Multi-task adaptive pooling enabled synergetic learning of RNA modification across tissue, type and species from low-resolution epitranscriptomes

Yiyou Song, Yue Wang, Xuan Wang, Daiyun Huang, Anh Nguyen and Jia Meng

Corresponding author: Yue Wang, Department of Mathematical Sciences, School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou 215123, PR China. E-mail: yue.wang@liverpool.ac.uk

Abstract

Post- and co-transcriptional RNA modifications are found to play various roles in regulating essential biological processes at all stages of RNA life. Precise identification of RNA modification sites is thus crucial for understanding the related molecular functions and specific regulatory circuitry. To date, a number of computational approaches have been developed for *in silico* identification of RNA modification sites; however, most of them require learning from base-resolution epitranscriptome datasets, which are generally scarce and available only for a limited number of experimental conditions, and predict only a single modification, even though there are multiple inter-related RNA modification types available. In this study, we proposed AdaptRM, a multi-task computational method for synergetic learning of multi-tissue, type and species RNA modifications from both high- and low-resolution epitranscriptome datasets. By taking advantage of adaptive pooling and multi-task learning, the newly proposed AdaptRM approach outperformed the state-of-the-art computational models (WeakRM and TS-m^{6}A-DL) and two other deep-learning architectures based on Transformer and ConvMixer in three different case studies for both high-resolution and low-resolution prediction tasks, demonstrating its effectiveness and generalization ability. In addition, by interpreting the learned models, we unveiled for the first time the potential association between different tissues in terms of epitranscriptome sequence patterns. AdaptRM is available as a user-friendly web server from http://www.rnamd.org/AdaptRM together with all the codes and data used in this project.

Keywords: low-resolution epitranscriptomes, multi-task learning, RNA modification, adaptive pooling, transformer, ConvMixer

INTRODUCTION

Post-transcriptional RNA modifications are found to play essential roles in epitranscriptome regulation for all types of RNAs and at all stages of RNA life [1–3]. Over 170 post-transcriptional modifications have been identified in living organisms [4], participating in various important biological processes such as fine-tuning RNA structures, regulation of gene expression and protein synthesis, response to environmental exposures, cell differentiation and mechanistic toxicology [5–10]. Recent studies suggest that RNA modifications have implications for human health and medical science [11–15]. To date, over 100 RNA modification enzyme mutations have been found to have an association with human diseases [16]. Abnormal presence or absence of RNA modifications can lead to human diseases, including various cancers, metabolic disorders and developmental illnesses [17]. It becomes clear that many exhilarating functions of RNA modifications remain to be explored.

Precise identification of modification sites serves as the basis for revealing their regulatory mechanisms and functions. The recent rapid development of high-throughput sequencing approaches has enabled the transcriptome-wide profiling of RNA modification sites [18]. MeRIP-seq (m^{6}A-seq) is the earliest and most widely used *in vitro* N6-methyladenosine (m^{6}A) identification method [19, 20]. It combined immunoprecipitation with next-generation sequencing: the fragmented RNAs containing signals of modification are isolated (immunoprecipitated) from total RNA by the m^{6}A-specific antibody and then purified for sequencing. The identified m^{6}A-containing regions (peaks) are around 100 nt long. On this basis, a refined m^{6}A MeRIP-seq (refined RIP-seq) has been developed, which requires a lower amount of input RNA samples and could be applied to analyze patient tumors [21]. Later, other modifications, such as 5-hydroxymethylcytidine (hm^{5}C), N4-acetylcytidine (ac^{4}C) and N7-methylguanosine (m^{7}G), could be profiled with similar antibody-based high-throughput

Yiyou Song received a Bachelor of Science degree from Xi'an Jiaotong-Liverpool. He is currently a PhD student at the Department of Computer Sciences, University of Liverpool. His research interests are bioinformatics and deep learning.

Yue Wang received a Bachelor of Science degree from Xi'an Jiaotong-Liverpool. She is currently a PhD student at the Department of Computer Sciences, University of Liverpool. Her research interests are bioinformatics, biostatistics and data mining.

Xuan Wang received a Bachelor of Science degree from Xi'an Jiaotong-Liverpool University. She is currently a Master student at the Department of Biological Sciences, Xi'an Jiaotong-Liverpool University. Her research interests are bioinformatics and databases.

Daiyun Huang received a PhD degree from the University of Liverpool. He is currently a research assistant in the Academy of Pharmacy, Xi'an Jiaotong-Liverpool University. His research is focused on deep learning, bioinformatics and computational biology.

Anh Nguyen is an assistant professor at the Department of Computer Science, University of Liverpool. His research is in the areas of medical imaging, medical robotics and deep learning.

Jia Meng is a professor at the Department of Biological Sciences, Xi'an Jiaotong-Liverpool University. His work focuses on epitranscriptome, bioinformatics and computational biology.

sequencing approaches hMeRIP-seq [22], acRIP-seq [23] and m^7G-MeRIP [24], respectively. Such immunoprecipitation-based sequencing methods generate low-resolution epitranscriptome data, i.e. peaks or sequences of varying lengths surrounding the true modification site, while the exact location of the sites remains undetermined.

As the field progresses, a series of high-resolution methods have been proposed, making it possible to identify modification sites at a single-nucleotide level in the genome, such as antibody-based methods miCLIP [25], m^6A-CLIP [26] and PA-m^6A-seq [27], and enzymatic methods MAZTER-Seq [28] and m^6A REF-seq [29]. However, although these methods can locate the specific position of modification sites, they still have several limitations. Corresponding wet-laboratory experiments usually require expensive costs, long experiment time and large amounts of input RNA materials, making modification detection unavailable to limited-quantity samples [30]. Another exciting technique, Oxford Nanopore Technologies, which allows direct RNA sequencing without prior amplification, is under development for the direct detection of RNA modifications but still faces significant challenges in interpreting raw signals and systematic errors [31–34]. As a consequence, although many advanced methods hold great promise for detecting RNA modification at single-base resolution, currently low-resolution methods are more often used than single-base methods, and the majority of available data is still of low resolution [35].

Computational methods are often considered an alternative avenue for epitranscriptome profiling, given the fact that wet-laboratory experiments are usually costly and labor intensive [3]. A variety of machine-learning or deep-learning approaches have been developed to predict putative RNA modification sites. SRAMP was one of the earliest predictors of mammalian m^6A sites from sequence-derived features enabled by random forests [36]. iRNA-m^6A was established to identify m^6A sites in different tissues via the support vector machine (SVM) [37]. BERMP achieved an area under the receiver operating characteristic curve (AUROC) of 0.817 for predicting m^6A in multiple species via integrating a bidirectional gated recurrent unit network with random forest [38]. Gene2vec adopted word2vec embedding for encoding m^6A sequences, combined with a convolution network and reached an AUROC of 0.843 [39]. DeepM6ASeq applied two layers of convolution and one layer of bidirectional long short-term memory to achieve an AUROC of 0.850 in m^6A prediction [40]. WHISTLE conducted an m^6A forecast utilizing the information of sequence and 35 genomic features, obtaining an AUROC of 0.98 on the full transcript and 0.904 on mRNA, which so far was among the highest accuracy obtained by the state-of-the-art methods [41]. Besides, although most prediction tools focused on m^6A modifications, it is worth noting that an increasing number of computational methods have also been applied to other modifications. For example, RAMPred [42], RNAm5CPred [43] and m7GHub [44] employed the SVM algorithm to predict m^1A, m^5C and m^7G sites, respectively. References [1, 45] provide a detailed overview of current tools for RNA modification prediction. Furthermore, multi-task learning (MTL) has been introduced for solving multiple-related problems in this field. MultiRM performed multitasking via attention-based multi-label neural networks for predicting different types of modifications [46].

It is a non-trivial task to pick suitable predictive strategies for learning datasets. However, all the methods mentioned above (and most of the current methods) are based only on high-resolution data, which is generally scarce and available for only very limited experimental conditions, and therefore have quite restricted applicable scopes. They cannot be directly applied to low-resolution data, which is the most abundant data type nowadays, since the datasets of low resolution are regions or peaks with varying sizes, and it is unclear the exact location of the true modification site within the region or peak. The varying length and the uncertain position of low-resolution data increase the difficulty of exploiting its informative features through learning. Thus, there is a great demand for computing methods to detect and analyze low-resolution epitranscriptome data. WeakRM is the first weakly supervised method of learning RNA modifications from low-resolution data [47]. It cuts the input sequence into several overlapping regions via sliding window and evaluates the probability of these regions containing modification of interest by a weakly supervised neural network. It achieved reliable performance on the single task of identifying hm^5C, ac^4C and m^7G low-resolution datasets. m6A-TSFinder followed exactly the weakly supervised method in WeakRM and conducted a tissue-specific m^6A prediction in 23 tissues [48]. However, none of them are applicable to both high- and low-resolution data.

In this study, we proposed AdaptRM, a multi-task deep-learning method for an integrated study of epitranscriptomes across tissues and modifications. It could operate on both base-resolution and high-resolution datasets without further preprocessing the input primary sequence. It was mainly enabled by adaptive pooling [49] and convolutional neural networks (CNNs) [50]. The adaptive pooling fed the most informative features to the downstream portion of CNNs, generated vectors of the same size without manually setting polling kernel or stride, and demonstrated a good generalization ability for processing sequences with varying lengths. Repeated standard convolutional blocks were implemented to exploit useful sequence features through learning. An MTL [51] was conducted using AdaptRM, allowing learning several tasks simultaneously so that each task could help all other tasks, effectively avoiding potential overfitting during model training.

We focused on three case studies, including (1) tissue-specific m^6A prediction from low-resolution dataset of human [52], (2) type-specific RNA modification prediction from low-resolution dataset of zebrafish [53] and (3) cross-species m^6A prediction from high-resolution m^6A dataset [54]. By formulating each of these three case studies as an integrated multi-tasking learning problem, we trained AdaptRM model and obtained impressive results. We compared it with four methods, including Transformer-Encoder [55], ConvMixer [56], WeakRM [47] and TS-m^6A-DL [54]. The Transformer and ConvMixer are recently proposed deep-learning models for neural machine translation and image classification. Since they can deal with texts or images of varying sizes, we revised and implemented these two advanced models to solve the problem stated in this study, which also needs the support for handling input objects with varying lengths. Furthermore, we compared AdaptRM with WeakRM on low-resolution datasets and TS-m^6A-DL on high-resolution datasets, following the data type they targeted. Both are state-of-the-art methods developed recently for RNA modification prediction. We demonstrated that, despite its simplicity, AdaptRM outperformed all these four competing methods in all three case studies, suggesting its effectiveness and excellent generalization capability.

In addition, we analyzed the correlation of weights between each task in the model to unveil the potential association between different modification types from single species and the association of single modification among different tissues. The general workflow of AdaptRM is shown in Figure 1. To facilitate access to our model, a user-friendly web server has also been developed and

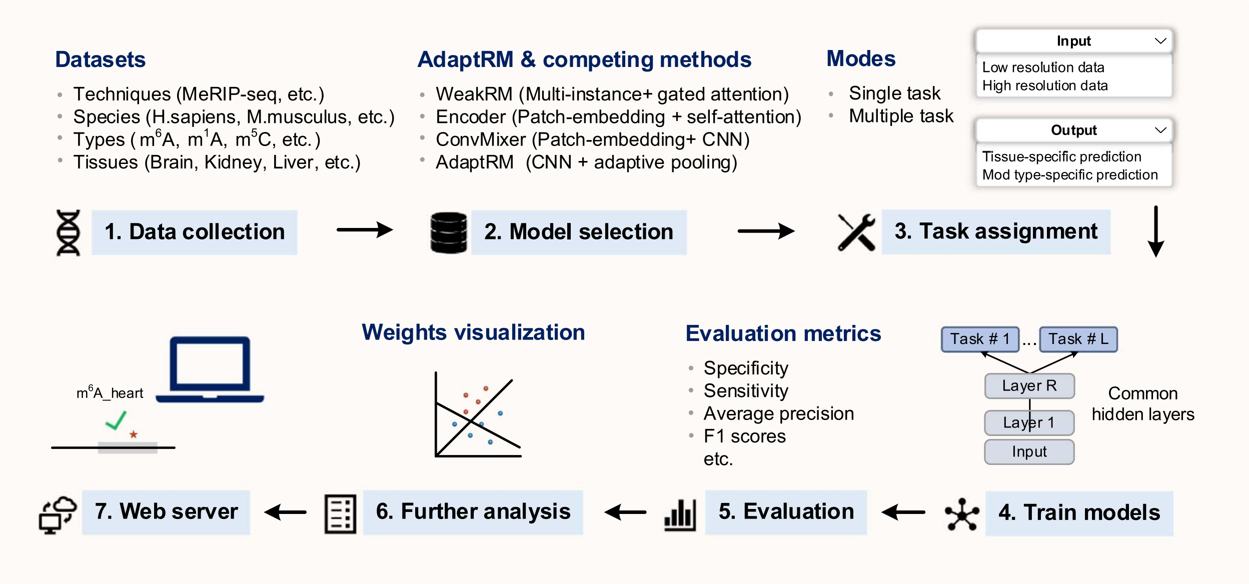

Datasets
• Techniques (MeRIP-seq, etc.)
• Species (H.sapiens, M.musculus, etc.)
• Types (m6A, m1A, m5C, etc.)
• Tissues (Brain, Kidney, Liver, etc.)
AdaptRM & competing methods
• WeakRM (Multi-instance+ gated attention)
• Encoder (Patch-embedding + self-attention)
• ConvMixer (Patch-embedding+ CNN)
• AdaptRM (CNN + adaptive pooling)
Modes
• Single task
• Multiple task
Input
Low resolution data
High resolution data
Output
Tissue-specific prediction
Mod type-specific prediction
1. Data collection
2. Model selection
3. Task assignment
m6A_heart
Weights visualization
Evaluation metrics
• Specificity
• Sensitivity
• Average precision
• F1 scores
etc.
Task # 1
...
Task # L
Layer R
Layer 1
Input
Common hidden layers
7. Web server
6. Further analysis
5. Evaluation
4. Train models

Figure 1. Workflow of developing AdaptRM. It entailed the following steps: (1) We collected RNA modifications derived from distinct techniques, which involved different types of modifications in various tissues and species. All were RNA primary sequences. (2) AdaptRM and other deep learning competing methods were implemented. (3) All the models were trained in both single-task and multi-task ways. (4) The MTL shared hidden layers between all the tasks. (5) The model performance was evaluated by different metrics. (6) The correlation of weights between each task in the model was visualized, unveiling the potential association between modifications. (7) A web server was developed to facilitate the use of our well-trained multi-task AdaptRM model.

made freely available at http://www.rnamd.org/AdaptRM. Our model is expected to be a useful tool for researchers of interest and provides insights into the computational study of both high- and low-resolution epitranscriptomes.

Table 1. Human tissue-specific m^{6}A data of low resolution

Human-m^{6}A	Total number of samples	Training	Testing
Adrenal gland	12 598	10 078	2520
Brainstem	18 176	14 540	3636
Cerebellum	3180	2544	636
Cerebrum	7472	5977	1495
Colon	14 722	11 777	2945
EndoC	18 732	14 985	3747
Endometrial	3148	2518	630
Heart	4794	3835	959
HSCs	11 054	8843	2211
Hypothalamus	18 640	14 912	3728
Islet	6706	5364	1342
Kidney	3930	3144	786
Liver	14 554	11 643	2911
Lung	6376	5100	1276
B-Lymphocyte	12 986	10 388	2598
Muscle	2876	2300	576
Ovary	4842	3873	969
Prostate	14 514	11 611	2903
Rectum	6388	5110	1278
RWPE-1	11 544	9235	2309
Skin	3978	3182	796
Stomach	4678	3742	936
Testis	13 056	10 444	2612
Thyroid gland	5496	4396	1100
Urinary bladder	3120	2496	624

HSCs = Hematopoietic stem cells.

METHODS

Benchmark dataset

The proposed AdaptRM framework was tested on three case studies, which were summarized in the following.

Table 1 shows a dataset of low-resolution m^{6}A-containing and non-modified RNA sequences in 25 human normal tissues detected using MeRIP-seq technology. The positive sequences contain m^{6}A sites, but the exact location of them is unknown. They were originally collected by Liu *et al.* [52]. We randomly picked the negative data from the non-peak regions on the same gene of the positive region. The negative samples were down-sampled and cut short to match the number and size of the positive samples. Sequences longer than 500 nt or shorter than 50 nt were removed to mitigate potential false-positive peaks caused by bioinformatics software. The whole dataset was merged, shuffled and split into training and testing datasets with a ratio of 8:2. This dataset is used for evaluating tissue-specific m^{6}A prediction from low-resolution epitranscriptome data.

Table 2 summarizes the RNA modification-containing and non-modified sequences of four different RNA modification types (m^{1}A, m^{5}C, m^{6}A and m^{7}G) in the zebrafish brain. The positive data were collected from a zebrafish methylation atlas [53] derived from the MeRIP-seq technique [19, 20, 37]. The negative data were generated in the same way previously described in Table 1. This dataset is used for testing cross-modification multi-task prediction from low-resolution epitranscriptome data.

Table 3 describes a cross-species cross-tissue m^{6}A dataset of base resolution. Its underlying technique is m^{6}A-REF-seq [29], a single-base antibody-independent sequencing method. Both positive and negative sequences are 41 nt long, containing an adenine in the middle corresponding to m^{6}A and non-m^{6}A in the positive and negative samples, respectively. This dataset is used to support cross-species cross-tissue m^{6}A prediction task from base-resolution epitranscriptome data.

Table 2. Low-resolution RNA modification data of multiple types in zebrafish brain

Modification	Total number of samples	Training	Testing
m^1A	11 557	9245	2312
m^5C	10 854	8683	2171
m^6A	7950	6360	1590
m^7G	1977	1581	396

Data and label encoding

One-hot encoding is one of the most prevalent encoding methods, which maps each input element into a vector. In this study, there are four types of nucleotides, i.e. A (adenine), C (cytosine), G (guanine) and U (uracil). Each nucleotide in sequences is assigned to a vector of 4 (A → [1,0,0,0], C → [0,1,0,0], G → [0,0,1,0], U → [0,0,0,1]).

The labels are represented by sparse encoding. We assigned each sequence for training a label vector of dimension T, where T is the number of assigned tasks. In this study, T should be 40 (25 tasks in the first case study +4 tasks in the second case +11 tasks in the third case). In the training step, a label vector consists of 1/0/−1, where 1 means a positive label to the target task and −1 means negative. 0 is assigned to irrelevant tasks for masked training and will be automatically exempt during cost calculation. Each label vector has only one element being 1 or −1 and all other elements are 0. Since input datasets vary in length, each training step only uses one sequence and its corresponding label vector.

Importantly, in the performance evaluation step, we only focused on the target tasks (ignored irrelevant tasks labeled as 0) and converted their corresponding labels from 1/−1 to 1/0 for the sake of comparing the predicted probabilities with them.

AdaptRM

Starting from the input layer, this sub-section presents a detailed description of the proposed AdaptRM framework (Figure 2). First, the input sequence of length *L* is converted to a matrix via one-hot encoding. Repeated standard convolutional blocks are implemented to exploit useful sequence features through learning. Each block contains a convolutional layer, a Dropout function and a PReLU activation function. The adaptive pooling in the middle reduces the spatial size of the features fed into the downstream portion of CNNs, generates vectors of the same size without manually setting the polling kernel or stride, and increases the generalization ability for processing sequences with varying lengths. Two convolution blocks before the adaptive pooling layer aim to extract local information, while the two convolution blocks after that aim to extract general information. A linear classifier is placed at the end, generating a vector of length T. Each element of it indicates the probability of each assigned task. The hyperparameter setting in AdaptRM is summarized in Table 4.

Multi-tasking learning [51] is applied during model training. MTL helps the model to generalize to multiple tasks simultaneously. The same hidden layers are shared between all tasks, meaning that the model is expected to find a general representation that captures the features from all tasks, reducing the risk of overfitting. Besides, this framework leads to implicit data augmentation. It allows the model to learn features that are easily found in one task and provides it to another task with a noisy pattern blurring this feature. Different tasks might share some features that are helpful to each other.

Pooling functions can play an important role in a model. It usually aims to capture the essential characteristics of input information, reduce the size of feature maps passed into the downstream neural networks and therefore increase model generalization ability. Most commonly used pooling functions include max pooling [57], average pooling [58], stochastic pooling [59] and variants of them. Adaptive pooling [49] is applied in AdaptRM, which is a variant of spatial pyramid pooling [60]. In such pooling, the output sizes are fixed no matter the length of its input layer. The stride and kernel sizes are automatically calculated to adapt to the output size. The adaptive pooling makes each spatial bin being processed proportional to the input vector size, therefore, maintaining the spatial information of the previous layer when capturing informative features.

The activation function PReLU [60], a generalized version of ReLU, is used in AdaptRM. PReLU has a learnable coefficient when the input element is less than 0, allowing different layers to have adjustable slopes in the negative part. A recent study [60] suggested in well-trained deep neural networks, the PReLU in earlier layers have larger positive slopes, while in deeper layers have smaller slopes, which means that the neural network in deeper layers tries to retain more information at earlier layers, demonstrating the effectiveness of PReLU during the training of deep-learning models.

For single-task training, binary cross-entropy is used as the cost function. For the multi-task training, a modified version of cross-entropy is utilized for masked training. It is calculated as follows:

$$\mathrm{Loss} = -\sum \left[\frac{\left(1 + y_{\mathrm{true}}\right) \times y_{\mathrm{true}}^2}{2} * \log y_{\mathrm{pred}} + \frac{\left(1 - y_{\mathrm{true}}\right) \times y_{\mathrm{true}}^2}{2} * \log\left(1 - y_{\mathrm{pred}}\right) \right], \quad (1)$$

where y_{true} represents the true label (one from 1,0 or −1) and y_{pred} represents the predictive probability. During the loss calculation, the y_{true}^2 term masks task with a label being 0, i.e. removing the task that the label is not given. The $\left(1 \pm y_{\mathrm{true}}\right)/2$ term maps the label from 1/−1 to the probability 1/0 so that the function is able to compare the predicted probability y_{pred} with its true label.

Competing methods

The competing methods include the Transformer-Encoder, ConvMixer, WeakRM (only for low resolution), TS-m^6A-DL and im^{6}A-TS-CNN (only for high resolution). All the methods were implemented with the Pytorch 1.12.

Transformer [61] is a Seq2seq model first proposed for neural machine translation and then has been successfully used in many NLP tasks [62–65]. It has an Encoder-Decoder architecture, enabled by multi-head attention mechanisms and feedforward neural networks. The Encoder reads a sequence vector and represents it with a high-dimensional vector, which is passed into the Decoder, generating another sequence in the target language. Taking advantage of the Encoder network of Transformer, the Vision Transformer [66] was proposed for image classification. It partitions the input image into several patches of the same size and adds an extra learnable token at the beginning of these patches for classification (CLS token). The input vector formed by sub-image patches and CLS token is then embedded by a neural network and fed into a stack of Encoders. The Encoder here is exactly the same concept as that in the Transformer model. The Encoder stack eventually generates a vector, where only the position of the CLS token is kept and utilized for image classification.

Table 3. Tissue-specific m^{6}A data of high resolution in multiple species

Species	Tissue	Total number of samples	Training	Testing
Human	Brain	18 418	9210	9208
	Kidney	18 294	9148	9146
	Liver	10 536	5268	5268
Mouse	Brain	32 100	16 050	16 050
	Heart	8802	4402	4400
	Kidney	15 810	7906	7904
	Liver	16 532	8266	8266
	Testis	18 826	9414	9412
Rat	Brain	9406	4704	4702
	Kidney	13 730	6866	6864
	Liver	7048	3524	3524

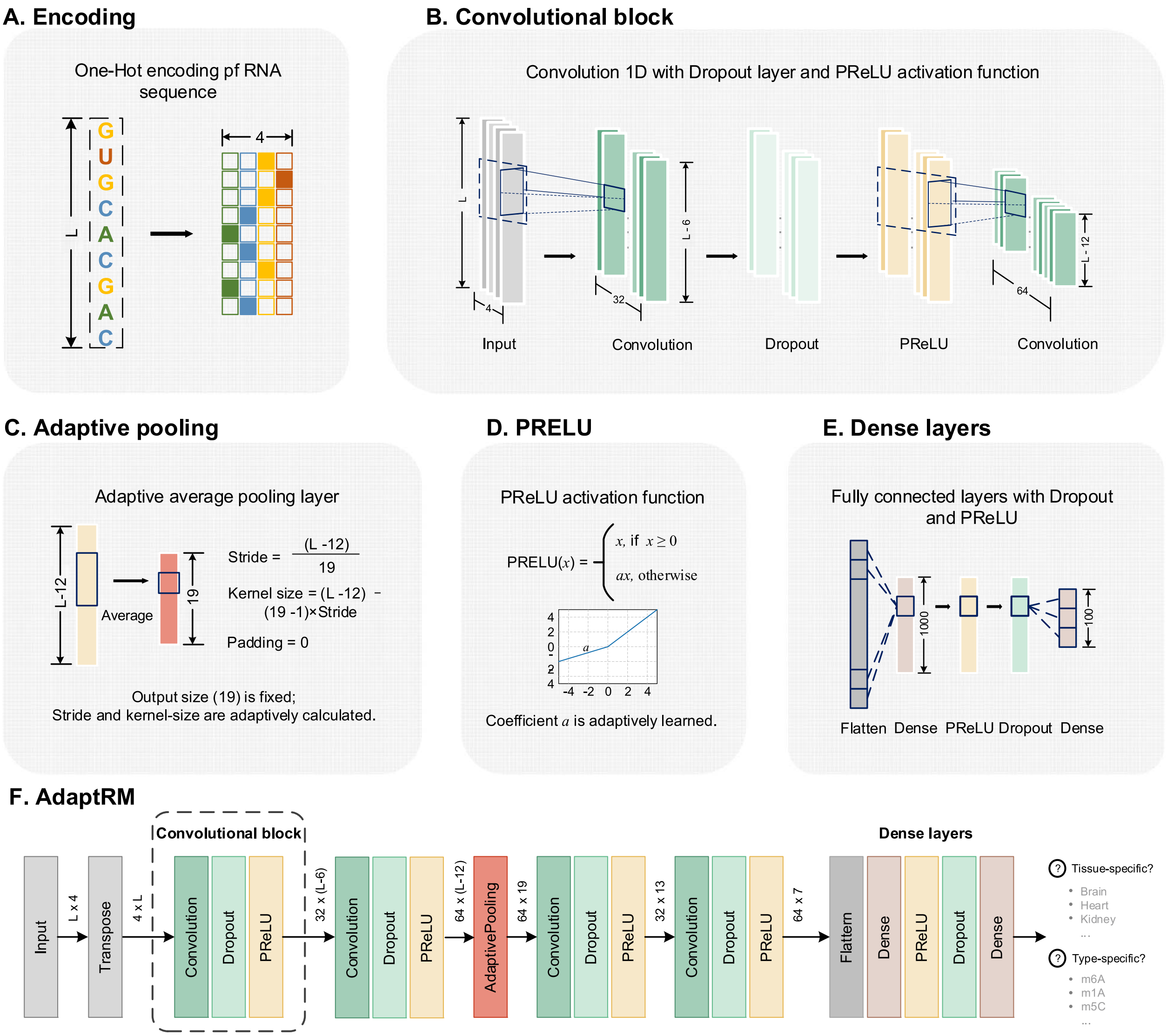


Figure 2. AdaptRM model architecture. (**A**) The input sequence is encoded by one-hot encoding. (**B**) A convolutional block in AdaptRM consists of a convolutional layer, dropout and activation function. (**C**) An adaptive pooling can produce vectors of fixed lengths. (**D**) The PReLU function introduces a learnable coefficient. (**E**) A linear classifier is placed at last. (**F**) An overview of AdaptRM. Taking a sequence as input, the output of AdaptRM is a vector, each element of which suggests the answer to the assigned task (true or false).

Based on the Encoder mechanisms proposed in the Transformer [61, 67], the CLS token and patch embedding structure utilized in Vision Transformer (ViT) [66], we developed a multi-task Encoder model for solving our problem stated previously. We chose the Transformer-Encoder as one of comparing methods for the following reasons. First, attention-based methods have achieved excellent performance in many fields, sometimes combining with or even replacing convolutional and

Table 4. Hyperparameter setting in AdaptRM

Layers	Settings	Output shape	Parameters
Input layer	One-hot encoding	(1, L, 4)	–
Conv1d_1	in_channels = 4, out_channels = 32, kernel_size = 7, stride = 1	(1, 32, L-6)	896
Conv1d_2	in_channels = 32, out_channels = 64, kernel_size = 7, stride = 1	(1, 64, L-12)	14 336
AdaptiveAvgPool1d	output_size = 19	(1, 64, 19)	64
Conv1d_3	in_channels = 64, out_channels = 32, kernel_size = 7, stride = 1	(1, 32, 13)	14 336
Conv1d_4	in_channels = 32, out_channels = 64, kernel_size = 7, stride = 1	(1, 64, 7)	14 336
Flatten	–	(1, 448)	–
Linear_1	in_features = 448, out_features = 1000, bias = True	(1, 1000)	448 000
Linear_2	in_features = 1000, out_features = 100, bias = True	(1, 100)	100 000
Sigmoid	–	(1, 100)	100
Dropout layers	P = [0.1, 0.2, 0.2, 0.2, 0.2]	–	1192
PReLU	num_parameters = 1	–	5
Adam	lr = 0.00005	–	–
Total	–	(1, T)	593 201

Note: *L is the length of the input sequence. T represents the number of assigned tasks. For a single-task model, T is just equal to 1.

recurrent neural networks [55]. Since most existing predictors for studying RNA modification are based on classic machine-learning algorithms, conventional CNNs or RNNs [1], we would like to explore the feature extraction ability of self-attention mechanisms on the low-resolution epitranscriptome datasets. Second, the patch embedding structure allows input with different sizes, which might be a good fit for low-resolution datasets with varying lengths. It partitions the input sequence into the same number of patches so that the subsequent layers can get a fixed number of embeddings. With patch embedding, the input sequences can be of any size. Please refer to supplementary materials for more details related to this model.

ConvMixer [56], a recently proposed neural network in the vision field, has a similar architecture as ViT. It consists of patch embedding and a stack of repeated convolutional blocks. Each block is formed by combining a depthwise convolutional layer (with a large kernel size) and a pointwise convolutional layer (with kernel 1). We chose ConvMixer as one of the competing methods because, similarly, the patch embedding structure allows input objects with varying sizes, and also, this method outperformed ViT and some of its variants, especially on small datasets. Mainly built upon the patch embedding and convolutional blocks stated in the ConvMixer, we developed a ConvMixer-based model to solve our multi-task problem. Please refer to supplementary materials for more details.

WeakRM [47] is the latest computational method identifying low-resolution RNA modifications. It formulated this problem as a multiple-instance learning (MIL)/weakly supervised task [68–72]. The MIL is a variation of supervised learning where one only knows the label of a ‘bag’ of instances, but the label of each instance is unclear. The task of MIL is to identify the label at the instance level. WeakRM adopted a gated-attention-based MIL [73]: it randomly cut the input sequence (bag) into several regions (instances), mapped these instances to an embedding by a convolutional neural network and finally calculated the instance-level features by a gated-attention mechanism. Since WeakRM is designed specially for peak calling data, we chose it as a competing method for case studies related to low-resolution datasets. We also extended it to multi-WeakRM to assess its performance on MTL. More details are provided in supplementary materials.

In addition, we compared our newly proposed models to some latest methods in terms of classifying high-resolution modification sites. Dao *et al.* [37] built a high-quality benchmark dataset of m^{6}A sites and utilized SVM to identify modifications in several tissues of humans, mice and rats. Later, trained on the same benchmark, CNN-enabled deep-learning methods im^{6}A-TS-CNN [74] and TS-m^{6}A-DL [54] were developed for a multi-tasking tissue-specific prediction. Both methods were constructed by a main convolutional body followed by a linear classifier. The difference is that TS-m^{6}A-DL concatenated the output of every convolutional block during training and fed them into the last linear classifier, rather than directly stacking these convolutional blocks. To further demonstrate generalization ability, AdaptRM was trained on this benchmark dataset and compared to im^{6}A-TS-CNN and TS-m^{6}A-DL.

RESULTS AND DISCUSSION

Case study 1: tissue-specific prediction from low-resolution data

To explore the model performance on tissue-specific datasets of low resolution, AdaptRM and three competing models were trained on the dataset summarized in Table 1. Two learning strategies, single-task learning and MTL, were implemented for each model. Figure 3 shows the distribution of AUROC results in each tissue task for these four deep-learning methods. Table 5 demonstrates the averaged results of 25 tissue tasks estimated by different metrics (the results of each specific tissue are shown in Tables S1–S4). The multi-task AdaptRM outperformed the competing classifiers and achieved the highest accuracy among all the tests. The performance evaluation demonstrated that MTL significantly improves model performance as we expected, leading to higher accuracy and better stability in all models. The results suggest that despite the existence of conditional-specific regulation of RNA methylation-related regulators, the methylation patterns from other tissues may still contribute significantly to the prediction in other tissues via the MTL framework.

Moreover, in order to explore the effectiveness of PReLU, we trained the above methods with ReLU and compared their performance to the training results with PReLU (Table S5). The results show that PReLU generally improved the accuracy of all methods. The degree of improvement might be related to the depth of neural networks. Since PReLU is mainly used to help the deeper layers retain more information from earlier layers, it has more evident effects on improving a deeper model. Therefore, it has significantly improved the Transformer-Encoder and ConvMixer, which contain stack of computational blocks (5 Encoder

Table 5. Model performance evaluation on tissue-specific m^{6}A prediction (low resolution)

Modes	Testing methods	Specificity	Sensitivity	Average precision	F1 scores	MCC	AUROC
Single task	WeakRM	0.7464	0.7096	0.7680	0.7342	0.4490	0.7964
	Encoder	0.7653	0.7082	0.7785	0.7438	0.4642	0.8047
	ConvMixer	0.7554	0.7140	0.7770	0.7431	0.4630	0.8032
	AdaptRM	0.7714	0.7477	0.8109	0.7622	0.5134	0.8340
Multi-task	WeakRM	0.8408	0.7300	0.8313	0.7922	0.5605	0.8554
	Encoder	0.8338	0.7595	0.8411	0.8022	0.5880	0.8676
	ConvMixer	0.8986	0.7586	0.8899	0.8301	0.6427	0.9027
	AdaptRM	0.9014	0.7634	0.8935	0.8336	0.6501	0.9060

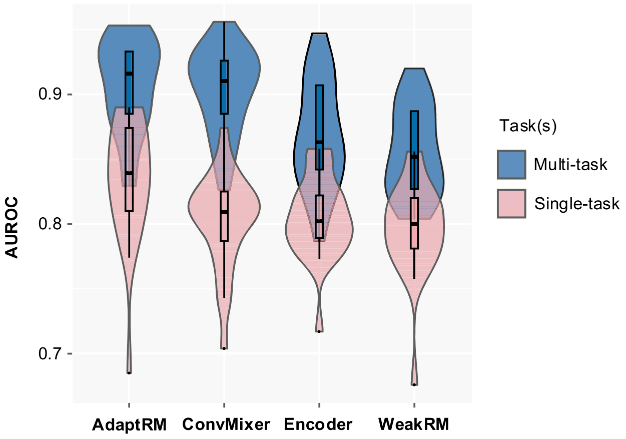


Figure 3. The AUROC of model performance on human tissue-specific datasets. AdaptRM and competing methods were evaluated on the refined RIP-seq datasets (low resolution) [52]. Multi-task versions of models presented better performance than the single-task models did. The AdaptRM trained in a multi-task way demonstrated the most accurate and stable performance among all the methods.

blocks and 10 convolutional blocks, respectively), but only slightly enhanced AdaptRM and WeakRM, which are relatively shallow models. Nonetheless, AdaptRM with PReLU performed better than all other methods in both modes.

Besides, since the number of samples in each task is quite diverse, we trained the multi-task version of all the above methods with a weighted cross entropy strategy. The weight assigned to the t-th task is shown as the following equation:

$$\text{weight}_t = \frac{\sum_{i=1}^{T} n_i}{T * n_t}, \qquad (2)$$

where T is the total number of tasks and n_t denotes the number of samples in the tth task. The weight is designed to be the inverse ratio of the sample size in target task times the average sample size of different tasks, which should be a value inversely proportional to the task sample number and close to 1. We expected the weighted cost function to eliminate the potential effects caused by the varying data sizes in different tasks. However, calculating the cost in this way slightly lowers the prediction accuracy compared to the binary cross-entropy mentioned before. The results are shown in Table S6. Therefore, we considered that directly summing the losses is a better strategy in this study.

Case study 2: modification-specific prediction from low-resolution data

To explore the ability to identify different RNA modifications from low-resolution epitranscriptomes, AdaptRM and competing models were trained on a zebrafish atlas dataset summarized in Table 2. There is no need to involve comparison to the Transformer-Encoder method here since it is declared to prefer adequately large-scale datasets [66], and it did not beat AdaptRM and ConvMixer in case study 1, not to say in this case with smaller datasets in case study 2. Table 6 summarizes the average AUROC scores of different well-trained models performing on the testing data. AdaptRM also outperformed both ConvMixer and WeakRM on the prediction of each type of modification. For all models, multi-task versions performed better than single-task versions, suggesting that the sequence patterns may be partially shared among different RNA modification types, and the predictive features identified for other modification types are useful as well via the MTL framework.

Case study 3: cross-species tissue-specific prediction on high-resolution data

We show in this example that, besides learning from low-resolution data, AdaptRM also has an impressive ability to learn from high-resolution datasets. AdaptRM was compared to ConvMixer and the state-of-the-art methods [46, 54, 74] for cross-species high-resolution modification prediction. WeakRM is not involved here since it is only for low-resolution epitranscriptomes. Table 7 and Table S7 present the AUROC scores obtained by different models trained on the m^{6}A-REF-seq datasets shown in Table 3. AdaptRM achieved better results than the tested competing methods. ConvMixer obtained similar AUROC as TS-m^{6}A-DL on the testing dataset but lower on the validation dataset. Estimated results demonstrated good generalization ability of AdaptRM and great power to handle multiple tasks simultaneously.

Potential similarity between tissues and modifications

In both tissue-specific and modification-specific predictions (case 1 and case 2), the multi-task strategy significantly improved the performance of all the models, suggesting putative links between these tasks. To gain insights into the potential association between modifications of distinct types or in different tissues, we extracted the correlation of weights between each task in the well-trained multi-task AdaptRM.

Specifically, we drew a heat map (Figure 4) to visualize potential relationship of m^{6}A methylation pattern among different human tissues, based on case study 1. It calculated coefficients between

Table 6. Model performance evaluation on modification-specific prediction (low resolution)

Tasks	Single task			Multi-task		
	WeakRM	ConvMixer	AdaptRM	WeakRM	ConvMixer	AdaptRM
m^1A	0.960	0.968	0.985	0.964	0.981	0.992
m^5C	0.948	0.974	0.979	0.959	0.983	0.989
m^6A	0.962	0.963	0.975	0.964	0.983	0.990
m^7G	0.923	0.914	0.946	0.946	0.960	0.978
Average	0.948	0.955	0.971	0.958	0.976	0.987

Table 7. Model performance evaluation on tissue-specific m^6A prediction (high resolution)

Multi-tasks		AdaptRM		ConvMixer		im^6A-TS-CNN		TS-m^6A-DL	
		Valid	Test	Valid	Test	Valid	Test	Valid	Test
Human	Brain	0.820	0.826	0.796	0.808	0.803	0.806	0.826	0.810
	Kidney	0.895	0.890	0.878	0.882	0.878	0.873	0.890	0.880
	Liver	0.893	0.899	0.889	0.891	0.881	0.881	0.914	0.878
Mouse	Brain	0.886	0.882	0.877	0.871	0.871	0.872	0.883	0.873
	Heart	0.854	0.847	0.846	0.827	0.812	0.816	0.850	0.823
	Kidney	0.915	0.901	0.899	0.891	0.884	0.886	0.908	0.889
	Liver	0.829	0.816	0.795	0.797	0.795	0.793	0.829	0.791
	Testis	0.874	0.856	0.836	0.843	0.838	0.847	0.863	0.843
Rat	Brain	0.876	0.878	0.844	0.869	0.847	0.852	0.876	0.854
	Kidney	0.916	0.918	0.905	0.911	0.902	0.908	0.907	0.908
	Liver	0.922	0.904	0.895	0.895	0.883	0.885	0.903	0.885
Average		0.880	0.874	0.860	0.862	0.854	0.856	0.877	0.858

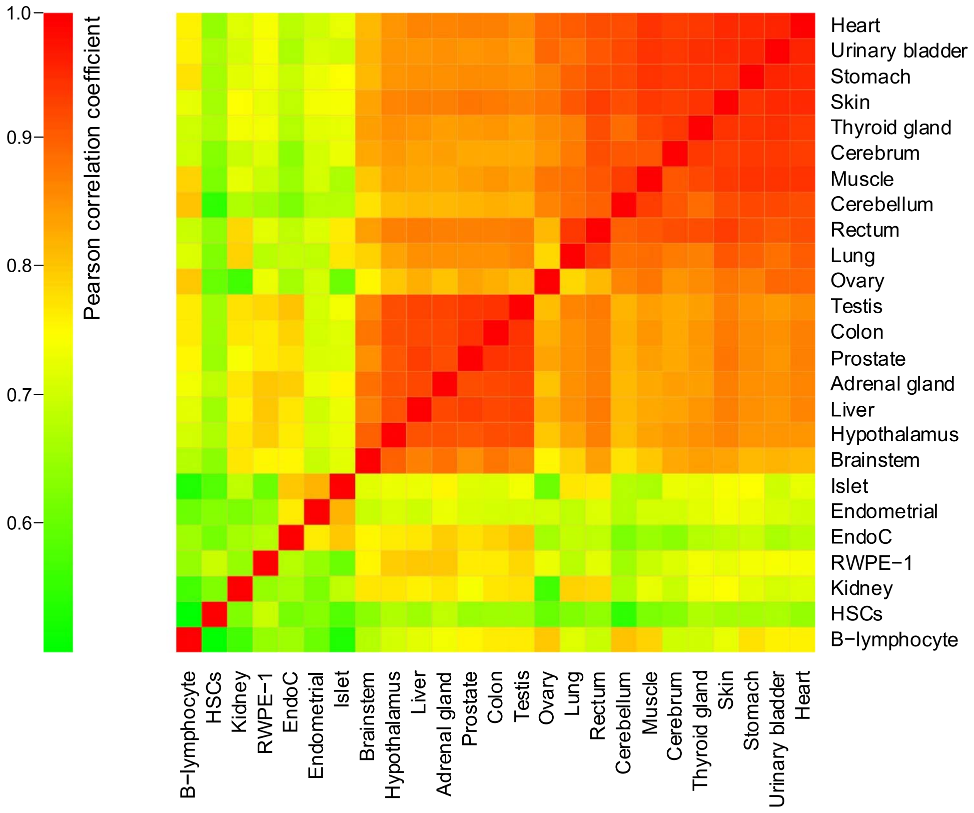


Figure 4. Association among different tissues in terms of m^6A sequence patterns.

weight vectors in each pair of tasks (tissue-specific predictions). A higher score represents higher similarity. The weight vectors were extracted from the second to last layer of the classifier. The diagonal element should be 1 since the correlation coefficient of a task and itself must be 1. m^6A modification in the heart, bladder and stomach are highly similar, while that in the ovary and kidney

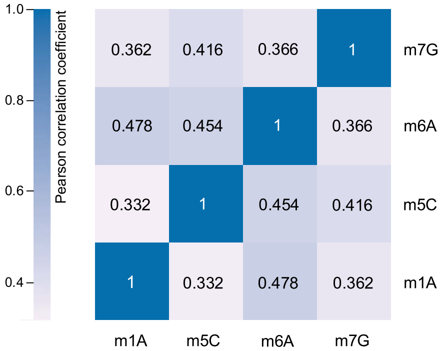


Figure 5. Association among different modification types in terms of sequence patterns.

differs. Overall, the model suggests that m^{6}A modification patterns over human tissues can be separated into two large groups and several outliers. The first group would be from the heart to the ovary, containing 11 tissues. The second group would be from the testis to the brainstem, containing seven tissues. Tissue/task pairs that share a low correlation coefficient use different information for prediction, which might indicate that modification between these tissues occurs in different conditions, but biological tests still need to be done to confirm it.

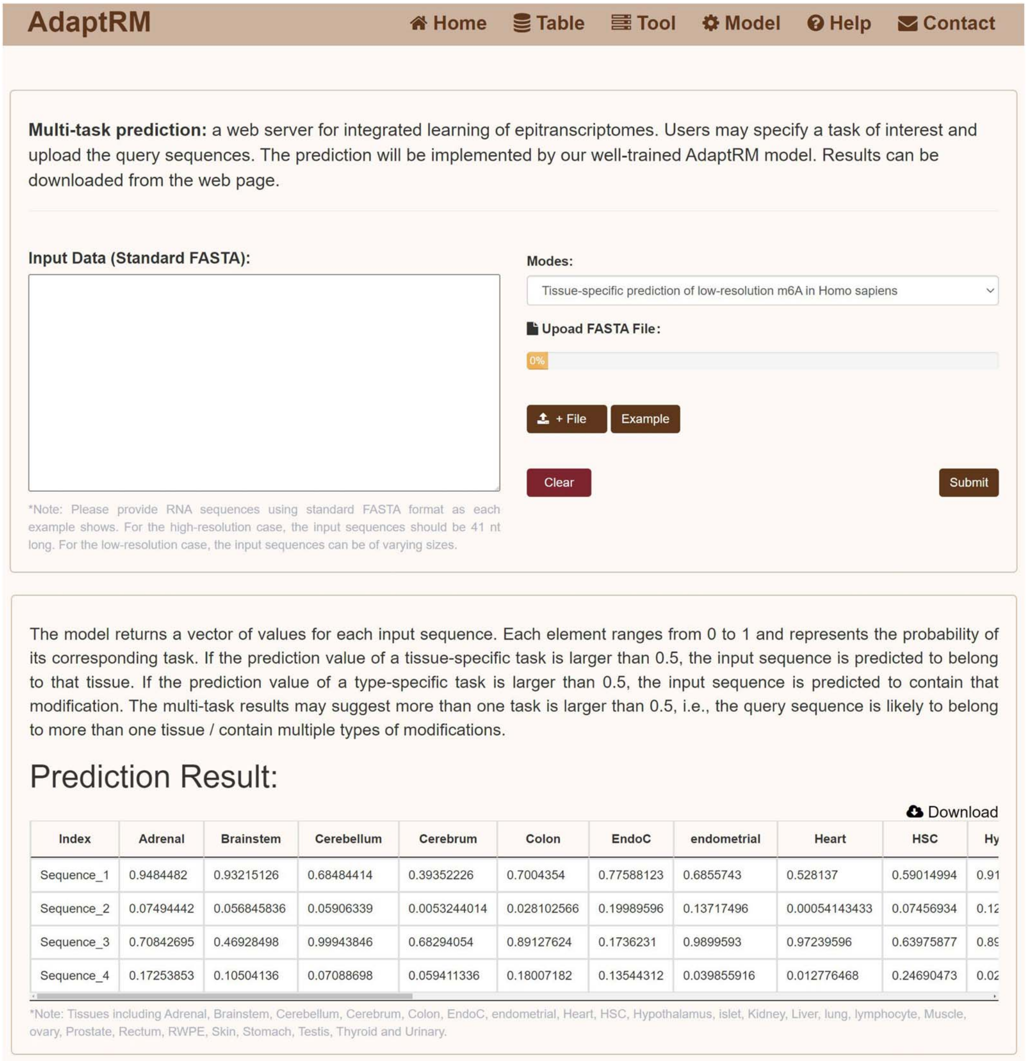
In addition, we visualized the potential relation between different modifications in zebrafish based on case study 2 in Figure 5. Interestingly, the strongest association was observed between m^{1}A and m^{6}A, which are both modifications to adenosine.

Web server

To facilitate access to our model, a user-friendly web server has been developed (Figure 6). It takes FASTA sequences as input. The length of input sequences should range from 40 to 500 nt. Users need to specify a task of interest first, paste the sequences into the text box or upload a FASTA file, and click the button to conduct RNA modification prediction automatically using multi-task AdaptRM. Results will be presented after a while.

CONCLUSION

Recent high-throughput sequencing techniques advanced the epitranscriptome study, and post- and co-transcriptional RNA modifications are found to play important roles in the regulation on all types of RNAs. Many computational approaches have been developed for predicting RNA modification sites, but most



Multi-task prediction: a web server for integrated learning of epitranscriptomes. Users may specify a task of interest and upload the query sequences. The prediction will be implemented by our well-trained AdaptRM model. Results can be downloaded from the web page.

Input Data (Standard FASTA):

Modes: Tissue-specific prediction of low-resolution m6A in Homo sapiens

*Note: Please provide RNA sequences using standard FASTA format as each example shows. For the high-resolution case, the input sequences should be 41 nt long. For the low-resolution case, the input sequences can be of varying sizes.

The model returns a vector of values for each input sequence. Each element ranges from 0 to 1 and represents the probability of its corresponding task. If the prediction value of a tissue-specific task is larger than 0.5, the input sequence is predicted to belong to that tissue. If the prediction value of a type-specific task is larger than 0.5, the input sequence is predicted to contain that modification. The multi-task results may suggest more than one task is larger than 0.5, i.e., the query sequence is likely to belong to more than one tissue / contain multiple types of modifications.

Prediction Result:

Index	Adrenal	Brainstem	Cerebellum	Cerebrum	Colon	EndoC	endometrial	Heart	HSC	Hy
Sequence_1	0.9484482	0.93215126	0.68484414	0.39352226	0.7004354	0.77588123	0.6855743	0.528137	0.59014994	0.91
Sequence_2	0.07494442	0.056845836	0.05906339	0.0053244014	0.028102566	0.19989596	0.13717496	0.00054143433	0.07456934	0.12
Sequence_3	0.70842695	0.46928498	0.99943846	0.68294054	0.89127624	0.1736231	0.9899593	0.97239596	0.63975877	0.8
Sequence_4	0.17253853	0.10504136	0.07088698	0.059411336	0.18007182	0.13544312	0.039855916	0.012776468	0.24690473	0.02

*Note: Tissues including Adrenal, Brainstem, Cerebellum, Cerebrum, Colon, EndoC, endometrial, Heart, HSC, Hypothalamus, islet, Kidney, Liver, lung, lymphocyte, Muscle, ovary, Prostate, Rectum, RWPE, Skin, Stomach, Testis, Thyroid and Urinary.

Figure 6. Screenshot of the AdaptRM web server. Users can specify a task of interest and upload the query sequences via our website. Results and models can be downloaded from the web page.

of them are computed only on high-resolution data. Although a series of advanced sequencing methods were proposed to generate high-resolution data, low-resolution methods are still more often used than single-base resolution methods because they are simpler, less expensive and have fewer requirements for the input RNA samples. Therefore, an integrated computational study for the peak-calling/low-resolution data is urgently needed.

In this study, we conducted MTL of low-resolution epitranscriptomes across tissues, types and species. We performed in-depth research for the model selection and utilized three novel methods, including AdaptRM (CNN + adaptive pooling), Transformer-Encoder (patch embedding + multi-head self-attention) and ConvMixer (patch embedding + CNN + adaptive pooling). The proposed methods were further compared to state-of-the-art approaches WeakRM (multiple-instance learning + CNN + gated-attention) on low-resolution datasets and TS-m^{6}A-DL (CNN) on high-resolution datasets.

We found that AdaptRM outperformed other competing methods tested in all the three case studies, including tissue-specific m^{6}A prediction in human (low resolution), modification-specific prediction in zebrafish (low resolution) and cross-species tissue-specific m^{6}A prediction (high resolution), demonstrating its effectiveness and excellent generalization ability. ConvMixer performed slightly worse than AdaptRM but beat the Transformer-Encoder and other methods. The results are consistent with the empirical conclusions summarized in recent studies that ConvMixer performed better than the Transformer-Encoder on small datasets, and the latter prefers adequately large-scale datasets.

The success of AdaptRM may primarily result from the application of adaptive pooling. For the peak calling data, the sequence itself varies in length, but classical machine-learning classifiers cannot handle inputs with unknown sizes. CNNs can handle inputs with unknown sizes but will return outputs with varying lengths. In adaptive pooling, the output size is fixed no matter the length of its previous input vector. The stride and kernel sizes are automatically calculated to adapt to the output setting. Even if the input sequences are of varying lengths, each spatial bin being focused is proportional to the input sequence size, allowing the pooling operator to extract informative features without spoiling the spatial information of the previous layer.

In addition, we found that MTL significantly improved the performance of all the models, suggesting clear links among various tasks, i.e. sequence patterns among different tissues and modification types. We visualized the correlation of weights between each task in multi-task AdaptRM, unveiling the potential association between modifications of distinct types or in different tissues, although such conclusions remain to be tested and verified by further experiments in the future.

Finally, a user-friendly web server was developed to facilitate access to our model. It is clear that our AdaptRM method can be easily extended to include more tissues, more modification types and more species for effectively learning from both low-resolution and high-resolution epitranscriptomes.

Key Points

- In this study, we proposed AdaptRM, a multi-task computational method for synergetic learning of multi-tissue, type and species RNA modifications from both high- and low-resolution epitranscriptome datasets.
- We showed that AdaptRM outperformed the state-of-the-art computational models (WeakRM and TS-m^{6}A-DL) and two other novel deep-learning architectures based on Transformer and ConvMixer in three different case studies for both high-resolution and low-resolution prediction tasks.
- We further interpreted the learned models and unveiled for the first time the potential association between different tissues in terms of epitranscriptome sequence patterns.
- To facilitate access to our model, a user-friendly web server has also been developed and made freely available at http://www.rnamd.org/AdaptRM.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

FUNDING

National Natural Science Foundation of China (32100519 and 31671373); XJTLU Key Program Special Fund (KSF-E-51 and KSF-P-02).

DATA AVAILABILITY

The AdaptRM framework was implemented with Pytorch 1.12, and the codes are available on GitHub: https://github.com/yiyousong/AdaptRM. The AdaptRM web server together with the completed datasets and the trained models can be freely accessed at http://www.rnamd.org/AdaptRM.

References

1. Liu L, Song B, Ma J, *et al.* Bioinformatics approaches for deciphering the epitranscriptome: recent progress and emerging topics. *Comput Struct Biotechnol J* 2020;**18**:1587–604.
2. McCown PJ, Ruszkowska A, Kunkler CN, *et al. Naturally occurring modified ribonucleosides.* WIREs. *RNA* 2020;**11**(5):e1595.
3. Jones JD, Monroe J, Koutmou KS. A molecular-level perspective on the frequency, distribution, and consequences of messenger RNA modifications. *WIREs RNA* 2020;**11**(4):e1586.
4. Boccaletto P, Stefaniak F, Ray A, *et al.* MODOMICS: a database of RNA modification pathways. 2021 update. *Nucleic Acids Res* 2022;**50**(D1):D231–5.
5. Batista PJ, Molinie B, Wang J, *et al.* m6A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell* 2014;**15**(6):707–19.
6. Delaunay S, Frye M. RNA modifications regulating cell fate in cancer. *Nat Cell Biol* 2019;**21**(5):552–9.
7. Yang C. ToxPoint: dissecting functional RNA modifications in responses to environmental exposure—mechanistic toxicology research enters a new era. *Toxicol Sci* 2020;**174**(1):1–2.
8. Pendleton KE, Chen B, Liu K, *et al.* The U6 snRNA m6A methyltransferase METTL16 regulates SAM synthetase intron retention. *Cell* 2017;**169**(5):824–835.e14.
9. Liu N, Dai Q, Zheng G, *et al.* N6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature* 2015;**518**(7540):560–4.

10. Geula S, Moshitch-Moshkovitz S, Dominissini D, *et al.* m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science* 2015;**347**(6225): 1002–6.
11. Esteve-Puig R, Bueno-Costa A, Esteller M. Writers, readers and erasers of RNA modifications in cancer. *Cancer Lett* 2020;**474**: 127–37.
12. Shulman Z, Stern-Ginossar N. The RNA modification N6-methyladenosine as a novel regulator of the immune system. *Nat Immunol* 2020;**21**:501–12.
13. Zaccara S, Ries RJ, Jaffrey SR. Reading, writing and erasing mRNA methylation. *Nat Rev Mol Cell Biol* 2019;**20**:608–24.
14. Fu Y, Dominissini D, Rechavi G, *et al.* Gene expression regulation mediated through reversible m6A RNA methylation. *Nat Rev Genet* 2014;**15**(5):293–306.
15. Zhang C, Fu J, Zhou Y. A review in research progress concerning m6A methylation and immunoregulation. *Front Immunol* 2019;**10**:922.
16. Jonkhout N, Tran J, Smith MA, *et al.* The RNA modification landscape in human disease. *RNA* 2017;**23**(1469–9001 (Electronic)): 1754–69.
17. McCown PJ, Ruszkowska A, Kunkler CN, *et al.* Naturally occurring modified ribonucleosides. *Wiley Interdiscip Rev RNA* 2020;**11**(5):e1595.
18. Li X, Xiong X, Yi C. Epitranscriptome sequencing technologies: decoding RNA modifications. *Nat Methods* 2017;**14**: 23–31.
19. Meyer KD, Saletore Y, Zumbo P, *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* 2012;**149**(7):1635–46.
20. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, *et al.* Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 2012;**485**(7397):201–6.
21. Zeng Y, Wang S, Gao S, *et al.* Refined RIP-seq protocol for epitranscriptome analysis with low input materials. *PLoS Biol* 2018;**16**(9):e2006092.
22. Delatte B, Wang F, Ngoc LV, *et al.* RNA biochemistry. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science* 2016;**351**(6270):282–5.
23. Arango D, Sturgill D, Alhusaini N, *et al.* Acetylation of cytidine in mRNA promotes translation efficiency. *Cell* 2018;**175**(7): 1872–1886.e24.
24. Zhang L-S, Liu C, Ma H, *et al.* Transcriptome-wide mapping of internal N7-methylguanosine methylome in mammalian mRNA. *Mol Cell* 2019;**74**(6):1304–1316.e8.
25. Linder B, Grozhik AV, Olarerin-George AO, *et al.* Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods* 2015;**12**(8):767–72.
26. Ke S, Alemu EA, Mertens C, *et al.* A majority of m6A residues are in the last exons, allowing the potential for 3′ UTR regulation. *Genes Dev* 2015;**29**(19):2037–53.
27. Chen K, Lu Z, Wang X, *et al.* High-resolution N(6)-methyladenosine (m(6) a) map using photo-crosslinking-assisted m(6) a sequencing. *Angew Chem Int Ed Engl* 2015;**54**(5): 1587–90.
28. Garcia-Campos MA, Edelheit S, Toth U, *et al.* Deciphering the “m(6)a code” via antibody-independent quantitative profiling. *Cell* 2019;**178**(3):731–47.
29. Zhang Z, Chen L-Q, Zhao Y-L, *et al.* Single-base mapping of m6A by an antibody-independent method. *Sci Adv* 2019;**5**(7):eaax0250.
30. Meyer KD. DART-seq: an antibody-free method for global m(6)a detection. *Nat Methods* 2019;**16**(12):1275–80.
31. Garalde D, Snell E, Jachimowicz D. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* 2018;**15**: 201–6.
32. Thomas NK, Poodari VC, Jain M, *et al.* Direct nanopore sequencing of individual full length tRNA strands. *ACS Nano* 2021;**15**(10): 16642–53.
33. Anreiter I, Mir Q, Simpson JT, *et al.* New twists in detecting mRNA modification dynamics. *Trends Biotechnol* 2021;**39**(1):72–89.
34. Liu H, Begik O, Lucas MC. Accurate detection of m6A RNA modifications in native RNA sequences. *Nat Commun* 2019;**10**:4079.
35. McIntyre ABR, Gokhale NS, Cerchietti L, *et al.* Limits in the detection of m(6)a changes using MeRIP/m(6)A-seq. *Sci Rep* 2020;**10**(1):6590.
36. Zhou Y, Zeng P, Li YH, *et al.* SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res* 2016;**44**(10):e91.
37. Dao FY, Lv H, Yang YH, *et al.* Computational identification of N6-methyladenosine sites in multiple tissues of mammals. *Comput Struct Biotechnol J* 2020;**18**:1084–91.
38. Huang Y, He N, Chen Y, *et al.* BERMP: a cross-species classifier for predicting m6A sites by integrating a deep learning algorithm and a random forest approach. *Int J Biol Sci* 2018;**14**(12): 1669–77.
39. Zou Q, Xing P, Wei L, *et al.* Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA* 2019;**25**:205–18.
40. Zhang Y, Hamada M. DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning. *BMC Bioinformatics* 2018;**19**(19):524.
41. Chen K, Wei Z, Zhang Q, *et al.* WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res* 2019;**47**(7):e41.
42. Chen W, Feng P, Tang H, *et al.* RAMPred: identifying the N1-methyladenosine sites in eukaryotic transcriptomes. *Sci Rep* 2016;**6**(1):31080.
43. Fang T, Zhang Z, Sun R, *et al.* RNAm5CPred: prediction of RNA 5-methylcytosine sites based on three different kinds of nucleotide composition. *Molecular Therapy - Nucleic Acids* 2019;**18**: 739–47.
44. Song B, Tang Y, Chen K, *et al.* m7GHub: deciphering the location, regulation and pathogenesis of internal mRNA N7-methylguanosine (m7G) sites in human. *Bioinformatics* 2020;**36**(11):3528–36.
45. El Allali A, Elhamraoui Z, Daoud R. Machine learning applications in RNA modification sites prediction. *Comput Struct Biotechnol J* 2021;**19**:5510–24.
46. Song Z, Huang D, Song B, *et al.* Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. *Nat Commun* 2021;**12**(1):4011.
47. Huang D, Song B, Wei J, *et al.* Weakly supervised learning of RNA modifications from low-resolution epitranscriptome data. *Bioinformatics* 2021;**37**:i222–30.
48. Song B, Huang D, Zhang, *et al.* m6A-TSHub: unveiling the context-specific m6A methylation and m6A-affecting mutations in 23 human tissues. *Genomics Proteomics Bioinformatics* 2022.
49. Abdu-Aguye M. G., Gomaa W., Makihara Y., *et al. Adaptive pooling is all you need: an empirical study on hyperparameter-insensitive human action recognition using wearable sensors*. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. Glasgow, UK: IEEE Computer Society, Piscataway, 2020.

50. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**(7553):436–44.
51. Ruder S. *An Overview of Multi-Task Learning in Deep Neural Networks*, 2017, CoRR:bs/1706.05098.
52. Liu J, Li K, Cai J, *et al.* Landscape and regulation of m(6)a and m(6)am methylome across human and mouse tissues. *Mol Cell* 2020;**77**(2):426–440 e6.
53. Li W, Li X, Ma X, *et al.* Mapping the m1A, m5C, m6A and m7G methylation atlas in zebrafish brain under hypoxic conditions by MeRIP-seq. *BMC Genomics* 2022;**23**(1):105.
54. Abbas Z, Tayara H, Zou Q, *et al.* TS-m6A-DL: tissue-specific identification of N6-methyladenosine sites using a universal deep learning model. *Comput Struct Biotechnol J* 2021;**19**: 4619–25.
55. Soydaner D. Attention mechanism in neural networks: where it comes and where it goes. *Neural Computing and Applications* 2022;**34**(16):13371–85.
56. Trockman A. and Zico Kolter J. *Patches Are All You Need?* 2022. arXiv:2201.09792.
57. Boureau Y. L., Ponce J. and Lecun Y. *A theoretical analysis of feature pooling in visual recognition*. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. Haifa, Israel: ACM, New York, 2010; 111–118.
58. Boureau Y. L., Roux N. L., Bach F., *et al. Ask the locals: multi-way local pooling for image recognition*. In: *2011 International Conference on Computer Vision*. Barcelona, Spain: IEEE Computer Society, Piscataway, 2011.
59. Zeiler MD, Fergus R. Stochastic pooling for regularization of deep convolutional neural networks, In: *International Conference on Learning Representations*. Scottsdale, Arizona, USA, 2013.
60. He K, Zhang X, Ren S, *et al.* Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 2015;**37**(9):1904–16.
61. Vaswani A., Shazeer N., Parmar N., *et al. Attention is all you need*. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, CA, USA: Curran Associates Inc., 2017.
62. Devlin J., Chang M.-W., Lee K., *et al. BERT: pre-training of deep bidirectional transformers for language understanding*. In: *Proceedings of NAACL-HLT*. Minneapolis, Minnesota, USA: ACL, Stroudsburg, 2019; **1**:4171–86.
63. Al-Rfou R, Choe D, Constant N, *et al. Character-level language modeling with deeper self-attention*. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Honolulu, Hawaii, USA: AAAI Press, Menlo Park, 2019;**33**:3159–66.
64. Maruf S, Martins AFT, Selective HG. *Attention for context-aware neural machine translation*. In: *Proceedings of NAACL-HLT*, Minneapolis, Minnesota, ACL, Stroudsburg, 2019;**1**:3092–2.
65. Dai Z., Yang Z., Yang Y., *et al. Transformer-XL: attentive language models beyond a fixed-length context*. In: *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*. Florence, Italy: ACL, Stroudsburg, 2019;2978–88.
66. Dosovitskiy A., Beyer L., Kolesnikov A., *et al. An image is worth 16x16 words: transformers for image recognition at scale*. In: *International Conference on Learning Representations*. Vienna, Austria: ICLR, 2021.
67. Rush A., *The annotated transformer*. In: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Melbourne, Australia: ACL, Stroudsburg, 2018, 52–60.
68. Amores J. Multiple instance classification: review, taxonomy and comparative study. *Artificial Intelligence* 2013;**201**:81–105.
69. Zhou Z-H. A brief introduction to weakly supervised learning. *Natl Sci Rev* 2018;**5**(1):44–53.
70. Wu J., Yinan Y., Chang H., *et al. Deep multiple instance learning for image classification and auto-annotation*. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, USA: IEEE Computer Society, Piscataway, 2015.
71. Zhang Q, Zhu L, Bao W, *et al.* Weakly-supervised convolutional neural network architecture for predicting protein-DNA binding. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**17**(2): 679–89.
72. Zhang Q, Shen Z, Huang DS. Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network. *Sci Rep* 2019;**9**(1):8484.
73. Ilse M, Tomczak JM, Welling M. *Attention-based deep multiple instance learning*. In: *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden: ACM, New York, 2018;**80**:2127–36.
74. Liu K, Cao L, Du P, *et al.* im6A-TS-CNN: identifying the N6-methyladenine site in multiple tissues by using the convolutional neural network. *Molecular Therapy - Nucleic Acids* 2020;**21**: 1044–9.