electronics

MDPI

Article

MLPPF: Multi-Label Prediction of piRNA Functions Based on Pretrained k-mer, Positional Embedding and an Improved TextRNN Model

Yajun Liu [1,*], Ru Li [1], Yang Lu [1], Aimin Li [1], Zhirui Wang [2] and Wei Li [1]

[1] Shaanxi Key Laboratory for Network Computing and Security Technology, School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China
[2] College of Life Sciences, Northwest A&F University, Yangling 712100, China
* Correspondence: liuyajun@xaut.edu.cn

Abstract: PIWI-interacting RNAs (piRNAs) are a kind of important small non-coding RNAs and play a vital role in maintaining the stability of genome. Previous studies have revealed that piRNAs not only silence transposons, but also mediate the degradation of a large number of mRNAs and lncRNAs. Existing computational models only focus on mRNA-related piRNAs and rarely concentrate on lncRNA-related piRNAs. In this study, we propose a novel method, MLPPF, which is designed for multi-label prediction of piRNA functions based on pretrained k-mer, positional embedding and an improved TextRNN model. First, a benchmark dataset, which contains two types of functional labels, namely mRNA-related and lncRNA-related piRNAs, was constructed by processing piRNA-function-annotated data and sequence data. Moreover, pretrained k-mer embedding fused with positional embedding was applied to get the sequence representation with biological significance. Finally, an improved textRNN model with Bi-GRU and an attention mechanism was employed for implementing the piRNA functional label prediction task. Experiments substantiate that our model can effectively identify the piRNA functional labels, reveal the key factors of its subsequences and be helpful for in-depth investigations into piRNA functions.

Keywords: piRNA functional label predictor; k-mer positional embedding; multi-label classification

Citation: Liu, Y.; Li, R.; Lu, Y.; Li, A.; Wang, Z.; Li, W. MLPPF: Multi-Label Prediction of piRNA Functions Based on Pretrained k-mer, Positional Embedding and an Improved TextRNN Model. *Electronics* **2024**, *13*, 92. https://doi.org/10.3390/electronics13010092

Academic Editor: Silvia Liberata Ullo

Received: 28 October 2023
Revised: 15 December 2023
Accepted: 19 December 2023
Published: 25 December 2023

1. Introduction

Genetic factors have important biological functions and are closely related to diseases. Identifying the biological functions of genetic factors is useful in analyzing the complex mechanisms of diseases and helpful in disease prevention, diagnosis, and treatment. piRNAs are a kind of small non-coding RNA (sncRNA) discovered in 2006 that bind to the PIWI subfamily of the Argonaute protein. They are slightly longer than miRNAs by approximately 24–32 nucleotides. They lack clear secondary structures but have two sequence preferences (the 5′ end uridine or 10th adenosine) [1,2].

Based on their origins, piRNAs are divided into three categories: piRNAs derived from inter-gene regions, piRNAs derived from mRNA, and piRNAs derived from lncRNA. Given their different origins, piRNAs have a variety of functions. Early studies found that piRNAs can recognize and silence transposons, thereby maintaining the stability of the genome structure [3]. As early as 2014, Gou et al. [4] found that piRNAs guided the large-scale elimination of mRNAs at the later stage of spermatogenesis in mice; that was the first research to demonstrate that piRNAs regulate mRNAs. In 2015, Watanabe et al. [5] discovered that piRNAs from transposons and pseudogenes mediate the degradation of large amounts of mRNA and lncRNA in mice. In 2019, Dai et al. [6] revealed that piRNAs are widely involved in mRNA translational activation in sperm cells and conducted in-depth and systematic research on the molecular mechanisms of positive piRNA regulation in target gene translation. Recent research on a *C. elegans* model showed that the regulatory

range of piRNAs involves almost all mRNAs in germ cells [7]. In summary, these breakthrough studies suggest that piRNA functions are important and diverse and that they mediate a highly complex RNA regulatory network.

Identifying piRNA functions is a fundamental and challenging problem. There are currently two existing recognition methods: biological experimental methods and computational methods. Many biological experimental techniques can be used to recognize piRNA functions, such as immunoprecipitation. However, experiments identifying piRNA functions are time-consuming, expensive, have only a minor flux, and require professional operators. With the establishment of piRNA-specific databases and the accumulation of piRNA-related functional data, it is possible to predict piRNA functional labels via computational methods [8].

Much research has been conducted in identifying the functional labels of miRNAs and lncRNAs, but prediction algorithm research on piRNA functional labels is still in its infancy. The current study focuses on mRNA-related piRNA using a miRNA-like mechanism [4,9,10], and lncRNA-related piRNA is disregarded. In 2017, Liu et al. [11] proposed 2L-piRNA, the first piRNA function predictor. By extracting features such as pseudo-nucleotide components (PseKNC) and using SVM classifiers, the authors built a two-layer ensemble classifier. The first layer is used to identify whether a query RNA molecule is piRNA or non-piRNA with 86.1% accuracy. The second layer identifies whether a piRNA is involved in mRNA deadenylation with 77.6% accuracy. In 2018, Li et al. [12] presented a 2L-piRNAPred algorithm based on features including nucleotide composition, positional specificity, physicochemical properties, and an F-value feature selection algorithm, achieving accuracies of 89% (first layer) and 84% (second layer). In 2019, Khan et al. proposed the first computational tool based on deep learning, 2L-piRNADNN [13]. This method constructs a deep neural network based on feature vectors containing dinucleotide autocovariance and physicochemical properties, with accuracies of 91.81% (first layer) and 84.52% (second layer). In 2020, Zuo et al. [14] proposed 2lpiRNAred using a sparse representation classifier (SRC) and a support vector machine with a radial basis function using Markov distance (SVMMDRBF). This study also proposed a new feature selection algorithm based on Luca fuzzy entropy and Gaussian membership function (LFE-GM), with accuracies of 88.72% (first layer) and 79.97% (second layer). Meanwhile, Khan et al. [15] proposed a 2L-PseKNC algorithm based on PseKNC and deep neural networks. This algorithm uses principal component analysis (PCA) to select features and can achieve accuracies of 94.73% (first layer) and 85.21% (second layer). In summary, such studies only focus on mRNA-related piRNA recognition and their performance still needs to be improved.

In addition to the above studies, only a few prediction tools for piRNA target loci are available. Gou et al. [4] were the first to identify the potential targeting sites of piRNAs within the three prime untranslated regions (3′ UTRs) of MIWI-associated mRNAs. In that study, miRanda [16], a miRNA target prediction tool, was used. Based on the experimental data of Gou et al. [17], pirnaPre was the first tool designed for piRNA target locus identification. It used mouse data and selected a combination of MIWI CLIP-seq-derived features and position-derived features to train an SVM classifier. A training area under the curve (AUC) of 0.87 was achieved, and 3781 mRNAs from 2587 protein-coding genes were predicted as potential piRNA targets. pirScan [18,19] is another tool for predicting piRNA target sites based on established targeting rules from *C. elegans* and *C. briggsae* data. Yang et al. [20] developed the first deep learning method based on multi-head attention, identifying piRNA targeting sites on *C. elegans* mRNAs and obtaining an AUC of 93.3% using an independent test set. However, these methods used experimentally validated data from specific species and cannot currently be extended to other organisms.

Recent studies have demonstrated that a piRNA can perform different functions in different times and spaces and can have multiple functional labels. For example, a piRNA can not only eliminate transposons during the ping-pong cycle but also mediate mRNA degradation in late cell stages [21]. Thus, there is a need for a careful investigation and dissection of piRNA functions. In this study, we propose a novel method, MLPPF, designed

for multi-label predictions of piRNA functions based on pretrained k-mer embedding, positional embedding, and an improved TextRNN model. First, a benchmark dataset was constructed by processing piRNA-function-annotated data and sequence data. Moreover, pretrained k-mer and positional embedding were applied to achieve biologically significant sequence representation. Finally, an end-to-end model that couples the discriminative power of representation with an improved textRNN was used to implement the piRNA functional label prediction task. Compared with the other three methods, MLPPF performed best and revealed the key factors of piRNA subsequences, thus demonstrating its effectiveness.

2. Materials and Methods

To identify the functional labels of piRNAs, our study builds a benchmark dataset, encodes a piRNA sequence with pretrained k-mer and positional embedding; and treats the piRNA function identification problem as an imbalanced multi-label learning issue by using an improved TextRNN model. The prediction process scheme is shown in Figure 1 and is divided into 4 main steps. They are benchmark dataset construction, piRNA sequence encoding, TextRNN model construction, and performance evaluation, which are explained in detail below.

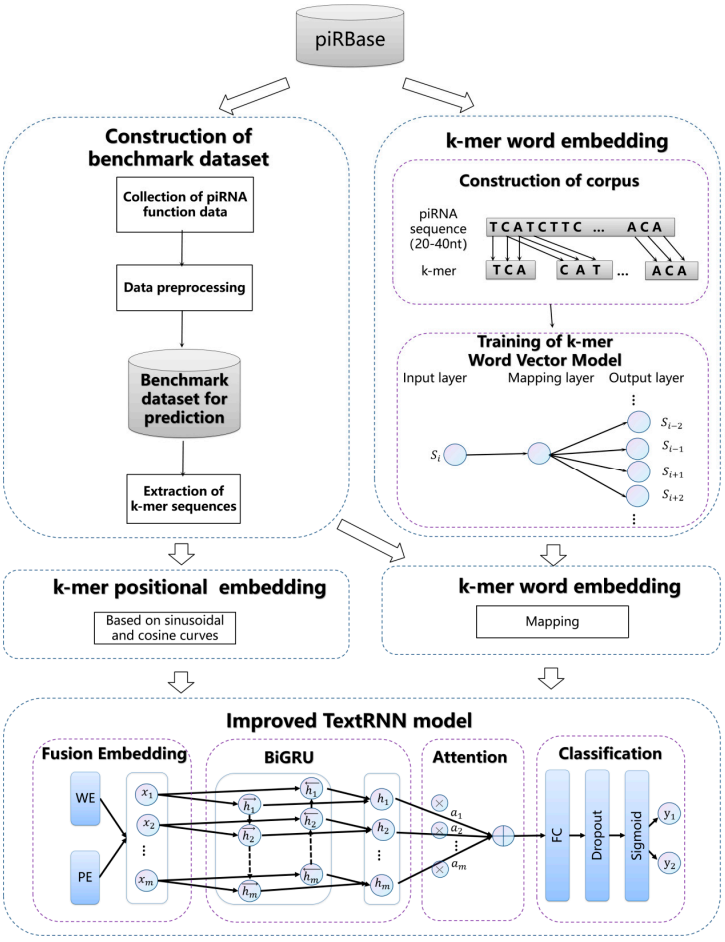


Figure 1. Framework illustration of piRNA functional label prediction.

2.1. Datasets

According to the NONCODE and piRBase databases, Liu et al. [11] constructed the first and most popular benchmark dataset for identifying mouse piRNA functional subsets in 2017, containing 1418 non-piRNA sequences; 709 piRNAs with the function of instructing target mRNA deadenylation; and 709 piRNAs without such a function. In 2020, they improved the dataset by increasing the contents of the three subsets to 798, 1257, and 2068,

respectively. Although fruit fly species were added to this improved dataset, most studies still focus on mRNA-related piRNAs and ignore lncRNA-related piRNAs.

In our study, mRNA-related and lncRNA-related piRNAs were considered. The piRNA target data were downloaded in 2022 from piRBase (http://bigdata.ibp.ac.cn/piRBase, accessed on 20 January 2022) and extracted from the published literature [22]. After eliminating redundant data, 2628 non-duplicate piRNA sequences were obtained. In a piRNA target record, piRBase provides a piRBase name linked to detailed information concerning the piRNA, a target gene/transcript ID, a target site, and other relevant information. According to the interaction characteristics of piRNAs, piRNA sequences are marked with functional tags. We focused on two types of functional labels: mRNA-related piRNAs and lncRNA-related piRNAs. The functional labels for specific piRNAs are based on predicted and experimentally validated information about piRNA target sites. piRNA target mRNAs in mouse tissues were extracted from the published literature and used to experimentally verify piRNA target relationships. The prediction of potential piRNA target lncRNAs was performed the same way as piRNA target mRNA cleavage in mouse testes is predicted [4,9,10].If the piRNA only targets an mRNA, the functional label of the piRNA is marked as [1, 0]. If the piRNA only targets a lncRNA, the functional label of the piRNA is marked as [0, 1]. If both piRNA-mRNA and piRNA-lncRNA interactions exist, the functional label of the piRNA is marked as [1, 1]. In the benchmark dataset, the distribution of functional piRNAs reveals that 1806 piRNAs are only related to mRNA, accounting for 68.7%. Additionally, 538 piRNAs are only related to lncRNA, accounting for 20.5%. Furthermore, 284 piRNAs are related to both mRNA and lncRNA, accounting for 10.8%.

2.2. Sequence Embedding

In our study, a fusion of k-mer word embedding and positional embedding was used to obtain a piRNA sequence representation with biological significance, resulting in high accuracy and reliability in the prediction.

2.2.1. k-mer Word Embedding

RNA sequences need to be encoded into a numeric vector before machine learning can take place. Most biological sequence analysis methods use one-hot encoding for their k-mer fragments. While such encoding is popular, it results in high sparsity and does not consider the biological significance of ribonucleic acids. It hardly reflects sequence similarities and differences.

Inspired by studies on natural language processing (NLP), we analogized a piRNA sequence to a sentence and its k-mer subsequence to a word. By using a word-embedding model, representations of piRNA sequences with biological significance were obtained.

Referenced to Zeng et al. [23], a piRNA sequence, S, can be denoted as

$$\mathrm{S} = \mathrm{N}_1, \mathrm{N}_2, \ldots, \mathrm{N}_n \tag{1}$$

where N_i is the i-th nucleotide, $\mathrm{N}_i \in \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}, i = 1, 2, 3, \ldots, n$, and n is the length of the piRNA sequence. The piRNA sequence is stored in a cDNA (complementary DNA) format.

We used a sliding window to extract m k-mer sequence fragments from a piRNA sequence S, as follows:

$$f(\mathrm{S}) = \mathrm{S}_{1:k}, \mathrm{S}_{2:2+k}, \ldots, \mathrm{S}_{m:m+k} \tag{2}$$

$$m = n - k + 1 \tag{3}$$

where $\mathrm{S}_{i:i+k}$ is the i-th k-mer subsequence, k is the width of the sliding window, and m is the number of k-mer subsequences in the piRNA sequence, S.

Word2vec [24] is a word-embedding technique designed to represent standard words in a corpus as vectors with word semantics. This technique has improved multiple natural language processing tasks. The Skip-gram model is a useful example of a Word2vec model;

it is used to obtain the embedding vectors of k-mers. piRNA sequences in the piRBase database serve as a corpus, and their k-mer sequence fragments, extracted by the sliding window, are used to pretrain vectors with biological meanings in an unsupervised manner. To maximize the probability of the target k-mer and the co-occurrence of the context sequence, the word vector e_i of the k-mer sequence fragment $S_{i:i+k}$ is obtained. The value of k ranged from 2 to 5 in our experiment, and the value selection is discussed in Section 3.1.

The set of m-many k-mer sequence fragments $f(S)$ corresponding to the piRNA sequence S is transformed into a pretrained embedding representation, which is expressed as follows:

$$WE(S) = [e_1, e_2, \ldots, e_m] \tag{4}$$

2.2.2. k-mer Positional Embedding

In natural language processing technology, positional vectors are used to represent the position information of words. To enhance our model in capturing the sequential characteristics of sequences, positional embedding (*PE*) was used to identify piRNA functional labels.

Since the binary representation of position vectors wastes substantial memory, positional embedding generated by sinusoidal and cosine curves of varying frequencies was applied in our study. The positional embedding of the *i*-th k-mer subsequence is a vector determined by both the k-mer position and the component position, calculated as follows:

$$PE_{i,2j} = \sin\left(i/10000^{2j/d}\right) \tag{5}$$

$$PE_{i,2j+1} = \cos\left(i/10000^{2j/d}\right) \tag{6}$$

where i represents the position of the k-mer subsequence in a piRNA, and j represents the component position of the k-mer positional embedding vector. The range of i is [1, m], the range of j is [0, $\left\lfloor \frac{d-1}{2} \right\rfloor$], and d represents the dimension of the positional embedding. $PE_{i,2j}$ and $PE_{i,2j+1}$ denote the value at even and odd component positions in the positional embedding vector of the i-th k-mer subsequence, respectively. This positional embedding is based on sinusoidal and cosine curves and provides our model with the ability to model the position of a k-mer subsequence and the distance of every two k-mer subsequences.

2.3. Improved TextRNN Network

In NLP, the TextRNN model [25] can capture the most important semantic information in a sentence and improve the performance of multi-label identification tasks. Hence, in our study, an improved TextRNN network was used to systematically recognize piRNA functional labels in four parts: a fusion-encoding layer, a sequence encoder, an attention layer, and a classification layer. The details of the above components are described in the following sections.

2.3.1. k-mer Encoding Layer

In the embedding layer, the final embedding matrix is an aggregate matrix $\boldsymbol{P} \in \mathbb{R}^{m \times d}$, which is accumulated by the k-mer word embedding matrix and the positional embedding matrix. d refers to the dimension of both positional embedding and the pretrained embedding representation. It is computed as follows:

$$\boldsymbol{P} = WE + PE \tag{7}$$

2.3.2. Sequence Encoder

TextRNN [25] is an RNN-based flexible neural network designed for text data processing. In multiple text tests, TextRNN has achieved good results. However, traditional RNNs are likely to face vanishing or exploding gradient problems. As a variant of recurrent

neural networks, gated recurrent units (GRUs) [25] can effectively address these problems. By simplifying the gating structure, GRUs become faster than Long Short-Term Memory (LSTM) and can better learn long-term dependencies. Therefore, a GRU can be successfully used to model sequence data in NLP, especially in sequence classification tasks and sequence annotation tasks. Typically, a GRU contains two gates, the update gate and the reset gate, which are used to determine the retention or disposal of information. The update gate z decides how much of the new input should be used to update the hidden state, while the reset gate r determines how much of the previous hidden state should be forgotten. The update gate z_i and reset gate r_i of the i-th hidden unit are computed by

$$z_i = \sigma(W_z x_i + U_z h_{i-1} + b_z) \tag{8}$$

$$r_i = \sigma(W_r x_i + U_r h_{i-1} + b_r) \tag{9}$$

where σ is the sigmoid function, x_i is the vector of the current-position k-mer fragment, and h_{i-1} is the final hidden state of the previous position k-mer fragment. The weight matrices W_z, W_r, U_z, and U_r are learned, and b_z and b_r are biases.

The hidden state h_i and candidate hidden state $\tilde{h}_i$ of the GRU are computed by

$$h_i = (1 - z_i) \odot h_{i-1} + z_i \odot \tilde{h}_i \tag{10}$$

$$\tilde{h}_i = \tanh(\mathrm{W}_h x_i + r_i \odot (U_h h_{i-1}) + b_h) \tag{11}$$

where $\odot$ is the dot product, W_h and U_h are weight matrices, and b_h is a bias.

Bi-GRU is a structure-reinforcing GRU neural network and provides the output layer with complete contextual information about the input data at each moment. The input sequence propagates through both a forward GRU and a backward GRU and then concatenates the outputs of both. The calculation process of the final hidden state of the current k-mer fragment, h_i, is as follows:

$$\overrightarrow{h}_i = \overrightarrow{\mathrm{GRU}}\left(x_i, \overrightarrow{h}_{i-1}\right) \tag{12}$$

$$\overleftarrow{h}_i = \overleftarrow{\mathrm{GRU}}\left(x_i, \overleftarrow{h}_{i+1}\right) \tag{13}$$

$$\begin{gathered} h_i = \left[\overrightarrow{h}_i, \overleftarrow{h}_i\right] \\ \boldsymbol{H} = [h_1, h_2, \ldots, h_m] \end{gathered} \tag{14}$$

where $\overrightarrow{h}_i$ and $\overleftarrow{h}_i$ are the forward and backward hidden states of the current k-mer fragment, respectively. $\overrightarrow{h}_{i-1}$ represents the forward hidden state of the previous position k-mer fragment. $\overleftarrow{h}_{i+1}$ represents the backward hidden state of the next position k-mer fragment. Finally, the output of the piRNA sequence encoder, $\boldsymbol{H} \in \mathbb{R}^{m \times h}$, is obtained by consolidating the encoding information from both directions of each k-mer fragment, where h is the dimension of the final hidden state.

2.3.3. Attention Layer

Recent studies have shown that piRNAs degrade mRNAs and lncRNAs through a miRNA-like mechanism. This suggests that a specific sequence fragment at a specific position is required for piRNA functions. The model should selectively focus on specific k-mer piRNA fragments that have biological functions. An attention mechanism can

enable neural networks to focus on a subset of their inputs, so an attention layer is used in our model.

The output of the sequence encoder serves as the input to this layer. As referenced by Yang et al. [26], a k-mer-level context vector, u_s, can be introduced, and a normalized importance weights, a_i, can be calculated using a softmax function. To measure the importance of k-mer subsequences, the computation process is as follows:

$$u_i = \tanh(W_s h_i + b_s) \tag{15}$$

$$a_i = \frac{\exp(u_i^T u_s)}{\sum_i \exp(u_i^T u_s)} \tag{16}$$

$$\widetilde{\boldsymbol{P}} = \sum_{i=1}^{m} a_i h_i \tag{17}$$

where W_s is a trainable weight; b_s is a bias, a_i is the attention weight of the i-th k-mer; and $\widetilde{\boldsymbol{P}}$ represents the new weighted representation of the piRNA, as calculated by the sum of element-wise multiplication between attention weights and the input data of that layer.

2.3.4. Classification Layer

We treated this piRNA function prediction task as a multi-label classification problem. We introduced a fully connected layer for linear transformation into the classification layer of the model, allowing each sample to have the possibility of multiple labels. To reduce the model's dependence on specific neurons and improve generalization capabilities, we performed a dropout operation on the output of the fully connected layer. Finally, a sigmoid activation function was used to generate the predicted probability of the corresponding label for each sample, ensuring they are independent of each other.

2.4. Evaluation Methods

In our research, 10-fold cross-validation was used to evaluate the performance of the model. Samples were randomly divided into 10 folds. Each fold was used to validate the model, with the remaining data used to train it.

Evaluating the performance of multi-label classification differs from multi-class classification. In multi-class classification, each sample only belongs to one class, and the class labels are mutually exclusive. Predictions can be either entirely correct or incorrect. In multi-label classification, each sample can belong to multiple classes, and evaluating multi-label classification methods is like evaluating of information retrieval methods.

In the present study, the multi-label corpus C is the piRNA sequence set with functional labels, where each sample in this set is represented as (a_i, B_i), $i = 1, 2, \ldots, |C|$, and $|C|$ is the total number of samples in C. a_i is a piRNA sample, B_i is the true associated label set of a_i, and Cl_i is a label set obtained with the prediction method for a_i. L represents the total number of labels.

To evaluate our piRNA function multi-label prediction method, two kinds of metrics were used. One is a sample-based metric, another is a label-based metric. Sample-based metrics include accuracy, precision, recall, F1 score, hamming loss, etc. To better highlight that our study falls within the domain of multi-label classification, we provide explanations for hamming loss, micro_F1, and macro_F1.

- Hamming loss

Hamming loss is used to estimate the incorrect classification ratio, indicating both the failure to predict correct labels and the prediction of incorrect labels. The lower the Hamming loss, the better the performance. The formula is as follows:

$$Hamming\ loss(h, Cl) = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{|B_i \Delta Cl_i|}{L} \tag{18}$$

where Δ represents the symmetric difference between two sets.

Two label-based metrics are also applied, and they are as follows:

- Micro_F1

Micro_F1 is the harmonic mean of *micro_precision* and *micro_recall*, calculated as follows:

$$Micro\_F1 = \frac{2 \times micro\_precision \times micro\_recall}{micro\_precision + micro\_recall} \tag{19}$$

where micro_precision represents the overall precision of all labels.

- Macro_F1

Macro_F1 is the harmonic mean of among multi-label precision and recall and is calculated as follows:

$$Macro\_F1 = \frac{1}{L}\sum_{j=1}^{L}\frac{2 \times precision_j \times recall_j}{precision_j + recall_j} \tag{20}$$

where precision_j represents the precision of the j-th label.

Apart from the above metrics, the receiver operating characteristic curve (ROC) and the precision–recall curve (PRC) are also used to evaluate the performance of methods.

3. Results

3.1. Hyperparameter Optimization

Many hyperparameters influence prediction performance. The grid search method is used for hyperparameter optimization. The best hyperparameters were selected based on a performance evaluation with the validation set. In the sequence-processing phase, k is a critical parameter of the current task, specifically representing the length of overlapping k-mer fragments obtained through sliding window extraction. If k is too small or too large, the prediction or computation performance is affected. In the pretrained k-mer-embedding phase, we utilized k-mer fragments with lengths ranging from two to five for word vector representation. The impact of the different tested k-mer lengths on performance is shown in Figure 2. The results indicate that F1, Micro_F1, and Macro_F1 slowly grow as k increases. Considering both the prediction performance and computation cost, k is set to 4, and the dimension of the pretrained vectors of k-mer is configured as 64.

In constructing the prediction model, BCEWithLogitsLoss was employed as the loss function, the Adam optimizer was chosen, the batch size was 50, the number of epochs was 10, the learning rate was 0.008, the dropout rate for hidden layers was 0.5, and the hidden layer dimension was set to 256.

3.2. Sequence Embedding Representation Comparison

This section describes four different sequence-encoding methods used to represent overlapping k-mer fragments, including randomly initialized k-mer embedding (Rand_E), pretrained k-mer embedding (Pre_E), randomly initialized k-mer embedding fused with k-mer positional embedding (Rand_E + Pos_E), and pretrained k-mer embedding fused with k-mer positional embedding (Pre_E + Pos_E).

To prove the effectiveness of the Pre_E + Pos_E representation, we compared the proposed method with the other three different embedding methods, as shown in Table 1. Comparing the performance of the Rand_E and Pre_E encoding representations, the overall

prediction performance of Pre_E was higher than Rand_E. Using Pos_E as the basis for Pre_E, pretrained k-mer embedding fused with k-mer positional embedding performed best among the seven evaluation metrics. The results indicate that the proposed embedding representation effectively captures the semantic information and relative positional relationships of the k-mer fragments.

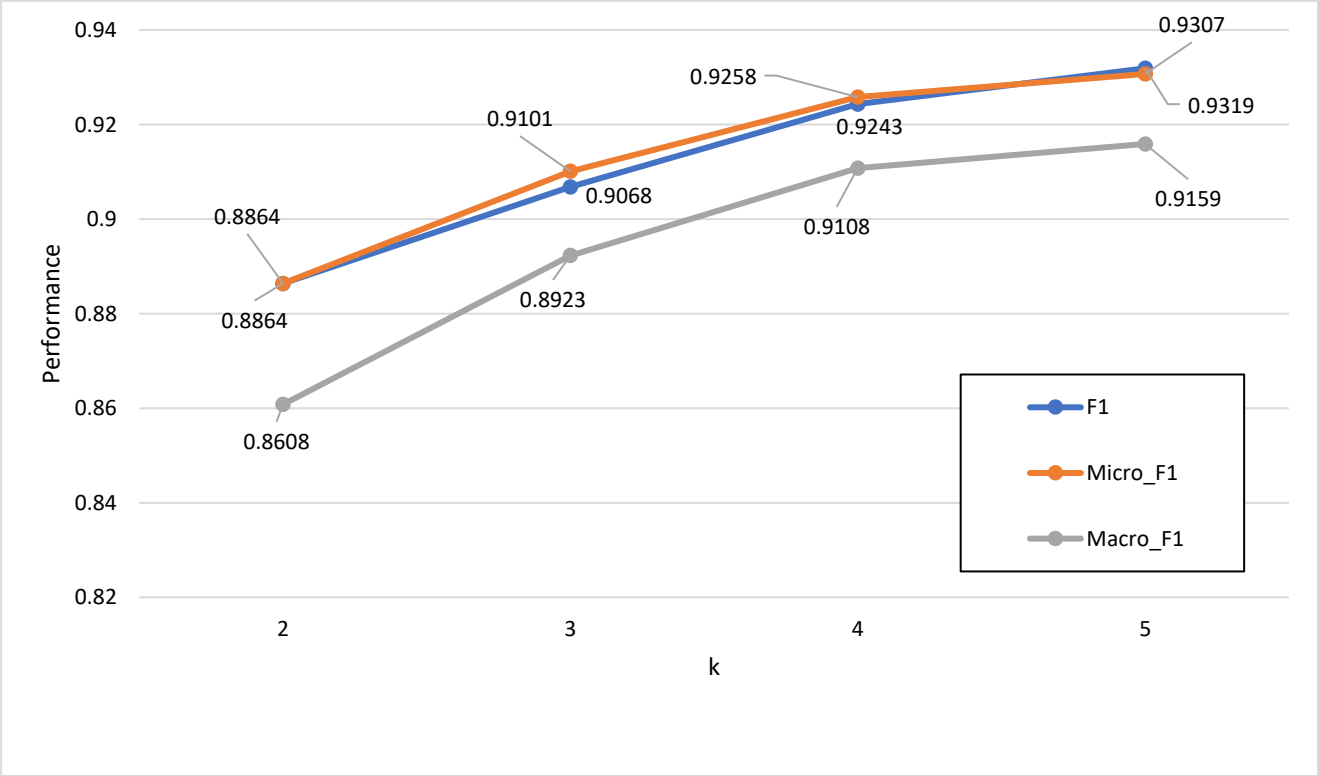


Figure 2. The predictive results with different k-mer lengths.

Table 1. Performance comparison of improved TextRNN model using four sequence-embedding methods.

Representation	Acc	Pre	Recall	F1	HL	Micro_F1	Macro_F1
Rand_E	0.9054	0.9189	0.9256	0.9167	0.0906	0.9185	0.9029
Pre_E	0.9085	0.9197	0.9252	0.9178	0.0854	0.9229	0.9079
Rand_E + Pos_E	0.9098	0.9229	0.9260	0.9196	0.0843	0.9237	0.9083
Pre_E + Pos_E	**0.9134**	**0.9231**	**0.9363**	**0.9243**	**0.0830**	**0.9258**	**0.9108**

Abbreviations: Acc, Accuracy; Pre, Precision; HL, Hamming Loss.

3.3. Performance Comparison

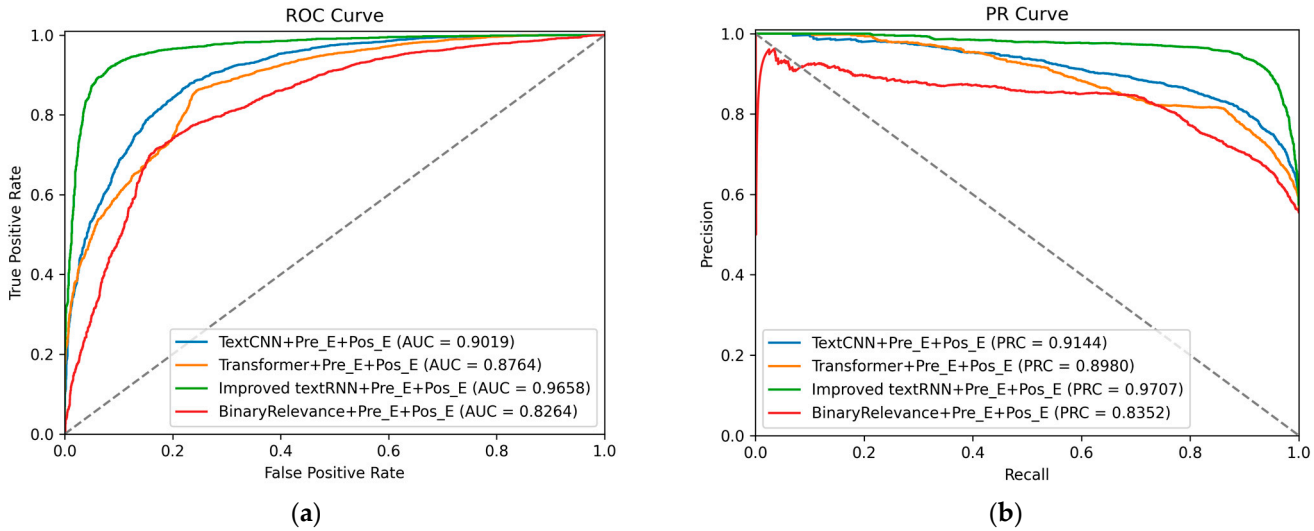
To better evaluate the performance of the improved TextRNN model in predicting piRNA functional multi-labels, we compared it with TextCNN, Transformer and Binary Relevance (BR) using pretrained k-mer embedding fused with positional embedding representation. TextCNN is a text classification model that can capture local sequence correlations well by using convolution cores and pooling operations. Transformer is a classic model proposed by Google, mainly used in NLP to capture the semantic relationships between sequences. Considering that non-deep methods achieve good performance [27], BR which decomposes the multi-label learning problem into multiple independent binary classification tasks is selected in this comparison. Using the BR method, we built an independent SVM model for each class label.

The performance of the four methods is shown in Table 2; seven performance metrics are listed. The results demonstrate that our method achieved the best performance; six indicators were above 90%, and the overall performance was higher than TextCNN, Transformer, or BinaryRelevance. Obviously, our proposed computational method performs better than other methods. In addition, ROC and PRC were used to evaluate classification performance using the Pre_E + Pos_E embedding method, as depicted in Figure 3a,b. Among the four comparison methods, the ROC curve based on the improved TextRNN was significantly higher than the other methods, and its AUC was the highest at 96.58%.

Table 2. Performance comparison of four methods based on pretrained k-mer embedding fused with k-mer positional embedding.

Method	Acc	Pre	Recall	F1	HL	Micro_F1	Macro_F1
BinaryRelevance	0.7584	0.8016	0.7723	0.7774	0.2321	0.7828	0.6536
TextCNN	0.8274	0.8383	0.8719	0.8459	0.1718	0.8495	0.8218
Transformer	0.8132	0.8288	0.8569	0.8329	0.1868	0.8355	0.8022
Improved TextRNN	**0.9134**	**0.9231**	**0.9363**	**0.9243**	**0.0830**	**0.9258**	**0.9108**

Abbreviations: Acc, Accuracy; Pre, Precision; HL, Hamming Loss.



(**a**) (**b**)

Figure 3. Performance curves of Four methods: (**a**) is the ROC curves of four models; (**b**) is the PRC curves of four models.

In summary, our method had the best predictive performance using the aggregate embedding method and the improved TextRNN model.

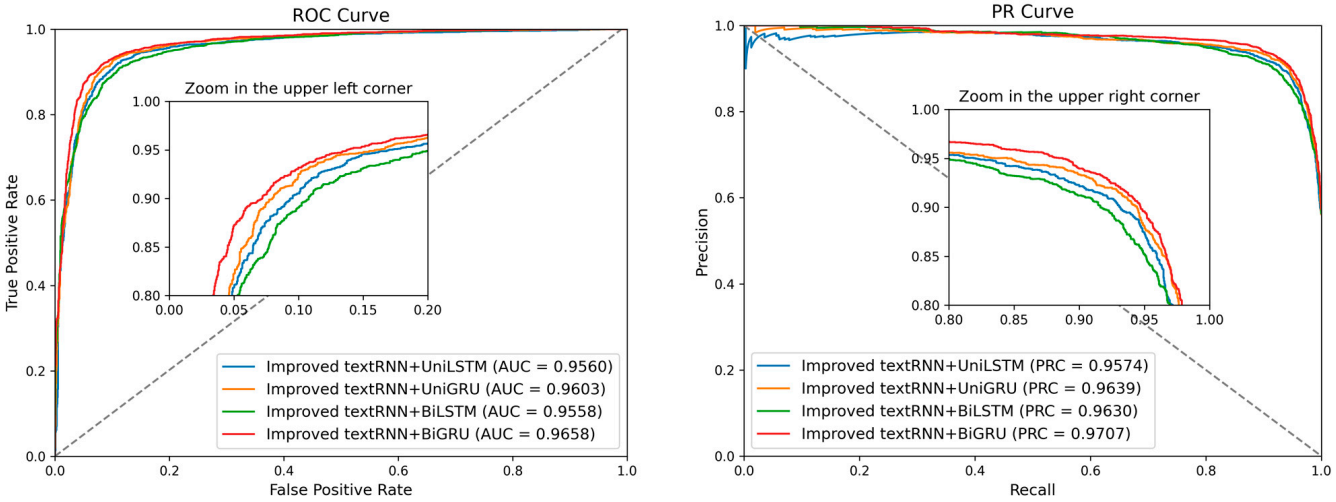
3.4. Performance Comparison of RNN Structure

RNN is a flexible network structure. The effects of different RNN structures on model performance are illustrated in Table 3. These structures include unidirectional GRU (UniGRU), bidirectional GRU (BiGRU), unidirectional LSTM (UniLSTM), and bidirectional LSTM (BiLSTM). The results show that the final hidden states obtained by BiGRU contained the most sequence context information and had the best performance among the six evaluation metrics. In UniGRU, the highest precision was achieved at 92.39%.

Table 3. Performance comparison of different RNN structures.

RNN	Acc	Pre	Recall	F1	HL	Micro_F1	Macro_F1
UniGRU	0.9123	**0.9239**	0.9317	0.9226	0.0854	0.9235	0.9068
BiGRU	**0.9134**	0.9231	**0.9363**	**0.9243**	**0.0830**	**0.9258**	**0.9108**
UniLSTM	0.9062	0.9184	0.9285	0.9177	0.0932	0.9167	0.9017
BiLSTM	0.8967	0.9104	0.9193	0.9088	0.1020	0.9086	0.8903

Next, we evaluated the ROC and PRC of our model in the different RNN structures, as shown in Figure 4a,b. To more clearly demonstrate the impact of different RNN structures on model performance, we choose to enlarge the *x*-axis range from 0.0 to 0.2 and the *y*-axis range from 0.8 to 1.0 in the ROC curve and enlarge the *x*-axis range from 0.8 to 1.0 and the *y*-axis range from 0.8 to 1.0 in the PR curve. The results reveal that all four RNN structures achieved AUC values surpassing 95%. BiGRU reached the highest at 96.58%, followed by UniGRU at 96.03%, UniLSTM at 95.60%, and BiLSTM at 95.58%. Through comparative analysis, we found that GRU performed slightly better than LSTM in predicting piRNA functional labels, and BiGRU proved more effective in capturing the semantic relationships within the piRNA sequences.



(**a**) (**b**)

Figure 4. (**a**) is the ROC curves of four RNN structures; (**b**) is the PRC curves of four RNN structures.

3.5. Attention Layer Analysis

At present, the attention mechanism is widely acknowledged as a crucial part of modern neural networks and has made significant advancements in various tasks. The attention mechanism was introduced to focus on key fragments of the piRNA sequence while ignoring irrelevant information.

First, ablation studies on the attention mechanism were performed, and the results are presented in Table 4. They indicate that the contribution of the attention layer in our model is not significant. Furthermore, we took a close look at the attention weights and plotted heat maps to reveal the key factors that are captured by the model using the attention mechanism. Different colors reflect the model's attention to different k-mer segments. The darker the color, the greater the contribution of the corresponding segment to the prediction task. The results of the piRNA demo are shown in Figure 5. For example, in the first piRNA in Figure 5, it seems that a focus is placed near the 5′ ends of the piRNA, and the k-mer fragment is "CCAG". We believe the attention mechanism helped us to mine key subsequences of functional piRNA as in the above case.

Table 4. Performance comparison of ablation studies on the attention layer.

Method	Acc	Pre	Recall	HL	F1	Micro_F1	Macro_F1
Our model without attention layer	0.9113	0.9229	0.9305	**0.0824**	0.9216	0.9257	**0.9117**
Our model with attention layer	**0.9134**	**0.9231**	**0.9363**	0.0830	**0.9243**	**0.9258**	0.9108

Figure 5. The Heatmap of piRNA sequences.

4. Conclusions and Future Directions

Compared with the prediction results for miRNA and lncRNA functional labels, research related to piRNAs is still in an early phase. In this study, we developed a computational method for predicting functional labels for piRNAs based on pretrained k-mer embedding, positional embedding, and an improved TextRNN model. First, we collected piRNA functional label data and sequence data and then constructed a benchmark dataset via data processing. Moreover, pretrained k-mer and positional embedding were applied to obtain sequence representation with biological significance. Finally, an end-to-end model that couples the discriminative power of representation with the improved textRNN was used to implement the piRNA functional label prediction task. In conclusion, our model can characterize piRNA functional labels and could be beneficial to researchers investigating piRNA functions.

Bioinformatics investigations necessitate extensive data. In contrast to the abundant databases on miRNA-mRNA and miRNA-lncRNA, functional annotated data on piRNA remain limited. Our ongoing studies revolve around a dedicated focus on contemporary advancements in piRNA research and will improve our work in the following future directions:

1. Expand and subdivide the functional labels of piRNA

Due to data limitations, we only focused on two types of functional labels for piRNA in this study. Existing studies have shown that piRNAs can not only degrade mRNA in a miRNA-like manner but also extensively activate mRNA translation. According to these biological facts, we will subdivide piRNA functional labels based on data accumulation. In addition, piRNA functions are diverse, as they can regulate various genetic factors such as transposable elements (TE). For these regulatory objects, we will expand the piRNA functional labels through data accumulation.

2. Identification of functional sites

Based on the transcriptome data, we will identify the functional sites of piRNA and explore piRNAs as potential biomarkers and drug targets using a computational method. This method can characterize the functional sites of piRNAs, helping researchers infer the potential regulatory functions of piRNAs and their binding mechanisms with other genetic factors.

Author Contributions: Conceptualization, Y.L. (Yajun Liu) and R.L.; methodology, software and writing, Y.L. (Yajun Liu), R.L. and Y.L. (Yang Lu); data collection, Z.W.; writing, review and editing, A.L. and W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62202374), the Natural Science Basic Research Program of Shaanxi Province of China (Program No. 2022JQ-719 and 2021JM-347), the China Postdoctoral Science Foundation (2021M693887).

Data Availability Statement: Denchmark dataset and source code for MLPPF are available at https://github.com/lralrac/MLPPF-master (accessed on 16 December 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aravin, A.; Gaidatzis, D.; Pfeffer, S.; Lagos-Quintana, M.; Landgraf, P.; Iovino, N.; Morris, P.; Brownstein, M.J.; Kuramochi-Miyagawa, S.; Nakano, T. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **2006**, *442*, 203–207. [CrossRef]
2. Liu, Y.; Li, A.; Zhu, Y.; Pang, X.; Hei, X.; Xie, G.; Wu, F.-X. piRSNP: A Database of piRNA-related SNPs and their Effects on Cancerrelated piRNA Functions. *Curr. Bioinform.* **2023**, *18*, 509–516. [CrossRef]
3. Zhang, S.; Pointer, B.; Kelleher, E.S. Rapid evolution of piRNA-mediated silencing of an invading transposable element was driven by abundant de novo mutations. *Genome Res.* **2020**, *30*, 566–575. [CrossRef]
4. Gou, L.T.; Dai, P.; Yang, J.H.; Xue, Y.C.; Hu, Y.P.; Zhou, Y.; Kang, J.Y.; Wang, X.; Li, H.R.; Hua, M.M.; et al. Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. *Cell Res.* **2014**, *24*, 680–700. [CrossRef]
5. Watanabe, T.; Cheng, E.C.; Zhong, M.; Lin, H. Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline. *Genome Res.* **2015**, *25*, 368. [CrossRef]
6. Dai, P.; Wang, X.; Gou, L.-T.; Li, Z.-T.; Wen, Z.; Chen, Z.-G.; Hua, M.-M.; Zhong, A.; Wang, L.; Su, H.; et al. A translation-activating function of MIWI/piRNA during mouse spermiogenesis. *Cell* **2019**, *179*, 1566–1581.e1516. [CrossRef]
7. Ramat, A.; Simonelig, M. Functions of PIWI Proteins in Gene Regulation: New Arrows Added to the piRNA Quiver. *Trends Genet.* **2021**, *37*, 188–200. [CrossRef]
8. Fei, R.; Wan, Y.; Hu, B.; Li, A.; Li, Q. A novel network core structure extraction algorithm utilized variational autoencoder for community detection. *Expert Syst. Appl.* **2023**, *222*, 119775. [CrossRef]
9. Zhang, P.; Kang, J.Y.; Gou, L.T.; Wang, J.J.; Xue, Y.C.; Skogerboe, G.; Dai, P.; Huang, D.W.; Chen, R.S.; Fu, X.D.; et al. MIWI and piRNA-mediated cleavage of messenger RNAs in mouse testes. *Cell Res.* **2015**, *25*, 193–207. [CrossRef]
10. Goh, W.S.S.; Falciatori, I.; Tam, O.H.; Burgess, R.; Meikar, O.; Kotaja, N.; Hammell, M.; Hannon, G.J. piRNA-directed cleavage of meiotic transcripts regulates spermatogenesis. *Genes Dev.* **2015**, *29*, 1032–1044. [CrossRef]
11. Liu, B.; Yang, F.; Chou, K.C. 2L-piRNA: A Two-Layer Ensemble Classifier for Identifying Piwi-Interacting RNAs and Their Function. *Mol. Ther. Nucleic Acids* **2017**, *7*, 267–277. [CrossRef]
12. Li, T.Y.; Gao, M.Y.; Song, R.Y.; Yin, Q.; Chen, Y. Support Vector Machine Classifier for Accurate Identification of piRNA. *Appl. Sci.* **2018**, *8*, 2204. [CrossRef]
13. Khan, S.; Khan, M.; Iqbal, N.; Hussain, T.; Khan, S.A.; Chou, K.C. A Two-Level Computation Model Based on Deep Learning Algorithm for Identification of piRNA and Their Functions via Chou's 5-Steps Rule. *Int. J. Pept. Res. Ther.* **2020**, *26*, 795–809. [CrossRef]
14. Zuo, Y.; Zou, Q.; Lin, J.; Jiang, M.; Liu, X. 2lpiRNApred: A two-layered integrated algorithm for identifying piRNAs and their functions based on LFE-GM feature selection. *RNA Biol.* **2020**, *17*, 892–902. [CrossRef]
15. Khan, S.; Khan, M.; Iqbal, N.; Khan, S.A.; Chou, K.-C. Prediction of piRNAs and their function based on discriminative intelligent model using hybrid features into Chou's PseKNC. *Chemom. Intell. Lab. Syst.* **2020**, *203*, 104056. [CrossRef]
16. John, B.; Enright, A.J.; Aravin, A.; Tuschl, T.; Sander, C.; Marks, D.S. Human MicroRNA targets. *PLoS Biol.* **2004**, *2*, e363. [CrossRef]
17. Yuan, J.; Zhang, P.; Cui, Y.; Wang, J.J.; Skogerbo, G.; Huang, D.W.; Chen, R.S.; He, S.M. Computational identification of piRNA targets on mouse mRNAs. *Bioinformatics* **2016**, *32*, 1170–1177. [CrossRef]
18. Wu, W.S.; Huang, W.C.; Brown, J.S.; Zhang, D.; Song, X.; Chen, H.; Tu, S.; Weng, Z.; Lee, H.C. pirScan: A webserver to predict piRNA targeting sites and to avoid transgene silencing in C. elegans. *Nucleic Acids Res.* **2018**, *46*, W43–W48. [CrossRef]

19. Zhang, D.; Tu, S.; Stubna, M.; Wu, W.S.; Huang, W.C.; Weng, Z.; Lee, H.C. The piRNA targeting rules and the resistance to piRNA silencing in endogenous genes. *Science* **2018**, *359*, 587–592. [CrossRef]
20. Yang, T.-H.; Shiue, S.-C.; Chen, K.-Y.; Tseng, Y.-Y.; Wu, W.-S. Identifying piRNA targets on mRNAs in C. elegans using a deep multi-head attention network. *BMC Bioinform.* **2021**, *22*, 503. [CrossRef]
21. Jensen, S.; Brasset, E.; Parey, E.; Crollius, H.R.; Sharakhov, I.V.; Vaury, C. Conserved Small Nucleotidic Elements at the Origin of Concerted piRNA Biogenesis from Genes and lncRNAs. *Cells* **2020**, *9*, 1491. [CrossRef]
22. Wang, J.; Shi, Y.; Zhou, H.; Zhang, P.; Song, T.; Ying, Z.; Yu, H.; Li, Y.; Zhao, Y.; Zeng, X.; et al. piRBase: Integrating piRNA annotation in all aspects. *Nucleic Acids Res.* **2022**, *50*, D265–D272. [CrossRef]
23. Zeng, M.; Wu, Y.; Lu, C.; Zhang, F.; Wu, F.X.; Li, M. DeepLncLoc: A deep learning framework for long non-coding RNA subcellular localization prediction based on subsequence embedding. *Brief. Bioinform.* **2022**, *23*, bbab360. [CrossRef]
24. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781. [CrossRef]
25. Liu, P.; Qiu, X.; Huang, X.J. Recurrent neural network for text classification with multi-task learning. *arXiv* **2016**, arXiv:1605.05101. [CrossRef]
26. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
27. Xing, L.; Lesperance, M.L.; Zhang, X. Simultaneous prediction of multiple outcomes using revised stacking algorithms. *Bioinformatics* **2020**, *36*, 65–72. [CrossRef]