algorithms

MDPI

Article

CSpredR: A Multi-Site mRNA Subcellular Localization Prediction Method Based on Fusion Encoding and Hybrid Neural Networks

Xiao Wang [1,2,*], Wenshuai Suo [1] and Rong Wang [3]

[1] School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450002, China; 332207050652@zzuli.edu.cn
[2] Henan Provincial Key Laboratory of Data Intelligence for Food Safety, Zhengzhou University of Light Industry, Zhengzhou 450002, China
[3] School of Electronic Information, Zhengzhou University of Light Industry, Zhengzhou 450002, China; wangrong@zzuli.edu.cn
* Correspondence: wangxiao@zzuli.edu.cn

Abstract: Current research widely acknowledges that the subcellular localization of mRNA is crucial for understanding its biological functions. However, current methods for mRNA subcellular localization based on k-mer frequency features may overlook the sequential information of the sequence, and a single encoding method may not adequately extract the sequence's features. This paper proposes a novel deep learning prediction method, CSpredR, specifically designed for predicting the subcellular localization of multi-site mRNAs. Unlike previous methods, CSpredR first employs k-mer to tokenize the mRNA sequences, then converts the tokenized sequences into de Bruijn graphs, thereby enabling a more precise capture of the structural information within the sequences. To mitigate the impact of lost sequential information and better capture sequence features, we combine word2vec and fasttext models to extract the features of each node in the graph and retain the sequence order. They can encode the k-mer units in the sequence into word vectors, thus serving as the node feature vectors of the graph. In this way, each node in the graph is assigned a feature vector containing rich semantic information. Subsequently, we utilize multi-scale convolutional neural networks and bidirectional long short-term memory networks to capture sequence features, respectively, and fuse the results as input for a multi-head attention mechanism model. The information from these heads is integrated into the node representations, and finally, the attention-processed data are fed into an MLP (Multi-Layer Perceptron) for prediction tasks. Extensive experiments reveal that CSpredR achieves a 2% improvement over the best existing predictors, offering a more effective tool for mRNA subcellular localization prediction.

Keywords: mRNA; subcellular localization; multi-label prediction; word2vec method; fasttext method; graph construction; deep learning; sequence-based prediction

check for updates

Academic Editors: Carla Piazza and Roberto Pagliarini

Received: 17 December 2024
Revised: 23 January 2025
Accepted: 24 January 2025
Published: 26 January 2025

Citation: Wang, X.; Suo, W.; Wang, R. CSpredR: A Multi-Site mRNA Subcellular Localization Prediction Method Based on Fusion Encoding and Hybrid Neural Networks. *Algorithms* **2025**, *18*, 67. https://doi.org/10.3390/a18020067

1. Introduction

The functionality of biomolecules within a cell largely depends on their localization within specific cellular compartments [1]. In other words, the role of biomolecules in a cell is closely related to their specific locations, as this directly affects their interactions with other molecules, thereby executing specific biological functions. The subcellular localization of RNA is of significant importance in the regulation of cellular functions, disease occurrence, and treatment. In particular, messenger RNA (mRNA), as a crucial participant in the protein-coding process, not only directly influences protein synthesis through its intracellular localization but is also closely linked to various biological processes such as cell signaling, metabolic regulation, and cell cycle modulation [2]. Studies have shown that different types of RNA molecules can be enriched in their respective specific cellular compartments, forming functional RNA-protein complexes (RNPs) [3,4]. These complexes play critical roles in regulating gene expression, maintaining cellular homeostasis, and responding to environmental changes [5]. Additionally, abnormal RNA localization is closely associated with the occurrence of various diseases. For instance, in several pathological conditions such as neurodegenerative diseases and cancer, incorrect RNA localization can lead to abnormal gene expression and protein synthesis disruption, thereby exacerbating disease progression [6–9]. Therefore, in-depth research into the mechanisms of RNA subcellular localization not only helps to uncover the fundamental principles of cell biology but may also provide new insights and methods for disease prevention, diagnosis, and treatment.

In recent years, the field of biology has successfully applied machine learning and deep learning technologies to solve numerous problems. In 2019, Yan et al. developed the first deep neural network-based mRNA localization prediction model, RNATracker [10]. This model integrated cutting-edge technologies, such as Convolutional Neural Networks (CNN) [11], Long Short-Term Memory networks (LSTM) [12], and attention mechanisms, and provided a method for detecting candidate sequences by masking 100 nt sequences at a time to evaluate their impact on the prediction. In 2020, Zhang et al. utilized binomial distribution and one-way ANOVA to identify the optimal nonamer combinations for effectively representing mRNA sequences [13]. Based on these combinations, they developed an SVM-based prediction model to accurately identify mRNA subcellular localization. This study emphasized the statistical selection of sequence features, ensuring that the model was grounded in biologically meaningful representations of mRNA sequences. In the same year, Garg et al. introduced a novel machine learning-based prediction model, mRNALoc [14], which utilized mRNA primary sequence information and SVM to predict its distribution across five subcellular locations. Their study also included the development of user-friendly software and a web server, enabling convenient access for researchers. This model provided an efficient and practical tool for subcellular localization prediction. Li et al. developed the SubLocEP prediction model [15], employing a two-layer ensemble prediction method to improve accuracy in predicting the subcellular locations of sequence samples. This method combined the outputs of multiple base classifiers, leveraging ensemble learning to enhance the model's robustness and prediction reliability. Tang et al. developed mRNALocater [16], a prediction model leveraging advanced machine learning techniques to predict mRNA localization in subcellular compartments. This model provided enhanced precision and efficiency, focusing on improving the accuracy of localization predictions for diverse compartments like the nucleus, cytoplasm, mitochondria, and endoplasmic reticulum. Wang et al. introduced DeepmRNALoc [17], a model that applies deep learning technologies to precisely predict mRNA localization in various subcellular compartments. By leveraging neural networks, this model demonstrated significant improvements in predictive performance and offered new opportunities for RNA localization research. Wang et al. proposed the DM3Loc model [18], focusing on multi-label mRNA subcellular localization prediction. This model addressed the challenge of predicting mRNA that localizes to multiple subcellular compartments simultaneously, introducing innovative strategies for multi-label learning in biological applications. Bi et al. introduced the Clarion model [19], another notable tool for multi-label mRNA localization prediction. The Weighted Series (WS) method was introduced as the ensemble framework for Clarion. This approach incorporates prior information about the labels during the model training process and improves prediction performance by optimizing the weight parameter, balancing the contributions of the non-label module and the fusion-label module. Currently, only two prediction models, DM3Loc by Wang et al. and Clarion by Bi et al., consider the possibility that mRNA may simultaneously localize to multiple subcellular locations.

Although some prediction models and computational methods have emerged in the related field, there are still many shortcomings in understanding the precise location and function of mRNA, especially in multi-site prediction. Additionally, traditional feature extraction methods are not effective in capturing and representing sequence information. To address these issues, this paper proposes a novel multi-site prediction model called CSpredR. This model converts sequences into graphs and uses a combination of two encoding methods to extract features from each graph node. It then employs a three-layer neural network model to extract high-level features from the graph. Finally, it uses an MLP to process the feature data and complete the prediction task. We conducted extensive experiments to evaluate the performance of CSpredR. By using different encoding methods and network structures, we demonstrated the effectiveness of CSpredR in predicting mRNA subcellular localization. CSpredR has two major advantages: (1) Using two encoding methods to extract features from graph nodes enriches the semantic information of the graph. (2) CNN effectively extracts local features, while Bi-LSTM excels at capturing long-term dependencies and global context information. Combining these two allows the model to focus on local details while considering the global context, leading to a more accurate understanding of the sequence data. By using the outputs of CNN and Bi-LSTM as inputs to the multi-head attention mechanism, the model can further integrate these features and generate a more robust and rich representation.

2. Materials and Methods

2.1. Datasets

In this study, we collected all the required subcellular localization annotations and mRNA sequence datasets from the RNALocate database. In the RNALocate database (version 2.0), this version provides more accurate localization annotations than the first version, facilitating the construction of a reliable benchmark dataset. If the experimental data in a paper confirms the specific localization of an mRNA, then this localization information will be annotated and assigned to the mRNA. If an mRNA has multiple localization annotations, these annotations will be merged, enabling the mRNA to have multiple labels. These datasets are consistent with those used by the Clarion method [19]. Specifically, we used 84,972 mRNA subcellular localization records from the RNALocate database as our initial dataset. After initial statistics, we found that these data covered 150 different subcellular localization types. Given the small number of entries and incomplete information for some subcellular localization types, we decided to exclude those with fewer than 3000 entries. After this screening, we obtained nine major unique transcripts, including exosomes, nuclei, cytoplasm, chromatin, nuclear matrix, ribosomes, nucleoli, and cell membranes. To reduce the impact of sequence redundancy on classifier performance, we further processed these transcripts using the CD-HIT-EST tool [20], setting an 80% sequence similarity threshold to reduce redundancy. This step ensured that the similarity between any two nucleotide sequences was below 80%. Finally, we obtained 36,971 mRNAs as the benchmark dataset for subsequent experimental analysis. The distribution of our dataset is shown in Table 1.

Table 1. Statistics on the distribution of mRNA in nine subcellular locations in the dataset.

Subcellular Sites	Sample Size
cytoplasm	4016
nucleus	21,439
ribosome	8680
Exosome	31,448
Nucleoplasm	14,237
chromatin	14,328
nucleolus	11,124
Cytosol	16,312
membrane	6739

Additionally, this study categorized the sequences based on the number of subcellular localization tags annotated for each sequence, as shown in Figure 1. Specifically, among the 36,971 mRNA sequences, there are 12,884 mRNA sequences with a single tag, 4060 mRNA sequences with two tags, 3442 mRNA sequences with three tags, 3165 mRNA sequences with four tags, 3518 mRNA sequences with five tags, 4258 mRNA sequences with six tags, 4079 mRNA sequences with seven tags, 1443 mRNA sequences with eight tags, and 122 mRNA sequences with all nine tags.

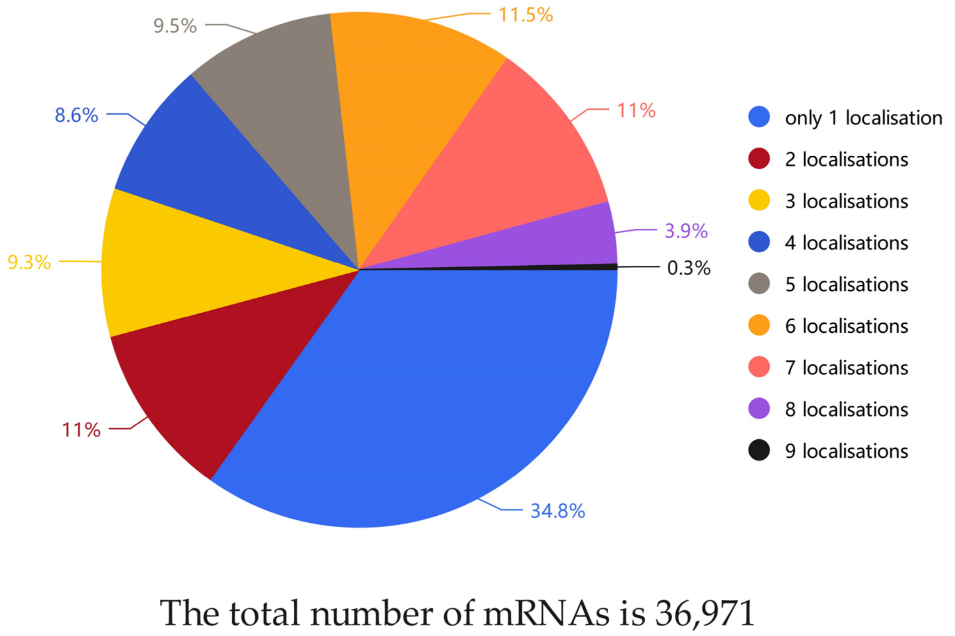

Figure 1. The relative percentages of mRNA with different labels in the benchmark dataset.

2.2. The Model Framework of CSpredR

In this study, we proposed a novel deep learning prediction model named CSpredR, which focuses on multi-site mRNA subcellular localization prediction. Figure 2 illustrates the overall framework of CSpredR. The main idea of CSpredR is to convert mRNA sequences into graphs and capture high-level features from these graphs using a multi-layer deep neural network structure. The prediction model comprises three main modules: feature encoding, model, and prediction. In the feature encoding module, the process of constructing the graph is included, which involves converting mRNA sequences into weighted de Bruijn graphs. The node feature extraction part mainly uses word2vec and fasttext techniques to generate features for each graph node. In the model part, we combined multi-layer deep neural network structures such as multi-scale convolutional neural networks, bidirectional long short-term memory networks, and multi-head attention mechanism models to capture high-level features of the graph structure. In the prediction module, we added a multi-layer perceptron (MLP), and for the output layer, we used the sigmoid activation function as a multi-label classifier to achieve multi-label classification.

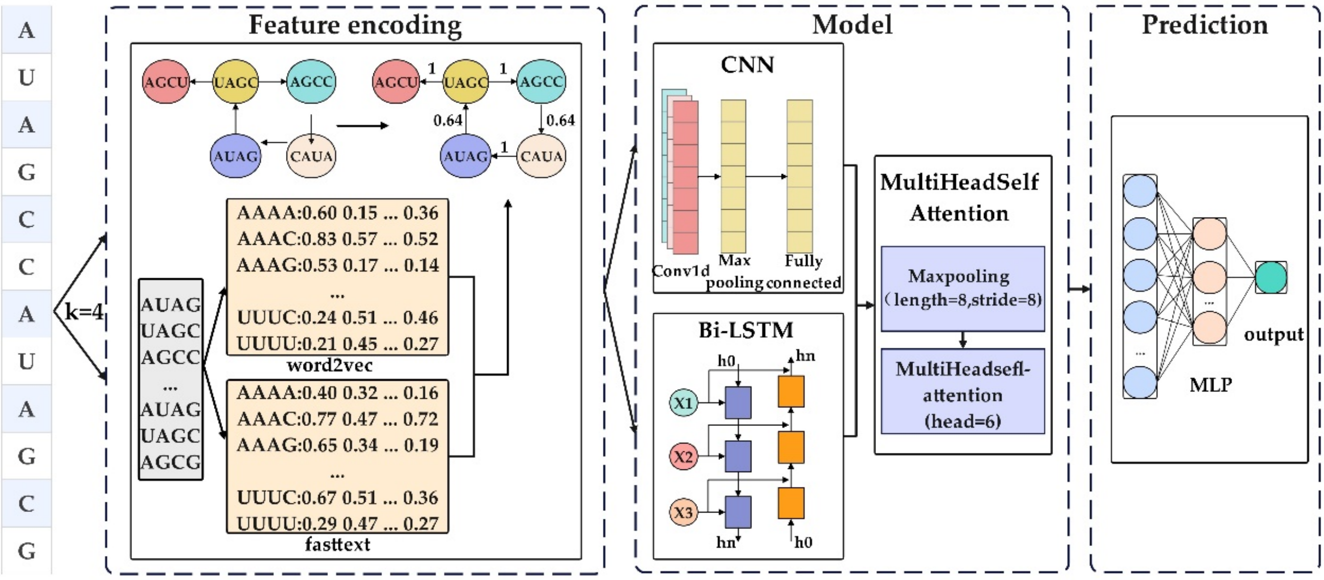

Figure 2. The framework of CSpredR consists of three parts: feature encoding, model, and prediction. First, the sequences are converted into de Bruijn graphs, and then two encoding methods are used to perform feature encoding for the graph nodes. In the model part, the graph features are separately input into CNN and Bi-LSTM to extract sequence features. The extracted sequence features are then fused and concatenated to serve as the input for the multi-head attention mechanism for further processing. Finally, multi-label classification predicting mRNA subcellular localization is achieved through an MLP combined with the sigmoid activation function.

2.3. Graph Construction

In the application of graph convolutional networks, constructing the graph is a crucial step [21]. During the graph construction process, the mRNA sequence is transformed into a directed graph. Given an mRNA sequence:

$$mRNA = R_1, R_2, R_3, \ldots, R_{L-1}, R_L \tag{1}$$

where L is the length of the sequence, Ri represents one of the four nucleotide bases (A, C, G, U) at position i of the mRNA sequence. To extract meaningful features from mRNA sequences, we employed k-mer representations, where each sequence is divided into overlapping substrings of length k. For example, a sequence “AUGCUG” with k = 3 is represented as [“AUG”, “UGC”, “GCU”, “CUG”]. By calculating the frequency of each k-mer across the sequence, we generated feature vectors that capture the sequence’s local compositional patterns. This method provides a computationally efficient way to represent sequence data while retaining biologically relevant information. Building upon this approach, a fixed window of length k is chosen, which slides along the mRNA sequence, extracting subsequences of length k each time, known as k-mer fragments. Therefore, the set of k-mers (using 4-mer as an example here) is $\{R_1R_2R_3R_4, R_2R_3R_4R_5, \ldots, R_{L-3}R_{L-2}R_{L-1}R_L\}$. Next, using these k-mer fragments, we construct a graph where each k-mer fragment corresponds to a node in the graph. Additionally, if two adjacent k-mer fragments in the sequence have matching prefixes and suffixes, a directed edge is added between these two nodes to represent the connectivity of k-mer fragments in the sequence. Ultimately, the original mRNA sequence is transformed into a de Bruijn graph, where nodes represent k-mer fragments and directed edges reveal the connections between these fragments. Then, we assign a weight to each directed edge. Following the approach used in GraphLncLoc [22], the weight of the k-th edge is determined by the frequency of the (k + 1)-mer, primarily influenced by the two nodes composing this edge. To mitigate the impact of absolute differences in edge frequencies, we normalize the edge weights in the graph. Formally, e_{ji} denotes the frequency weight of the edge from node j to node i, and $N(i)$ represents the set of neighboring nodes of node i. We normalize the frequency weights using the following formula:

$$Wnorm = \frac{e_{ji}}{\sqrt{\sum_{q \in N(j)} e_{jq} \sum_{q \in N(i)} e_{qi}}} \tag{2}$$

2.4. Feature Encoding

This study employs two techniques, word2vec [23] and fasttext [24], for feature encoding of mRNA sequences. Word2vec is a method used to learn word vector representations from large-scale text data, transforming words into vectors to capture semantic relationships between them in a high-dimensional space. Fasttext, on the other hand, learns word embeddings by considering character n-grams within words and uses these vector representations of n-grams to construct word vectors. This approach captures internal structures of words, thereby enhancing the model's generalization ability. The CSpredR model utilizes K-mer segmented fragments as features for the nodes in the de Bruijn graph. These K-mer fragments can be analogized to individual words, and directed edges naturally represent relationships between these "words". Thus, we view all mRNA sequences in the dataset as a corpus, where each mRNA sequence represents a sentence in this corpus. Furthermore, we consider all possible 4^k K-mer fragments as the vocabulary for constructing these sentences. We then encode these K-mer fragments using word2vec and fasttext techniques separately and fuse them as features for the nodes in the graph. **We chose weighted averaging as the fusion method for these two encoding approaches because it preserves their complementary characteristics while avoiding the computational complexity and overfitting issues associated with high-dimensional concatenation, thereby further optimizing feature representation and computational efficiency.** Experiments indicate that combining these two methods maximally enriches the semantic information of the de Bruijn graph.

2.5. Convolutional Neural Network Method

Traditional Convolutional Neural Networks (CNNs) excel in processing two-dimensional image data [25]. In the context of TextCNN, text is treated as a special type of one-dimensional sequence, analogous to a one-dimensional image. This means we can use one-dimensional CNNs to extract features from text sequences. TextCNN employs one-dimensional convolutional layers and max-pooling layers to extract sequence features. Specifically, if we have n sub-sequences, each represented as D-dimensional vectors, they can be combined into an n × D matrix to represent the entire sequence. For mRNA sequences, their representation can be conceptualized as a one-dimensional image with width n, height 1, and d channels. To extract advanced features, we use three convolutional kernels (sizes = 1, 3, 5) in TextCNN to capture the correlations between adjacent nucleotides. These kernels operate across all channels and are followed by a max-pooling layer to extract the most significant features and reduce the dimensionality of the output vectors. Finally, the output vectors from the max-pooling layer are concatenated together and serve as inputs to a fully connected layer, resulting in a processed feature vector.

2.6. Bidirectional Long Short-Term Memory Method

To capture long-term dependencies and sequence information in text, this study also employs Bidirectional Long Short-Term Memory networks (Bi-LSTM). Bi-LSTM consists of two LSTM (Long Short-Term Memory) layers: one processes the text sequence in the forward direction, and the other processes it in the backward direction, thereby capturing both forward and backward contextual information simultaneously. We set the number of hidden states in the Bi-LSTM units to 128 and feed the word embedding vectors of the text data as inputs to the Bi-LSTM layers. As information passes through each layer, the LSTM units update their internal states at each time step to capture relationships between the current word and its preceding and succeeding words. By stacking multiple LSTM layers, we can capture more complex sequence patterns. Finally, we apply an average pooling over the entire sequence output to obtain a fixed-length feature vector, which represents the encoded form of the text sequence.

2.7. Synergistic Model: CNN and Bi-LSTM for Capturing Sequence Features

After processing the text data separately through CNN and Bi-LSTM, we obtain two different sources of feature representations. CNN captures local and hierarchical features of the text, while Bi-LSTM captures long-term dependencies and sequence information within the text. To enrich the semantic information in the feature vector, we employ a feature fusion and concatenation strategy. Assuming CNN transforms the data into a 2D tensor after max pooling, with a shape of (batch_size, num_filters), and Bi-LSTM outputs a 2D tensor with a shape of (batch_size, hidden_size), we can directly concatenate these two 2D tensors. This results in a more comprehensive feature vector with a shape of (batch_size, num_filters + hidden_size). This concatenated vector contains rich information from both CNN and Bi-LSTM, providing a powerful input for subsequent attention mechanisms.

2.8. Multi-Head Attention

The attention mechanism was originally proposed to enhance the model's ability to focus on key parts of input data, particularly in machine translation tasks. It has been widely applied in fields such as image processing and natural language processing and has shown significant effectiveness in extracting important information from input data [26]. In our study, we have obtained feature vectors through feature fusion and concatenation, but these vectors may not be enough for direct use in prediction tasks. Therefore, we introduced a multi-head attention mechanism to further process the fused and concatenated feature vectors. Through the multi-head self-attention mechanism, we can parallelly capture dependencies at different positions in the sequence, thereby enhancing the model's representation capability. In bioinformatics, attention mechanisms are often used in conjunction with structures like RNNs and have been demonstrated to achieve competitive performance in a wide range of biological sequence analysis problems [27]. Therefore, we chose to apply the attention mechanism to identify crucial information for predicting mRNA subcellular localization. We use the higher-level fused features learned from the Bi-LSTM and CNN layers as inputs to the attention layer. Through the attention mechanism [28,29], we assign different weights to these higher-level features to highlight the most critical influences, thereby helping the model more accurately predict the 9 subcellular locations of mRNA.

2.9. Performance Evaluation Metrics

In the context of multi-label mRNA subcellular localization prediction studies, the diversity of the multi-label space compared to the single-label space leads to increased complexity in evaluating multi-label learning algorithms. Therefore, to comprehensively assess the learning capability of multi-label models, we need to employ multiple evaluation metrics. We have selected six key evaluation metrics to assess the performance of the model: Hamming loss, One-error, Accuracy, Coverage, Average precision, and Ranking loss [30,31]. Accuracy reflects the proportion of correctly predicted samples by the model and is suitable for measuring overall classification performance; Average precision combines precision and recall, making it particularly suitable for imbalanced datasets; One-error focuses on whether the most relevant label predicted by the model is in the true label set, with lower values indicating better performance; Ranking loss evaluates the accuracy of the model's label ranking, with smaller values indicating more reasonable ranking; Hamming loss measures the proportion of incorrectly predicted labels in multi-label classification tasks, with lower values indicating better model performance. These metrics collectively provide a multi-dimensional evaluation of model performance, catering to different task requirements and data characteristics. For a test set S = {(x_1, Y_1), (x_2, Y_2), ..., (x_p, Y_p)} containing multiple samples, we will use the following evaluation formulas to comprehensively assess the model's performance.

$$Hamming\ Loss = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{q}|R_i \triangle Y_i| \tag{3}$$

$$One - error = \frac{1}{p}\sum_{i=1}^{p}\left\{\left[arg_{y'\in y_i} max f\left(x_i, y'\right)\right] \in Y_i\right\} \tag{4}$$

$$Accuracy = \frac{1}{p}\sum_{i=1}^{p}\frac{R_i \cap Y_i}{R_i \cup Y_i} \tag{5}$$

$$Coverage = \frac{1}{p}\sum_{i=1}^{p} max_{y'\in Y_i} rank_f\left(x_i, y'\right) - 1 \tag{6}$$

$$Average\ precision = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i|}\frac{\left|\left\{y'\middle|rank_f(x_i, y') \le rank_f(x_i, y), y' \in Y_i\right\}\right|}{rank_f(x_i, y)} \tag{7}$$

$$Ranking\ loss = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i|\left|\overline{Y}_i\right|}\left|\left\{(y_1, y_2)|f(x_i, y_1) \le f(x_i, y_2), (y_1, y_2) \in Y_i \times \overline{Y}_i\right\}\right| \tag{8}$$

where Δ represents the symmetric difference between two sets, f(·) is the multi-label classifier, $rank_f$ represents the rank of y in Y based on the descending order, q is the cardinality of Y_i, Y_i is the complement of $\overline{Y}_i$, Y_i and R_i are, respectively, the true label set and the predicted label set for a sequence.

2.10. Hyperparameter Optimization

In this study, there are many hyperparameters that influence the model's performance, such as the k value of the k-mer nodes, the dimensions of the pre-trained word2vec and fasttext embedding vectors, the number of hidden neurons in the CNN and Bi-LSTM, and more. These hyperparameters play important roles in constructing the de Bruijn graph, building feature representations, and adjusting the model's complexity. Using a grid search strategy, we tried different combinations of hyperparameters and selected the best combination based on the model's performance. Specifically, we experimented with k values for k-mer nodes in the range of {2, 3, 4, 5, 6}, dimensions of pre-trained word2vec and fasttext embedding vectors in the range of {32, 64, 128, 256}, and the number of hidden neurons in the CNN and Bi-LSTM in the range of {32, 64, 128, 256}.

Additionally, we paid special attention to the number of attention heads in the multi-head attention mechanism, a critical parameter that significantly affects the model's expressive power and complexity. By adjusting the number of attention heads, we can more effectively capture the complex relationships between nodes, thereby improving the model's performance. Therefore, we experimented with different numbers of attention heads, specifically {2, 4, 6, 8}, and evaluated the model's performance under each.

By comparing the model performance under different hyperparameters and numbers of attention heads, we successfully determined the optimal parameter combination.

Through grid search [32], we found that the number of attention heads in the multi-head attention mechanism significantly impacts model performance. During the hyperparameter adjustment process, we discovered that the model achieves optimal performance when the K-mer node value is set to 4, the dimensions of the pre-trained word2vec and fasttext embedding vectors are set to 64, the number of hidden neurons in both CNN and Bi-LSTM is 128, and the number of attention heads is 6.

3. Results and Discussion

3.1. Comparison of Different k-mer Features

In this study, we used a 5-fold cross-validation method to reliably evaluate the performance of our CSpredR predictor [33]. Specifically, the benchmark dataset was divided into five equally sized subsets. Four of these subsets were used for model training, while the remaining subset was used as test data for model performance evaluation. This process was repeated five times, with each subset being selected as the test data in turn. In our experiments, we found that the value of k for k-mers had a significant impact on the results. To determine the most suitable k-mer value, we conducted a series of experiments using different k-mer lengths: k = 1, 2, 3, 4, 5, and 6.

The experimental results, as shown in Figure 3, indicate that the model's prediction performance is relatively poor with smaller k values, such as k = 1, k = 2, and k = 3. This may be because these shorter k-mers fail to fully capture the critical information in the RNA sequences, leading to lower model accuracy. However, the best prediction performance was observed with k = 4, suggesting that a k-mer length of 4 can better capture the contextual information in RNA sequences, thereby improving model accuracy. Consequently, we chose k = 4 as the optimal k-mer value for further analysis.

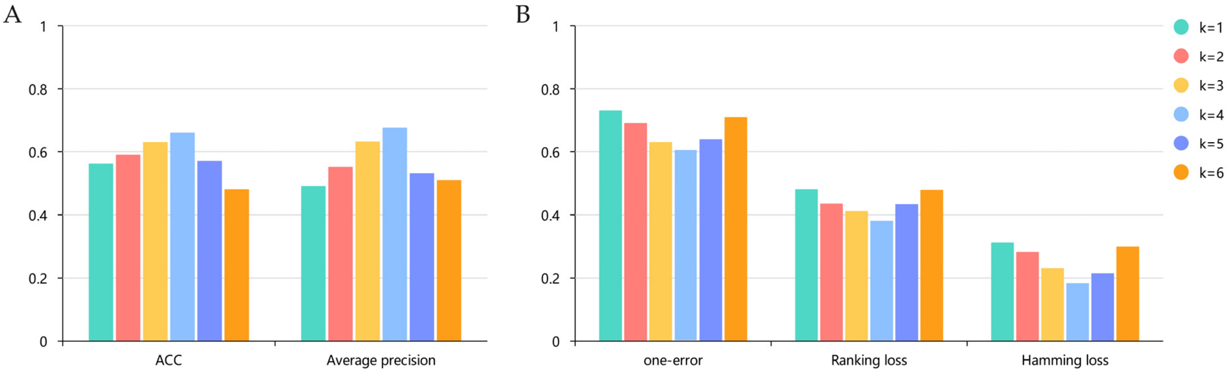

Figure 3. Experimental results with different k values. (**A**) Performance evaluation of different k-values in terms of ACC and Average precision. (**B**) Performance evaluation of different k-values in terms of one-error, Ranking loss and Hamming loss.

It is noteworthy that using longer k-mer values, such as k = 5 and k = 6, did not further improve the model's prediction performance. This might be because longer k-mers could introduce noise or excessive features, reducing the model's generalization ability. Choosing k = 4 is a trade-off that allows for effective capture of contextual information in RNA sequences while controlling the complexity of the de Bruijn graph to ensure computational efficiency. It also reduces the information insufficiency caused by smaller k values and the noise and cost introduced by larger k values. Therefore, k = 4 is a reasonable and efficient choice. In summary, in this study, we ultimately determined that a k-mer length of k = 4 is the optimal choice for mRNA subcellular localization prediction based on word2vec and fasttext encoding. The research results also indicate that when k = 4, the k-mer length performs best in balancing information capture and feature dimension control [34].

3.2. Ablation Experiment

CSpredR consists of three parts, with the first two modules using different encoding methods and neural network models for feature extraction. To verify the effectiveness of certain structures proposed in the CSpredR model, we conducted ablation experiments on the first two modules separately. In the encoding part, we compared the performance of word2vec and fasttext individually and their combined encoding results while keeping the neural network model module unchanged. As shown in Table 2, we found that fasttext's performance was not as good as word2vec when used alone—this could be attributed to the limitations of using fastText encoding alone compared to word2Vec. While fastText excels at handling out-of-vocabulary words by breaking them into subword units, it struggles to effectively capture word order information, which is critical for sequence-based tasks. In contrast, word2Vec directly encodes semantic relationships between words in a way that better preserves contextual information, making it more suitable for tasks requiring fine-grained sequence representation. but the performance improved when the two were combined. This may be because the two methods complement each other in learning word vectors and sequence representation, resulting in richer text representations. Additionally, we performed Wilcoxon tests to evaluate the predictive results of different encoding methods. As shown in Table 3, CSpredR significantly outperforms other methods across all evaluation metrics ($p < 0.05$). Notably, CSpredR shows significant advantages in Accuracy and One-error, demonstrating higher accuracy and better label selection capability, with other metrics such as Hamming-loss, Coverage, Average precision, and Ranking loss also exhibiting relative superiority.

Table 2. Ablation experimental results based on different encoding methods.

	Hamming Loss	One-Error	ACC	Coverage	Average Precision	Ranking Loss
Word2vec	0.219	0.659	0.625	6.211	0.650	0.445
Fasttext	0.274	0.716	0.588	6.279	0.648	0.482
CSpredR	0.182	0.605	0.657	6.035	0.675	0.380

Table 3. Wilcoxon test results comparing the performance of different encoding methods for subcellular localization prediction.

	Word2vec	Fasttext
p-value of Hamming loss	0.03125	0.00484
p-value of One-error	0.03659	0.03496
p-value of ACC	0.03805	0.03778
p-value of Coverage	0.00272	0.00278
p-value of Average precision	0.04256	0.04231
p-value of Ranking loss	0.03805	0.04121

In the neural network model part, we compared the performance of five combinations of network model structures while keeping the two encoding methods unchanged: CNN, CNN + attention, Bi-LSTM, Bi-LSTM + attention, and CSpredR. The results are shown in Table 4. We found that using CNN alone yielded unsatisfactory results, possibly because CNN only focuses on local information in the text. After adding the multi-head self-attention mechanism, the model's performance improved, likely due to the model's enhanced ability to process key information. However, Bi-LSTM outperformed CNN because it can capture long-term dependencies in the text, providing a more comprehensive text representation. When attention mechanisms were integrated into Bi-LSTM, the model's performance improved significantly. The attention mechanisms allowed the model to focus on the most relevant parts of the sequence, enhancing both predictive power and interpretability. By combining Bi-LSTM's sequence modeling capabilities with attention mechanisms, CSpredR effectively leverages the strengths of both approaches. Finally, CSpredR demonstrated the highest performance, combining the advantages of both CNN and Bi-LSTM. It can capture local features of the text and handle long-term dependencies, while the multi-head attention mechanism captures dependencies in the sequence from multiple perspectives, resulting in better performance [35]. This combination leverages the strengths of different components, ensuring a balanced approach to capturing both local and global information, which is essential for complex sequence prediction tasks. Subsequently, we conducted Wilcoxon tests on the predictive results of different network structures. As shown in Table 5, CSpredR also significantly outperforms other methods ($p < 0.05$). Particularly in Accuracy and Ranking loss, CSpredR exhibits higher prediction precision and better label ranking ability, further validating its superiority.

Table 4. Ablation experimental results based on different network model structures.

	Hamming Loss	One-Error	ACC	Coverage	Average Precision	Ranking Loss
CNN	0.322	0.710	0.437	7.496	0.496	0.517
Bi-LSTM	0.289	0.656	0.512	6.942	0.530	0.471
CNN + attention	0.276	0.671	0.525	6.828	0.579	0.429
Bi-LSTM + attention	0.235	0.632	0.578	6.440	0.646	0.402
CSpredR	0.182	0.605	0.657	6.035	0.675	0.380

Table 5. Wilcoxon test results for performance comparison across different network structures in subcellular localization prediction.

	CNN	Bi-LSTM	CNN+ Attention	Bi-LSTM+ Attention
p-value of Hamming loss	0.01354	0.01964	0.01978	0.02169
p-value of One-error	0.03141	0.03783	0.04918	0.04571
p-value of ACC	0.03192	0.03794	0.04126	0.04780
p-value of Coverage	0.00429	0.01468	0.02779	0.03014
p-value of Average precision	0.03837	0.03994	0.04108	0.04296
p-value of Ranking loss	0.02711	0.02846	0.02971	0.03249

3.3. Comparison with Other Single Label Multi-Class Classification Methods

To verify the superior performance of the CSpredR model, we made key adjustments to the classification part while ensuring that the feature extraction and model construction mechanisms of the CSpredR model remained unchanged. Specifically, we changed the originally designed multi-label classification task model structure to a single-label multi-class classification task. This modification allowed us to directly compare the CSpredR model with existing single-label multi-class models.

By using the single-label multi-class task as an evaluation metric, we could more clearly understand the performance of the CSpredR model in different application scenarios and directly compare it with other popular multi-label classification models. For this purpose, we constructed a single-label classification benchmark dataset by selecting 12,884 mRNA sequences from the 36,971 entries derived after preprocessing the initial dataset. These sequences were further refined to ensure consistency and suitability for single-label classification tasks. Using a 5-fold cross-validation setup, we trained the model on four subsets and tested it on the remaining one, ensuring a robust and reliable evaluation. This helped us to comprehensively evaluate the applicability and performance differences in the CSpredR model in single-label, multi-label, and various task contexts.

We used accuracy, precision, recall, and F1 score as evaluation metrics to assess the classification performance, prediction accuracy, and performance in multi-class problems of the CSpredR model. Table 6 shows the comparison of CSpredR based on an independent test set with other predictors. The results indicate that the CSpredR model performed excellently in terms of accuracy, reaching 0.671, surpassing other predictors. Its precision was 0.755, recall was 0.592, and F1 score was 0.643, also showing significant advantages in multi-class problems. This demonstrates that the CSpredR model not only has a competitive edge in prediction accuracy but also exhibits excellent performance in handling multi-class problems.

Table 6. Performance comparison of the CSpredR predictor based on the independent test set with other methods.

Predictors	ACC	Precision	Recall	F1 Score
lncLocator	0.421	0.374	0.325	0.289
iLoc-lncRNA	0.509	0.524	0.470	0.474
Locate-R	0.368	0.362	0.321	0.321
GraphLncLoc	0.579	0.736	0.557	0.584
CSpredR	0.671	0.755	0.592	0.643

This paper further compares the performance of CSpredR in single-label prediction with various advanced methods, all of which are accessible via web servers, including iLoc-mRNA, mRNALoc, mRNALocator, and DM3loc. To ensure fairness in the comparison, mRNA sequences from the independent test set were submitted to each server to obtain their predicted subcellular localization results, which were then compared with the true labels. Detailed experimental results are presented in Table 7.

Table 7. Performance comparison of the CSpredR predictor on single subcellular localization based on an independent test set compared to other methods.

	iLoc-mRNA	mRNALoc	mRNALocator	DM3Loc	Clarion	CSpredR
chromatin	--	--	--	--	81.47%	81.50%
cytoplasm	--	54.88%	38.90%	--	91.29%	94.62%
Cytosol	--	--	--	57.37%	79.77%	83.55%
Exosome	--	--	--	70.00%	92.10%	95.46%
membrane	--	--	--	70.92%	89.15%	91.12%
nucleolus	--	--	--	--	83.74%	83.88%
Nucleoplasm	--	--	--	--	80.74%	81.20%
nucleus	--	55.18%	57.42%	69.52%	79.23%	80.06%
ribosome	73.41%	--	--	69.03%	84.74%	86.42%

In predicting key subcellular locations such as cytoplasm, cytosol, exosome, membrane, nucleus, and ribosome, CSpredR demonstrated significant performance advantages, with its prediction accuracy notably surpassing that of other methods. This further confirms the outstanding effectiveness and reliability of CSpredR in mRNA subcellular localization prediction. Notably, in prediction tasks involving nine subcellular locations, CSpredR achieved accuracy rates exceeding 80%. These comprehensive comparative results once again highlight the exceptional performance of CSpredR in single-site prediction tasks.

3.4. The Comparison of CSpredR with Other Prediction Models

To comprehensively evaluate the performance of CSpredR, we conducted a comparative analysis with existing mRNA subcellular localization prediction methods, DM3Loc and Clarion. As shown in Figure 4, CSpredR demonstrates superior performance across multiple key metrics. Firstly, CSpredR excels in accuracy, indicating its enhanced precision in predicting mRNA subcellular localization. This advantage stems from the model's optimized design and training process, further validating its efficiency in this field. Secondly, CSpredR outperforms other methods in Average precision. The high Average precision reflects its ability to accurately distinguish true localizations from incorrect predictions, thereby improving the reliability and practicality of the results.

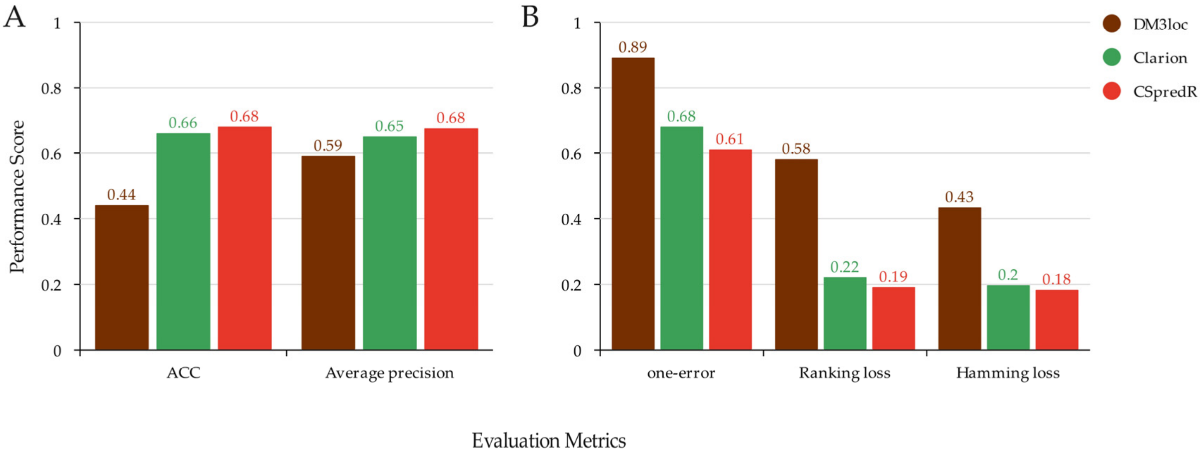

Figure 4. Performance comparison of prediction models, (**A**) Performance evaluation of the model in terms of ACC and Average precision. (**B**) Performance evaluation of the model in terms of one-error, Ranking loss and Hamming loss.

Additionally, CSpredR also shows advantages in metrics such as One-error, Ranking loss, and Hamming loss. Its low One-error value indicates a reduced misclassification rate, while its performance in Ranking loss and Hamming loss highlights its stability and accuracy in multi-label classification and true label prediction tasks.

In summary, compared to DM3Loc and Clarion, CSpredR demonstrates significant performance advantages in accuracy, Average precision, and other key metrics. These results not only confirm the reliability and efficiency of CSpredR but also support its application in bioinformatics and biomedical research.

4. Conclusions

This study presents an innovative deep learning predictor, CSpredR, specifically designed for multi-site mRNA subcellular localization prediction. The method utilizes graph construction combined with two encoding techniques to perform fusion encoding on graph nodes, which are then input into a hybrid neural network model for feature extraction and final prediction. For feature encoding, CSpredR transforms sequences into an advanced graph structure and applies word2vec and fasttext encoding techniques to process graph nodes. This approach enhances the richness of sequence features, providing high-quality inputs for the subsequent neural network. The neural network component employs a parallel architecture with CNN and Bi-LSTM to extract key features from sequences, which are then fused and fed into a multi-head attention mechanism to capture dependencies between sequences. The multi-head attention mechanism achieves parallel computation, improving efficiency and enhancing text processing performance. Finally, the output from the MLP is mapped to multiple labels, enabling multi-label classification. Evaluation results show that CSpredR outperforms existing methods across most metrics on both benchmark and independent test datasets, demonstrating superior performance and robustness. By effectively integrating advanced techniques and neural network models, CSpredR has made notable progress in the field of mRNA subcellular localization. However, we believe there is still room for improvement. In the future, we will focus on developing new tools to further enhance the accuracy of mRNA subcellular localization prediction, aiming for a deeper understanding and more precise prediction of mRNA functions and roles.

Author Contributions: Conceptualization, W.S.; methodology, W.S.; supervision, X.W. and R.W.; validation, X.W.; writing—original draft, W.S.; writing—review and editing, X.W. and R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by funds from the Key Research Project of Colleges and Universities of Henan Province (No. 22A520013, No. 23B520004), the Key Science and Technology Development Program of Henan Province (No. 232102210020, No. 202102210144), and the Training Program of Young Backbone Teachers in Colleges and Universities of Henan Province (No. 2019GGJS132).

Data Availability Statement: The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare there are no conflicts of interest.

References

1. Di Liegro, C.M.; Schiera, G.; DI Liegro, I. Regulation of mRNA transport, localization and translation in the nervous system of mammals. *Int. J. Mol. Med.* **2014**, *33*, 747–762. [CrossRef] [PubMed]
2. Meyer, C.; Garzia, A.; Tuschl, T. Simultaneous detection of the subcellular localization of RNAs and proteins in cultured cells by combined multicolor RNA-FISH and IF. *Methods* **2017**, *118–119*, 101–110. [CrossRef] [PubMed]
3. Liu-Yesucevitz, L.; Bassell, G.J.; Gitler, A.D.; Hart, A.C.; Klann, E.; Richter, J.D.; Warren, S.T.; Wolozin, B. Local rna translation at the synapse and in disease. *J. Neurosci.* **2011**, *31*, 16086–16093. [CrossRef] [PubMed]
4. O'Rourke, J.R.; Swanson, M.S. Mechanisms of RNA-mediated Disease. *J. Biol. Chem.* **2009**, *284*, 7419–7423. [CrossRef] [PubMed]
5. Chin, A.; Lécuyer, E. RNA localization: Making its way to the center stage. *Biochim. Biophys. Acta Gen. Subj.* **2017**, *1861*, 2956–2970. [CrossRef] [PubMed]
6. Zhu, Y.; Zhu, L.; Wang, X.; Jin, H. RNA-based therapeutics: An overview and prospectus. *Cell Death Dis.* **2022**, *13*, 644. [CrossRef] [PubMed]
7. Uemura, M.; Zheng, Q.; Koh, C.M.; Nelson, W.G.; Yegnasubramanian, S.; De Marzo, A.M. Overexpression of ribosomal RNA in prostate cancer is common but not linked to rDNA promoter hypomethylation. *Oncogene* **2011**, *31*, 1254–1263. [CrossRef] [PubMed]
8. Dolezal, J.M.; Dash, A.P.; Prochownik, E.V. Diagnostic and prognostic implications of ribosomal protein transcript expression patterns in human cancers. *BMC Cancer* **2018**, *18*, 275. [CrossRef] [PubMed]
9. Sprenkle, N.T.; Sims, S.G.; Sánchez, C.L.; Meares, G.P. Endoplasmic reticulum stress and inflammation in the central nervous system. *Mol. Neurodegener.* **2017**, *12*, 42. [CrossRef]
10. Yan, Z.; Lécuyer, E.; Blanchette, M. Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics* **2019**, *35*, i333–i342. [CrossRef] [PubMed]
11. Alshubaily, I. TextCNN with attention for text classification. *arXiv* **2021**, arXiv:2108.01921.
12. Jang, B.; Kim, M.; Harerimana, G.; Kang, S.U.; Kim, J.W. Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. *Appl. Sci.* **2020**, *10*, 5841. [CrossRef]
13. Meher, P.K.; Rai, A.; Rao, A.R. mLoc-mRNA: Predicting multiple sub-cellular localization of mRNAs using random forest algorithm coupled with feature selection via elastic net. *BMC Bioinform.* **2021**, *22*, 342. [CrossRef] [PubMed]
14. Garg, A.; Singhal, N.; Kumar, R.; Kumar, M. mRNALoc: A novel machine-learning based in-silico tool to predict mRNA subcellular localization. *Nucleic Acids Res.* **2020**, *48*, W239–W243. [CrossRef] [PubMed]
15. Li, J.; Zhang, L.; He, S.; Guo, F.; Zou, Q. SubLocEP: A novel ensemble predictor of subcellular localization of eukaryotic mRNA based on machine learning. *Brief. Bioinform.* **2021**, *22*, bbaa401. [CrossRef]
16. Tang, Q.; Nie, F.; Kang, J.; Chen, W. mRNALocater: Enhance the prediction accuracy of eukaryotic mRNA subcellular localization by using model fusion strategy. *Mol. Ther.* **2021**, *29*, 2617–2623. [CrossRef] [PubMed]

17. Wang, S.; Shen, Z.; Liu, T.; Long, W.; Jiang, L.; Peng, S. DeepmRNALoc: ANovelPredictor of Eukaryotic mRNA Subcellular Local-ization Based on Deep Learning. *Molecules* **2023**, *28*, 2284. [CrossRef]
18. Wang, D.; Zhang, Z.; Jiang, Y.; Mao, Z.; Wang, D.; Lin, H.; Xu, D. DM3Loc: Multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism. *Nucleic Acids Res.* **2021**, *49*, e46. [CrossRef]
19. Bi, Y.; Li, F.; Guo, X.; Wang, Z.; Pan, T.; Guo, Y.; I Webb, G.; Yao, J.; Jia, C.; Song, J. Clarion is a multi-label problem transformation method for identifying mRNA subcellular localizations. *Briefings Bioinform.* **2022**, *23*, bbac467. [CrossRef]
20. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef]
21. Musleh, S.; Arif, M.; Alajez, N.M.; Alam, T. Unified mRNA Subcellular Localization Predictor based on machine learning tech-niques. *BMC Genom.* **2024**, *25*, 151. [CrossRef] [PubMed]
22. Li, M.; Zhao, B.; Yin, R.; Lu, C.; Guo, F.; Zeng, M. GraphLncLoc: Long non-coding RNA subcellular localization prediction using graph convolutional networks based on sequence to graph transformation. *Brief. Bioinform.* **2022**, *24*, bbac565. [CrossRef] [PubMed]
23. Tsukiyama, S.; Hasan, M.M.; Fujii, S.; Kurata, H. LSTM-PHV: Prediction of human-virus protein-protein interactions by LSTM with word2vec. *Brief. Bioinform.* **2021**, *22*, bbab228. [CrossRef] [PubMed]
24. Le, N.Q.K.; Yapp, E.K.Y.; Nagasundaram, N.; Yeh, H.-Y. Classifying Promoters by Interpreting the Hidden Information of DNA Sequences via Deep Learning and Combination of Continuous FastText N-Grams. Front. *Bioeng. Biotechnol.* **2019**, *7*, 305. [CrossRef]
25. Chauhan, R.; Ghanshala, K.K.; Joshi, R.C. Convolutional neural network (CNN) for image detection and recognition. In Proceedings of the 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 15–17 December 2018; pp. 278–282.
26. Abdin, O.; Nim, S.; Wen, H.; Kim, P.M. PepNN: A deep attention model for the identification of peptide binding sites. *Commun. Biol.* **2022**, *5*, 503. [CrossRef]
27. Hong, Z.; Zeng, X.; Wei, L.; Liu, X. Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics* **2019**, *36*, 1037–1043. [CrossRef]
28. Park, S.; Koh, Y.; Jeon, H.; Kim, H.; Yeo, Y.; Kang, J. Enhancing the interpretability of transcription factor binding site prediction using attention mechanism. *Sci. Rep.* **2020**, *10*, 13413. [CrossRef] [PubMed]
29. Zou, Z.; Tian, S.; Gao, X.; Li, Y. mlDEEPre: Multi-functional enzyme function prediction with hierarchical multi-label deep learning. *Front. Genet.* **2019**, *9*, 714. [CrossRef]
30. Ghamrawi, N.; McCallum, A. Collective multi-label classification. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, Bremen, Germany, 31 October–5 November 2005.
31. Gopal, S.; Yang, Y. Multilabel classification with meta-level features. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland, 19–23 July 2010.
32. Bai, T.; Yan, K.; Liu, B. DAmiRLocGNet: miRNA subcellular localization prediction by combining miRNA–disease associations and graph convolutional networks. *Brief. Bioinform.* **2023**, *24*, bbad212. [CrossRef]
33. Quinn, J.J.; Chang, H.Y. Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* **2016**, *17*, 47–62. [CrossRef]
34. Zhang, Z.Y.; Yang, Y.H.; Ding, H.; Wang, D.; Chen, W.; Lin, H. Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief. Bioinform.* **2021**, *22*, 526–535. [CrossRef] [PubMed]
35. Quang, D.; Xie, X. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA se-quences. *Nucleic Acids Res.* **2016**, *44*, e107. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.