Computational and Structural Biotechnology Journal 23 (2024) 3175–3185

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Domain-knowledge enabled ensemble learning of 5-formylcytosine (f5C) modification sites

Jiaming Huang [a,c,1], Xuan Wang [c,1], Rong Xia [c,d,1], Dongqing Yang [b], Jian Liu [a], Qi Lv [a], Xiaoxuan Yu [e], Jia Meng [c,f,g], Kunqi Chen [h], Bowen Song [b,*], Yue Wang [a,*]

[a] Jiangsu Key Laboratory for Functional Substance of Chinese Medicine, School of Pharmacy, Nanjing University of Chinese Medicine, Nanjing 210023, China
[b] Department of Public Health, School of Medicine, Nanjing University of Chinese Medicine, Nanjing 210023, China
[c] Department of Biological Sciences, School of Science, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[d] School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[e] Department of Pharmacology, School of Medicine, Nanjing University of Chinese Medicine, Nanjing 210023, China
[f] AI University Research Centre, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[g] Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L7 8TX, United Kingdom
[h] Key Laboratory of Ministry of Education for Gastrointestinal Cancer, School of Basic Medical Sciences, Fujian Medical University, Fuzhou 350004, China

ARTICLE INFO

Keywords:
RNA modification
Ensemble learning
5-formylcytidine
Epitranscriptomic marks
Genomic features

ABSTRACT

5-formylcytidine (f^5C) is a unique post-transcriptional RNA modification found in mRNA and tRNA at the wobble site, playing a crucial role in mitochondrial protein synthesis and potentially contributing to the regulation of translation. Recent studies have unveiled that the f^5C modifications may drive mitochondrial mRNA translation to power cancer metastasis. Accurate identification of f^5C sites is essential for further unraveling their molecular functions and regulatory mechanisms, but there are currently no computational methods available for predicting their locations. In this study, we introduce an innovative ensemble approach, successfully enabling the computational recognition of *Saccharomyces cerevisiae* f^5C. We conducted a comprehensive model selection process that involved multiple basic machine learning and deep learning algorithms such as recurrent neural networks, convolutional neural networks and Transformer-based models. Initially trained only on sequence information, these individual models achieved an AUROC ranging from 0.7104 to 0.7492. Through the integration of 32 novel domain-derived genomic features, the performance of individual models has significantly improved to an AUROC between 0.7309 and 0.8076. To further enhance accuracy and robustness, we then constructed the ensembles of these individual models with different combinations. The best performance attained by our ensemble models reached an AUROC of 0.8391. Shapley additive explanations were conducted to explain the significant contributions of genomic features, providing insights into the putative distribution of f^5C across various topological regions and potentially paving the way for revealing their functional relevance within distinct genomic contexts. A freely accessible web server that allows real-time analysis of user-uploaded sites can be accessed at: www.rnamd.org/Resf5C-Pred.

1. Introduction

Chemical modification plays a vital role in controlling the function of biological macromolecules such as DNA, RNA and proteins [1]. To date, over 170 different RNA post-transcriptional modifications have been identified [2]. These modified residues have been identified in all RNA types, including mRNA, rRNA, tRNA, and snRNA [3]. Among these modifications, 5-formylcytidine (f^5C) was first observed at position 34 of mammalian mitochondrial methionine transfer RNA (mt-tRNAMet) [4]. The biosynthetic pathway of f^5C34 begins with the methylation of C34 by the NSUN RNA methyltransferase (NSUN2 or NSUN3), leading to the formation of 5-methylcytosine (m^5C). Following this, ALKBH1 hydroxylates m^5C to produce hydroxymethylcytosine (hm^5C), which is then further oxidized to generate f^5C [5,6].

f^5C plays a crucial role in mitochondrial protein synthesis and may contribute to the regulation of translation. For example, tRNA modified

* Corresponding authors.
E-mail addresses: bowen.song@njucm.edu.cn (B. Song), yue.wang@njucm.edu.cn (Y. Wang).
[1] Contributed equally to this work

https://doi.org/10.1016/j.csbj.2024.08.004
Received 11 May 2024; Received in revised form 7 August 2024; Accepted 7 August 2024
Available online 8 August 2024

with f^5C in the anticodon facilitates the translation of both AUA and AUG as methionine [7]. Besides its presence in tRNA, f^5C has also been identified in mRNA, which further verified its physiological significance. Moreover, recent studies have revealed that deficiencies in f^5C34 are associated with multiple diseases [8,9]. In the context of aggressive and metastatic cancers, m^5C and f^5C drive the translation of mitochondrial mRNA, thus promoting metastasis. Inhibiting mitochondrial mRNA translation through targeting these specific RNA modifications could potentially serve as a therapeutic strategy to combat metastasis [10].

Recent advancements in high-throughput sequencing techniques have greatly facilitated comprehensive analysis of post-transcriptional RNA modifications [11]. Recently, Wang et al. developed f5C-seq, which successfully reduces f^5C to 5,6-dihydrouracil through treatment with pyridine borane [12]. Inspired by a C-to-T transition technique that detected DNA modifications 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) [13], this process involved the oxidation of 5mC and 5hmC to 5-carboxylcytosine (5caC), followed by the conversion of 5caC into dihydrouracil (DHU) using pyridine borane, and conversion to thymine (T) during PCR amplification. This technique enabled the precise mapping of f^5C in the transcriptome through an f^5C-to-T transition, thus providing a single-base resolution identification of f^5C sites across the entire transcriptome for the first time.

Although wet-lab experiments have advanced the study of epitranscriptome, they can be expensive and time-consuming. Therefore, computational approaches often served as a cost-effective avenue [14,15]. Numerous epitranscriptome databases [16–21] and computational approaches [22–29] have been developed for the large-scale collection or prediction of different types of RNA modifications. Among them, machine learning techniques like Support Vector Machine (SVM) [30–46], Random Forest (RF) [47], Logistic Regression (LR) [48] and eXtreme Gradient Boosting (XGB) [49] have been used for model development and comparative analysis. For example, iRNA-m^6A [50], RAMPred [51], RNAm^5CPred [52] and m^7GHub [53] utilized SVM to predict m^6A, m^1A, m^5C and m^7G sites respectively. Furthermore, deep learning-based approaches have also emerged as powerful tools for computational identification of RNA modification sites [54–58]. Gene2vec adopted word2vec embedding to encode m^6A sequences, combined with a convolution network and achieved an AUROC of 0.843 [59]. MultiRM performed multi-task learning using attention-based multi-label neural networks for predicting different types of modifications simultaneously [60]. AdaptRM was also introduced as a multi-task computational method designed for a synergetic learning of multi-tissue, type, and species RNA modifications from both high and low-resolution epitranscriptome datasets [58]. Besides these sequence-only methods, some approaches incorporated genomic features to enhance prediction accuracy and facilitate model interpretation. WHISTLE conducted an m^6A forecast utilizing the information of both sequence and genomic features, obtaining an AUROC of 0.904 on mRNA [61]. Geo2vec introduced novel encoding schemes for RNA transcripts, capturing sub-molecular geographic information and thereby enhancing the accuracy of tissue-specific prediction of m^6A RNA methylation sites [62]. These advancements have greatly facilitated the *in silico* identification of modified residues. However, to best our knowledge, there are still no prediction tools available for the computational identification of f^5C locations, limiting the efficient selection of putative f^5C sites from multiple regions of interest.

In addition to the previously mentioned individual machine learning and deep learning method, ensemble learning has emerged as an innovative approach that combines multiple models to improve overall performance. There are feature-based and model-based ensemble learning. In feature-based ensemble learning, different models are trained on distinct representations of input features, each focusing on capturing information of specific pattern. For example, MVIL6 integrated two output representations from the MG-BERT and the Transformer encoder to make the prediction of IL-6 induced peptide with their fusion output [63]. 4mCBERT encoded DNA sequence segments with Transformer and employed a CatBoost to identify DNA 4mC sites [64]. MNNEL-DTA model incorporated various neural networks, such as graph neural network (GNN), long and short memory (LSTM), and convolutional neural network (CNN) to identify potential therapeutic agents for Alzheimer's disease [65]. On the other hand, in model-based ensemble learning, multiple algorithms are independently trained on the same set of features, and their predictions are aggregated to enhance model performance and robustness. For example, EnsembleDL-ATG investigated different combinations of deep neural network (DNN), CNN and LSTM to predict autophagy proteins from protein sequence and evolutionary information [66]. A hybrid neural network (HNN) model was proposed to predict the drug-target affinity, which integrated multiple basic models such as DNN, CNN, LSTM and Transformer to obtain embedding features of drugs and targets [67]. While ensemble learning has shown promise in various research domains, its application in epitranscriptome analysis remains relatively limited.

Here, we took advantage of model-based ensemble deep learning and incorporated 32 novel genomic features to develop a computational recognition method for f^5C modification sites. Our research comprehensively investigated the mainstream machine learning algorithms (SVM, LR, XGB) and deep learning algorithms (CNN, ResNet, LSTM, Transformer). Among them, the Resf^{5}C-Pred, inspired by ResNet block, achieved an AUROC around 0.7492 with sequence input data, which is the highest among the algorithms. Subsequently, through integration with 32 domain features, it demonstrated significant improvement, reaching an AUROC of 0.8076 in both 5-fold cross-validation and on an independent dataset. Furthermore, by incorporating ensembles of multiple algorithms, proposed methods finally achieved an impressive AUROC of 0.8391. Fig. 1 shows the development process of our method, starting from data and feature generation (Fig. 1 A- C), training individual models (Fig. 1-D), and performing ensemble learning (Fig. 1-E). To interpret proposed models, Shapley value analysis was conducted to analyze how different features contribute to the overall predictions and provide insights into the putative distribution of f^5C across various topological regions. Moreover, to facilitate the identification of putative f^5C sites, a freely accessible web server is developed that allows real-time analysis of user-uploaded sites. The web server can be accessed at: www.rnamd.org/Resf5C-Pred.

2. Methods

2.1. Benchmark dataset

To develop our f^5C prediction framework, the experimentally-validated f^5C sites were obtained from a recently published technique f^5C-seq [12] (See Fig. 1-**A**). The positive datasets (f^5C sites) were downloaded from Gene Expression Omnibus (GEO) under accession number GSE133138. The annotation file of the yeast full genome utilized in this study is sacCer3 [68]. A total of 3784 base-resolution f^5C sites were extracted from yeast mRNAs (dataset P). The negative data (dataset N) was randomly selected from unmodified Cs located on the same transcripts of f^5C modification sites with 1:1 P-to-N ratio. The dataset was randomly split into the training and testing part with a ratio of 4:1.

2.2. Sequence characteristics

The length of 41 nt flanking window has been widely utilized for extracting sequence information in numerous previous studies [69–71]. In our study, we employed the combination of one-hot encoding and nucleotide density to extract the sequence characteristics of the 41 nt flanking window surrounding both f^5C and unmodified Cs (See Fig. 1-**B**). One-hot encoding converted each nucleotide into a vector. Specifically, four different types of nucleotides adenine (A), cytosine (C), guanine (G) and uracil (U) can be assigned to: (A → [1, 0, 0, 0], C → [0, 1, 0, 0], G → [0, 0, 1, 0], U → [0, 0, 0, 1]). Nucleotide density (ND) encodes the

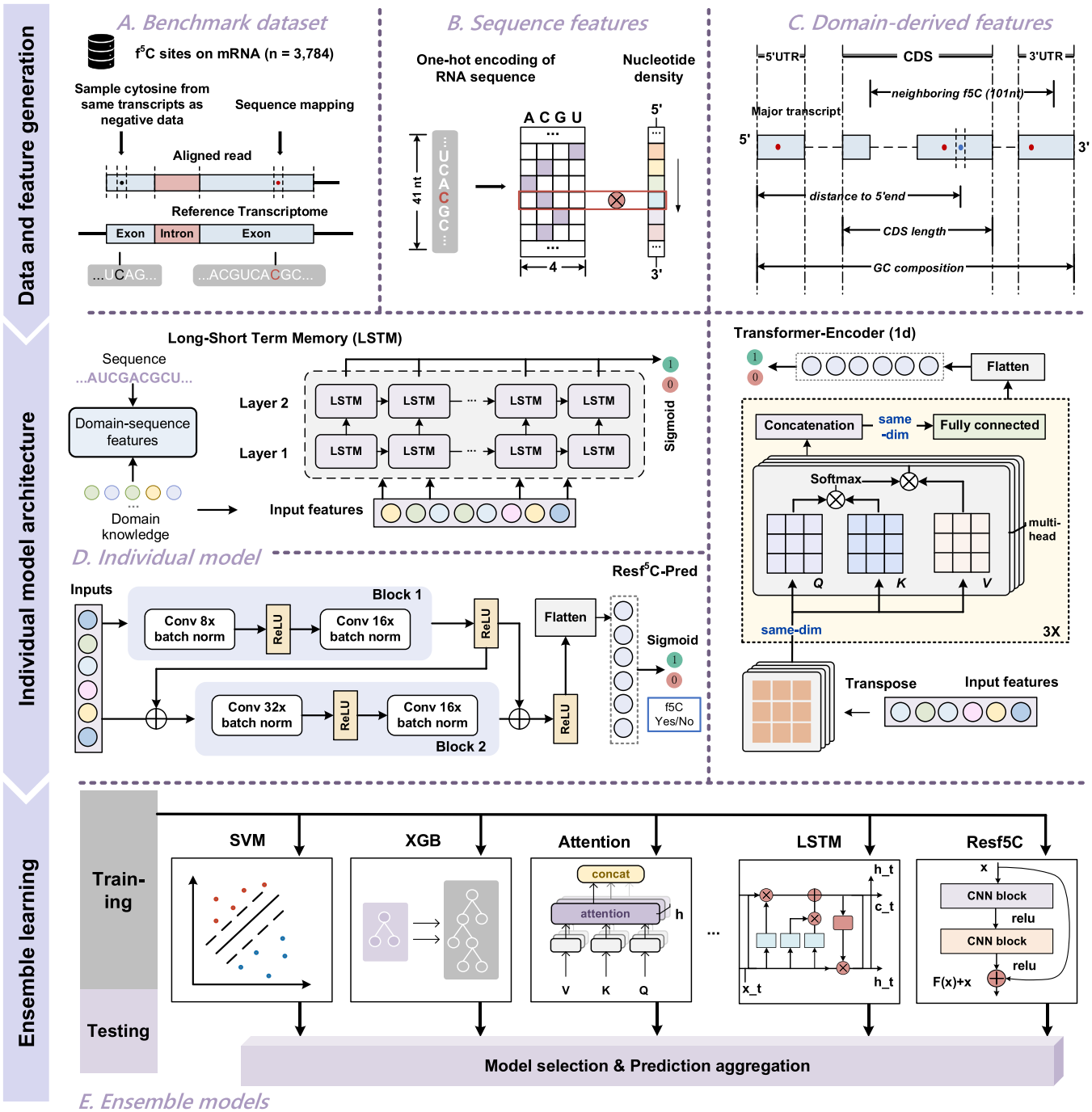

Fig. 1. Workflow of developing individual and ensemble models for f^{5}C identification. It entailed the following steps: (**A**) We collected f^{5}C modification sites derived from f^{5}C-seq [12] and sampled cytosine from the same transcripts as negative data. (**B**) We extracted the RNA primary sequence of 41 nt containing cytosine in the middle, which is a potential f^{5}C modification to be evaluated. Each nucleotide in sequences was encoded into a discrete vector consisting of a one-hot representation plus nucleotide density. (**C**) For each potential f^{5}C modification, we generated 32 additional domain-derived features that may contribute to the prediction. A complete list of these genomic features can be found in Table S1. (**D**) The model architecture of individual LSTM, Resf^{5}C-Pred, and Transformer-based method utilized for the computational identification of f^{5}C. (**E**) The ensemble model was constructed by aggregating the predictions of selected machine learning and deep learning methods.

nucleotides' distribution and cumulative frequency at each position [72]. The density of the i-th nucleotide was defined as follows:

$$d_i = \frac{\sum_{k=1}^{i} f(s_k)}{i} \quad (1)$$

where

$$f(s_k) = \begin{cases} 1 \text{if} s_k = s_i \\ 0 \text{otherwise} \end{cases} \quad (2)$$

Taking RNA sequence 'AUUCGCU' as an example, the nucleotide density of A at first position can be encoded as 1 (1/1), and the Us at the second and third positions as 0.5 (1/2) and 0.67 (2/3), respectively.

2.3. Domain-derived features

To improve the performance of the model in classifying modified or unmodified residues, domain-derived information has been included and encoded [73]. In our research, we extracted 32 domain (genomic) features for both f^{5}C modified and unmodified base-resolution sites (See Fig. 1-**C**). The first 13 features were dummy variables which indicate whether the f^{5}C sites are overlapped with specific topological regions such as start/stop codons, transcription start sites (TSS), exons, and introns on the major RNA transcripts. Features 14–15 represented the relative position of the target f^{5}C on two region types (i.e., coding sequence, exon). Genomic features 16–22 calculated the length of multiple region types, including mature transcript length, coding sequence length, full gene length, exon transcript length and full transcript length. Genomics features 23–26 recorded nucleotide distances toward the splicing junctions, nearest neighboring sites and cytosine residues. Genomic feature 27 described the count of adjacent neighboring cytosine. Lastly, features 28–32 represented genomic properties of the genes or transcripts related to the putative f^{5}C sites, such as GC composition. For more details about these genomic features for model construction, please refer to Supplementary Table S1.

2.4. Model training and method development overview

We evaluated three traditional machine learning methods: SVM, LR, and XGB alongside four deep learning methods: CNN, LSTM, Transformer-Encoder, and Resf^{5}C-Pred (Residual block + CNN). Initially, these methods were trained and evaluated using sequence information encoded in one-hot format and nucleotide density. To enhance performance, we incorporated 32 domain characteristics into the training and evaluation of these methods. Both sequence-only and sequence-domain integrated approaches were employed using a benchmark dataset with a balanced positive-negative ratio of 1:1. The dataset was split into a 4:1 ratio for training and testing, respectively, with five-fold cross-validation utilized during training.

Subsequently, the top five performing methods from the sequence-domain integrated models were selected to construct ensemble models, where the final result was obtained by aggregating the predicted values from each model. Five different combinations of these selected methods were evaluated and compared. The same testing dataset was utilized for evaluating each individual combination.

2.5. Resf^{5}C-Pred

We developed Resf^{5}C-Pred (Fig. 1-D), a model composed by convolutional blocks and residual structures to determine whether a putative site is f^{5}C modified or not. To be specific, both convolutional blocks were configured as 1-dimensional with a kernel size of 3, stride of 1, and padding of 1. Within the first convolutional block, the data passes through a convolutional layer, followed by a batch normalization and a ReLU activation. Subsequently, the processed data undergoes another convolutional layer and a batch normalization. The number of channels of each convolutional layer is set to [8,16]. The second block has similar architecture as the first convolutional block. The number of channels of each convolutional layer is set to [16,32]. Two residual blocks are implemented after two convolutional blocks, though a shortcut connection, adding the output from the previous layer to the output of the current layer. Next, the output is flattened and passed through dense layers, and fed into a sigmoid activation function for binary classification (f^{5}C Yes/No). Here is the formula for sigmoid activation function, where z is the input in the last two neuron before activation.

$$Sigmoid(z) = \frac{1}{1+e^{-z}} \tag{3}$$

The loss function employed in Resf^{5}C-Pred is binary cross entropy (i.e. BCELoss):

$$\text{BCELoss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \bullet \log p_i + (1-y_i) \bullet \log(1-p_i)] \tag{4}$$

where N is the total number of samples. y_i is the label of each individual sample, with a value of either 0 or 1. p_i is the predicted probability that the sample belongs to class 1, as predicted by the model for each sample. Log denotes the natural logarithm. The loss can also be expressed as

$$\text{BCELoss} = \begin{cases} -\log(sigmoid(z)) if y_1 = 1 \\ -\log(1-sigmoid(z)) if y_1 = 0 \end{cases} \tag{5}$$

The Adam optimizer [74] was implemented with a starting learning rate of 0.0001, which takes advantage of the momentum strength and adaptive learning rate simultaneously. The epoch for the training is set to be 50.

2.6. Transformer-encoder method

Transformer [75] is a prevalent Seq2seq model first proposed for neural machine translation and subsequently has been applied to many NLP tasks [76–79]. It functions through an Encoder-Decoder architecture. Since it has the potential to be adapted for various types of sequence-based tasks, we involved it in our study. Here, we only employed the Encoder module of the Transformer model. The model is mainly comprised by three multi-head self-attention layers. In each multi-head attention layer, multiple self-attention heads are parallel attention structures operating on the input sequence. For each attention head, three linear transformations of the input embeddings are derived: Query (Q), Key (K), and Value (V). These vectors are then used to compute the attention scores as follows:

$$Attention(Q, \quad K, \quad V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

where d_k is the column number of those equal-sized three matrixes, i.e. the matrix dimension. A softmax function is used to weight the corresponding value vectors. The concatenation of the resulting projection of this kind of attention head is then linearly transformed to finally produce the output of the multi-head attention layer. The model architecture is displayed in Fig. 1-D.

Details of hyperparameters used in this study were set as follow. The input to this model is a vector incorporating 205 sequence features (4 ×41 nt) and 32 domain features. Additionally, three padding elements are added to these features, resulting in a total feature length of 240. The input 1-d vector was transformed into a 10×24 matrix, making the input fit the model's dimensional requirement. In each multi-head self-attention layer, the head number is set to be 8. The dimension of each Query, Key, and Value matrix is $(24/8) \times 10$. The number of Encoder layers is set to 3. The final output of the Encoder module is flattened and passed through a fully connected layer with an output dimension of 2 to predict whether a site is f^{5}C modified or not. The loss function employed in our Transformer Encoder is cross entropy. Stochastic Gradient Descent (SGD) is selected as the optimizer with a learning rate of 0.005 and without momentum mechanism. The epoch for training is set to be 15.

2.7. Long short-term memory network method

LSTM [80] is a specialized variant of Recurrent Neural Network (RNN), which has been widely used in NLP tasks such as text categorization [81–83], machine translation [84–86] and language models [87,88]. LSTMs effectively addressed the challenge of vanishing or exploding gradients encountered by conventional RNNs when processing long sequences of data. In our study, the input of LSTM was set to be a 1-dimentional input vector, consisting of 205 dimensions for sequence-only mode and 237 dimensions for sequence-domain

integrated mode. The hidden size was set to be 256. We utilized two LSTM layers for model construction, followed by a linear layer to reduce the output to a single value. A sigmoid activation function was applied at the end to obtain a binary output. The LSTM model architecture is displayed in Fig. 1-D. For training, we employed binary cross-entropy loss as the loss function and Adam optimizer. The epoch was set to be 50.

The competing methods also included SVM, LR, XGB and CNN. All the methods were implemented using Pytorch 2.1.0. For details of the remaining methods, please refer to our Github repository.

2.8. Ensemble learning

Ensemble learning is a technique that combines the power of multiple predictive models to improve overall performance. The ensemble model was selected as one of our comparison methods for three reasons including reduced overfitting, improved accuracy, and improved robustness. Ensemble model (Fig. 1-E) considered multiple combinations of methods to give a comprehensive evaluation. Sequence-domain integrated methods with high AUROC in cross validation set and independent test set were selected for constructing ensemble model. In machine learning methods, LR and XGB were selected, while in deep learning methods, CNN, LSTM and Resf^{5}C were selected. Here we considered five combinations following their AUROC: 1. CNN, LSTM, LR, XGB and Resf^{5}C-Pred (top 5); 2. CNN, LR, XGB and Resf^{5}C-Pred (top 4); 3. CNN, LR and XGB (top 2 in ML and top 2 in DL); 4. LR, XGB and Resf^{5}C-Pred (top 2 in ML and top 1 in DL); 5. XGB and Resf^{5}C-Pred (top 2 in all methods). Those ensemble models were compared with each other under sequence-domain integrated context. The ensemble method aggregates predictions from these different models by taking an average of their output as a fusion result. Ideally, due to the combination of multiple models, the integrated model tends not to be overly dependent on specific features of the training data, thus reducing the risk of overfitting. The integrated model is less sensitive to small changes and outliers in the data.

2.9. Performance evaluation metrics

The following evaluation metrics were applied. We used the Receiver Operating Characteristic (ROC) curve (sensitivity against 1-specificity) and the area under the ROC curve (AUROC). Besides AUROC, sensitivity (Sn), specificity (Sp), overall accuracy (ACC), F1 score, and Matthew's Correlation Coefficient (MCC) were also included. A 5-fold cross-validation was applied on 80 % of the data as training datasets, while the rest of 20 % were used as testing datasets for independent testing.

$$S_n = \frac{TP}{TP + FN} \tag{7}$$

$$S_p = \frac{TN}{TN + FP} \tag{8}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{10}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{11}$$

Among them, *TP* represents the number of true positives, while *TN* represents true negatives; *FP* stands for the number of false positives, and *FN* stands for the number of false negatives.

2.10. Estimating the likelihood ratio of putative f^5C modification sites

For each predicted putative f^{5}C sites, we calculated the likelihood ratio to represent how extreme the site can be a true f^{5}C residue as follows:

$$LR = \frac{P(observation|f^5C)}{P(observation|C)} \tag{12}$$

Following the above calculation, a site would be classified as a putative f^{5}C site if its predictive value exceeded 0.5, i.e., a minimum LR value of 1. A higher likelihood ratio value of a predicted result suggests that it has a greater likelihood of being a f^{5}C residue. In our proposed f^{5}C prediction website, positive sites (prediction probability > 0.5) are categorized as follows: those with a probability greater than 0.9 are classified as high-confidence, between 0.8 and 0.9 as medium-confidence, and between 0.5 and 0.8 as low-confidence. Correspondingly, the likelihood ratio intervals assigned to these categories are > 9 for indicating high confidence in predicting an f^{5}C site;.

> 4 and < 9 for a medium confidence, > 1 and < 4 for a low confidence.

3. Results

3.1. Sequence-only methods

To develop our prediction method, we first evaluate the performance of different sequence-encoding approaches for f^{5}C prediction using both machine learning and deep learning methods. Three different machine learning classifiers were included: SVM, LR and XGB. Three deep learning methods were included: CNN, LSTM and Resf^{5}C-Pred. Performance evaluation of these sequence-only methods using a 41 nt sequence has been summarized in Table 1. The comparison of model performance with varying sequence lengths is shown in Table S2. Methods achieved highest accuracy when using a 41 nt sequence as input. Additionally, we demonstrated a combined ROC figure for cross validation (Fig. 2**a**) and independent test (Fig. 2**b**). The results indicated that deep learning methods generally outperformed machine learning methods. However, relying solely on sequence information may not be sufficient for achieving highly accurate predictions. Even the best sequence-only method, Resf^{5}C-Pred only achieved an AUROC around 0.7492 in the independent test. To improve f^{5}C prediction, we tried to integrate sequence-based features with domain-derived characteristics in the subsequent section.

3.2. Sequence-domain integrated methods

We incorporated the sequence-based information (i.e. one-hot encoding and nucleotide density) with 32 newly derived features from the genome domain to create the final model for f^{5}C site prediction. The performance of all the methods was evaluated utilizing both sequence and domain features as input. The related performance evaluation is demonstrated in Table 2. (The performance evaluation of each method using only 32 features as input is shown in Table S3.) The ROC curves for cross validation and independent test are shown in Fig. 3a and b. By combining domain-derived features, the prediction performance improved significantly compared to using sequence alone across all testing methods. Among these methods, Resf^{5}C-Pred achieved the best prediction performance, with an average AUC of 0.8074 and 0.8076 tested on 5-fold cross-validation and the independent dataset, respectively.

3.3. Ensemble models

From Table 2, we can see that LR, XGB, CNN, LSTM and Resf^{5}C-Pred are the top 5 methods with the best performance. Next, we constructed ensemble models by applying different combinations of models from those five methods. The performance evaluation toward those methods in independent test is summarized in Table 3 and a combined ROC figure

Table 1
Performance evaluation of sequence-only methods.

Methods & Modes		Sn (%)	Sp (%)	ACC (%)	F1	MCC	AUROC
SVM	Cross Validation	73.85	63.72	66.57	0.6841	0.3374	0.7320
	Independent Test	70.23	64.41	66.17	0.6720	0.3254	0.7104
LR	Cross Validation	69.62	63.55	66.01	0.6645	0.3228	0.7303
	Independent Test	72.58	67.63	67.63	0.6941	0.3534	0.7262
XGB	Cross Validation	83.98	58.39	64.33	0.6889	0.3293	0.7317
	Independent Test	85.53	63.33	67.16	0.7277	0.3611	0.7265
CNN	Cross Validation	65.57	67.10	66.33	0.6622	0.3268	0.7313
	Independent Test	63.60	66.77	65.18	0.6477	0.3039	0.7240
LSTM	Cross Validation	68.19	70.76	69.47	0.6921	0.3897	0.7478
	Independent Test	67.20	64.39	65.78	0.6605	0.3160	0.7197
Resf^{5}C-Pred	Cross Validation	74.24	62.86	68.48	0.6992	0.3733	0.7574
	Independent Test	79.47	55.26	67.32	0.7079	0.3578	0.7492

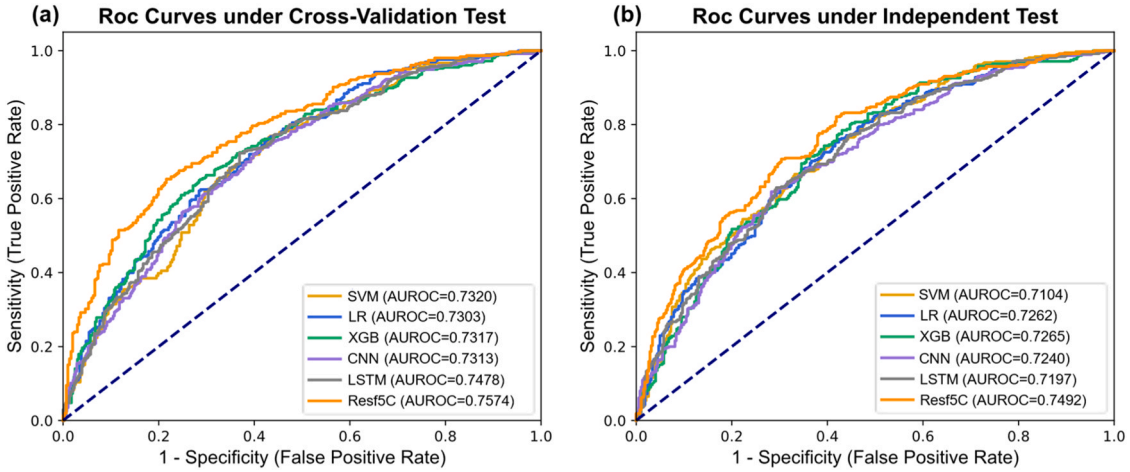

Fig. 2. The ROC curves of sequence-only methods. (**A**) The ROC curves of multiple approaches for identifying f^{5}C modification sites under the 5-fold cross-validation test. (**B**) The ROC curves of identifying f^{5}C modification sites under the independent dataset test.

Table 2
Performance evaluation of sequence-domain integrated methods.

Methods & Modes		Sn (%)	Sp (%)	ACC (%)	F1	MCC	AUROC
SVM	Cross Validation	68.46	63.42	66.97	0.6584	0.3406	0.7311
	Independent Test	70.38	65.97	65.84	0.6810	0.3149	0.7309
LR	Cross Validation	71.91	66.03	68.64	0.6885	0.3754	0.7680
	Independent Test	73.13	65.78	67.76	0.6926	0.3580	0.7509
XGB	Cross Validation	78.42	68.97	68.97	0.7089	0.3914	0.7865
	Independent Test	82.18	66.02	70.14	0.7322	0.4162	0.7850
CNN	Cross Validation	77.73	62.73	69.96	0.7138	0.4082	0.7617
	Independent Test	76.59	58.79	67.63	0.7015	0.3594	0.7566
LSTM	Cross Validation	73.97	64.01	68.81	0.6956	0.3810	0.7566
	Independent Test	71.27	60.62	65.91	0.6750	0.3208	0.7408
Transformer-Encoder	Cross Validation	73.13	64.30	68.69	0.6988	0.3758	0.7396
	Independent Test	71.80	63.77	67.76	0.6887	0.3569	0.7319
Resf5C-Pred	Cross Validation	81.50	63.05	71.94	0.7368	0.4517	0.8074
	Independent Test	81.11	60.62	70.80	0.7340	0.4262	0.8076

for independent test is demonstrated in Fig. 3**c**. The ensemble method aggregated predictions from these different models by fusing their output to make a final decision. All ensemble methods demonstrated improvement compared to the average AUROC of individual methods (0.7309–0.8076). The ensemble model of XGB and Resf^{5}C-Pred combination outperformed other combinations with an AUROC value of 0.8391.

3.4. Model interpretation

Gaining insights into the key input features and the underlying mechanisms behind model decisions are crucial for making further improvements to the model. However, interpreting deep-leaning models can be challenging. To overcome this limitation, we adopted an approach to interpret the role of domain-derived features using machine learning alternative, following a previously published work [89]. The Shapley additive explanations [90] were employed to assess the relative importance of each input feature in the prediction.

We conducted SHAP analysis for model interpretation of the top 3 domain-integrated methods XGB, CNN, and Resf^{5}C-Pred. Based on their SHAP values, top 10 important features of each method are shown in Fig. 4. Similar sets of top important features can be found across different individual methods. Several important features reoccur in these three methods, which include dist_C_p200 (distance to nearest

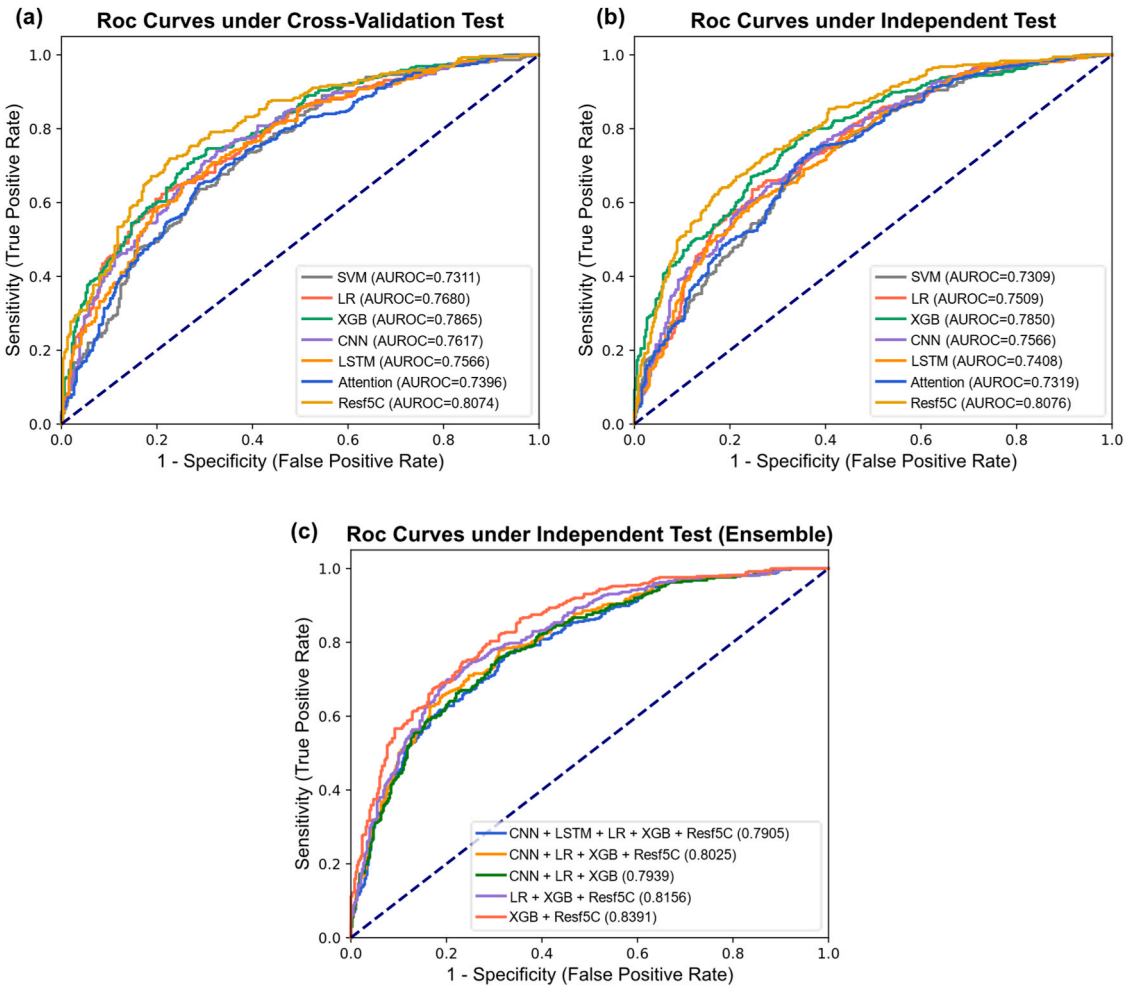

Fig. 3. The ROC curves of sequence-domain integrated methods. (**A**) The ROC curves of multiple approaches for identifying f^{5}C modification sites under the 5-fold cross-validation test. (**B**) The ROC curves of identifying f^{5}C modification sites under the independent dataset test. (**C**) The ROC curves of ensemble models with different method combinations under the independent dataset test.

Table 3
Performance evaluation of sequence-domain integrated ensemble methods.

Ensemble Methods	Sn (%)	Sp (%)	ACC (%)	F1	MCC	AUROC	Improvement
CNN + LSTM + LR + XGB + Resf^{5}C-Pred	78.72	63.77	71.20	0.7308	0.4296	0.7905	↑2.69 %
CNN + LR + XGB + Resf^{5}C- Pred	82.93	62.76	72.91	0.6971	0.4669	0.8025	↑4.14 %
CNN + LR + XGB	78.45	63.51	70.93	0.7283	0.4243	0.7939	↑3.10 %
LR + XGB + Resf^{5}C-Pred	80.58	63.58	71.99	0.7408	0.4473	0.8156	↑5.68 %
XGB + Resf^{5}C-Pred	86.17	63.25	74.63	0.7714	0.5073	0.8391	↑8.32 %

neighboring C site within 201 bp), last_exon_400 bp (whether the site is overlapped with 400 bp of the last exon), GC_cont_101bp (GC composition of 101 bp region around the site) and pos_cds (relative position of the site on coding sequence). The results indicate that these features might significantly influence the predictive occurrence of f^{5}C modifications. On the other hand, other dummy variables, i.e., features indicating whether the site overlaps with specific transcript regions, such as start_codons, stop_codons (genomic regions surrounding start codons and stop codons), and TSS (transcription start sites), exhibit relatively lower SHAP values, suggesting the influence of these particular transcript regions on the prediction results may be limited and their functional relevance related to f^{5}C might be weak. Fig. 5.

High SHAP values indicate the importance of a feature in predicting f^{5}C modifications and its statistical association with f^{5}C. However, it does not necessarily provide direct evidence of the biological functions or regulatory mechanisms in which the feature is involved. Specific experiments need to be conducted to further interpreting the significance of a feature in a complex biological system. Moreover, the synergy between these features is not yet discussed in this study. Additionally, we conducted a SHAP analysis for an ensemble learning model (Resf^{5}C-Pred + XGB). The result is shown in Fig. S1. We found similar top important features across individual models and their ensemble, suggesting consistency of feature importance and also model robustness. For detailed information about these feature items, please refer to Supplementary Table S1.

3.5. Web server

A user-friendly online platform has been developed for sharing our findings (Fig. 3). Users can easily access this platform by either inputting site coordinates or uploading a text file containing chromosome, position, and strand information for the query sites. Our Resf^{5}C-Pred

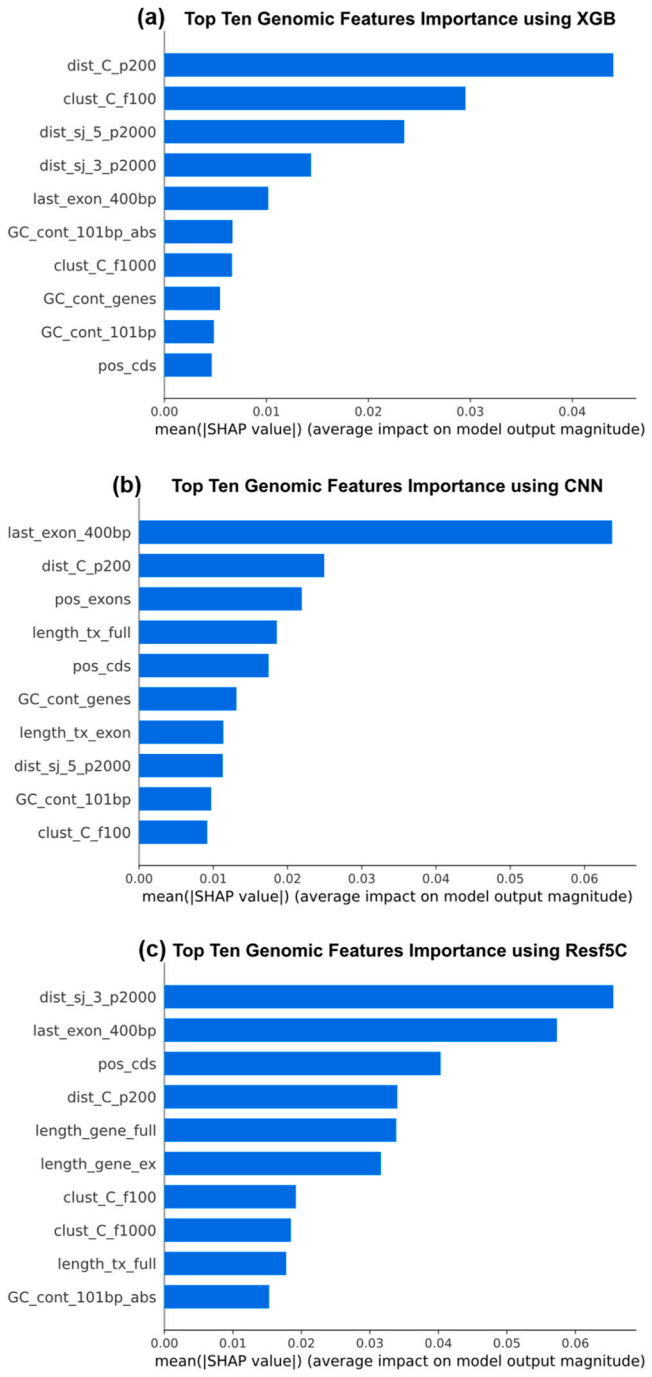

Fig. 4. Model interpretation. The significance of each domain feature in the prediction was assessed using Shapley additive explanations. The top ten genomic features in XGB, CNN, and Resf^{5}C-Pred were demonstrated in (**A**), (**B**) and (**C**) respectively.

webserver provides two modes of f^{5}C modification site prediction. One mode is "Resf^{5}C-Pred Only" and another mode is "Ensemble Model (Resf^{5}C-Pred and XGB)". By utilizing our platform, the web server will provide predictions for the probability of the site being an f^{5}C: whether the site is f^{5}C or not (using a default threshold of 0.5) and the corresponding confidence level. The results can be downloaded as an excel table. The website is now available at www.rnamd.org/Resf5C-Pred.

4. Discussion

5-formylcytidine (f^{5}C) is a unique modification observed post-transcriptionally in mRNA and tRNA, playing a vital role in mitochondrial protein synthesis and possibly influencing translation regulation. Identifying f^{5}C modification sites is essential to investigate its molecular roles and regulatory mechanisms, yet there has been a lack of computational tools for such predictions. To investigate the computational identification of f^{5}C sites, we investigated traditional machine learning approaches, deep learning approaches and their ensembles. Firstly, we trained our models with one-hot encoded 41nt sequence information but found they only exhibited an AUROC value between 0.7104 and 0.7492. To further improve prediction, 32 additional genomic features were included, enabling the best model achieved AUROC 0.8047 and 0.8076 in both 5-fold cross-validation and independent dataset evaluations respectively. We also developed ensemble models. The best ensemble model which fuses prediction result from both XGB and Resf^{5}C-Pred, achieving an AUROC of 0.8391. Lastly, we explained our models using Shapley additive explanations. The results suggested that the distance to 5′ and 3′ splicing junctions emerged as the top two predictive features, followed by the distance to known f^{5}C sites, demonstrating potential association between splicing junctions and the clustering effect of f^{5}C modification. This possibly implied certain specific transcripts regions with a high likelihood of f^{5}C occurrence. Additionally, almost the same time as our work, Wang et al. independently developed an f^{5}C predictor based on a multi-head attention framework. They utilized five distinct feature extraction methods to achieve an integrated learning of f^{5}C and reported AUROCs of 0.807 and 0.827 on 10-fold cross-validation and independent tests, respectively [91]. It should be noted that direct comparison of AUROC is inappropriate due to the different strategies of constructing benchmark datasets in both studies.

There is a theoretical possibility of confusing f^{5}C (5-formylcytidine) with 5caC (5-carboxylcytosine) since the experiment profiling of both modifications undergoes a similar conversion to dihydrouracil (DHU) using pyridine borane. The source f^{5}C data we utilized in this study may contain sites arise from 5caC in RNA [12]. Therefore, proposed prediction model in our study might also wrongly predict 5caC sites as f^{5}C. Lack of labeled data makes it difficult to address this issue in the current stage. Golden standard f^{5}C data sites are expected to be established in future to minimize the risk of such misidentification.

Another limitation of this study is its focus solely on f^{5}C sites in *Saccharomyces cerevisiae* mRNAs. Multi-organisms or -species predictions are not conducted due to the limited availability of specific wet-experiment f^{5}C epitranscriptome data. Besides the work conducted by Wang *et al.* that reported f^{5}C sites in yeast mRNA [12], a recent study by Lyu et al. utilized f^{5}C-seq to map f^{5}C modifications in the tRNA and chromatin-associated RNA (caRNA) across the HeLa cells and mouse embryonic stem cells (mESCs) [92]. However, the number of identified f^{5}C sites was relatively limited (13 in HeLa tRNA, 11 in mESC tRNA, 3 in HeLa caRNA, and 3 in mESC caRNA), which is insufficient for training a predictive model. Additionally, f^{5}C has previously been identified within human mRNA, but a comprehensive dataset providing transcriptome-wide profiling of high-confidence f^{5}C sites is still lacking. Despite this, these findings serve as valuable references for future studies exploring the biological functions and mechanisms of the epitranscriptomic mark of f^{5}C.

5. Conclusion

In this study, we incorporated three traditional machine learning approaches (SVM, LR and XGB), four deep learning approaches (CNN, LSTM, Transformer-Encoder and Resf^{5}C-Pred) and their ensembles, to investigate the computational identification of f^{5}C sites. The integration of 32 novel genomic features derived from f^{5}C-related domain knowledge and the ensemble learning framework significantly improved prediction accuracy compared to the sequence-only methods. Shapley additive explanations were conducted to explain the significant contributions of specific genomic feature. A free-to-use web server has been established, enabling users to conduct real-time analyses of potential f^{5}C

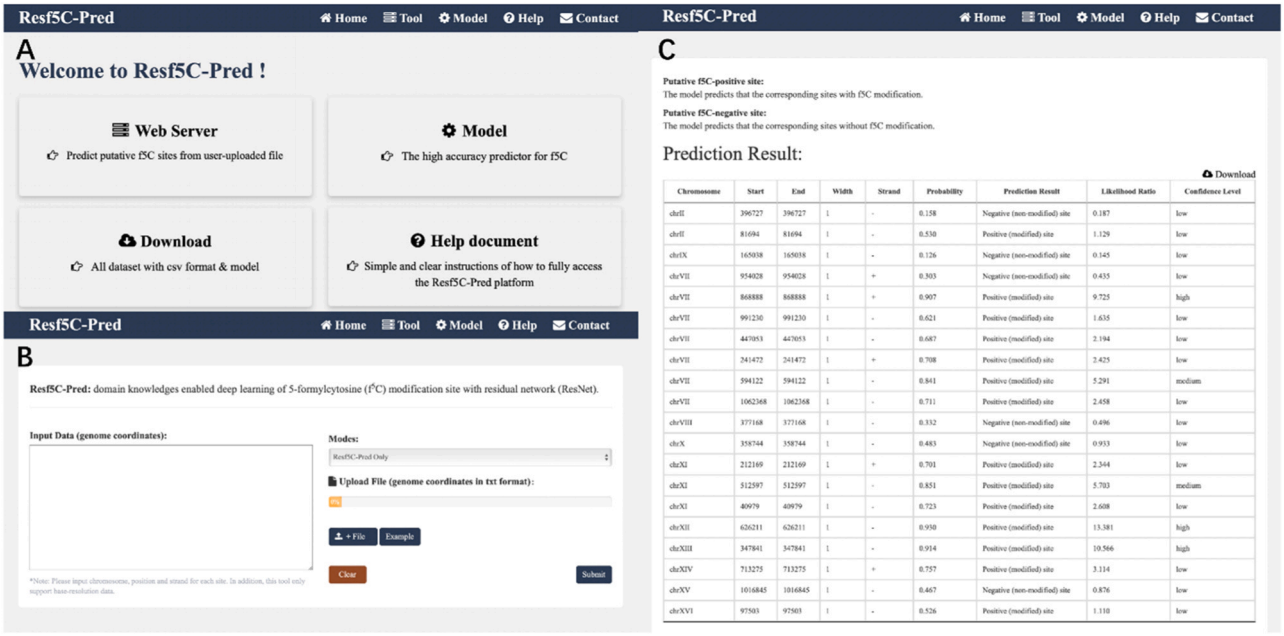

Fig. 5. Screenshot of the Resf^{5}C-Pred webserver. Users can upload query f^{5}C sites information. Results can be downloaded from the web page as a csv file.

sites. The work only focuses on the f^{5}C sites in *Saccharomyces cerevisiae* mRNAs due to data limitation. Future work could include training tissue-specific models when more data is available and incorporating features such as secondary structures and other RNA types to enhance the model's robustness and generalizability.

Code availability

The Resf^{5}C-Pred, LSTM, Transformer-Encoder and all ensemble models were implemented with Pytorch 2.1.0. Codes for model construction can be available at: https://github.com/Jiaming21/F5C-codes.git. The web server together with the completed datasets and the parameters of two trained models (XGB and Resf^{5}C-Pred) can be freely accessed at www.rnamd.org/Resf5C-Pred.

Funding

Nanjing University of Chinese Medicine (Grant No. 013038019029 and 013038030001); National Natural Science Foundation of China [82204443]; XJTLU Key Program Special Fund [KSF-E-51 and KSF-P-02]; Natural Science Foundation of Jiangsu province (BK20210693), General Project of Basic Science in Colleges and Universities of Jiangsu Province (21KJB310013), Jiangsu Provincial Health Commission (Z2021064). This work is supported by the Supercomputing Platform of Xi'an Jiaotong-Liverpool University.

CRediT authorship contribution statement

Jiaming Huang: Methodology, Software, Writing original draft. **Xuan Wang:** Website, Software. **Rong Xia:** Data curation. **Dongqing Yang:** Resources, Supervision. **Jian Liu:** Resources, Supervision. **Qi Lv:** Resources, Supervision. **Xiaoxuan Yu:** Resources, Supervision. **Jia Meng:** Conceptualization, Resources, Supervision. **Kunqi Chen:** Resources, Supervision. **Bowen Song:** Data curation, Writing original draft, Resources, Supervision. **Yue Wang:** Conceptualization, Writing original draft, Methodology, Software, Resources, Supervision.

Declaration of Competing Interest

All authors 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. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.08.004.

References

[1] Barbieri I, Kouzarides T. Role of RNA modifications in cancer. Nat Rev Cancer 2020;20(6):303–22.
[2] Machnicka MA, Milanowska K, Osman Oglou O, et al. MODOMICS: a database of RNA modification pathways–2013 update. Nucleic Acids Res 2013;41(Database issue):D262–7.
[3] Helm M, Alfonzo JD. Posttranscriptional RNA Modifications: playing metabolic games in a cell's chemical Legoland. Chem Biol 2014;21(2):174–85.
[4] Lusic H, Gustilo EM, Vendeix FA, et al. Synthesis and investigation of the 5-formylcytidine modified, anticodon stem and loop of the human mitochondrial tRNAMet. Nucleic Acids Res 2008;36(20):6548–57.
[5] Kawarada L, Suzuki T, Ohira T, et al. ALKBH1 is an RNA dioxygenase responsible for cytoplasmic and mitochondrial tRNA modifications. Nucleic Acids Res 2017;45(12):7401–15.
[6] Haag S, Sloan KE, Ranjan N, et al. NSUN3 and ABH1 modify the wobble position of mt-tRNAMet to expand codon recognition in mitochondrial translation. EMBO J 2016;35(19):2104–19.
[7] Takemoto C, Spremulli LL, Benkowski LA, et al. Unconventional decoding of the AUA codon as methionine by mitochondrial tRNAMet with the anticodon f5CAU as revealed with a mitochondrial in vitro translation system. Nucleic Acids Res 2009;37(5):1616–27.
[8] Nakano S, Suzuki T, Kawarada L, et al. NSUN3 methylase initiates 5-formylcytidine biogenesis in human mitochondrial tRNA(Met). Nat Chem Biol 2016;12(7):546–51.
[9] Van Haute L, Dietmann S, Kremer L, et al. Deficient methylation and formylation of mt-tRNA(Met) wobble cytosine in a patient carrying mutations in NSUN3. Nat Commun 2016;7:12039.
[10] Delaunay S, Pascual G, Feng B, et al. Mitochondrial RNA modifications shape metabolic plasticity in metastasis. Nature 2022;607(7919):593–603.
[11] Boccaletto P, Baginski B. MODOMICS: an operational guide to the use of the RNA modification pathways database. Methods Mol Biol 2021;2284:481–505.
[12] Wang Y, Chen Z, Zhang X, et al. Single-base resolution mapping reveals distinct 5-formylcytidine in Saccharomyces cerevisiae mRNAs. ACS Chem Biol 2022;17(1):77–84.
[13] Liu Y, Siejka-Zielinska P, Velikova G, et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nat Biotechnol 2019;37(4):424–9.
[14] Chen X., Sun Y.Z., Liu H., et al., RNA methylation and diseases: experimental results, databases, Web servers and computational models. Brief Bioinform, 2017: p. bbx142-bbx142.

[15] Chen Z, Zhao P, Li F, et al. Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. Brief Bioinform 2019.
[16] Boccaletto P, Machnicka MA, Purta E, et al. MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res 2018;46(D1):D303–7.
[17] Xuan JJ, Sun WJ, Lin PH, et al. RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. Nucleic Acids Res 2018;46 (D1):D327–34.
[18] Song B, Chen K, Tang Y, et al. ConsRM: Collection and large-scale prediction of the evolutionarily conserved RNA methylation sites, with implications for the functional epitranscriptome. Brief Bioinforma 2021.
[19] Bao X, Zhang Y, Li H, et al. RM2Target: a comprehensive database for targets of writers, erasers and readers of RNA modifications. Nucleic Acids Res 2023;51(D1): D269–79.
[20] Song B, Wang X, Liang Z, et al. RMDisease V2.0: an updated database of genetic variants that affect RNA modifications with disease and trait implication. Nucleic Acids Res 2022.
[21] Luo X, Li H, Liang J, et al. RMVar: an updated database of functional variants involved in RNA modifications. Nucleic Acids Res 2021;49(D1):D1405–12.
[22] Qiu WR, Jiang SY, Xu ZC, et al. iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget 2017;8(25):41178–88.
[23] Chen W, Song X, Lv H, et al. iRNA-m2G: Identifying N(2)-methylguanosine sites based on sequence-derived information. Mol Ther Nucleic Acids 2019;18:253–8.
[24] Zhai J, Song J, Cheng Q, et al. PEA: an integrated R toolkit for plant epitranscriptome analysis. Bioinformatics 2018;34(21):3747–9.
[25] Liang Z, Zhang L, Chen H, et al. m6A-Maize: Weakly supervised prediction of m(6) A-carrying transcripts and m(6)A-affecting mutations in maize (Zea mays). Methods 2021.
[26] Körtel N, Rücklé C, Zhou Y, et al. Deep and accurate detection of m6A RNA modifications using miCLIP2 and m6Aboost machine learning. Nucleic Acids Res 2021.
[27] Xiong Y, He X, Zhao D, et al. Modeling multi-species RNA modification through multi-task curriculum learning. Nucleic Acids Res 2021.
[28] Yao J, Hao C, Chen K, et al. Pseudouridine Identification and Functional Annotation with PIANO. Methods Mol Biol 2023;2624:153–62.
[29] Wang Y, Wang X, Cui X, et al. Self-attention enabled deep learning of dihydrouridine (D) modification on mRNAs unveiled a distinct sequence signature from tRNAs. Mol Ther Nucleic Acids 2023;31:411–20.
[30] Chen W, Feng P, Ding H, et al. iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. Anal Biochem 2015;490:26–33.
[31] Chen W, Feng P, Ding H, et al. Identifying N (6)-methyladenosine sites in the Arabidopsis thaliana transcriptome. Mol Genet Genom 2016;291(6):2225–9.
[32] Chen W, Tang H, Lin H. MethyRNA: a web server for identification of N(6)-methyladenosine sites. J Biomol Struct Dyn 2017;35(3):683–7.
[33] Feng P, Ding H, Yang H, et al. iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. Mol Ther Nucleic Acids 2017;7:155–63.
[34] Hou Y, Bao J, Song Y, et al. Integration of clinicopathologic identification and deep transferrable image feature representation improves predictions of lymph node metastasis in prostate cancer. EBioMedicine 2021;68:103395.
[35] Chen W, Ding H, Zhou X, et al. iRNA(m6A)-PseDNC: Identifying N(6)-methyladenosine sites using pseudo dinucleotide composition. Anal Biochem 2018; 561-562:59–65.
[36] Chen W, Xing P, Zou Q. Detecting N(6)-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines. Sci Rep 2017;7:40242.
[37] Xing P, Su R, Guo F, et al. Identifying N(6)-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. Sci Rep 2017;7:46757.
[38] Xiang S, Liu K, Yan Z, et al. RNAMethPre: a web server for the prediction and query of mRNA m6A Sites. PLoS One 2016;11(10):e0162707.
[39] Chen W, Feng P, Yang H, et al. iRNA-3typeA: identifying three types of modification at RNA's Adenosine Sites. Mol Ther Nucleic Acids 2018;11:468–74.
[40] Wei L, Chen H, Su R. M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. Mol Ther Nucleic Acids 2018;12: 635–44.
[41] Liu Z, Xiao X, Yu DJ, et al. pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties. Anal Biochem 2016;497:60–7.
[42] Li GQ, Liu Z, Shen HB, et al. TargetM6A: Identifying N(6)-methyladenosine sites from RNA sequences via position-specific nucleotide propensities and a support vector machine. IEEE Trans Nanobioscience 2016;15(7):674–82.
[43] Xiang S, Yan Z, Liu K, et al. AthMethPre: a web server for the prediction and query of mRNA m(6)A sites in Arabidopsis thaliana. Mol Biosyst 2016;12(11):3333–7.
[44] Akbar S, Hayat M. iMethyl-STTNC: Identification of N(6)-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences. J Theor Biol 2018;455:205–11.
[45] Jia CZ, Zhang JJ, Gu WZ. RNA-MethylPred: a high-accuracy predictor to identify N6-methyladenosine in RNA. Anal Biochem 2016;510:72–5.
[46] Tu G, Wang X, Xia R, et al. m6A-TCPred: a web server to predict tissue-conserved human m6A sites using machine learning approach. BMC Bioinforma 2024;25(1): 127.
[47] Zhou Y, Zeng P, Li YH, et al. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res 2016;44(10):e91.
[48] Zhuang YY, Liu HJ, Song X, et al. A linear regression predictor for identifying N(6)-methyladenosine sites using frequent gapped K-mer pattern. Mol Ther Nucleic Acids 2019;18:673–80.
[49] Zhao Z, Peng H, Lan C, et al. Imbalance learning for the prediction of N(6)-Methylation sites in mRNAs. BMC Genom 2018;19(1):574.
[50] Dao FY, Lv H, Yang YH, et al. Computational identification of N6-methyladenosine sites in multiple tissues of mammals. Comput Struct Biotechnol J 2020;18: 1084–91.
[51] Chen W, Feng P, Tang H, et al. RAMPred: identifying the N1-methyladenosine sites in eukaryotic transcriptomes. Sci Rep 2016;6(1):31080.
[52] Fang T, Zhang Z, Sun R, et al. RNAm5CPred: prediction of RNA 5-methylcytosine sites based on three different kinds of nucleotide composition. Mol Ther - Nucleic Acids 2019;18:739–47.
[53] Wang X, Zhang Y, Chen K, et al. m7GHub V2.0: an updated database for decoding the N7-methylguanosine (m7G) epitranscriptome. Nucleic Acids Res 2024;52(D1): D203–12.
[54] Zou Q, Xing P, Wei L, et al. Gene2vec: gene subsequence embedding for prediction of mammalian N (6)-methyladenosine sites from mRNA. RNA 2019;25(2):205–18.
[55] Chen Z, Zhao P, Li F, et al. Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. Brief Bioinform 2020;21(5):1676–96.
[56] Huang D, Song B, Wei J, et al. Weakly supervised learning of RNA modifications from low-resolution epitranscriptome data. Bioinformatics 2021.
[57] Song B, Huang D, Zhang Y, et al. m6A-TSHub: unveiling the context-specific m(6)A methylation and m6A-affecting mutations in 23 human tissues. Genom Proteom Bioinforma 2022.
[58] Song Y, Wang Y, Wang X, et al. Multi-task adaptive pooling enabled synergetic learning of RNA modification across tissue, type and species from low-resolution epitranscriptomes. Brief Bioinforma 2023;24(3):bbad105.
[59] Zou Q, Xing P, Wei L, et al. Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. RNA 2019;25:205–18.
[60] Song Z, Huang D, Song B, et al. Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. Nat Commun 2021;12(1):4011.
[61] Chen K, Wei Z, Zhang Q, et al. WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. Nucleic Acids Res 2019;47(7):e41.
[62] Huang D, Chen K, Song B, et al. Geographic encoding of transcripts enabled high-accuracy and isoform-aware deep learning of RNA methylation. Nucleic Acids Res 2022;50(18):10290–310.
[63] Wang R, Feng Y, Sun M, et al. MVIL6: accurate identification of IL-6-induced peptides using multi-view feature learning. Int J Biol Macromol 2023;246:125412.
[64] Yang S, Yang Z, Yang J. 4mCBERT: a computing tool for the identification of DNA N4-methylcytosine sites by sequence- and chemical-derived information based on ensemble learning strategies. Int J Biol Macromol 2023;231:123180.
[65] Zhao L, Li Z, Chen G, et al. Multi-perspective neural network for dual drug repurposing in Alzheimer's disease. Knowl-Based Syst 2024;283:111195.
[66] Yu L, Zhang Y, Xue L, et al. EnsembleDL-ATG: identifying autophagy proteins by integrating their sequence and evolutionary information using an ensemble deep learning framework. Comput Struct Biotechnol J 2023;21:4836–48.
[67] Lv Q, Chen G, He H, et al. TCMBank: bridges between the largest herbal medicines, chemical ingredients, target proteins, and associated diseases with intelligence text mining. Chem Sci 2023;14(39):10684–701.
[68] TBD T., BSgenome.Scerevisiae.UCSC.sacCer3: Saccharomyces cerevisiae (Yeast) full genome (UCSC version sacCer3). R package version 1.4.0., 2014.
[69] Liu K, Chen W, Lin H. XG-PseU: an eXtreme gradient boosting based method for identifying pseudouridine sites. Mol Genet Genom 2020;295(1):13–21.
[70] Yang H, Lv H, Ding H, et al. iRNA-2OM: a sequence-based predictor for identifying 2′-O-methylation sites in Homo sapiens. J Comput Biol 2018;25(11):1266–77.
[71] Yang X, Yang Y, Sun BF, et al. 5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an m(5)C reader. Cell Res 2017;27(5): 606–25.
[72] Chen W, Yang H, Feng P, et al. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics 2017;33(22):3518–23.
[73] Liu L, Song B, Chen K, et al. WHISTLE server: a high-accuracy genomic coordinate-based machine learning platform for RNA modification prediction. Methods 2021.
[74] Kingma D., and Ba, J., Adam: a method for stochastic optimization. In International Conference on Learning Representations (ICLR). 2014.
[75] Vaswani A., Shazeer N., Parmar N., et al. Attention is All You Need. in Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. Long Beach, California, USA: Curran Associates Inc.
[76] Devlin J., Chang M.-W., Lee K., et al. BERT: Pre-training of deep bidirectional transformers for language understanding. in Proceedings of NAACL-HLT. 2019.
[77] Al-Rfou R, Choe D, Constant N, et al. Character-level language modeling with deeper self-attention. *AAAI Conf Artif Intell* 2019.
[78] Maruf S., Martins A.F.T. and Haffari G. Selective Attention for Context-aware Neural Machine Translation. in Proceedings of NAACL-HLT. 2019. Minneapolis, Minnesota.
[79] Dai Z, Yang Z, Yang Y, et al. Transformer-XL: attentive language models beyond a fixed-length context. *Proc 57th Annu Meet Assoc Comput Linguist* 2019.
[80] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8): 1735–80.
[81] Kim Y. Convolutional neural networks for sentence classification. arXiv 2014.
[82] Cheng J, Dong L, Lapata M. Long short-term memory-networks for machine reading. arXiv 2016.

[83] Huang Z., Xu W. and Yu K., Bidirectional LSTM-CRF Models for Sequence Tagging. 2015, arXiv.
[84] Sutskever I., Vinyals O. and Le Q V., Sequence to Sequence Learning with Neural Networks. 2014, arXiv.
[85] Bahdanau D., Cho K. and Bengio Y., Neural Machine Translation by Jointly Learning to Align and Translate. 2016, arXiv.
[86] Cho K., van Merrienboer B., Gulcehre C., et al., Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. 2014, arXiv.
[87] Zaremba W., Sutskever I. and Vinyals O., Recurrent Neural Network Regularization. 2015, arXiv.
[88] Jozefowicz R., Vinyals O., Schuster M., et al., Exploring the Limits of Language Modeling. 2016, arXiv.
[89] Huang D, Chen K, Song B, et al. Geographic encoding of transcripts enabled high-accuracy and isoform-aware deep learning of RNA methylation. Nucleic Acids Res 2022;50(18):10290–310.
[90] Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56–67.
[91] Wang G, Liu T, Lyu H, Liu Z. F5C-finder: An Explainable and Ensemble Biological Language Model for Predicting 5-Formylcytidine Modifications on mRNA. arXiv, 2024.
[92] Lyu R., Pajdzik K., Sun H.-L., et al., A Quantitative Sequencing Method for 5-Formylcytosine in RNA. 2024. 64(3–4): p. e202300111.