a list of real methylation percentages and $\widehat{Y}_i$ is a list of predicted methylation percentages. To prevent overfitting, we applied 20% dropout [32] for all layers (except for the output layer).

Methylated levels in human genomes are not uniform but highly imbalanced, with CpG sites predominantly exhibiting either high or low methylation, as shown in Fig. 3 for chromosome 1. This imbalance poses a significant challenge for model training. To mitigate this issue, undersampling was applied by randomly selecting positions within methylated sites for training.

2.4. Comparison with state-of-the-art methods

Although several existing methods have been developed to predict 5mC status, only iPromoter-5mC and 5mC_Pred released well-trained models that were publicly available for downloading and testing (as of December 2024). Consequently, these two methods were the only options for comparison in this study, despite their limitations as ideal benchmarks.

iPromoter-5mC employs a straightforward deep forward network to identify methylated positions. It extracts diverse features from 41 bp subsequences, such as one-hot vectors for nucleotides, and deoxynucleotide property and frequency. iPromoter-5mC was trained and tested on data collected from the encyclopedia of cancer cell line [33] and UCSC genome browser [34]. It was specifically utilized for detecting 5mC modifications in promoter regions of cancer cells.

5mC_Pred leveraged natural language processing models to predict 5mC methylation percentages. It utilized a fastText [35] model to generate representation vectors of 1-mers to 3-mers from 41 bp subsequences, and then used various machine learning models (XGBoost, random forest, deep forest, and deep feed forward network) for methylation prediction. 5mC_Pred was trained and tested on the same dataset of iPromoter-5mC.

2.5. Evaluation metrics

Since each site can be considered to be either methylated or unmethylated, we used various classification metrics to evaluate prediction performance of Deep5mC and of the state-of-art methods. These metrics include sensitivity, specificity, precision, F1-score, accuracy, and Matthew's correlation coefficient (MCC), which are defined as follows:

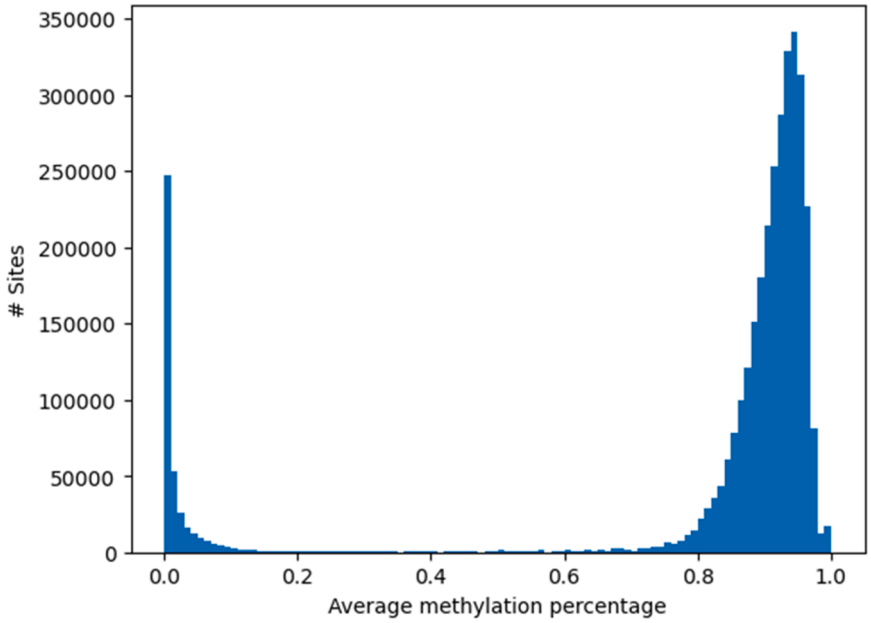

Fig. 3. Distribution of methylation percentages in Chromosome 1 after preprocessing, illustrating an imbalance of highly methylated and low methylated sites. Other chromosomes show similar methylation distributions.

$$Sensitivity = \frac{TP}{(TP + FN)}$$

$$Specificity = \frac{TN}{(TN + FP)}$$

$$Precision = \frac{TP}{(TP + FP)}$$

$$F1 - score = \frac{2 \times Sensitivity \times Precision}{(Sensitivity + Precision)}$$

$$Accuracy = \frac{TP + TN}{(TP + FN + FP + TN)}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Where TP is true positive (methylated) predictions of methylated sites, TN is true negative (unmethylated) predictions of unmethylated sites, FN is false negative (unmethylated) predictions of methylated sites and FP is false positive (methylated) predictions of unmethylated sites.

For regression assessment where each site is associated with a methylation percentage, we used Pearson correlation coefficient (r)

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \bullet \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Where x_i is a list of predicted values, $\overline{x}$ is the mean of predicted values, y_i is a list of expected values and $\overline{y}$ is the mean of the expected values.

3. Results

3.1. Prediction comparison with the state-of-the-art methods

We evaluated Deep5mC against two state-of-the-art approaches: iPromoter-5mC [19] and 5mC_Pred [19]. These two methods were the only ones with released well-trained models for 5mC prediction (as of 2024), although both were mainly tested on promoter regions. To ensure a fair comparison, we constructed a testing dataset using a subset of filtered sites located within 2000 bp upstream regions of transcription start sites in chromosomes 21 and 22. In addition, different existing methods used varying thresholds to define methylated (τ_m) and unmethylated (τ_{um}) sites: a site is classified as methylated if its methylation percentage is larger than τ_m or unmethylated if its methylation percentage is smaller than τ_{um}. To conduct a comprehensive evaluation, we applied three threshold pairs: $\tau_m = 0.95$ and $\tau_{um} = 0.05$ (95–5 dataset, similar to iPromoter-5mC), $\tau_m = 0.8$ and $\tau_{um} = 0.2$(80–20 dataset), as well as $\tau_m = 0.5$ and $\tau_{um} = 0.5$ (50–50 dataset, used by 5mC_Pred). With these threshold pairs, the number of methylated and unmethylated sites in testing datasets was shown in Table 1, and prediction performance was presented in Table 2.

As shown in Table 2, Deep5mC consistently outperformed the other models across all three datasets. Matthew's correlation coefficient (MCC) for Deep5mC was approximately 0.87, significantly higher than that achieved by iPromoter-5mC (by 0.88) and 5mC_Pred (by 0.87). Additionally, Deep5mC achieved markedly higher F1-scores, with average improvements of 0.59 over iPromoter-5mC and 0.94 over 5mC_Pred. These results suggest the superior capability of Deep5mC in accurately predicting both methylated and unmethylated sites.

Specifically, Deep5mC demonstrated substantial improvements in nearly all evaluated metrics. For example, it achieved superior sensitivity, with an average improvement of 0.67 over iPromoter-5mC and 0.90 over 5mC_Pred, underscoring its effectiveness in identifying methylated sites. While 5mC_Pred exhibited marginally higher specificity compared to Deep5mC, with improvements of 0.06, 0.03, and 0.07

Table 1
Summary of the testing datasets.

Threshold	# Methylated sites	# Unmethylated sites
95–5	8683	4272
80–20	31,730	4948
50–50	32,354	4948

Threshold: $\tau_m - \tau_{um}$. A site is classified as methylated if its methylation percentage is larger than τ_m or unmethylated if its methylation percentage is smaller than τ_{um}.

Table 2
Comparison of the performance of the proposed model with iPromoter-5mC and 5mC_Pred.

Measure	Deep5mC			iPromoter-5mC			5mC_Pred		
	95–5	80–20	50–50	95–5	80–20	50–50	95–5	80–20	50–50
Sensitivity	0.89	0.91	0.91	0.21	0.24	0.24	0.00	0.00	0.00
Specificity	0.95	0.92	0.91	0.65	0.66	0.66	0.98	0.98	0.98
Precision	0.97	0.98	0.98	0.56	0.82	0.82	0.31	0.57	0.57
Accuracy	0.91	0.91	0.91	0.36	0.30	0.30	0.32	0.13	0.13
MCC	0.88	0.87	0.86	−0.13	−0.07	−0.07	−0.06	−0.06	−0.06
F1-score	0.93	0.95	0.95	0.31	0.37	0.38	0.00	0.00	0.00

MCC: Matthew's correlation coefficient

for the 80–20, 95–5, and 50–50 datasets, respectively, Deep5mC achieved a balanced performance across both sensitivity and specificty.

Overall, the findings demonstrate the markedly superior efficacy of Deep5mC in discerning methylation patterns, revealing a strong connection between nucleotide context and methylation status in the human genomes.

3.2. Chromosome-level performance evaluation of Deep5mC

Beyond the ability to predict methylation status in promoter regions, Deep5mC can predict methylation status of CpG sites in other genomic regions. To assess its performance at a chromosome level, we evaluated the prediction of Deep5mC on all filtered CpG sites from two chromosomes 21 and 22, while training Deep5mC on chromosomes 1 through 20. The total number of CpG sites from chromosomes 21 and 22 are 1383,361 with 1197,784 methylated and 185,577 unmethylated positions using $\tau_m = 0.8$ and $\tau_{um} = 0.2$. On this dataset, Deep5mC achieved an accuracy of 0.87, an AUC of 0.93, showcasing the proficiency of Deep5mC in predicting methylation status. Moreover, predicted methylation percentages of Deep5mC correlated strongly with experimental methylation percentages (Pearson correlation=0.71), as illustrated in Fig. 4. This strong correlation demonstrates Deep5mC's ability to accurately predict methylated sites. Please note that a diagonal line does not appear in Fig. 4 because real genomic methylation data is dominated by highly methylated or highly unmethylated sites.

To further validate the robustness of Deep5mC's predictions, we randomly selected 4 chromosomes (chromosomes 7, 4, 17 and 11 with 8124,119 CpG sites) to train Deep5mC and then tested Deep5mC on the remaining 18 chromosomes with 31,440,860 CpG sites. The results demonstrate that Deep5mC maintained high prediction accuracy, achieving an MCC of 0.61, an accuracy of 0.9 and an F1-score of 0.9.

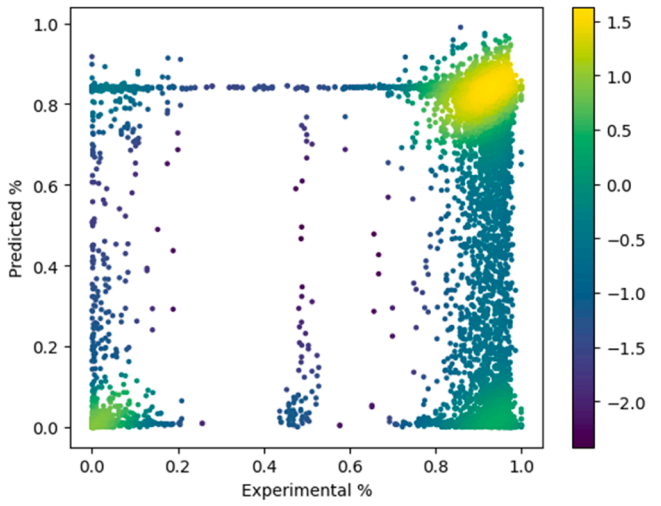

Fig. 4. Correlation of Deep5mC predicted methylation percentages (%) against experimental methylation percentages (%). Light yellow: high density of dots; Blue: lower density.

3.3. Effect of sequence length on the methylation prediction by Deep5mC

Deep5mC utilized 2561 bp sequences to detect nucleotide context information, while most existing methods rely on shorter (41 bp) sequences. To evaluate how sequence length affects Deep5mC's performance, we conducted investigations using various sequence lengths, ranging from 11 bp to 1281 bp, as summarized in Table 3. In these investigations, Deep5mC models were trained without CNN layers. The results revealed that longer sequences usually generated better performances. For example, training on 161 bp sequences improved an AUC to 0.74, while training on 11 bp sequences only achieved an AUC of 0.63. However, we also found that the best performance was achieved with the length of 321, and longer sequences slightly decreased the performance. Since this best performance is still lower than that of Deep5mC with CNN layers, more investigations are needed to determine an optimal sequence length for our model with CNN layers.

Nevertheless, these investigations suggest that longer sequences, rather than 41 bp sequences used in existing works, enable Deep5mC to capture long-range context information for predicting methylation profiles. These findings support a strong association between 5mC and long-range context in the human genome.

3.4. Long-range sequence dependency of 5mC

To visualize this dependency of 5mC status on long-range sequence context, we plotted the trained weights in the transformer model of Deep5mC in Fig. 5, which presents the learned positional attention patterns in Deep5mC. This visualization illustrated that Deep5mC captured information from bases located further away from the central position, supporting our conclusion of long-range dependency of 5mC. In addition, Deep5mC can detect 5mC patterns from genomic sequences, which can be extended to study 5mC-sequence dependency in human diseases.

3.5. Association of generated vectors and 5mC methylations

To better understand how Deep5mC predicts 5mC status, we analyzed the representation output of the transformer layer and the output of the first fully connected layer. We used uniform manifold approximation and projection (UMAP) [36] to create those vectors for

Table 3
Performance of Deep5mC for predicting methylation percentage on CpG sites in chromosome 21 and 22 with shorter sequences as input for Deep5mC. CNN layer was not used. Correlation: Pearson correlation coefficient.

Length (#bp)	Accuracy	Correlation	AUC
11	0.54	0.32	0.63
21	0.52	0.36	0.67
41	0.55	0.44	0.69
81	0.59	0.54	0.71
161	0.70	0.58	0.74
321	0.80	0.52	0.85
641	0.75	0.46	0.82
1281	0.71	0.43	0.78

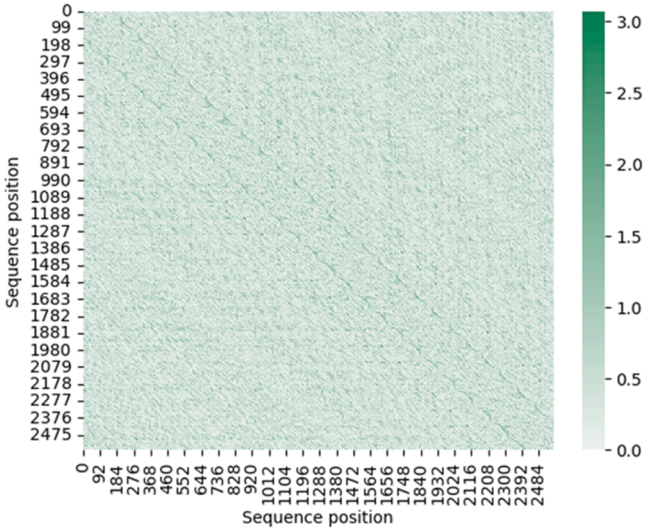

Fig. 5. The heatmap of self-attention scores of the positional encoding layer from head 0. Green indicates high weights, while white indicates low weights.

2D visualization, and the results were shown in Fig. 6. For this analysis, we used sequences from chromosome 22.

Fig. 6(a) clearly shows that the transformer layer effectively generated a meaningful representation of methylated and unmethylated sites based on genomic sequences. Following the processing through a fully connected layer, Fig. 6(b) presented two clusters: one corresponding to methylated sites and the other to unmethylated sites, although some methylated sites were misclassified. Since Deep5mC uses only genomic sequences as input to generate the representation, the clear separation of methylated and unmethylated sites in Fig. 6 suggests that 5mC methylations are strongly influenced by genomic sequence context, rather than occur randomly.

3.6. *Effect of transformer sizes on methylation prediction*

Deep5mC relies on a transformer with 16 layers and 16 multi-heads. While this hyperparameter setting has not been extensively optimized, we investigated the effect of different transformer sizes on methylation prediction performance. We trained a smaller transformer with 6 layers and 8 multi-heads, and the results were shown in Table 4. The smaller transformer achieved an accuracy of 0.58, Pearson correlation of 0.35, and an AUC of 0.86, which were significantly lower than the performance achieved by a transformer with 16 layers and 16 multi-heads. These results confirmed that larger transformer models improve 5mC predictions, as complex transformer models can capture more intricate patterns hidden in genomic sequences. However, training larger transformer models requires substantially more GPU resources. Given the excellent performance of the current model, Deep5mC offers a practical and effective solution for methylation predictions.

Table 4
Performance of Deep5mC for predicting methylation percentage on CpG sites in chromosome 21 and 22 with different transformer layer sizes. Correction: Pearson correlation coefficient.

#Layers	#Multi-Heads	Accuracy	Correlation	AUC
6	8	0.58	0.35	0.86
16	16	**0.87**	**0.71**	**0.93**

4. Discussion

In this study, we developed Deep5mC, an innovative deep learning-based method to identify 5mC methylations through biological language learning models. Using solely DNA sequences as input, Deep5mC discerned methylation patterns by integrating local dependencies via convolutional neural networks, long-range dependencies through transformer layers and relative position embedding. Trained on a comprehensive database encompassing nearly 10 million sequences pertinent to DNA 5mC methylations, Deep5mC demonstrated the capability to accurately predict methylation status across the majority of CpG sites in the human genome.

Despite relying on genomic sequences, Deep5mC generated accurate prediction of methylation status, compellingly suggesting that Deep5mC could be instrumental in exploring that 5mC methylations do not randomly occur in the human genome but are associated with genomic context. In addition, the investigations with different transformer configurations revealed that (1) the association of 5mC and genomic sequences is complicated, which needs to be learned by larger transformers with more layers and more multi-heads, and (2) long-range genomic context (>300 bp) is necessary to accurately detect methylation status. These observations provide valuable recommendations for designing DNA sequence-based transformer models in future.

Also, DNA methylations may be regulated by chromatin

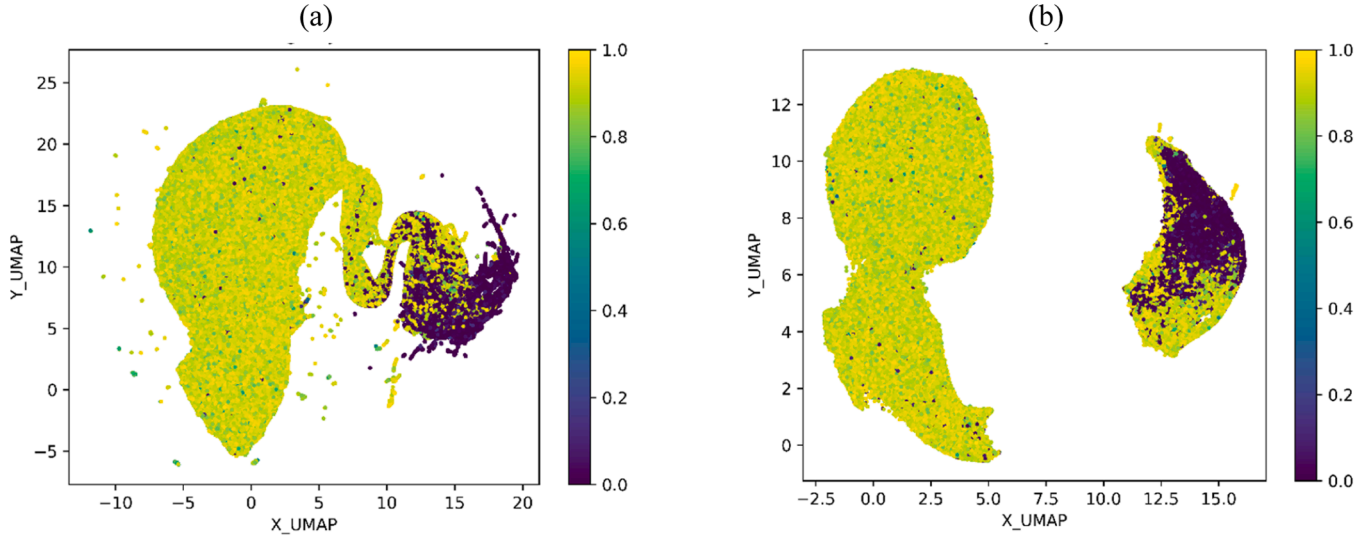

Fig. 6. Visualization after dimensionality reduction with UMAP for representation vector generated by the transformer layer (a) and for the output generated by the first fully connected layer (b). The instances were from chromosome 22. Yellow: methylated; purple: unmethylated.

accessibility, histone modifications and transcription factors. These factors offer dynamic regulations in response to various stimuli and environments. Thus, when more multi-omics datasets are available for individual genomes, we will integrate these regulatory factors into Deep5mC, not only improving 5mC prediction performance but enabling cell- and tissue-specific methylation detection. In this multi-omics framework, we will design specific transformer models to detect chromatin accessibility, histone modification and transcription factor binding, and then integrate output of Deep5mC and those of these specific transformer models to make the final methylation prediction. Furthermore, while Deep5mC focuses on the human genome, it is highly beneficial to extend Deep5mC for methylation detection in other eukaryotic genomes, such as mouse and zebrafish, where 5mC is a common chemical modification. When developing across-eukaryotic-genome Deep5mC, transfer learning is a promising approach for addressing the issue of limited training data in some species. Moreover, across-species studies could also uncover species-specific patterns of DNA methylations.

Finally, Deep5mC has several limitations. First, Deep5mC currently does not differentiate cell-type-specific or disease-associated methylation patterns. Further improvements will incorporate cell types, diseases or other epigenetic information as additional embedded input to generate condition-specific methylations. Second, Deep5mC is designed to predict 5mC methylations and does not cover other DNA methylation types. We will extend Deep5mC to predict other types of methylations, such as 6 mA and 4mC. Given the scarcity of labeled training data for those methylation types, transfer learning can be used to adapt 5mC models. Third, GPU resources are recommended to train Deep5mC and use Deep5mC, as it is time-consuming to train and test a larger transformer model. For example, on the training set with 76,363,332 genomic sites, it took approximately 72 h to train Deep5mC on a NVIDIA A100 GPU. On the testing dataset consisting of 37,302 genomic sites (Table 1, 50–50 threshold), Deep5mC prediction takes about 3 h on a NVIDIA A100 GPU to generate predictions with a batch size of 16. Without GPU acceleration, training and inference times would be significantly longer, limiting accessibility.

5. Conclusion

In this work, we introduced and rigorously evaluated Deep5mC, an advanced transformer-based model designed to predict 5mC methylation status within the human genome. Comparative analyses across diverse evaluation strategies demonstrated that Deep5mC significantly outperformed existing methods in detecting 5mC methylation patterns. These findings compellingly suggested that Deep5mC could empower the study of the association of 5mC methylations and genomic sequences. Moreover, Deep5mC can be extended to other eukaryotic genomes and applied to disease-specific studies.

CRediT authorship contribution statement

Kinnear Evan: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Data curation. **Liu Qian:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Zhao Zhongming:** Writing – review & editing, Validation, Data curation. **Derbel Houssemeddine:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis.

Declaration of Competing Interest

The authors claim that there is no conflict of interest.

Acknowledgements

The authors would like to thank the lab members for their valuable discussion and the reviewers for valuable comments. This research was funded by the National Institute of General Medical Sciences grant number P20GM121325. ZZ was partially supported by National Institutes of Health grants (R01LM012806, R01LM012806-07S1, U01AG079847) and the Cancer Prevention and Research Institute of Texas grant (CPRIT RP240610 and RP180734). The funders had no role in the study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

References

[1] Bird A. DNA methylation patterns and epigenetic memory. Genes Dev 2002;16:6–21. https://doi.org/10.1101/gad.947102.
[2] Chowdhury B, Cho I-H, Hahn N, Irudayaraj J. Quantification of 5-methylcytosine, 5-hydroxymethylcytosine and 5-carboxylcytosine from the blood of cancer patients by an enzyme-based immunoassay. Anal Chim Acta 2014;852:212–7. https://doi.org/10.1016/j.aca.2014.09.020.
[3] Jones PA, Takai D. The role of DNA methylation in mammalian epigenetics. Science 2001;293:1068–70. https://doi.org/10.1126/science.1063852.
[4] Bestor TH. The DNA methyltransferases of mammals. Hum Mol Genet 2000;9:2395–402. https://doi.org/10.1093/hmg/9.16.2395.
[5] Ming X, Zhang Z, Zou Z, Lv C, Dong Q, He Q, Yi Y, Li Y, Wang H, Zhu B. Kinetics and mechanisms of mitotic inheritance of DNA methylation and their roles in aging-associated methylome deterioration. Cell Res 2020;30:980–96. https://doi.org/10.1038/s41422-020-0359-9.
[6] Bartolomei MS, Ferguson-Smith AC. Mammalian genomic imprinting. a002592–a002592 Cold Spring Harb Perspect Biol 2011;3. https://doi.org/10.1101/cshperspect.a002592.
[7] Su Z, Han L, Zhao Z. Conservation and divergence of DNA methylation in eukaryotes. Epigenetics 2011;6:134–40. https://doi.org/10.4161/epi.6.2.13875.
[8] Wang Q, Jia P, Cheng F, Zhao Z. Heterogeneous DNA methylation contributes to tumorigenesis through inducing the loss of coexpression connectivity in colorectal cancer. Genes Chromosomes Cancer 2015;54:110–21. https://doi.org/10.1002/gcc.22224.
[9] Nishiyama A, Nakanishi M. Navigating the DNA methylation landscape of cancer. Trends Genet 2021;37:1012–27. https://doi.org/10.1016/j.tig.2021.05.002.
[10] Ballestar E, Sawalha AH, Lu Q. Clinical value of DNA methylation markers in autoimmune rheumatic diseases. Nat Rev Rheumatol 2020;16:514–24. https://doi.org/10.1038/s41584-020-0470-9.
[11] Li Y, Tollefsbol TO. DNA methylation detection: bisulfite genomic sequencing analysis. In: Tollefsbol TO, editor. Epigenetics Protocols Methods in Molecular Biology. Humana Press; 2011. p. 11–21. https://doi.org/10.1007/978-1-61779-316-5_2.
[12] Booth Michael J, Ost Tobias WB, Beraldi Dario, Bell Neil M, Branco Miguel R, Reik Wolf, Balasubramanian Shankar. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nat Protoc 2013;8:1841–51.
[13] Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res 2018;46:2159–68. https://doi.org/10.1093/nar/gky066.
[14] Lin B, Hui J, Mao H. Nanopore technology and its applications in gene sequencing. Biosensors 2021;11:214. https://doi.org/10.3390/bios11070214.
[15] Bhasin M, Zhang H, Reinherz EL, Reche PA. Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS Lett 2005;579:4302–8. https://doi.org/10.1016/j.febslet.2005.07.002.
[16] Fang F, Fan S, Zhang X, Zhang MQ. Predicting methylation status of CpG islands in the human brain. Bioinformatics 2006;22:2204–9. https://doi.org/10.1093/bioinformatics/btl377.
[17] Liu Z, Xiao X, Qiu W-R, Chou K-C. iDNA-methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal Biochem 2015;474:69–77. https://doi.org/10.1016/j.ab.2014.12.009.
[18] Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Their Appl 1998;13:18–28.
[19] Zhang L, Xiao X, Xu Z-C. iPromoter-5mC: a novel fusion decision predictor for the identification of 5-methylcytosine sites in genome-wide DNA promoters. Front Cell Dev Biol 2020;8:614. https://doi.org/10.3389/fcell.2020.00614.
[20] Cheng X, Wang J, Li Q, Liu T. BiLSTM-5mC: a bidirectional long short-term memory-based approach for predicting 5-methylcytosine sites in genome-wide DNA promoters. Molecules 2021;26:7414. https://doi.org/10.3390/molecules26247414.
[21] Nguyen T-T-D, Tran T-A, Le N-Q-K, Pham D-M, Ou Y-Y. An extensive examination of discovering 5-methylcytosine sites in genome-wide DNA promoters using machine learning based approaches. IEEE/ACM Trans Comput Biol Bioinform 2022;19:87–94. https://doi.org/10.1109/TCBB.2021.3082184.
[22] Jin J, Yu Y, Wang R, Zeng X, Pang C, Jiang Y, Li Z, Dai Y, Su R, Zou Q, et al. iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. Genome Biol 2022;23:219. https://doi.org/10.1186/s13059-022-02780-1.

[23] Tsukiyama S, Hasan MM, Deng H-W, Kurata H. BERT6mA: prediction of DNA N6-methyladenine site using deep learning-based approaches. Brief Bioinform 2022; 23:bbac053. https://doi.org/10.1093/bib/bbac053.
[24] Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol 2017;18:67. https://doi.org/10.1186/s13059-017-1189-z.
[25] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. Attention is All you Need.
[26] Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317–30. https://doi.org/10.1038/nature14248.
[27] Lund JB, Li S, Christensen K, Mengel-From J, Soerensen M, Marioni RE, Starr J, Pattie A, Deary IJ, Baumbach J, et al. Age-dependent DNA methylation patterns on the Y chromosome in elderly males. Aging Cell 2020;19:e12907. https://doi.org/10.1111/acel.12907.
[28] Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, Dupre Y, Antonarakis SE. DNA methylation profiles of human active and inactive X chromosomes. Genome Res 2011;21:1592–600. https://doi.org/10.1101/gr.112680.110.
[29] Devall M, Soanes DM, Smith AR, Dempster EL, Smith RG, Burrage J, Iatrou A, Hannon E, Troakes C, Moore K, et al. Genome-wide characterization of mitochondrial DNA methylation in human brain. Front Endocrinol 2023;13: 1059120. https://doi.org/10.3389/fendo.2022.1059120.
[30] Shaw, P., Uszkoreit, J., and Vaswani, A. (2018). Self-Attention with Relative Position Representations. Preprint at arXiv.
[31] Huang, C.-Z.A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A.M., Hoffman, M.D., Dinculescu, M., and Eck, D. (2018). Music Transformer. Preprint at arXiv.
[32] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014; 15:1929–58.
[33] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483: 603–7. https://doi.org/10.1038/nature11003.
[34] Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN, et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res 2019;47:D853–8. https://doi.org/10.1093/nar/gky1095.
[35] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017;5:135–46. https://doi.org/10.1162/tacl_a_00051.
[36] McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at arXiv.