Computers in Biology and Medicine 171 (2024) 108129

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

PSAC-6mA: 6mA site identifier using self-attention capsule network based on sequence-positioning

Zheyu Zhou, Cuilin Xiao, Jinfen Yin, Jiayi She, Hao Duan, Chunling Liu, Xiuhao Fu, Feifei Cui, Qi Qi, Zilong Zhang *

School of Computer Science and Technology, Hainan University, Haikou, 570228, China

ARTICLE INFO

Keywords:
Deep learning
N6-methyladenine
Capsule network
Self-attention

ABSTRACT

DNA N6-methyladenine (6mA) modifications play a pivotal role in the regulation of growth, development, and diseases in organisms. As a significant epigenetic marker, 6mA modifications extensively participate in the intricate regulatory networks of the genome. Hence, gaining a profound understanding of how 6mA is intricately involved in these biological processes is imperative for deciphering the gene regulatory networks within organisms. In this study, we propose PSAC-6mA (Position-self-attention Capsule-6mA), a sequence-location-based self-attention capsule network. The positional layer in the model enables positional relationship extraction and independent parameter setting for each base position, avoiding parameter sharing inherent in convolutional approaches. Simultaneously, the self-attention capsule network enhances dimensionality, capturing correlation information between capsules and achieving exceptional results in feature extraction across multiple spatial dimensions within the model. Experimental results demonstrate the superior performance of PSAC-6mA in recognizing 6mA motifs across various species.

1. Introduction

N6-methyladenine (6mA) stands out as a pivotal epigenetic modification in DNA. Recent studies have shed light on its indispensable role in biological growth, development, and disease regulation [1,2]. This modification intricately intertwines with fundamental life processes, including gene expression [3], transcription [4], replication [5], DNA repair [6], and the cell cycle [7]. Attaining a profound understanding of these biological phenomena relies heavily on the precise identification of 6mA modification sites. Hence, the significance of accurate 6mA identification in the realm of biology cannot be overstated.

Currently, various experimental techniques are utilized to detect 6mA loci, including liquid chromatography-mass spectrometry (LC-MS/MS) [8,9], single-molecule real-time sequencing (SMRT-seq) [10], as well as antibody-based immunoblotting (Immunoblotting) [11] and immunoprecipitation co-sequencing (DIP-seq) [12–15]. However, the sensitivity of these methods decreases when the amount of 6mA in the sample is low. Additionally, they are vulnerable to microbial contaminants (such as mycoplasma and bacteria) as well as DNA contamination within the sample. The interference of these factors poses a significant challenge to the accurate detection of 6mA. Moreover, these experimental approaches are not only time-consuming and labor-intensive but also financially burdensome. Therefore, the development of computerized methods for the precise detection of 6mA loci holds immense relevance and value [16–20].

In the realm of DNA methylation prediction, significant strides have been made through various deep learning models [21–23]. One such innovation, SNNRice6mA [24], was introduced by Yu et al., in 2019, tailored specifically for predicting 6mA sites within the rice genome. Despite its architectural simplicity utilizing convolutional neural networks (CNN) [25], the model's predictive accuracy remained susceptible to biases due to the limited size of the training dataset. In a parallel development, Li and colleagues devised the Deep6mA [26] framework in 2021, a hybrid deep learning network integrating CNN and long short-term memory (LSTM) [27] technologies. While this approach enhanced prediction accuracy, the model's computational complexity and specific parameter design posed constraints, potentially limiting its optimization potential. Addressing the challenge of species variability, Tsukiyama et al. introduced BERT6mA [28] in 2022, a deep learning model trained across 11 diverse species. Although this model, alongside Deep6mA, employed pre-training and fine-tuning methods, their effectiveness relied heavily on datasets exhibiting high sequence similarity. A

* Corresponding author.
E-mail address: zhangzilong@hainanu.edu.cn (Z. Zhang).

https://doi.org/10.1016/j.compbiomed.2024.108129
Received 19 December 2023; Received in revised form 6 February 2024; Accepted 6 February 2024
Available online 7 February 2024

Table 1
The sample distribution in the ten datasets used in the research.

Datasets	Training data		Testing data	
	positive	negative	positive	negative
A.thaliana	685289	685289	685246	685246
C.elegans	171181	171181	171138	171138
C.equisetifpolia	130417	130417	130417	130417
D.melanogaster	240626	240626	240583	240583
F.vesca	66691	66691	66691	66691
H.sapiens	394222	394222	394179	394179
R.chinensis	898	898	898	898
S.cerevisiae	81397	81397	81397	81397
T.thermophile	2313398	2313398	2313398	2313398
Xoc BLS256	370142	370142	370099	370101

recent innovation, CNN6mA [29] proposed in 2023, introduced a novel approach utilizing one-dimensional convolutional layers and cross-interaction networks for 6mA prediction. Despite these advancements, there remains a critical need for further refinement. Enhancing the accuracy of these models is imperative, necessitating continuous optimization efforts and the exploration of novel methodologies to bolster the precision and reliability of 6mA locus identification.

Mobeen Ur Rehman and colleagues introduced i6mA-Caps [30] in 2022 study, a model predicated on capsule network architectures designed for the recognition of DNA methylation. Nevertheless, the training process of capsule networks typically necessitates extensive datasets. Moreover, the inherent complexity of the dynamic routing mechanism in conventional capsule networks exacerbates the training challenges. This complexity consequently escalates the overall computational cost of the model. We introduce PSAC-6mA, a sequence localization self-attention capsule network. In this model, we replace the convolutional layer of the original capsule network with a Position linear Layer. We extract bottom-dimensional features of the sequence by employing the Position linear Layer with various window sizes, capturing not only the features but also the positional relationships between bases. Building upon the original capsule network [31], we incorporate the concept of self-attention capsules [32,33]. This involves dividing the capsules using one-dimensional group convolution and employing the self-attention mechanism to extract correlations between the capsules, resulting in high-dimensional features of the bases. Multi-dimensional feature extraction is achieved by conducting diverse dimension extractions through the Position linear Layer and capsule layer. This enables the multi-dimensional analysis of DNA sequences and captures positional relationships effectively. Consequently, PSAC-6mA surpasses existing state-of-the-art models in numerous species.

2. Materials and methods

2.1. Datasets

To ensure a fair comparison with existing tools, we utilize various datasets obtained from the web application of Lv et al. [34]. To demonstrate the model's ability to recognize 6mA across different species, we conduct experiments on our model using sequences from *A. thaliana*, *C.elegans*, *C.equisetifolia*, *D.melanogaster*, *F.vesca*, *H.sapiens*, *R. chinensis*, *S.cerevisiae*, *T.thermophile*, and *Xoc BLS256* (41dp) (containing both 6mA and non-6mA sequences, the exact number of samples is shown in Table 1). The dataset is available at the following http://lin-group.cn/server/iDNA-MS.

2.2. Neural network architecture

The neural network architecture comprises four key components: Coding Layer, Position linear Layers, PrimaryCap Layer and Routing Layer. To address the concern that self-attention might overlook the positional relationships of vectors, we replace the convolutional layer with the Position linear Layer to preserve these positional relationships. We propose a self-attention capsule network for sequence localization, as illustrated in Fig. 1. The input DNA sequence undergoes positional feature extraction through the Position linear Layer, allowing parameter updates through back-propagation for effective localization optimization. Simultaneously, the capsule layer and self-attention layer extract base correlations, enabling the determination of the presence or absence of 6mA modifications.

In the Encode layers, we utilize embedding in PyTorch to encode DNA sequences. Each base is mapped as follows: 'A' -> 0, 'C' -> 1, 'G' -> 2, 'T'-> 3. When encoding random input DNA sequences with a length of $L = 41$ and the number of features set to $F = 128$, we obtain a matrix X of size $L \times F$ ($X = [x_{i,j}](1 \le i \le L, 1 \le j \le F)$) . In the Position linear Layer, $W = w \times F(w = 3, 5, 7)$ F feature vectors are generated for each Position linear Layer, where the size of W represents the localization weights. Here, w denotes the window size for each layer. Each position is designated for feature extraction and localization of an individual base point. Location information is updated through backpropagation, with each parameter $x_{i,j}$ corresponding to a specific location within the generated matrix.

$$S = \left[s_{i,k}(W)\right] \tag{1}$$

$$s_{i,k}(W) = \sum_{l=1}^{W} \sum_{j=1}^{F} f_{i,k}(W)_{l,j} \otimes x_{i-W/2+l-1,j} \tag{2}$$

Following the linear function, we apply the ReLU activation function and introduce a dropout mechanism to randomly mask neurons, mitigating the risk of overfitting. The dropout rate is set to 0.5. Subsequently, three linear layers are employed, each with window sizes of 3, 5 and 7, respectively, to capture features from different positions and sizes.

$$n_{i,k}(W) = Dropout\left(\mathrm{ReLU}\left(s_{i,k}(W)\right)\right) \tag{3}$$

We use the visualization of position-attention relationships in DNA sequences in the 1D SENet [35,36] module, where we first input data via Position linear Layer into 1D SENet. The model captures the global relationships between features through operations within the SE block and assigns the appropriate weights to each base pair to realize the representation of attention. The core of this process is to visualize the intensity of the attention of each methylation site, which is shown by the change in the color of the heat map. In this way, we are able to visually identify the methylation regions that the model regards as important, providing guidance for subsequent biological analysis.

In the PrimaryCaps Layer and Routing Layer, we use one-dimensional grouped convolution to realize the segmentation of capsules and increase the dimension to raise the two-dimensional matrix to three bits, to realize the capsule segmentation of sequences. At the same time, we set up a self-attention layer at the capsule level to obtain the relationship between capsules, and to realize the extraction of high-dimensional features. We use Einstein summation to express the formula, where I represents the input matrix, W represents the and in the self-attention mechanism, where u denotes the summation in the self-attention mechanism, and get the correlation coefficient.

$$u_{k,j,i} = \sum_{j=1}^{j} \sum_{i=1}^{i} I_{\ldots j,i} \times \overline{W}_{k,j,i} \tag{4}$$

$$c_{\ldots j} = \frac{1}{F} \sum_{j=1}^{j} \left(u_{k,j,i}\right)^2 \tag{5}$$

$$S_{\ldots i} = \sum_{j=1}^{j} \left(\frac{e^{c_{\ldots j}}}{\sum_{j=1}^{j} e^{c_{\ldots j}}} + b \right) \times u_{k,i,j} \tag{6}$$

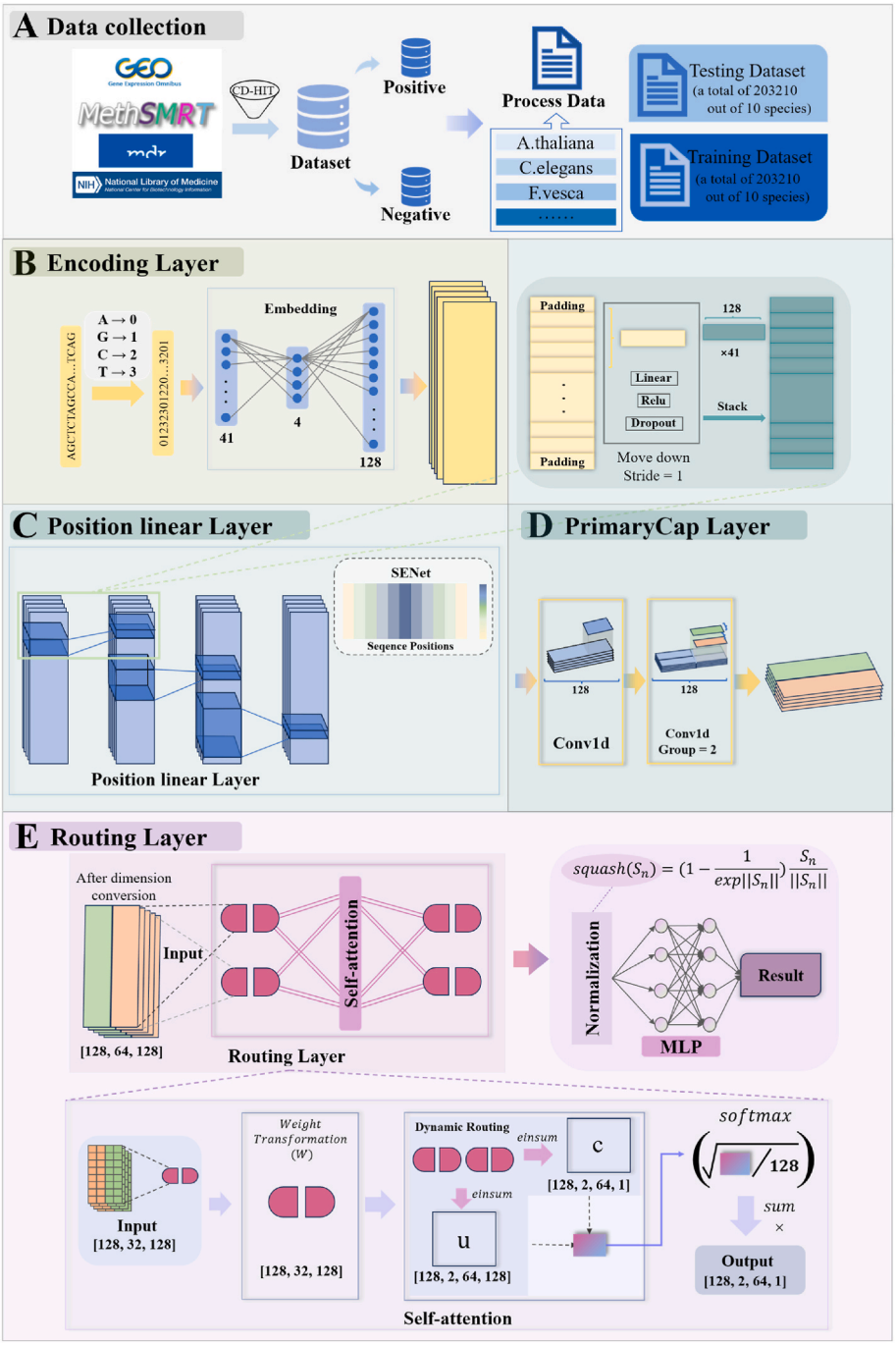

Fig. 1. The flowchart of PSAC-6mA. A. Data Collection: Data are sourced from databases such as GEO and categorized into positive and negative samples to form the training and validation sets. B. Encoding Layer: The input sequence is mapped and encoded using the embedding function from the torch library. C. Position linear Layer: The sequence is processed through three linear layers with window sizes of 3, 5 and 7, respectively. Linear layers and activation functions like ReLU extract features from the sequence while preserving the stable positional relationships within the DNA sequence. D. PrimaryCap Layer: One-dimensional grouped convolution is applied to segment capsules. E. Routing layer: The self-attention mechanism is employed to determine the relevance between capsules, compressing the inputs via a squash function. The sequence is then fed into a Multi-Layer Perceptron (MLP) to obtain the final output.

Following this, we employ the Squash function to compress and activate each capsule.

$$squash(S_n) = \left(1 - \frac{1}{exp||S_n||}\right) \frac{S_n}{||S_n||} \tag{7}$$

In the design of the multilayer perceptron (MLP [37]), we employ three fully connected layers with output sizes of 64, 32 and 1, respectively. Following the output of each fully connected layer, we introduce a Batch Normalization layer to normalize the output matrix, enhancing network stability and training. Subsequently, the ReLU activation function is applied to nonlinearly map the normalized output, introducing nonlinear features to enhance the network's expressive power. Between layers, we incorporate a Dropout layer with a dropout rate set to 0.3, mitigating network overfitting and enhancing its generalization ability. Finally, in the Output layer, we select the maximum value as the final probabilistic output to obtain the ultimate prediction result.

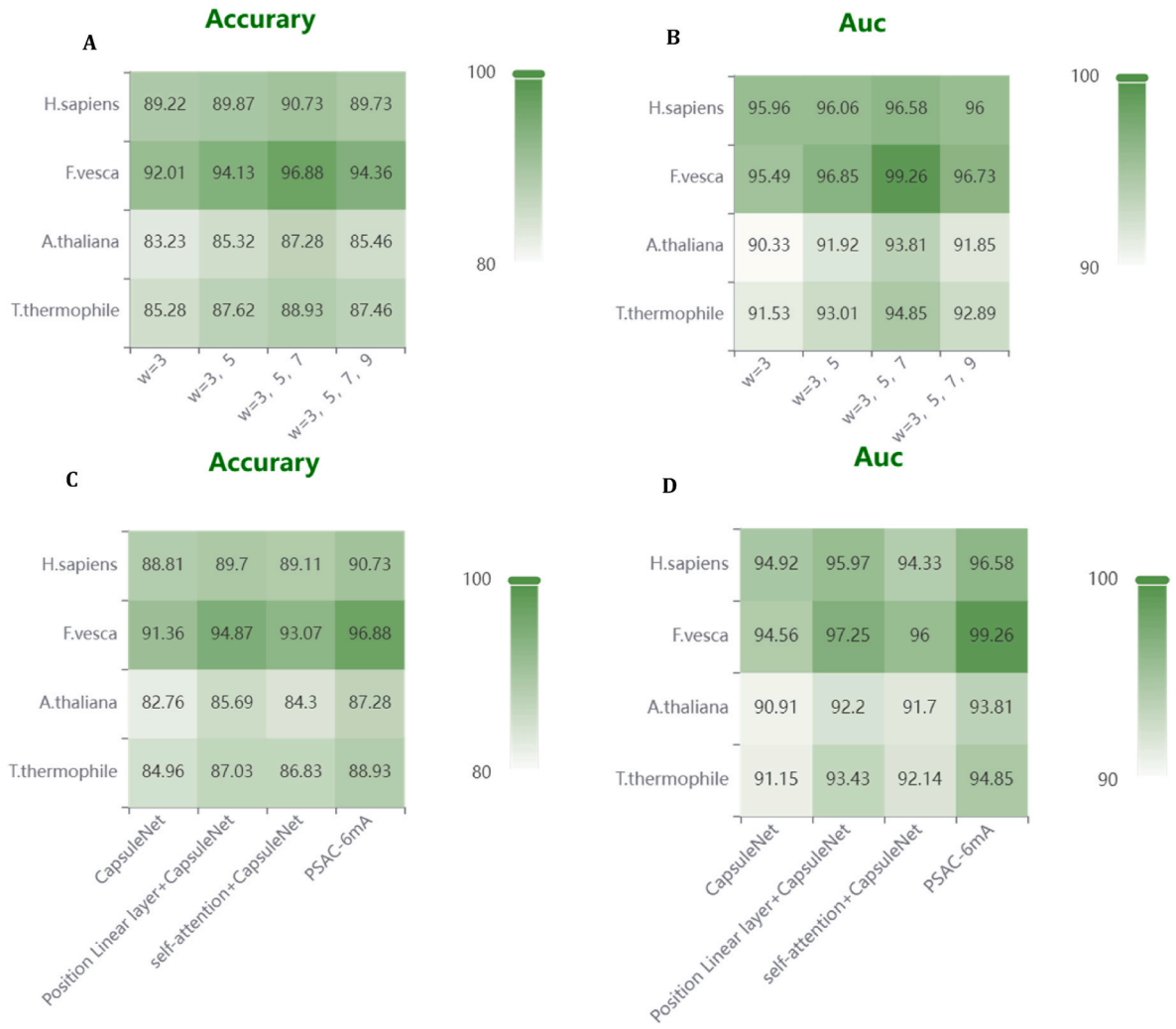

Fig. 2. Comparison of performance between different configurations of capsule neural network models and multi-scale filter combinations. A and B: The impact of multi-scale filter combinations on the performance of PSAC-6mA across various datasets is assessed. A range of filters with varying window sizes is designed, including a single linear filter with a window size of 3, a combination of two linear filters with window sizes of 3 and 5, a combination of three linear filters with window sizes of 3, 5, and 7, and a combination of four linear filters with window sizes of 3, 5, 7, and 9. The performance of PSAC-6mA on different datasets is evaluated in light of these combinations; C and D: The influence of different configurations of capsule neural network models on the performance of PSAC-6mA is evaluated. Specifically, a comparative analysis is conducted between capsule networks with and without Position Linear Layers, as well as self-attention capsule networks with and without enhanced location layers.

2.3. Training and evaluation

In our experiments, we employ a 5-fold cross-validation method [38–42] for training and predicting neural networks, and validate it on an independent test set. Throughout the training process, we utilize a small batch stochastic gradient descent optimization approach, setting the batch size (batch_size) to 128. The learning rate (lr) is configured to 1.0, and we employ the Adadelta optimization method [43] with a mean coefficient of 0.9 to optimize the model.

To evaluate the model's performance comprehensively and objectively, we employ four metrics [44–49]: Sn (sensitivity), Sp (specificity), ACC (accuracy), and MCC (Matthews correlation coefficient). These metrics provide a thorough assessment of the model's effectiveness.

$$\mathrm{SN} = \frac{TP}{TP + FN} \tag{8}$$

$$SP = \frac{TN}{TN + FP} \tag{9}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN) \times (TP + FP) \times (TN + FP) \times (TP + FN)}} \tag{11}$$

Sn represents the model's classification effectiveness on positive samples, while Sp denotes the model's classification effectiveness on negative samples. ACC signifies the prediction accuracy of the samples, and MCC reflects the overall performance of the model. In these metrics, TP stands for the number of correctly predicted true sites, TN for the number of correctly identified non-sites, FP for the number of predicted true non-sites, and FN represents the number of true sites predicted as non-sites. Positive and negative thresholds for the samples are set to 0.5, and the final results are obtained by averaging the metrics from the five generated models.

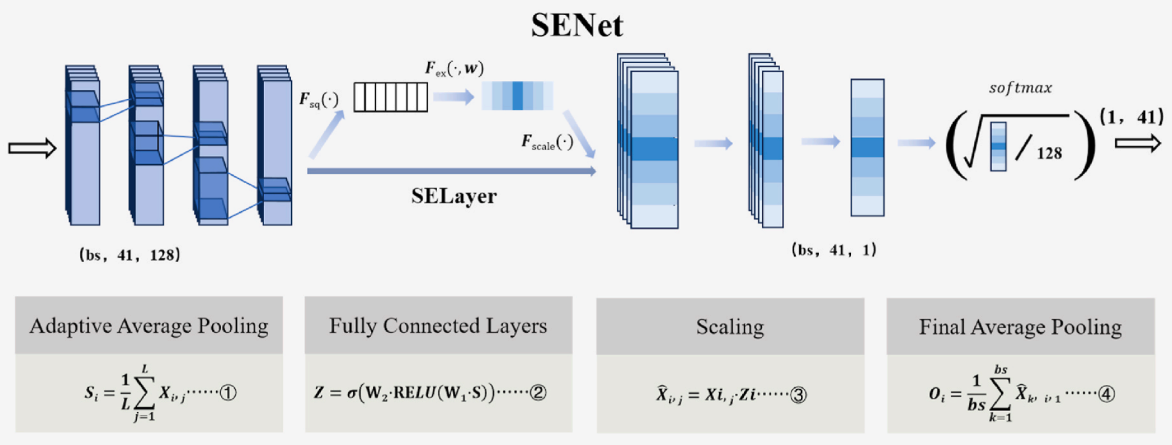

Fig. 3. The specific framework of SENet. The dimension of the given input tensor X is (bs,C,L), where bs is the batch size, C is the number of channels, and $L$is the length. ①Here S is the output of size (bs,C), where i is the channel index and j is the length index. ② $W_1 \in R^{C\times(\frac{C}{r})}$, and $W_2 \in R^{(\frac{C}{r})\times C}$ (This means that W_1 is a matrix that reduces the dimensionality from a C-dimensional space to a $\frac{C}{r}$ -dimensional space, while W_2 is a matrix that restores the dimensionality from a $\frac{C}{r}$ -dimensional space back to a C- dimensional space), both are the weights of the fully connected layer, r is the dimensionality reduction coefficient, σ is the sigmoid activation function, and ***ReLU*** is the linear rectification function. ③Here is the scaled output, which is obtained by multiplying it element by element Z with the original input X. The features of the base pairs were subjected to average pooling to generate a heat map that illustrates the distribution of the model's focus on the base pairs.

3. Results and discussion

3.1. Impact of coding style on the model

In our model, we initially utilize the torch library's embedding for encoding. However, we observe that the model is unable to make predictions when switching to one-hot encoding. Based on this, we conduct a study on encoding methods. Our model employs linear layers instead of convolution to extract positional relationships. Using one-hot encoding results in a sparse matrix where each word is independent, potentially causing the model to lose significant semantic information, which in turn affects the transmission and learning of parameters in the linear layer. On the other hand, using embeddings for encoding generates a low-dimensional, dense vector representation, which can learn similarities between words and better capture positional relationship information. Furthermore, we find that the mapping parameters of the embedding have a significant impact on the training of the model.

In Table S1, we alter the 'embedding_dim' parameter and train the model using 5-fold cross-validation. It is observed that a too-small mapping space prevents the model from learning effectively, leading to underfitting. Conversely, an excessively large mapping space does not yield significant improvements in the model's performance, resulting in unnecessarily high training costs.

3.2. Ablation experiment

To assess the effectiveness of the Position linear Layer in the context of location-based positioning and Performance Improvement due to Self-Attention Mechanism, our study conducts a series of comprehensive experiments within its research framework. Utilizing four different capsule neural network models, our study applies a rigorous 5-fold cross-validation strategy across four distinct datasets: *Arabidopsis thaliana* (*A.thaliana*), *Thermus thermophilus* (*T.thermophilus*), *Homo sapiens* (*H.sapiens*), and *Fragaria vesca* (*F.vesca*). The evaluation involves a comparative analysis of performance metrics across various model architectures, including the area under the ROC curve (AUC) and accuracy (ACC).

As depicted in Fig. 2, Capsule neural networks, when augmented with Position linear Layers, exhibit remarkable efficacy across all four examined datasets. This highlights the critical role of Position linear Layers in base position localization tasks. These layers provide essential positional insights about the bases, thereby enhancing the model's ability to interpret spatial relationships and locational nuances within the input data more accurately. Additionally, the integration of a self-attentive Routing Layer enables capsule networks to analyze the data attributes from multiple perspectives, facilitating a multi-dimensional data interpretation. Consequently, this approach not only allows for precise model recognition but also significantly boosts the overall model performance.

3.3. Optimal combinations of multi-scale filters

To comprehensively examine the impact of various window sizes and Position linear Layers on the localization of base position features, a series of ablation experiments are conducted. In these experiments, combinations of linear filters with different window sizes and Position linear Layers are utilized, training and evaluating on four datasets: *Arabidopsis thaliana* (*A.thaliana*), *Thermus thermophilus* (*T.thermophilus*), *Homo sapiens* (*H.sapiens*), and *Fragaria vesca* (*F.vesca*). Specifically, we design multi-scale Position linear Layers, including linear filters with window sizes of 3, 5, and 7. Each model underwent 5-fold cross-validation to assess performance across different datasets.

The results from Fig. 2 collectively demonstrate that models equipped with linear filters of window sizes 3, 5, and 7, as well as multi-scale Position linear Layers, exhibit outstanding performance in the task of base position localization.

3.4. Position linear layers can extract position relations

We assess Position linear Layers' efficacy in discerning positional details within DNA sequences. To enrich our understanding of the Position linear Layers capabilities, we have integrated a one-dimensional Squeeze-and-Excitation Network (SENet), as shown in Fig. 3. This addition aims to bolster the analysis of Position linear Layers, particularly in enhancing the visualization of their sensitivity to specific locational nuances.

The kpLogo tool is utilized for the visual analysis of datasets from *Arabidopsis thaliana* (*A.thaliana*), *strawberry* (*F.vesca*), *human* (*H.sapiens*), and *Chinese rowan* (*R.chinensis*) [50]. This tool highlights the nucleotide preferences surrounding the 6mA sites. Analysis of these datasets revealed the distribution characteristics of nucleotide frequencies in the vicinity of the 6mA sites, where a higher frequency of specific letters

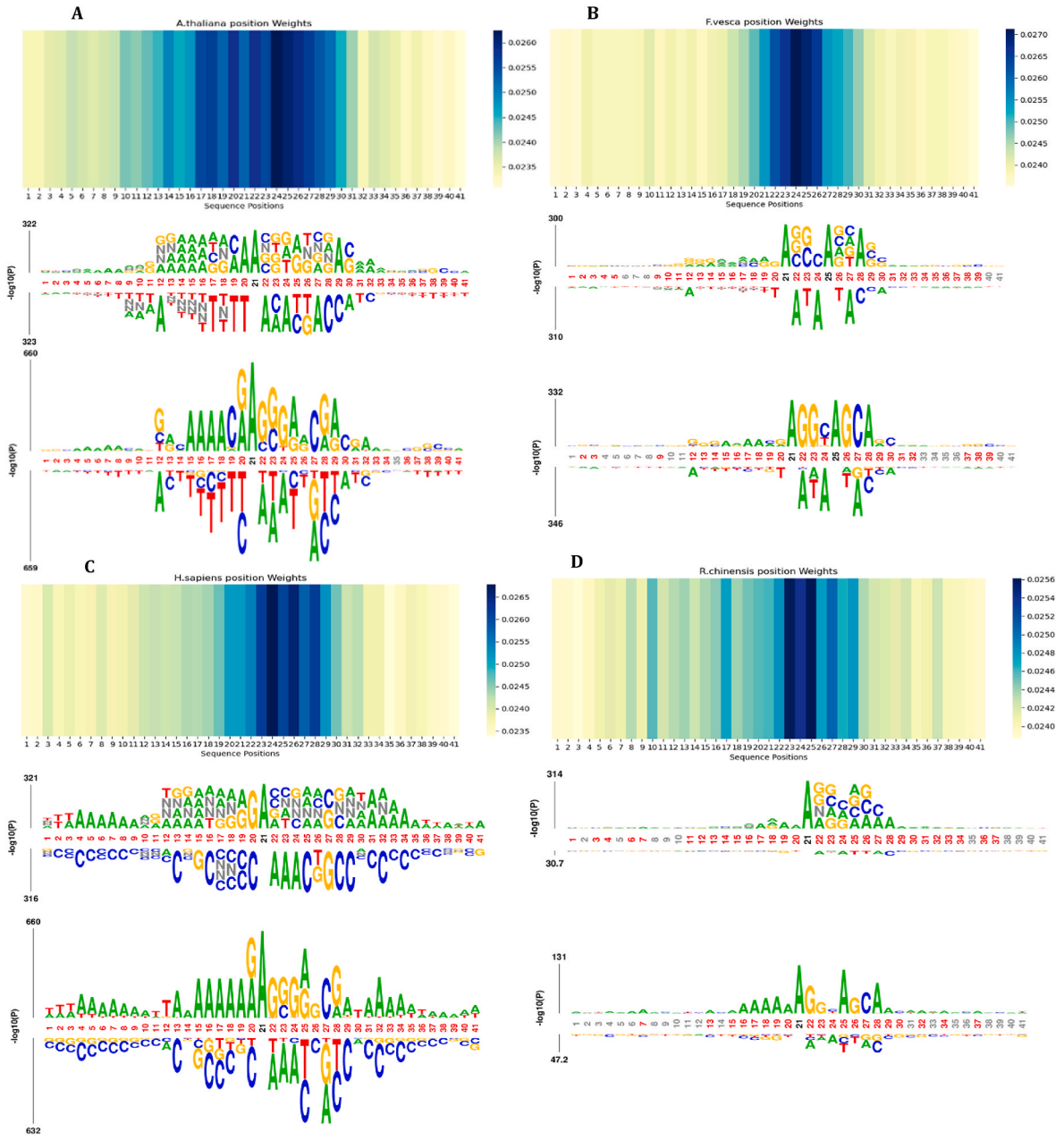

Fig. 4. Model Attention Maps and DNA Infographics for DNA Sequence Positions. The kpLogo tool is utilized for the visual analysis of datasets from *Arabidopsis thaliana* (*A. thaliana*), *strawberry* (*F. vesca*), *human* (*H. sapiens*), and *Chinese rowan* (*R. chinensis*). This tool highlights the nucleotide preferences surrounding the 6mA sites. Analysis of these datasets revealed the distribution characteristics of nucleotide frequencies in the vicinity of the 6mA sites, where a higher frequency of specific letters indicates a greater occurrence, reflecting lower information entropy and higher information content. Corresponding heatmaps underscore the focus of Position Linear layers on these critical positions, illustrating their targeted sensitivity.

indicates a greater occurrence, reflecting lower information entropy and higher information content. Corresponding heatmaps underscore the focus of Position linear Layers on these critical positions, illustrating their targeted sensitivity, as shown in Fig. 4.

DNA methylation, a key epigenetic modification, primarily occurs at the CpG dinucleotide sites, comprising cytosine (C) followed by guanine (G). In this study, we utilize the kpLogo tool for visualizing datasets from *Arabidopsis thaliana* (*A.thaliana*), *strawberry* (*F.vesca*), *human* (*H.sapiens*), and *Chinese rowan* (*R.chinensis*), highlighting the frequency and preference of nucleotides surrounding the 6mA sites. Furthermore, we introduce Position linear Layers in conjunction with one-dimensional Squeeze-and-Excitation Networks (SENet) to assess their capacity for capturing positional information. Comparing our findings with the outputs of kpLogo, we observe that Position linear Layers exhibit heightened sensitivity in detecting and characterizing CpG sites within DNA sequences. Notably, through SENet visualization, it is found that the focus of Position linear Layers progressively intensifies, thereby enhancing the recognition of potential methylation sites. This suggests that our model structure can effectively identify sequence features associated with methylation, providing more accurate information for predicting DNA methylation states.

3.5. Excellent positive and negative sample discrimination

We employ the popular feature analysis strategy, Uniform Manifold Approximation and Projection (UMAP) [51], for dimensionality reduction to analyze the distribution characteristics of samples. Training is conducted on three datasets: *Arabidopsis thaliana* (*A.thaliana*), *Thermus*

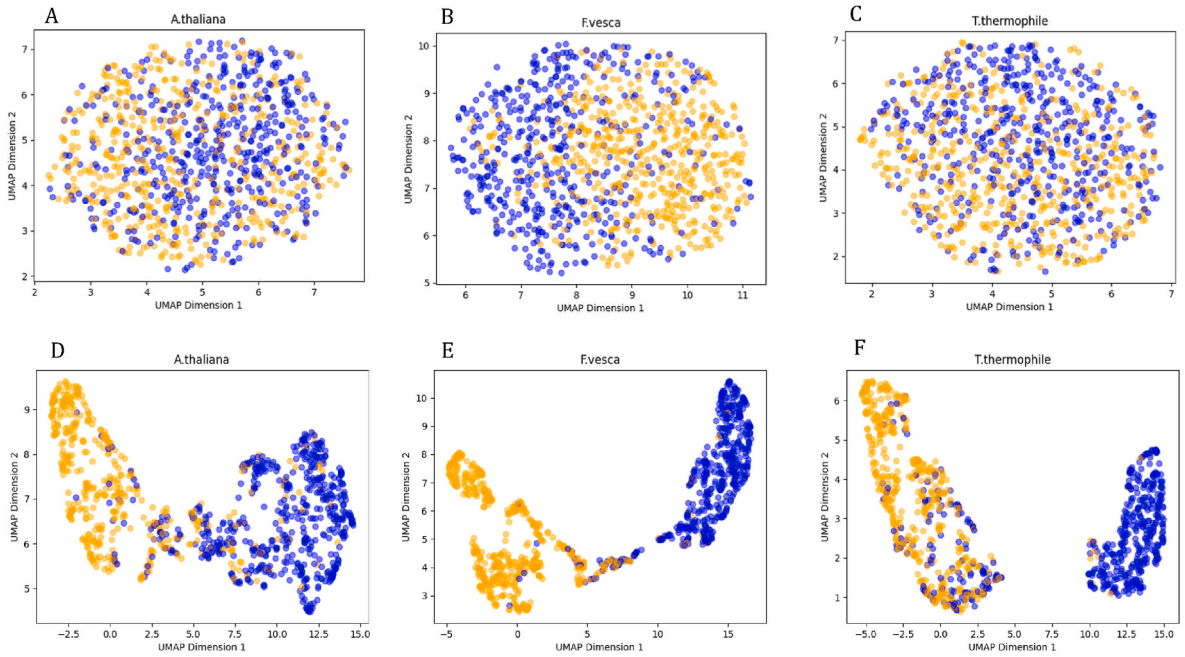

Fig. 5. The UMAP visualization of the feature extraction distribution. The figure presents the UMAP visualization of the feature extraction distribution for the *Arabidopsis thaliana* (*A.thaliana*), *Thermus thermophilus* (*T.thermophilus*), and *Fragaria vesca* (*F.vesca*) datasets. In the figure, orange dots represent positive samples, while blue dots signify negative samples. Panels A, B, and C show the distribution before training, whereas panels D, E, and F depict the distribution after training.

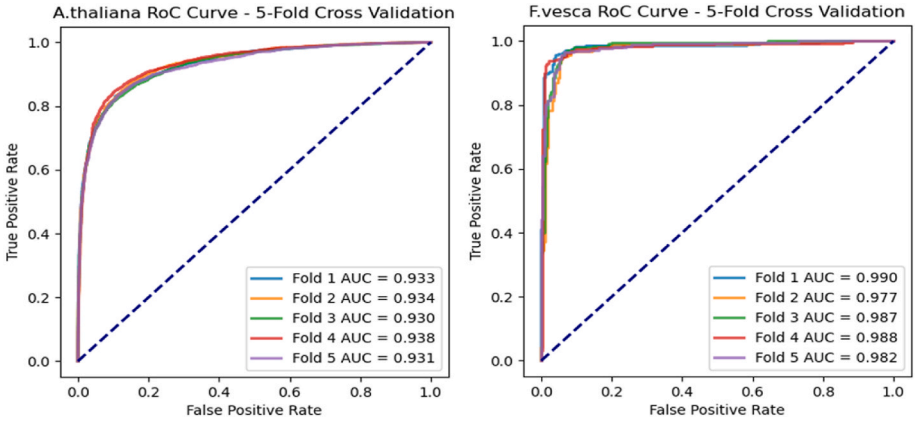

Fig. 6. ROC curves and their areas under *A.thaliana* and *F. vesca*.

thermophilus (*T.thermophilus*), and *Fragaria vesca* (*F.vesca*), each exhibiting distinctly different sample sizes. The selection of these datasets underscores the model's proficiency in performing well across datasets of varying sizes. Fig. 5 illustrates the feature distribution before and after training.

The figure clearly shows that for all three datasets, there is a distinct separation between positive and negative samples after training. Particularly in the case of the *Fragaria vesca* (*F.vesca*) dataset, there is almost no overlap between positive and negative samples, indicating that the information we extract is highly effective in distinguishing between methylated and unmethylated sequences. These observations suggest that our model performs well across different datasets, effectively differentiating between methylated and unmethylated sequences.

3.6. Model stability

Our model's performance is comprehensively evaluated through 5-fold cross-validation on 10 different biological datasets. During this process, two representative datasets, *Arabidopsis thaliana* (*A.thaliana*) and *Fragaria vesca* (*F.vesca*), are selected for detailed demonstration. In Fig. 6, The Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC) are employed as intuitive metrics to assess the model. On the *A.thaliana* dataset, the model exhibits consistently exceptional performance across five independent folds, with AUC values ranging from 0.930 to 0.938, highlighting its outstanding predictive capability. Similarly, on the *F.vesca* dataset, the model demonstrates even higher AUC values, ranging from 0.977 to 0.990, further affirming its exemplary performance and robustness. This profound evidence of the model's reliability adds significant credibility. Such a comprehensive and rigorous evaluation method lays a reliable foundation for our research, further supporting the practicality and application potential of our model.

3.7. Model demonstrates robust efficacy on datasets with limited size

In this study, we compare the performance differences in terms of accuracy (ACC) and area under the curve (AUC) between our model and other leading models listed in the table. Despite our model having a

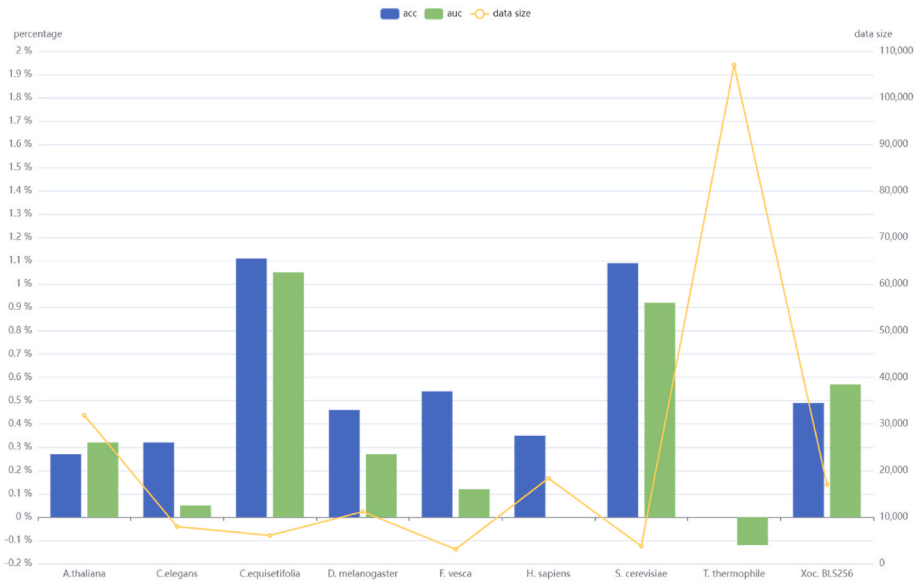

Fig. 7. Enhancement of our model on datasets of different data sizes. In the figure, the line graph represents the sample size of each dataset, with values indicated on the vertical axis to the right. The bar graph illustrates the performance enhancement of our model in terms of Accuracy (ACC) and Area Under the Receiver Operating Characteristic Curve (AUC) in comparison to the current state-of-the-art models.

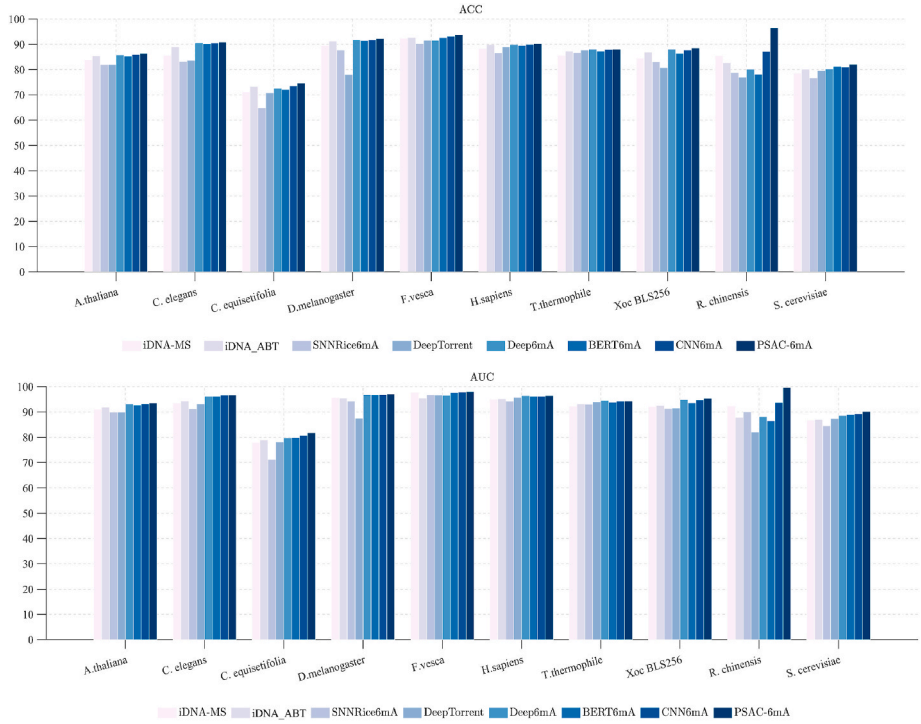

Fig. 8. The performance of our model across various datasets. The lower part of the figure shows experimental results and comparative analysis of PSAC-6mA against seven state-of-the-art models on an independent test set on data sets from 10 different species.

higher number of parameters, it demonstrates exceptional performance on smaller datasets when applying the same hyperparameter values across all prediction tasks. Our method is particularly notable for its learning effectiveness in smaller datasets.

Fig. 7 illustrates the performance of our model across various datasets. The improvements in the PSAC-6mA model for the ACC and AUC metrics are indicated on the left axis of the graph. The right vertical axis displays the sample sizes of the 10 different datasets, with a line graph representing the sample size for each dataset. Datasets with sample sizes below 10,000 are defined as small datasets. Notably, in the *R.chinensis* dataset, due to significant improvements of 9.28% in ACC and 5.99% in AUC, these data are omitted from the graph for clarity.

In most datasets, there is an enhancement in performance metrics (see Fig. 7). Specifically, in the *R.chinensis* dataset, the PSAC-6mA model achieves the highest improvements in ACC and AUC, at 9.28% and 5.99%, respectively. Significant gains are also observed in the *C.equisetifolia* and *S.cerevisiae* datasets. In *C.equisetifolia*, ACC and AUC improve by 1.11% and 1.05% respectively, while in *S.cerevisiae*, they increase by 0.79% and 0.92%. In terms of ACC, the PSAC-6mA either outperforms or matches other models in all datasets. Specifically, the

Table 2
Experimental results and comparative analysis of PSAC-6mA alongside seven state-of-the-art models from ten different species on test datasets.

Species	Model	SN(%)	SP(%)	ACC(%)	AUC(%)	Species	Model	SN(%)	SP(%)	ACC(%)	AUC(%)
A.thaliana	iDNA-MS	82.40	85.10	83.80	91.10	*H.sapiens*	iDNA-MS	86.30	90.50	88.40	95.00
	iDNA_ABT	82.30	88.40	85.40	91.80		iDNA_ABT	89.40	90.20	89.80	95.10
	SNNRice6mA	79.30	84.60	82.00	89.90		SNNRice6mA	86.20	87.10	86.60	94.10
	DeepTorrent	76.70	87.20	82.00	89.90		DeepTorrent	86.90	91.10	89.00	95.70
	Deep6mA	82.30	**89.40**	85.80	93.10		Deep6mA	**90.80**	88.80	89.80	**96.30**
	BERT6mA	**84.60**	85.90	85.30	92.70		BERT6mA	89.10	90.10	89.60	96.20
	CNN6mA	**84.60**	87.30	86.00	93.20		CNN6mA	89.60	90.20	89.90	96.20
	PSAC-6mA	83.50	89.04	**86.27**	**93.52**		PSAC-6mA	88.90	**91.60**	**90.25**	**96.30**
C.elegans	iDNA-MS	86.80	84.40	85.60	93.50	*R.chinensis*	iDNA-MS	88.00	82.90	85.50	92.40
	iDNA_ABT	88.20	89.90	89.00	94.30		iDNA-ABT	80.90	84.30	82.60	87.90
	SNNRice6mA	87.20	79.20	83.20	91.30		SNNRice6mA	77.20	80.40	78.80	90.00
	DeepTorrent	72.80	**94.60**	83.70	93.10		DeepTorrent	80.60	73.20	76.90	82.00
	Deep6mA	91.30	89.80	90.60	96.20		Deep6mA	76.90	83.30	80.10	88.20
	BERT6mA	90.80	89.50	90.20	96.20		BERT6mA	74.30	81.90	78.10	86.50
	CNN6mA	92.40	88.60	90.50	96.60		CNN6mA	89.80	84.70	87.20	93.70
	PSAC-6mA	**94.22**	90.99	**90.92**	**96.65**		PSAC-6mA	**94.98**	**97.99**	**96.48**	**99.69**
C.equisetifolia	iDNA-MS	71.80	70.50	71.10	77.90	*S.cerevisiae*	iDNA-MS	75.40	81.70	78.60	86.80
	iDNA_ABT	68.90	77.70	73.30	79.00		iDNA_ABT	72.40	**87.90**	80.10	87.10
	SNNRice6mA	64.50	65.10	64.80	71.20		SNNRice6mA	76.20	77.10	76.60	84.60
	DeepTorrent	59.00	82.30	70.70	78.10		DeepTorrent	76.40	82.90	79.60	87.30
	Deep6mA	**72.80**	72.30	72.60	79.80		Deep6mA	74.50	85.90	80.20	88.60
	BERT6mA	70.70	73.60	72.10	79.90		BERT6mA	80.10	82.50	81.30	89.00
	CNN6mA	69.60	77.40	73.50	80.70		CNN6mA	79.50	82.40	81.00	89.30
	PSAC-6mA	66.79	**82.42**	**74.61**	**81.75**		PSAC-6mA	**81.24**	82.93	**82.09**	**90.22**
D.melanogaster	iDNA-MS	89.00	90.30	89.60	95.60	*T.thermophile*	iDNA-MS	**95.80**	75.50	85.60	92.20
	iDNA_ABT	90.40	92.10	91.20	95.40		iDNA_ABT	93.30	81.50	87.40	93.10
	SNNRice6mA	87.10	88.30	87.70	94.20		SNNRice6mA	94.00	79.30	86.70	93.00
	DeepTorrent	63.00	**92.70**	77.90	87.50		DeepTorrent	93.90	81.60	87.70	94.00
	Deep6mA	92.20	91.30	91.80	96.80		Deep6mA	94.10	82.10	**88.10**	**94.40**
	BERT6mA	91.30	91.70	91.50	96.70		BERT6mA	92.50	**82.30**	87.40	93.80
	CNN6mA	91.50	92.10	91.80	96.80		CNN6mA	93.80	82.00	87.90	94.20
	PSAC-6mA	**92.92**	91.59	**92.26**	**97.07**		PSAC-6mA	95.53	80.66	**88.10**	94.28
F.vesca	iDNA-MS	**93.90**	90.60	92.30	97.70	*Xoc BLS256*	iDNA-MS	82.50	86.50	84.50	92.10
	iDNA_ABT	92.30	93.00	92.70	95.40		iDNA_ABT	88.90	84.90	86.90	92.60
	SNNRice6mA	91.40	89.20	90.30	96.70		SNNRice6mA	85.00	81.20	83.10	91.40
	DeepTorrent	90.30	92.80	91.60	96.60		DeepTorrent	**94.10**	66.70	80.70	91.50
	Deep6mA	92.00	91.20	91.60	96.50		Deep6mA	87.60	88.50	88.10	94.90
	BERT6mA	92.50	92.60	92.60	97.60		BERT6mA	84.80	87.80	86.30	93.60
	CNN6mA	93.80	92.60	93.20	97.80		CNN6mA	88.00	87.30	87.70	94.70
	PSAC-6mA	93.68	**93.81**	**93.74**	**97.92**		PSAC-6mA	86.74	**90.44**	**88.59**	**95.27**

Note: The bold face represents the highest achieved results.

largest and smallest improvements in ACC are observed in *R.chinensis* (9.28% increase) and *T.thermophile* (0% increase), respectively. For the AUC metric, the largest and smallest changes are in *R.chinensis* (5.99% increase) and *T.thermophile* (0.12% decrease). These results indicate that our model performs more excellently in smaller datasets.

3.8. Comparison of PSAC-6mA with the existing model

In this section, the proposed PSAC-6mA model is compared with seven other existing popular 6mA site prediction tools, including iDNA-MS, iDNA_ABT [52], SNNRice6mA, DeepTorrent [53], Deep6mA, BERT6mA, and CNN6mA. To comprehensively assess the learning and generalization capabilities of these models across different biological species, we evaluate these eight models (including PSAC-6mA) on datasets from 10 different species. Independent test data is used to ensure the reliability of the performance assessment. Detailed experimental results and comparative analysis are presented in Fig. 8 and Table 2.

From the results shown in Table 2, it is evident that PSAC-6mA consistently performs exceptionally well across all 10 datasets, achieving the highest AUC values. This superior performance can primarily be attributed to the unique design of PSAC-6mA, which includes innovative structures such as the positional layer and self-attention capsule layers. These design elements endow PSAC-6mA with enhanced robustness against noise and intra-sequence variations. Even in the presence of errors or mutations in the sequences, PSAC-6mA is still able to make relatively accurate predictions. Furthermore, the consistency and high performance of PSAC-6mA across multiple datasets further validate its excellent generalization ability. This demonstrates its effectiveness in handling biological sequence data from various sources with diverse characteristics while avoiding overfitting. Additionally, despite its complex network architecture, PSAC-6mA performs remarkably well even on smaller datasets. This can be mainly attributed to the precision in its structure and module design, allowing it to capture key pattern information in scenarios with limited data. For instance, the performance improvement of PSAC-6mA on the *R. chinensis* dataset is particularly noteworthy, especially in comparison with other models, including CNN6mA and iDNA-MS, which also show good performance on this dataset.

3.9. Web server implementation

To facilitate research in the fields of epigenetics and genomic analysis, we have developed the PSAC-6ma Web server application. You can freely access the application through the following link: www.bioai-lab.com/PSAC-6mA.

The main functionality of the application involves receiving DNA sequence files of length 41 bp, provided in txt format. The application will assess whether each sequence is methylated, aiding researchers in the identification of 6mA modifications. Detailed information about the application and usage instructions can be found on the website's help page. We encourage users to carefully read this page to ensure the correct usage and understanding of the application's features.

Please note that, for an enhanced user experience and to ensure the

smooth operation of the application, it is recommended to use the latest version of web browsers. We appreciate your attention and support, and we hope that this application will serve as a powerful tool for researchers in the field of Position-based 6mA identification. If you have any questions or suggestions, please feel free to contact us.

4. Conclusion

6mA modification serves as a crucial regulatory mechanism within cellular processes. Due to the inherent limitations in our understanding of 6mA sites, the precise detection of 6mA remains a challenging task. Our PSAC-6mA model adopts an innovative network architecture, specifically the self-attention capsule neural network for sequence localization, aiming at addressing key challenges in predicting 6mA sites. This architecture incorporates two pivotal components: the Position linear Layer and the Routing Layer. The Position linear Layer is designed to capture positional information regarding 6mA modifications within DNA sequences. By precisely locating and extracting the specific positions of each 6mA modification, this layer goes beyond merely analyzing gene sequences. Such a design helps maintain the stability of spatial relationships within DNA sequences, enabling the model to gain a more accurate understanding of the distribution of 6mA modifications within the genome. On the other hand, the Routing Layer allows the model to simultaneously consider relationships between different positions within the sequence. It accomplishes this by assigning varying attention weights to capture correlations between different 6mA modification sites. This empowers the model to comprehensively analyze and learn multi-dimensional features of the sequence. By cleverly combining the Position linear Layer and the Routing Layer, our model successfully extracts both the positions and spatial relationships of 6mA modifications while delving into the multi-layer features of the sequence. As demonstrated through extensive experiments on five-fold cross-validation and validation on an independent test set, PSAC-6mA has outperformed current state-of-the-art methods across multiple species. We believe that the successful application of this model not only provides a deeper and more accurate tool for biological research but also paves the way for new directions in related fields, offering robust support for future research in the life sciences.

Data availability

The source code for PSAC-6mA is freely available on the Zheyu1115/PSAC-6mA-DNA-6mA-Modification-Recognition-Model (github.com).

Funding

The work is supported by the National Natural Science Foundation of China (No. 62101100, No. 62262015).

CRediT authorship contribution statement

Zheyu Zhou: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Cuilin Xiao:** Writing – original draft, Visualization, Investigation, Formal analysis, Data curation. **Jinfen Yin:** Writing – original draft, Resources, Investigation. **Jiayi She:** Writing – original draft, Investigation, Conceptualization. **Hao Duan:** Writing – original draft, Investigation. **Chunling Liu:** Writing – original draft. **Xiuhao Fu:** Visualization. **Feifei Cui:** Supervision. **Qi Qi:** Supervision. **Zilong Zhang:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Zilong Zhang reports was provided by National Natural Science Foundation of China (No. 62101100, No. 62262015). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108129.

References

[1] W. Tang, S. Wan, Z. Yang, A.E. Teschendorff, Q. Zou, Tumor origin detection with tissue-specific miRNA and DNA methylation markers, Bioinformatics 34 (2017) 398–406.
[2] J. Jin, Y. Yu, R. Wang, X. Zeng, C. Pang, Y. Jiang, Z. Li, Y. Dai, R. Su, Q. Zou, K. Nakai, L. Wei, iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations, Genome Biol. 23 (2022) 219.
[3] D.A. Low, N.J. Weyand, M.J. Mahan, Roles of DNA adenine methylation in regulating bacterial gene expression and virulence, Infect. Immun. 69 (2001) 7197–7204.
[4] J.L. Robbins-Manke, Z.Z. Zdraveski, M. Marinus, J.M. Essigmann, Analysis of global gene expression and double-strand-break formation in DNA adenine methyltransferase- and mismatch repair-deficient Escherichia coli, J. Bacteriol. 187 (2005) 7027–7037.
[5] J.L. Campbell, N. Kleckner, E. coli, oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork, Cell 62 (1990) 967–979.
[6] K.G. Au, K. Welsh, P. Modrich, Initiation of methyl-directed mismatch repair, J. Biol. Chem. 267 (1992) 12142–12148.
[7] H. Lv, F.Y. Dao, D. Zhang, H. Yang, H. Lin, Advances in mapping the epigenetic modifications of 5-methylcytosine (5mC), N6-methyladenine (6mA), and N4-methylcytosine (4mC), Biotechnol. Bioeng. 118 (2021) 4204–4216.
[8] W. Huang, J. Xiong, Y. Yang, S.-M. Liu, B.-F. Yuan, Y.-Q. Feng, Determination of DNA adenine methylation in genomes of mammals and plants by liquid chromatography/mass spectrometry, RSC Adv. 5 (2015) 64046–64054.
[9] B. Liu, X. Liu, W. Lai, H. Wang, Metabolically generated stable isotope-labeled deoxynucleoside code for tracing DNA N(6)-methyladenine in human cells, Anal. Chem. 89 (2017) 6202–6209.
[10] B.A. Flusberg, D.R. Webster, J.H. Lee, K.J. Travers, E.C. Olivares, T.A. Clark, J. Korlach, S.W. Turner, Direct detection of DNA methylation during single-molecule, real-time sequencing, Nat. Methods 7 (2010) 461–465.
[11] D.I. Stott, Immunoblotting and dot blotting, J. Immunol. Methods 119 (1989) 153–187.
[12] Y. Fu, G.Z. Luo, K. Chen, X. Deng, M. Yu, D. Han, Z. Hao, J. Liu, X. Lu, L.C. Dore, X. Weng, Q. Ji, L. Mets, C. He, N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas, Cell 161 (2015) 879–892.
[13] E.L. Greer, M.A. Blanco, L. Gu, E. Sendinc, J. Liu, D. Aristizábal-Corrales, C.H. Hsu, L. Aravind, C. He, Y. Shi, DNA methylation on N6-adenine in C. elegans, Cell 161 (2015) 868–878.
[14] G. Zhang, H. Huang, D. Liu, Y. Cheng, X. Liu, W. Zhang, R. Yin, D. Zhang, P. Zhang, J. Liu, C. Li, B. Liu, Y. Luo, Y. Zhu, N. Zhang, S. He, C. He, H. Wang, D. Chen, N6-methyladenine DNA modification in Drosophila, Cell 161 (2015) 893–906.
[15] T.P. Wu, T. Wang, M.G. Seetin, Y. Lai, S. Zhu, K. Lin, Y. Liu, S.D. Byrum, S. G. Mackintosh, M. Zhong, A. Tackett, G. Wang, L.S. Hon, G. Fang, J.A. Swenberg, A.Z. Xiao, DNA methylation on N(6)-adenine in mammalian embryonic stem cells, Nature 532 (2016) 329–333.
[16] Y. Wang, Y. Zhai, Y. Ding, Q.J.a.e.-p. Zou, SBSM-pro: Support Bio-Sequence Machine for Proteins, 2023 arXiv:2308.10275.
[17] F. Cui, Z. Zhang, Q. Zou, Sequence representation approaches for sequence-based protein prediction tasks that use deep learning, Briefings in functional genomics 20 (2021) 61–73.
[18] C. Ao, X. Ye, T. Sakurai, Q. Zou, L. Yu, m5U-SVM: identification of RNA 5-methyluridine modification sites based on multi-view features of physicochemical features and distributed representation, BMC Biol. 21 (2023) 93.
[19] L. Wang, Y. Ding, P. Tiwari, J. Xu, W. Lu, K. Muhammad, V.H.C. de Albuquerquee, F. Guo, A deep multiple kernel learning-based higher-order fuzzy inference system for identifying DNA N4-methylcytosine sites, Inf. Sci. 630 (2023) 40–52.
[20] Q. Huang, J. Zhang, L. Wei, F. Guo, Q. Zou, 6mA-RicePred: a method for identifying DNA N 6-methyladenine sites in the rice genome based on feature fusion, Front. Plant Sci. 11 (2020) 4.
[21] C. Ao, S. Jiao, Y. Wang, L. Yu, Q. Zou, Biological sequence classification: a review on data and general methods, Research (2022), 2022:0011.
[22] J. Qiao, J. Jin, H. Yu, L. Wei, Towards retraining-free RNA modification prediction with incremental learning, Inf. Sci. 660 (2024) 120105.
[23] M. Liu, C. Li, R. Chen, D. Cao, X. Zeng, Geometric deep learning for drug discovery, Expert Syst. Appl. 240 (2024) 122498.
[24] H. Yu, Z. Dai, SNNRice6mA: a deep learning method for predicting DNA N6-methyladenine sites in rice genome, Front. Genet. 10 (2019) 1071.
[25] L.O. Chua, Cnn: a paradigm for complexity, in: J.L. HUERTAS, et al. (Eds.), Visions of Nonlinear Science in the 21st Century, World Scientific Publishing Co. Pte. Ltd, 1999, pp. 529–837. Published by.

[26] Z. Li, H. Jiang, L. Kong, Y. Chen, K. Lang, X. Fan, L. Zhang, C. Pian, Deep6mA: a deep learning framework for exploring similar patterns in DNA N6-methyladenine sites across different species, PLoS Comput. Biol. 17 (2021) e1008767.
[27] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: LSTM cells and network architectures, Neural Comput. 31 (2019) 1235–1270.
[28] S. Tsukiyama, M.M. Hasan, H.W. Deng, H. Kurata, BERT6mA: prediction of DNA N6-methyladenine site using deep learning-based approaches, Briefings Bioinf. 23 (2022).
[29] S. Tsukiyama, M.M. Hasan, H. Kurata, CNN6mA: interpretable neural network model based on position-specific CNN and cross-interactive network for 6mA site prediction, Comput. Struct. Biotechnol. J. 21 (2023) 644–654.
[30] M.U. Rehman, H. Tayara, Q. Zou, K.T. Chong, i6mA-Caps: a CapsuleNet-based framework for identifying DNA N6-methyladenine sites, Bioinformatics 38 (2022) 3885–3891.
[31] S. Sabour, N. Frosst, E. Hinton, Dynamic Routing between Capsules, 2017 09829 arXiv:1710.
[32] V. Mazzia, F. Salvetti, M. Chiaberge, Efficient-CapsNet: capsule network with self-attention routing, Sci. Rep. 11 (2021) 14634.
[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, Polosukhin, Attention Is All You Need, 2017 03762 arXiv:1706.
[34] H. Lv, F.Y. Dao, D. Zhang, Z.X. Guan, H. Yang, W. Su, M.L. Liu, H. Ding, W. Chen, H. Lin, iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes, iScience 23 (2020) 100991.
[35] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-Excitation Networks, 2017 01507 arXiv:1709.
[36] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-net: Efficient Channel Attention for Deep Convolutional Neural Networks, 2019 03151 arXiv:1910.
[37] H. Taud, J.F. Mas, Multilayer perceptron (MLP), in: M.T. Camacho Olmedo, M. Paegelow, J.-F. Mas, F. Escobar (Eds.), Geomatic Approaches for Modeling Land Change Scenarios, Springer International Publishing, Cham, 2018, pp. 451–455.
[38] H. Yang, Y.M. Luo, C.Y. Ma, T.Y. Zhang, T. Zhou, X.L. Ren, X.L. He, K.J. Deng, D. Yan, H. Tang, H. Lin, A gender specific risk assessment of coronary heart disease based on physical examination data, NPJ digital medicine 6 (2023) 136.
[39] F.Y. Dao, H. Lv, M.J. Fullwood, H. Lin, Accurate identification of DNA replication origin by fusing epigenomics and chromatin interaction information, Research 2022 (2022) 9780293.
[40] H. Li, Y. Pang, B. Liu, BioSeq-BLM: a platform for analyzing DNA, RNA, and protein sequences based on biological language models, Nucleic Acids Res. 49 (2021) e129.
[41] X. Fu, Y. Yuan, H. Qiu, H. Suo, Y. Song, A. Li, Y. Zhang, C. Xiao, Y. Li, L. Dou, Z. Zhang, F. Cui, AGF-PPIS: a protein–protein interaction site predictor based on an attention mechanism and graph convolutional networks, Methods 222 (2024) 142–151.
[42] H. Duan, Y. Zhang, H. Qiu, X. Fu, C. Liu, X. Zang, A. Xu, Z. Wu, X. Li, Q. Zhang, Z. Zhang, F. Cui, Machine learning-based prediction model for distant metastasis of breast cancer, Comput. Biol. Med. 169 (2024) 107943.
[43] M.D. Zeiler, ADADELTA: an Adaptive Learning Rate Method, 2012 arXiv: 1212.5701.
[44] H. Zhu, H. Hao, L. Yu, Identifying disease-related microbes based on multi-scale variational graph autoencoder embedding Wasserstein distance, BMC Biol. 21 (2023) 294.
[45] X. Zou, L. Ren, P. Cai, Y. Zhang, H. Ding, K. Deng, X. Yu, H. Lin, C. Huang, Accurately identifying hemagglutinin using sequence information and machine learning methods, Front. Med. 10 (2023) 1281880.
[46] W. Zhu, S.S. Yuan, J. Li, C.B. Huang, H. Lin, B. Liao, A first computational frame for recognizing heparin-binding protein, Diagnostics 13 (2023).
[47] Y. Tang, Y. Pang, B. Liu, IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning, Bioinformatics 36 (2021) 5177–5186.
[48] X. Zeng, F. Wang, Y. Luo, S.-g. Kang, J. Tang, F.C. Lightstone, E.F. Fang, W. Cornell, R. Nussinov, F.J.C.R.M. Cheng, Deep generative molecular design reshapes drug discovery, Cell Reports Medicine 4 (2022) 100794.
[49] R. Liu, Z. Zhang, X. Fu, S. Yan, F. Cui, AIPPT: predicts anti-inflammatory peptides using the most characteristic subset of bases and sequences by stacking ensemble learning strategies, IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2023) 23–29, 2023.
[50] X. Wu, D.P. Bartel, kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences, Nucleic Acids Res. 45 (2017) W534–w538.
[51] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2018 03426 arXiv:1802.
[52] Y. Yu, W. He, J. Jin, G. Xiao, L. Cui, R. Zeng, L. Wei, iDNA-ABT: advanced deep learning model for detecting DNA methylation with adaptive features and transductive information maximization, Bioinformatics 37 (2021) 4603–4610.
[53] Q. Liu, J. Chen, Y. Wang, S. Li, C. Jia, J. Song, F. Li, DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites, Briefings Bioinf. (2021) 22.