



# DeepOMe: A Web Server for the Prediction of 2'-O-Me Sites Based on the Hybrid CNN and BLSTM Architecture

Hongyu Li<sup>1,2</sup>, Li Chen<sup>1</sup>, Zaoli Huang<sup>1</sup>, Xiaotong Luo<sup>1</sup>, Huiqin Li<sup>1</sup>, Jian Ren<sup>1\*</sup> and Yubin Xie<sup>1\*</sup>

<sup>1</sup> School of Life Sciences, Precision Medicine Institute, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China, <sup>2</sup> School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

## OPEN ACCESS

### Edited by:

Jia Meng,  
Xi'an Jiaotong-Liverpool University,  
China

### Reviewed by:

Kunqi Chen,  
University of Liverpool,  
United Kingdom  
Lin Zhang,  
China University of Mining  
and Technology, China

### \*Correspondence:

Yubin Xie  
xieyb6@mail.sysu.edu.cn  
Jian Ren  
renjian@sysucc.org.cn

### Specialty section:

This article was submitted to  
Epigenomics and Epigenetics,  
a section of the journal  
Frontiers in Cell and Developmental  
Biology

**Received:** 28 March 2021

**Accepted:** 23 April 2021

**Published:** 14 May 2021

### Citation:

Li H, Chen L, Huang Z, Luo X,  
Li H, Ren J and Xie Y (2021)  
DeepOMe: A Web Server for  
the Prediction of 2'-O-Me Sites  
Based on the Hybrid CNN  
and BLSTM Architecture.  
*Front. Cell Dev. Biol.* 9:686894.  
doi: 10.3389/fcell.2021.686894

2'-O-methylations (2'-O-Me or Nm) are one of the most important layers of regulatory control over gene expression. With increasing attentions focused on the characteristics, mechanisms and influences of 2'-O-Me, a revolutionary technique termed Nm-seq were established, allowing the identification of precise 2'-O-Me sites in RNA sequences with high sensitivity. However, as the costs and complexities involved with this new method, the large-scale detection and in-depth study of 2'-O-Me is still largely limited. Therefore, the development of a novel computational method to identify 2'-O-Me sites with adequate reliability is urgently needed at the current stage. To address the above issue, we proposed a hybrid deep-learning algorithm named DeepOMe that combined Convolutional Neural Networks (CNN) and Bidirectional Long Short-term Memory (BLSTM) to accurately predict 2'-O-Me sites in human transcriptome. Validating under 4-, 6-, 8-, and 10-fold cross-validation, we confirmed that our proposed model achieved a high performance (AUC close to 0.998 and AUPR close to 0.880). When testing in the independent data set, DeepOMe was substantially superior to NmSEER V2.0. To facilitate the usage of DeepOMe, a user-friendly web-server was constructed, which can be freely accessed at <http://deepome.renlab.org>.

**Keywords:** CNN, BLSTM, web service, RNA modification, 2'-O-methylation

## INTRODUCTION

To date, hundreds of different RNA modifications have been identified in human transcriptome, and found to be critical in the regulation of various transcriptional events (Behm-Ansmant et al., 2011). Among those, 2'-O-methylation (2'-O-Me) is one of the most abundant RNA modifications, presenting in transfer RNAs (tRNAs) (Somme et al., 2014), ribosomal RNAs (rRNAs) (Rebane and Metspalu, 2002), small nuclear/small nucleolar RNAs (snRNAs/snoRNAs) (Darzacq et al., 2002), microRNAs (miRNAs) (Li, 2005)/Piwi-interacting RNAs (piRNAs) (Yu et al., 2005), and some messenger RNAs (mRNAs) (Dai et al., 2017). The addition of methyl groups on the ribose moiety can affect sterical properties, hydrogen-bonding potential, and structural rigidity of the target RNA (Kierzek et al., 2009; Hengesbach and Schwalbe, 2014), and orchestrating the biogenesis (Ojha et al., 2020), metabolism (Salem et al., 2019), and functions (Choi et al., 2018) of these RNA molecules.

Given its functional importance, the precise detection and functional analysis of 2'-O-Me are important research topics in the community.

Recently, several experimental techniques were developed to pinpoint the precise 2'-O-Me sites. For example, perchloric acid (HClO<sub>4</sub>) hydrolysis (Baskin and Dekker, 1967), periodate oxidation hydrolysis (Trim and Parker, 1972), chromatography and mass-spectrometry (Abbate and Rottman, 1972; Sardana and Fuke, 1980; Fenghe and McCloskey, 1999; Kirpekar et al., 2005). At present, high-throughput techniques that established based on deep sequencing were also reported. Typical examples included RiboMethSeq (Krogh et al., 2016; Eraldes et al., 2017; Sharma et al., 2017; Zhou et al., 2017), 2'-OMe-Seq (Incarnato et al., 2017), RibOxi-Seq (Zhu et al., 2017), and Nm-seq (Dai et al., 2017; Hsu et al., 2019).

Although the previous mentioned high-throughput techniques can provide single-nucleotide mapping of 2'-O-Me sites at transcriptome level, the experimental procedure is still expensive and labor-exhausting. Therefore, there is still an urgent need of a computational model to mine the sequence feature of 2'-O-Me sites and identify the 2'-O-Me sites *in silico*. So far, several computational methods such as iRNA-2methyl (Qiu et al., 2017), Deep-2'-O-Me (Mostavi et al., 2018), iRNA-2OM (Yang et al., 2018), NMSEER V2.0 (Zhou et al., 2019), and iRNA-PseKNC (Tahir et al., 2019) have been developed. However, many issues remain in these methods, leaving plenty of room for improvement. Firstly, 2'-O-Me can occur in all types of RNA nucleotides, resulting an extremely imbalanced dataset between positive and negative samples. The traditional classification algorithm, which aims at the overall classification accuracy, pays too much attention to the major class, leading to poor performances in minor class and high false positives. Secondly, previous studies have randomly sampled subsequences near experimentally identified 2'-O-Me sites as negative sequences. This procedure can produce a high degree of similarity between extracted positive and negative sequences in training dataset, which limits the accuracy of traditional sequence-based models. Third, many tools lack a convenient webserver, hindering their widespread use in biological scenario. Therefore, the development of a reliable prediction tool that can not only extract useful features from the primary mRNA sequences but also produce high-precision results is still an important problem to be solved.

The performance of traditional machine learning algorithms relies heavily on data representations. However, features are typically designed by human engineers with extensive domain expertise, and identifying which features are more appropriate for the given task remains difficult. Thanks to the ability of deep learning architectures in automatically extracting high-representation information in the raw data, the application of deep learning framework is a promising way to address the above issues. In recent years, many attempts have been made to apply deep learning algorithms in biological research. For example, DeepBind (Alipanahi et al., 2015) for predicting DNA- and RNA-binding specificity, AlphaFold (Senior et al., 2020) for predicting protein structure, scDeepCluster (Tian et al., 2019) for clustering

single cell RNA-seq data, DeepCpG (Angermueller et al., 2017) for predicting single-cell DNA methylation state. Considering the characteristics of 2'-O-Me, deep learning algorithms are more suitable to analyses the patterns of 2'-O-Me and thus may greatly improve the prediction performance.

In this article, we present DeepOMe, a web server based on a hybrid deep learning architecture for predicting 2'-O-Me sites in Human mRNA. To our best knowledge, our work is the first effort to use the combination of CNN with RNN in the prediction of mRNA modification sites under the sequence-to-sequence mode. Moreover, a webserver was further developed and makes it easier for researchers and experimenters to use our proposed model.

## MATERIALS AND METHODS

### Dataset Collection

The training and test dataset of DeepOMe was constructed from the recently developed Nm-seq experiment (Dai et al., 2017) which comprised of 4,481 2'-O-Me sites in human transcriptome. The site data were first preprocessed and split into training and independent test set using the scheme presented in **Supplementary Figure 1**. Firstly, 2'-O-Me sites in intergenic region were removed. Due to the reason that our collected data had two coordinates versions (GRCh37 and GRCh38), we next converted original GRCh37 coordinates to GRCh38 coordinates using LiftOver and further mapped it to human transcripts for better transcriptome annotation. Transcript sequences with at least one mapped 2'-O-Me site were extracted according to the corresponding gene set annotation. If the same 2'-O-Me site located in multiple transcripts, the longest transcript were retained in our dataset. Finally, we collected 2,285 RNA sequences with 3,052 2'-O-Me sites as the final data set. We randomly selected 10% of the collected RNA sequences as independent testing set, and the remaining sequences were regarded as training set. As the result, we assembled 2,046 sequences with 2,743 2'-O-Me sites as the training set, and 239 sequences with 309 2'-O-Me sites as the testing set.

### Extract Features From Primary mRNA Sequences

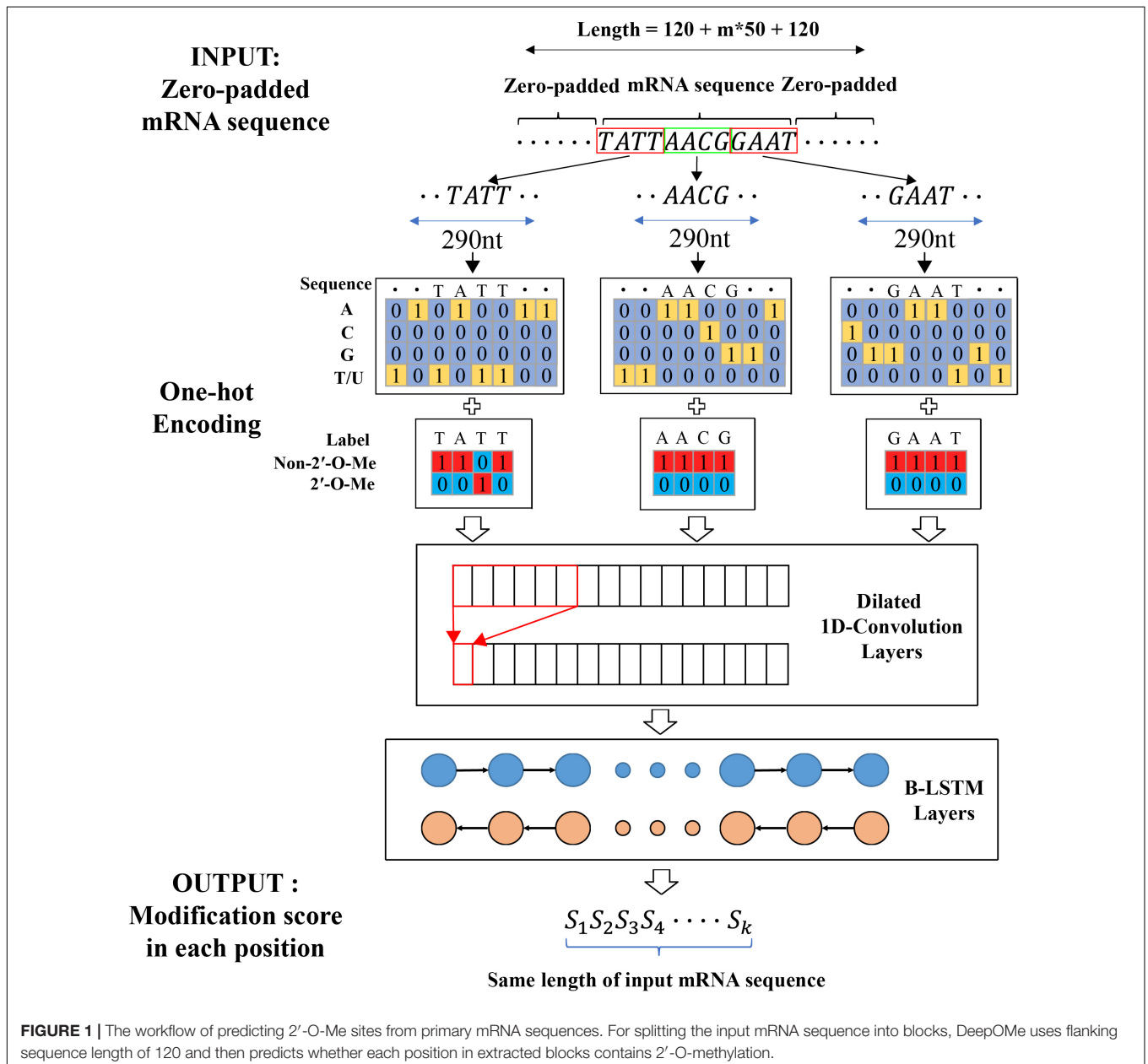
As mentioned above, previous studies extracted the flanking region of specific length around each 2'-O-Me site as the positive sequences for the training process. To create non-2'-O-Me sites or negative sequences, they randomly selected the non-modified RNA sites around known 2'-O-Me sites and captured its surrounding nucleotide sequences as negative sequences. This procedure suffers from several pitfalls.

First of all, the training set would contain overlapping sequences if the randomly selected negative sites were adjacent to 2'-O-Me sites. This would result in high similarity between positive sequences and negative sequence. The similar sequences would generate many redundant sequence-based features and thus make sequence-based machine learning algorithms difficult to train a validity predictor. To avoid such a scenario, a positive-to-negative ratio (1:10 in NmSEER V2.0; 1:1 in iRNA-2OM; 1:4 in Deep-2'-O-Me) in training set should be manually set. Since

in natural transcripts the number of 2'-O-Me sites and non-2'-O-Me sites are highly imbalanced, this kind of operations may always generate many false positives.

To solve these problems, the similar procedure from Jaganathan et al. (2019) was chosen to extract input features and output labels from primary mRNA sequences (as shown in Figure 1). Firstly, the transcript sequence was one-hot encoded as follows: A, C, G, T/U mapped to [1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1], respectively. Next, the one-hot encoded sequence was zero-padded until the length became a multiple of 50 in order to successfully split into non-overlapped blocks of 50 nt. To capture sequence dependent features, such mRNA sequence was further zero-padded at the 5'- and 3'-end with a flanking sequence of

length  $L$ . At last, the padded sequence was split into blocks in such a way that the  $i^{th}$  block consisted of nucleotide positions from  $50(i-1)-L + 1$  to  $50i + L$ . Therefore, the 50nt center regions in the  $i^{th}$  block and  $(i + 1)^{th}$  block had no overlapping sequence in original mRNA sequence. Similarly, the modification output label sequence was one-hot encoded as follows: 2'-O-Me modification and non-2'-O-Me modification were mapped to [0,1] and [1,0] respectively. The one-hot encoded label sequence was zero-padded until the length became a multiple of 50 and then further zero-padded at the start and the end with a flanking sequence of length  $L$ . The padded label sequence was split into blocks using the same procedure as described for the inputted mRNA sequence. The extracted one-hot encoded



nucleotides sequences and the corresponding one-hot encoded label sequences were used as inputs and the target outputs to train and evaluate our model.

## Architecture of the CNN-BLSTM Model

**Figure 2A** shows our proposed CNN-BLSTM architecture. DeepOMe is composed of 10 layers of CNN and 2 layers of B-LSTM (Schuster and Paliwal, 1997). The model structure consists of input layer, CNN layers, BLSTM layers, fully connected layer, and the output layer. The input layer can receive one-hot encoded sequence data. In the CNN layers, we first enriched the representation in the Stem Block (**Supplementary Figure 2A**) by computing multiple feature maps with different kernel sizes (Szegedy et al., 2015). Then, we stacked three residual blocks (He et al., 2016) for local feature extraction. The convolution operation in Stem block and ResBlock was 1D-convolution with kernel size of 10 and dilation rate of 2 (**Supplementary Figure 2B**). The CNN layers were used as preprocessing step to extract the deep spatial features from the input sequences. Then, these deep features were fed into two BLSTM layers with 32 units for learning of sequence-dependent features. The last layer in our model is a fully-connected layer with softmax activation, which was used to generate the final prediction score. The detailed architectural information was listed in **Supplementary Table 1**.

## Experimental Setup

The proposed model was implemented with the TensorFlow library (Abadi et al., 2016) in Python and trained on an NVIDIA GTX2080 GPU. The proposed model was trained through 100 epochs using batch size of 200. The categorical cross entropy loss between the target and the predicted outputs was minimized using Adam optimizer (Kingma and Ba, 2014). The initial learning rate of 0.001 was used to train the model. Early-stopping (Caruana et al., 2001) was used to control overfitting. We monitored the validation loss at each epoch. When the validation loss has not improved after ten epochs, training is interrupted.

## Evaluation Metrics

Testing set was used to validate our proposed model comparing with available prediction tools after cross-validation. The performance was evaluated based on several metrics, namely area under Precision-Recall Curve (AUPR), area under Receiver Operating Characteristic Curve (AUC), sensitivity (Sn), specificity (Sp), precision (Pr), accuracy (Acc), and Matthew's correlation coefficient (Mcc).

When evaluating model's performance in full mRNA sequence, an accuracy metric was largely ineffective since most of the positions in mRNA sequence are not 2'-O-Me sites. The prediction model was like the recommender systems which was to suggest the most proper modification sites in mRNA sequences. Thus, top-k accuracy was more appropriate in such situation. When comparing among different methods, we evaluated the top-k accuracy besides the AUC and AUPR metrics. The top-k accuracy is defined as follows: Suppose the test set has  $k$  positions that belong to the right class which is 2'-O-Me site. We choose the threshold so that exactly  $k$  test set positions are

predicted as belonging to the right class. The fraction of these  $k$  predicted positions that truly belong to the right class is reported as the top-k accuracy.

## RESULTS

### Flanking Sequence Length Analysis

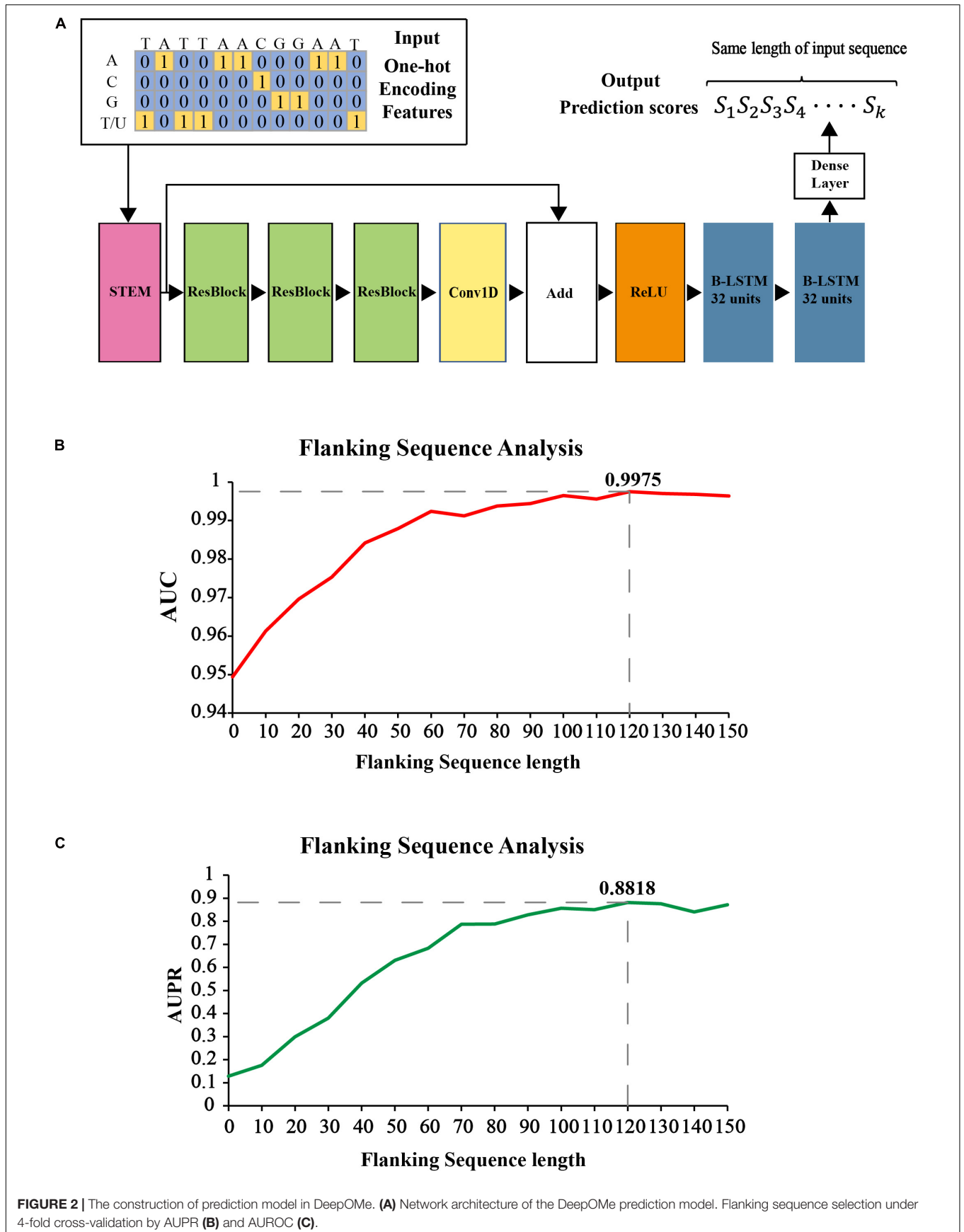
It is necessary to determine the optimal flanking sequence length  $L$  of input sequences for identifying 2'-O-Me sites. Generally speaking, if the flanking sequence around the known 2'-O-Me site is too short, it may not carry enough information for prediction and will lead to poor performance. Otherwise, if the flanking sequence is too long, it may carry too much redundant information, leading to poor generalization. Thus, we first analyzed the averaged AUC and AUPR of the proposed model with different flanking sequence length under 4-fold cross-validation. As shown in **Figures 2B,C** the search step size for flanking sequence length was 10 nt, with a range of 0 and 150. According to the evaluation results, when the flanking sequence length equals to 120 nt and block length equals to 290 nt, the performance generated by our proposed model was the best (AUC = 0.9975, AUPR = 0.8818). Therefore, we selected the flanking sequence with length of 120.

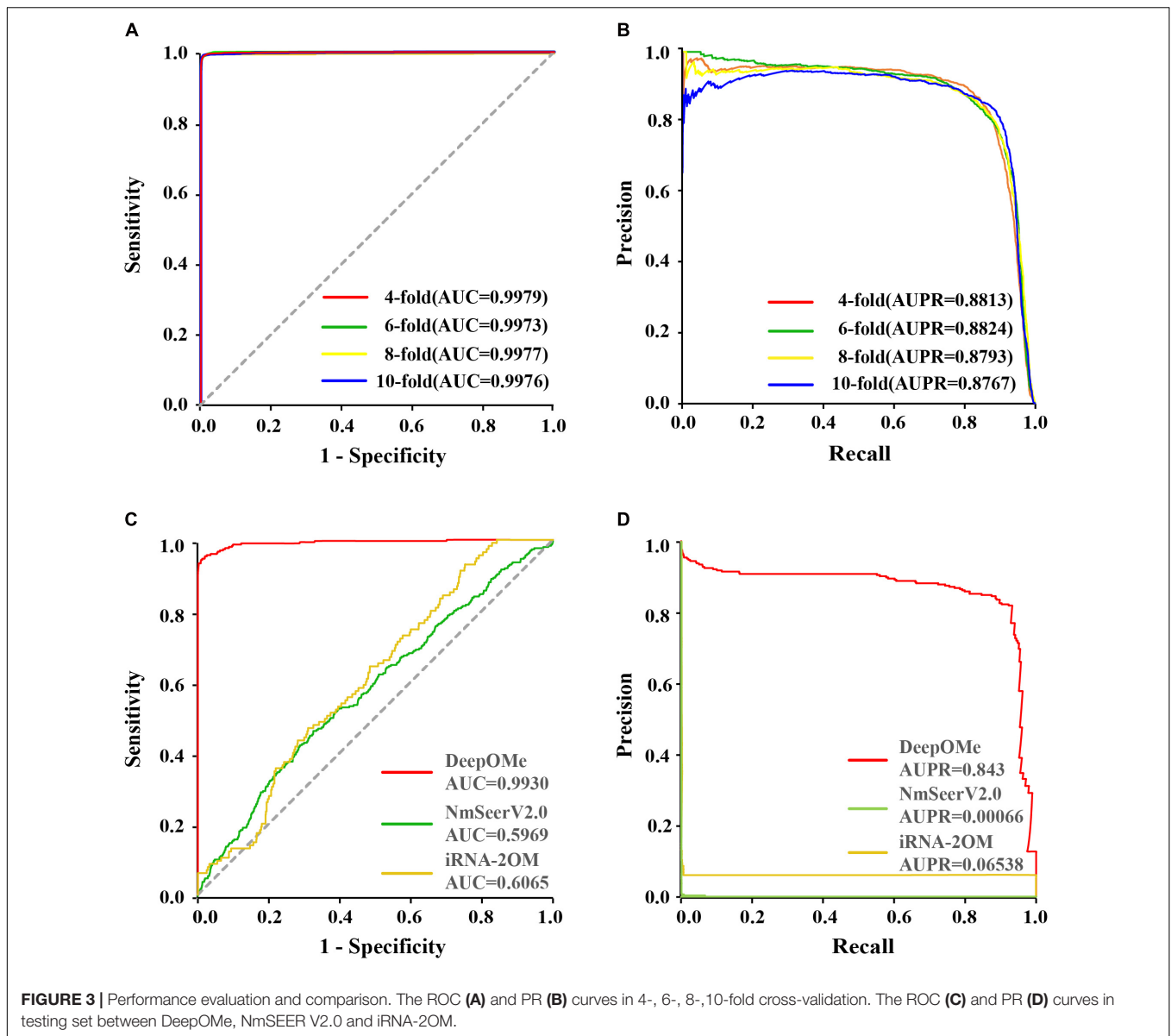
### Evaluation of the Prediction Performance

To evaluate the prediction performance of DeepOMe, we performed 4-, 6-, 8-, and 10-fold cross-validation of the training set. **Figures 3A,B** shows the ROC and PR curves of our proposed CNN-BLSTM model under 4-, 6-, 8-, and 10-fold cross-validations with flanking sequence length of 120. As a result, DeepOMe showed an acceptable performance in  $n$ -fold cross-validations with the area under the ROC curves (AUROC) close to 0.998 and area under the PR curves (AUPR) close to 0.880.

To rigorously evaluate the prediction and generalizability performance of DeepOMe, we next compared it with other state-of-art predictors using the independent test set. Since only iRNA-2OM, iRNA-2methyl, and NmSEER V2.0 provided web-server or standalone package for usage, the comparison will only perform between them. During the comparison, we further found that there were no responses in the webservers of iRNA-2methyl, hence, the final comparison only performs between DeepOMe, NmSEER V2.0, and iRNA-2OM.

**Figures 3C,D** presented the comparison results in ROC curves and PR curves. The results showed that the DeepOMe achieved a better performance (AUROC = 0.993, 95%CI:0.993-0.993; AUPR = 0.843) in the testing set than NmSEER V2.0 (AUROC = 0.5969, 95%CI:0.599-0.600; AUPR = 0.00066) and iRNA-2OM (AUROC = 0.6065, 95%CI:0.601-0.612; AUPR = 0.06538). When testing in full mRNA sequences, we further compared the top-k accuracy between DeepOMe, NmSEER V2.0, and iRNA-2OM. The comparison results in **Table 1** suggested that DeepOMe (Top-1 Acc = 0.8602, Top-100 Acc = 0.9563) was more sensitive and robust than NmSEER V2.0 (Top-1 Acc = 0.0, Top-100 Acc = 0.1004) and iRNA-2OM (Top-1 Acc = 0.0, Top-100 Acc = 0.1087).





To evaluate the sequence similarity between the predicted sites and the detected sites in transcriptome, sequence logos were generated using WebLogo (Crooks et al., 2004) in training and testing set. **Supplementary Figure 3** presented the graphical representation of sequence similarity. The results showed that the predicted sites under different thresholds were similar to the detected sites both in training set and testing set, proving that our proposed model could precisely identify 2'-O-Me sites.

## Web-Server

To facilitate the use of our prediction models, we next developed an online predictor called DeepOMe for the community. The predictor is freely available at <http://deepome.renlab.org>. DeepOMe only requires mRNA sequences to run a prediction. Multiple mRNA sequences can be input into the text area or uploaded as a single FASTA file. For users' convenience, we

selected three thresholds based on the 10-fold cross-validation results (**Figure 4A**), which correspond to the false discovery rate of 0.10, 0.15, and 0.20. The detailed performance values under these three thresholds are shown in **Supplementary Table 2**. Besides, users can select the threshold by setting the false discovery rate in advanced option menu. After the query sequences are submitted to DeepOMe, users can check its running status in the result panel in real time. When the prediction is complete, the button that links out to the result page will be clickable (**Figure 4B**). **Figure 4C** provides a snapshot for the result page of the example mRNA sequence. The prediction position, score and prediction threshold were listed in an interactive table, which allows the users to easily search and sort the results. Remarkably, to facilitate a further analysis of the protein function and RNA structure, we also implemented an automatic pipeline for visualizing the prediction

**TABLE 1** | Comparison of Top-k Accuracy between DeepOMe, NmSEER V2.0, and iRNA-2OM in testing set.

Top-k Accuracy	DeepOMe	NmSEER V2.0	iRNA-2OM
Top-1 Accuracy	0.8602	0.0	0.0
Top-3 Accuracy	0.9039	0.0131	0.0
Top-5 Accuracy	0.9082	0.0175	0.0
Top-10 Accuracy	0.9126	0.0175	0.0043
Top-20 Accuracy	0.9257	0.0306	0.0130
Top-30 Accuracy	0.9344	0.0611	0.0130
Top-40 Accuracy	0.9476	0.0699	0.0261
Top-50 Accuracy	0.9520	0.0699	0.0478
Top-60 Accuracy	0.9520	0.0830	0.0696
Top-70 Accuracy	0.9520	0.0917	0.0739
Top-80 Accuracy	0.9563	0.0961	0.0826
Top-90 Accuracy	0.9563	0.1004	0.0870
Top-100 Accuracy	0.9563	0.1004	0.1087

results. By integrating IBS (Liu et al., 2015), InterProScan (Jones et al., 2014), and ViennaRNA (Lorenz et al., 2011) into the web server, DeepOMe can present the graphical representation of the input mRNA sequence together with their predicted sites in the visualization panel. **Figures 4D,E** provide snapshots for the visualization results of RNA secondary structure and protein domain organization. The diagrams can be saved as a vector graphic (SVG) for further analysis.

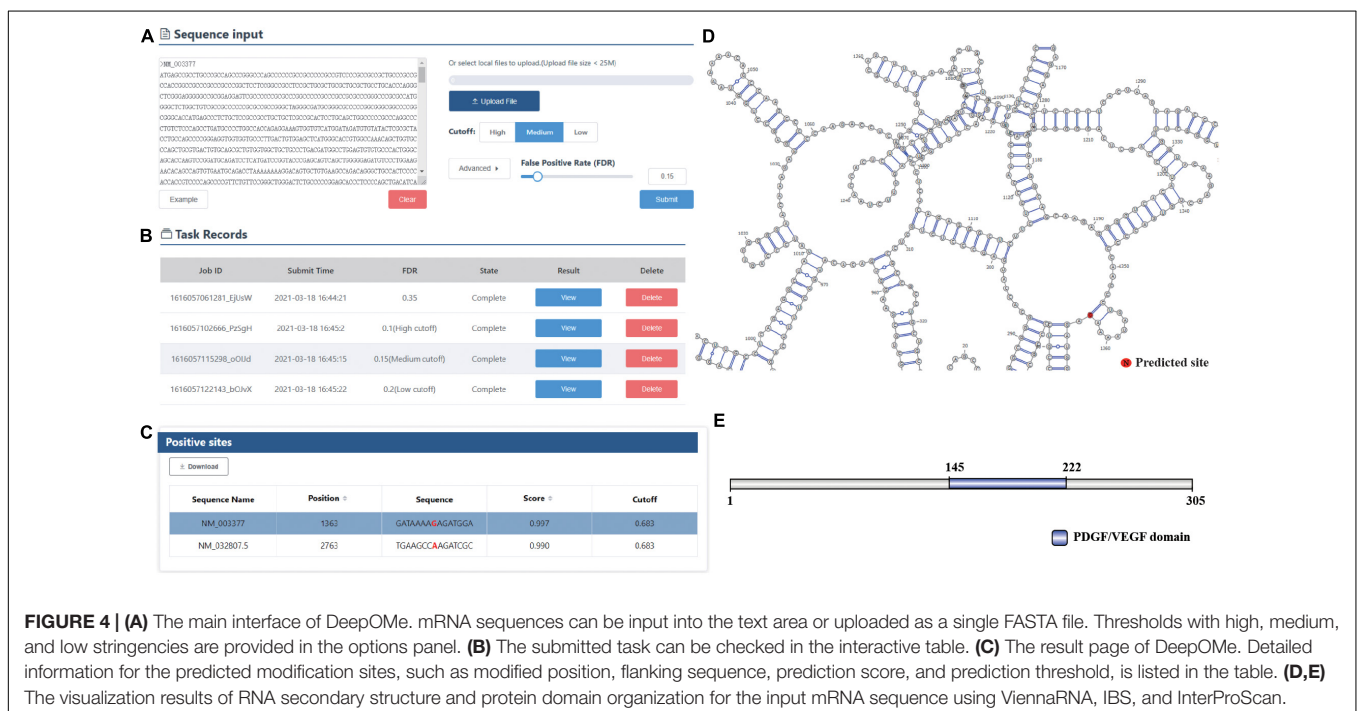
## DISCUSSION

2'-O-methylation plays critical roles in regulating gene expressions at the post-transcriptional levels. Thus, proper

identification of the 2'-O-Me site is essential to understand the mechanism of RNA metabolisms. 2'-O-Me can occur in any base on the mRNA sequence. Given a mRNA sequence, we need to get an output sequence with the same length of input sequence. The score in each position of output sequence represents whether this position in the input mRNA contains 2'-O-Me. Therefore, the 2'-O-Me site prediction problem can be considered as a many-to-many prediction problem.

However, previous studies tried to train the prediction model based on a Many-to-One mode. They had to randomly select non-2'-O-Me sites around the known 2'-O-Me sites as negative samples, which resulted in high sequence similarity between positive and negative sequences. Besides, to reduce the degree of imbalance in their training data set, the negative sites were manually down-sampled to obtain a relatively small positive-to-negative ratio. However, in reality, the positive-to-negative ratio in a given RNA sequence was always extremely high, and thus caused their models to have poor generalization ability in unseen data. These were the two main reasons why their models received very poor performance in our testing set.

Unlike the previous works that use handcrafted features for classification, DeepOMe could automatically extract the deep features from primary mRNA sequences by CNN layers. DeepOMe was proven to be more efficient than the available method in terms of all evaluation metrics. We found several factors that may explain the high performance achieved by our proposed model. Firstly, the procedure we used to train and test the models was the many-to-many mode. Thus, there was no need to manually balance the training dataset in our model, allowing to learn sufficient information between 2'-O-Me and non-2'-O-Me sites and achieving lower false positives. Secondly, the use of the dilated 1D CNN compared to the



traditional 1D CNN allowed our model to cover more relevant information by increasing the receptive field of the filters. Additionally, the stacked Resblocks used in our model allowed us to build a deeper network and take advantages of the powerful representational ability of deep neural network. At last, the application of bidirectional LSTM was able to exploit meaningful representations from upstream and downstream sequences. The comparison results suggest that the combination of CNN and BLSTM can successfully capture the key features of the entire mRNA sequences. Although promising performance was obtained in DeepOMe, a number of future improvements are expected. First of all, we have designed only a relatively simple CNN-based neural network model in our current version. Various deeper and wider CNN architectures were awaited exploration in the future. Secondly, attention mechanism will be introduced to achieve better representation for contextual information in the future version. Last but not least, we have now trained the model only based on the experimental 2'-O-Me data for *Homo sapiens*. The prediction models for other species such as *Mus musculus* will be established in the future.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

## REFERENCES

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "TensorFlow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, (California: USENIX Association), 265–283.
- Abbate, J., and Rottman, F. (1972). Gas chromatographic method for determination of 2-O-methylation in RNA. *Anal. Biochem.* 47, 378–388. doi: 10.1016/0003-2697(72)90131-5
- Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831–838. doi: 10.1038/nbt.3300
- Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* 18:67.
- Baskin, F., and Dekker, C. A. (1967). A rapid and specific assay for sugar methylation in ribonucleic acid. *J. Biol. Chem.* 242, 5447–5449. doi: 10.1016/s0021-9258(18)99445-7
- Behm-Ansmant, I., Helm, M., and Motorin, Y. (2011). Use of Specific Chemical Reagents for Detection of Modified Nucleotides in RNA. *J. Nucleic Acids* 2011:408053.
- Caruana, R., Lawrence, S., and Giles, C. L. (2001). *Overfitting In Neural Nets: Backpropagation, Conjugate Gradient, And Early Stopping*, In: *Advances In Neural Information Processing Systems*. Massachusetts: MIT Press, 402–408.
- Choi, J., Indrisiunaitė, G., Demirci, H., Jeong, K.-W., Wang, J., Petrov, A., et al. (2018). 2'-O-methylation in mRNA disrupts tRNA decoding during translation elongation. *Nat. Struct. Mol. Biol.* 25, 208–216. doi: 10.1038/s41594-018-0030-z
- Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190. doi: 10.1101/gr.849004
- Dai, Q., Moshitch-Moshkovitz, S., Han, D., Kol, N., Amariglio, N., Rechavi, G., et al. (2017). Nm-seq maps 2'-O-methylation sites in human mRNA with base precision. *Nat. Methods.* 14, 695–698. doi: 10.1038/nmeth.4294

## AUTHOR CONTRIBUTIONS

HL implemented the DeepOMe algorithm and wrote the manuscript. LC and XL manually collected 2'-O-Me data from published literatures and performed data pre-processing. ZH and HL are respectively responsible for the front-end page display and back-end logic design of the DeepOMe website. JR was responsible for supervision, funding acquisition, and writing-review. YX supervised this work, reviewed and edited the manuscript. All authors have read and approved the manuscript.

## FUNDING

This work was supported by the National Natural Science Foundation of China [91753137, 31771462, 81772614, U1611261, 31801105, and 81802438], the National Key R&D Program of China [2017YFA0106700], the Program for Guangdong Introducing Innovative and Entrepreneurial Teams [2017ZT07S096], and the Guangdong Basic and Applied Basic Research Foundation [2018A030313323 and 2020A1515010220].

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcell.2021.686894/full#supplementary-material>

- Darzacq, X., Jády, B. E., Verheggen, C., Kiss, A. M., Bertrand, E., Kiss, T. (2002). Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNA. *EMBO J.* 21, 2746–2756. doi: 10.1093/emboj/21.11.2746
- Erales, J., Marchand, V., Panthu, B., Gillot, S., Belin, S., and Ghayad, S. E. (2017). Evidence for rRNA 2'-O-methylation plasticity: control of intrinsic translational capabilities of human ribosomes. *Proc. Natl. Acad. Sci. U. S. A.* 114, 12934–12939. doi: 10.1073/pnas.1707674114
- Fenghe, Q., and McCloskey, J. A. (1999). Selective detection of ribose-methylated nucleotides in RNA by a mass spectrometry-based method. *Nucleic Acids Res* 27:e20.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (New York: IEEE), 770–778.
- Hengesbach, M., and Schwalbe, H. (2014). Structural Basis for Regulation of Ribosomal RNA 2-O-Methylation. *Angew. Chem.* 53, 1742–1744. doi: 10.1002/anie.201309604
- Hsu, P. J., Fei, Q., Dai, Q., Shi, H., Dominissini, D., Ma, L., et al. (2019). Single base resolution mapping of 2'-O-methylation sites in human mRNA and in 3' terminal ends of small RNAs. *Methods* 156, 85–90. doi: 10.1016/j.ymeth.2018.11.007
- Incarnato, D., Anselmi, F., Morandi, E., Neri, F., Maldotti, M., and Rapelli, S. (2017). High-throughput single-base resolution mapping of RNA 2'-O-methylated residues. *Nucleic Acids Res.* 45, 1433–1441.
- Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., Mcrae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 176:e524.
- Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., Mcanulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031
- Kierzek, E., Pasternak, A., Pasternak, K., Gdaniec, Z., Yildirim, I., Turner, D. H., et al. (2009). Contributions of stacking, preorganization, and hydrogen bonding



- to the thermodynamic stability of duplexes between RNA and 2'-O-methyl RNA with locked nucleic acids. *Biochemistry* 48, 4377–4387 doi: 10.1021/bi9002056
- Kingma, D., and Ba, J. (2014). “Adam: A Method for Stochastic Optimization,” in *International Conference on Learning Representations*. Cornell University: New York
- Kirpekar, F., Hansen, L. H., Rasmussen, A., Poehlsgaard, J., and Vester, B. (2005). The archaeon *Haloarcula marismortui* has few modifications in the central parts of its 23 S ribosomal RNA. *J. Mol. Biol.* 348, 563–573 doi: 10.1016/j.jmb.2005.03.009
- Krogh, N., Jansson, M. D., Häfner, S. J., Tehler, D., Birkedal, U., and Christensen-Dalsgaard, M. (2016). Profiling of 2'-O-Me in human rRNA reveals a subset of fractionally modified positions and provides evidence for ribosome heterogeneity. *Nucleic Acids Res.* 44, 7884–7895. doi: 10.1093/nar/gkw482
- Li, J. (2005). Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in *Arabidopsis*. *Curr. Biol.* 15, 1501–1507. doi: 10.1016/j.cub.2005.07.029
- Liu, W., Xie, Y., Ma, J., Luo, X., Nie, P., Zuo, Z., et al. (2015). IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics* 31, 3359–3361. doi: 10.1093/bioinformatics/btv362
- Lorenz, R., Bernhart, S. H., Zu Siederdisen, C. H., Tafer, H., Flamm, C., Stadler, P. F., et al. (2011). ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 6, 1–14.
- Mostavi, M., Salekin, S., and Huang, Y. (2018). “Deep-2'-O-Me: Predicting 2'-O-methylation sites by Convolutional Neural Networks. *Annu.Int. Conf. IEEE Eng. Med. Biol. Soc.* 2018, 2394–2397.
- Ojha, S., Malla, S., and Lyons, S. M. (2020). snoRNPs: functions in Ribosome Biogenesis. *Biomolecules* 10:783. doi: 10.3390/biom10050783
- Qiu, W. R., Jiang, S. Y., Sun, B. Q., Xiao, X., Cheng, X., and Chou, K. C. (2017). iRNA-2methyl: identify RNA 2'-O-methylation Sites by Incorporating Sequence-Coupled Effects into General PseKNC and Ensemble Classifier. *Med. Chem.* 13, 734–743.
- Rebane, A., and Metspalu, R. (2002). Locations of several novel 2'-O-methylated nucleotides in human 28S rRNA. *BMC Mol. Biol.* 3:1. doi: 10.1186/1471-2199-3-1
- Salem, E. S., Vonberg, A. D., Borra, V. J., Gill, R. K., and Nakamura, T. (2019). RNAs and RNA-binding proteins in immuno-metabolic homeostasis and diseases. *Front. Cardiovasc. Med.* 6:106. doi: 10.3389/fcvm.2019.00106
- Sardana, M. K., and Fuke, M. (1980). A rapid procedure to determine the content of 2'-O-methylation in RNA by homochromatography. *Anal. Biochem.* 103, 285–288. doi: 10.1016/0003-2697(80)90611-9
- Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Proc.* 45, 2673–2681. doi: 10.1109/78.650093
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710.
- Sharma, S., Marchand, V., Motorin, Y., and Lafontaine, D. (2017). Identification of sites of 2'-O-methylation vulnerability in human ribosomal RNAs by systematic mapping. *Sci. Rep.* 7, 1–15.
- Somme, J., Van Laer, B., Roovers, M., Steyaert, J., Versées, W., and Droogmans, L. (2014). Characterization of two homologous 2'-O-methyltransferases showing different specificities for their tRNA substrates. *RNA* 20, 1257–1271. doi: 10.1261/rna.044503.114
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). “Going deeper with convolutions,” in *In: Proceedings Of The Ieee Conference On Computer Vision And Pattern Recognition*, (New York: IEEE), 1–9.
- Tahir, M., Tayara, H., and Chong, K. T. (2019). iRNA-PseKNC(2methyl): identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components. *J. Theor. Biol.* 465, 1–6. doi: 10.1016/j.jtbi.2018.12.034
- Tian, T., Wan, J., Song, Q., and Wei, Z. (2019). Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nat. Mach. Intell.* 1, 191–198. doi: 10.1038/s42256-019-0037-0
- Trim, A. R., and Parker, J. E. (1972). Nucleotide sequence in fourteen dinucleotides, modified by 2'-O-methylation, from yeast ribonucleic acid, determined by periodate degradation and by pentose analysis. *Anal. Biochem.* 46, 482–488. doi: 10.1016/0003-2697(72)90322-3
- Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). iRNA-2OM: a Sequence-Based Predictor for Identifying 2'-O-Methylation Sites in Homo sapiens. *J. Comput. Biol.* 25, 1266–1277. doi: 10.1089/cmb.2018.0004
- Yu, B., Yang, Z., Li, J., Minakhina, S., Yang, M., Padgett, R. W., et al. (2005). Methylation as a Crucial Step in Plant microRNA Biogenesis. *Science* 307, 932–935. doi: 10.1126/science.1107130
- Zhou, F., Liu, Y., Rohde, C., Pauli, C., Gerloff, D., Köhn, M., et al. (2017). AML1-ETO requires enhanced C/D box snoRNA/RNP formation to induce self-renewal and leukaemia. *Nat. Cell Biol.* 19, 844–855. doi: 10.1038/ncb3563
- Zhou, Y., Cui, Q., and Zhou, Y. (2019). NmSEER V2.0: a prediction tool for 2'-O-methylation sites based on random forest and multi-encoding combination. *BMC Bioinformatics* 20:690. doi: 10.1186/s12859-019-3265-8
- Zhu, Y., Pirnie, S. P., and Carmichael, G. G. (2017). High-throughput and site-specific identification of 2'-O-methylation sites using ribose oxidation sequencing (RibOxi-seq). *RNA* 23, 1303–1314. doi: 10.1261/rna.061549.117

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Li, Chen, Huang, Luo, Li, Ren and Xie. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.