*GigaScience*, 2023, **12**, 1–11
DOI: 10.1093/gigascience/giad054
Research

# MuLan-Methyl—multiple transformer-based language models for accurate DNA methylation prediction

Wenhuan Zeng [1], Anupam Gautam [1,2,3] and Daniel H. Huson [1,2,3,*]

[1]Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany
[2]International Max Planck Research School "From Molecules to Organisms", Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany
[3]Cluster of Excellence: EXC 2124: Controlling Microbes to Fight Infection, University of Tübingen, 72076 Tübingen, Germany
*Correspondence address. Daniel H. Huson, Sand 14, University of Tübingen 72076 Germany. E-mail: daniel.huson@uni-tuebingen.de

## Abstract

Transformer-based language models are successfully used to address massive text-related tasks. DNA methylation is an important epigenetic mechanism, and its analysis provides valuable insights into gene regulation and biomarker identification. Several deep learning–based methods have been proposed to identify DNA methylation, and each seeks to strike a balance between computational effort and accuracy. Here, we introduce MuLan-Methyl, a deep learning framework for predicting DNA methylation sites, which is based on 5 popular transformer-based language models. The framework identifies methylation sites for 3 different types of DNA methylation: N6-adenine, N4-cytosine, and 5-hydroxymethylcytosine. Each of the employed language models is adapted to the task using the "pretrain and fine-tune" paradigm. Pretraining is performed on a custom corpus of DNA fragments and taxonomy lineages using self-supervised learning. Fine-tuning aims at predicting the DNA methylation status of each type. The 5 models are used to collectively predict the DNA methylation status. We report excellent performance of MuLan-Methyl on a benchmark dataset. Moreover, we argue that the model captures characteristic differences between different species that are relevant for methylation. This work demonstrates that language models can be successfully adapted to applications in biological sequence analysis and that joint utilization of different language models improves model performance. Mulan-Methyl is open source, and we provide a web server that implements the approach.

**Keywords:** DNA methylation, natural language processing, model ensemble, model explainability, web server

**Key Points:**

- MuLan-Methyl aims at identifying 3 types of DNA methylation sites.
- It uses an ensemble of 5 transformer-based language models, which were pretrained and fine-tuned on a custom corpus.
- The self-attention mechanism of transformers give rise to importance scores, which can be used to extract motifs.
- The method performs favorably in comparison to existing methods.
- The implementation can be applied to chromosomal sequences to predict methylation sites.

## Introduction

DNA methylation is an important biological process. It facilitates epigenetic regulation of gene expression, is associated with various medical disorders [1–3], and has other applications, such as a marker in metagenomic binning [4]. While DNA methylation is a dynamic process, existing machine learning techniques are able to predict DNA methylation states from genomic sequence with some degree of accuracy.

There are several types of DNA methylation that differ by which methyl group is attached to which type of nucleotide in the sequence. Here, we focus on 6-methyladenine (6mA), 5-hydroxymethylcytosine (5hmC), and 4-methylcytosine (4mC) methylation [5–7]. Different organisms exhibit different patterns of methylation, and this gives rise to the computational problem of predicting the location of methylation sites for a given genome sequence. While much algorithmic work has been done on the question, recent work has focused on the application of deep learning methods [8, 9]. However, there is room for improvement of accuracy and comprehensiveness.

A large number of studies address the problem of identifying methylation sites, but most of them focus on a specific form of modification [10–29], and only a few methods address all 3 types of methylation mentioned above [30–34], in particular iDNA-MS, iDNA-ABT, and iDNA-ABF. The database presented in [31, 35] is now widely used as a benchmark dataset for assessing model performance [21, 23, 32–34].

While different deep learning–based methods all address the same goal, they differ in the details of the features employed and the model structure. Input features include an encoding of the sequence, of course, but may also include biochemical properties [10, 12] or a DNA molecular graph representation [22]. Utilized model structures include convolutional neural networks, graph convolutional neural networks, bidirectional encoder representation from transformers (BERT) [36], and other types of machine learning algorithms. The specific choices made during feature engineering and model selection determine performance and are key to proposing a new framework.

**Received:** February 19, 2023. **Revised:** May 9, 2023. **Accepted:** July 18, 2023

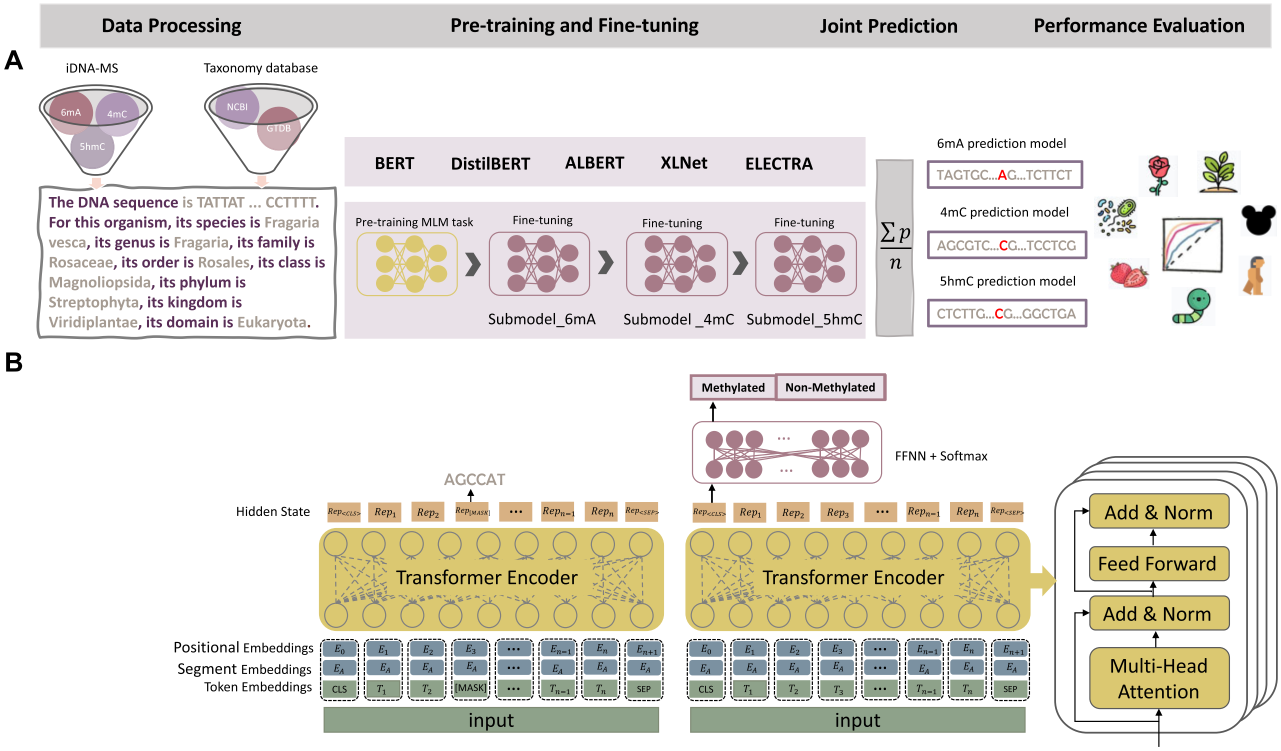

Data Processing
Pre-training and Fine-tuning
Joint Prediction
Performance Evaluation
A
iDNA-MS
6mA
4mC
5hmC
Taxonomy database
NCBI
GTDB
The DNA sequence is TATTAT ... CCTTTT. For this organism, its species is Fragaria vesca, its genus is Fragaria, its family is Rosaceae, its order is Rosales, its class is Magnoliopsida, its phylum is Streptophyta, its kingdom is Viridiplantae, its domain is Eukaryota.
BERT
DistilBERT
ALBERT
XLNet
ELECTRA
Pre-training MLM task
Fine-tuning
Fine-tuning
Fine-tuning
Submodel_6mA
Submodel _4mC
Submodel_5hmC
$\sum p / n$
6mA prediction model
TAGTGC...AG...TCTTCT
4mC prediction model
AGCGTC...CG...TCCTCG
5hmC prediction model
CTCTTG...CG...GGCTGA
B
AGCCAT
Hidden State
$Rep_{<CLS>}$ $Rep_1$ $Rep_2$ $Rep_{[MASK]}$ ⋯ $Rep_{n-1}$ $Rep_n$ $Rep_{<SEP>}$
Transformer Encoder
Positional Embeddings
Segment Embeddings
Token Embeddings
$E_0$ $E_1$ $E_2$ $E_3$ ⋯ $E_{n-1}$ $E_n$ $E_{n+1}$
$E_A$
CLS $T_1$ $T_2$ [MASK] ⋯ $T_{n-1}$ $T_n$ SEP
input
Methylated
Non-Methylated
FFNN + Softmax
$Rep_{<CLS>}$ $Rep_1$ $Rep_2$ $Rep_3$ ⋯ $Rep_{n-1}$ $Rep_n$ $Rep_{<SEP>}$
Transformer Encoder
CLS $T_1$ $T_2$ $T_3$ ⋯ $T_{n-1}$ $T_n$ SEP
input
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention

**Figure 1:** The MuLan-Methyl workflow. (A) The framework employs 5 fine-tuned language models for joint identification of DNA methylation sites. Methylation datasets (obtained from iDNA-MS) were processed as sentences that describe the DNA sequence as well as the taxonomy lineage, giving rise to the processed training dataset and the processed independent set. For each transformer-based language model, a custom tokenizer was trained based on a corpus of the processed training dataset and taxonomy lineage data. Pretraining and fine-tuning were both conducted on each methylation site-specific training subset separately. During model testing, the prediction of a sample in the processed independent test set is defined as the average prediction probability of the 5 fine-tuned models. We thus obtained 3 methylation type-wise prediction models. We evaluated the model performance on the genome type that contained the corresponding methylation type-wise dataset, respectively. In total, we evaluated 17 combinations of methylation types and taxonomic lineages. (B) The general transformer-based language model architecture for pretraining and fine-tuning. The model was pretrained using the masked language modeling (MLM) task and then fine-tuned on the methylation type-wise processed training dataset.

Here, we phrase DNA methylation site detection as a natural language processing (NLP) problem and propose a novel framework to address it. Previous studies for identifying methylation sites usually use BERT, a classic NLP approach, or, in the context of DNA sequences, the variant DNABERT [37], either as a model that accepts embeddings from Word2vec or as an encoder that generates embeddings for input to a deep neural network [23, 25, 32, 33, 38]. Only few published approaches aim at predicting multiple DNA modification sites. Moreover, many do not use taxonomic information as explicit features, although the taxonomic identity of an organism has an impact on DNA methylation [39]. Here we address both shortcomings by providing a new framework that uses a set of collective training language models, including but not limited to BERT, to predict 3 types of methylation sites from DNA sequences and taxonomic information.

Combining the transformer-based language model BERT with the "pretrain and fine-tune" paradigm has become the method of choice in NLP applications. In the pretraining step, self-supervised learning of the masked language modeling (MLM) task and the next sequence prediction task is usually performed on a corpus consisting of Wikipedia and books. This allows the transformer-based language model to capture the semantics of text input and contextual information exceptionally well. Transformer-based language models dynamically learn the input's representation through a multihead self-attention mechanism [40], and this leads to an improvement of prediction over classification models constructed using static embedding approaches [41]. The fine-tuning step involves supervised training of the pretrained language model to adapt to specific downstream tasks, here the prediction of 3 different types of methylation sites. Using BERT as a starting point, and then varying the network architecture and parameters, one can obtain 5 different language models, [42–46]. By pretraining on a domain-specific custom corpus, BERT can be adapted to a specific application scenario [47–50]. While the analysis of DNA sequences can be considered an application of NLP, using language models that are trained on human languages will not do well at capturing nucleotide rules. To address this, several approaches, such as BERTax, DNABERT, and LOGO [37, 51, 52], use large amounts of genomic sequence, instead of Wikipedia, as a corpus or similar structure.

The main aim of this article is to introduce MuLan-Methyl, a novel deep learning framework that combines 5 transformer-based language models to collectively predict sites for 3 different types of methylation (see Fig. 1A). In this approach, each methylation site sample is written as a sentence that represents the surrounding DNA sequence and the taxonomic identity of the corresponding genome. The output of our model is based on the average of the prediction probabilities obtained by 5 transformer-based language models: BERT [36], DistilBERT [42], ALBERT [46], XLNet [44], and ELECTRA [45].

Each of the 5 language models is trained according to the "pretrain and fine-tune" paradigm. For this, we used a custom corpus that contains the processed training dataset and taxonomic lineage information downloaded from NCBI [53] and GTDB [54]. For each language model, we trained a custom tokenizer on the custom corpus, using the same configuration as the model's default tokenizer. We use a customized tokenizer to ensure that the represented DNA sequences and taxonomic information associated with each sample are captured effectively. Each language model was pretrained by training the MLM task on the processed training dataset. We then obtained the 6mA model by fine-tuning the pretrained language model using the 6mA training dataset. Next, the 4mC prediction model was obtained by fine-tuning the 6mA prediction model using the 4mC training dataset. Finally, the 5hmC prediction model was obtained by fine-tuning the 4mC prediction model using the 5hmC training dataset. In addition, we compared the performance of all models contained in MuLan-Methyl.

A main contribution of this work is that we use both DNA sequence and taxonomic identity as explicit features in the model. Using the iDNA-MS [31] independent test set as a benchmark, our approach shows improved performance over previous methods, especially for certain genomes. MuLan-Methyl is capable of making accurate predictions for genomes whose taxonomy lineage is not present in the training dataset. The interpretability of MuLan-Methyl facilitates the discovery of DNA motifs that are associated with DNA methylation and potential correlations between specific methylation sites and taxonomic lineages.

To the best of our knowledge, this is the first application in biology that achieves improved prediction performance by integrating multiple transformer-based language models. This shows that adding features to a model is not the only way to improve the accuracy of predictions.

# Materials and Methods

## Data processing

### *Data collection*

We downloaded a DNA methylation dataset from the iDNA-MS web server [55]. This is an open resource that was published with the iDNA-MS method [31] and is widely used for benchmarking. The dataset contains 3 main types of DNA methylation sites (6mA, 4mC, and 5hmC) across 12 genomes (1 bacteria and 11 eukaryotes), in total 250,599 positive samples. In addition, the dataset provides the same number of nonmethylation sequences as negative samples.

The dataset is partitioned into a training set and an independent test set at a 1:1 ratio. The training dataset provides samples for methylation type 6mA present in 11 different species. In more detail, the numbers are 53,800 for *T. thermophile*, 15,937 for *Arabidopsis thaliana*, 9,168 for *Homo sapiens*, 8,608 for *Xoc. BLS256*, 5,596 for *Drosophila melanogaster*, 3,981 for *Caenorhabditis elegans*, 3,033 for *Casuarina equisetifolia*, 1,893 for *Saccharomyces cerevisiae*, 1,690 for *Tolypocladium*, 1,551 for *Fragaria vesca*, and 300 for *Rosa chinensis*. The 4mC methylation type is present in 4 species, where the numbers of samples are 7,899 for *F. vesca*, 7,664 for *Tolypocladium*, 990 for *S. cerevisiae*, and 183 for *C. equisetifolia*. Finally, the numbers of samples for the type 5hmC are 1,840 for *Mus musculus* sequences and 1,172 for *H. sapiens*. More detailed statistics are provided in Supplementary Table S1.

Each sample is a DNA segment of length 41, which is centered on an experimentally verified methylation site, in the case of a positive sample.

### *Dataset preparation*

We processed each sample (DNA sequence of length 41) as follows. Using a sliding window of length 6, we extracted $36 = 41 - 6 + 1$ individual 6-mers from the DNA sequence and embedded these into a sentence, together with a description of the taxonomic lineage of the corresponding organism, which was phrased as follows: "For this organism, its species is *species*, its genus is *genus*, its family is *family*, its order is *order*, its class is *class*, its phylum is *phylum*, its kingdom is *kingdom*, its domain is *domain*." We refer to a set of sentences obtained from a set of samples as a "processed dataset." The full processed training dataset, containing all 3 types of methylation sites, was put into the custom corpus. For purposes of fine-tuning, both the processed training dataset and the processed independent test set were split into 3 sets by methylation type.

### *Corpus generation*

We require a custom corpus for pretraining each language model to allow the model to learn and capture domain-specific words, which are not contained in a text corpus such as Wikipedia. The custom corpus contains the processed training dataset, as mentioned above. In addition, to enable the language to detect words about taxonomy, we incorporated all taxonomic lineages from the NCBI and GTDB taxonomies that are not already contained in the training datasets. In total, the corpus contains 2,440,894 sentences and uses a vocabulary of 25,000 words.

### *External dataset*

We downloaded DNA methylation data published with the Hyb4mC method [16] and with the i6mA-pred method [56], respectively. As these "external" data are not contained in our training or independent datasets, nor do the associated taxonomic lineages coincide, it is ideal for evaluating the performance of MuLan-Methyl more broadly. In more detail, these data consist of samples (DNA sequences of length 41) representing 320 4mC site sequences in *Escherichia coli*, 1,926 4mC site sequences in *Geobacter pickeringii*, and 880 6mA site sequences in *Oryza sativa* L., each with the same number of negative samples, respectively.

## Training transformer-based language models

We pretrained and fine-tuned 5 transformer-based language models. In the following, we first describe the architecture of each of the 5 employed language models. We then discuss the details of the training process for the first method, BERT (RRID:SCR_018008), including tokenization, pretraining, and fine-tuning (see Fig. 1B). The other 4 languages are trained in a similar way.

All code is written in Python 3.10, using the Pytorch and the Huggingface Transformers library [57]. The experiments were run on a Linux Virtual Machine (Ubuntu 20.04 LTS) equipped with 4 GPUs provided by de.NBI (flavor: de.NBI RTX6000 4 GPU medium).

### *Transformer-based language models*

Our approach uses 5 transformer-based language models, which we introduce in the following.

(1) BERT is capable of modeling bidirectional contexts, using denoising and autoencoding-based pretraining. The transformer architecture of $\text{BERT}_{\text{base}}$ uses 12 layers in the encoder stack, 768 hidden units for feed-forward networks, and 12 attention heads; it has 110 million parameters in total.

(2) A distilled version of BERT, DistilBERT, is obtained by decreasing the number of layers. It has 40% the size of BERT and is 60% faster, while only being 3% less accurate.

(3) ALBERT adopts a cross-layer parameter sharing technique for 12 transformer encoder blocks and imports embedding

factorization between vocabulary and the hidden layer in order to reduce the parameter size of BERT.

(4) XLNet uses an innovative pretraining step; its generalized autoregressive pretraining method enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order, overcoming the issues caused by BERT's neglect of dependency between masked positions.

(5) Using a different architecture, ELECTRA trains 2 transformer models; a generator replaces tokens in a sequence and a discriminator tries to identify which tokens were replaced by the generator, instead of training on the MLM task.

### *Custom tokenizer*

A tokenizer must be used to convert samples into the format that is expected by the transformer block of a language. In our study, such a tokenizer is obtained by training the language's default tokenizer on our custom corpus. Once trained, the tokenizer can capture any sample represented by a sentence consisting of 6-mer DNA words and a textual description of taxonomic lineage.

After tokenization, each input sample is represented by a list of tokens, starting and ending with special tokens [CLS] and [SEP], respectively, and padded to a length of 100 using padding tokens [PAD].

### *Model pretraining*

The BERT language model is pretrained by performing unsupervised training of the MLM task on the custom corpus. Pretraining was conducted on the model using an architecture that is the same as *bert-base-uncase* but with setting the embedding size of input to 25,000 to match the vocabulary size of the corpus.

During training of the MLM task, 15% of all WordPiece tokens of a sample are selected at random as masking candidates. Of these, 80% are replaced by a special token [MASK], and 10% are replaced by a random token. Then the original tokens are predicted.

Pretraining was conducted using 8 epochs, a batch size of 64 per GPU, and a learning rate of 5e-4, which is achieved after 100 steps of warmup.

### *Model fine-tuning*

Fine-tuning is performed for each of the 3 methylation site types separately, and so the processed training dataset is split into 3 training subsets, 6mA, 4mC, and 5hmC, listed in order of decreasing size. Each training subset is split into a training set and a validation set at a ratio of 8:2. The target model used to be fine-tuned depended on the subset's size. First, for the 6mA subset, we simply fine-tuned the pretrained language model that was trained on the custom corpus. Second, the 4mC fine-tuned model was then obtained by fine-tuning the 6mA fine-tuned model. Finally, the 5hmC fine-tuned model was obtained by fine-tuning the 4mC fine-tuned model. We fine-tune the fine-tuned models in this way to make the predictions more accurate on the smaller training subsets.

In all 3 cases, fine-tuning was performed using an early-stopping strategy, with a maximum of 32 epochs, a batch size of 64 per GPU, and a learning rate of 1e-5, which is achieved after 100 steps of warmup.

## Multilanguage model

For each of the 3 types of methylation sites, 5 language models are trained and then the MuLan-Methyl framework integrates these, computing prediction probabilities that are obtained by averaging over the probabilities returned by the 5 models.

## Interpretability of MuLan-Methyl

Transformer-based language models learn different and distant dependencies in the input, by virtue of the multihead self-attention mechanisms that are present in each encoding layer. For example, BERT contains 12 encoder layers containing 12 attention heads each. For 1 layer, the multihead self-attention can be described as

$$\mathrm{MultiHead}\left(Q, K, V\right) = \mathrm{Concat}\left(head_1, \cdots, head_n\right) W^O.$$

Here, the *i*th single attention head is computed as

$$\mathrm{head_i} = \mathrm{Attention}\left(QW_i{}^Q, KW_i{}^K, VW_i{}^V\right),$$

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

where the projections represent parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$. *Attention* $= \{a_{ij}\}$ is a scoring matrix, in which $a_{ij}$ denotes the attention weight that the *Query* token $t_i$ gets from then *Key* token $t_j$. This matrix is widely used for representing and exploring the binding between tokens [33, 50, 58].

While the language models are fine-tuned on the methylation sites prediction task, in the last layer of our model, a softmax function that acts as a classifier is placed on the special token [CLS] that is present at the beginning of each input sentence.

For each token, we sum the attention weights assigned to [CLS] over the 12 heads and regard this as the token's contribution to sample prediction.

To analyze the impact of the DNA sequence of a sample on the taxonomic lineage of the sample, we extract the attention weights assigned by the DNA tokens to the taxonomic hierarchy tokens.

Note that the WordPiece algorithm, which is used by the tokenizer employed in BERT, DistilBERT, and ELECTRA, provides word-wise tokens, so it makes sense to view the attention weights of tokens as contribution scores.

Here we conduct the above computation on the 3 fine-tuned models of each methylation type in MuLan-Methyl, respectively. The token importance score for MuLan-Methyl is obtained as the average score achieved on each of the 3 site-specific models.

# Results

## Comparison with encoders from language models

To illustrate the effectiveness of the approaches we proposed for training language models for DNA-based applications, we compared the encoder of our pretrained language model with that of both BERT and DNABERT (see Fig. 2A). Each pretrained language model was applied to 10% of the positive DNA sequences in the independent test set, obtaining their sentence representation by extracting the embedding of [CLS], with a dimension of (1, 768). The samples were then clustered and visualized using Uniform Manifold Approximation and Projection (UMAP) technique, colored by taxonomic lineage.

Since the original corpus that BERT is trained on does not explicitly includes DNA fragments, during tokenization, BERT will represent each DNA 6-mer with the special symbol [UNK], or cuts it into small pieces, unaware that it is a biological sequence. Consequently, the DNA sequences are embedded into a sparse space distribution by this encoder, with a poor ability to distinguish different species.

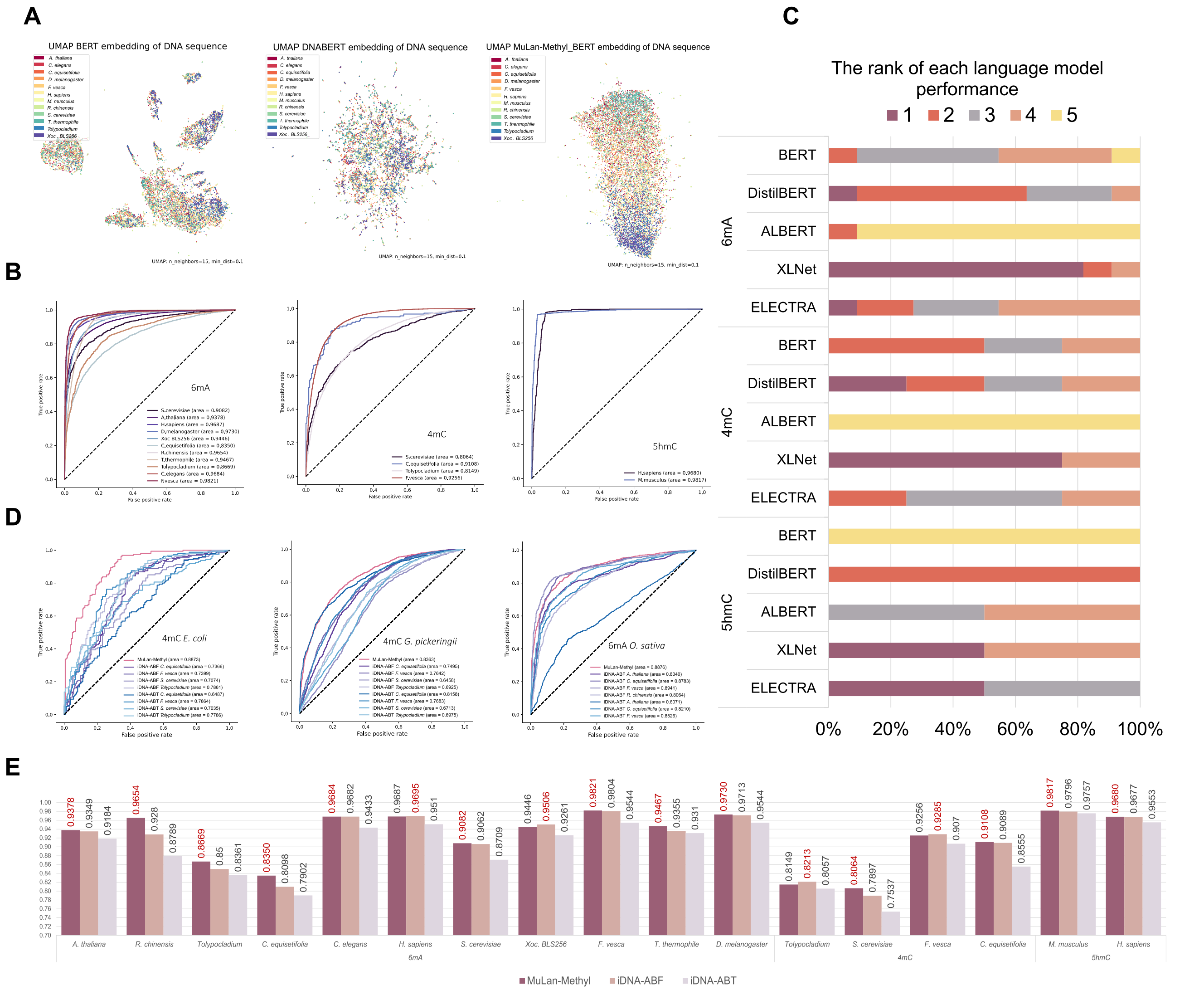


**Figure 2:** Model analysis and performance comparison of MuLan-Methyl. (A) UMAP clustering of sample representations encoded by different pretrained models: BERT, DNABERT, and MuLan-Methyl_BERT (from left to right). Samples are colored by taxonomic lineage. (B) For MuLan-Methyl predictions of the 3 methylation site types, 6mA, 4mC, and 5hmC, we present ROC curves for each of the 12 taxonomic types in the dataset. The AUC values are shown in brackets. (C) For each of the 3 methylation site types and each of the 5 language models, BERT, DistilBERT, ALBERT, XLNet, and ELECTRA, we show the ranking of models over all taxonomic lineages in terms of AUC scores. Moreover, the frequency with which each fine-tuned model appeared is indicated as the width of the corresponding block. (D) Comparison of MuLan-Methyl against 2 published methods, iDNA-ABF and iDNA-ABT, on an additional dataset that only contains taxonomic lineages that were not used to train the methods. From left to right, we show the ROCs obtained for the prediction of 4mC sites in *E. coli*, 4mC sites in *G. pickeringii* data, and 6mA sites in *O. sativa* L. data, respectively. (E) Comparison of MuLan-Methyl against iDNA-ABF and iDNA-ABT, on the iDMA-MS independent test set. We display the AUC scores for all 3 methods, for each of the 3 methylation site types and each of the 12 taxonomic lineages.

DNABERT is trained on genome sequences and has a better ability to capture DNA sequence features, as reflected in the absence of significant gaps between the distribution of DNA sequence representation obtained by its encoder. In the UMAP visualization, colors representing different taxonomic lineages appear to be randomly distributed.

In comparison, the MuLan-Methyl-BERT encoder is better at identifying DNA fragments and differentiating sequences by taxonomic lineage. Colors representing different taxonomic lineages exhibit a gradient from top to bottom. This suggests that pretraining the language model using a custom corpus that contains both DNA 6-mers and taxonomic lineages significantly improves the model's ability to capture potential information in this application scenario.

## Comparison with single language submodels

The MuLan-Methyl framework uses 5 language models. In this section, we establish that the average prediction probability of this integrated approach is better than using any of the individual submodels, by comparing model performance using area under the curve (AUC) values. In summary, MuLan-Methyl outperforms the submodels, displaying the highest AUC across different taxonomic lineages and for each methylation site type.

In more detail, for 6mA site prediction, MuLan-Methyl is most beneficial when predicting on *Tolypocladium* and *S. cerevisiae*, with an AUC gain of 1.4% over the AUC calculated by ALBERT, which was the best-performing submodel. The average increase of AUC compared to the taxonomic lineage–specific best submodel is 0.7%. For 4mC site prediction, the average gain of AUC computed from MuLan-Methyl is 0.8%, where the biggest improvement using MuLan-Methyl happened on *S. cerevisiae*, with an AUC increase of 1.1% over XLNet, the best submodel for this taxonomic lineage. MuLan-Methyl performs as well as taxonomic lineage–specific submodels at identifying 5hmC sites on both of the genomes. Moreover, we assessed the performance of MuLan-Methyl for each methylation site type and report on this for each taxonomic lineage using multiple metrics, including accuracy, F1-score, recall, area under the precision-recall curve (AUPR) and AUC (see Tables 1–3), as well as their receiver operating characteristic (ROC) curve (see Fig. 2B).

For each of the 3 methylation site types and for each of the 5 submodels included in MuLan-Methyl, we evaluated the performance of submodels on the corresponding independent test set. For each of the 12 taxonomic lineages, we ranked the given submodels based on their AUC values. Also, we determined the occurrence frequency of each submodel at each rank. This is shown in Fig. 2C and Supplementary Fig. S1.

We observed that XLNet usually shows better AUC than the other submodels for predicting 6mA sites, ranked first for 9 lineages. In contrast, ALBERT performs very poorly.

XLNet also performed best in 4mC site predictions, achieving the highest AUC on 3 of 4 taxonomic lineages. The lowest AUC from 4 taxonomic types concentratedly results from ALBERT. XLNet and ELECTRA performed best on the 5hmC site; BERT performs worst.

## Comparison with existing methods

To demonstrate the advantage of MuLan-Methyl over existing methods, we compared the method against iDNA-ABF and iDNA-ABT, 2 state-of-the-art methods, that are both able to predict methylation sites for all 3 types, across different taxonomic lineages. (Note that all 3 frameworks were trained on the same training dataset, provided by iDNA-MS.) For this, we used the iDNA-MS independent test set, which is considered a benchmark dataset. We report the AUC scores in Fig. 2E, and more comprehensive evaluation metrics are displayed in Supplementary Table S2.

In this study, MuLan-Methyl outperforms the other 2 methods on 13 of 17 combinations of methylation types and taxonomic lineages. First, for 6mA site prediction, MuLan-Methyl improves over the other methods by between 0.02% and 3.74% AUC, whereas for *R. chinensis*, *C. equisetifolia*, *Tolypocladium*, and *T. thermophile*, the improvement is by more than 1%. Second, for 4mC site prediction, our method shows an increase of 1.67% and 0.02% AUC, on *S. cerevisiae* and *C. equisetifolia*, respectively. Finally, for 5hmC site prediction, our method shows an increase of 0.21% and 0.03% on *M. musculus* and *H. sapiens*, respectively.

The iDNA-ABF method has higher AUC scores in the remaining 4 cases—namely, for 6mA site prediction on *H. sapiens* and *Xoc.* BLS256, with an improvement of 0.08% and 0.6%, and for 4mC site prediction on *Tolypocladium* and *F. vesca*, with an improvment of 0.64%, and 0.3%, respectively, over MuLan-Methyl. A cursory comparison suggests that MuLan-Methyl and iDNA-ABF have similar reported runtimes (albeit using different GPUs), whereas iDNA-ABT runs about 10 times faster.

## Explainability of MuLan-Methyl aids motifs discovery

To assess the contribution of each token toward correct methylation site detection, we use the average attention weight assigned by each token to [CLS] in the fine-tuned submodel, based on the positive sample from the independent test set.

The importance scores of each position in a DNA sequence has a Gaussian distribution across 17 different combinations of methylation site types and taxonomic lineages (see Fig. 3D–F and Supplementary Fig. S3). Positions of higher importance are concentrated around the center of the samples, and the central position always has high significance.

This observation supports the rationale used for constructing the iDNA-MS dataset—namely, to use, as positive samples, DNA segments of length 41 that are each centered on an experimentally verified methylation site. It also suggests the existence of DNA motifs that are closely associated with DNA methylation.

We also observed, for all 17 combinations, that the importance score starts low and then reaches a local maximum at position ±15. It then steadily increases from ±16 to the center of each sample (of length 41). This suggests that 41 is an ideal sample length for methylation detection, neither wasting resources to store unimportant positions nor missing important sequence.

The 6-mers with high importance may be considered to be DNA methylation "motifs" (see Fig. 3A–C and Supplementary Fig. S2). For a fixed taxonomic lineage, the 3 different methylation site types each have different motifs. However, for a fixed methylation site type, some motifs occur across different taxonomic lineages.

For example, the motif CGAAGT is important for 6mA methylation for several taxonomic lineages—namely, *S. cerevisiae*, *Tolypocladium*, and *Xoc.* BLS256. Note that the former 2 are eukaryotes, whereas the latter is a bacteria. Moreover, for 5hmC methylation, *H. sapiens* and *M. musculus* share many motifs. Similarly, for 4mC methylation, *C. equisetifolia* and *F. vesca* share many motifs.

## Explainability of MuLan-Methyl reveals relationships between DNA sequence and taxonomic lineage

Integrating DNA sequences with taxonomic lineage as an explicit feature adds information and thus increases detection accuracy. Moreover, during fine-tuned model prediction, the association between DNA sequence and taxonomy can be measured by extracting the attention weights assigned from DNA tokens to the tokens that represent taxonomic lineage (see Fig. 3G–I and Supplementary Fig. S4).

The impact of DNA sequence on taxonomic lineage varies across the 17 combinations of methylation site types and taxonomic lineages. Generally, sequence locations that determine taxonomic lineage are concentrated around the center of samples, where the discussed DNA methylation-associated motifs are also clustered.

Of the 8 taxonomic ranks used to specify taxonomic lineage, the highest (kingdom) and lowest rank (species), in particular, are assigned larger attention weights by a wide range of positions in the sequence.

However, not all combinations follow this rule. For example, the impact of DNA sequence on species is weaker than on genus and family for the combinations 6mA + *D. melanogaster* and 5hmC + *M. musculus*. On combinations 6mA + *R. chinensis*, 6mA + *S. cerevisiae*, 6mA + *C. elegans*, 4mC + *S. cerevisiae*, and 5hmC + *H. sapiens*, we observed that the high scores assigned to the taxonomy lineages are quite sparsely distributed over the different ranks.

These observations demonstrate that the explainability of MuLan-Methyl can shed light on the relationships between DNA sequences and taxonomic lineage.

**Table 1:** MuLan-Methyl prediction performance on 6mA sites

| Lineage | AUC | Accuracy | F1 | Recall | AUPR |
|---|---|---|---|---|---|
| *T. thermophile* | 0.9467 | 0.8840 | 0.8923 | 0.9611 | 0.9321 |
| *A. thaliana* | 0.9378 | 0.8649 | 0.8615 | 0.8401 | 0.9423 |
| *H. sapiens* | 0.9687 | 0.9077 | 0.9068 | 0.8975 | 0.9721 |
| *Xoc.* BLS256 | 0.9446 | 0.8742 | 0.8712 | 0.8511 | 0.9421 |
| *D. melanogaster* | 0.9730 | 0.9276 | 0.9275 | 0.9258 | 0.9761 |
| *C. elegans* | 0.9684 | 0.9131 | 0.9138 | 0.9219 | 0.9674 |
| *C. equisetifolia* | 0.8350 | 0.7590 | 0.7481 | 0.7158 | 0.8492 |
| *S. cerevisiae* | 0.9082 | 0.8325 | 0.8233 | 0.7802 | 0.9198 |
| *Tolypocladium* | 0.8669 | 0.7895 | 0.7824 | 0.7567 | 0.8730 |
| *F. vesca* | 0.9821 | 0.9407 | 0.9403 | 0.9336 | 0.9831 |
| *R. chinensis* | 0.9654 | 0.9164 | 0.9167 | 0.9197 | 0.9691 |

**Table 3:** MuLan-Methyl prediction performance on 5hmC sites

| Lineage | AUC | Accuracy | F1 | Recall | AUPR |
|---|---|---|---|---|---|
| *M. musculus* | 0.9817 | 0.9649 | 0.9651 | 0.9685 | 0.9782 |
| *H. sapiens* | 0.9680 | 0.9484 | 0.9500 | 0.9787 | 0.9485 |

**Table 2:** MuLan-Methyl prediction performance on 4mC sites

| Lineage | AUC | Accuracy | F1 | Recall | AUPR |
|---|---|---|---|---|---|
| *C. equisetifolia* | 0.9108 | 0.8333 | 0.8272 | 0.7978 | 0.9221 |
| *F. vesca* | 0.9256 | 0.8522 | 0.8554 | 0.8739 | 0.9144 |
| *S. cerevisiae* | 0.8064 | 0.7376 | 0.7253 | 0.6926 | 0.8215 |
| *Tolypocladium* | 0.8149 | 0.7380 | 0.7285 | 0.7031 | 0.8089 |

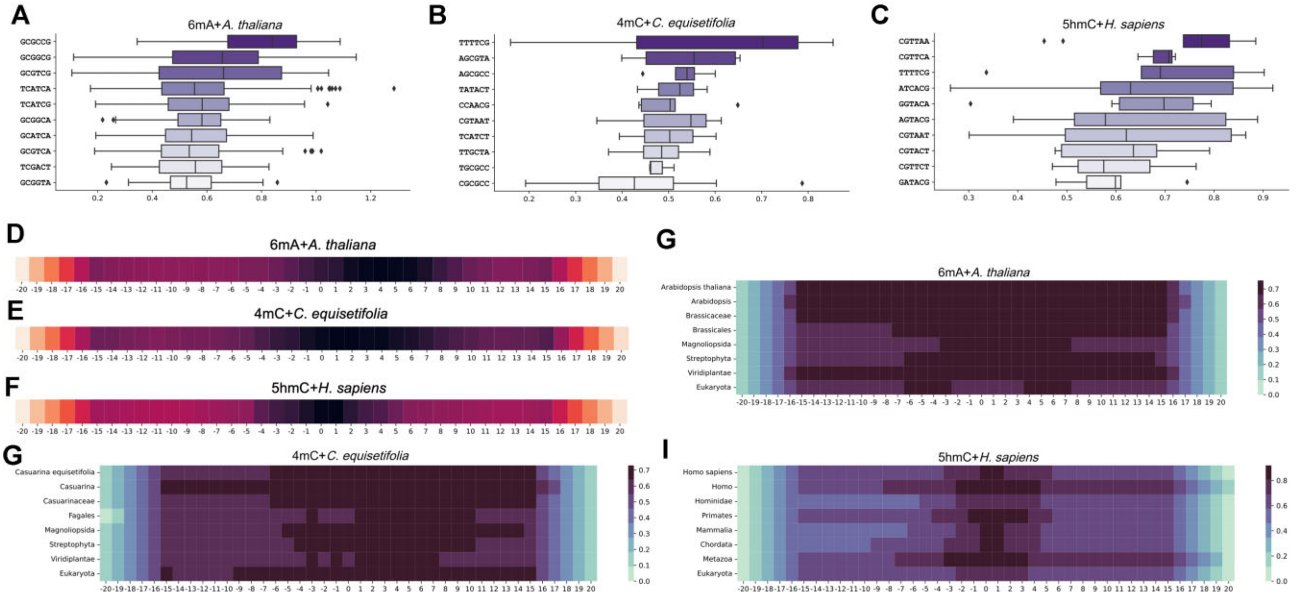
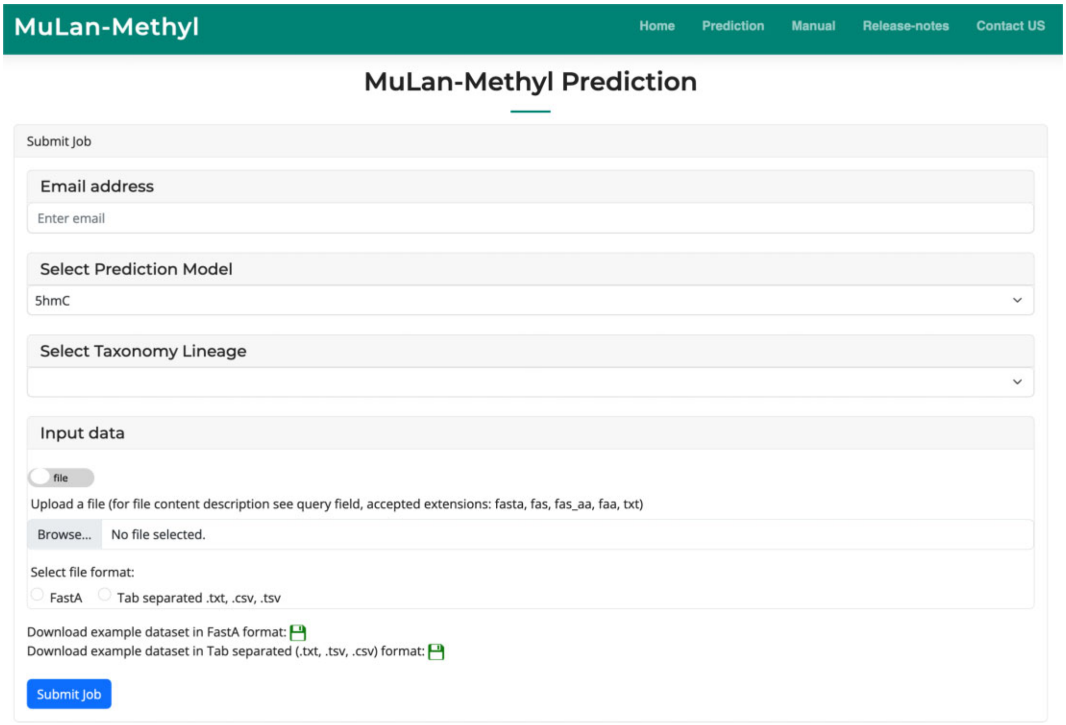


**Figure 3:** Interpretation of MuLan-Methyl by attention weights resulting from transformer self-attention mechanism. (A–C) Boxplots show the distribution of attention weights for the ten 6-mer of highest average importance scores, for the combinations 6mA + *A. thaliana*, 5mC + C. *equisetifolia* and 5hmC + *H. sapiens*, respectively. (D–F) We indicate the importance score for each position in the DNA sequences of length 41, obtained by merging 6-mer fragments, for the same 3 combinations listed above, respectively. (G–I) For each taxonomic rank of a lineage, we indicate the attention weight assigned by MuLan-Methyl to each position of the sequence for generating the taxon of the given rank, for the same 3 combinations listed above, respectively.

## Performance on the external dataset

MuLan-Methyl was trained on 17 combinations of DNA methylation site types and taxonomic lineages. Fine-tuned models aim at performing well on input whose distribution is consistent with the training dataset but are not guaranteed to perform well on other data.



MuLan-Methyl Prediction

Submit Job

Email address: Enter email

Select Prediction Model: 5hmC

Select Taxonomy Lineage:

Input data

file

Upload a file (for file content description see query field, accepted extensions: fasta, fas, fas_aa, faa, txt)

No file selected.

Select file format:
○ FastA ○ Tab separated .txt, .csv, .tsv

Download example dataset in FastA format:
Download example dataset in Tab separated (.txt, .tsv, .csv) format:

**Figure 4:** The Mulan-Methyl server hosted at [59] allows upload of DNA sequences and will perform methylation prediction along the whole sequence.

To explore the performance of MuLan-Methyl on other data, we applied the approach to an external dataset that contains 3 combinations of methylation types and taxonomic lineages—namely, 4mC + *E. coli*, 4mC + *G. pickeringii*, and 6mA + *O. sativa* L. Note that these 3 taxonomic lineages do not appear in the iDNA-MS datasets.

For the sake of comparison, we also calculated predictions using the servers provided by iDNA-ABF and iDNA-ABT. Since both approaches provide independent models for each combination, we ran all taxon-wise models for 4mC site detection and the appropriate ones for 6mA site detection.

MuLan-Methyl performed much better than the other 2 models on the 4mC + *E. coli* combination, achieving an AUC of 0.89, more than 10% better than the others. Our method also performed best on the 4mC + *G. pickeringii* combination, with an advantage of 2.05% over iDNA-ABT (using its *C. equisetifolia* model). On the third combination, 6mA + *O. sativa* L, MuLan-Methyl performed slightly worse (0.65%) than iDNA-ABF (using its *F. vesca* model). See Fig. 2D.

## MuLan-Methyl server

We provide an implementation of MuLan-Methyl as a web server (see Fig. 4). Like other deep learning–based methylation services, this allows the user to upload DNA samples of length 41 and select the closet taxonomic lineage and the type of methylation site of interest. The uploaded samples will then be classified as methylation sites or not.

We also allow upload of longer DNA sequences, and in this case, the server will provide a list of all methylation sites that are predicted in the uploaded sequence.

To implement this extended functionality, we first extract all samples of length 41 that are centered on a nucleotide of the appropriate type (e.g., C when predicting 4mC or 5hmC sites) and then perform Mulan-Methyl prediction on these. The predicted positive samples are then filtered by feature importance analysis to resolve overlapping predictions. In more detail, we only retain samples for which the importance scores are highest at the center of the sample. Output is the list of all predicted methylation positions.

## Discussion

Previous studies have focused on adapting BERT to specific biological tasks using the pretrain and fine-tune paradigm, with the aim of applying this popular NLP approach to tasks in genomics, phylogenetics, and other areas of computational biology.

However, BERT is not the only transformer-based language model, and it is important to choose the best model for a given task. Our proposed framework, MuLan-Methyl, consists of 5 transformer-based language models for identifying 3 types of DNA methylation sites across several taxonomic lineages, including both Eukaryota and Bacteria. With this work, we extend the list of transformed-based language models that have been successfully adapted to tasks involving biological sequences.

Each submodel in MuLan-Methyl is pretrained and fine-tuned on the training dataset, which then collectively predicts methylation sites on an independent test dataset. The performance of MuLan-Methyl was evaluated by multiple metrics and in comparison with 2 existing approaches, and the method showed very good performance.

Our study also indicates that models with enhanced algorithms in the pretraining step, such as XLNET, and models with fewer parameters and less memory consumption, such as DistilBERT, are more appropriate than BERT when storage or computational resources are limited.

In contrast to other biological domain adaption language models, the custom corpus that we trained MuLan-Methyl on contains multimodal data, consisting of both DNA sequences from iDNA-MS and taxonomy lineage in text format from the NCBI and GTDB taxonomies. To the best of our knowledge, MuLan-Methyl is the first language model framework to take taxonomy information into consideration.

This improves model accuracy and feature contribution analysis. The DNA methylation motifs found by MuLan-Methyl greatly benefited from the self-attention mechanism of transformer structure. In addition, the attention weights assigned to taxonomic lineages by DNA sequences help to analyze the relationship between nucleotide sequences and taxonomy lineage.

Previous approaches build a separate classifier for each taxonomic lineage and each methylation site type, giving rise to 17 different classifiers, for the data used here. In contrast, MuLan-Methyl considers taxonomic lineage as a feature and so only gives rise to 3 classifiers, one for each type of methylation site.

This study demonstrates that BERT is not the only choice when one wants to adapt a transformer-based language model to a specific domain; one should also consider its variants. It also shows that integrating multiple language models can offset the deficiencies of the individual models, to some extent, so as to obtain an improved ensemble prediction performance.

In conclusion, we have proposed a framework that integrates 5 popular NLP approaches to solve an important biological problem. MuLan-Methyl can be used to detect DNA methylation sites reliably for DNA sequences of 41 bp length from known taxonomic lineages, especially when closely related to the lineages involved in training, with slightly better performance than current state-of-the-art methods.

In practical applications, the input will usually be a chromosome or a set of assembled contigs, and the desired output will be a list of putative methylation sites. To address this, we designed a 2-step validation strategy for false-positive rate controlling and implemented it in the MuLan-Methyl server.

## Availability of Source Code and Requirements

Project name: MuLan-Methyl
Project homepage: http://ab.cs.uni-tuebingen.de/software/mulan-methyl
Code GitHub: https://github.com/husonlab/mulan-methyl
Operating system(s): Platform independent
Programming language: Code: Python (3.10.6); WebServer: HTML5, Bootstrap, PHP (7.2.24), JavaScript
Other requirements: For WebServer MySQL (5.7.39)
License: Apache-2.0.
Any restrictions to use by nonacademics: None
Biotools ID: https://bio.tools/MuLan-Methyl
RRID: SCR_023591

## Additional Files

**Supplementary Fig. S1.** Heatmap of Kendall tau distance matrix for exploring the ranking correlation, where the rank is obtained by comparing AUC with the other submodels on each methylation site type of each submodel.
**Supplementary Fig. S2.** Boxplot of top 10 tokens with the highest average attention scores for the remaining combinations of methylation types and taxonomic lineage.
**Supplementary Fig. S3.** Heatmap of the average importance score for each position of a 41-bp DNA sequence obtained by merging 6-mer fragments for each remaining combination of methylation types and taxonomic lineage.
**Supplementary Fig. S4.** Heatmap of the impact between DNA sequence and its taxonomy lineage for each remaining combination of methylation types and taxonomic lineage.
**Supplementary Table S1.** iDNA-MS dataset statistics.
**Supplementary Table S2.** The comparison of model performance on the iDNA-MS independent test set between MuLan-Methyl and its submodels, as well as with the previous studies.

## Abbreviations

4mC: 4-methylcytosine; 5hmC: 5-hydroxymethylcytosine; 6mA: 6-methyladenine; AUC: area under the curve; BERT: bidirectional encoder representation from transformers; GTDB: Genome Taxonomy Database; AUPR: area under the precision-recall curve; NLP: natural language processing; MLM: masked language modeling; NCBI: The National Center for Biotechnology Information; ROC: receiver operating characteristic.

## Data Availability

The benchmark dataset used in this study is available here [31, 55]. The processed dataset used for training MuLan-Methyl and the source code are available at [60]. A web server implementing the MuLan-Methyl approach is freely accessible at [59]; see also [61], RRID: SCR_023591. All supporting data and materials are available in the *GigaScience* GigaDB database [62].

## Competing Interests

The authors declare that they have no competing interests.

## Funding

We acknowledge support by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). We acknowledge support by the Open Access Publishing Fund of the University of Tübingen.

## Authors' Contributions

W.Z. and D.H.H. conceived the project. W.Z. collected and processed the dataset for the project, designed and implemented the architecture and algorithms of MuLan-Methyl, and conducted model analysis. A.G. and W.Z. designed and implemented the web server of MuLan-Methyl. W.Z., D.H.H., and A.G. contributed to the manuscript.

## References

1. Robertson KD, Wolffe AP. DNA methylation in health and disease. Nat Rev Genet 2000;1(1):11–9.
2. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology 2013;38(1):23–38.
3. Armstrong MJ, Jin Y, Allen EG, et al. Diverse and dynamic DNA modifications in brain and diseases. Hum Mol Genet 2019;28(R2):R241–53.
4. Tourancheau A, Mead EA, Zhang XS, et al. Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. Nat Methods 2021;18(5):491–8.
5. O'Brown ZK, Boulias K, Wang J, et al. Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. BMC Genomics 2019;20(1):1–15.
6. Ito S, Shen L, Dai Q, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science 2011;333(6047):1300–3.

7. Bilyard MK, Becker S, Balasubramanian S. Natural, modified DNA bases. Curr Opin Chem Biol 2020;57:1–7.
8. Rauluseviciute I, Drabløs F, Rye MB. DNA methylation data by sequencing: experimental approaches and recommendations for tools and pipelines for data analysis. Clin Epigenet 2019;11(1): 1–13.
9. Ye P, Luan Y, Chen K, et al. MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. Nucleic Acids Res 2016;45(D1):D85–89.
10. Xu H, Jia P, Zhao Z. Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. Brief Bioinform 2021;22(3):bbaa099.
11. Zeng R, Cheng S, Liao M. 4mCPred-MTL: accurate identification of DNA 4mC sites in multiple species using multi-task deep learning based on multi-head attention mechanism. Front Cell Dev Biol 2021;9:664669.
12. Liu Q, Chen J, Wang Y, et al. DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. Brief Bioinform 2021;22(3):bbaa124.
13. Hasan MM, Manavalan B, Shoombuatong W, et al. i4mC-Mouse: improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes. Comput Struct Biotech J 2020;18:906–12.
14. Jin J, Yu Y, Wei L. Mouse4mC-BGRU: Deep learning for predicting DNA N4-methylcytosine sites in mouse genome. Methods 2022;204:258–62.
15. Zulfiqar H, Sun ZJ, Huang QL, et al. Deep-4mCW2V: a sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli. Methods 2022;203:558–63.
16. Liang Y, Wu Y, Zhang Z, et al. Hyb4mC: a hybrid DNA2vec-based model for DNA N4-methylcytosine sites prediction. BMC Bioinformatics 2022;23(1):1–18.
17. Tran TA, Pham DM, Ou YY, et al. An extensive examination of discovering 5-methylcytosine sites in genome-wide DNA promoters using machine learning based approaches. IEEE/ACM T Comput Biol Bioinform 2021;19(1):87–94.
18. Cheng X, Wang J, Li Q, et al. BiLSTM-5mC: a bidirectional long short-term memory-based approach for predicting 5-methylcytosine sites in genome-wide DNA promoters. Molecules 2021;26(24):7414.
19. Li Z, Jiang H, Kong L, et al. Deep6mA: a deep learning framework for exploring similar patterns in DNA N6-methyladenine sites across different species. PLoS Comput Biol 2021;17(2): e1008767.
20. Rehman MU, Tayara H, Zou Q, et al. i6mA-Caps: a CapsuleNet-based framework for identifying DNA N6-methyladenine sites. Bioinformatics 2022;38(16):3885–91.
21. Zeng R, Liao M. 6mAPred-MSFF: a deep learning model for predicting DNA N6-methyladenine sites across species based on a multi-scale feature fusion mechanism. Appl Sci 2021;11(16):7731.
22. Liu M, Sun ZL, Zeng Z, et al. MGF6mARice: prediction of DNA N6-methyladenine sites in rice by exploiting molecular graph feature and residual block. Brief Bioinform 2022;23(3): bbac082.
23. Tsukiyama S, Hasan MM, Deng HW, et al. BERT6mA: prediction of DNA N6-methyladenine site using deep learning-based approaches. Brief Bioinform 2022;23(2):bbac053.
24. Tahir M, Hayat M, Ullah I, et al. A deep learning-based computational approach for discrimination of DNA N6-methyladenosine sites by fusing heterogeneous features. Chemometr Intell Lab Syst 2020;206:104151.
25. Le NQK, Ho QT. Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes. Methods 2022;204:199–206.
26. Tang X, Zheng P, Li X, et al. Deep6mAPred: a CNN and Bi-LSTM-based deep learning method for predicting DNA N6-methyladenosine sites across plant species. Methods 2022;204:142–50.
27. Hasan MM, Basith S, Khatun MS, et al. Meta-i6mA: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. Brief Bioinform 2021;22(3):bbaa202.
28. Chen J, Zou Q, Li J. DeepM6ASeq-EL: prediction of human N6-methyladenosine (m6A) sites with LSTM and ensemble learning. Front Comput Sci 2022;16(2):1–7.
29. Zhang Y, Liu Y, Xu J, et al. Leveraging the attention mechanism to improve the identification of DNA N6-methyladenine sites. Brief Bioinform 2021;22(6):bbab351.
30. Yang X, Ye X, Li X, et al. iDNA-MT: identification DNA modification sites in multiple species by using multi-task learning based a neural network tool. Front Genet 2021;12:663572.
31. Lv H, Dao FY, Zhang D, et al. iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. Iscience 2020;23(4):100991.
32. Yu Y, He W, Jin J, et al. iDNA-ABT: advanced deep learning model for detecting DNA methylation with adaptive features and transductive information maximization. Bioinformatics 2021;37(24):4603–10.
33. Jin J, Yu Y, Wang R, et al. iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. Genome Biol 2022;23(1):1–23.
34. Zheng Z, Le NQK, Chua MCH. MaskDNA-PGD: an innovative deep learning model for detecting DNA methylation by integrating mask sequences and adversarial PGD training as a data augmentation method. Chemometr Intell Lab Syst 2023;232:104715.
35. Lv H, Dao FY, Zhang D, et al. Supporting data for "iDNA-MS: An Integrated Computational Tool for Detecting DNA Modification Sites in Multiple Genomes." GigaScience Database. 2023. Http://dx.doi.org/10.5524/102395.
36. Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. Vol. 1. Association for Computational Linguistics.Minneapolis, Minnesota.2019;4171–86.
37. Ji Y, Zhou Z, Liu H, et al. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics 2021;37(15):2112–20.
38. Zhang Yz, Yamaguchi K, Hatakeyama S, et al. On the application of BERT models for NanoPore methylation detection. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Houston, Texas: IEEE, 2021;320–7.
39. Seong HJ, Han SW, Sul WJ. Prokaryotic DNA methylation and its functional roles. J Microbiol 2021;59(3):242–8.
40. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neur Inf Process Syst 2017;30:5998–6008.
41. Zeng W, Gautam A, Huson DH. DeepToA: an ensemble deep-learning approach to predicting the theater of activity of a microbiome. Bioinformatics 2022;38(20):4670–6.
42. Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:191001108. 2019. https://arxiv.org/abs/1910.01108.
43. Liu Y, Ott M, Goyal N, et al. Roberta: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:190711692. 2019. https://arxiv.org/abs/1907.11692.

44. Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding. Adv Neur Inf Process Syst 2019;32:5754–64.
45. Clark K, Luong MT, Le QV, et al. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:200310555.2020. https://arxiv.org/abs/2003.10555.
46. Lan Z, Chen M, Goodman S, et al. Albert: a lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:190911942. 2019. https://arxiv.org/abs/1909.11942.
47. Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:200410964. 2020. https://arxiv.org/abs/2004.10964.
48. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.
49. Conneau A, Lample G. Cross-lingual language model pretraining. Adv Neur Inf Process Syst 2019;32:7059–69.
50. Lupo U, Sgarbossa D, Bitbol AF. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. Nat Commun 2022;13(1):6298.
51. Mock F, Kretschmer F, Kriese A, et al. Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. Proc Natl Acad Sci 2022;119(35):e2122636119.
52. Yang M, Huang L, Huang H, et al. Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. Nucleic Acids Res 2022;50(14):e81.
53. Schoch CL, Ciufo S, Domrachev M, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) 2020;2020. http://doi.org/10.1093/database/baaa062.
54. Parks DH, Chuvochina M, Rinke C, et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res 2022;50.D785–94
55. iDNA-MS web server. 2020. http://lin-group.cn/server/iDNA-MS/download.html.
56. Chen W, Lv H, Nie F, et al. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics 2019;35(16):2796–800.
57. Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg, Pennsylvania: Association for Computational Linguistics, 2020;38–45.
58. Yamada K, Hamada M. Prediction of RNA-protein interactions using a nucleotide language model. Bioinform Adv 2022;2(1):vbac023.
59. MuLan-Methyl web server. 2023. http://ab.cs.uni-tuebingen.de/software/mulan-methyl/.
60. GitHub repository of MuLan-Methyl. 2023. https://github.com/husonlab/mulan-methyl.
61. Biotools link of MuLan-Methyl. https://bio.tools/MuLan-Methyl.
62. Zeng W, Gautam A, Huson DH. Supporting data for "MuLan-Methyl—Multiple Transformer-Based Language Models for Accurate DNA Methylation Prediction." GigaScience Database. 2023. http://dx.doi.org/10.5524/102402.

**Received:** February 19, 2023. **Revised:** May 9, 2023. **Accepted:** July 18, 2023