BMC Bioinformatics

RESEARCH Open Access

DeepMiRBP: a hybrid model for predicting microRNA-protein interactions based on transfer learning and cosine similarity

Sasan Azizian[1] and Juan Cui[1*]

*Correspondence:
jcui@unl.edu

[1] School of Computing, University of Nebraska-Lincoln, 1400 R St, Lincoln, NE 68588-0115, USA

Abstract

Background: Interactions between microRNAs and RNA-binding proteins are crucial for microRNA-mediated gene regulation and sorting. Despite their significance, the molecular mechanisms governing these interactions remain underexplored, apart from sequence motifs identified on microRNAs. To date, only a limited number of microRNA-binding proteins have been confirmed, typically through labor-intensive experimental procedures. Advanced bioinformatics tools are urgently needed to facilitate this research.

Methods: We present DeepMiRBP, a novel hybrid deep learning model specifically designed to predict microRNA-binding proteins by modeling molecular interactions. This innovation approach is the first to target the direct interactions between small RNAs and proteins. DeepMiRBP consists of two main components. The first component employs bidirectional long short-term memory (Bi-LSTM) neural networks to capture sequential dependencies and context within RNA sequences, attention mechanisms to enhance the model's focus on the most relevant features and transfer learning to apply knowledge gained from a large dataset of RNA-protein binding sites to the specific task of predicting microRNA-protein interactions. Cosine similarity is applied to assess RNA similarities. The second component utilizes Convolutional Neural Networks (CNNs) to process the spatial data inherent in protein structures based on Position-Specific Scoring Matrices (PSSM) and contact maps to generate detailed and accurate representations of potential microRNA-binding sites and assess protein similarities.

Results: DeepMiRBP achieved a prediction accuracy of 87.4% during training and 85.4% using testing, with an F score of 0.860. Additionally, we validated our method using three case studies, focusing on microRNAs such as miR-451, -19b, -23a, -21, -223, and -let-7d. DeepMiRBP successfully predicted known miRNA interactions with recently discovered RNA-binding proteins, including AGO, YBX1, and FXR2, identified in various exosomes.

Conclusions: Our proposed DeepMiRBP strategy represents the first of its kind designed for microRNA-protein interaction prediction. Its promising performance underscores the model's potential to uncover novel interactions critical for small RNA sorting and packaging, as well as to infer new RNA transporter proteins. The

methodologies and insights from DeepMiRBP offer a scalable template for future small RNA research, from mechanistic discovery to modeling disease-related cell-to-cell communication, emphasizing its adaptability and potential for developing novel small RNA-centric therapeutic interventions and personalized medicine.

Keywords: MicroRNAs, RNA binding proteins, Interaction prediction, RNA sorting, Deep learning

Introduction

RNA-binding proteins (RBPs), with their ability to bind directly to single and double-stranded RNA molecules, play a central role in RNA processing and various cellular activities linked to RNA's function [7, 10, 26]. Recently, interactions between proteins and small non-coding RNAs, specifically microRNAs(miRNAs), have garnered significant attention due to their profound impact on gene expression regulation [4]. Mature miRNA molecules, approximately 22 nucleotides in length, assemble into the RNA-induced silencing complex (RISC) comprised of Ago2, TRBP, PACT, and Dicer, and activate the complex to target messenger RNA (mRNAs) specified by the miRNA, leading to mRNA degradation or translational repression [13]. Beyond maintaining cellular homeostasis, disruptions in miRNA regulation have been implicated in many diseases, ranging from cancer to cardiovascular and neurological disorders [22].

Recent studies have revealed that miRNA molecules can be selectively incorporated into multivesicular bodies (MVBs) and subsequently released as exosomes, known as exomiRNAs. This process hints at RNA transporter proteins and specific sequence motifs that might play a role in miRNA sorting [16, 23, 27]. The selective packaging and dispatching of miRNAs to circulation and their subsequent integration into recipient cells coordinate biological processes across different tissues and organs, demonstrating the precision and complexity of cellular communication [48]. For example, releasing miR-105 in breast cancer exosomes promotes tumor growth in distant tissues like the lungs and brain [14, 52]. Understanding miRNA sorting mechanisms has therapeutic potential and implications in disease progression, though the exact mechanisms remain understudied, beckoning further exploration.

Prior studies have demonstrated that short sequence motifs of miRNA are responsible for its secretion [15, 48]. For instance, bioinformatics analysis has identified conserved 4-mers among exosomal miRNAs such as [AGU]G[AG]G in human T cells and [CGU][UA][GU]G in colon cancer cells [15, 45]. Experiments show that the mutating these motif sequences significantly decreased miRNA levels in exosomes versus cells compared to the wild type, indicating that exomiR sorting depends on the presence of these motifs. Current research has also identified miRNA-binding proteins responsible for sorting miRNAs with specific motifs, such as hnRNPA2B1 in human primary T cells [48] and Sdpr and Fus in adipocyte cells [16].

To further elucidate the protein-mediated miRNA sorting and packaging beyond motif analysis, we need an efficient discovery tool that can enable the systematic study of miRNA-protein interactions in an automated and high-throughput manner by leveraging the massive amounts of omics data on sequence, structure, and (mi)RNA-protein interactome available in the field. The advent of deep learning has revolutionized the prediction landscape of RNA-protein interactions. A slew of models, including DeepBind

[1], DeeperBind [17], and models by Zeng et al. [11], have harnessed the prowess of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. Other models, ranging from iDeep [37], iDeepV [38], iDeepE [39], to iDeeps [40], and DanQ [21], have showcased the potential of other architectures in capturing intricate sequence and structure motifs, thereby enhancing the precision of RNA-protein interaction predictions. While extensive research has been conducted on RNA-binding proteins, the specific domain of miRNA-binding proteins is relatively new. Although a general beRBP model based on random forests was applied to explore protein binding sites on miRNA precursors [9, 51], none of the existing models were designed for small RNA analysis. This gap in the research landscape underscores this work's novelty and pioneering nature. Venturing into this nascent domain, we aim to bridge the existing knowledge gap and contribute seminal insights into the intricate dynamics of miRNA-protein interactions.

Although miRNA-protein binding prediction is burgeoning, it is fraught with significant challenges, predominantly due to the sparse availability of specialized miRNA-protein binding datasets. These datasets are crucial for the training, testing, and validation of predictive models, and their scarcity could impede the development of reliable and accurate prediction algorithms [5]. The complexity of miRNA-protein interactions, which exhibit considerable variability across different biological contexts, further complicates the prediction process. In contrast, validated RBP-RNA interactions from ENCODE RIP-chip, eCLIP, and iCLIP experiments include many RBP binding sites on RNA, including miRNA precursors [31]. To some extent, such data is expected to capture the intrinsic RNA-protein binding features important for small RNA analysis.

To address these challenges, we introduce *DeepMiRBP*, a new multimodal deep neural network for miRNA-protein Binding prediction, which integrates sequence and structural information from both (mi)RNA and RBPs. It comprises two main components. The first component leverages transfer learning and cosine similarity [46, 50] for effectively predicting RBP candidates by utilizing the available (mi)RNA-protein interactome datasets. The second component, after obtaining the RBP candidates, expands these candidates by finding new similar proteins based on structural information. Together, both parts offer precise predictions of miRNA-protein interactions. Subsequent sections will delve into the model's details and its implications in molecular biology.

Materials and methods

The overall design

DeepmiRBP is designed to predict miRNA-protein interactions and identify new miRNA-binding proteins. The architecture is divided into two primary components, as illustrated in Fig. 1.

- *First Component:* This component utilizes transfer learning and cosine similarity to identify RBP candidates. The transfer learning module includes the source and target domains for predicting miRNA-protein binding interactions. The source domain is trained using RNA sequences (known as RBP binding sites) to identify features within the RNA sequences that facilitate RBP binding. Once the source domain is adequately trained, the acquired knowledge is transferred to the target domain,

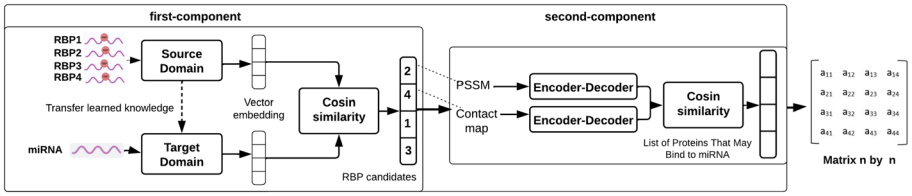

Fig. 1 The schematic workflow of the DeepMiRBP model. In the first component, the source-domain model is trained based on RNA sequences related to known binding sites of different RBPs(RNA-binding proteins). After this training phase, the learned parameters are transferred to the target domain using a transfer learning approach. The target model is then retrained using sequences of protein-interacting miRNAs as input. A cosine similarity measure is applied to identify and rank RBP sequences from the source domain that are most similar to the given miRNA, resulting in a ranked list of candidate proteins. The candidate proteins identified in the first component undergo further analysis in the second component. Position-Specific Scoring Matrices (PSSM) and contact maps are utilized for each candidate protein to perform a more comprehensive similarity assessment. This step enhances the understanding of miRNA-protein interactions, thereby improving the model's prediction accuracy

which takes miRNA sequences as input. In both domains, embedding layers convert the sequences into unique 128-dimensional vectors. Cosine similarity is employed to identify RNA representations most similar to the miRNA representation, leading to a ranked list of candidate RBPs based on similarity scores.

- *Second Component:* This component processes the Position-Specific Scoring Matrix (PSSM) and the protein structure contact map for each RBP candidate identified in the first component, using CNNs to encode these matrices. The primary objective is to compare all RBP candidates with a comprehensive set of proteins to determine which proteins exhibit the highest similarity to the RBP candidates based on sequence and structural information. Cosine similarity is then utilized to evaluate the similarity. The resulting output is an $n \times n$ matrix, where n denotes the number of proteins. Each cell in this matrix represents the similarity between two proteins. From this matrix, we derive a ranked list of proteins with a high probability of binding to the miRNA sequence based on their similarity scores.

The first component provides a comprehensive representation of RBP-RNA interactions and RBP recommendations. Subsequently, the second component refines these predictions by incorporating structural information, ensuring a robust and accurate identification of miRNA-binding proteins.

To ensure the selection of the optimal model architecture and meticulous hyperparameter optimization, over 45 different architectures were initially explored, including various LSTM, CNN, and hybrid models with attention mechanisms. Key hyperparameters were fine-tuned, including embedding dimensions (32, 64, 128, 256, 512, 1024), LSTM units (32, 64, 128, 256, 512, 1024), dropout rates (0.1, 0.2, 0.5), batch sizes (32, 64, 128), and learning rates (0.001, 0.0001). This thorough optimization ensured the model's robustness and high performance across diverse input data types. The choice of 128 dimensions was also fine-tuned through hyperparameter optimization to achieve optimal performance, which renders a balance between capturing detailed information and maintaining computational efficiency. In the following sections, we detail the datasets collected, data preprocessing and refinement techniques, embedding representations, and the design intricacies of the model architectures.

Data collection and preprocessing

This study's foundation is underpinned by meticulously curated datasets encompassing RNA, protein, and miRNA sequences. Herein, we detail the sources and specifics of the data utilized.

- *RNA Sequences:* Our study utilizes a comprehensive dataset of RNA-binding site sequences and their corresponding RNA-binding proteins (RBPs). The primary dataset consists of RNA binding site sequences associated with 154 RBPs, sourced from the benchmark dataset used in RBPSuite [41]. Additionally, sequences for 31 RBPs were extracted from the dataset employed in iDeepS [40], obtained from ENCODE. We further incorporated 65,301 interactions involving 147 RBPs and 1,494 miRNAs, downloaded from the EVPsort database [9], and 18,515 AGO-related human miRNA and mRNA sequences from the CLASH dataset by Helwak et al. [18]. To enhance the robustness of our model, we utilized an extensive dataset comprising 18,380,117 sequences, of which 8,822,297 contain binding sites, while 9,557,820 do not. Each protein file typically contains around 120,000 binding-site sequences, with approximately 60,000 labeled as positive and 60,000 as unfavorable. This dataset provides a substantial foundation for analyzing RNA-protein interactions. The data is further divided into two domains: the *source domain*, containing the broader dataset for model training and testing, and the *target domain*, focused on miRNA sequences, which offers a smaller, specialized dataset for investigating miRNA-protein interactions. Details of the source and target domain datasets are summarized in Table 1, illustrating the sequence distribution and highlighting the distinction between RNA and miRNA datasets.
- *Protein Sequences and Structures:* Protein sequences were primarily derived from the UniProt database [47] and the NCBI Protein Database [44]. To augment our research with protein structural insights, we extracted the contact map information from AlphaFold [43] and ResPRE [28].

The preprocessing of the sequencing-based RNA-RBP binding dataset encompassed several stages. Initially, we merged the binding site peak files for each RBP to consolidate the data. Regions overlapping with the reference gene were selected using the intersect Bed function of bedtools [42]. We extended these regions for gene-overlapped regions with less than 101 base pairs (bp) downstream and upstream to ensure they qualified as positive regions for RBPs. Negative RBP binding regions, each 101 bp, were generated using shuffleBed from bedtools. Fasta files for positive

Table 1 Overview of Source and Target Domain Datasets

Domain	Data	Positive (Sequence)	Negative (Sequence)
Source	Training	7,940,067	8,602,038
	Testing	882,230	955,782
Target	Training	16,665	15,801
	Testing	1850	1755

and negative regions were retrieved using fastaFromBed of bedtools. To maintain a balanced dataset, only 60,000 positive and 60,000 negative sites for each RBP were retained if the extracted samples exceeded this number; otherwise, all extracted samples were utilized.

Input representation using embeddings

Different embedding techniques were applied to represent RNA and protein sequences, including the following.

- *RNA Embedding:* For RNA sequences of 101 characters, each character is transformed into a unique 128-dimensional vector, resulting in a matrix of size 101×128. To obtain a single embedding for the entire sequence, vectors are summed along the columns:

$$\mathbf{v}_{\text{RNA}} = \sum_{i=1}^{101} \mathbf{v}_i$$

  miRNA sequences, typically shorter than 25 characters, are padded with zeros to match the required input length of 101 characters. Each character is transformed into a 128-dimensional vector. For miRNA embedding, we sum only up to the original sequence length:

$$\mathbf{v}_{\text{miRNA}} = \sum_{i=1}^{\text{Len(miRNA)}} \mathbf{v}_i$$

- *Position-Specific Scoring Matrix (PSSM)* was utilized for a more nuanced representation of protein sequences, capturing evolutionary information and sequence conservation. Derived from multiple sequence alignments of related proteins, PSSMs provide log-odds scores for each amino acid at specific positions. These scores indicate the significance of observing a particular amino acid at a specific position relative to its expected frequency [2, 19, 20, 25]. The formula for the log-odds score is:

$$\text{Log-Odds Score (a,i)} = \log_2 \left(\frac{\text{Frequency of a at position i}}{\text{Background frequency of a}} \right)$$

- *Protein Structure Contact Map (PSCM):* To incorporate the spatial relationships between amino acid residues and the tertiary structure into our model, we utilized PSCMs as a two-dimensional matrix representation of the three-dimensional protein structure. The ResPRE algorithm, a deep learning-based method for predicting residue-level contacts [28], was integrated to generate these maps as part of our pipeline. Additionally, contact maps predicted from AlphaFold were downloaded. In the PSCM, residues are marked as '1' if the distance between them falls below a defined threshold, typically within 6–8 Ångstroms, indicating they are in contact. Otherwise, the matrix cell is marked as '0'. This approach results in a symmetric matrix representation of protein structures.

These embedding techniques play a crucial role in our study, transforming raw sequence data into formats more amenable to analysis and interpretation by our deep learning models.

Model architecture

Figure 2 shows the detailed architecture of the multimodal deep-learning framework of DeepMiRBP. Primary methods are described in the following sections.

First component: transfer learning framework

In a transfer learning framework, the source domain is trained using a large set of RNA binding sites known to interact with RBPs. The knowledge acquired from this source domain is transferred to the target domain, where miRNA sequences serve as input. The supplementary document provides detailed explanations of the architecture, including the embedding layer, LSTM, attention mechanisms, and the training process.

Sampling methodology

Following the initial training of the first model component, it was essential to evaluate the similarity between RNA-binding proteins (RBPs) and microRNAs (miRNAs) using cosine similarity. Due to the vast number of sequences associated with each RBP, computing similarity across all RNA sequences presented substantial computational challenges. To address this, we employed a targeted sampling strategy.

To ensure efficiency and relevance in our analysis, we exclusively sampled 1000 RNA sequences for each RBP, focusing solely on those confirmed to bind to proteins (positive sequences). This focus on positive sequences was crucial, as our primary objective was to identify similarities between RNA sequences with known binding activity and miRNAs. Including only positive sequences allowed us to maintain the relevance of the similarity analysis, avoiding any noise introduced by non-binding (negative) sequences.

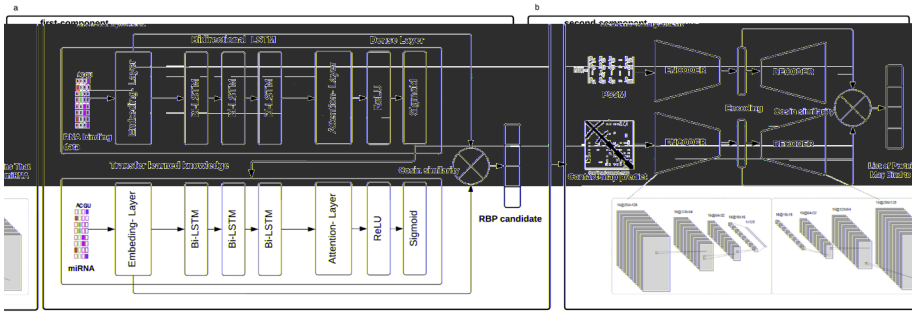

Fig. 2 The detailed architecture of DeepMiRBP in both components for predicting microRNA-protein interactions. **a** First component architecture: This model focuses on training RNA sequences that bind to RBPs to capture intricate features of RNA-protein interactions. Initially, the model learns from RNA sequences bound by RBPs and transfers this knowledge to the target domain. Here, miRNA sequences serve as input, generating embedding codes. Cosine similarity is then applied to identify RNA sequences most similar to the miRNA sequences. **b** Second component architecture: In this model, each RBP candidate identified in the first part is processed using PSSM and contact maps. CNN layers and max-pooling are employed to encode these matrices. Subsequently, cosine similarity is calculated to compare RBP candidates with other proteins, resulting in a matrix that identifies proteins with a higher probability of binding to the miRNA sequence

This sampling strategy was designed to capture the essential characteristics and variability of the dataset without exhaustive computations, ensuring the robustness and representativeness of the model. The Central Limit Theorem (CLT) underpins this approach, guaranteeing that the distribution of sample means approximates a normal distribution when the sample size is sufficiently large. In our context, the CLT is expressed as:

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

where $\bar{X}_n$ represents the sample mean, μ denotes the population mean, σ^2 indicates the population variance, and n is the sample size.

By selecting 1000 positive RNA sequences for each RBP, we ensured that the sample size was sufficient to approximate the underlying data distribution's normality. To achieve a comprehensive understanding and capture the variability within the dataset, we repeated this sampling process multiple times, we conducting 30 independent sampling iterations for each RBP. This repeated sampling approach helps capture a wide range of possible data variations, providing a statistically sound basis for subsequent analyses. By concentrating on positive samples, we avoided the inclusion of irrelevant negative samples, which could introduce noise and dilute the accuracy of our findings.

This targeted sampling approach not only streamlined the computational process but also preserved the statistical integrity of the dataset, ensuring that our model's training and similarity assessments were based on a representative and meaningful subset of the data.

After completing the sampling and training phases, we computed the similarity between the sampled RBP-binding RNA sequences and miRNAs. This method substantially reduced computational overhead while retaining the accuracy and effectiveness of our similarity assessments by focusing on RNA sequences with confirmed binding activity, aligning precisely with the study's objectives.

Similarity calculation

The main objective of the first component is to identify RNA sequences similar to the miRNA using cosine similarity. Cosine similarity is a metric that measures the similarity between two non-zero vectors in an inner product space. It is calculated as the cosine of the angle between the vectors, providing a measure to evaluate the degree of similarity between sets of embeddings. This metric is particularly advantageous in high-dimensional spaces where traditional Euclidean distance may not accurately capture subtle nuances of vector similarity.

The cosine similarity between two vectors A and B is given by:

$$\text{cosine\_similarity} = \frac{A \cdot B}{\|A\|\|B\|}$$

where $A \cdot B$ denotes the dot product of vectors A and B, and $\|A\|$ and $\|B\|$ represent the Euclidean norms (magnitudes) of the vectors. Mathematically, the dot product $A \cdot B$ is calculated as:

$$A \cdot B = \sum_{i=1}^{n} A_i B_i$$

The Euclidean norm of a vector A is calculated as:

$$\|A\| = \sqrt{\sum_{i=1}^{n} A_i^2}$$

Similarly, the Euclidean norm of a vector B is:

$$\|B\| = \sqrt{\sum_{i=1}^{n} B_i^2}$$

The embedding codes derived from RBP sequences (Source Domain) and miRNA sequences (Target Domain) serve as input vectors for calculating cosine similarity. This computation generates a list indicating which RBPs are most likely to bind to the miRNA, summarized as a vector across all trained data.

Cosine similarity measures the cosine of the angle between two vectors. A cosine similarity of 1 indicates maximum similarity (if the vectors are identical), while a similarity of 0 suggests no similarity (if the vectors are orthogonal, at a 90-degree angle). This angular measure provides a normalized similarity score independent of the vectors' magnitudes, which is particularly advantageous in applications where vector scales vary significantly.

Model evaluations

The performance of the source domain in the first component of our model was evaluated using a dataset consisting of 188 RBP sequences. We employed a 90/10 split for data division, allocating 90% of the data for training and 10% for testing. To ensure robustness and reliability, we implemented a 10-fold cross-validation approach, repeated ten times to mitigate the impact of any potential randomness in the training process. The Adam optimizer was utilized for optimization during the training process. To evaluate the model's performance on training and testing datasets, we employed the following metrics:

- *Accuracy* measures the overall correctness of the model by calculating the ratio of correctly predicted interactions (both true positives and true negatives) to the total observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- *Precision* indicates the quality of positive predictions by measuring the ratio of correctly predicted positive interactions to the total predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- *Recall* (sensitivity) measures the model's ability to identify all relevant positive interactions by calculating the ratio of correctly predicted positive observations to all observations in the actual class.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- *F1 Score* is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between precision and recall.

$$F1\ Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Design of the case studies

To comprehensively evaluate DeepmiRBP's performance in identifying miRNA-binding proteins, we designed three case studies:

- Case Study 1: based on miRNA interactions with RBPs that are included in the model. The source domain comprises 188 RBPs. We curated new miRNA interactions validated with RBPs from recent literature. This case study aims to assess whether the model accuractely identifies the binding proteins for these miRNAs based on its training data.
- Case Study 2: based on miRNAs interactions with new RBPs that are excluded in the model. In this scenario, we focus on miR-223 known to interact with exosomal protein YBX1. YBX1, although not included in our training data, plays a crucial role in packaging miR-223 into exosomes through liquid-liquid phase separation, as evidenced by Liu et al. [29]. This case study tests DeepmiRBP's ability to generalize to new RBPs not encountered during training.
- Case Study 3: to identify novel miRNA sorting proteins for selected exosomes. This case study aims to illustrate how to use DeepmiRBP to identify miRNA transporter proteins in exosomes of interest, e.g., from cancer cells by leveraging miRNA and protein profiles of cancer-derived exosomes. Taking let-7 as an example, this miRNA family has been extensively studied for its tumor-suppressive properties. According to Johnson et al. [24], the miR-let-7 represses cell proliferation pathways in human cells, highlighting its potential as a therapeutic target. Furthermore, Nwaeburu et al. [35] demonstrated that the up-regulation of miRNA-let-7c by quercetin inhibits pancreatic cancer progression by activating Numbl. These findings underscore the critical role of the let-7 family in combating cancers. We utilized EVPsort [9] and public data of miRNA and protein profiles specific to cancer-derived exosomes to obtain data for this test case. This case study highlights the importance of combining public and user data to advance our understanding of miRNA-protein interactions in disease contexts. We aim to uncover novel miRNA transporter proteins that could serve as potential cancer therapeutic targets.

We will discuss these case studies in the next section, focusing on the model's performance evaluation and its implications for predicting miRNA-protein interactions.

Table 2 Performance metrics for source and target models

Domain	Data	Accuracy	Precision	Recall	F1
Source	Training	0.862	0.849	0.885	0.867
	Testing	0.824	0.811	0.851	0.831
Target	Training	0.874	0.864	0.896	0.880
	Testing	0.854	0.843	0.877	0.860

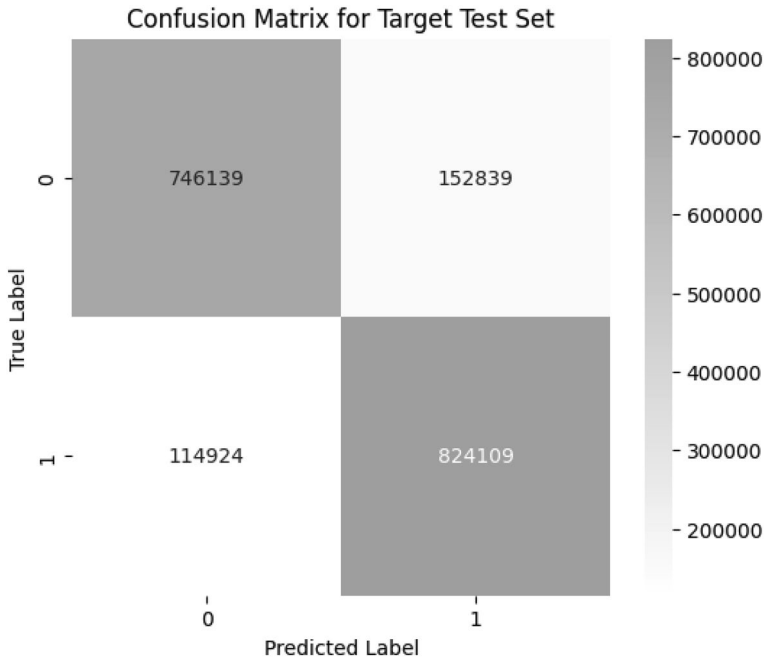

Fig. 3 Confusion matrix for test data in the source domain

Results

Model performance

After training, the comprehensive evaluation of the source and target domains, as summarized in Table 2, indicates DeepMiRBP's robust capability and effectiveness in predicting RNA-binding proteins.

In the source domain, the model demonstrated commendable performance metrics, with an accuracy of 82.4% on the test dataset across all RBPs (see Table 2), indicating the model's robust capability to identify RNA-binding sites correctly. A precision of 81.1% reflects the model's proficiency in accurately detecting true positive interactions while minimizing false positives. A recall of 85.1% highlights the model's ability to identify a substantial proportion of true interactions. Last, the F1 score of 0.831, balances precision and recall, confirming the model's overall reliability and robustness.

The confusion matrix for the source domain test data (Fig. 3) further illustrates the model's performance, providing a detailed view of the true positive, true negative, false positive, and false negative predictions. This visualization reinforces the quantitative metrics in Table 2 and offers deeper insight into the model's prediction accuracy.

Following the training of the source domain, the acquired knowledge and parameters were transferred to the target domain through transfer learning. In this phase, miRNAs known to bind to AGO family proteins were input to ensure the comprehensive functionality of the entire framework. The target domain's performance, with an accuracy of 85.4% on the test data, demonstrates the successful integration and efficacy of both the source and target domains.

Additionally, the source domain's training set achieved an accuracy of 86.2%, a precision of 84.9%, a recall of 88.5%, and an F1 score of 0.867. The target domain's training set reported an accuracy of 87.4%, a precision of 86.4%, a recall of 89.6%, and an F1 score of 0.880. These metrics highlight the model's strong performance across both domains.

In summary, the results from the source and target domains establish a solid foundation for our model, demonstrating its effectiveness in accurately predicting RNA-binding proteins. The high accuracy and balanced performance metrics in both domain validate the model's reliability. The subsequent sections will present the results from the three case studies, further illustrating the model's application and performance in real-world scenarios.

Comparing DeepMiRBP (Source Domain) with other state-of-the-art methods

To evaluate the performance of the initial component of DeepMiRBP, we conducted experiments using a dataset comprising 248,000 binding sites and 992,000 non-binding sites from 31 RNA-binding proteins (RBPs), as previously utilized by iDeepS. According to the original study, the dataset for each protein was divided into 24,000 instances for training, 6,000 instances for model optimization and validation, and 10,000 instances for independent testing. This setup ensured a fair and consistent comparison with several advanced models, including iDeepS, DeepBind, DeeperBind, Oli [30], GraphProt [32], and iDeepV.

Given the significant class imbalance in the dataset, where non-binding sites vastly outnumber binding sites, we employed specific techniques to address this issue without altering the dataset. We applied a combination of Focal Loss and Class Weights Adjustment to ensure robust model training while maintaining the original data distribution. Focal Loss was used to focus the model's learning on the minority class (binding sites) by dynamically down-weighting the loss contribution of well-classified examples, thereby enhancing learning from hard-to-classified instances. Simultaneously, Class Weights Adjustment was implemented to assign higher weights to the minority class in the loss function, ensuring that predictions for binding sites were treated with greater importance during training.

By applying these techniques, DeepMiRBP consistently achieved better or comparable Area Under the Curve (AUC) values than the state-of-the-art models, confirming its effectiveness as a powerful tool for advancing our understanding of RNA-binding mechanisms. As detailed in Table 3, DeepMiRBP achieved an average AUC of 0.865, surpassing iDeepS (0.861), DeepBind (0.854), DeeperBind (0.857), Oli (0.767), GraphProt (0.819), and iDeepV (0.840). Notably, DeepMiRBP demonstrated superior performance for 17 out of the 31 proteins, including TAF15 and MutFUS, where it attained AUC values of 0.981 and 0.979, respectively-outperforming the other models.

Furthermore, DeepMiRBP outperformed sequence-only models such as iDeepV, DeepBind, and Oli, which have shown competitive performance against methods incorporating both sequence and structural information, like iDeepS and GraphProt. For example, while iDeepV achieved an average AUC slightly lower than that of DeepMiRBP, our results suggest that the advanced deep learning architecture of DeepMiRBP more effectively captures essential features, even without integrating structural information. However, for some proteins, such as SRSF1, DeepMiRBP's performance was marginally lower than that of sequence-structure models, reflecting the complex nature of miRNA-mediated RNA-protein interactions.

The receiver operating characteristic (ROC) analysis across the 31 experiments (see Fig. 4) indicated variability in performance, with AUC values ranging from 0.678 for hnRNPL-2 to 0.981 for TAF15. These findings suggest that DeepMiRBP provides a robust alternative to existing models, particularly for challenging proteins where traditional sequence- or structure-based models may underperform. The strategic use of Focal Loss combined with Class Weights Adjustment effectively addressed the data imbalance issue without altering the dataset, reinforcing DeepMiRBP's potential as a valuable tool for biological research.

Extended ablation study and comparison with baseline models

We conducted an extended ablation study to thoroughly evaluate the contribution of each component in the DeepMiRBP model and compare its performance with other baseline models. Specifically, we evaluate how the attention mechanism, LSTM units, dropout layers, embedding dimensions, and the choice between bidirectional and

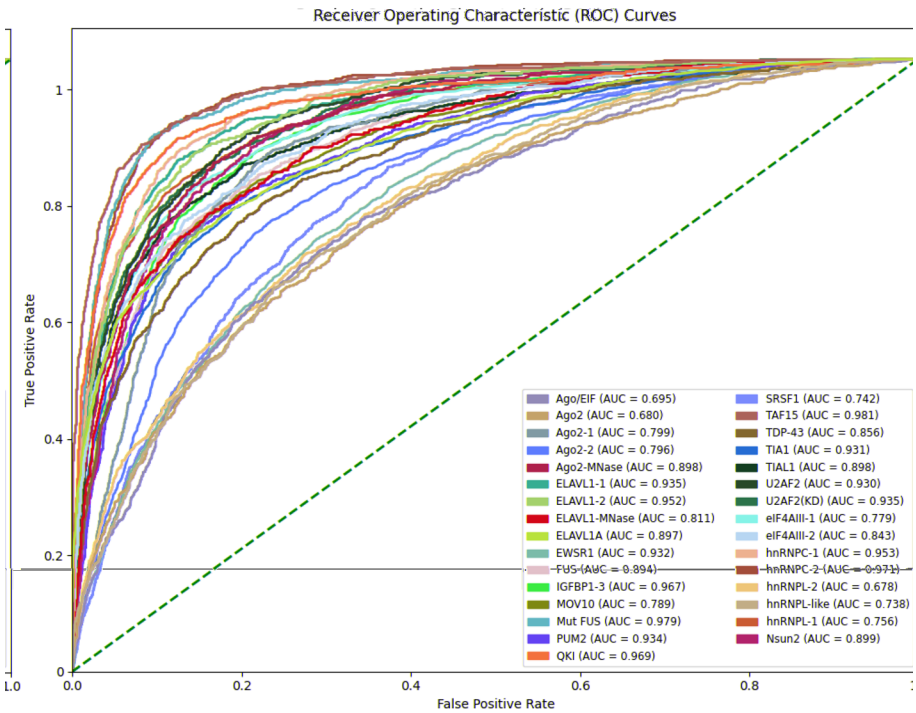

Fig. 4 ROC Performance. The ROC curve for predicting RNA-protein binding sites on 31 experiment datasets

Table 3 The performance of iDeeps, iDeepV, DeepBind, DeeperBind, Oli, and GraphProt are taken from original papers [40] on the same datasets with DeepMiRBP

Protein	DeepMiRBP	iDeepS	DeepBind	DeeperBind	Oli	iDeepV	GraphProt
Ago/EIF	0.695	**0.773**	0.713	0.740	0.610	0.732	0.691
Ago2-MNase	**0.898**	0.591	0.595	0.606	0.512	0.571	0.595
Ago2-1	0.799	**0.865**	0.849	0.857	0.803	0.844	0.817
Ago2-2	0.796	**0.868**	0.830	**0.868**	0.800	0.832	0.823
Ago2	**0.680**	0.634	0.628	0.630	0.534	0.615	0.633
eIF4AIII-1	0.779	**0.950**	0.938	**0.950**	0.919	0.943	0.918
eIF4AIII-2	0.843	0.953	0.950	**0.954**	0.929	0.942	0.931
ELAVL1-1	**0.935**	0.932	0.924	0.930	0.889	0.912	0.915
ELAVL1-MNase	**0.811**	0.600	0.613	0.614	0.491	0.590	0.591
ELAVL1A	**0.897**	0.893	0.886	0.893	0.843	0.891	0.867
ELAVL1-2	**0.952**	0.919	0.914	0.919	0.875	0.922	0.895
ESWR1	**0.932**	0.917	0.912	0.915	0.808	0.900	0.840
FUS	0.894	0.934	**0.942**	0.939	0.846	0.931	0.860
Mut FUS	**0.979**	0.958	0.953	0.957	0.822	0.950	0.853
IGFBP1-3	**0.967**	0.717	0.702	0.713	0.569	0.661	0.697
hnRNPC-1	0.953	**0.960**	0.957	0.959	0.885	0.955	0.930
hnRNPC-2	0.971	0.975	0.973	**0.976**	0.941	0.970	0.953
hnRNPL-1	0.756	0.756	**0.771**	0.746	0.392	0.761	0.698
hnRNPL-2	0.678	0.747	**0.769**	0.746	0.474	0.750	0.708
hnRNPL-like	**0.738**	0.708	0.711	0.679	0.562	0.700	0.650
MOV10	0.789	**0.813**	0.804	0.812	0.783	0.771	0.803
Nsun2	**0.899**	0.835	0.803	0.801	0.754	0.850	0.779
PUM2	0.934	**0.962**	0.950	0.955	0.939	0.954	0.914
QKI	**0.969**	0.966	0.962	0.961	0.924	0.966	0.932
SRSF1	0.742	**0.887**	0.874	0.875	0.839	0.864	0.838
TAF15	**0.981**	0.964	0.956	0.963	0.804	0.951	0.850
TDP-43	0.856	**0.930**	0.926	**0.930**	0.883	0.911	0.907
TIA1	**0.931**	0.930	0.924	0.926	0.842	0.922	0.896
TIAL1	**0.898**	0.893	0.888	0.895	0.831	0.614	0.858
U2AF2	0.930	**0.953**	0.941	0.945	0.861	0.946	0.873
U2AF2(KD)	**0.935**	0.931	0.923	0.930	0.840	0.926	0.883
Averages	**0.865**	0.861	0.854	0.857	0.767	0.840	0.819

The boldface indicates this performance is the best among the compared methods

single-directional LSTM impact the model's overall performance, including its accuracy and AUC.

Overall, the ablation study demonstrates that each component of the DeepMiRBP model significantly contributes to its high performance while the detailed explanations are provided in the Supplementary Materials. In particular, the attention mechanism is crucial for identifying and focusing on key sequence features. The optimal number of LSTM units is essential for capturing long-term dependencies, while an appropriate dropout rate prevents overfitting, enhancing the model's robustness. Additionally, the choice of embedding dimensions is critical for maintaining a detailed representation of input sequences, and the use of a bidirectional LSTM further improves context learning. These findings collectively affirm the robustness and

effectiveness of the proposed DeepMiRBP architecture, as evidenced by its superior performance compared to all ablated variants.

Furthermore, we evaluated several different model architectures to understand their effectiveness on the input data used in this study. The architectures tested included various convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models combining CNN and LSTM layers. Some specific architectures are outlined in the Supplementary Materials.

The performance of these models, including metrics like accuracy, AUC, and other relevant measures, was evaluated on the test data. The results demonstrate that the proposed DeepMiRBP model outperforms these baseline models, indicating its superiority in predicting miRNA-binding proteins. Table 4 provides detailed performance metrics and comparison results. These experiments provide valuable insights into which types of models are most effective for handling the input data used in this study. By selecting several basic models as baselines and comparing them with our proposed method, we illustrate the advantages of the DeepMiRBP architecture over simpler alternatives.

Validation on miR-451, miR-19b, miR-23a, and miR-21 (Case Study 1)

After the first component was completely trained, we validated the model using miRNA interactions with RBPs in the training domain. We tested [3] the model with several miRNAs and experimental data to ensure its robustness and accuracy.

- *miR-451:* According to Dueck et al. [12], miR-451 is directly processed by AGO2, which is unusual because AGO2 is not typically involved in miRNA processing; it usually just helps with the sorting and function of miRNAs that have already been processed by Dicer. After processing, miR-451 remains associated with AGO2, which acts as a form of sorting since miR-451 is specifically bound to AGO2. We first obtained samples from each RBP within our domain to validate this and saved the embedding code for each RBP sequence. Next, we provided miR-451 as input to the target domain, calculated the embedding code, and utilized cosine similarity to determine which RBP sequences were most similar to miR-451. The results, shown in Table 5, list the top 10 RBPs with the highest similarity scores: As illustrated, AGO2 has the top score in the table, confirming its exclusive association with miR-451. Interestingly, AGO1, with a score of −0.394, appears much lower in

Table 4 Performance comparison of different architectures and baseline models

Model	Accuracy	Precision	Recall	F1 Score
DeepMiRBP (Proposed)	0.824	0.811	0.851	0.831
Advanced CNN-BiLSTM with regularization	0.814	0.801	0.832	0.812
Modified CNN-BiLSTM model	0.804	0.791	0.831	0.811
CNN-LSTM Hybrid model	0.784	0.771	0.811	0.791
Baseline LSTM model	0.754	0.741	0.771	0.756
Baseline RNN model	0.714	0.701	0.741	0.720
Baseline CNN model	0.704	0.691	0.721	0.706

The table compares the proposed DeepMiRBP model against several tested architectures and baseline models, demonstrating the superiority of DeepMiRBP in terms of accuracy, AUC, and other relevant metrics

Table 5 Top 10 RBPs with highest scores for miR-451

RBP	Similarity score
AGO2	0.67
KHDRBS1	0.04
SFPQ	−0.16
PRPF8	−0.82
SF3B4	−0.87
QKI	−0.92
KHSRP	−0.12
SF3A3	−0.17
HNRNPK	−0.21
SF3B1	−0.24

Table 6 Top common RBPs with highest scores for miR-19b, miR-23a, and miR-21

RBP	Similarity score
AGO1	0.55
AGO2	0.45
HNRNPK	0.26
SERBP1	0.25
NIP7	0.24
PCBP2	0.19
FKBP4	0.17
PCBP1	0.16
PHF6	0.14
IGF2BP3	0.13

the table in the 29th row. This result validates that DeepmiRBP functions correctly in identifying known interactions for miR-451.

- *miR-19b, miR-23a, and miR-21:* According to Dueck et al. [12], miR-19b, miR-23a, and miR-21 are known to associate with Argonaute protein families in vivo, indicating they are processed by Dicer and are not limited to a specific Ago protein. We repeated the sampling and embedding process for these miRNAs to validate our model further. As predicted by our model, the high similarity scores with various Argonaute proteins confirm the expected associations and demonstrate the model's accuracy in predicting miRNA-RBP interactions across multiple miRNAs. The results are shown in Table 6, listing the top RBPs with the highest similarity scores for miR-19b, miR-23a, and miR-21: AGO1 and AGO2 stand at the top, confirming the model's effectiveness. However, it is essential to note that the model provides a list of candidate RBPs ranked by similarity score, ensuring comprehensive identification of potential interactions.

These validation results demonstrate the robustness and reliability of the DeepmiRBP model in accurately predicting miRNA-RBP interactions. The successful

identification of known interactions for miR-451, miR-19b, miR-23a, and miR-21 reinforces the model's effectiveness and lays a solid foundation for further studies.

Validation on miR-223 (Case Study 2)

We used miR-223 [9] as input to our source domain to test the model's ability to predict interactions for miRNAs excluded from the training dataset. miR-223 is known to bind to the YBX1 protein [29], which was not included in our training data. Initially, we provided miR-223 as input to the target domain to identify which RNA sequences that bind to RBPs are more similar to miR-223 sequences. The first component of the model generated a list of candidate RBPs with sequences similar to miR-223. In the subsequent step, we utilized PSSM and contact maps for each candidate from the first component. We then provided each candidate's PSSM and contact map as input to the second component, generating a list of final candidate proteins to which miR-223 could potentially bind.

For miR-223, we identified 25 RBPs from the 188 total RBPs used for training the first component, with similarity scores greater than zero. Table 7 shows the top 10 similarity scores: With this list of candidate RBPs similar to YBX1, we provided each candidate's PSSM and contact map as input to the second component. The second component computed the similarity between each protein, resulting in an $n \times n$ matrix. Table 8 presents the similarity scores for the top 15 proteins, including YBX1. The matrix shows that the top three highest scores are associated with SERBP1, CSDE1, and TIAL1, along with YBX1. This indicates that these proteins would be selected as candidates to which miR-223 could potentially bind. These high similarity scores suggest a strong likelihood of interaction between miR-223 and these candidate RBPs, thereby validating the model's efficacy in predicting miRNA-protein interactions for proteins excluded from the training dataset.

These case studies illustrate the efficacy of our model in predicting miRNA-RBP interactions, even for miRNAs not included in the training domain. The comprehensive approach of combining sequence similarity and structural information through PSSM and contact maps ensures accurate and reliable predictions.

Table 7 Top 10 RBPs with highest scores for miR-223

RBP	Similarity score
TIAL1	0.12
CPEB4	0.12
CSDE1	0.12
SLBP	0.11
SERBP1	0.11
NIPBL	0.11
METAP2	0.11
SDAD1	0.11
APOBEC3C	0.11
ZNF800	0.10

Table 8 Cosine similarity matrix for final candidate proteins for miR-223

	TIAL1	CPEB4	SSB	SLBP	SERBP1	NIPBL	METAP2	SDAD1	APOBEC3C	ZNF800	CSDE1	YBX1	IGF2BP1	SYNCRIP	HSPA1B
TIAL1	1.00	0.35	0.27	0.19	0.44	0.36	0.31	0.32	0.24	0.31	0.40	0.62	0.39	0.45	0.28
CPEB4	0.35	1.00	0.28	0.25	0.38	0.47	0.22	0.38	0.30	0.27	0.39	0.51	0.30	0.37	0.31
SSB	0.27	0.28	1.00	0.41	0.21	0.32	0.37	0.26	0.31	0.28	0.36	0.06	0.38	0.29	0.35
SLBP	0.19	0.25	0.41	1.00	0.35	0.39	0.28	0.32	0.27	0.26	0.32	0.17	0.42	0.24	0.33
SERBP1	0.44	0.38	0.21	0.35	1.00	0.28	0.39	0.31	0.32	0.43	0.42	0.66	0.49	0.43	0.35
NIPBL	0.36	0.47	0.32	0.39	0.28	1.00	0.34	0.39	0.36	0.32	0.34	0.12	0.35	0.36	0.27
METAP2	0.31	0.22	0.37	0.28	0.39	0.34	1.00	0.41	0.32	0.28	0.31	0.02	0.38	0.32	0.30
SDAD1	0.32	0.38	0.26	0.32	0.31	0.39	0.41	1.00	0.36	0.31	0.41	0.08	0.34	0.32	0.39
APOBEC3C	0.24	0.30	0.31	0.27	0.32	0.36	0.32	0.36	1.00	0.29	0.34	0.14	0.38	0.30	0.33
ZNF800	0.31	0.27	0.28	0.26	0.43	0.32	0.28	0.31	0.29	1.00	0.31	0.06	0.30	0.35	0.34
CSDE1	0.40	0.39	0.36	0.32	0.42	0.34	0.31	0.41	0.34	0.31	1.00	0.63	0.34	0.39	0.37
YBX1	0.62	0.51	0.06	0.17	0.66	0.12	0.02	0.08	0.14	0.06	0.63	1.00	0.51	0.58	0.42
IGF2BP1	0.39	0.30	0.38	0.42	0.49	0.35	0.38	0.34	0.38	0.30	0.34	0.51	1.00	0.49	0.40
SYNCRIP	0.45	0.37	0.29	0.24	0.43	0.36	0.32	0.32	0.30	0.35	0.39	0.58	0.49	1.00	0.39
HSPA1B	0.28	0.31	0.35	0.33	0.35	0.27	0.30	0.39	0.33	0.34	0.37	0.42	0.40	0.39	1.00

Discovery on miR-let-7d (Case Study 3)

To illustrate how DeepMiRBP identifies novel candidates for miRNA sorting in exosomes, we focused on let-7d, an exosomal miRNA found in colon cancer cells [34] and pancreatic cancer cells [49]. Our goal was to determine which RBPs miRNA hsa-let-7d would bind.

Using let-7d as input to the target domain, we obtained the similarity scores indicating the affinity of various RBPs to this miRNA, as shown in Table 9.

Although the model evaluated 22 RBP candidates, the table presents the top 10 candidates. Notably, IGF2BP2 and FXR2 emerged as top candidates, with similarity scores of 0.65 and 0.61, respectively. Both proteins have been identified as exosomal proteins in colorectal cancer cells [6], aligning with their potential roles in exosome-mediated RNA transport.

This result corroborates the experimental data from VEPsort, where FXR2 is known to bind to let-7d precursors. The identification of FXR2 among the top candidates for let-7d, coupled with their presence in exosomes, underscores FXR2's role in RNA binding and exosomal RNA sorting. It highlights DeepMiRMP's utility in providing reliable insights into miRNA-RBP interactions, which is crucial for understanding gene regulation mechanisms and developing targeted therapeutic strategies.

Discussion

Introducing the DeepmiRBP model into RNA research has provided a profound leap forward in our understanding of miRNA-protein interactions. The results presented in this study underscore the effectiveness and reliability of the DeepMiRBP model in predicting (mi)RNA-RBP interactions, even for miRNAs not included in the training domain. The model's ability to generalize to novel miRNA-RBP interactions is particularly significant, as it demonstrates the potential for discovering new miRNA-binding proteins and elucidating the mechanisms underlying miRNA sorting.

The promising performance of the DeepmiRBP model in predicting binding sites for AGO, YBX1, and FXR2 proteins is noteworthy. These proteins play a pivotal role in the post-transcriptional regulation of gene expression [8]. Identifying let-7d interactions with FXR2 and other RBPs emphasizes the model's utility in identifying miRNA-protein interactions relevant to cancer biology and indicates its potential in pinpointing critical

Table 9 Top 10 RBPs with highest scores for let-7d

RBP	Similarity score
NIP7	0.66
IGF2BP2	0.65
FXR2	0.61
IGF2BP3	0.49
XRN2	0.47
SLTM	0.36
SERBP1	0.34
BCCIP	0.27
SRSF9	0.17
FAM120A	0.15

regulatory nodes within complex disease networks. The high accuracy achieved in these predictions suggests that the model could serve as a valuable tool for identifying novel RNA-centric therapeutic targets.

DeepmiRBP has demonstrated effectiveness in elucidating the complex interplay between miRNAs and proteins and underscores the power of deep learning, which has been increasingly recognized for its ability to decipher complex biological systems. However, the challenges inherent in applying cosine similarity and transfer learning to such a complex biological problem should not be underestimated. The specificity required for accurate RNA-protein interaction prediction necessitates a tailored approach to model training and validation. It is important to note that the DeepmiRBP model does not predict which miRNA binds to an RBP; rather, it generates a candidate list based on cosine similarity scores, where higher scores indicate a greater likelihood of binding. The model creates candidate lists using cosine similarity with LSTM, CNN, and transfer learning. Another challenge faced was the volume of data and the preparation required, which was demanding and complex. [33].

The potential of transfer learning, as demonstrated by the DeepmiRBP model, is immense. It offers a promising avenue for enhancing the predictive performance of computational models in scenarios characterized by limited data availability or high biological complexity. Nonetheless, the application of this technique must be carefully calibrated to capture the nuances of each protein-miRNA interaction and avoid overfitting to particular datasets or scenarios [36].

Integrating multi-omic data, including genomics, transcriptomics, and proteomics, is expected to refine the predictive accuracy of models like DeepmiRBP further. By incorporating a broader spectrum of biological data, researchers can hope to capture the full complexity of RNA-mediated cell signaling and communication and their regulatory roles in human diseases. This holistic approach will likely pave the way for the next generation of precision medicine, where targeted therapies are developed based on a comprehensive understanding of the molecular underpinnings.

Overall, the DeepMiRBP model provides a robust and scalable framework for predicting miRNA-RBP interactions, offering valuable insights into the molecular mechanisms of miRNA sorting. The model's adaptability to new datasets and its potential for identifying novel miRNA-binding proteins make it a powerful tool for advancing small RNA research. Future work will focus on expanding the model's capabilities, incorporating additional datasets, and validating predictions experimentally to refine further our understanding of miRNA-protein interactions and their implications in disease contexts.

Conclusion

This investigation into miRNA-protein interactions has illuminated the intricate nature of RNA sorting and showcased the efficacy of the DeepmiRBP model in elucidating understudied biological processes. By integrating LSTM, CNN, transfer learning, cosine similarity, and encoding techniques, DeepmiRBP has demonstrated exceptional precision in identifying miRNA-protein binding sites, underscoring the transformative potential of computational approaches in RNA research.

The model's adeptness, particularly in pinpointing binding sites for proteins such as AGO, YBX1, and FXR2, holds profound implications for understanding regulatory

mechanisms in cancer and other diseases where miRNA functionality is pivotal. Integrating PSSM and contact map data via CNN has enriched the model's interpretive depth, advancing our grasp of miRNA-mediated cell signaling. The model's ability to capture the nuanced expression of miRNAs across biological conditions presents challenges and opportunities.

While DeepmiRBP focuses on the predictive analysis of miRNA binding proteins, the methodologies, and insights gleaned offer a scalable template for future studies across various RNA applications and human diseases like cancers. The adaptable nature of this model, informed by its success in the current study, primes it for exploratory applications in RNA-centric targeted therapies.

In conclusion, the DeepmiRBP model significantly advances our ability to predict miRNA-protein binding sites and understand the regulatory mechanisms in cancer. The insights gained from this research contribute to a richer understanding of the complex interplay between miRNAs and proteins and highlight the potential for deep learning to revolutionize bioinformatics. Future research should continue to build upon these findings, leveraging the power of computational models to unravel the complexities of cancer biology and guide the development of new therapeutic strategies.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05985-2.

Supplementary file 1.

Acknowledgements
The authors would like to thank the members of the SBBI Lab (https://sbbi.unl.edu/) for their helpful and constructive discussion and advice. We also appreciate the UNL Holland Computing Center for providing the computational facility.

Author Contributions
SA implemented the system, conducted the analysis, and wrote the manuscript; JC conceived this study and wrote the manuscript.

Funding
Not applicable.

Availability of data and materials
DeepMiRBP is available as open-source from GitHub at: https://github.com/sbbi-unl/DeepmiRBP.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 6 August 2024 Accepted: 12 November 2024
Published online: 18 December 2024

References

1. Alipanahi B, Delong A, Weirauch MT, et al. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8.
2. Altschul SF, Madden TL, Schäffer AA, et al. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.

3. Azizian S. A data-driven discovery system for studying extracellular micro rna sorting and rna-protein interactions. PhD thesis, The University of Nebraska-Lincoln. 2024.
4. Bartel DP. Micrornas: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281–97.
5. Bartel DP. Micrornas: target recognition and regulatory functions. Cell. 2009;136(2):215–33.
6. Beckler MD, Higginbotham JN, Franklin JL, et al. Proteomic analysis of exosomes from mutant kras colon cancer cells identifies intercellular transfer of mutant kras. Mol Cell Proteom. 2013;12(2):343–55.
7. Borgonetti V, Coppi E, Galeotti N. Targeting the rna-binding protein hur as potential thera-peutic approach for neurological disorders: Focus on amyo-trophic lateral sclerosis (als), spinal muscle atrophy (sma) and multiple sclerosis. Int J Mol Sci. 2021;22(19):10394.
8. Burroughs AM, Ando Y, de Hoon MJ, et al. A comprehensive survey of 3′ animal mirna modification events and a possible role for 3′ adenylation in modulating mirna targeting effectiveness. Genome Res. 2010;20(10):1398–410.
9. Chen Hc, Wang J, Coffey RJ, et al. Evpsort: An atlas of small ncrna profiling and sorting in extracellular vesicles and particles. J Mol Biol. 2024;168571
10. Das A, Sinha T, Shyamal S, et al. Emerging role of circular rna-protein interactions. Non-coding RNA. 2021;7(3):48.
11. Davis CA, Hitz BC, Sloan CA, et al. The encyclopedia of dna elements (encode): data portal update. Nucleic Acids Res. 2018;46(D1):D794–801.
12. Dueck A, Ziegler C, Eichner A, et al. microrna associated with the different human argonaute proteins. Nucleic Acids Res. 2012;40(19):9850–62.
13. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microrna: are the answers in sight? Nat Rev Genet. 2008;9(2):102–14.
14. Fong MY, Zhou W, Liu L, et al. Breast-cancer-secreted mir-122 reprograms glucose metabolism in premetastatic niche to promote metastasis. Nat Cell Biol. 2015;17(2):183–94.
15. Gao T, Shu J, Cui J. A systematic approach to rna-associated motif discovery. BMC Genom. 2018;19:1–17.
16. Garcia-Martin R, Wang G, Brandão BB, et al. Microrna sequence codes for small extracellular vesicle release and cellular retention. Nature. 2022;601(7893):446–51.
17. Hassanzadeh HR, Wang MD. Deeperbind: Enhancing prediction of sequence specificities of dna binding proteins. In: 2016 IEEE International conference on bioinformatics and biomedicine (BIBM), IEEE, 2016;178–183.
18. Helwak A, Kudla G, Dudnakova T, et al. Mapping the human mirna interactome by clash reveals frequent noncanonical binding. Cell. 2013;153(3):654–65.
19. Henikoff JG, Henikoff S. Using substitution probabilities to improve position-specific scoring matrices. Bioinformatics. 1996;12(2):135–43.
20. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci. 1992;89(22):10915–9.
21. Hu S, Ma R, Wang H. An improved deep learning method for predicting dna-binding proteins based on contextual features in amino acid sequences. PLoS ONE. 2019;14(11): e0225317.
22. Iorio MV, Croce CM. Microrna dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. EMBO Mol Med. 2012;4(3):143–59.
23. Janas T, Janas MM, Sapoń K, et al. Mechanisms of rna loading into exosomes. FEBS Lett. 2015;589(13):1391–8.
24. Johnson CD, Esquela-Kerscher A, Stefani G, et al. The let-7 microrna represses cell proliferation pathways in human cells. Can Res. 2007;67(16):7713–22.
25. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Bioinformatics. 1992;8(3):275–82.
26. Kajitani N, Schwartz S. Role of viral ribonucleoproteins in human papillomavirus type 16 gene expression. Viruses. 2020;12(10):1110.
27. Koppers-Lalic D, Hackenberg M, Bijnsdorp IV, et al. Nontemplated nucleotide additions distinguish the small rna composition in cells from exosomes. Cell Rep. 2014;8(6):1649–58.
28. Li Y, Hu J, Zhang C, et al. Respre: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. Bioinformatics. 2019;35(22):4647–55.
29. Liu XM, Ma L, Schekman R. Selective sorting of microrna into exosomes by phase-separated ybx1 condensates. Elife. 2021;10: e71982.
30. Livi CM, Blanzieri E. Protein-specific prediction of mrna binding using rna sequences, binding motifs and predicted secondary structures. BMC Bioinform. 2014;15:1–11.
31. Luo Y, Hitz BC, Gabdank I, et al. New developments on the encyclopedia of dna elements (encode) data portal. Nucleic Acids Res. 2020;48(D1):D882–9.
32. Maticzka D, Lange SJ, Costa F, et al. Graphprot: modeling binding preferences of rna-binding proteins. Genome Biol. 2014;15:1–18.
33. Muppirala UK, Honavar VG, Dobbs D. Predicting rna-protein interactions using only sequence information. BMC Bioinform. 2011;12:1–11.
34. Noh GT, Kwon J, Kim J, et al. Verification of the role of exosomal microrna in colorectal tumorigenesis using human colorectal cancer cell lines. PLoS ONE. 2020;15(11): e0242057.
35. Nwaeburu CC, Bauer N, Zhao Z, et al. Up-regulation of microrna let-7c by quercetin inhibits pancreatic cancer progression by activation of numbl. Oncotarget. 2016;7(36):58367.
36. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2009;22(10):1345–59.
37. Pan X, Shen HB. Rna-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. BMC Bioinform. 2017;18:1–14.
38. Pan X, Shen HB. Learning distributed representations of rna sequences and its application for predicting rna-protein binding sites with a convolutional neural network. Neurocomputing. 2018;305:51–8.
39. Pan X, Shen HB. Predicting rna-protein binding sites and motifs through combining local and global deep convolutional neural networks. Bioinformatics. 2018;34(20):3427–36.
40. Pan X, Rijnbeek P, Yan J, et al. Prediction of rna-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genom. 2018;19(1):1–11.

41. Pan X, Fang Y, Li X, et al. Rbpsuite: Rna-protein binding sites prediction suite based on deep learning. BMC Genom. 2020;21:1–8.
42. Quinlan AR, Hall IM. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
43. Ruff KM, Pappu RV. Alphafold and implications for intrinsically disordered proteins. J Mol Biol. 2021;433(20): 167208.
44. Sherry ST, Ward MH, Kholodov M, et al. dbsnp: the ncbi database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.
45. Shu J, Chiang K, Zempleni J, et al. Computational characterization of exogenous microrna that can be transferred into human circulation. PLoS ONE. 2015;10(11): e0140587.
46. Singhal A, et al. Modern information retrieval: A brief overview. IEEE Data Eng Bull. 2001;24(4):35–43.
47. Uniprot: the universal protein knowledgebase. Nucleic Acids Research. 2017;45: D158–D169
48. Villarroya-Beltri C, Gutiérrez-Vázquez C, Sánchez-Cabo F, et al. Sumoylated hnrnpa2b1 controls the sorting of mirnas into exosomes through binding to specific motifs. Nat Commun. 2013;4(1):2980.
49. Wang F, Zhou C, Zhu Y, et al. The microrna let-7 and its exosomal form: Epigenetic regulators of gynecological cancers. Cell Biol Toxicol. 2024;40(1):42.
50. Xia P, Zhang L, Li F. Learning similarity with cosine similarity ensemble. Inf Sci. 2015;307:39–52.
51. Yu H, Wang J, Sheng Q, et al. berbp: binding estimation for human rna-binding proteins. Nucleic Acids Res. 2019;47(5):e26–e26.
52. Zhou W, Fong MY, Min Y, et al. Cancer-secreted mir-105 destroys vascular endothelial barriers to promote metastasis. Cancer Cell. 2014;25(4):501–15.

Publisher's Note