scientific reports

Check for updates

OPEN

LGLoc as a new language model-driven graph neural network for mRNA localization

Saeedeh Akbari Rokn Abadi[1,3], Aref Shahbakhsh[2,3] & Somayyeh Koohi[2]✉

The localization of mRNA is crucial for the synthesis of functional proteins and plays a significant role in cellular processes. Understanding mRNA localization can enhance applications in disease diagnosis (e.g., cancer, Alzheimer's) and drug development. While numerous methods have been developed for this purpose, existing approaches face challenges: experimental methods are often costly and time-consuming, while computational methods may lack accuracy and efficiency. To address these limitations, we propose LGLoc, a machine learning-based approach designed to improve the accuracy of mRNA localization predictions with low computational overhead. LGLoc employs a Graph Neural Network encoder that utilizes the RNA's secondary structure, complemented by a BERT encoder focused on the primary RNA sequence. Additionally, it integrates k-mer and nucleotide frequency-based encoders to capture essential sequence characteristics. Feature selection is conducted using an analysis of variance, and classification is performed through a one-vs-rest Naïve Bayes classifier tailored for mRNA classification. Our results indicate that LGLoc significantly outperforms existing methods, such as mRNALoc and MSLP, across key performance metrics including Accuracy, Sensitivity, Specificity, F1-score, AUC, and MCC. Notably, LGLoc achieves over 49% improvement in average F1-score and 26% in average MCC compared to mRNALoc, demonstrating its reliability and effectiveness in mRNA subcellular localization.

Keywords mRNA, Machine learning, Graph neural network, Language model, BERT, Subcellular localization

mRNA localization involves complex interactions between mRNA molecules, motor proteins, and cytoskeletal elements. These interactions are crucial for transporting mRNA to specific subcellular locales, ensuring protein synthesis occurs at the appropriate time and location within the cell[1]. This spatial localization of mRNA transcripts is essential for gene expression and protein synthesis, significantly influencing cellular functions and contributing to pathological states when dysregulated[2,3]. So, mislocalization of mRNA is associated with severe problems in cellular processes, such as diseases including spinal muscular atrophy, Alzheimer's disease, and various cancers[4]. In response, the pharmaceutical industry has harnessed the targeting of mRNA within different subcellular compartments. This strategy enables the suppression of defective genes through innovative treatments like oligonucleotide therapy and macrophage-targeted therapy, offering promising avenues for addressing these diseases[5–7]. As a result of this importance, several types of methods were developed to identify and predict the location of mRNA within cells.

Experimental techniques such as RNA fluorescent in situ hybridization (RNA-FISH) are among the most reliable and accurate methods for determining mRNA localization. Despite this, their cost and time consumption limit their use. These methods are also primarily limited to specific tissues and demand extensive resources, highlighting the need for alternative approaches[8,9]. Consequently, the development of computational tools for in silico prediction of mRNA subcellular localization has gained increasing importance[10].

In recent years, similar to other biological problems such as predicting the function of other types of RNA (e.g., piRNA)[11] or identifying sumoylation sites[12], innovative computational models have significantly advanced the prediction of mRNA localization. The first approach, RNATracker[13], employed Recurrent Neural Networks (RNNs) to integrate mRNA sequence data with predicted secondary structures using one-hot encoding. Although this method laid the groundwork for further advancements, its reliance on noisy datasets and inadequate secondary structure representation led to unreliable predictions. The mRNALoc[14] tool was

[1]Division of Computational Science and Technology, School of Electrical Engineering and Computer Science (EECS), KTH Royal Institute of Technology, Stockholm, Sweden. [2]Computer Engineering Department, Sharif University of Technology, Tehran, Iran. [3]Saeedeh Akbari Rokn Abadi and Aref Shahbakhsh contributed equally to this work. ✉email: Koohi@sharif.edu

developed to address these shortcomings, utilizing the comprehensive and reliable RNALocate[15] dataset and the PseKNC encoder. This approach sought to overcome the limitations of RNATracker by employing five binary SVM classifiers with specific thresholds and PseKNC encoding. Subsequently, mRNALocator[16] was introduced to enhance mRNALoc's performance by incorporating an additional encoder, Eseiip, along with PseKNC and using various classifiers such as XGBoost, CatBoost, and LightGBM.

Further advancements led to the development of iLoc-mRNA[17], demonstrating that an intermediate feature selection phase could significantly boost performance. iLoc-mRNA employed ANOVA and forward search for feature selection and utilized four binary SVM classifiers for classification. Building on these advances, the SubLocEP[18] tool aimed to improve mRNA subcellular localization prediction by incorporating nine encoders for richer feature representation, followed by feature selection using LightGBM. Lastly, MSLP[19] introduced an approach that uses specific features tailored to each class rather than a broad range of features for a single mRNA sequence. MSLP employed k-mer and PseKNC features for classes Cytoplasm, Nucleus, Endoplasmic Reticulum, and physicochemical properties, Z-curve features for classes Extracellular Region, Mitochondria, followed by the SHAP algorithm for feature selection and five binary CatBoost classifiers for classification. Of course, there are other efforts in related tasks, such as the Deep-piRNA model, which incorporates a bi-layered prediction framework relying on discriminative features encoded for PIWI-interacting RNA (piRNA) prediction. This model showcases the power of encoder-enhanced architectures for sequence feature extraction[12] and highlights the growing importance of using specific, contextually relevant encoders to improve feature representation and prediction accuracy—an approach that can inspire solutions to the challenges addressed in this work.

Despite all the efforts in developing tools, the accuracy of them for all categories still needs to be improved. A summary of these methods is presented in Table 1. This table reveals that proposed methods have paid less attention to encoding, feature representation, and feature selection phases of mRNA subcellular localization prediction. So, our efforts are focused on enhancing encoding strategies to develop more robust and informative features.

In this study, we employ three methods together to distinguish effective patterns of RNA sequences in their location. For the first level, we enhance the sequence encoding process by incorporating secondary structure data, representing mRNA molecules as graphs to effectively capture critical structural information often lost in linear sequence analysis. This approach is precious because the secondary structure of RNA plays a crucial role in various cellular functions, including transcription, splicing, translation, and localization, thereby improving prediction accuracy[20]. Additionally, for the second level, we employ Natural Language Processing (NLP) models to process the primary sequence, capturing hierarchical relationships within the sequence crucial for understanding complex biological processes. Hierarchical feature representation enhances the ability of models to learn from multi-level patterns within the data, leading to more precise and reliable predictions[21]. As the last level, we also integrate k-mer-based features, which are particularly effective in improving prediction Accuracy for specific localization categories. By combining these diverse encoding methods, we aim to address the limitations of previous approaches, which could be better features from RNA, and develop a more accurate and reliable model for predicting mRNA subcellular localization.

In this paper, we begin with the Method section, which details the methodologies employed in our research, including a comprehensive explanation of the feature encoding processes. Within the Feature Encoding subsection, we explore three approaches: mRNA Sequence Encoding using Graph Neural Networks (GNNs), Splice BERT, and k-mer frequency combined with CKSNAP. Following this, the Feature Selection and Classification section elaborates on the techniques we applied for feature selection and classification. The Results section presents the findings from our experiments, accompanied by an in-depth analysis in the Discussions section. Finally, the Conclusion section summarizes the key contributions of our study and outlines potential directions for further research.

Methods

The LGLoc model, Language Model-Driven Graph Neural Network for mRNA Localization, is designed to predict mRNA subcellular localization by integrating advanced machine-learning techniques. Figure 1 shows that the LGLoc model is created in three stages: Feature Encoding, Feature Selection, and Classification. Each stage incorporates sophisticated methodologies to ensure robust and accurate predictions.

Tool name	Model	Encoding algorithm	Feature selection
RNATracker[13]	Convolutional network, long short-term memory network, attention mechanism	One-hot encoding	None
mRNALoc[14]	Support vector machine	Pse-KNC (k = 2,…6)	None
mRNALocater[16]	LightGBM, XGBoost, CatBoost	Pse-KNC (k = 2,…,6) + PseEIIP	None
iLoc-mRNA[17]	Support vector machine	k-mer (k = 9)	Anova + forward search
SubLocEP[18]	LightGBM	PseEIIP, TNC, DNC, CKSNAP, DNC, PCPseDNC, PCPseTNC, SCPseDNC, SCPseTNC, DACC	Lightgbm
MSLP[19]	CatBoost	k-mer (k = 2,…,5), Pse-KNC (k = 2,…,5), (PseEIIP, DPCP, TPCP), Z-curve	Shap values

Table 1. Summary of mRNA localization prediction methods.

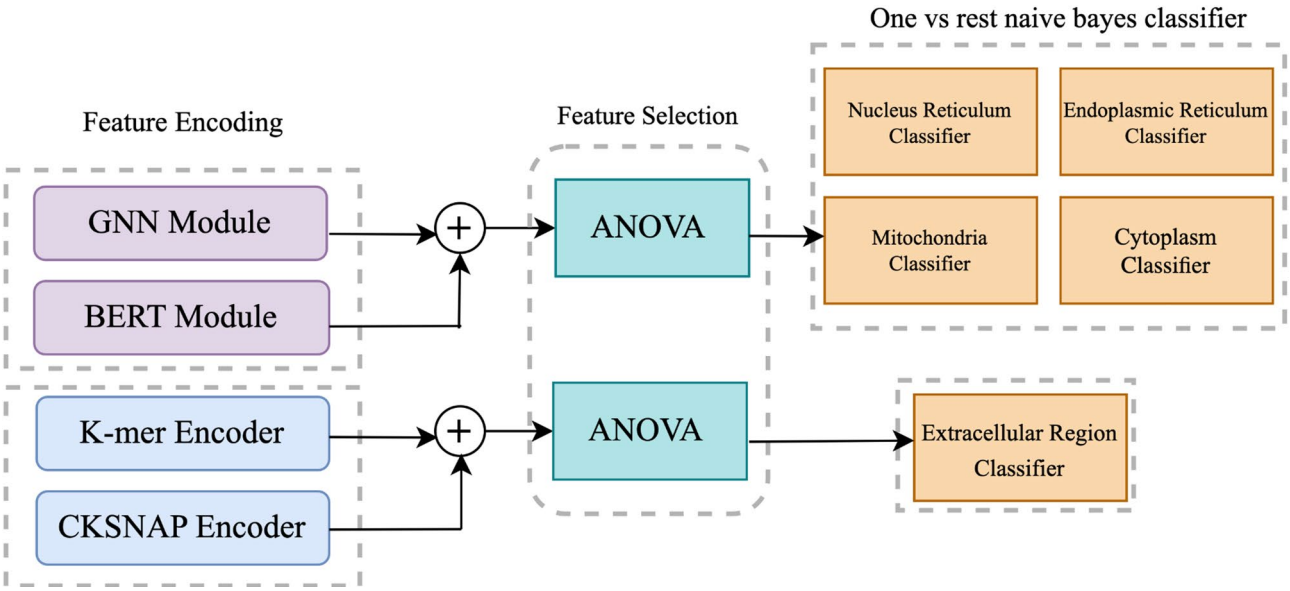

Figure 1. The LGLoc Architecture; LGLoc consists of the two main blocks: (1) Encoding block, and (2) Feature selection and Classification block.

The encoding stage leverages the secondary structure of mRNA through a GNN and the primary sequence using the Splice BERT model, an NLP technique, and k-mer frequency. The inclusion of secondary structure information provides additional contextual features that enhance prediction Accuracy by capturing spatial and structural dependencies within the mRNA molecule[22]. Additionally, using Splice BERT, a transformer-based model, offers significant advantages in understanding the primary sequence by effectively capturing long-range dependencies and contextual relationships within the mRNA sequence[23]. This comprehensive approach captures the nuanced features of mRNA necessary for precise localization predictions. It should be noted that, in order to select each module of the feature encoding stage in LGLoc, we conducted extensive tests using several popular Large Language Models (LLMs) and GNNs. The best-performing model was selected for each module. The results of these tests are provided in the supplementary materials.

In the LGLoc model, the feature selection stage uses the Anova algorithm to use only efficient and probabilistic features by using Anova analysis. Then, the last stage employs classification by one-vs-rest using the naive Bayes algorithm. In the following, we explain each of the three sections of the LGLoc model by incorporating every component of each part.

Feature Encoding section

This section describes implementing four distinct encoding methods for predicting mRNA subcellular localization. The first method utilizes a GNN to encode features based on the RNA's secondary structure. The second method employs the SpliceBERT[24] model, which encodes features derived from the primary structure of RNA. The third method involves k-mer frequency analysis, which calculates the normalized frequencies of k-mers within the RNA sequence. The fourth method uses the CKSNAP algorithm, which, like k-mer frequency analysis, considers k-mers but includes specific gaps between them. It should be noted for the LGloc approach, we concatenate the GNN and SpliceBERT encoders to classify mRNA into Cytoplasmic, Endoplasmic Reticulum, Mitochondrial, and Nuclear categories. Additionally, k-mer frequency analysis and the CKSNAP algorithm are applied to predict extracellular localization.

Encoding by Secondary Structure and Graph Neural Network

Once the secondary structure of mRNA is predicted by the RNAfold tool from the Vienna RNA package[25], an mRNA graph is constructed based on this structure and the primary sequence. Each vertex has a 10-dimension feature vector concatenated from two one-hot vectors (4D + 6D) that specify the type of nucleotide and its substructure based on the forgi package[26]. This package breaks down RNA secondary structure into six distinct substructures:

1. Five prime (f): The unpaired nucleotides at the 5' end of the RNA strand. These are the first unpaired bases and are marked with an "f" (e.g., 'f0').
2. Three prime (t): The unpaired nucleotides at the 3' end of the strand, always labeled with a "t" (e.g., 't0').
3. Stem (s): Continuous regions of canonical Watson-Crick base-paired nucleotides. These stems are usually composed of consecutive base pairs and are labeled as 's0', 's1', 's2', etc.
4. Interior loop (i): The bulged-out nucleotides and loops found in the interior of the structure, flanked by paired regions. These are marked as 'i0', 'i1', 'i2', etc.

5. Multiloop segment (m): Single-stranded regions between two stems, and pseudo-knots or exterior loops are treated as multiloop segments in this context (e.g., 'm0', 'm1', etc.).
6. Hairpin loop (h): Short loop regions that form hairpin structures, denoted by 'h'.

Each of these six substructures, along with the nucleotide types (A, C, G, T), are encoded using one-hot encoding. The one-hot encoding scheme for nucleotides is as follows: each nucleotide (A, C, G, T) is represented by a 4-dimensional vector, where one element is '1' indicating the nucleotide, and the others are '0'. For substructures, we use a 6-dimensional vector indicating the type of substructure that each nucleotide belongs to.

Once the RNA secondary structure is predicted and encoded, we proceed to create the RNA graph. Each nucleotide in the sequence is represented as a node in the graph, and the connections between nucleotides are represented by edges. As mentioned previously, the nodes are labeled with a 10-dimensional feature vector, which is the concatenation of the one-hot encoded nucleotide vector (4D) and the one-hot encoded substructure vector (6D).

Subsequently, a graph neural network, as seen in Figure 2, is trained. At the core of the graph neural network architecture, there is a GenConv[27] layer that applies the message-passing formula described in Eq. 1.

$$x_i' = MLP(x_i + AGG(\{ReLU\,(x_j + e_{ji}) + \epsilon : j \in\ N\,(i)\}\,)) \tag{1}$$

where, x_i denotes a node-level embedding, $N\,(i)$ represents the set of neighbors for node i, e_{ji} indicates the edge features from node j to i, AGG is the aggregation function, ReLU is the activation function. MLP stands for Multi-Layer Perceptron Neural Networks. The readout layer is the concatenation of mean and max scatter, aggregating the graph into a 200-dimensional vector.

Encoding by splice Bert

Following encoding the secondary structure with the GNN Encoder, the primary sequence is encoded using Splice Bert. Splice Bert is a BERT-based model that was pre-trained on pre-mRNA sequences. We fine-tuned Splice Bert on our dataset, using sequences longer than 1024 nucleotides by splitting them into two halves, each containing 512 nucleotides of the first and last part of the sequence and then concatenated to each other. Sequences shorter than 1024 nucleotides were pad to 1024 and used directly with Splice Bert. For handling the unbalancing matter in the dataset, LGLoc utilized the Focal Loss function[28] as the loss function, which is represented by Eq. 2.

$$FL\,(p_t) = -(1 - p_t)^{\gamma}\,log\,(p_t) \tag{2}$$

In this equation, p_t is the predicted probability of the true class. The term $(1 - p_t)^{\gamma}$ acts as a modulating factor, where γ is a focusing parameter that reduces the relative loss for well-classified examples, allowing the model to focus more on hard, misclassified examples.

Encoding by k-mer Frequency and CKSNAP

LGLoc uses k-mer feature-based encoders; k-mer represents the normalized occurrence frequencies of *k* neighboring base pairs in the DNA or RNA sequence. Eq. 3 shows the 2-mer descriptor as an example.

$$f_t = \frac{m\,(t)}{N}, t \in\ \{AA, AC, AG, \ldots\ , TT\} \tag{3}$$

where *m*(*t*) represents the total number of the k-mer type *t*, *N* denotes the sequence length. For LGLoc, *k* is set as 2, 3, 4, 5, and 6 and combined, resulting in a 5456 (= 4^2+ 4^3+ 4^4+4^5+4^6)-dimensional (1360D) feature vector. Because these vectors are high-dimensional, we delete the highly correlated features by eliminating features that correlate up to 0.8. Following this, CKSNAP converts a DNA/RNA sequence into a numerical feature vector by computing the occurrence frequency of all possible *k*-spaced Nucleotide Pairs (KNP) along the sequence. For instance, in the sequence 'AXXTXXXG,' 'AT,' and 'TG' represent two- and three-spaced nucleotide pairs. The frequency of KNP can be defined as Eq. 4:

$$f\,(KNP) = \frac{m\,(KNP)}{N - k - 1}, k \in\ [0, k_{max}] \tag{4}$$

Where $m\,(KNP)$ represents the number of KNP along the sequence, and (N-k-l) represents the number of *KNP* along a sequence with length *N*. We kept k_{max} = 5, which generated a 96D feature vector. We Compute k-mer and CKSNAP features using iLearnPlus[29] software.

Feature Selection and Classification

In the feature selection phase, LGLoc employs the ANOVA algorithm based on the F-test. LGLoc calculates the F-score for each feature and then selects the top k features. These selected features are subsequently fed into the classifier stage, which comprises five Naive Bayes classifiers. These classifiers are trained using a one-vs-rest approach, and the final label is determined by selecting the class with the highest probability.

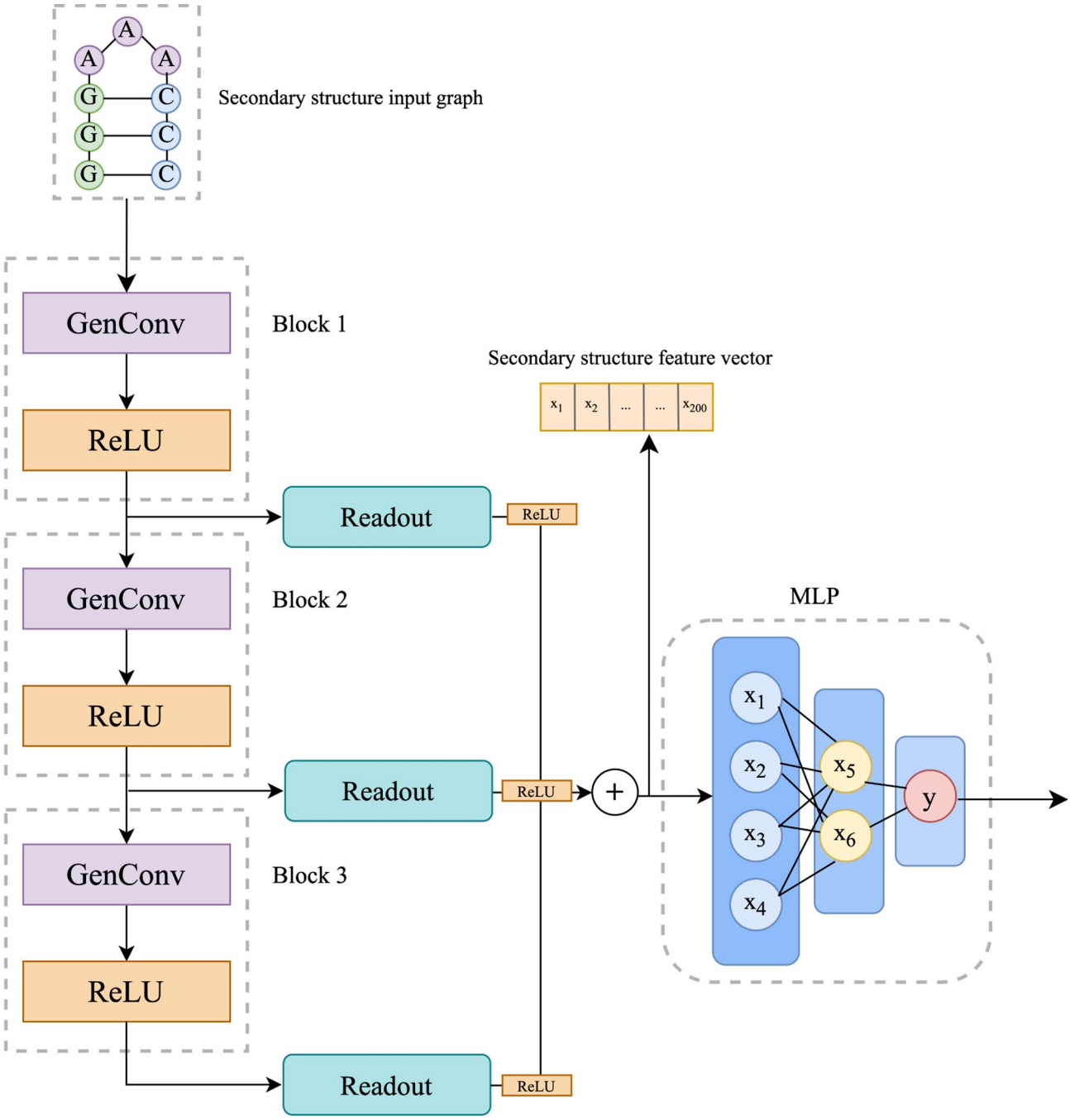

Figure 2. The GNN architecture for mRNA sequence graph embedding.

Assessments and Results

Assessment Conditions and Database

The present work uses the RNALocate database (version 2.0)[15] as the primary data source. This comprehensive repository encompasses location information pertinent to various RNA molecules, including mRNA, miRNA, and lncRNA. This database extracts an initial dataset comprising 28,829 mRNA sequences, each characterized by localization to one or more subcellular compartments. To ensure the precision of subcellular localization predictions and to facilitate optimal model design, the study is restricted to those mRNA sequences uniquely localized to a single compartment.

To mitigate the potential confounding effects of sequence homology and redundancy, mRNALoc utilized the NCBI BLASTCLUST program on this dataset. mRNALoc configured the application to select sequences exhibiting more than 70% full-length coverage and less than 40% homology, employing the '-S 40' and '-L 0.7' options. We used this final dataset, which has five categories of Cytoplasm, Endoplasmic region, Extra Cellular region, Mitochondria, and Nucleus, to ensure consistency and reliability in selecting sequences. Figure 3

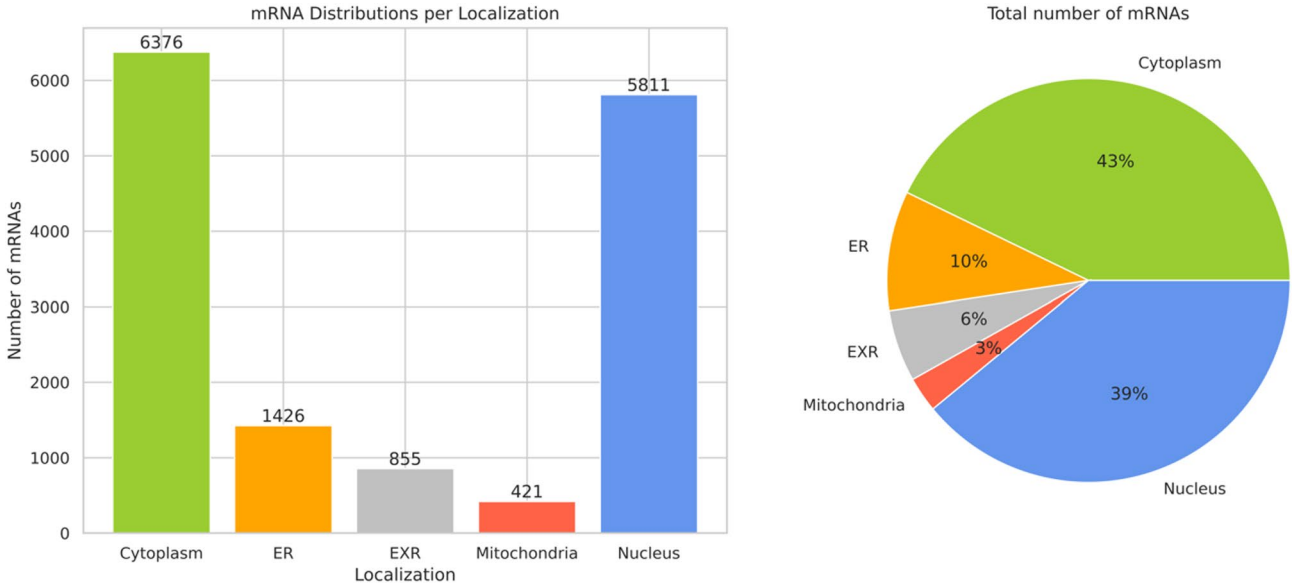

Figure 3. Details about the RNALocate dataset.

illustrates the mRNA distribution per subcellular localization, providing a visual summary of the dataset composition used in this work.

Assessment's Criteria

In the following, we compare LGLoc against two leading methodologies, mRNALoc and MSLP, using the performance evaluation metrics: Sensitivity (Eq. 5), Specificity (Eq. 6), Accuracy (Eq. 7), F1-score (Eq. 8), Area Under the Curve (ACC), and Matthews Correlation Coefficient (MCC) (Eq. 9). These metrics were chosen for their ability to quantify different aspects of model performance, ensuring a comprehensive evaluation. Sensitivity and Specificity measure the model's ability to correctly identify positive and negative cases. Accuracy provides an overall measure of correctness, while the F1-score balances precision and recall. MCC offers a balanced view, even in cases of class imbalance. The formulas for each metric are provided for clarity.

$$Accuracy = \frac{tp + tn}{tp + fn + fp + tn} \tag{5}$$

$$Sensitivity\ (Recall) = \frac{tp}{tp + fn} \tag{6}$$

$$Specificity = \frac{tn}{fp + tn} \tag{7}$$

$$F1 - score = \frac{2 \times\ Precision \times\ Recall}{Precision + Recall} \tag{8}$$

$$MCC = \frac{tp \times\ tn - fp \times\ fn}{\sqrt{(tp + fp)\,(tp + fn)\,(tn + fp)\,(tn + fn)}} \tag{9}$$

Where tp stands for true positive, tn for true negative, fp for false positive, and fn for False Negative.

Assessment Result

Accuracy

This section reports the Accuracy of three methods: mRNAloc, LGLoc, and MSLP. Based on Figure 4, in the Cytoplasm, LGLoc performs better than MSLP and mRNALoc, indicating its strong capability in this area. In the Endoplasmic Reticulum, LGLoc performs slightly better than the MSLP, but mRNALoc outperforms them. Indeed, LGLoc reduces the gap between MSLP and mRNALoc, but it still needs improvement. In the Extracellular Region, LGLoc maintains comparable Accuracy to MSLP but slightly outperforms it. At the same time, both of them outperform mRNALoc with a high difference. In the Mitochondria, LGLoc demonstrates high Accuracy, only slightly below mRNALoc, but both are better than MLSP. Finally, in the Nucleus, LGLoc exhibits impressive Accuracy, which is superior to MLSP and mRNALoc, reflecting its robustness and consistency. All explanations are summarized in Table 2 by computing each method's average and standard deviation of accuracy. This result shows that LGLoc outperforms the other methods, MLSP and mRNALoc, by increasing average Accuracy with a lower standard deviation.

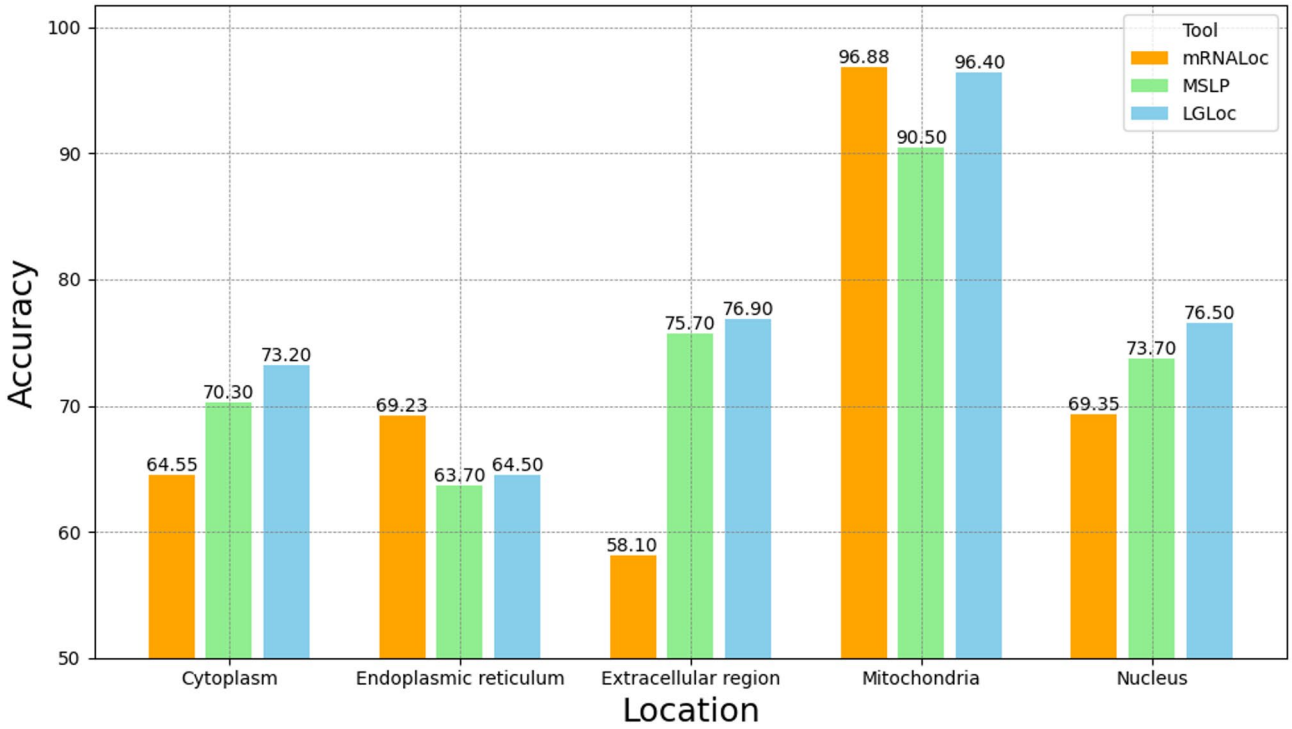

Figure 4. Accuracy comparison of mRNALoc, MSLP, and LGLoc across different cellular locations.

Methods	Average of accuracy	Standard deviation of accuracy
mRNALoc	71.622	14.85
MSLP	74.78	9.90
LGLoc	77.36	11.70

Table 2. Average and standard deviation of accuracy for each method.

Sensitivity

This section reports sensitivity, which shows the ability to detect the correct label for each class. According to Figure 5, in the Cytoplasm, LGLoc performs better than MSLP and slightly better than mRNALoc. For the Endoplasmic Reticulum, LGLoc again achieves better Sensitivity, but MSLP is overtaking mRNALoc this time. All of them are close together in the Extracellular Region, but LGLoc is in third place. The tool exhibits exceptionally high Sensitivity in the Mitochondria, closely matching MSLP and surpassing mRNALoc, highlighting its reliability in this critical category. In the Nucleus, LGLoc's Sensitivity, with a slight difference, occupies the second place after MSLP, but both outperform mRNALoc with observable differences.

In general, according to Table 3, the average and standard deviation of Sensitivity show that LGLoc takes the first place, and it can keep its Sensitivity in all classes in an acceptable range without any sharp decrementing in a class. It should be noted that while the difference between MLSP and LGLoc in terms of average Sensitivity is not high, and LGLoc improves it slightly, the critical point about LGLoc is that it improves the standard deviation of Sensitivity, too. So, in terms of Sensitivity, LGLoc generally makes classification more robust.

Specificity

For Specificity, according to Figure 6, in three categories, Cytoplasm, Endoplasmic Reticulum, and Mitochondria, LGLoc takes second place among ranking. Still, for other categories, including Extracellular Region and Nucleus, LGLoc is the vanguard. It should be noted that for claas Mitochondria, while LGLoc is in place two, its amount is so high and in the excellent range. Generally, according to Table 4, LGLoC has comparable and acceptable specificity for all classes compared to other methods; its overall Specificity is better than the other two methods, mRNALoc and MLSP, with relatively good standard deviation. So generally, LGLoc demonstrates a good balance in minimizing false positives across different contexts.

F1-score

The F1-score is a crucial metric that balances precision and recall, providing a single measure of a model's accuracy in scenarios where class distribution is imbalanced. It is beneficial in evaluating the performance of our tool, as it captures both the relevance and completeness of the classifications made. Since the F1-score was

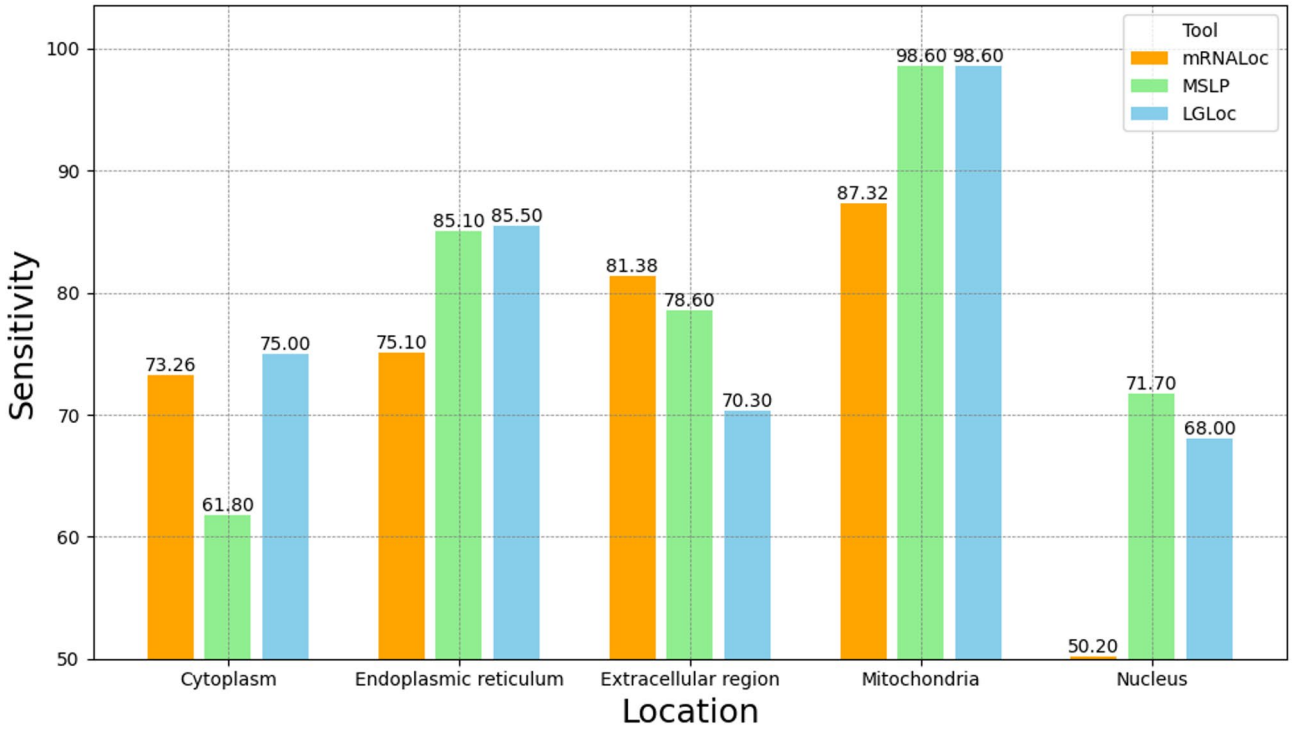

Figure 5. Sensitivity comparison of mRNALoc, MSLP, and LGLoc across different cellular locations.

Methods	Average of sensitivity	Standard deviation of sensitivity
mRNALoc	73.45	14.13
MSLP	79.16	13.88
LGLoc	81	11.67

Table 3. Average and standard deviation of sensitivity for each method.

just reported for mRNALoc, we also compared LGLoc with it. Based on Figure 7, in all five categories, LGLoc can surpass the mRNALoc method in the perspective of the F1-score, and among them for three categories, including Cytoplasm, Mitochondria, and Nucleus, its F1-score is in the acceptable range (more than 0.7). This superpassing can be observed in Table 5, too. According to this data, it can be seen that not only does LGLoc have a better average F1-score, but its lower standard deviation of F1-score also shows its reliability among all categories. Its strong F1-scores across these categories demonstrate its capability to provide balanced and accurate predictions, making it an effective tool for various applications where precision and recall are both critical.

AUC

The AUC is a vital metric in classification tasks, as it quantifies the model's ability to distinguish between classes across all possible thresholds, providing a comprehensive measure of performance that reflects both sensitivity and specificity. As in the previous section, because of reporting AUC just by mRNALoc, we compare LGLoc with this method. According to Figure 8, from the perspective of AUC, LGLoc outperforms mRNALoc in all categories again, but with the difference that both are in the acceptable range of AUC in all categories (more than 0.7). As far as the AUC of both of them can even be reached up to 0.98 in the Mitochondria category. While both methods have a good performance here, as mentioned and can be seen in Table 6, LGLoc has better results with higher reliability among all categories compared to mRNALoc.

MCC

As in the two previous sections, we compare LGLoc with mRNALoc here, but in this turn, it is from an MCC perspective. According to Figure 9, LGLoc outperforms mRNALoc in all categories. Of course, in some categories, such as Mitochondria, mRNALoc achieves MCC near LGLoc, but overall, as also mentioned in Table 7, LGLoc has better results with higher reliability again. Despite the huge performance improvement by LGLoc here, based on the nature of MCC criteria that reflect various aspects of methods, one thing that can be concluded from Figure 9 is that some categories, such as Endoplasmic Reticulum, need more improvement that can be considered in the future works.

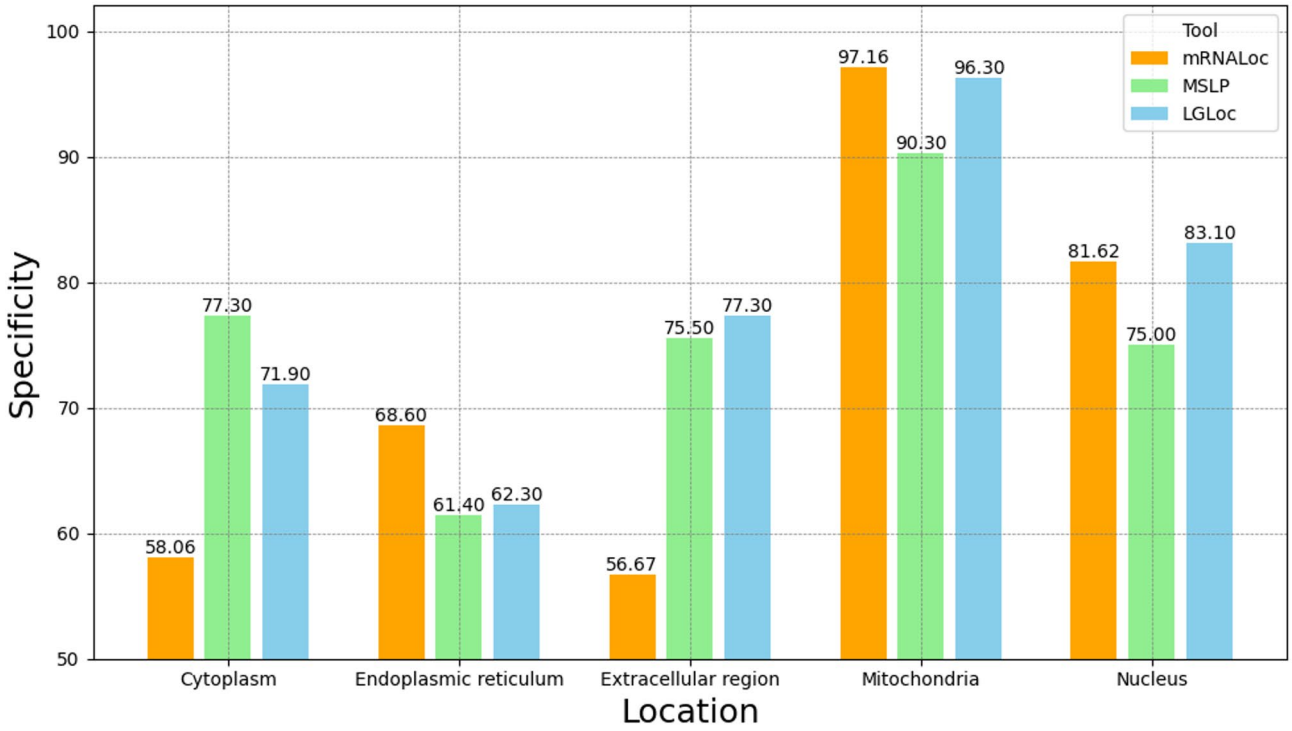

Figure 6. Specificity comparison of mRNALoc, MSLP, and LGLoc across different cellular locations.

Methods	Average of specificity	Standard deviation of specificity
mRNALoc	72.422	17.07
MSLP	75.9	10.25
LGLoc	77.92	12.73

Table 4. Average and standard deviation of specificity for each method.

Discussion and Conclusion

This paper proposes LGLoc, a classifier to predict mRNA sequences' quintet locations within a cell. LGLoc employs distinct feature sets tailored to each class, enhancing the model's performance compared with state-of-the-art methods. Indeed, LGLoc leverages hierarchical and structural features derived from BERT, and GNN encoders are utilized for classes including Cytoplasm, Endoplasmic Reticulum, Mitochondria, and Nucleus. For the class Extracellular Region, it incorporates k-mer features alongside CKSNAP. This targeted feature selection for each class, coupled with ANOVA as a feature selection approach and a one-vs-rest strategy, has proven effective for achieving optimal classification results. Actually, LGLoc can improve average performance in all classes and reduce the harsh differences between performance levels among classes. In other words, it has stable performance for all classes compared to state-of-the-art methods.

Quantitatively, LGLoc demonstrates an increase in the average of all six criteria—Accuracy, Sensitivity, Specificity, F1-score, AUC, and MCC—when compared to popular methods like mRNALoc and MSLP. Notably, LGLoc enhances average F1-score and average MCC, two metrics that assess the balanced performance of classification methods, by over 49% and 26%, respectively, compared to mRNALoc.

Despite all these improvements, more than the highest performance obtained by LGLoc is needed for some categories, such as the Endoplasmic Reticulum and Extracellular Region. So they need more attention in future works. For future work, several avenues for enhancement can be pursued based on the current model. Expanding the dataset by incorporating additional data, especially from diverse biological conditions and species, could significantly improve the model's predictive performance and broaden its applicability across various biological contexts. This would allow the model to generalize more effectively and provide more robust predictions.

Moreover, enhancing the explainability of the model is crucial for gaining deeper insights into the biological factors influencing mRNA localization. By increasing the model's transparency, we can better understand the underlying biological mechanisms, thus bridging the gap between computational predictions and biological interpretation. In addition, we plan to extend our model to support multi-label classification, enabling the prediction of multiple mRNA localization outcomes simultaneously.

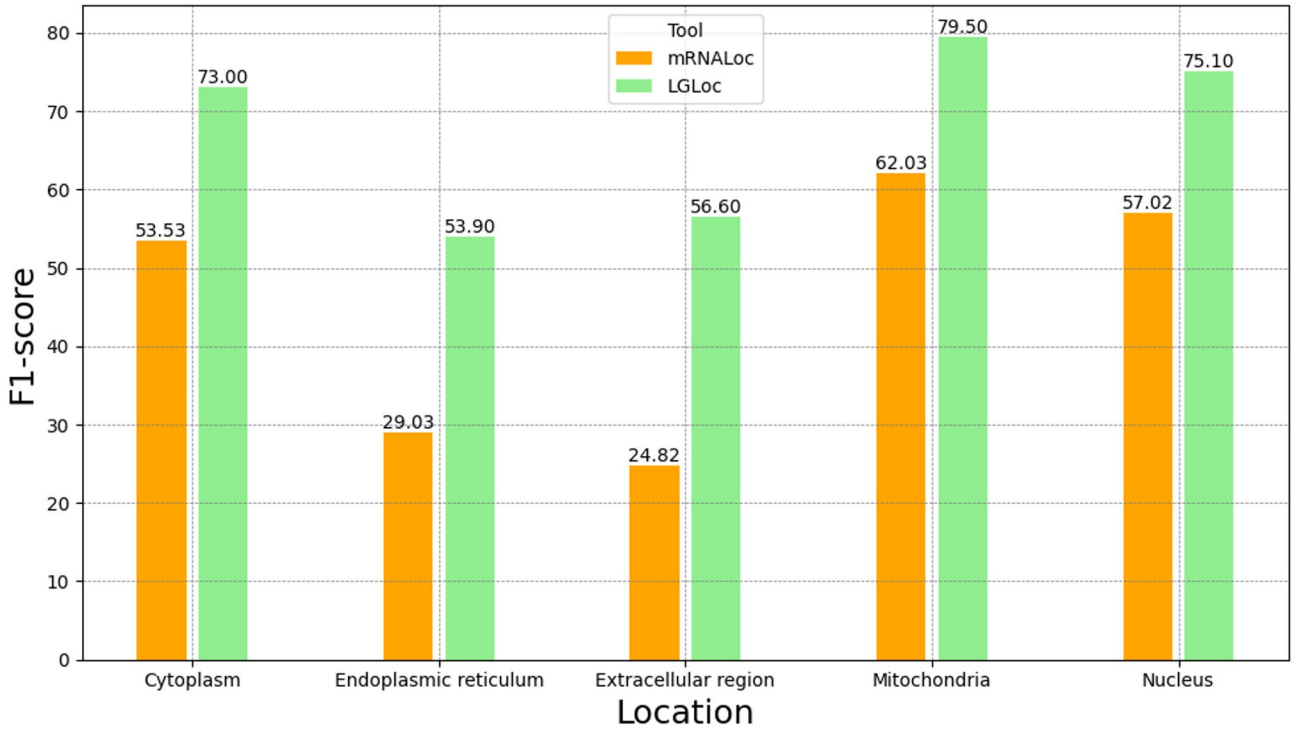

Figure 7. Comparison of F1-score between LGLoc and mRNALoc across various cellular locations.

Methods	Average of F1-score	Standard deviation of F1-score
mRNALoc	45.29	17.10
LGLoc	67.62	11.57

Table 5. Average and standard deviation of F1-score for each method.

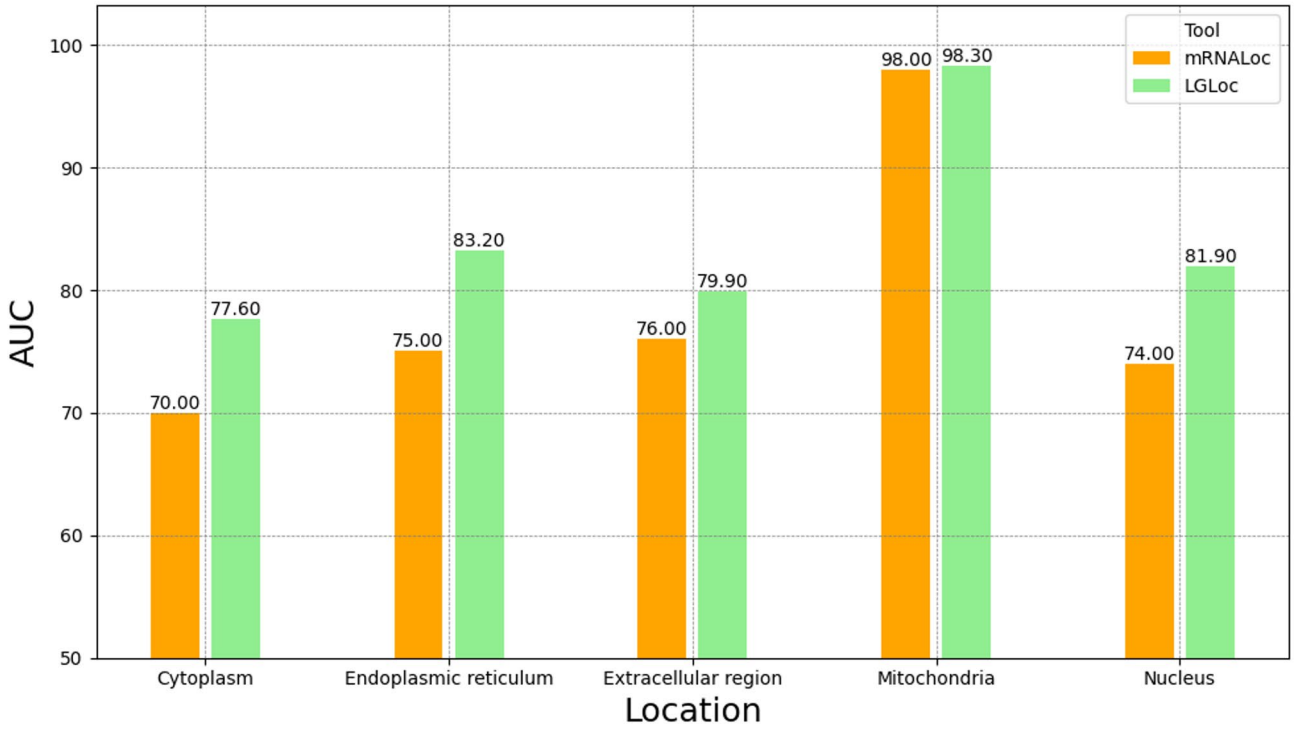

Figure 8. Comparison of AUC between LGLoc and mRNALoc across various cellular locations.

Methods	Average of AUC	Standard deviation of AUC
mRNALoc	78.6	11.08
LGLoc	84.96	7.84

Table 6. Average and standard deviation of AUC for each method.

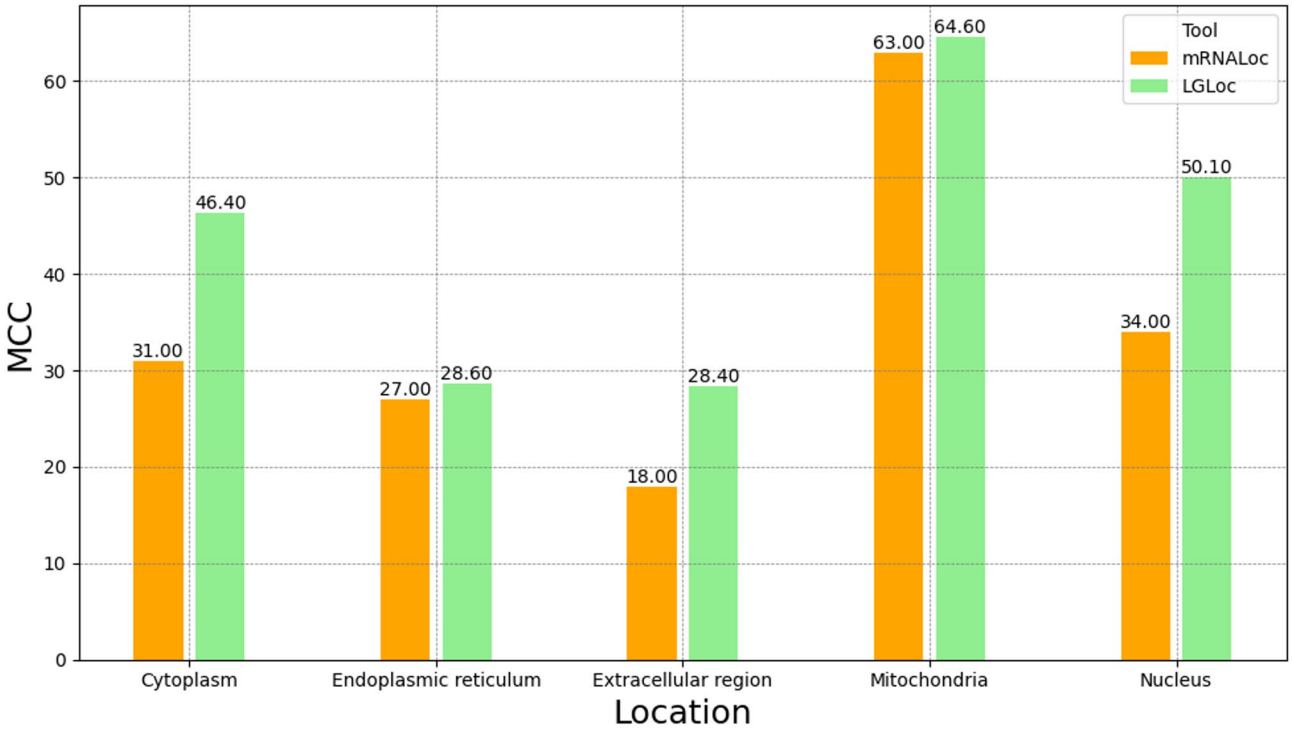

Figure 9. Comparison of MCC between LGLoc and mRNALoc across various cellular locations.

Methods	Average of MCC	Standard deviation of MCC
mRNALoc	34.60	16.98
LGLoc	43.62	15.39

Table 7. Average and standard deviation of MCC for each method.

Data availability

The code and supplementary materials for this paper are available in the GitHub repository at https://github.com/aref-shahbakhsh/LGLoc and mRNA localization_supplementary materials.docx file.

Received: 29 October 2024; Accepted: 20 May 2025
Published online: 28 May 2025

References

1. Engel, K. L., Arora, A., Goering, R., Lo, H. G. & Taliaferro, J. M. Mechanisms and consequences of subcellular RNA localization across diverse cell types. *Traffic* **21**, 404–418. https://doi.org/10.1111/tra.12730 (2020).
2. Martin, K. C. & Ephrussi, A. mRNA localization: Gene expression in the spatial dimension. *Cell* **136**, 719–730. https://doi.org/10.1016/j.cell.2009.01.044 (2009).
3. Jung, H., Yoon, B. C. & Holt, C. E. Axonal mRNA localization and local protein synthesis in nervous system assembly, maintenance and repair. *Nat. Rev. Neurosci.* **13**, 308–324. https://doi.org/10.1038/nrn3210 (2012).
4. Cooper, T. A., Wan, L. & Dreyfuss, G. RNA and disease. *Cell* **136**, 777–793. https://doi.org/10.1016/j.cell.2009.02.011 (2009).
5. Didiot, M. C. et al. Nuclear localization of Huntingtin mRNA is specific to cells of neuronal origin. *Cell. Rep.* **24**, 2553–2560. https://doi.org/10.1016/j.celrep.2018.07.106 (2018).
6. Pelekanou, V., Villarroel-Espindola, F., Schalper, K. A., Pusztai, L. & Rimm, D. L. CD68, CD163, and matrix metalloproteinase 9 (MMP-9) co-localization in breast tumor microenvironment predicts survival differently in ER-positive and -negative cancers. *Breast Cancer Res.* **20**, 154. https://doi.org/10.1186/s13058-018-1076-x (2018).
7. Liu, H. et al. DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Res.* **1** (1). https://doi.org/10.1093/nar/gkz1007 (2019).
8. Chen, J., McSwiggen, D. & Ünal, E. Single molecule fluorescence hybridization (smFISH) analysis in budding yeast vegetative growth and meiosis. *J. Visualized Experiments.* **135**, 774. https://doi.org/10.3791/57774 (2018).
9. 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* **118–119**, 101–110. https://doi.org/10.1016/j.ymeth.2016.09.010 (2017).
10. Alam, T., Al-Absi, H. R. H. & Schmeier, S. Deep learning in LncRNAome: Contribution, challenges, and perspectives. *Noncoding RNA* **6**, 47. https://doi.org/10.3390/ncrna6040047 (2020).
11. Khan, S., Khan, M., Iqbal, N., Khan, S. A. & Chou, K. C. Prediction of piRNAs and their function based on discriminative intelligent model using hybrid features into Chou's PseKNC. *Chemometr. Intell. Lab. Syst.* **203**, 104056. https://doi.org/10.1016/j.chemolab.2020.104056 (2020).

12. Khan, S., AlQahtani, S. A., Noor, S. & Ahmad, N. PSSM-Sumo: deep learning based intelligent model for prediction of sumoylation sites using discriminative features. *BMC Bioinform.* **25**, 284. https://doi.org/10.1186/s12859-024-05917-0 (2024).
13. Yan, Z., Lécuyer, E. & Blanchette, M. Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics* **35**, i333–i342. https://doi.org/10.1093/bioinformatics/btz337 (2019).
14. Garg, X. A., Singhal, N., Kumar, R. & Kumar, M. mRNALoc: a novel machine-learning based in-silico tool to predict mRNA subcellular localization. *Nucleic Acids Res.* **48**, W239–W243. https://doi.org/10.1093/nar/gkaa385 (2020).
15. Cui, T. et al. RNALocate v2.0: an updated resource for RNA subcellular localization with increased coverage and annotation. *Nucleic Acids Res.* **50**, D333–D339. https://doi.org/10.1093/nar/gkab825 (2022).
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.* **29**, 2617–2623. https://doi.org/10.1016/j.ymthe.2021.04.004 (2021).
17. Zhang, Z. Y. et al. Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief Bioinform.* **22**, 526–535. https://doi.org/10.1093/bib/bbz177 (2021).
18. 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.* **22**, 5. https://doi.org/10.1093/bib/bbaa401 (2021).
19. Musleh, S., Islam, M. T., Qureshi, R., Alajez, N. M. & Alam, T. MSLP: mRNA subcellular localization predictor based on machine learning techniques. *BMC Bioinform.* **24**, 109. https://doi.org/10.1186/s12859-023-05232-0 (2023).
20. Bartys, N., Kierzek, R. & Lisowiec-Wachnicka, J. The regulation properties of RNA secondary structure in alternative splicing. *Biochim. Biophys. Acta Gene Regul. Mech.* **1862**, 194401. https://doi.org/10.1016/j.bbagrm.2019.07.002 (2019).
21. Iuchi, H. et al. Representation learning applications in biological sequence analysis. *Comput. Struct. Biotechnol. J.* **19**, 3198–3208. https://doi.org/10.1016/j.csbj.2021.05.039 (2021).
22. Rivas, E. & Eddy, S. R. The language of RNA: a formal grammar that includes pseudoknots. *Bioinformatics* **16**, 334–340. https://doi.org/10.1093/bioinformatics/16.4.334 (2000).
23. Devlin, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North* 4171–4186. https://doi.org/10.18653/v1/N19-1423 (Association for Computational Linguistics, 2019).
24. Chen, K. et al. Self-supervised learning on millions of primary RNA sequences from 72 vertebrates improves sequence-based RNA splicing prediction. *Brief Bioinform* **25**, 3. https://doi.org/10.1093/bib/bbae163 (2024).
25. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26. https://doi.org/10.1186/1748-7188-6-26 (2011).
26. Thiel, B. C., Beckmann, I. K., Kerpedjiev, P. & Hofacker, I. L. 3D based on 2D: Calculating helix angles and stacking patterns using forgi 2.0, an RNA Python library centered on secondary structure elements. *F1000Research* **8**, 287. https://doi.org/10.12688/f1000research.18458.2 (2019).
27. Li, G., Xiong, C., Qian, G., Thabet, A. & Ghanem, B. DeeperGCN: training deeper GCNs with generalized aggregation functions. *IEEE Trans. Pattern Anal. Mach. Intell.* **1**, 1–12. https://doi.org/10.1109/TPAMI.2023.3306930 (2023).
28. Lin, T. Y., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 318–327. https://doi.org/10.1109/TPAMI.2018.2858826 (2020).
29. Chen, Z. et al. iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Res.* **49**, e60. https://doi.org/10.1093/nar/gkab122 (2021).

Author contributions

S. Koohi, S. Akbari Rokn Abadi, and A. Shahbakhsh conceived of the presented idea and developed the theory, and designed assessments. A. Shahbakhsh implemented the code. S. Akbari Rokn Abadi and A. Shahbakhsh analyzed the results and wrote the first version of the write-up. S. Akbari Rokn Abadi wrote the final version of the manuscript and S. Koohi revised it. All authors discussed the results and contributed to the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-03485-8.

Correspondence and requests for materials should be addressed to S.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.