



## EnDeep4mC predicts DNA $M^4$ -methylcytosine sites using a dual-adaptive feature encoding framework in deep ensembles

Shuyu Zhang, Quan Zou, Mengting Niu, et al.

*Genome Res.* 2026 36: 589-599 originally published online February 17, 2026

Access the most recent version at doi:[10.1101/gr.280977.125](https://doi.org/10.1101/gr.280977.125)

---

**References** This article cites 52 articles, 6 of which can be accessed free at:  
<http://genome.cshlp.org/content/36/3/589.full.html#ref-list-1>

**Open Access** Freely available online through the *Genome Research* Open Access option.

**Creative Commons License** This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

---

## Method

# EnDeep4mC predicts DNA $N^4$ -methylcytosine sites using a dual-adaptive feature encoding framework in deep ensembles

Shuyu Zhang,<sup>1</sup> Quan Zou,<sup>1,2</sup> Mengting Niu,<sup>1,2</sup> Zhibin Lv,<sup>3</sup> Antony Stalin,<sup>1</sup> and Ximei Luo<sup>1,2</sup>

<sup>1</sup>Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, Sichuan, China; <sup>2</sup>Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324000, Zhejiang, China; <sup>3</sup>College of Biomedical Engineering, Sichuan University, Chengdu 610041, Sichuan, China

DNA  $N^4$ -methylcytosine (4mC), a key epigenetic modification regulating DNA repair and replication, requires efficient computational detection methods due to experimental limitations. Although machine learning predictors have been proposed, their performance could be enhanced through systematic optimization of feature encoding schemes. Here, we propose EnDeep4mC, a dual-adaptive framework integrating species-specific modeling with ensemble deep learning architectures to systematically optimize feature encoding schemes. Evaluated across six species, EnDeep4mC demonstrates commendable prediction performance and significantly outperforms current state-of-the-art predictors. Cross-species validation confirms its robust transferability from animal to microbe groups. Evolutionary analysis further uncovers the functional differentiation of 4mC sequences in biological evolution: Prokaryotic 4mC relies on stable patterns, whereas eukaryotes achieve regulatory plasticity through dynamic sequence combinations, which provides experimental evidence for species-adaptive encoding strategies.

[Supplemental material is available for this article.]

DNA methylation is a form of DNA epigenetic modification, which modulates the genetic performance without changing the DNA sequence (Moore et al. 2013). Numerous studies have shown that DNA methylation can alter chromatin structure, DNA conformation, DNA stability, and DNA–protein interaction to regulate gene expression (Razin and Riggs 1980; Cheng 1995; Jones and Takai 2001; Robertson 2005; Wang et al. 2024b).  $N^6$ -methyladenine (6mA), 5-methylcytosine (5mC), and  $N^4$ -methylcytosine (4mC) are three common DNA methylation modifications. Compared with 5mC and 6mA, 4mC has received comparatively less attention, primarily owing to existing costly and time-consuming experimental methods (Flusberg et al. 2010; Yu et al. 2015). First identified in 1983, 4mC is ubiquitous in prokaryotes (such as bacteria) but relatively rare in eukaryotes (Beaulaurier et al. 2019). Similar to 6mA and 5mC, 4mC plays critical roles in many biological processes, including DNA repair and replication. It operates as a core component in the antiphage immune system of bacteria, mainly reflected in the participation of 4mC in the bacterial restriction–modification system (RM system). This system safeguards bacteria against phage and other invasive foreign DNA (O’Brown et al. 2019; Yu et al. 2021; Xiong et al. 2022). Consequently, 4mC is recognized as a key regulator of gene expression, DNA repair, and antiphage immunity in prokaryotes, particularly bacteria, highlighting its potential value for microbiology and antimicrobial drug development (Vandenbussche et al. 2021; Yu et al. 2023; Zheng et al. 2023). To systematically investigate its biological functions and mechanisms, developing re-

liable detection methods is essential. Although recent advances in nanopore sequencing allow for the direct detection of 5mC and 4mC during basecalling (Liu et al. 2021b; Wang et al. 2021), conventional experimental approaches remain expensive, time consuming, and labor intensive (Flusberg et al. 2010; Yu et al. 2015). It is also important to note that nanopore-based detection methods depend on machine learning (ML) models to interpret sequencing signals accurately (Liu et al. 2019; Wang et al. 2021; Chen et al. 2025; Galeone et al. 2025). Therefore, developing bioinformatics tools for the large-scale, accurate, and efficient 4mC site identification is a growing trend and an effective supplement to laboratory methods (Wang et al. 2024a).

In recent years, considerable progress has been made in predicting DNA methylation sites. Researchers have developed a series of effective DNA methylation site predictors based on traditional ML and recent deep learning (DL) algorithms (Feng et al. 2016; Zhou et al. 2016; Jin et al. 2019; Bonet et al. 2022). Among them, Chen et al. (2017) proposed the first SVM-based predictor iDNA4mC and screened the MethSMRT database (Ye et al. 2017) to construct the benchmark data set. Most of the subsequent studies based on ML algorithms also continued to adopt this data set, such as 4mCPred (He et al. 2019), Meta-4mCpred (Manavalan et al. 2019), 4mCPred-SVM (Wei et al. 2019a), 4mCPred-IFL (Wei et al. 2019b), etc. Subsequently, Khanal et al. (2019) proposed 4mCCNN, the first predictor based on the DL method, convolutional neural networks (CNN), for 4mC identification. Based on this, Liu et al. (2021a) proposed DeepTorrent by combining multiple feature encoding schemes with CNN and Bi-LSTM algorithms. Xu et al. (2021) proposed Deep4mC to further improve the model

**Corresponding author:** [luoximei@uestc.edu.cn](mailto:luoximei@uestc.edu.cn)

Article published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.280977.125>. Freely available online through the *Genome Research* Open Access option.

© 2026 Zhang et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

performance by building a stacked model based on the prediction probabilities of six different ML algorithms. Liu et al. (2022) addressed a sample imbalance of species via scale-aware learning, integrating class weights into cross-entropy loss to develop MSNet-4mC. Based on Hyb4mC (Liang et al. 2022), iDNA-MS (Lv et al. 2020), and DeepTorrent (Liu et al. 2021a), Li et al. (2023) constructed a more robust data set and proposed EpiTEAmDNA, the first stacked model integrating traditional ML algorithms and CNN. It achieved better performance than other DL-based algorithms on small sample data sets. Yao et al. (2024) used a new data set proposed by Zeng et al. (Zeng and Liao 2020) to leverage existing models through transfer learning and combined multiple ensemble learning techniques to build DeepSF-4mC, which achieved enhanced prediction performance on three specific species.

However, the above methods still exhibit two critical constraints. First, traditional methods usually adopt static encoding schemes in feature engineering that neglect the distinct characteristics of different species and models, resulting in the one-size-fits-all encoding dilemma. Second, prevalent predictors mostly rely on homogeneous architectures, which are difficult to simultaneously capturing local conserved patterns, long-range context dependence, and global attention correlations of DNA sequences.

To address these limitations, we propose EnDeep4mC, a dual-adaptive optimization framework. The core innovations of EnDeep4mC include (1) a species-model collaborative mechanism that dynamically selecting feature encodings via cross-species/model performance quantification and incremental feature selection and (2) a three-tier probability fusion architecture in which Tier-I employs models (CNN, Bi-LSTM, Transformer) to parallelly generate prediction initial probabilities, Tier-II concatenates probabilities via XGBoost/LightGBM for feature enhancement, and Tier-III applies elastic-net logistic regression to achieve a robust and regularized decision fusion.

Compared with existing predictors, EnDeep4mC achieves three advances: (1) It transforms traditional feature selection into a dual-optimization problem integrating species specificity and model architecture, thereby overcoming the fixed-feature limitations in cross-species/model scenarios; (2) it achieves 94.72% average AUC across benchmark data sets, outperforming SOTA predictors with robust cross-species transferability; and (3) it systematically elucidates species-feature associations and reveals 4mC sequence functional differentiation in biological evolution from a eukaryote/prokaryote perspective, offering potential biological insights that could guide future adaptive encoding strategies.

The following sections first detail the Methods used to construct the EnDeep4mC framework and then present a comprehensive evaluation of its performance and biological findings.

## Methods

### Benchmark data sets

A variety of benchmark data sets have been proposed in the field of 4mC site prediction. Most of them before 2021 were designed for ML models with relatively few samples (Chen et al. 2017; Ye et al. 2017; He et al. 2019; Manavalan et al. 2019; Wei et al. 2019a,b). However, reasonable and sufficient data sets are essential for training DL models. The 4mC data used in this experiment are from EpiTEAmDNA (Li et al. 2023). To the best of our knowledge, this is the most comprehensive resource of its kind. This data set contains three modification types: 4mC, 5mC, and 6mA. The 4mC modification data are integrated from data sets established in Hyb4mC (Liang et al. 2022), iDNA-MS (Lv et al. 2020), and

DeepTorrent (Liu et al. 2021a). All collected 4mC sequences (41 bp) have a methylated cytosine (C) at the center, whereas negative samples are nonmethylated, as confirmed by SMRT sequencing. All input DNA sequences are processed into 41 bp fixed-length samples. Sequences <41 bp are symmetrically padded with N's. Longer sequences are segmented into consecutive, overlapping 41 bp sliding windows. This ensures the model operates within its validated parameters (Supplemental Fig. S1; for details, see Supplemental Methods). Furthermore, to ensure the consistency of comparison, the CD-HIT tool was used to remove redundant sequences. The final data set comprises 11 model organisms. For model training, we mainly focused on six species: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Escherichia coli*, *Geobacillus subterraneus*, *Geobacter pickeringii*. The respective positive sample counts for these species are 125,628, 68,917, 109,289, 2068, 14,877, and 5687, respectively. Training and independent test data sets were split as shown in Supplemental Table S1.

### Overview of the EnDeep4mC architecture

The EnDeep4mC framework employs a dual-phase architecture for species-model co-optimization: (1) a dynamic feature selection (DFS) module identifies optimal species-model feature combinations via dual adaptation, and (2) a heterogeneous ensemble fusion module constructs a metalearning system using deep probabilistic predictions (Fig. 1B).

The workflow initiates with feature-enhanced base models (CNN, Bi-LSTM, Transformer), each trained on species/model-specific optimized encodings to generate probabilistic outputs. These outputs are hierarchically fused via stacking, processed by XGBoost/LightGBM as base learners and finalized through elastic network logistic regression, to yield robust classification.

### Feature-enhanced base model architecture

Guided by ensemble learning principles that emphasize the importance of error diversity among base models (Dietterich 2000; Kuncheva and Whitaker 2003), we constructed three DL networks—CNN, Bi-LSTM, and Transformer—to collaboratively capture multiscale sequence features (for visual effect, Supplemental Fig. S2). This principled selection, validated by subsequent quantitative diversity analysis (Supplemental Table S2), ensures that the ensemble leverages distinct and complementary predictive behaviors rather than redundant signals.

The CNN model adopts a hierarchical convolution architecture with 256 1D convolution kernels (kernel\_size = 1) and separable convolution layers (128 filters, kernel\_size = 3). The sequence feature compression is realized by the global max pooling layer. L2 regularization ( $\lambda = 0.001$ ) and dropout (rate = 0.3) were introduced to mitigate overfitting.

Bi-LSTM model employs bidirectional recursive structure with two LSTM layers: The first (128 units) captures local temporal patterns via sequential information retention, and the second (64 units) performs feature abstraction. The recurrent dropout (rate = 0.1) and batch normalization were introduced to enhance the generalization.

The Transformer model features a multilayer encoder with multihead attention (two heads) and a feed-forward network (FFN\_dim = 64) per layer. Training stability is ensured via layer normalization (LayerNorm  $\epsilon = 1 \times 10^{-6}$ ) and adaptive learning rate scheduling (initial lr =  $1 \times 10^{-3}$ , decay rate = 0.9).

All base models were trained on species-specific data sets using fivefold cross-validation, optimized with Adam (gradient clipping threshold = 1.0) and sigmoid activation. The DFS module was integrated into each base model to dynamically optimize encoding combinations during training.



specify all layers, regularization strategies, optimizer settings, and the complete training protocol, enabling exact replication of the model architectures and experimental results.

### Feature encoding schemes

More than 24 common feature encoding schemes are widely used in epigenetic site prediction. In the work of EpiTEAmDNA, Li et al. (2023) demonstrated that 10 feature encoding schemes such as pseKNC, pseDNC, and DAC are unsuitable for the 4mC site prediction task based on DL architecture based on computational efficiency and contribution to accuracy improvement. Ultimately, this study utilizes the remaining 14 feature encoding schemes—ENAC, Binary, NCP, EIIP, *k*-mer, CKSNAP, PseEIIP, TNC, RCK-mer, SCPseTNC, PCPseTNC, ANF, NAC, and TAC—as our candidate set. The key characteristics of the 14 selected encoding schemes are summarized in Supplemental Table S6. Based on the encoding schemes defined in the iLearn toolkit (Chen et al. 2020), we provide their detailed descriptions in the Supplemental Methods.

### DFS framework

To enhance the universality of feature encoding schemes, we proposed a DFS framework to optimize the feature selectivity of species and the characteristics of model architecture simultaneously. As shown in Figure 1A, the framework achieves end-to-end optimization through two stages of feature-model cross-evaluation and dynamic feature fusion. The core innovation of the framework is to establish a 3D selection space of species-model-features and to realize the dynamic adaptation of feature combinations through a two-layer optimization mechanism.

The first stage is feature-model cross-evaluation. The purpose of this stage is to realize the joint analysis of species model traits by constructing a 3D evaluation matrix. Specifically, for each species  $s \in \{A. thaliana, C. elegans, \dots\}$  and DL base model  $k \in \{\text{CNN, BLSTM, Transformer}\}$ , the utility of 14 candidate feature encoding methods  $f_i$  is evaluated independently. The evaluation process was carried out in the following steps: First, the performance of a single feature was quantified by calculating the independent classification accuracy  $Acc(s, k, f_i)$  of a feature  $f_i$  on the species  $s$  using a five-fold cross-validation strategy, which was calculated as follows:

$$Acc(s, k, f_i) = \frac{1}{N_{test}} \sum_{m=1}^{N_{test}} I(y_{test}^{(m)} = \hat{y}_{test}^{(m)}), \quad (1)$$

where  $I$  is the indicator function, and  $y_{test}$ ,  $\hat{y}_{test}$  represent the true label and the prediction result, respectively.

Second, generate feature utility ranking  $R(s, k) = [f_{(1)}, f_{(2)}, \dots, f_{(i)}, \dots, f_{(14)}]$  for pairs determined by species and model  $(s, k)$ , where  $f_{(i)}$  denotes the optimal features for  $(s, k)$  pairs, in  $Acc(s, k, f_i)$  descending order.

Through feature-model cross-evaluation, we establish a 3D decision space containing  $6 \times 3 \times 14$  evaluation units, thereby quantifying the difference of feature selectivity between different model architectures on different species data.

The second stage is dynamic feature fusion. In this stage, the incremental feature combination optimization is carried out based on the ranking results to realize the dual-adaptive mechanism. For each  $(s, k)$  pair, 14 candidate feature subsets  $\{F_n^{(s,k)}\}_{n=1}^{14}$  are constructed in order  $R(s, k)$ , where  $F_n^{(s,k)} = \bigcup_{i=1}^n f_{(i)}$  denotes the union of the previous  $n$  features. Based on model adaptation, the prediction performance  $Acc(F_n^{(s,k)})$  of each  $F_n^{(s,k)}$  is evaluated on the validation set, and the  $n^*$  that maximizes the validation accuracy is selected:

$$n^* = \arg \max_{n \in [1, 14]} Acc(F_n^{(s,k)}). \quad (2)$$

Based on species adaptability, we allow different species to obtain different  $n^*$  values when using different models for feature encoding. For example, for the CNN model, the  $n^*$  of *A. thaliana* is six, whereas that of *C. elegans* is nine. Likewise, there are significant differences in the optimal feature combinations selected by different models under the same species (Supplemental Table S7).

Finally, the following objective function was used as the species-model joint optimization objective:

$$\max \sum_{s=1}^6 \sum_{k=1}^3 Acc(F_{opt}^{(s,k)}), \quad (3)$$

where  $S$  represents six species, and  $K$  represents for three models. By joint optimization, the objectives selected by  $n^*$  satisfies  $F_{opt}^{(s,k)} = F_{n^*}^{(s,k)}$ .

The above computational framework incorporates phylogenetic feature selectivity and architecture-specific feature optimization through its dual-adaptive mechanism. The selected optimal features are provided in Supplemental Table S7. From the perspective of species, the framework establishes a species-specific feature ranking  $R(s, k)$ , which can capture the property differences of DNA sequences of different species. For example, experimental results show that *A. thaliana* showed maximal compatibility with physicochemical features (e.g., NCP, EIIP), whereas *G. subterraneus* demonstrated optimal performance with sequence composition features (e.g., CKSNAP, *k*-mer). For the model perspective, our proposed feature selection framework optimizes the feature processing ability of different DL base models. For example, in the experimental results, for *A. thaliana*, CNN and Bi-LSTM only select six out of 14 candidate encodings, whereas Transformer selects 13 candidate encodings. The optimal encodings of different base models show significant difference.

### Evaluation metrics

Comparative benchmarking with existing predictors was conducted through six evaluation metrics: accuracy (ACC), sensitivity (SN), specificity (SP), Matthews correlation coefficient (MCC), area under the curve (AUC), and F1-score. The above metrics are defined as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \quad (4)$$

$$SN = \frac{TP}{TP + FN}, \quad (5)$$

$$SP = \frac{TN}{TN + FP}, \quad (6)$$

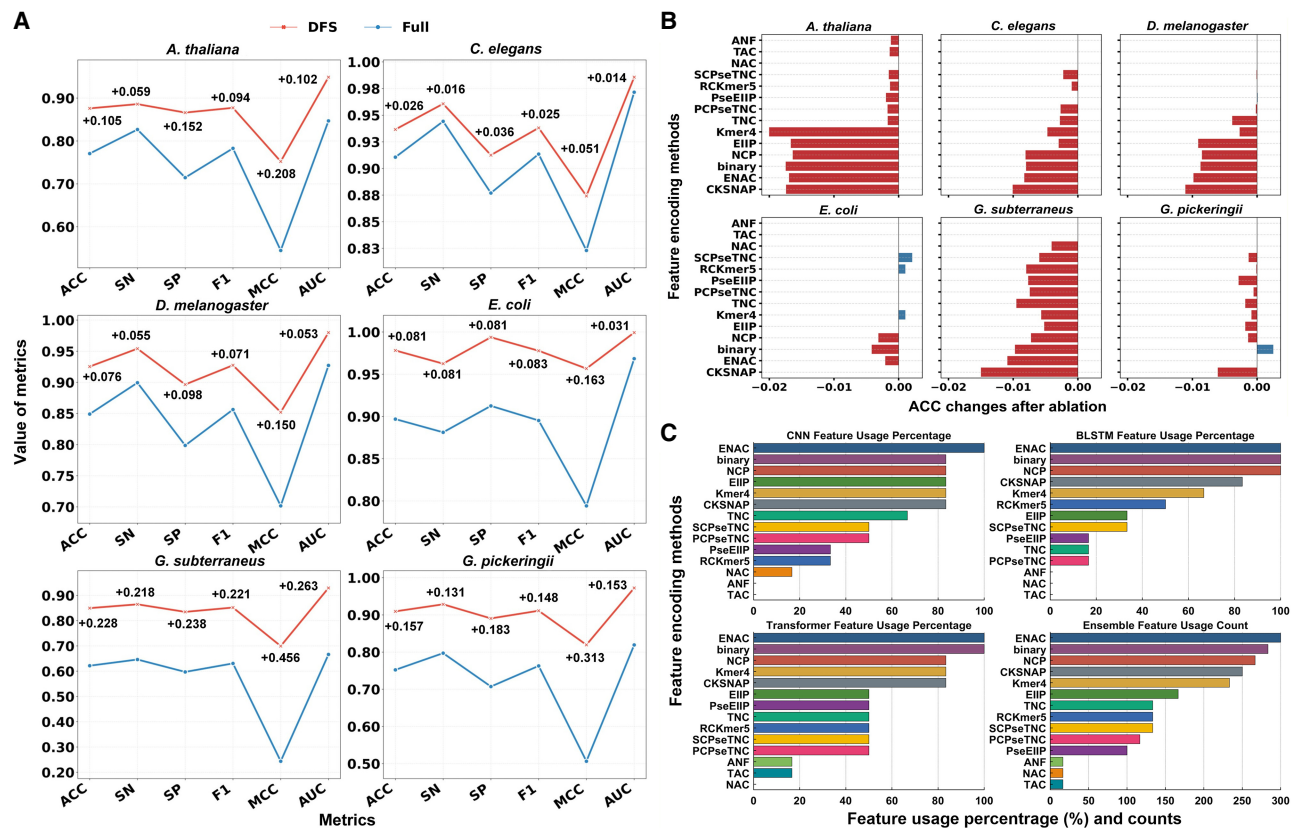
$$F1 \text{ Score} = 1 - \frac{TP + TN}{2 \times TP + FP + FN}, \quad (7)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}, \quad (8)$$

where TP, TN, FP, and FN represent the number of true-positive, true-negative, false-positive, and false-negative samples, respectively. AUC is a commonly used metric to evaluate the performance of binary classifiers. The  $x$ -axis of the ROC curve is the false-positive rate (FPR), and the  $y$ -axis is the true-positive rate (TPR). AUC is the area enclosed by the ROC curve against the  $x$ -axis. Using these metrics and the framework described above, we proceeded to evaluate the performance of EnDeep4mC, as detailed in the Results section.

## Results

EnDeep4mC framework has been evaluated through a series of experiments designed to validate the efficacy of its core components:



**Figure 2.** The impact of feature selection and encoding on EnDeep4mC's performance. (A) Metrics of the ensemble model using DFS versus full 14 encodings on the data sets of six species. (B) ACC changes of features selected by the DFS framework after ablation (Supplemental Tables S8–S10). (C) The selection rates of 14 feature encoding schemes on different base models and the total usage frequencies on ensemble model.

the DFS mechanism, the heterogeneous ensemble architecture, and its overall predictive performance. The results are presented as follows.

### DFS of EnDeep4mC

To validate the performance gains of the DFS framework, we conducted a systematic comparison against the baseline implementation using the 14 encoding schemes listed in Supplemental Table S6. As Figure 2A shows, the dual-adaptive DFS framework exhibited systematic performance gains on the data sets of six species. Specifically, the proposed framework improved the average AUC and ACC of the ensemble model by 11.84% (DFS: 0.9691 vs. full: 0.8666) and 14.05% (DFS: 91.25% vs. full: 80.01%), respectively. All the six metrics were positively improved in all species (+17.43% on average) (for feature selection details, Supplemental Fig. S3).

The results show that the feature selection framework can eliminate redundant and noisy features (such as TAC, ANF, etc.) while strengthening species-specific feature expression, as evidenced by NCP selectivity in *A. thaliana* and CKSNAP selectivity in *G. subterraneus*, through co-optimization of species-specific feature selectivity and model architectural characteristics. Collectively, these findings confirm that the proposed dual-adaptive framework enables the model to focus on the key discriminative features of 4mC sites, which significantly improves the prediction performance of EnDeep4mC.

To evaluate the impact of feature encoding schemes on EnDeep4mC, we conducted ablation experiments on the 14 candi-

date encoding schemes. We removed each feature selected in different DL base models through the DFS framework to examine the performance changes of the model on data sets of all six species. As Figure 2B shows, the average performance variation results of the cross-model analysis. The removal of 83.6% of the features led to a significant decline in model performance, with an average ACC decrease of 0.54%, demonstrating the efficacy of the DFS framework. Among them, features such as CKSNAP, ENAC, EIIP, and *k*-mer exhibited universal cross-species contributions. When these features were removed in cross-species scenarios, ACC decreased by an average of 0.55% to 1.19%. Particularly, the removal of *k*-mer resulted in an average decrease of 2.00% in ACC on the *A. thaliana* data set. Meanwhile, the *k*-mer encoding was also the one and only important encoding that ranked among the top six in all base models and all species. Notably, performance improvements were observed in *E. coli* upon removing SCPseTNC, RCKmer, and *k*-mer, potentially owing to feature redundancy in small sample contexts.

Furthermore, model-specific effects were observed. In the *G. pickeringii* data set, ablating Binary encoding reduced CNN accuracy by 0.39% but increased Bi-LSTM accuracy by 0.80%, underscoring the architecture-dependent utility of encoding schemes. These findings verify the necessity of the species-model collaborative optimization framework, which helps to mitigate adverse effects in heterogeneous models.

We also systematically analyzed the feature selection patterns of the DFS framework across all base models (Fig. 2C). ENAC, Binary, and NCP exhibited dominance, with average selection

rates of 100%, 94.44%, and 88.89%, respectively. Especially, ENAC achieved the maximum theoretical selection frequency (18 instances), highlighting its universal representational ability for cross-species 4mC prediction. In contrast, encoding schemes such as ANF, NAC, and TAC were not selected in most species, suggesting that their representational capabilities might be limited by specific sequence context environments.

### Comparison with single base models

In this section, we further evaluated the performance comparison of EnDeep4mC and its three base models with fivefold cross-validation on different species data sets, with key metrics reported as mean  $\pm$  standard deviation. The narrow 95% confidence intervals (typically within  $\pm 0.5$ –1.5% of the mean) confirm the stability of the results across different data partitions. As summarized in Table 1 and detailed in Supplemental Tables S11 through S16 and Supplemental Figure S4, the EnDeep4mC ensemble outperformed the other individual models across all species, with average ACC and AUC of 94.99% and 0.9859, respectively, representing 4.52% and 2.12% improvements over the best base model (Transformer). Notably, EnDeep4mC achieved near-optimal performance on *E. coli*, with an ACC of 99.71% and an AUC of 0.9999, approaching theoretical perfection.

Among the base models, CNN and Bi-LSTM showed comparable performance. The Transformer was demonstrated the strongest overall performance (average ACC = 90.47%), outperforming CNN (90.30%) and Bi-LSTM (90.45%). It also exhibited stable performance on *C. elegans* and *D. melanogaster*, with ACC reaching 92.96% and 92.46%, respectively. Notably, Transformer achieved superior sensitivity (SN = 0.9545) on *D. melanogaster*, even surpassing EnDeep4mC by 0.0037, which might be because of the stronger ability of Transformer to capture the global features of specific sequential patterns.

Through cross-species comparative analysis, it was also found that the standard deviation of ACC of EnDeep4mC was 3.62%, which was lower than that of the base models (3.85%–4.12%), demonstrating that its ensemble strategy effectively mitigated data bias sensitivity. In addition, the AUC of EnDeep4mC in *A. thaliana* and *G. pickeringii* reached 0.9697 and 0.9999, respectively, which further verified its generalization ability. The performance improvements were particularly significant on microbial species: EnDeep4mC achieved 7.98% and 5.49% ACC gains over the best base model on *G. subterraneus* and *G. pickeringii*, respectively, highlighting the framework's robustness in handling phylogenetically diverse data sets. Therefore, the EnDeep4mC framework is feasible on all benchmark data sets of DNA 4mC, as well as its superiority over traditional single DL models in terms of prediction accuracy, robustness, and cross-species adaptability.

### Ablation experiment of EnDeep4mC model structure

The probabilistic fusion architecture of EnDeep4mC relies on the complementary integration of CNN, Bi-LSTM, and Transformer. To validate their synergistic effects, we systematically excluded each single base model and retrain the ensemble system to quantify the performance change on independent test sets of six species (Fig. 3). The results demonstrated that the removal of any single base model consistently degraded the ensemble's performance, and the decrease effect is species specific. As shown in Figure 3, excluding the Bi-LSTM had the largest negative impact on average ACC (−0.78%), AUC (−0.34%), and SP (−0.58%), with a notable effect on microbial groups such as *G. subterraneus* (−0.43% ACC).

The removal of Transformer significantly impacted the average MCC (−0.56%) and F1-score (−0.39%), especially in animal groups (e.g., *C. elegans* with −0.79% in MCC). Notably, excluding CNN and Transformer from the ensemble surprisingly slightly improved ACC in *G. pickeringii*, potentially owing to their local-global feature redundancy.

The analysis above highlight the complementary roles of the three base models. CNN provided the most stable performance for prokaryotes owing to its ability to parse short-range local patterns. Bi-LSTM effectively captured the sequence dependence of negative samples through bidirectional time series modeling, and its removal led to an average reduction of 0.58% in specificity (SP). The Transformer relied on the self-attention mechanism to strengthen the global correlation discrimination, which significantly improved the prediction accuracy of class imbalance samples. In summary, by integrating these heterogeneous base models, EnDeep4mC dynamically adapted to multiscale sequence features, providing actionable insights into ensemble architecture design in DL frameworks.

### Comparison with existing predictors

To verify the performance of EnDeep4mC, we systematically compared it against five DL models (4mCCNN, Deep4mC, Hyb4mC, DeepSF-4mC, and EpiTEAmDNA) and one traditional ML model (4mCPred-SVM). To ensure a fair and reproducible comparison with existing predictors, our benchmarking analysis adhered to the following principles: All models were evaluated on the identical benchmark data sets curated from EpiTEAmDNA, employing the same fivefold cross-validation splits and evaluation metrics. Comprehensive benchmarking analyses (Fig. 3B) demonstrate that EnDeep4mC's significant outperformance across key evaluation metrics. EnDeep4mC outperformed all baselines with average ACC of 95.28% and AUC of 0.9863, surpassing the suboptimal model EpiTEAmDNA by 2.76% (ACC) and 1.21% (AUC). Notably, EnDeep4mC achieved exceptional performance on *E. coli* (ACC = 99.73%, AUC = 0.9999), outperforming EpiTEAmDNA by 0.67% in ACC. This represents one of the highest reported accuracy rates, surpassing 90% across all benchmark species.

To ensure a statistically rigorous comparison, we conducted a side-by-side evaluation of EnDeep4mC against two prominent ensemble-based baselines—Hyb4mC and the current state-of-the-art EpiTEAmDNA—under consistent experimental conditions (identical data sets, fivefold CV). As detailed in Supplemental Table S20, EnDeep4mC achieves consistent and often substantial improvements across all six species. For instance, it attains a notable AUC gain of 11.2% over Hyb4mC on *G. subterraneus*, while maintaining a stable advantage over EpiTEAmDNA even on high-performing data sets like that of *E. coli*. These results validate the effectiveness of the species-model collaborative feature selection strategy, which enhances the prediction accuracy and robustness by adaptively fusing local conservation (CNN), long-range dependence (Bi-LSTM), and global correlation features (Transformer).

To assess the compatibility of EnDeep4mC with emerging detection platforms, we also evaluated the framework on independent data derived from a latest Oxford Nanopore Technologies (ONT) sequencing study (Galeone et al. 2025). Using motif-based labeling of 4mC and non-4mC sites from *Enterococcus faecium*, *Klebsiella pneumoniae*, and *Listeria monocytogenes*, we constructed a FASTA-format data set of variable length (Supplemental Table S21). EnDeep4mC and its base models all delivered robust performance on this set (Supplemental Table S22), with the ensemble attaining an average AUC more than 0.9999 across the three species.

**Table 1.** Performance comparison between EnDeep4mC ensemble model and each base model trained by fivefold cross-validation on six species data sets

Data set	Algorithm	ACC	SN	SP	MCC	AUC	F1-score
<i>C. elegans</i>	CNN	0.9252	0.9559	0.8946	0.8529	0.9812	0.9278
	Bi-LSTM	0.9231	0.9514	0.8948	0.8461	0.9792	0.9246
	Transformer	0.9296	0.9589	0.9003	0.859	0.9829	0.9308
	EnDeep4mC	<b>0.9571</b>	<b>0.9594</b>	<b>0.9548</b>	<b>0.9142</b>	<b>0.9914</b>	<b>0.9572</b>
<i>D. melanogaster</i>	CNN	0.9191	0.9530	0.8851	0.8401	0.9758	0.9216
	Bi-LSTM	0.9169	0.9502	0.8836	0.8379	0.9738	0.9205
	Transformer	0.9246	<b>0.9540</b>	0.8951	0.8512	0.9778	0.927
	EnDeep4mC	<b>0.9412</b>	0.9508	<b>0.9316</b>	<b>0.8826</b>	<b>0.9842</b>	<b>0.9418</b>
<i>A. thaliana</i>	CNN	0.8727	0.8933	0.8522	0.7489	0.9453	0.8758
	Bi-LSTM	0.8702	0.8781	0.8623	0.7423	0.9423	0.8727
	Transformer	0.8789	0.8962	0.8617	0.7587	0.9491	0.8811
	EnDeep4mC	<b>0.9133</b>	<b>0.9205</b>	<b>0.9061</b>	<b>0.8267</b>	<b>0.9697</b>	<b>0.9133</b>
<i>E. coli</i>	CNN	0.9619	0.9513	0.9725	0.9438	0.9933	0.9718
	Bi-LSTM	0.9644	0.9437	0.9850	0.9251	0.9972	0.9623
	Transformer	0.9675	0.9613	0.9737	0.9376	0.9947	0.9686
	EnDeep4mC	<b>0.9973</b>	<b>0.9969</b>	<b>0.9976</b>	<b>0.9945</b>	<b>0.9999</b>	<b>0.9973</b>
<i>G. subterraneus</i>	CNN	0.8525	0.8725	0.8325	0.7102	0.9334	0.852
	Bi-LSTM	0.8551	0.8545	0.8557	0.7142	0.9309	0.8578
	Transformer	0.8514	0.8527	0.8500	0.7024	0.9282	0.8513
	EnDeep4mC	<b>0.9349</b>	<b>0.9323</b>	<b>0.9375</b>	<b>0.8698</b>	<b>0.9786</b>	<b>0.9347</b>
<i>G. pickeringii</i>	CNN	0.9075	0.9437	0.8712	0.8233	0.9722	0.9138
	Bi-LSTM	0.9180	0.9384	0.8976	0.8345	0.9742	0.9186
	Transformer	0.9107	0.9346	0.8868	0.8241	0.9709	0.9127
	EnDeep4mC	<b>0.9729</b>	<b>0.9740</b>	<b>0.9718</b>	<b>0.9457</b>	<b>0.9937</b>	<b>0.9729</b>

Bold values indicate the best performance achieved within each comparison group.

These results indicate that our sequence-based predictor retains competitive accuracy even when applied to data generated by state-of-the-art ONT sequencing, highlighting its potential as a rapid, cost-efficient screening tool that can usefully complement experimental profiling efforts.

### Cross-predictions between different species

To evaluate the cross-species generalization capability of EnDeep4mC, we performed systematic prediction across a broader range of species. The cross-prediction performance was visualized in a heatmap using AUC and ACC as indicators (Fig. 4A; Supplemental Fig. S5), which revealed the differences in the transfer ability. Notably, we adopted a strict feature inheritance protocol: For zero-shot transfer, target species sequences were encoded exclusively using the feature encoding scheme optimized for the source species during its training.

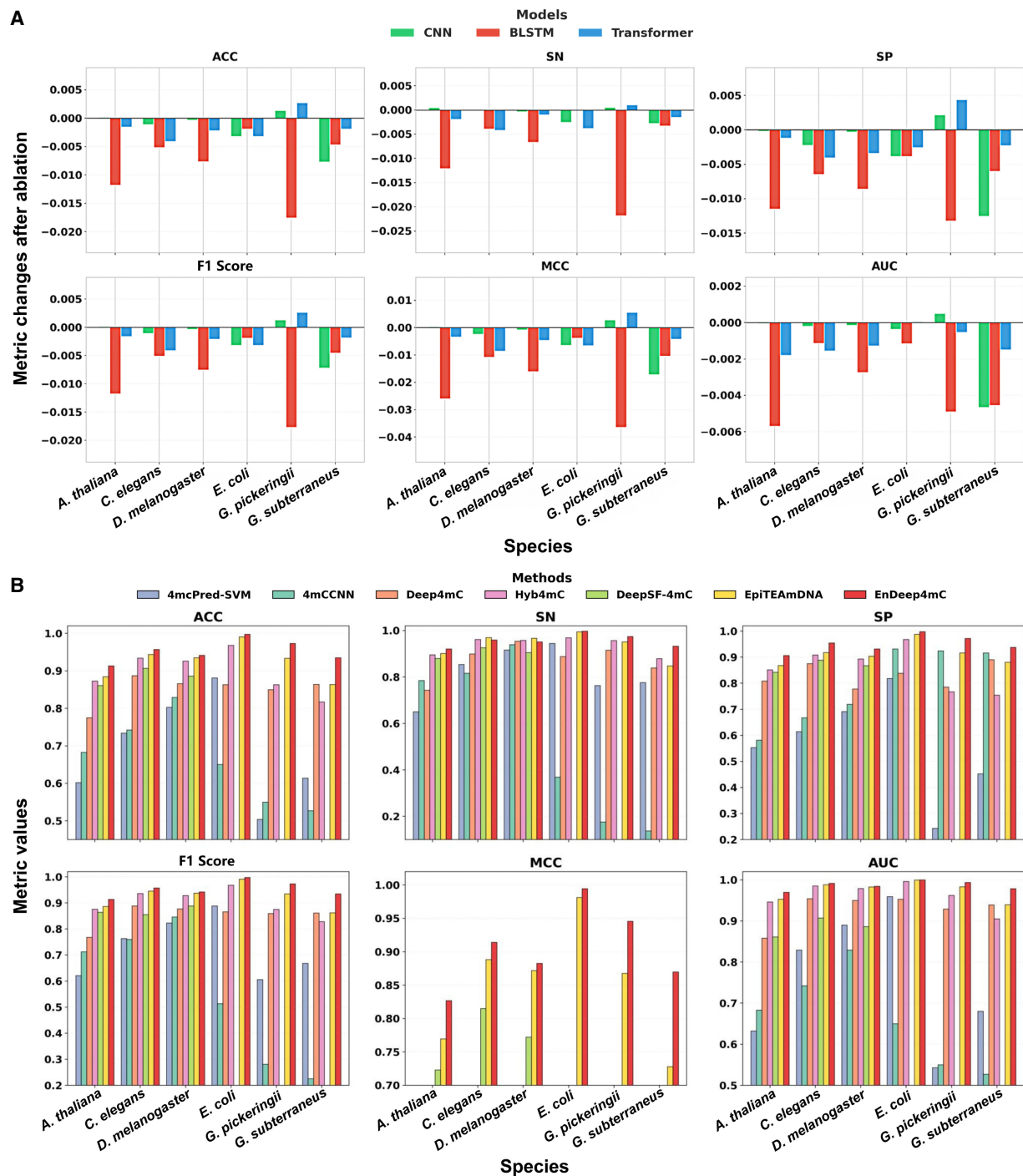
As shown in Figure 4A, the 4mC data sets from EpiTEAMDNA were categorized into plant, animal, and microbial groups. Cross-prediction results showed that the diagonal species achieved the highest performance (average ACC=83.36%, AUC=0.9138), validating intraspecies transfer reliability. Among them, *D. melanogaster* exhibited the best cross-species generalization ability in our model, achieving AUC of 0.9639 for *E. coli* and 0.9090 for *G. pickeringii*. In contrast, cross-group predictions were markedly less accurate. For instance, plant-to-microbial transfers showed

an average AUC of 0.6851, a 25.0% reduction compared to intra-species predictions.

Notably, *Fragaria vesca* achieved an AUC of 0.8028 in predicting *E. coli*—4.3% higher than intragroup microbial transfer (0.7696), indicating conserved methylation features in its sequences. The performance of intramicrobial prediction outperformed other groups with an average AUC of 0.7696, with *Tolypocladium-to-E. coli* prediction achieving AUC=0.9216, which might be related to the homology of the modification system in prokaryotes.

These observed performance patterns, strong intragroup predictability alongside attenuated cross-group transfer, directly support the utility of more targeted training strategies. Our results indicate that models trained specifically for a particular biological group (e.g., prokaryotes or plants) are likely to achieve superior performance when predicting 4mC sites in newly sequenced genomes from phylogenetically related organisms, compared with a single universal model (Zemach et al. 2010; Beaulaurier et al. 2019).

Furthermore, we identified specific species as acting as “knowledge hubs.” The model trained on *D. melanogaster* achieved AUC >0.7 in 80% cross-predictions and exhibited superior global transfer stability (standard deviation = 0.1269) compared with other species (average standard deviation = 0.1526). This property could be related to the global methylation pattern of DNA, and its sequences might capture more universal features of 4mC modification. These findings imply that integrating feature encoding schemes from multiple species is expected to further improve the prediction stability of the model in cross-group species.



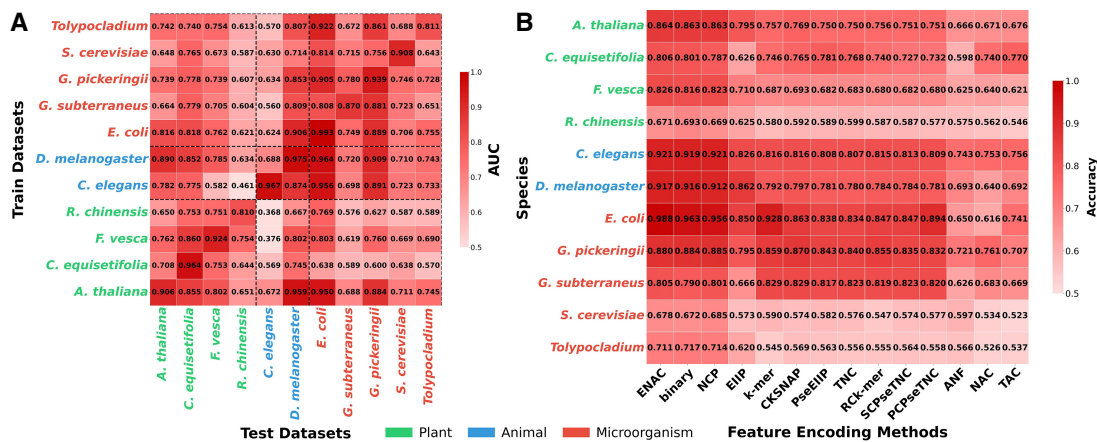
**Figure 3.** Comparative performance analysis of 4mC prediction methods. (A) Metrics changes of the system after removing each base model, respectively, in the EnDeep4mC framework (Supplemental Tables S17–S19). The vertical axis represents the relative change values of each metric. (B) Comprehensive performance comparison of seven 4mC prediction methods across multiple species and evaluation metrics. Blank values indicate that the corresponding metric was not reported in the original publication.

### Selectivity of different species for feature encoding schemes

Species-specific selectivity for feature encoding schemes was further analyzed across 11 species. We encoded all the 11 data sets from EpiTEAmDNA by 14 feature encoding schemes and evaluated ACC via three DL base models on independent test sets, thus

quantifying feature selectivity across species and models to uncover encoding specificity. For cross-group analysis, 11 species were still categorized into three groups: plants, animals, and microbes.

The results revealed significant feature encoding selectivity across three biological groups (as shown in Fig. 4B; Supplemental Fig. S6). Plant species showed optimal performance with global



**Figure 4.** Prediction performance on different species groups, in which species names are green for plant group, blue for animal group, and red for microbial group. (A) Cross-species prediction performance (AUC) evaluated on 11 species, including held-out subsets from the six benchmark species to assess zero-shot transfer to previously unseen data. (B) Prediction performance on independent test data sets of 11 species using 14 feature encodings respectively, taking CNN as an example (Supplemental Fig. S6).

statistical encodings: *A. thaliana* achieved 86.39% ACC with ENAC, whereas *F. vesca* attained 82.58% with the same encoding. Local sequential patterns like CKSNAP showed moderate performance (average ACC=70.48% across four plant species), with *Casuarina equisetifolia* reaching 76.50% ACC, indicating plants' preference for global sequence features (Zilberman et al. 2007).

Animal species exhibited strong preference for nucleotide chemical property (NCP) and global encodings. *C. elegans* achieved 92.11% ACC with NCP, whereas *D. melanogaster* attained 91.66% with ENAC. Notably, NCP encoding outperformed local *k*-mer encoding by 10.19% in *C. elegans*, suggesting animals' reliance on nucleotide-level chemical features for methylation recognition (Feng et al. 2010).

Microbial groups displayed significant prokaryotic-eukaryotic differentiation. Prokaryotes (*E. coli*, *G. subterraneus*, *G. pickeringii*) excelled with both global and local encodings: *E. coli* achieved exceptional 98.75% ACC with ENAC, whereas *G. pickeringii* attained 85.85% with *k*-mer. Eukaryotes (*S. cerevisiae*, *Tolypocladium*) showed lower performance across all encodings, with *S. cerevisiae* achieving only 67.80% ACC via NCP, 22.95% below the prokaryotic average (90.73%), reflecting sequence heterogeneity in eukaryotic epigenetic regulation (Jones 2012).

In addition, group stability analysis revealed animals exhibited the lowest coefficient of variation (CV=0.094), whereas plants (CV=0.120) and microbes (CV=0.200) showed higher volatility, consistent with feature granularity differences. These findings provide optimization guidelines for species-feature adaptation: Plant groups benefit from global statistical encodings (ENAC, binary); animal groups perform best with NCP encodings; and microbial groups require differentiation between prokaryotic and eukaryotic encoding schemes.

These findings provide a new perspective for optimization of species-feature adaptation. On the whole, plant groups show a preference for global statistical encodings (e.g., ENAC, binary) rather than local patterns, suggesting global sequence composition is more critical for methylation recognition in plants. Animal groups consistently perform best on NCP encodings and global statistical features. Microbial groups maintain significant prokaryotic-eukaryotic differentiation, necessitating distinct encoding strategies for prokaryotic versus eukaryotic species.

## Discussion

EnDeep4mC achieves state-of-the-art prediction performance and provides novel insights into the evolutionary dynamics of DNA 4mC modification. A key insight is its enhanced intraclass transferability for prokaryotes, as evidenced in cross-species predictions. These observations are consistent with the hypothesis that this stems from the homology of their modification systems. However, some eukaryotic species exhibited global transfer stability, possibly reflecting their global DNA methylation pattern. Furthermore, cross-group *k*-mer spectrum analysis (Supplemental Methods) revealed an evolutionary divergence: Prokaryotes maintain the functional stability of the RM system through conserved short *k*-mers, whereas eukaryotes achieve epigenetic plasticity through diverse sequence combinations. Additionally, Supplemental Figure S7 shows vertically compact, horizontally dispersed warm-colored regions, revealing two evolutionary insights. First, multiple significant *k*-mers at a specific *k*-value indicate the strong selection pressure at that motif length. For example, the proportion of warm colors at *k*=3 was higher than that at other *k*-values, implying that the scale conservation of 4mC sequences was strong under 3-mer. Second, fluctuations in significance levels as *k*-values vary reveal length-dependent selection. For the *k*-mer sequences of "AAA," the significance level gradually decreased when the *k*-value increased from three to five, indicating the specificity of 4mC sequences in terms of length.

We next examined *k*-mer enrichment differences between prokaryotes and eukaryotes (Supplemental Fig. S8). At the *k*=3 scale, prokaryotes exhibited highly specific *k*-mer enrichment, with motifs such as "GGC" and "CGG" showing significant positive differences. The strong conservation of these motifs in prokaryota suggests that their 4mC modification system may be under strict coevolutionary selection pressure to ensure the exact match between methylation sites and restriction sites in their host RM system (Vasu and Nagaraja 2013). This finding aligns with the observed intraclass transferability advantage, suggesting that conserved sequence features could enhance the cross-species prediction generalizability.

Unlike prokaryotes, eukaryotes showed dynamic sequence selectivity across different *k*-values. For example, in Supplemental Figure S7, the enrichment intensity of "AAA"

showed a gradient decline with the increase of  $k$ -value, suggesting that eukaryotes may achieve the improvement of regulatory accuracy through the expansion of sequence length. Notably, most of the significant  $k$ -mers showed a bidirectional distribution (e.g., “CGG” was enriched in prokaryotic but absent in eukaryotic at  $k = 3$ ). This polymorphism may reflect the adaptive requirement of 3D chromatin structure (Zhou et al. 2019) and could indicate that regions rich in repetitive sequences (such as poly(A)) are more likely to form open chromatin conformations, thus facilitating the accessibility of methyltransferases (Schübeler 2015).

Frequency difference polarization revealed a clear functional divergence between prokaryotes and eukaryotes: 83.3% of significant  $k$ -mers in prokaryotes showed unidirectional enrichment, whereas 93.3% of significant  $k$ -mers in eukaryotes showed background selectivity. These patterns are consistent with the hypothesis of different evolutionary selection pressures. Prokaryotes appear to be subject to the functional constraints of the host RM system to maintain the sequence conservation of methylation sites, whereas eukaryotes, potentially driven by developmental regulation, may dynamically modify heterogeneous DNA regions (such as gene spacer regions) to achieve spatiotemporal-specific epigenetic encoding (Suzuki and Bird 2008). Furthermore,  $k$ -mers frequently found in eukaryotic negative samples (e.g., “CCC,” “GGG,” “GCC,” etc.) potentially act as an “antimethylation” signal, analogous to the methylation suppression phenomenon of CpG islands in mammals (Deaton and Bird 2011).

Despite the robust cross-species predictive performance of EnDeep4mC, our approach is subject to certain limitations. Specifically, the inference of methylation profiles across species inherently assumes conserved regulatory mechanisms mediated by methyltransferases. In prokaryotes, in which methylation is often governed by RM systems with highly specific sequence motifs, the absence of prior knowledge regarding the active methyltransferase repertoire in a target species may constrain prediction accuracy. Our cross-species validation, while demonstrating promising transferability within microbial groups, also underscores the necessity of phylogenetic context when applying the model to distantly related organisms. For instance, although *E. coli* and *G. pickeringii* (both prokaryotes) exhibit strong intragroup predictability, extrapolation to uncharacterized prokaryotes without methyltransferase annotation remains challenging.

Looking forward, as genomic and epigenomic data sets continue to expand, covering increasingly diverse species and methylation systems, we anticipate that models like EnDeep4mC will gain further robustness and biological interpretability. The integration of methyltransferase annotation, structural motifs, and evolutionary context into future frameworks will be crucial to systematically explore the variation of modification patterns across species, ultimately uncovering deeper mechanisms of epigenetic evolution.

## Code availability

An integrated web server combining the ensemble framework and DFS module was implemented for public access at <http://lab.malab.cn/~lxm/EnDeep4mC>. The platform supports species selection and accepts DNA 4mC sequence inputs either through FASTA-formatted text entry or through file upload. Prediction results are generated as interactive web reports or downloadable text documents, providing systematic analysis of 4mC modifications. The server architecture ensures user-friendly operation while maintaining the computational accuracy demonstrated in our methodology. The

source code and data sets are publicly available at GitHub (<https://github.com/RaySYZhang/EnDeep4mC>) (release v1.0), with detailed implementation and usage described in the [Supplemental Methods](#). EnDeep4mC source code is also available as [Supplemental Code](#).

## Competing interest statement

The authors declare no competing interests.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 62450002 to Q.Z.; no. 62371347, no. 62573090 to X.L.).

*Author contributions:* S.Z. and X.L. conceived the study. X.L. designed the model. S.Z. implemented the model, performed the experiments, and analyzed the data. Q.Z. and X.L. supervised the project. S.Z. drafted the manuscript with input from all authors. X.L. led the revision of the manuscript. A.S. polished the language and improved the clarity of the manuscript. All authors reviewed and approved the final manuscript.

## References

- Beaulaurier J, Schadt EE, Fang G. 2019. Deciphering bacterial epigenomes using modern sequencing technologies. *Nat Rev Genet* **20**: 157–172. doi:10.1038/s41576-018-0081-3
- Bonet J, Chen MD, Dabad M, Heath S, Gonzalez-Perez A, Lopez-Bigas N, Lagergren J. 2022. DeepMP: a deep learning tool to detect DNA base modifications on nanopore sequencing data. *Bioinformatics* **38**: 1235–1243. doi:10.1093/bioinformatics/btab745
- Chen W, Yang H, Feng PM, Ding H, Lin H. 2017. iDNA4mC: identifying DNA N<sup>4</sup>-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* **33**: 3518–3523. doi:10.1093/bioinformatics/btx479
- Chen Z, Zhao P, Li FY, Marquez-Lago TT, Leier A, Revote J, Zhu Y, Powell DR, Akutsu T, Webb G, et al. 2020. *llearn*: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform* **21**: 1047–1057. doi:10.1093/bib/bbz041
- Chen Z, Ni P, Wang JX. 2025. Identifying DNA methylation types and methylated base positions from bacteria using nanopore sequencing with multi-scale neural network. *Bioinformatics* **41**: btaf397. doi:10.1093/bioinformatics/btaf397
- Cheng X. 1995. DNA modification by methyltransferases. *Curr Opin Struct Biol* **5**: 4–10. doi:10.1016/0959-440x(95)80003-j
- Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. *Genes Dev* **25**: 1010–1022. doi:10.1101/gad.2037511
- Dietterich TG. 2000. Ensemble methods in machine learning. In *Multiple classifier systems* (ed. Kittler J, Roli F), Vol. 1857, pp. 1–15. Springer-Verlag, Berlin. doi:10.1007/3-540-45014-9\_1
- Feng SH, Cokus SJ, Zhang XY, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci* **107**: 8689–8694. doi:10.1073/pnas.1002720107
- Feng PM, Ding H, Chen W, Lin H. 2016. Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions. *Mol Biosyst* **12**: 3307–3311. doi:10.1039/c6mb00471g
- Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**: 461–472. doi:10.1038/nmeth.1459
- Galeone V, Dabernig-Heinz J, Lohde M, Brandt C, Kohler C, Wagner GE, Hölzer M. 2025. Decoding bacterial methylomes in four public health-relevant microbial species: nanopore sequencing enables reproducible analysis of DNA modifications. *BMC Genomics* **26**: 394. doi:10.1186/s12864-025-11592-z
- He WY, Jia CZ, Zou Q. 2019. 4mCPred: machine learning methods for DNA N<sup>4</sup>-methylcytosine sites prediction. *Bioinformatics* **35**: 593–601. doi:10.1093/bioinformatics/bty668
- Jin QG, Meng ZP, Pham TD, Chen Q, Wei LY, Su R. 2019. DUNet: a deformable network for retinal vessel segmentation. *Knowledge-Based Syst* **178**: 149–162. doi:10.1016/j.knsys.2019.04.025

- Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**: 484–492. doi:10.1038/nrg3230
- Jones PA, Takai D. 2001. The role of DNA methylation in mammalian epigenetics. *Science* **293**: 1068–1070. doi:10.1126/science.1063852
- Khanal J, Nazari I, Tayara H, Chong KT. 2019. 4mCCNN: identification of N4-methylcytosine sites in prokaryotes using convolutional neural network. *IEEE Access* **7**: 145455–145461. doi:10.1109/access.2019.2943169
- Kuncheva LI, Whitaker CJ. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach Learn* **51**: 181–207. doi:10.1023/a:1022859003006
- Li F, Liu S, Li KW, Zhang YQ, Duan MY, Yao ZM, Zhu GC, Guo YT, Wang Y, Huang L, et al. 2023. EpiTEAmDNA: sequence feature representation via transfer learning and ensemble learning for identifying multiple DNA epigenetic modification types across species. *Comput Biol Med* **160**: 107030. doi:10.1016/j.combiomed.2023.107030
- Liang Y, Wu YA, Zhang ZQ, Liu NN, Peng J, Tang JJ. 2022. Hyb4mC: a hybrid DNA2vec-based model for DNA N4-methylcytosine sites prediction. *BMC Bioinformatics* **23**: 258. doi:10.1186/s12859-022-04789-6
- Liu Q, Fang L, Yu GL, Wang DP, Xiao CL, Wang K. 2019. Detection of DNA base modifications by deep recurrent neural network on Oxford nanopore sequencing data. *Nat Commun* **10**: 2449. doi:10.1038/s41467-019-10168-2
- Liu QZ, Chen JX, Wang YZ, Li SQ, Jia CZ, Song JN, Li FY. 2021a. DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. *Brief Bioinform* **22**: bbaa124. doi:10.1093/bib/bbaa124
- Liu Y, Rosikiewicz W, Pan ZW, Jillette N, Wang P, Taghbalout A, Foox J, Mason C, Carroll M, Cheng A, et al. 2021b. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biol* **22**: 295. doi:10.1186/s13059-021-02510-z
- Liu CT, Song JN, Ogata H, Akutsu T. 2022. MSNet-4mC: learning effective multi-scale representations for identifying DNA N4-methylcytosine sites. *Bioinformatics* **38**: 5160–5167. doi:10.1093/bioinformatics/btac671
- Lv H, Dao FY, Zhang D, Guan ZX, Yang H, Su W, Liu ML, Ding H, Chen W, Lin H. 2020. iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. *iScience* **23**: 100991. doi:10.1016/j.isci.2020.100991
- Manavalan B, Basith S, Shin TH, Wei LY, Lee G. 2019. Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol Ther Nucl Acids* **16**: 733–744. doi:10.1016/j.omtn.2019.04.019
- Moore LD, Le T, Fan GP. 2013. DNA methylation and its basic function. *Neuropsychopharmacology* **38**: 23–38. doi:10.1038/npp.2012.112
- O’Brown ZK, Boulias K, Wang J, Wang SY, O’Brown NM, Hao ZY, Shibuya H, Fady PE, Shi Y, He C, et al. 2019. Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC Genomics* **20**: 445. doi:10.1186/s12864-019-5754-6
- Razin A, Riggs AD. 1980. DNA methylation and gene function. *Science* **210**: 604–610. doi:10.1126/science.6254144
- Robertson KD. 2005. DNA methylation and human disease. *Nat Rev Genet* **6**: 597–610. doi:10.1038/nrg1655
- Schübeler D. 2015. Function and information content of DNA methylation. *Nature* **517**: 321–326. doi:10.1038/nature14192
- Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**: 465–476. doi:10.1038/nrg2341
- Vandenbussche I, Sass A, Van Nieuwerburgh F, Pinto-Carbó M, Mannweiler O, Eberl L, Coenye T. 2021. Detection of cytosine methylation in *Burkholderia cenocepacia* by single-molecule real-time sequencing and whole-genome bisulfite sequencing. *Microbiology (Reading)* **167**: 001027. doi:10.1099/mic.0.001027
- Vasu K, Nagaraja V. 2013. Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol Mol Biol Rev* **77**: 53–72. doi:10.1128/mmb.00044-12
- Wang YH, Zhao Y, Bollas A, Wang YR, Au KF. 2021. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**: 1348–1365. doi:10.1038/s41587-021-01108-x
- Wang X, Du Q, Wang R. 2024a. Mus4mCPred: accurate identification of DNA N4-methylcytosine sites in mouse genome using multi-view feature learning and deep hybrid network. *Processes* **12**: 1129. doi:10.3390/pr12061129
- Wang YZ, Zhai YX, Ding YJ, Zou Q. 2024b. SBSM-Pro: support bio-sequence machine for proteins. *Sci China-Inf Sci* **67**: 212106. doi:10.1007/s11432-024-4171-9
- Wei LY, Luan S, Nagai LAE, Su R, Zou Q. 2019a. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* **35**: 1326–1333. doi:10.1093/bioinformatics/bty824
- Wei LY, Su R, Luan SS, Liao ZJ, Manavalan B, Zou Q, Shi XL. 2019b. Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics* **35**: 4930–4937. doi:10.1093/bioinformatics/btz408
- Xiong J, Wang P, Shao WX, Li GJ, Ding JH, Xie NB, Wang M, Cheng QY, Xie CH, Feng YQ, et al. 2022. Genome-wide mapping of N<sup>4</sup>-methylcytosine at single-base resolution by APOBEC3A-mediated deamination sequencing. *Chem Sci* **13**: 9960–9972. doi:10.1039/d2sc02446b
- Xu HD, Jia PL, Zhao ZM. 2021. Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. *Brief Bioinform* **22**: bbaa099. doi:10.1093/bib/bbaa099
- Yao ZM, Li F, Xie WM, Chen JM, Wu JZ, Zhan Y, Wu XD, Wang ZG, Zhang GX. 2024. DeepSF-4mC: a deep learning model for predicting DNA cytosine 4mC methylation sites leveraging sequence features. *Comput Biol Med* **171**: 108166. doi:10.1016/j.combiomed.2024.108166
- Ye PH, Luan YZ, Chen KN, Liu YZ, Xiao CL, Xie Z. 2017. MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic Acids Res* **45**: D85–D89. doi:10.1093/nar/gkw950
- Yu M, Ji LX, Neumann DA, Chung DH, Groom J, Westpheling J, He C, Schmitz RJ. 2015. Base-resolution detection of N<sup>4</sup>-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite-sequencing. *Nucleic Acids Res* **43**: gkv738. doi:10.1093/nar/gkv738
- Yu YY, He WJ, Jin JR, Xiao GB, Cui LZ, Zeng R, Wei LY. 2021. iDNA-ABT: advanced deep learning model for detecting DNA methylation with adaptive features and transductive information maximization. *Bioinformatics* **37**: 4603–4610. doi:10.1093/bioinformatics/btab677
- Yu X, Ren J, Cui YN, Zeng R, Long HX, Ma CH. 2023. DRSN4mCPred: accurately predicting sites of DNA N4-methylcytosine using deep residual shrinkage network for diagnosis and treatment of gastrointestinal cancer in the precision medicine era. *Front Med* **10**: 1187430. doi:10.3389/fmed.2023.1187430
- Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**: 916–919. doi:10.1126/science.1186366
- Zeng R, Liao M. 2020. Developing a multi-layer deep learning based predictive model to identify DNA N4-methylcytosine modifications. *Front Bioeng Biotechnol* **8**: 274. doi:10.3389/fbioe.2020.00274
- Zheng PJ, Zhang GY, Liu YW, Huang GH. 2023. MultiScale-CNN-4mCPred: a multi-scale CNN and adaptive embedding-based method for mouse genome DNA N4-methylcytosine prediction. *BMC Bioinformatics* **24**: 21. doi:10.1186/s12859-023-05135-0
- Zhou Y, Zeng P, Li YH, Zhang ZD, Cui QH. 2016. SRAMP: prediction of mammalian N<sup>6</sup>-methyladenosine (m<sup>6</sup>A) sites based on sequence-derived features. *Nucleic Acids Res* **44**: e91. doi:10.1093/nar/gkw104
- Zhou KD, Gaullier G, Luger K. 2019. Nucleosome structure and dynamics are coming of age. *Nat Struct Mol Biol* **26**: 3–13. doi:10.1038/s41594-018-0166-x
- Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2007. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genet* **39**: 61–69. doi:10.1038/ng1929

Received June 8, 2025; accepted in revised form January 8, 2026.