OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae627
https://doi.org/10.1093/bib/bbae627
Problem Solving Protocol

The improved de Bruijn graph for multitask learning: predicting functions, subcellular localization, and interactions of noncoding RNAs

Yuxiao Wei[1], Qi Zhang[2], Liwei Liu[2,*]

[1]College of Software, Dalian Jiaotong University,794 Huanghe Road, Dalian 116028, China
[2]College of Science, Dalian Jiaotong University, 794 Huanghe Road, Dalian 116028, China

*Corresponding author. Liwei Liu, College of Science, Dalian Jiaotong University, 794 Huanghe Road, Dalian 116028, China. E-mail: liutree80@163.com

Abstract

Noncoding RNA refers to RNA that does not encode proteins. The lncRNA and miRNA it contains play crucial regulatory roles in organisms, and their aberrant expression is closely related to various diseases. Traditional experimental methods for validating the interactions of these RNAs have limitations, and existing prediction models exhibit relatively limited functionality, relying on isolated feature extraction and performing poorly in handling various types of small sample tasks. This paper proposes an improved de Bruijn graph that can inject RNA structural information into the graph while preserving sequence information. Furthermore, the improved de Bruijn graph enables graph neural networks to learn broader dependencies and correlations among data by introducing richer edge relationships. Meanwhile, the multitask learning model, DVMnet, proposed in this paper can handle multiple related tasks, and we optimize model parameters by integrating the total loss of three tasks. This enables multitask prediction of RNA interactions, disease associations, and subcellular localization. Compared with the best existing models in this field, DVMnet has achieved the best performance with a 3% improvement in the area under the curve value and demonstrates robust results in predicting diseases and subcellular localization. The improved de Bruijn graph is also applicable to various scenarios and can unify the sequence and structural information of various nucleic acids into a single graph.

Keywords: improved de Bruijn graph algorithm; multitask; deep learning; long noncoding RNA

Introduction

With the discovery of many long noncoding RNAs (lncRNAs) and microRNAs (miRNAs) in the early part of the first decade of the 21st century, it became widely recognized that RNA can function as a regulatory molecule [1]. LncRNA, with a length >200 nucleotides, has been implicated in the development and progression of various diseases through its abnormal expression [2]. MiRNA is a class of RNA molecules ∼18–22 nucleotides long, widely present in eukaryotes as noncoding RNAs (ncRNAs) [3]. Their subcellular localization may determine their specific functions, for instance, regulating gene transcription in the nucleus or controlling mRNA translation in the cytoplasm [4]. Furthermore, lncRNAs and miRNAs exhibit complex and intimate interrelationships within cells [5]. For example, lncRNAs can attenuate the direct impact of miRNAs on mRNAs, and there may exist positive or negative feedback regulatory relationships between lncRNAs and miRNAs [6]. In summary, both lncRNAs and miRNAs play pivotal roles in the onset and progression of diseases [7–11].

Traditional experimental approaches to validate the interactions between lncRNAs and miRNAs, as well as to identify RNA–disease associations, are often time-consuming and labor-intensive, limiting their application to only a few specific lncRNAs and miRNAs [12–14]. RNA subcellular localization relies on smFISH, making computational methods for predicting RNA interactions, associations, and localization valuable. Proposed methods use 1D or 2D RNA features with CNNs, DNNs, or transformers due to limited 3D data [15–19]. Zhou *et al.* studied the relationship between ncRNA and proteins based on the multihead attention mechanism, effectively utilizing the sequence information of ncRNA and proteins, but did not incorporate other richer feature information from additional modalities [20]. Zhang *et al.* employed the fusion of multisource relationships in the functional prediction of lncRNA and proposed strategies of path masking and degree regression [21]. Additionally, some models constructed heterogeneous graphs with RNAs, diseases, and other entities and employed graph neural networks (GNNs) for node feature representation [22–24]; Wei *et al.* proposed a framework based on GNNs to predict ncRNA–protein interactions. This study leveraged the topological information of graph data and employed a more refined sampling strategy. However, it is highly dependent on the specific dataset used and lacks general applicability [25]. Graph representation of RNA captures its structure and interactions, aiding RNA information extraction and supporting function prediction and disease analysis [26–28].

Received: September 16, 2024. **Revised:** November 13, 2024. **Accepted:** November 15, 2024

Currently, the most common prediction classifiers fall into two categories. One approach integrates the features of heterogeneous entities such as RNAs into a unified dimension and constructs an end-to-end prediction framework using networks like multilayer perceptrons or fully connected layers [29, 30]. This method uses machine learning algorithms (SVMs, logistic regression, random forests) to build classifiers for RNA feature–based prediction [31, 32]. Despite the progress made by these methods, there are still two major challenges. The first challenge is the effective representation of RNA features. RNA contains various stem–loop structures, such as hairpins, bulges, and loops [33]. The de Bruijn map has been used to construct an RNA map. However, it can't reflect the structural characteristics of RNA. [34, 35]. The second challenge involves enhancing the learning efficiency for small datasets of ncRNAs. By leveraging the potential correlations and shared features between lncRNA–miRNA interactions and RNA characteristics, the performance of tasks involving small sample sizes can be significantly improved through the application of multitask learning [36, 37].

To address these issues, we propose DVMnet, which is tailored for predicting the attributes and interactions associated with lncRNAs and miRNAs. We first co-express RNA features through an improved de Bruijn graph and sequence features and then utilize a GNN to extract these features. This approach captures the complex interactions between RNA molecules, thereby providing a more comprehensive understanding of RNA function and mechanisms within cells. Subsequently, we construct corresponding multilayer neural network prediction classifiers for each task and optimize the model as well as update its parameters by integrating the total loss across all three tasks. Finally, we conduct various evaluation experiments on DVMnet and perform a comprehensive performance comparison with state-of-the-art prediction methods. The results demonstrate that our method achieves the best overall performance compared to state-of-the-art approaches, providing a more accurate and effective tool for lncRNA and miRNA research.

In summary, DVMnet has accomplished the following tasks: (i) we have proposed an improved de Bruijn graph specifically designed for capturing RNA structural information. It is capable of handling sequences of different lengths by introducing structural information and constructing a large number of edge relationships; it enables effective information propagation between originally disconnected nodes, allowing the model to learn broader dependencies and correlations among bases. This enhanced learning capability helps the model perform better on unseen data and can also appropriately mitigate overfitting. This makes its application scenarios more extensive and better suited for multitask scenarios. (ii) It simultaneously learns multiple related tasks through a multitask learning approach, enhancing overall performance and providing a platform for handling small data tasks. (iii) We also employ GNNs for feature extraction, enabling the model to better capture key information and potential interaction patterns within RNA sequences. These innovations give DVMnet significant advantages in predicting properties and interactions related to lncRNAs and miRNAs, offering new ideas and methods for research in related fields.

Methods

Datasets

LncACTdb [38] (http://bio-bigdata.hrbmu.edu.cn/LncACTdb/) is a comprehensive database where all lncRNA–miRNA associations have been experimentally validated. Initially, we obtained 1057 lncRNA–miRNA associations from LncACTdb, involving 284 lncRNAs and 520 miRNAs. Subsequently, we derived the associations between lncRNAs and diseases. From the data on lncRNA–disease relationships, we filtered out 187 lncRNAs that intersected with the lncRNA–miRNA association dataset and were related to nine types of cancer. RNAlocate [39] (http://www.rna-society.org/rnalocate/) is a search tool for RNA subcellular localization. We selected 454 miRNAs that intersect with the lncRNA–miRNA association dataset and retrieved the subcellular localization information for these 454 miRNAs from this database. The sequence information for lncRNAs and miRNAs was sourced from LNCipedia [40], NONCODE [41], and miRbase [42]. LNCipedia and NONCODE are fundamental noncoding RNA database platforms specifically designed to store lncRNA sequences and annotation information, integrating lncRNA data from multiple sources. On the other hand, miRBase is an online miRNA database dedicated to collecting, storing, and managing various miRNA sequences. For the main task, we allocated 80% and 20% of the entire dataset as the training set and validation set, respectively. For the two auxiliary tasks, we allocated 85% and 15% of the entire dataset to their respective training sets and validation sets.

DVMnet framework

Figure 1a illustrates the proposed DVMnet framework. Given an input lncRNA–miRNA pair, firstly, the RNA sequence data undergo preprocessing to extract key features. Then, a dual-view model is constructed using an improved de Bruijn graph and a heterogeneous graph to accurately represent RNA sequence information and potential interactions. Subsequently, an advanced graph attention network is employed to extract and encode RNA features. By capturing crucial information within RNA sequences and underlying interaction patterns, it provides a powerful feature representation for subsequent prediction tasks. Next, based on the extracted features, three interrelated multilayer neural network prediction classifiers are constructed for predicting RNA interactions, disease associations, and subcellular localization, respectively. By sharing underlying feature representations and jointly optimizing the overall loss across multiple tasks, the overall prediction performance is enhanced.

Construction of de Bruijn graph

The secondary structure of ncRNA is crucial for its normal function. Its sequence is mainly composed of four types of bases, following pairing rules such as A-U, C-G, or G-U. During computation, examining the pairing probability of each base with other bases can not only obtain accurate secondary structure information but also derive biologically meaningful features by capturing potential pairing information.

The traditional de Bruijn graph is a directed graph that represents the overlapping relationships between symbolic sequences. The difference between the improved de Bruijn graph and the traditional one lies in that it represents the pairing relationships between symbolic sequences, rather than the overlapping relationships. In the graph construction part, we transform the lncRNA sequence into a graph. Specifically, the nodes in the graph are the same as those in the traditional de Bruijn graph. When constructing the edges, instead of connecting adjacent base modules, we assign weight data between bases that can match the given matching rules. The lncRNA sequence is as follows:

$$LncRNA = N_1, N_2, N_3 \ldots\ldots N_{L-1}, N_L.$$

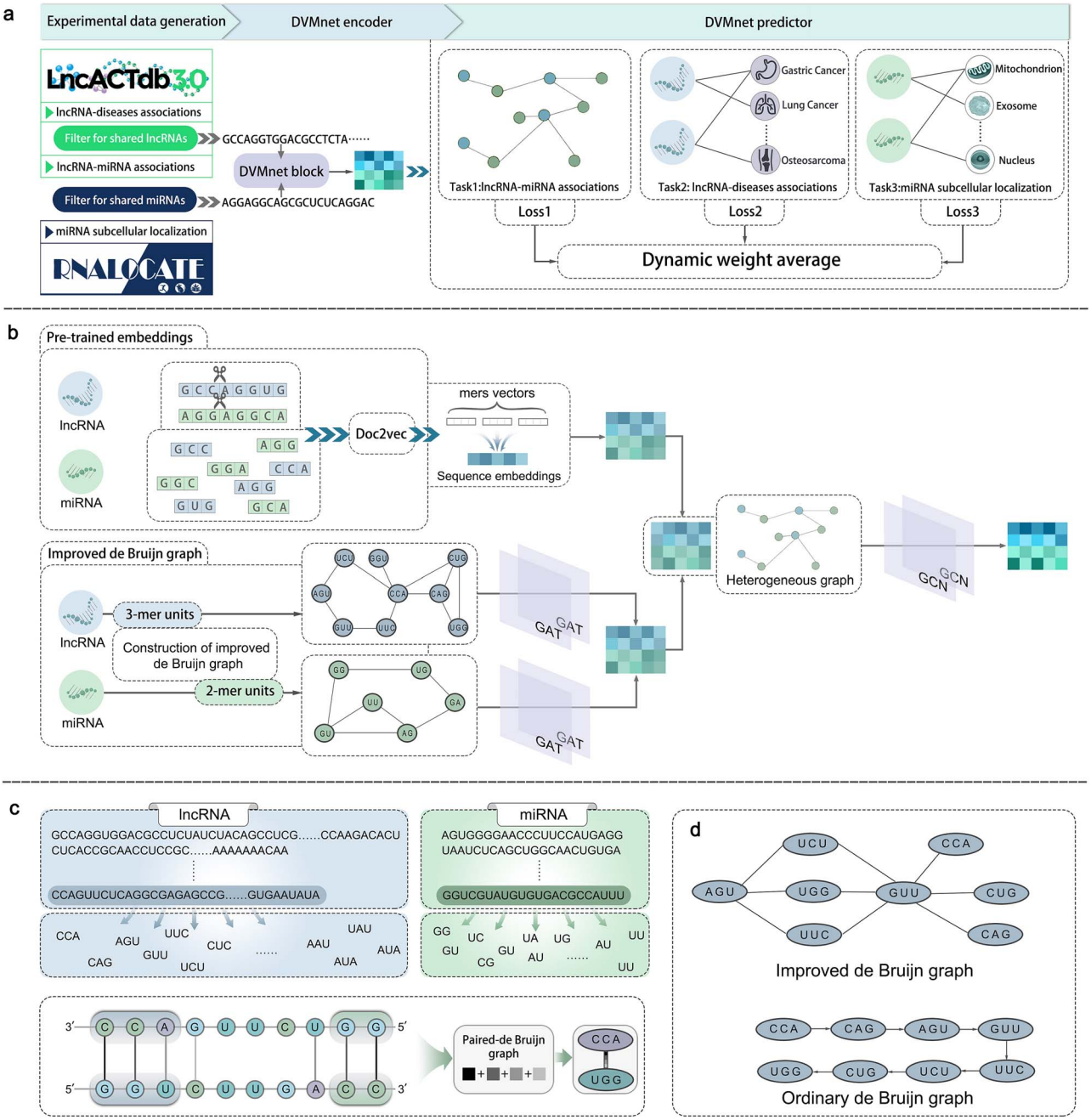

Figure 1. DVMnet flowchart. (a) Model framework. After the data processing stage, we obtained the sequences of lncRNA and miRNA, as well as data on lncRNA–miRNA interactions, lncRNA–disease associations, and subcellular localization of miRNA. Firstly, we feed the RNA sequences into the DVMnet block to update their features. Then, based on the output of DVMnet, we construct three predictors and adopt a dynamic task prioritization method to integrate the losses of the three predictors for parameter updating and model optimization. Finally, the trained DVMnet model will accomplish the tasks of predicting lncRNA–miRNA interactions, lncRNA–disease associations, and subcellular localization of miRNA. (b) DVMnet block. This is the module used to update the features of each RNA. Initially, an improved de Bruijn graph is constructed, and a graph attention network is employed to extract latent structural features within the RNA molecules. Subsequently, a heterogeneous graph with lncRNA and miRNA as nodes is built, and a graph convolutional neural network is utilized to capture the complex relationships and interactions among RNAs. (c) Construction of improved Debruin diagram. The RNA sequence is decomposed into subsequences of length k, where the k value of lncRNA is 3, and the k value of miRNA is 2. Taking the segmented sequence as a node, according to the principle of base pairing, the pairing of each base and its nearby bases with other bases in the sequence is traversed, and the corresponding weights are given to generate edges. (d) Comparing the improved Debruin diagram with the traditional Debruin diagram, it is obvious that the improved Debruin diagram has more related relations.

Taking base N_1 as an example, we first consider its matching relationship with base N_L. In Tab. S1, we illustrate the weights of base pairings in the improved de Bruijn graph construction, and the pairing weight is set as follows:

$$S = \begin{cases} 2, (if\ (N_a = A\ and\ N_b = U)\ \ or\ \ (N_a = U\ and\ N_b = A)) \\ 3, (if\ (N_a = G\ and\ N_b = C)\ \ or\ \ (N_a = C\ and\ N_b = G)) \\ 1, (if\ (N_a = G\ and\ N_b = U)\ \ or\ \ (N_a = U\ and\ N_b = G)) \\ 0, else \end{cases}.$$

If the pairing is successful, we continue to consider the pairing relationship between base N_2 and the next base N_{L-1}. If it fails, the matching stops; if it succeeds, it continues until it reaches V_{MAX} steps. V_{MAX} is the maximum number of steps we define. Each successful step forward earns a matching score. Based on the type of matched bases and the number of steps taken, we incorporate the idea of locally weighted linear regression and add a Gaussian function as a weight. This way, we obtain the total sum of matching scores:

$$W_{score} = S \cdot e^{-0.5x^2},$$

$$W_{sum} = \sum W_{score},$$

where W_{score} represents the weight score for each step, x represents the number of matching steps, and W_{sum} represents the total weight score. Ultimately, W_{sum} is used as the edge weight and inserted into the de Bruijn graph matrix.

As shown in Fig. 1c, we used 3-mers to reconstruct lncRNAs and 2-mers to reconstruct miRNAs, thereby constructing an improved de Bruijn graph. As depicted in Fig. 1d, the original de Bruijn graph builds graph associations based on the sequence order (from left to right), which merely describes the 1D spatial information of RNA based on its nucleotide sequence. However, the improved de Bruijn graph can capture the structural information of RNA and construct more graph–space associations. In addition, we have also presented a heatmap of the characteristics of the improved de Bruijn graph in Supporting Fig. S1.

DVMnet encoder

In the DVMnet encoder, we first construct two types of graphs separately. Besides the improved de Bruijn graph, we also build a heterogeneous graph containing two types of nodes through the lncRNA–miRNA interaction network. The nodes represent lncRNAs and miRNAs.

Next, we will introduce the process of extracting RNA features in each DVMnet block within the DVMnet encoder. As shown in Fig. 1b, For the de Bruijn graph of RNA, our model employs a graph attention network to capture the interactions of each node within the graph. Specifically, we allow each trinucleotide (dinucleotide) in the RNA to send attention information along the edges of the de Bruijn graph and aggregate all information from its neighboring nodes. Firstly, we calculate the attention coefficients between each node and its neighboring nodes in the graph using the following equation:

$$\hat{e}_{ij} = \hat{a}^{(l)^T}\left[W_k^{(l)}\bar{h}_i^{(l)} \| W_{k'}^{(l)}\bar{h}_j^{(l)}\right],$$

$$\alpha_{ij} = \frac{\exp\left(LeakyReLU\left(\hat{e}_{ij}\right)\right)}{\sum_{s\in N_i\cup\{i\}}\exp\left(LeakyReLU\left(\hat{e}_{is}\right)\right)},$$

where $\hat{e}_{ij}$ represents the attention weight between adjacent nodes, $\hat{a}^{(l)^T}$, $W_k^{(l)}$, and $W_{k'}^{(l)}$ are trainable parameters, and $\|$ denotes concatenation operation. α_{ij} represents the attention score between node i and node j, softmax denotes the activation function, and N_i represents the set of neighboring nodes of node i. $\bar{h}_i^{(l)}$ and $\bar{h}_j^{(l)}$ represent the feature representations of node i and node j at the lth layer in the de Bruijn graph, respectively. When $l = 0$, the feature of each node is represented by a one-hot encoding vector. After obtaining the attention score α_{ij}, we use the following equation to aggregate the features of neighboring nodes and complete the node update:

$$\bar{h}_i^{(l+1)} = \sigma\left(\sum_{j\in N_i\cup\{i\}}\left(\frac{1}{c_i}\alpha_{ij}W_{k''}^{(l)}\bar{h}_i^{(l)} + \hat{b}^{(l)}\right)\right),$$

where $\bar{h}_i^{(l+1)}$ represents the updated feature representation of node i in the de Bruijn graph. σ denotes the activation function, c_i is the normalization constant, and $c_i = |N_i| + 1$. Both $W_{k''}^{(l)}$ and $\hat{b}^{(l)}$ are trainable parameters.

Unlike the node representation in the de Bruijn graph, each node in the heterogeneous graph is represented by a complete RNA sequence, and the adjacency matrix in the heterogeneous graph is described as the association between two types of RNA. For the heterogeneous graph composed of lncRNA and miRNA, our model employs a graph convolutional neural network to capture the interactions of each node within the graph. Specifically, the feature update process for node i in the heterogeneous graph can be formulated as follows:

$$\overset{=}{h}_i^{(l+1)} = \sigma\left(\sum_{j\in\tilde{N}_i}\left(\frac{1}{\tilde{c}_i}W_{rk''}^{(l)}\overset{=}{h}_i^{(l)} + \tilde{b}^{(l)}\right)\right),$$

where $\overset{=}{h}_i^{(l+1)}$ denotes the updated feature representation of node i in the heterogeneous graph, $\tilde{c}_i = |\tilde{N}_i| + 1$, and both $W_{rk''}^{(l)}$ and $\tilde{b}^{(l)}$ are trainable parameters. $\tilde{N}_i$ represents the set of neighboring nodes for node i. $\overset{=}{h}_i^{(l)}$ indicates the feature representation of node i at the l-th layer in the heterogeneous graph. When $l = 0$, the features of each node are represented by the output features of the de Bruijn graph and Doc2vec encoding.

After passing through the DVMnet Block shown in Fig. 1b, we obtain the feature representation for each RNA:

$$g = BN\left(\sum_{i=1}^{n}\overset{=}{h}_i\right),$$

where BN() represents the batch normalization layer.

DVMnet predictor

Through the DVMnet encoder, we have extracted the embedded features of RNA from two perspectives. In this section, we use these embedded features to make predictions.

Firstly, we adopt a prediction function to determine the likelihood of lncRNA–miRNA interactions. The prediction function is as follows:

$$P\left(g_{lnc}, g_{mi}\right) = \sigma\left({g_{lnc}}^T W_{lm} g_{mi}\right),$$

where σ denotes the activation function, and g_{lnc} and g_{mi} represent the embedded features of lncRNA and miRNA, respectively. W_{lm} is the trainable weight matrix of the prediction function. $P\left(g_{lnc}, g_{mi}\right)$ is the predicted value for the interaction between lncRNA and miRNA.

In addition, we predict lncRNA–disease associations using a multilayer perceptron (MLP) composed of a three-layer neural network as the predictor. The prediction process can be formulated as follows:

$$P\left(g_{lnc}, D_r\right) = \sigma\left(MLP\left(g_{lnc}, W_r, b_r\right)\right),$$

where σ denotes the activation function, MLP represents the predictor, and g_{lnc} indicates the embedded features of lncRNA. W_r is the trainable weight matrix associated with disease type r. $P\left(g_{lnc}, D_r\right)$ is the predicted value for the association between lncRNA and diseaseD_r of type r.

Similar to the predictor for lncRNA–disease associations, we have constructed a classifier for miRNA subcellular localization. The prediction process can be formulated as follows:

$$P\left(g_{mi}, C_e\right) = \sigma\left(MLP\left(g_{mi}, W_e, b_e\right)\right),$$

where σ denotes the activation function, MLP represents the predictor, and g_{mi} indicates the embedded features of miRNA. W_e is the trainable weight matrix associated with subcellular type e. $P\left(g_{mi}, C_e\right)$ is the predicted value for the localization of miRNA in the subcellular compartment C_e of type e.

We use binary cross-entropy loss as the loss function for training the lncRNA–miRNA interaction task. For the prediction tasks of lncRNA–disease associations and miRNA subcellular localization, we use the cross-entropy loss function:

$$Loss_B = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(p_i\right) + \left(1 - y_i\right)\log\left(1 - p_i\right),$$

$$Loss_C = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(p_i\right),$$

where y_i represents the true label value of the ith sample, p_i represents the predicted value of the i-th sample, and N denotes the batch size. After obtaining the losses for the three tasks, we employ dynamic task prioritization to integrate them and select the Adam optimizer based on the total loss for parameter updates and model optimization.

$$\gamma_k(t) = \frac{Kexp\left(w_k\left(t-1\right)/T\right)}{\sum_i \exp\left(w_k\left(t-1\right)/T\right)},$$

$$w_k\left(t-1\right) = \frac{L_k\left(t-1\right)}{L_k\left(t-2\right)},$$

where γ_k represents the weight of each task, w_k is the ratio of losses from the previous and second-to-last rounds, indicating the learning rate for different tasks. T serves as a smoothing factor for task weights; the larger T is, the more evenly distributed the weights of different tasks will be. The total loss is the weighted average of the losses from all tasks:

$$Loss = \sum_k \gamma_k L_k.$$

Experimental settings

DVMnet is trained using PyTorch1.11.0 in Python3.8 (ubuntu20.04) on the following hardware configurations:

r CPU: 16 vCPU AMD EPYC 9654 96-Core Processor.

r GPU: RTX 4090(24GB) ∗ 1.

We use the Adam optimizer for parameter training. In addition, we present the AUC values for each epoch and the changes in loss values for each task in the supporting materials, specifically in Fig. S2 and Fig. S3. Furthermore, in Tab. S2 of the supporting materials, we provide a comparison of memory usage and time for the improved de Bruijn graph.

Results

Evaluation strategies and metrics

To comprehensively assess the model's performance, we selected the F1 score, area under the curve (AUC) value, average precision (AP), and normalized discounted cumulative gain (NDCG) as evaluation metrics. In terms of evaluation metrics, the F1 score represents the harmonic mean of precision and recall, which allows for a comprehensive assessment of both the classification accuracy and completeness of the model. The AUC value is commonly used to evaluate the performance of binary classification models. AP calculates the area under the precision–recall curve, which better reflects the model's performance at high recall and high precision rates, and NDCG is frequently utilized to assess the performance of ranking tasks.

Performance comparison on benchmark datasets

Here, we compare DVMnet with two state-of-the-art baselines: preMLI [43] and SPGNN [44]. The focus of the comparison lies in the lncRNA–miRNA interaction prediction task. We compared DVMnet with the following two models for lncRNA–miRNA interaction prediction: (i) preMLI, which combines rna2vec pretraining technology with a deep feature mining mechanism for training. On benchmark datasets, it has demonstrated performance superior to previous methods. Although it employs word embedding models at different positions similar to DVMnet, it lacks the structural representation of DVMnet. Therefore, we compare it with DVMnet. (ii) SPGNN, a GNN based on sequence pretraining that generates high-quality pretrained embeddings from the entire RNA sequence corpus during the pretraining phase using sequence-to-vector conversion techniques. This model is the latest model at present and uses different GNNs from DVMnet in different positions, so it is selected as the baseline model.

As shown in the confusion matrix in Fig. 2a, DVMnet achieves more accurate predictions of lncRNA and miRNA interactions compared to preMLI and SPGNN. Figure 2b displays the F1 score, AUC value, AP value, and NDCG value of the three models. Among them, the area under the receiver operating characteristic (ROC) curve for each model is shown in Fig. 2c. Experimental results demonstrate that the DVMnet model exhibits significant advantages across all four evaluation metrics. This indicates that the DVMnet model can more effectively reduce false positives and false negatives compared to the preMLI and SPGNN models, exhibiting higher classification accuracy and stability. This is primarily attributed to the dual-view multitask learning framework adopted by the DVMnet model, along with advanced feature extraction and encoding methods.

Ablation study

Here, we conducted a series of ablation studies to investigate the impact of each feature representation and extraction method within DVMnet on the model's final prediction performance. The left side of Fig. 3a represents the length of the decomposed subsequences in the lncRNA de Bruijn graph, while the right side shows the length of the decomposed subsequences in the miRNA de Bruijn graph. DVMnet performs best in all three tasks using 3-mer for lncRNA and 2-mer for miRNA, possibly due to longer lncRNA sequences expressing more dependencies. The

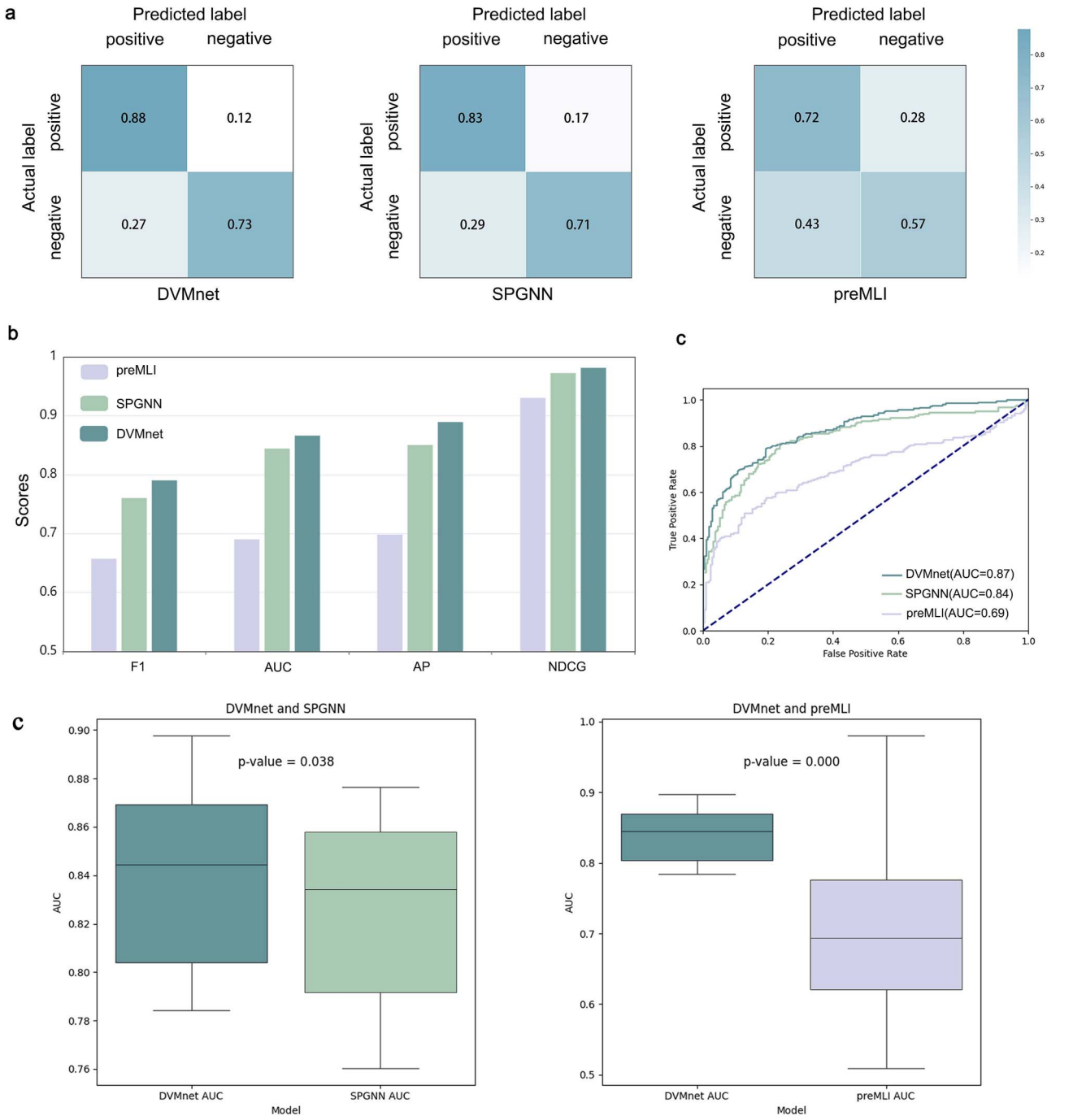

Figure 2. Comparative evaluation of DVMnet with other advanced baseline models. (a) Confusion matrix for predicting lncRNA–miRNA interactions using benchmark datasets. (b) Evaluation metrics including F1, AUC, AP, and NDCG and (c) AUROC curves. (c) To assess the statistical significance of the performance differences between DVMnet and other baseline methods, we adopted the Wilcoxon test. The figure presents the results for the task of predicting lncRNA–miRNA interactions on the benchmark dataset.

improved de Bruijn graph represents a significant innovation in this study. To explore its advantages, we compared it with various encoding methods, including Text2vec, RNABERT, and Doc2vec, as shown in Fig. 3b. Initially, when we used each of these three feature encoding methods separately, the performance of Text2vec and RNABERT was significantly inferior to that of the improved de Bruijn graph. Doc2vec's performance was slightly lower than improved de Bruijn graph (IDG)'s. Combining them improved results, with IDG + Doc2vec performing best. Two GNNs were used for feature extraction from improved graphs, a key step in DVMnet's RNA feature extraction. To explore the impact of various GNNs on the model's performance, we selected four GNNs specifically for the de Bruijn graph and six for the heterogeneous graph, integrating them into the DVMnet model under the same experimental conditions. By comparing the performance of different models in predicting lncRNA–miRNA interactions (Task 1), we found that, as shown in Fig. 3c, graph attention network (GAT) excelled on the improved de Bruijn graph, while graph convolution network (GCN) performed best on the heterogeneous graph, possibly due to their respective strengths matching the

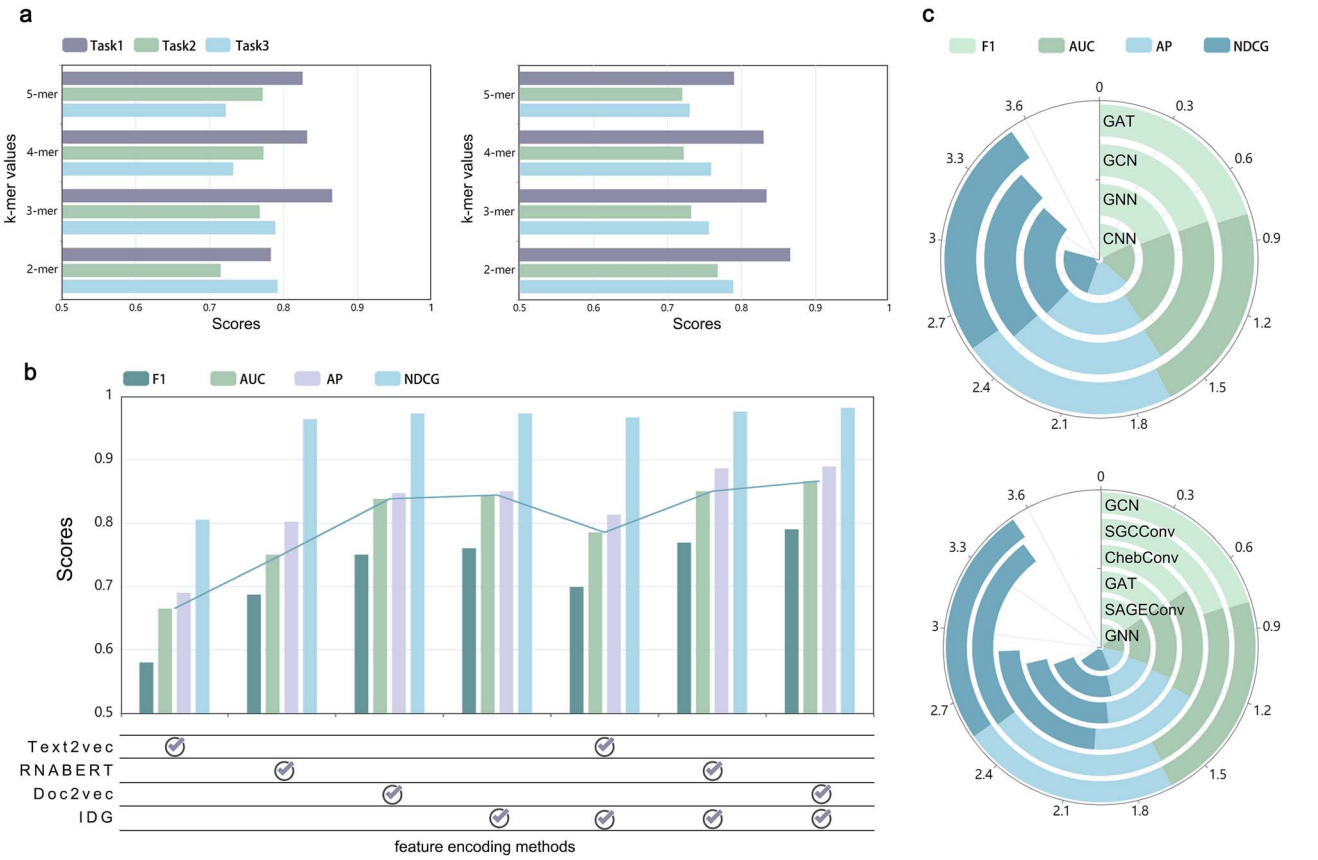

Figure 3. Partial ablation experimental results of DVMnet. (a) The influence of k-mer values in the de Bruijn graph on three tasks. (b) The impact of different feature encoding methods and combinations on Task 1. (c) The effect of various GNN approaches on Task 1 for both the de Bruijn graph and the heterogeneous graph.

graph characteristics. The de Bruijn graph suits GAT's attention mechanism, while the complex heterogeneous graph favors GCN.

Multitask effect

The prediction of lncRNA–disease interactions and the subcellular localization of miRNAs are secondary tasks within a multitask system. We connect these topics into a multitask system through the interactions between lncRNAs and miRNAs (as shown in Fig. 4a), and the impact of different predictors on the performance of these two secondary tasks is illustrated in Fig. 4b. When comparing the performance of various predictors, we found that the MLP demonstrated superior performance across all evaluation metrics. Compared to the five other methods, including bidirectional gated recurrent unit (BiGRU), recurrent neural network (RNN), the combination of BIGRU and MLP, the combination of RNN and MLP, and random forest, MLP exhibited the best predictive results. In contrast, BiGRU and RNN, while capable of processing sequential data and capturing temporal dependencies, may be affected by gradient vanishing or exploding issues in certain cases, limiting their performance. The combined approaches of BiGRU + MLP and RNN + MLP, while incorporating the strengths of both recurrent neural networks and MLPs to some extent, may suffer from overfitting or reduced computational efficiency due to an excess of parameters or high model complexity. As shown in Fig. 4c, we utilized the t-SNE algorithm to reduce the dimensionality of the model's predicted outputs for the two secondary tasks. The resulting data points were then plotted in a 3D coordinate system. Negative samples are represented in green, while positive samples are indicated in purple. This visualization approach allows us to intuitively observe the spatial distribution of data points belonging to different categories. Meanwhile, to mitigate gradient conflicts and poor model performance arising from low correlation among multiple tasks, we experimented with several methods for loss weight allocation (as illustrated in Fig. 4d). Among them, dynamic weight average [45] demonstrated the best performance, significantly surpassing gradient normalization and using uncertainty to weigh losses. Furthermore, for the task of miRNA subcellular localization, we compared the performance of our model with other models on smaller datasets and included these comparisons in Fig. S4 in the supporting materials.

Application of the improved de Bruijn graph in predicting tasks for more types of RNA

The improved de Bruijn graph proposed in this paper incorporates RNA structural information and biological significance based on RNA sequence information. This approach obtains more edge information and is able to learn more comprehensive correlations between bases, making it universally applicable to various types of RNA and DNA sequences.

To test the performance of the improved de Bruijn graph in other nucleic acids, we selected 15 well-known circRNA-RBP datasets [46] (https://circinteractome.nia.nih.gov/). In this dataset, we obtained circRNA fragments with a 1:1 ratio of positive and negative samples, along with labels indicating whether they are binding sites for target proteins. We applied the improved de Bruijn graph to this prediction task and evaluated the improvements it brought when added to a regular model that uses Doc2vec for feature extraction. On the other hand, we

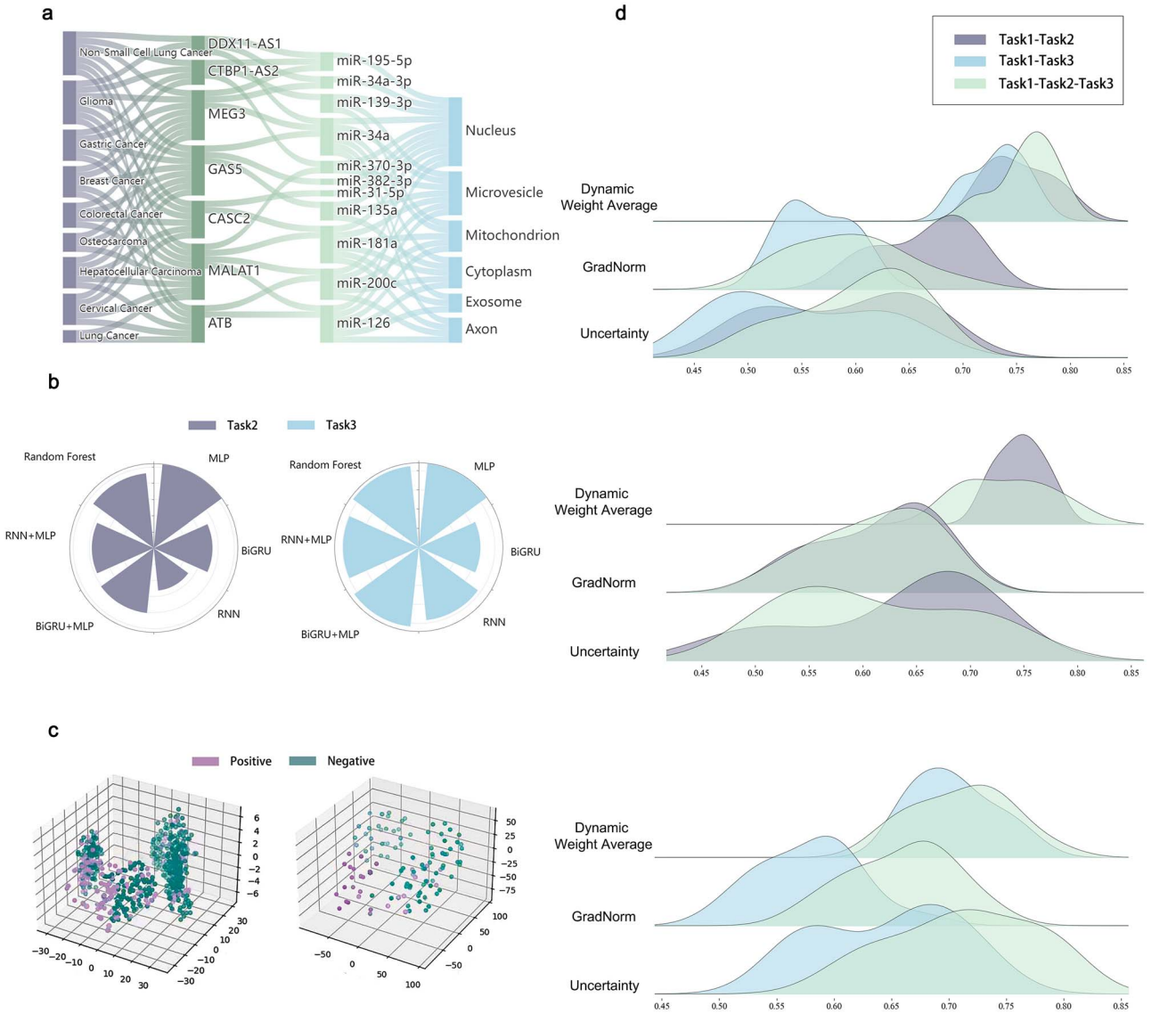

Figure 4. (a) Visualize the interrelationships between lncRNA, miRNA, diseases, and subcellular localization in the dataset using a Sankey diagram. (b) Performance comparison of different predictors in the tasks of lncRNA–disease association (Task 2) and miRNA subcellular localization (Task 3). (c) Classification visualization for the tasks of lncRNA–disease association and miRNA subcellular localization. (d) Ridge plot displaying the optimization strategy for multitask loss.

verified the DNA sequence. We used the dataset from Liu *et al.* [47] and got the same number of enhancer and nonenhancer samples. We applied the improved de Bruijn graph to this prediction task. Similarly, we compared the effect of using Doc2vec alone with that of adding the improved de Bruijn graph. Figure 5a and b shows that the improved de Bruijn graph has contributed to these two tasks. In the task of identifying the binding site of circRNA-RBP, after adding the improved de Bruijn graph as feature extraction, the AUC has increased by 18% on average. In the judgment task of enhancer, the AUC value increased by 7%.

Furthermore, as shown in Fig. 5c, to further understand the contribution of the improved de Bruijn graph to the prediction results and observe the contribution of each base pair recognition under this method, we conducted an interpretability analysis of the circRNA-RBP binding site recognition task using the integrated gradients attribution algorithm. In the AUF1 dataset, among all the samples, we selected 128 for statistical analysis and found that bases located in the stem regions of RNA contributed 21% more on average to the discrimination compared to those in the loop regions. This fully demonstrates the constructive role of the improved de Bruijn graph in structural information extraction.

Discussion

An increasing number of lncRNAs have their specific research foci, and the literature is replete with related content. However, some common themes are emerging: the complex regulatory networks formed among ncRNAs jointly regulate cellular gene expression and function. There is a current need for more precise information extraction methods and more versatile predictive frameworks. A major challenge lies in the fact that high-precision RNA 3D structures account for only a small fraction of known RNA sequences. Due to the vast differences in sequence length, it is quite challenging to obtain RNA structural features using these high-precision 3D structural RNAs or existing RNA structure prediction models. Therefore, we propose an improved de Bruijn

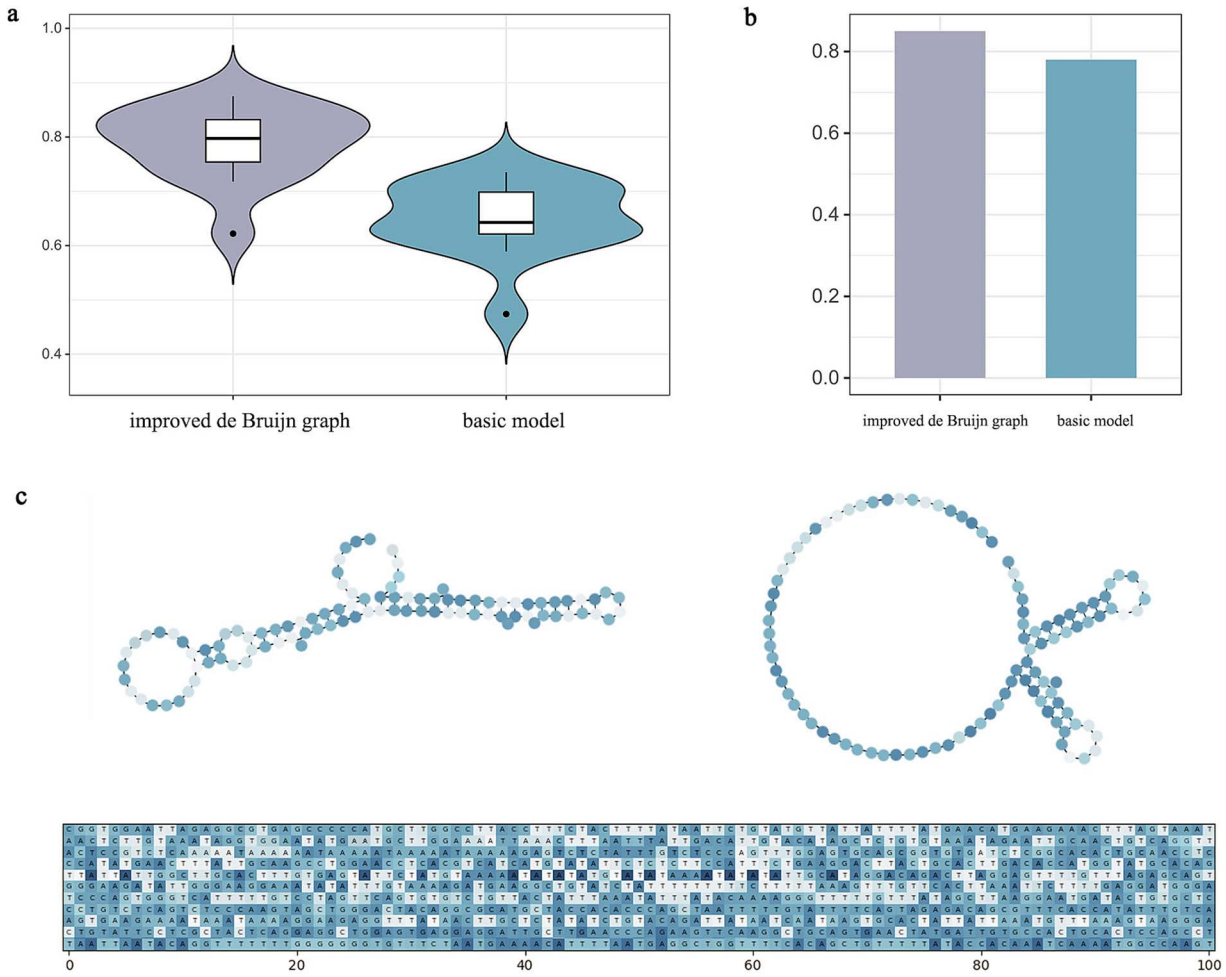

Figure 5. (a) Using violin plots to represent the comparison of predicting target protein binding sites in 15 circRNA datasets with and without using the improved de Bruijn graph. (b) Comparison of using the improved de Bruijn graph in enhancer identification tasks. (c) Visual display of the contribution of each base pair to the recognition task (taking hsa_circ_0028545 and hsa_circ_0001575 as examples) and the interpretability heatmap for partial data in the AUF1 dataset.

graph that can represent RNA sequence information, including base types and positions, and can also encompass potential structural information such as hairpins, bulges, or loop structures. This method is applicable to all nucleotide chains and can be extended to other interaction prediction problems, including enhancer prediction, transcription factor binding site prediction, and RNA-RBP binding site prediction. Secondly, the key to the model's success lies in its network design. At the micro level, it delves into the local patterns and structures of RNA sequences, while, at the macro level, it captures the global relationships among RNA features. On the other hand, when dealing with personalized small datasets, directly fine-tuning pretrained large models often fails to achieve the desired results. Researchers opt for experimental exploration of learning patterns on small datasets, which not only consumes significant time and effort but also has a low success rate. We adopt a multitask training approach to further optimize the model, treating multiple real-world attributes of ncRNAs as different learning tasks and designing a unified model to handle these tasks simultaneously. Furthermore, the multitask model reduces the number of parameters required for training each task separately by sharing underlying parameters. This not only helps to decrease the complexity of the model and effectively avoid overfitting but also allows the model to learn more general feature representations that can be generalized across different tasks, thereby facilitating the model's adaptation to new data and prediction and enhancing the model's learning ability and generalization performance on small datasets. Compared to other models, DVMnet consistently demonstrates better predictive performance across multiple tasks. The potential application prospects of DVMnet are extensive. The versatility of the improved de Bruijn graph suggests its wide applicability in other nucleic acid–related tasks, where it can serve as a foundational feature extraction algorithm for a multitude of applications. Furthermore, the model's multitask training strategy allows for the integration of various task types, such as transcription-factor binding site prediction and RNA-RBP binding site prediction. This not only facilitates knowledge sharing across different tasks but also significantly boosts the model's learning capability and generalization performance, especially when dealing with small datasets.

Key Points

- This paper proposes an improved de Bruijn graph algorithm, which is particularly suitable for nucleic acid sequences based on base pairing principles and can better express the structural characteristics of nucleic acids that implicate biological information.

- Multitask learning is employed, training by coordinating the losses of multiple tasks and providing insights for enhancing the learning ability of small datasets.
- For RNA sequences, learning is approached from two aspects: sequence information and structural information, with learning conducted from different viewpoints or perspectives.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work is supported by the Fundamental Research Funds for the Provincial Universities of Liaoning (No.LJ212410150016).

Data availability

The source code and implementation details of DVMnet are freely available at the GitHub repository (https://github.com/liuliwei1980/DVMnet).

References

1. Guan D, Zhang W, Liu G. *et al.* Switching cell fate, ncRNAs coming to play. *Cell Death Dis* 2013;**4**:e464–4. https://doi.org/10.1038/cddis.2012.196.
2. Cao Y, Geddes TA, Yang JYH. *et al.* Ensemble deep learning in bioinformatics. *Nat Mach Intell* 2020;**2**:500–8. https://doi.org/10.1038/s42256-020-0217-y.
3. Zhang Z-Y, Ning L, Ye X. *et al.* iLoc-miRNA: extracellular/intracellular miRNA prediction using deep BiLSTM with attention mechanism. *Brief Bioinform* 2022;**23**:bbac395. https://doi.org/10.1093/bib/bbac395.
4. Wang J, Li J, Yue K. *et al.* NMCMDA: neural multicategory MiRNA–disease association prediction. *Brief Bioinform* 2021;**22**:bbab074. https://doi.org/10.1093/bib/bbab074.
5. Chen Q, Meng X, Liao Q. *et al.* Versatile interactions and bioinformatics analysis of noncoding RNAs. *Brief Bioinform* 2019;**20**:1781–94. https://doi.org/10.1093/bib/bby050.
6. Franco-Zorrilla JM, Valli A, Todesco M. *et al.* Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat Genet* 2007;**39**:1033–7. https://doi.org/10.1038/ng2079.
7. Du H, Zhao Y, Wen J. *et al.* LncRNA DCRT protects against dilated cardiomyopathy by preventing NDUFS2 alternative splicing by binding to PTBP1. *Circulation* 2024;**150**:1030–49. https://doi.org/10.1161/CIRCULATIONAHA.123.067861.
8. Barnett MM, Reay WR, Geaghan MP. *et al.* miRNA cargo in circulating vesicles from neurons is altered in individuals with schizophrenia and associated with severe disease. *Sci Adv* 2023;**9**:eadi4386. https://doi.org/10.1126/sciadv.adi4386.
9. Chen X, Zhu CC, Yin J. Ensemble of decision tree reveals potential miRNA-disease associations. *PLoS Comput Biol* 2019;**15**:e1007209. https://doi.org/10.1371/journal.pcbi.1007209.
10. Chen X, Li TH, Zhao Y. *et al.* Deep-belief network for predicting potential miRNA-disease associations. *Brief Bioinform* 2021;**22**:bbaa186. https://doi.org/10.1093/bib/bbaa186.
11. Wang CC, Han CD, Zhao Q. *et al.* Circular RNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2021;**22**:bbab286. https://doi.org/10.1093/bib/bbab286.
12. Chen X, Yan CC, Zhang X. *et al.* Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2017;**18**:558–76. https://doi.org/10.1093/bib/bbw060.
13. Sun F, Sun J, Zhao Q. A deep learning method for predicting metabolite–disease associations via graph neural network. *Brief Bioinform* 2022;**23**:bbac266. https://doi.org/10.1093/bib/bbac266.
14. Xu M, Chen Y, Xu Z. *et al.* MiRLoc: predicting miRNA subcellular localization by incorporating miRNA–mRNA interactions and mRNA subcellular localization. *Brief Bioinform* 2022;**23**:bbac044. https://doi.org/10.1093/bib/bbac044.
15. Xiao Y, Cai J, Yang Y. *et al.* Prediction of microrna subcellular localization by using a sequence-to-sequence model. *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, Singapore, 2018;1332–7. https://doi.org/10.1109/ICDM.2018.00181.
16. Yang Y, Fu X, Qu W. *et al.* MiRGOFS: a GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA–disease association. *Bioinformatics* 2018;**34**:3547–56. https://doi.org/10.1093/bioinformatics/bty343.
17. Wang W, Guan X, Khan MT. *et al.* LMI-DForest: a deep forest model towards the prediction of lncRNA-miRNA interactions. *Comput Biol Chem* 2020;**89**:107406. https://doi.org/10.1016/j.compbiolchem.2020.107406.
18. Kang Q, Meng J, Cui J. *et al.* PmliPred: a method based on hybrid model and fuzzy decision for plant miRNA–lncRNA interaction prediction. *Bioinformatics* 2020;**36**:2986–92. https://doi.org/10.1093/bioinformatics/btaa074.
19. Zhou Z, Zhuo L, Fu X. *et al.* Joint masking and self-supervised strategies for inferring small molecule-miRNA associations. *Molecular Therapy-Nucleic Acids* 2024;**35**:102103. https://doi.org/10.1016/j.omtn.2023.102103.
20. Zhou Z, Du Z, Wei J. *et al.* MHAM-NPI: predicting ncRNA-protein interactions based on multi-head attention mechanism. *Comput Biol Med* 2023;**163**:107143. https://doi.org/10.1016/j.compbiomed.2023.107143.
21. Zhang X, Liu M, Li Z. *et al.* Fusion of multi-source relationships and topology to infer lncRNA-protein interactions. *Molecular Therapy-Nucleic Acids* 2024;**35**:102187. https://doi.org/10.1016/j.omtn.2024.102187.
22. Bai T, Yan K, Liu B. DAmiRLocGNet: miRNA subcellular localization prediction by combining miRNA–disease associations and graph convolutional networks. *Brief Bioinform* 2024;**24**:bbad212. https://doi.org/10.1093/bib/bbad212.
23. Peng L, Huang L, Su Q. *et al.* LDA-VGHB: identifying potential lncRNA–disease associations with singular value decomposition, variational graph auto-encoder and heterogeneous Newton boosting machine. *Brief Bioinform* 2024;**25**:bbad466. https://doi.org/10.1093/bib/bbad466.
24. Li M, Zhao B, Yin R. *et al.* GraphLncLoc: Long non-coding RNA subcellular localization prediction using graph convolutional networks based on sequence to graph transformation. *Brief Bioinform* 2023;**24**:bbac565. https://doi.org/10.1093/bib/bbac565.
25. Wei J, Zhuo L, Pan S. *et al.* Headtailtransfer: an efficient sampling method to improve the performance of graph neural network method in predicting sparse ncrna–protein interactions. *Comput Biol Med* 2023;**157**:106783. https://doi.org/10.1016/j.compbiomed.2023.106783.
26. Liu L, Wei Y, Zhang Q. *et al.* SSCRB: predicting circRNA-RBP interaction sites using a sequence and structural

feature-based attention model. *IEEE J Biomed Health Inform* 2024;**3**:1762–72.
27. Zhang S, Zhou J, Hu H. *et al.* A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res* 2016;**44**:e32–2. https://doi.org/10.1093/nar/gkv1025.
28. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 2003;**31**:3406–15. https://doi.org/10.1093/nar/gkg595.
29. Wang W, Zhang L, Sun J. *et al.* Predicting the potential human lncRNA–miRNA interactions based on graph convolution network with conditional random field. *Brief Bioinform* 2022;**23**:bbac463. https://doi.org/10.1093/bib/bbac463.
30. Chen X, Sun L-G, Zhao Y. NCMCMDA: miRNA–disease association prediction through neighborhood constraint matrix completion. *Brief Bioinform* 2021;**22**:485–96. https://doi.org/10.1093/bib/bbz159.
31. Yuan G-H, Wang Y, Wang G-Z. *et al.* RNAlight: a machine learning model to identify nucleotide features determining RNA subcellular localization. *Brief Bioinform* 2023;**24**:bbac509. https://doi.org/10.1093/bib/bbac509.
32. Amin N, McGrath A, Chen YPP. Evaluation of deep learning in non-coding RNA classification. *Nat Mach Intell* 2019;**1**:246–56. https://doi.org/10.1038/s42256-019-0051-2.
33. Gong J, Xu K, Ma Z. *et al.* A deep learning method for recovering missing signals in transcriptome-wide RNA structure profiles from probing experiments. *Nat Mach Intell* 2021;**3**:995–1006. https://doi.org/10.1038/s42256-021-00412-0.
34. Sze S-H, Tarone AM. A memory-efficient algorithm to obtain splicing graphs and de novo expression estimates from de Bruijn graphs of RNA-Seq data. *BMC Genomics* 2014;**15**:1–12.
35. Ye Y, Tang H. Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis. *Bioinformatics* 2019;**32**:1001–8. https://doi.org/10.1093/bioinformatics/btv510.
36. Sener O, Koltun V. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems* 2018;**31**.
37. Swersky K, Snoek J, Adams RP. Multi-task bayesian optimization. *Advances in neural information processing systems* 2013;**26**.
38. Wang P, Guo Q, Qi Y. *et al.* LncACTdb 3.0: an updated database of experimentally supported ceRNA interactions and personalized networks contributing to precision medicine. *Nucleic Acids Res* 2022;**50**:D183–9. https://doi.org/10.1093/nar/gkab1092.
39. Zhang T, Tan P, Wang L. *et al.* RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res* 2017;**45**:D135–8. https://doi.org/10.1093/nar/gkw728.
40. Volders P-J, Anckaert J, Verheggen K. *et al.* LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res* 2019;**47**:D135–9. https://doi.org/10.1093/nar/gky1031.
41. Zhao Y, Li H, Fang S. *et al.* NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res* 2016;**44**:D203–8. https://doi.org/10.1093/nar/gkv1252.
42. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 2019;**47**:D155–62. https://doi.org/10.1093/nar/gky1141.
43. Yu X, Jiang L, Jin S. *et al.* preMLI: a pre-trained method to uncover microRNA–lncRNA potential interactions. *Brief Bioinform* 2022;**23**:bbab470. https://doi.org/10.1093/bib/bbab470.
44. Wang Z, Liang S, Liu S. *et al.* Sequence pre-training-based graph neural network for predicting lncRNA-miRNA associations. *Brief Bioinform* 2023;**24**:bbad317. https://doi.org/10.1093/bib/bbad317.
45. Liu S, Johns E, Davison AJ. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, 2019;1871–80.
46. Witten JT, Ule J. Understanding splicing regulation through RNA splicing maps. *Trends Genet* 2011;**27**:89–97. https://doi.org/10.1016/j.tig.2010.12.001.
47. Liu B, Fang L, Long R. *et al.* iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics* 2016;**32**:362–9. https://doi.org/10.1093/bioinformatics/btv604.

Briefings in Bioinformatics, 2025, **26(1)**, bbae627
https://doi.org/10.1093/bib/bbae627
Problem Solving Protocol