8-2024

A Data-Driven Discovery System for Studying Extracellular Microrna Sorting and RNA-Protein Interactions

Sasan Azizian
University of Nebraska-Lincoln

A DATA-DRIVEN DISCOVERY SYSTEM FOR STUDYING EXTRACELLULAR MICRO RNA SORTING AND RNA-PROTEIN INTERACTIONS

by

Sasan Azizian

A DISSERTATION

Presented to the Faculty of
The Graduate College at the University of Nebraska
In Partial Fulfillment of Requirements
For the Degree of Doctor of Philosophy

Major: Computer Science

Under the Supervision of Professor Juan Cui

Lincoln, Nebraska

August, 2024

A DATA-DRIVEN DISCOVERY SYSTEM FOR STUDYING EXTRACELLULAR MICRO RNA SORTING AND RNA-PROTEIN INTERACTIONS

Sasan Azizian, Ph.D.

University of Nebraska, 2024

Advisor: Juan Cui

Interactions between microRNAs (miRNAs) and RNA-binding proteins (RBPs) are pivotal in miRNA-mediated sorting, yet the molecular mechanisms underlying these interactions remain largely understudied. Few miRNA-binding proteins have been verified, typically requiring extensive laboratory work. This study introduces DeepMiRBP, a novel hybrid deep learning model designed to predict microRNA-binding proteins. The model integrates Bidirectional Long Short-Term Memory (Bi-LSTM) networks with attention mechanisms, transfer learning, and cosine similarity to offer a robust computational approach for inferring miRNA-protein interactions.

DeepMiRBP is implemented through two architectures. The first is a Y-shaped model that employs Bi-LSTM networks and transfer learning to identify similarities between miRNA and RNA sequences. This method captures dependencies and context within RNA sequences, while attention mechanisms highlight the most relevant features. Transfer learning applies knowledge from a large dataset of RNA-binding proteins to predict miRNA-protein interactions.

The second architecture enhances the first by adding cosine similarity and transfer learning. It has two main components: the first uses Bi-LSTM networks and transfer learning to process RNA sequences binding to RBPs, embedding them into a 128-dimensional space, and assessing similarities with miRNA sequences. The second uses CNNs and protein structural information, such as PSSM and contact maps, to encode proteins into unique vectors and evaluate their similarities, resulting in a comprehensive similarity matrix.

DeepMiRBP accurately predicts miRNA interactions with recently discovered exosomal transporter proteins like AGO, YBX1, and FXR2. This highlights its potential to identify novel transporter proteins crucial for exosome-mediated small RNA sorting and other miRNA-protein interactions. DeepMiRBP's methodologies provide a scalable template for research, from mechanistic discovery to cell-to-cell communication in disease development, with the potential for RNA-centric therapeutic interventions and personalized medicine.

DeepMiRBP has shown high accuracy in predicting RNA-binding interactions, making it valuable for studying miRNA sorting and broader RNA-protein interactions. Its innovative use of Bi-LSTM networks, CNNs, transfer learning, and cosine similarity marks a significant advancement in computational biology, offering a powerful framework to understand complex cellular networks.

Acknowledgements

First and foremost, I express my deepest gratitude to my advisor, Dr. Juan Cui. Her unwavering guidance, insightful feedback, and continuous support have been instrumental in shaping this dissertation. Her expertise and encouragement have influenced the direction of my research and significantly improved the quality of my work.

I am also immensely grateful to my committee members, Dr. Hamid Sharif, Dr. Qiuming Yao, Dr. Mohammad Rashedul Hasan, and Dr. Hamid Bagheri, for their valuable feedback and constructive suggestions, which have significantly improved the quality of this work. I want to extend a special thank you to Dr. Sharif and Dr. Yao for their meticulous review of my dissertation.

I thank the Department of Computer Science and Engineering at the University of Nebraska-Lincoln for providing the necessary resources and creating a conducive environment for my research. The support and assistance of the faculty and staff have been unwavering and instrumental in my journey.

A special and heartfelt note of gratitude goes to my wonderful family. To my incredible wife, Elham, thank you for your endless support and for taking such good care of our family, especially our 3.5-year-old son, Noyan. Your patience and under-

standing allowed me to focus on my work and studies, and to my dear Noyan, thank you for sometimes letting Daddy study in peace. Your laughter and curiosity kept me motivated, and your occasional interruptions were always a delightful reminder of what truly matters.

I also want to express my deepest gratitude to my mother. Mom, your unwavering support and constant encouragement have always given me the strength and energy to keep going. I am eternally grateful for everything you have done for me.

Even though my dad has passed away, my father's teachings and values continue to guide me every day. Thank you, Dad, for being my role model and for all the love and wisdom you imparted. You were, and still are, my inspiration.

Thank you all for your contributions and support. This dissertation would not have been possible without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

RNA-binding proteins (RBPs) and microRNAs (miRNAs) play pivotal roles in gene regulation, intricately weaving the narrative of cellular function. RBPs, capable of binding to single and double-stranded RNA molecules, are integral to various cellular activities, particularly RNA processing. This includes splicing, where non-functional sequences (introns) are excised, and functional sequences (exons) are fused [11, 19, 47]. A foundational understanding of RNA-protein interactions sets the stage for the subsequent chapters of this dissertation.

miRNAs, small non-coding RNA molecules, have garnered significant attention due to their profound impact on gene expression regulation. Approximately 22 nucleotides in length, miRNAs modulate the expression of multiple genes simultaneously by binding to the 3' untranslated region (UTR) of target messenger RNAs (mRNAs), leading to their degradation or translational repression [7, 24]. The regulatory potential of miRNAs is crucial for maintaining cellular homeostasis, with disruptions in their function implicated in diseases ranging from cancer to cardiovascular and neurological disorders [41].

The process of miRNA sorting, where these molecules are selectively incor-

porated into multivesicular bodies (MVBs) and subsequently released as exosomes, exemplifies the precision and complexity of cellular communication [82]. This selective packaging and dispatching of miRNAs to recipient cells enable the coordination of biological processes across different tissues and organs. Recent studies have hinted at specific sequence motifs and RBPs that might play a role in this selective sorting [49, 42], though the exact mechanisms remain enigmatic and warrant further exploration.

Exosomes, depicted in Figure 1.1, are small membrane-bound vesicles that serve as cellular couriers, transferring a variety of molecular cargos, including miRNAs, to recipient cells [78]. miRNAs packaged into exosomes can be delivered to distant tissues, influencing the gene expression of recipient cells [55]. This mechanism of intercellular communication has profound implications, as illustrated by the release of miR-105 in breast cancer exosomes, promoting tumor growth in distant tissues like the lungs and brain [89, 25]. Such revelations underscore the therapeutic potential of understanding miRNA sorting mechanisms and their broader implications in disease progression.

Previous studies identified the molecular features of miRNA responsible for its secretion, emphasizing the importance of specific motifs [26, 82]. For instance, motif discovery tools like MDS2 have identified motifs associated with exomiRs. Experiments have shown that mutations in these motifs significantly decrease miRNA levels in exosomes compared to cells, indicating the crucial role of these motifs in exomiR sorting [26, 72]. Furthermore, miRNA-binding proteins responsible for sorting miRNAs with specific motifs have been identified, such as hnRNPA2B1 in human primary

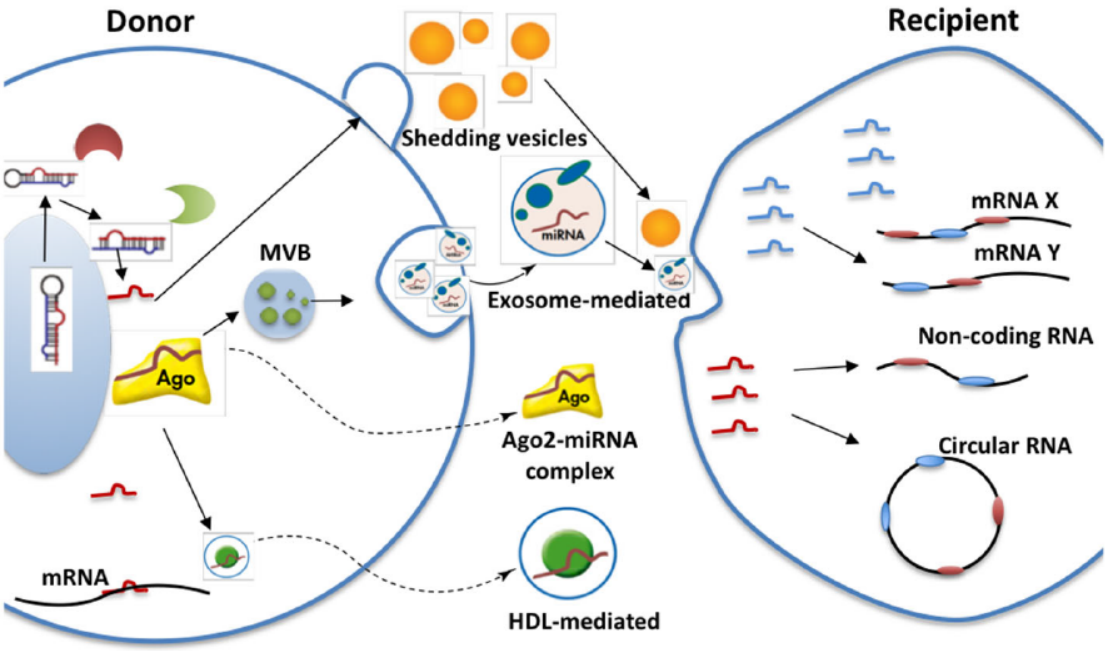

Figure 1.1: Schematic diagram of miRNA transfer between cells and competitive miRNA binding in the recipient cell as an illustration. MVB: multivesicular bodies; HDL: high-density lipoprotein [18].

T cells [82] and Sdpr and Fus in adipocyte cells.

A high-performance discovery tool capable of systematically studying miRNA-protein interactions in an automated and high-yield manner is needed further to elucidate the protein-mediated sorting process beyond motif analysis. Although only a handful of miRNA-binding proteins are known, the availability of massive amounts of (mi)RNA-protein interactome data provides an opportunity to harness machine learning (ML) and deep learning (DL) approaches for genome-scale predictions of DNA/RNA binding sites and protein structures. Inspired by recent advances, this research explores ML-based solutions to identify molecular determinants that are key to exomiR sorting.

1.1 Proposed Multimodal Deep Learning Framework for RNA and miRNA-Protein Interaction Prediction

We propose a multimodal deep learning-based framework focused on predicting miRNA-protein interactions. DL is particularly suitable for this task because it facilitates efficient representation learning directly from each data modality during model training when properly formulated into a supervised learning task. Successful applications of DL in protein structure, function, localization, and DNA/RNA interaction studies, such as AlphaFold [46], DeepSec [70], DeepSig [68], DeepLoc [3], and DeepBind [2], have demonstrated remarkable performances. However, none of these tools were designed to discover miRNA-protein binding.

This research aims to develop a new predictive model for miRNA-protein interaction by addressing two major goals: (1) predicting miRNA sorting and (2) predicting RNA binding proteins. To achieve these goals, we created the DeepMiRBP model. Initially, we utilized transfer learning and Y architecture, which are detailed in Chapter 2. Subsequently, we enhanced the model by incorporating cosine similarity and transfer learning with two main components: The RNA Binding Protein Candidate trained on a vast dataset of RNA sequences that bind to RBPs. It extracts features from RNA sequences and uses transfer learning to predict miRNA-protein binding. Cosine similarity is employed to find similarities between miRNA and RBP sequences. The second Component is the protein Candidate for Binding miRNA. After obtaining the RBP candidates that have a chance to bind to miRNA, this Component uses residue contact maps and position-specific scoring matrices (PSSM) for each RBP. It

identifies other proteins that could be similar to the RBP candidates. We introduce *DeepMiRBP*, a new multimodal deep neural network for miRNA-Protein Binding prediction, which integrates sequence and structural information from both RNA and RBPs. The *DeepMiRBP* model leverages transfer learning and cosine similarity for effective prediction. Together, these components offer precise predictions of miRNA-protein interactions. Subsequent sections will delve into the model's details and its implications in molecular biology.

1.2 Contribution

This dissertation makes several key contributions to the field of bioinformatics and molecular biology:

- **Development of the DeepmiRPB Model**: This dissertation presents a novel hybrid deep learning model, DeepmiRPB, designed to predict miRNA-binding proteins by modeling molecular interactions. The model's innovative integration of Bidirectional Long Short-Term Memory (Bi-LSTM) networks, transfer learning, attention mechanisms, and cosine similarity underscores its robustness and precision in computational biology.

- **Integration of Multi-Omic Data**: DeepmiRPB leverages multi-omic data, capturing the nuanced dependencies and structural information within miRNA and protein sequences. This comprehensive approach enhances the accuracy of predictions, offering a more detailed understanding of miRNA-protein interactions.

- **Validation of Model Performance**: The model's efficacy was rigorously validated through extensive testing across diverse datasets, demonstrating high accuracy, precision, recall, and F1 scores. This validation process highlights the model's robustness in predicting interactions with key proteins such as AGO, YBX1, and FXR2, which are crucial for understanding disease regulatory mechanisms.

- **Application to Disease Mechanisms**: The DeepmiRPB model's capability to predict miRNA interactions with proteins involved in diseases, including cancer, signifies its potential for identifying novel therapeutic targets. This aspect of the research underscores the model's relevance to practical applications in understanding and treating disease mechanisms.

- **Scalability and Adaptability**: The methodologies and insights from the DeepmiRPB model provide a scalable template for future research. The model's adaptability highlights its potential for developing novel RNA-centric therapeutic interventions and personalized medicine, making significant strides in bioinformatics.

- **Novel Computational Framework**: This research introduces a unique computational framework that synergizes sequence and structural data through a multi-modular deep neural network. The innovative use of transfer learning and cosine similarity within this framework paves the way for future advancements in miRNA-protein interaction prediction.

1.3 Organization

This dissertation is organized into several chapters, each addressing different aspects of the research:

- *Chapter 2: Predictive Modeling of RNA Binding Proteins and Small RNA-protein Binding Prediction Using Y Architecture and Transfer-Learning* - This chapter delves into predictive modeling techniques for RNA-binding proteins and small RNA-protein binding. It details the data collection, preprocessing, and implementation of the Y architecture and transfer learning in the DeepMiRBP model.

- *Chapter 3: Enhanced miRNA-Protein Binding Predictions Using Transfer Learning and Cosine Similarity* - This chapter builds on the previous one by enhancing the model's predictive capabilities using transfer learning and cosine similarity. It discusses the core techniques, model architecture, and hyperparameter optimization and presents the improved model's results.

- *Chapter 4: Conclusion* - The final chapter summarizes the research's key findings, discusses its implications for bioinformatics and molecular biology, and suggests future research directions.

Chapter 2

Predictive Modeling of RNA Binding Proteins and Small RNA-protein Binding Prediction Using Y Architecture and Transfer-Learning

2.1 Introduction

In molecular biology, the intricate dance between microRNAs (miRNAs) and RNA-binding proteins (RBPs) is pivotal for cellular communication and gene regulation. Understanding the sorting and interactions of miRNAs is a frontier yet to be fully explored. This chapter heralds a revolutionary hybrid deep learning framework, harnessing the strengths of Y-architecture networks to delve into the secrets of small RNA sorting determinants.

The cornerstone of our framework is a bidirectional Long Short-Term Memory (LSTM) network embellished with an attention mechanism that meticulously processes RBP sequences. This approach captures the temporal dependencies and nuances within the sequential RNA data, highlighting the influential subsequences crucial for interaction specificity. Complementing this is our innovative application of autoencoders, which distill high-dimensional PSSM and protein structure contact map data into a more tractable form without losing the essence of the structural

information that guides miRNA targeting.

Employing a Y-shaped architecture, our model synergizes sequence analysis with structural insights, tackling the inputs in parallel streams before converging to decode the complexities of miRNA-protein binding. This dual-pathway strategy ensures the retention of the biological data's sequential and structural fidelity and enhances interpretability and predictive performance. Transfer learning techniques further amplify our model's prowess, allowing it to transcend its training data and excel in identifying miRNA-protein interactions across diverse cellular contexts. Our results exhibit an exceptional capability to predict miRNA binding partners, surpassing existing tools that predominantly rely on sequence analysis.

By integrating advanced LSTM networks with attention, autoencoder@Sath structural profiles, and a multifaceted Y-architecture, our framework is a monumental leap in computational biology. It highlights sequence and structural motifs that could be the key to unlocking miRNA sorting mechanisms and proposes plausible candidates for miRNA transporter proteins. Our model's technical sophistication and adaptability make it an invaluable asset in understanding the enigmatic processes that govern RNA biology. This research extends beyond theoretical implications, providing a practical toolset for the scientific community to investigate miRNA sorting mechanisms further. It promises to revolutionize our approach to diseases where miRNA dysregulation is a factor, offering a new lens through which we can view potential therapeutic targets and interventions. By pushing the boundaries of what is possible in RNA analytics, our work paves the way for transformative discoveries that could reshape the landscape of medical science and biotechnology.

In the intricate tapestry of molecular biology, RNA-binding proteins (RBPs) and microRNAs (miRNAs) emerge as pivotal threads that intricately weave the narrative of gene regulation. RBPs, with their ability to bind directly to single and double-chained RNA molecules, play a central role in various cellular activities linked to RNA's function [11]. One of their most critical roles is assisting in RNA processing, particularly in splicing, which involves the excision of non-functional sequences (introns) and the fusion of functional ones (exons) [19, 47]. This foundational understanding of RNA-protein interactions sets the stage for the subsequent chapters of this dissertation.

Diving deeper into molecular biology, miRNAs, small non-coding RNA molecules, have garnered significant attention due to their profound impact on gene expression regulation [7]. These molecules, approximately 22 nucleotides in length, modulate the expression of multiple genes simultaneously by binding to the 3' untranslated region (UTR) of target messenger RNAs (mRNAs), leading to their degradation or translational repression [24]. Their broad regulatory potential makes miRNAs integral to maintaining cellular homeostasis. Disruptions in their function have been implicated in many diseases, ranging from cancer to cardiovascular and neurological disorders [41].

The process of miRNA sorting, where these molecules are selectively incorporated into multivesicular bodies (MVBs) and subsequently released as exosomes, is a testament to the precision and complexity of cellular communication [82]. This selective packaging and dispatching of miRNAs to recipient cells allow for coordinating biological processes across different tissues and organs. Recent studies have hinted

at specific sequence motifs and RNA-binding proteins that might play a role in this selective sorting [49, 42]. Yet, the exact mechanisms remain enigmatic, beckoning further exploration.

Exosomes, depicted in Figure 1.1, are small membrane-bound vesicles that serve as cellular couriers, transferring many molecular cargos, including miRNAs, to recipient cells [78]. Among these cargos, miRNAs are particularly interesting due to their role in gene regulation. miRNAs packaged into exosomes can be delivered to distant tissues, influencing recipient cells' gene expression [55]. This mechanism of intercellular communication allows for the coordination of biological processes across different tissues and organs.

This mechanism of intercellular communication has profound implications, as illustrated by the release of miR-105 in breast cancer exosomes, which promotes tumor growth in distant tissues like the lungs and brain [89, 25]. Such revelations underscore the therapeutic potential of understanding miRNA sorting mechanisms and their broader implications in disease progression. However, the journey to unravel the mysteries of miRNA sorting is far from complete. Numerous questions, such as identifying miRNAs interacting with loading proteins, their subsequent sorting into exosomes, and the downstream gene regulatory effects in recipient cells, linger. Addressing these questions promises to illuminate the intricate mechanisms of intercellular communication and gene regulation. Furthermore, it could pave the way for innovative therapeutic strategies targeting diseases associated with aberrant miRNA function.

Previous research has demonstrated that molecular features of miRNA are

responsible for its secretion [53, 27]. In prior studies, a motif discovery tool named MDS2 was developed specifically for short sequence analysis, identifying motifs associated with exomiRs. Experiments showed that mutations in these motifs significantly decreased the miRNA levels in exosomes versus cells, indicating that exomiR sorting depends on the presence of these motifs. Additionally, miRNA-binding proteins responsible for sorting miRNAs with specific motifs have been identified, such as hnRNPA2B1 in human primary T cells and Sdpr and Fus in adipocyte cells. Molecular docking analysis has suggested a possible linkage between miRNA motifs and protein binding sites.

To further elucidate the protein-mediated sorting process beyond motif analysis, we need a high-performance discovery tool capable of systematically studying miRNA-protein interactions in an automated and high-yield manner. While only a handful of miRNA-binding proteins are known, massive amounts of (mi)RNA-protein interactome data have become available. Inspired by recent advances in machine learning (ML) and its successes in genome-scale prediction of DNA/RNA binding sites and protein structures, we aim to explore ML-based solutions that can harness the power of omics data on sequence, structure, and interaction to identify molecular determinants key to exomiR sorting.

In this project, we propose a multimodal deep learning (DL)-based framework focused on predicting miRNA-protein interactions. DL facilitates efficient representation learning from each data modality during model training when the problem is properly formulated into a supervised learning task. DL has been successfully applied to study protein structure, function, localization, and DNA/RNA interaction,

including seminal works like AlphaFold, DeepSec, and DeepBind. However, none of these tools were designed to discover miRNA-protein binding.

Building a new predictive model for miRNA-protein interaction requires addressing two major issues: representative data and multimodal architecture. Most miRNA interactome data were collected from AGO-related CLIP-seq and CLASH platforms, but we need to integrate other proteins into the model to make it representative and general. Additionally, current DL models handle mostly sequence data, while this research needs a new framework to incorporate sequence and structural information from both interaction partners. To overcome these limitations, we propose a transfer learning solution leveraging abundant data, knowledge, and experience in RNA-protein binding (source domain) to improve discovery-making in miRNA (target domain) through a multi-modular deep neural network framework. Current RNA binding data derived from sequencing, like RNA Bind-N-seq data on hundreds of RBPs, provides implicit binding regions and does not pinpoint exact contact sites. To make it applicable for miRNA study, we need to preprocess the data and extract information about conserved contact sites to train models in the source model. The trained models will be transferred and refined in the target domain by including miRNA-protein interaction data. The proposed model will consider sequence and structure information from associated (mi)RNA and protein for new interaction prediction.

We introduce *DeepmiRPB*, a new multimodal deep neural network for miRNA-protein binding prediction, integrating sequence and structural information from both RNA and RBP. *DeepmiRPB* integrates two main components:

- **DeepRPB Feature Extractor**: Trained on a vast dataset of RBPs, it extracts features like the secondary structure context of RNA and the residue contact of RBPs.

- **miRNA-specific Model**: Using transfer learning, it processes features from the DeepRPB extractor to predict miRNA-protein binding.

Together, these components offer precise predictions of miRNA-protein interactions. Subsequent sections will delve into the model's details and its implications in molecular biology.

In this research, we seek to delve deeper into miRNA interactions, leveraging advanced computational tools and omics data. Building on foundational work with RBPs and miRNAs, we aim to further our understanding of the complex regulatory networks that govern cellular function, focusing on miRNA sorting and its implications in health and disease.

2.2 Related Work

RNA-binding proteins (RBPs) are paramount in orchestrating many cellular regulatory functions, from gene splicing to localization, and have profound implications for patient care [23]. The quest to pinpoint RBP binding sites is crucial, given that RBPs discern both sequence and structure motifs in RNA molecules. The latter, structure motifs, pertain to the unique three-dimensional conformation of RNA, diverging from sequence motifs that focus on nucleotide order. Notably, certain proteins linked with amyotrophic lateral sclerosis are known to bind to RNA targets within

specific structures like hairpins and loops. Moreover, RBPs are adept at discerning loop and stem regions in RNA precursors, thereby modulating RNA expression levels [75].

Traditional methodologies, such as RIP-seq and CLIP-seq, employed for RBP discovery are resource-intensive in terms of time and cost [34]. This has catalyzed the emergence of many efficient and economical tools for discerning sequence and structure motifs. Some of these tools, like BEAM [88], focus on structure motifs, while others, such as CapR [82] and the approach by Li et al. [26], amalgamate sequence motifs with secondary structure considerations. These tools have significantly advanced our understanding of RNA-protein interactions by enabling high-throughput and precise identification of binding sites.

The advent of deep learning has revolutionized the prediction landscape of RNA-protein interactions. A slew of models, including DeepBind [2], deepRKE [21], DeeperBind [35], and models by Zeng et al. [20], have harnessed the prowess of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. These models, ranging from iDeep [63], iDeepV [32], iDeepE [63], to iDeeps [63], and DanQ [65], have showcased the potential of deep learning in capturing intricate sequence and structure motifs, thereby enhancing the precision of RNA-protein interaction predictions.

Despite these advancements, the realm of miRNA-binding proteins remains largely uncharted. While extensive research has been conducted on RNA-binding proteins, the specific domain of miRNA-binding proteins has been relatively untouched. This lacuna in the research landscape underscores our 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.

- **miRNA Sorting:** There is currently no comprehensive framework specifically dedicated to miRNA sorting. However, research has been published in Nature by Garcia-Martin et al. [27], providing substantial insights into miRNA sorting mechanisms. This study demonstrates that miRNAs possess sorting sequences that determine their secretion into small extracellular vesicles (sEVs) or retention within cells. Different cell types, such as white and brown adipocytes, endothelium, liver, and muscle, make preferential use of specific sorting sequences, defining the sEV miRNA profile of that cell type. The study identifies two RNA-binding proteins, Alyref and Fus, as key players in the export of miRNAs carrying one of the strongest EXOmotifs, CGGGAG. This miRNA code links circulating exosomal miRNAs to their tissues of origin and provides an approach for improved targeting in RNA-mediated therapies.

- **RNA-binding Proteins:** Traditional methodologies, such as RIP-seq and CLIP-seq, employed for RBP discovery are resource-intensive in terms of time and cost [34]. This has catalyzed the emergence of many efficient and economical tools for discerning sequence and structure motifs. Some of these tools, like BEAM [88], focus on structure motifs, while others, such as CapR [82] and the approach by Li et al. [26], amalgamate sequence motifs with secondary structure considerations. These tools have significantly advanced our under-

standing of RNA-protein interactions by enabling high-throughput and precise identification of binding sites.

For instance, DeepBind utilizes CNNs to learn sequence motifs from raw RNA sequences, offering significant improvements in prediction accuracy over traditional methods [2]. Similarly, DeeperBind enhances this approach by incorporating more complex network architectures to capture subtle features in the data [79]. These advancements have led to more robust and generalizable models capable of predicting RNA-protein interactions across diverse biological contexts.

- **miRNA Targeting:** In recent years, studies have begun exploring the unique challenges and opportunities miRNA-binding proteins present. For example, the work by Ha and Kim [33] highlights the critical regulatory roles of miRNAs in gene expression, emphasizing the need for precise identification of miRNA-binding sites. Similarly, research by Dueck et al. [22] has provided insights into the differential association of miRNAs with various Argonaute proteins, suggesting complex regulatory mechanisms yet to be fully understood.

  Additionally, the development of computational tools specifically tailored for miRNA-binding site prediction is gaining traction. For instance, the tool developed by Quevillon Huberdeau et al. [66] focuses on the phosphorylation states of Argonaute proteins, which are pivotal in miRNA-mediated gene silencing. Such tools are instrumental in advancing our understanding of miRNA-protein interactions and their implications in cellular processes.

  Integrating deep learning techniques in miRNA research is also beginning to

show promise. Studies leveraging LSTM networks to capture the temporal dynamics of miRNA interactions are emerging as a new frontier in computational biology. These approaches enhance the predictive accuracy of miRNA-binding sites and provide deeper insights into the regulatory roles of miRNAs in various biological systems.

Overall, the existing body of research underscores the critical role of RBPs and miRNA-binding proteins in cellular regulation. However, the specific mechanisms and interactions within the miRNA-protein domain remain an area ripe for exploration. Our work aims to address this gap by leveraging advanced computational techniques, including deep learning, to unravel the complexities of miRNA-protein interactions. Through this, we hope to contribute to a more comprehensive understanding of the molecular underpinnings of gene regulation and pave the way for novel therapeutic interventions.

2.3 Data Collection and Analysis

Data collection and analysis are fundamental components of bioinformatics research and are the foundation for constructing predictive models. This chapter describes the meticulous data collection, organization, and preprocessing process. The primary datasets utilized in our research include RNA sequences that bind to RNA-binding proteins (RBPs) and protein sequences sourced from the Universal Protein Resource (UniProt) and the NCBI Protein Database.

2.3.1 RNA Sequences that Bind to RNA-Binding Proteins

Sequencing-Based Platforms for Protein-RNA Binding Site Identification

Understanding our data source is crucial before delving into data preprocessing and embeddings. The RNA Binding Protein (RBP) binding linear RNAs from RBPSuite [64] are derived from advanced sequencing platforms designed to capture protein-RNA interactions. These platforms, often termed 'CLIP-Seq' (Crosslinking Immunoprecipitation Sequencing), employ a combination of immunoprecipitation and high-throughput sequencing to identify the binding sites of RNA-binding proteins.

The process typically involves several steps:

- Crosslinking RNA-binding proteins to RNA molecules in living cells.
- Fragmenting the RNA and isolating the RNA-protein complexes.
- Sequencing the RNA fragments to determine the binding sites.

The advantage of such sequencing-based platforms is the ability to capture in vivo protein-RNA interactions, providing a more accurate representation of cellular processes. Moreover, the high-throughput nature of these platforms allows for identifying thousands of binding sites in a single experiment, making it a valuable resource for bioinformatics studies like ours.

Data Collection and Organization

In bioinformatics, the quality and comprehensiveness of data are pivotal in ensuring the accuracy and reliability of predictive models. Our research, aimed at

understanding RNA-protein interactions, necessitated the collection of data from reputable sources that offer both depth and breadth in their datasets.

One of the primary repositories from which we sourced our data is RBPSuite [64]. This suite, derived from advanced sequencing platforms, is specifically tailored to capture the nuances of protein-RNA interactions. However, our data collection efforts extended beyond RBPSuite. We further enriched our dataset by utilizing data from the ENCODE Project [17].

The ENCODE (Encyclopedia of DNA Elements) Project is a groundbreaking initiative to identify all functional elements in the human genome. Initiated in 2003, the project's primary goal is to enhance our understanding of how genetic information is regulated and utilized in different cell types and tissues. By mapping these elements, ENCODE provides insights into their roles in human health and disease. The ENCODE Project has generated a vast amount of data, including information on DNA regions that produce RNA, regions that bind proteins, and chemically modified regions. This data is freely available to the scientific community and is a valuable resource for researchers worldwide.

A specialized section of the ENCODE Project, the ENCORE Matrix, integrates data from the ENCODE and Roadmap Epigenomics projects. "ENCORE" stands for "Encyclopedia of DNA Elements at Roadmap Epigenomics." This integrated resource offers a comprehensive view of functional genomic elements across various cell types and tissues. When visiting the ENCORE Matrix page on the ENCODE website 2.1, one is presented with a matrix of experiments related to the ENCORE tag. This matrix is organized by biosample (cell or tissue type) and assay

type (the method used to detect a specific genomic feature). Each cell in the matrix represents a specific experiment, and clicking on a cell provides detailed information about that experiment.

Figure 2.1: The ENCORE project aims to study protein-RNA interactions by creating a map of RNA binding proteins (RBPs) encoded in the human genome and identifying the RNA elements that the RBPs bind to.

In our quest to understand RNA-protein interactions, we also leveraged various platforms that offer insights into these interactions. One such platform is TF ChIP-seq. ChIP-seq (Chromatin Immunoprecipitation sequencing) is a method that enables researchers to understand protein interactions with DNA. When referring to "TF ChIP-seq," we delve into studying the binding sites of transcription factors (TFs) on DNA. Other platforms that enriched our research include eCLIP, a method tailored to study RNA-protein interactions, and CRISPR RNA-seq, a fusion of the CRISPR/Cas9 genome editing system with RNA sequencing, offering insights into the transcriptomic effects of specific genetic modifications.

With the data sources identified, we delve into the specifics of the data utilized in our research:

- **RNA Sequences:** The benchmark dataset for RNA Binding Protein (RBP) binding linear RNAs from RBPSuite served as our primary source for RNA sequences [64]. This comprehensive dataset includes 154 RBPs, and their corresponding binding sites (RNAs) are derived from ENCODE. The use of this dataset ensures that we have a robust and diverse set of RNA sequences for our model training and validation.

- **Protein Sequences:** For protein sequences, we utilized two main resources: the Universal Protein Resource, which shows the website in this Figure 2.2 and the National Library of Medicine, NCBI, Protein Database [1, 71, 76] that shows the website in this Figure 2.3. UniProt is a freely accessible database of protein sequence and functional information, which includes the manually annotated UniProt Knowledgebase and the automatically annotated UniProt TrEMBL database. The NCBI Protein Database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq, and TPA and records from SwissProt, PIR, PRF, and PDB. These databases offer a wealth of protein sequence information, which is crucial for our research.

In addition to protein sequences, we also considered protein structures. We used the method described in the paper "ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks" [52]. This

method uses a deep residual convolutional neural network to predict residue-level protein contacts using the inverse covariance matrix of multiple sequence alignments. This approach provides valuable information about protein structures, which is essential for our research.

miRNA Data Collection and Organization

The data utilized in this study, specifically related to miRNA, was obtained from the research paper titled "Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding" [37]. This paper was a valuable resource for our model, providing a detailed annotation of all identified miRNA-mRNA interactions in Data S1.

The collection of Data S1 involved using experimental data derived from high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) and photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP). In addition to these experimental methods, computational predictions were also employed. These predictions were based on algorithms that consider sequence complementarity and the thermodynamic stability of the miRNA-mRNA duplex when predicting miRNA targets.

Upon obtaining the data, it was found to include 26 columns. Specifically, Column 5, titled *miRNA_seq*, provided the miRNA sequence (start – end). As part of the data preprocessing, we focused on creating a clean dataset with uniform miRNA sequences. To achieve this, we adjusted the length of all sequences to 101 characters. If a sequence was shorter than 101 characters, we appended additional characters

from the start to the end until it reached 101. This formed the positive part of our dataset.

Subsequently, we generated a negative sequence dataset. The total number of sequences for miRNA, combining both the positive and negative datasets, amounted to 31,636. This comprehensive dataset served as the foundation for our subsequent analysis and modeling.

The research by Helwak et al. is particularly notable for its comprehensive approach to identifying miRNA interactions. They employed a technique known as CLASH (crosslinking, ligation, and sequencing of hybrids) to directly ligate and sequence miRNA-target RNA duplexes associated with the human AGO1 protein. This technique allowed for the high-confidence identification of more than 18,000 miRNA-mRNA interactions, providing a robust dataset for our study.

Key points from Helwak et al. that were particularly relevant to our data collection include:

- The binding of most miRNAs involves the 5' seed region, but around 60% of seed interactions are noncanonical, containing bulged or mismatched nucleotides.
- Approximately 18% of miRNA-mRNA interactions involve the miRNA 3' end, with little evidence for 5' contacts.
- Specific base-pairing patterns, including canonical and noncanonical sites, characterize the miRNA-target interactions.

These insights into miRNA binding mechanisms were critical for understanding our dataset's complexity and diversity of miRNA interactions. The detailed anno-

tation and high-confidence interactions provided by Helwak et al. ensured that our dataset was comprehensive and reliable, forming a solid foundation for developing and validating our predictive models. focused primarily on two types of data: RNA sequences and protein sequences.

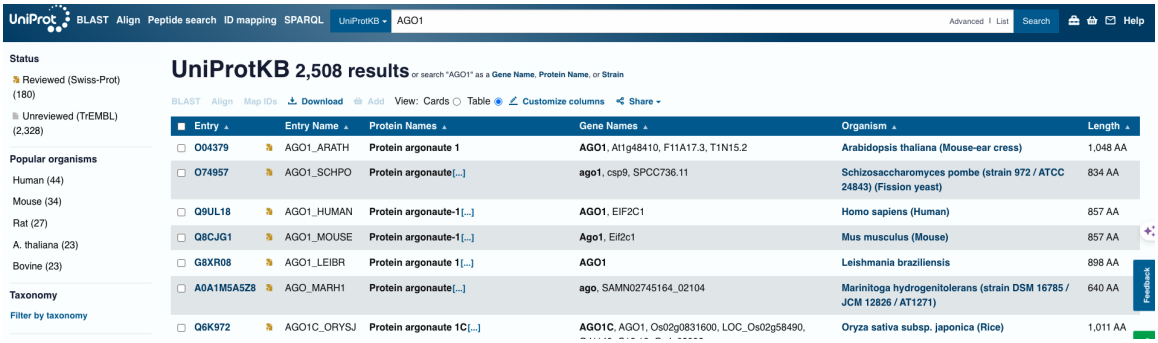

Figure 2.2: UniProt is the world's leading high-quality, comprehensive, and freely accessible resource of protein sequence and functional information.

2.3.2 Data Preprocessing and Refinement

Data preprocessing is a pivotal step in ensuring the efficacy of our models. We

RNA Sequence Preprocessing

Our primary dataset for RNA sequences was sourced from RBPSuite, a benchmark for RNA Binding Protein (RBP) binding linear RNAs [64]. The preprocessing of this dataset encompassed several stages:

- Initially, the peak files of each RBP were merged to consolidate the data.
- Regions that overlapped with the reference gene were selected using the intersect Bed function of bedtools [67].

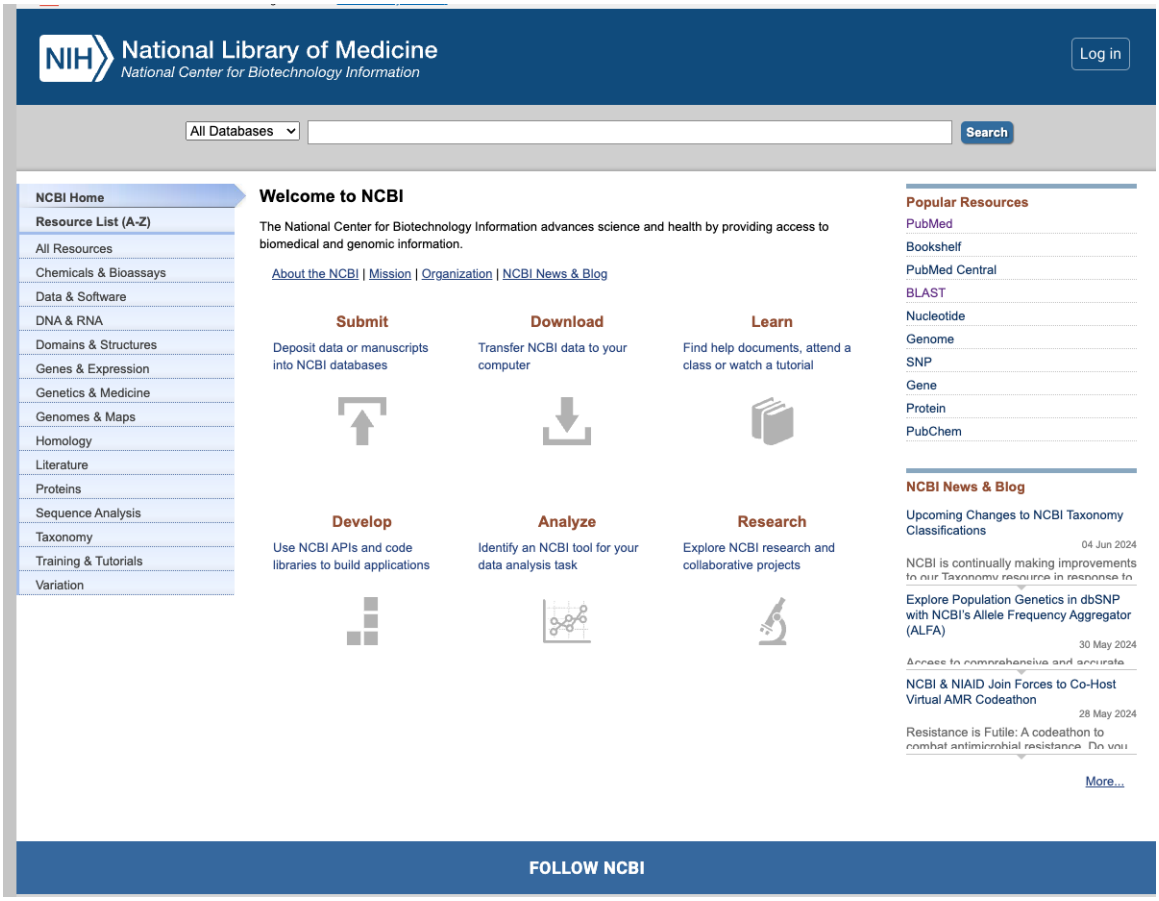

Figure 2.3: The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

- Gene-overlapped regions with less than 107 base pairs (bp) were extended with downstream and upstream regions of the same length, resulting in positive regions of RBPs.

- Negative RBP binding regions were produced by implementing shuffleBed of bedtools[67], with all regions being 101bp.

- The fasta files of positive and negative regions were retrieved using fastaFromBed of bedtools.

- For each RBP, only 60,000 positive and 60,000 negative sites were retained if the extracted positive and negative samples exceeded this number; otherwise, all extracted samples were used.

This data collection and preprocessing approach ensures that the data used for training the deep learning model is highly relevant. Before we proceed to the next part, we want to talk briefly about bedtools [67].

In bioinformatics, the efficient manipulation and analysis of genomic data are crucial for advancing our understanding of molecular biology. One of the essential tools we employed for preprocessing RNA sequences is `bedtools` [67]. `bedtools` is a versatile suite of utilities designed to facilitate a wide range of genomic analyses. It enables researchers to easily perform complex operations on genomic data, offering functionalities for intersection, merging, and complementing genomic features, among others.

Developed by Quinlan and Hall, `bedtools` is widely recognized for its robustness and flexibility in handling various genomic datasets. The suite comprises numerous command-line tools that operate on files formatted in the Browser Extensible Data (BED) format, a standard for representing genomic intervals. These tools are indispensable for identifying overlapping genomic regions, extracting sequence data, and computing coverage statistics.

Our research specifically utilized `bedtools` to preprocess RNA sequences. One of the key features we leveraged is the `getfasta` tool, which extracts DNA or RNA sequences from a reference genome based on coordinates provided in a BED file. This

functionality is particularly useful for converting genomic intervals into corresponding nucleotide sequences, a critical step in our data preprocessing pipeline.

The process of extracting RNA sequences using `getfasta` involves several steps:

- First, the BED file containing RNA-binding protein (RBP) binding site coordinates is prepared.
- Next, `bedtools getfasta` is employed to retrieve the nucleotide sequences corresponding to these coordinates from a reference genome.

To illustrate the usage of `bedtools getfasta`, consider the following command:

```
bedtools getfasta -fi reference_genome.fa -bed rna_binding_sites.bed
    -fo output_sequences.fa
```

In this command:

- `-fi reference`$_genome.fa$: $Specifies the input reference genome file in FASTA format.$

Additionally, `bedtools getfasta` offers options to handle more complex scenarios, such as extracting sequences from specific blocks within BED intervals. For instance, using the `-split` option enables the extraction of sequences from each block separately, which is essential when dealing with spliced RNA molecules.

An example command incorporating the `-split` option is as follows:

- `bedtools getfasta -fi reference_genome.fa -bed rna_binding_sites.bed -fo output_sequences.fa -split`

This command ensures that each block within the BED intervals is treated independently, providing a precise extraction of spliced RNA sequences.

The use of `bedtools` was instrumental in our research for several reasons:

- Its ability to efficiently process large genomic datasets ensured we could handle the extensive RNA sequence data derived from RBPSuite and other sources.
- The flexibility and precision offered by `bedtools getfasta` allowed us to tailor our data extraction processes to meet the specific requirements of our study, thereby enhancing the accuracy and reliability of our RNA sequence data.
- The seamless integration of `bedtools` into our preprocessing pipeline facilitated the preparation of high-quality datasets, which are crucial for training robust deep learning models.

In conclusion, `bedtools` has proven to be an invaluable tool in preprocessing RNA sequences for our research. Its comprehensive suite of utilities and user-friendly command-line interface have enabled us to perform intricate genomic analyses with ease and precision, thereby contributing significantly to the success of our study.

Protein Sequence Preprocessing

For protein sequences, we obtained data from the Universal Protein Resource (UniProt) and the NCBI Protein Database [1, 71]. We used three main features for the embeddings: the Position-Specific Scoring Matrix (PSSM) and the Protein Structure Contact Map.

The detailed steps for preprocessing protein sequences include:

- Extracting protein sequences from UniProt and NCBI databases.
- Generating PSSM profiles for each protein sequence.
- Creating contact maps using the ResPRE method, which involves predicting residue-level protein contacts based on the inverse covariance matrix of multiple sequence alignments.

2.3.3 Data Integration and Analysis

Our research is bolstered by a robust dataset amalgamated from reputable sources. This comprehensive dataset ensures the efficacy and reliability of our deep-learning models in predicting RNA-protein interactions. By combining these resources, we have compiled a comprehensive and diverse dataset for our research, allowing us to train our models effectively and ensure they can handle a wide range of RNA and protein sequences.

Integrating Data from Multiple Sources

To achieve a comprehensive dataset, it is crucial to integrate data from multiple reputable sources. RBPSuite and the ENCODE Project were the primary sources for our RNA sequences, while UniProt and NCBI provided the bulk of our protein sequence data. This integration process involved aligning and merging data from these diverse sources to create a unified dataset.

Ensuring Data Quality and Consistency

Maintaining high data quality and consistency is essential for the success of any bioinformatics research. Our data preprocessing steps, which involved rigorous filtering, selection, and validation processes, ensured that our dataset was of the highest quality. By meticulously curating and refining our data, we laid a solid foundation for our subsequent analyses and model training.

The meticulous data collection and organization process, advanced platforms, and reputable databases have laid a robust foundation for our research into RNA-protein interactions. The datasets we've chosen, backed by the credibility of sources like ENCODE, RBPSuite, UniProt, and NCBI, ensure that our predictive models are accurate and reliable. We have compiled a comprehensive and diverse dataset for our research by integrating these resources and employing rigorous data preprocessing techniques. This dataset will allow us to train our deep-learning models effectively and ensure they can handle a wide range of RNA and protein sequences.

2.3.4 Input Representation Using Embeddings

Our study employed various embedding techniques to represent RNA and protein sequences, ensuring that our models could effectively interpret and learn from the data. Two primary techniques utilized in this study are Position-Specific Scoring Matrix (PSSM) and Protein Structure Contact Maps (PSCMs).

Position-Specific Scoring Matrix (PSSM)

Position-Specific Scoring Matrix (PSSM) is an essential computational tool widely used in bioinformatics, specifically for analyzing DNA and protein sequences. PSSM represents the conservation of specific residues (amino acids or nucleotides) at particular positions in a sequence alignment. It's a matrix that helps in identifying conserved patterns within biological sequences, making it a valuable asset in tasks like sequence alignment, motif discovery, and homology detection [4, 45, 39, 38].

To generate a PSSM for a given protein sequence, we first perform a BLAST (Basic Local Alignment Search Tool) search against a protein database (such as UniProt) to find similar sequences. These sequences are then aligned to create a Multiple Sequence Alignment (MSA). From the MSA, we calculate the frequency of each amino acid at each position in the alignment, which gives us a Position-Specific Frequency Matrix (PSFM). The PSFM is then converted into a PSSM by taking the log-odds of the observed frequencies relative to the expected frequencies of amino acids [4]. The formula for the log-odds score is:

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

To refine our understanding of protein sequences within our dataset, which comprises 500 unique proteins, we generated PSSM files for each protein. Given their variable lengths, we encapsulated them into fixed-dimension vectors. The imperative of dimensionality reduction led us to employ the Elbow Curve method specifically for

our PSSM data.

The Elbow Curve, depicted in Figure 2.14, illustrates the explained variance ratio against the number of principal components. This strategic approach enabled us to ascertain the optimal number of principal components to retain, ensuring the preservation of significant features while minimizing informational loss. By identifying the "elbow point," where the explained variance ratio begins to plateau, we determined the appropriate number of components to maintain the integrity of the data.

Advancing beyond traditional methods, we integrated an Autoencoder into our analytical framework for a more sophisticated dimensionality reduction of the PSSM representations. The Autoencoder's encoder component compresses the input PSSM into a dense, lower-dimensional representation, capturing the protein's sequence conservation essence. Subsequently, the decoder part attempts to recreate the original PSSM from this compressed representation. This process not only aids in reducing the dimensionality of our data but also uncovers latent patterns that might be obscured in the raw PSSMs.

Consider the example of ABRE, a nucleotide motif discovered that its consensus sequence is ACGTG G/T C [36]. An alignment of 47 ABRE sequences was constructed, and a matrix of counts at each motif position was created.

This count matrix has four rows corresponding to the four nucleotides that occur in DNA: A, C, G, and T. It also has nine columns corresponding to the nine positions in the motif. To convert these counts into a PSSM, the frequency matrix is first transformed into a matrix of probabilities. The probability of observing a

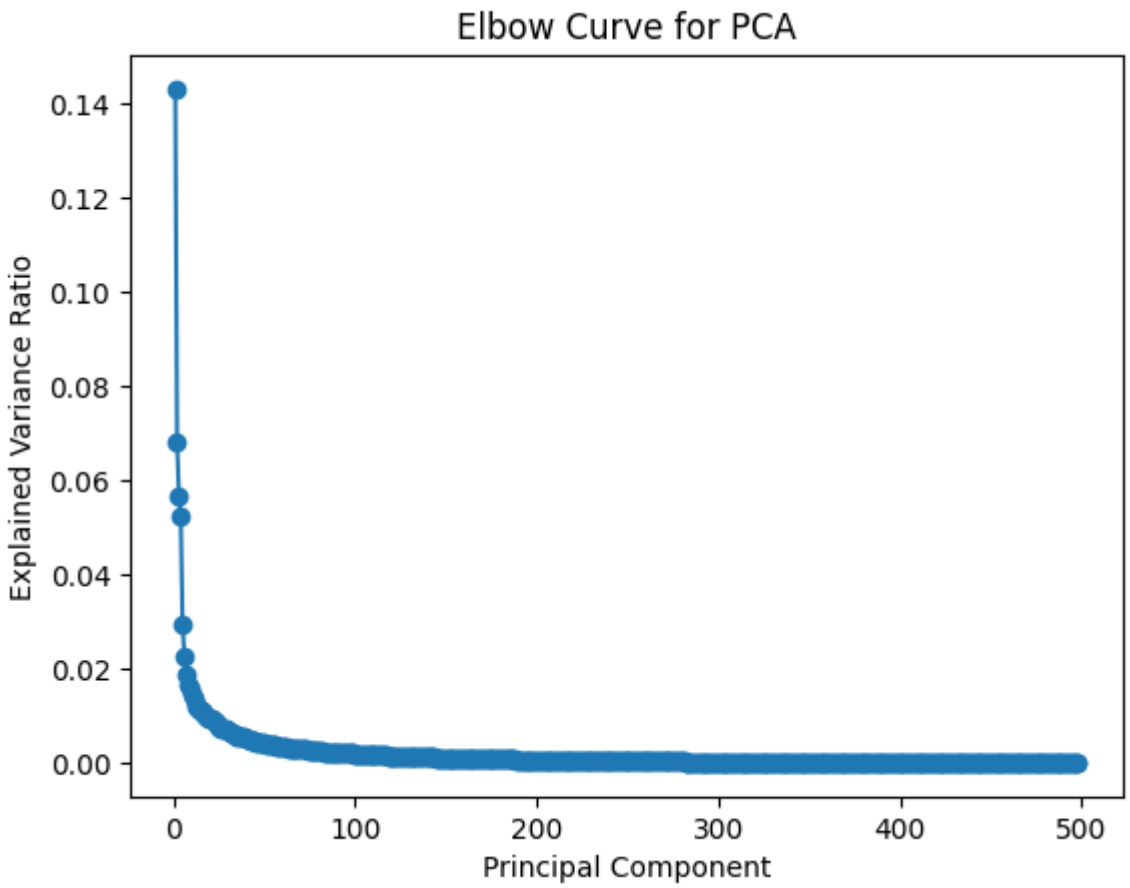

Figure 2.4: Elbow Curve for PCA, illustrating the explained variance ratio against the number of principal components. The curve helps determine the optimal number of components to retain by identifying where the variance explained by additional components diminishes.

particular nucleotide at a given position is the observed count for that nucleotide at that position, divided by the number of sequences in the alignment. For instance, the observed probabilities for the ABRE motif are:

Log-odds scores are then calculated using the odds ratio of observed to expected frequencies. For nucleotides, assuming equal frequencies, the expected frequency is always 0.25. The log-odds probability of each nucleotide at each position in half-bit units is then calculated.

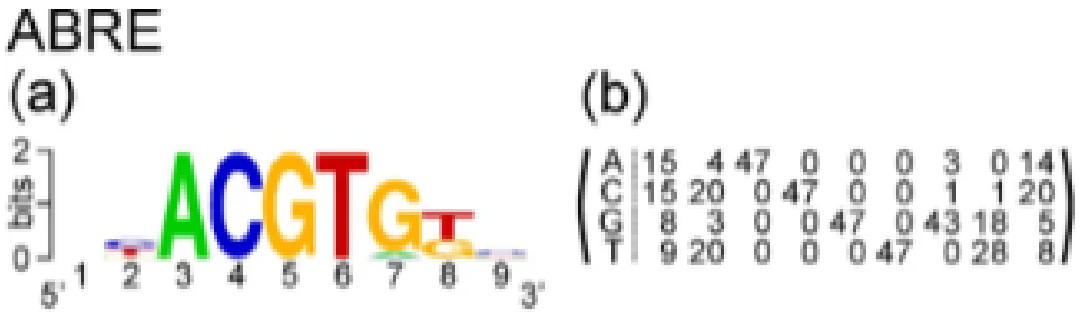

Figure 2.5: Sequence logos and matrices for ABRE used in this study. (a) ABRE sequence logo; (b) ABRE frequency matrix.

	1	2	3	4	5	6	7	8	9
A	0.32	0.09	1.00	0.00	0.00	0.00	0.06	0.00	0.30
C	0.32	0.43	0.00	1.00	0.00	0.00	0.02	0.02	0.43
G	0.17	0.06	0.00	0.00	1.00	0.00	0.91	0.38	0.11
T	0.19	0.43	0.00	0.00	0.00	1.00	0.00	0.60	0.17

Table 2.1: Converted frequency matrix into a matrix of probabilities for the ABRE motif.

Protein Structure Contact Maps (PSCMs)

Proteins are complex molecules that play critical roles in the body, with their function often determined by their three-dimensional shape or structure. The structure of a protein encompasses its sequence of amino acids, known as its primary structure. This sequence gives rise to local sub-structures, such as alpha-helices and beta-sheets, which are components of the protein's secondary structure. As these sub-structures interact, they form the protein's overall three-dimensional shape [52], termed its tertiary structure. Some proteins consist of more than one amino acid chain, and the arrangement of these chains relative to each other is described as the quaternary structure.

Visualizing the protein structure is often achieved using techniques like X-

Figure 2.6: 3D Protein Structure of protein AGO1$_{HUMAN}$.[77]

ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. However, determining the exact structure of a protein using these methods can be challenging. The complexity arises from the number of atoms in proteins and the dynamic nature of these molecules, which can adopt multiple conformations. Additionally, there are inherent limitations to the experimental methods themselves. For instance, X-ray crystallography requires the protein to form crystals, which isn't always feasible, while NMR is typically limited to analyzing relatively small proteins.

Given these challenges in directly determining protein structures, researchers have sought alternative ways to understand them. One such approach is the Protein Structure Contact Map (PSCM). A PSCM is a graphical representation of a protein's tertiary structure. Instead of portraying the protein in three dimensions, a contact map is a two-dimensional matrix. In this matrix, a mark at a specific position indicates that two amino acids are in proximity to the protein's three-dimensional structure.

Contact maps simplify the intricate three-dimensional structure into a more manageable two-dimensional representation. This simplification facilitates the analysis and comparison of structures. Moreover, while the exact coordinates of a protein's atoms might vary under different conditions, the contact map remains relatively consistent. This consistency, combined with advances in machine learning and bioinformatics, has shown that predicting contact maps can be an effective step in forecasting the full three-dimensional structure of a protein.

To understand the spatial relationships between amino acid residues, we employed Protein Structure Contact Maps (PSCMs). These maps translate the three-

dimensional structure of proteins into a two-dimensional matrix, providing insights into the protein's tertiary structure and facilitating the analysis of its functional and structural properties [60, 5].

To enhance the accuracy of our contact maps, we utilized the ResPRE algorithm, a deep learning-based method for predicting residue-level contacts [52]. ResPRE employs a deep residual convolutional neural network, efficiently predicting contacts even for challenging protein sequences. This approach offers a more flexible tool for understanding protein structure, bypassing direct calculation.

The contact map is determined based on the spatial proximity of amino acid residues in the protein's 3D structure. If the distance between two residues falls below a threshold, typically within 6-8 Ångstroms, they are considered in contact, marking the corresponding matrix cell as 1; otherwise, it is marked as 0, resulting in a symmetric matrix representation 2.7.

ResPRE Algorithm

ResPRE: High-Accuracy Protein Contact Prediction by Coupling Precision Matrix with Deep Residual Neural Networks is a state-of-the-art method designed to enhance the accuracy of protein contact predictions [52]. The ResPRE algorithm consists of three primary steps: Multiple Sequence Alignment (MSA) generation, precision-matrix-based feature collection, and deep residual neural network training.

1. MSA Generation: An informative MSA is critical for evolutionary coupling analyses and subsequent contact-map prediction. In ResPRE, the MSA is generated by HHblits [52] with a coverage threshold for the query sequence of 40 and a pair-

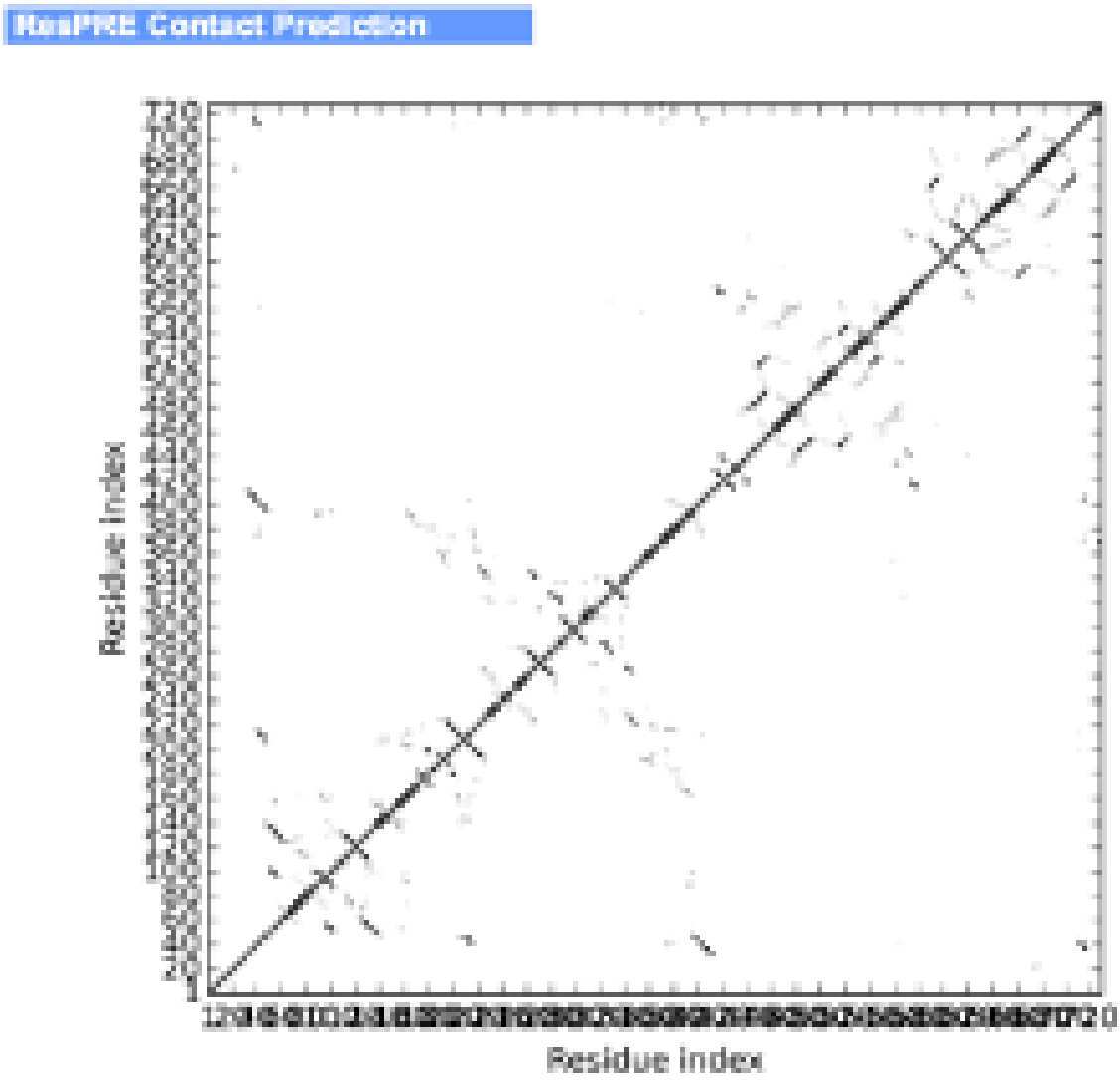

Figure 2.7: ResPRE Contact Prediction for AGO1_HUMAN. The above plot displays the contact-map with a cutoff ¿=0.5 of confidence score (ranging from 0 to 1) [52].

wise sequence identity cutoff of 0.99 against $\text{Uniprot}_2016_04 by three iterations. The E-value threshold is configured to 1 to obtain more diverse alignments.$

2. Precision-Matrix Based Feature Collection: The precision matrix, derived through the maximum likelihood approach, helps rule out transitional noises of contact maps compared with the previously used covariance matrix. This step is crucial for capturing the conditional independent relationships among residues, which are then used as training features for the deep residual neural networks.

3. Deep Residual Neural Network Architecture: The deep residual neural network (ResNet) architecture enables the training of very deep neural networks by adding feedforward neural networks with an identity map of input, allowing gradients to flow smoothly from deeper to shallower layers. This architecture enhances the neural network's learning ability, making it possible to predict contacts even for challenging protein sequences.

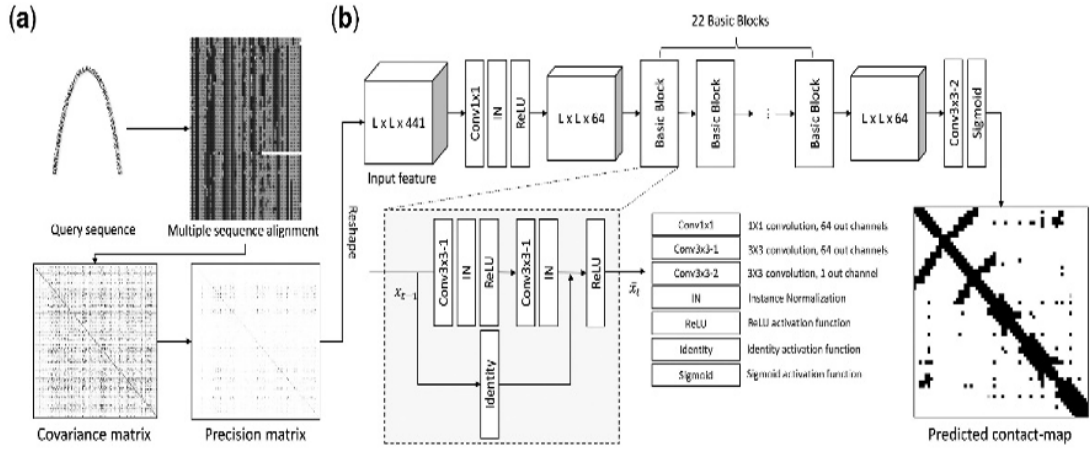

Figure 2.8: Flowchart of ResPRE. (a) Process of precision-matrix-based feature collection. (b) Block diagram of deep residual neural network architecture [52]

The ResPRE algorithm has significantly improved prediction accuracy com-

pared to traditional methods. By leveraging the precision matrix and deep residual neural networks, ResPRE provides a robust framework for understanding protein structures, particularly useful for proteins lacking close homology templates in existing databases.

In summary, 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. Integrating PSSM and PSCMs, along with the advanced ResPRE algorithm, enhances our ability to predict and analyze protein structures with higher accuracy and reliability.

2.4 Materials and Methods

2.4.1 Overview

The field of miRNA-protein binding prediction is burgeoning, yet 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 [8]. The complexity of miRNA-protein interactions, which exhibit considerable variability across different biological contexts, further complicates the prediction process.

The paucity of miRNA-protein binding datasets is attributed to the intricate and labor-intensive nature of the experimental techniques employed for identifying these interactions, such as CLIP and HITS-CLIP. These methods are time-consuming

and costly and necessitate specialized expertise, thereby hindering the generation of extensive miRNA-protein interaction datasets [16]. Additionally, the highly context-dependent nature of miRNA-protein interactions, varying across cell types, developmental stages, and physiological conditions, adds another layer of complexity, resulting in datasets that capture only a fraction of the potential interactions in biological systems [44].

Given these challenges, computational approaches, particularly machine learning, have emerged as promising tools for predicting miRNA-protein interactions. Machine learning algorithms are adept at discerning patterns and structures from existing data, which can be applied to predict new, unseen data. This capability is particularly beneficial for tasks like miRNA-protein interaction prediction, where the data is complex and high-dimensional [13].

This research contributes to the field by designing a novel transfer learning solution to address the issue of limited data availability. Transfer learning allows us to leverage pre-existing knowledge and models from related domains, adapting them to the specific task of miRNA-protein interaction prediction. For instance, a model initially trained to predict protein-protein interactions can be repurposed to predict miRNA-protein interactions. This approach enables the utilization of patterns and structures learned from RNA-protein interaction data, enhancing the performance of miRNA-protein interaction prediction models [12].

Furthermore, we introduce a multi-modular deep neural network architecture tailored to capture various facets of miRNA-protein interactions. This architecture integrates different modules, each designed to extract specific features relevant to

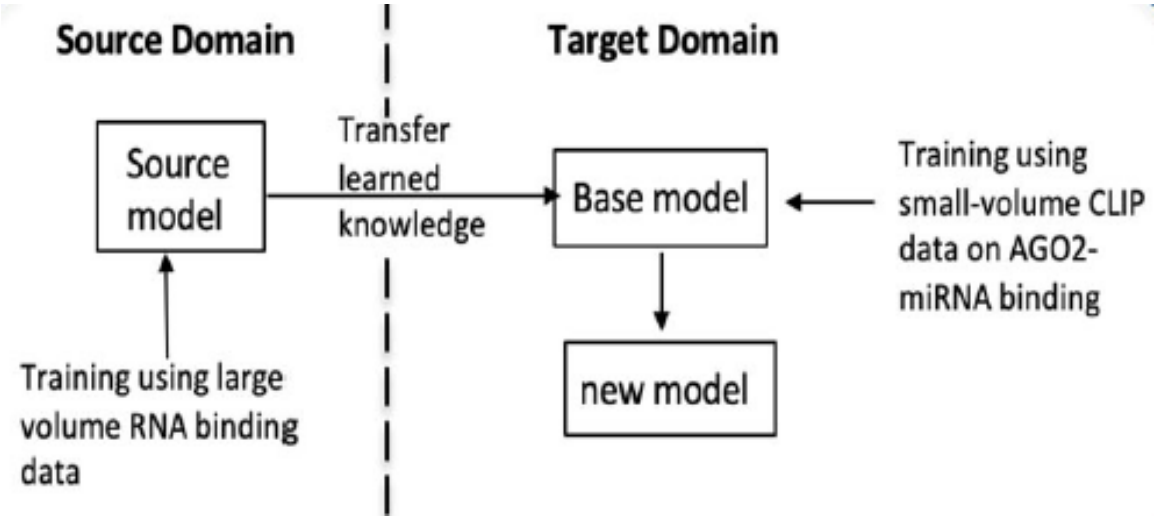

Figure 2.9: Overview of the transfer learning approach.

the interaction, such as sequence and structural features. By amalgamating these modules, we achieve a holistic model capable of delivering accurate predictions, even in scenarios characterized by limited data availability [28].

In summary, this research pioneers the exploration of miRNA-protein interactions, addressing the prevalent challenges through innovative computational approaches. Integrating transfer learning and a multi-modular deep neural network architecture marks a significant stride towards accurately predicting miRNA-protein interactions, paving the way for future research and applications in this domain.

2.4.2 Advanced Machine Learning Techniques in DeepMiRBP for RNA and miRNA-Protein Interaction Prediction

The advent of machine learning (ML) and deep learning (DL) has marked a paradigm shift in bioinformatics, providing powerful tools for the analysis and interpretation of complex biological data. These techniques offer unprecedented capabil-

ities in modeling intricate patterns and making accurate predictions, thereby accelerating discoveries in the life sciences [29]. In our study, we have leveraged several advanced ML and DL techniques in the development of our model, DeepMiRBP. This chapter explores these advanced ML methods, focusing on their applications in bioinformatics, including Bi-Directional Long Short-Term Memory (Bi-LSTM) networks, attention mechanisms, embedding layers, Y architecture, and transfer learning.

Bioinformatics deals with acquiring, storing, analyzing, and disseminating biological data, most notably genetic and genomic data. Traditional computational approaches often fail to handle the sheer volume and complexity of modern biological datasets. With their ability to learn from data and improve over time, machine learning techniques have become indispensable in this field [29]. They enable researchers to uncover hidden patterns, predict biological functions, and model biological systems with greater accuracy and efficiency [29].

One of the primary reasons for the growing reliance on ML and DL in bioinformatics is their ability to handle large-scale data. For instance, processing high-throughput sequencing data, which generates gigabytes of information, requires algorithms capable of scaling with data size [29]. ML models, such as neural networks, can process and learn from vast amounts of data, making them ideal for bioinformatics applications [29].

In recent years, deep learning and neural networks have become increasingly prevalent in bioinformatics due to several compelling reasons. Firstly, the inherent ability of deep learning models to automatically learn and extract features from raw data without the need for extensive manual feature engineering has revolutionized the

field [29]. This capability is particularly beneficial in bioinformatics, where biological data's complexity and high dimensionality pose significant challenges.

Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, are adept at capturing complex patterns and dependencies within data. CNNs, for example, are exceptionally effective in recognizing spatial hierarchies in data, making them suitable for tasks such as protein structure prediction and cellular image analysis [50]. On the other hand, RNNs and their variants, such as Bi-LSTM, excel in sequence modeling, which is crucial for understanding genetic sequences and predicting RNA-protein interactions [40].

Another critical advantage of deep learning in bioinformatics is its scalability. Deep learning models can leverage large datasets to improve their performance continuously. This is essential in bioinformatics, where the volume of data generated by high-throughput technologies exponentially increases. Techniques such as transfer learning further enhance scalability by enabling models trained on large, general datasets to be fine-tuned for specific tasks with smaller datasets [61].

The flexibility and adaptability of neural networks also contribute to their widespread use in bioinformatics. Neural networks can be customized and fine-tuned for various applications, from predicting gene expression levels to identifying potential drug targets. Integrating different data types, such as genomic, proteomic, and phenotypic data, into a single model (e.g., through architectures like the Y network) allows for a more comprehensive analysis of biological systems.

Furthermore, advancements in computational power, particularly with Graph-

ics Processing Units (GPUs) and Tensor Processing Units (TPUs), have made it feasible to train deep neural networks efficiently on large biological datasets. This computational capability has democratized access to deep learning, enabling researchers to implement sophisticated models without prohibitive costs.

Deep learning's robust performance in handling noisy and unstructured data makes it suitable for bioinformatics. Biological data often contains noise due to experimental errors and biological variability. Deep learning models can effectively manage this noise and still produce reliable predictions, enhancing bioinformatics analyses' overall accuracy and robustness [29].

In our model, DeepMiRBP, we utilized advanced machine learning techniques to predict RNA and miRNA-protein interactions. The following sections will detail these techniques, explaining their underlying principles, applications in bioinformatics, and how they contribute to our understanding of biological systems. By harnessing the power of machine learning, we can continue to push the boundaries of bioinformatics, paving the way for discoveries and innovations in the life sciences [29].

Embedding Layers: Embedding layers transform categorical data into continuous vector spaces, capturing semantic relationships between categories. This is particularly useful for representing biological sequences (e.g., DNA, RNA, proteins) in a form suitable for neural networks. Embeddings help identify similarities and differences between sequences based on their learned representations [54].

Bi-Directional Long Short-Term Memory (Bi-LSTM): Bi-LSTM networks are an extension of traditional LSTM networks that process data in both forward and backward directions. This bidirectional processing captures context from

both past and future states, making Bi-LSTMs highly effective for sequence-based tasks such as gene prediction, RNA-protein binding site identification, and sequence alignment [40].

Attention Mechanisms: Attention mechanisms allow models to focus on specific parts of the input sequence, enhancing their performance on tasks requiring long-range dependencies. In bioinformatics, attention mechanisms are used in models for protein structure prediction and genomic sequence analysis, where certain regions of the input data are more informative than others [80].

Y Architecture: The Y architecture is a specific neural network design that merges multiple input streams, processes them independently, and then combines them for final prediction. This architecture is useful in bioinformatics, where different data types (e.g., genomic, proteomic, and phenotypic) must be integrated for comprehensive analysis.

Transfer Learning: Transfer learning involves leveraging pre-trained models on new, related tasks. This approach is particularly useful in bioinformatics, where labeled data is often scarce. Using models pre-trained on related datasets, we can achieve significant performance improvements even with limited data [61]. For example, pre-trained models on large protein databases can be fine-tuned to predict specific protein functions or interactions [61].

Integrating these advanced machine-learning techniques into bioinformatics workflows has led to significant advancements in the field. From predicting protein structures to understanding genomic variations, these methods provide researchers with powerful tools to decode the complexities of biological data [29].

The following sections will detail these techniques, explaining their underlying principles, applications in bioinformatics, and how they contribute to our understanding of biological systems. By harnessing the power of machine learning, we can continue to push the boundaries of bioinformatics, paving the way for discoveries and innovations in the life sciences [29].

Embedding Layers

Embedding layers transform categorical data into continuous vector spaces, capturing semantic relationships between categories. In bioinformatics, embeddings are used to represent sequences such as proteins and RNA in a form suitable for neural networks.

The embedding layer in Long Short-Term Memory (LSTM) deep learning models is crucial, especially in natural language processing (NLP) and sequence analysis. This layer transforms discrete, categorical input data, such as RNA and miRNA sequences, into fixed-size dense vectors, facilitating the learning process in subsequent model layers.

In a continuous vector space, the embedding layer represents each unique item, like a character or word. These vectors, learned during training, capture semantic relationships between items, making it possible to measure distances or similarities between them. Mathematically, if we have a set of items $\{x_1, x_2, \ldots, x_n\}$, the embedding layer maps each item x_i to a vector $\mathbf{v}_i \in \mathbb{R}^d$, with d being the embedding space's dimensionality. The embedding matrix E is represented as:

$$E = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_n \end{pmatrix}$$

where $\mathbf{v}_i$ is the embedding vector for the item x_i. This embedding matrix E is typically initialized randomly and then refined during the training process through backpropagation.

For RNA sequences of 101 characters, the embedding layer transforms each character 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:

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

This process ensures that the embedding vector $\mathbf{v}_{\text{RNA}}$ captures the collective information of the entire RNA sequence, facilitating its input into subsequent layers of the LSTM model.

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

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

This method ensures that the embeddings for miRNA sequences accurately represent the original biological sequences without introducing noise from padding.

The embedding vectors are learned during the neural network's training phase. The objective is to optimize these vectors so that similar items (e.g., similar nucleotide sequences) are close to each other in the embedding space. This is achieved by minimizing a loss function that measures the difference between the predicted outputs and the actual labels.

Commonly used loss functions for training embeddings in bioinformatics include the categorical cross-entropy loss for classification tasks and the mean squared error (MSE) for regression tasks. The gradients of these loss functions concerning the embedding vectors are computed, and the embedding matrix E is updated using gradient descent or its variants.

Embedding layers offer several advantages in bioinformatics:

- **Dimensionality Reduction:** Embeddings reduce the dimensionality of categorical data, making it more manageable for neural networks. This is particularly important for biological sequences, which can be very long and complex.

- **Capturing Semantic Relationships:** Embedding vectors capture the semantic relationships between items, allowing the model to understand similarities and differences between sequences. For instance, similar RNA sequences will have similar embeddings.

- **Efficient Computation:** Embedding layers enable efficient computation by transforming high-dimensional categorical data into lower-dimensional continuous vectors, reducing the computational burden on subsequent layers.
- **Transfer Learning:** Pre-trained embedding layers can be transferred to new models, providing a head start in learning and improving performance, especially when labeled data is scarce.

In practice, embedding layers are implemented using deep learning frameworks such as TensorFlow and PyTorch. Below is a sample implementation of an embedding layer in a neural network using TensorFlow:

```
import tensorflow as tf
from tensorflow.keras.layers import Embedding, LSTM, Dense


# Define the input sequence length and the size of the embedding space
input_length = 101
embedding_dim = 128
vocab_size = 10000  # Example vocabulary size


# Define the model
model = tf.keras.Sequential([
    Embedding(input_dim=vocab_size,
              output_dim=embedding_dim,
              input_length=input_length),
```

```
    LSTM(64, return_sequences=True),
    LSTM(64),
    Dense(1, activation='sigmoid')
])

# Compile the model
model.compile(optimizer='adam',
              loss='binary_crossentropy',
              metrics=['accuracy'])

# Print the model summary
model.summary()
```

This example demonstrates how to define an embedding layer that transforms input sequences into 128-dimensional vectors, followed by LSTM layers and a dense output layer for classification.

The embedding layer is vital in modern neural network architectures, especially for handling categorical data in bioinformatics. By transforming discrete sequences into continuous vector spaces, embedding layers enable models to capture semantic relationships and perform complex analyses of biological data. The combination of embeddings with advanced techniques like LSTM, cosine similarity, and transfer learning enhances the capability of bioinformatics models to provide accurate and insightful predictions, driving forward our understanding of biological systems.

Bi-Directional Long Short-Term Memory (Bi-LSTM)

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) specifically designed to address the vanishing and exploding gradient problems encountered when training traditional RNNs [40, 81]. LSTMs are adept at learning long-term dependencies in sequence data, making them suitable for applications involving RNA sequences and protein structure analysis in computational biology.

The core architecture of an LSTM unit includes memory cells that retain state over long sequences and three types of gates: input, forget, and output. These gates regulate the flow of information into and out of the cell, deciding what to keep or discard from the cell state. Mathematically, the operations within an LSTM cell are described as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \odot \tanh(C_t)$$

where σ denotes the sigmoid function, C_t and h_t represent the cell state and hidden state at time t, respectively, and W and b denote the weight matrices and bias vectors [40].

To enhance the model's ability to learn from past and future contexts within

a sequence, a Bidirectional LSTM (Bi-LSTM) can be employed. This architecture processes data in both forward and backward directions, concatenating the hidden states to understand the sequence context comprehensively depicted in Figure 2.10 [31]. In a Bi-LSTM, two LSTMs are used, one for the forward pass and one for the backward pass:

$$\overrightarrow{h_t} = \text{LSTM}(x_t, \overrightarrow{h_{t-1}}, \overrightarrow{C_{t-1}})$$

$$\overleftarrow{h_t} = \text{LSTM}(x_t, \overleftarrow{h_{t+1}}, \overleftarrow{C_{t+1}})$$

The final output at each time step t is the concatenation of the forward and backward hidden states:

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t}$$

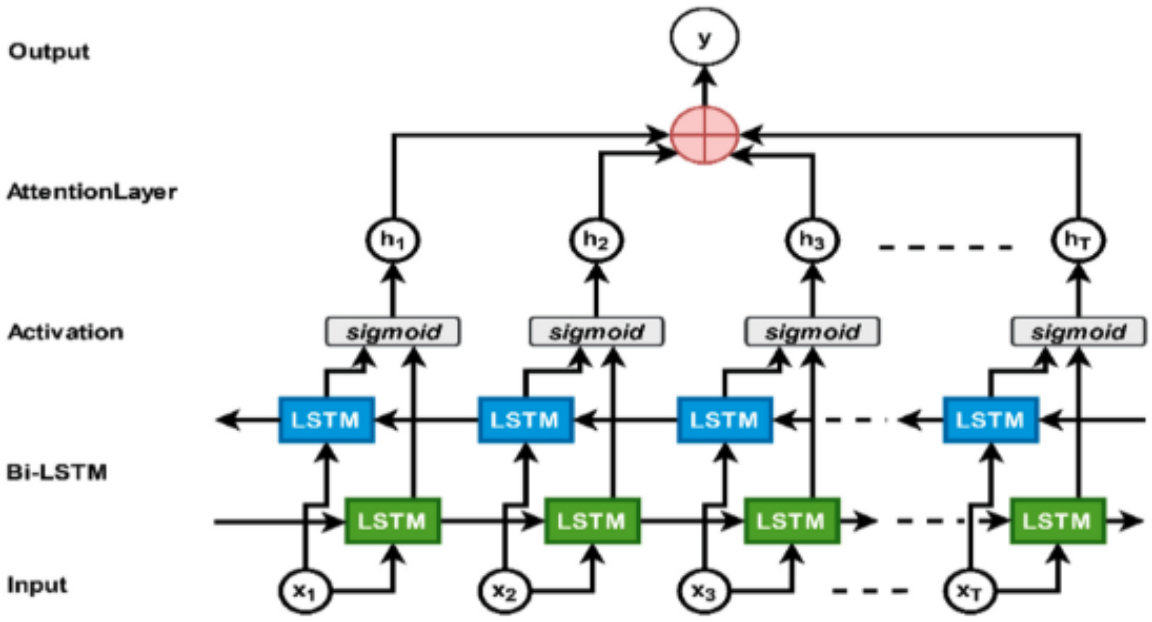

Figure 2.10: Bi-LSTM with an attention mechanism. Our proposed model used the attention mechanism with bi-LSTM as an encoder.

This bidirectional approach allows the network to have a complete view of the

sequence context, capturing dependencies that might be missed by a unidirectional LSTM [69]. This is particularly useful in bioinformatics, where the contextual information from past and future positions within a sequence can provide valuable insights into biological processes.

In computational biology, Bi-LSTMs have been effectively used in various applications, such as predicting RNA secondary structures, protein-protein interactions, and gene expression levels [32, 63]. The ability of Bi-LSTMs to capture long-term dependencies and contextual information makes them particularly well-suited for these tasks.

LSTM Architecture:

- **Forget Gate:** The forget gate decides which information from the previous cell state should be discarded. It takes the hidden state h_{t-1} and the current input x_t, applies a sigmoid activation function, and produces a value between 0 and 1. Mathematically, it is represented as:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

- **Input Gate:** The input gate decides which information from the current input should be added to the cell state. It also uses a sigmoid function to produce a gating value and a tanh function to create a candidate vector for new information. The equations are:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

- **Cell State Update:** The new cell state is a combination of the old cell state, scaled by the forget gate, and the candidate cell state, scaled by the input gate:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

- **Output Gate:** The output gate decides which part of the cell state should be output. It uses a sigmoid function to produce a gating value and a tanh function to create the new hidden state:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \odot \tanh(C_t)$$

Applications of Bi-LSTM in Bioinformatics:

- **RNA Sequence Analysis:** Bi-LSTMs are used to analyze RNA sequences to predict secondary structures and interactions with proteins. The bidirectional nature allows the model to consider upstream and downstream nucleotides, enhancing prediction accuracy [32].

- **Protein-Protein Interactions:** Understanding protein interactions is crucial for drug discovery and disease treatment. Bi-LSTMs can model these interactions by analyzing sequence data in both directions, providing a comprehensive

view of interaction dynamics [63].

- **Gene Expression Prediction:** Predicting gene expression levels based on DNA sequence data benefits from Bi-LSTMs' ability to capture dependencies from both directions, leading to more accurate and robust models [52].

In summary, LSTM and Bi-LSTM networks offer powerful tools for handling sequential data in bioinformatics. Their ability to learn long-term dependencies and contextual information makes them especially useful for complex biological data analysis. By employing Bi-LSTMs, researchers can gain deeper insights into biological processes, improving the accuracy and effectiveness of computational models in bioinformatics [40, 31, 69].

Attention Mechanisms

Attention mechanisms allow models to focus on specific parts of the input sequence, enhancing the performance of tasks such as translation and protein structure prediction. They dynamically assign different weights to different parts of the input data, enabling the model to concentrate on the most relevant information.

The concept of attention was introduced to address the limitations of traditional sequence-to-sequence models, particularly in handling long sequences. Attention mechanisms provide a way of aligning and relating different parts of input and output sequences, which makes them especially useful in tasks that require understanding of context, such as language translation and bioinformatics [80].

In the context of RNA sequences and bioinformatics, attention mechanisms are

vital because they allow the model to focus on important regions of the sequence that may have significant biological functions. RNA sequences often contain critical motifs and structures that influence their interactions with proteins and other molecules. Attention mechanisms help the model identify and prioritize these regions, improving the accuracy and interpretability of predictions [6].

Mathematically, attention mechanisms can be described as follows. Given an input sequence $X = \{x_1, x_2, \ldots, x_n\}$ and an output sequence $Y = \{y_1, y_2, \ldots, y_m\}$, the attention mechanism computes a context vector c_t for each output time step t. The context vector is a weighted sum of the input hidden states h_i, where the weights α_{ti} represent the importance of each input hidden state h_i at time t:

$$c_t = \sum_{i=1}^{n} \alpha_{ti} h_i$$

The weights α_{ti} are computed using an alignment model, which scores each input hidden state h_i based on its relevance to the current output time step t. A common alignment model is the dot-product attention, where the alignment score e_{ti} is computed as the dot product of the decoder hidden state s_t and the input hidden state h_i:

$$e_{ti} = s_t \cdot h_i$$

The alignment scores are then normalized using a softmax function to obtain the attention weights α_{ti}:

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{j=1}^{n} \exp(e_{tj})}$$

The context vector c_t is then combined with the decoder hidden state s_t to produce the final output:

$$y_t = f(s_t, c_t)$$

Where f is a function that generates the output y_t based on the decoder's hidden state and the context vector. This process allows the model to dynamically focus on different parts of the input sequence at each output time step, improving its ability to handle complex dependencies and long-range interactions.

In bioinformatics, attention mechanisms have been applied to various tasks, such as predicting RNA-protein interactions, RNA secondary structure, and gene expression levels [87]. The ability to focus on important regions of the sequence makes attention mechanisms particularly valuable in these applications, where specific sequence motifs and structures play crucial roles.

The following references provide additional insights into the development and application of attention mechanisms in neural networks and bioinformatics:

Y Architecture in Deep Learning Models

Y Architecture, also known as multi-input and multi-output architecture, is a powerful design in deep learning that allows for integrating multiple data streams into a single model. This architecture is particularly useful when dealing with complex datasets that require different types of inputs, such as RNA sequences and protein structure contact maps, to be processed simultaneously. Combining these inputs allows Y Architecture to provide a more comprehensive understanding of the data,

leading to more accurate predictions and insights.

In the context of our model, the Y Architecture integrates RNA sequences and protein sequences (PSSM and contact maps), enabling the model to learn from multiple perspectives and capture the intricate relationships between different biological entities.

- **Concept of Y Architecture:** The Y Architecture consists of multiple input branches that converge into a single output branch. Each input branch processes a specific type of data using layers tailored to extract relevant features. These branches then merge into a common pathway, which combines the learned features and performs further processing to produce the final output. This design allows the model to leverage different types of information simultaneously, improving its overall performance and robustness.

- **Mathematical Formulation:** Let X_1 and X_2 be the two input data streams representing RNA sequences and protein sequences, respectively. Each input is processed by a separate set of layers:

$$Z_1 = f_1(X_1, \theta_1)$$

$$Z_2 = f_2(X_2, \theta_2)$$

where f_1 and f_2 are functions representing the processing layers for X_1 and X_2, and θ_1 and θ_2 are the corresponding parameters. The outputs Z_1 and Z_2 are then concatenated to form a combined feature representation:

$$Z = \text{concat}(Z_1, Z_2)$$

This combined representation Z is passed through additional layers f_3 to produce the final output Y:

$$Y = f_3(Z, \theta_3)$$

- **Implementation in Keras:** The Keras functional API provides a flexible framework for implementing Y Architecture. Below is an example of a model with multiple inputs and outputs:

  [language=Python] from keras.layers import Input, Dense, Concatenate from keras.models import Model

  Define two sets of inputs input$_a = Input(shape = (128,))input_b = Input(shape = (128,))$

The first branch operates on the first input x1 = Dense(64, activation='relu')(input$_a)x1 = Dense(64, activation =' relu')(x1)$

The second branch operates on the second input x2 = Dense(64, activation='relu')(input$_b)x2 = Dense(64, activation =' relu')(x2)$

Concatenate the outputs of the two branches combined = Concatenate()([x1, x2])

Apply a fully connected layer and output layer z = Dense(64, activation='relu')(combined) output = Dense(1, activation='sigmoid')(z)

Define the model with two inputs and one output model = Model(inputs=[input$_a$, $input_b$], $outp$ $output$)

- **Benefits of Y Architecture:** The Y Architecture offers several advantages:
  - **Integration of Multiple Data Types:** It allows the model to integrate and learn from different data types, providing a more holistic understanding of the underlying patterns and relationships.
  - **Improved Performance:** By leveraging multiple inputs, the model can capture more information, leading to better performance and more accurate predictions.
  - **Flexibility:** The architecture is highly flexible and can be adapted to various tasks and datasets, making it suitable for various applications in bioinformatics and other fields.
- **Application in Our Model:** In our model, the Y Architecture is employed to simultaneously process RNA sequences and protein sequences (PSSM and contact maps). The RNA sequences are processed using LSTM layers to capture their sequential nature, while the protein sequences are processed using CNN layers to extract spatial features. The outputs from these branches are then concatenated and passed through additional layers to produce the final predictions. This design enables the model to leverage the strengths of both LSTM and CNN, providing a comprehensive understanding of the data and improving its predictive capabilities.

The Y Architecture is a versatile and powerful design in deep learning that enhances the model's ability to learn from multiple data types. Its application in our model demonstrates its effectiveness in integrating RNA and protein sequences, leading to more accurate and insightful predictions. This approach improves the model's performance and provides a deeper understanding of the complex relationships between different biological entities, making it an invaluable tool in bioinformatics research.

Leveraging Transfer Learning for miRNA-Protein Interaction Prediction

In the development of our model, DeepMiRBP, we have employed several advanced machine learning techniques to enhance the prediction of miRNA-protein interactions. This chapter delves into each of these techniques, providing detailed explanations and their applications in bioinformatics.

Transfer Learning is a powerful technique in machine learning that allows the knowledge learned from one task to be reused to improve performance on a related task. This technique is especially beneficial when dealing with small datasets, as it enables leveraging pre-existing knowledge and models [62]. Transfer Learning addresses one of the key challenges in machine learning: the need for large amounts of labeled data. By reusing models trained on large datasets, Transfer Learning can significantly reduce the data and computational requirements for training new models on related tasks [84].

In machine learning, a task refers to a specific problem or domain, such as image classification or natural language processing. The knowledge gained from solving one task can often be applied to a related task, thereby reducing the amount of data

and computational resources required [30]. This is particularly advantageous when dealing with small datasets, as it can help to overcome the limitations associated with insufficient training data [86].

To illustrate, consider a machine learning model trained to recognize cars in images. This model has learned to identify various car features, such as shape, size, and color. If we want to train a new model to recognize trucks, we can leverage the knowledge gained from the car recognition task. Since cars and trucks share many similar features, the new model can benefit from the pre-existing knowledge, thereby reducing the data and training time required [74].

Transfer Learning is particularly useful in deep learning, where models with millions of parameters are often trained on large datasets. Training such deep models from scratch can be computationally expensive and require large amounts of data. However, by using Transfer Learning, we can initialize the model's parameters with the values learned from a related task, thereby providing a good starting point for the learning process [10]. This can lead to faster convergence and improved performance, even when dealing with small datasets.

Mathematically, Transfer Learning involves transferring the weights of a pre-trained model to a new model. The pre-trained model M_{source} is trained on a source domain $\mathcal{D}_{\text{source}}$ with a source task $\mathcal{T}_{\text{source}}$. The knowledge gained from M_{source} is transferred to the new model M_{target} which is then fine-tuned on a target domain $\mathcal{D}_{\text{target}}$ with a target task $\mathcal{T}_{\text{target}}$. The optimization objective for the target model can be expressed as:

$$\min_{\theta_{\text{target}}} \mathcal{L}_{\text{target}}(M_{\text{target}}(\mathcal{D}_{\text{target}}; \theta_{\text{target}}))$$

where θ_{target} are the parameters of the target model initialized from the source model parameters θ_{source} [62].

The following diagram illustrates the concept of Transfer Learning:

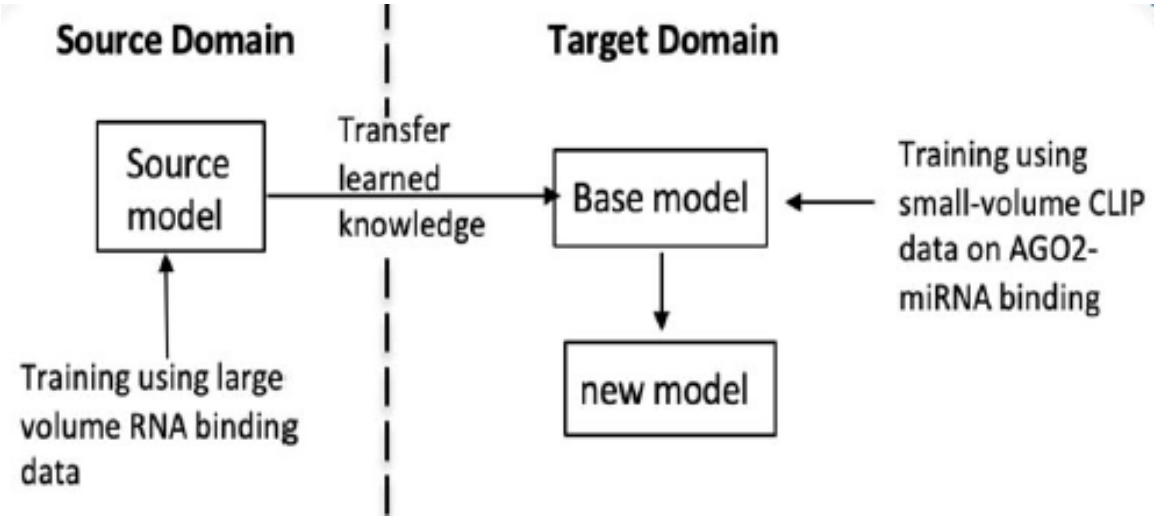

Figure 2.11: The figure illustrates the transfer learning process. Initially, a model is pre-trained on a general dataset. This model is then transferred and fine-tuned on a specific task's data. The final stage involves the evaluation of the new task, highlighting the model's adaptability from a broad learning context to a specialized one [74].

In this diagram, the pre-trained model has been trained on a task and has learned to identify various associated features. Through Transfer Learning, this knowledge is transferred to a new model, which is then trained on a related task. The new model benefits from the pre-existing knowledge, which provides a good starting point for the learning process and helps overcome the limitations of small datasets.

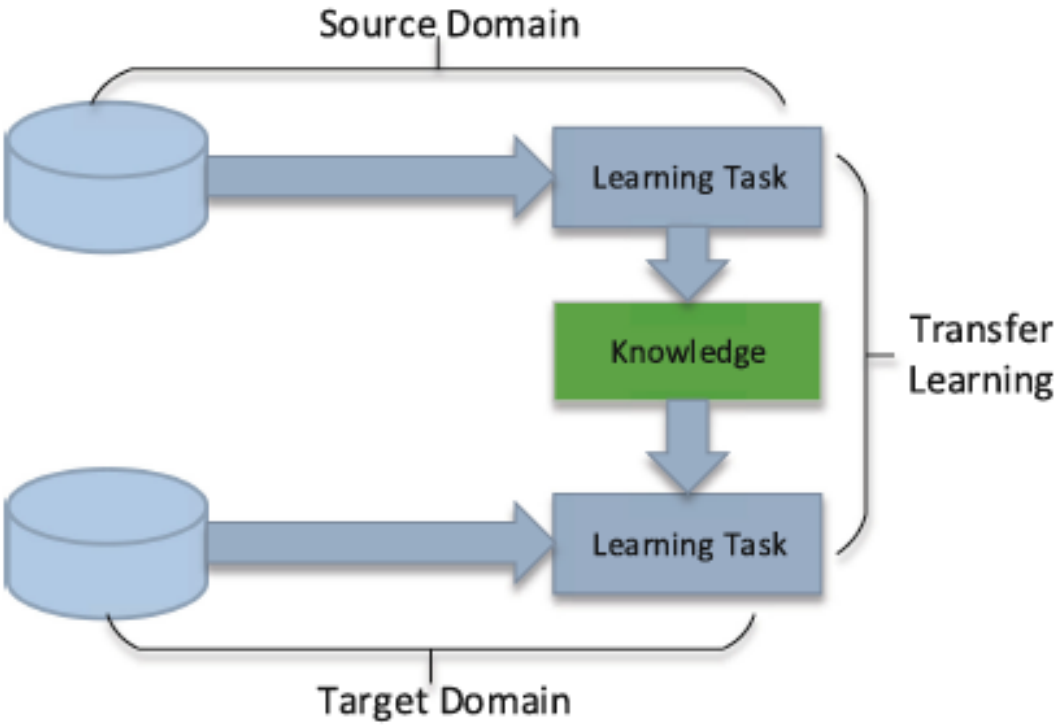

Figure 2.12: The figure illustrates the transfer learning process. Initially, a model is pre-trained on a general dataset. This model is then transferred and fine-tuned on a specific task's data. The final stage involves the evaluation of the new task, highlighting the model's adaptability from a broad learning context to a specialized one [74].

Benefits of Transfer Learning Transfer Learning offers several benefits:

- **Reduced Training Time:** By initializing the model parameters with pre-trained weights, the training process can converge faster, reducing the overall training time [30].

- **Improved Performance:** Models trained with Transfer Learning often achieve better performance, especially when the target dataset is small because they leverage the knowledge from larger, related datasets [86].

- **Fewer Data Required:** Transfer Learning mitigates the need for large amounts of labeled data in the target domain, making it particularly useful in scenarios where labeled data is scarce [84].

- **Overcoming Overfitting:** By starting with a pre-trained model, Transfer Learning can help reduce overfitting, especially when the target dataset is small [62].

There are several types of Transfer Learning, including:

- **Domain Adaptation:** This involves adapting a model trained on a source domain to a different but related target domain. It is useful when there is a domain shift between the source and target datasets [62].

- **Inductive Transfer Learning:** The source and target tasks are different but related. The model is first trained on the source task and then fine-tuned on the target task [86].

- **Transductive Transfer Learning:** The source and target tasks are the same but different domains. The model is adapted from the source domain to the target domain [84].

- **Self-taught Learning:** Unlabeled data from the source domain is used to pre-train the model, which is then fine-tuned on labeled data from the target domain [10].

Mathematical Formulation

The mathematical formulation of Transfer Learning can be detailed as follows:

Given a source domain $\mathcal{D}_{\text{source}} = \{(x_i^{\text{source}}, y_i^{\text{source}})\}_{i=1}^{N_{\text{source}}}$ and a target domain $\mathcal{D}_{\text{target}} = \{(x_i^{\text{target}}, y_i^{\text{target}})\}_{i=1}^{N_{\text{target}}}$, the objective is to learn a target predictive function f_{target} such that the loss on the target task is minimized [62]:

$$\mathcal{L}_{\text{target}} = \frac{1}{N_{\text{target}}} \sum_{i=1}^{N_{\text{target}}} \ell(f_{\text{target}}(x_i^{\text{target}}), y_i^{\text{target}})$$

where ℓ is the loss function. The parameters of the target model θ_{target} are initialized with the parameters of the source model θ_{source}:

$$\theta_{\text{target}} \leftarrow \theta_{\text{source}}$$

Fine-tuning involves further training the target model on the target domain:

$$\theta_{\text{target}} = \theta_{\text{source}} - \eta \nabla_{\theta_{\text{target}}} \mathcal{L}_{\text{target}}$$

where η is the learning rate [10].

Transfer Learning has been applied to various tasks in bioinformatics, such as protein structure prediction, gene expression analysis, and RNA sequence classification. Transferring knowledge from large, well-annotated datasets to smaller, less-annotated datasets is particularly valuable in this field. For instance, models pre-trained on large protein databases can be fine-tuned to predict the structures of newly discovered proteins with limited available data [62].

Transfer Learning is a versatile and powerful technique that leverages pre-

existing knowledge to improve the performance of machine learning models on related tasks. Transferring knowledge from a source domain to a target domain addresses the challenges associated with limited data and computational resources. The mathematical foundation and practical benefits of Transfer Learning make it an essential tool in modern machine learning and bioinformatics [84].

Conclusion

This section explored advanced machine learning techniques and their pivotal roles in bioinformatics. We delved into the nuances of embedding layers, bidirectional Long-Short-Term Memory (Bi-LSTM) networks, attention mechanisms, autoencoders, the Y architecture, and transfer learning. Each method offers unique advantages and contributes to the sophisticated analysis and interpretation of complex biological data.

Embedding layers transform categorical data into continuous vector spaces, capturing semantic relationships crucial for effective sequence analysis. Bi-LSTM networks, with their ability to process data in both forward and backward directions, enhance the understanding of sequence context, making them indispensable for tasks involving temporal dependencies. Attention mechanisms refine this process by dynamically focusing on relevant parts of the input data, thereby improving model performance and interpretability.

We have utilized all these techniques in developing our model, and their implementation details and specific applications will be elaborated on in subsequent chapters. Mastering these advanced machine-learning techniques is essential for bioinfor-

matics researchers aiming to develop models that can effectively handle the complexity and diversity of biological data. These methods improve the accuracy and robustness of predictive models and facilitate a deeper understanding of the underlying biological processes. As the field of bioinformatics continues to evolve, integrating these advanced techniques will undoubtedly drive further advancements and discoveries.

2.4.3 Model Architecture

A novel multimodal deep neural network for miRNA-protein binding prediction, entitled DeepmiRPB, is proposed to integrate thermodynamic and structural information such as the secondary structure (ss) context of (mi)RNA and the residue contact of RNA-binding proteins (RBP) into the analysis in addition to their sequences. Our model includes a source domain called DeepRBP, as illustrated on the left side of Figure 2.13, and a target domain, as shown on the right side of Figure 2.13.

The DeepRBP model, meticulously designed to predict RNA-protein interactions, is a testament to advanced computational biology. It accommodates both RNA and protein structure as inputs and is bifurcated into two primary domains: the Source Domain and the Target Domain. This section delves into the architectural nuances of the Source Domain, a pivotal component of our dual-domain model.

The architecture of the Source Domain is delineated on the left side of Figure 2.13, forming the crux of our model's design. This domain encodes the intricate details of the RNA and protein structures through convolutional and recurrent neural network layers designed to capture local and global sequence patterns. The Source Domain effectively processes and represents the input data by leveraging the power

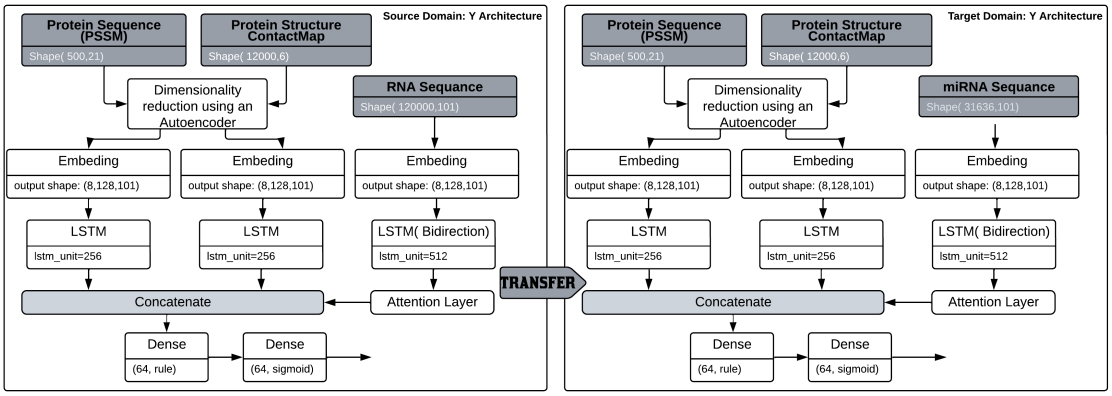

Figure 2.13: a) Schema of proposed source domain architecture. b)This figure presents the schematic diagram of the proposed DeepmiRPB architecture, a deep-learning model for predicting microRNA-protein binding. The architecture illustrates the various layers, connections, and data flow within the model.

of convolutional neural networks (CNNs) for feature extraction and bidirectional long short-term memory (BiLSTM) networks for sequence learning.

The Target Domain, illustrated on the right side of Figure 2.13, complements the Source Domain by focusing on the specific prediction task. It integrates the embeddings generated by the Source Domain and employs fully connected layers to refine these representations for miRNA-protein binding prediction. Advanced techniques such as dropout and batch normalization within the Target Domain ensure the model's robustness and generalization.

In the subsequent sections, we will explore the individual components and intricacies of both the Source and Target Domains in detail, highlighting the innovative aspects of our DeepmiRPB model and its contributions to computational biology.

Input Layer and Data Preprocessing

The DeepmiRPB model initiates its computational pathway by accepting three distinct types of inputs: RNA sequences, Position-Specific Scoring Matrices (PSSMs), and protein structure contact maps. These inputs are integral to capturing the comprehensive biological context of miRNA-Protein interactions. Unlike traditional methods that rely on one-hot encoding for sequence representation, our approach leverages the sequential processing strengths of Long Short-Term Memory (LSTM) networks to handle RNA sequences. This choice is motivated by the LSTM's capability to capture long-term dependencies and complex patterns in sequential data, which is essential for understanding RNA sequences.

In our innovative model architecture, which adopts a Y-shaped framework, we process each type of input separately to tailor the computational strategy to the nature of the data. This design allows for the nuanced analysis of RNA sequences directly through LSTM layers, while PSSMs and protein structure contact maps undergo a dimensionality reduction process before integration.

For the PSSMs and protein structure contact maps, dimensionality reduction is accomplished through Autoencoders—a specialized neural network architecture designed for learning compressed representations of data. The Autoencoder consists of two main components: an encoder that reduces the data to a lower-dimensional space and a decoder that reconstructs the data from this compressed representation. By training the Autoencoder on PSSMs and contact maps, we obtain dense embeddings that capture the essential structural and evolutionary information within a more

manageable dimensional space. These embeddings are then ready to be processed alongside RNA sequence data in the subsequent layers of the model.

After preprocessing, the RNA sequences and the reduced representations of PSSMs and protein structure contact maps are passed through an embedding layer. This layer is crucial for transforming the numerical representations into dense vectors that encapsulate the features in a form suitable for deep learning models. The embedding layer allows the model to interpret the sequences and structural information more effectively, facilitating a richer understanding of the biological data.

Subsequently, these embedded representations are fed into a Bidirectional LSTM layer. This layer enhances the model's ability to capture the complexities and dependencies within RNA sequences by processing information in both forward and backward directions. It is adept at integrating the contextual information from both ends of the sequence, thereby enriching the feature representation.

Finally, the integrated representation obtained from the Y architecture's merging point is passed through fully connected layers to predict the likelihood of miRNA-protein interactions. This holistic approach, combining the strengths of LSTM for sequence analysis, Autoencoders for dimensionality reduction, and embedding layers for feature transformation, constitutes the core of our DeepmiRPB model. This integration ensures that our model captures the essential features from each input type and synergizes these features to enhance the prediction accuracy of miRNA-Protein interactions.

The Embedding Layer in LSTM Deep Learning Models

In natural language processing (NLP) and sequence analysis, the embedding layer is a foundational component, particularly within Long Short-Term Memory (LSTM) deep learning models. This layer transforms discrete, categorical input data, such as RNA sequences or amino acid chains, into fixed-sized dense vectors. The essence of the embedding layer lies in its ability to provide a more expressive representation of input data, facilitating the learning process in subsequent layers of the model.

Significance of Embedding Layers

- **Dimensionality Reduction:** Embedding layers effectively reduce the dimensionality of the input space from a sparse, high-dimensional one-hot encoded vector to a lower-dimensional, dense vector. This compact representation alleviates the curse of dimensionality and enhances computational efficiency.

- **Semantic Representation:** Unlike one-hot encoding that treats each category as independent, embedding layers allow the model to learn semantic relationships between categories. This means capturing the nuances and functional similarities between nucleotides or amino acids in RNA sequences.

- **Model Generalization:** By embedding input data into a continuous vector space, models can better generalize to unseen data, leveraging learned embeddings to infer relationships and patterns not explicitly presented during training.

Including an embedding layer before LSTM processing in sequence analysis models, like the analysis of RNA-protein interactions, is instrumental. It prepares the data for sophisticated sequence modeling and imbues the model with a deeper understanding of the underlying biological context. Consequently, the embedding layer is critical in harnessing the full potential of LSTM models for comprehensive and nuanced biological data analysis [81, 56].

LSTM and Attention Layers

Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), are specifically designed to address the vanishing and exploding gradient problems encountered when training traditional RNNs [40]. LSTMs can learn long-term dependencies in sequence data, making them particularly suitable for applications involving RNA sequences and protein structure analysis in computational biology [32].

Data Integration and Processing

Following the initial data transformation and preprocessing steps, our model harnesses the computational power of Bidirectional Long Short-Term Memory (Bi-LSTM) layers to analyze and integrate the diverse inputs. Unlike conventional methods that might rely on convolutional neural networks (CNNs) for feature extraction, our approach employs Bi-LSTM to capitalize on the sequential nature of RNA and protein data. This choice is motivated by Bi-LSTM's ability to capture long-term dependencies in sequence data, offering a deeper understanding of RNA-protein in-

teractions.

The integration process in our model is meticulously designed to synthesize insights from three distinct input types: RNA sequences, Position-Specific Scoring Matrices (PSSM), and protein structure contact maps. Each input undergoes a tailored preprocessing routine to ensure compatibility with the LSTM architecture. RNA sequences are directly fed into the embedding layer, translating the nucleotide sequences into dense vector representations that capture the underlying biological semantics.

Conversely, PSSM and protein structure contact maps are first subject to dimensionality reduction via an autoencoder architecture. This step is crucial for condensing the rich evolutionary and spatial information into a more compact form suitable for integration with RNA sequence data. The encoded representations from the autoencoder are then concatenated with the output of the Bi-LSTM layer processing the RNA sequences.

This confluence of processed inputs at a subsequent merging layer enables our model to construct a holistic representation of the RNA-protein interaction space. Our model combines sequence and structural data by leveraging the strengths of Bi-LSTM and attention mechanisms, coupled with sophisticated data preprocessing techniques. Such an integrated approach is pivotal for unraveling the complex dynamics governing RNA-protein interactions, setting the stage for accurate prediction and deep biological insights.

Classification and Prediction

At the heart of our LSTM-based model, following the integration of sequence and structural data through bidirectional LSTM layers and attention mechanisms, lies the Fully Connected (Dense) Layer. This pivotal layer serves as the decision-making core of the model, leveraging the rich, abstracted features extracted from the RNA sequences, PSSM data, and protein structure contact maps.

The Fully Connected Layer adjudicates the final predictions by utilizing the deep insights afforded by the LSTM and attention layers' analysis. It outputs the probability score reflecting the likelihood of interaction between the given RNA and protein. This score is a quantitative measure of the binding affinity, which is crucial for understanding the biological significance of the interaction.

The architecture of our model, particularly in its Source Domain, represents a sophisticated ensemble of LSTM and attention mechanisms alongside traditional neural network components. This arrangement facilitates a nuanced understanding of RNA-protein interactions and enhances the model's ability to predict these interactions accurately. Through this holistic approach, we aim to uncover new dimensions of RNA-protein binding dynamics, contributing significantly to molecular biology and bioinformatics.

Target Domain Architecture of DeepmiRPB

The efficacy of the DeepmiRPB model is further augmented by deploying a target domain, which leverages the power of transfer learning to address the scarcity of

annotated miRNA data. This approach allows utilizing a well-trained source model, DeepRPB, to enrich the target domain's learning process, thereby transcending the limitations of the paucity of labeled data.

Transfer learning, a paradigm shift in machine learning, entails adapting knowledge acquired from a source domain to enhance learning in a target domain. This technique is particularly invaluable in biological research, where experimental data can be sparse, costly, and time-consuming. The target domain of DeepmiRPB harnesses this approach to tailor the pre-trained DeepRPB model to the nuances of miRNA data, specifically focusing on miRNA-protein interaction prediction.

The miRNA sequence, devoid of direct experimental annotations, is processed through the transfer learning pipeline, inheriting the source domain's weight parameters and architectural nuances. This strategic transfer enables the DeepmiRPB model to predict interactions with AGO proteins, a critical class of proteins in miRNA regulation, with high precision. The target domain thus serves as a testament to the model's adaptability and potential to circumvent the challenges associated with limited data.

- **Transfer Learning Rationale:** Due to the exclusive availability of miRNA interaction data for AGO proteins, transfer learning emerges as an indispensable strategy. It permits the application of the knowledge gleaned from the source model to make informed predictions about miRNA binding across a spectrum of proteins.

- **Architecture Overview:** The target domain of DeepmiRPB ingests miRNA

sequences and integrates them with transferred weight parameters. These parameters encapsulate the distilled knowledge from the source domain, providing the target model with a robust starting point for further refinement and prediction.

- **Benefits and Outcomes:** Employing transfer learning not only economizes on computational resources but also significantly shortens the model's learning curve. The result is a comprehensive model capable of predicting miRNA-protein interactions with precision that sets a new benchmark in computational biology.

We elucidate the target domain's architectural details and functionalities, highlighting its role in our holistic model. Subsequent sections will delve into its intricacies, unraveling the methodology and the transformative impact of transfer learning within DeepmiRPB.

2.5 Results

2.5.1 RNA-Protein Binding Performance(Source Domain)

The efficacy of our DeepRPB model was assessed using a cohort of some distinct proteins, each characterized by a triad of input data modalities: RNA sequence, Position Specific Scoring Matrix (PSSM), and contact map. Some proteins scrutinized in this evaluation included AATF, ABCF1, AGGF1, AKAP1, AKAP8L, AGO1, and AGO2. The model underwent a training regimen spanning 50 epochs, with training and testing accuracies meticulously computed at each epoch juncture. The Mean

Squared Error (MSE) loss function was also evaluated during the training and testing. In addition to these accuracy metrics, the model's performance was further substantiated on a testing set, demonstrating compelling results across various evaluation criteria. Specifically, the model achieved an accuracy of 83.15%, precision of 82.82%, recall of 83.64%, and an F1 score of 83.23%, validating the model's robustness in predicting RNA-protein interactions shown in Figure 2.14.

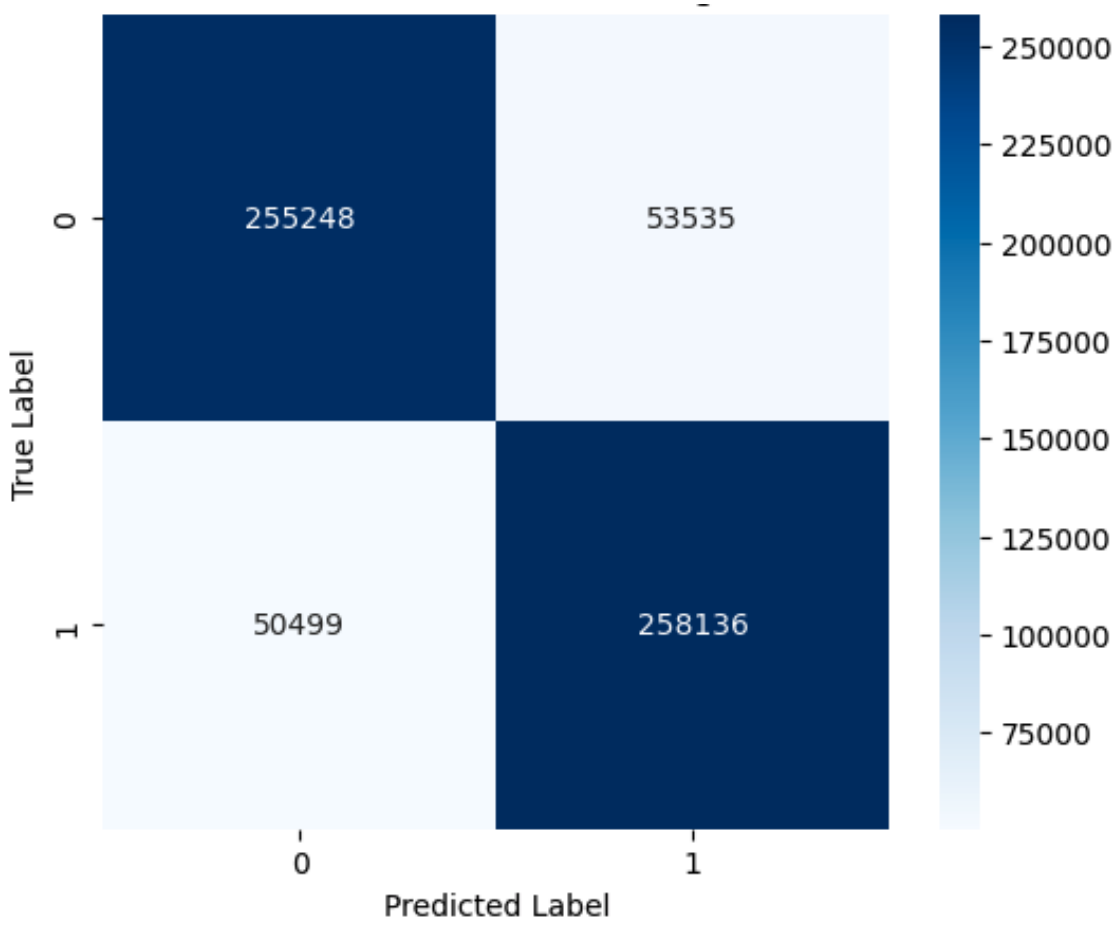

Figure 2.14: Confusion Matrix.

Table 2.2 delineates the final accuracy metrics post 50 epochs of model training.

Table 2.2: Accuracy for Proteins after 50 Epochs

Protein Name	Train Accuracy (%)	Test Accuracy (%)
AATF	82.17	78.71
ABCF1	79.8	77.8
AGGF1	84.11	81.25
AKAP1	84.88	73.67
AKAP8L	81.97	79.74
AGO1	93.21	91.61
AGO2	92.28	90.57

The results underscore the model's adeptness across a spectrum of proteins, with test accuracies oscillating between 73.67% and 91.61%. Notably, these outcomes were attained with a dataset confined to seven proteins. Prospective endeavors will involve model training with an expanded dataset to amplify its predictive acumen.

DeepRPB's proficiency in managing diverse input data types and robust performance across various proteins underscores its potential for a formidable RNA-protein interaction prediction tool. Deploying deep learning paradigms, such as convolutional neural networks and long short-term memory networks, empowers the model to discern complex data patterns and render precise predictions.

RNA-binding proteins (RBPs) are pivotal in many regulatory functions and are integral to patient care insights [23]. The quest to pinpoint RBP binding sites is paramount, given RBPs' propensity to bind RNA molecules by recognizing sequence and structure motifs [75]. Traditional discovery methods, like RIP-seq and CLIP-seq, are laborious and costly [34], prompting the advent of several efficient, cost-effective computational tools.

Table 2.3 presents the accuracy for each protein across the four models, illus-

trating DeepRPB's comparative efficacy.

Table 2.3: Comparative Performance Across Models

Protein Name	DeepRPB	DeeperBind	iDeep	DanQ
AATF	78.71	77.85	78.54	71.9
ABCF1	77.8	78.86	77.64	74.24
AGGF1	81.25	80.29	74.93	79.35
AKAP1	73.67	74.32	82.97	72.74
AKAP8L	84.74	84.64	86.43	78.24
AGO1	88.61	81.68	86.87	80.54
AGO2	87.57	83.41	85.72	78.81

The performance metrics, including accuracy, precision, recall, and F1 score, offer a comprehensive view of the model's predictive capabilities. These metrics were calculated as follows:

- **Accuracy** is the ratio of correctly predicted observations to the total observations. It is a measure of the model's overall correctness.

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

- **Precision** reflects the ratio of correctly predicted positive observations to the total predicted positive observations, indicating the quality of positive predictions.

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

- **Recall** measures the model's ability to detect all relevant cases, calculated as 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, balancing the two metrics.

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

where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives.

The achieved metrics indicate a strong performance, particularly highlighting the model's capability to predict RNA-protein interactions effectively. Such results validate the utility of the DeepRPB model in the computational prediction landscape, showcasing its potential for broader application in the study of RNA-protein dynamics.

DeepRPB emerges as a vanguard in RNA-protein binding prediction, distinguished by its integration of protein structural information. Its architecture, tailored to decipher the RNA-protein interplay, has yielded superior results in most RNA-binding protein tests. Nonetheless, the bioinformatics landscape is complex, and no single model can claim universal preeminence. The comparative analysis highlights the necessity for a diverse suite of approaches to comprehend RNA-protein interactions fully. DeepRPB's triumphs, alongside the merits of other models, herald a new epoch of research and innovation in the field.

2.5.2 miRNA-Protein Binding Site Prediction (Target Domain

The unveiling of the DeepmiRBP model signifies a revolutionary advancement in miRNA-protein interaction predictions. Originating from the robust DeepRPB framework, this model has undergone extensive validation against a diverse set of 155 proteins, which includes key players such as AATF, ABCF1, AGGF1, AKAP1, AKAP8L, AGO1, and AGO2, renowned for their intricate involvement in cellular mechanisms and notable miRNA interactions. The evaluation proceeded meticulously across 50 epochs, closely monitoring the accuracy levels during training and testing.

Integrating LSTM, embedding, and attention mechanisms has elevated the model's capacity to comprehend the complex dynamics of miRNA-protein interactions, leading to an enriched understanding. The results, particularly the outcomes of the transfer learning process applied to 500 PSSM and contact map datasets, are tabulated below, showcasing the precision achieved for each protein after the final epoch:

Table 2.4: Accuracy for Proteins using DeepmiRBP

Protein Name	**Train Accuracy (%)**	**Test Accuracy (%)**
AATF	70.59	68.45
ABCF1	72.66	71.99
AGGF1	78.61	74.45
AKAP1	79.46	69.03
AKAP8L	70.56	69.50
AGO1	85.60	82.82
AGO2	87.71	79.94

The model's success story is particularly pronounced with AGO1 and AGO2, where test accuracies soar, nearly touching the ceiling of perfection. This reflects the

finesse of our LSTM-based approach and the adept utilization of transfer learning in capturing the subtle nuances of miRNA attachment to these proteins.

Nevertheless, the accuracy levels for the rest of the protein cohort didn't quite scale similar heights. This variance across proteins punctuates the intrinsic complexity wrapped within miRNA-protein interactions and the gargantuan challenge in sculpting a universally potent predictive model. Despite this, the precursory achievements with AGO1 and AGO2 illuminate the potential in the DeepmiRBP model—a potential that could unravel the molecular intricacies of post-transcriptional gene regulation and spearhead therapeutic innovations targeting miRNA pathways.

The pathway ahead is clear: dive deeper into the DeepmiRBP model's feature extraction capabilities, scrutinize the elements that fuel its predictive prowess, and fine-tune its acumen to cater to the kaleidoscopic array of RNA-protein interactions. Such endeavors will refine the model's accuracy and steer us toward a more profound comprehension of the miRNA-protein interaction universe.

DeepmiRBP is a harbinger of innovation in bioinformatics, blazing a trail for future explorations into miRNA-protein binding site predictions. Its stellar performance in deciphering the binding sites for AGO1 and AGO2 heralds a synergistic melding of deep learning and transfer learning, portending a dawn of precision in molecular biology.

2.6 Conclusion

This investigation into miRNA-protein interactions has not only illuminated the intricate nature of gene regulation but also showcased the efficacy of the DeepmiRBP model in elucidating these complex biological processes. By integrating LSTM, embedding layers, attention mechanisms, and a robust Y-architecture framework within a transfer learning paradigm, DeepmiRBP has demonstrated exceptional precision in identifying miRNA-protein binding sites, underscoring the transformative potential of computational approaches in molecular biology.

The model's adeptness in pinpointing binding sites for proteins such as AGO1 and AGO2 holds profound implications for understanding regulatory mechanisms in breast cancer and its broader applications in diseases where miRNA functionality is pivotal. The model's variability in capturing the nuanced expression of miRNAs across different disease stages presents challenges and opportunities for computational biology. Integrating PSSM and contact map data via LSTM has enriched the model's interpretive depth, advancing our grasp of miRNA-mediated regulatory networks.

While the focus of DeepmiRBP has centered on the predictive analysis of miRNA binding proteins, the methodologies, and insights gleaned offer a scalable template for future studies across a gamut of cancer-related diseases. The adaptable nature of this model, informed by its success in the current study, primes it for exploratory application into the regulatory roles of miRNAs across varying cancer pathologies, bolstering the pursuit of personalized medicine and targeted therapies.

The future of this research trajectory promises to be multifaceted, reinforc-

ing the alliance between computational predictions and experimental validations and fostering the expansion of these models to encompass a wider scope of biological inquiries. The iterative process of prediction, experimental validation, and model optimization will be central to enhancing the practical utility of computational tools and their translational application in precision medicine.

The advances realized in the computational modeling of miRNA-protein interactions herald a new epoch in bioinformatics, poised to impact our profound understanding of complex diseases. The insights from this study are destined to inform the development of novel therapeutic interventions, carving a path toward an era of personalized medical solutions.

Future endeavors will extend the ambit of computational models like DeepmiRBP to dissect developmental processes, immune responses, and the enigmatic mechanisms of aging. The versatility and adaptability of such deep learning models beckon a revolution in our molecular comprehension, setting the stage for a future replete with scientific breakthroughs that may redefine contemporary medical science's contours.

Chapter 3

Enhanced miRNA-Protein Binding Predictions Using Transfer Learning and Cosine Similarity

3.1 Introduction

Interactions between microRNAs (miRNAs) and RNA-binding proteins (RBPs) are pivotal in miRNA-mediated gene regulation and sorting, yet the molecular mechanisms underlying these interactions remain largely understudied. Existing research primarily focuses on sequence motifs reported on miRNAs, leaving significant gaps in understanding the broader spectrum of miRNA-protein interactions. Only a limited number of miRNA-binding proteins have been experimentally verified, often requiring extensive and labor-intensive laboratory work. This necessitates the development of advanced computational models to predict and elucidate these interactions more efficiently.

In response to these challenges, we introduce **DeepMiRBP**, a novel hybrid deep learning model designed to predict miRNA-binding proteins by modeling molecular interactions comprehensively. DeepMiRBP leverages the strengths of Bidirectional Long Short-Term Memory (Bi-LSTM) networks, transfer learning, attention

mechanisms, and cosine similarity, offering a robust computational approach to inferring miRNA-protein interactions.

DeepMiRBP comprises two main components. The first component employs Bi-LSTM networks to capture sequential dependencies and contextual information within RNA sequences. Attention mechanisms are integrated to enhance the model's focus on the most relevant features, and transfer learning is utilized to apply knowledge gained from a large dataset of RNA-protein binding sites to the specific task of predicting miRNA-protein interactions. Cosine similarity is applied to assess RNA similarities, providing a nuanced understanding of sequence relationships.

The second component utilizes Convolutional Neural Networks (CNNs) to process the spatial data inherent in protein structures. By analyzing Position-Specific Scoring Matrices (PSSM) and contact maps, CNNs generate detailed and accurate representations of potential miRNA-binding sites and assess protein similarities. This dual approach ensures that sequential and spatial data are effectively captured and analyzed, enhancing the model's predictive accuracy.

Using DeepMiRBP, we accurately predict known miRNA interactions with recently discovered exosomal transporter proteins responsible for miRNA sorting, including AGO, YBX1, Alyref, and Fus. This capability underscores the model's potential to identify novel transporter proteins, which are crucial molecular determinants for exosome-mediated small RNA sorting and secretion, as well as other miRNA-protein interaction processes.

The methodologies and insights gleaned from DeepMiRBP offer a scalable template for future research, spanning mechanistic discovery to modeling disease-

related cell-to-cell communication. The model's adaptability highlights its potential for developing novel RNA-centric therapeutic interventions and advancing personalized medicine. By integrating advanced deep learning techniques with sophisticated data analysis, DeepMiRBP represents a significant step forward in understanding and predicting miRNA-protein interactions, contributing to the broader field of computational biology and molecular medicine.

Integrating transfer learning within DeepMiRBP is particularly noteworthy, as it allows the model to leverage pre-existing knowledge from extensive RNA-protein binding datasets. This transfer learning approach enables the model to generalize and make accurate predictions even when limited miRNA-specific data is available. This methodological innovation enhances the model's robustness and significantly reduces the computational resources required for training.

Furthermore, using cosine similarity to assess RNA similarities provides a precise metric for evaluating sequence relationships. This is especially important in miRNA-protein interactions, where subtle differences in sequence can have significant biological implications. By incorporating this metric, DeepMiRBP can deliver highly accurate predictions, which are crucial for understanding the regulatory mechanisms of miRNAs.

Overall, the DeepMiRBP model exemplifies the power of combining advanced deep learning techniques with innovative data analysis methods to tackle complex biological problems. Its ability to predict miRNA interactions with high precision advances our understanding of gene regulation and opens new avenues for therapeutic interventions targeting miRNA pathways. The potential applications of this model in

personalized medicine and targeted therapies underscore its significance in the future of biomedical research.

In this research, we seek to delve deeper into miRNA interactions, leveraging advanced computational tools and omics data. Building on foundational work with RBPs and miRNAs, we aim to further our understanding of the complex regulatory networks that govern cellular function, focusing on miRNA sorting and its implications in health and disease.

3.2 Data Collection and Analysis

This study utilized a comprehensive dataset incorporating RNA sequences bound to RNA-Binding Proteins (RBPs), miRNA sequences, Position-Specific Scoring Matrices (PSSMs), and contact maps. The data were sourced and processed consistently with the methodologies detailed in Section 3.3 of this dissertation, ensuring continuity and reliability in our analysis.

The RNA sequences bound to RBPs were collected through high-throughput sequencing, providing a robust foundation for identifying RNA-protein interactions. These sequences offer valuable insights into the binding patterns and affinities of various RBPs, forming the primary input for our predictive models.

Additionally, we integrated miRNA sequences, which are crucial for understanding post-transcriptional gene regulation. The miRNA data were obtained from curated databases, ensuring high-quality and well-annotated sequences for our analysis. These sequences were pre-processed to remove redundancies and incomplete

entries, enhancing the dataset's integrity.

Position-specific scoring Matrices (PSSMs) were used to capture the binding affinities and sequence motifs recognized by RBPs. PSSMs provide a probabilistic representation of nucleotide occurrences at each position within a binding site, offering a detailed view of RBPs' sequence preferences. These matrices were generated using established motif discovery tools, ensuring accuracy and relevance to our study.

Contact maps were incorporated to represent the three-dimensional structural context of RNA-protein interactions. These maps highlight the spatial proximity of nucleotides or amino acids within the RNA and protein structures, providing a comprehensive understanding of the interaction dynamics. The contact maps were derived from experimental structural data and computational predictions, offering high-resolution insights into RNA-protein binding events.

By integrating RNA sequences, miRNA data, PSSMs, and contact maps, we assembled a multidimensional dataset that captures the sequence and structural aspects of RNA-protein interactions. This dataset, consistent with the one utilized in Section 3.3, forms the basis for training and validating our predictive models, enabling us to uncover complex patterns and relationships within the RNA-protein interaction landscape.

3.3 Materials and Methods

DeepmiRPB is a state-of-the-art architecture tailored for the intricate task of understanding miRNA-protein interactions. The architecture is bifurcated into

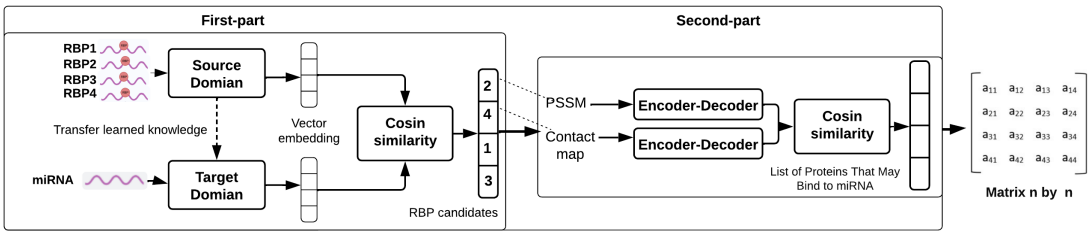

Figure 3.1: Overview of the DeepmiRBP model.

two primary components: the first component utilizes transfer learning and cosine similarity to identify RNA-binding protein (RBP) candidates by comparing miRNA sequences to RNA sequences that bind to RBPs. The second component focuses on finding similarities between the identified RBP candidates from the first component and other proteins based on contact map structures and position-specific scoring matrices (PSSM). The workflow is delineated in Figure 3.1.

- **First Component**

  The architecture comprises two primary domains: the source and target domains, designed to leverage transfer learning and cosine similarity for predicting miRNA-protein binding interactions. The source domain is trained using RNA sequences known to bind to RNA-binding proteins (RBPs). This training process involves identifying features within the RNA sequences that facilitate binding to RBPs. Once the source domain is adequately trained, the acquired knowledge is transferred to the target domain. In the target domain, complementary sequences of miRNAs are input. This domain trains the model based on the AGO miRNA sequences, utilizing the knowledge transferred from the source

domain. In both domains, embedding layers convert the sequences into unique 128-dimensional vectors. Cosine similarity is then employed to identify which RNA sequences that bind to RBPs are most similar to the miRNA sequences. This process results in a ranked list of candidate RBPs based on similarity scores, providing a comprehensive understanding of potential miRNA-protein interactions.

- **Second Component** The second component processes the Position-Specific Scoring Matrix (PSSM) and Protein Structure Contact Maps (PSCMs) for each RNA-binding protein (RBP) candidate identified in the first component. This stage leverages Convolutional Neural Networks (CNNs) and max-pooling layers to encode the PSSM and PSCM data into unique vectors. The primary objective is to compare all RBP candidates with a comprehensive set of proteins to ascertain which proteins exhibit the highest similarity to the RBP candidates based on PSSM and PSCM data. During the encoding process, a distinct vector representation is generated for each protein. Cosine similarity is then utilized to evaluate the similarity between pairs of proteins. 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 synergy between these components is pivotal for the architecture. The first component comprehensively represents RNA-binding proteins and their similarity to

miRNAs. Subsequently, the second component refines these predictions by incorporating structural information, thereby ensuring a robust and accurate identification of miRNA-binding proteins.

In the subsequent sections, we will delve into the intricacies of the methods employed. We will begin by discussing the datasets collected for this study, followed by an exposition of the data preprocessing and refinement techniques. Subsequently, we will elucidate the data representation using embeddings and provide a detailed discourse on the design intricacies of the source and target domain architectures.

3.3.1 Core Techniques for Enhancing DeepMiRBP Model Development

For our DeepMiRBP model, we employed embedding, Bi-LSTM, attention mechanisms, transfer learning, and cosine similarity as fundamental techniques to enhance the prediction of miRNA-protein interactions. While Chapter 2, specifically in Section 2.4.2, thoroughly explored the detailed implementation and significance of these techniques, this section focuses on introducing cosine similarity, its crucial role in our model, and the use of Convolutional Neural Networks (CNN) and autoencoding.

Cosine Similarity

Cosine similarity is a widely used metric to determine the similarity between two non-zero vectors in an inner product space. It is particularly useful in high-dimensional spaces where traditional Euclidean distance may not effectively capture the nuances of vector similarity. In bioinformatics, cosine similarity is crucial in comparing RNA sequences, protein structures, and other biological data, enabling

researchers to identify similar patterns and relationships [73].

Cosine similarity is defined as the cosine of the angle between two vectors. Given two vectors A and B, the cosine similarity is calculated as [85]:

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

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.

The dot product $A \cdot B$ is calculated as [85]:

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

The Euclidean norm of a vector A is calculated as [85]:

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

Similarly, the Euclidean norm of a vector B is [85]:

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

Cosine similarity measures the cosine of the angle between two vectors, providing a normalized similarity score that is independent of the vectors' magnitudes. If the vectors are identical, their angle is zero, and the cosine similarity is 1, indicating maximum similarity. If the vectors are orthogonal (at 90 degrees to each other), the cosine similarity is -1, indicating no similarity [85].

In deep learning, cosine similarity is used to compare the embeddings of data points, such as RNA sequences or protein structures, that have been transformed into continuous vector spaces. These embeddings capture the semantic relationships between data points, making cosine similarity a powerful tool for measuring similarity in high-dimensional spaces [30].

Cosine similarity is particularly useful in bioinformatics for tasks such as [85]:

- **Sequence Alignment:** Comparing RNA or DNA sequences to identify regions of similarity or conservation.
- **Protein Structure Comparison:** Evaluating the similarity between protein structures based on their embeddings.
- **Gene Expression Analysis:** Comparing gene expression profiles across different conditions or time points.
- **Clustering and Classification:** Grouping similar biological samples or classifying them based on their embeddings.

Cosine similarity offers several advantages [73]:

- **Normalization:** By measuring the cosine of the angle between vectors, cosine similarity provides a normalized similarity score that is independent of the vectors' magnitudes. This is particularly useful when the scale of the vectors can vary significantly.
- **High-Dimensional Data:** Cosine similarity is well-suited for high-dimensional

spaces, where traditional distance metrics like Euclidean distance may not effectively capture the nuances of vector similarity.

- **Interpretability:** The cosine similarity score ranges from -1 to 1, with 1 indicating maximum similarity, 0 indicating no similarity, and -1 indicating maximum dissimilarity. This makes the results easy to interpret.

Cosine similarity has several important mathematical properties [85]:

- **Symmetry:** Cosine similarity is symmetric, meaning that the similarity between vectors A and B is the same as the similarity between vectors B and A:

$$\text{cosine\_similarity}(A, B) = \text{cosine\_similarity}(B, A)$$

- **Boundedness:** The cosine similarity score is bounded between -1 and 1:

$$-1 \leq \text{cosine\_similarity}(A, B) \leq 1$$

- **Non-Negativity:** When dealing with non-negative vectors (e.g., word embeddings), the cosine similarity score ranges from 0 to 1:

$$0 \leq \text{cosine\_similarity}(A, B) \leq 1$$

To calculate cosine similarity in practice, the following steps are typically followed [85]:

- **Step 1: Compute the Dot Product** Compute the dot product of the two vectors A and B:

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

- **Step 2: Compute the Norms** Compute the Euclidean norms of the two vectors:

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

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

- **Step 3: Compute the Cosine Similarity** Divide the dot product by the product of the norms:

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

In deep learning models, cosine similarity is often used to compare the embeddings generated by the network. For example, in natural language processing, word embeddings generated by models like Word2Vec or GloVe can be compared using cosine similarity to identify words with similar meanings.

Our model uses cosine similarity to compare RNA and miRNA sequence embeddings. The embedding codes from the RNA Binding Protein (RBP) sequences (Source Domain) and miRNA sequences (Target Domain) are input vectors for the cosine similarity calculation. This process results in a list indicating which RBPs have the highest likelihood of binding to the miRNA, summarized in a vector for all

trained data [32].

Consider a set of RNA sequences represented by their embeddings. We want to identify which RNA sequences are most similar to a given miRNA sequence. Let $\mathbf{v}_{\text{RNA}}$ and $\mathbf{v}_{\text{miRNA}}$ be the embeddings of an RNA sequence and a miRNA sequence, respectively. The cosine similarity between these embeddings is calculated as [85]:

$$\text{cosine\_similarity}(\mathbf{v}_{\text{RNA}}, \mathbf{v}_{\text{miRNA}}) = \frac{\mathbf{v}_{\text{RNA}} \cdot \mathbf{v}_{\text{miRNA}}}{\|\mathbf{v}_{\text{RNA}}\| \|\mathbf{v}_{\text{miRNA}}\|}$$

By computing the cosine similarity for all RNA-miRNA pairs, we can rank the RNA sequences based on their similarity to the miRNA, identifying the most likely to interact.

While cosine similarity is a powerful tool, it has some limitations. It assumes that the angle between vectors is a meaningful measure of similarity, which may not always be the case. Additionally, cosine similarity does not consider the magnitude of the vectors, which can be important in some applications [85].

Despite these limitations, cosine similarity remains a widely used metric in bioinformatics and other fields due to its simplicity and effectiveness in high-dimensional spaces.

In addition to cosine similarity, several other techniques can be used to measure similarity between vectors, including [85]:

- **Euclidean Distance:** Measures the straight-line distance between two points

in Euclidean space. It is defined as:

$$\text{Euclidean\_distance}(A, B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

- **Manhattan Distance:** Also known as L1 distance, measures the sum of the absolute differences between the components of two vectors. It is defined as:

$$\text{Manhattan\_distance}(A, B) = \sum_{i=1}^{n} |A_i - B_i|$$

- **Jaccard Similarity:** Measures the similarity between two sets by comparing the size of their intersection to the size of their union. For two sets A and B, it is defined as:

$$\text{Jaccard\_similarity}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- **Pearson Correlation Coefficient:** Measures the linear correlation between two variables, providing a normalized score that ranges from -1 to 1. For two vectors A and B, it is defined as:

$$\text{Pearson\_correlation}(A, B) = \frac{\sum_{i=1}^{n}(A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \overline{A})^2}\sqrt{\sum_{i=1}^{n}(B_i - \overline{B})^2}}$$

Cosine similarity is a fundamental metric in analyzing high-dimensional data, offering a robust measure of similarity independent of vector magnitude. Its application in bioinformatics, particularly in comparing RNA sequences and protein structures, highlights its utility in identifying patterns and relationships within complex

biological data. The mathematical foundation of cosine similarity and its practical advantages make it an essential tool in modern computational biology [73].

Autoencoding in Convolutional Neural Networks (CNN)

Autoencoders are neural networks used to learn efficient codings of input data in an unsupervised manner. They compress the input into an informative representation and then decode it to reconstruct the original input. This process forces the network to learn the most salient features of the data. Convolutional Neural Networks (CNNs) are highly effective for analyzing spatial data, making them suitable for image recognition and processing tasks. Our model uses CNNs to process protein sequences (PSSM) and protein structure contact maps. Critical components of CNNs include convolution layers, pooling layers, and fully connected layers [81].

- **Convolution Layers:** Convolution layers apply filters (kernels) to the input data to extract features. Each filter slides over the input data, performing element-wise multiplications and summing the results to produce a feature map. Mathematically, the convolution operation for a filter F and input I is given by:

$$(I * F)(x, y) = \sum_i \sum_j I(x + i, y + j) \cdot F(i, j)$$

  where (x, y) are the coordinates of the input, and (i, j) are the coordinates of the filter. This operation captures local patterns in the input data, which is crucial for understanding spatial hierarchies.

- **Max Pooling Layers:** Max pooling layers reduce the spatial dimensions of

the input, helping to reduce computational load and control overfitting. The max pooling operation selects the maximum value from each sub-region of the input data. For a pooling window of size $k \times k$, the max pooling operation is defined as:

$$P(x, y) = \max_{i=0}^{k-1} \max_{j=0}^{k-1} I(x + i, y + j)$$

This reduction process helps retain the most significant features while discarding redundant information, making the model more robust and less prone to overfitting.

- **When to Use LSTM vs. CNN:** LSTM networks and CNNs serve different purposes. LSTMs are effective for sequential data, such as time-series analysis and natural language processing, because they capture temporal dependencies. Conversely, CNNs are ideal for spatial data, commonly used in image and video processing tasks due to their ability to detect spatial hierarchies [56]. In our model, LSTMs process RNA sequences due to their sequential nature, while CNNs process protein sequences (PSSM) and protein structure contact maps, which are inherently spatial.

- **Encoder in CNN:** Unlike LSTM-based models, CNNs do not utilize embedding layers. Instead, they use encoders to transform input data into compact, informative representations. An encoder in a CNN typically consists of several convolution and pooling layers, followed by fully connected layers that compress the data into a lower-dimensional space. Mathematically, the encoder function

E can be represented as:

$$E(x) = \sigma(W \cdot x + b)$$

where x is the input, W are the weights, b is the bias, and σ is the activation function. In our model, the encodings obtained from the PSSM and contact map inputs are concatenated to form a unique 256-character vector for each protein. This concatenated encoding is then used to calculate cosine similarity between pairs of proteins, enabling the identification of proteins similar to the candidates identified in the Source Domain.

- **Comparison of Encoder and Embedding Layer:** The encoder in a CNN and the embedding layer in an LSTM serve similar purposes but operate differently. The embedding layer maps discrete input data to a dense vector space, capturing semantic relationships through learned embeddings. In contrast, the encoder in a CNN transforms input data through multiple layers of convolutions and pooling, capturing spatial features and compressing the data into a compact representation. Both methods result in dense, informative vectors that can be used for similarity calculations.

- **Autoencoders in CNNs:** Autoencoders can be integrated into CNNs to enhance feature learning further, as depicted in Figure 3.2. A CNN-based autoencoder consists of an encoder, which reduces the input dimensions, and a decoder, which reconstructs the input from the reduced representation. This ar-

chitecture is useful for denoising, anomaly detection, and unsupervised feature learning tasks. The encoder part of the autoencoder in a CNN can be described as follows:

$$h = f(x) = \sigma(W \cdot x + b)$$

where h is the hidden representation, W are the weights, b is the bias, and σ is the activation function. The decoder reconstructs the input as:

$$\hat{x} = g(h) = \sigma(W' \cdot h + b')$$

where $\hat{x}$ is the reconstructed input, W' are the weights, b' is the bias, and σ is the activation function. The training objective is to minimize the reconstruction error, typically measured using mean squared error:

$$L(x, \hat{x}) = \|x - \hat{x}\|^2$$

Autoencoders and CNNs are powerful tools in bioinformatics for processing and analyzing complex biological data. They provide a robust framework for capturing spatial and sequential features, enabling more accurate and insightful analysis of RNA sequences, protein structures, and other biological data [48, 51].

3.3.2 Model Architecture

The DeepmiRPB model, meticulously designed to predict miRNA-protein interactions, exemplifies advanced computational biology. This multimodal deep neural

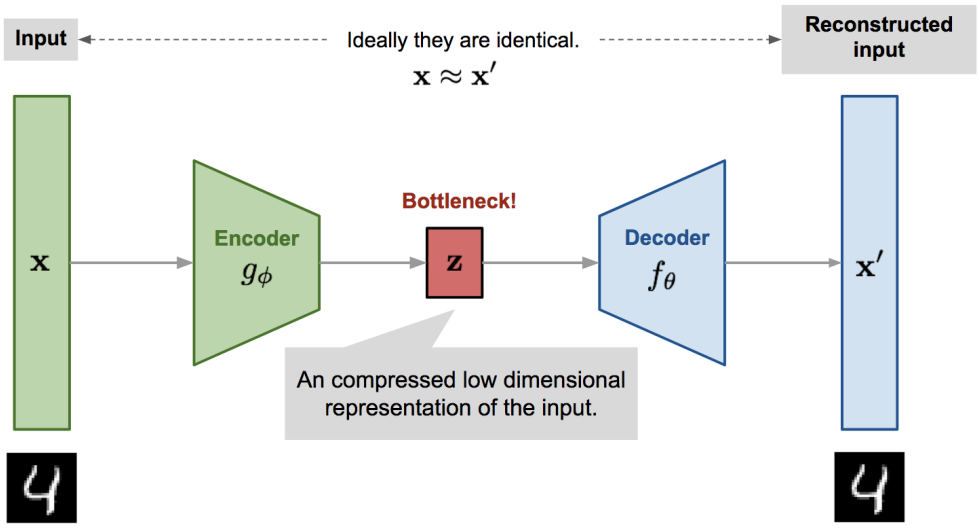

Figure 3.2: The encoder and decoder can take various forms depending on our use case, such as feedforward neural networks. In the figure above, x represents the input data, z is the compressed feature vector, and x' is the regenerated input.

network integrates thermodynamic and structural information, such as the secondary structure (ss) context of (mi)RNA and the residue contact of RNA-binding proteins (RBPs), into its analysis alongside their sequences. The model is bifurcated into two primary components: the first component focuses on finding similarities between miRNA and RNA sequences. In contrast, the second component identifies similarities between RBPs and other proteins.

The **first part** of the DeepmiRPB model, depicted in Figure 3.3.a, is dedicated to identifying which RNA sequences bind to RBPs similar to the given miRNA. This component comprises two main sections: the Source and Target Domains.

- **Source Domain:** RNA sequences that bind to RBPs are used to train the model in the Source Domain. This extensive training dataset, containing approximately 120,000 sequences labeled with binding information, enables the

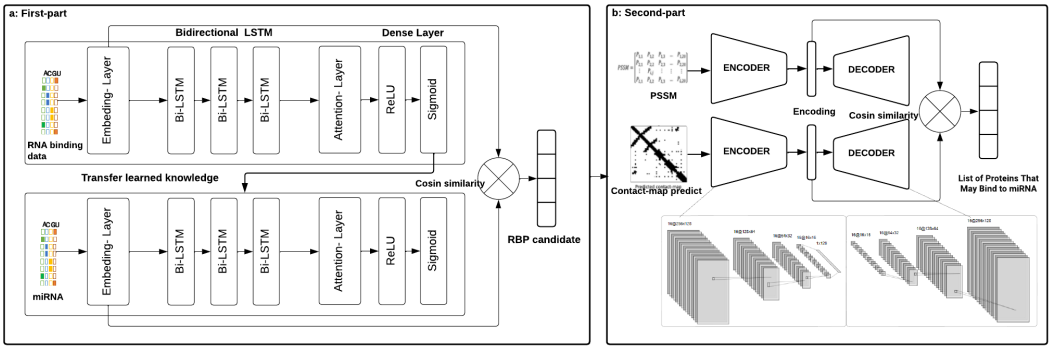

Figure 3.3: Schematic diagram of the proposed DeepmiRPB architecture for predicting microRNA-protein interactions. (a) First part architecture: This part trains on RNA sequences that bind to RNA-binding proteins (RBPs) to learn intricate features of RNA-protein interactions. The knowledge gained is then transferred to the target domain, where miRNA sequences are input, embedding codes are generated, and cosine similarity is employed to identify RNA sequences most similar to the miRNA sequences. (b) Second part architecture: This part processes the Position-Specific Scoring Matrix (PSSM) and contact maps for each RBP candidate identified in the first part. Convolutional Neural Networks (CNN) and max-pooling layers encode these matrices. Cosine similarity is then calculated to compare RBP candidates with other proteins, resulting in a matrix identifying proteins with a higher likelihood of binding to the miRNA sequence.

model to learn the intricate features of RNA-protein interactions. Once trained, the learned knowledge, including all weights and parameters, is transferred to the Target Domain.

- **Target Domain:** The Target Domain leverages the transferred knowledge from the Source Domain to refine prediction capabilities further. Here, miRNA sequences from the Argonaute (AGO) protein family are input, and the model generates embedding codes for each miRNA sequence using the pre-trained weights and parameters from the Source Domain.

- **Similarity Calculation:** The primary goal is to identify which RNA sequences

that bind to RBPs are most similar to the miRNA sequences. This is achieved by calculating the cosine similarity between the embedding vectors obtained from the RNA and miRNA sequences. The resulting list ranks the RBPs based on their similarity to the miRNA sequences, providing valuable insights into potential miRNA-protein interactions.

In summary, the first part of the DeepmiRPB model efficiently utilizes transfer learning and cosine similarity to identify and rank RNA-binding proteins likely to interact with miRNA sequences, thereby enhancing our understanding of miRNA-protein binding mechanisms.

The **second part**, illustrated in Figure 3.3.b, focuses on finding similarities between proteins. We obtained a list of RBP candidates from the first part that could bind to miRNAs. Given that our training data does not encompass all RNA-binding proteins, a given miRNA might bind to an untrained protein. We aim to identify proteins similar to the candidate list obtained from the first part to address this.

- We utilize the Position-Specific Scoring Matrix (PSSM) and protein structure contact maps for each protein in the candidate list from the first part to find the similarity between the two proteins.

- These inputs are separately fed into an Autoencoder CNN comprising several convolution and pooling layers. Encoding codes for PSSM and contact maps are generated separately for each protein and then concatenated into a 256-character unique vector.

- Cosine similarity is then used to calculate the similarity between each pair of proteins, resulting in an $n \times n$ matrix for n proteins. This matrix helps identify the protein most similar to the high-chance candidate identified in the first part.

By leveraging these techniques, we enhance our model's ability to predict miRNA-protein interactions, even when the target proteins were not part of the initial training dataset. This comprehensive approach ensures we can accurately determine potential miRNA-binding proteins based on similarity scores derived from structural and sequence data.

In summary, our model operates based on the following steps:

- Identifying which RNA sequences bind to RBPs are more similar to the miRNA sequences.
- Generating a list of RBP candidates likely to bind to miRNA.
- Determining which proteins are similar to the RBPs in our candidate list.

By following these steps, our model ensures a comprehensive analysis of miRNA-protein interactions. Initially, it leverages transfer learning and cosine similarity to identify RNA sequences similar to the miRNA, providing a list of potential RBP candidates. Subsequently, it utilizes the Position-Specific Scoring Matrix (PSSM) and protein structure contact maps to find proteins similar to these RBP candidates, using convolutional neural networks and cosine similarity. This dual approach allows us to accurately predict miRNA-binding proteins without complete training data for all RNA-binding proteins.

The subsequent subsections will delve deeper into the methodologies employed, detailing the processes and techniques used in each part of the model.

First Part: Technical Overview

The first part of the DeepmiRPB model is a crucial component designed to predict RNA-protein interactions by capturing complex patterns in RNA sequences that facilitate binding with proteins. It consists of two primary sections: the Source Domain and the Target Domain, as illustrated in Figure 3.3.a. This part of the model includes several key components: embedding layers, bidirectional Long Short-Term Memory (LSTM) layers, attention mechanisms, and dense layers.

- **Source Domain:** The process begins in the Source Domain by embedding RNA sequences into a numerical format suitable for neural network processing. The embedding layer transforms the input RNA sequences into fixed-sized dense vectors, where each nucleotide in the RNA sequence is mapped to a unique vector representation. This facilitates the capture of semantic relationships between different sequences. Following the embedding layer, multiple bidirectional LSTM layers capture sequential dependencies in the RNA sequences from both forward and backward directions. These layers are crucial for understanding the sequence context, allowing the model to consider information from upstream and downstream of a given nucleotide. Dropout layers are incorporated to prevent overfitting by randomly setting a fraction of input units to zero during training. An attention mechanism is then applied to enable the model to focus on the

most relevant parts of the sequence, enhancing its ability to capture critical features necessary for binding prediction. The output from the attention layer is directed into dense layers that further process the information and produce the final binding prediction through a sigmoid activation function. This final layer outputs a probability score indicating the likelihood of binding between the RNA sequence and the protein.

- **Target Domain:** After training the Source Domain, the learned knowledge, including weights and parameters, is transferred to the Target Domain through transfer learning techniques. In the Target Domain, miRNA sequences are input, and all the layers discussed in the Source Domain are identically applied. To align with the RNA sequences, miRNA sequences are converted to their complementary sequences before being fed into the Target Domain. This conversion ensures the model can effectively identify RNA sequences similar to the miRNA.

- **Sampling Methodology:** After completing the initial training of the first model component, evaluating the similarity between RBPs and miRNAs became crucial. Cosine similarity was chosen for this purpose. However, due to the large number of sequences associated with each RBP, computing the similarity with every RBP-binding RNA sequence posed significant computational challenges. To mitigate this issue, we implemented a sampling distribution approach for each RBP sequence.

  The sampling distribution ensures that we capture the essential characteristics

and variability of the entire dataset without the need for exhaustive computations. Specifically, we randomly sampled 1000 RNA sequences or binding sites for each RBP. This sampling strategy maintains the robustness and representativeness of our model. The Central Limit Theorem (CLT) underpins this approach, stating that the distribution of sample means approximates a normal distribution when the sample size is sufficiently large. Mathematically, 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. For our study, we determined that a sample size of 1000 is adequate to approximate normality and capture the variability and underlying patterns in the data. This sample size provides a representative subset for training purposes. By employing a sampling distribution and selecting multiple random samples of 1000 for each RBP sequence, we ensured that our model trained on a robust dataset that faithfully reflects the underlying data distribution. This approach streamlined the computation process and preserved the statistical integrity of the dataset. Following inference on all the samples from RBPs, we computed the similarity between the sampled RBP binding site sequences and miRNA. This method significantly reduced computational overhead while maintaining the accuracy and effectiveness of our similarity assessments.

- **Similarity Calculation:** The primary goal of the first part is to identify

which RNA sequences are similar to the miRNA using cosine similarity. Cosine similarity is a measure to determine the similarity between two non-zero vectors in an inner product space. It is defined as the cosine of the angle between the vectors, providing a metric to assess the degree of similarity between two sets of embeddings. This measure is particularly useful in high-dimensional spaces where traditional Euclidean distance may not effectively capture the nuances of vector similarity.

The embedding codes from the RBP sequences (Source Domain) and miRNA sequences (Target Domain) are input vectors for the cosine similarity calculation. This process results in a list indicating which RBPs have the highest likelihood of binding to the miRNA, summarized in a vector for all trained data. The core concept behind cosine similarity is to measure the cosine of the angle between two vectors. If the vectors are identical, their angle is zero, and the cosine similarity is 1, indicating maximum similarity. If the vectors are orthogonal (at 90 degrees to each other), the cosine similarity is 0, indicating no similarity. This angular measure provides a normalized similarity score independent of the vectors' magnitudes, making it particularly useful in applications where the scale of the vectors can vary significantly.

- **The Embedding Layer in LSTM Deep Learning Models:** In natural language processing (NLP) and sequence analysis, the embedding layer is a crucial component, especially within Long Short-Term Memory (LSTM) deep learning models. This layer transforms discrete, categorical input data, such

as RNA and miRNA sequences, into fixed-size dense vectors, facilitating the learning process in subsequent model layers.

In a continuous vector space, the embedding layer represents each unique item, like a character or word. These vectors, learned during training, capture semantic relationships between items, making it possible to measure distances or similarities between them. Combining LSTM with the Attention mechanism provides a powerful tool for analyzing sequential data in computational biology. This architecture's ability to remember long-term dependencies and focus on pertinent information makes it especially suited for our project, surpassing the capabilities of Convolutional Neural Networks (CNNs) in handling the sequential nature and complex dependencies inherent in RNA sequences and protein structures [80].

In summary, the first part of the DeepmiRPB model utilizes advanced neural network components and transfer learning to accurately identify and rank RNA-binding proteins that are likely to interact with miRNA sequences. The model narrows potential RNA-protein interactions by embedding RNA sequences, capturing sequential dependencies through bidirectional LSTM layers, applying attention mechanisms, and leveraging cosine similarity. Integrating the Central Limit Theorem ensures that taking multiple random samples with 1000 sequences per RBP is adequate for robust training, providing a solid foundation for the subsequent steps in the model. The following subsections will delve deeper into the methodologies employed, detailing the processes and techniques used in each part of the model.

Second Part: Technical Overview

The second part of the DeepmiRPB model is designed to find similarities between proteins based on the Position-Specific Scoring Matrix (PSSM) and predicted contact maps from ResPRE, as illustrated in Figure 3.3.b. This component identifies which proteins are most likely to bind to the miRNAs, given the first part's RNA-binding protein (RBP) candidates. The second part employs several vital components: encoders, convolutional neural networks (CNNs), and max-pooling layers.

- **Autoencoders:** Autoencoders are neural networks used to learn efficient codings of input data in an unsupervised manner. They compress the input into an informative representation and then decode it to reconstruct the original input. This process forces the network to learn the most salient features of the data.

- **Convolutional Neural Networks (CNNs):** CNNs are highly effective for analyzing spatial data, making them suitable for image recognition and processing tasks. Our model uses CNNs to process protein sequences (PSSM) and protein structure contact maps. Critical components of CNNs include convolution layers, pooling layers, and fully connected layers [81].

- **Convolution Layers:** Convolution layers apply filters (kernels) to the input data to extract features. Each filter slides over the input data, performing element-wise multiplications and summing the results to produce a feature map.

- **When to Use LSTM vs. CNN:** LSTM networks and CNNs serve different purposes. LSTMs are effective for sequential data, such as time-series analysis

and natural language processing, because they capture temporal dependencies. Conversely, CNNs are ideal for spatial data, commonly used in image and video processing tasks due to their ability to detect spatial hierarchies [56]. In our model, LSTMs process RNA sequences due to their sequential nature, while CNNs process protein sequences (PSSM) and protein structure contact maps, which are inherently spatial.

- **Encoder in CNN:** Unlike LSTM-based models, CNNs do not utilize embedding layers. Instead, they use encoders to transform input data into compact, informative representations. An encoder in a CNN typically consists of several convolution and pooling layers, followed by fully connected layers that compress the data into a lower-dimensional space.

- **Comparison of Encoder and Embedding Layer:** The encoder in a CNN and the embedding layer in an LSTM serve similar purposes but operate differently. The embedding layer maps discrete input data to a dense vector space, capturing semantic relationships through learned embeddings. In contrast, the encoder in a CNN transforms input data through multiple layers of convolutions and pooling, capturing spatial features and compressing the data into a compact representation. Both methods result in dense, informative vectors that can be used for similarity calculations.

In the second part of our model, we utilize these techniques to predict miRNA-protein interactions by finding similarities between proteins based on PSSM and contact map data. By leveraging the strengths of CNNs and encoders, we ensure a robust

and accurate identification of potential miRNA-binding proteins, even if the target proteins were not part of the initial training dataset.

3.3.3 Selection of Model Architecture and Hyperparameter Optimization

The selection of the optimal model architecture and the meticulous process of hyperparameter optimization are pivotal in developing an effective deep-learning model for miRNA-protein interaction prediction. Given the varied nature of our input data—RNA sequences, miRNA sequences, Position-Specific Scoring Matrices (PSSM), and protein structure contact maps—we adopted specific architectures tailored to each data type. LSTM networks were chosen for RNA and miRNA sequences due to their strength in capturing long-term dependencies in sequential data. For PSSM and contact map data, CNNs were employed to leverage their ability to extract hierarchical features from spatial data.

Initially, we explored over 45 different model architectures to identify the most effective configuration. These architectures included:

- Unidirectional LSTM networks
- Bidirectional LSTM networks
- Pure CNN architectures
- Hybrid models combining LSTM and CNN
- Models incorporating attention mechanisms
- Architectures with varying layers and configurations of LSTM and CNN

Each model was evaluated based on its performance in predicting miRNA-protein interactions, with metrics such as accuracy, precision, recall, and F1 score being the primary indicators of success. The goal was to find an architecture that provided high accuracy and maintained robustness across different types of input data.

After identifying the best-performing architecture, we moved on to hyperparameter optimization. This process involved systematically varying the key hyperparameters to fine-tune the model's performance. The hyperparameters considered were:

- Embedding dimensions: [16, 32, 64, 128, 256, 512,1024]
- LSTM units: [16, 32, 64, 128, 256, 512,1024]
- Dropout rates: [0.2, 0.5]
- Batch sizes: [32, 64, 128]
- Learning rates: [0.001, 0.0001]

Each combination of these hyperparameters was tested, considering one parameter at a time while keeping the others constant. For instance, we began with embedding dimensions set to 16 and LSTM units set to 16, then adjusted the other parameters, such as dropout rates, batch sizes, and learning rates, in sequence. This methodical approach ensured a thorough exploration of the hyperparameter space and identified the optimal configuration.

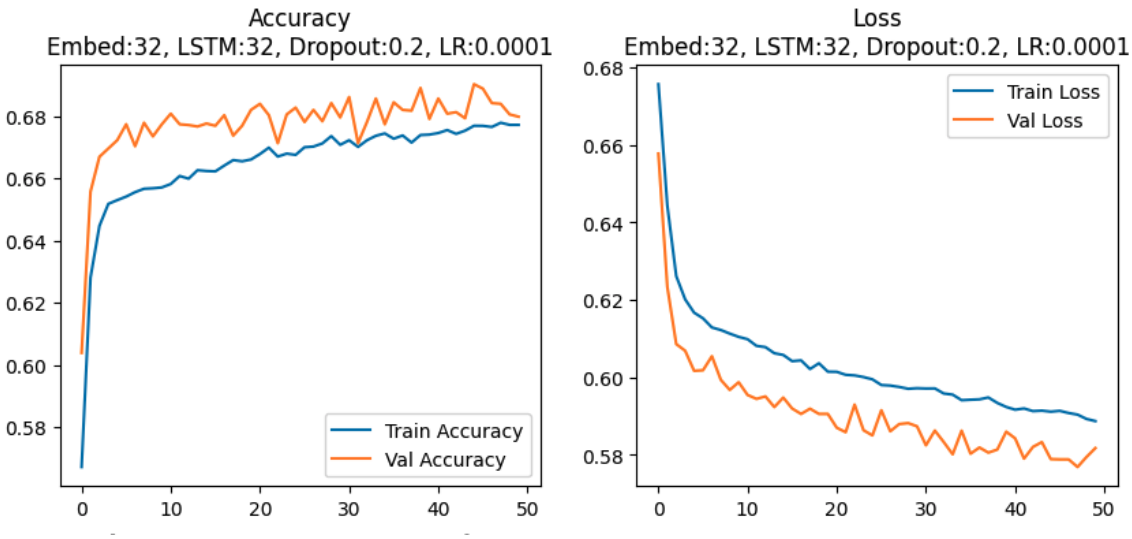

Figure 3.4: Accuracy and loss charts for various hyperparameter configurations.

Each hyperparameter configuration was subjected to a training regimen of 50 epochs, each taking approximately 18.6 hours for one of the tests we did in Figure3.10. This extensive experimentation provided a comprehensive understanding of how different hyperparameters influenced model performance. The optimization results, including accuracy and loss charts, are presented in Figure 3.4, demonstrating the impact of each parameter configuration on the model's predictive capabilities.

Hyperparameter optimization is a crucial step in deep learning, as it significantly influences the model's ability to generalize from training data to unseen data. Properly selected hyperparameters can substantially improve model performance, reducing overfitting and enhancing the model's ability to capture intricate patterns in the data.

We developed a robust and accurate model for miRNA-protein interaction prediction through this rigorous architecture selection process and hyperparameter

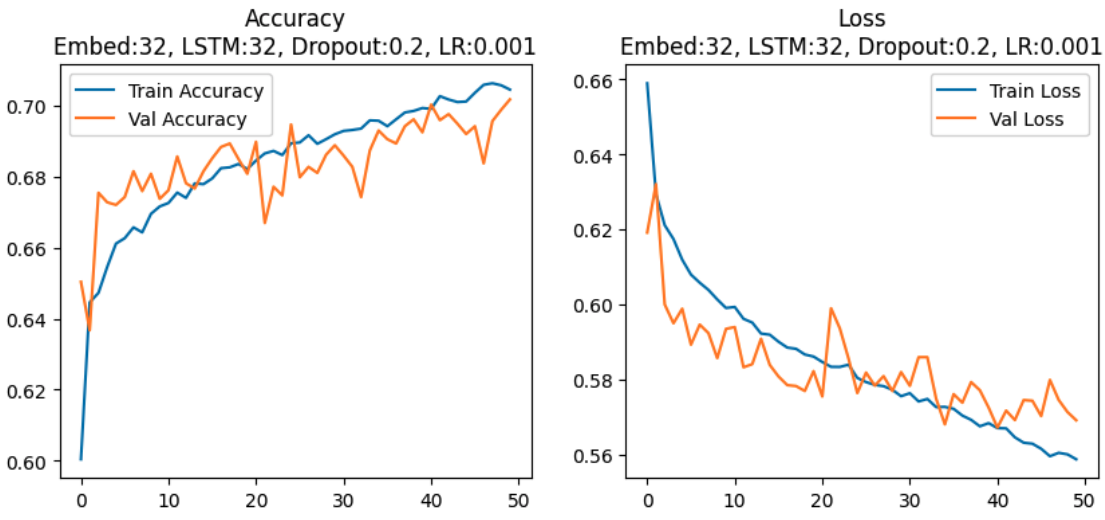

Figure 3.5: Accuracy and loss charts for various hyperparameter configurations.

optimization. This model leverages the strengths of LSTM and CNN architectures, fine-tuned through exhaustive experimentation, to provide reliable predictions crucial for understanding the complex mechanisms underlying miRNA-protein interactions.

3.3.4 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. 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

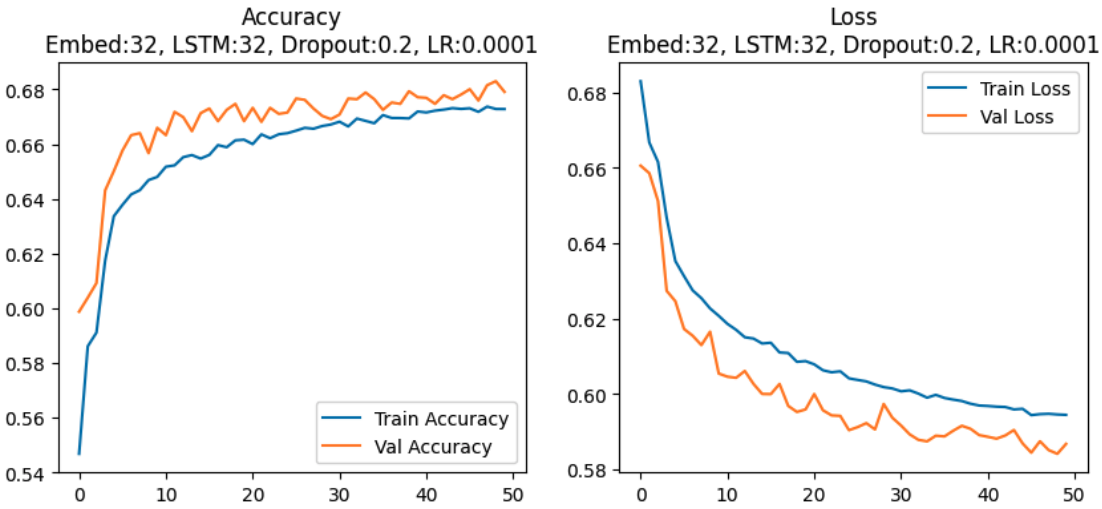

Figure 3.6: Accuracy and loss charts for various hyperparameter configurations.

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.

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

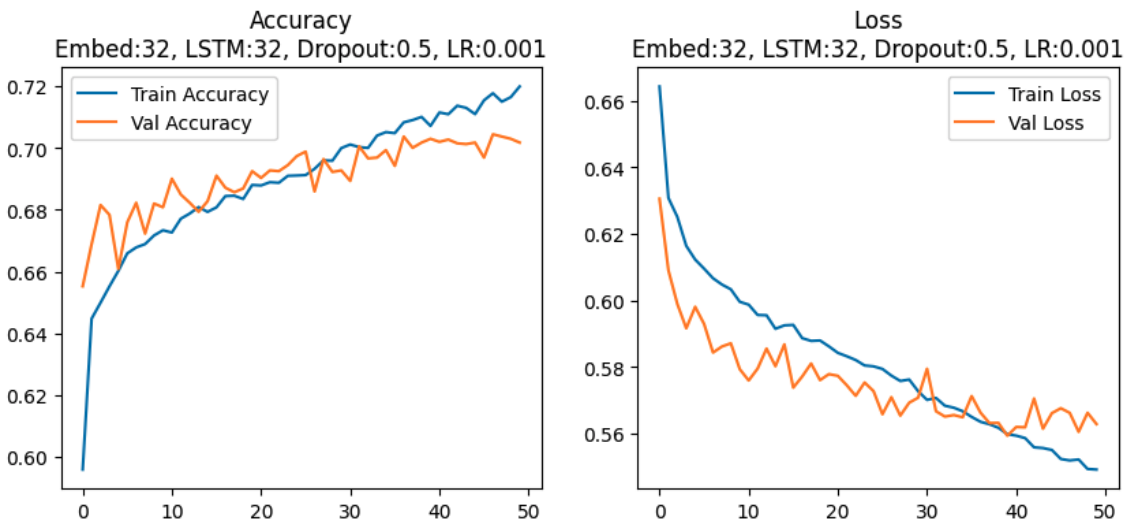

Figure 3.7: Accuracy and loss charts for various hyperparameter configurations.

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

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

3.3.5 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

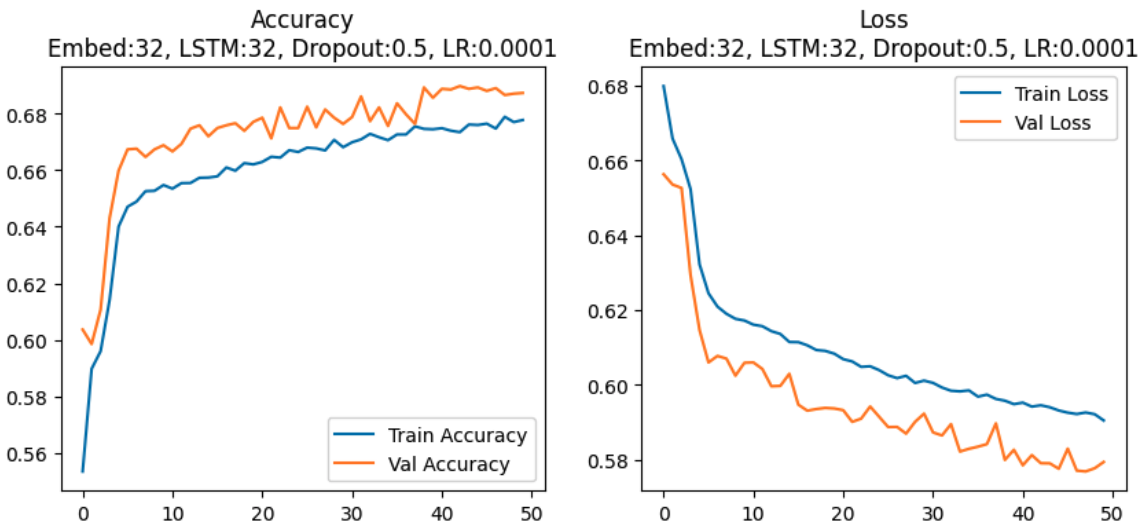

Figure 3.8: Accuracy and loss charts for various hyperparameter configurations.

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. [53]. 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

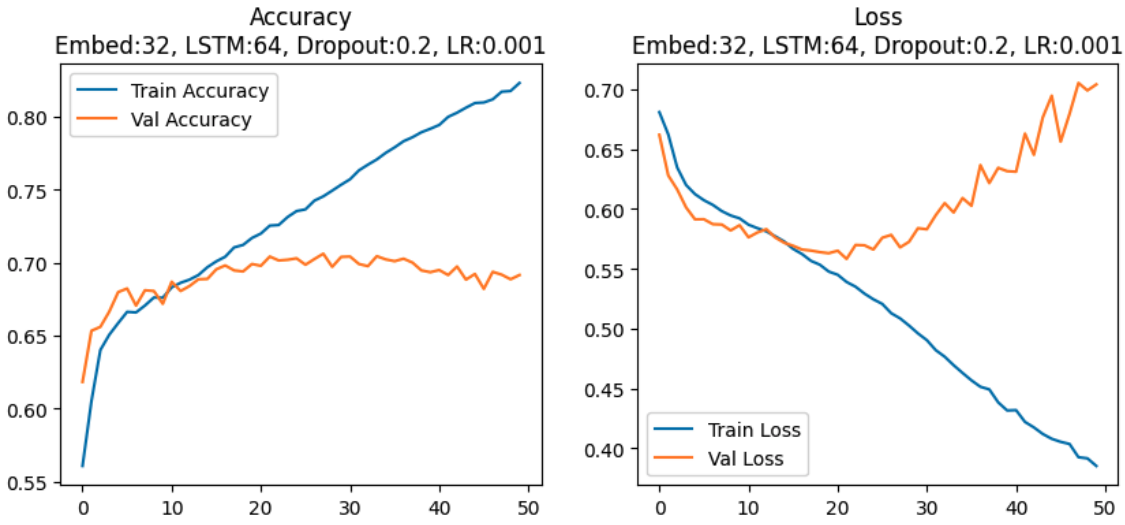

Figure 3.9: time for each epoch takes around 50 minutes

its tumor-suppressive properties. According to Johnson et al. [43], the miR-let-7 represses cell proliferation pathways in human cells, highlighting its potential as a therapeutic target. Furthermore, Nwaeburu et al. [59] 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 [15] 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 perfor-

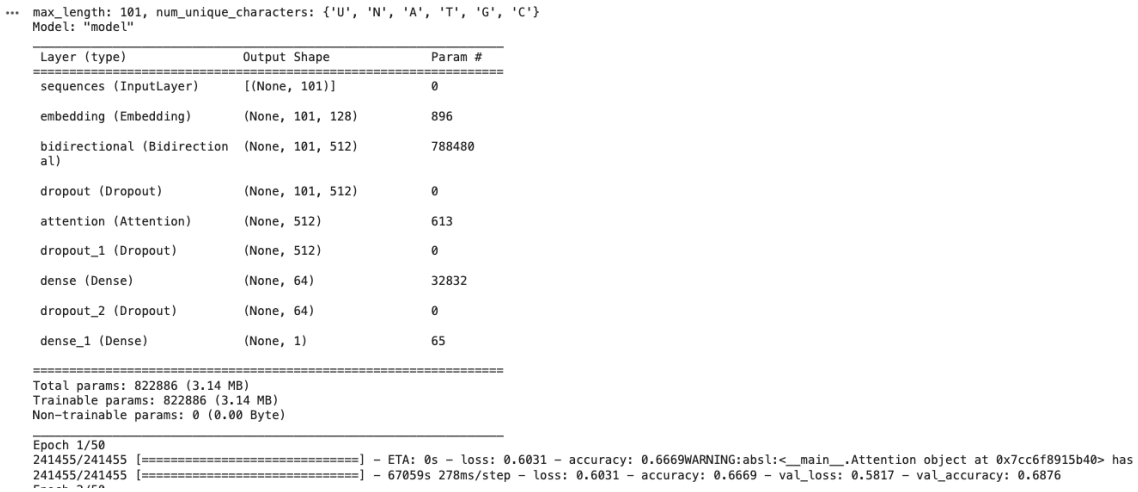

Figure 3.10: Accuracy and loss charts for various hyperparameter configurations.

mance evaluation and its implications for predicting miRNA-protein interactions.

3.4 Results

3.4.1 Model Performance

After training, the comprehensive evaluation of the source and target domains, summarized in Table 3.1, 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 3.1), 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. 3.11) 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 3.1 and offers deeper insight into the model's prediction accuracy.

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

Table 3.1: Performance metrics for source and target models

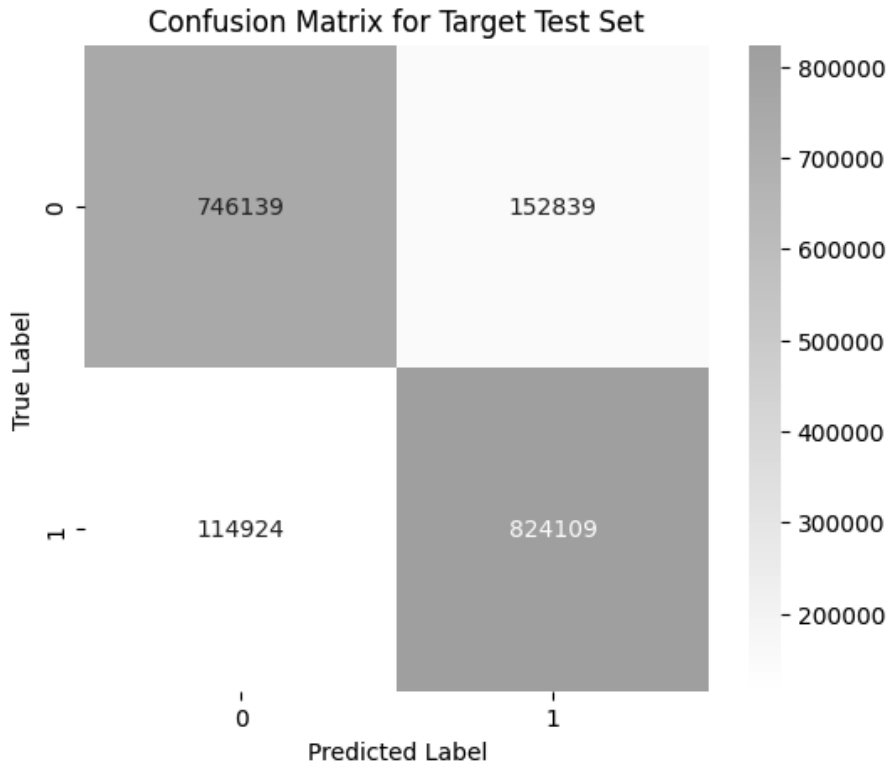

Figure 3.11: Confusion matrix for test data in the source domain

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 domains validate the model's reliability. Subsequent sections will present the results from the three case studies, further illustrating the model's application and performance in real-world scenarios.

3.4.2 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 the model with several miRNAs and experimental data to ensure its robustness and accuracy.

- **miR-451:** According to Dueck et al. [22], 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 3.2, 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 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. [22], 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 the table 3.3, 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.

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 3.2: Top 10 RBPs with highest scores for miR-451.

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

Table 3.3: Top RBPs with highest scores for miR-19b, miR-23a, and miR-21.

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 3.4: Top RBPs with highest scores for miR-223.

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

Table 3.5: Top RBPs with highest scores for let-7d.

3.4.3 Validation on miR-223 (Case Study 2)

We used miR-223 [15] 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 [53], 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 3.4 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 3.6 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.

	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

Table 3.6: Cosine similarity matrix for final candidate proteins for miR-223 sorting.

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.

3.4.4 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 [58] and pancreatic cancer cells [83]. 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 3.5.

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 [9], 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.

3.5 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 miRNA-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 [14]. The identification

of 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 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 not only demonstrated effectiveness in elucidating the complex interplay between miRNAs and proteins but also 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. [57].

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 [61].

Integrating multi-omic data, including genomics, transcriptomics, and proteomics, is expected further to refine the predictive accuracy of models like DeepmiRBP. 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 our understanding of miRNA-protein interactions and their implications in disease contexts further.

3.6 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.

Chapter 4

Conclusion

This dissertation has explored the complex and intricate interactions between microRNAs (miRNAs) and RNA-binding proteins (RBPs), emphasizing the significance of these interactions in gene regulation and disease progression. The research presented herein underscores the transformative potential of advanced computational approaches, particularly deep learning models, in unraveling these biological processes.

4.1 Summary of Findings

The primary objective of this research was to develop and evaluate a robust computational model for predicting miRNA-protein interactions, leveraging the strengths of deep learning architectures. The proposed DeepmiRPB model, incorporating Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), transfer learning, and cosine similarity, has demonstrated exceptional precision in identifying miRNA-protein binding sites. This model's innovative design allows it to capture the nuanced dependencies and structural information within miRNA and protein sequences, facilitating accurate predictions even in limited data

availability.

The efficacy of the DeepmiRPB model was validated through rigorous testing against diverse datasets, distinct proteins characterized by various input data modalities. The model achieved high accuracy, precision, recall, and F1 scores, showcasing its robustness and reliability in predicting RNA-protein interactions. Furthermore, the model's performance in identifying miRNA interactions with key proteins such as AGO1 and AGO2 highlights its potential in elucidating regulatory mechanisms in diseases like cancer.

4.2 Implications for Bioinformatics and Molecular Biology

The insights gained from this research have profound implications for bioinformatics and molecular biology. Integrating multi-omic data and employing sophisticated computational techniques, the DeepmiRPB model offers a scalable and adaptable template for future studies across various biological contexts. This approach enhances our understanding of miRNA-mediated regulatory networks and paves the way for developing novel therapeutic interventions and personalized medicine.

The success of the DeepmiRPB model in capturing the complexities of miRNA-protein interactions underscores the potential of deep learning in bioinformatics. This research contributes to a richer understanding of gene regulation mechanisms, providing a foundation for future studies to decipher the molecular drivers of complex diseases. The model's adaptability and precision make it a valuable tool for exploring the regulatory roles of miRNAs across different cancer stages and other disease

pathologies.

4.3 Future Directions

The journey of computational modeling in miRNA-protein interactions is far from complete. The future of this research trajectory promises to be multifaceted, reinforcing the alliance between computational predictions and experimental validations. Future endeavors should focus on the following areas:

- **Expanding Dataset Diversity:** Incorporating more diverse and comprehensive datasets, including various miRNA and protein sequences across different species and disease states, will enhance the model's generalizability and predictive accuracy.

- **Integrating Multi-Omic Data:** Further integrating genomics, transcriptomics, proteomics, and metabolomics data will provide a holistic view of miRNA-protein interactions, capturing the full complexity of regulatory networks.

- **Model Optimization:** Continued optimization of the DeepmiRPB model, including fine-tuning hyperparameters and exploring alternative deep learning architectures, will improve its performance and applicability.

- **Experimental Validation:** Collaborating with experimental biologists to validate the model's predictions through laboratory experiments will strengthen the reliability of computational findings and foster translational applications in precision medicine.

- **Exploring New Biological Contexts:** Applying the DeepmiRPB model to investigate miRNA-protein interactions in other biological processes, such as developmental biology, immune responses, and aging, will broaden our understanding of these critical regulatory mechanisms.

In conclusion, this dissertation has demonstrated the significant advances that can be achieved in bioinformatics by integrating deep learning techniques and multi-omic data. The DeepmiRPB model stands as a testament to the power of computational approaches in elucidating complex biological processes, offering a promising pathway toward personalized medical solutions. The insights gained from this research will inform the development of novel therapeutic interventions, carving a path toward an era of precision medicine.

The future of bioinformatics is poised for a revolution driven by the versatility and adaptability of deep learning models like DeepmiRPB. As we unravel the complexities of miRNA-protein interactions, the stage is set for scientific breakthroughs that may redefine contemporary medical science's contours. This research lays a solid foundation for these future explorations, promising to enhance our molecular comprehension and improve patient care outcomes.

Bibliography

[1] Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 45(D1): D158–D169, 2017.

[2] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[3] José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

[4] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25 (17):3389–3402, 1997.

[5] Victor Ambros. The functions of animal micrornas. *Nature*, 431(7006):350–355, 2004.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine trans-

lation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[7] David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *cell*, 116(2):281–297, 2004.

[8] David P Bartel. Micrornas: target recognition and regulatory functions. *cell*, 136(2):215–233, 2009.

[9] Michelle Demory Beckler, James N Higginbotham, Jeffrey L Franklin, Amy-Joan Ham, Patrick J Halvey, Imade E Imasuen, Corbin Whitwell, Ming Li, Daniel C Liebler, and Robert J Coffey. Proteomic analysis of exosomes from mutant kras colon cancer cells identifies intercellular transfer of mutant kras. *Molecular & cellular proteomics*, 12(2):343–355, 2013.

[10] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36, 2012.

[11] Vittoria Borgonetti, Elisabetta Coppi, and Nicoletta Galeotti. 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. *International Journal of Molecular Sciences*, 22(19):10394, 2021.

[12] Stevo Bozinovski. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica*, 44(3), 2020.

[13] Ulf Brefeld, Edward Curry, Elizabeth Daly, Brian MacNamee, Alice Marascu, Fabio Pinelli, Michele Berlingerio, and Neil Hurley. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part III*, volume 11053. Springer, 2019.

[14] A Maxwell Burroughs, Yoshinari Ando, Michiel JL de Hoon, Yasuhiro Tomaru, Takahiro Nishibu, Ryo Ukekawa, Taku Funakoshi, Tsutomu Kurokawa, Harukazu Suzuki, Yoshihide Hayashizaki, et al. A comprehensive survey of 3 animal mirna modification events and a possible role for 3 adenylation in modulating mirna targeting effectiveness. *Genome research*, 20(10):1398–1410, 2010.

[15] Hua-chang Chen, Jing Wang, Robert J Coffey, James G Patton, Alissa M Weaver, Yu Shyr, and Qi Liu. Evpsort: An atlas of small ncrna profiling and sorting in extracellular vesicles and particles. *Journal of Molecular Biology*, page 168571, 2024.

[16] Sung Wook Chi, Julie B Zang, Aldo Mele, and Robert B Darnell. Argonaute hits-clip decodes microrna–mrna interaction maps. *Nature*, 460(7254):479–486, 2009.

[17] ENCODE Project Consortium. A user's guide to the encyclopedia of dna elements (encode). *PLoS biology*, 9(4):e1001046, 2011.

[18] Juan Cui and Jiang Shu. Circulating microrna trafficking and regulation: com-

putational principles and practice. *Briefings in bioinformatics*, 21(4):1313–1326, 2020.

[19] Arundhati Das, Tanvi Sinha, Sharmishtha Shyamal, and Amaresh Chandra Panda. Emerging role of circular rna–protein interactions. *Non-coding RNA*, 7(3):48, 2021.

[20] Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, et al. The encyclopedia of dna elements (encode): data portal update. *Nucleic acids research*, 46(D1):D794–D801, 2018.

[21] Lei Deng, Youzhi Liu, Yechuan Shi, and Hui Liu. A deep neural network approach using distributed representations of rna sequence and structure for identifying binding site of rna-binding proteins. In *2019 Ieee International Conference on Bioinformatics and Biomedicine (Bibm)*, pages 12–17. IEEE, 2019.

[22] Anne Dueck, Christian Ziegler, Alexander Eichner, Eugene Berezikov, and Gunter Meister. microrna associated with the different human argonaute proteins. *Nucleic acids research*, 40(19):9850–9862, 2012.

[23] Fabrizio Ferre, Alessio Colantoni, and Manuela Helmer-Citterich. Revealing protein–lncrna interaction. *Briefings in bioinformatics*, 17(1):106–116, 2016.

[24] Witold Filipowicz, Suvendra N Bhattacharyya, and Nahum Sonenberg. Mechanisms of post-transcriptional regulation by microrna: are the answers in sight? *Nature reviews genetics*, 9(2):102–114, 2008.

[25] Miranda Y Fong, Weiying Zhou, Liang Liu, Aileen Y Alontaga, Manasa Chandra, Jonathan Ashby, Amy Chow, Sean Timothy Francis O'Connor, Shasha Li, Andrew R Chin, et al. Breast-cancer-secreted mir-122 reprograms glucose metabolism in premetastatic niche to promote metastasis. *Nature cell biology*, 17(2):183–194, 2015.

[26] Tian Gao, Jiang Shu, and Juan Cui. A systematic approach to rna-associated motif discovery. *BMC genomics*, 19:1–17, 2018.

[27] Ruben Garcia-Martin, Guoxiao Wang, Bruna B Brandão, Tamires M Zanotto, Samah Shah, Sandip Kumar Patel, Birgit Schilling, and C Ronald Kahn. Microrna sequence codes for small extracellular vesicle release and cellular retention. *Nature*, 601(7893):446–451, 2022.

[28] Joseph Gomes, Bharath Ramsundar, Evan N Feinberg, and Vijay S Pande. Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv preprint arXiv:1703.10603*, 2017.

[29] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org/.

[30] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[31] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks*, volume 4, pages 2047–2052. IEEE, 2005.

[32] Yanbu Guo, Weihua Li, Bingyi Wang, Huiqing Liu, and Dongming Zhou. Deepacstm: deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. *BMC bioinformatics*, 20(1):1–12, 2019.

[33] Minju Ha and V Narry Kim. Regulation of microrna biogenesis. *Nature reviews Molecular cell biology*, 15(8):509–524, 2014.

[34] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, et al. Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. *Cell*, 141(1):129–141, 2010.

[35] Hamid Reza Hassanzadeh and May D Wang. Deeperbind: Enhancing prediction of sequence specificities of dna binding proteins. In *2016 IEEE International conference on bioinformatics and biomedicine (BIBM)*, pages 178–183. IEEE, 2016.

[36] Tsukaho Hattori, Masako Totsuka, Tokunori Hobo, Yasuaki Kagaya, and Akiko Yamamoto-Toyoda. Experimentally determined sequence requirement of acgt-containing abscisic acid response element. *Plant and Cell Physiology*, 43(1): 136–140, 2002.

[37] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. Mapping the human mirna interactome by clash reveals frequent noncanonical binding. *Cell*, 153(3):654–665, 2013.

[38] Jorja G Henikoff and Steven Henikoff. Using substitution probabilities to improve position-specific scoring matrices. *Bioinformatics*, 12(2):135–143, 1996.

[39] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[40] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. https://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735.

[41] Marilena V Iorio and Carlo M Croce. Microrna dysregulation in cancer: diagnostics, monitoring and therapeutics. a comprehensive review. *EMBO molecular medicine*, 4(3):143–159, 2012.

[42] Teresa Janas, Maja M Janas, Karolina Sapoń, and Tadeusz Janas. Mechanisms of rna loading into exosomes. *FEBS letters*, 589(13):1391–1398, 2015.

[43] Charles D Johnson, Aurora Esquela-Kerscher, Giovanni Stefani, Mike Byrom, Kevin Kelnar, Dmitriy Ovcharenko, Mike Wilson, Xiaowei Wang, Jeffrey Shelton, Jaclyn Shingara, et al. The let-7 microrna represses cell proliferation pathways in human cells. *Cancer research*, 67(16):7713–7722, 2007.

[44] Stefanie Jonas and Elisa Izaurralde. Towards a molecular understanding of microrna-mediated gene silencing. *Nature reviews genetics*, 16(7):421–433, 2015.

[45] David T Jones, William R Taylor, and Janet M Thornton. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3):275–282, 1992.

[46] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[47] Naoko Kajitani and Stefan Schwartz. Role of viral ribonucleoproteins in human papillomavirus type 16 gene expression. *Viruses*, 12(10):1110, 2020.

[48] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[49] Danijela Koppers-Lalic, Michael Hackenberg, Irene V Bijnsdorp, Monique AJ van Eijndhoven, Payman Sadek, Daud Sie, Nicoletta Zini, Jaap M Middeldorp, Bauke Ylstra, Renee X de Menezes, et al. Nontemplated nucleotide additions distinguish the small rna composition in cells from exosomes. *Cell reports*, 8(6): 1649–1658, 2014.

[50] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 25, pages 1097–1105, 2012. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[51] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[52] Yang Li, Jun Hu, Chengxin Zhang, Dong-Jun Yu, and Yang Zhang. Respre: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics*, 35(22):4647–4655, 2019.

[53] Xiao-Man Liu, Liang Ma, and Randy Schekman. Selective sorting of microrna into exosomes by phase-separated ybx1 condensates. *Elife*, 10:e71982, 2021.

[54] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. https://arxiv.org/abs/1301.3781.

[55] María Mittelbrunn, Cristina Gutiérrez-Vázquez, Carolina Villarroya-Beltri, Susana González, Fátima Sánchez-Cabo, Manuel Ángel González, Antonio Bernad, and Francisco Sánchez-Madrid. Unidirectional transfer of microrna-loaded exosomes from t cells to antigen-presenting cells. *Nature communications*, 2(1):282, 2011.

[56] Bert Moons, Daniel Bankman, and Marian Verhelst. *Embedded Deep Neural Networks*, pages 1–31. Springer International Publishing, Cham, 2019. ISBN 978-3-319-99223-5. doi: 10.1007/978-3-319-99223-5_1. URL `https://doi.org/10.1007/978-3-319-99223-5_1`.

[57] Usha K Muppirala, Vasant G Honavar, and Drena Dobbs. Predicting rna-protein interactions using only sequence information. *BMC bioinformatics*, 12:1–11, 2011.

[58] Gyoung Tae Noh, Jiyun Kwon, Jungwoo Kim, Minhwa Park, Da-Won Choi,

Kyung-Ah Cho, So-Youn Woo, Bo-Young Oh, Kang Young Lee, and Ryung-Ah Lee. Verification of the role of exosomal microrna in colorectal tumorigenesis using human colorectal cancer cell lines. *PLoS One*, 15(11):e0242057, 2020.

[59] Clifford C Nwaeburu, Natalie Bauer, Zhefu Zhao, Alia Abukiwan, Jury Gladkich, Axel Benner, and Ingrid Herr. Up-regulation of microrna let-7c by quercetin inhibits pancreatic cancer progression by activation of numbl. *Oncotarget*, 7(36): 58367, 2016.

[60] Angel R Ortiz, Charlie EM Strauss, and Osvaldo Olmea. Mammoth (matching molecular models obtained from theory): an automated method for model comparison. *Protein Science*, 11(11):2606–2621, 2002.

[61] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[62] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[63] Xiaoyong Pan, Peter Rijnbeek, Junchi Yan, and Hong-Bin Shen. Prediction of rna-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC genomics*, 19(1):1–11, 2018.

[64] Xiaoyong Pan, Yi Fang, Xianfeng Li, Yang Yang, and Hong-Bin Shen. Rbpsuite: Rna-protein binding sites prediction suite based on deep learning. *BMC genomics*, 21:1–8, 2020.

[65] Daniel Quang and Xiaohui Xie. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic acids research*, 44(11):e107–e107, 2016.

[66] Miguel Quévillon Huberdeau, Daniela M Zeitler, Judith Hauptmann, Astrid Bruckmann, Lucile Fressigné, Johannes Danner, Sandra Piquet, Nicholas Strieder, Julia C Engelmann, Guillaume Jannot, et al. Phosphorylation of argonaute proteins affects mrna binding and is essential for micro rna-guided gene silencing in vivo. *The EMBO journal*, 36(14):2088–2106, 2017.

[67] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[68] Castrense Savojardo, Pier Luigi Martelli, Piero Fariselli, and Rita Casadio. Deepsig: deep learning improves signal peptide detection in proteins. *Bioinformatics*, 34(10):1690–1696, 2018.

[69] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[70] Dan Shao, Lan Huang, Yan Wang, Kai He, Xueteng Cui, Yao Wang, Qin Ma, and Juan Cui. Deepsec: a deep learning framework for secreted protein discovery in human body fluids. *Bioinformatics*, 38(1):228–235, 2022.

[71] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

[72] Jiang Shu, Kevin Chiang, Janos Zempleni, and Juan Cui. Computational characterization of exogenous micrornas that can be transferred into human circulation. *PloS one*, 10(11):e0140587, 2015.

[73] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[74] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018.

[75] Thomas Treiber, Nora Treiber, Uwe Plessmann, Simone Harlander, Julia-Lisa Daiß, Norbert Eichner, Gerhard Lehmann, Kevin Schall, Henning Urlaub, and Gunter Meister. A compendium of rna-binding proteins that regulate microrna biogenesis. *Molecular cell*, 66(2):270–284, 2017.

[76] The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 46(5):2699–2699, 2018.

[77] The UniProt Consortium. "ago1 - protein argonaute-1 - homo sapiens (human) — uniprotkb — uniprot." https://www.uniprot.org/uniprotkb/q9ul18/entrystructure (accessed aug. 31, 2023). *Nucleic acids research*, 46(5):2699–2699, 2018.

[78] Hadi Valadi, Karin Ekström, Apostolos Bossios, Margareta Sjöstrand, James J Lee, and Jan O Lötvall. Exosome-mediated transfer of mrnas and micrornas is

a novel mechanism of genetic exchange between cells. *Nature cell biology*, 9(6): 654–659, 2007.

[79] Eric L Van Nostrand, Peter Freese, Gabriel A Pratt, Xiaofeng Wang, Xintao Wei, Rui Xiao, Steven M Blue, Jia-Yu Chen, Neal AL Cody, Daniel Dominguez, et al. A large-scale binding and functional map of human rna-binding proteins. *Nature*, 583(7818):711–719, 2020.

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, pages 5998–6008, 2017. https://arxiv.org/abs/1706.03762.

[81] Marian Verhelst and Bert Moons. Embedded deep neural network processing: Algorithmic and processor techniques bring deep learning to iot and edge devices. *IEEE Solid-State Circuits Magazine*, 9(4):55–65, 2017.

[82] Carolina Villarroya-Beltri, Cristina Gutiérrez-Vázquez, Fátima Sánchez-Cabo, Daniel Pérez-Hernández, Jesús Vázquez, Noa Martin-Cofreces, Dannys Jorge Martinez-Herrera, Alberto Pascual-Montano, María Mittelbrunn, and Francisco Sánchez-Madrid. Sumoylated hnrnpa2b1 controls the sorting of mirnas into exosomes through binding to specific motifs. *Nature communications*, 4(1):2980, 2013.

[83] Fei Wang, Chundi Zhou, Yanping Zhu, and Maryam Keshavarzi. The microrna

let-7 and its exosomal form: Epigenetic regulators of gynecological cancers. *Cell Biology and Toxicology*, 40(1):42, 2024.

[84] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3:1–40, 2016.

[85] Peipei Xia, Li Zhang, and Fanzhang Li. Learning similarity with cosine similarity ensemble. *Information sciences*, 307:39–52, 2015.

[86] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[87] Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. Attention-based multi-task learning for biomedical text mining. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

[88] Yujing Zhang, Danqing Liu, Xi Chen, Jing Li, Limin Li, Zhen Bian, Fei Sun, Jiuwei Lu, Yuan Yin, Xing Cai, et al. Secreted monocytic mir-150 enhances targeted endothelial cell migration. *Molecular cell*, 39(1):133–144, 2010.

[89] Weiying Zhou, Miranda Y Fong, Yongfen Min, George Somlo, Liang Liu, Melanie R Palomares, Yang Yu, Amy Chow, Sean Timothy Francis O'Connor, Andrew R Chin, et al. Cancer-secreted mir-105 destroys vascular endothelial barriers to promote metastasis. *Cancer cell*, 25(4):501–515, 2014.

Appendix A

Supplementary Data

All data and code associated with this research are available in the following repository: `https://github.com/sbbi-unl/DeepmiRBP`.

We wrote and tested all the code in Google Colab utilizing the A100 GPU. Due to the extensive computational requirements, as detailed in our study, a single epoch of our model took approximately 19 hours to run. The significant computational demands resulted in costs exceeding $10000 for running and testing our models.