Molecular Therapy
Nucleic Acids
Original Article

lncRNA localization and feature interpretability analysis

Jing Li,[1,2] Ying Ju,[3] Quan Zou,[4] and Fengming Ni[5]

[1]Department of Microbiology, University of Hong Kong, Hong Kong, China; [2]School of Biomedical Sciences, University of Hong Kong, Hong Kong, China; [3]School of Informatics, Xiamen University, Xiamen, China; [4]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324000, Zhejiang, China; [5]Department of Gastroenterology, The First Hospital of Jilin University, Changchun, China

Subcellular localization is crucial for understanding the functions and regulatory mechanisms of biomolecules. Long non-coding RNAs (lncRNAs) have diverse roles in cellular processes, and their localization within specific subcellular compartments provides insights into their biological functions and implications in health and disease. The nucleolus and nucleoplasm are key hubs for RNA metabolism and cellular regulation. We developed a model, LncDNN, for identifying the localization of lncRNAs in the nucleolus and nucleoplasm. LncDNN uses three different encoding schemes and employs Shapley Additive Explanations for feature analysis and selection. The results show that LncDNN is more accurate than other models. Additionally, an interpretable analysis of the features influencing the model was conducted. LncDNN is applicable for identifying the localization of lncRNA in the nucleolus and nucleoplasm, aiding in the understanding and in-depth study of related biological processes and functions.

INTRODUCTION

Subcellular localization plays a crucial role in understanding the functions and regulatory mechanisms of biomolecules within the complex cellular environments. In recent years, long non-coding RNAs (lncRNAs) have emerged as key players in cellular processes, exhibiting diverse functions ranging from gene regulation to structural organization.[1–4] The localization of these transcripts within specific subcellular compartments provides critical insights into their biological roles and their potential implications in health and disease.[5–8] Among the various subcellular compartments, the nucleolus and nucleoplasm are key hubs for RNA metabolism and cellular regulation.[9–12] The nucleolus, a distinct subnuclear organelle, is primarily associated with RNA processing and the regulation of cell-cycle progression.[13–15] In contrast, the nucleoplasm encompasses the entire nuclear content excluding the nucleolus and is involved in diverse cellular processes, including transcription, RNA splicing, and DNA repair.[16,17]

Aberrant lncRNA localization patterns of lncRNAs have been implicated in various diseases, highlighting the clinical significance of elucidating subcellular distribution.[5,18–21] Under normal conditions, the localization of lncRNAs in specific subcellular compartments reflects their specific functional roles. However, when lncRNAs are aberrantly localized to inappropriate subcellular regions, it can lead to a series of dysregulations in cellular processes.[22] Such abnormal localization is often associated with various diseases, including cancer, neurodegenerative disorders, and other complex diseases, highlighting the importance of precise subcellular regulation in maintaining normal cellular function.[23] For instance, aberrant lncRNA localization can lead to dysregulated ribosome biogenesis, which is a hallmark of cancers.[24]

Under normal conditions, the localization of lncRNAs in specific subcellular compartments reflects their specific functional roles. However, when lncRNAs are mislocalized to inappropriate subcellular regions, it can lead to a series of dysregulations in cellular processes. Such abnormal localization is often associated with various diseases, including cancer, neurodegenerative disorders, and other complex diseases, highlighting the importance of precise subcellular regulation in maintaining normal cellular function.

Understanding the subcellular localization patterns of lncRNAs in relation to the nucleolus and nucleoplasm is of paramount importance.[25,26] The localization of these transcripts to specific subcellular compartments reflects the functional relevance and regulatory roles within the cell. The nucleolus is crucial for ribosomal RNA synthesis and ribosome assembly, which are fundamental to cellular protein production.[27] Localization of lncRNAs in the nucleolus often indicates their involvement in ribosome biogenesis and cell-cycle regulation. Nucleolar lncRNAs are involved in ribosome biogenesis and have been found to be overexpressed in certain cancers, suggesting their role in promoting tumor cell growth.[28,29] Nucleoplasmic lncRNAs play a critical role in transcriptional regulation and RNA splicing, and their dysregulation is associated with the pathology of neurodegenerative diseases.[30,31] The localization of lncRNAs in nucleolus or nucleoplasm thus reflects distinct regulatory mechanisms and biological functions.

Traditional laboratory methods for obtaining the localization of lncRNA in the nucleolus and nucleoplasm typically require a

Received 26 September 2024; accepted 10 December 2024;
https://doi.org/10.1016/j.omtn.2024.102425.

Correspondence: Fengming Ni, Department of Gastroenterology, The First Hospital of Jilin University, Changchun, China.
E-mail: n_fengming@jlu.edu.cn

significant amount of time, manpower, and resources. For example, fluorescence *in situ* hybridization requires specialized equipment. Moreover, due to experimental conditions and resource limitations, it may not be feasible to handle large-scale samples. In contrast, machine learning-based models for predicting lncRNA localization in the nucleolus and nucleoplasm can make faster predictions, saving time and human resources. Additionally, machine learning methods do not require large amounts of experimental materials and equipment, resulting in relatively lower costs. Importantly, machine learning methods can handle large-scale datasets, thereby providing more comprehensive analyses.[32] Through machine learning methods, the exploration of the potential mechanisms and correlations of lncRNA in the nucleolus and nucleoplasm can help elucidate the cellular functions and mechanisms of disease occurrence.

In recent years, several studies on RNA subcellular localization have been published. Zhang et al.[26] published a model for RNA subcellular localization. The model was trained using a support vector machine approach, integrating a mutual information algorithm and an incremental feature selection strategy to address issues like low discriminative power and overfitting. Tree-based stacking approach for cell-specific lncRNA subcellular localization (TACOS)[33] combines tree-based classifiers and feature descriptors to predict the subcellular localization of human lncRNAs across 10 cell types. TACOS integrates AdaBoost baseline models through a stacking approach for improved prediction accuracy. SubLocEP[23,34] is a two-layer integrated prediction model that comprehensively considers additional feature attributes and combines them with LightGBM for accurate and stable prediction of eukaryotic mRNA subcellular localization, overcoming the limitations of existing models.

RNAlight,[35] developed using LightGBM, identifies nucleotide k-mers linked to subcellular localizations of mRNAs and lncRNAs. It employs the Tree SHAP (Shapley Additive Explanations) algorithm to extract features determining RNA localization, revealing the sequence basis. Additionally, RNAlight maps features to known RNA-binding protein motifs, uncovering associations with distinct subcellular localizations. DeepLncLoc,[36] a deep learning framework, predicts lncRNA subcellular localization by introducing a novel subsequence embedding method to retain sequence order information. It divides sequences into consecutive subsequences, extracts patterns from each subsequence, and combines those subsequences to represent the full sequence. GraphLncLoc,[37] a graph convolutional network-based deep learning model, predicts lncRNA subcellular localization by transforming lncRNA sequences into de Bruijn graphs. GraphLncLoc then employs graph convolutional networks to extract high-level features from these graphs and uses a fully connected layer for prediction. DeepmRNALoc,[38] a deep neural network-based method, predicts eukaryotic mRNA subcellular localization by utilizing a two-stage feature extraction strategy. This approach involves bimodal information splitting and fusing in the first stage and convolutional neural network module in the second stage. GM-lncLoc,[39] a novel prediction model for lncRNA subcellular localization, combines low-level sequence information with graph structure information to extract high-level features. Additionally, GM-lncLoc utilizes meta-learning to obtain meta-parameters, enabling rapid learning of parameters for similar tasks and addressing the few-samples problem. So far, no dedicated tool for identifying the subcellular localization of mRNA in the nucleolus and nucleoplasm has been developed. Our primary objective in this study is to fill this gap.

In this study, we performed a comprehensive analysis of lncRNA localization patterns. Using machine learning approaches, we have developed predictive models for the classifying lncRNAs in nucleolus or nucleoplasm. By integrating the features from different levels, we trained LncDNN, and the advantages are as follows:

(1) This model focuses on identifying the localization of lncRNA in the nucleolus and nucleoplasm.
(2) By analyzing features extracted at different levels, it was found that the 12 features obtained from feature fusion and feature selection of enhanced nucleic acid composition (ENAC),[40] binary,[40] and accumulated nucleotide frequency (ANF)[40] can effectively identify the localization of lncRNA in the nucleolus and nucleoplasm.
(3) Experimental results demonstrate that LncDNN has good predictive and generalization ability, with area under the curve (AUC) scores of 0.873 and 0.831 on the validation and test sets, respectively.

RESULTS

Model development

To effectively identify the localization of lncRNA in the nucleolus and nucleoplasm, features at different levels of lncRNA were extracted. These features include nucleic acid composition, pseudo-nucleotide composition, nucleic acid composition, and binary. To ensure the validity of the features, feature selection was employed. Considering the ability of random forest to handle high-dimensional data and capture nonlinear relationships,[41] we chose random forest to identify the localization of lncRNA in the nucleolus and nucleoplasm. Owing to the slight imbalance present in this dataset, the AUC was selected as the primary optimization objective for model performance on the validation dataset.[42–45] Using random forest to evaluate different subsets of features with AUC as the optimization objective, it was found that ENAC, binary, and ANF had the highest AUC among the single features. Therefore, ENAC, binary, and ANF were subsequently retained for further analysis (Figure 1B).

Feature fusion can help capture information that a single feature cannot express. Feature selection helps remove redundant and irrelevant features, reducing noise. The combination of feature fusion and feature selection can help improve model performance, reduce overfitting, and enhance model interpretability. We fused ENAC, binary, and ANF to form a more comprehensive feature set. Then, SHAP[46] was used to analyze the features. After ranking features by mean absolute SHAP values[47] (MASVs), it was found that the model trained with the top 12 features (Figure 1C) had strong predictive power and generalization ability. This model is called LncDNN.

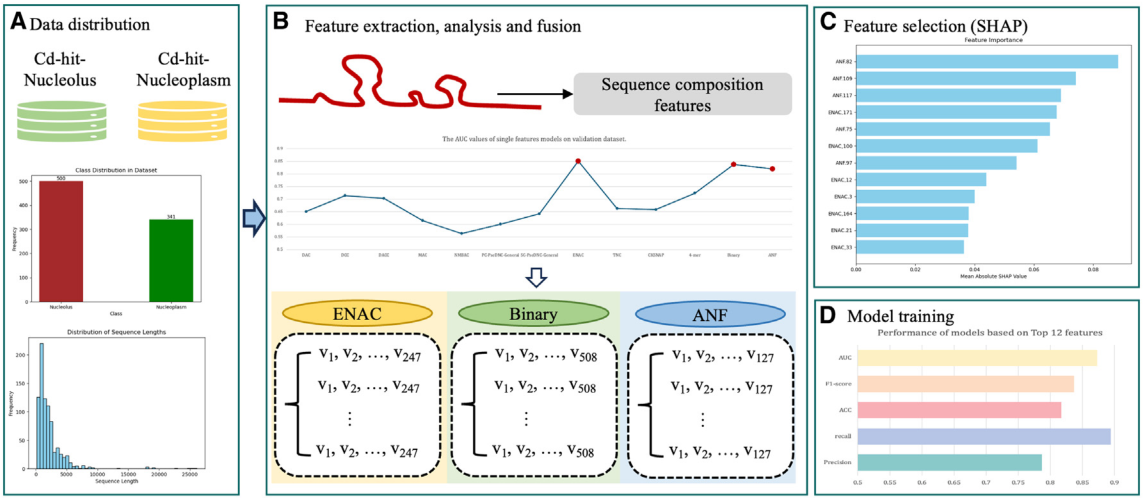

Figure 1. The overall framework of LncDNN

(A) Data process. (B) Feature extraction, analysis, and fusion. (C) Feature selection by using SHAP (top 12 features). (D) Model training (top 12 features).

Model performance and validation

Performance of individual features

In the evaluation of feature subsets in the autocorrelation feature group, metrics such as precision, recall, accuracy (ACC), F1 score, and AUC were calculated. The average AUC for sub-features in the autocorrelation group was 0.649 (Table 1). In the pseudo-nucleotide composition feature group, the average AUC for general parallel correlation pseudo-dinucleotide composition (PC-PseDNC-General) and general series correlation pseudo-dinucleotide composition (SC-PseDNC-General) was 0.622 (Table 1). In the nucleic acid composition group, the average performance metrics for sub-features were precision: 0.634, recall: 0.909, ACC: 0.678, F1 score: 0.747, and AUC: 0.743. The AUC for ENAC and ANF were 0.101 and 0.076 higher than the average AUC, respectively. Additionally, the AUC for binary was 0.838 (Table 1). In summary, the AUCs for the single-feature models of ENAC, binary, and ANF all exceeded 0.80.

Model performance after feature fusion and feature selection

We directly fused ENAC, binary, and ANF to train the model, and found that the AUC of the model on the validation set was 0.827, which is 0.011 lower than the AUC of ENAC on the validation set (Table 1). Then, we used SHAP to perform feature selection on the combined feature set of ENAC, binary, and ANF (Figure 1C). Figure 2 presents the top 20 MASV features. To obtain a better-performing model, we analyzed and evaluated the top 20 features and found that selecting the top 12 features (LncDNN) resulted in the highest AUC on the validation set, reaching 0.873, which exceeded all other models (Figures 2 and 3; Table 2).

The proposed LncDNN outperforms other models on independent test set

To demonstrate the advantages of LncDNN in identifying the localization of lncRNA in the nucleolus and nucleoplasm, we compared LncDNN with single-feature models (including ENAC, binary, and ANF) and models with MASV top 13/18/20 features on an independent test set (Table 3). In the single-feature models, the average AUC of ENAC, binary, and ANF on the validation set was 0.795, which is 0.036 lower than LncDNN (Table 3). The AUCs of models with MASV top 12/18/20 features (Figure 2) on the independent test set were 0.813, which is 0.018 lower than LncDNN (Table 3). In summary, LncDNN achieved a higher AUC on both the validation and independent test sets compared to other models, indicating its superior performance in identifying the localization of lncRNA in the nucleolus and nucleoplasm.

DISCUSSION

Our study optimized the model by applying feature fusion and feature selection on ENAC, binary, and ANF. This approach not only addressed the limitations of individual features but also removed redundant features, significantly improving model accuracy. The proposed model, LncDNN, demonstrates superior performance, with an AUC of 0.873 on the validation set and 0.831 on the independent test set, outperforming other models including those based on single features and various top feature subsets. This indicates the robustness and generalization ability of LncDNN in identifying the subcellular localization of lncRNAs. Our findings contribute to the understanding of subcellular localization patterns of lncRNAs, which is crucial for elucidating their functional roles and regulatory mechanisms within the cell. Additionally, this research provides valuable insights for studies related to diseases associated with aberrant lncRNA localization. However, we acknowledge that there are also limitations to this work. Although LncDNN applies feature fusion and selection, which effectively enhances model performance, there is still the issue of insufficient information. The training and testing data for our model come from RNALocate version 2.0, which, while being a comprehensive database, still has limitations regarding data sources. The model may perform poorly under different contexts and experimental

Table 1. The performance of single features of random forest models on valid dataset

Features	Precision	Recall	ACC	F1 score	AUC
Autocorrelation					
DAC	0.584	0.894	0.611	0.707	0.651
DCC	0.630	0.879	0.669	0.734	0.714
DACC	0.617	0.879	0.651	0.725	0.703
MAC	0.593	0.773	0.603	0.671	0.615
NMBAC	0.574	0.818	0.587	0.675	0.564
Pseudo-nucleotide composition					
PC-PseDNC-General	0.584	0.788	0.595	0.671	0.601
SC-PseDNC-General	0.621	0.818	0.6429	0.706	0.642
Nucleic acid composition					
ENAC	0.670	0.924	0.722	0.777	0.851[a]
TNC	0.621	0.894	0.659	0.733	0.663
CKSNAP	0.606	0.864	0.635	0.713	0.659
4-mer	0.617	0.879	0.651	0.725	0.724
ANF	0.657	0.985	0.722	0.788	0.820[a]
Binary encoding					
Binary	0.670	0.985	0.738	0.798	0.838[a]
Only feature fusion					
ENAC + binary + ANF	0.649	0.955	0.706	0.773	0.827

[a]The single-feature model's AUCs on the validation set exceeds 80%.

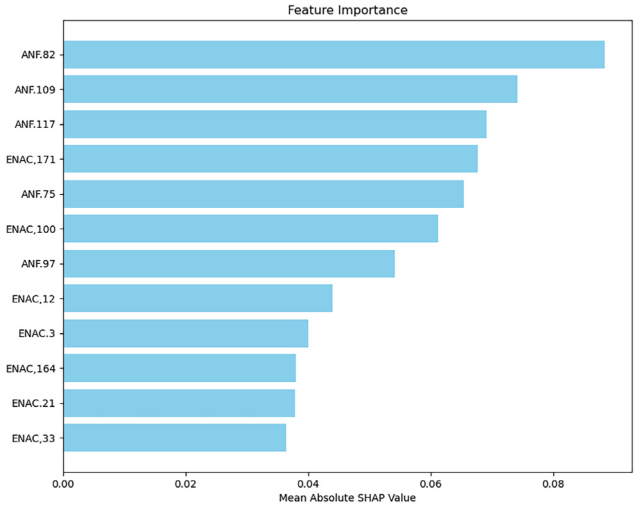

Figure 2. The importance of the MASV top 20

This figure displays the importance of features as meansured by Mean Absolute SHAP Value. SHAP values (SHapley Additive exPlanations) quantify the contribution of each featres to the model's prediction, providing a meansure of feature importance.

conditions, posing a risk of insufficient generalizability. In addition, our study lacks wet-lab experiments to support the model's findings. Therefore, in the future, we will focus on expanding feature types, improving feature extraction and enhancing dataset diversity. We aim to further improve model performance by adding new types of features and employing more advanced feature extraction methods. We will also work on improving the model's generalizability to different contexts by incorporating lncRNA data from various species and tissues. Additionally, if laboratory conditions permit, we will conduct experiments to validate the superiority of the LncDNN.

MATERIALS AND METHODS

Data description

This study harnessed the extensive RNALocate 2.0 database,[48] which provides a comprehensive collection of nucleic acid sequences, including lncRNAs, mRNAs, small nucleolar RNAs, and small noncoding RNAs, across a multitude of species. This database is accessible at http://www.rnalocate.org/ or http://www.rna-society.org/rnalocate/. RNALocate 2.0 is an RNA subcellular localization database that integrates data from various sources and experimental validation. It covers over 213,000 RNA localization entries across 104 species. These data are derived from manual curation of the literature, five additional databases, and 35 RNA sequencing datasets. Although RNALocate 2.0 offers a broad dataset, this study specifically focused on the lncRNA sequences because of their significant roles in various biological processes. Within the scope of lncRNA analysis, we focused on data from two key species: humans and mice. The datasets compiled for this investigation comprised 11,616 human samples and 343 mouse samples, thereby allowing us to delve into the investigation of lncRNA functions and their evolutionary aspects in these organisms (Figure 4).

In a comprehensive analysis of lncRNA data, a varied distribution of lncRNA localizations across cellular components was observed, reflecting the complexity and diversity of lncRNA functions. The dataset encompasses a broad range of locations, including the exosome (5,574 samples); nucleus (3,804 samples); nucleoplasm (969 samples); nucleolus (700 samples); chromatin (897 samples); cytoplasmic regions such as the cytoplasm (912 samples), cytosol (760 samples), and insoluble cytoplasm (96 samples); and other specific sites like the membrane (327 samples) and ribosome (143 samples). Given the significant role of lncRNAs in gene regulation, their localization can offer valuable insights into their functional mechanisms within the cells (Figure 4).

Given the pivotal roles played by the nucleolus and nucleoplasm in the regulation of genetic material and RNA processing, we focused on the localization of lncRNAs within these two cellular compartments. The nucleolus, known as the site of ribosomal RNA synthesis, and the nucleoplasm, which serve as the milieu for a variety of nuclear processes, are critical for understanding the regulatory landscapes of lncRNAs.[49,50] This study narrowed the focus to the nucleolus and nucleoplasm, accounting for 1,669 samples in the dataset, and further refined the data through sequence similarity clustering. Using Cluster Database at High Identity with Tolerance (cd-hit)[51] with a threshold set of 90% to ensure the uniqueness of the sequences, the dataset was

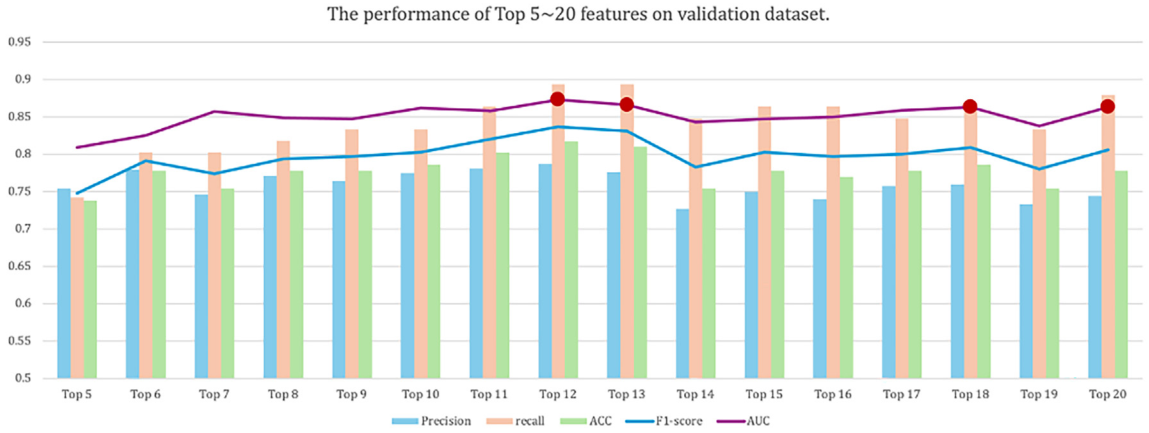

Figure 3. The performance of the model with the MASV top 5 to 20 features on the validation dataset

This figure illustrates the performance of the model across various subsets of features from top 5 to 20 as ranked by MASV, including metrics such as Precision, Recall, Accuracy (ACC), F1-score, and Area Under the Curve (AUC). Precision: The proportion of positive identifications that were actually correct. Recall: The proportion of actual positives that were correctly identified. ACC: The ratio of correctly predicted observations to the total observations. F1-score: The harmonic mean of Precision and Recall. AUC: The area under ROC curve, reflecting the overall performance of the model.

effectively distilled into a more manageable and representative subset. This process resulted in the identification of 500 unique samples localized to the nucleolus and 341 unique samples associated with the nucleoplasm (Figures 1A and 4). The data supportintg this study are availabel at http://github.com/lijingtju/LncDNN, providing across to the refined datasets and associated analysis. In this study, we conducted a detailed statistical analysis of the lengths of lncRNA sequences in the dataset to better understand the data characteristics. Figure 1A shows the distribution of sequence lengths. The results indicate that most lncRNA sequences are between 500 and 2,000 bases in length, with the highest frequency around 1,000 bases. Additionally, some longer sequences, up to 25,000 bases, were observed, although they are fewer in number. We divided the data processed by cd-hit into training, validation, and test sets in a ratio of 70%, 15%, and 15%, respectively. This partitioning ensures that the model can learn and be evaluated on different datasets, thereby improving its prediction capability and reliability on new data.

Feature extraction

Feature extraction techniques were used to analyze the lncRNA sequences and their subcellular localization. These techniques were divided into three hierarchical levels to capture the intrinsic and compositional features of the lncRNAs. Employing these established feature extraction techniques, we aimed to construct a multidimensional feature space that accurately represented the complexities of lncRNA sequences for effective classification in a biological context.

Nucleic acid composition features

Tri-nucleotide composition (TNC)[40] reflects tetra-nucleotide frequencies. ENAC[40] enriches the representation by considering positional information. ENAC captures the overall proportion of each nucleotide (A, U, G, C) in the nucleic acid sequence, reflecting the basic chemical characteristics and composition of the sequence. k-Spaced nucleic acid pairs (CKSNAP)[40] captures complex patterns and spacing effects by counting pairs of nucleotides separated by k spaces within a sequence. The 4-mer[40] represents all possible combinations of four adjacent nucleotides in the lncRNA sequences. These 4-mer patterns serve as features to capture important sequence information, enabling the characterization and classification of lncRNA sequences based on the underlying nucleotide composition. ANF[40] creates a cumulative frequency profile across the sequence. ANF is a cumulative nucleotide frequency-encoding method that calculates

Table 2. The performance of selected features on invalidation dataset

Features	Precision	Recall	ACC	F1 score	AUC
Top 5	0.754	0.742	0.738	0.748	0.809
Top 6	0.779	0.803	0.778	0.791	0.825
Top 7	0.746	0.803	0.754	0.774	0.857
Top 8	0.771	0.818	0.778	0.794	0.849
Top 9	0.764	0.833	0.778	0.797	0.847
Top 10	0.775	0.833	0.786	0.803	0.862
Top 11	0.781	0.864	0.802	0.820	0.858
LncDNN = top 12	0.787	0.894	0.817	0.837	0.873[a]
Top 13	0.776	0.894	0.810	0.831	0.866[a]
Top 14	0.727	0.848	0.754	0.783	0.843
Top 15	0.75	0.864	0.778	0.803	0.847
Top 16	0.740	0.864	0.770	0.797	0.850
Top 17	0.757	0.848	0.778	0.800	0.859
Top 18	0.76	0.864	0.786	0.809	0.863[a]
Top 19	0.733	0.833	0.754	0.780	0.838
Top 20	0.744	0.879	0.778	0.806	0.863[a]

[a] The model's AUC on validation set exceeds 86%.

Table 3. The performance of selected features on independent test set

Features	Precision	Recall	ACC	F1 score	AUC
Single-features models					
ENAC	0.706	0.935	0.724	0.804	0.814
ANF	0.716	0.948	0.740	0.816	0.790
Binary	0.682	0.948	0.701	0.793	0.780
Models based on top 12/13/18/20 features					
LncDNN = top 12	0.75	0.818	0.724	0.783	0.831[a]
Top 13	0.759	0.779	0.717	0.769	0.822
Top 18	0.733	0.818	0.709	0.773	0.822
Top 20	0.703	0.831	0.685	0.762	0.794

[a]LncDNN achieves the highest AUC o the test set.

the cumulative occurrences of each nucleotide in the sequence to reflect the global characteristics of the sequence. The cumulative distribution of nucleotides provides additional information on nucleotide distribution trends, thus providing richer features for the model. This is crucial for capturing the global patterns of the sequence.

Autocorrelation features

At an initial stage, autocorrelation features were extracted to quantify the internal structure and inter-nucleotide relationships within the lncRNA sequences. Dinucleotide-based auto covariance (DAC)[40] measures the variation in the properties of dinucleotide pairs within a sequence over distances, thereby capturing the structural tendencies of sequence. Dinucleotide-based cross-covariance (DCC)[40] evaluates the covariance between different dinucleotide properties or sequences, which is useful for understanding sequence interactions. Dinucleotide-based auto-cross-covariance (DACC)[40] combines DAC and DCC, analysis within-sequence variations and between-sequence interactions, and offers a detailed view of the sequence features. Moran autocorrelation (MAC)[52] was utilized to evaluate spatial correlations, and normalized Moreau-Broto autocorrelation (NMBAC)[52] normalized these autocorrelations to enable comparison across varying sequence lengths.

Pseudo-nucleotide composition features

The third level focuses on the global sequence order information by computing the pseudo-nucleotide composition features. This included PC-PseDNC-General,[52] which encompasses dinucleotide composition and the physicochemical properties of nucleotides, and SC-PseDNC-General,[52] which integrates sequence order effects with the dinucleotide composition to capture the sequential correlation information.

Binary encoding

The binary encoding[40] scheme represents each nucleic acid as a four-dimensional binary vector. Binary encoding captures the positional information of each nucleotide in the sequence, preserving the sequential characteristics of the sequence.

Feature analysis

Using random forest to train single features, it was found that among single-feature models, ENAC, binary, and ANF had AUCs exceeding 0.8 (Figure 1B; Table 1). Subsequently, ENAC, binary, and ANF were concatenated horizontally to generate a comprehensive feature set (Figure 1B). However, the AUC of the model directly trained on the concatenated features was only 0.827 on the validation set, lower than the AUC of the ENAC single-feature model (Table 1). This performance decline may be due to redundancy among features, which reduces the effective information density of the model. To obtain a better-performing model, SHAP was used to analyze the ENAC + binary + ANF feature set (Figure 1C). Based on the MASVs, we selected the top 20 features for further analysis (Figure 2). The figure shows the training results of the top 5 to 20 features, and it was found that the top 12 features (Figures 1C and 2) based on MASV (LncDNN) exhibited the highest AUC of 0.87 on the validation set (Figure 3; Table 2). Similarly, according to Table 3, LncDNN also demonstrated the highest AUC on the independent test set. This indicates that these features exhibit strong complementarity, effectively enhancing the performance of the model. Through an in-depth analysis of feature complementarity, we found that ENAC, binary, and ANF capture different sequence information. ENAC provides the overall composition, binary retains positional

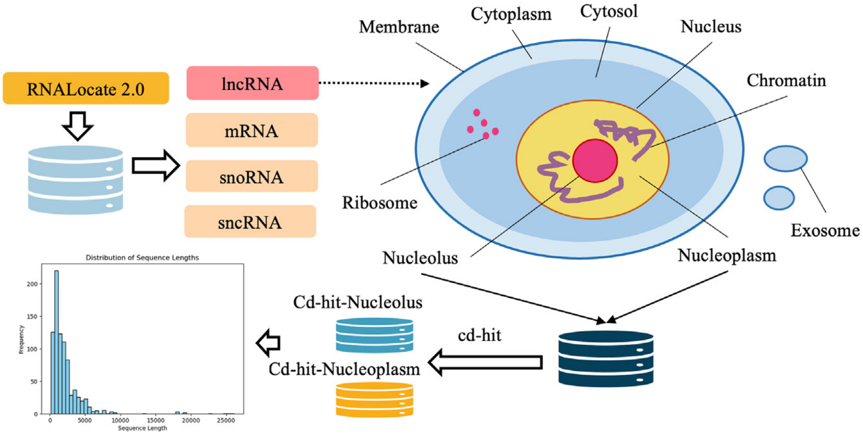

Figure 4. Data collection and process

information, and ANF offers a cumulative perspective. Their combination helps the model understand sequence characteristics from multiple dimensions, thereby improving prediction accuracy.

We used SHAP to analyze the combined features of ANF + binary + ENAC and found that the top 12 features with the highest MASVs (Figure 4), ANF.82, ANF.75, ANF.109, ANF.97, and ANF.117, represent the cumulative nucleotide frequencies up to positions 83, 76, 110, 172, 98, and 118 in the RNA sequence, with MASV values of 0.054, 0.0252, 0.0241, 0.0155, and 0.0151. This indicates that the cumulative nucleotide frequencies at these positions play a crucial role in our model. Additionally, the top 12 MASVs also included ENAC.33, ENAC.77, ENAC.164, ENAC.100, ENAC.3, ENAC.21, and ENAC.12, which represent the nucleotide frequencies at positions 34, 78, 165, 101, 4, 22, and 13 in the RNA sequence, respectively. According to SHAP analysis, the MASVs for ENAC.33, ENAC.77, ENAC.164, ENAC.100, ENAC.3, ENAC.21, and ENAC.12 were 0.044, 0.027, 0.0213, 0.0183, 0.0161, 0.0157, and 0.0135, which are significantly higher than those of other features. In other words, the MASVs top 12 features not only improve the AUC of the model but also reduce model complexity.

Evaluation metrics

Several evaluation indicators were employed to assess the performance of the classifier, including precision, recall, ACC, F1 score, and AUC.[53–56] These indicators comprehensively evaluate the ability of the classifier to discriminate between the nucleolus and the nucleoplasm. Precision measures the accuracy of the positive predictions of a classifier. It quantifies the ratio of correctly predicted nucleolus samples to all the samples predicted as nucleolus. Precision quantifies the accuracy of the positive predictions of the LncDNN within the nucleolus samples. Recall measures the ability of a classifier to identify nucleolus samples correctly. F1 score is the harmonic mean of the precision and recall, providing a balance between the two indicators. The AUC represents the area under the receiver operating characteristic curve, which plots the true positive rate (sensitivity) against the false positive rate (1 − specificity). This provides an overall measure of the ability of the classifier to discriminate between the nucleolus and nucleoplasm across different threshold values.[57,58]

$$Precision = \frac{TP}{TP+FP} \quad \text{(Equation 1)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{(Equation 2)}$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(Equation 3)}$$

$$F1\ score = \frac{2 * Precision * SN}{Precision+SN} \quad \text{(Equation 4)}$$

where *TP* represents true positives, the number of samples correctly classified as nucleolus; *FN* represents false negatives, the number of samples incorrectly classified as nucleoplasm; *TN* represents true negatives, the number of samples correctly classified as nucleoplasm; and *FP* represents false positives, the number of samples incorrectly classified as nucleolus.

DATA AND CODE AVAILABILITY

The data, code, and models from this study are openly accessible on Github (https://github.com/lijingtju/LncDNN.git). This repository enables researchers to access the datasets, utilize the code, invoke the models developed, and conduct predictions.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grant nos. 62102063 and 62072385).

AUTHOR CONTRIBUTIONS

F.N. and Q.Z. were the initiators and key authors of the manuscript. J.L. contributed significantly to the design and execution of the project and was the primary author of the paper. Y.J. assisted in manuscript preparation.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. Nat. Rev. Mol. Cell Biol. *22*, 96–118.
2. Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. Nat. Rev. Genet. *15*, 7–21.
3. Akhade, V.S., Pal, D., and Kanduri, C. (2017). Long noncoding RNA: genome organization and mechanism of action. Adv. Exp. Med. Biol. *1008*, 47–74.
4. Qiao, J., Jin, J., Yu, H., and Wei, L. (2024). Towards Retraining-free RNA Modification Prediction with Incremental Learning. Inf. Sci. *660*, 120105.
5. Bridges, M.C., Daulagala, A.C., and Kourtidis, A. (2021). LNCcation: lncRNA localization and function. J. Cell Biol. *220*, e202009045.
6. Cao, C., Shao, M., Zuo, C., Kwok, D., Liu, L., Ge, Y., Zhang, Z., Cui, F., Chen, M., Fan, R., et al. (2024). RAVAR: a curated repository for rare variant-trait associations. Nucleic Acids Res. *52*, D990–D997.
7. Cao, C., Wang, J., Kwok, D., Cui, F., Zhang, Z., Zhao, D., Li, M.J., and Zou, Q. (2022). webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. Nucleic Acids Res. *50*, D1123–D1130.
8. Li, P., Tiwari, P., Xu, J., Qian, Y., Ai, C., Ding, Y., and Guo, F. (2022). Sparse regularized joint projection model for identifying associations of non-coding RNAs and human diseases. Knowl. Base Syst. 258.
9. Ilık, İ.A., and Aktaş, T. (2022). Nuclear speckles: dynamic hubs of gene expression regulation. FEBS J. *289*, 7234–7245.
10. Carotenuto, P., Pecoraro, A., Palma, G., Russo, G., and Russo, A. (2019). Therapeutic approaches targeting nucleolus in cancer. Cells *8*, 1090.
11. Zhou, H., Wang, H., Tang, J., Ding, Y., and Guo, F. (2022). Identify ncRNA Subcellular Localization via Graph Regularized k-Local Hyperplane Distance Nearest Neighbor Model on Multi-Kernel Learning. IEEE ACM Trans. Comput. Biol. Bioinf *19*, 3517–3529.
12. Li, H., and Liu, B. (2023). BioSeq-Diabolo: Biological sequence similarity analysis using Diabolo. PLoS Comput. Biol. *19*, e1011214.
13. Iarovaia, O.V., Minina, E.P., Sheval, E.V., Onichtchouk, D., Dokudovskaya, S., Razin, S.V., and Vassetzky, Y.S. (2019). Nucleolus: a central hub for nuclear functions. Trends Cell Biol. *29*, 647–659.
14. Olson, M.O.J., and Dundr, M. (2005). The moving parts of the nucleolus. Histochem. Cell Biol. *123*, 203–216.
15. Boisvert, F.-M., Van Koningsbruggen, S., Navascués, J., and Lamond, A.I. (2007). The multifunctional nucleolus. Nat. Rev. Mol. Cell Biol. *8*, 574–585.
16. Grummt, I. (2013). The nucleolus—guardian of cellular homeostasis and genome integrity. Chromosoma *122*, 487–497.

17. Tajrishi, M.M., Tuteja, R., and Tuteja, N. (2011). Nucleolin: The most abundant multifunctional phosphoprotein of nucleolus. Commun. Integr. Biol. *4*, 267–275.

18. Carlevaro-Fita, J., and Johnson, R. (2019). Global positioning system: understanding long noncoding RNAs through subcellular localization. Mol. Cell. *73*, 869–883.

19. Lin, Z., Xie, G., Jiang, Z., Gu, G., Sun, Y., Su, Q., Cui, J., and Zhang, H. (2023). DHOSGR: lncRNA-disease Association Prediction Based on Decay High-order Similarity and Graph-regularized Matrix Completion. Curr. Bioinf. *18*, 92–104.

20. Wu, H., Liang, Q., Zhang, W., Zou, Q., El-Latif Hesham, A., and Liu, B. (2022). iLncDA-LTR: Identification of lncRNA-disease associations by learning to rank. Comput. Biol. Med. *146*, 105605.

21. Jin, J., Yu, Y., Wang, R., Zeng, X., Pang, C., Jiang, Y., Li, Z., Dai, Y., Su, R., Zou, Q., et al. (2022). iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. Genome Biol. *23*, 219.

22. Wu, Y.-Y., and Kuo, H.-C. (2020). Functional roles and networks of non-coding RNAs in the pathogenesis of neurodegenerative diseases. J. Biomed. Sci. *27*, 49.

23. Li, J., Zou, Q., and Yuan, L. (2023). A review from biological mapping to computation-based subcellular localization. Mol. Ther. Nucleic Acids *32*, 507–521.

24. Hwang, S.-P., and Denicourt, C. (2024). The impact of ribosome biogenesis in cancer: from proliferation to metastasis. NAR Cancer *6*, zcae017.

25. Zhang, Z.Y., Ning, L., Ye, X., Yang, Y.H., Futamura, Y., Sakurai, T., and Lin, H. (2022). iLoc-miRNA: extracellular/intracellular miRNA prediction using deep BiLSTM with attention mechanism. Briefings Bioinf. *23*, bbac395.

26. Zhang, Z.Y., Sun, Z.J., Yang, Y.H., and Lin, H. (2022). Towards a better prediction of subcellular location of long non-coding RNA. Front. Comput. Sci. *16*, 165903.

27. Pederson, T. (2011). The nucleolus. Cold Spring Harbor Perspect. Biol. *3*, a000638.

28. Huarte, M. (2015). The emerging role of lncRNAs in cancer. Nat. Med. *21*, 1253–1261.

29. Peng, L., Yuan, X., Jiang, B., Tang, Z., and Li, G.-C. (2016). LncRNAs: key players and novel insights into cervical cancer. Tumour Biol. *37*, 2779–2788.

30. Clemson, C.M., Hutchinson, J.N., Sara, S.A., Ensminger, A.W., Fox, A.H., Chess, A., and Lawrence, J.B. (2009). An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol. Cell. *33*, 717–726.

31. Riva, P., Ratti, A., and Venturin, M. (2016). The long non-coding RNAs in neurodegenerative diseases: novel mechanisms of pathogenesis. Curr. Alzheimer Res. *13*, 1219–1231.

32. L'heureux, A., Grolinger, K., Elyamany, H.F., and Capretz, M.A.M. (2017). Machine learning with big data: Challenges and approaches. IEEE Access *5*, 7776–7797.

33. Jeon, Y.-J., Hasan, M.M., Park, H.W., Lee, K.W., and Manavalan, B. (2022). TACOS: a novel approach for accurate prediction of cell-specific long noncoding RNAs subcellular localization. Briefings Bioinf. *23*, bbac243.

34. Li, J., Zhang, L., He, S., Guo, F., and Zou, Q. (2021). SubLocEP: a novel ensemble predictor of subcellular localization of eukaryotic mRNA based on machine learning. Briefings Bioinf. *22*, bbaa401.

35. Yuan, G.-H., Wang, Y., Wang, G.-Z., and Yang, L. (2023). RNAlight: a machine learning model to identify nucleotide features determining RNA subcellular localization. Briefings Bioinf. *24*, bbac509.

36. Zeng, M., Wu, Y., Lu, C., Zhang, F., Wu, F.-X., and Li, M. (2022). DeepLncLoc: a deep learning framework for long non-coding RNA subcellular localization prediction based on subsequence embedding. Briefings Bioinf. *23*, bbab360.

37. Li, M., Zhao, B., Yin, R., Lu, C., Guo, F., and Zeng, M. (2023). GraphLncLoc: long non-coding RNA subcellular localization prediction using graph convolutional networks based on sequence to graph transformation. Briefings Bioinf. *24*, bbac565.

38. Wang, S., Shen, Z., Liu, T., Long, W., Jiang, L., and Peng, S. (2023). DeepmRNALoc: A Novel Predictor of Eukaryotic mRNA Subcellular Localization Based on Deep Learning. Molecules *28*, 2284.

39. Cai, J., Wang, T., Deng, X., Tang, L., and Liu, L. (2023). GM-lncLoc: LncRNAs subcellular localization prediction based on graph neural network with meta-learning. BMC Genom. *24*, 52.

40. Chen, Z., Zhao, P., Li, F., Marquez-Lago, T.T., Leier, A., Revote, J., Zhu, Y., Powell, D.R., Akutsu, T., Webb, G.I., et al. (2020). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Briefings Bioinf. *21*, 1047–1057.

41. Inture, A.R., Nadh, B.S.S., Sha, A., Abhishek, S., Anjali, T., and Mullapudi, T.V. (2023). Leveraging Random Forests for Ovarian Cancer Detection and Precision Prediction. In 2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (IEEE).

42. Tahir, M.A.U.H., Asghar, S., Manzoor, A., and Noor, M.A. (2019). A classification model for class imbalance dataset using genetic programming. IEEE Access *7*, 71013–71037.

43. Sayed, G.I., Soliman, M.M., and Hassanien, A.E. (2021). A novel melanoma prediction model for imbalanced data using optimized SqueezeNet by bald eagle search optimization. Comput. Biol. Med. *136*, 104712.

44. Ai, C., Yang, H., Ding, Y., Tang, J., and Guo, F. (2023). Low Rank Matrix Factorization Algorithm Based on Multi-Graph Regularization for Detecting Drug-Disease Association. IEEE ACM Trans. Comput. Biol. Bioinf *20*, 3033–3043.

45. Zhu, H., Hao, H., and Yu, L. (2023). Identifying disease-related microbes based on multi-scale variational graph autoencoder embedding Wasserstein distance. BMC Biol. *21*, 294.

46. Vega García, M., and Aznarte, J.L. (2020). Shapley additive explanations for NO2 forecasting. Ecol. Inf. *56*, 101039.

47. Baptista, M.L., Goebel, K., and Henriques, E.M. (2022). Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. Artif. Intell. *306*, 103667.

48. Cui, T., Dou, Y., Tan, P., Ni, Z., Liu, T., Wang, D., Huang, Y., Cai, K., Zhao, X., Xu, D., et al. (2022). RNALocate v2. 0: an updated resource for RNA subcellular localization with increased coverage and annotation. Nucleic Acids Res. *50*, D333–D339.

49. Quinn, J.J., and Chang, H.Y. (2016). Unique features of long non-coding RNA biogenesis and function. Nat. Rev. Genet. *17*, 47–62.

50. Nair, L., Chung, H., and Basu, U. (2020). Regulation of long non-coding RNAs and genome dynamics by the RNA surveillance machinery. Nat. Rev. Mol. Cell Biol. *21*, 123–136.

51. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

52. Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.-C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. *43*, W65–W71.

53. Zulfiqar, H., Guo, Z., Ahmad, R.M., Ahmed, Z., Cai, P., Chen, X., Zhang, Y., Lin, H., and Shi, Z. (2023). Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings. Front. Med. *10*, 1291352.

54. Zhu, W., Yuan, S.S., Li, J., Huang, C.B., Lin, H., and Liao, B. (2023). A First Computational Frame for Recognizing Heparin-Binding Protein. Diagnostics *13*, 2465.

55. Ma, T., Lin, X., Song, B., Philip, S.Y., and Zeng, X. (2023). Kg-mtl: Knowledge graph enhanced multi-task learning for molecular interaction. IEEE Trans. Knowl. Data Eng. *35*, 7068–7081.

56. Xu, J., Xu, J., Meng, Y., Lu, C., Cai, L., Zeng, X., Nussinov, R., and Cheng, F. (2023). Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data. Cell Rep. Methods *3*, 100382.

57. Tang, Y.J., Pang, Y.H., and Liu, B. (2021). IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning. Bioinformatics *36*, 5177–5186.

58. Zou, X., Ren, L., Cai, P., Zhang, Y., Ding, H., Deng, K., Yu, X., Lin, H., and Huang, C. (2023). Accurately identifying hemagglutinin using sequence information and machine learning methods. Front. Med. *10*, 1281880.