Molecular Therapy
Nucleic Acids
Original Article

Staem5: A novel computational approach for accurate prediction of m5C site

Di Chai,[1] Cangzhi Jia,[1] Jia Zheng,[1] Quan Zou,[2] and Fuyi Li[3]

[1]School of Science, Dalian Maritime University, Dalian 116026, China; [2]Yangtze Delta Region Institute (Quzhou), Quzhou, China; [3]Department of Microbiology and Immunology, The Peter Doherty Institute for Infection and Immunity, The University of Melbourne, 792 Elizabeth Street, Melbourne, VIC 3000, Australia

5-Methylcytosine (m5C) is an important post-transcriptional modification that has been extensively found in multiple types of RNAs. Many studies have shown that m5C plays vital roles in many biological functions, such as RNA structure stability and metabolism. Computational approaches act as an efficient way to identify m5C sites from high-throughput RNA sequence data and help interpret the functional mechanism of this important modification. This study proposed a novel species-specific computational approach, Staem5, to accurately predict RNA m5C sites in *Mus musculus* and *Arabidopsis thaliana*. Staem5 was developed by employing feature fusion tactics to leverage informatic sequence profiles, and a stacking ensemble learning framework combined five popular machine learning algorithms. Extensive benchmarking tests demonstrated that Staem5 outperformed state-of-the-art approaches in both cross-validation and independent tests. We provide the source code of Staem5, which is publicly available at https://github.com/Cxd-626/Staem5.git.

INTRODUCTION

There are more than 170 types of RNA chemical modifications (RCMs) that have been found in transfer RNAs (tRNAs), ribosome RNAs (rRNAs), mRNAs, and non-coding RNAs.[1–5] The RCMs is determined by three coordinating factors, including methyltransferase, RNA binding protein, and demethylase.[3,6,7] Among all RCMs, 5-cytosine-methylation (m5C) is one of the most important modifications in mRNA. However, it is challenging to identify m5C accurately. Because of the instability of mRNA molecules, high-throughput sequencing technologies usually fail to accurately identify m5C sites at single-nucleotide resolution.[6,8–10] Therefore, computational approaches that can accurately identify m5C sites would be highly valuable and may provide insights into the functional roles of this important RNA modification.

A number of computational approaches based on sequence-derived information and machine learning algorithms have been developed to predict m5C sites of four species, including *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae*, and *Arabidopsis thaliana*. These approaches can be classified into two categories according to the machine learning algorithm they applied: (1) support vector machine-based predictors, including m5C-PseDNC,[11] M5C-HPCR,[12] pM^{5}CS-Comp-mRMR,[13] RNAm5CPred,[14] m5CPred-SVM,[15] and iRNA-m5C_SVM[16]; (2) random forest (RF)-based approaches, including PEA-m5C,[17] RNAm5Cfinder,[18] and iRNA-m5C.[19] In addition, some studies developed computational methods for predicting multiple types of RNA modifications, including m5C. For example, Liu and Chen developed iMRM[20] based on extreme gradient boosting (XGBoost) to recognize five types of RNA modifications. Song et al.[8] developed an attention-based multi-label neural network, MultiRM, to predict 12 types of RNA modifications simultaneously. Table 1 summarizes these two categories of predictors specially designed for m5C in several aspects, including the feature extraction, performance evaluation strategy, species, webserver or software availability, and benchmark datasets. We found that most of the methods were developed for *H. sapiens*, and only a few predictors were designed and tested for m5C sites of *M. musculus* and *A. thaliana*, such as iRNA-m5C, iRNA-m5C_SVM, RNAm5Cfinder, and m5CPred-SVM.[15,16,19] In addition, the predictive performance of the m5C site in *M. musculus* and *A. thaliana* is unsatisfactory compared with that in *H. sapiens*. For example, m5CPred-SVM, iRNA-m5C_SVM, and iRNA-m5C were developed on the same benchmark dataset of *A. thaliana* and achieved 71.8%, 73.06%, and 70.7% in terms of average accuracy of the cross-validation tests, respectively. The reason is probably that these predictors were developed based on a single RF or SVM algorithm. With recent advances in ensemble learning strategies used in bioinformatics to develop robust prediction models, we were motivated to leverage the ensemble learning techniques to improve m5C prediction in *M. musculus* and *A. thaliana*.

In this study, we introduce Staem5, a stacked ensemble model for predicting m5C sites in *A. thaliana* and *M. musculus*. Staem5 was developed based on four types of sequence features, such as position-specific propensity, *k*-mer, electron-ion interaction pseudo potentials of trinucleotide, and parallel correlation pseudo dinucleotide composition. The base models to build the optimal stacked model of

Received 5 July 2021; accepted 6 October 2021;
https://doi.org/10.1016/j.omtn.2021.10.012.

Correspondence: Fuyi Li, PhD, Department of Microbiology and Immunology, The Peter Doherty Institute for Infection and Immunity, The University of Melbourne, 792 Elizabeth Street, Melbourne, VIC 3000, Australia.
E-mail: fuyi.li@unimelb.edu.au
Correspondence: Cangzhi Jia, PhD, School of Science, Dalian Maritime University, Dalian 116026, China.
E-mail: cangzhijia@dlmu.edu.cn

Table 1. A comprehensive list of the existing methods for m5C site prediction

Method classification	Tools[a]	Webserver/software availability[b]	Features	Evaluation strategy	Species	Benchmark dataset (positive: negative)
SVM-based	m5C-PseDNC[11]	no	PseDNC$^{(3)}$	Jackknife test	*H. sapiens*	120: 120
	M5C-HPCR[12]	no	PseDNC$^{(23)}$	Jackknife test	*H. sapiens*	Met1320 (120: 1,200); Met1900 (475: 1,425)
	pM^{5}CS-Comp-mRMR[13]	no	DNC, TriNC, TetraNC	Jackknife test	*H. sapiens*	120:120
	RNAm5CPred[14]	yes	KNF, KSNPF, PseDNC$^{(3)}$	Jackknife test, 10-fold CV and independent test	*H. sapiens*	Met935 (127: 808); Met240 (120: 120) Met1900 (475: 1,425) Test1157 (157: 1,000)
	m5CPred-SVM[15]	decommissioned	KNF, PseDNC$^{(3)}$, KSNPF, PSNP, KSPSDP, CPD	10-fold CV and independent test	*H. sapiens* *M. musculus* *A. thaliana*	269: 269 5,563: 5,563 6,289: 6,289
	iRNA-m5C_SVM[16]	no	PSP, k-mer, PseEIIP, PCPseDNC$^{(22)}$	10-fold CV and independent test	*A. thaliana*	6,289: 6,289
RF-based	iRNAm5C-PseDNC[21]	yes	PseDNC(10)	Jackknife test	*H. sapiens*	475: 1,425
	PEA-m5C[17]	decommissioned	binary, k-mer, PseDNC$^{(3)}$	10-fold CV and independent test	*A. thaliana*	DatasetCV (1,196: 11,960) DatasetHT (100: 100) DatasetT1 (79: 79) DatasetT2 (73: 73)
	RNAm5Cfinder[18]	yes	binary	10-fold CV and independent test	*H. sapiens* *M. musculus*	Three m5C datasets from GSE90963, GSE93749, GSE83432 database
	iRNA-m5C[19]	yes	k-mer, binary, natural vector, PseKNC	Jackknife test, 10-fold CV and independent test	*H. sapiens* *M. musculus* *S. cerevisiae* *A. thaliana*	120: 120 97: 97 211: 211 6,289: 6,289

CV, cross-validation; PseDNC$^{(m)}$, pseudo dinucleotide composition, m is the number of physical-chemical properties; DNC, dinucleotide; TriNC, trinucleotide; TetraNC, tetranucleotide; KNF, K-nucleotide frequency; KSNPF, K-spaced nucleotide pair frequency; PseKNC, pseudo K-tuple nucleotide frequency component; KSNPF, K-spaced nucleotide pair frequency; PSNP, position-specific nucleotide propensity; KSPSDP, K-spaced position-specific dinucleotide propensity; CPD, chemical property with density; PSP, PSNP, PSDP, and PSTP, associated with frequencies of nucleotides, dinucleotides, and trinucleotides; PseEIIP, electron-ion interaction pseudo potential of trinucleotide; PCPseDNC, general parallel correlation pseudo dinucleotide composition.

[a]The URL addresses for the listed tools are as follows: iRNAm5C-PseDNC, http://www.jci-bioinfo.cn/iRNAm5C-PseDNC; PEA-m5C, https://github.com/cma2015/PEA-m5C;RNAm5Cfinder, http://www.rnanut.net/rnam5cfinder; RNAm5CPred, http://zhulab.ahu.edu.cn/RNAm5CPred/; iRNA-m5C, http://lin-group.cn/server/iRNA-m5C/service.html; m5CPred-SVM, https://zhulab.ahu.edu.cn/m5CPred-SVM.

[b]Yes: the publication is accompanied with a webserver/softpackage and it is still functional; decommissioned: the webserver/softpackage is no longer available; no: the publication has no webserver or softpackage.

each species were selected from five popular machine learning algorithms, and the feature selection strategies were employed to further optimize the predictive performance. The cross-validation and independent tests demonstrate that Staem5 achieved competitive predictive performance compared with state-of-the-art approaches.

RESULTS

In this work, we propose a novel computational method, Staem5, to identify m5C sites for both *A. thaliana* and *M. musculus*. The model integrates four kinds of encoding schemes, i.e., position-specific propensity (PSP), *k*-mer (k = 1, 2, and 3), parallel correlation pseudo dinucleotide composition (PCPseDNC), and electron-ion interaction pseudo potentials of trinucleotide (PseEIIP). Bayesian optimization was applied to tune parameters for each classifier. Then, we evaluated the different combinations of base classifiers, including SVM, XGBoost, light gradient boosting machine (LightGBM), extremely randomized trees (ExtraTree), and gradient boosting decision tree (GBDT), by stacked tactics to identify the optimal ensemble model for *A. thaliana* and *M. musculus*, respectively. Meanwhile, the F score is used to reduce the dimension of features and computing time. Compared with training and independent datasets, Staem5 exhibits its superiority to other existing approaches. The source code of Staem5 can be found at https://github.com/Cxd-626/Staem5.git.

DISCUSSION

Nucleotide preferences of the m5C site

This section analyzes the nucleotide preferences of the sequence fragments containing m5C sites using the Two Sample Logo (http://www.twosamplelogo.org/).[22] The sequence logos of *A. thaliana* and *M. musculus* generated by Two Sample Logo are presented in Figures 1A and 1B, respectively. As observed, cytidine (C) was enriched upstream of the m5C sites of *A. thaliana*, especially at positions −18 to −10 and −7 to −1. In contrast, adenine (A) and guanine (G) are abundant upstream of the non-m5C sequence fragments, especially at positions −19, −18, −15, −12, −11, −9, −7, −6, −3, and −1. For *M. musculus*, C and G have relatively higher frequencies than

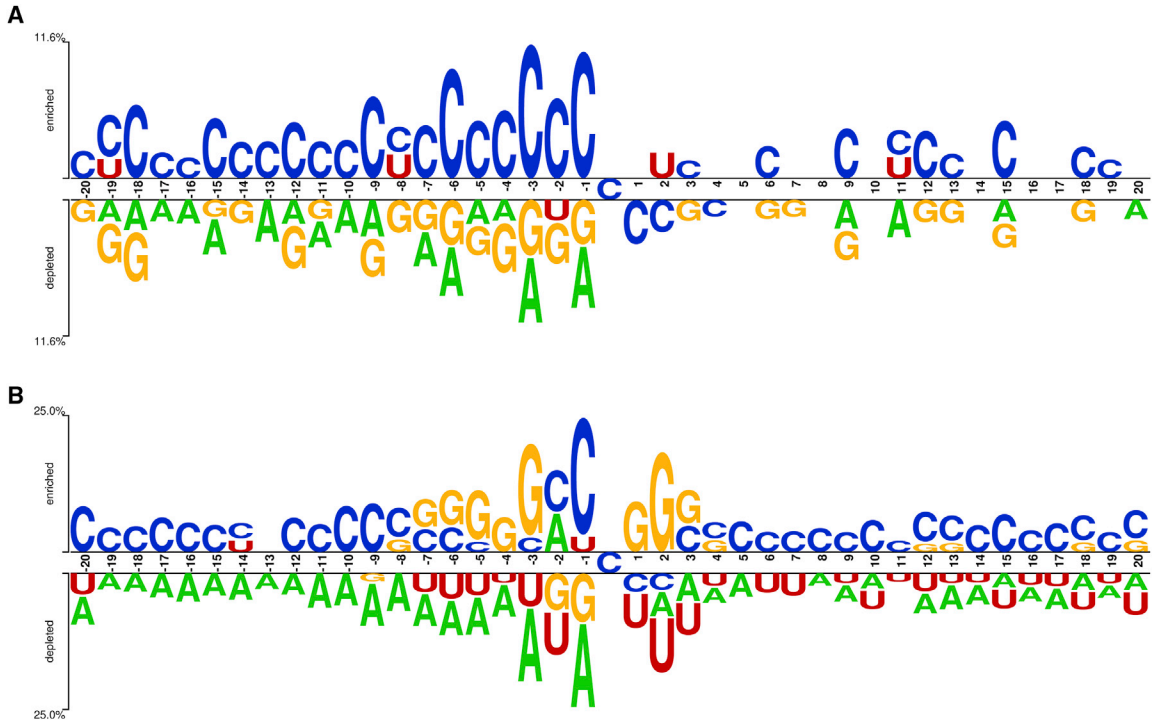

Figure 1. Nucleotide preferences of the fragments with m5C sites and without m5C sites in the center
(A) For *A. thaliana* and (B) for *M. musculus*.

the other two nucleic acids, especially at positions −20, −10, −9, −5, −3, −2, −1, and 2. Also, the non-m5C sequences had a frequent A and uridine (U) pattern at positions −9, −6, −5, −3 to −1, and 1 to 3 of the corresponding sequence segments. These results demonstrate that m5C sites in *A. thaliana* and *M. musculus* do not have notable sequence motifs compared with non-m5C sites, and that the sequence segments have different nucleotide preferences in these two species. Therefore, it could be difficult to develop a general model for cross-species prediction, and it is necessary to set up species-specific models.

The effectiveness of parameter optimization

In this section, we evaluate the predictive performance of five popular machine learning algorithms, i.e., SVM,[23] GBDT,[24] XGBoost,[25] LightGBM,[26] and ExtraTree,[27] for m5C site prediction in *A. thaliana* and *M. musculus*. For each classification algorithm, the hyperparameters were pre-set according to previous experience[28–30] and optimized by the Bayesian optimization,[31] which has effectiveness in many prediction tasks in bioinformatics.[29,30,32–37] We searched the optimal combination of hyperparameters according to the value of accuracy based on the 10-fold cross-validation tests. The performance comparison results in terms of accuracy of the five base classifiers before and after parameter optimization on the 10-fold cross-validation tests are shown in Figure 2 (the detailed values of other performance metrics are provided in Table S1), and the selected parameters are listed in Table S2. We can observe that the performance of the five base classifiers enhanced after parameter optimization and performance improvement of SVM was the largest among the five base classifiers. The accuracy of the SVM model of *A. thaliana* increased from 62.40% to 73.62%. In addition, the accuracy of the GBDT model also witnessed an increase from 65.69% to 71.77%. In comparison, ExtraTree had the most negligible performance improvement with a 0.05% increase in terms of accuracy.

In addition, from Figure 2 and Table S1, we can see that SVM achieved the best accuracy (73.62%), Matthew's correlation coefficient (MCC) (0.476), specificity (Sp) (79.41%), Pre (76.71%), and AUC (0.807); while LightGBM achieved the best sensitivity (Sn) (71.19%); and XGBoost secured the best F1 score (0.723) among these five classifiers for *A. thaliana*. Considering all five performance metrics, SVM achieved the best predictive performance on the 10-fold cross-validation tests for *A. thaliana*. In comparison, XGBoost achieved the best accuracy (76.68%), MCC (0.534), and F1 score (0.769) for *M. musculus*.

Processing of building ensemble model

There are two levels in the stacking ensemble learning strategy, and the classifiers in these two levels are referred to as base and meta-classifiers, respectively. In the first level, a set of base classifiers generate the probability values, which are subsequently used as the input for the meta-classifier. In this study, we used logistic regression as the meta-model to ensemble the base classifiers into a stacked model.

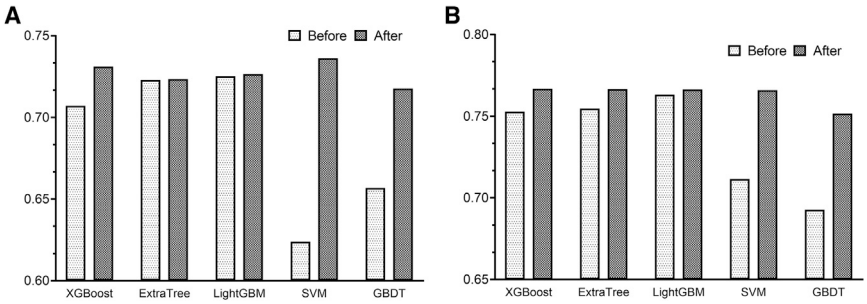

Figure 2. Performance comparison results in terms of accuracy of the five base classifiers before and after parameter optimization on the 10-fold cross-validation tests

(A) For *A. thaliana* and (B) for *M. musculus*.

The stacking strategy was implemented in the "*mlxtend*" package[38] in Python. The selection of base classifiers was based on the accuracy of the model. Take *A. thaliana* for an example, and the process of stacking is as follows: we first ensembled the top two best-performed classifiers, SVM and XGBoost, and evaluated whether the model performance in accuracy increased or not. We identified that the stacked model achieved increased accuracy compared with support vector machine (SVM) only from 73.62% to 73.85% on the 10-fold cross-validation. Therefore, we further integrated the third-ranked classifier LightGBM to the stacked model, and the accuracy further improved by 73.85%–73.89%. However, when combined with the fourth-ranked classifier GBDT and ExtraTree, the accuracy decreased in varying degrees. Therefore, we accordingly selected SVM, XGBoost, and LightGBM as the base classifiers for the stacked model, which achieved 73.89% accuracy and 0.479 MCC. Figure 3 illustrates the performance comparison results of different base classifiers' combinations in accuracy and MCC (the detailed results are provided in Table S3).

Subsequently, we also compared the stacked strategy with the voting strategy, which is another popular ensemble learning strategy. To ensure the fairness of the comparison, the voting models were constructed according to the same principle as the stacked model (with logistic regression). The performance comparison results of different classifier combinations are provided in Table S4, and we summarize the performance comparison results between the best stacking and voting models in Table 2. The results demonstrated that the stacking model achieved better predictive performance, which is more suitable for m5C site prediction in *A. thaliana* and *M. musculus*.

Feature selection analysis

To remove redundant information caused by high-dimensional input features and further optimize the meta-models, we evaluated three popular feature selection algorithms, including maximum-relevance-maximum-distance (MRMD),[39] Pearson correlation coefficient (PCC) feature selection,[40] and F score, to find the optimal feature subset. We first ranked all features by each feature selection algorithm and then reduced the dimension of the feature set by step of 50. The performance comparison results of three feature selection algorithms are provided in Table S5. The results suggested that these three feature selection approaches did not further improve the predictive performance of m5C sites in *A. thaliana*; however, the selected features enhanced the model performance of the *M. musculus* model on 10-fold cross-validation tests. For the model performance during feature selection by F score, the average accuracy first increased and then decreased with the decrease of features, and at the best average accuracy 77.26% at the feature dimension of 180. In contrast, the best average accuracies for MRMD and PCC are same, which were achieved to 77.21% at the dimension of 230 for MRMD and 280 for PCC, respectively. These results demonstrated F score achieved slightly better performance compared with MRMD and PCC. Therefore, we used F score to select the optimal features and reduce the feature dimension by setting a smaller step of 5 and provided the feature selection results at the feature dimension of 165–195 in Table S6. From Table S6, we can see the best performance in terms of accuracy (77.42%) and AUC (0.855) achieved with 185 features. Finally, we further selected the optimal feature subsets by step of 1 on the feature dimension of 175–190 and report the results in Table S7. The results doubly confirm that the feature subset with 185 features can secure the best performance in accuracy and AUC. Therefore, these 185 features were used as the input features for the stacked model to predict m5C sites in *M. musculus*. In addition, the performance comparison

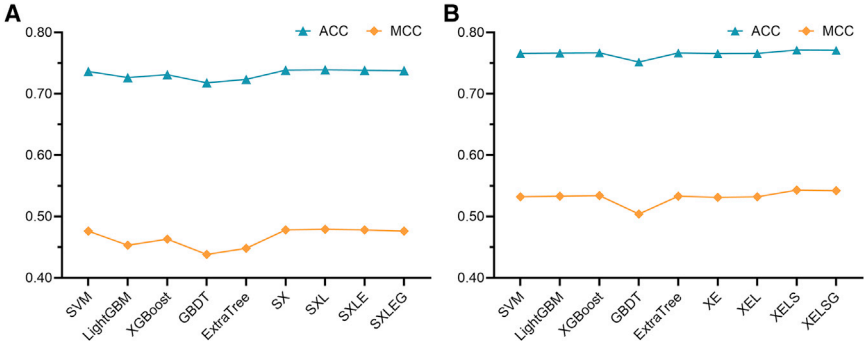

Figure 3. Performance comparison results of different base classifiers' combinations in accuracy and MCC

(A) For *A. thaliana* and (B) for *M. musculus*. Note: SX, SVM + XGBoost; SXL, SVM + XGBoost + LightGBM; SXLE, SVM + XGBoost + LightGBM + ExtraTree; SXLEG, SVM + XGBoost + LightGBM + ExtraTree + GBDT; XE, XGBoost + ExtraTree; XEL, XGBoost + ExtraTree + LightGBM; XELS, XGBoost + ExtraTree + LightGBM + SVM; XELSG, XGBoost + ExtraTree + LightGBM + SVM + GBDT.

Table 2. Performance comparison results between two ensemble strategies in 10-fold cross-validation tests on the training dataset

Species	Strategy	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
A. thaliana	voting strategy	70.85	76.52	73.68	0.474	0.809
	stacking	**70.98**	**76.80**	**73.89**	**0.479**	**0.810**
M. musculus	voting strategy	**78.11**	75.56	76.83	0.537	0.852
	stacking	77.97	**76.26**	**77.12**	**0.543**	**0.854**

Bold numbers indicate the highest values in each column.

Table 3. Performance comparison results between Staem5 and existing methods in 10-fold cross-validation tests on the training dataset

Species	Methods	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
A. thaliana	iRNA-m5C	65.70	75.70	70.70	0.420	0.770
	m5CPred-SVM	68.10	75.50	71.80	0.437	0.782
	iRNA-m5C_SVM	66.42	**79.70**	73.06	0.470	0.800
	Staem5	**70.98**	76.80	**73.89**	**0.479**	**0.810**
M. musculus	m5CPred-SVM	75.70	72.80	74.30	0.486	0.822
	Staem5	**78.28**	**76.55**	**77.42**	**0.549**	**0.855**

Bold numbers indicate the highest values in each column.

results of the models before and after selection on the independent test dataset are provided in Table S8.

Performance comparison with state-of-the-art methods

In this section, we compare the predictive performance of Staem5 with several state-of-the-art predictors on the same training and independent test datasets of *A. thaliana* and *M. musculus*. For *A. thaliana*, we compared Staem5 with iRNA-m5C,[19] m5CPred-SVM,[15] and iRNA-m5C_SVM[16]; while for *M. musculus*, we compared Staem5 with m5CPred-SVM. The performance comparison results on training and independent test datasets are provided in Tables 3 and 4, respectively. From Table 3, we can see that Staem5 achieved the best performance on the training dataset of both *A. thaliana* and *M. musculus* for almost all the evaluation metrics with the only exception that iRNA-m5C_SVM achieved the best Sp of *A. thaliana*. The independent test results in Table 4 show that Staem5 was inferior to iRNA-m5C_SVM and m5CPred-SVM on the independent test set of *A. thaliana*. However, Staem5 achieved better predictive performance than m5CPred-SVM on the independent test set of *M. musculus*. Although the Staem5's performance on the independent test set of *A. thaliana* was slightly lower than iRNA-m5C_SVM and m5CPred-SVM, the training and testing performance differences were lower compared with these two approaches. The independent test results of iRNA-m5C_SVM and m5CPred-SVM were much higher than their performance on the training dataset. Instead, Staem5 showed similar performance on the independent dataset and training dataset, e.g., 73.70% versus 73.89% in terms of accuracy, which indicates that Staem5 is more robust and stable compared with others. Therefore, we can conclude that Staem5 can accurately predict *M. musculus* and *A. thaliana* m5C sites.

MATERIALS AND METHODS

Benchmark datasets

The schematic flowchart of Staem5 is shown in Figure 4. There are four major steps, including data collection, feature extraction, feature selection, and model construction. In the first step, the training and independent test datasets of *A. thaliana* were collected from the datasets constructed by Chen et al.[15] The m5C site data of *A. thaliana* was derived from the NCBI Gene Expression Omnibus (GEO) database http://www.ncbi.nlm.nih.gov/geo/ using accession number GEO: GSE94065 http://www.ncbi.nlm.nih.gov/geo/, while the *M. musculus* dataset was collected from Yang et al.[6] A statistical summary of the training and independent test datasets of *A. thaliana* and *M. musculus* is provided in Table S9. The *A. thaliana* dataset contains 5,298 positive and 5,298 negative training samples, 1,000 positive and 1,000 negative testing samples. In comparison, the *M. musculus* dataset has 4,563 positive and 4,563 negative training samples, 1,000 positive and 1,000 negative testing samples.

Sequence encoding schemes

In this study, we employed four types of sequence encoding schemes, including parallel correlation pseudo dinucleotide composition (PCPseDNC), position-specific propensity (PSP), *k*-mer, and electron-ion interaction pseudo potentials of trinucleotide (PseEIIP). PCPseDNC was calculated by *iLearn*,[41] and there are 38 physicochemical properties in PCPseDNC. PSP, *k*-mer and PseEIIP have been extensively applied in the field of prediction RNA N6-methyladenosine (m6A) sites, protein S-sulfenylation sites, and identifying N4-acetylcytidine (ac4C) sites in mRNA.[42–45] We provide the detailed definitions and formulas in the supplemental information.

Stacked ensemble learning framework

There are two levels in the stacking ensemble learning strategy, and the classifiers in these two levels are referred to as base classifiers and meta-classifier, respectively. In this work, we explored five popular machine learning algorithms, including SVM,[23] GBDT,[24] XGBoost,[25] LightGBM,[26] and ExtraTree,[27] as the base classifiers, and applied the logistic regression[46] algorithm as the meta-classifier to build the stacked ensemble model. The base classifiers were built using the scikit-learn package,[47,48] and the model stacking was implemented using the "*mlxtend*" package.[38]

In this study, we employed the radial basis kernel function in SVM and optimized the regularization parameter C and kernel parameter γ to find the most suitable hyperparameters.[14,23,49] GBDT is a tree-based boosting algorithm that learns directly from mistake residual errors rather than updating the weight of the data. It uses the gradient descent algorithm to minimize training error.[24,50] XGBoost improves GBDT by employing parallel learning techniques and regularization terms, which makes the model more efficient and robust. XGBoost achieved great success in many bioinformatics tasks, such

Table 4. Performance comparison results between Staem5 and existing methods on the independent test dataset

Species	Classifier	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
A. thaliana	iRNA-m5C	72.40	75.60	74.10	0.481	–
	m5CPred-SVM	75.50	76.10	75.50	0.516	0.836
	iRNA-m5C_SVM	**79.40**	**80.90**	**80.15**	**0.600**	**0.880**
	Staem5	74.80	72.60	73.70	0.474	0.829
M. musculus	m5CPred-SVM	**67.90**	74.90	71.40	0.429	0.775
	Staem5	66.10	**77.80**	**71.95**	**0.442**	**0.787**

Bold numbers indicate the highest values in each column.

as protein/DNA/RNA functional sites prediction.[25,51–54] LightGBM is a further extension of XGBoost, which improves training speed and reduce memory consumption by applying a histogram algorithm.[26] In addition, LightGBM proposes gradient-based one-side sampling, exclusive feature bundling, and leaf-wise growth strategy to obtain better accuracy and efficient computation. Meanwhile, it also adopted limiting maximum depth parameters to mitigate over-fitting[55,56] and LightGBM has been widely used in bioinformatics.[57,58] ExtraTree is also a tree-based algorithm that was proposed by Pierre Geurts et al.[27] in 2006. Although ExtraTree is very similar to RF, there are two major differences between them. First, RF is a bagging method, while ExtraTree uses all the training samples to train the decision tree. Second, the RF gets the best bifurcation feature in a random subset; while ExtraTree performs a completely random bifurcation.[59]

Model evaluation

To evaluate and compare Staem5 with existing approaches, 10-fold cross-validation and independent tests were conducted based on training and testing datasets, respectively. We applied five commonly used evaluation metrics for model evaluation, including Sn, Sp, accuracy (Acc), MCC, and area under the receiver operating characteristic curve (AUC), defined as:

$$Sn = \frac{TP}{TP + FN}, \quad \text{(Equation 1)}$$

$$Sp = \frac{TN}{TN + FP}, \quad \text{(Equation 2)}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \quad \text{(Equation 3)}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}, \quad \text{(Equation 4)}$$

where TP, TN, FP, and FN indicate the number of true-positive, true-negative, false-positive, and false-negative sequences, respectively.

Experimental environment

The experiments were conducted on a PC with a 64-bit Windows 10 operating system. The PC is equipped with an Intel(R) Core (TM) i7-7700 CPU and 16 GB physical memory; the CPU's main frequency is 3.60 GHz. Staem5 was developed based on Python 3.7,

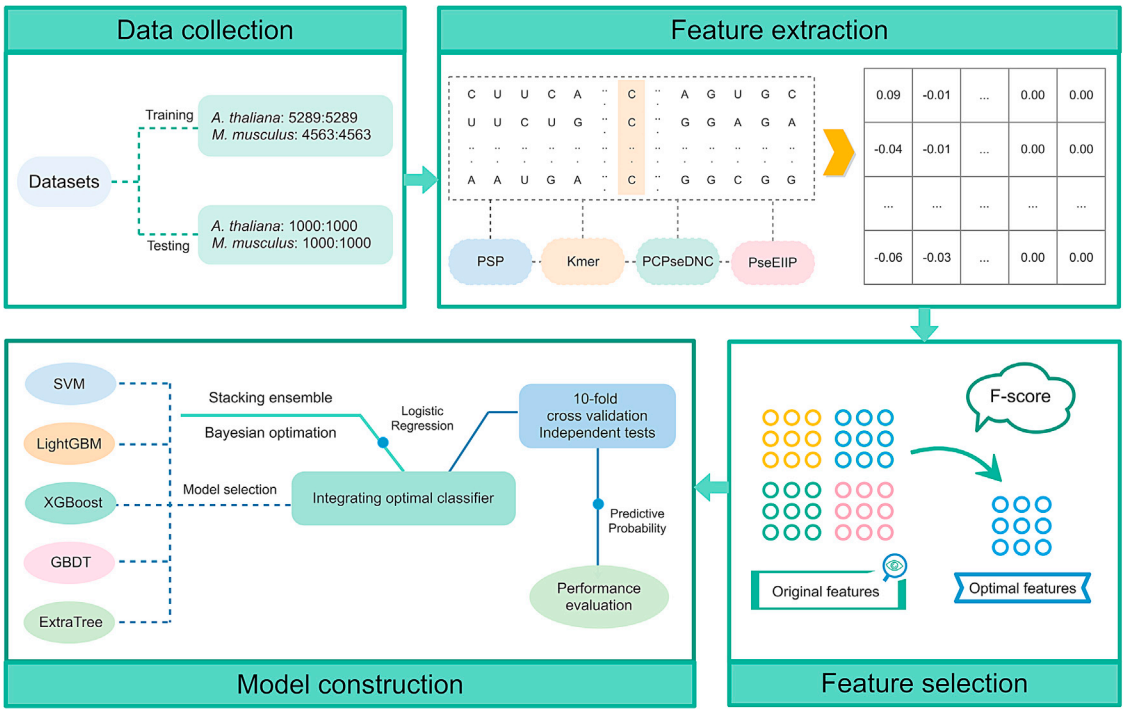

Figure 4. The schematic flowchart of Staem5

and it requires approximately 5.7 s to predict 1,000 enquiry sequence segments with 41 bp.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2021.10.012.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China no. 62071079.

AUTHOR CONTRIBUTIONS

C.J. and F.L. conceived the initial idea and designed the methodology. J.Z. and D.C. implemented the algorithm, conducted the experiments, and processed the results. All authors drafted, revised, and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Xuan, J.J., Sun, W.J., Lin, P.H., Zhou, K.R., Liu, S., Zheng, L.L., Qu, L.H., and Yang, J.H. (2018). RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. Nucleic Acids Res. *46*, D327–D334. https://doi.org/10.1093/nar/gkx934.

2. Dubin, D.T., and Taylor, R.H. (1975). The methylation state of poly A-containing messenger RNA from cultured hamster cells. Nucleic Acids Res. *2*, 1653–1668. https://doi.org/10.1093/nar/2.10.1653.

3. Frye, M., Harada, B.T., Behm, M., and He, C. (2018). RNA modifications modulate gene expression during development. Science *361*, 1346–1349. https://doi.org/10.1126/science.aau1646.

4. Squires, J.E., Patel, H.R., Nousch, M., Sibbritt, T., Humphreys, D.T., Parker, B.J., Suter, C.M., and Preiss, T. (2012). Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Res. *40*, 5023–5033. https://doi.org/10.1093/nar/gks144.

5. Boccaletto, P., Machnicka, M.A., Purta, E., Piatkowski, P., Baginski, B., Wirecki, T.K., de Crecy-Lagard, V., Ross, R., Limbach, P.A., Kotter, A., et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res. *46*, D303–D307.

6. Yang, X., Yang, Y., Sun, B.F., Chen, Y.S., Xu, J.W., Lai, W.Y., Li, A., Wang, X., Bhattarai, D.P., Xiao, W., et al. (2017). 5-methylcytosine promotes mRNA export-NSUN2 as the methyltransferase and ALYREF as an m(5)C reader. Cell Res. *27*, 606–625. https://doi.org/10.1038/cr.2017.55.

7. Zheng, G.Q., Dahl, J.A., Niu, Y.M., Fedorcsak, P., Huang, C.M., Li, C.J., Vagbo, C.B., Shi, Y., Wang, W.L., Song, S.H., et al. (2013). ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. Mol. Cell *49*, 18–29. https://doi.org/10.1016/j.molcel.2012.10.015.

8. Song, Z., Huang, D., Song, B., Chen, K., Song, Y., Liu, G., Su, J., Magalhaes, J.P.d., Rigden, D.J., and Meng, J. (2021). Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. Nat. Commun. *12*, 4011. https://doi.org/10.1038/s41467-021-24313-3.

9. Khoddami, V., and Cairns, B.R. (2013). Identification of direct targets and modified bases of RNA cytosine methyltransferases. Nat. Biotechnol. *31*, 458. https://doi.org/10.1038/nbt.2566.

10. Hussain, S., Tuorto, F., Menon, S., Blanco, S., Cox, C., Flores, J.V., Watt, S., Kudo, N.R., Lyko, F., and Frye, M. (2013). The mouse cytosine-5 RNA methyltransferase NSun2 is a component of the chromatoid body and required for testis differentiation. Mol. Cell. Biol. *33*, 1561–1570. https://doi.org/10.1128/mcb.01523-12.

11. Feng, P., Ding, H., Chen, W., and Lin, H. (2016). Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions. Mol. Biosyst. *12*, 3307–3311. https://doi.org/10.1039/c6mb00471g.

12. Zhang, M., Xu, Y., Li, L., Liu, Z., Yang, X.B., and Yu, D.J. (2018). Accurate RNA 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble. Anal. Biochem. *550*, 41–48. https://doi.org/10.1016/j.ab.2018.03.027.

13. Sabooh, M.F., Iqbal, N., Khan, M., Khan, M., and Maqbool, H.F. (2018). Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC. J. Theor. Biol. *452*, 1–9. https://doi.org/10.1016/j.jtbi.2018.04.037.

14. Fang, T., Zhang, Z.Z., Sun, R., Zhu, L., He, J.J., Huang, B., Xiong, Y., and Zhu, X.L. (2019). RNAm5CPred: prediction of RNA 5-methylcytosine sites based on three different kinds of nucleotide composition. Mol. Ther. Nucleic Acids *18*, 739–747. https://doi.org/10.1016/j.omtn.2019.10.008.

15. Chen, X., Xiong, Y., Liu, Y.B., Chen, Y.Q., Bi, S.D., and Zhu, X.L. (2020). m5CPred-SVM: a novel method for predicting m5C sites of RNA. BMC Bioinformatics *21*, 489. https://doi.org/10.1186/s12859-020-03828-4.

16. Dou, L.J., Li, X.L., Ding, H., Xu, L., and Xiang, H.K. (2020). Prediction of m5C modifications in RNA sequences by combining multiple sequence features. Mol. Ther. Nucleic Acids *21*, 332–342. https://doi.org/10.1016/j.omtn.2020.06.004.

17. Song, J., Zhai, J., Bian, E., Song, Y., Yu, J., and Ma, C. (2018). Transcriptome-wide annotation of m(5)C RNA modifications using machine learning. Front. Plant Sci. *9*, 519. https://doi.org/10.3389/fpls.2018.00519.

18. Li, J.W., Huang, Y., Yang, X.Y., Zhou, Y.R., and Zhou, Y. (2018). RNAm5Cfinder: a web-server for predicting RNA 5-methylcytosine (m5C) sites based on random forest. Sci. Rep. *8*, 17299. https://doi.org/10.1038/s41598-018-35502-4.

19. Lv, H., Zhang, Z.M., Li, S.H., Tan, J.X., Chen, W., and Lin, H. (2020). Evaluation of different computational methods on 5-methylcytosine sites identification. Brief Bioinform. *21*, 982–995. https://doi.org/10.1093/bib/bbz048.

20. Liu, K.W., and Chen, W. (2020). iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications. Bioinformatics *36*, 3336–3342. https://doi.org/10.1093/bioinformatics/btaa155.

21. Qiu, W.R., Jiang, S.Y., Xu, Z.C., Xiao, X., and Chou, K.C. (2017). iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget *8*, 41178–41188. https://doi.org/10.18632/oncotarget.17104.

22. Vacic, V., Iakoucheva, L.M., and Radivojac, P. (2006). Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics *22*, 1536–1537. https://doi.org/10.1093/bioinformatics/btl151.

23. Cortes, C., Cortes, C., Vapnik, V., Llorens, C., Vapnik, V.N., Cortes, C., and Côrtes, M. (1995). Support-vector networks[J].

24. Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Ann. Stat. *29*, 1189–1232. https://doi.org/10.1214/aos/1013203451.

25. Chen, T.Q., Guestrin, C., and Assoc Comp, M. (2016). XGBoost: A Scalable Tree Boosting System (Assoc Computing Machinery). https://doi.org/10.1145/2939672.2939785.

26. Ke, G.L., Meng, Q., Finley, T., Wang, T.F., Chen, W., Ma, W.D., et al. (2017). LightGBM: a highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems 30, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, and S. Vishwanathan, et al., eds. (Neural Information Processing Systems (Nips)), pp. 1–9.

27. Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. Mach. Learn. *63*, 3–42. https://doi.org/10.1007/s10994-006-6226-1.

28. Yi, H.C., You, Z.H., Wang, M.N., Guo, Z.H., Wang, Y.B., and Zhou, J.R. (2020). RPI-SE: a stacking ensemble learning framework for ncRNA-protein interactions prediction using sequence information. BMC Bioinformatics *21*, 60. https://doi.org/10.1186/s12859-020-3406-0.

29. Li, F., Chen, J., Ge, Z., Wen, Y., Yue, Y., Hayashida, M., Baggag, A., Bensmail, H., and Song, J. (2021). Computational prediction and interpretation of both general and specific types of promoters in *Escherichia coli* by exploiting a stacked ensemble-learning framework. Brief Bioinform. *22*, 2126–2140. https://doi.org/10.1093/bib/bbaa049.

30. Mei, S., Li, F., Xiang, D., Ayala, R., Faridi, P., Webb, G.I., Illing, P.T., Rossjohn, J., Akutsu, T., Croft, N.P., et al. (2021). Anthem: a user customised tool for fast and accurate prediction of binding between peptides and HLA class I molecules. Brief Bioinform. https://doi.org/10.1093/bib/bbaa415.

31. Snoek, J., Larochelle, H., and Adams, R.P. (2012). Practical Bayesian optimization of machine learning algorithms. Adv. Neural Inf. Process. Syst. *4*, 1–9.

32. Zhu, Y., Li, F., Xiang, D., Akutsu, T., Song, J., and Jia, C. (2020). Computational identification of eukaryotic promoters based on cascaded deep capsule neural networks. Brief Bioinform. https://doi.org/10.1093/bib/bbaa299.

33. Li, F., Chen, J., Leier, A., Marquez-Lago, T., Liu, Q., Wang, Y., Revote, J., Smith, A.I., Akutsu, T., Webb, G.I., et al. (2020). DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. Bioinformatics *36*, 1057–1065. https://doi.org/10.1093/bioinformatics/btz721.

34. Li, F., Leier, A., Liu, Q., Wang, Y., Xiang, D., Akutsu, T., Webb, G.I., Smith, A.I., Marquez-Lago, T., Li, J., and Song, J. (2020). Procleave: predicting protease-specific substrate cleavage sites by combining sequence and structural information. Genomics Proteomics Bioinformatics *18*, 52–64. https://doi.org/10.1016/j.gpb.2019.08.002.

35. Liu, Q., Chen, J., Wang, Y., Li, S., Jia, C., Song, J., and Li, F. (2021). DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. Brief Bioinform. *22*. https://doi.org/10.1093/bib/bbaa124.

36. Li, F., Guo, X., Jin, P., Chen, J., Xiang, D., Song, J., and Coin, L.J.M. (2021). Porpoise: a new approach for accurate prediction of RNA pseudouridine sites. Brief Bioinform. https://doi.org/10.1093/bib/bbab245.

37. Jia, C., Bi, Y., Chen, J., Leier, A., Li, F., and Song, J. (2020). PASSION: an ensemble neural network approach for identifying the binding sites of RBPs on circRNAs. Bioinformatics *36*, 4276–4282. https://doi.org/10.1093/bioinformatics/btaa522.

38. Raschka, S. (2018). MLxtend: providing machine learning and data science utilities and extensions to Python's scientific computing stack. J. Open Source Softw. 3, 638.

39. Zou, Q., Zeng, J.C., Cao, L.J., and Ji, R.R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing *173*, 346–354. https://doi.org/10.1016/j.neucom.2014.12.123.

40. Guha, R., Ghosh, K.K., Bhowmik, S., and Sarkar, R. (2020). Mutually Informed Correlation Coefficient (MICC) - a New Filter Based Feature Selection Method (In 6th IEEE Calcutta Conference (CALCON)).

41. Chen, Z., Zhao, P., Li, F.Y., Marquez-Lago, T.T., Leier, A., Revote, J., Zhu, Y., Powell, D.R., Akutsu, T., Webb, G.I., et al. (2020). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief Bioinform. *21*, 1047–1057. https://doi.org/10.1093/bib/bbz041.

42. Huang, Q.F., Zhang, J., Wei, L.Y., Guo, F., and Zou, Q. (2020). 6mA-RicePred: a method for identifying DNA N (6)-methyladenine sites in the rice genome based on feature fusion. Front. Plant Sci. *11*, 4. https://doi.org/10.3389/fpls.2020.00004.

43. Alam, W., Tayara, H., and Chong, K.T. (2020). XG-ac4C: identification of N4-acetylcytidine (ac4C) in mRNA using eXtreme gradient boosting with electron-ion interaction pseudopotentials. Sci. Rep. *10*, 20942. https://doi.org/10.1038/s41598-020-77824-2.

44. Zhang, L., Qin, X., Liu, M., Xu, Z., and Liu, G. (2021). DNN-m6A: a cross-species method for identifying RNA N6-methyladenosine sites based on deep neural network with multi-information fusion. Genes *12*, 354. https://doi.org/10.3390/genes12030354.

45. Wang, M.H., Cui, X.W., Yu, B., Chen, C., Ma, Q., and Zhou, H.Y. (2020). SulSite-GTB: identification of protein S-sulfenylation sites by fusing multiple feature information and gradient tree boosting. Neural Comput. Appl. *32*, 13843–13862. https://doi.org/10.1007/s00521-020-04792-z.

46. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Science & Business Media), p. 33.

47. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

48. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., et al. (2013). API design for machine learning software: experiences from the scikit-learn project. arXiv, 1–14, preprint arXiv:1309.0238.

49. Zhu, X.L., He, J.J., Zhao, S.H., Tao, W., Xiong, Y., and Bi, S.D. (2019). A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of *Saccharomyces cerevisiae*. Brief. Funct. Genomics *18*, 367–376. https://doi.org/10.1093/bfgp/elz018.

50. Gao, J.B., Zhang, L.F., Yu, G.Q., Qu, G.Q., Li, Y.F., and Yang, X.B. (2020). Model with the GBDT for colorectal adenoma risk diagnosis. Curr. Bioinformatics *15*, 971–979. https://doi.org/10.2174/1574893614666191120142005.

51. Zheng, R., Li, M., Chen, X., Wu, F.-X., Pan, Y., and Wang, J. (2019). BiXGBoost: a scalable, flexible boosting-based method for reconstructing gene regulatory networks. Bioinformatics *35*, 1893–1900. https://doi.org/10.1093/bioinformatics/bty908.

52. Yu, J.L., Shi, S.P., Zhang, F., Chen, G.D., and Cao, M. (2019). PredGly: predicting lysine glycation sites for *Homo sapiens* based on XGboost feature optimization. Bioinformatics *35*, 2749–2756. https://doi.org/10.1093/bioinformatics/bty1043.

53. Yu, X., Zhou, J., Zhao, M., Yi, C., Duan, Q., Zhou, W., and Li, J. (2020). Exploiting XG boost for predicting enhancer-promoter interactions. Curr. Bioinformatics *15*, 1036–1045. https://doi.org/10.2174/1574893615666200120103948.

54. Bi, Y., Xiang, D., Ge, Z., Li, F., Jia, C., and Song, J. (2020). An interpretable prediction model for identifying N(7)-methylguanosine sites based on XGBoost and SHAP. Mol. Ther. Nucleic Acids *22*, 362–372. https://doi.org/10.1016/j.omtn.2020.08.022.

55. Chen, C., Zhang, Q.M., Ma, Q., and Yu, B. (2019). LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion. Chemometr. Intell. Lab. Syst. *191*, 54–64. https://doi.org/10.1016/j.chemolab.2019.06.003.

56. Maiti, S., Hassan, A., and Mitra, P. (2020). Boosting phosphorylation site prediction with sequence feature-based machine learning. Proteins *88*, 284–291. https://doi.org/10.1002/prot.25801.

57. Zhang, G.S., Deng, Y.Y., Liu, Q.Y., Ye, B.X., Dai, Z.M., Chen, Y.W., and Dai, X.H. (2020). Identifying circular RNA and predicting its regulatory interactions by machine learning. Front. Genet. *11*, 655. https://doi.org/10.3389/fgene.2020.00655.

58. Liu, P., Song, J., Lin, C.-Y., and Akutsu, T. (2021). ReCGBM: a gradient boosting-based method for predicting human dicer cleavage sites. BMC Bioinformatics *22*, 63. https://doi.org/10.1186/s12859-021-03993-0.

59. Heddam, S., Ptak, M., and Zhu, S.L. (2020). Modelling of daily lake surface water temperature from air temperature: extremely randomized trees (ERT) versus Air2Water, MARS, M5Tree, RF and MLPNN. J. Hydrol. *588*, 125130. https://doi.org/10.1016/j.jhydrol.2020.125130.