Journal Pre-proof

A comprehensive review of computational methods for predicting DNA N4-methylcytosine sites

Zihang Wang, Aoyun Geng, Junlin Xu, Yajie Meng, Zilong Zhang, Leyi Wei, Quan Zou, Feifei Cui

PII: S0141-8130(25)08778-1

DOI: https://doi.org/10.1016/j.ijbiomac.2025.148221

Reference: BIOMAC 148221

To appear in: *International Journal of Biological Macromolecules*

Received date: 19 August 2025

Revised date: 27 September 2025

Accepted date: 8 October 2025

Please cite this article as: Z. Wang, A. Geng, J. Xu, et al., A comprehensive review of computational methods for predicting DNA N4-methylcytosine sites, *International Journal of Biological Macromolecules* (2024), https://doi.org/10.1016/j.ijbiomac.2025.148221

A Comprehensive Review of Computational Methods for Predicting DNA N⁴-Methylcytosine Sites

Zihang Wang[1, a], Aoyun Geng[1, a], Junlin Xu[2], Yajie Meng[3], Zilong Zhang[1], Leyi Wei[4], Quan Zou[5], Feifei Cui[1, *]

[1]*School of Computer Science and Technology, Hainan University, Haikou, 570228, China*

[2]*School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, 430081, Hubei, China*

[3]*School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, 430200, Hubei, China*

[4]*Centre for Artificial Intelligence driven Drug Discovery, Faculty of Applied Science, Macao Polytechnic University, Macao SAR, China*

[5]*Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China*

* Corresponding author: Feifei Cui (feifeicui@hainanu.edu.cn)

Abstract

N4-methylcytosine (4mC) is a distinct form of DNA methylation that plays a critical role in various biological processes by protecting bacterial DNA from degradation and participating in the regulation of gene expression. With advances in technology, computational approaches have increasingly replaced traditional experimental methods, which are often associated with high costs, prolonged processing times, and labor-intensive workflows. Over the past five years, a growing number of machine learning (ML) and deep learning (DL) models have been developed to predict 4mC sites. In this review, we provide a systematic overview of these computational methods, focusing on model architectures and comparing the strengths and limitations of ML- and DL-based approaches. To facilitate future tool development, we have collected and organized commonly used databases and benchmark datasets relevant to 4mC prediction. In addition, we compared several recently proposed methods to highlight their respective strengths and capabilities. Finally, we highlight the current challenges and opportunities in the field, aiming to facilitate the development of more accurate and robust predictive frameworks for 4mC methylation.

Keywords: N4-methylcytosine; Deep learning; Algorithm benchmarking

Introduction

DNA methylation is a fundamental epigenetic process that plays a crucial role in the proper regulation of transcription, embryonic development, genomic imprinting, genome stability, and chromatin structure [1, 2]. With the advancement of molecular biology techniques, researchers have been able to investigate DNA methylation more deeply and subsequently discovered its intricate connections with other epigenetic modifications, such as histone modifications. These findings suggest that there are complex interactions among different epigenetic regulatory mechanisms [3]. Aberrant DNA methylation is closely associated with

[a] These authors contributed equally to this work.

the onset and progression of various diseases, particularly cancer [4], where alterations in DNA methylation patterns have emerged as potential early biomarkers [5]. As a result, increasing attention has been directed toward the study of DNA methylation. Currently, based on the position of the methyl group attachment, DNA methylation can be broadly categorized into three major types: N4-methylcytosine (4mC), 5-methylcytosine (5mC), and N6-methyladenine (6mA) [6]. Every kind of methylation modification plays a distinct biological role. 5mC is established through the enzymatic transfer of a methyl group from S-adenosylmethionine (SAM) to the fifth carbon position of cytosine, a reaction catalyzed by DNA methyltransferases (DNMTs). This modification is typically associated with transcriptional repression; in promoter regions, 5mC often interferes with transcription factor binding, ultimately resulting in gene silencing [7]. 6mA refers to the methylation occurring at the nitrogen-6 (N6) position of the adenine (A) base. This modification plays crucial roles in transcriptional regulation, nucleosome positioning, and cell cycle control [8-10]. Another form of DNA methylation, 4mC, involves methylation at the nitrogen-4 (N4) amino group of the cytosine base. Functioning as a key component of restriction-modification systems, 4mC protects bacterial DNA from degradation and also contributes to the regulation of gene expression. In addition, DNA methylation can alter the physical properties of the DNA molecule, which may in turn influence its interactions with proteins and other biomolecules, thereby modulating fundamental DNA metabolic processes, such as replication, transcription, and repair [11, 12]. Compared to 5mC and 6mA, 4mC is more prevalent in prokaryotic genomes. Figure 1A illustrates the formation mechanisms of the 4mC methylation. However, due to its unique chemical properties, 4mC cannot be effectively detected using conventional bisulfite sequencing methods [13]. Therefore, developing novel approaches to predict 4mC methylation is of great importance.

In 2010, Single Molecule Real-Time (SMRT) [14] sequencing, a third-generation sequencing technology, played a significant role in the identification of 4mC methylation. SMRT sequencing enables the direct detection of DNA modifications, including 4mC, without requiring bisulfite conversion or amplification, achieving single-base resolution. In the subsequent years, various experimental approaches have been developed to profile 4mC methylation. However, these traditional methods are often hindered by high costs, prolonged processing times, and labor-intensive workflows. Meanwhile, the advent and ongoing advancement of machine learning and deep learning technologies have revolutionized numerous fields [15-17]. Machine learning has been extensively employed in bioinformatics. By transforming biological sequences into numerical feature representations and applying classification algorithms, substantial progress has been made in identifying 4mC methylation sites [18]. In 2017, iDNA4mC [19] introduced the first SVM-based predictor for identifying 4mC sites. It used features derived from nucleotide chemical properties to train the classifier. Later, several models continued to employ SVMs while exploring new feature encoding methods. For example, 4mcPred-SVM [20] combined k-mer-based dinucleotide frequencies, binary encoding of mono- and dinucleotides, and localized position-specific dinucleotide frequencies. In contrast, 4mCPred [21] used position-specific trinucleotide propensity (PSTNP) and EIIP as feature descriptors. In 2019, 4mcPred-IFL [22] introduced an iterative feature representation learning algorithm. This approach progressively optimized features to better distinguish 4mC from non-4mC sites. Meta-4mCpred [23] enhanced prediction by integrating outputs from multiple base predictors, forming an SVM-based meta-predictor. This ensemble approach improved both accuracy and generalization. At the same time, 4mCpred-EL [24] combined four machine learning algorithms with seven types of feature encodings. It achieved the first genome-wide prediction of 4mC sites in the mouse. In 2020, DNA4mC-LIP [25] proposed a linear ensemble method. It iteratively combined multiple existing 4mC predictors to build a stronger model. In the same year, i4mC-Mouse [26] used six different DNA sequence encoding schemes and

a random forest (RF) algorithm for prediction. Beyond traditional machine learning, deep learning has also been applied to 4mC prediction. DeepTorrent [27] combined a CNN with a BiLSTM network and added an attention mechanism. This design improved performance in predicting 4mC methylation sites. Deep4mcPred [28] was the first to combine a residual network (ResNet) with a BiLSTM architecture, and further incorporated an attention mechanism to construct a multi-layer deep learning framework for 4mC site identification. Although these models have demonstrated good performance in predicting 4mC methylation, most were specifically designed for a single type of DNA methylation and thus lack generalizability. To address this limitation, iDNA-MS [29] was the first model developed to simultaneously predict three types of DNA methylation: 5hmC, 6mA, and 4mC. This model extracted features using three encoding schemes—k-tuple nucleotide frequency components (KNFC), nucleotide chemical property and nucleotide frequency (NCPNF), and mono-nucleotide binary encoding (MNBE)—and employed an RF algorithm as the classifier. Following these developments, an increasing number of computational models have been proposed for predicting 4mC methylation sites, highlighting the urgent need for a systematic organization and classification of these methods.

In this study, we systematically summarize and categorize the 4mC methylation site prediction tools developed in the past five years based on their methodological approaches. We also conduct benchmarking experiments on six representative models to compare their performance and analyze their respective strengths and weaknesses. Additionally, we discuss the current challenges and potential opportunities in the prediction of 4mC methylation sites.

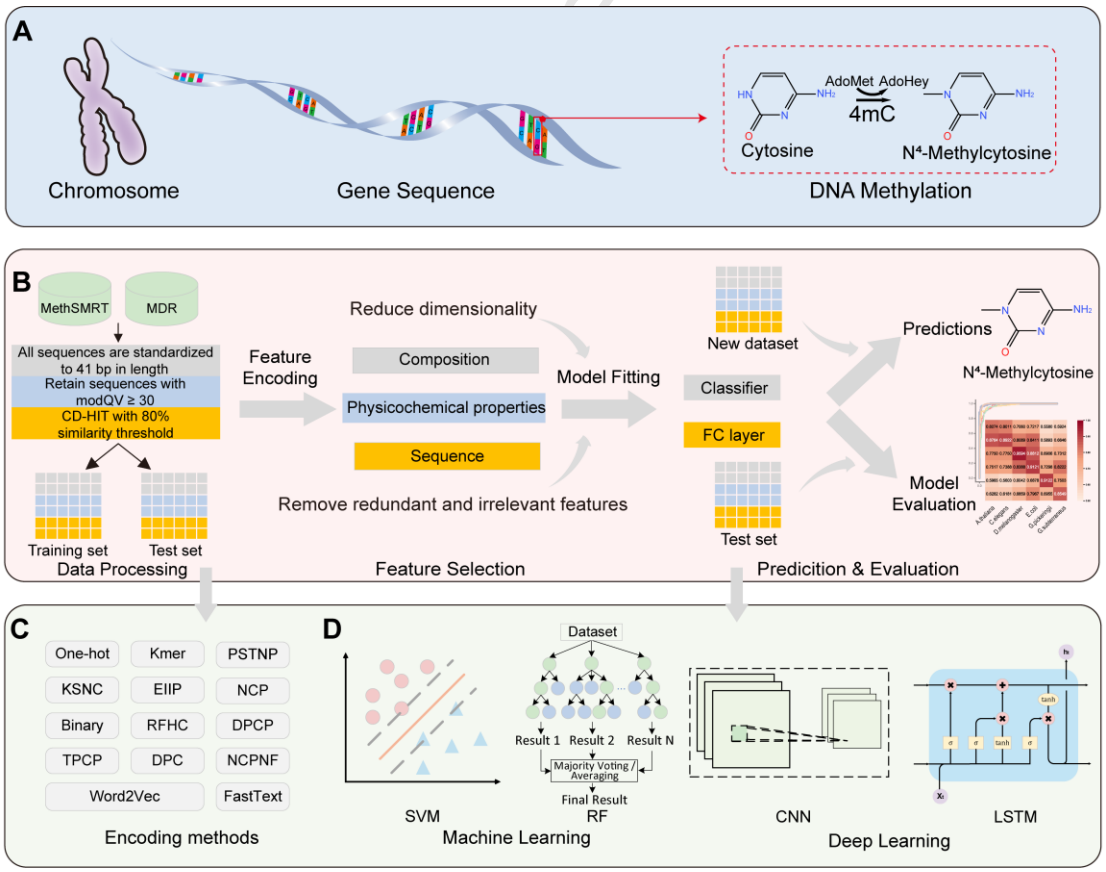

Figure 1. An integrated framework for predicting 4mC methylation using machine learning and deep learning approaches. **A.** The formation processes of 4mC in eukaryotic cells. **B.** Detailed workflow for 4mC methylation prediction, including data preprocessing, feature encoding and selection, model training, prediction, and evaluation. **C-D.** Commonly used encoding schemes and predictive models, respectively.

Computational Approaches for 4mC Site Prediction

In the context of 4mC methylation site prediction, the task is generally formulated as a binary classification problem, aiming to accurately determine whether a given DNA sequence contains a 4mC modification. The computational strategies employed for this task are generally divided into two major groups: conventional machine learning techniques and deep learning approaches. Traditional machine learning algorithms commonly employed include RF, SVM, and CatBoost [30, 31], which rely on handcrafted features and statistical learning strategies [32]. In contrast, deep learning models such as CNNs, RNNs, Transformers, and pretrained language models have shown great potential in automatically learning complex representations from raw DNA sequences [33-35]. Figures 1B–D illustrate the standard computational pipeline for predicting 4mC sites, including frequently adopted feature encoding techniques and representative machine learning and deep learning models. In this study, we classify the models into two categories for a more systematic discussion: machine learning-based models and deep learning-based models. This taxonomy provides a comprehensive perspective on methodological developments in 4mC site prediction, facilitating a clearer understanding of their respective advantages, design choices, and performance characteristics.

Machine learning-based models

Traditional ML methods for 4mC site prediction typically follow a four-step workflow. First, an appropriate dataset is selected. Second, DNA sequences in the dataset are transformed into numerical representations using various feature encoding techniques. Third, the resulting feature vectors are fed into a classifier for model training and prediction. Finally, the models are assessed to determine their performance using standard evaluation metrics. Over the past five years, relatively few studies have employed traditional machine learning approaches for predicting 4mC sites. Below, we summarize the representative models that have been proposed using these methods.

***MvLapKSRC-HSIC* (2023)**

MvLapKSRC-HSIC [36] model was evaluated across multiple species. The study employed three feature extraction approaches—PSTNP, Nucleotide Chemical Property (NCP), and DNA Physicochemical Properties (DPP)—which were integrated into a multi-view representation to enhance the model's discriminative capability. To capture the nonlinear relationships among sequences, the model incorporated a kernel sparse representation classifier (KSRC). In addition, the $L_{2,1}$ norm, Laplacian graph regularization, and Hilbert–Schmidt Independence Criterion (HSIC) were introduced as regularization terms to promote sparsity, preserve local structural information, and encourage view-wise independence, respectively.

***MMC-KHFIS* (2023)**

MMC-KHFIS [37] model extracts features from DNA sequences using PSTNP to generate feature vectors, which are subsequently processed with the Kernelized High-order Fuzzy Inference System (KHFIS) [38]. Additionally, the model incorporates the Maximum multi-correntropy (MMC) to adapt complex error distributions.

***SSR-RVFL* (2024)**

SSR-RVFL [39] model is a shallow machine learning approach built on the Random Vector Functional Link

(RVFL) network. It first employs PSTNP to extract sequence features, and then enhances representation by integrating features generated through multiple activation functions. To select features and improve robustness against noise, it introduces structured sparse regularization terms. These terms include the $L_{2,1}$ matrix norm and G_1 norm, which help capture dependencies among different features.

***DNA-MP* (2023)**

DNA-MP [40] model introduces a new sequence encoding method called POCD-ND. This approach calculates normalized differences in k-mer occurrence frequencies between modified and unmodified classes to capture position-related distribution patterns, and the resulting features are input into a Deep Forest classifier for DNA modification prediction. Deep Forest is an ensemble learning-based method that constructs multi-layered forest structures to enhance classification performance. Notably, DNA-MP is capable of predicting three types

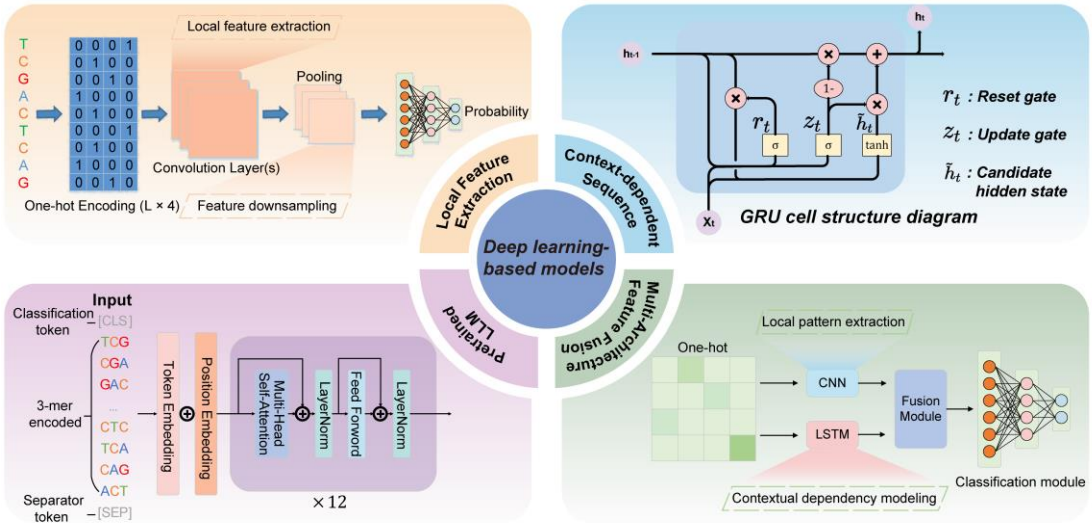

Figure 2. Deep learning model classification diagram. The figure categorizes deep learning-based models into four distinct types: those focusing on local feature extraction, those capturing contextual sequence features, those leveraging pre-trained large language models, and those employing multi-architecture fusion. Representative model architectures are illustrated under each category.

of DNA methylation modifications.

***4mCBERT* (2023)**

4mCBERT [41] model utilized 4mC site data from the MethSMRT database [6]. The data were divided into training, validation, and independent test sets for evaluation. The model adopted six encoding methods—One-hot, EIIP, NCP, Word2Vec, PCP, and Chemical BERT—to represent DNA sequences. The encoded feature vectors were concatenated and input into a CatBoost classifier for prediction.

***DMKL-HFIS* (2023)**

DMKL-HFIS [42] model extracts position-specific trinucleotide features using PSTNP. It employs a KHFIS to construct multiple fuzzy kernel matrices, capturing the nonlinear relationships among sequences. Each fuzzy kernel matrix corresponds to an individual fuzzy rule and is generated via kernel functions. Furthermore, the model integrates deep learning by leveraging a multilayer neural network to perform nonlinear fusion of the multiple fuzzy kernel matrices. The final fused kernel matrix is then utilized as the input kernel function for classification.

Deep learning-based models

Unlike traditional machine learning methods that rely on manually crafted features, deep learning models are capable of automatically extracting both local and global sequence features, enabling effective discrimination between methylated and unmethylated sites [43]. Through backpropagation, these models iteratively optimize their parameters, achieving high prediction accuracy even in large and complex biological sequences. Given that the performance of deep learning models is closely tied to how effectively they extract discriminative features from DNA sequences, feature extraction strategies have become a critical differentiator among models. Therefore, in the following sections, we categorize deep learning models developed over the past five years based on their distinct feature extraction methods. An overview of this categorization is illustrated in Figure 2.

Models Based on Local Feature Extraction

In 4mC methylation site prediction, local feature extraction is typically performed using CNNs. The general workflow begins by encoding DNA sequences into two-dimensional matrices via one-hot encoding or other embedding schemes. One-dimensional convolutional filters are applied to scan the sequence and extract local motif-like patterns, followed by pooling operations that reduce dimensionality while retaining salient features within each receptive field. The resulting feature maps are flattened and passed through one or more fully connected layers to integrate the information, ultimately yielding a prediction probability for methylation status. The core operation of a CNN is the convolution, which extracts local features by sliding a convolutional kernel across the input sequence [44]. Mathematically, for a one-dimensional sequence $X = [x_1, x_2, \dots, x_L]$, the convolution output at position i can be expressed as:

$$y_i = f\left(\sum_{j=0}^{k-1} W_j \cdot x_{i+j} + b\right) \quad (1)$$

where W_j denotes the j-th weight of the convolutional kernel, b is the bias term, $f(\cdot)$ is a nonlinear activation function.

***i4mC-Deep* (2021)**

i4mC-Deep [45] model integrates NCP and nucleotide density (ND) into a unified feature vector, which is then fed into a CNN for classification. The model is trained with the Adam optimizer and evaluated through 10-fold cross-validation to assess its predictive performance.

***DCNN-4mC* (2021)**

DCNN-4mC [46] employs one-hot encoding, mapping each nucleotide to a unique 4-dimensional binary vector, which is then input into a CNN. To more comprehensively capture sequence features, the model incorporates skip connections that concatenate shallow and deep feature representations. It utilizes a custom loss function that combines the Dice Loss Coefficient (DLC) with Weighted Cross-Entropy (WCE), and optimizes model parameters using the Stochastic Gradient Descent (SGD) algorithm.

***MSNet-4mC* (2022)**

MSNet-4mC [47] model uses one-hot encoding to represent DNA sequences. It first feeds the encoded vectors into an initial convolutional layer. After that, the vectors pass through multiple parallel convolutional branches with different kernel sizes. This design allows the model to extract features at various receptive fields and capture sequence dependencies across multiple scales. The model also includes residual connections to facilitate gradient flow and network optimization.

***Deep-4mCGP* (2022)**

Deep-4mCGP [48] model is designed for predicting 4mC sites in *G. pickeringii*. It first uses k-mer

nucleotide composition and binary encoding to extract features. The model refines the feature representation using correlation analysis, gradient boosting decision trees (GBDT), and incremental feature selection (IFS). Then, it feeds the optimized feature vectors into a 1D-CNN for training and prediction.

***GS-MLDS* (2023)**

GS-MLDS [49] model employs a k-mer strategy to fragment DNA sequences into 3-mers. It then embeds these 3-mers into 100-dimensional vectors using the Word2Vec algorithm. These embeddings are input into a 1D-CNN for feature extraction. Moreover, the model introduces a Multilayer Dynamic Ensemble System (MLDS) [50]. Each layer correctly predicted samples and the samples are passed to the training set of the next layer, while misclassified samples are moved to its test set. This approach allows effective knowledge transfer across layers. Additionally, grid search is applied at each layer to optimize the combination of model weights, enhancing predictive performance.

***fastText+CNN based model* (2024)**

Nguyen et al [51] tested three embedding strategies for DNA sequences: Word2Vec, fastText, and variable k-mer embeddings. And they used CNNs, RNNs, and their variants, such as LSTM and GRU for prediction. After trying various combinations, the study ultimately selected fastText with CNN as the final model architecture.

***Hyb4mC* (2022)**

Hyb4mC [52] model employs DNA2vec to segment DNA sequences into 6-mers, which are then mapped into 100-dimensional vectors using a pretrained DNA2vec model to build the feature matrix. Depending on the sample size, the model chooses different training and prediction strategies. For small datasets, it employs a capsule network with dynamic routing. For large datasets, it combines attention mechanisms with CNNs.

Models Based on Context-dependent Sequential Feature Extraction

These models are built to capture context-dependent relationships in DNA sequences [53]. They often uses architectures such as RNNs and their variants, such as LSTMs and GRUs. Transformer models are also used to capture long-range dependencies. In the context of 4mC methylation site prediction, widely adopted architectures include Bi-GRU and Transformer. Below, we provide a thorough explanation of these two architectures:

BiGRU is an extension of the standard gated recurrent unit (GRU) architecture, designed to capture more comprehensive contextual information by processing the input sequence in both forward and backward directions simultaneously [54, 55]. For an input sequence $\{x_1, x_2, \dots, x_T\}$, BiGRU computes both the forward hidden state $\overrightarrow{h_t}$ and the backward hidden state $\overleftarrow{h_t}$, and concatenates them to represent the current position:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{2}$$

The forward GRU computes the hidden state at time step t as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{3}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{4}$$

$$\widetilde{h_t} = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{5}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h_t} \tag{6}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, $\odot$ indicates element-wise multiplication, W_* and U_* are the weight matrices for the current input and previous hidden state, respectively, and b_* represents the bias vectors. The backward GRU follows the same computation process but operates in the reverse temporal direction.

The Transformer encoder is a deep neural network architecture based on self-attention mechanisms,

capable of modeling dependencies between any positions in a sequence [56, 57]. For DNA sequences, the input is typically segmented into k-mer tokens, each of which is embedded into a vector x_i, and combined with a positional encoding p_i to form the input representation:

$$\widetilde{x_i} = x_i + p_i \tag{7}$$

The self-attention mechanism computes pairwise relationships between all positions in the sequence. For a single attention head, the query, key, and value matrices are calculated as:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{8}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

Multi-head attention performs several self-attention operations in parallel and concatenates the results:

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \tag{10}$$

After passing through multiple stacked Transformer encoder layers, the model outputs a set of context-aware representations for each token: $\{h_1, h_2, \dots, h_n\}$.

iDNA-MT **(2021)**

iDNA-MT [58] model first segments nucleotide sequences into overlapping 2-grams, which are then converted into high-dimensional vectors using one-hot encoding. These encoded vectors are fed into a BiGRU network. Leveraging a multi-task learning framework, the model is capable of simultaneously predicting both 4mC and 6mA methylation sites across multiple species.

Mouse4mC-BGRU **(2022)**

Mouse4mC-BGRU [59] model incorporates an adaptive embedding module to enhance sequence representation. DNA sequences are first converted into k-mer representations—in this study, 1-mer is used—by segmenting the sequence into individual nucleotides. Each 1-mer is then mapped to a randomly initialized vector, which is updated through backpropagation during training. These embedded vectors are subsequently fed into a BiGRU module for further feature extraction and information integration.

i4mC-GRU **(2023)**

i4mC-GRU [60] model integrates sequence embedding with a BiGRU architecture. DNA sequences are first represented using overlapping 2-mers, with each k-mer assigned a unique index. A sliding window moves across the sequence to generate an index vector. This vector is then fed into a BiGRU network.

4mCPred-MTL **(2021)**

4mCPred-MTL [61] model segments DNA sequences into 2-gram nucleotide subsequences. It then converts these subsequences into one-hot encoded feature vectors. The model feeds these vectors into a shared module that uses a Transformer encoder. This study also adds positional encodings based on trigonometric functions to help the model learn positional information.

iDNA-ABT **(2021)**

iDNA-ABT [62] model employs an Adaptive Embedding Module to convert DNA sequences into numerical vectors. These vectors are then input into a BERT encoder. The multi-head attention mechanism in the BERT captures contextual relationships and global dependencies within the sequence. After that, the extracted feature vectors are passed through a fully connected network to generate prediction probabilities. Additionally, the model also applies the Transductive Information Maximization (TIM) loss function during training.

StableDNAm **(2023)**

StableDNAm [63] model segments DNA sequences into k-mers at first, including 3-mer, 4-mer, 5-mer, and 6-mer, to create initial sequence features. Then, features are extracted using BERT encoders with different dropout rates. The model linearly weights and fuses the multi-scale features. A 2D-SENET module [64]

adaptively recalibrates the features to adjust their importance. The model also introduces a contrastive learning strategy. In this strategy, positive and negative sample pairs are constructed using different dropout rates. During training, the model applies the InfoNCE loss function to maximize similarity between positive pairs and minimize it between negative pairs.

Models Based on Feature Extraction Using Pretrained Large Language Models

In recent years, pretrained language models have attracted attention in 4mC methylation site prediction. DNABERT [65] is commonly used in this context. By fine-tuning DNABERT on labeled 4mC datasets, the model can learn both local and global sequence features associated with 4mC modifications. Compared to traditional deep learning models, DNABERT often shows better generalization capabilities because of its initial training on large-scale datasets. During pretraining, DNA sequences are usually segmented into contiguous k-mers. A special [CLS] token is added at the beginning to represent the overall sequence classification information, and a [SEP] token is added at the end to indicate sequence termination. Additional special tokens such as [PAD] (padding), [UNK] (unknown), and [MASK] (masking) are also introduced. Following the masked language modeling (MLM) strategy used in BERT, 15% of the k-mer tokens in the sequence are randomly masked. After masking, the model uses a Transformer encoder with multi-head attention to capture global contextual information. It is trained in a self-supervised manner on a large-scale dataset constructed from the human reference genome[66]. Through this process, DNABERT learns to predict the masked k-mers, and acquires a generalized representation of DNA sequences. To enable DNABERT to be more effectively applied to 4mC prediction, it is necessary to fine-tune the model. Specifically, the labeled 4mC datasets are segmented into k-mers and converted into token sequences, which are then input into the pre-trained DNABERT model for training. Through backpropagation, the model parameters are updated to adapt to the 4mC prediction task. Ultimately, the high-dimensional vectors carrying the global sequence features of 4mC are fed into a fully connected layer for prediction. At present, most large language models applied in 4mC prediction are based on DNABERT. In our investigation, we also identified DNABERT-2 [67], GeneBERT [68], Evo [69] and Nucleotide Transformer [70], both of which are designed specifically for DNA sequences. If applied appropriately to 4mC prediction, these models may achieve promising performance.

iDNA-ABF **(2022)**

iDNA-ABF [71] model tokenizes DNA sequences using both 3-mer and 6-mer representations. Each representation is independently encoded by pretrained DNABERT encoders. The model then integrates these multi-scale representations through a gated fusion module. It also applies adversarial training in the classification module using the Fast Gradient Method (FGM) to reduce overfitting.

iDNA-ITLM **(2024)**

iDNA-ITLM [72] model first converts nucleotide bases into numerical indices based on a mapping dictionary, then introduces a data augmentation technique termed self-replication, which concatenates short DNA sequences (41 bp) repeatedly to form longer sequences. An adaptive embedding method is employed to embed these sequences into two-dimensional matrices, effectively enlarging the receptive field. Subsequently, pretrained DNABERT is utilized to further extract contextual features before training.

iDNA-OpenPrompt **(2024)**

iDNA-OpenPrompt [73] model, built on the OpenPrompt framework, reformulates DNA methylation site prediction as a cloze-style classification task. The model constructs a DNA vocabulary comprising k-mer combinations with k ranging from 1 to 6. Using a BERT tokenizer, sequences are converted into token sequences containing [MASK] tokens. The central base of the methylation site (C for 4mC/5hmC and A for

6mA) is used to generate label tokens. The model maps BERT's predictions for the [MASK] tokens to biological labels indicating methylated or unmethylated states, leveraging a pretrained BERT model for final prediction.

***Methyl-GP* (2025)**

Methyl-GP [74]model fine-tunes a pretrained DNABERT to extract features from DNA sequences. Four k-mer tokenization schemes (k = 3, 4, 5, 6) segment the sequences into tokens of varying lengths. Datasets from multiple species are mixed and input into DNABERT for joint training and fine-tuning. The fine-tuned feature extractors produce embedding vectors corresponding to each k-mer scale. To address the challenge of integrating multi-view embeddings, the model introduces a feature fusion module. During decoding, a relation matrix W^v combined with structured sparse regularization is employed to capture both consistent information shared across embeddings and complementary information unique to each embedding, generating an information-complete auxiliary representation $\hat{z}_n$. In the encoding phase, guided by $\hat{z}_n$ and W^v, the original embeddings are mapped to a compact integrated representation z_n.

Models Based on Multi-Architecture Feature Fusion

Different neural network architectures possess unique strengths in feature extraction [75, 76]. Consequently, multi-architecture fusion approaches have been naturally adopted in 4mC methylation site prediction. By integrating these diverse structures, such models can simultaneously capture local patterns, global contextual information, and sequential dependencies, thereby enhancing their discriminative ability for methylation site identification.

***MultiScale-CNN-4mCPred* (2023)**

MultiScale-CNN-4mCPred [77] model first maps nucleotides to integers, which are then passed through an adaptive embedding layer to generate dense feature vectors of size 41×8. A Bi-LSTM layer is employed to capture contextual semantic information, followed by multi-scale CNNs that extract local features at varying scales. This combination enables accurate prediction of 4mC sites.

***Mus4mCPred* (2024)**

Mus4mCPred [78] model utilizes four encoding schemes: character encoding, token encoding, EIIP encoding, and Word2Vec encoding. It adopts a multi-view learning framework comprising three branches: a Word2Vec-BiLSTM branch, a token-based residual CNN-BiLSTM branch, and a character/EIIP-CNN branch. Feature vectors extracted from these branches are concatenated directly and fed into a fully connected layer for prediction.

***DeepSF-4mC* (2024)**

DeepSF-4mC [11] model utilizes the LazyPredict package to select four optimal encoding schemes—Binary, EIIP, ENAC, and NCP—which are concatenated into a hybrid feature vector and input into a pretrained deep neural network (DNN). Simultaneously, one-hot encoded feature vectors are fed separately into a pretrained CNN and LSTM model. The models' prediction probabilities are used as meta-features. These features are then combined using a linear regression model to generate the final prediction.

***DeepPGD* (2024)**

DeepPGD [79] model integrates temporal convolutional networks (TCN) and BiLSTM. TCN captures local structural features, while BiLSTM models global sequence dependencies. The model also uses a multi-head attention mechanism to dynamically weight and fuse the outputs from TCN and BiLSTM. This unified model is capable of predicting methylation sites across three different species.

In summary, we have reviewed models developed over the past five years for predicting 4mC methylation sites. A comprehensive overview is provided in Table 1 for reference by researchers in the field.

Comparative Analysis of Machine Learning and Deep Learning Approaches

Traditional machine learning models often perform well on small sample datasets due to their reliance on handcrafted feature extraction guided by domain knowledge. These models can flexibly combine various encoding schemes and feature selection methods, allowing for easier parameter tuning and architectural

Table 1 Existing Models for 4mC Site Prediction

Methods	Category	Year	Model	Source code/ Website
DMKL-HFIS	ML	2023	KHFIS + DMKL	
MvLapKSRC-HSIC		2023	Kernel Sparse Representation + Graph Regularization	https://github.com/guofei-tju/MvLapKSRC_HSIC
4mCBERT		2023	sequence-based and chemical-based encoding + CatBoost.	https://github.com/abcair/4mCBERT
MMC-KHFIS		2023	KHFIS + MMC	
DNA-MP		2023	POCD-ND + RF	https://sds_genetic_analysis.opendfki.de/DNA_Modifications/
SSR-RVFL		2024	RVFL + SSR	https://github.com/Hao010418/SSR-RVFL
i4mC-Deep	DL	2021	CNN	https://github.com/waleed551/i4mC-Deep
DCNN-4mC		2021	DCNN	
4mCPred-MTL		2021	Transformer	
iDNA-MT		2021	BiGRU	
iDNA-ABT		2021	BERT	https://github.com/YUYING07/iDNA_ABT
Mouse4mC-BGRU		2022	BiGRU	
MSNet-4mC		2022	CNN	https://github.com/LIU-CT/MSNet-4mC
Hyb4mC		2022	CNN + Attention,CapsNet	https://github.com/YingLiangjxau/Hyb4mC
Deep-4mCGP		2022	CNN	https://github.com/hasanzulfiqar/Deep-4mCGP

iDNA-ABF		2022	Pretrained BERT	https://github.com/FakeEnd/iDNA_ABF

Table 1 (continued)

DRSN4mCPred		2023	MS-CAM+Bi-LSTM	
Methods	**Category**	**Year**	**Model**	**Source code/ Website**
MultiScale-CNN-4mCPred		2023	BiLSTM+CNN	https://github.com/paomian97/MultiScale_CNN_4mCPred
i4mC-GRU		2023	BiGRU	https://github.com/mldlproject/2022-i4mC-GRU
GS-MLDS		2023	CNN	https://github.com/rajib1346/GS-MLDS
StableDNAm		2023	BERT+2D-SENET	https://github.com/wrab12/StableDNAm
Mus4mCPred		2024	BiLSTM, CNN	https://github.com/meloaedy/Mus4mCPred
iDNA-ITLM		2024	Pretrained BERT	https://github.com/Yyxx-1987/iDNA-ITLM/tree/master/iDNA-ITLM/data
DeepSF-4mC		2024	DNN+CNN+LSTM	https://github.com/754131799/DeepSF-4mC
iDNA-OpenPrompt		2024	Pretrained BERT	https://github.com/Yyxx-1987/iDNA-OpenPrompt/
DeepPGD		2024	TCN+BiLSTM+Attention	https://github.com/FROZEN160/DeepPGD
fasttext+CNN based model		2024	CNN	https://github.com/khanhlee/4mC

Note: In the *Source code/Website* column, only the DNA-MP model provides a website, while all others provide source code.

adjustments, while also offering a certain degree of interpretability. However, manual feature engineering struggles to capture complex sequence relationships, and the inclusion of excessive features can lead to high dimensionality, necessitating effective feature selection strategies.

In contrast, deep learning approaches have gradually become dominant in recent years for 4mC methylation prediction by automatically learning complex and hierarchical features from raw sequences. Local feature extraction models, such as CNNs, are good at capturing short-range dependencies. They use relatively simple architectures and train quickly. However, they struggle to model long-range dependencies and global contextual information. Contextual sequence models like LSTM and BiGRU capture sequential context more effectively. But these models often have larger parameter sets, require longer training times, and can overfit on small datasets.

Pretrained large language models (LLMs) benefit from training on massive datasets. They show strong transferability and generalization for 4mC prediction tasks. At the same time, they need high computational resources during training and inference.

Multi-architecture fusion models are designed to integrate diverse hierarchical features from DNA sequences, thereby yielding more comprehensive representations. Nonetheless, devising effective fusion strategies remains a significant challenge. Simple concatenation or weighted averaging can miss complex feature interactions. This may cause redundancy, weaken critical signals, or introduce noise, reducing overall performance.

Therefore, careful selection and design of models and fusion strategies are essential to improve prediction accuracy in 4mC methylation site identification.

Model Validation

Model validation is essential for objectively evaluating a model's predictive performance. Common approaches include independent testing and k-fold cross-validation. In independent testing, the model is trained on the training set and then applied to the test set. A series of evaluation metrics are calculated to assess its effectiveness. In k-fold cross-validation, the dataset is randomly divided into k subsets. The model trains on k-1 folds and tests on the remaining fold. This process repeats k times, ensuring that each subset is used for testing once [80, 81]. Compared to independent testing, k-fold cross-validation uses the data more efficiently and is suitable for smaller datasets. However, it may not fully reflect the model's ability to generalize to completely unseen data.

In 4mC methylation prediction, performance and generalization are evaluated using several validation metrics [82]. Since this is a binary classification problem, evaluation is usually based on the confusion matrix. The confusion matrix includes four elements: true positives (TP), which are correctly predicted positive samples; false positives (FP), which are negative samples incorrectly predicted as positive; true negatives (TN), which are correctly predicted negative samples; and false negatives (FN), which are positive samples incorrectly predicted as negative [83]. Based on these values, common performance metrics include accuracy, sensitivity, specificity, and the Matthews correlation coefficient (MCC). The corresponding formulas are as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{11}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{12}$$

$$Specificity = \frac{TN}{FP+TN} \tag{13}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (14)$$

Accuracy shows how often the model makes correct predictions. Sensitivity indicates the model's capability to detect positive samples accurately. Specificity assesses how well the model recognizes negative samples. MCC is considered a reliable metric for evaluating the overall performance of binary classification models. Its value ranges from -1 to 1. A value close to 1 indicates excellent predictive performance. A value near 0 suggests performance similar to random guessing. A value of -1 implies completely incorrect predictions. In addition, the area under the receiver operating characteristic curve (AUC) is also commonly used in 4mC prediction tasks. The ROC curve is constructed by plotting the true positive rate against the false positive rate across different thresholds, thereby reflecting the model's ability to discriminate between positive and negative samples. An AUC value closer to 1 indicates stronger discriminative performance, whereas a value approaching 0.5 suggests weak predictive ability, equivalent to random guessing.

Database and Dataset

MethSMRT [6] is a specialized database for storing and analyzing DNA 6mA and 4mC methylation data generated by SMRT [14] sequencing technology. Commonly used 4mC methylation datasets in MethSMRT include species such as *Geobacter pickeringii*, *Geoalkalibacter subterraneus*, *Escherichia coli*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Tolypocladium* sp. SUP5-1, and *Mus musculus*. MDR [84] is the first dedicated database for displaying and storing SMRT sequencing-derived 6mA and 4mC methylation data in Rosaceae plants. The most commonly used species in MDR include *Casuarina equisetifolia*, *Fragaria vesca*, *Rosa chinensis*, and *Saccharomyces cerevisiae*.The Lin_2017 dataset [19] comprises positive samples exclusively collected from the MethSMRT [6] database, encompassing six species: *C. elegans*, *D. melanogaster*, *A. thaliana*, *E. coli*, *G. subterraneus*, and *G. pickeringii*. All sequences are standardized to 41 bp in length. Negative samples were selected as 41-bp sequences centered on cytosines not detected as methylated by SMRT sequencing technology. High- confidence positive sequences with modification quality values (modQV) ≥ 30 were retained [85], and redundancy was removed using CD-HIT [86] with an 80% sequence similarity threshold. Due to the disproportionately larger number of negative samples, an equal number of negative sequences were randomly sampled from each species to balance the dataset. The Li_2020 dataset [27] retained the same species as Lin_2017 but expanded the dataset size. For positive sample selection, sequences with modQV ≥ 30 were considered candidates, while those with modQV ≥ 50 were reserved as an independent test set. Moreover, the redundancy removal threshold was tightened to 70%. The Hyb_2021 dataset [52] further updated the data, incorporating additional samples and improvements. Lv et al. [29] constructed their dataset by integrating methylation data—including 4mC, 5hmC, and 6mA sites—collected from multiple sources such as the MethSMRT database [6], MDR database [84], GEO database [87], NCBI Genome database [88] and relevant published literature. For jump-seq data, positive samples were selected with a 5hmC proportion exceeding 95%. For SMRT data, sequences with modification quality values (modQV) ≥ 30 were retained as positive samples; however, if the sample size was insufficient, the modQV filtering criterion was relaxed. All sequences were standardized to 41 bp in length. Redundancy removal was performed using the CD-HIT tool with an 80% sequence similarity threshold. Negative samples were constructed as 41-bp sequences centered on cytosine or adenine residues that were experimentally unverified as methylated. Finally, the benchmark dataset was randomly split into two equal parts, one used for model training and the other for performance evaluation.

Table 2 summarizes datasets utilized for 4mC methylation prediction over the past five years.

Table 2 Summary of datasets used in the past five years

Data	Training sets	Test sets	Database	Model	Link
Lin_2017	A. thaliana:P-1,978/N-1,978 C. elegans:P-1,554/N-1,554 D. melanogaster:P-1,769/N-1,769 E. coli:P-388/N-388 G. pickeringii:P-569/N-569 G. subterraneus:P-906/N-906		MethSMRT	DMKL-HFIS、 MvLapKSRC-HSIC MMC-KHFIS、SSR-RVFL、 i4mC-Deep、GS-MLDS、 MSNet-4mC	Unavailable
Xu et al.	A. thaliana:P-111,927 / N-111,927 C. elegans:P-60,662 / N-60,662 D. melanogaster:P-90,333 / N-90,333 E. coli:P-2,067 / N-10,335 G. Pickering:P-5,727 / N-28,635 G. subterraneous:P-15,135 / N-75,675	A. thaliana:P-1,978 / N-1,978 C. elegans:P-1,554 / N-1,554 D. melanogaster:P-1,769 / N-1,769 E. coli:P-388 / N-388 G. Pickering:P-569 / N-569 G. subterraneous:P-905 / N-905	MethSMRT	4mCBERT	https://bioinfo.uth.edu/Deep4mC/Download.php

Lv et al.	C. equisetifolia:P-183 / N-183 F. vesca:P-7,899 / N-7,899 S. cerevisiae:P-990 / N-990 Tolypocladium:P-7,664 / N-7,664	C. equisetifolia:P-183 / N-183 F. vesca:P-7,899 / N-7,899 S. cerevisiae:P-990 / N-990 Tolypocladium:P-7,664 / N-7,664	MethSMRT and MDR	DNA-MP、iDNA-MT、iDNA-ABT、iDNA-ABF、iDNA-ITLM、iDNA-OpenPrompt StableDNAm、DeepPGD Methyl-GP	http://lin-group.cn/server/iDNA-MS/
Zeng et al.	A. thaliana:P-16,000 / N-16,000 C. elegans:P-16,000 / N-16,000 D. melanogaster:P-16,000 / N-16,000	A. thaliana:P-4,000 / N-4,000 C. elegans:P-4,000 / N-4,000 D. melanogaster:P-4,000 / N-4,000		4mCPred-MTL DeepSF-4mC	

Data	Training sets	Test sets	Database	Model	Link
Li_2020	A. thaliana:P-63,720/N-63,720 C. elegans:P-55,729/N-55,729 D. melanogaster:P-53,970 /N-53,970 E. coli:P-1,941/N-1,941 G. pickeringii:P-4,514 /N-4,514 G. subterraneus:P-9,934/N-9,934	A. thaliana:P-11,307/N-11,307 C. elegans:P-2,667/N-2,667 D. melanogaster:P-3684 /N-3684 E. coli:P-126/N-126 G. pickeringii:P-1,210 /N-1,210 G. subterraneus:P-5,263/N-5,263	MethSMRT	Hyb4mC MSNet-4mC	Unavailable

Rehman et al.	A. thaliana:P-81,143 / N-85,456 C. elegans:P-7939 / N-82,033 D. melanogaster:P-72,127 / N-75,460 E. coli:P-1,959 / N-2,156 G. pickeringii:P-4,703 / N-4,900 G. subterraneus:P-10,583 / N-10,780 Mus musculus:P-800 / N-800 C. equisetifolia:P-183 / N-183 S. cerevisiae:P-990 / N-990 Tolypocladium sp.:P-7,664 / N-7,664 F. vesca:P-12,298 / N-12,152 R. chinensis:P-2,337 / N-2,337	A. thaliana:P-10,388 / N-11,172 C. elegans:P-2,352 / N-2,660 D. melanogaster:P-3,332 / N-3,521 E. coli:P-126 / N-126 G. pickeringii:P-1,210 / N-1,210 G. subterraneus:P-5,263 / N-5,263 Mus musculus:P-180 / N-180 C. equisetifolia:P-183 / N-183 S. cerevisiae:P-989 / N-989 Tolypocladium sp.:P-7,663 / N-7,663 F. vesca:P-8,819 / N-9,015 R. chinensis:P-779 / N-779	MethSMRT and MDR	DCNN-4mC	https://nsclbio.jbnu.ac.kr/tools/DCNN-4mC/
Nguyen et al.	F. vesca:P-3,457 / N-3,457 R. chinensis:P-1,938 / N-1,938			fasttext+CNN based model	

Data	Training sets	Test sets	Database	Model	Link
Zulfiqar et al.	G. pickeringii:P-569 / N-569	G. pickeringii:P-200 / N-200	MethSMRT	Deep-4mCGP	https://github.com/BionicsAI/Deep-4mCGP
Nguyen-Vo et al.	Mus musculus:P-5,304/N-5,816(training) Mus musculus:P-1,136/N-1,247(validation)	Mus musculus:P-1,136/N-1,209	MethSMRT	i4mC-GRU	https://github.com/mldlproject/2022-i4mC-GRU
Hyb_2021	A. thaliana:P-74,662 / N-74,662 C. elegans:P-56,770 / N-56,770 D. melanogaster:P-81,289 / N-81,289	A. thaliana:P-28,000 / N-28,000 C. elegans:P-12,147 / N-12,147 D. melanogaster:P-50,966 /	MethSMRT	Hyb4mC DRSN4mCPred	https://github.com/YingLiangjxau/Hyb4mC

	E. coli:P-1,908 / N-1,908	N-50,966
	G. pickeringii:P-3,761 / N-3,761	E. coli:P-160 / N-160
	G. subterraneous:P-7,064 / N-7,064	G. pickeringii:P-1,926 / N-1,926
		G. subterraneous:P-7,813 / N-7,813

Benchmark Evaluation of Representative 4mC Prediction Models

To enable a more detailed evaluation of the models described in this study, we selected several of the most recent models based on publicly available reproducibility and ultimately chose six representative models for a series of experiments to comprehensively assess their performance. The models include MultiScale-CNN-4mCPred [77]、4mCBERT [41]、DeepPGD [79]、fastText+CNN based model [51]、i4mC-GRU [60] and Hyb4mC [52]. The experiments were conducted using the Hyb_2021 dataset [52], which, as summarized in Table 2, comprises six species: *A. thaliana*, *C. elegans*, *D. melanogaster*, *E. coli*, *G. pickeringii*, and *G. subterraneus*. The number of positive samples in the training set for each species is 74,662, 56,770, 81,289, 1,908, 3,761, and 7,064, respectively. For the test set, the corresponding numbers of positive samples are 28,000, 12,147, 50,966, 160, 1,926, and 7,813. Each test set includes the same number of negative samples, keeping the binary classification task balanced.

We evaluated and analyzed the models from multiple perspectives. Overall, the above experiments indicate that DeepPGD [79] demonstrates strong overall performance, achieving satisfactory predictive results across species with varying dataset sizes. However, its cross-species generalization is limited, making it more suitable for prediction tasks within a single species. For cross-species prediction, model performance generally declines, and considering this, MultiScale-CNN-4mCPred [77] is a more appropriate choice for such experiments. Hyb4mC [52], with architectures specifically designed for datasets of different sizes, can be considered when experimental datasets vary greatly across multiple species. Nevertheless, it is worth noting that, according to the experiments, Hyb4mC [52] does not perform optimally when the training data are of moderate size. In contrast, the fastText+CNN-based model [51] exhibits relative insensitivity to dataset size, suggesting that it can achieve favorable results under conditions of limited data availability.

Performance Comparison Across Species under Five-Fold Cross-Validation

In this section, we systematically evaluate the performance of six models across six species-specific datasets using five-fold cross-validation. To this end, all samples in the training sets were used for cross-validation. For each fold, 80% of the data were chosen for training, and the remaining 20% were reserved for validation. The overall performance was calculated by taking the average of the results from the five folds. To highlight the influence of dataset size on model performance, we compared results on two species with the most pronounced difference in sample size, as shown in Figure 3. The performance metrics of each model on the remaining species are provided in Supplementary Figures S1 and S2. The DeepPGD model [79], which integrates TCN and BiLSTM architectures, effectively captures multi-scale local features as well as contextual information, achieving consistently strong performance across both species. However, for most models, a performance drop was observed when applied to *E. coli*, the species with the smallest dataset, compared to *D. melanogaster*, which has the largest. Notably, the 4mCBERT model [41] exhibited the most pronounced performance decline on small-scale datasets, likely because the

use of chemical BERT for feature encoding causes the model to capture excessive sequence details, including noise, leading to overfitting. In contrast, Hyb4mC [52], which was specifically designed with two architectural variants to adapt to datasets of different sizes, showed improved performance on the smaller *E. coli* dataset.

Performance Comparison Across Species on the Independent Test Sets

Next, we evaluated the performance of the models trained using five-fold cross-validation on the independent

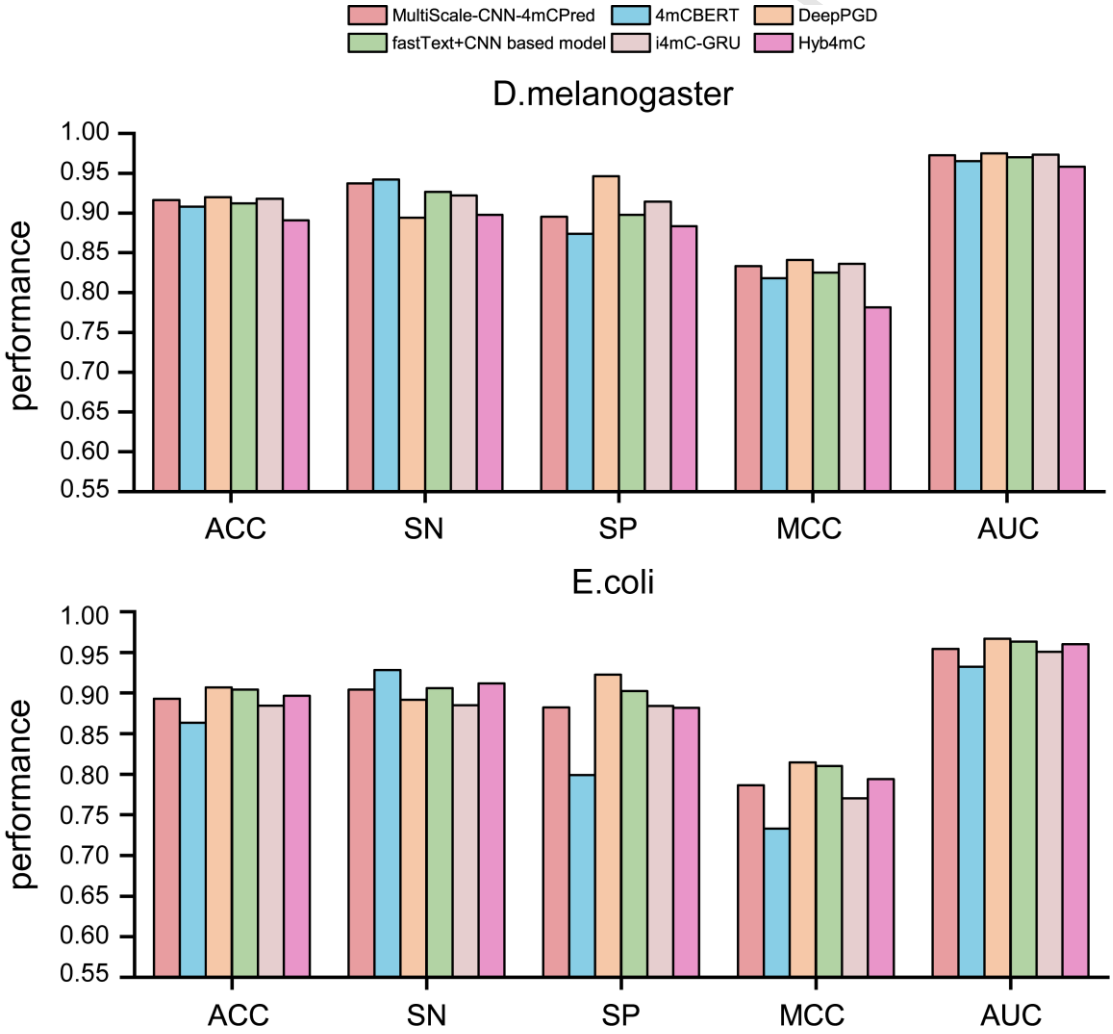

Figure 3. Overall performance of the model on *D. melanogaster* and *E. coli* under five-fold cross-validation. The performance of six models was compared using evaluation metrics such as ACC, SN, SP, MCC, and AUC.

test sets. First, based on the predictions from the fastText+CNN based model [51], we explored the relationship

between evolutionary proximity and predictive accuracy across the six species (Figure 4A). Notably, *C. elegans* and *D. melanogaster*, which are evolutionarily close, achieved relatively high accuracy values of

0.9321 and 0.9251, respectively. Similarly, for *G. pickeringii* and *G. subterraneus*, which also share close evolutionary proximity, the model achieved comparable accuracy values of 0.8284 and 0.8348. Among them, 4mCBERT [41] exhibited noticeably weaker performance on *G. subterraneus* compared to the other models. In this study, we placed particular emphasis on sensitivity, as it reflects the model's ability to correctly identify positive samples. A high sensitivity ensures the detection of as many potential 4mC sites as possible, which is critical for comprehensive identification in 4mC prediction. To further investigate sensitivity differences, we analyzed the proportion of true positives (TP) and false negatives (FN) relative to the total number of positive samples, as shown in Figure 4C. Interestingly, DeepPGD [79] , with its TCN+BiLSTM architecture, is capable of capturing more detailed and well-defined sequence dependencies within species-specific datasets, resulting in a stricter and more conservative discrimination between positive and negative samples, which led to reduced sensitivity on datasets with larger sample sizes. In contrast, the fastText+CNN-based model, owing to the limited number of parameters in fastText, was unable to adequately learn the subtle differences among individual DNA fragments in smaller datasets, leading to insufficient feature representation and consequently underperforming in terms of sensitivity on species with limited data.

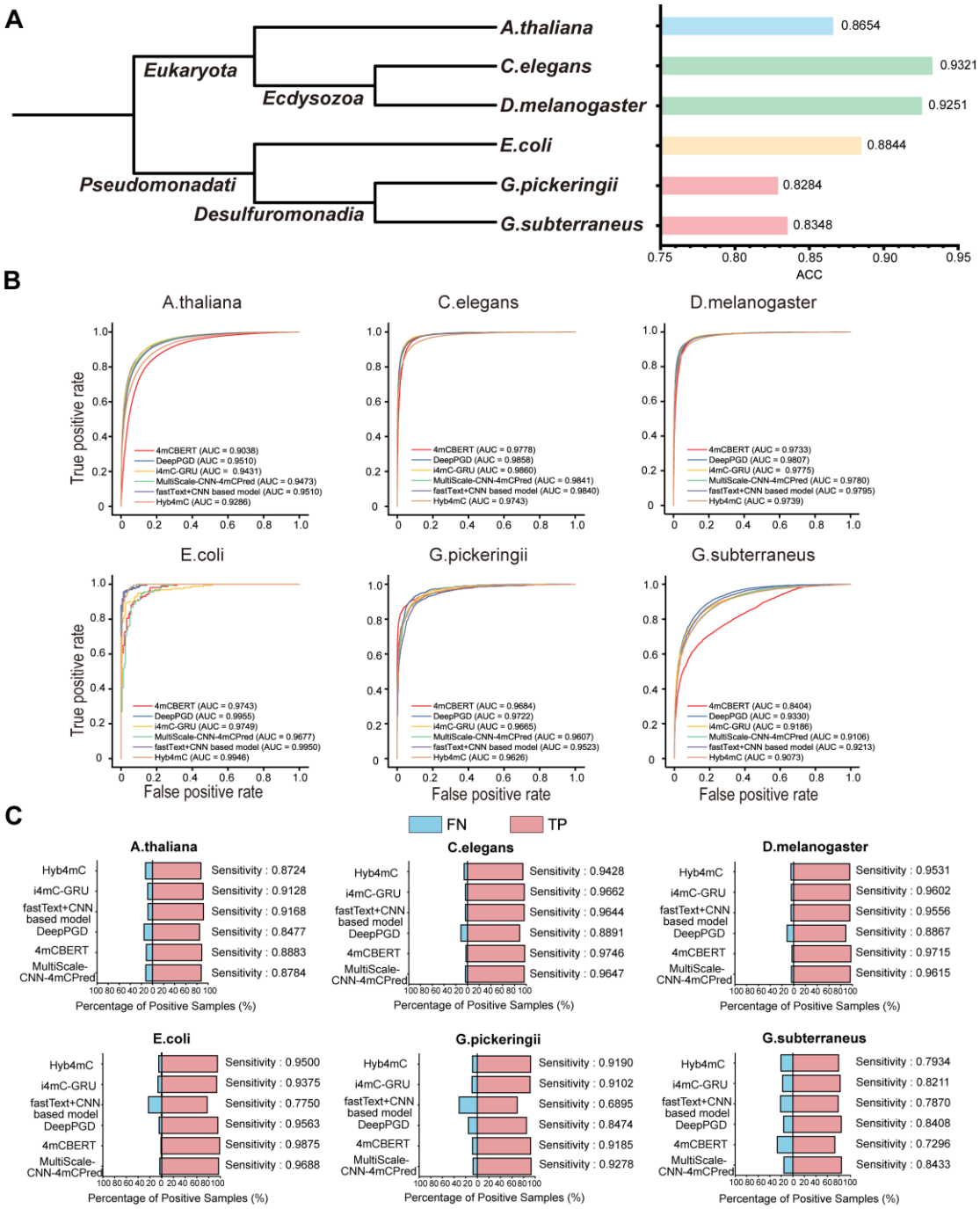

Figure 4. Performance comparison of the predictive model on the test sets across multiple species. **A** Phylogenetic tree of the six species along with their prediction accuracies. Species with similar evolutionary relationships are marked with the same color. **B** ROC curves of the predictive model on six different species. **C** Comparison of true positive (TP) and false negative (FN) counts across species, reflecting the sensitivity of the model. Comprehensive performance metrics of each predictive model on the independent test set are provided in Supplementary Tables S1–S6.

Cross-Species Performance Evaluation

To investigate whether a model trained on one species can effectively generalize to others, we designed a

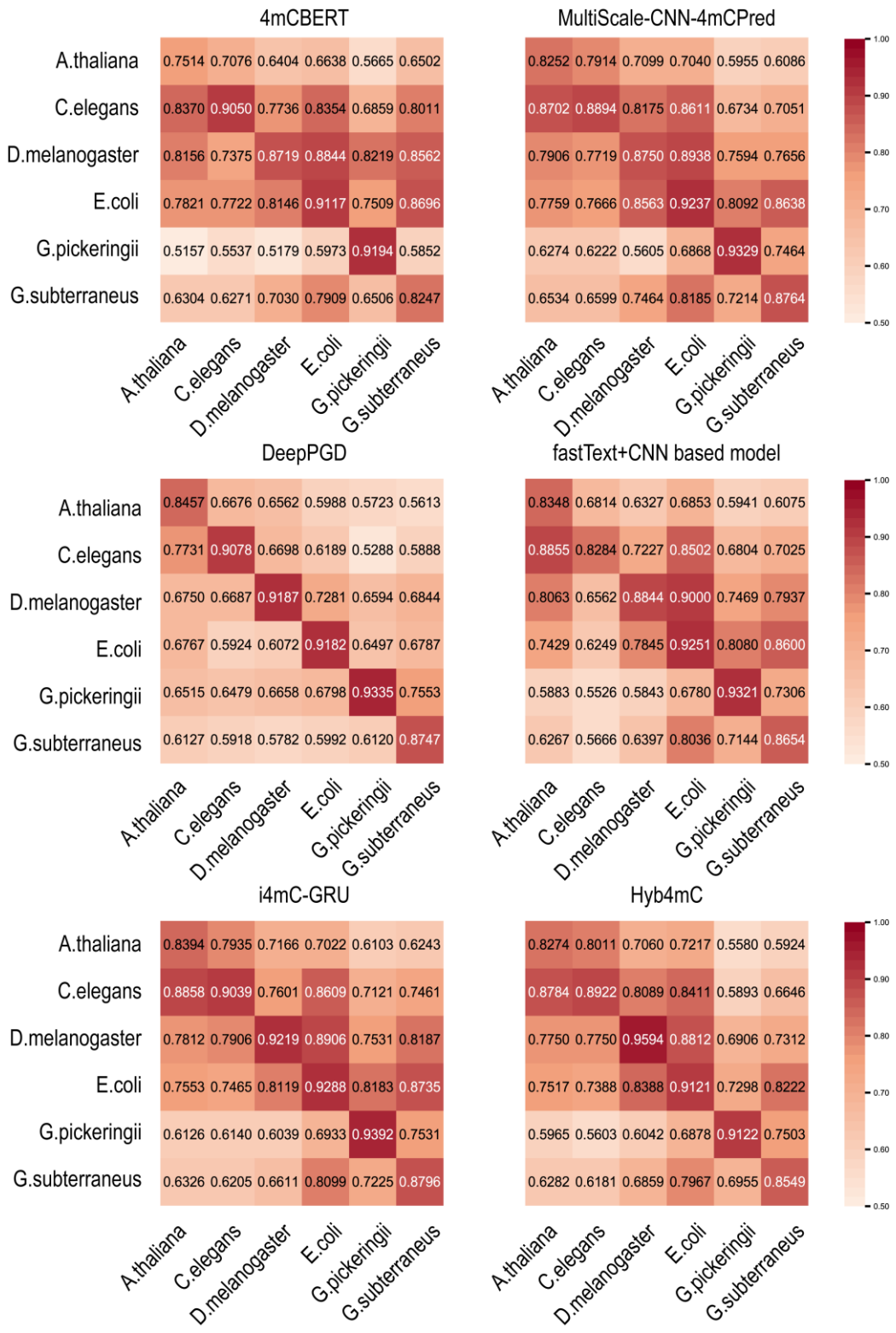

Figure 5. Cross-species performance evaluation of the predictive model. The vertical axis represents the species used for training, while the horizontal axis denotes the species used for testing. In the heatmap, darker colors indicate higher prediction accuracy.

cross-species prediction experiment to evaluate the transferability of each method. Specifically, for each model, we trained six species-specific models using the respective training sets and subsequently evaluated

them on the test sets of all six species, including the source species. The ACC results of all six prediction methods are visualized in Figure 5 as a heatmap, where the vertical axis represents the training species and the horizontal axis denotes the test species. The heatmap reveals several notable patterns. Overall, all models exhibited a clear performance drop when applied to species other than the one they were trained on, underscoring the challenge of cross-species prediction. While DeepPGD [79] achieved strong performance when the training and test species were the same, the model tends to extract comprehensive information from the sequences and,

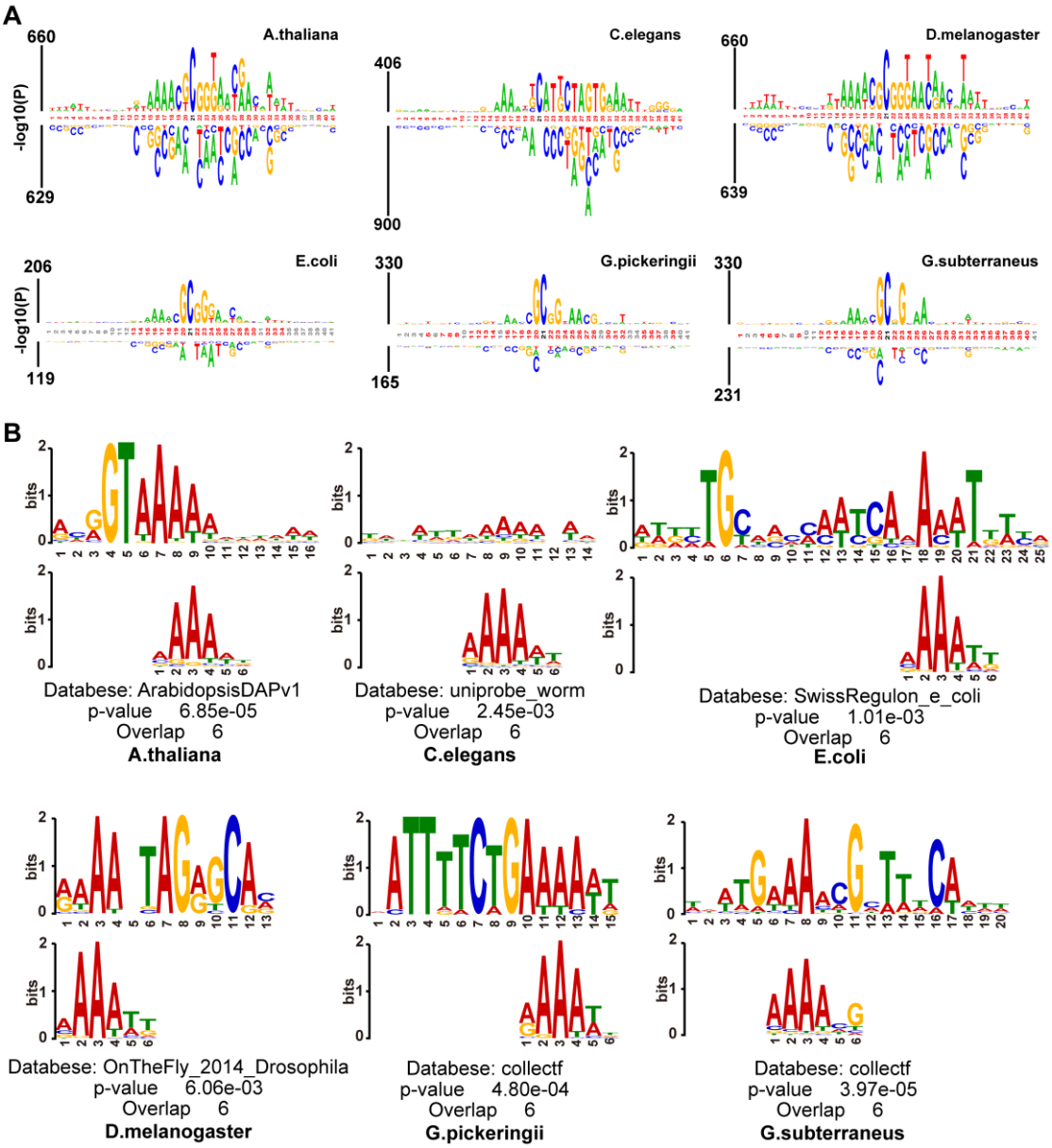

Figure 6. Cross-species model performance from the perspective of biological sequences. A Probability-based kpLogo visualizations of motifs across six species, where taller logos indicate higher residue occurrence probabilities at specific positions. B Motifs extracted from positions 15–20 in the six species using STREME and compared against existing motif databases through Tomtom. The database denotes the reference database used for comparison. The *P*-value is calculated by Tomtom.

during training within a single species, predominantly fits the distribution of the training data rather than learning features that are generalizable across species. As a result, its transferability to other species was

relatively limited. In contrast, other models demonstrated varying degrees of cross-species generalization. Additionally, models trained on species with larger datasets generally performed better in cross-species settings, suggesting that training data volume is critical for enhancing model generalization. Moreover, all models demonstrated a notable drop in sensitivity on *G. subterraneus*, indicating potential challenges in accurately detecting methylation sites in this species.

By examining the heatmap in Figure 5, we observed that when *G. pickeringii* was used as the training set, the overall cross-species generalization ability of the model was relatively weak. To further elucidate the relationship between the model's cross-species performance and underlying biological contexts, we employed kpLogo [89] to generate probability-based motif logos for the six species, as shown in Figure 6A. Interestingly, we found that all six species shared a similar motif at positions 15–18, characterized by an AAAA enrichment, while *G. pickeringii* exhibited the weakest sequence conservation at this region. This prompted us to explore its biological implications more deeply. Previous studies have demonstrated that poly(dA:dT) sequences strongly disfavor nucleosome formation, leading to nucleosome-depleted regions in and around these sequences. Such regions tend to expose nearby transcription factor binding sites, thereby facilitating their accessibility and enhancing transcriptional activity [90, 91]. To further validate the biological relevance of this observation, we extracted the subsequences spanning positions 15–20 from all six species and applied STREME [92] to identify motifs. The resulting motifs all showed statistically significant enrichment, with *P*- values less than 1.0×10^{-3}. Notably, among the six species, *G. pickeringii* exhibited the lowest percentage of positive samples matching the identified motif (6.5%), whereas the other five species each exceeded 20%. Furthermore, we compared the extracted motifs against transcription factor motif databases using Tomtom, as illustrated in Figure 6B. The analysis revealed a degree of similarity to known transcription factor motifs, supporting the notion that A-rich sequences may be associated with transcription factor binding and thus hold biological significance. However, given the relatively low sequence conservation in *G. pickeringii*, the model may fail to effectively capture this feature, which could explain the limited cross-species generalization when trained on this species.

Accuracy Evaluation under Varying Training Set Sizes

To examine how the size of the training dataset influences the performance of each prediction method, we conducted a controlled downsampling experiment. For each species, the training set was progressively reduced by randomly removing 4mC sequences in 10% increments from 0% to 90%, while the test set remained unchanged. Each model was trained and evaluated across all six species using ten different levels of downsampling, resulting in a total of 60 experiments. The outcomes are summarized in Figure 7. As expected, model performance declined as training data decreased. Notably, The fastText+CNN based model [51], with its small parameter size and simple feature extraction scheme, was less prone to overfitting and therefore maintained stable performance, showing only limited degradation under reduced data conditions. In contrast, the Hyb4mC [52] model showed higher sensitivity to data reduction, particularly in large datasets, likely due to its architecture being optimized for large-scale datasets in these species.

Challenges and Opportunities of ML and DL in 4mC Methylation Site Prediction

Improving Cross-Species Prediction Performance

Most current 4mC prediction models are trained and evaluated within a single species, which often results in

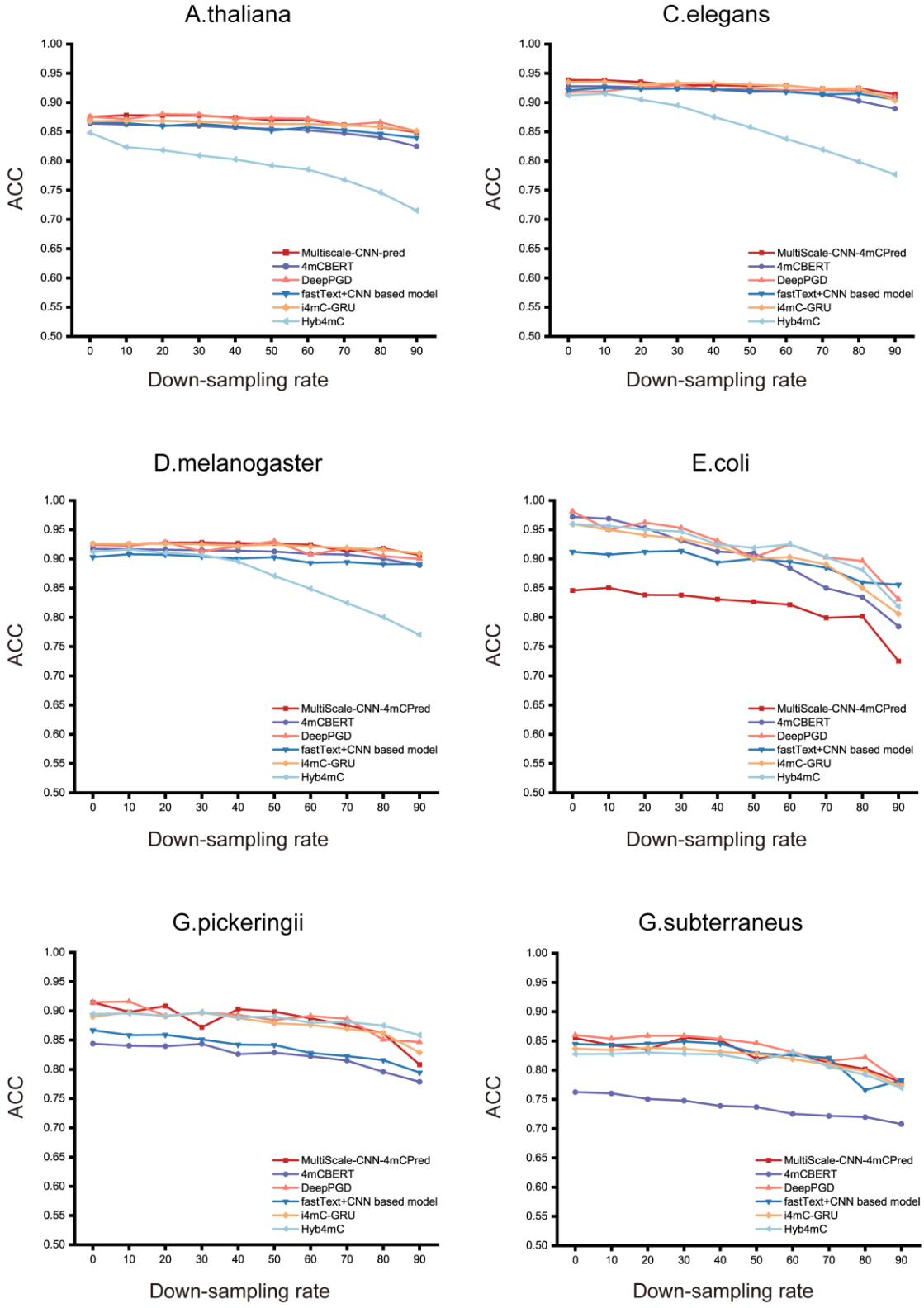

Figure 7. Accuracy (ACC) comparison of different prediction methods with varying training data sizes. The horizontal axis represents the downsampling rate of training samples, ranging from 0% to 90% in decrements of 10%, while the vertical axis

shows the prediction accuracy of the model on the test set.

a significant drop in performance when applied to other species. This species-specific bias limits the generalizability and practical applicability of such models. To address this issue, integrating datasets from multiple species during training can enhance the diversity and representation of methylation patterns, thereby improving model robustness. In addition, the incorporation of adversarial learning strategies has shown potential for improving both model robustness and cross-species generalization by reducing domain discrepancies between species.

Challenges in Modeling Long-Range Dependencies

Most existing 4mC datasets are constructed using fixed-length sequences, typically centered on the methylation site with a window size of 41 bp. As a result, current models predominantly focus on the local context surrounding the modification site, while largely neglecting potential long-range interactions. Long-range regulatory interactions, such as enhancer–promoter contacts, play crucial roles in gene regulation and may also influence DNA methylation patterns. Thus, constructing datasets with longer genomic sequences and developing models that can capture these distal dependencies are essential for advancing 4mC prediction. Incorporating long-range genomic context may provide a more comprehensive understanding of the epigenetic mechanisms underlying 4mC methylation.

Opportunities for Multimodal Feature Integration

Current 4mC methylation prediction models predominantly rely on nucleotide sequence information alone, without incorporating additional data modalities. However, integrating complementary modalities—such as DNA secondary structure features and electrical signal data from SMRT sequencing—has the potential to enrich the feature space and improve prediction performance [93]. Multimodal fusion can enable models to capture diverse biological cues that are not evident from sequence alone, thereby enhancing both accuracy and interpretability. The incorporation of such heterogeneous data sources represents a promising direction for building more robust and biologically informed methylation prediction frameworks.

Development of Online Web Servers

Currently, most 4mC methylation prediction models rely on local deployment, requiring users to download source code, configure computational environments, and load pretrained models for inference. This process presents a significant technical barrier for many biological researchers, limiting the widespread adoption and practical application of these tools. In contrast, the development of online web servers can substantially enhance model accessibility and user-friendliness, allowing researchers to make rapid predictions through intuitive web interfaces without the need for a complex setup. Such platforms also facilitate broader dissemination and collaborative usage, accelerating epigenetic research.

Incorporating DNA Structural Features

Current studies on 4mC prediction have predominantly relied on sequence-derived features, while DNA

structural characteristics are also closely associated with the occurrence of modifications. Since the formation of 4mC requires recognition by specific enzymes, information on protein–DNA interactions represents an important feature. Variations in the spatial conformation of DNA-binding proteins can alter their binding affinity to DNA, thereby influencing the degree of methylation. Capturing the three-dimensional structural patterns of DNA-binding proteins is thus beneficial for the accurate identification of 4mC sites. Moreover, the local structural characteristics of binding sites also play a significant role in determining the methylation propensity of 4mC. In lysine acetylation site prediction, for example, the SIPSC-Kac [94] model effectively improved computational accuracy and predictive efficiency by integrating protein three-dimensional structural features extracted from AlphaFold, providing important insights for our field. Inspired by this, an important future direction lies in the effective extraction of DNA structural features and the incorporation of protein–DNA interaction information into predictive frameworks.

Conclusion

In this study, we emphasized the growing significance of computational methods in the prediction of N4-methylcytosine methylation sites and provided a comprehensive review of models developed over the past five years. We systematically categorized existing prediction tools into ML-based and DL-based models, summarizing their methodological characteristics, strengths, and limitations. Furthermore, we compiled commonly used databases and benchmark datasets to support further research and tool development in this field.

To facilitate practical application and model selection, we conducted a benchmarking analysis of six representative models, comparing their predictive performance, feature encoding strategies, and computational efficiency. We also discussed the current challenges and future opportunities in 4mC site prediction, including limited cross-species generalization, insufficient modeling of long-range dependencies, challenges in integrating multimodal biological data, and the lack of accessible and user-friendly web tools.

Addressing these challenges will require continuous methodological innovation, better interpretability, and interdisciplinary collaboration. With ongoing advancements, machine learning and deep learning approaches are expected to further enhance the accuracy, robustness, and applicability of 4mC methylation prediction, paving the way for breakthroughs in epigenetics, disease mechanism discovery, and precision medicine.

Data availability

We have compiled the methods with publicly available code mentioned in this study, along with the benchmark datasets used in this work, on GitHub at https://github.com/narissu/4mC-prediction-methods.

Funding

The work is supported by the National Natural Science Foundation of China (No. 62450002), Hainan Provincial Natural Science Foundation of China (324MS009) and Science and Technology special fund of

Hainan Province (ZDYF2024GXJS018).

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Robertson KD. DNA methylation and human disease. Nat Rev Genet. 2005;6:597-610.

[2] Wei R, Zhang L, Zheng H, Xiao M. A Systematic Review of the Application of Machine Learning in CpG Island (CGI) Detection and Methylation Prediction. Current Bioinformatics. 2024;19:235-49.

[3] Mattei AL, Bailly N, Meissner A. DNA methylation: a historical perspective. Trends Genet. 2022;38:676-707.

[4] Tang W, Wan S, Yang Z, Teschendorff AE, Zou Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics. 2018;34:398-406.

[5] Kulis M, Esteller MJAig. DNA methylation and cancer. 2010;70:27-56.

[6] Ye P, Luan Y, Chen K, Liu Y, Xiao C, Xie Z. MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. Nucleic Acids Res. 2016;45:D85-D9.

[7] Branco MR, Ficz G, Reik W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. Nat Rev Genet. 2011;13:7-13.

[8] Boulias K, Greer EL. Means, mechanisms and consequences of adenine methylation in DNA. Nat Rev Genet. 2022;23:411-28.

[9] Zhou Z, Xiao C, Yin J, She J, Duan H, Liu C, et al. PSAC-6mA: 6mA site identifier using self-attention capsule network based on sequence-positioning. Computers in Biology and Medicine. 2024;171:108129.

[10] Peng X, Cui W, Kong X, Huang Y, Li J. DMR_Kmeans: Identifying Differentially Methylated Regions Based on k-means Clustering and Read Methylation Haplotype Filtering. Current Bioinformatics. 2024;19:490-501.

[11] Yao Z, Li F, Xie W, Chen J, Wu J, Zhan Y, et al. DeepSF-4mC: A deep learning model for predicting DNA cytosine 4mC methylation sites leveraging sequence features. Comput Biol Med. 2024;171:108166.

[12] Zhao Z, Zhang X, Chen F, Fang L, Li J. Accurate prediction of DNA N4-methylcytosine sites via boost-learning various types of sequence features. BMC Genomics. 2020;21.

[13] Barros-Silva D, Marques CJ, Henrique R, Jerónimo C. Profiling DNA Methylation Based on Next-Generation Sequencing Approaches: New Insights and Clinical Applications. Genes. 2018;9.

[14] Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 2018;46:2159-68.

[15] Navlakha S, Bar-Joseph Z. Algorithms in nature: the convergence of systems biology and computational thinking. Mol Syst Biol. 2011;7.

[16] Chelly Dagdia Z, Avdeyev P, Bayzid MS. Biological computation and computational biology: survey, challenges, and discussion. Artificial Intelligence Review. 2021;54:4169-235.

[17] Peng X, Cui W, Zhang W, Li Z, Zhu X, Yuan L, et al. A Metric to Characterize Differentially Methylated Region Sets Detected from Methylation Array Data. Current Bioinformatics. 2024;19:571-83.

[18] Luo X, Wang Y, Zou Q, Xu L. Recall DNA methylation levels at low coverage sites using a CNN model in WGBS. Plos Computational Biology. 2023;19:1011205.

[19] Chen W, Yang H, Feng P, Ding H, Lin H. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics. 2017;33:3518-23.

[20] Wei L, Luan S, Nagai LAE, Su R, Zou Q. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. Bioinformatics. 2018;35:1326-33.

[21] He W, Jia C, Zou Q. 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. Bioinformatics. 2018;35:593-601.

[22] Wei L, Su R, Luan S, Liao Z, Manavalan B, Zou Q, et al. Iterative feature representations improve N4-methylcytosine site prediction. Bioinformatics. 2019;35:4930-7.

[23] Manavalan B, Basith S, Shin TH, Wei L, Lee G. Meta-4mCpred: A Sequence-Based Meta-Predictor for Accurate DNA 4mC Site Prediction Using Effective Feature Representation. Molecular Thera py - Nucleic Acids. 2019;16:733-44.

[24] Manavalan B, Basith S, Shin TH, Lee DY, Wei L, Lee G. 4mCpred-EL: An Ensemble Learning Framework for Identification of DNA N4-Methylcytosine Sites in the Mouse Genome. Cells. 2019;8:1332.

[25] Tang Q, Kang J, Yuan J, Tang H, Li X, Lin H, et al. DNA4mC-LIP: a linear integration method to identify N4-methylcytosine site in multiple species. Bioinformatics. 2020;36:3327-35.

[26] Hasan MM, Manavalan B, Shoombuatong W, Khatun MS, Kurata H. i4mC-Mouse: Improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes. Comput Struct Biotechnol J. 2020;18:906-12.

[27] Liu Q, Chen J, Wang Y, Li S, Jia C, Song J, et al. DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. Briefings Bioinf. 2020;22.

[28] Zeng R, Liao M. Developing a Multi-Layer Deep Learning Based Predictive Model to Identify DNA N4-Methylcytosine Modifications. Front Bioeng Biotechnol. 2020;8.

[29] Lv H, Dao F-Y, Zhang D, Guan Z-X, Yang H, Su W, et al. iDNA-MS: An Integrated Computational Tool for Detecting DNA Modification Sites in Multiple Genomes. iScience. 2020;23:100991.

[30] Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Briefings Bioinf. 2023;24.

[31] Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. Journal of Big Data. 2020;7:94.

[32] Cui F, Zhang Z, Zou Q. Sequence representation approaches for sequence-based protein prediction tasks that use deep learning. Briefings in Functional Genomics. 2021;20:61-73.

[33] Zhou J, Chen Q, Braun PR, Perzel Mandell KA, Jaffe AE, Tan HY, et al. Deep learning predicts DNA methylation regulatory variants in the human brain and elucidates the genetics of psychiatric disorders. Proceedings of the National Academy of Sciences. 2022;119.

[34] Chen H-X, Liu Z-D, Bai X, Wu B, Song R, Yao H-C, et al. Accurate cross-species 5mC detection for Oxford Nanopore sequencing in plants with DeepPlant. Nat Commun. 2025;16.

[35] Liu J, Shen H, Chen K, Li X. Large language model produces high accurate diagnosis of cancer from end-motif profiles of cell-free DNA. Briefings Bioinf. 2024;25.

[36] Ai C, Tiwari P, Yang H, Ding Y, Tang J, Guo F. Identification of DNA N4-methylcytosine Sites via Multiview Kernel Sparse Representation Model. IEEE Trans Artif Intell. 2023;4:1236-45.

[37] Ding Y, Tiwari P, Guo F, Zou Q. Multi-correntropy fusion based fuzzy system for predicting DNA N4-methylcytosine sites. Information Fusion. 2023;100:101911.

[38] Ding Y, Tiwari P, Zou Q, Guo F, Pandey HM. C-Loss Based Higher Order Fuzzy Inference Systems for Identifying DNA N4-Methylcytosine Sites. IEEE Trans Fuzzy Syst. 2022;30:4754-65.

[39] Xie H, Ding Y, Qian Y, Tiwari P, Guo F. Structured Sparse Regularization based Random Vector Functional Link Networks for DNA N4-methylcytosine sites prediction. Expert Syst Appl. 2024;235:121157.

[40] Nabeel Asim M, Ali Ibrahim M, Fazeel A, Dengel A, Ahmed S. DNA-MP: a generalized DNA modifications predictor for multiple species based on powerful sequence encoding method. Briefings Bioinf. 2022;24.

[41] Yang S, Yang Z, Yang J. 4mCBERT: A computing tool for the identification of DNA N4-methylcytosine sites by sequence- and chemical-derived information based on ensemble learning strategies. Int J Biol Macromol. 2023;231:123180.

[42] Wang L, Ding Y, Tiwari P, Xu J, Lu W, Muhammad K, et al. A deep multiple kernel learning-based higher-order fuzzy inference system for identifying DNA N4-methylcytosine sites. Information Sciences. 2023;630:40-52.

[43] Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017;18:67.

[44] Pawar S, Kanika, S S, Gopalakrishnan S, Alsalami Z, T MP. Integrating Convolutional Neural Networks for Genomic Sequence Analysis: Deep Learning Applications in Genomics and Bioinformatics. 2024 International Conference on Advances in Computing, Communication and

Materials (ICACCM): IEEE; 2024. p. 1-5.

[45] Alam W, Tayara H, Chong KT. i4mC-Deep: An Intelligent Predictor of N4-Methylcytosine Sites Using a Deep Learning Approach with Chemical Properties. Genes: MDPI AG; 2021. p. 1117.

[46] Rehman MU, Tayara H, Chong KT. DCNN-4mC: Densely connected neural network based N4-methylcytosine site prediction in multiple species. Comput Struct Biotechnol J. 2021;19:6009-19.

[47] Liu C, Song J, Ogata H, Akutsu T. MSNet-4mC: learning effective multi-scale representations for identifying DNA N4-methylcytosine sites. Bioinformatics. 2022;38:5160-7.

[48] Zulfiqar H, Huang Q-L, Lv H, Sun Z-J, Dao F-Y, Lin H. Deep-4mCGP: A Deep Learning Approach to Predict 4mC Sites in Geobacter pickeringii by Using Correlation-Based Feature Selection Technique. Int J Mol Sci. 2022;23:1251.

[49] Halder RK, Uddin MN, Uddin MA, Aryal S, Islam MA, Hossain F, et al. A Grid Search-Based Multilayer Dynamic Ensemble System to Identify DNA N4—Methylcytosine Using Deep Learning Approach. Genes. 2023;14:582.

[50] Uddin MN, Halder RK. An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach. Inf Med Unlocked. 2021;24:100584.

[51] Nguyen V-N, Ho T-T, Doan T-D, Le NQK. Using a hybrid neural network architecture for DNA sequence representation: A study on N^4-methylcytosine sites. Comput Biol Med. 2024;178:108664.

[52] Liang Y, Wu Y, Zhang Z, Liu N, Peng J, Tang J. Hyb4mC: a hybrid DNA2vec-based model for DNA N4-methylcytosine sites prediction. BMC Bioinf. 2022;23.

[53] Pflughaupt P, Abdullah Adib A, Masuda K, Sahakyan Aleksandr B. Towards the genomic sequence code of DNA fragility for machine learning. Nucleic Acids Res. 2024;52:12798-816.

[54] Dey R, Salem FM. Gate-variants of Gated Recurrent Unit (GRU) neural networks. 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS): IEEE; 2017. p. 1597-600.

[55] Yassi M, Chatterjee A, Parry M. Application of deep learning in cancer epigenetics through DNA methylation analysis. Briefings Bioinf. 2023;24.

[56] Jeong Y, Gerhäuser C, Sauter G, Schlomm T, Rohr K, Lutsik P. MethylBERT enables read-level DNA methylation pattern identification and tumour deconvolution using a Transformer-based model. Nat Commun. 2025;16.

[57] Yuan J, Wang Z, Pan Z, Li A, Zhang Z, Cui F. DPNN-ac4C: a dual-path neural network with self-attention mechanism for identification of N4-acetylcytidine (ac4C) in mRNA. Bioinformatics. 2024;40:btae625.

[58] Yang X, Ye X, Li X, Wei L. iDNA-MT: Identification DNA Modification Sites in Multiple Species by Using Multi-Task Learning Based a Neural Network Tool. Front Genet. 2021;12.

[59] Jin J, Yu Y, Wei L. Mouse4mC-BGRU: Deep learning for predicting DNA N4-methylcytosine sites in mouse genome. Methods. 2022;204:258-62.

[60] Nguyen-Vo T-H, Trinh QH, Nguyen L, Nguyen-Hoang P-U, Rahardja S, Nguyen BP. i4mC-GRU: Identifying DNA N4-Methylcytosine sites in mouse genomes using bidirectional gated recurrent unit and sequence-embedded features. Comput Struct Biotechnol J. 2023;21:3045-53.

[61] 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.

[62] Yu Y, He W, Jin J, Xiao G, Cui L, Zeng R, et al. iDNA-ABT: advanced deep learning model for detecting DNA methylation with adaptive features and transductive information maximization. Bioinformatics. 2021;37:4603-10.

[63] Zhuo L, Wang R, Fu X, Yao X. StableDNAm: towards a stable and efficient model for predicting DNA methylation based on adaptive feature correction learning. BMC Genomics. 2023;24:742.

[64] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition: IEEE; 2018.

[65] Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 2021;37:2112-20.

[66] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In:

Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171-86.

[67] Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu HJae-p. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. 2023. p. arXiv:2306.15006.

[68] Mo S, Fu X, Hong C, Chen Y, Zheng Y, Tang X, et al. Multi-modal Self-supervised Pre-training for Large-scale Genome Data. NeurIPS 2021 AI for Science Workshop2021.

[69] Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, Li DB, et al. Sequence modeling and design from molecular to genome scale with Evo. Science.386:eado9336.

[70] Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Lopez Carranza N, Grzywaczewski AH, Oteri F, et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. Nat Methods. 2025;22:287-97.

[71] Jin J, Yu Y, Wang R, Zeng X, Pang C, Jiang Y, et al. iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. Genome Biol. 2022;23.

[72] Yu X, Yani C, Wang Z, Long H, Zeng R, Liu X, et al. iDNA-ITLM: An interpretable and transferable learning model for identifying DNA methylation. PLoS One. 2024;19:e0301791.

[73] Yu X, Ren J, Long H, Zeng R, Zhang G, Bilal A, et al. iDNA-OpenPrompt: OpenPrompt learning model for identifying DNA methylation. Front Genet. 2024;15.

[74] Xie H, Wang L, Qian Y, Ding Y, Guo F. Methyl-GP: accurate generic DNA methylation prediction based on a language model and representation learning. Nucleic Acids Res. 2025;53.

[75] Young T, Hazarika D, Poria S, Cambria E. Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. IEEE Comput Intell Mag. 2018;13:55-75.

[76] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436-44.

[77] Zheng P, Zhang G, Liu Y, Huang G. MultiScale-CNN-4mCPred: a multi-scale CNN and adaptive embedding-based method for mouse genome DNA N4-methylcytosine prediction. BMC Bioinf. 2023;24.

[78] Wang X, Du Q, Wang R. Mus4mCPred: Accurate Identification of DNA N4-Methylcytosine Sites in Mouse Genome Using Multi-View Feature Learning and Deep Hybrid Network. Processes: MDPI AG; 2024. p. 1129.

[79] Teragawa S, Wang L, Liu Y. DeepPGD: A Deep Learning Model for DNA Methylation Prediction Using Temporal Convolution, BiLSTM, and Attention Mechanism. Int J Mol Sci. 2024;25:8146.

[80] Wong TT, Yeh PY. Reliable Accuracy Estimates from k-Fold Cross Validation. IEEE Trans Knowl Data Eng. 2020;32:1586-94.

[81] Rodriguez JD, Perez A, Lozano JA. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. IEEE Trans Pattern Anal Mach Intell. 2010;32:569-75.

[82] Wang Y, Zhai Y, Ding Y, Zou Q. SBSM-Pro: support bio-sequence machine for proteins. Science China Information Sciences. 2024;67:212106.

[83] Raza A, Uddin J, Akbar S, Alarfaj FK, Zou Q, Ahmad A. Comprehensive Analysis of Computational Methods for Predicting Anti-inflammatory Peptides. Arch Comput Methods Eng. 2024;31:3211-29.

[84] Liu Z-Y, Xing J-F, Chen W, Luan M-W, Xie R, Huang J, et al. MDR: an integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae. Hortic Res. 2019;6.

[85] Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010;7:461-5.

[86] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150-2.

[87] Wang Z, Monteiro CD, Jagodnik KM, Fernandez NF, Gundersen GW, Rouillard AD, et al. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. Nat Commun. 2016;7.

[88] Coordinators NR. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2013;41:D8-D20.

[89] Wu X, Bartel DP. kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences. Nucleic Acids Res.

2017;45:W534-W8.

[90] Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, et al. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat Genet. 2012;44:743-50.

[91] Rajewska M, Wegrzyn K, Konieczny I. AT-rich region and repeated sequences – the essential elements of replication origins of bacterial replicons. FEMS Microbiol Rev. 2012;36:408-34.

[92] Bailey TL. STREME: accurate and versatile sequence motif discovery. Bioinformatics. 2021;37:2834-40.

[93] Schoenfelder S, Fraser P. Long-range enhancer–promoter contacts in gene expression control. Nat Rev Genet. 2019;20:437-55.

[94] Yao Z, Shangguan H, Xie W, Liu J, He S, Huang H, et al. SIPSC-Kac: Integrating swarm intelligence and protein spatial characteristics for enhanced lysine acetylation site identification. Int J Biol Macromol. 2024;282:137237.