JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

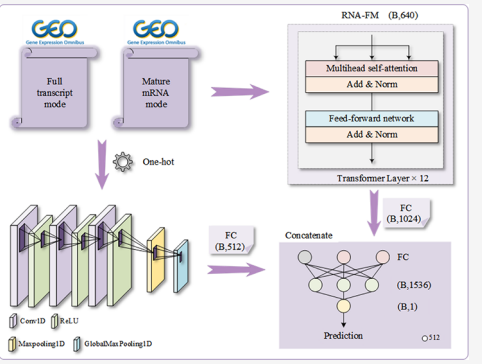
m5U-HybridNet: Integrating an RNA Foundation Model with CNN Features for Accurate Prediction of 5-Methyluridine Modification Sites

Xinyu Li, Zhenjie Luo, Jingwei Lv, Chao Yang, Shankai Yan, Junlin Xu, Yajie Meng, Leyi Wei, Zilong Zhang, Quan Zou, and Feifei Cui*

Cite This: J. Chem. Inf. Model. 2025, 65, 8079−8096 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The 5-methyluridine (m5U) modification in RNA is vital for numerous biological processes, making its precise identification a key focus in computational biology. However, traditional wet-lab detection methods are cumbersome and time-consuming, whereas existing machine learning and deep learning computational prediction models still have room for improvement. Consequently, this study introduces m5U-HybridNet, an innovative framework that strategically integrates an RNA foundation model (RNA-FM) for deep semantic feature extraction with convolutional neural network-derived characteristics, attaining unparalleled success in identifying RNA m5U modification sites. Simultaneously, when compared with other existing models across different cell types and experimental techniques, it exhibits outstanding generalization capabilities. The m5U-HybridNet web server, accessible at http://www.bioai-lab.com/m5U, offers an effective and reliable platform for predicting RNA modification sites. It not only implies the diverse potential applications of pretrained models in the analysis of biological sequences but also enhances the application of data-driven machine intelligence in the realm of molecular biophysics principles.



1. INTRODUCTION

As a core research area of epitranscriptomics, post-transcriptional modifications of RNA have ascended to significance as a focal point in the field of epigenetics. To date, scientists have identified approximately 200 distinct chemical modifications. These discoveries have significantly enhanced our understanding of the diversity of RNA molecules. Among the diverse array of post-transcriptional modifications, RNA methylation stands out as the most prevalent form.[1] It is ubiquitously present in various RNA molecules, including transfer RNA (tRNA), ribosomal RNA (rRNA), messenger RNA (mRNA), small nuclear RNA (snRNA), and small nucleolar RNA, showcasing a rich and diverse spectrum of modification types.[1] These modifications encompass 5-methyluridine (m5U), N1-methyladenosine, pseudouridine, N6-methyladenosine, 2'-O-methylation, N7-methylguanine, 5-methylcytosine, 5-hydroxymethylcytosine, and adenosine-to-inosine editing.[2] m5U modification is a type of pyrimidine modification. It occurs through the methylation of the 5-carbon position of uridine. This process is catalyzed by pyrimidine methyltransferase, which acts on the C5 position of uridine to yield m5U.[3] m5U modification is prevalently found in rRNA and tRNA.[4] The enzymes responsible for catalyzing m5U modifications exhibit species-specific variations. For instance, mammals rely on TRMT2A and TRMT2B,[5,6] whereas the *Saccharomyces cerevisiae*[6] and *Escherichia coli* utilize TrmA for this process.[7,8] Notably, aberrant expression of m5U-modified ribonucleosides has been strongly linked to the pathogenesis of human diseases, including systemic lupus erythematosus, breast cancer, as well as plant stress responses and developmental regulation.[9,10] These findings underscore the determinative role of m5U modifications in diverse biological pathways. Consequently, comprehensive investigation of m5U modification sites is indispensable for unraveling the biological processes and functional mechanisms governed by m5U, offering potential insights into therapeutic strategies for related disorders.

In recent years, in contrast to other more extensively studied RNA modifications, the identification and functional study of m5U modification sites remain relatively insufficient. Although wet-lab methods such as miCLIP (m5U individual-nucleotide resolution Cross-Linking), iCLIP (individual-nucleotide resolution UV cross-linking and immuno precipitation), and FICC (ImmunoPrecipitation and Fluorouracil-Induced Cross-linking

Received: May 29, 2025
Revised: July 4, 2025
Accepted: July 14, 2025
Published: July 22, 2025

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

and Capture Sequencing) have been developed for the identification of m5U modification sites,[11] these techniques generally face limitations such as high cost and time-consuming operations. In order to surmount this constraint, investigators have concocted a plethora of computational methodologies predicated on machine learning or deep learning algorithms to discern the m5U modification sites within the full transcript mode, as well as within mature mRNA mode, derived from disparate cellular states in human subjects. Among them, Feng and Chen[12] employed the iRNA-m5U model, a computational model based on machine learning that is designed to predict m5U sites in RNA. iRNA-m5U utilizes a feature encoding scheme that manually encodes the features of nucleotide density and nucleotide chemical property to transform RNA samples into a discrete feature space.[12] Subsequently, the Support Vector Machine (SVM) from the realm of machine learning is used as a classifier, enabling it to accurately predict m5U sites.[12] Xu et al.[13] established the m5U-GEPred model, a predictive framework that combines sequence-derived features with graph embedding-based information to identify m5U sites. Yu et al.[14] conducted a comparative evaluation of various deep learning models in epigenetic research with the help of autoBioSeqpy and developed an improved Convolutional Recurrent Neural Network model, Deepm5U, to predict m5U sites in RNA sequences, further employing techniques such as LayerUMAP and DeepSHAP to explain the advantages of deep neural networks. In a recent breakthrough, the Deep-m5U[15] was proposed for predicting m5U modification sites by integrating one-hot encoding with convolutional neural networks (CNN) and features extracted from tetranucleotides. Deep-m5U applies a Fully Connected (FC) layer followed by 10-fold cross-validation for binary classification of m5U modification sites, making it one of the most advanced approaches currently available.[15] Moreover, both the training and test datasets are balanced datasets with a positive-to-negative sample ratio of 1:1, which differs from real-world scenarios. As a result, this approach lacks sufficient persuasiveness. In the context of imbalanced datasets, Jiang et al.[16] were the initial to introduce m5UPred, a web server for the computational identification of m5U sites in primary sequences, which is established based on a SVM and a biochemical encoding scheme. Ao et al.[17] developed a novel predictor model named m5U-SVM, which integrates multiview features with machine learning for accurate identification of RNA m5U modification sites. m5U-SVM employs a two-step LightGBM strategy combined with incremental feature selection methodology to process four conventional physicochemical descriptors: Kmer, nucleic acid composition, pseudo dinucleotide composition, and composition of k-spaced nucleic acid pairs.[17] Through systematic feature extraction and fusion, m5U-SVM generates enhanced multiview representations of sequence characteristics, with the final SVM-optimized framework achieving strong predictive performance.[17] However, the m5U-SVM model shows limitations in generalization capability for precise modification site prediction, primarily due to inherent similarities between m5U and non-m5U sequences. These constraints may stem from the methodology's reliance on conventional machine learning approaches, which could potentially overlook subtle discriminative patterns in the complex RNA modification landscape.

Therefore, in view of the current situation and limitations of existing prediction methods, in this study, we have explored the state-of-the-art method of feature extraction using large models and an advanced interpretable technology, m5U-HybridNet. This is a method of postfusion of features extracted from two branches, RNA Foundation Model (RNA-FM)[18] and CNN.[19] In this method, the RNA-FM pretrained model based on a 12 layer transformer biencoder module is introduced for research on feature encoding. This self-supervised learning method is built upon the bidirectional transformer language model proposed in the BERT[20] architecture, followed by an unsupervised training approach. Additionally, efficient one-hot encoding is used to convert RNA sequences into discrete feature vectors, which are then processed through three layers of CNN for feature extraction. Certainly, before the feature fusion of the two branches, the FC layers are required to continuously adjust the internal weights, and then feature fusion is carried out. After the fusion, the features are used by the FC layers to achieve accurate prediction of the m5U modification sites, significantly enhancing the overall performance. Notably, the method demonstrated robust performance across diverse validation scenarios. It achieved accuracies of 92.85% and 94.11% on independent test sets for full transcript and mature mRNA datasets, respectively. In the cross-validation analyses within the context of molecular biology research, the model achieved an accuracy of 90.85% for full transcript samples, whereas it showed a reduced ACC of 74.41% for mature mRNA samples. More remarkably, when evaluated through cross-validation across different cell types (human embryonic kidney cells293 and near-haploid leukemia cells1) and technical platforms (FICC and miCLIP), the method exhibited exceptional performance with accuracies of 95.94%, 94.97%, 97.30%, and 95.26%, respectively. These results not only confirm the model's enhanced predictive capability for RNA m5U modification sites but also effectively demonstrated its cross-technological universality, cross-cellular generalizability, and robustness, thereby further highlighting its potential as a versatile analytical framework. The successful integration of multiview feature engineering with advanced machine learning paradigms provides novel methodological insights and computational tools for future investigations into RNA modification mechanisms.

2. MATERIALS AND METHODS

2.1. Data Acquisition. The original dataset of RNA m5U modification sites, as reported by Jiang et al.,[16] was sourced and downloaded from the Gene Expression Omnibus (GEO) database under the accession number GSE109184. These data encompass two distinct cell lines, namely human embryonic kidney cells293 and near-haploid leukemia cells1 (HEK 293 and HAP1). This dataset was generated through the advanced FICC and miCLIP, known for their high precision in capturing m5U modifications. The data primarily encompass two distinct modalities: the global transcript pattern and the mature mRNA pattern, which are generated through the sophisticated FICC and miCLIP techniques, applied to two diverse human cellular models, namely HEK293 and HAP1. The full transcript mode represents the precursor mRNA or the unprocessed RNA, which constitutes a comprehensive ensemble of mRNAs and noncoding RNAs, whereas the mature mRNA mode specifically pertains to the processed, translation-competent mRNA forms. This latter category exclusively comprises mature mRNAs that have undergone 5′ capping, 3′ tailing, and splicing to become coding RNAs, thereby reflecting the

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

intricate post-transcriptional modifications that are quintessential for mRNA maturation and functionality. In this study, the same preprocessing method as used by Ao et al.[17] was applied to the dataset. Specifically, positive samples were experimentally processed to obtain modified sequences of 41 nucleotides in length, with the modified uridine (U) site located at the center of the sequence. Negative samples were determined by randomly selecting unmodified U sites from the same transcripts as the positive samples. Consequently, we established two distinct datasets: a full transcript mode dataset comprising 3,696 positive/negative samples and a mature mRNA mode dataset containing 1,232 positive/negative samples. To prevent sequence homology bias from compromising model performance, rigorous homology reduction was implemented using CD-HIT with an 80% identity threshold cutoff. This stringent filtering process yielded biologically representative imbalanced datasets that better reflect real-world distribution patterns. Hence, the training set in the full transcript mode contains 1,534 positive samples and 2,862 negative samples of m5U modification sites, and the test set contains 500 positive samples and 247 negative samples of m5U modification sites. The training set in the mature mRNA mode contains 1,534 positive samples and 2,862 negative samples of m5U modification sites, and the test set contains 500 positive samples and 247 negative samples of m5U modification sites. This strategic dataset construction, as visually summarized in Table 1, ensures both biological relevance and methodological rigor. The intentional preservation of natural class imbalance enhances model generalizability while maintaining sufficient positive examples for productive pattern recognition.

Table 1. Benchmark Datasets of Full Transcript Mode and Mature mRNA Mode

Datasets	labels	train	test
Full transcript mode	positive	1,535	500
	negative	2,862	731
Mature mRNA mode	positive	983	245
	negative	985	247

2.2. Feature Extraction. The foundational stage in developing m5U-HybridNet for precise m5U modification recognition involves sophisticated RNA sequence encoding that prioritizes maximal information retention. This methodological innovation represents a critical advancement in computational epigenomics, establishing a robust computational framework for deciphering RNA modification patterns while maintaining biological fidelity through comprehensive sequence feature preservation. To enhance the accuracy and reliability of m5U modification site prediction, m5U-HybridNet integrates the feature information extracted from the transformer-based RNA-FM pretrained model with that obtained from the sequence information processed by one-hot encoding and then input into the CNN. By adopting the dual-branch sequence feature extraction method of the RNA-FM model and CNN, the approach effectively identifies and represents intricate patterns within RNA sequences, enabling the capture of complex interactions and features. This approach demonstrates great advantages in terms of feature extraction, enabling a more comprehensive capture of sequence features.

2.2.1. One-Hot Encoding. To convert RNA sequences into a format compatible for input data, the technique of one-hot encoding is utilized. This method enables the representation of the four RNA bases—adenine (A), cytosine (C), guanine (G), and U—as binary vectors made up of 0s and 1s. In particular, during this transformation, each nucleotide is encoded as a distinct binary vector with a length of 4. The position corresponding to the specific nucleotide is marked as 1, while the other positions are marked as 0. In detail, the one-hot encodings for the four nucleotides are (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), and (0, 0, 0, 1).[21] Therefore, this scheme provides a solid guarantee for more convenient and accurate input of data into the CNN for subsequent feature processing.

2.3. Model Framework. In this study, an integrated learning framework named m5U-HybridNet is designed, as depicted in Figure 1. On one hand, the RNA-FM model based on a 12-layer transformer is primarily employed for feature extraction. Subsequently, a FC layer is utilized to elevate the dimensionality to 1,024, aiming to further capture the nonlinear relationships within the data and enrich the feature information. On the other hand, the m5U sequences are converted into binary vectors via the application of one-hot encoding, facilitating the input of the acquired sequence information into a three-layer CNN with a 512-dimensional space. This enables a more refined and accurate capture of local information features. Subsequently, through a FC layer, there is a further enhancement in the comprehensive acquisition of feature information. Once the features from the two branches are integrated into a 1,536-dimensional space, the final task of binary classification for m5U sites is carried out via the FC layers and dropout layers. In addition, we adopted the 10-fold cross-validation approach to fully exploit the information on each data point, ensuring the robustness of the model and preventing overfitting. We also tested the model's robustness and generalization ability. Specifically, we conducted cross-validation at the biomolecular level using the full transcript pattern and the mature mRNA dataset. Moreover, with different technologies and cell types as preconditions, we further evaluated the model, thereby highlighting its substantial advantages and remarkable generalization capabilities. The distinctive aspect of the ultimate strategy further encompasses the development of an innovative, approachable web-based equipped with a highly intuitive graphical interface. Through this platform, users can conveniently input m5U sequences and swiftly obtain highly accurate and efficient predictions for m5U modification sites.

2.3.1. RNA-FM Model. In addressing the challenge of m5U modification site prediction, we leveraged the RNA-FM architecture as a feature extraction engine. RNA-FM employs a bidirectional transformer encoder framework, pretrained through self-supervised learning on a corpus of 23 million noncoding RNA sequences sourced from the RNAcentral repository. Drawing inspiration from BERT's language modeling paradigm,[20,22] the architecture incorporates 12 stacked transformer layers, each comprising a multihead self-attention module and a position-wise feedforward network with a latent dimension of 640, culminating in a total of 99 million parameters. Prelayer normalization[20] and residual skip connections are systematically integrated across all layers to stabilize gradient propagation. For an input RNA sequence of variable length L, the model computes a context-aware embedding matrix of dimensions $L \times 640$, which is subsequently condensed into a fixed-length 640-dimensional

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

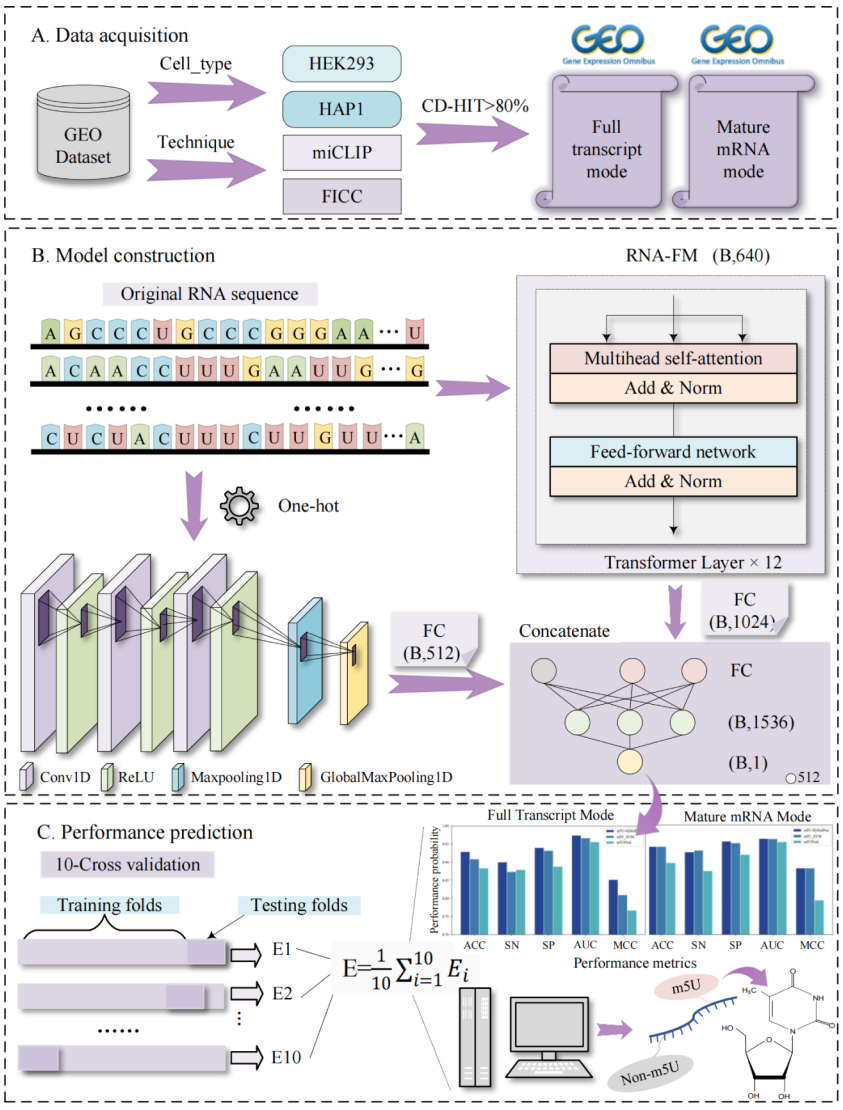


Figure 1. m5U-HybridNet model framework. (A) Data acquisition. By leveraging diverse experimental techniques and datasets from different cell types provided by Ao et al.[17] we obtained a total of 3,696 positive/negative full transcript mode datasets and 1,232 positive/negative mature mRNA mode datasets. These datasets were further processed using 0.8 cd-hit to generate imbalanced datasets. (B). Model construction. The acquired RNA sequences are processed through the RNA-FM model to extract 640-dimensional features, which are then passed through a 1,024-dimensional FC layer for dimensional transformation (Table S1). Simultaneously, the RNA sequences are converted into binary vectors via one-hot encoding. These vectors undergo local feature extraction through the three-layer 512-dimensional CNN and activation functions. The resulting features are globally integrated via a FC layer. The features obtained from this branch are fused with those from the RNA-FM branch. Finally, the combined features are passed through the FC layers to produce a one-dimensional output, which is used for binary classification via a sigmoid function. (C) Performance evaluation. The model undergoes performance evaluation using 10-fold cross-validation. Furthermore, a web-based platform Furthermore, a web-based platform has been created to streamline the prediction of m5U modification sites.

representation via global average pooling. This hierarchical feature abstraction mechanism enables robust identification of both local sequence motifs and global structural patterns critical for m5U site recognition.

Preliminary step, one of the core components of the transformer based RNA-FM model is the multihead attention mechanism. By parallelly computing multiple attention heads, it captures the multidimensional aspects of the input sequence and long-range dependencies, significantly enhancing the model's ability to parse and comprehend RNA sequences. Specifically, each self-attention layer fundamentally operates through query-key-value projections. In this process, the output is a context-aware weighted sum of value vectors, which is dictated by the affinity between the query and the key. The fundamental process of attention calculation can be described in the following manner:

$$\text{Multi-Head}(Q, K, V) = \text{Concat}(\text{head}_i, ..., \text{head}_h)W^o \quad (1)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

In the aforementioned formulas, Q denotes the query matrix, served to calculate similarity with a K matrix to determine which positions' information to concentrate on. K denotes the key matrix, the basis for calculating attention weight in

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

alignment with query matrix. V denotes the value matrix, store the actual information to be extracted, which is the aggregated value weighted by attention weights. W^o represents the weight matrix of the output layer, which maps the concatenated multihead results to the last output dimension. head_i conveys an attention layer, W_i^Q, W_i^K,W_i^V each symbolizes projecting the input into the query space of the i-th head, mapping into the key space, and projecting into the value space. The dot product between the query matrix Q and key matrix K^T generates an affinity matrix, which undergoes softmax normalization to produce attention weights. These weights quantify the pairwise relevance of each query-key vector pair, reflecting their contextual interdependence within the attention mechanism. d_k denotes the dimension of the key vector. The scaling factor d_k serves to prevent gradient vanishing or explosion during the optimization process. Lastly, the attention weights are applied to the value matrix V via a weighted summation, synthesizing a contextually enriched embedding for each positional representation.

During the subsequent step of the pretraining phase, the RNA-FM model follows the self-supervised learning strategy akin to BERT.[20] In other words, roughly 15% of the nucleotide tokens are arbitrarily substituted with unique masked tokens. There are three replacement methods. Upon the selection of the i-th token, it undergoes a transformation based on a probabilistic scheme: it is replaced with a [MASK] token 80% of the time, substituted with an arbitrary token 10% of the time, and retains its original form as the i-th token in the remaining 10% of instances. Following this, the technique of Masked Language Modeling (MLM)[20] is utilized to refine the model's predictive capabilities by estimating the original obscured tokens, leveraging cross-entropy as the loss function. Subsequently, with each masked token, as the context is furnished by the masked sequence, the objective function is designed to reduce the negative logarithmic probability of the authentic nucleotide in question. By employing this methodology, the model adeptly discerns the interrelations among the obscured segments and the remaining elements of the input sequence, thereby facilitating precise determination of the masked positions. Consequently, RNA-FM effortlessly attains a profound comprehension and nuanced portrayal of each successive token.

To further enhance contextual modeling, RNA-FM integrates multihead self-attention mechanisms where each attention head operates in parallel to capture distinct sequence interaction patterns. By computing scaled dot-product attention across all sequence positions, the model dynamically weights intertoken relationships, effectively resolving both short- and long-range dependencies. These learned representations are then aggregated through residual connections and layer normalization[20] forming context-aware embeddings that serve as robust feature inputs for downstream tasks.

Interestingly, in our quest to elucidate the rationale behind selecting the RNA-FM model for extracting features associated with m5U modification sites, we conducted a series of comparative experiments, the results of which are detailed in Table S2. Specifically, we juxtaposed the RNA-FM model against another pretrained RNA language model, RNAErnie[23] which is similarly predicated on a 12-layer Transformer architecture. To our pleasant surprise, the RNA-FM model demonstrated a marginal yet notable superiority over its counterpart. We posit that this disparity in performance may be attributable to the pronounced differences in their pretraining strategies. The RNA-FM model employs a foundational MLM strategy, which involves the stochastic masking of a portion of the input sequence (typically 15% of the tokens) and subsequently prompts the model to predict the obscured segments. This approach enables the model to concurrently capture bidirectional information from both preceding and succeeding positions relative to the masked tokens, thereby endowing it with robust generalizability and computational efficiency. Conversely, the RNAErnie model adopts a motif-aware multilevel masking strategy, augmented by RNA type label guidance. This method integrates three tiers of masking—basic, subsequence, and motif-level—yet it necessitates prior knowledge of motif positions, thereby incurring higher computational costs and a stronger dependency on data. Given these considerations, we elected to utilize the RNA-FM model to generate high-quality sequence embeddings, which effectively capture the long-range dependencies and global structural information inherent in RNA sequences.

Concurrently, we undertook a rudimentary fine-tuning of the RNA-FM model methodology, specifically experimenting with freezing the final layer or the penultimate and final layers for feature extraction. However, upon observation, it was discerned that the model's performance metrics were subpar in comparison to the unaltered RNA-FM model (Table S2). Thus, we ultimately resolved to utilize the pristine RNA-FM model as a pivotal branch for feature extraction within our overarching model architecture.

Certainly, although the RNA-FM model is specifically designed for noncoding sequences, relevant studies have demonstrated its robust generalization capabilities.[24] Moreover, we conducted comparative experiments with the mRNA-FM model,[20] which is specialized for mature mRNA data, and found that the performance metrics of the RNA-FM model slightly outperformed those of the mRNA-FM model (Table S2). We hypothesize that this outcome may stem from the fact that the mRNA-FM model primarily employs a 12-layer deep neural network architecture, with each layer comprising 1,280 hidden units. However, the quantity of mature mRNA pattern data at our disposal is relatively scarce, and the dimensionality, compared to 640 dimensions, may be insufficient to capture the intricate information and relationships within mRNA sequences.

2.3.2. Convolutional Neural Networks. CNN demonstrate remarkable performance in capturing the local characteristic information on sequences.[25] Therefore, we utilize the CNN architecture to gradually extract more intricate local features. The CNN, a widely used deep learning architecture, is commonly employed for processing data with a mesh-like structure. Its core concept involves using local receptive fields, weight sharing, and spatial down-sampling operations to extract the local features of input data.[26] In this research, through multiple layers of convolution and pooling operations, it forms complex feature representations. Ultimately, the FC layers is used for classification tasks, enabling the optimized extraction of hierarchical features from complex data. The input layer, which is one of the branches of the first layer of the m5U-HybridNet model, primarily functions to transmit the one-hot-encoded binary data to the network. This facilitates subsequent convolutional operations aimed at extracting the local features of the sequence.

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

For this study, the input sequence $X \in \mathbb{R}^{L \times d}$ has a length L of 43 and a feature dimension d of 640. Note that padding the sequence tail with two spaces to reach a length to 43 (Table S3) is intended to align with the sequence length processed by the RNA-FM model. When using the RNA-FM model to process the data, the sequence information on ⟨CLS⟩ and ⟨EOS⟩ is retained. A three-layer CNN is employed.[20] Each layer has 512 filters with a kernel width of 9 (Tables S4, S5, and S6). Subsequently, batch normalization is applied to mitigate issues such as gradient vanishing or explosion. The data then passes through the CNN module consisting of a max-pooling layer, a global max-pooling layer, and a ReLU activation function. The CNN module distills the time-domain features into fixed-dimensional embedding space, thereby reducing the number of parameters while retaining the most crucial features.

2.3.3. Feature Fusion. Feature fusion serves as a pivotal enabler in contemporary classification and regression tasks, especially when dealing with complex problems.[27−29] It is crucial for enhancing model performance and generalization capabilities.[30] In this study, we employ an ingenious feature fusion approach that combines global semantic features extracted by the RNA-FM model with local sequence features identified by CNN, utilizing a concatenation operation followed by a FC layer. The concatenation operation adeptly merges features from disparate sources along the channel dimension, thereby circumventing the potential information loss associated with feature summation. This is particularly advantageous for the heterogeneous feature integration of local and global RNA features. Subsequently, the FC layer further elucidates the interaction patterns of the concatenated features through nonlinear transformations, offering a flexibility that far surpasses the rudimentary feature fusion methods of addition or multiplication. Moreover, this method imposes no mandatory alignment requirements on the input feature channels, thereby eschewing the additional overhead of 1 × 1 convolutional operations necessitated by addition or multiplication to adjust channel dimensions. To demonstrate the superiority of our approach, we project the output channels of the two distinct branches onto the same dimension and compare them with the methods of feature addition and multiplication (Table S7). Consequently, from a holistic perspective, our concatenation and FC fusion approach generally outperforms alternative feature integration techniques in terms of performance metrics. This flexible fusion strategy provides ample complementary information, thereby augmenting the model's decision-making capacity, mitigating model bias, and facilitating the creation of a more robust and accurate model. It also simplifies the handling of diverse input scenarios, thereby enhancing the overall performance of the model.

Using Rectified Linear Unit (ReLU) as the activation function throughout the feature extraction and feature fusion processes is also advantageous for several reasons. ReLU not only alleviates the vanishing gradient problem and helps prevent overfitting, but also features a computationally efficient design, enabling the model to perform binary classification tasks more effectively and conveniently. Meanwhile, Adam[31,32] is selected as the optimization evaluator. Widely adopted in deep-learning optimization algorithms, Adam combines the concepts of the momentum method and RMSProp.[32] It implements an adaptive learning rate adjustment mechanism by dynamically computing the first-order moment estimates (mean) and second-order moment estimates (uncentered variance) of gradients, substantially improving training convergence rates through parameter-specific optimization. It adjusts the learning rate of each parameter by computing the first-order moment estimate and the second-order moment estimate of the gradients, thereby enabling more efficient network training.

2.3.4. Loss Function. Meanwhile, the binary cross-entropy loss function is employed in the model's binary classification tasks and measures the difference between the model's predicted probabilities and the actual labels, thereby enabling more precise classification. The equation for the binary cross-entropy loss function is presented below:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)] \tag{4}$$

In this context, N represents the sample count and is used to normalize the subsequent summation term, ensuring that the loss calculation is based on an average value. This normalization prevents the sample size from affecting the loss measurement. The index i denotes the iteration from 1 to N, serving as the index for the summation. y_i indicates the actual label of the i-th sample, while $\hat{y}_i$ signifies the predicted likelihood that the i-th sample is in the affirmative category. In our study, we also adopted an early stopping mechanism with a patience threshold of 20. When the performance on the validation set no longer improves, training is promptly halted to prevent the model from overfitting due to continued learning. For detailed parameter settings, please refer to Table S8 in the supplementary document.

2.3.5. Model Evaluation Metrics. The systematic quantification of model efficacy metrics constitutes a critical determinant for achieving high-precision identification of m5U modification loci. We therefore employ six widely adopted decisive metrics to assess model performance and reliability:[33,34] (1) Sensitivity (SN) measures the true positive rate, prioritizing minimization of false negatives to ensure comprehensive detection of m5U sites. (2) Specificity (SP) evaluates true negative recognition, critical for reducing spurious predictions in nonmodified regions. (3) F1-Score (F1) serves as the balance point between precision and recall, addressing the inherent trade-off between these two metrics. (4) Accuracy (ACC) reflects the intuitive proportion of correctly classified samples. While naturally influenced by class distribution, we mitigate this limitation through the class_-weight parameter to balance loss function weights across categories. (5) Matthews Correlation Coefficient (MCC) offers a holistic assessment of classification performance by taking all elements of the confusion matrix into account. (6) Area Under the ROC Curve (AUC) demonstrates particular value through its insensitivity to class distribution, excelling in assessing a model's ranking capability.[35−37] These performance metrics can be formally expressed using the following definitions, where TP, FN, TN, and FP respectively denote counts of true positives, false negatives, true negatives, and false positives. The mathematical expressions for these metrics are as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

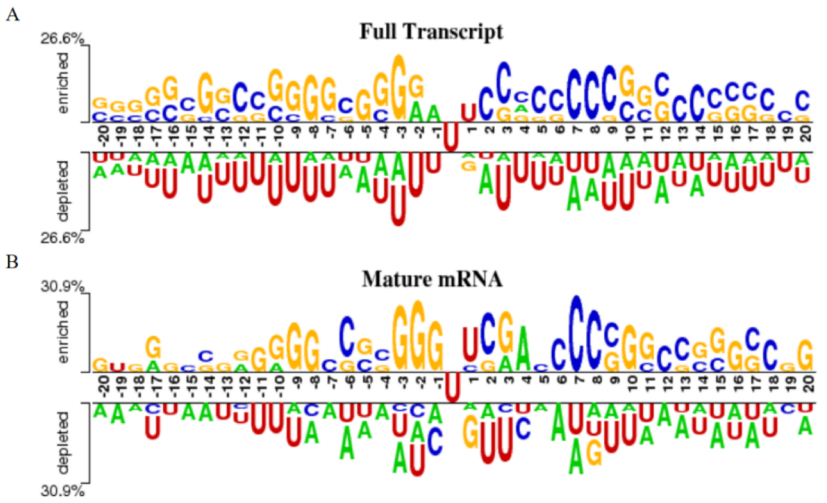


Figure 2. An in-depth examination of nucleotide composition is conducted across both the full transcript and mature mRNA modalities. The upper section of the panel highlights the locations of m5U modification sites, while the lower section delineates regions devoid of m5U modifications.

Table 2. Assessment of the Impact of Single-Branch Models on Predictive Outcomes in Full Transcript Mode and Mature mRNA Mode[a]

Mode	Methods	ACC	SN	SP	AUC	MCC	F1
Full transcript mode	RNA-FM	0.8781	0.8320	0.9097	0.9399	0.7463	0.8320
	CNN	0.9212	0.8820	**0.9480**	0.9739	0.8361	0.9009
	m5U-HybridNet	**0.9285**	**0.9000**	0.9400	**0.9745**	**0.8514**	**0.9109**
Mature mRNA mode	RNA-FM	0.9309	0.8980	**0.9636**	0.9591	0.8636	0.9283
	CNN	0.9390	0.9265	0.9514	0.9613	0.8745	0.9356
	m5U-HybridNet	**0.9411**	**0.9265**	0.9555	**0.9631**	**0.8825**	**0.9400**

[a]Bold indicates the best performance of the method.

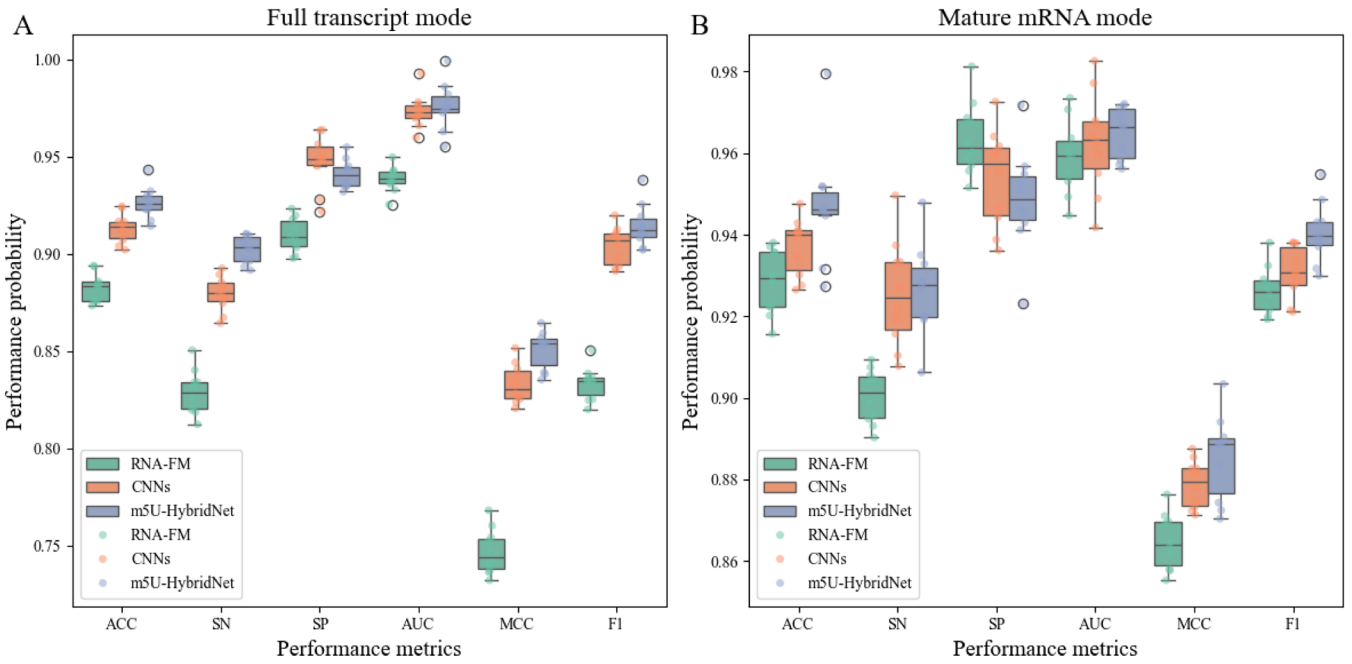


Figure 3. Impact of single-branch models on the evaluation of prediction results. The horizontal axis describes various performance metrics, including metrics such as ACC, SN, SP, AUC, MCC, and F1. The *y*-axis illustrates represents the performance scores predicted by the model. Green represents the single-branch RNA-FM method, orange denotes the single-branch CNN method, and blue signifies the m5U-HybridNet model. (A) Full transcript mode. (B) Mature mRNA mode.

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

$$SN = \frac{TP}{TP + FN} \quad (7)$$

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

Table 3. Comparison of the Performance between Full Transcript Mode and Mature mRNA Mode with Existing Deep Learning Models[a]

test	Methods	ACC	SN	SP	AUC	MCC	F1
Full transcript mode	BiGRU	0.8814	0.8540	0.9001	0.9454	0.7541	0.8540
	BiLSTM	0.8920	0.8440	0.9248	0.9503	0.7750	0.8639
	RNN	0.8733	0.8400	0.8960	0.9517	0.7370	0.8434
	Transformer	0.8765	0.8300	0.9083	0.9410	0.7429	0.8452
	m5U-HybridNet	**0.9285**	**0.9000**	**0.9400**	**0.9745**	**0.8514**	**0.9109**
Mature mRNA mode	BiGRU	0.9370	0.9102	0.9636	0.9553	0.8752	0.9350
	BiLSTM	0.9390	0.9143	0.9636	0.9557	0.8791	0.9372
	RNN	0.9126	0.8857	0.9393	0.9557	0.8263	0.9099
	Transformer	0.9329	0.9020	**0.9636**	0.9600	0.8674	0.9305
	m5U-HybridNet	**0.9411**	**0.9265**	0.9555	**0.9631**	**0.8825**	**0.9400**

[a]The bold-faced entries represent the methods achieving the best performance.

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

$$F1 = 2 \times \frac{\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times \mathrm{SN}}{\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} + \mathrm{SN}} \tag{9}$$

Regarding our methodology of employing "class_weight" to address the issue of data imbalance, we conducted comparative experiments with untreated imbalanced data approaches and oversampling techniques for imbalanced datasets (Table S9). The outcomes demonstrated that the method utilizing "class_weight" nearly outperformed the other methods across all metrics. We surmise that the underlying reason may be that the class weighting approach automatically assigns a higher weight to the minority classes in conjunction with the loss function, thereby enhancing the predictive prowess of the model.

Cross-validation differs somewhat from the ordinary division of dataset ratios. Primarily employed for assessing and substantiating the accuracy of data classification models, cross-validation entails partitioning a dataset into several segments and successively training and evaluating the model on these subsets multiple times. This methodology evaluates the model's robustness and capacity for generalization, with the goal of mitigating the propensity for overfitting throughout the training phase and bolstering the model's efficacy when applied to novel, unseen datasets. In this investigation, the 10-fold cross-validation technique is meticulously applied to the training dataset, which is divided into 10 nearly equal subsets, with each subset serving as the validation set in rotation while the remaining subsets are used for training, a process repeated 10 times to generate comprehensive results that are aggregated to compute an average value for benchmarking training and test set evaluation, thereby enhancing outcome reliability and model robustness and establishing a solid foundation for assessing the model's generalization capabilities on diverse and unseen datasets.

3. RESULTS AND DISCUSSION

3.1. Nucleotide Composition Analysis. Nucleotide composition is not only an indispensable component of biological genetic information and the carrier of genetic cryptography but also the central hub for regulating life activities. Therefore, to investigate whether there are specific biases in the composition of nucleotides at m5U modification sites, we employed the Two-Sample Logo software.[38] Employing this software, we conducted a comparative analysis to discern the disparities between sequences harboring m5U modification sites and those devoid of such modifications. This enabled us to distinctly illustrate the residues that are markedly enriched or diminished within the m5U-modified segments of the two-sample logos. These visualizations facilitate in-depth observation and analysis. As shown in Figure 2, a t test was conducted to carry out the statistical analysis, establishing a threshold of significance at $p < 0.05$. This allows for the observation of a clear distinction in the resulting plots. The full transcript mode is presented in Figure 2A, while Figure 2B illustrates the nucleotide composition distribution of the mature mRNA model sequences, both of which include m5U and non-m5U modification sites. In Figure 2, shared motifs are observed between the full transcript and mature mRNA modes. For instance, the motif "UUC" is consistently present at positions 0−2 in both cases. Interestingly, the figure also reveals a position-specific nucleotide enrichment pattern for m5U modification sites in both full transcripts and mature mRNA. For instance, U predominantly accumulates at position +1, G clusters at positions −8 and −3, cytosine C exhibits a significant abundance at the +7 and +8 positions. A shows distinct localization patterns: it primarily accumulates at position −1 in full transcripts, while in mature mRNAs, it exhibits a highly concentrated and dominant distribution at position +6. This differential analysis precisely reveals the positional specificity of m5U modification sites across distinct sequence contexts. Notably, at position +19, transcript-wide modification patterns predominantly feature C enrichment, whereas mature mRNA modification sites exhibit guanine G predominance. Similarly, at position −1, transcript-wide patterns demonstrate A enrichment in contrast to the G preference observed in mature mRNA modification sites. The analysis of nucleotide deviations, which demonstrated the high effectiveness of sequence information in identifying RNA m5U modification sites, collectively provided our nucleotide composition analysis with critical biological insights. Further still, these insights were instrumental in informing model design and enhancing the interpretation of predictive features.

3.2. Ablation Experiment Identified the Potency of m5U-HybridNet. Ablation analyses are indispensable for gaining insights, refining, and advancing the development of models. Their inclusion serves to substantiate the efficacy of the model's constituent elements, optimize model performance, and enhance interpretability and generalization capa-

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

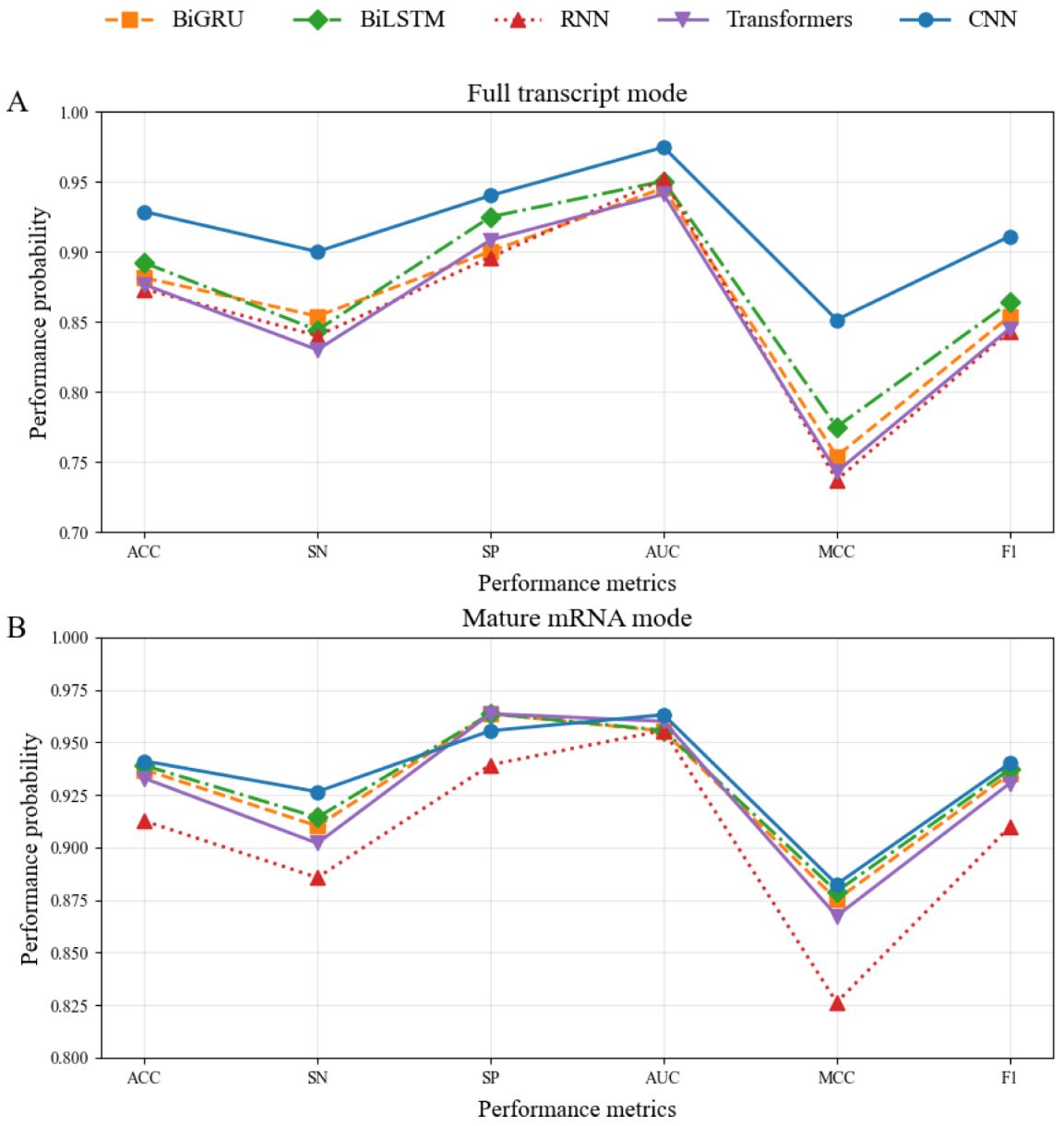


Figure 4. Examining the efficacy of an array of deep learning techniques on a separate test dataset, we've conducted a performance analysis in both the full transcript and mature mRNA configurations. The horizontal axis encompasses a suite of performance indicators, namely ACC, SN, SP, AUC, MCC, and F1 Score, while the vertical axis quantifies the performance outcomes for these metrics. The color blue is designated for the CNN-based feature processing paradigm. Distinct color codings-orange for BiGRU, green for BiLSTM, red for RNN, and purple for the Transformer model are used to differentiate the various deep learning methodologies. The findings are presented in two distinct panels: Panel A for the full transcript configuration and Panel B for the mature mRNA configuration.

bilities.[35] Therefore, in this paper, we employed ablation experiments to systematically remove or modify the components of these modules one by one, aiming to comprehensively assess the overall contribution of each module in the m5U-HybridNet model to its final performance. In the subsequent research, we will conduct separate model training on the dual-branch hybrid model based on RNA-FM and CNN feature processing. We will utilize the most commonly used critical metrics, including SN, SP, F1, ACC, MCC, and AUC. These metrics will allow us to closely scrutinize and compellingly showcase the influence that each branch exerts on the model's overall efficacy. Surprisingly, the performance metrics of the model based on global feature information extracted from the RNA-FM large model are slightly lower than those derived from local feature information processed by CNN.

This discrepancy may be attributed to the fact that the extracted global features might introduce redundant information, thereby diluting the critical information within the sequence. In contrast, the local features extracted by CNN can precisely identify key functional sites, suggesting that the characteristic information on m5U modification sites might be predominantly governed by specific local sequence patterns. As observed in Table 2, across both full transcript and mature mRNA datasets, the competence metrics of the three-layer CNN consistently exhibit slightly higher values than those of the RNA-FM model. This clearly indicates that fusing features extracted by the RNA-FM model with those processed by the three-layer CNN via the FC layers can yield superior model

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

Table 4. An Assessment of Comparative Efficacy between the Full Transcript and Mature mRNA Modalities against Current Cutting-Edge Methodologies is Conducted[a]

Datesets_test	methods	ACC	SN	SP	AUC	MCC
Full transcript mode	m5UPred	0.8432	0.7480	0.9083	0.9253	0.6723
	m5U_SVM	0.9082	0.8740	0.9316	0.9667	0.8092
	Deep-m5U	0.6044	0.6560	0.5691	0.5682	0.2213
	m5UGEPred	0.8261	0.8262	0.8043	0.8819	0.6374
	m5U-HybridNet	**0.9285**	**0.9000**	**0.9400**	**0.9745**	**0.8514**
Mature mRNA mode	m5UPred	0.8923	0.8531	0.9312	0.9477	0.7868
	m5U_SVM	0.9411	**0.9306**	0.9514	0.9621	0.8823
	Deep-m5U	0.7012	0.7224	0.6802	0.7165	0.4029
	m5UGEPred	0.8720	0.8719	0.8719	0.9233	0.7439
	m5U-HybridNet	**0.9411**	0.9265	**0.9555**	**0.9631**	**0.8825**

[a]The entries highlighted in bold signify those approaches that have attained the pinnacle of performance.

performance metrics. Consequently, by examining the specific distributions of box plots and the variations in scatter plot distributions shown in Figure 3, distinct disparities in performance metrics between the single-branch configurations of our model and m5U-HybridNet can be systematically observed.

It is noteworthy that the branch of the RNA-FM model which extracts features is not directly interpretable with respect to our dataset; hence, our focus will be primarily on the CNN branch. Observations from Figure S1 in the supplementary document reveal that the CNN branch is still capable of effectively utilizing certain motifs, and it has also identified discriminative features beyond those encompassed by the combinatorial analysis (i.e., areas not surrounded by dashed boxes in the figure). Through this analysis, we have been able to discern that the model has adeptly captured certain patterns indicative of m5U modification sites. Consequently, these insights not only substantiate that one CNN branch of our model can learn from compositional biases identified in nucleotide analysis but also further elucidate that our dual-branch model will be more conducive to capturing sequential information on nucleotide composition and uncovering additional discriminative features identified in compositional analysis.

In summary, by leveraging the feature fusion approach of the RNA-FM and CNN dual-branch architecture, we can more comprehensively capture the characteristic information on m5U modification sites, which greatly aids in our in-depth understanding of RNA sequence information.

3.3. Comparison with Other Deep Learning Methods. Currently, there is a wide variety of deep learning models available. Therefore, determining which deep learning model to use for the m5U dataset, and selecting the most suitable model from numerous candidates for optimal performance on this dataset, is of great importance. In order to assess the efficacy of various deep learning models in handling sophisticated and nuanced features, we executed comparative trials encompassing CNN,[26] Bidirectional Gated Recurrent Unit (BiGRU),[39] Transformer,[22] Recurrent Neural Network (RNN),[40] and Bidirectional Long Short-Term Memory (BiLSTM).[41] These experiments were designed to highlight the advantageous characteristics of CNN, and we performed clear and explicit numerical comparisons using the six widely adopted performance evaluation metrics mentioned above. As shown in Table 3, the table systematically presents a comparison of various deep learning models with CNN across different evaluation metrics. While every deep learning model demonstrates a satisfactory level of proficiency in identifying local correlations and sequential patterns, they still encounter certain constraints when confronted with the multifaceted and intricate attributes of the dataset. To exemplify, while BiLSTM and BiGRU enable bidirectional information capture, their strong reliance on processing high-dimensional features and temporal dependencies introduces slight limitations. While RNN perform exceptionally well in processing sequential data, they still lag slightly behind CNN, which excel innately at managing local features and spatial hierarchies. Although the transformer model is characterized by powerful representational capabilities and parallel computing advantages, refusing it with the 12-layer transformer-based RNA-FM model can result in over-representation, degrade model interpretability, and potentially induce overfitting issues. Therefore, we employ the CNN model to capture local feature information in Figure 4, which is then combined with the global features extracted by the RNA-FM model. This integration further enhances the model's performance and generalization capability, effectively improving its capacity to differentiate m5U modification sites from those that are non-m5U modified.

3.4. Comparison of Existing Premier Techniques. In order to evaluate the efficacy of our novel m5U-HybridNet model for the prediction of m5U modification sites, we have chosen two existing models for comparative analysis, m5UPred,[16] m5U_SVM,[17] Deep-m5U,[15] m5UGEPred,[13] for comparison. Notably, both baseline models were evaluated using imbalanced datasets, ensuring a consistent framework for performance assessment under comparable data distribution challenges. A summary of the performance comparisons with existing models is presented in Table 4. From the table, we observe that for the full transcript mode, the ACC, SN, SP, MCC, and AUC evaluation metrics are 92.85%, 90.00%, 94.00%, 85.14%, and 97.45%, respectively. For the mature mRNA mode, the corresponding ACC, SN, SP, MCC, and AUC evaluation metrics are 94.11%, 92.65%, 95.55%, 88.25%, and 96.20%, respectively. As shown in Figure 5, it is evident that our model outperforms other models across almost all metrics, both in full transcript mode and the mature mRNA mode. Thus, the comparisons demonstrate that our model consistently outperforms existing methods slightly in nearly every aspect, highlighting its substantial advantages.

Such results not only underscore the superiority of our proposed model but also provide a valuable and practical computational tool for identifying m5U modification sites.

3.5. Model Visualization Demonstrates the Performance of the Model. The kernel density estimation (KDE)

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

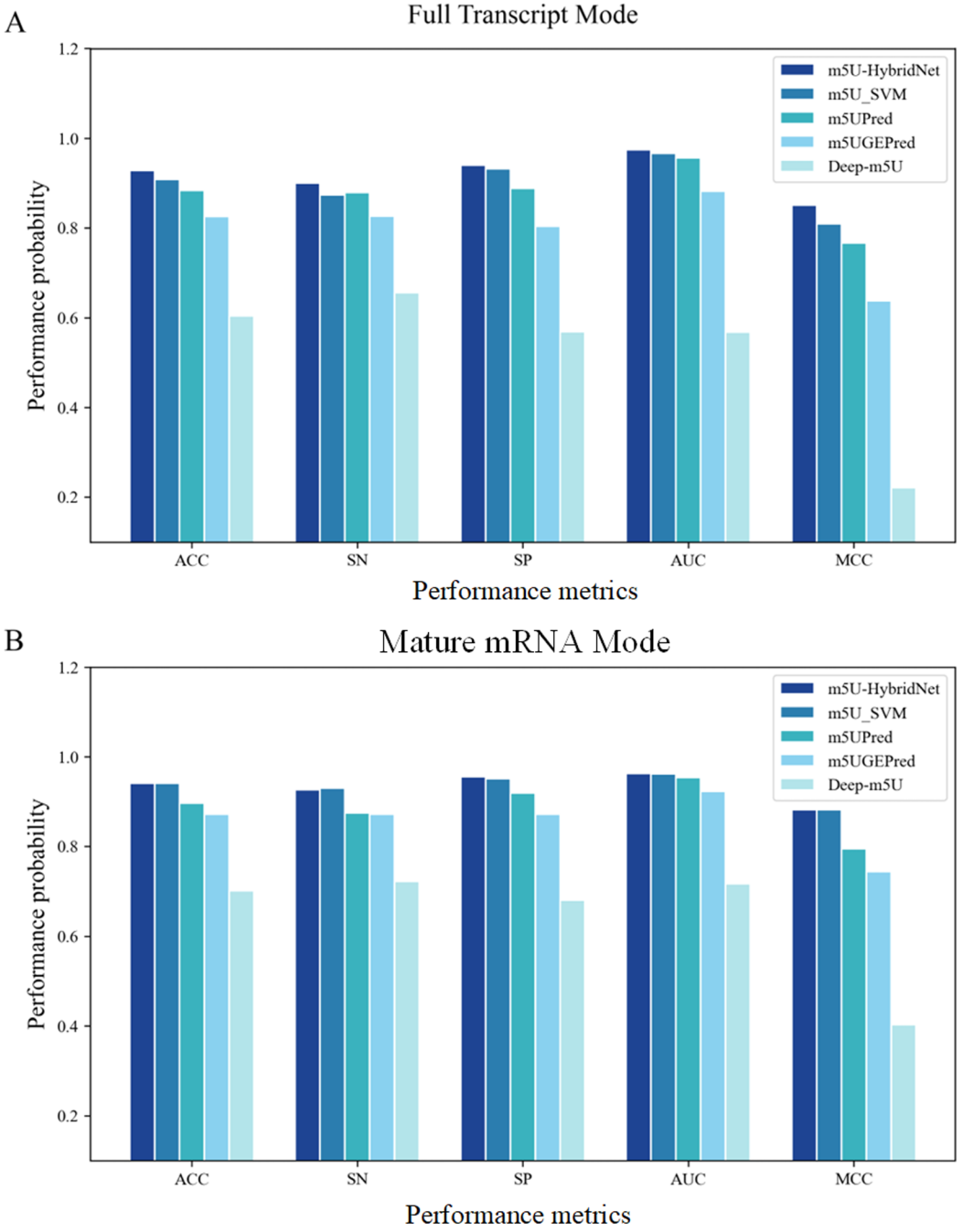


Figure 5. An exhaustive performance evaluation of diverse predictive models for identifying m5U modification sites is presented, encompassing both the full transcript and mature mRNA modalities, as assessed on an independent test dataset. The *x*-axis encompasses a spectrum of metrics, including ACC, SN, SP, AUC, MCC, while the *y*-axis delineates the respective performance metrics. In stark contrast to the m5UPred and m5U_SVM models, our innovative m5U-HybridNet model exhibits pronounced superiority and distinctive attributes. Panel A illustrates the findings for the full transcript modality, whereas Panel B corresponds to the mature mRNA modality.

plot, a specialized nonparametric estimation method for visualizing probability density distributions, employs smoothed curves to approximate the underlying distribution of data.[42] As illustrated in Figure 6, the affirmative and negative specimens in both full transcript mode and mature mRNA mode are remarkably well-separated. The positive samples predominantly concentrate near 1.0, while negative samples cluster primarily around 0.0. The magnitude of density distributions in panels A and B elegantly demonstrates a substantial predominance of negative samples over positive ones, highlighting the model's exceptional classification accuracy across both validation and test sets. Furthermore, comparative analysis between validation and test sets reveals that while the full transcript test set exhibits partial overlap within the 0.4−0.6 range compared to its validation counterpart, the mature mRNA test set maintains nearly identical KDE patterns to its validation set. Crucially, however, the majority of test samples in both modes remain correctly discriminated, with dense aggregation at their respective extremal regions, thereby reinforcing the model's robust generalization capability and precise categorical differentiation.

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

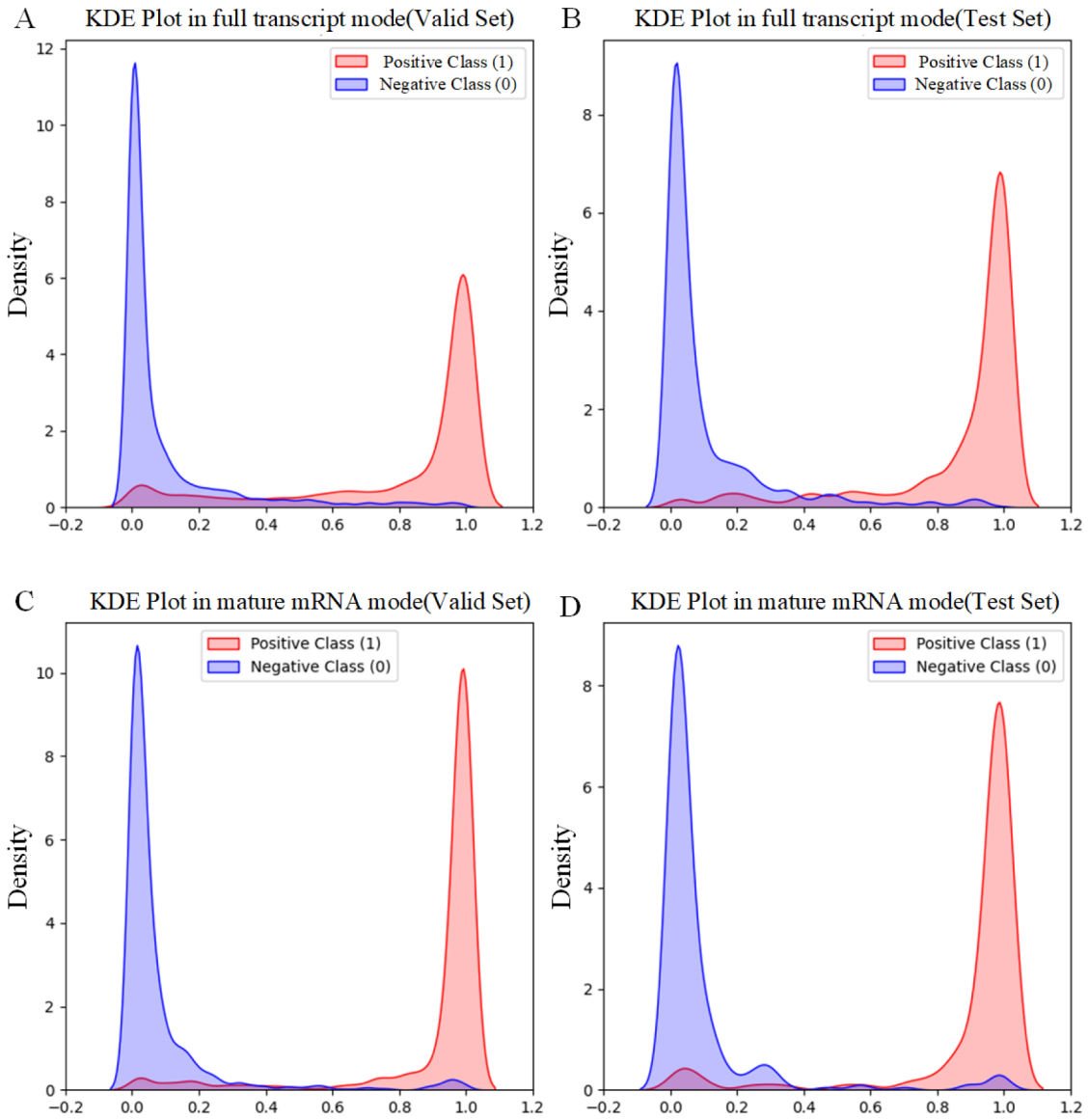


Figure 6. Model Visualization. Panels A and B display the KDE plots of the average prediction probabilities from 10-fold cross-validation and the test set under the full transcript mode. Panels C and D showcase the Kernel Density Estimation (KDE) plots, which represent the average prediction probabilities derived from a 10-fold cross-validation process and the test dataset, both conducted within the mature mRNA mode. The *x*-axis delineates the predicted probabilities assigned by the model to both positive and negative samples, which predominantly span the interval from −0.2 to 1.2. The *y*-axis quantifies the sample density across various predicted probability ranges. The blue and red curves correspond to the model's predicted probability distributions for positive and negative samples, respectively.

3.6. Cross Validation. The m5U dataset in full transcript mode encompasses all types of RNA (e.g., tRNA, mRNA, snRNA, pre-RNA, etc.),[43] capturing the dynamic modifications throughout the entire process of RNA synthesis to maturation. In contrast, the m5U datset in mature mRNA mode focuses specifically on mRNA molecules that have undergone polyadenylation and splicing,[44] emphasizing aspects such as translation efficiency, reliability, subcellular localization,[45] and functional roles in diseases. Here, we aim to investigate the modification patterns and functional divergence of m5U across different stages of the RNA life cycle. Additionally, we will explore the modification patterns and functional differentiation of m5U modification during different stages of the RNA lifecycle, including different cell types (HEK293 and HAP1) and different experimental techniques (miCLIP and FICC).

3.6.1. Cross-Validation between Full Transcript Mode and Mature mRNA Mode. In this investigation, we conducted rigorous cross-validation analyses between full transcript profiling and mature mRNA sequencing modalities for m5U modification characterization. The scatterplots in Figure 7A,B demonstrate distinct distribution patterns of prediction outcomes under different training-testing paradigms. Notably, when employing the full transcript dataset as the test set (Figure 7A), both positive and negative samples predominantly clustered near the 0−1 decision boundaries, revealing suboptimal predictive performance when utilizing mature mRNA profiles for model training. Conversely, the mature mRNA test set exhibited superior classification accuracy across various training configurations, with samples being accurately segregated into their respective categories (Figure 7B).

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

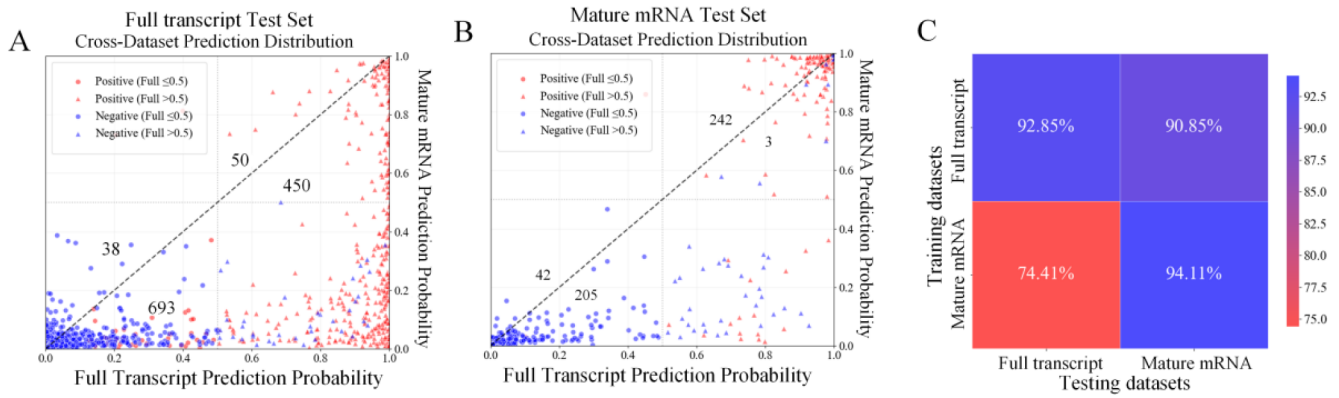


Figure 7. Panels A and B depict the test sets of the full transcript mode and mature mRNA modalities, respectively. The *x*-axis of both panels represents the training set of the full transcript mode, whereas the *y*-axis corresponds to the training dataset of the mature mRNA modality. Figure C presents a heatmap depicting the cross-validation between the full transcript and mature mRNA modalities. It is noteworthy that the full transcript and mature mRNA modalities are denoted as the test and training datasets on the *x*-axis and *y*-axis, respectively. The greater the likelihood, the deeper the blue, indicating a closer similarity between them.

Table 5. Performance Comparison of Cross-Technology and Cross-Cell Type Validation in the Full Transcript Mode and Mature mRNA Mode[a]

		Cross_technique			Cross_cell_type		
Datasets	Model	miCLIP_acc (%)	FICC_ acc (%)	Avg_ acc (%)	HEK293_acc (%)	HAP1_acc (%)	Avg_acc (%)
Full transcript mode	m5UPred	82.29	73.29	77.79	86.20	73.99	80.10
	m5U-GEPred	86.26	90.83	88.55	74.82	78.71	76.77
	Deepm5U	95.64	93.90	94.77	97.10	94.33	95.72
	m5U-HybridNet	**95.94**	**94.97**	**95.46**	**97.30**	**95.26**	**96.28**
Mature mRNA mode	m5UPred	90.46	68.16	79.31	92.17	68.57	80.37
	m5U-GEPred	83.20	71.19	77.20	77.88	63.20	70.54
	Deepm5U	98.28	95.06	96.67	98.62	94.88	96.75
	m5U-HybridNet	**99.16**	**96.31**	**97.74**	**99.03**	**95.89**	**97.46**

[a]miCLIP_acc: the accuracy percentage evaluated using miCLIP as the training dataset and FICC as the testing dataset. FICC_acc, HEK293_acc, HAP1_acc, and so on follow the same logic. Avg_acc represents the average percentage of ACC values across different technologies or cell types.

The heatmap visualization in Figure 7C provides a comprehensive comparison of model performance metrics, highlighting substantial variations in generalization capabilities across experimental conditions. Strikingly, our 10-fold cross-validation achieved exceptional accuracy (ACC = 90.85%) when training on full transcript data and testing on mature mRNA profiles. The inverse configuration yielded diminished performance (ACC = 74.41%), potentially attributable to the significant disparity in dataset scale between full transcript and mature mRNA samples, which may introduce training-testing distributional shifts.

These systematic evaluations substantiate that our computational framework maintains remarkable robustness and generalizability across heterogeneous testing scenarios, particularly demonstrating enhanced predictive power when leveraging larger-scale full transcript data for model training. The observed performance differentials underscore the critical importance of dataset compatibility in cross-modal biological sequence analysis.

3.6.2. Multimethod and Multicellular Validation. Throughout this part, cross-validation experiments were conducted for m5U modification sites using different cell types (HEK293 and HAP1) and techniques (miCLIP and FICC). Of particular interest is the fact that our dataset features a positive-to-negative sample ratio of 1:10 (Table S10). This dataset originates from Jiang et al.,[16] where the negative samples were generated by randomly selecting unmodified uridine sites located on the same transcript as the positive m5U sites, resulting in the creation of ten distinct negative datasets. Each of these negative sets was then combined with the positive dataset to construct ten separate datasets, each maintaining the same ratio of positive to negative samples. Building on that we not only employed 10-fold cross-validation to evaluate performance via ACC scores during the assessment phase but also compared the generalization ability with state-of-the-art models including m5UPred,[16] m5U-GEPred,[13] and Deepm5U.[14] The results clearly demonstrate in Table 5 that our model exhibits significantly superior generalization capability over other models, particularly for the HEK293 and HAP1 cell types. As an illustration, the full transcript mode of the m5U-HybridNet model achieved accuracies of 95.94% and 94.97% on a detached test dataset using miCLIP and FICC techniques, respectively, with an average accuracy of 95.46%. On the independent test dataset for the HEK293 and HAP1 cell types, the model's accuracy reached 97.30% and 95.26%, respectively, averaging 96.28%. And the mature mRNA mode achieved accuracies of 99.16% and 96.31% on a detached test dataset using miCLIP and FICC techniques, respectively, with an average accuracy of 97.74%. On the independent test dataset for the HEK293 and HAP1 cell types, the model's accuracy reached 99.03% and 95.89%, respectively, averaging 97.46%. In comparison with the top-performing existing frameworks, this represents an improvement of approximately 19% for cell type-

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

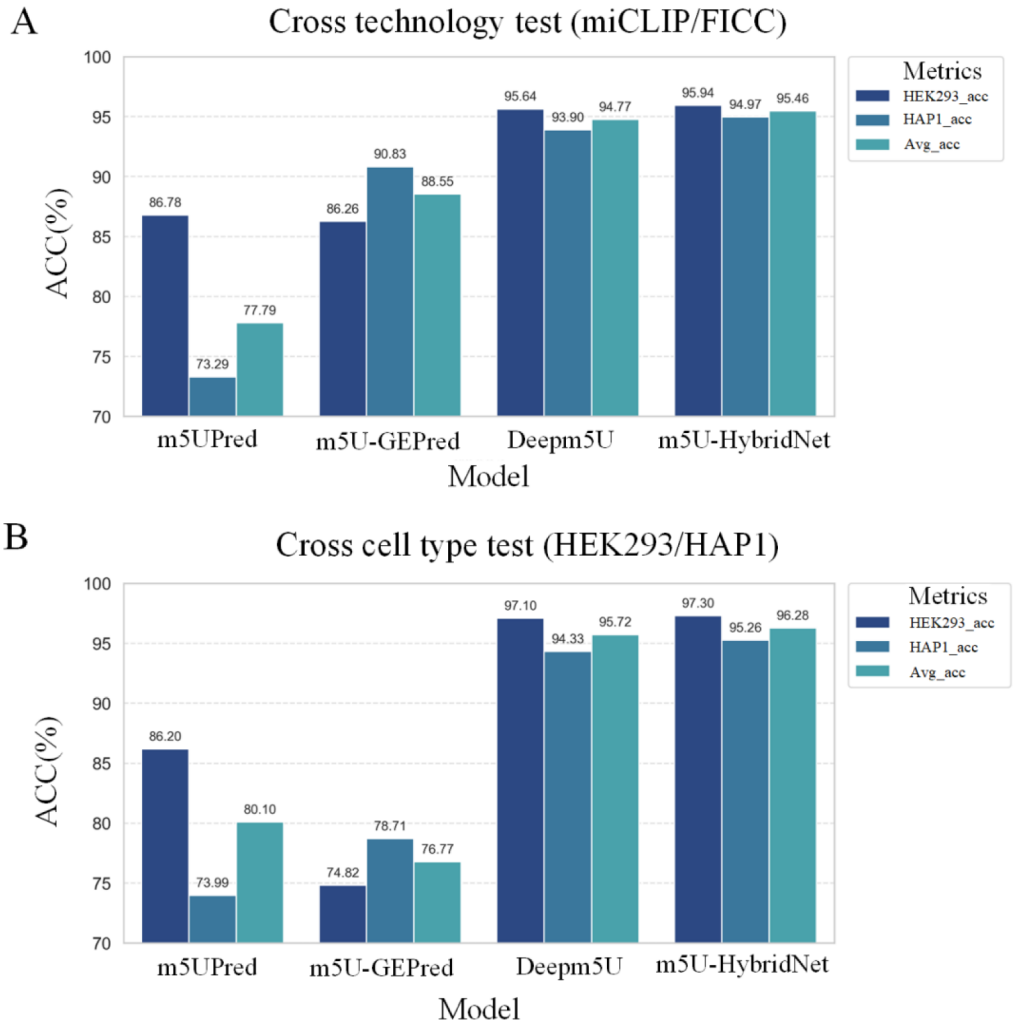


Figure 8. Bar chart of the performance comparison of cross-cell type and cross-technology validation. The *x*-axis shows various m5U prediction models, including m5UPred, m5U-GEPred, Deepm5U, and the proposed m5U-HybridNet. The *y*-axis indicates the percentage probability of ACC. miCLIP_acc and FICC_acc represent the accuracy values of m5U modification site prediction using different techniques. HEK293_acc and HAP1_acc describe the accuracy values for different cell types. Avg_acc is the average accuracy across techniques or cell types.

specific predictions. Likewise, the above-mentioned results can also be distinctly demonstrated in the bar chart of Figure 8.

t-SNE visualization, a nonlinear dimensional reduction technique, effectively discloses potential high-dimensional data distributions. For a more systematic model performance evaluation, this study adopts a multidimensional feature visualization strategy. It creates 3D t-SNE projections for features extracted solely from the CNN branch, solely from the RNA-FM branch, and from the m5U-HybridNet model's fused feature space. As shown in Figures 9 and S2, the visualizations demonstrate that, compared to single-branch models, m5U-HybridNet exhibits a superior ability to separate positive and negative samples in the t-SNE space. The increased distance between the two sample clusters and the reduced overlap highlight its advantages. This finding robustly confirms the m5U-HybridNet model's strong generalization capability across multiple technical platforms and heterogeneous cell systems. It also underscores the model's precise discrimination in identifying epitranscript modification sites.

Given the current absence of definitive and uniform data regarding the positive-to-negative sample ratio of m5U modification sites in nature, we have ventured to postulate a scenario where the originally near 1:2 imbalanced dataset is further partitioned into a significantly more skewed ratio of 1:100, albeit without any pertinent references to corroborate this assumption. Intriguingly, the m5U-HybridNet model has been found to retain its remarkable performance in terms of the ACC metric under such extreme conditions (Table S11). Specifically, for the full transcript mode, the ACC evaluation metric stands at an impressive 97.57%, while for the mature mRNA mode, the corresponding ACC is a commendable 95.45%. This suggests that the m5U-HybridNet model may well possess the capability to deliver accurate and superior predictions even when confronted with highly imbalanced datasets. Furthermore, in addition to these findings, we have also conducted comparative experiments with other models, namely Deep-m5U[15] and m5U-GEPred,[13] under an imbalance ratio of 1:100. In terms of the ACC prediction metric, our model's performance remains unparalleled, thereby underscoring its robust generalization ability and resilience.

In summary, the proposed m5U-HybridNet model demonstrates highly efficient and superior generalization capabilities across these two different approaches.

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

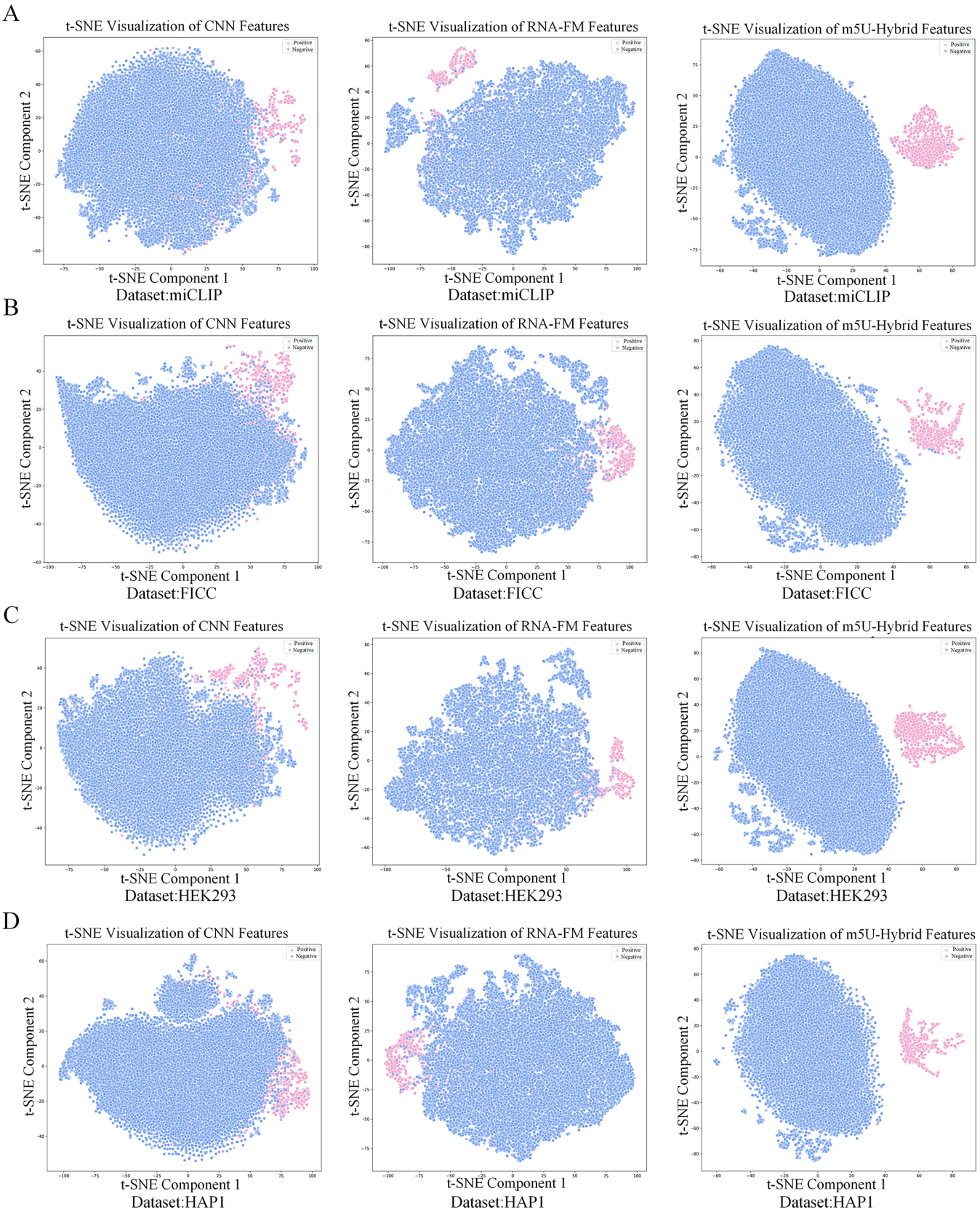


Figure 9. t-SNE visualization in the full transcript mode. Each column depicts the t-SNE visualization of features extracted exclusively through the CNN branch, exclusively through the RNA-FM branch, and exclusively through the HybridNet model. (A) miCLIP signifies that miCLIP is utilized as the training dataset, while FICC serves as the test dataset for model training. (B) FICC denotes the model training with FICC as the training dataset and miCLIP as the test dataset. (C) HEK293 represents the model training with HEK293 as the training dataset and HAP1 as the test dataset. (D) In the HAP1 scenario, the model undergoes rigorous training on the HAP1 dataset, followed by a comprehensive evaluation on

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

Figure 9. continued

the HEK293 dataset to assess its performance. The horizontal and vertical coordinates correspond to distinct principal directions of variation captured within the high-dimensional space. Pink denotes positive samples, and blue denotes negative samples.

4. WEB SERVER DEVELOPMENT

To make RNA modification research more accessible and enhance the accessibility of the m5U-HybridNet model technology, we've developed a smart, user-friendly online prediction platform. It requires no programming or computational modeling expertise.

Just two simple steps to classify m5U modification sites: Initially, upon accessing the website, please engage the "Try the Prediction Tool" feature. Subsequently, opt for distinct dataset modes or upload a FASTA file, and then activate the "Submit for Prediction" process, whereby you will promptly receive m5U modification site predictions within a matter of seconds. Alternatively, visitors may elect to engage the "Load Sample FASTA" functionality to execute a binary classification prediction of m5U modification sites. Click the Web server link http://www.bioai-lab.com/m5U to access it.

5. CONCLUSION

m5U modification sites result from the methylation of uridine at the fifth carbon atom. Our research presents m5U-HybridNet, a novel model for the binary classification prediction of these modification sites, with the objective of broadly applying and promoting the model to newly sequenced transcriptomes for de novo identification of m5U sites, thereby facilitating a more efficient and accurate reevaluation of existing large-scale datasets. The model employs a two-branch architecture for feature processing. In one branch, the RNA-FM model is utilized to extract global features of m5U, which are subsequently processed through a FC layer to achieve feature dimension compression. Another sophisticated approach involves employing conventional one-hot encoding to extract sequence-related information, subsequently feeding it into a three-layer CNN for local feature mining, and ultimately utilizing a FC layer to carry out feature integration. The local features obtained from this branch are fused with the global features from the RNA-FM branch. The amalgamated features are then passed through the FC layers to perform binary classification for performance prediction. Through 10-fold cross-validation and independent testing, the m5U-HybridNet model exhibits outstanding performance in all evaluated criteria. The innovation lies in its integration of two key strategies: it not only employs the RNA-FM pretrained model to distill semantic insights from RNA sequences but also fuses the captured global features with local features derived from the CNN, forming a novel hybrid network architecture. This approach aligns with contemporary methods for modifying site prediction, significantly enhancing the model's resilience and adaptability ability particularly with its exceptional advantage in generalization standing out prominently as a hallmark of m5U-HybridNet. Despite the remarkable success of the m5U-HybridNet model in predicting m5U modification sites, the current framework is still subject to significant limitations. For instance, the model's proficiency may not generalize well to different types of modification sites or across diverse species.[46,47] Expanding the model's prediction scope to encompass multiple RNA methylation modification sites could substantially enhance its generalizability, robustness, and efficacy, thereby enabling it to better address complex biomedical challenges. Moreover, although the framework has achieved powerful computational performance, its translational applicability remains constrained due to the lack of experimental validation in biologically relevant systems. This limitation may impede its functional utility in real-world applications. Future work maybe focus on conducting more in-depth biological experiments and extending beyond mere prediction of modification sites to include quantitative analysis of these sites, thereby enhancing the model's biological relevance. Specifically, we aim to predict the degree and proportion of methylation modifications on RNA molecules, providing a more comprehensive understanding of the biological functions of these modifications and offering crucial references for experimental design and disease research. Additionally, future research can delve into the adaptive learning methods for RNA across different lifecycle stages and species, both in the context of full transcript and mature mRNA modes. By developing advanced adaptive learning algorithms, we hope to better capture the dynamic changes of RNA across diverse species and lifecycle stages, thereby offering a more comprehensive perspective for the study of RNA function and regulatory mechanisms. Through these future research directions, we anticipate providing more powerful tools and deeper insights for RNA-related research, propelling the advancement of this field.

■ ASSOCIATED CONTENT

Data Availability Statement

The data and source code of m5U-HybridNet can be obtained on the website (http://www.bioai-lab.com/m5U) or GitHub (https://github.com/xinyuli-Judy/m5U-hybridNet). The download links for the weights of RNA-FM and RNAErnie models are as follows: RNA-FM: https://huggingface.co/multimolecule/rnafm. RNAEnire: https://zenodo.org/records/10847621.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c01237.

Table S1: dimensionality changes in the FC layer; Table S2: comparison of different models and various model fine-tuning approaches; Table S3: padding at the head or tail with different methods; Table S4−S6: changes in input channels, kernel size, and number of modules of CNN; Table S7: comparison of different fusion methods; Table S8: main parameters; Table S9: different methods for handling imbalanced datasets; Table S10: the number of positive to negative samples in different cell types and techniques; Table S11: imbalanced dataset 1:100 ratio split; Figure S1: analyzing the interpretability of CNN branches; Figure S2: t-SNE visualization in the mature mRNA mode (PDF)

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096

AUTHOR INFORMATION

Corresponding Author

Feifei Cui − *School of Computer Science and Technology, Hainan University, Haikou 570228, China;* orcid.org/0000-0001-7055-3813; Email: feifeicui@hainanu.edu.cn

Authors

Xinyu Li − *School of Computer Science and Technology, Hainan University, Haikou 570228, China*

Zhenjie Luo − *School of Computer Science and Technology, Hainan University, Haikou 570228, China*

Jingwei Lv − *School of Computer Science and Technology, Hainan University, Haikou 570228, China*

Chao Yang − *School of Computer Science and Technology, Hainan University, Haikou 570228, China*

Shankai Yan − *School of Computer Science and Technology, Hainan University, Haikou 570228, China*

Junlin Xu − *School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, China*

Yajie Meng − *School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China*

Leyi Wei − *Centre for Artificial Intelligence driven Drug Discovery, Faculty of Applied Science, Macao Polytechnic University, Macao SAR 999078, China; School of Informatics, Xiamen University, Xiamen 361000, China*

Zilong Zhang − *School of Computer Science and Technology, Hainan University, Haikou 570228, China;* orcid.org/0000-0002-4934-1258

Quan Zou − *Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China;* orcid.org/0000-0001-6406-1142

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c01237

Author Contributions

X.L.: writing-original draft, designed and experiments. Z.L.: reproduce and review. J.L.: make website. C.Y.: writing-review and editing. S.Y.: writing-review and editing, resource. J.X.: writing-review and editing, resource, supervision. Y.M.: writing-review and editing, resource, supervision. L.W.: writing-review and editing, resource, supervision. Z.Z.: writing-review and editing, resource, supervision. Q.Z.: writing-review and editing, resource, supervision. F.C.: writing-review and editing, resource, supervision, funding acquisition.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The study received funding from the National Natural Science Foundation of China (No. 62450002).

REFERENCES

(1) Chen, Y.; Jiang, Z.; Yang, Y.; Zhang, C.; Liu, H.; Wan, J. The functions and mechanisms of post-translational modification in protein regulators of RNA methylation: Current status and future perspectives. *Int. J. Biol. Macromol.* **2023**, *253*, 126773.

(2) Khanal, J.; Tayara, H.; Zou, Q.; Chong, K. T. Identifying DNA N4-methylcytosine sites in the rosaceae genome with a deep learning model relying on distributed feature representation. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1612−1619.

(3) Bujnicki, J. M.; Feder, M.; Ayres, C. L.; Redman, K. L. Sequence−structure−function studies of tRNA: m5C methyltransferase Trm4p and its relationship to DNA: m5C and RNA: m5U methyltransferases. *Nucleic Acids Res.* **2004**, *32*, 2453−2463.

(4) Chou, K.-C. Progresses in predicting post-translational modification. *Int. J. Pept. Res. Ther.* **2020**, *26*, 873−888.

(5) Powell, C. A.; Minczuk, M. TRMT2B is responsible for both tRNA and rRNA m5U-methylation in human mitochondria. *RNA Biol.* **2020**, *17*, 451−462.

(6) Chang, Y.-H.; Nishimura, S.; Oishi, H.; Kelly, V. P.; Kuno, A.; Takahashi, S. TRMT2A is a novel cell cycle regulator that suppresses cell proliferation. *Biochem. Biophys. Res. Commun.* **2019**, *508*, 410−415.

(7) Ny, T.; Björk, G. Cloning and restriction mapping of the trmA gene coding for transfer ribonucleic acid (5-methyluridine)-methyltransferase in *Escherichia coli* K-12. *J. Bacteriol.* **1980**, *142*, 371−379.

(8) Urbonavičius, J.; Jäger, G.; Björk, G. R. Amino acid residues of the *Escherichia coli* tRNA (m5U54) methyltransferase (TrmA) critical for stability, covalent binding of tRNA and enzymatic activity. *Nucleic Acids Res.* **2007**, *35*, 3297−3305.

(9) Jonkhout, N.; Tran, J.; Smith, M. A.; Schonrock, N.; Mattick, J. S.; Novoa, E. M. The RNA modification landscape in human disease. *RNA* **2017**, *23*, 1754−1769.

(10) Guo, G.; Wang, H.; Shi, X.; Ye, L.; Yan, K.; Chen, Z.; Zhang, H.; Jin, Z.; Xue, X. Disease activity-associated alteration of mRNA m5 C methylation in CD4+ T cells of systemic lupus erythematosus. *Front. Cell Dev. Biol.* **2020**, *8*, 430.

(11) Carter, J.-M.; Emmett, W.; Mozos, I. R.; Kotter, A.; Helm, M.; Ule, J.; Hussain, S. FICC-Seq: a method for enzyme-specified profiling of methyl-5-uridine in cellular RNA. *Nucleic Acids Res.* **2019**, *47* (19), No. e113.

(12) Feng, P.; Chen, W. iRNA-m5U: a sequence based predictor for identifying 5-methyluridine modification sites in saccharomyces cerevisiae. *Methods* **2022**, *203*, 28−31.

(13) Xu, Z.; Wang, X.; Meng, J.; Zhang, L.; Song, B. m5U-GEPred: prediction of RNA 5-methyluridine sites based on sequence-derived and graph embedding features. *Front. Microbiol.* **2023**, *14*, 1277099.

(14) Yu, L.; Zhang, Y.; Xue, L.; Liu, F.; Jing, R.; Luo, J. Evaluation and development of deep neural networks for RNA 5-Methyluridine classifications using autoBioSeqpy. *Front. Microbiol.* **2023**, *14*, 1175925.

(15) Alam, W.; Tahir, M.; Hussain, S.; Gul, S.; Hayat, M.; Irshad, R. R.; Pallonetto, F. Unveiling the potential pattern representation of rna 5-methyluridine modification sites through a novel feature fusion model leveraging convolutional neural network and tetranucleotide composition. *IEEE Access* **2024**, *12*, 10023−10035.

(16) Jiang, J.; Song, B.; Tang, Y.; Chen, K.; Wei, Z.; Meng, J. m5UPred: a web server for the prediction of RNA 5-methyluridine sites from sequences. *Molecular Therapy. Nucleic Acids* **2020**, *22*, 742−747.

(17) Ao, C.; Ye, X.; Sakurai, T.; Zou, Q.; Yu, L. m5U-SVM: identification of RNA 5-methyluridine modification sites based on multi-view features of physicochemical features and distributed representation. *BMC Biol.* **2023**, *21* (1), 93.

(18) Chen, J.; Hu, Z.; Sun, S.; Tan, Q.; Wang, Y.; Yu, Q.; Zong, L.; Hong, L.; Xiao, J.; Shen, T. et al.Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv: 2204.00300*, **2022**.

(19) Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 6999−7019.

(20) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, **2019**.

(21) Jia, J.; Qin, L.; Lei, R. DGA-5mC: A 5-methylcytosine site prediction model based on an improved DenseNet and bidirectional GRU method. *Math. Biosci. Eng.* **2023**, *20* (6), 9759−9780.

(22) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv:1706.03762*, **2017**.
(23) Wang, N.; Bian, J.; Li, Y.; Li, X.; Mumtaz, S.; Kong, L.; Xiong, H. Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nat. Mach. Intell.* **2024**, *6*, 548−557.
(24) Li, G.; Jiang, F.; Zhu, J.; Cui, H.; Wang, Z.; Chen, W. HydraRNA: a hybrid architecture based full-length RNA language model. *bioRxiv*, **2025**.
(25) Liang, Y.; You, X.; Zhang, Z.; Qiu, S.; Li, S.; Fu, L. MGFmiRNAloc: Predicting miRNA Subcellular Localization Using Molecular Graph Feature and Convolutional Block Attention Module. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2024**, *21*, 1348−1357.
(26) Tan, M.; Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv:1905.11946*, **2019**.
(27) Fu, X.; Duan, H.; Zang, X.; Liu, C.; Li, X.; Zhang, Q.; Zhang, Z.; Zou, Q.; Cui, F. Hyb_SEnc: An Antituberculosis Peptide Predictor Based on a Hybrid Feature Vector and Stacked Ensemble Learning. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2024**, *21* (6), 1897−1910.
(28) Lv, J.; Geng, A.; Pan, Z.; Wei, L.; Zou, Q.; Zhang, Z.; Cui, F. iBitter-GRE: A Novel Stacked Bitter Peptide Predictor with ESM-2 and Multi-View Features. *J. Mol. Biol.* **2025**, *437*, 169005.
(29) Liang, Y.; Li, S.; You, X.; Guo, Y.; Tang, J. Stacking-Kcr: A Stacking Model for Predicting the Crotonylation Sites of Lysine by Fusing Serial and Automatic Encoder. *Curr. Bioinf.* **2023**, *19* (7), 674−686.
(30) Liang, Y.; Yin, X.; Zhang, Y.; Guo, Y.; Wang, Y. Predicting lncRNA−protein interactions through deep learning framework employing multiple features and random forest algorithm. *BMC Bioinf.* **2024**, *25* (1), 108.
(31) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv: 1412.6980*, **2014**.
(32) Elshamy, R.; Abu-Elnasr, O.; Elhoseny, M.; Elmougy, S. Improving the efficiency of RMSProp optimizer by utilizing Nestrove in deep learning. *Sci. Rep.* **2023**, *13* (1), 8814.
(33) Wang, Y.; Zhai, Y.; Ding, Y.; Zou, Q. SBSM-Pro: support bio-sequence machine for proteins. *Sci. China Inf. Sci.* **2024**, *67* (11), 212106.
(34) Qiu, S.; Liu, R.; Liang, Y. GR-m6A: Prediction of N6-methyladenosine sites in mammals with molecular graph and residual network. *Comput. Biol. Med.* **2023**, *163*, 107202.
(35) Wu, H.; Zhang, P.; Ai, Z.; Wei, L.; Zhang, H.; Yang, F.; Cui, L. StackTADB: a stacking-based ensemble learning model for predicting the boundaries of topologically associating domains (TADs) accurately in fruit flies. *Briefings Bioinf.* **2022**, *23* (2), bbac023.
(36) Jiao, S.; Ye, X.; Sakurai, T.; Zou, Q.; Liu, R. Integrated convolution and self-attention for improving peptide toxicity prediction. *Bioinformatics* **2024**, *40* (5), btae297.
(37) Ai, C.; Yang, H.; Liu, X.; Dong, R.; Ding, Y.; Guo, F. MTMol-GPT: De novo multi-target molecular generation with transformer-based generative adversarial imitation learning. *PLoS Comput. Biol.* **2024**, *20*, No. e1012229.
(38) Vacic, V.; Iakoucheva, L. M.; Radivojac, P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **2006**, *22*, 1536−1537.
(39) Dey, R.; Salem, F. M. Gate-variants of gated recurrent unit (GRU) neural networks. *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*; IEEE: Boston, MA, USA, 2017, 1597−1600.
(40) Feng, L.; Tung, F.; Hajimirsadeghi, H.; Ahmed, M. O.; Bengio, Y.; Mori, G. Attention as an RNN. *arXiv: 2405.13956*, **2024**.
(41) Graves, A.; Graves, A. Long short-term memory*Supervised Sequence Labelling With Recurrent Neural Networks* Springer, 2012, *385*, 37−45.
(42) Lu, Q.; Hao, P.; Curcin, V.; He, W.; Li, Y.-Y.; Luo, Q.-M.; Guo, Y.-K.; Li, Y.-X. KDE Bioscience: platform for bioinformatics analysis workflows. *J. Biomed. Inf.* **2006**, *39*, 440−450.
(43) Desrosiers, R.; Friderici, K.; Rottman, F. Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. *Proc. Int. Acad. Sci.* **1974**, *71*, 3971−3975.
(44) Wang, X.; Lu, Z.; Gomez, A.; Hon, G. C.; Yue, Y.; Han, D.; Fu, Y.; Parisien, M.; Dai, Q.; Jia, G.; et al. N 6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **2014**, *505* (7481), 117−120.
(45) Guzzi, N.; Cieśla, M.; Ngoc, P. C. T.; Lang, S.; Arora, S.; Dimitriou, M.; Pimková, K.; Sommarin, M. N. E.; Munita, R.; Lubas, M.; et al. Pseudouridylation of tRNA-derived fragments steers translational control in stem cells. *Cell* **2018**, *173* (5), 1204−1216. E26.
(46) Song, Z.; Huang, D.; Song, B.; Chen, K.; Song, Y.; Liu, G.; Su, J.; de Magalhães, J. P.; Rigden, D. J.; Meng, J. Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. *Nat. Commun.* **2021**, *12* (1), 4011.
(47) Huang, S.; Wylder, A. C.; Pan, T. Simultaneous nanopore profiling of mRNA m6A and pseudouridine reveals translation coordination. *Nat. Biotechnol.* **2024**, *42*, 1831−1835.

https://doi.org/10.1021/acs.jcim.5c01237
J. Chem. Inf. Model. 2025, 65, 8079−8096