MuLan-Methyl - Multiple Transformer-based Language Models for Accurate DNA Methylation Prediction

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

[1]Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, 72076, Germany and [2]International Max Planck Research School "From Molecules to Organisms", Max Planck Institute for Biology Tübingen, Max-Planck-Ring 5, Tübingen, 72076, Germany and [3]Cluster of Excellence: EXC 2124: Controlling Microbes to Fight Infection, University of Tübingen, Tübingen, Germany

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 a MuLan-Methyl, a deep learning framework for predicting DNA methylation sites, which is based on multiple (five) popular transformer-based language models. The framework identifies methylation sites for three different types of DNA methylation, namely N6-adenine (6mA), N4-cytosine (4mC), and 5-hydroxymethylcytosine (5hmC). Each of the five employed language models is adapted to the task using the "pre-train and fine-tune" paradigm. Pre-training is performed on a custom corpus consisting of DNA fragments and taxonomy lineages using self-supervised learning. Fine-tuning then aims at predicting the DNA-methylation status of each type. The five models are used to collectively predict the DNA methylation status. We report excellent performance of MuLan-Methyl on a benchmark dataset. Moreover, we show 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 this domain of application and that joint utilization of different language models improves model performance.

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). There are several different types of DNA methylation, depending on which methyl group is attached to which type of nucleotide in the sequence. Here, we focus on 6-methyadenine (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 significant room for improvement of accuracy and comprehensiveness.

There is a large number of papers that address the problem of identifying methylation sites, however, most of them focus on specific form of modification (10–29), and only a few methods address all three types of methylation mentioned above (30–34), including iDNA-MS, iDNA-ABT, and iDNA-ABF. Note that the database presented in (31) 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), say. Utilized model structures include Convolutional Neural Networks (CNN), Graph Convolutional Neural Networks (GCN), Bidirectional Encoder Representation from Transformers (BERT) (35), as well as machine learning algorithms. The specific way that an approach combines feature engineering and model structure determines its performance, and is key to proposing a new framework.

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 (36), 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, 37).

Only few published approaches aim at predicting multiple DNA modification sites. Moreover, many do not use

*To whom correspondence should be addressed. Tel: +49 7071 2970450; Email: daniel.huson@uni-tuebingen.de

taxonomic information as explicit features, although the taxonomic identity of an organism has an impact on DNA methylation (38). Here we seek to address both shortcomings by providing a new framework that uses a set of collectively training language models, including, but not limited to BERT, to predict three types of methylation sites from DNA sequences and taxonomic information.

Combining the transformer-based language model BERT with the "pre-train and fine-tune" paradigm has become the method of choice in NLP applications. In the pre-training step, self-supervised learning of the Masked Language Modelling (MLM) task and the Next Sequence Prediction (NSP) task are initially 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 multi-head self-attention mechanism (39) and this leads to improved prediction over classification models constructed using static embedding approaches (40).

The fine-tuning step involves supervised training of the pre-trained language model to adapt to specific downstream tasks, here the prediction of three different types of methylation sites. Using BERT as a starting point, and then varying the network architecture and parameters, one can obtain five different language models, (41–45). By pre-training on a domain-specific custom corpus, BERT can be adapted to a specific application scenario (46–49). 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. Hence, several approaches, such as BERTax, DNABERT and LOGO (36, 50, 51), use large amounts of genomic sequence, instead of Wikipedia, as a corpus or similar structure.

The main aim of this paper is to introduce MuLan-Methyl, a novel deep learning framework that combines five transformer-based language models to collectively predict sites for three different types of methylation (see Figure 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 five transformer-based language models, namely BERT (35), DistilBERT (41), ALBERT (45), XLNet (43) and ELECTRA (44).

Each of the five language models is trained according to the "pre-train and fine-tune" paradigm. For this, we used a custom corpus that contains the processed training dataset and taxonomic lineage information downloaded from NCBI (52) and GTDB (53). 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 is captured effectively.

Each language model was pre-trained by training the MLM task on the processed training dataset. We then obtained the 6mA model by fine-tuning the pre-trained 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 facilitiates the discovery of DNA methylation motifs and potential associations between specific methylation sites and taxonomic lineages.

This work demonstrates that adding features to a model is not the only way to improve the accuracy of predictions. 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.

MATERIALS AND METHODS

Data processing

Data collection We downloaded a DNA methylation dataset from http://lin-group.cn/server/iDNA-MS/download.html. This is an open resource that was published with the iDNA-MS method (31) and is now widely used for benchmarking. The dataset contains three main types of DNA methylation sites - 6mA, 4mC and 5hmC - across 12 genomes (one bacteria and 11 eukaryotes), in total 250,599 positive samples. In addition, the dataset provides the same number of non-methylation sequences as negative samples.

The dataset is partitioned into a training set and a independent test set at a 1:1 ratio. In the training dataset, 11 species contain samples associated with methylation type 6mA, in more detail, the numbers are 53,800 for *T. thermophile*, 15,937 for *A. thaliana*, 9,168 for *H. sapiens*, 8,608 for *Xoc. BLS256*, 5,596 for *D. melanogaster*, 3,981 for *C. elegans*, 3,033 for *C. equisetifolia*, 1,893 for *S. cerevisiae*, 1,690 for *Tolypocladium*, 1,551 for *F. vesca* and 300 for *R. chinensis*. The type 4mC 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 *M. musculus* sequences and 1,172 for *H. sapiens*.

The samples are DNA segments of length 41; a positive sample is always centered on an experimentally-verified methylation site, whereas a negative sample is not.

Dataset preparation We processed each sample (a DNA sequence of length 41) as follows. Using a sliding window of length 6, we extract $36 = 41 - 6 + 1$ individual 6-mers from the DNA sequence, and embed these within a sentence, together with a description of the taxonomic lineage of the corresponding organism, 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

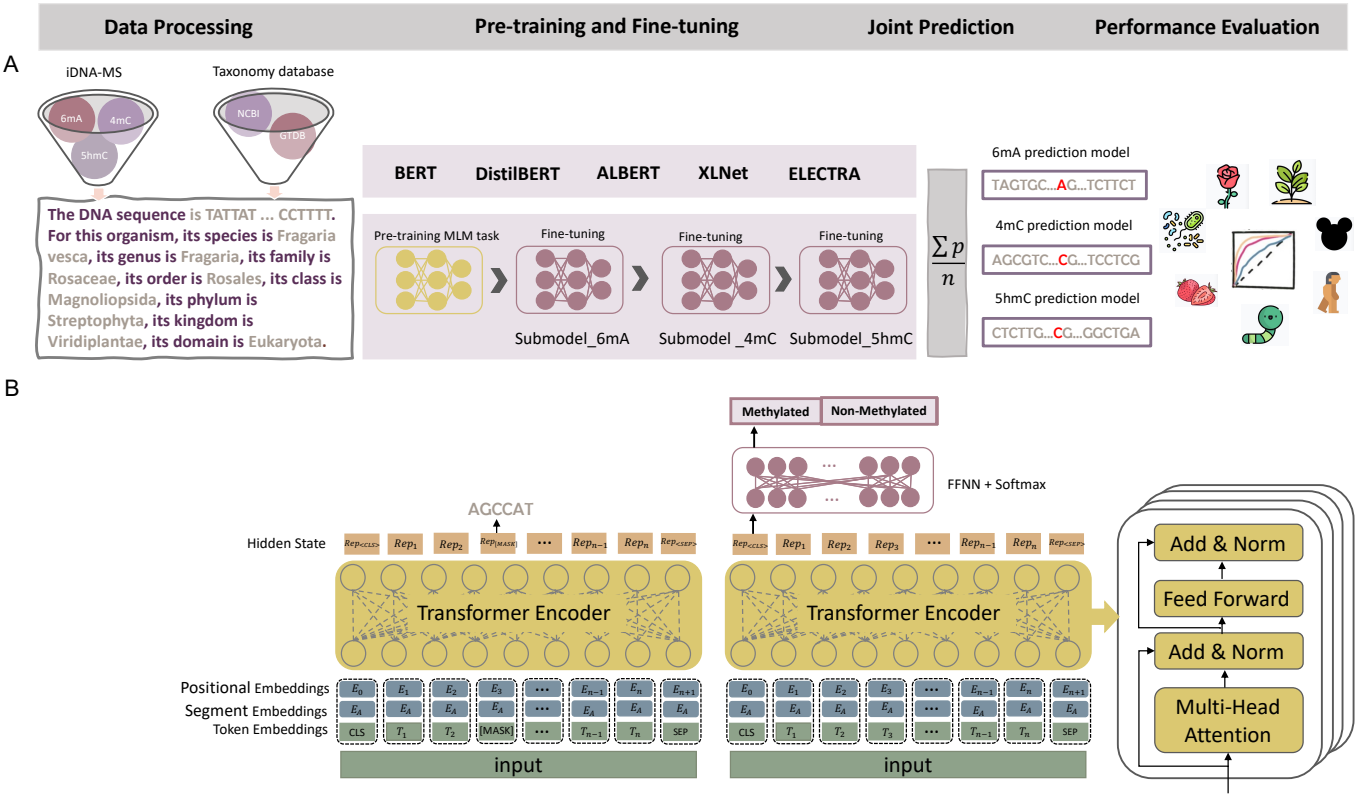

Figure 1. The MuLan-Methyl workflow. (A) The framework employs five fine-tuned language models for joint identification of DNA methylation sites. Methylation datasets (obtained from iDNA-MS) are processed as sentences that describe the DNA sequence as well as the taxonomy lineage, giving rise to the processed training dataset and the processed independent set. For each transformer-based language model, a custom tokenizer is trained based on a corpus that consists of the processed training dataset and taxonomy lineage data from NCBI and GTDB. Pre-training and fine-tuning are 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 five fine-tuned models. We thus obtain three methylation type-wise prediction models. We evaluated the model performance according to the genome type that contained in 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 pre-training and fine-tuning. The transformer-based language model is pre-trained using the masked language modeling (MLM) task and then fine-tuned on the methylation type-wise processed training dataset.

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 three types of methylation sites, is used to generate the custom corpus. For purposes of fine-tuning, both the processed training dataset and the processed independent test set were split into three sets by methylation types.

Corpus generation We require a custom corpus for pre-training each language model to allow it 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, which consists of sentences containing DNA 6-mers and a description of the associated taxonomic lineage. In addition, to enable the language to detect words about taxonomy, we incorporated all taxonomic lineages from the NCBI and GTDB taxonomies. 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 (54). As this data is not contained in the 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, this data consists of sequence-based samples that were processed using the above mentioned methods, including 320 4mC-site sequences from *E. coli*, 1,926 4mC-site sequences from *G. pickeringii*, and 880 6mA-site sequences from *Oryza sativa* L., along with the same number of corresponding non-methylated sequences.

Training transformer-based language models

We pre-trained and fine-tuned five transformer-based language models. In the following, we first describe the architecture of each of the five employed language models. We then discuss the details of the training process for the first method, BERT, including tokenization, pre-training, and fine-tuning (see Figure 1B). The other four languages are trained in a similar way.

All code is written in Python 3.10, using the Pytorch and Huggingface Transformers library (55). 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 five transformer-based language models, which we introduce in the following.
(1) BERT is capable of modelling bidirectional contexts, using denoising and autoencoding-based pre-training. For the transformer architecture of $\mathrm{BERT}_{\mathrm{base}}$, it use 12 layers in the encoder stack, 768 hidden units for feedforward networks and 12 attention heads; in total 110M parameters.
(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 pre-training step; its generalized autoregressive pre-training 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) In contrast to the other architectures, ELECTRA trains two 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 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 pre-training The BERT language model is pre-trained by performing unsupervised training of the MLM task on the custom corpus. Pre-training 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.

Pre-training was conducted by 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 three methylation-site types separately, and so the processed training dataset is split into three 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 depend on the subset's size. First, for the 6mA subset, we simply fine-tuned the pre-trained 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, 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 three cases, fine-tuning is 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.

Multi-language model

For each of the three types of methylation sites, five 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 five models.

Interpretability of MuLan-Methyl

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

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

where $Query(Q)\in\mathbb{R}^{n\times d_k}$, $Key(K)\in\mathbb{R}^{n\times d_k}$, $Value(V)\in\mathbb{R}^{n\times d_v}$ and $W^O\in\mathbb{R}^{hd_v\times d_{model}}$.

The i_{th} self-attention head is computed as

$$\mathrm{head_h}=\mathrm{Attention}\left(QW_h{}^Q,KW_h{}^K,VW_h{}^V\right)V \text{ and}$$

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

where $Attention_h=\{a_{ij}\}$ 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, 49, 56).

Whereas 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 these three fine-tuned models of each methylation type in MuLan-Methyl, respectively, and the tokens importance score of MuLan-Methyl is evaluated as the average score of sub-models.

The token importance score for MuLan-Methyl is obtained as the average score achieved on each of the three 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 compare the encoder of our pre-trained language model with that of both BERT and DNABERT (see Figure 2A).

Each pre-trained language model is 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 are then clustered and visualized using UMAP, 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, unware 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 .

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. However, the cluster groups representing different species are mixed.

In comparison, the MuLan-Methyl-BERT encoder is better at identifying DNA fragments and differentiating sequences by taxonomic lineage. This shows that pre-training the language model using a custom corpus that contains both DNA 6-mers and taxonomic lineages, significantly improves the models ability to capture potential information in this application scenario.

Comparison with single language sub-models

The MuLan-Methyl framework uses five language models. In this section, we establish that the average prediction probability of this integrated approach is better than using any of the individual sub-models, by comparing model performance using AUC values.

In summary, MuLan-Methyl outperforms the sub-models, displaying the highest AUC across different taxonomic lineages and for each methylation-site type.

In more detail, for 6mA-site prediction, MuLan-Methyl had the most significant benefit while predicting on *Tolypocladium*, with an AUC gain of 1.7% over the AUC calculated by ALBERT, which was the best-performing sub-model. The average increase of AUC compared to the taxonomic-lineage-specific best sub-model is 0.68%. For 4mC-site prediction, the average gain of AUC computed from MuLan-Methyl is 0.85%, where the biggest improvement using MuLan-Methyl happened on *S. cerevisiae*, with an AUC increase of 1.48% over XLNet, the best sub-model for this taxonomic lineage. Moreover, MuLan-Methyl performed slightly better than ELECTRA at identifying 5hmC-sites on the *H. sapiens* genome, with a 0.05% AUC rise. 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 and precision, and AUC (see Table 1, Table 2, Table 3), as well as their ROC curve (see Figure 2B).

For each of the three methylation-site types, and for each of the five sub-models included in MuLan-Methyl, we evaluated the performance of sub-models on the corresponding independent test set. For each of the 12 taxonomic lineages, we ranked the give sub-models based on their AUC values. Also, we determined the occurrence frequency of each sub-model at each rank. This is shown in Figure 2C.

We observed that XLNet most frequently shows better AUC than the other sub-models for predicting 6mA-sites, ranked first for 6 lineages. In contrast, BERT and ELECTRA both perform very poorly.

XLNet also performs best in 4mC-site predictions, achieving the highest AUC on 3 out of 4 taxonomic lineages. The lowest AUC from 4 taxonomic types are distributed equally over four other models. XLNet and ELECTRA perform best on 5hmC-site. Again, 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

Table 1. MuLan-Methyl prediction performance on 6mA-sites

Lineage	AUC	Accuracy	F1	Recall	AUPR
T. thermophile	0.9473	0.8849	0.8924	0.9543	0.9334
A. thaliana	0.9385	0.8654	0.8621	0.8419	0.9427
H. sapiens	0.9694	0.9076	0.9072	0.9028	0.9726
Xoc. BLS256	0.9485	0.8799	0.8764	0.8515	0.9459
D. melanogaster	0.9739	0.9299	0.9296	0.9251	0.9768
C. elegans	0.9701	0.9171	0.9181	0.9294	0.9692
C. equisetifolia	0.8371	0.7618	0.7540	0.7300	0.8510
S. cerevisiae	0.9117	0.8328	0.8246	0.7861	0.9230
Tolypocladium	0.8711	0.7928	0.7845	0.7543	0.8773
F. vesca	0.9839	0.9429	0.9426	0.9362	0.9852
R. chinensis	0.9671	0.9114	0.9118	0.9164	0.9704

Table 2. MuLan-Methyl prediction performance on 4mC-sites

Lineage	AUC	Accuracy	F1	Recall	Precision
C. equisetifolia	0.9091	0.8361	0.8315	0.8087	0.9231
F. vesca	0.9262	0.8522	0.8562	0.8801	0.9142
S. cerevisiae	0.8100	0.7371	0.7294	0.7088	0.8270
Tolypocladium	0.8161	0.7385	0.7325	0.7160	0.8108

Table 3. MuLan-Methyl prediction performance on 5hmC-sites

Lineage	AUC	Accuracy	F1	Recall	AUPR
M. musculus	0.9824	0.9649	0.9651	0.9685	0.9810
H. sapiens	0.9688	0.9488	0.9503	0.9787	0.9533

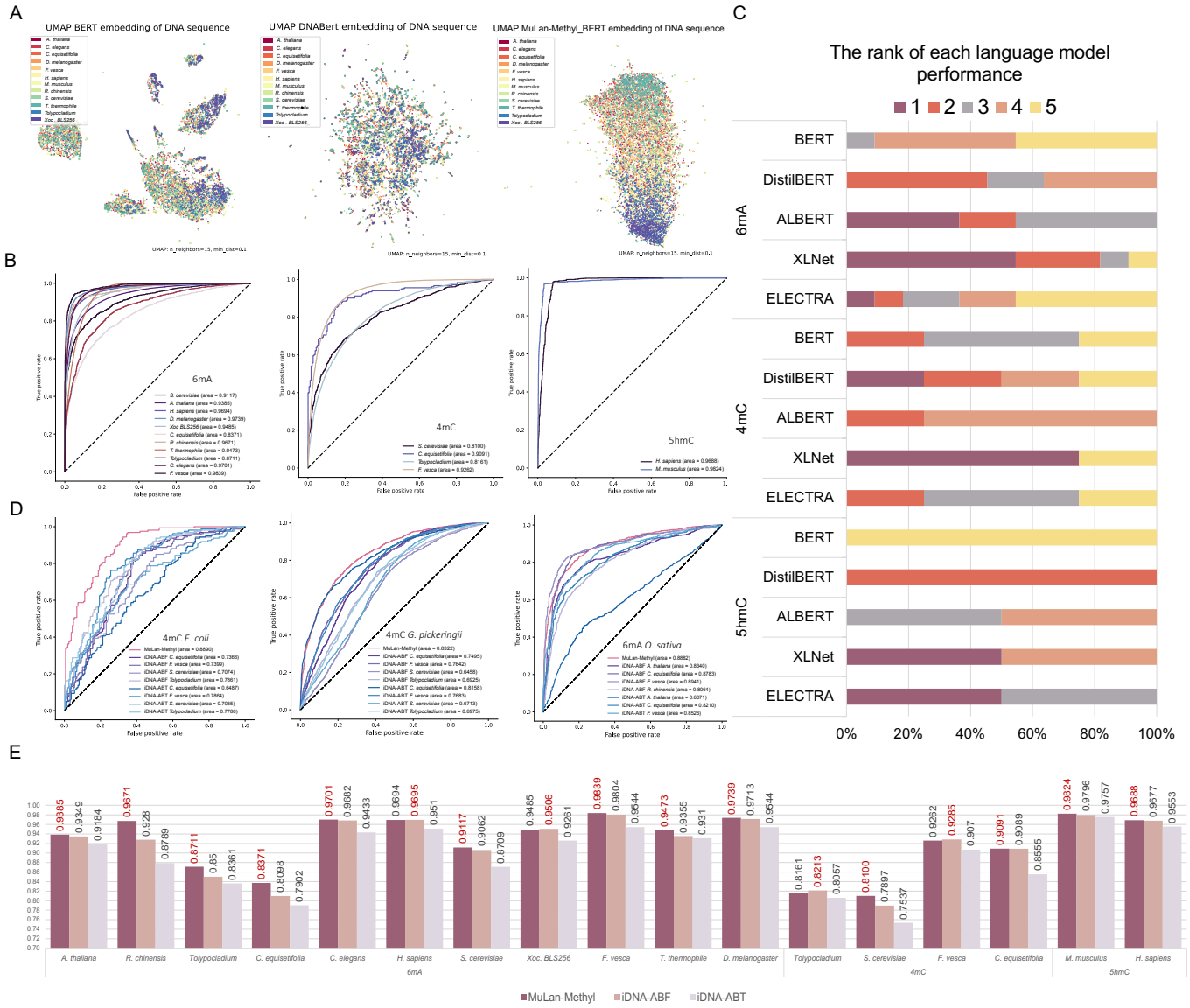

Figure 2. Model analysis and performance comparison of MuLan-Methyl. **(A)** UMAP clustering of sample representations encoded by different pre-trained models, namely BERT, DNABERT, and MuLan-Methyl_BERT (from left to right). Samples are colored by taxonomic lineage. **(B)** For MuLan-Methyl predictions of the three 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 three methylation-site types, and each of the five 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 in indicated as the width of the corresponding block. **(D)** Comparion of MuLan-Methyl against two 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. pickeringiin* 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 three methods, for each of the three methylation-site types and each of the 12 taxonomic lineages.

iDNA-ABT, two state-of-the-art methods, that are both able to predict methylation-sites for all three types, across different taxonomic lineages. For this, we used the iDNA-MS independent test set, which is considered a benchmark dataset. We report the AUC scores in Figure 2E.

In this study, MuLan-Methyl outperforms the other two methods on 13 out of 17 combinations of methylation types and taxonomic lineages. First, for 6mA-site prediction, MuLan-Methyl improves over the other methods by between 0.19% to 3.91% 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 2.03% and 0.02% AUC, on *S. cerevisiae* and *C. equisetifolia*, respectively. Finally, for 5hmC-site prediction, our method show an increase of 0.28% and 0.11% 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 0.01% and 0.21%, and for 4mC-site prediction on *Tolypocladium* and *F. vesca*, with an improvment of 0.52%, and 0.23%, respectively, over MuLan-Methyl.

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 sub-model, 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 Figure 3D-F). Positions of higher importance are concentrated around the center of the samples, and the central position always has high significance.

This observation underlines 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 observe, 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 DNA-methylation "motifs" (see Figure 3A-C). For a fixed taxonomic lineage, the three different methylation-site types each have different motifs. However, for a fixed methylation-type-site, 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 two are eukaryotes, whereas the latter is bacterial. 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 Figure 3G-I).

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

Of the eight 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.

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, however are not guaranteed to perform well on other data.

To explore the performance of MuLan-Methyl on other data, we applied the approach to the external dataset that contains three combinations of methylation types and taxonomic lineages, namely 4mC + *E. coli*, 4mC + *G. pickeringi* and 6mA + *O. sativa* L. Note that these three 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 run all taxon-wise models for 4mC-site detection, and the appropriate ones for 6mA-site detection.

MuLan-Methyl performed much better than the other two 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. pickeringi* combination, with an advantage of 1.64% over iDNA-ABT (using its *C. equisetifolia* model). On the third combination, 6mA + *O. sativa* L, MuLan-Methyl performed slightly worse (0.59%) than iDNA-ABF (using its *F. vesca* model). See Figure 2D.

DISCUSSION AND CONCLUSION

Previous studies have focused on adapting BERT to specific biological tasks using the pre-train 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 five transformer-based language models for identifying three 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

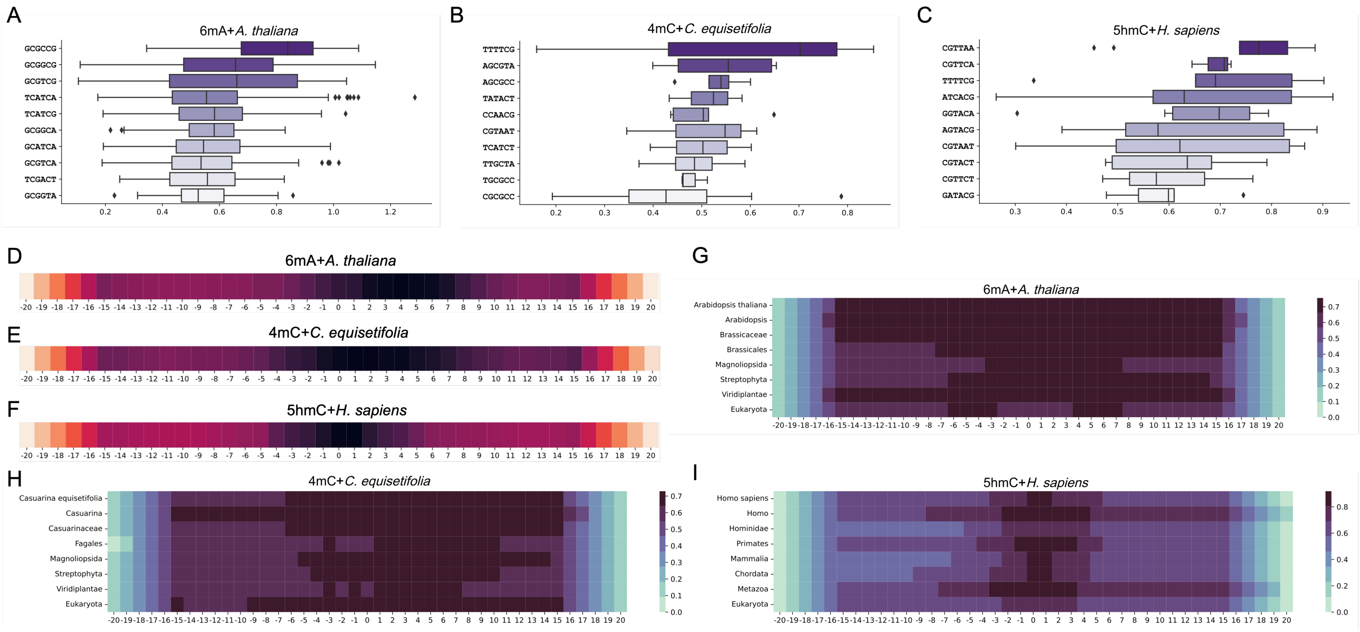

Figure 3. Interpretation of MuLan-Methyl by attention weights resulting from transformer self-attention mechanism. In **(A)**–**(C)**, we use box-plot to show the distribution of attention weights for the ten 6-mer of hightest average importance scores, for the combinations 6mA + *A. thaliana*, 5mC + *C. euisetifolia* and 5hmC + *H. sapiens*, respectively. In **(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 three combinations listed above, respectively. In **(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 three combinations listed above, respectively.

have been successfully adapted to tasks involving biological sequences.

Each sub-model in MuLan-Methyl is pre-trained and fine-tuned on the training dataset, and they then collectively predict methylation sites on an independent test dataset. The performance of MuLan-Methyl was evaluated by multiple metrics and in comparison with two existing approaches, and the method showed very good performance.

Our study also indicates that models with enhanced algorithms in the pre-training step, such as XLNET, and models with fewer parameters and less memory consumption, such as ALBERT, are more appropriate than BERT in situations with limited storage and computational resources.

In contrast to other biological domain-adaption language models, the custom corpus that we trained MuLan-Methyl on contains multi-modal 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 lineage 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 three classifiers, one for each type of methylation-site.

In conclusion, we have proposed a framework that integrates five popular NLP approaches to solve an important biological problem. MuLan-Methyl is able to detect DNA methylation sites reliably for DNA sequences from known taxonomic lineages, with slightly better performance than current state-of-the-art methods.

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 individuals models, to some extent, so as to obtain an improved ensemble prediction performance.

DATA AVAILABILITY

The benchmark dataset used in this study is accessible via the link http://lin-group.cn/server/iDNA-MS/download.html. The processed dataset used for training MuLan-Methyl and the source code are available at https://github.com/husonlab/mulan-methyl. A web server implementing the MuLan-Methyl approach will be made freely available at http://ab.cs.uni-tuebingen.de/software/mulan-methyl/.

ACKNOWLEDGEMENTS

We acknowledge support of 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).

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

FUNDING

Not Applicable

Conflict of interest statement. None declared.

REFERENCES

1. Robertson, K. D. and Wolffe, A. P. (2000) DNA methylation in health and disease. *Nature reviews genetics,* **1**(1), 11–19.
2. Moore, L. D., Le, T., and Fan, G. (2013) DNA methylation and its basic function. *Neuropsychopharmacology,* **38**(1), 23–38.
3. Armstrong, M. J., Jin, Y., Allen, E. G., and Jin, P. (2019) Diverse and dynamic DNA modifications in brain and diseases. *Human Molecular Genetics,* **28**(R2), R241–R253.
4. Tourancheau, A., Mead, E. A., Zhang, X.-S., and Fang, G. (2021) Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nature methods,* **18**(5), 491–498.
5. O'Brown, Z. K., Boulias, K., Wang, J., Wang, S. Y., O'Brown, N. M., Hao, Z., Shibuya, H., Fady, P.-E., Shi, Y., He, C., et al. (2019) Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC genomics,* **20**(1), 1–15.
6. Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A., He, C., and Zhang, Y. (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science,* **333**(6047), 1300–1303.
7. Bilyard, M. K., Becker, S., and Balasubramanian, S. (2020) Natural, modified DNA bases. *Current Opinion in Chemical Biology,* **57**, 1–7.
8. Rauluseviciute, I., Drabløs, F., and Rye, M. B. (2019) DNA methylation data by sequencing: experimental approaches and recommendations for tools and pipelines for data analysis. *Clinical epigenetics,* **11**(1), 1–13.
9. Ye, P., Luan, Y., Chen, K., Liu, Y., Xiao, C., and Xie, Z. (2016) MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic acids research,* p. gkw950.
10. Xu, H., Jia, P., and Zhao, Z. (2021) Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. *Briefings in Bioinformatics,* **22**(3), bbaa099.
11. Zeng, R., Cheng, S., and Liao, M. (2021) 4mCPred-MTL: accurate identification of dna 4mc sites in multiple species using multi-task deep learning based on multi-head attention mechanism. *Frontiers in Cell and Developmental Biology,* **9**, 664669.
12. Liu, Q., Chen, J., Wang, Y., Li, S., Jia, C., Song, J., and Li, F. (2021) DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. *Briefings in bioinformatics,* **22**(3), bbaa124.
13. Hasan, M. M., Manavalan, B., Shoombuatong, W., Khatun, M. S., and Kurata, H. (2020) i4mC-Mouse: Improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes. *Computational and structural biotechnology journal,* **18**, 906–912.
14. Jin, J., Yu, Y., and Wei, L. (2022) Mouse4mC-BGRU: Deep learning for predicting DNA N4-methylcytosine sites in mouse genome. *Methods,*.
15. Zulfiqar, H., Sun, Z.-J., Huang, Q.-L., Yuan, S.-S., Lv, H., Dao, F.-Y., Lin, H., and Li, Y.-W. (2022) Deep-4mCW2V: A sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli. *Methods,* **203**, 558–563.
16. Liang, Y., Wu, Y., Zhang, Z., Liu, N., Peng, J., and Tang, J. (2022) Hyb4mC: a hybrid DNA2vec-based model for DNA N4-methylcytosine sites prediction. *BMC bioinformatics,* **23**(1), 1–18.
17. Tran, T.-A., Pham, D.-M., Ou, Y.-Y., et al. (2021) An extensive examination of discovering 5-Methylcytosine Sites in Genome-Wide DNA Promoters using machine learning based approaches. *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* **19**(1), 87–94.
18. Cheng, X., Wang, J., Li, Q., and Liu, T. (2021) BiLSTM-5mC: A Bidirectional Long Short-Term Memory-Based Approach for Predicting 5-Methylcytosine Sites in Genome-Wide DNA Promoters. *Molecules,* **26**(24), 7414.
19. Li, Z., Jiang, H., Kong, L., Chen, Y., Lang, K., Fan, X., Zhang, L., and Pian, C. (2021) Deep6mA: a deep learning framework for exploring similar patterns in DNA N6-methyladenine sites across different species. *PLoS computational biology,* **17**(2), e1008767.
20. Rehman, M. U., Tayara, H., Zou, Q., and Chong, K. T. (2022) i6mA-Caps: a CapsuleNet-based framework for identifying DNA N6-methyladenine sites. *Bioinformatics,* **38**(16), 3885–3891.
21. Zeng, R. and Liao, M. (2021) 6mAPred-MSFF: a deep learning model for predicting DNA N6-Methyladenine sites across species based on a multi-scale feature fusion mechanism. *Applied Sciences,* **11**(16), 7731.
22. Liu, M., Sun, Z.-L., Zeng, Z., and Lam, K.-M. (2022) MGF6mARice: prediction of DNA N6-methyladenine sites in rice by exploiting

molecular graph feature and residual block. *Briefings in Bioinformatics,* **23**(3), bbac082.

23. Tsukiyama, S., Hasan, M. M., Deng, H.-W., and Kurata, H. (2022) BERT6mA: prediction of DNA N6-methyladenine site using deep learning-based approaches. *Briefings in Bioinformatics,* **23**(2), bbac053.
24. Tahir, M., Hayat, M., Ullah, I., and Chong, K. T. (2020) A deep learning-based computational approach for discrimination of dna n6-methyladenosine sites by fusing heterogeneous features. *Chemometrics and Intelligent Laboratory Systems,* **206**, 104151.
25. Le, N. Q. K. and Ho, Q.-T. (2022) Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes. *Methods,* **204**, 199–206.
26. Tang, X., Zheng, P., Li, X., Wu, H., Wei, D.-Q., Liu, Y., and Huang, G. (2022) Deep6mAPred: A CNN and Bi-LSTM-based deep learning method for predicting DNA N6-methyladenosine sites across plant species. *Methods,*.
27. Hasan, M. M., Basith, S., Khatun, M. S., Lee, G., Manavalan, B., and Kurata, H. (2021) Meta-i6mA: an interspecies predictor for identifying DNA N 6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Briefings in Bioinformatics,* **22**(3), bbaa202.
28. Chen, J., Zou, Q., and Li, J. (2022) DeepM6ASeq-EL: prediction of human N6-methyladenosine (m6A) sites with LSTM and ensemble learning. *Frontiers of Computer Science,* **16**(2), 1–7.
29. Zhang, Y., Liu, Y., Xu, J., Wang, X., Peng, X., Song, J., and Yu, D.-J. (2021) Leveraging the attention mechanism to improve the identification of DNA N6-methyladenine sites. *Briefings in Bioinformatics,* **22**(6), bbab351.
30. Yang, X., Ye, X., Li, X., and Wei, L. (2021) iDNA-MT: identification DNA modification sites in multiple species by using multi-task learning based a neural network tool. *Frontiers in genetics,* **12**, 663572.
31. Lv, H., Dao, F.-Y., Zhang, D., Guan, Z.-X., Yang, H., Su, W., Liu, M.-L., Ding, H., Chen, W., and Lin, H. (2020) iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. *Iscience,* **23**(4), 100991.
32. Yu, Y., He, W., Jin, J., Xiao, G., Cui, L., Zeng, R., and Wei, L. (2021) iDNA-ABT: advanced deep learning model for detecting DNA methylation with adaptive features and transductive information maximization. *Bioinformatics,* **37**(24), 4603–4610.
33. Jin, J., Yu, Y., Wang, R., Zeng, X., Pang, C., Jiang, Y., Li, Z., Dai, Y., Su, R., Zou, Q., et al. (2022) iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. *Genome biology,* **23**(1), 1–23.
34. Zheng, Z., Le, N. Q. K., and Chua, M. C. H. (2022) MaskDNA-PGD: An innovative deep learning model for detecting DNA methylation by integrating mask sequences and adversarial PGD training as a data augmentation method. *Chemometrics and Intelligent Laboratory Systems,* p. 104715.
35. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805,*.
36. Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. (2021) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics,* **37**(15), 2112–2120.
37. Zhang, Y.-z., Yamaguchi, K., Hatakeyama, S., Furukawa, Y., Miyano, S., Yamaguchi, R., and Imoto, S. (2021) On the application of BERT models for nanopore methylation detection. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* IEEE pp. 320–327.
38. Seong, H. J., Han, S.-W., and Sul, W. J. (2021) Prokaryotic DNA methylation and its functional roles. *Journal of Microbiology,* **59**(3), 242–248.
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017) Attention is all you need. *Advances in neural information processing systems,* **30**.
40. Zeng, W., Gautam, A., and Huson, D. H. (2022) DeepToA: an ensemble deep-learning approach to predicting the theater of activity of a microbiome. *Bioinformatics,* **38**(20), 4670–4676.
41. Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108,*.
42. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019) Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692,*.
43. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019) Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems,* **32**.
44. Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020) Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555,*.
45. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019) Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942,*.
46. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020) Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964,*.
47. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics,* **36**(4), 1234–1240.
48. Lample, G. and Conneau, A. (2019) Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291,*.
49. Lupo, U., Sgarbossa, D., and Bitbol, A.-F. (2022) Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *bioRxiv,*.
50. Mock, F., Kretschmer, F., Kriese, A., Böcker, S., and Marz, M. (2022) Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. *Proceedings of the National Academy of Sciences,* **119**(35), e2122636119.
51. Yang, M., Huang, L., Huang, H., Tang, H., Zhang, N., Yang, H., Wu, J., and Mu, F. (2022) Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. *Nucleic acids research,* **50**(14), e81–e81.
52. Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., and Karsch-Mizrachi, I. (Jan, 2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools.. *Database (Oxford),* **2020**.
53. Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., and Hugenholtz, P. (2021) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res,* **10**.
54. Chen, W., Lv, H., Nie, F., and Lin, H. (2019) i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics,* **35**(16), 2796–2800.
55. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (October, 2020) Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* Online: Association for Computational Linguistics pp. 38–45.
56. Yamada, K. and Hamada, M. (2022) Prediction of RNA–protein interactions using a nucleotide language model. *Bioinformatics Advances,* **2**(1), vbac023.