CellCircLoc: Deep Neural Network for Predicting and Explaining Cell line-specific CircRNA Subcellular Localization

Min Zeng, Jingwei Lu, Yiming Li, Chengqian Lu, Shichao Kan, Fei Guo, and Min Li *Member, IEEE*

***Abstract*— The subcellular localization of circular RNAs (circRNAs) is crucial for understanding their functional relevance and regulatory mechanisms. CircRNA subcellular localization exhibits variations across different cell lines, demonstrating the diversity and complexity of circRNA regulation within distinct cellular contexts. However, existing computational methods for predicting circRNA subcellular localization often ignore the importance of cell line specificity and instead train a general model on aggregated data from all cell lines. Considering the diversity and context-dependent behavior of circRNAs across different cell lines, it is imperative to develop cell line-specific models to accurately predict circRNA subcellular localization. In the study, we proposed CellCircLoc, a sequence-based deep learning model for circRNA subcellular localization prediction, which is trained for different cell lines. CellCircLoc utilizes a combination of convolutional neural networks, Transformer blocks, and bidirectional long short-term memory to capture both sequence local features and long-range dependencies within the sequences. In the Transformer blocks, CellCircLoc uses an attentive convolution mechanism to capture the importance of individual nucleotides. Extensive experiments demonstrate the effectiveness of CellCircLoc in accurately predicting circRNA subcellular localization across different cell lines, outperforming other computational models that do not consider cell line specificity. Moreover, the interpretability of CellCircLoc facilitates the discovery of important motifs associated with circRNA subcellular localization. The CellCircLoc web server is available at http://csuligroup.com:8000/cellcircloc. The source code can be obtained from https://github.com/CSUBioGroup/CellCircLoc.**

Manuscript received XXX XXX, 2022; revised XXX XX, XXX; accepted XXX XXX, 2024. Date of publication XXX XXX, 2024. This work was supported by the National Key Research and Development Program of China (No. 2022YFC3400300), the National Natural Science Foundation of China under Grants (No. 62102457), Hunan Provincial Natural Science Foundation of China under Grant (No. 2023JJ40763), Hunan Provincial Science and Technology Program under Grant (No. 2021RC4008), the Fundamental Research Funds for the Central Universities of Central South University (No. 2023ZZTS0627). This work was carried out in part using computing resources at the High Performance Computing Center of Central South University.

Min Zeng, Jingwei Lu, Yiming Li, Shichao Kan, Fei Guo, Min Li are with the School of Computer Science and Engineering, Central South University, Changsha 410083, China (e-mail: zengmin@csu.edu.cn; 224712201@csu.edu.cn; lym1998@csu.edu.cn; kanshichao@csu.edu.cn; guofei@csu.edu.cn; limin@mail.csu.edu.cn)

Chengqian Lu is with School of Computer Science, Key Laboratory of Intelligent Computing and Information Processing, Xiangtan University, Xiangtan, Hunan, 411105, China (e-mail: chengqlu@xtu.edu.cn).

*Corresponding author: Min Li

***Index Terms*— CircRNA Subcellular Localization Prediction, Cell Line Specificity, Deep Learning.**

I. Introduction

CIRCULAR RNAs (circRNAs) are a special family of non-coding RNA molecules characterized by their closed loop structure that lacks 5' or 3' ends. They have attracted significant interest in the field of RNA therapeutics [1]–[4]. CircRNAs have diverse functions and unique properties, including high stability, tissue-specific expression, and coding potential [5]–[10]. Dysregulated circRNAs play crucial roles in various aspects of cancer biology, including proliferation, metastasis, apoptosis, angiogenesis, immune evasion, and drug resistance [11]–[15].

The subcellular localization of circRNAs refers to their distribution and movement within different cellular compartments. Understanding the subcellular localization of circRNAs is crucial for unraveling their biogenesis, function, and regulation in diverse biological and pathophysiological contexts [16], [17]. Nuclear-localized circRNAs can regulate transcription and splicing processes, thereby influencing the expression of host genes or other genes [16]–[18]. Cytoplasmic-localized circRNAs can act as miRNA sponges, interact with proteins or other RNAs, and even serve as templates for translation [17], [18]. Although biological experimental methods like single-molecule fluorescent in situ hybridization (smFISH) provide accurate determinations of circRNA subcellular localization, they are limited in terms of cost, time, and scalability [19]. Computational methods can complement biological experimental methods by providing rapid predictions for large-scale datasets.

To investigate the subcellular localization of circRNAs, some databases have been developed. RNAlocateV2.0 is a comprehensive database that collects various types of RNA subcellular localization data [20]. CSCD2 is another comprehensive database, containing more than two million circRNAs from about one thousand samples [21], which is a useful resource for studying cancer-specific circRNAs. Although these databases have been developed, only a limited number of computational methods has been proposed for circRNA subcellular localization prediction. To the best of our knowledge, Circ-LocNet is the first computational method that focuses on circRNA subcellular localization prediction. It incorporates various sequence descriptors, including residue frequency-

based, residue order and frequency-based, and physicochemical property-based descriptors, and uses five machine learning classifiers to make predictions [22]. Subsequently, RNAlight was developed to predict the subcellular localization of multiple types of RNAs, including lncRNAs, mRNAs, and circRNAs [23]. It extracts k-mer features and utilizes LightGBM to predict the subcellular localizations of RNAs.

Further research is necessary to explore and improve the understanding of circRNA subcellular localization prediction. Although machine learning methods have shown promise in the field, it is important to note that deep learning methods have demonstrated remarkable capabilities in capturing complex patterns from large-scale biological data [24]. Their ability to automatically extract features and learn high-level representations from biological sequences can provide more accuracy and generalizable performance in predicting circRNA subcellular localization. In addition, circRNAs exhibit significant specificity in their expression patterns, showing high levels of cell type and tissue specificity, indicating circRNA subcellular localization is highly related to cellular environments [25]–[27]. Certain circRNAs exhibit varying subcellular localization patterns across different cell types [28]. For example, circZEB1, extensively studied in brain development, shows dynamic subcellular localization, primarily in the cytoplasm at embryonic day 60 but exclusively in the nucleus at embryonic day 80 [29]. However, current computational methods do not take cell line-specificity into account. They often aggregate data from all cell lines to train a general model, which cannot predict circRNA subcellular localization in diverse cell lines.

In the study, we proposed CellCircLoc, a cell line-specific deep learning model, to predict circRNA subcellular localization. Considering the differences of subcellular localization patterns between different cell lines, we collected cell line-specific circRNA subcellular localization data, consisting of 121,266 circRNA sequences from seven distinct cell lines. CellCircLoc utilizes a combination of convolutional neural networks (CNN), Transformer blocks, bidirectional long short-term memory (Bi-LSTM) to capture both sequence local features and long-range dependencies within the sequences. In the Transformer blocks, we used an attentive convolution mechanism to capture the importance of nucleotides. Finally, a fully connected layer is used to perform the circRNA subcellular localization prediction task.

To evaluate the performance of CellCircLoc, we conducted a comparative analysis against baseline models and existing predictors. The experimental results show that CellCircLoc outperforms both the baseline models and existing predictors in terms of accuracy, AUC, and F1-score. These results highlight the ability of CellCircLoc to accurately predict the subcellular localization of circRNAs. In addition to its predictive capabilities, CellCircLoc provides a nucleotide-level biological interpretation, which allows for the identification of nucleus-localized motifs. Furthermore, case studies show the advantages of considering cell line-specificity. Moreover, we conducted an ablation study and developed a user-friendly web server for CellCircLoc. Finally, we discussed the limitations of CellCircLoc and outlined potential future directions for its improvement.

II. Materials and Methods

A. Benchmark Dataset

To develop a reliable predictor, it is crucial to establish a reliable benchmark dataset. The Cancer-Specific CircRNA Database (CSCD) is a comprehensive resource that contains a large number of cancer-specific circRNAs identified from various human cancer and normal cell lines. We collected circRNA data with subcellular localization information from the CSCD 2.0 database (http://gb.whu.edu.cn/CSCD2) [21]. We generated a benchmark dataset to train and test our model by the following steps:

1. We retrieved all circRNA data from the CSCD 2.0 database and selected only those with subcellular localization information.

2. We chose circRNAs that only have one subcellular localization since only a very small number of circRNAs have multiple subcellular localizations.

3. In this study, we focused on two major subcellular localizations: the nucleus and the cytoplasm. CircRNAs in the cytosolic, insoluble cytoplasmic, and membrane fractions were considered as cytoplasmic localizations, while circRNAs in the nuclear, nucleolus, nucleoplasmic, and chromatin fractions were considered as nuclear localizations. We focus on the two primary localization classes is driven by both biological significance and the availability of reliable data. The nucleus and cytoplasm are the major compartments where circRNAs are most frequently identified and extensively studied. Moreover, existing datasets provide more comprehensive and higher-quality annotations for circRNAs localized in the two compartments, while data for other subcellular locations, such as chromatin or exosomes, remains limited.

4. We used the cd-hit-est tool with a cutoff of 80% to reduce data redundancy [30].

Finally, our benchmark dataset consists of 121,266 circRNAs across seven cell lines. Table I shows the distribution of circRNA subcellular localization of across different cell lines. From Table I, first, we can observe that the number of circRNAs varies greatly among different cell lines. Specifically, the K562 and HepG2 cell lines have a large amount of data, while the Hela-S3, HUVEC, Keratinocyte, GM12878, and H1-hESC cell lines exhibit a small amount of data. Second, it is important to note that the distribution of subcellular localization is imbalanced across all cell lines. Except for the HepG2 cell line, the ratio of nucleus to cytoplasm varies by more than five-fold in the other cell lines. In addition, we analyzed the distribution of circRNA sequence lengths in the benchmark dataset (see Supplementary Table S1). Approximately 80% of the sequences in the dataset have a length of fewer than 1000 nucleotides.

B. Model architecture

CellCircLoc is a sequence-based deep learning model designed for circRNA subcellular localization prediction. The architecture of CellCircLoc is shown in Figure 1. It consists of five main components: (1) Sequence coding, (2) CNN layer, (3) Transformer blocks with multi-head attentive convolution mechanism, (4) Bi-LSTM, (5) Prediction layer.

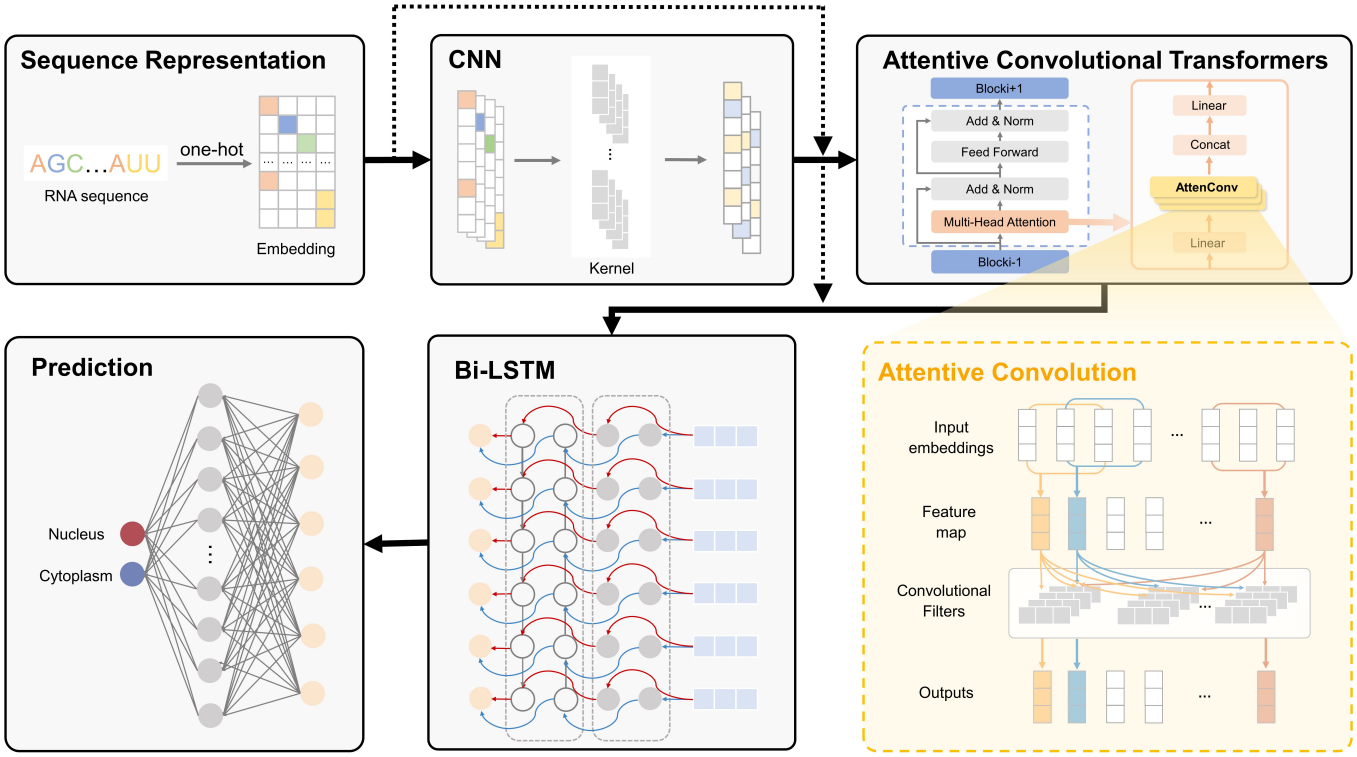

Fig. 1. Architecture of the CellCircLoc model. CellCircLoc accepts a circRNA sequence as input, and transforms it into a vector representation using one-hot encoding method. After sequence encoding, a CNN layer is used to extract local features from the encoded sequence. Subsequently, Transformer blocks with an attentive convolution mechanism is employed to capture the importance of nucleotides in the sequence. Next, a bi-LSTM layer is utilized to capture long-range dependencies and sequential information in the circRNA sequence. Finally, the subcellular localization prediction task is performed through a fully connected layer.

TABLE I
THE DISTRIBUTION OF CIRCRNA SUBCELLULAR LOCALIZATION OF DIFFERENT CELL LINE DATASETS.

Cell line	# of Nucleus	# of Cytoplasm	Total
K562	80,523	10,622	91,145
HepG2	11,660	16,022	27,682
Hela-S3	830	154	984
HUVEC	651	439	490
Keratinocyte	42	327	369
GM12878	248	53	301
H1-hESC	48	247	295

1) Sequence coding: The sequence coding part of CellCircLoc utilizes one-hot encoding to transform input circRNA sequences into vector representations. One-hot encoding is a widely used and simple coding method that serves as a fundamental representation in sequence analysis and plays an essential role in various computational biology tasks. With one-hot encoding, each nucleotide in the sequence is represented as a 4-dimensional vector, where 4 represents the four types of nucleotides. All elements of the vector are set to 0 except for the position corresponding to the nucleotide, which is set to 1. Specifically, A, C, G and U are encoded with a one-hot vector of (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0) and (0, 0, 0, 1), respectively. This encoding method creates a unique vector representation for each nucleotide in the sequence.

2) Convolutional neural network layer: The CNN layer is a fundamental component of CellCircLoc, which is designed to extract local features from the input circRNA sequence representation, similar to the utilization of input k-mers in some deep learning models. We chose CNN because deep learning models should have the ability to automatically learn features without the need for handcrafted feature engineering. The CNN's ability to automatically learn relevant features from the data makes it efficient and effective, eliminating the requirement for handcrafted features. In CellCircLoc, a 1-dimensional convolutional layer is employed to process the circRNA sequence representation. The length of the input circRNA sequence is denoted as l, and each nucleotide is represented by an embedding of dimension d, which is also treated as input channels. The convolutional filters with a kernel size of k slide across the circRNA sequence, capturing local features from the input circRNA sequence.

3) Transformer blocks with multi-head attentive convolution mechanism: The Transformer is a groundbreaking neural network architecture that has revolutionized natural language processing (NLP) tasks [31]. It introduces a self-attention mechanism that allows the model to focus on different parts of the input sequence when making predictions. The attention mechanism enables the model to attend to relevant words or tokens, giving it the ability to understand the context and relationships within a sentence or document. Li, et al. [32] proposed the attentive convolution mechanism, a variant of

the self-attention mechanism that has demonstrated remarkable performance in NLP. In the study, we used the attentive convolution mechanism to replace the original self-attention mechanism in CellCircLoc. Notably, the attentive convolution mechanism employed in the CellCircLoc offers several advantages over the original self-attention mechanism. One of its main advantages is its smaller model size, which results in more efficient and faster inference speed. Additionally, the attentive convolution mechanism does not worsen the predictions. Despite its reduced complexity, it consistently achieves comparable results, and in some cases, even better than the original self-attention mechanism. In a word, the attentive convolution mechanism provides a promising alternative to the original self-attention mechanism, offering the benefits of a smaller model size and faster inference without sacrificing predictive accuracy. Suppose the input embedding is $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_l]$. Perform convolution over $\mathbf{Q}$ using m convolution kernel $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_m]$, where $\mathbf{f}_i \in \mathbb{R}^{kd}$, $\mathbf{f}_i$ is the i-th convolution kernel, k is the width of the convolution kernel, and d is the embedding dimension. Perform n-gram convolution over the input embeddings using convolutional filters and obtain a feature map matrix $\mathbf{M}$:

$$\mathbf{M} = \mathbf{Q} \oplus \mathbf{F} \tag{1}$$

where $\oplus$ represents the convolution operation of $\mathbf{f}_i$ over $\mathbf{Q}$. The value in the feature map, denoted as m_{ij}, is computed using the formula:

$$m_{ij} = f\left(\mathbf{f}_i^T \cdot Concate\left(q_j, q_{j+1}, \cdots q_{j+k-1}\right) + b\right) \tag{2}$$

where $Concate$ represents concatenation, f is a non-linear activation function, and b is a bias term.

The feature map matrix $\mathbf{M}$ represents the semantic relevance between the k-mers and the filters. A k-mer refers to a subsequence of length k extracted from a larger sequence. For instance, in "ACGUAG", the 3-mers would be "ACG","CGU","GUA" and "UAG". K-mers are closely linked to RNA function by encoding sequence motifs and structural elements that play crucial roles in processes such as binding, regulation, and localization [33]. Within the feature map, values represent the importance or relevance of different k-mers to the task. These values can be calculated as attention weights, indicating the attention each k-mer should receive. The attentive convolution mechanism uses a set of convolutional filters $\mathbf{F}$, which aim to capture diverse features within the input sequence. By using the attention weights derived from the feature map and combining the convolutional filters accordingly, the output $\mathbf{O}$ of the attentive convolution mechanism can be calculated as follows:

$$\mathbf{O} = \mathbf{F} \cdot \mathbf{M} \tag{3}$$

where $\mathbf{O} = [o_1, o_2, \ldots, o_l] \in \mathbb{R}^{kd \times l}$.

In the original self-attention mechanism, the output space (which refers to the possible outputs or representations that a model can generate for a given input) is dynamic and highly dependent on the specific input. The self-attention mechanism assigns different weights to each component based on its importance in the context of other elements within the same input, leading to varying outputs for different inputs. In contrast, the output space in the attentive convolution mechanism is built using convolutional filters trained on the entire dataset, ensuring consistency across all inputs. Since the convolutional filters are learned globally and are invariant to the inputs, they capture general patterns and features that are relevant across different sequences, including k-mers. When certain k-mer are important for circRNA subcellular localization, their corresponding filters in the output space will have higher values. The feature map generated by the attentive convolution mechanism represents the semantic relevance between k-mers and convolutional filters, allowing the model to identify critical k-mers that play a decisive role in subcellular localization.

The attentive convolution mechanism also has a multi-head structure, which allows the model to attend to different parts of the input sequence simultaneously, enabling it to capture diverse and complementary information. Moreover, the multi-head structure can be used to evaluate the importance of different n-grams and convolutional filters. Assuming there are h heads in the model, the multi-head attentive convolution can be calculated as follows:

$$MultiHead\left(\mathbf{Q}\right) = \mathbf{W}^O cat\left(\mathbf{O}_1, \mathbf{O}_2, \cdots, \mathbf{O}_k\right) \tag{4}$$

$$\mathbf{O}_i = AttenConv\left(\mathbf{W}_i^Q \mathbf{Q}\right) \tag{5}$$

where $AttenConv$ represents the attentive convolution operation, $\mathbf{W}_i^Q \in \mathbb{R}^{(d/h)\times d}$ and $\mathbf{W}^O \in \mathbb{R}^{d \times kd}$.

As an important component of Transformer, a feed-forward neural network applies non-linear transformations to the intermediate representations generated by the attentive convolution mechanism. This approach enhances the model's ability to capture complex and higher-level features. Additionally, layer normalization is applied to stabilize the learning process and improve the model's generalization. The residual connection helps mitigate the potential vanishing gradient problem and enables the model to learn more effectively by preserving important information from the input. We employed the max-pooling operation to highlight the most important and relevant information within the input data. $\mathbf{F}$ and $\mathbf{M}$ are obtained from the last Transformer block, we calculated the global representation as follows:

$$\mathbf{g} = \mathbf{F} \cdot max\left(\mathbf{M}\right) \tag{6}$$

Assuming $\mathbf{O}_{last}$ is the output of the multi-head attention module from the last Transformer block, the attention weights α are calculated based on both $\mathbf{O}_{last}$ and $\mathbf{g}$.

$$\alpha = {\mathbf{O}_{last}}^T \cdot \mathbf{g} \tag{7}$$

where T represents the transpose operation.

4) Bi-LSTM: CellCircLoc incorporates Bidirectional Long Short-Term Memory (Bi-LSTM) layers as an essential component for capturing long-term dependencies in the input sequences [34]. By incorporating Bi-LSTM in the model, the need for positional encoding in the Transformer is eliminated as LSTM can effectively handle sequential information. The Bi-LSTM network consists of two LSTM sub-networks: one processes the input sequence in a forward direction, while

the other processes it in a backward direction. By utilizing hidden states and cell states in both directions, the Bi-LSTM can capture long-term dependencies and preserve sequential information. The Bi-LSTM layers in the CellCircLoc model enhance its ability to capture the intricate relationships within the input sequences, contributing to its subcellular localization prediction performance. Assuming the input to Bi-LSTM is $\mathbf{X} = [x_1, x_2, \ldots, x_l] \in \mathbb{R}^{d \times l}$:

$$f_t = \sigma\left(\mathbf{W}_f x_t + \mathbf{U}_f c_{t-1} + b_f\right) \tag{8}$$

$$i_t = \sigma\left(\mathbf{W}_i x_t + \mathbf{U}_i c_{t-1} + b_i\right) \tag{9}$$

$$o_t = \sigma\left(\mathbf{W}_o x_t + \mathbf{U}_o c_{t-1} + b_o\right) \tag{10}$$

$$g_t = tanh\left(\mathbf{W}_g x_t + \mathbf{U}_g c_{t-1} + b_g\right) \tag{11}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ g_t \tag{12}$$

$$h_t = o_t \cdot tanh\left(c_t\right) \tag{13}$$

where f_t, i_t, o_t stand for three gates, c_t represents the cell state and h_t stands for the hidden state at position t of the input sequence, respectively. The symbol σ represents the sigmoid function and $\circ$ represents the Hadamard product operation. In a Bi-LSTM, both a forward LSTM and a backward LSTM operate for each time step. The hidden states of these two LSTMs at all positions collectively form the output of the Bi-LSTM, which is then forwarded to subsequent layers for further processing.

5) Prediction: In the prediction module, we used a combination of a max-pooling layer and a fully connected layer to perform the circRNA subcellular localization prediction task. After the Bi-LSTM layer, a max-pooling layer is applied to extract the most important features from the sequence representation. The max-pooling operation selects the maximum value from each dimension of the sequence representation. The pooling operation helps reduce the dimensionality of the representation while retaining the most relevant information. Subsequently, a fully connected layer is employed to perform a linear transformation on the pooled features. Finally, the final neural network outputs are converted into class probabilities for each subcellular localization class of circRNAs.

C. Baseline methods

In this study, we constructed CellCircLoc, an advanced deep learning model for circRNA subcellular localization prediction. To demonstrate the effectiveness of CellCircLoc, we compared it with several baseline models. These baseline models include both traditional machine learning models and deep learning models.

1. k-mer + SVM, the model utilizes k-mer features and trains a SVM classifier.

2. k-mer + RF, the model utilizes k-mer features and trains a random forest (RF) classifier.

3. k-mer + LR, the model utilizes k-mer features and trains a linear regression (LR) classifier.

4. k-mer + NN, the model utilizes k-mer features and trains a neural network (NN) classifier.

5. word2vec + LSTM + MLP, the model represents circRNA sequences using word2vec embeddings, processes them with LSTM to capture sequential information, and employs an MLP for classification.

6. word2vec + TextCNN, the model represents circRNA sequences using word2vec embeddings, and applies multiple filters of varying sizes in the TextCNN to capture different n-gram features for subcellular localization prediction.

7. one-hot + Transformer + MLP, the model represents circRNA sequences using one-hot encoding, and employs a Transformer with the basic attention mechanism to capture contextual information, and then feeds the encoded sequence into an MLP for classification.

D. Evaluation Metrics

To evaluate the performance of CellCircLoc in predicting circRNA subcellular localization, we reported some well-known evaluation metrics, including accuracy (ACC), F-measure (F1 score), area under the receiver operating characteristic curve (AUC), and area under the precision-recall curve (AUPR). These evaluation metrics are widely used for RNA subcellular localization prediction and provide a comprehensive assessment of the model's performance.

E. Implementation details

CellCircLoc was implemented using PyTorch. To standardize the lengths of all circRNA sequences, a preprocessing step involving truncation and zero-padding was performed. A 5-fold cross-validation (CV) method was employed to evaluate the performance of the models. In this process, the dataset was divided into five subsets, with each subset being used as the test set once, while the remaining four subsets were used for training. The output values for the test set were calculated as the average score of the five models' outputs. During the training process, a grid search strategy was applied over a predefined search space to determine the optimal hyperparameters. Supplementary Table S2 provides the search space and the selected optimal hyperparameters. To address the challenge of class imbalance in the dataset, CellCircLoc incorporated a focal loss function during training, which places more emphasis on difficult-to-classify samples [35].

$$Focal\ Loss = -\frac{1}{m}\Sigma_{i=i}^{m}\left(\alpha y_i log(\hat{y}_i) + (1-y_i)\, log\left(1-\hat{y}_i\right)\right) \tag{14}$$

where m is the number of training samples, y_i and $\hat{y}_i$ are the true label and predictive score of sample i. The imbalance parameter α is defined as the ratio of the majority class to the minority class, where the majority class is labeled as 0 and the minority class is labeled as 1. During the training process, a transfer learning strategy was employed, where the model parameters trained on K562 data was used to replace the random initialization parameters for training on HUVEC, Keratinocyte, H1-hESC, and GM12878 cell lines. This was done since these four cell lines have limited data, and they can benefit from the shared characteristics and similarities between different cell lines, potentially enhancing the prediction results.

III. Comparison with Deep Learning Baseline Models

A. Comparison with baseline models

In order to compare the performance of CellCircLoc with the baseline models, a 5-fold CV was employed. Specifically, the benchmark dataset was split into training (90%) and hold-out test (10%) sets. Within the training set, we conducted a 5-fold CV, where 80% of the data was used for training and the remaining 20% for validation. This process was repeated five times, with each fold serving as the validation set, resulting in five distinct sets of results. We calculated the average performance across these five folds to obtain a robust evaluation.

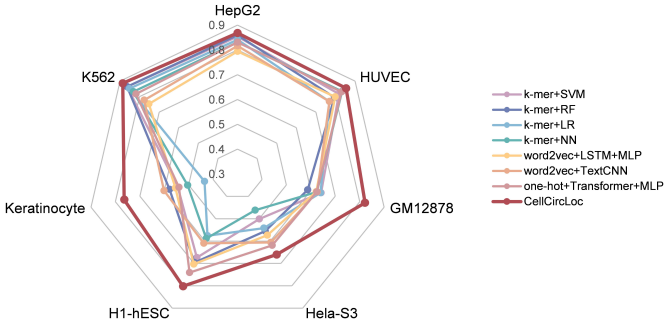

Fig. 2. AUCs for CellCircLoc and baseline models across seven cell lines.

Figure 2 displays a spider chart that presents the AUCs for CellCircLoc and the baseline models across seven cell lines. By comparing the AUC values of the baseline models, we can observe that their performances vary across different cell lines. For example, the k-mer+RF model achieves the highest AUC in all baseline models (AUC of 0.859 for HepG2 and AUC of 0.864 for K562). However, it shows relatively lower performance in other cell lines. In contrast, CellCircLoc consistently outperforms the baseline models across all cell lines. Particularly in the GM12878 and Keratinocyte cell lines, CellCircLoc exhibits significantly superior performance compared to the other baseline models.

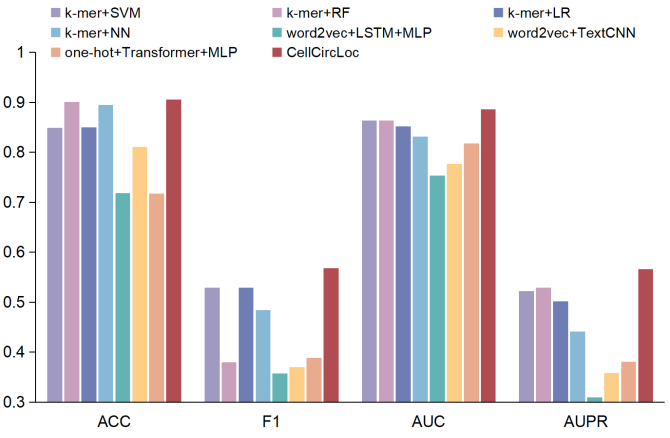

Fig. 3. The prediction performance of CellCircLoc and baseline models for the K562 cell line.

In addition, we took K562 cell line to show the effectiveness of Cell-CircLoc (the results of other six cell lines can be seen in Supplementary Figure S1). The choice of using the K562 cell line is based on several key factors. K562 is known to have the largest amount of circRNA data among the considered cell lines. This abundance of data allows for a more comprehensive and representative analysis, reducing potential biases that could arise from smaller datasets. Furthermore, selecting K562 as the representative cell line is essential in showcasing the models' ability to handle varying data distributions and class imbalance. Figure 3 shows the prediction performance of CellCircLoc and the baseline models for the K562 cell line. From Figure 3, we can observe that CellCircLoc outperforms the other baseline models in terms of ACC (0.906), F1 score (0.568), AUC (0.886), and AUPR (0.566). The results clearly demonstrate that CellCircLoc outperforms all the baseline models for the K562 cell line.

B. Comparison with existing predictors

In addition, to further evaluate the performance of Cell-CircLoc in predicting circRNA subcellular localizations, we compared it with some existing state-of-the-art methods. To the best of our knowledge, Circ-LocNet is the only method that focuses on circRNA subcellular localization prediction. However, we were unable to compare the performance of Circ-LocNet due to the unavailability of its source code and some issues with their web server, which prevented us from accessing their prediction results [22]. Therefore, we selected RNAlight [23], which has been demonstrated to be highly effective in predicting the subcellular localization of multiple types of RNAs, including lncRNAs, mRNAs, and circRNAs. Figure 4 displays a spider chart that presents the

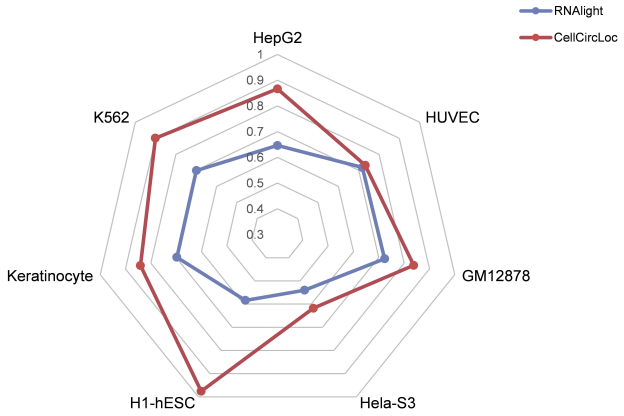

Fig. 4. AUCs for CellCircLoc and RNAlight across seven cell lines.

AUCs for CellCircLoc and RNAlight across seven cell lines (the confusion matrices and ROC curves of CellCircLoc with RNAlight on the test set can be seen in Supplementary Figure S2 and S3). The results clearly indicate that CellCircLoc outperforms RNAlight.

Similarly, we took the K562 cell line as an example to show the effectiveness of CellCircLoc (the results of other six cell lines can be found in Supplementary Figure S4).

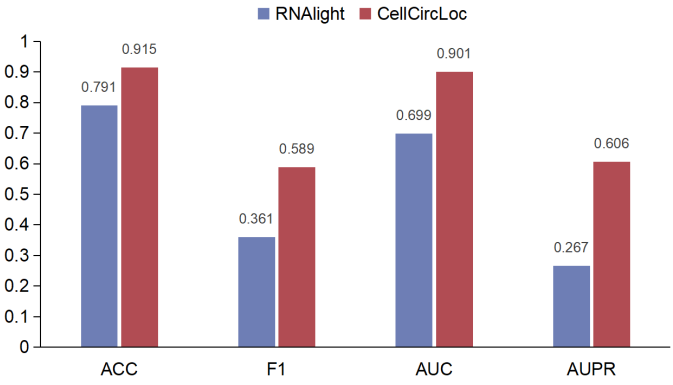

Fig. 5. The prediction performance of CellCircLoc and RNAlight for the K562 cell line.

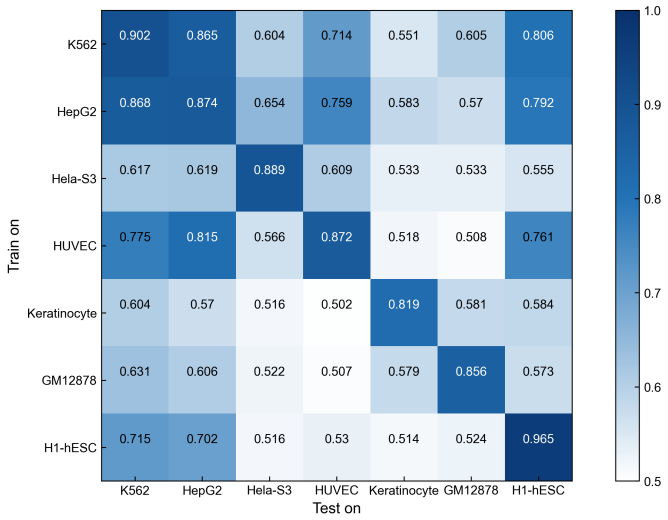

Fig. 6. The heatmap of AUCs for the cross-cell line analysis. The darker blue colors indicate the higher AUC values. The horizontal axis represents the cell line used for model training, and the vertical axis represents the cell line used for evaluation.

Figure 5 shows the prediction performance of CellCircLoc and RNAlight for the K562 cell line. From Figure 5, CellCircLoc achieves superior performance over RNAlight, as evidenced by its higher values for ACC (0.915 vs. 0.791), F1 (0.589 vs. 0.361), AUC (0.901 vs. 0.699), and AUPR (0.606 vs. 0.267). These results demonstrate that CellCircLoc is more effective in predicting the subcellular localization of circRNAs.

C. Cross-cell line analysis

To demonstrate the significance of predicting circRNA subcellular localization across different cell lines, we conducted a cross-cell line analysis. In this analysis, we trained individual models using the dataset from a specific cell line and evaluated them on the datasets of the remaining six cell lines. This approach allowed us to investigate whether a model trained on one cell line could generalize effectively to other cell lines.

Figure 6 presents a heatmap illustrating the AUC values obtained during the cross-cell line analysis. In Figure 6, the horizontal axis represents the cell line utilized for training the model, while the vertical axis represents the cell line where the model's performance was assessed. From the figure, several key observations can be made: while certain models perform well across different cell types (e.g., the model trained on the K562 cell line performs well on the HepG2 cell line), the majority of models show a notable decline in performance when applied to diverse cell lines. For instance, the model trained on the HUVEC cell line performs poorly when applied to the Keratinocyte cell line, achieving an AUC of only 0.502. These results underscore the importance of cell line specificity in circRNA subcellular localization prediction, highlighting the need to account for distinct cellular environments for robust and accurate model generalization.

D. Interpretability analysis

CellCircLoc can provide nucleotide-level interpretability, enabling the discovery of important motifs. Here, we took circRNA "hsa_circBIRC6_008" as an example to show the interpretability of CellCircLoc. A recent study found that the nuclear enrichment of circRNA is driven by the binding of the RNA-binding protein SRSF1 [36]. The study demonstrated that circRNA containing SRSF1 binding motifs were enriched in the nuclear fraction. Upon knockdown of SRSF1, a significant shift from nuclear to cytoplasmic localization was observed for six circRNAs (including circBIRC6), which contained a relatively higher number of SRSF1 eCLIP clusters and/or predicted binding motifs. Transcriptome-wide analysis further revealed that circRNA with SRSF1 binding motifs exhibited increased nuclear localization, and SRSF1 knockdown led to a decrease in the nuclear-to-cytoplasmic ratios of bound circRNA. We used RBPmap to predict SRSF1 binding sites for hsa_circBIRC6_008 [37], [38]. In the RBPmap prediction, "GGAUGA" at position 62 and "GAAGGA" at position 86 get the two highest Z-scores, indicating that they are the most likely binding motifs. Figure 7 shows the prediction heatmap of hsa_circBIRC6_008. The red regions in the heatmap denote contributions to the nuclear subcellular localization, while the blue regions indicate contributions to the cytoplasm subcellular localization. In the heatmap generated by CellCircLoc, we can observe that the "GGAUGA" and "GAAGGA" regions are marked in red. This level of interpretability in CellCircLoc allows us to identify crucial regions that play a significant role in determining the circRNA's localization, thereby enhancing our understanding of the underlying biological mechanisms.

E. Case Study

In order to demonstrate the effectiveness of CellCircLoc in predicting cell line-specific circRNA subcellular localization, we conducted a case study using the circRNA hsa_circ_0054124. The circRNA is identified in the K562 and HelaS3 cell lines, with PRKD3 serving as its host gene [39]. Previous study reported that circRNA isoforms of PRKD3 exhibit dynamic subcellular localization, with predominant nuclear localization in the K562 cell line and cytoplasmic localization in the Hela-S3 cell line [28]. We used CellCircLoc and RNAlight to predict the subcellular localization of hsa_circ_0054124 and Figure 8 displays the prediction results.

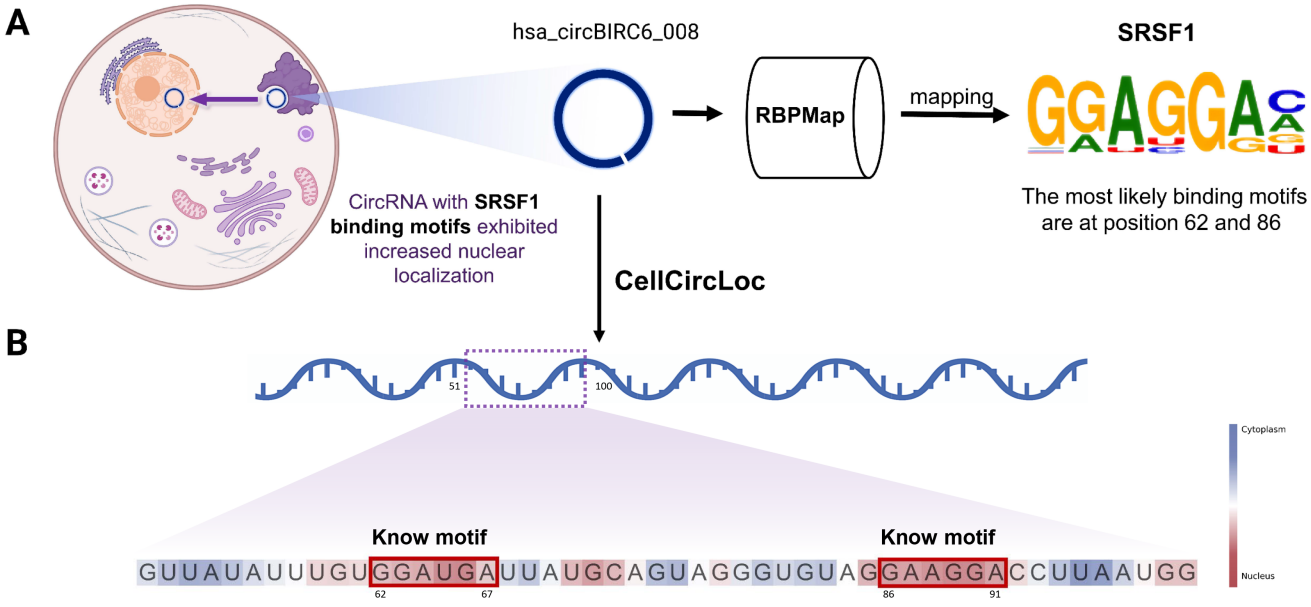

Fig. 7. Utilizing RBPmap to predict SRSF1 binding motif sites for hsa_circBIRC6_008. This figure displays the prediction heatmap of hsa_circBIRC6_008, where the red regions indicate contributions to the nucleus subcellular localization, and the blue regions indicate contributions to the cytoplasm subcellular localization. In CellCircLoc's heatmap, the "GGAUGA" and "GAAGGA" regions are marked in red.

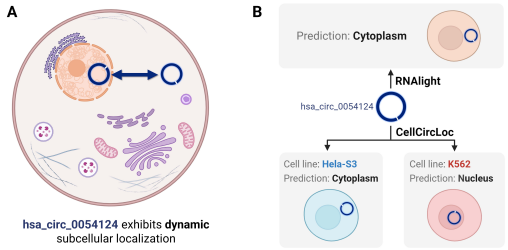

Fig. 8. Cell line-specific subcellular localization prediction of circRNA hsa_circ_0054124 using CellCircLoc and RNAlight. (A) The circRNA exhibits dynamic subcellular localization. (B) CellCircLoc successfully predicts the subcellular localization of hsa_circ_0054124 in both the K562 cell line (nucleus) and the Hela-S3 cell line (cytoplasm). However, RNAlight only provide a general prediction result for hsa_circ_0054124, without cell line specificity.

From Figure 8, we can observe that CellCircLoc successfully predicted the subcellular localization of hsa_circ_0054124 in both the K562 cell line (nucleus) and the Hela-S3 cell line (cytoplasm). However, RNAlight, as a general predictor, only provided an single prediction result for hsa_circ_0054124 because RNAlight does not support cell line-level prediction. These results further demonstrate the advantage of considering cell line specificity in circRNA subcellular localization prediction.

F. Ablation study

To thoroughly evaluate the contributions of each component within CellCircLoc, we conducted an ablation study. This ablation study involved training and evaluating various variant models, each excluding a key module, to discern their impact on overall performance. These variations included versions of CellCircLoc without the CNN module, without the Transformer blocks, without the Bi-LSTM module, and one that replaced the attentive convolution mechanism with a basic attention mechanism. Subsequently, each model was retrained and tested using 5-fold cross-validation on the HepG2 cell line, and the results are summarized in Table II.

The ablation study results clearly demonstrated that removing any component from CellCircLoc led to a performance decline across all evaluation metrics. We believe that the improved performance in CellCircLoc stems from the synergy between its various components. The CNN module effectively captures local sequence patterns, which are essential for identifying circRNA sequence motifs and features. Notably, the removal of the Bi-LSTM module had the most significant impact on performance, given its vital role in modeling long-term dependencies within circRNA sequences. Moreover, the attentive convolution mechanism enhances both interpretability and accuracy by utilizing globally trained convolutional filters to detect biologically significant patterns and k-mers. This mechanism outperforms basic attention by maintaining structured output space, ensuring consistent identification of key sequence motifs across input sequences. In summary, the combination of these components provides a more comprehensive representation of circRNA sequences, resulting in superior performance.

G. Web server

To facilitate the use of CellCircLoc, we developed a user-friendly web server, http://csuligroup.com:8000/cellcircloc.

TABLE II
THE PERFORMANCES OF CELLCIRCLOC AND ITS VARIANT MODELS IN THE HEPG2 CELL LINE.

Model	Acc	F1	AUC	AUPR
Without CNN	0.791	0.819	0.857	0.872
Without Transformer	0.783	0.813	0.847	0.864
Without Bi-LSTM	0.771	0.806	0.828	0.845
Using basic attention	0.795	0.824	0.862	0.875
CellCircLoc	0.802	0.832	0.868	0.881

The web server allows users to input a circRNA sequence and obtain prediction results. By submitting a circRNA sequence through the web server, CellCircLoc is applied to analyze and predict the subcellular localization of the circRNA. Notably, CellCircLoc can predict subcellular localization when the cell line information is unknown. By using a weighted voting mechanism across the results of 7 established cell line models, CellCircLoc generates prediction results for circRNAs with unknown cell line information. This user-friendly web server provides a convenient way for researchers and users to gain insights into the potential subcellular localization of circRNAs of interest.

IV. CONCLUSION

The identification of circRNA subcellular localization can provide valuable insights into the biological functions of circRNAs. However, existing computational methods do not take the cell line specificity into consideration, which is crucial for accurate prediction of circRNA subcellular localization. In the study, we proposed CellCircLoc, a cell line-specific circRNA subcellular localization prediction model. By combining various deep learning components, including CNN, Transformer blocks with attentive convolution mechanism, Bi-LSTM, and fully connected layers, CellCircLoc achieves remarkable performance in accurately predicting the subcellular localization of circRNAs. The main contributions of CellCircLoc are summarized as follows:

1) CellCircLoc considers the diversity and context-dependent behavior of circRNAs across different cell lines by taking cell line specificity into account.

2) CellCircLoc provides interpretability analysis, offering valuable insights into the importance of motifs that contribute to the predictions.

3) CellCircLoc demonstrates superior performance compared to baseline models and existing predictors, highlighting its effectiveness in accurately predicting circRNA subcellular localization.

4) The availability of a free webserver for CellCircLoc enhances accessibility and usability, making it a useful tool for researchers in the field.

Although the results demonstrate that CellCircLoc could serve as an effective predictor for cell line-specific circRNA subcellular localization, there are still limitations that influence its performance. One limitation is its dependence on data availability, particularly in different cell lines. The model's performance is better in cell lines with more data, such as K562 and HepG2 cell lines, while it exhibits relatively poorer performance in cell lines with limited data. Therefore, expanding the dataset to include more diverse cell lines would enhance the model's generalizability and improve its performance across different cellular contexts. Another potential improvement is the utilization of pre-training and fine-tuning techniques using other RNA data sources. By leveraging pre-trained models on large-scale RNA datasets, such as lncRNAs or mRNAs, and then fine-tuning them for circRNA subcellular localization, CellCircLoc could achieve better performance. Additionally, the current version of CellCircLoc focuses on predicting circRNAs with a single subcellular localization. However, many circRNAs shuttle between multiple locations, performing distinct biological functions. While biologically significant, the limited availability of high-confidence data on circRNAs with multiple localization annotations poses a challenge. Future versions of CellCircLoc could be adapted for multi-label classification, allowing each circRNA to be linked with multiple subcellular locations. Furthermore, future research could explore additional advanced techniques to tackle class imbalance issues prevalent in circRNA subcellular localization datasets. Implementing more sophisticated techniques [40], such as ensemble learning, cost sensitive learning, or data augmentation strategies, could mitigate this imbalance and improve the model's performance.

REFERENCES

[1] X. Liu, Y. Zhang, S. Zhou, L. Dain, L. Mei, and G. Zhu, "Circular rna: An emerging frontier in rna therapeutic targets, rna therapeutics, and mrna vaccines," *Journal of Controlled Release*, vol. 348, pp. 84–94, 2022.

[2] Z. Ward, J. Pearson, S. Schmeier, V. Cameron, and A. Pilbrow, "Insights into circular rnas: their biogenesis, detection, and emerging role in cardiovascular disease," *RNA biology*, vol. 18, no. 12, pp. 2055–2072, 2021.

[3] C. Lu, M. Zeng, F. Zhang, F.-X. Wu, M. Li, and J. Wang, "Deep matrix factorization improves prediction of human circrna-disease associations," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 3, pp. 891–899, 2020.

[4] C. Lu, M. Zeng, F.-X. Wu, M. Li, and J. Wang, "Improving circrna–disease association prediction by sequence and ontology representations with convolutional and recurrent neural networks," *Bioinformatics*, vol. 36, no. 24, pp. 5656–5664, 2020.

[5] R. Chen, S. K. Wang, J. A. Belk, L. Amaya, Z. Li, A. Cardenas, B. T. Abe, C.-K. Chen, P. A. Wender, and H. Y. Chang, "Engineering circular rna for enhanced protein production," *Nature Biotechnology*, vol. 41, no. 2, pp. 262–272, 2023.

[6] Y.-J. Kang, D.-C. Yang, L. Kong, M. Hou, Y.-Q. Meng, L. Wei, and G. Gao, "Cpc2: a fast and accurate coding potential calculator based on sequence intrinsic features," *Nucleic acids research*, vol. 45, no. W1, pp. W12–W16, 2017.

[7] S. Khan, A. Jha, A. C. Panda, and A. Dixit, "Cancer-associated circrna–mirna–mrna regulatory networks: A meta-analysis," *Frontiers in Molecular Biosciences*, vol. 8, p. 671309, 2021.

[8] A. R. Sonawane, J. Platig, M. Fagny, C.-Y. Chen, J. N. Paulson, C. M. Lopes-Ramos, D. L. DeMeo, J. Quackenbush, K. Glass, and M. L. Kuijjer, "Understanding tissue-specific gene regulation," *Cell reports*, vol. 21, no. 4, pp. 1077–1088, 2017.

[9] R. A. Wesselhoeft, P. S. Kowalski, and D. G. Anderson, "Engineering circular rna for potent and stable translation in eukaryotic cells," *Nature communications*, vol. 9, no. 1, p. 2629, 2018.

[10] C. Lu, L. Zhang, M. Zeng, W. Lan, G. Duan, and J. Wang, "Inferring disease-associated circrnas by multi-source aggregation based on heterogeneous graph neural network," *Briefings in Bioinformatics*, vol. 24, no. 1, p. bbac549, 2023.

[11] L. Chen and G. Shan, "Circrna in cancer: Fundamental mechanism and clinical potential," *Cancer letters*, vol. 505, pp. 49–57, 2021.

[12] L. S. Kristensen, T. Jakobsen, H. Hager, and J. Kjems, "The emerging roles of circrnas in cancer and oncology," *Nature reviews Clinical oncology*, vol. 19, no. 3, pp. 188–206, 2022.

[13] J. Zhang, X. Zhang, C. Li, L. Yue, N. Ding, T. Riordan, L. Yang, Y. Li, C. Jen, S. Lin *et al.*, "Circular rna profiling provides insights into their subcellular distribution and molecular characteristics in hepg2 cells," *RNA biology*, vol. 16, no. 2, pp. 220–232, 2019.

[14] L. Zhang, C. Lu, M. Zeng, Y. Li, and J. Wang, "Crmss: predicting circrna-rbp binding sites based on multi-scale characterizing sequence and structure features," *Briefings in Bioinformatics*, vol. 24, no. 1, p. bbac530, 2023.

[15] Y. Guo, X. Lei, Y. Pan, and R. Su, "An encoding-decoding framework based on cnn for circrna-rbp binding sites prediction," *Chinese Journal of Electronics*, vol. 33, no. 1, pp. 256–263, 2024.

[16] Z. Li, C. Huang, C. Bao, L. Chen, M. Lin, X. Wang, G. Zhong, B. Yu, W. Hu, L. Dai *et al.*, "Exon-intron circular rnas regulate transcription in the nucleus," *Nature structural & molecular biology*, vol. 22, no. 3, pp. 256–264, 2015.

[17] S. Misir, N. Wu, and B. B. Yang, "Specific expression and functions of circular rnas," *Cell Death & Differentiation*, vol. 29, no. 3, pp. 481–491, 2022.

[18] A. J. van Zonneveld, M. Kölling, R. Bijkerk, and J. M. Lorenzen, "Circular rnas in kidney disease and cancer," *Nature Reviews Nephrology*, vol. 17, no. 12, pp. 814–826, 2021.

[19] S. Kwon, "Single-molecule fluorescence in situ hybridization: quantitative imaging of single rna molecules," *BMB reports*, vol. 46, no. 2, p. 65, 2013.

[20] T. Cui, Y. Dou, P. Tan, Z. Ni, T. Liu, D. Wang, Y. Huang, K. Cai, X. Zhao, D. Xu *et al.*, "Rnalocate v2. 0: an updated resource for rna subcellular localization with increased coverage and annotation," *Nucleic acids research*, vol. 50, no. D1, pp. D333–D339, 2022.

[21] J. Feng, W. Chen, X. Dong, J. Wang, X. Mei, J. Deng, S. Yang, C. Zhuo, X. Huang, L. Shao *et al.*, "Cscd2: an integrated interactional database of cancer-specific circular rnas," *Nucleic Acids Research*, vol. 50, no. D1, pp. D1179–D1183, 2022.

[22] M. N. Asim, M. A. Ibrahim, M. Imran Malik, A. Dengel, and S. Ahmed, "Circ-locnet: A computational framework for circular rna sub-cellular localization prediction," *International Journal of Molecular Sciences*, vol. 23, no. 15, p. 8221, 2022.

[23] G.-H. Yuan, Y. Wang, G.-Z. Wang, and L. Yang, "Rnalight: a machine learning model to identify nucleotide features determining rna subcellular localization," *Briefings in Bioinformatics*, vol. 24, no. 1, p. bbac509, 2023.

[24] M. Zeng, Y. Wu, C. Lu, F. Zhang, F.-X. Wu, and M. Li, "Deeplncloc: a deep learning framework for long non-coding rna subcellular localization prediction based on subsequence embedding," *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab360, 2022.

[25] X. Li, L. Yang, and L.-L. Chen, "The biogenesis, functions, and challenges of circular rnas," *Molecular cell*, vol. 71, no. 3, pp. 428–442, 2018.

[26] L. Verduci, E. Tarcitano, S. Strano, Y. Yarden, and G. Blandino, "Circrnas: role in human diseases and potential use as biomarkers," *Cell death & disease*, vol. 12, no. 5, p. 468, 2021.

[27] C.-Y. Yu and H.-C. Kuo, "The emerging roles and functions of circular rnas and their generation," *Journal of biomedical science*, vol. 26, no. 1, pp. 1–12, 2019.

[28] T.-J. Chuang, Y.-J. Chen, C.-Y. Chen, T.-L. Mai, Y.-D. Wang, C.-S. Yeh, M.-Y. Yang, Y.-T. Hsiao, T.-H. Chang, T.-C. Kuo *et al.*, "Integrative transcriptome sequencing reveals extensive alternative trans-splicing and cis-backsplicing in human cells," *Nucleic Acids Research*, vol. 46, no. 7, pp. 3671–3691, 2018.

[29] M. T. Venø, T. B. Hansen, S. T. Venø, B. H. Clausen, M. Grebing, B. Finsen, I. E. Holm, and J. Kjems, "Spatio-temporal regulation of circular rna expression during porcine embryonic brain development," *Genome biology*, vol. 16, pp. 1–17, 2015.

[30] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "Cd-hit suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[32] P. Li, P. Zhong, K. Mao, D. Wang, X. Yang, Y. Liu, J. Yin, and S. See, "Act: an attentive convolutional transformer for efficient text classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 15, 2021, pp. 13 261–13 269.

[33] J. M. Kirk, S. O. Kim, K. Inoue, M. J. Smola, D. M. Lee, M. D. Schertzer, J. S. Wooten, A. R. Baker, D. Sprague, D. W. Collins *et al.*, "Functional classification of long non-coding rnas by k-mer content," *Nature genetics*, vol. 50, no. 10, pp. 1474–1482, 2018.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[36] M. Ron and I. Ulitsky, "Context-specific effects of sequence elements on subcellular localization of linear and circular rnas," *Nature communications*, vol. 13, no. 1, p. 2481, 2022.

[37] M. Liu, Q. Wang, J. Shen, B. B. Yang, and X. Ding, "Circbank: a comprehensive database for circrna with standard nomenclature," *RNA biology*, vol. 16, no. 7, pp. 899–905, 2019.

[38] I. Paz, I. Kosti, M. Ares Jr, M. Cline, and Y. Mandel-Gutfreund, "Rbpmap: a web server for mapping binding sites of rna-binding proteins," *Nucleic acids research*, vol. 42, no. W1, pp. W361–W367, 2014.

[39] P. Glažar, P. Papavasileiou, and N. Rajewsky, "circbase: a database for circular rnas," *Rna*, vol. 20, no. 11, pp. 1666–1670, 2014.

[40] M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, "Effective prediction of three common diseases by combining smote with tomek links technique for imbalanced medical data," in *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*. IEEE, 2016, pp. 225–228.