

KYOTO UNIVERSITY
GRADUATE SCHOOL OF INFORMATICS

Deep Learning Approaches Leveraging Sequence-Structure Representations for Human microRNA Precursors Cleavage Site Prediction

配列・構造表現を活用した深層学習に基づくヒト・マイクロ RNA 前駆体
の切断部位予測

Lixuan Mu

Supervisor: Tatsuya Akutsu

A dissertation presented for the degree of
Doctor of Philosophy



京都大学
KYOTO UNIVERSITY

DEPARTMENT OF INTELLIGENCE SCIENCE AND TECHNOLOGY
Graduate School of Informatics

February, 2026

Abstract

Bioinformatics integrates computational and information science methodologies to extract insights from complex biological data. Although deep learning has enabled significant advances in biological prediction tasks, its application remains challenged by limited dataset size, data heterogeneity, and the need for models capable of capturing complex structural and contextual patterns. To address these issues, this dissertation introduces two deep learning architectures for predicting human Dicer cleavage sites.

MicroRNAs (miRNAs) are a class of small non-coding RNAs essential for post-transcriptional gene regulation, and their precise maturation depends on accurate cleavage of precursor miRNAs (pre-miRNAs) by Dicer. Experimental identification of Dicer cleavage sites is costly and labor-intensive, underscoring the importance of reliable computational prediction.

In Chapter 3, we present DiCleave, a CNN-based model that predicts Dicer cleavage sites within pre-miRNA segments (Cleavage Patterns). DiCleave integrates sequence, structural features, and learned secondary structure representations, achieving performance comparable to the state-of-the-art in binary classification and superior accuracy in multi-class classification.

In Chapter 4, we introduce DiCleavePlus, which overcomes key limitations of Cleavage Pattern-based models by allowing cleavage sites to occur at arbitrary positions and incorporating full-length pre-miRNA context. Using Transformer-based encoders, DiCleavePlus demonstrates accurate and robust performance, offering a more flexible and powerful framework for Dicer cleavage site prediction.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Tatsuya Akutsu, for his continuous guidance, encouragement, and support throughout my doctoral studies.

I am also deeply grateful to Dr. Tomoya Mori and Dr. Russell Young Neches for their invaluable advice, insightful comments, and generous assistance during my research.

Finally, I would like to extend my heartfelt thanks to my parents, all members of the Akutsu Laboratory, and my friends. Their unwavering support has sustained me both academically and personally.

Contents

List of Figures	v
List of Tables	x
1 Introduction	1
1.1 Background	1
1.2 Related work	3
1.3 Contributions	5
1.4 Contents organization	7
2 Preliminaries	8
2.1 Biogenesis of microRNAs	8
2.2 Deep neural networks	9
2.3 Convolutional neural networks	10
2.4 Transformer	10
3 DiCleave: a deep learning model for predicting human Dicer cleavage sites	13
3.1 Background	13
3.2 Methods	15
3.2.1 Concept of Cleavage Pattern	15
3.2.2 Dataset preparation	16
3.2.3 Architecture of autoencoder	19
3.2.4 Architecture of DiCleave	20
3.2.5 Ablation models	23
3.2.6 Performance evaluation	23
3.3 Results	24
3.3.1 Performance of binary classification models	24
3.3.2 Performance of the multi-class classification model	26
3.4 Discussion	31

4	DiCleavePlus: a Transformer-based model to detect human Dicer cleavage sites within Cleavage Pattern	35
4.1	Background	35
4.2	Methods	37
4.2.1	Extended Cleavage Pattern	37
4.2.2	Data preparation	38
4.2.3	Data processing	42
4.2.4	Model architecture	43
4.2.5	Attentional feature fusion	45
4.2.6	Comparison with ablation models	45
4.2.7	Comparison with other models	48
4.2.8	Details of training procedure	48
4.2.9	Performance metrics	49
4.3	Results	52
4.3.1	Performance of cleavage site position prediction	52
4.3.2	Performance of positive-negative binary classification	55
4.3.3	Performance comparison with ablation models	57
4.3.4	Performance comparison with other models	58
4.4	Discussion	63
5	Conclusion and Future Directions	67
5.1	Conclusion	67
5.2	Future directions	68
	List of Publications	71
	References	72

List of Figures

1.1	Schematic summary of miRNA research in bioinformatics.	4
2.1	Canonical miRNA biogenesis pathway. Primary miRNAs (pri-miRNAs) are transcribed from genomic DNA in the nucleus and subsequently cleaved by the RNase III enzyme Drosha to generate precursor miRNAs (pre-miRNAs). The pre-miRNAs are then exported to the cytoplasm, where they are further processed by Dicer to produce a miRNA duplex. One strand of the duplex, known as the guide strand, is loaded into the AGO protein to form the miRNA-induced silencing complex (miRISC), while the opposite strand, referred to as the passenger strand, is degraded.	9
2.2	Convolutional neural network. A classical convolutional neural network architecture is shown. A convolutional kernel slides over the input tensor to compute dot products and generate feature maps. A subsequent pooling operation is applied to these feature maps to reduce their spatial dimensions and complexity.	11
2.3	Self-attention mechanism. Given an input matrix X , the Query (Q), Key (K), and Value (V) matrices are computed using learnable weight matrices W^Q , W^K , and W^V , respectively. The attention weights are obtained by applying a softmax function to the scaled dot-product QK^T . The final output of the self-attention operation is computed as the weighted sum of V , where the weights are given by the normalized attention scores.	12

3.1	Illustration of Cleavage Patterns. The predicted secondary structure of the pre-miRNA hsa-mir-548ar is shown as an example. (a) The 5' Cleavage Pattern is enclosed in a box and spans 14 nucleotides. The Dicer cleavage site is indicated by a red triangle positioned at the center of this region. The complementary sequence corresponding to the 5' Cleavage Pattern is shown below. Notably, the fifth base and the last two bases (U, A, and A, respectively) are unpaired; these positions are represented using the symbol "O". (b) The 3' Cleavage Pattern is similarly highlighted, with the 3' Dicer cleavage site marked by a red triangle at its center. The complementary sequence is displayed below, and the unpaired guanine in this region is denoted by the symbol "O".	16
3.2	Overview of dataset preprocessing and partitioning for the classification models.	19
3.3	Architecture of the autoencoder. A fully connected autoencoder is used to generate the representations of pre-miRNA secondary structures. The input consists of one-hot encoded dot-bracket representations standardized to a fixed length of 200 nucleotides. The encoder compresses this representation into a 64-dimensional latent vector, which serves as the representation for the pre-miRNA secondary structure. The encoder and decoder consist of three FC layers, which gradually project the input to 64-dimensional vector and the latent representation to the reconstructed sequence, respectively.	20

3.4	Architecture of DiCleave. The model architecture is divided into two sequential modules. Module 1 receives a 13-dimensional input tensor composed of one-hot encoded Cleavage Pattern features, complementary sequence features, and secondary structure features. The output of Module 1 is the representation of Cleavage Pattern inputs, and is then concatenated with the representation of the corresponding pre-miRNA secondary structure and passed to Module 2 for further processing. The final output layer is configured to support either binary or multi-class classification depending on the prediction task.	22
3.5	Binary performance of DiCleave. (a) Receiver-operating characteristic (ROC) curves and (b) precision-recall (PR) curves for DiCleave in the binary classification task. Blue and green lines represent performance of the classifiers for 5' and 3' Cleavage Patterns, respectively.	25
3.6	Binary performance comparison between DiCleave and existing models. Comparison of DiCleave with previously proposed methods for binary classification of (a) 5' cleavage site prediction and (b) 3' cleavage site prediction. For DiCleave and ReCGBM, the values represent the average performance across 10 independent replications. The performance metrics for PHDcleav and LBSIZEcleav are taken from their original publications (indicated by asterisks).	28
3.7	Confusion matrix for the multi-class classification model. Rows and columns correspond to the true and predicted labels, respectively. Values along the diagonal represent the number of correctly classified instances for each class.	30
3.8	Multi-class classification performance of DiCleave. (a) Receiver-operating characteristic (ROC) curves and (b) precision-recall (PR) curves for the multi-class classification model. For evaluation, the model was converted into a one-vs-all classifier by treating one class as the positive class and the remaining two classes as negative.	31

4.1	Concept of extended Cleavage Pattern.	Conventional Cleavage Patterns (upper, black box) restrict the Dicer cleavage site (red triangle) to the central position. In DiCleavePlus, Cleavage Patterns are generated using a sliding window approach along the pre-miRNA sequence, allowing the cleavage site to appear at any position within the pattern (lower, blue and green boxes). As examples, the blue pattern is labeled as class 12 and the green pattern as class 5.	37
4.2	Linear representation of human pre-miRNA hsa-let-7a.	The blue-shaded regions correspond to the two mature miRNAs, hsa-let-7a-5p and hsa-let-7a-3p. The positions of the Dicer cleavage sites (red triangles) and Drosha cleavage sites (blue triangles) are inferred from the boundaries of these mature miRNAs.	38
4.3	Workflow for dataset construction.		40
4.4	Class distribution of patterns across different datasets.	The x-axis represents class labels, and the y-axis indicates the number of patterns. In Dataset-1, the number of negative patterns exceeds that of each positive class, while the total number of positive patterns is approximately equal to the number of negative patterns. In Dataset-2, the number of negative patterns is nearly equal to that of each individual positive class; however, the total number of positive patterns exceeds the number of negative patterns.	41
4.5	Data processing workflow of DiCleavePlus.	(a) Pre-miRNA sequence (blue blocks) and their corresponding secondary structures (green blocks) are tokenized into 3-mers and encoded using 32-dimensional embeddings. (b) The sequence embeddings are concatenated with the secondary structure embeddings to generate the input features of DiCleavePlus. . . .	44

4.6	Overview of the DiCleavePlus architecture. The grey boxes represent the Transformer-based Encoder Units, which are used to extract features from Cleavage Pattern and full-length pre-miRNA sequences. Two feature integration strategies were implemented. Models employing the attentional feature fusion (AFF) block to integrate processed pattern and sequence features are referred to as DCP-AFF, whereas models that merge these features using simple concatenation are referred to as DCP-Concat.	46
4.7	Architecture of attentional feature fusion (AFF) block. The orange module represents the multi-scale channel attention block. The symbol $\oplus$ denotes element-wise addition, and $\otimes$ indicates element-wise multiplication. The dash line indicates the computation of complementary weight $(1 - w)$	47
4.8	Confusion matrix heatmaps of DiCleavePlus models trained and evaluated on Dataset-1 with different pattern sizes. (a) DCP-AFF with 14-nt patterns; (b) DCP-Concat with 14-nt patterns; (c) DCP-AFF with 18-nt patterns; (d) DCP-Concat with 18-nt patterns. The x-axis denotes the predicted class; the y-axis denotes the true class. Class 0 corresponds to negative patterns (i.e., patterns that do not contain a cleavage site). . .	53
4.9	Confusion matrix heatmaps of DiCleavePlus models trained and evaluated on Dataset-2 with different pattern sizes. (a) DCP-AFF with 14-nt patterns; (b) DCP-Concat with 14-nt patterns; (c) DCP-AFF with 18-nt patterns; (d) DCP-Concat with 18-nt patterns. The x-axis denotes the predicted class; the y-axis denotes the true class. Class 0 corresponds to negative patterns (i.e., patterns that do not contain a cleavage site). . .	57

List of Tables

1.1	Summary of existing cleavage site prediction models	6
3.1	Configuration of the main dataset	17
3.2	Performance of DiCleave and its ablation models	26
3.3	Description of class-unbalanced datasets and pre-miRNA sequence similarity cutoff datasets	27
3.4	Performance of the best DiCleave model on unbalanced test sets	28
3.5	Performance of DiCleave and ReCGBM for 10 replications . .	29
3.6	Performance of models trained from similarity cutoff datasets	33
4.1	Dataset configuration	39
4.2	Average training time of DiCleavePlus (minutes)	49
4.3	DiCleavePlus cleavage site prediction performance using different pattern sizes	54
4.4	DiCleavePlus binary performance using different pattern sizes	56
4.5	Performance comparison between DiCleavePlus and its ablation models	59
4.6	Binary performance comparison between DiCleavePlus and its ablation models	60
4.7	Performance comparison between DiCleavePlus and other models	61
4.8	Binary performance comparison between DiCleavePlus and other models	62

Chapter 1

Introduction

1.1 Background

Biology seeks to elucidate the fundamental principles that govern living organisms, from the molecular mechanisms underlying cellular processes to the dynamic interactions shaping ecological systems. Although historically a descriptive discipline, biology was transformed in the 20th century by advances in molecular biology, biochemistry, and genetics, evolving into a mechanistic science capable of dissecting the architecture and function of biological systems. The development of high-throughput sequencing, mass spectrometry-based proteomics, genome engineering technologies such as CRISPR-Cas systems, and single-cell analysis has enabled the examination of biological processes with unprecedented depth and precision. These technological breakthroughs have greatly expanded our understanding of life's molecular complexity and accelerated progress across medicine, pharmacy, agriculture, and ecology.

The rapid advancement of experimental technologies has led to an explosion of biological data spanning the field of genomics, transcriptomics, proteomics, metabolomics, and epigenomics. Modern research routinely generates data of a scale and complexity that exceed the capacity of manual analysis. Extracting meaningful insights from such data requires systematic computational

methodologies for processing, statistical modeling, and multi-omics integration. Bioinformatics emerged to address these needs, integrating concepts from computer science, mathematics, and statistics with biological knowledge to enable comprehensive data analysis. It now plays an essential role in identifying molecular signatures, characterizing regulatory networks, and uncovering mechanisms that would otherwise remain inaccessible through experimental approaches alone. As biological data continue to expand, bioinformatics has become indispensable in modern biological research.

Machine learning has emerged as a powerful paradigm for identifying patterns in large and complex datasets, making it particularly suitable for biological applications. Early machine learning methods, including support vector machines (SVM), random forests (RF), and probabilistic graphical models, achieved considerable success in tasks such as gene expression classification, motif detection, structural annotation, and variant prioritization. Building on these foundations, deep learning has revolutionized the field by enabling models to learn hierarchical representations directly from raw biological data. In particular, deep learning has been widely applied to sequence-based problems, where DNA, RNA, and protein sequences are modeled directly through architectures such as convolutional neural networks (CNNs), recurrent neural network (RNNs), and Transformer-based models. These approaches have yielded substantial advances in predicting transcription factor binding sites, identifying RNA secondary structures, recognizing splice junctions, modeling chromatin accessibility, and inferring variant effects from genomic sequences.

Computational methods have been extensively developed for various aspects of microRNA (miRNA) analysis. Machine learning approaches, in particular, have been widely applied to miRNA discovery and precursor identification. An early SVM-based method employing local contiguous structure-sequence features achieved approximately 90% accuracy in identifying human miRNA precursors (pre-miRNAs) [1]. MiPred [2] is another SVM-based framework that distinguishes real pre-miRNAs from pseudo hairpin sequences using a Gaussian radial basis function kernel and 29 global and intrinsic hairpin folding features. MiRRim [3] is a hidden Markov model that integrates evolutionary conservation and secondary structure information for miRNA detection, while miR-KDE [4] applies a Naive Bayes-based kernel density estimation approach to classify pre-miRNAs based on structural characteristics.

With the advent of deep learning, miRNA discovery and classification have further advanced through improved representation learning. Park et al. [5] propose a RNN-based architecture for pre-miRNA identification. More recently, miRe2e [6] introduced a fully end-to-end framework based on the Transformer architecture [7], enabling the identification of pre-miRNA sequences directly from genomic data.

Another major research direction in computational miRNA analysis is miRNA target prediction. TargetScan [8] is a classic algorithm that predicts conserved miRNA targets across multiple genomes by modeling RNA duplex thermodynamics and comparative sequence analysis. TarPmiR [9] integrates crosslinking ligation and sequencing of hybrids (CLASH) data with a random forest model to improve target prediction accuracy. In addition, deep learning-based approaches have been proposed to further enhance miRNA target prediction by integrating diverse sequence- and structure-derived features [10, 11].

In addition to these aspects, because miRNA maturation requires precise cleavage on precursor sequences by RNase enzymes, several computational methods have been proposed to predict cleavage sites on pre-miRNAs [12, 13, 14, 15].

This dissertation focuses on the development of accurate end-to-end computational methodologies for pre-miRNA cleavage site prediction (specifically, Dicer cleavage site prediction), aiming to advance both predictive performance and practical applicability in miRNA biogenesis studies.

1.2 Related work

PHDcleav [12] is the first SVM-based approach developed to predict human Dicer cleavage sites on pre-miRNAs. In this method, the authors define a Dicer Cleavage Pattern as a segment of the pre-miRNA sequence. A pattern is labeled as positive when the Dicer cleavage site is located at the center of the segment. The input features of PHDcleav integrate primary sequence information with partial secondary structure information, specifically focusing on loop/bulge structures, which are known to play an important role in determining Dicer cleavage positions [16]. One-hot encoding is used to represent the combined sequence and structural features.

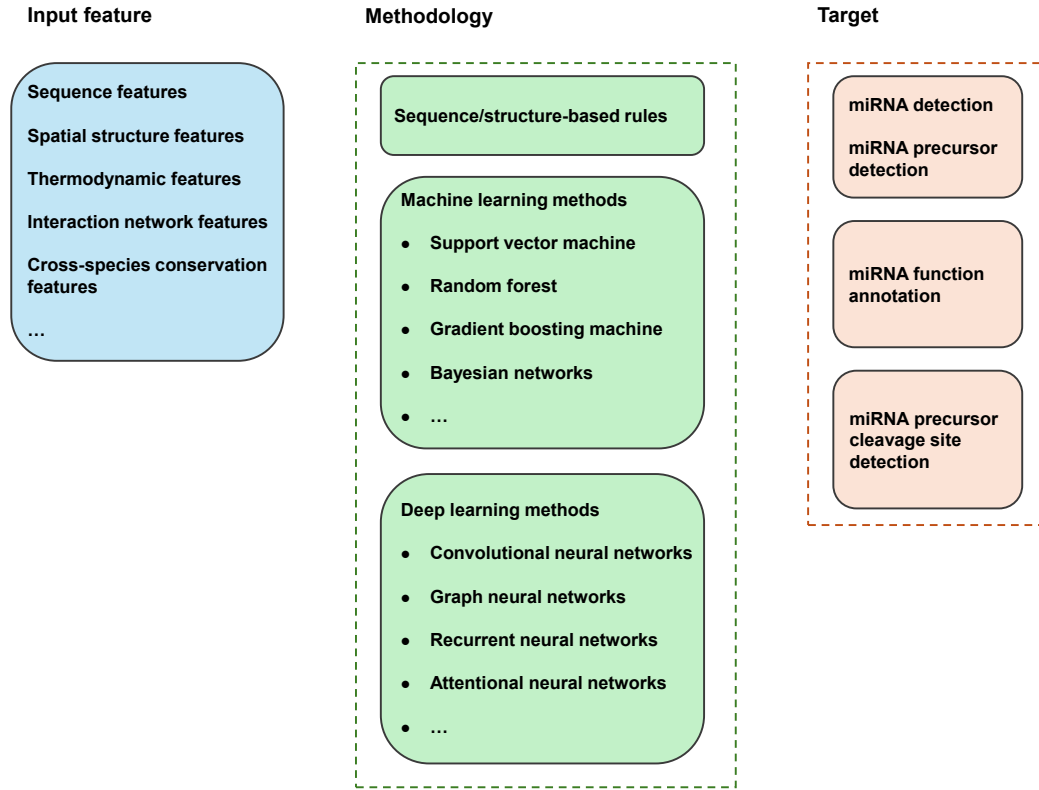


Figure 1.1: **Schematic summary of miRNA research in bioinformatics.**

LBSIZEcleav [13] is another SVM-based model that follows a similar design principle. Like PHDcleav, it incorporates loop/bulge structural information along with sequence features. However, LBSIZEcleav extends this feature by considering not only the position but also the length of loop/bulge regions. Although its feature representation is more complex, LBSIZEcleav achieves superior performance compared with PHDcleav.

ReCGBM [14] represents a more recent architecture that also uses Cleavage Patterns for predicting human Dicer cleavage sites. Based on the secondary structure of pre-miRNAs, the authors extract complementary sequence of Cleavage Patterns and use the character “O” to denote unpaired nucleotides, thereby forming complementary pattern sequences. Assuming that similar inputs should yield similar predictions, affinity propagation clustering [17] is applied to assign cluster labels to samples as auxiliary information. ReCGBM then employs a gradient boosting machine as the final predictor, achieving state-of-the-art performance in cleavage site prediction.

While PHDcleav, LBSIZEcleav, and ReCGBM rely on Cleavage Patterns as their primary input, DeepMirCut [15] is an RNN-based model that predicts

both Dicer and Drosha cleavage sites directly from RNA sequences. Leveraging the strong representation capabilities of deep neural networks, it eliminates the need for complex handcrafted feature engineering and supports multi-output prediction, enabling simultaneous identification of cleavage sites for both Dicer and Drosha.

1.3 Contributions

In this dissertation, we introduce two deep learning architectures for predicting human Dicer cleavage sites.

In the first study, we introduce DiCleave, a CNN-based approach for predicting cleavage sites within Cleavage Patterns. DiCleave is the first Cleavage Pattern-based deep neural network model, which integrates multiple sources of information, including the sequence of the Cleavage Pattern, its secondary structure, and the corresponding complementary pattern sequence. In addition, we incorporate a pre-trained autoencoder to extract secondary structure representations from the full-length pre-miRNA from which each pattern is derived. Unlike previous machine learning methods that require separate models for patterns from the 5' and 3' arms, DiCleave employs a unified multi-classification layer to determine whether a pattern contains a cleavage site and, if so, from which arm it originates. This design eliminates the need for prior knowledge about the input pattern, thereby enhancing its practicality in real-world applications. Benchmarking experiments demonstrate that DiCleave achieves performance comparable to that of ReCGBM in binary classification tasks.

In the secondary study, we introduce DiCleavePlus to overcome the limitations inherent in models that apply Cleavage Pattern. Conventional definition considers a Cleavage Pattern positive only when the cleavage site lies at its center. We extend this definition by treating all intervals within the pattern as distinct classes, thereby increasing both the modeling flexibility and the effective size of the training data. DiCleavePlus incorporates both the extended Cleavage Pattern and its associated full-length pre-miRNA as joint inputs and employs Transformer-based encoders to learn features from both sequence and secondary structure representations. By enabling flexible cleavage site positioning and integrating multi-level RNA information, DiCleavePlus maximizes

Table 1.1: Summary of existing cleavage site prediction models

Model	Use of CP	Input Feature	Classifier	Type of Output
PHDcleav	yes	• pattern sequence	SVM	binary
LBSizeCleav	yes	• pattern secondary structure • pattern sequence • pattern secondary structure • loop/bulge counting	SVM	binary
ReCGBM	yes	• pattern sequence • complementary pattern sequence • pattern similarity label	GBM	binary
DeepMirCut	no	• miRNA precursor sequence • miRNA precursor structure	RNN	multiple

CP: Cleavage Pattern
 GBM: Gradient boosting machine

the utilization of structural and contextual features. These improvements underscore its potential as a practical tool to guide experimental design, reduce the search space for candidate cleavage sites, and facilitate downstream investigations into miRNA biogenesis and function.

1.4 Contents organization

In Chapter 2, we provide a concise introduction to the biological background of miRNA biogenesis as well as the foundational deep learning techniques that underpin the proposed models.

In Chapter 3, we present our first study, which focuses on identifying human Dicer cleavage sites using a CNN-based framework (DiCleave). We describe the model architecture in detail and compare its performance with state-of-the-art methods. The results demonstrate the effectiveness of applying deep learning to Dicer cleavage site prediction.

In Chapter 4, we discuss the major limitations of existing Cleavage Pattern-based prediction models and introduce a new architecture, DiCleavePlus. This framework incorporates an extended Cleavage Pattern and employs Transformer-based encoders to extract features from both sequence and structural representations. Comprehensive experiments are conducted to evaluate the accuracy and robustness of the proposed model.

In Chapter 5, we summarize the findings of Chapters 3 and 4 and discuss potential directions for future research.

Chapter 2

Preliminaries

2.1 Biogenesis of microRNAs

MicroRNAs (miRNAs) are small non-coding RNAs with approximately 20-25 nucleotides in length. They participate in a wide range of cellular processes and play pivotal roles in animal development [18]. Mature miRNAs associate with Argonaute (AGO) proteins, guiding them to the 3' untranslated regions (UTRs) of target messenger RNAs (mRNAs). Through base pairing with complementary sequences within the mRNA 3' UTR, miRNAs direct AGO-mediated silencing of target transcripts [19, 20].

In the canonical miRNA biogenesis pathway, primary miRNAs (pri-miRNAs) are transcribed in the nucleus. These transcripts are processed by Drosha, an RNase III enzyme and a component of the microprocessor complex, which cleaves near the base of the hairpin structure to generate precursor miRNAs (pre-miRNAs) [21]. The resulting pre-miRNAs are subsequently exported to the cytoplasm and further processed by Dicer, another RNase III endonuclease, which removes the terminal loop to produce a mature miRNA duplex [21, 22, 23]. MiRNAs derived from the 5' arm of the pre-miRNA hairpin are referred to as the 5' strands, while those originating from the 3' arm are termed the 3' strands.

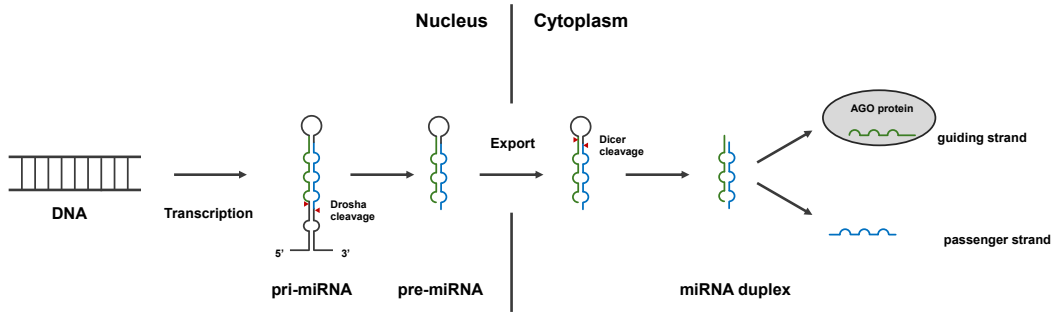


Figure 2.1: **Canonical miRNA biogenesis pathway.** Primary miRNAs (pri-miRNAs) are transcribed from genomic DNA in the nucleus and subsequently cleaved by the RNase III enzyme Drosha to generate precursor miRNAs (pre-miRNAs). The pre-miRNAs are then exported to the cytoplasm, where they are further processed by Dicer to produce a miRNA duplex. One strand of the duplex, known as the guide strand, is loaded into the AGO protein to form the miRNA-induced silencing complex (miRISC), while the opposite strand, referred to as the passenger strand, is degraded.

One strand of the miRNA duplex is selectively loaded into an AGO protein to form the miRNA-induced silencing complex (miRISC). The relative proportion of 5' strand and 3' strand incorporated into AGO protein varies across cell types and cellular conditions [24]. The strand incorporated into AGO is designated the guide strand, whereas the opposing strand, known as the passenger strand, is unwound from the duplex and subsequently degraded [25].

2.2 Deep neural networks

Deep neural networks (DNN) are computational models composed of multiple layers of interconnected neurons designed to learn hierarchical representations from data. Each neuron serves as a basic computational unit that computes a weighted sum of its inputs. Modern neuron implementations further incorporate nonlinear activation functions, enabling the network to model complex relationships and allowing a neuron to produce an output only when its input exceeds a certain threshold [26]. DNNs are typically trained using backpropagation, in which gradients of an objective function are propagated through the network to iteratively update and optimize model parameters.

DNNs have achieved remarkable success across a wide range of domains, including image recognition, speech recognition, potential drug molecules prediction, particle accelerator data analysis, and predicting the effects of mutations in non-coding DNA [27]. Building on the classical DNN paradigm, numerous architectural variants have been developed to improve data representation, preserve information more effectively, or enhance computational efficiency.

2.3 Convolutional neural networks

Convolutional neural networks (CNNs) are an important class of DNNs. Inspired by the organization of the mammalian visual cortex [28], neurons in a CNN layer are connected only to a local subset of neurons in the preceding layer. This locality is implemented through convolutional kernels, which are small matrices that slide across the input and compute dot products with local sub-regions of the data. Stacking multiple convolutional layers enables hierarchical feature extraction, i.e., kernels in early layers detect concrete, simple, localized patterns, while kernels in deeper layers capture more abstract and complex features.

Another characteristic of CNNs is weight sharing, where the parameters of each kernel remain fixed as it moves across the input. This significantly reduces the number of trainable parameters compared with fully-connected layers [29]. CNNs commonly incorporate pooling operations following convolution, which downsample feature maps and further reduce computational cost while maintaining robustness to small spatial variations. Through their ability to extract hierarchical representations effectively and with relatively few parameters, CNNs have become powerful feature extractors across many areas of informatics.

2.4 Transformer

Transformer [7] is a revolutionary deep learning architecture that advanced sequence modeling tasks. Similar to earlier neural sequence transduction models [30, 31], Transformer adopts an encoder-decoder framework. However, unlike recurrent or convolutional architectures, Transformer relies entirely on the self-attention mechanism to compute contextualized representations of both its input and output sequences.

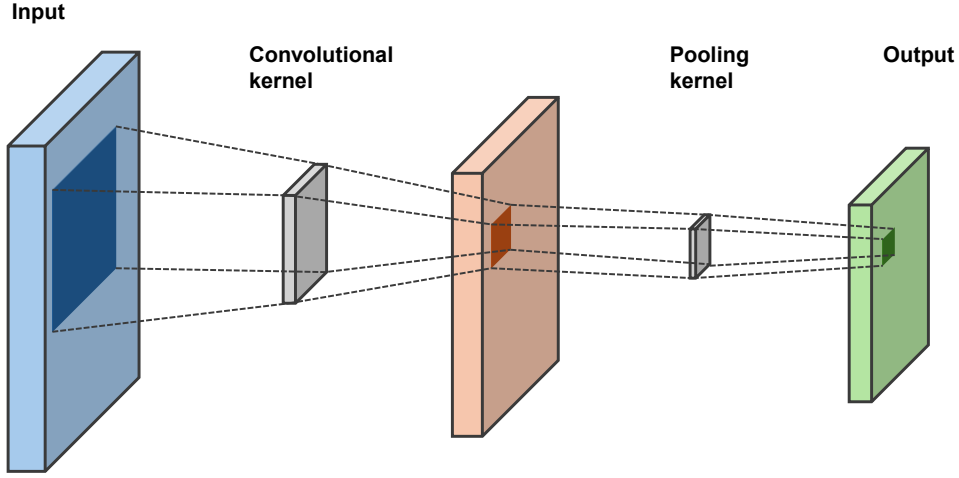


Figure 2.2: **Convolutional neural network.** A classical convolutional neural network architecture is shown. A convolutional kernel slides over the input tensor to compute dot products and generate feature maps. A subsequent pooling operation is applied to these feature maps to reduce their spatial dimensions and complexity.

An attention mechanism can be described as a function that maps a query and a set of key-value pairs to an output. The output is computed as a weighted sum of the values, where the weight for each value is determined by a compatibility function between the query and the corresponding key [7]. Transformer encoder implements self-attention mechanism, in which the queries, keys and values are all derived from the same input sequence. Through this mechanism, each element in the sequence attends to all other elements, allowing the model to capture long-range dependencies and intrinsic relationships within the input. The attention mechanism can be mathematically described as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2.1)$$

where d_k is the dimension of Q and K , and $\frac{1}{\sqrt{d_k}}$ serves as a scaling factor.

Because self-attention alone does not encode positional information, Transformer incorporates sinusoidal positional encodings to preserve the order of the sequence. These encodings provide each token with a unique, continuous representation of its position, which enables Transformer to effectively distinguish between sequence elements even when they are far apart.

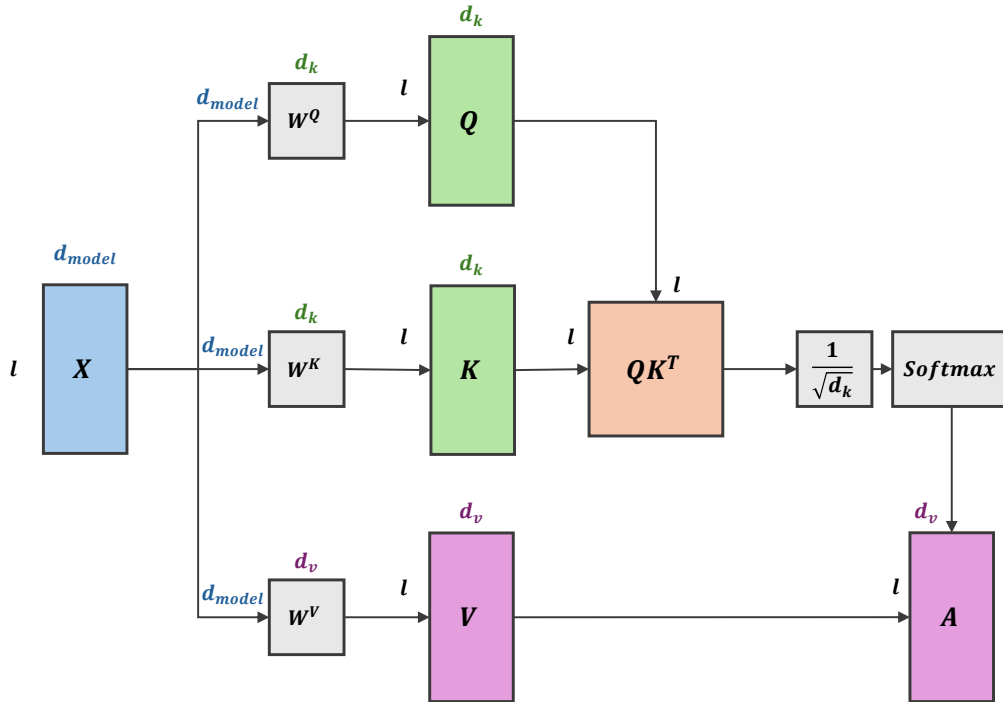


Figure 2.3: **Self-attention mechanism.** Given an input matrix X , the Query (Q), Key (K), and Value (V) matrices are computed using learnable weight matrices W^Q , W^K , and W^V , respectively. The attention weights are obtained by applying a softmax function to the scaled dot-product QK^T . The final output of the self-attention operation is computed as the weighted sum of V , where the weights are given by the normalized attention scores.

Chapter 3

DiCleave: a deep learning model for predicting human Dicer cleavage sites

3.1 Background

MicroRNAs (miRNAs) are a class of non-coding RNAs that function as gene expression regulators across diverse animal systems [32]. Typically, miRNAs bind to complementary sequences located in the 3' untranslated regions (UTRs) of target messenger RNAs (mRNAs), leading to mRNA degradation and post-transcriptional gene silencing [25]. Dysregulation of miRNA expression has been closely associated with a wide range of human diseases, including cancer [33, 34, 35], cardiovascular disorders [36, 37], inflammatory diseases [38, 39, 40], neurodevelopmental disorders [41, 42, 43], and other pathological conditions [44].

Mature miRNAs are approximately 20-25 nucleotides in length. In the canonical miRNA biogenesis pathway, miRNA genes are first transcribed into long primary transcripts known as primary miRNAs (pri-miRNAs), which may span several kilobases [45]. Pri-miRNAs are then processed in the nucleus by

the Microprocessor complex, which consists of the RNase III enzyme Drosha and its associated cofactors [46]. This processing step produces precursor miRNAs (pre-miRNAs), short stem-loop structures that are subsequently exported to the cytoplasm and further cleaved by the RNase III endonuclease Dicer, generating mature miRNA duplexes [32]. Dicer cleavage produces two miRNA strands—one derived from the 5' arm and the other from the 3' arm of the pre-miRNA hairpin [47].

Dicer is a key RNase III enzyme essential for the proper maturation of miRNAs [48, 25, 49]. It contains multiple functional domains [50]. Among which the RNase IIIa and RNase IIIb domains are pivotal for cleaving precursor RNA molecules [51, 49], while the PAZ domain participates in determining cleavage positions [50, 51]. Since miRNAs regulate gene expression through base pairing with target mRNAs, precise Dicer cleavage is of importance, e.g., inaccurate cleavage can generate miRNAs with altered seed regions [8], thereby impacting the proper function of miRNAs. Experimental identification of Dicer cleavage sites in relation to pre-miRNA sequence and secondary structure relies on techniques such as Northern blotting, primer extension, isotopic or chemical labeling, in vitro RNA synthesis, RNA immunoprecipitation, and high-throughput RNA sequencing [52, 23, 53, 54, 55, 51]. These methods, however, are label-intensive, time-consuming, and costly. Therefore, an accurate computational predictive model of Dicer cleavage sites can greatly facilitate experimental design and biological discovery.

Numerous machine learning-based approaches have been developed for protein and RNA cleavage site prediction [56, 57, 58, 59, 60, 61, 62, 63, 64, 65]. For Dicer-specific cleavage site prediction in humans, two notable support vector machine (SVM) models have been proposed [12, 13]. More recently, Liu et al. introduced a gradient boosting machine-based model, ReCGBM [14], which predicts the presence of a cleavage site within the central position of a 14-nucleotide (nt) sequence (Cleavage Pattern). Although ReCGBM surpasses earlier SVM-based methods, several limitations remain. First, ReCGBM is restricted to binary classification, i.e., it determines whether a Cleavage Pattern contains a cleavage site but cannot identify the originating arm (5' or 3'), though each pre-miRNA contains two Dicer cleavage sites. This limitation reduces its utility in more complex or integrated prediction pipelines. Second,

ReCGBM does not fully exploit the secondary structural information of pre-miRNAs, despite the critical role of RNA structure in processing and function [51, 66].

Deep neural networks (DNNs) have demonstrated remarkable success across domains such as computer vision and natural language processing, and they have been increasingly applied to bioinformatics tasks. Their capacity to extract latent features directly from raw data makes them powerful alternatives to traditional feature-engineered models. Computational approaches also offer practical and cost-effective substitutes for labor-intensive biological experiments, including RNA secondary structure prediction [67, 68] and drug-drug interaction prediction [69, 70]. Numerous studies have explored the use of DNNs in bioinformatics and computational biology [71, 72, 73].

Motivated by the limitations of ReCGBM, we propose DiCleave, a DNN-based model that integrates pre-miRNA sequence segments with secondary structure information using an autoencoder-derived representations. DiCleave operates as both a binary and multi-class model. Our results demonstrate that its performance is comparable to ReCGBM in binary classification and exhibits strong performance in multi-class classification.

3.2 Methods

3.2.1 Concept of Cleavage Pattern

ReCGBM utilizes a 14-nt segment of pre-miRNA as its input, which we refer to as the Cleavage Pattern. A Cleavage Pattern is labeled as positive if a Dicer cleavage site is located at its central position; otherwise, it is designated as negative. Because pre-miRNAs fold into characteristic hairpin structures, each Cleavage Pattern has a corresponding complementary sequence. If a nucleotide in the Cleavage Pattern is paired with another nucleotide in the pre-miRNA structure, the complementary base is represented using its standard nucleotide symbol (“A”, “C”, “G”, or “U”). If the base is unpaired, the corresponding position in the complementary sequence is denoted by “O”. An illustrative example of a Cleavage Pattern and its complementary sequence is shown in Figure 3.1.

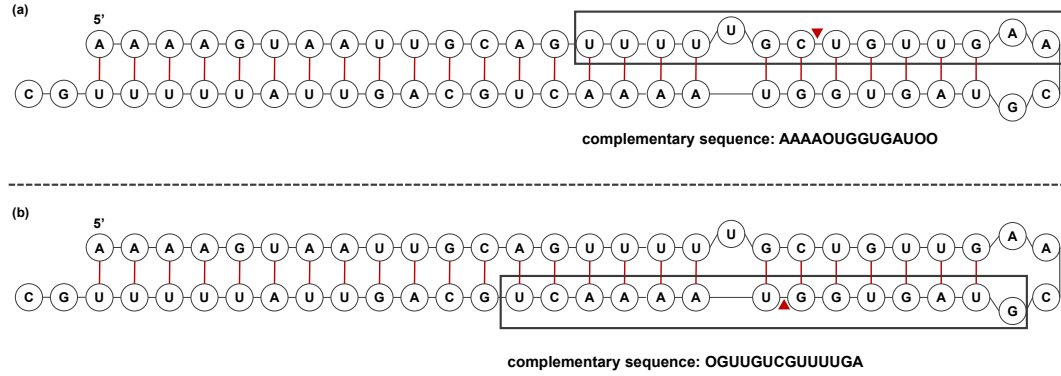


Figure 3.1: **Illustration of Cleavage Patterns.** The predicted secondary structure of the pre-miRNA hsa-mir-548ar is shown as an example. **(a)** The 5' Cleavage Pattern is enclosed in a box and spans 14 nucleotides. The Dicer cleavage site is indicated by a red triangle positioned at the center of this region. The complementary sequence corresponding to the 5' Cleavage Pattern is shown below. Notably, the fifth base and the last two bases (U, A, and A, respectively) are unpaired; these positions are represented using the symbol “O”. **(b)** The 3' Cleavage Pattern is similarly highlighted, with the 3' Dicer cleavage site marked by a red triangle at its center. The complementary sequence is displayed below, and the unpaired guanine in this region is denoted by the symbol “O”.

3.2.2 Dataset preparation

Main dataset

We used a total of 956 human pre-miRNA sequences from ReCGBM [14], originally obtained from the miRBase database (Release 22.1) [74], to construct the main dataset for this study. Each pre-miRNA contains two positive Cleavage Patterns: one derived from the 5' arm and the other from the 3' arm. In addition, we randomly selected two 14-nt sequences from each pre-miRNA that did not contain a Dicer cleavage site—one from the 5' arm and the other from the 3' arm—and designated them as negative Cleavage Patterns. Consequently, each pre-miRNA contributed four Cleavage Patterns, resulting in a total of 3,824 patterns.

We further retrieved the full-length sequences of these pre-miRNAs from miRBase (Release 22.1) and predicted their secondary structures using RNAfold from the ViennaRNA package [75]. RNAfold outputs base-pairing information in dot-bracket notation, which we used as the structural representation. An overview of the main dataset is shown in Table 3.1.

Autoencoder dataset

To extract secondary structure representations of pre-miRNAs, we constructed a separate dataset for training an autoencoder. We collected all human pre-miRNA sequences from miRBase (Release 22.1) and removed those included in the main dataset. The remaining sequences were filtered by length, retaining only those between 40 and 200 nucleotides. Their secondary structures were then predicted using RNAfold, yielding a final collection of 36,190 pre-miRNA secondary structures for autoencoder training.

Data preprocessing

RNAfold outputs secondary structures in dot-bracket notation using three symbols: “(” and “)” to denote paired bases; “.” to denote unpaired bases.

To represent this structures, we applied one-hot encoding. For the autoencoder, each secondary structure was encoded as a 4×200 tensor, where the fourth symbol “N” served as a padding indicator for sequences shorter than 200 nt.

Cleavage Pattern sequences were also one-hot encoded. As shown in Figure 3.1, unpaired positions in the complementary sequence were assigned the symbol “O”. Thus, both Cleavage Patterns and the complementary sequences were encoded using five symbols: “A”, “C”, “G”, “U”, and “O”. Dot-bracket secondary structures for the Cleavage Patterns were encoded using only “(”, “)”, and “.” without the “N” padding symbol.

Datasets separation

To train and evaluate the binary classification models, we divided the main dataset into two subsets based on origin arm of the patterns: one containing all patterns from the 5’ arm and the other containing all patterns from the 3’ arm. From each subset, 800 positive and 800 negative patterns were randomly selected to form the training sets, *DS-5p-train* and *DS-3p-train*, respectively. The remaining 312 patterns in each arm were reserved as independent test sets (*DS-5p-test* and *DS-3p-test*).

For the multi-class classification model, we reassigned pattern labels: negative = 0, 5’ positive = 1, and 3’ positive = 2. We then randomly selected 800 patterns from each class (5’ positive, 3’ positive, and negative patterns

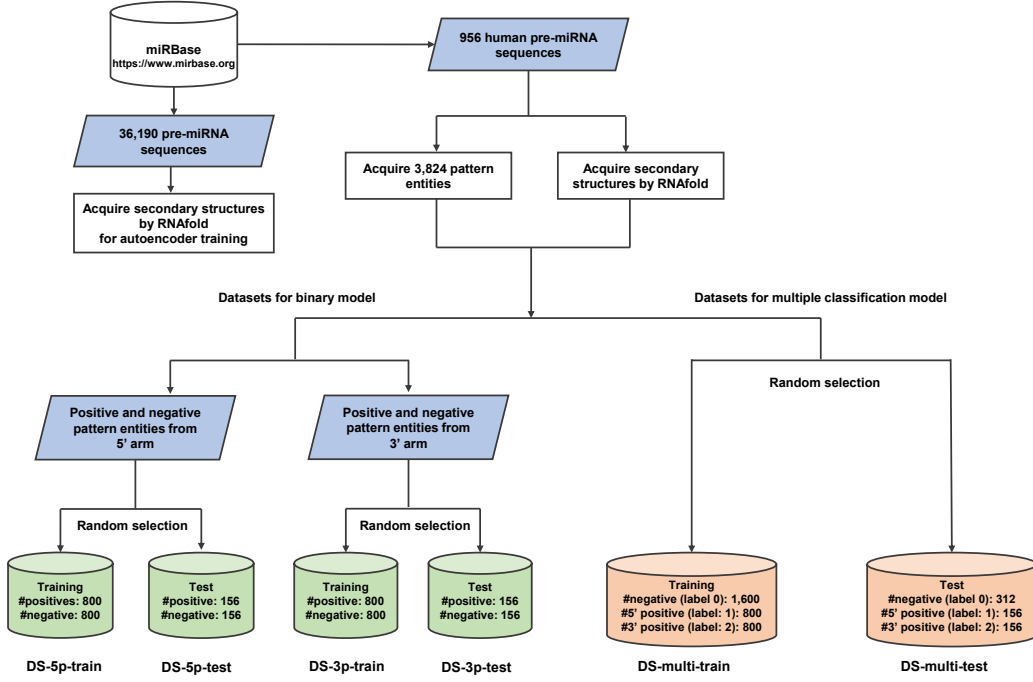


Figure 3.2: Overview of dataset preprocessing and partitioning for the classification models.

from both arms), resulting in 3,200 patterns for the training set (*DS-multi-train*) and 624 patterns for the independent test set (*DS-multi-test*). Figure 3.2 provides an overview of the data processing and splitting procedures.

3.2.3 Architecture of autoencoder

We trained an autoencoder to extract representations of pre-miRNA secondary structures. The overall architecture of the autoencoder is shown in Figure 3.3. The input of autoencoder is the dot-bracket pre-miRNA secondary structure, one-hot encoded as a 4×200 tensor (each dimension represents left bracket, right bracket, dot, and padding symbol “N”, respectively). The input tensor is first flattened by fully connected layers, after which the encoder transforms it into a 64-dimensional latent vector. This intermediate vector serves as the representation of pre-miRNA secondary structure. Then the decoder reconstructs the original input sequence tensor from the latent vector using fully connected layers.

All 36,190 pre-miRNA secondary structures were used for training. For each input, we computed the reconstruction error as the mean squared error (MSE) between the original structure and the decoded output. It is worth

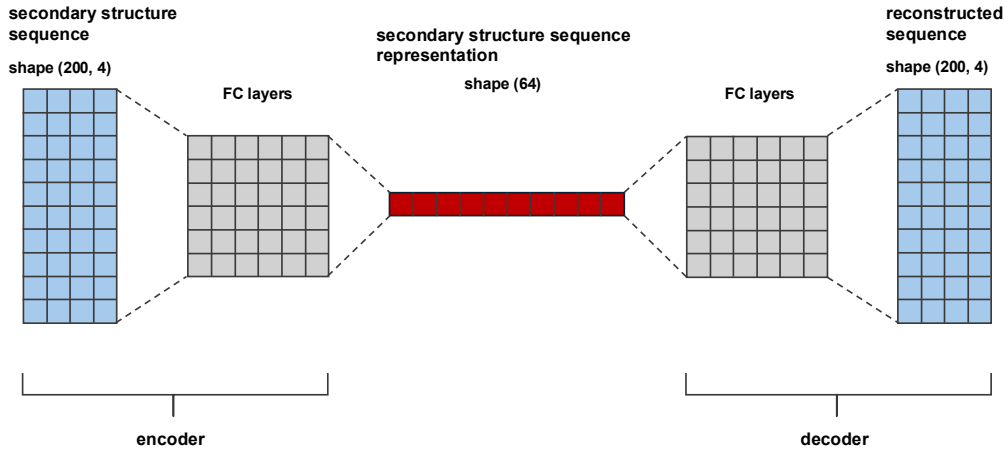


Figure 3.3: **Architecture of the autoencoder.** A fully connected autoencoder is used to generate the representations of pre-miRNA secondary structures. The input consists of one-hot encoded dot-bracket representations standardized to a fixed length of 200 nucleotides. The encoder compresses this representation into a 64-dimensional latent vector, which serves as the representation for the pre-miRNA secondary structure. The encoder and decoder consist of three FC layers, which gradually project the input to 64-dimensional vector and the latent representation to the reconstructed sequence, respectively.

noting that the goal of the autoencoder is to learn smooth and compact latent representations, rather than to model discrete nucleotide distributions. Using MSE allows the autoencoder to preserve the geometric structure of the input feature space. The decoder is only used during the training, and no biological interpretation is assigned to the reconstructed sequences.

Model optimization was performed using the Adam algorithm [76]. To promote generalization and prevent overfitting, the autoencoder was trained for 10 epochs.

3.2.4 Architecture of DiCleave

The architecture of the DiCleave consists of two modules, each comprising Convolution Unit 1 (CU1), Convolution Unit 2 (CU2), and a fully connected (FC) layer. Both binary and multi-class models share the same overall architecture; they differ only in the number of FC neurons and the configuration of the output layer. The input of DiCleave is a 13×14 tensor, which is a concatenation of Cleavage Pattern sequence (5×14), the complementary pattern sequence (5×14), and pattern secondary structure (3×14). The full

architecture is illustrated in Figure 3.4.

Module 1 is used to acquire the representation of Cleavage Pattern features. The input tensor is first processed by CU1, which is composed of three basic units. Each unit includes a one-dimensional convolutional layer followed by an activation (ACT) layer. Padding is applied in each convolution to maintain the same output dimensions as the input. A skip connection is then used to combine the output of CU1 with its input, preserving original information.

The output of CU1 is subsequently passed to CU2, which contains two basic units. Each unit consists of a convolutional layer followed by batch normalization (BN) [77] and an ACT layer. The first and second convolutional layers contain 16 and 32 filters, respectively. After feature extraction, the output is fed into an FC layer that reshapes the features into a 64-dimensional vector, which is the latent feature of concatenated Cleavage Pattern input. A dropout layer is included within the FC layer to reduce overfitting [29]. This 64-dimensional vector is then stacked with the 64-dimensional pre-miRNA secondary structure representation generated by the pre-trained autoencoder, producing a combined 2×64 tensor.

The stacked 2×64 tensor is subsequently processed by Module 2, which replicates the structure of Module 1. In this module, however, the CU2 and FC layers differ slightly between the binary and multi-class classification settings. For the multi-class classification model, CU2 includes an additional convolutional layer to project the features into a higher-dimensional latent space. This layer is followed sequentially by BN, max pooling, and ACT operations (Figure 3.4). The FC layer then transforms the features into the appropriate shape for classification.

The final output layer varies depending on the task. For binary classification, a sigmoid activation is used. Binary DiCleave model outputs 0, 1, where 0 indicates negative Cleavage Patterns, and 1 indicates positive patterns. For multi-class classification, a softmax activation is used. Multi-class DiCleave model outputs 0, 1, 2, where 0 indicates negative patterns, 1 indicates positive patterns from the 5' arm, and 2 indicates positive patterns from the 3' arm.

During training of the binary models for 5' and 3' patterns, 20% of the training patterns were set aside as a validation set, and the remaining 80% were used for training. Due to the small dataset size, we adopted a mini-batch size of 20. Binary cross-entropy (BCE) was used as the loss function. Training

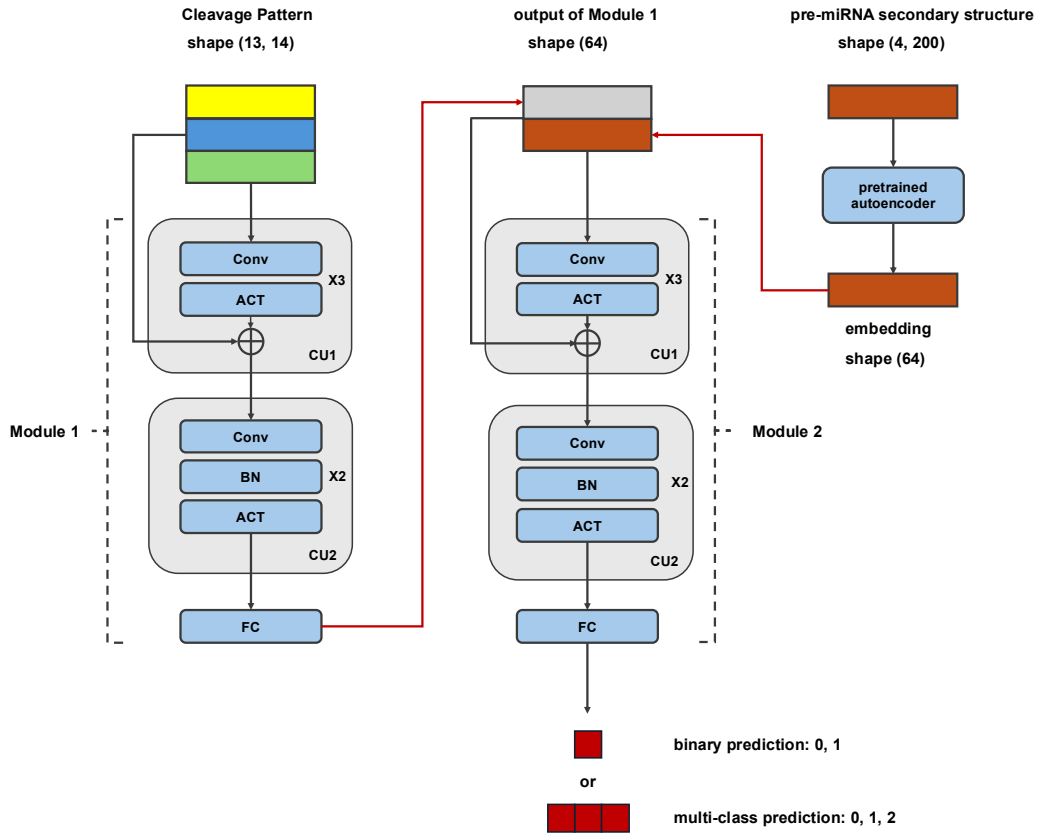


Figure 3.4: **Architecture of DiCleave.** The model architecture is divided into two sequential modules. Module 1 receives a 13-dimensional input tensor composed of one-hot encoded Cleavage Pattern features, complementary sequence features, and secondary structure inputs. The output of Module 1 is the representation of Cleavage Pattern inputs, and is then concatenated with the representation of the corresponding pre-miRNA secondary structure and passed to Module 2 for further processing. The final output layer is configured to support either binary or multi-class classification depending on the prediction task.

was performed using the Adam optimizer with a learning rate of 0.005 and a weight decay of 0.001. Training proceeded for a maximum of 50 epochs with early stopping to prevent overfitting. We retained the three models that achieved the highest validation accuracy.

The training procedure for the multi-class classification model was similar, with two notable differences. First, BCE was replaced with the negative log-likelihood (NLL) loss. Second, class weights were applied to address class imbalance: negative patterns were assigned a weight of 0.5, while positive classes retained a weight of 1.0.

3.2.5 Ablation models

To evaluate the contribution of each component in the proposed architecture, we constructed two ablation models. In Ablation Model 1 (AM1), the output of the first module was passed directly to the second module without incorporating the pre-miRNA secondary structure representations. This ablation was designed to assess the contribution of structural information extracted by the autoencoder.

In Ablation Model 2 (AM2), all CU1 components were removed from both modules. Consequently, the original inputs to each module were fed directly into CU2 without the enhancement provided by CU1. This design allowed us to evaluate the importance of CU1 in preserving and refining input information prior to convolutional processing.

Both AM1 and AM2 followed the same training procedure as the full proposed model, with only minor adjustments to the layer parameters where necessary.

3.2.6 Performance evaluation

For the binary classification task, we evaluated model performance on the independent test sets using several commonly adopted metrics: accuracy (Acc), specificity (Spe), sensitivity (Sen), F1 score, and Matthews correlation coefficient (MCC) [78]. The metrics are defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.1)$$

$$Spe = \frac{TN}{TN + FP} \quad (3.2)$$

$$Sen = \frac{TP}{TP + FN} \quad (3.3)$$

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (3.4)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (3.5)$$

where TP , TN , FP , and FN denote the counts of true positives, true negatives, false positives, and false negatives, respectively.

For the multi-class classification task, we used macro-averaged Acc, Spe, Sen, and F1 score to compute the average performance across all classes. In addition, we adopted a multi-class extension of MCC [79] defined as:

$$MCC_{\text{multi}} = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2) \times (s^2 - \sum_k^K t_k^2)}} \quad (3.6)$$

where t_k is the number of occurrences for class k ; p_k signifies the number of predictions for class k ; c denotes the total number of samples correctly predicted; s represents the total number of samples.

3.3 Results

3.3.1 Performance of binary classification models

The performance of the best models across multiple training replications is summarized in Table 3.2. The independent test sets were balanced, containing equal numbers of positive and negative patterns. The models trained on 5' and 3' patterns exhibited comparable performance: both achieved accuracy, specificity, sensitivity, and F1 scores of 0.90 or higher, and an MCC of 0.82. The receiver operating characteristic (ROC) curves and precision-recall (PR) curves for the binary classification task are shown in Figure 3.5. Although the AUC score for the 3' pattern model was slightly higher than that of the 5'

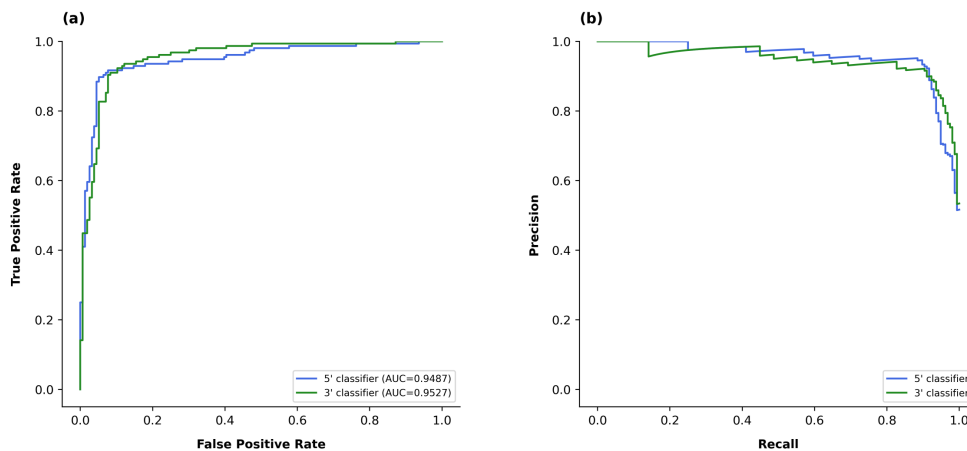


Figure 3.5: **Binary performance of DiCleave.** (a) Receiver-operating characteristic (ROC) curves and (b) precision-recall (PR) curves for DiCleave in the binary classification task. Blue and green lines represent performance of the classifiers for 5' and 3' Cleavage Patterns, respectively.

model, the overall performance of the two models can be regarded as essentially equivalent.

When evaluated on an unbalanced test set, which was constructed by randomly selecting 50 positive patterns and including all negative patterns, the model performance showed a modest decline. Nevertheless, both the F1 score and MCC remained above 0.80 and 0.75, respectively (Table 3.3 and Table 3.4), indicating that the model retained strong discriminative ability under class imbalance.

The proposed model consistently outperformed both ablation models across all classification tasks (Table 3.2). AM1, which excluded the pre-miRNA secondary structure representations, showed the lowest accuracy across tasks. Although AM1 exhibited the highest specificity in the binary classification analysis, its sensitivity was the lowest, indicating weaker ability to detect positive patterns. AM2, which removed all CU1 components, showed improved sensitivities due to the inclusion of secondary structure representations. However, its performance still lagged behind that of the full DiCleave model. This decline can be attributed to the absence of CU1, which plays a key role in preserving and enhancing input information before convolution.

To further assess the effectiveness of the proposed model, we compared

Table 3.2: Performance of DiCleave and its ablation models

		Accuracy	Specificity	Sensitivity	F1 Score	MCC
5' prediction	DiCleave	0.914	0.910	0.917	0.914	0.827
	AM1	0.878	0.917	0.840	0.873	0.759
	AM2	0.891	0.878	0.904	0.892	0.782
3' prediction	DiCleave	0.910	0.917	0.904	0.910	0.820
	AM1	0.878	0.930	0.827	0.872	0.760
	AM2	0.894	0.897	0.891	0.894	0.788
Multi-class	DiCleave	0.894	0.940	0.890	0.893	0.830
	AM1	0.869	0.927	0.869	0.868	0.790
	AM2	0.873	0.930	0.887	0.874	0.803

AM1: Ablation model without pre-miRNA secondary structure representations

AM2: Ablation model without convolution unit 1 (CU1)

DiCleave with ReCGBM. We computed the average performance across 10 training replications with different random initializations. As shown in Figure 3.6 and Table 3.5, DiCleave surpassed the average performance of ReCGBM in predicting 5' patterns. Moreover, while DiCleave demonstrated comparable performance in predicting 5' and 3' patterns, ReCGBM exhibited a performance gap, showing inferior accuracy for 5' pattern prediction relative to 3' pattern prediction. Figure 3.6 also presents performance comparisons with PHDcleav [12] and LBSIZEcleav [13]. PHDcleav uses binary sequence features and a 14-nt window [12], whereas LBSIZEcleav incorporates loop/bulge size information using $k = 1$ and a 14-nt window [13]. We included these models because their input feature representations are most similar to those used in our architecture. As illustrated in Figure 3.6, both SVM-based models performed substantially worse than ReCGBM and DiCleave.

3.3.2 Performance of the multi-class classification model

Our model was extended to perform multi-class classification by modifying the output layer to predict whether a Cleavage Pattern corresponds to a 5' pattern, a 3' pattern, or a negative pattern. The performance of the multi-class classification model is summarized in Table 3.2. The model achieved high overall accuracy and specificity, as well as a macro-averaged F1 score of 0.89. When evaluated on an unbalanced dataset, the model maintained strong performance, with both the F1 score and MCC exceeding 0.85 and 0.75, respectively (Table 3.3 and Table 3.4).

Table 3.3: Description of class-unbalanced datasets and pre-miRNA sequence similarity cutoff datasets

Dataset	# of Total Samples	# of Positive Samples	# of 5' Samples	# of 3' Samples	# of Negative Samples
DS-5p-ub	206	50	50	0	156
DS-3p-ub	206	50	0	50	156
DS-multi-ub	412	100	54	46	312
DS-80%-5p-train	1256	666	666	0	599
DS-80%-5p-test	300	166	166	0	134
DS-80%-5p-test-ub	184	50	50	0	134
DS-80%-3p-train	1235	641	0	641	594
DS-80%-3p-test	300	164	0	164	136
DS-80%-3p-test-ub	186	50	0	50	136
DS-80%-multi-train	2650	1401	719	682	1249
DS-80%-multi-test	450	236	113	123	214
DS-80%-multi-test-ub	314	100	53	47	214

DS-5p-ub: unbalanced test set derived from the original 5' model test set; DS-3p-ub: unbalanced test set derived from the original 3' model test set; DS-multi-ub: unbalanced test set derived from the original multi-classification model test set; DS-80%-5p-train: training set with 80% similarity cutoff of 5' model; DS-80%-5p-test: test set with 80% similarity cutoff of 5' model; DS-80%-5p-test-ub: unbalanced test set with 80% similarity cutoff of 5' model; DS-80%-3p-train: training set with 80% similarity cutoff of 3' model; DS-80%-3p-test: test set with 80% similarity cutoff of 3' model; DS-80%-3p-test-ub: unbalanced test set with 80% similarity cutoff of 3' model; DS-80%-multi-train: training set with 80% similarity cutoff of multi-classification model; DS-80%-multi-test: test set with 80% similarity cutoff of multi-classification model; DS-80%-multi-test-ub: unbalanced test set with 80% similarity cutoff of multi-classification model.

Table 3.4: Performance of the best DiCleave model on unbalanced test sets

Dataset	Accuracy	Specificity	Sensitivity	F1 Score	MCC
DS-5p-ub	0.903	0.910	0.880	0.815	0.753
DS-3p-ub	0.922	0.917	0.940	0.854	0.808
DS-multi-ub	0.903	0.937	0.894	0.854	0.776

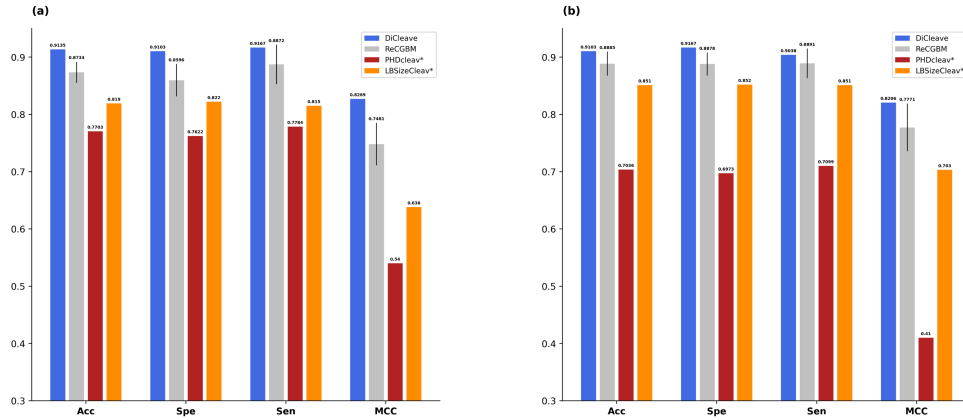


Figure 3.6: **Binary performance comparison between DiCleave and existing models.** Comparison of DiCleave with previously proposed methods for binary classification of (a) 5' cleavage site prediction and (b) 3' cleavage site prediction. For DiCleave and ReCGBM, the values represent the average performance across 10 independent replications. The performance metrics for PHDcleav and LBSIZEcleav are taken from their original publications (indicated by asterisks).

Table 3.5: Performance of DiCleave and ReCGBM for 10 replications

Model		Accuracy	Specificity	Sensitivity	F1 Score	MCC
DiCleave	5' prediction	Best	0.942	0.910	0.913	0.827
		Avg.	0.885 (0.011)	0.870 (0.028)	0.883 (0.030)	0.772 (0.028)
	3' prediction	Best	0.949	0.910	0.913	0.820
		Avg.	0.880 (0.014)	0.865 (0.040)	0.880 (0.030)	0.762 (0.040)
ReCGBM	Multi-class prediction	Best	0.941	0.895	0.891	0.828
		Avg.	0.863 (0.015)	0.861 (0.018)	0.862 (0.015)	0.781 (0.024)
	5' prediction	Best	0.904	0.930	-	0.802
		Avg.	0.873 (0.018)	0.887 (0.034)	-	0.748 (0.037)
	3' prediction	Best	0.930	0.930	-	0.859
		Avg.	0.888 (0.021)	0.889 (0.026)	-	0.777 (0.042)

Numbers in parentheses indicate standard deviation.

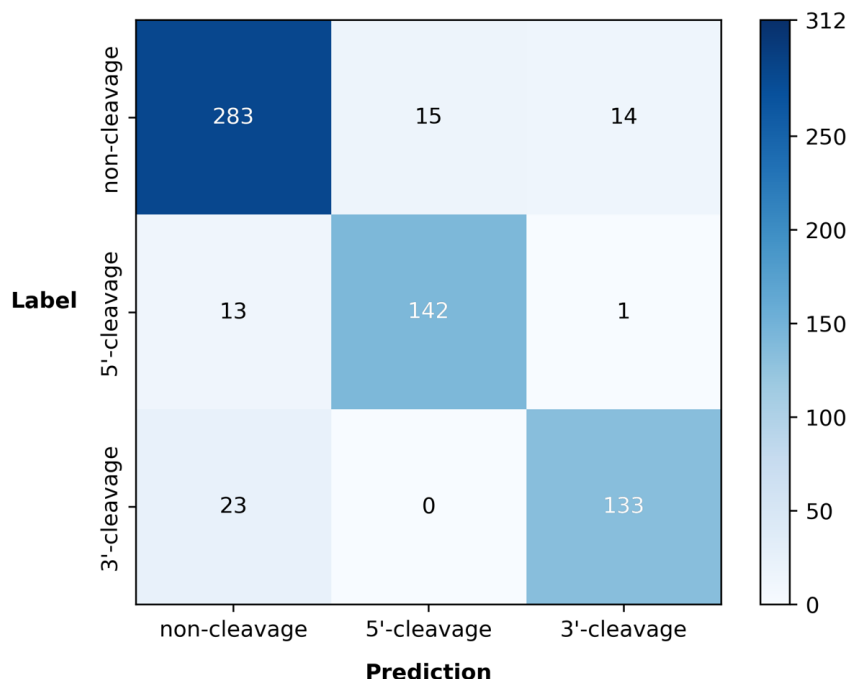


Figure 3.7: **Confusion matrix for the multi-class classification model.** Rows and columns correspond to the true and predicted labels, respectively. Values along the diagonal represent the number of correctly classified instances for each class.

The confusion matrix for the multi-class classification model is shown in Figure 3.7. Only a single 5' pattern was misclassified as a 3' pattern, and no 3' patterns were misclassified as 5' patterns, demonstrating that the model effectively distinguishes between patterns originating from the two arms of pre-miRNAs. However, misclassification between positive and negative patterns was more common, particularly the misclassification of 3' patterns as non-cleavage patterns, indicating remaining challenges in differentiating subtle structural or sequence features between these classes.

Figure 3.8 presents the ROC and PR curves for the multi-class model using a one-vs-all evaluation strategy. The AUC scores for both the 5' and 3' classes reached 0.97, outperforming the negative class (AUC = 0.93). In the PR analysis, precision for 3' patterns declined more rapidly as recall increased compared with the 5' and negative classes, reflecting the difficulty in identifying 3' Cleavage Patterns relative to the other classes.

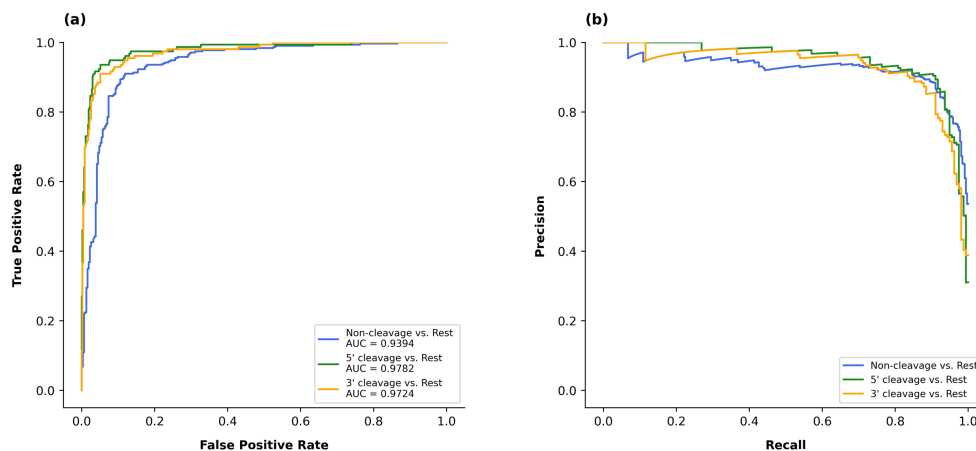


Figure 3.8: **Multi-class classification performance of DiCleave.** (a) Receiver-operating characteristic (ROC) curves and (b) precision-recall (PR) curves for the multi-class classification model. For evaluation, the model was converted into a one-vs-all classifier by treating one class as the positive class and the remaining two classes as negative.

3.4 Discussion

In this study, we introduced a deep neural network model—DiCleave—for predicting the presence of a human Dicer cleavage site within a short sequence segment defined as a Cleavage Pattern. Our model integrates one-hot encoded representations of nucleotide sequences, complementary sequences, and secondary structure information. Rather than merging these components into twelve dimer-based features combining nucleotide symbols with secondary structure symbols (“A”, “C”, “G”, “U” $\times$ “(”, “:”, “)”), we opted to stack the encoded matrices. This design avoids creating a sparse input space that may lead to unstable or biologically uninformative convolutional outputs. The incorporation of secondary structure representations extracted by the autoencoder improved the discriminative capability of the model, enabling DiCleave to achieve accuracies of 0.91 and 0.89 in the binary and multi-class classification tasks, respectively. These results demonstrate that leveraging pre-miRNA secondary structure and the skip-connection-based CU1 architecture substantially enhances predictive performance. Notably, although the multi-class model was trained on an imbalanced dataset where negative patterns outnumbered positive patterns, it still achieved strong classification performance with only

minimal loss function weight adjustment applied to the negative class.

Our computational experiments involved random training-test splits of the main dataset. To further assess the robustness of the model, we filtered sequences using CD-HIT-EST [80, 81] to remove human pre-miRNAs sharing $\geq 80\%$ sequence similarity. Even under this stricter dataset configuration, DiCleave maintained strong performance: the best model among 10 independent replications achieved F1 scores above 0.85 for both binary and multi-class tasks under balanced conditions, and above 0.80 under unbalanced conditions (Table 3.6). This suggests that the model generalizes well even when trained on more diverse and less redundant datasets.

However, when comparing performance across 10 replications, we observed that although the best-performing DiCleave model outperformed the best ReCGBM model, there was no significant difference in average performance between the two methods. This observation suggests that further improvements are still needed to enhance prediction accuracy and model robustness.

One limitation of this study is the relatively small size of the human pre-miRNA dataset available for training. This constraint contributed to overfitting during early experiments. To mitigate this, we adopted a small mini-batch size of 20, as larger batches often caused the model to converge to poor local minima. We selected a learning rate of 0.005, which is higher than the PyTorch default, after observing that lower learning rates (e.g., 10^{-4} or 5×10^{-5}) hindered convergence, while higher learning rates increased overfitting risk. As for data augmentation, although it is a common approach to address small or imbalanced datasets, effective augmentation strategies for nucleotide sequences require careful biological consideration and were beyond the scope of this study.

Another limitation concerns the definition of the Cleavage Pattern itself. In this work, a Cleavage Pattern is defined as a 14-nt sequence with the cleavage site fixed at the central position. In real biological contexts, however, cleavage may occur at variable positions within a segment. Considering all possible cleavage positions would substantially expand the training dataset and potentially improve model robustness, but at the cost of a more complex data generation process.

Finally, Liu et al. reported that features near the center of pre-miRNAs

Table 3.6: Performance of models trained from similarity cutoff datasets

Dataset		Accuracy	Specificity	Sensitivity	F1 Score	MCC
DS-80%-5p-test	Best	0.893	0.955	0.843	0.897	0.794
	Avg.	0.860 (0.015)	0.867 (0.061)	0.855 (0.044)	0.871 (0.015)	0.723 (0.031)
DS-80%-5p-test-ub	Best	0.929	0.955	0.860	0.869	0.820
	Avg.	0.871 (0.036)	0.866 (0.061)	0.882 (0.037)	0.791 (0.040)	0.711 (0.055)
DS-80%-3p-test	Best	0.867	0.853	0.878	0.878	0.731
	Avg.	0.838 (0.012)	0.821 (0.032)	0.852 (0.031)	0.851 (0.013)	0.674 (0.024)
DS-80%-3p-test-ub	Best	0.882	0.860	0.940	0.810	0.742
	Avg.	0.849 (0.020)	0.821 (0.032)	0.926 (0.030)	0.768 (0.022)	0.684 (0.030)
DS-80%-multi-test	Best	0.876	0.934	0.885	0.878	0.807
	Avg.	0.850 (.018)	0.920 (0.010)	0.859 (0.019)	0.852 (0.018)	0.769 (0.026)
DS-80%-multi-test-ub	Best	0.882	0.914	0.840	0.842	0.757
	Avg.	0.836 (0.028)	0.908 (0.011)	0.850 (0.020)	0.805 (0.025)	0.701 (0.036)

Numbers in parentheses indicate standard deviation.

play an especially important role in cleavage site determination [14], suggesting the presence of long-range or context-dependent interactions within the sequence. In light of this, future work will explore Transformer-based architectures, which can explicitly model such interactions through attention mechanisms [7]. Additionally, we plan to investigate end-to-end generative models capable of producing mature miRNA sequences directly from pre-miRNAs, building on the cleavage prediction framework developed in this study.

Chapter 4

DiCleavePlus: a Transformer-based model to detect human Dicer cleavage sites within Cleavage Pattern

4.1 Background

In Chapter 3, we introduced DiCleave, a deep learning model for predicting human Dicer cleavage sites, and demonstrated the effectiveness of deep neural networks in identifying the presence of cleavage sites within short pre-miRNA subsequences. Similar to previous approaches [12, 13, 14], DiCleave relies on the concept of Cleavage Pattern, typically defined as a 14-nucleotide subsequence extracted from a pre-miRNA (Figure 4.1). Models based on Cleavage Patterns are trained to classify whether a given subsequence contains a Dicer cleavage site. In contrast, DeepMirCut [15], an RNN-based framework, bypasses the Cleavage Pattern design and directly predicts Drosha and Dicer cleavage sites from full-length pre-miRNA sequences. However, despite its broader input scope, its prediction accuracy remains suboptimal.

Despite Cleavage Pattern-based models demonstrate superior performance, they have a critical limitation: these models are restricted to detecting Dicer cleavage sites only when they occur at the central position of the patterns (Figure 4.1, upper, black box). For instance, the human pre-miRNA hsa-let-7a, which consists of 80 nucleotides (nt), can generate 67 Cleavage Patterns when the pattern length is set to 14 nt (Figure 4.2). Among these, 26 patterns actually contain a Dicer cleavage site, yet existing models can identify only 2 patterns where the cleavage site lies at the center (Figure 4.2, black boxes). Consequently, these models oversimplify the prediction as a binary classification task (i.e., whether a pattern contains a cleavage site), without fully exploring the positional diversity of cleavage sites. Moreover, despite reporting binary accuracies around 0.90, these models still face a non-negligible risk of misidentifying true cleavage sites. Therefore, there remains considerable scope for improving Dicer cleavage site prediction.

Transformer [7] applies the multi-head attention mechanism to alleviate the gradient vanishing problem associated with long input sequences. With its success in natural language processing, Transformer has become a popular architecture in bioinformatics [82]. Recent works from [83], [84], and [85] employed the encoder part of Transformer as feature extractors to exploit hidden features from input data, taking advantage of the strong capacity to capture internal relationships.

In this study, we introduce DiCleavePlus, a Transformer-based model for human Dicer cleavage site prediction. Compared with previous works, DiCleavePlus extends Cleavage Pattern by allowing cleavage sites to appear at any position within the pattern, which maximizes the informational content extracted from pre-miRNA sequences. DiCleavePlus incorporates both the extended Cleavage Pattern and the corresponding full-length pre-miRNA as joint inputs and employs Transformer encoders to extract features from both sequence and secondary structure representations. Benchmarking experiments demonstrate that DiCleavePlus achieves high accuracy and robustness in cleavage site prediction. By allowing flexible cleavage site positioning and integrating multi-level RNA information, DiCleavePlus maximizes the utility of pre-miRNA features. These improvements highlight its potential as a practical tool for guiding experimental design, narrowing the search space of candidate sites, and ultimately facilitating downstream studies of miRNA biogenesis and

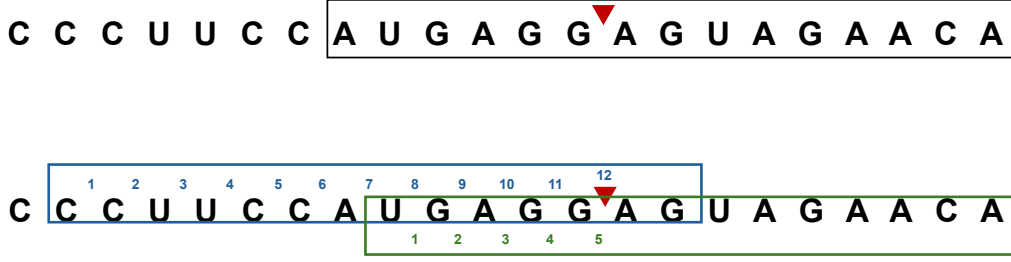


Figure 4.1: **Concept of extended Cleavage Pattern.** Conventional Cleavage Patterns (upper, black box) restrict the Dicer cleavage site (red triangle) to the central position. In DiCleavePlus, Cleavage Patterns are generated using a sliding window approach along the pre-miRNA sequence, allowing the cleavage site to appear at any position within the pattern (lower, blue and green boxes). As examples, the blue pattern is labeled as class 12 and the green pattern as class 5.

function.

4.2 Methods

4.2.1 Extended Cleavage Pattern

A conventional Cleavage Pattern is defined as a short subsequence of a pre-miRNA sequence. Patterns in which the cleavage site is located at the central position are considered positive, whereas those do not contain cleavage site are labeled negative. Typically, a Cleavage Pattern is 14 nucleotides in length, resulting in 13 intervals between adjacent nucleotides (Figure 4.1). Previous models based on conventional Cleavage Patterns can only detect one type of positive pattern where the cleavage site is positioned at the central interval (Figure 4.1, upper, black box).

To overcome this limitation, we expanded the definition of positive Cleavage Patterns by allowing the cleavage sites to appear at any intervals within the pattern (Figure 4.1, lower, blue and green boxes). In this study, each interval index is regarded as a unique class. Specifically, class 0 represents negative patterns (i.e., patterns with no cleavage site), while classes 1 to k correspond to cleavage sites located at interval positions within the patterns.

With this extension, Dicer cleavage site prediction task is converted into a multi-class classification problem, where the model predicts the index of the interval containing the cleavage site or assigns class 0 if no cleavage site is present. To investigate the impact of pattern size on model performance,

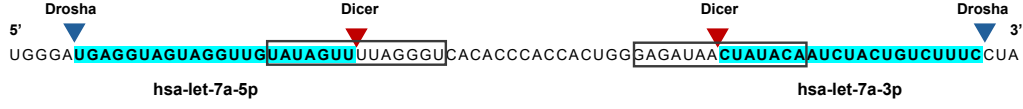


Figure 4.2: **Linear representation of human pre-miRNA hsa-let-7a.** The blue-shaded regions correspond to the two mature miRNAs, hsa-let-7a-5p and hsa-let-7a-3p. The positions of the Dicer cleavage sites (red triangles) and Drosha cleavage sites (blue triangles) are inferred from the boundaries of these mature miRNAs.

Cleavage Patterns of 10, 12, 14, 16 and 18 nt were generated and used in subsequent experiments.

4.2.2 Data preparation

Data collection

A total of 956 human pre-miRNAs that explicitly contain two mature miRNAs were retrieved from the miRBase database (Release 22.1). As illustrated in Figure 4.2, the positions of Dicer cleavage sites can be reliably inferred from the annotated mature miRNAs regions within these pre-miRNAs. To reduce sequence redundancy, CD-HIT-EST [80, 81] was applied using an 80% similarity threshold. After filtering, 812 pre-miRNA sequences were retained, each containing two Dicer cleavage sites.

For each pre-miRNA sequence, all possible Cleavage Patterns were generated in a sliding window manner. Specifically, if the pre-miRNA has a length of L and the pattern size is denoted by s , then it yields in total of $L - s + 1$ Cleavage Patterns. Patterns that inadvertently encompassed two cleavage sites, which occurred when the 5' and 3' cleavage sites were in close proximity, were removed to ensure that each pattern contained at most one cleavage site. If a pattern contained a cleavage site at its k -th interval, it was assigned to class k ; otherwise, it was labeled as class 0. In our datasets, each sample entry is a Cleavage Pattern, which is associated with other information, including the corresponding pattern secondary structure, the full-length pre-miRNA sequence from which it originates, and the secondary structure of that pre-miRNA. The detailed configuration of dataset is presented in Table 4.1.

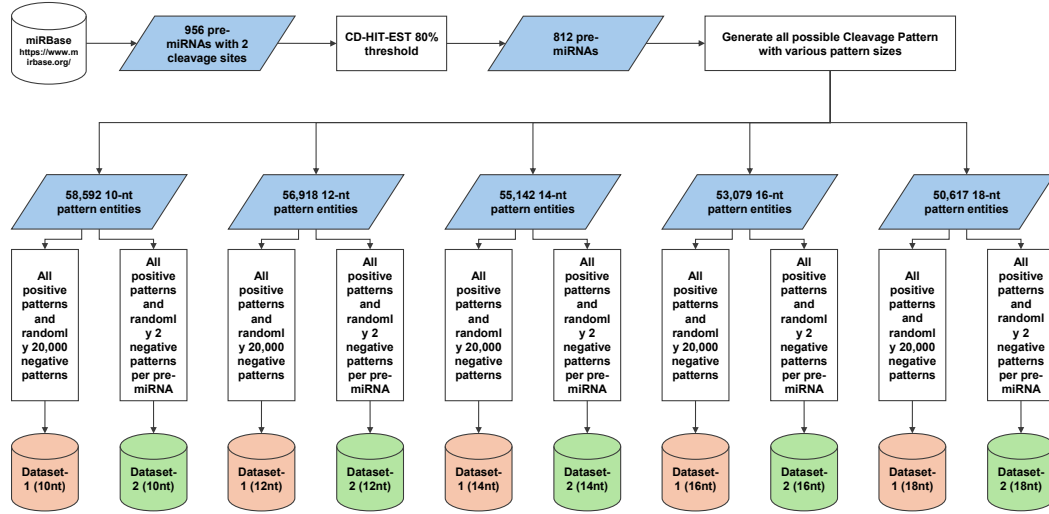


Figure 4.3: **Workflow for dataset construction.**

To construct Dataset-1, all positive patterns were retained, and 20,000 negative patterns were randomly sampled. Because negative class substantially outnumbered each positive class in Dataset-1, a class-balanced dataset, Dataset-2, was constructed by randomly selecting two negative patterns per pre-miRNA. To investigate the influence of pattern size on model performance, we set $s \in \{10, 12, 14, 16, 18\}$, resulting in five variants of Dataset-1 and Dataset-2 using different pattern size s . A visual overview of the data collection and processing pipeline is provided in Figure 4.3. Figure 4.4 summarizes the number of patterns per class in each variant of Dataset-1 and Dataset-2.

Cross-validation strategy

A 5-fold cross-validation strategy was employed to evaluate the performance of DiCleavePlus. The dataset was randomly partitioned into five equal subsets at the sample (Cleavage Pattern) level, meaning that each individual Cleavage Pattern was treated as an independent unit when assigning samples to subsets. Each Cleavage Pattern was characterized by four components: the pattern sequence, the corresponding pattern secondary structure, the full-length pre-miRNA sequence from which it originates, and the secondary structure of that pre-miRNA. DiCleavePlus was designed to incorporate all four components as its inputs, ensuring that no two input samples are identical.

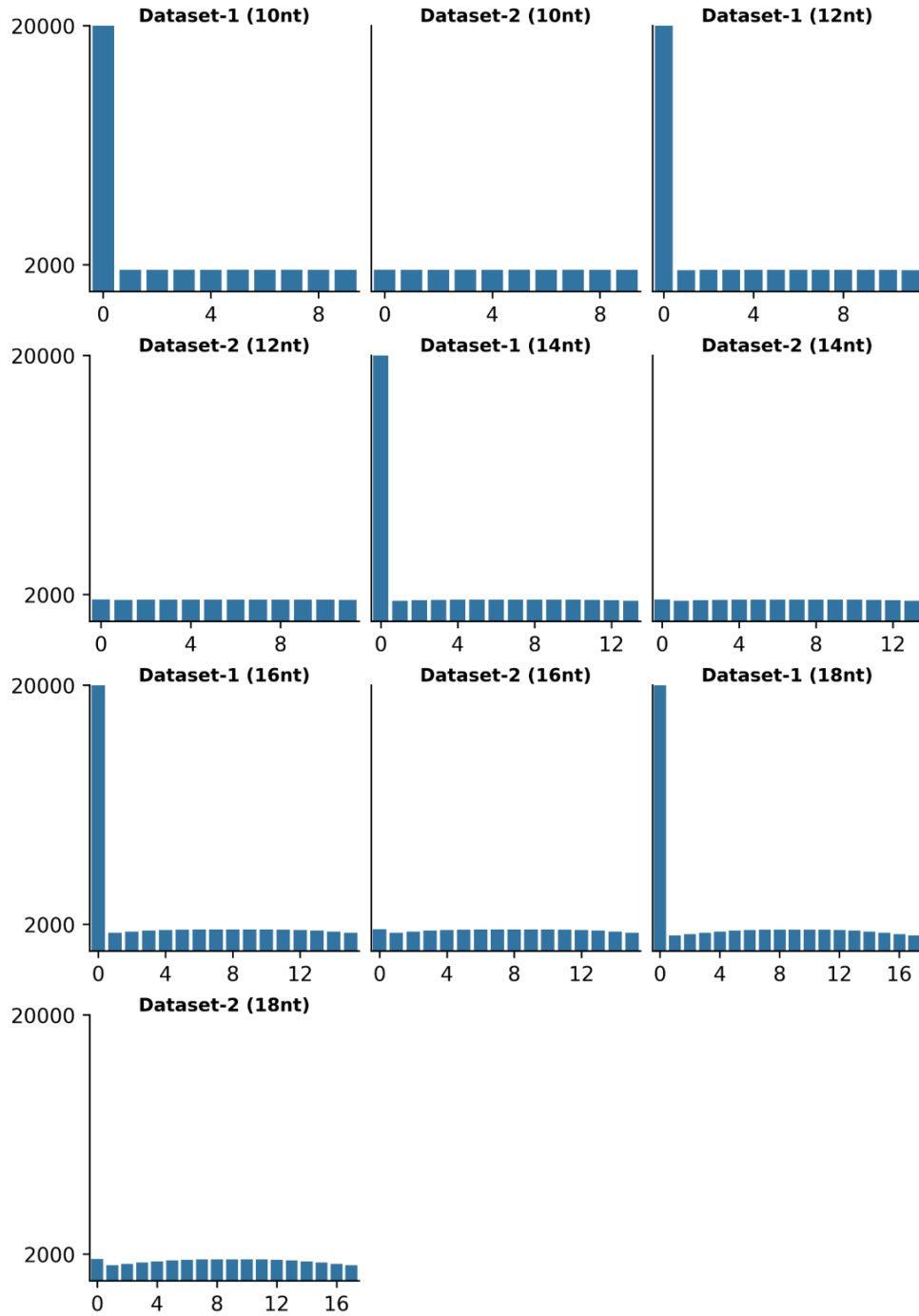


Figure 4.4: **Class distribution of patterns across different datasets.** The x-axis represents class labels, and the y-axis indicates the number of patterns. In Dataset-1, the number of negative patterns exceeds that of each positive class, while the total number of positive patterns is approximately equal to the number of negative patterns. In Dataset-2, the number of negative patterns is nearly equal to that of each individual positive class; however, the total number of positive patterns exceeds the number of negative patterns.

In each fold, one subset was held out as the test set, while the remaining four subsets were further split into a training set and a validation set, with 90% of the data allocated for training and 10% for validation to monitor the training process. This process was repeated five times such that each subset served once as the test set, and the final performance was reported as the average across the five folds.

4.2.3 Data processing

Pre-miRNA secondary structure

The secondary structures of pre-miRNAs were predicted using RNAfold [75], and represented in dot-bracket notation. Because RNAfold predicts pseudo-knot-free structures, only three types of symbols are used in the output: an opening bracket “(” and a closing bracket “)” to represent base-paired nucleotides, and a dot “.” to indicate unpaired nucleotides. An example of this representation is shown in Figure 4.5 (a), green part.

K-mer encoding

We utilized 3-mer token to represent the pre-miRNA sequences and the corresponding secondary structures. For each sequence, overlapping 3-mer tokens were generated. Since the last two elements of a sequence cannot form complete 3-mers, placeholder symbols were appended to ensure that the total number of 3-mer tokens matched the original sequence length. For instance, the penultimate and the final nucleotides of a pre-miRNA sequence are represented as “NN<PH>” and “N<PH><PH>”, respectively, where “N” denotes any nucleotide, and “<PH>” indicates a placeholder.

Based on this scheme, a vocabulary of 84 unique 3-mer tokens was constructed for nucleotide sequences, while the dot-bracket secondary structures formed a vocabulary of 39 unique 3-mer tokens. Each 3-mer token was subsequently embedded into a 32-dimensional continuous vector space and used as the input representation for both sequence and secondary structure (Figure 4.5 (a))

Input features of DiCleavePlus

To ensure a consistent input length across samples, the embeddings vectors of pre-miRNA sequences and their corresponding secondary structure were padded to a fixed length of 200. The sequence embeddings (32×200) were then concatenated with the corresponding secondary structure embeddings (32×200) to form the sequence feature tensors (64×200). To preserve positional information that is essential for sequential modeling, sinusoid positional encodings [7] were added to the combined embeddings.

A similar procedure was applied to the Cleavage Pattern inputs. The embeddings of each Cleavage Pattern ($32 \times s$, s denotes pattern size) were concatenated with their respective secondary structure embeddings ($32 \times s$), followed by the addition of sinusoid positional encodings. These Cleavage Pattern features ($64 \times s$), together with the pre-miRNA sequence features, served as the dual input streams to DiCleavePlus (Figure 4.5 (b)).

4.2.4 Model architecture

The overall architecture of DiCleavePlus is illustrated in Figure 4.6. DiCleavePlus utilizes Transformer-based Encoder Units to process Cleavage Pattern features and full-length pre-miRNA sequence features, respectively. Each Encoder Unit begins with three standard Transformer encoder layers, each consisting of a multi-head self-attention layer followed by a feed-forward layer [7]. Following the Transformer layers, two convolutional blocks are applied to further refine the representations. Each block is composed of a 1-dimensional convolutional layer followed by a leaky ReLU activation layer [86]. Residual connections [87] are added between the output of the Transformer layer and each convolutional block to facilitate gradient flow and preserve feature information. A final linear layer is applied to project the processed features into a fixed-dimensional representation.

To integrate the processed pattern features and pre-miRNA sequence features, we designed two fusion strategies. In the first strategy, we implemented the attentional feature fusion (AFF) module from [88] to integrate these features. Models that used this strategy are hereafter denoted as DCP-AFF. In the alternative strategy, we employed concatenation operation to merge pattern features and pre-miRNA features. Models adopting this strategy are referred to as DCP-Concat. It is worth noting that the dimensionality of the

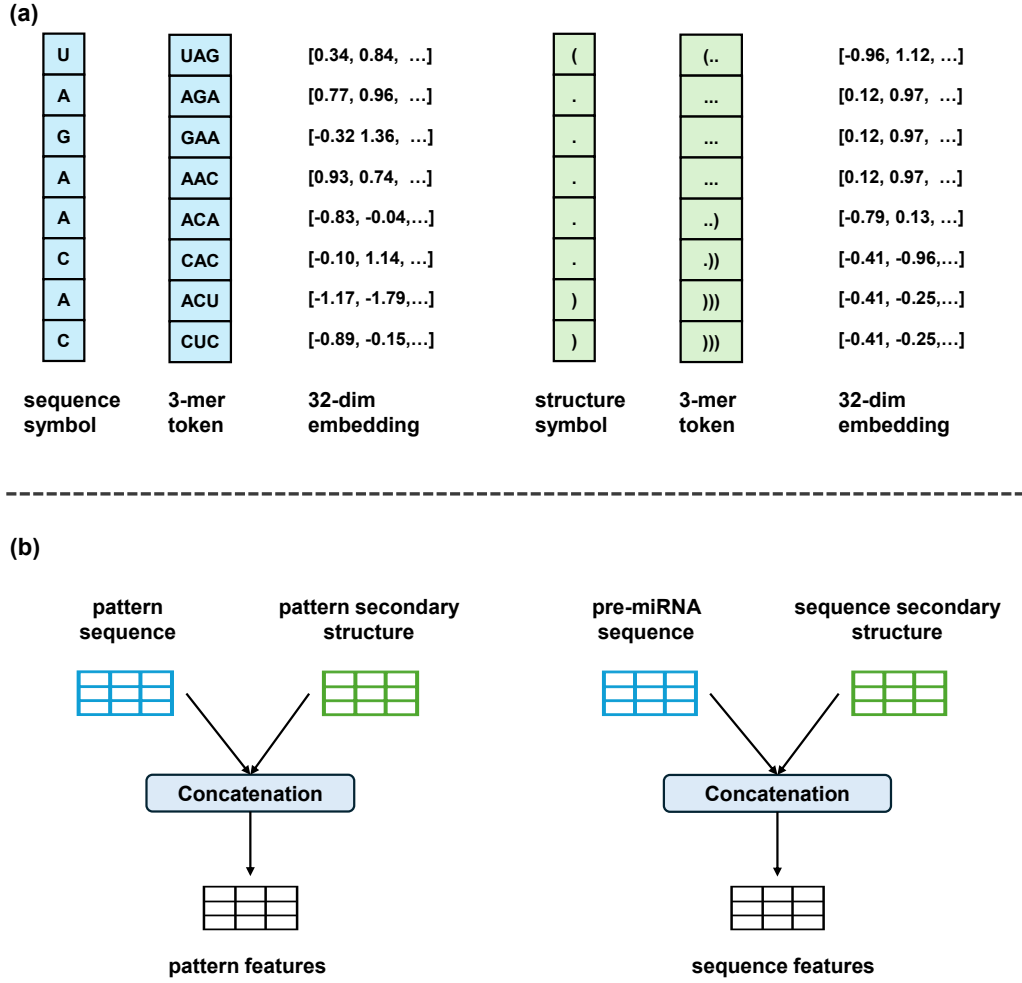


Figure 4.5: **Data processing workflow of DiCleavePlus.** (a) Pre-miRNA sequence (blue blocks) and their corresponding secondary structures (green blocks) are tokenized into 3-mers and encoded using 32-dimensional embeddings. (b) The sequence embeddings are concatenated with the secondary structure embeddings to generate the input features of DiCleavePlus.

outputs of AFF module and concatenation operation is different. Assume that the two input features have dimension d , since the essence of AFF is weighted sum, the output of AFF also has dimension d . Concatenation operation stacks two input features and therefore outputs $2d$ tensor as merged feature.

After feature fusion, the combined representation is passed through a fully connected block and then fed into the output layer. A softmax function is applied to predict the index of interval where the Dicer cleavage site is located in the Cleavage Pattern (or class 0 for negative patterns).

4.2.5 Attentional feature fusion

Attentional feature fusion (AFF) from [88] was utilized to integrate Cleavage Pattern features and pre-miRNA sequence features processed by the Encoder Unit. The core component of AFF, termed the multi-scale channel attention block, computes global and local contexts of input features separately. These contexts are subsequently aggregated and passed through a sigmoid function to calculate the attention weight. AFF then fuses the two weighted input features by summing them.

$$Z = M(X \oplus Y) \otimes X + (1 - M(X \oplus Y)) \otimes Y \quad (4.1)$$

where Z denotes the output fused feature; X and Y indicate two input features; $M()$ denotes multi-scale channel attention block. The output of $M()$ is a tensor with value in the range of $[0, 1]$, which is used as the weight for X ; $1 - M()$ is used as the weight of Y , which is complementary to the weight of X ; $\oplus$ denotes element-wise addition; $\otimes$ denotes element-wise multiplication.

The AFF module implemented in DiCleavePlus is illustrated in Figure 4.7.

4.2.6 Comparison with ablation models

To evaluate the contribution of Transformer-based feature extractors within DiCleavePlus, we constructed three ablation models by replacing the Encoder Unit with alternative architectures: a multi-layer perceptron (MLP), a convolutional neural network (CNN), and a recurrent neural network (RNN). For the CNN-based ablation model, sinusoidal positional encodings were additionally incorporated to provide explicit positional information for the input sequences, which is not inherently captured by standard convolutional operations. All

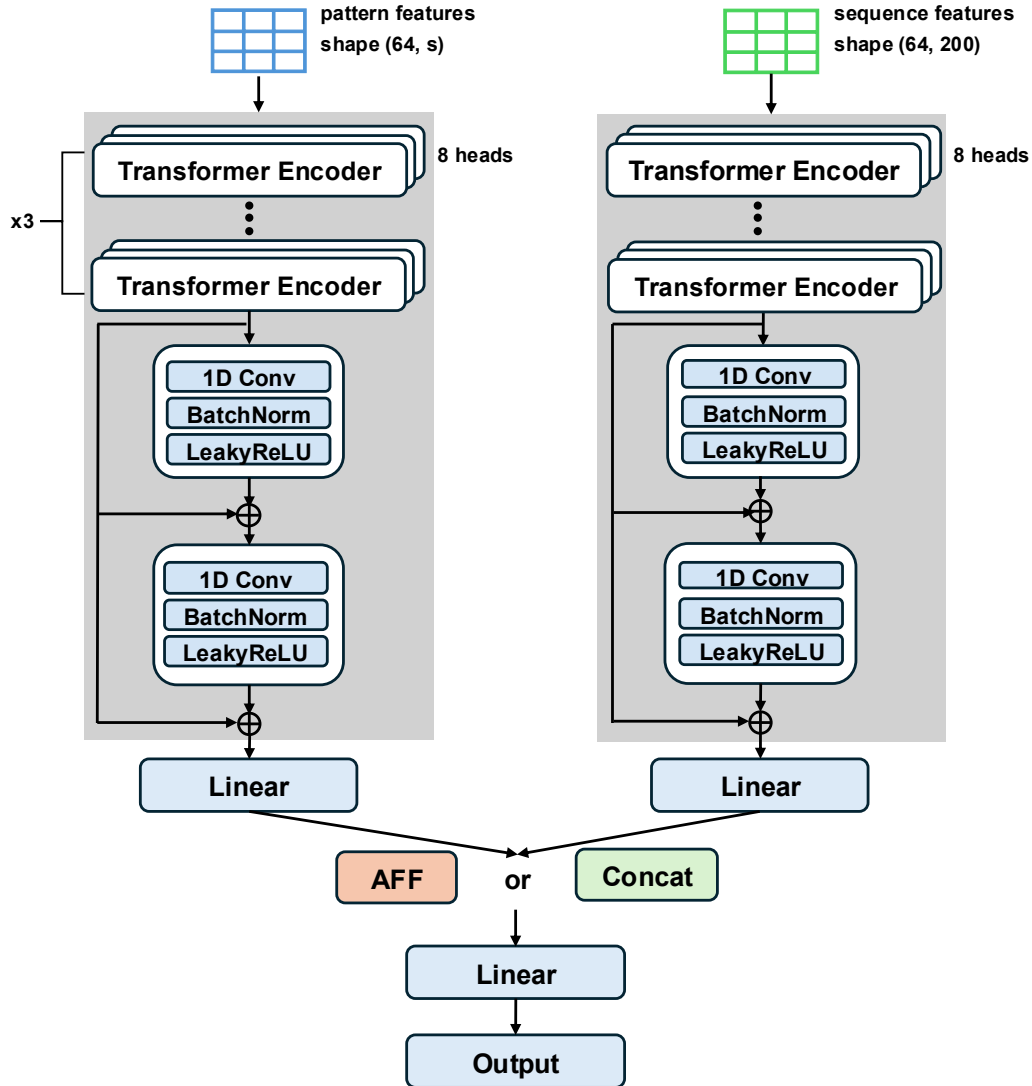


Figure 4.6: **Overview of the DiCleavePlus architecture.** The grey boxes represent the Transformer-based Encoder Units, which are used to extract features from Cleavage Pattern and full-length pre-miRNA sequences. Two feature integration strategies were implemented. Models employing the attentional feature fusion (AFF) block to integrate processed pattern and sequence features are referred to as DCP-AFF, whereas models that merge these features using simple concatenation are referred to as DCP-Concat.

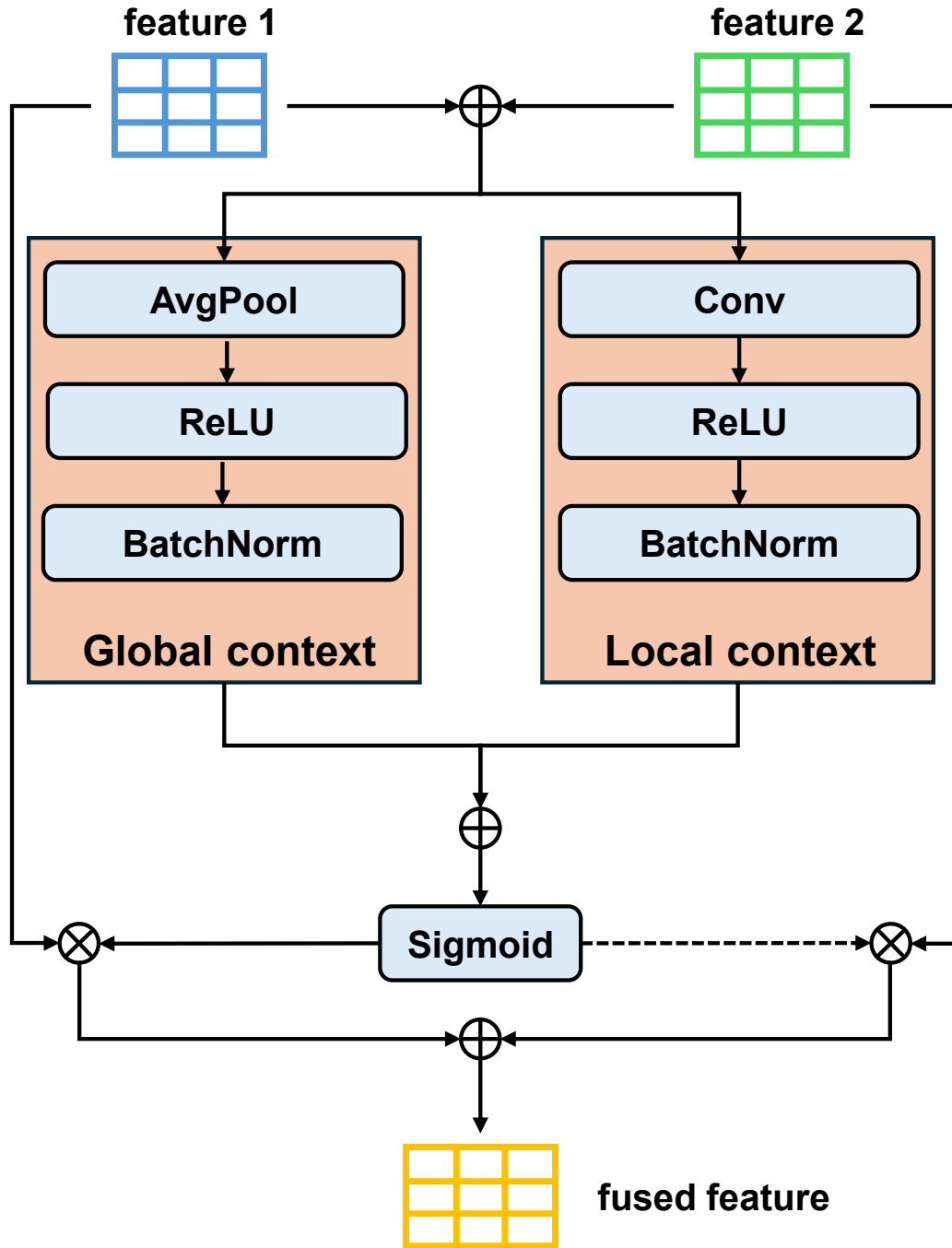


Figure 4.7: **Architecture of attentional feature fusion (AFF) block.** The orange module represents the multi-scale channel attention block. The symbol $\oplus$ denotes element-wise addition, and $\otimes$ indicates element-wise multiplication. The dash line indicates the computation of complementary weight $(1 - w)$.

ablation models were trained and evaluated under the same experimental settings as DiCleavePlus, using both 14-nt and 18-nt datasets to ensure a fair comparison.

4.2.7 Comparison with other models

Although DiCleave employs a multi-class output layer, its prediction task is essentially binary, as it only predicts whether a Dicer cleavage site is located at the center of a given Cleavage Pattern. To ensure a fair comparison with DiCleavePlus, we modified DiCleave by replacing its output layer with that of our proposed model, thereby adapting it to the multi-class setting of this study.

Similarly, we adapted ReCGBM, another state-of-the-art architecture, by modifying its source code to enable multi-class classification, as the original model was designed exclusively for binary predictions.

Since DiCleave and ReCGBM were originally designed to process Cleavage Patterns of 14 nt, for comparative evaluation, we retrained and evaluated these models using the same 14-nt datasets employed in this study. All training procedures and hyperparameter configurations for DiCleave and ReCGBM were preserved as in their original implementations, ensuring the validity and fairness of performance comparison with DiCleavePlus.

4.2.8 Details of training procedure

A 5-fold cross-validation strategy was employed to evaluate model performance. The dataset was randomly partitioned into five equal subsets. In each fold, one subset was held out as the test set, while the remaining four subsets were used for model development. Among these four subsets, 10% of the samples were randomly selected to form the validation set, and the remaining 90% were used for training. This process was repeated five times to ensure that each subset served once as the test set.

The model was trained with a mini-batch size of 128, which was chosen as a compromise between computational efficiency and model stability. Optimization was performed using the AdamW optimizer [89], with an initial learning rate of 0.001 and weight decay set to 10^{-5} .

The maximum number of training epochs was set to 100. At the end of each epoch, both the training and validation sets were shuffled prior to the

Table 4.2: Average training time of DiCleavePlus (minutes)

	Pattern Size	Dataset-1	Dataset-2
DCP-AFF	10	100.4	22.8
	12	174.6	37.8
	14	207.0	61.4
	16	231.2	79.0
	18	250.0	87.6
DCP-Concat	10	90.8	20.4
	12	173.6	29.6
	14	206.0	60.2
	16	231.6	78.4
	18	250.0	89.0

next epoch to avoid learning biases from fixed sample orders. A learning rate scheduler was applied to reduce the learning rate by the factor of 0.5 when the validation loss value failed to decrease. In addition, an early-stopping mechanism based on validation loss value was employed to terminate training when no further improvement was observed.

For each fold, the final checkpoint of the trained model and up to three checkpoints corresponding to the highest validation performance were saved for subsequent analysis. The checkpoint with the best performance on the test set was selected as the representative model for that fold.

All experiments were conducted on an NVIDIA H100 Tensor Core GPU with 64 GB of memory. The computational time required for training is summarized in Table 4.2.

4.2.9 Performance metrics

Metrics to assess multi-class classification performance

DiCleavePlus is designed as a multi-class classification model. Accordingly, its performance was evaluated using top-1 accuracy, macro F1 score, weighted F1 score, perfect match fraction (PMF), positional shift error (PSE), and top-3 accuracy.

Top-1 accuracy was defined as the proportion of patterns for which the predicted class exactly matched the labeled class:

$$top1\ accuracy = \frac{\sum \mathbf{1}(pred = lbl)}{N} \quad (4.2)$$

where N indicates the total number of patterns; $pred$ denotes predicted cleavage site position; lbl denotes actual cleavage site position; $\mathbf{1}(pred = lbl)$ is an indicator variable, which is set to 1 when the prediction matches the label.

To account for performance across all classes, we calculated the macro F1 score:

$$macro\ F1 = \frac{1}{N_C} \sum F1\ score_c \quad (4.3)$$

$$F1\ score_c = \frac{2 \times TP_c}{2 \times TP_c + (FP_c + FN_c)} \quad (4.4)$$

where N_C indicates total number of classes; $F1\ score_c$ indicates the F1 score per-class; TP_c , FP_c and FN_c denote the numbers of true positives, false positives and false negatives for each class, respectively.

The weighted F1 score was also computed to account for class imbalance:

$$weight\ F1 = \sum w_c \times F1\ score_c \quad (4.5)$$

$$w_c = \frac{\sum \mathbf{1}(c_i \in c)}{N} \quad (4.6)$$

where w_c is defined as the proportion of the number of patterns in class c to the total number of patterns; $\mathbf{1}(c_i \in c)$ is an indicator variable, which is set to 1 when the class of pattern i belongs to class c .

In this study, each class represents a specific interval within a Cleavage Pattern corresponding to a potential Dicer cleavage site. To more precisely assess the positional accuracy of cleavage site prediction, we employed the perfect match fraction (PMF) and positional shift error (PSE) [15]. Since class 0 represents negative patterns with no cleavage site, PSE is undefined for sample where the true or predicted class is 0. Therefore, PMF and PSE were calculated only for true positives:

$$PMF = \frac{\sum \mathbf{1}(pred = lbl)}{N_{TP}} \quad pred \neq 0, lbl \neq 0 \quad (4.7)$$

$$PSE = \frac{\sum |pred - lbl|}{N_{TP}} \quad pred \neq 0, lbl \neq 0 \quad (4.8)$$

where N_{TP} indicates the total number of true positive patterns; $pred$ denotes predicted cleavage site position; lbl denotes actual cleavage site position; $\mathbf{1}(pred = lbl)$ is an indicator variable, which is set to 1 when the prediction matches the label. PMF assesses the proportion of predictions in which the predicted class exactly matches the labeled class, while PSE quantifies the positional shift between predicted and labeled class.

Finally, to assess the practical utility of the model, we reported the top-3 accuracy, which measures the proportion of samples for which the correct class is ranked among the three most probable model predictions. This metric is especially relevant in experimental scenarios, where a small set of high-confidence candidate sites is often sufficient for further validation.

Metrics to assess binary classification performance

In addition to multi-class evaluation, we further assessed the ability of Di-CleavePlus to distinguish positive patterns from negative ones. For this purpose, conventional binary classification metrics were employed, including accuracy (Acc), specificity (Spe), sensitivity (Sen), and the Matthews correlation coefficient (MCC).

Accuracy was defined as the proportion of correctly classified samples:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.9)$$

Specificity measures the proportion of true negative samples that were correctly classified:

$$Spe = \frac{TN}{TN + FP} \quad (4.10)$$

Sensitivity quantifies the proportion of true positive samples that were correctly identified:

$$Sen = \frac{TP}{TP + FN} \quad (4.11)$$

The Matthews correlation coefficient was used as a balanced metric to evaluate the overall quality of binary classification, particularly under class-imbalanced conditions:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (4.12)$$

where TP , TN , FP , and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

4.3 Results

4.3.1 Performance of cleavage site position prediction

We evaluated the performance of DiCleavePlus using 5-fold cross-validation on both datasets and across different pattern sizes. As shown in Table 4.3 A, performance of DCP-AFF and DCP-Concat improved consistently as the pattern size increased. When using the empirical 14-nt pattern size on Dataset-1, both models exhibited similar performance. Although the top-1 accuracy was approximately 0.86, the top-3 accuracy was remarkably high (DCP-AFF: 0.979; DCP-Concat: 0.980), suggesting high practical reliability. Both models also achieved high perfect match fraction (PMF) scores and low positional shift error (PSE) scores, implying that most errors arose from misclassification between positive and negative patterns rather than incorrect localization of cleavage sites.

When the pattern size increased to 18 nt, both models showed further performance improvement, particularly in top-1 accuracy and F1 scores (macro and weighted). The confusion matrix heatmaps of the best-performing models on Dataset-1 are displayed in Figure 4.8. As shown, cleavage sites positioned near the center of the patterns were predicted accurately, whereas most misclassifications were associated with negative patterns or with positive patterns where cleavage sites were near the pattern boundaries.

The results for Dataset-2 are summarized in Table 4.3 B. Similar to Dataset-1, longer pattern size consistently led to improved performance. Notably, when patterns sizes were ≥ 14 nt, DiCleavePlus achieved high accuracy on both Dataset-1 and Dataset-2 despite their distinct class distributions, demonstrating the robustness of the proposed architecture. Furthermore, no substantial performance difference was observed between DCP-AFF and DCP-Concat, especially when larger pattern size were used.

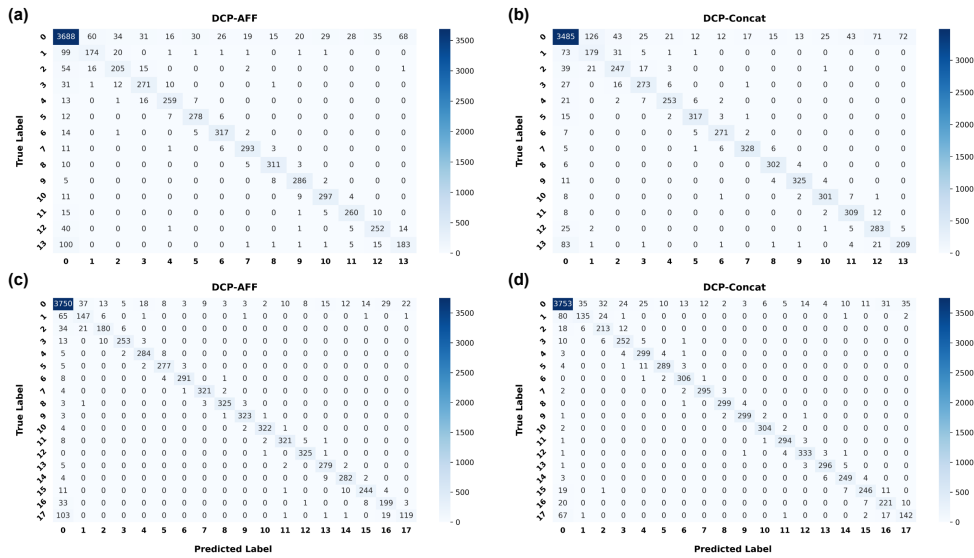


Figure 4.8: **Confusion matrix heatmaps of DiCleavePlus models trained and evaluated on Dataset-1 with different pattern sizes.** (a) DCP-AFF with 14-nt patterns; (b) DCP-Concat with 14-nt patterns; (c) DCP-AFF with 18-nt patterns; (d) DCP-Concat with 18-nt patterns. The x-axis denotes the predicted class; the y-axis denotes the true class. Class 0 corresponds to negative patterns (i.e., patterns that do not contain a cleavage site).

Table 4.3: DiCleavePlus cleavage site prediction performance using different pattern sizes

A. Prediction performance on Dataset-1

	Pattern Size	Top-1 Accuracy	Macro F1 Score	Weighted F1 Score	Top-3 Accuracy	PMF	PSE↓
DCP-AFF	10	0.700 (0.037)	0.378 (0.062)	0.685 (0.049)	0.904 (0.036)	0.700 (0.083)	0.406 (0.091)
DCP-Concat	10	0.721 (0.002)	0.414 (0.006)	0.715 (0.005)	0.926 (0.006)	0.735 (0.012)	0.358 (0.018)
DCP-AFF	12	0.807 (0.017)	0.641 (0.017)	0.803 (0.016)	0.960 (0.008)	0.888 (0.016)	0.168 (0.030)
DCP-Concat	12	0.822 (0.010)	0.663 (0.012)	0.820 (0.010)	0.970 (0.004)	0.905 (0.010)	0.144 (0.013)
DCP-AFF	14	0.861 (0.009)	0.828 (0.011)	0.857 (0.010)	0.979 (0.003)	0.931 (0.004)	0.100 (0.010)
DCP-Concat	14	0.864 (0.012)	0.835 (0.018)	0.863 (0.013)	0.980 (0.006)	0.928 (0.014)	0.104 (0.017)
DCP-AFF	16	0.872 (0.021)	0.864 (0.037)	0.884 (0.020)	0.979 (0.009)	0.924 (0.037)	0.106 (0.039)
DCP-Concat	16	0.854 (0.020)	0.837 (0.030)	0.868 (0.018)	0.973 (0.002)	0.895 (0.034)	0.145 (0.035)
DCP-AFF	18	0.916 (0.008)	0.914 (0.015)	0.926 (0.009)	0.988 (0.002)	0.958 (0.010)	0.064 (0.013)
DCP-Concat	18	0.917 (0.005)	0.916 (0.006)	0.928 (0.004)	0.986 (0.003)	0.955 (0.009)	0.068 (0.015)

B. Prediction performance on Dataset-2

	Pattern Size	Top-1 Accuracy	Macro F1 Score	Weighted F1 Score	Top-3 Accuracy	PMF	PSE↓
DCP-AFF	10	0.553 (0.084)	0.396 (0.061)	0.555 (0.087)	0.832 (0.031)	0.639 (0.099)	0.591 (0.155)
DCP-Concat	10	0.573 (0.020)	0.410 (0.013)	0.574 (0.018)	0.832 (0.019)	0.654 (0.024)	0.603 (0.060)
DCP-AFF	12	0.756 (0.004)	0.646 (0.005)	0.755 (0.005)	0.918 (0.006)	0.836 (0.007)	0.286 (0.034)
DCP-Concat	12	0.736 (0.006)	0.630 (0.007)	0.735 (0.007)	0.911 (0.009)	0.816 (0.008)	0.320 (0.026)
DCP-AFF	14	0.847 (0.008)	0.845 (0.007)	0.846 (0.006)	0.956 (0.003)	0.912 (0.007)	0.167 (0.013)
DCP-Concat	14	0.847 (0.029)	0.846 (0.027)	0.847 (0.028)	0.955 (0.009)	0.915 (0.021)	0.170 (0.036)
DCP-AFF	16	0.892 (0.005)	0.899 (0.005)	0.902 (0.004)	0.969 (0.003)	0.943 (0.005)	0.133 (0.009)
DCP-Concat	16	0.892 (0.006)	0.898 (0.006)	0.900 (0.006)	0.966 (0.004)	0.940 (0.006)	0.144 (0.021)
DCP-AFF	18	0.915 (0.016)	0.919 (0.013)	0.922 (0.014)	0.976 (0.005)	0.956 (0.011)	0.088 (0.020)
DCP-Concat	18	0.911 (0.008)	0.915 (0.006)	0.919 (0.007)	0.974 (0.003)	0.955 (0.007)	0.103 (0.015)

Average results of 5-fold cross-validation. Numbers in parentheses indicate standard deviation.
↓ indicates the lower value the better performance.
Best result is marked in bold.

4.3.2 Performance of positive-negative binary classification

Although DiCleavePlus is designed as a multi-class classification model, we further evaluated its capability to distinguish positive patterns from negative ones. In this binary evaluation, the original multi-class output layer was retained. Predictions assigned to class 0 were regarded as negative, whereas predictions of any non-zero class were considered positive. The binary classification performance of DiCleavePlus is summarized in Table 4.4. On Dataset-1, using the empirical 14-nt pattern size, DCP-AFF and DCP-Concat exhibited comparable performance, with DCP-Concat showing slightly higher sensitivity. When the pattern size increased to 18 nt, both architectures achieved stronger performance in binary classification task (Table 4.4 A).

In contrast, Dataset-2 presents an inherently imbalanced binary classification scenario, as negative patterns constitute only approximately 7% of all samples ($1624/22282 \approx 7.29\%$ for 14-nt patterns and $1624/26224 \approx 6.19\%$ for 18-nt patterns). Under this imbalance, for 14-nt patterns, DCP-AFF achieved a sensitivity of 0.962 and DCP-Concat achieved 0.955. However, their specificity scores were substantially lower (0.462 and 0.507, respectively), which consequently led to low Matthews correlation coefficient (MCC) scores of 0.436 and 0.450. Models trained on 18-nt patterns showed overall improvement, yet their specificity remained limited, resulting in suboptimal MCC scores (Table 4.4 B).

It is important to emphasize that Dataset-2 is highly imbalanced across classes and the actual outputs of DiCleavePlus remain intrinsically multi-class. Therefore, binary evaluation metrics alone do not fully capture the model’s predictive behavior. To obtain a more comprehensive understanding, confusion matrix heatmaps of the best-performing models are presented in Figure 4.9. All models exhibited similar classification patterns across different pattern sizes. Negative patterns were frequently misclassified as positive patterns with cleavage sites located near boundary intervals. Specifically, mispredictions were concentrated at intervals 1, 2, 13, and 14 in 14-nt patterns (Figure 4.9 (a) and (b)), and at intervals 1, 2, 17 and 18 in 18-nt patterns (Figure 4.9 (c) and (d)). Conversely, positive patterns with cleavage sites at these boundary intervals were often misclassified as negative. Within the positive classes, most errors were also concentrated at these boundary regions rather than in the central intervals. Cleavage sites located near the center of the pattern were

Table 4.4: DiCleavePlus binary performance using different pattern sizes

A. Prediction performance on Dataset-1

	Pattern Size	Accuracy	Specificity	Sensitivity	MCC
DCP-AFF	10	0.781 (0.030)	0.866 (0.022)	0.665 (0.100)	0.548 (0.063)
DCP-Concat	10	0.804 (0.004)	0.850 (0.017)	0.742 (0.029)	0.597 (0.008)
DCP-AFF	12	0.850 (0.013)	0.866 (0.029)	0.832 (0.029)	0.700 (0.025)
DCP-Concat	12	0.860 (0.007)	0.864 (0.021)	0.857 (0.015)	0.721 (0.012)
DCP-AFF	14	0.892 (0.009)	0.893 (0.009)	0.891 (0.011)	0.784 (0.017)
DCP-Concat	14	0.897 (0.007)	0.882 (0.011)	0.912 (0.008)	0.794 (0.014)
DCP-AFF	16	0.909 (0.006)	0.901 (0.016)	0.915 (0.019)	0.816 (0.012)
DCP-Concat	16	0.905 (0.005)	0.909 (0.013)	0.901 (0.018)	0.809 (0.010)
DCP-AFF	18	0.938 (0.005)	0.934 (0.009)	0.940 (0.007)	0.874 (0.009)
DCP-Concat	18	0.940 (0.003)	0.933 (0.013)	0.946 (0.009)	0.879 (0.006)

B. Prediction performance on Dataset-2

	Pattern Size	Accuracy	Specificity	Sensitivity	MCC
DCP-AFF	10	0.847 (0.022)	0.341 (0.095)	0.904 (0.029)	0.227 (0.065)
DCP-Concat	10	0.856 (0.006)	0.369 (0.039)	0.911 (0.013)	0.262 (0.029)
DCP-AFF	12	0.897 (0.006)	0.411 (0.065)	0.941 (0.011)	0.345 (0.039)
DCP-Concat	12	0.894 (0.007)	0.407 (0.037)	0.939 (0.010)	0.335 (0.021)
DCP-AFF	14	0.926 (0.005)	0.462 (0.051)	0.962 (0.007)	0.436 (0.021)
DCP-Concat	14	0.923 (0.013)	0.507 (0.039)	0.955 (0.014)	0.450 (0.044)
DCP-AFF	16	0.945 (0.003)	0.482 (0.025)	0.978 (0.003)	0.511 (0.024)
DCP-Concat	16	0.946 (0.004)	0.486 (0.054)	0.979 (0.004)	0.521 (0.040)
DCP-AFF	18	0.955 (0.007)	0.511 (0.043)	0.985 (0.004)	0.572 (0.051)
DCP-Concat	18	0.952 (0.003)	0.522 (0.040)	0.980 (0.005)	0.553 (0.018)

Average results of 5-fold cross-validation. Numbers in parentheses indicate standard deviation.
Best result is marked in bold.

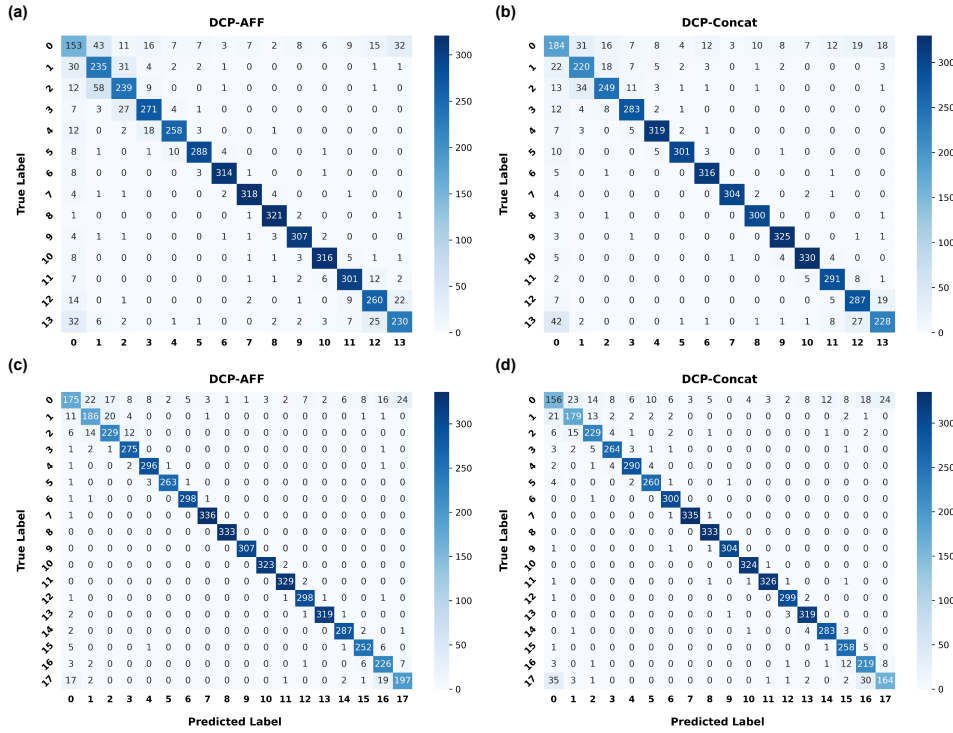


Figure 4.9: **Confusion matrix heatmaps of DiCleavePlus models trained and evaluated on Dataset-2 with different pattern sizes.** (a) DCP-AFF with 14-nt patterns; (b) DCP-Concat with 14-nt patterns; (c) DCP-AFF with 18-nt patterns; (d) DCP-Concat with 18-nt patterns. The x-axis denotes the predicted class; the y-axis denotes the true class. Class 0 corresponds to negative patterns (i.e., patterns that do not contain a cleavage site).

predicted with high accuracy, which is consistent with the high PMF and low PSE scores reported in Table 4.3.

Notably, despite operating in a larger classification space, models trained on 18-nt patterns outperformed their 14-nt counterparts in distinguishing between negative patterns and boundary positive patterns. This indicates that longer patterns provide richer contextual cues that enhance discriminative performance, making the 18-nt configuration the best-performing models in this study.

4.3.3 Performance comparison with ablation models

The multi-class classification performances of DiCleavePlus were compared with three ablation models, namely a multi-layer perceptron (AM-MLP), a

convolutional neural network (AM-CNN), and a recurrent neural network (AM-RNN). As shown in Table 4.5, DiCleavePlus consistently outperformed all ablation models on both Dataset-1 and Dataset-2, under both the empirical pattern size (14 nt) and the extended pattern size (18 nt). Among the ablation models, AM-RNN achieved the best performance, suggesting that the ability to capture positional dependencies within nucleotide sequences contributes positively to cleavage site prediction.

Table 4.6 presents the binary classification results of DiCleavePlus and the ablation models. The ablation models demonstrated high specificity but low sensitivity on Dataset-1, indicating a tendency to classify most inputs as negative. In contrast, on Dataset-2, where negative patterns represent only a small minority, these models achieved high sensitivity but markedly low specificity, revealing their difficulty in distinguishing minority classes under imbalanced distributions. DiCleavePlus, by comparison, maintained a more stable balance between sensitivity and specificity. Notably, it successfully identified approximately 50% of the minority class samples in Dataset-2, despite the strong class imbalance in the dataset (Figure 4.4).

These results demonstrate the importance of Transformer-based feature extractors used in DiCleavePlus. The superior performance, particularly in detecting minority classes, indicates that attention-based mechanisms enable more effective modeling of both local and long-range sequence-structure dependencies compared with simpler MLP-, CNN-, or RNN-based architectures.

4.3.4 Performance comparison with other models

We compared DiCleavePlus with two other models, ReCGBM [14] and DiCleave. The results of multi-class classification and binary classification are summarized in Table 4.7 and Table 4.8, respectively. Across both tasks, DiCleavePlus consistently outperformed the competing models. As a deep learning model, DiCleave achieved better performance than ReCGBM on Dataset-1 and demonstrated comparable performance on Dataset-2. However, when compared with the three ablation models (Table 4.5 and Table 4.6), DiCleave outperformed only the MLP-based ablation model, whereas it fell behind the CNN- and RNN-based ablation models.

The limited performance of DiCleave can be attributed to several factors. First, DiCleave employs one-hot encoding to represent input sequence. This

Table 4.5: Performance comparison between DiCleavePlus and its ablation models

A. Prediction performance on Dataset-1

	Pattern Size	Top-1 Accuracy	Macro F1 Score	Weighted F1 Score	Top-3 Accuracy	PMF	PSE↓
DCP-AFF	14	0.861 (0.009)	0.828 (0.011)	0.857 (0.010)	0.979 (0.003)	0.931 (0.004)	0.100 (0.010)
DCP-Concat	14	0.864 (0.012)	0.835 (0.018)	0.863 (0.013)	0.980 (0.006)	0.928 (0.014)	0.104 (0.017)
AM-MLP	14	0.512 (0.020)	0.176 (0.015)	0.439 (0.016)	0.714 (0.011)	0.242 (0.024)	1.384 (0.075)
AM-CNN	14	0.631 (0.060)	0.420 (0.148)	0.589 (0.088)	0.843 (0.059)	0.558 (0.175)	0.646 (0.342)
AM-RNN	14	0.687 (0.020)	0.557 (0.031)	0.674 (0.019)	0.908 (0.009)	0.693 (0.045)	0.395 (0.055)
DCP-AFF	18	0.916 (0.008)	0.914 (0.015)	0.926 (0.009)	0.988 (0.002)	0.958 (0.010)	0.064 (0.013)
DCP-Concat	18	0.917 (0.005)	0.916 (0.006)	0.928 (0.004)	0.986 (0.003)	0.955 (0.009)	0.068 (0.015)
AM-MLP	18	0.488 (0.010)	0.168 (0.008)	0.451 (0.004)	0.689 (0.012)	0.216 (0.012)	1.544 (0.102)
AM-CNN	18	0.599 (0.079)	0.389 (0.161)	0.587 (0.096)	0.820 (0.079)	0.477 (0.196)	0.852 (0.470)
AM-RNN	18	0.682 (0.045)	0.596 (0.088)	0.706 (0.054)	0.901 (0.022)	0.674 (0.083)	0.448 (0.126)

B. Prediction performance on Dataset-2

	Pattern Size	Top-1 Accuracy	Macro F1 Score	Weighted F1 Score	Top-3 Accuracy	PMF	PSE↓
DCP-AFF	14	0.847 (0.008)	0.845 (0.007)	0.846 (0.006)	0.956 (0.003)	0.912 (0.007)	0.167 (0.013)
DCP-Concat	14	0.847 (0.029)	0.846 (0.027)	0.847 (0.028)	0.955 (0.009)	0.915 (0.021)	0.170 (0.036)
AM-MLP	14	0.208 (0.005)	0.203 (0.007)	0.202 (0.007)	0.540 (0.008)	0.221 (0.006)	1.798 (0.014)
AM-CNN	14	0.456 (0.095)	0.444 (0.094)	0.445 (0.095)	0.798 (0.046)	0.486 (0.103)	0.927 (0.238)
AM-RNN	14	0.593 (0.028)	0.592 (0.028)	0.593 (0.027)	0.880 (0.014)	0.668 (0.030)	0.518 (0.059)
DCP-AFF	18	0.915 (0.016)	0.919 (0.013)	0.922 (0.014)	0.976 (0.005)	0.956 (0.011)	0.088 (0.020)
DCP-Concat	18	0.911 (0.008)	0.915 (0.006)	0.919 (0.007)	0.974 (0.003)	0.955 (0.007)	0.103 (0.015)
AM-MLP	18	0.200 (0.008)	0.185 (0.020)	0.187 (0.019)	0.529 (0.006)	0.203 (0.012)	1.998 (0.036)
AM-CNN	18	0.495 (0.088)	0.502 (0.095)	0.511 (0.097)	0.817 (0.040)	0.522 (0.093)	0.865 (0.162)
AM-RNN	18	0.617 (0.124)	0.645 (0.167)	0.654 (0.165)	0.884 (0.047)	0.671 (0.136)	0.546 (0.254)

Average results of 5-fold cross-validation. Numbers in parentheses indicate standard deviation.
 ↓ indicates the lower value the better performance.
 Best result is marked in bold.

Table 4.6: Binary performance comparison between DiCleavePlus and its ablation models

A. Prediction performance on Dataset-1

	Pattern Size	Accuracy	Specificity	Sensitivity	MCC
DCP-AFF	14	0.892 (0.009)	0.893 (0.009)	0.891 (0.011)	0.784 (0.017)
DCP-Concat	14	0.897 (0.007)	0.882 (0.011)	0.912 (0.008)	0.794 (0.014)
AM-MLP	14	0.711 (0.011)	0.912 (0.031)	0.517 (0.047)	0.466 (0.016)
AM-CNN	14	0.759 (0.043)	0.915 (0.027)	0.607 (0.115)	0.550 (0.064)
AM-RNN	14	0.805 (0.008)	0.859 (0.017)	0.751 (0.018)	0.614 (0.016)
DCP-AFF	18	0.938 (0.005)	0.934 (0.009)	0.940 (0.007)	0.874 (0.009)
DCP-Concat	18	0.940 (0.003)	0.933 (0.013)	0.946 (0.009)	0.879 (0.006)
AM-MLP	18	0.741 (0.014)	0.935 (0.024)	0.583 (0.044)	0.541 (0.014)
AM-CNN	18	0.784 (0.031)	0.933 (0.019)	0.662 (0.070)	0.608 (0.043)
AM-RNN	18	0.821 (0.025)	0.879 (0.008)	0.773 (0.049)	0.650 (0.042)

B. Prediction performance on Dataset-2

	Pattern Size	Accuracy	Specificity	Sensitivity	MCC
DCP-AFF	14	0.926 (0.005)	0.462 (0.051)	0.962 (0.007)	0.436 (0.021)
DCP-Concat	14	0.923 (0.013)	0.507 (0.039)	0.955 (0.014)	0.450 (0.044)
AM-MLP	14	0.909 (0.016)	0.129 (0.060)	0.970 (0.019)	0.135 (0.037)
AM-CNN	14	0.924 (0.007)	0.173 (0.073)	0.983 (0.016)	0.255 (0.034)
AM-RNN	14	0.873 (0.006)	0.394 (0.037)	0.911 (0.005)	0.252 (0.034)
DCP-AFF	18	0.955 (0.007)	0.511 (0.043)	0.985 (0.004)	0.572 (0.051)
DCP-Concat	18	0.952 (0.003)	0.522 (0.040)	0.980 (0.005)	0.553 (0.018)
AM-MLP	18	0.915 (0.015)	0.277 (0.119)	0.957 (0.022)	0.242 (0.048)
AM-CNN	18	0.932 (0.007)	0.268 (0.077)	0.976 (0.011)	0.304 (0.046)
AM-RNN	18	0.909 (0.014)	0.369 (0.084)	0.944 (0.018)	0.288 (0.047)

Average results of 5-fold cross-validation. Numbers in parentheses indicate standard deviation.
Best result is marked in bold.

Table 4.7: Performance comparison between DiCleavePlus and other models

A. Prediction performance on Dataset-1

	Pattern Size	Top-1 Accuracy	Macro F1 Score	Weighted F1 Score	Top-3 Accuracy	PMF	PSE↓
DCP-AFF	14	0.861 (0.009)	0.828 (0.011)	0.857 (0.010)	0.979 (0.003)	0.931 (0.004)	0.100 (0.010)
DCP-Concat	14	0.864 (0.012)	0.835 (0.018)	0.863 (0.013)	0.980 (0.006)	0.928 (0.014)	0.104 (0.017)
DiCleave	14	0.515 (0.017)	0.192 (0.030)	0.452 (0.016)	0.723 (0.030)	0.243 (0.046)	1.507 (0.310)
ReCGBM	14	0.499 (0.007)	0.129 (0.006)	0.380 (0.008)	NA	0.294 (0.021)	1.948 (0.083)

B. Prediction performance on Dataset-2

	Pattern Size	Top-1 Accuracy	Macro F1 Score	Weighted F1 Score	Top-3 Accuracy	PMF	PSE↓
DCP-AFF	14	0.847 (0.008)	0.845 (0.007)	0.846 (0.006)	0.956 (0.003)	0.912 (0.007)	0.167 (0.013)
DCP-Concat	14	0.847 (0.029)	0.846 (0.027)	0.847 (0.028)	0.955 (0.009)	0.915 (0.021)	0.170 (0.036)
DiCleave	14	0.222 (0.028)	0.210 (0.039)	0.209 (0.039)	0.576 (0.054)	0.242 (0.027)	1.750 (0.333)
ReCGBM	14	0.224 (0.002)	0.224 (0.002)	0.224 (0.002)	NA	0.244 (0.005)	2.766 (0.051)

Average results of 5-fold cross-validation. Numbers in parentheses indicate standard deviation.

↓ indicates the lower value the better performance.

Best result is marked in bold.

Table 4.8: Binary performance comparison between DiCleavePlus and other models

A. Prediction performance on Dataset-1

	Pattern Size	Accuracy	Specificity	Sensitivity	MCC
DCP-AFF	14	0.892 (0.009)	0.893 (0.009)	0.891 (0.011)	0.784 (0.017)
DCP-Concat	14	0.897 (0.007)	0.882 (0.011)	0.912 (0.008)	0.794 (0.014)
DiCleave	14	0.724 (0.018)	0.913 (0.023)	0.541 (0.054)	0.488 (0.021)
ReCGBM	14	0.565 (0.008)	0.959 (0.003)	0.183 (0.006)	0.224 (0.012)

B. Prediction performance on Dataset-2

	Pattern Size	Accuracy	Specificity	Sensitivity	MCC
DCP-AFF	14	0.926 (0.005)	0.462 (0.051)	0.962 (0.007)	0.436 (0.021)
DCP-Concat	14	0.923 (0.013)	0.507 (0.039)	0.955 (0.014)	0.450 (0.044)
DiCleave	14	0.878 (0.017)	0.165 (0.039)	0.934 (0.017)	0.101 (0.048)
ReCGBM	14	0.873 (0.006)	0.189 (0.032)	0.926 (0.007)	0.109 (0.026)

Average results of 5-fold cross-validation. Numbers in parentheses indicate standard deviation.
Best result is marked in bold.

encoding scheme is insufficient to capture nuanced differences between highly similar patterns, especially in this study where most Cleavage Patterns, particularly the positive ones, substantially overlap. Consequently, the model may fail to discriminate between patterns that differ only in subtle yet functionally important positions. Second, DiCleave processes Cleavage Patterns and full-length pre-miRNA sequences separately. The features of pre-miRNA sequences are derived from a pre-trained autoencoder, and the model does not jointly process the Cleavage Pattern alongside its original pre-miRNA context. The isolated processing hinders the model from leveraging the contextual relationship between a pattern and its structural environment within the pre-miRNA. Third, the CNN-based feature extractor in DiCleave does not incorporate positional encoding in its input representation. Without explicit positional information, the model may struggle to learn the spatial arrangement of nucleotides within Cleavage Patterns, thereby limiting its representational capacity, especially for tasks that require precise localization of cleavage sites.

In contrast, DiCleavePlus addresses these limitations by integrating embedding-based representations, joint processing of Cleavage Patterns and full-length pre-miRNA sequences, and the use of Transformer encoders with positional encoding. As a result, DiCleavePlus achieves superior performance in both multi-class and binary prediction tasks and better captures positional and contextual dependencies critical for Dicer cleavage site prediction.

4.4 Discussion

Precise Dicer cleavage is essential for the biogenesis of mature miRNAs. Extensive *in vivo* and *in vitro* studies have been conducted to elucidate the underlying principles of Dicer processing. The studies have demonstrated that Dicer does not cleave randomly but instead follows defined rules, such as the 3' counting rules [90, 91] and the 5' counting rules [92]. Subsequent findings further revealed that both the secondary structure of pre-miRNAs and the presence of specific sequence motifs play crucial roles in determining cleavage sites [16, 51, 49]. In addition to its role in miRNA maturation, Dicer also participates in the production of small interfering RNAs (siRNAs) by cleaving double-stranded RNAs [93, 94, 95, 96]. Given that siRNAs are considered

a promising therapeutic modality [97, 98, 99, 100], a comprehensive understanding of Dicer cleavage specificity is of both biological and clinical importance. Although previous studies have provided substantial insights into the mechanism governing Dicer cleavage, experimental approaches remain labor-intensive, time-consuming, and limited in scale. Consequently, the development of accurate computational models for Dicer cleavage site prediction is indispensable and will facilitate research on miRNA biogenesis, discovery of potential miRNA products, and the rational design of siRNA precursors.

Several computational approaches have been proposed to predict human Dicer cleavage sites. PHDcleav, LBSIZEcleav, and ReCGBM are Cleavage Pattern-based machine learning models. Due to their limited use of contextual information, Cleavage Patterns derived from the 5' and 3' arms of the same pre-miRNA are considered independent entities in these models. In contrast, DiCleave integrates full-length pre-miRNA sequences and therefore requires no additional prior knowledge of the input Cleavage Patterns, making it a more practical tool. However, DiCleave only recognizes cleavage sites located at the central position of the patterns, essentially reducing the task to binary classification and failing to fully utilize positional diversity within Cleavage Patterns.

In this study, we proposed DiCleavePlus to address these limitations. By treating each interval within a Cleavage Pattern as a unique class, we reformulated the task as a multi-class classification problem. In this new setting, Dicer cleavage sites can occur at any interval within the patterns, rather than being restricted to the central position. This design not only allows for a more comprehensive representation of Dicer cleavage sites but also effectively increases the number of training samples. For example, when using a 14-nt pattern, a single cleavage site yields 13 distinct positive patterns, whereas conventional Cleavage Pattern generates only one. In this sense, DiCleavePlus introduces an implicit form of data augmentation, mitigating the limitation of insufficient training data for deep learning architecture. We represented pre-miRNA sequences and dot-bracket secondary structures using 3-mer tokens with 32-dimensional embedding and employed Transformer encoders to extract features from both Cleavage Patterns and full-length pre-miRNA sequences. To integrate the information from these two sources, an attentional feature fusion module was implemented.

Our experimental results demonstrate that DiCleavePlus achieved promising predictive performance. Its accuracy improved with longer pattern size, which is consistent with observations reported in PHDcleav and LBSIZEcleav. While 14-nt patterns yielded satisfactory results, the best performance was achieved using 18-nt patterns, indicating that extended patterns provide additional structural and sequence-based contextual information that facilitates precise cleavage site localization. DiCleavePlus achieved approximately 91% top-1 accuracy and 98% top-3 accuracy with 18-nt patterns, suggesting that correct predictions were almost always ranked among the top candidates. This high level of accuracy underscores the robustness and practical applicability of the model.

Notably, we employed three Transformer encoder layers to process both Cleavage Pattern features and full-length pre-miRNA sequence features, each of which include dot-bracket secondary structure information. A study by [101] suggests that soft-attention networks with two layers are theoretically sufficient to represent hierarchical languages of the form $Dyck_{K,D}$. Although pseudoknot-free dot-bracket secondary structures resemble the $Dyck_1$ language in terms of single-type bracket matching, the presence of unpaired bases carries additional biologically meaningful information that extends beyond simple bracket pairing. Moreover, the secondary structure information embedded in Cleavage Pattern features is asymmetric and heterogeneous. As a result, the dot-bracket sequences in this study cannot be simplified as Dyck formal language. Nevertheless, the theoretical analysis presented in [101] provides a lower bound on the representational capacity required for Transformer architectures, offering valuable insight into potential directions for further optimizing model depth and architecture design.

The main limitation of DiCleavePlus lies in its difficulty in distinguishing negative patterns from positive patterns in which cleavage sites are located near the boundary intervals. In contrast, cleavage sites located near the center of the patterns were predicted with high accuracy, suggesting that the flanking regions surrounding the cleavage site play an essential role in determining cleavage site. Hence, boundary-localized cleavage sites lack complete upstream or downstream flanking region information, which likely contributes to their frequent misclassification (Figure 4.8 and Figure 4.9). This phenomenon suggests that the current pattern-based representation still has inherent constraints in

capturing the full structural landscape required for cleavage site recognition.

Another source of potential limitation arises from the RNA secondary structures used in this study. All pre-miRNA secondary structures were predicted by RNAfold [75], which does not allow pseudoknot structures. Since substrate structures have been reported to influence Dicer processing [91, 102], pseudoknot structures may likewise play an important role in determining Dicer cleavage site. In case where the 5' and 3' cleavage sites are spatially adjacent, pseudoknot-like interactions could impose local structural constraints that guide Dicer activity. The absence of such structural information in the current dataset may therefore limit the representational capacity of DiCleavePlus to fully capture these context-dependent Cleavage Patterns.

We also observed that no substantial difference in performance between models adopting the two feature fusion strategies, DCP-AFF and DCP-Concat. Further analysis of attention weights revealed that the AFF module exhibited a strong bias toward pattern features, with attention contributions ranging from 0.95 to 0.99, suggesting that the current architecture may not fully capture or utilize information from full-length pre-miRNA sequences. Future improvements could involve more effective sequence embedding strategies, such as multi-view multi-scale encoding [103, 85], to enhance the contribution of sequence context. Additionally, the AFF module adopted in this work corresponds to the basic implementation described by [88]. Their iterative attentional feature fusion (iAFF) architecture, which repeatedly computes feature contributions by stacking multiple multi-scale channel attention blocks, may enhance the model's capacity to integrate heterogeneous features. Expanding DiCleavePlus to incorporate such advanced feature fusion strategies could further improve prediction accuracy and robustness.

Chapter 5

Conclusion and Future Directions

5.1 Conclusion

Accurate identification of Dicer cleavage sites is essential for understanding miRNA biogenesis, as cleavage precision directly determines the sequence of mature miRNAs and, consequently, their regulatory functions. In this dissertation, we investigated the problem of human Dicer cleavage site prediction within pre-miRNA sequences, with a particular emphasis on effective feature extraction from sequence-based inputs while minimizing the reliance on prior biological assumptions. To this end, we proposed two deep learning-based frameworks, DiCleave and DiCleavePlus, that address key limitations of existing computational approaches. Our results demonstrate that, when coupled with appropriate data preprocessing and representation learning strategies, deep learning models constitute effective architectures for sequence-based bioinformatics problems.

In Chapter 3, we proposed DiCleave, a deep-learning based framework designed for Dicer cleavage site prediction. Unlike previous machine learning

approaches that depend on handcrafted features and additional prior knowledge about the input data, DiCleave directly takes Cleavage Patterns and their corresponding originating pre-miRNAs as input. Notably, DiCleave is the first method to explicitly integrate full-length pre-miRNA secondary structure information into the prediction framework. Experimental results show that DiCleave achieves performance comparable to or superior to existing state-of-the-art methods, validating the effectiveness of deep learning-based modeling for this task.

In Chapter 4, we introduced DiCleavePlus, a novel framework designed to overcome the intrinsic limitations of Cleavage Pattern-based models, including DiCleave. DiCleavePlus employs extended Cleavage Patterns that allow cleavage sites to occur at arbitrary positions within a pattern, thereby increasing modeling flexibility. The framework jointly incorporates extended Cleavage Patterns, full-length pre-miRNA sequences, and secondary structure information as input, enabling more comprehensive exploitation of sequence context. In addition, sequence inputs are tokenized into 3-mer units and mapped to 32-dimensional embeddings, providing richer representations than conventional one-hot encoding. By leveraging Transformer-based encoders to jointly learn representations from both sequence and structural information, DiCleavePlus achieves accurate and robust performance in Dicer cleavage site prediction. These results indicate that integrating multi-level contextual information with flexible positional modeling substantially enhances predictive capability.

Overall, this dissertation demonstrates that carefully designed deep learning architectures, informed by biological structure and contextual constraints, can significantly improve the reliability of Dicer cleavage site prediction. The proposed models provide practical computational tools that may assist experimental design and facilitate further studies of miRNA biogenesis.

5.2 Future directions

Although DiCleavePlus advances Dicer cleavage site prediction to a new level, several aspects remain open for further refinement. First, while Transformer-based feature extractors are employed to process Cleavage Pattern features and full-length pre-miRNA sequence features, these two feature streams are handled by separate encoders. As a result, no direct attention flow exists between pattern-level features and global pre-miRNA context. We observe

that the AFF module exhibits a strong bias toward pattern features, suggesting that the current architecture does not yet fully exploit information from the full-length pre-miRNA sequence. This limitation indicates that deeper cross-attention or feature interaction mechanisms may be necessary to achieve tighter integration between local and global sequence representations.

Another aspect that warrants improvement concerns the biological features used as model input. The current framework relies on predicted, pseudoknot-free RNA secondary structures. While such representations are computationally convenient, they do not fully capture all biologically relevant structural properties. Additionally, we observe that most misclassification by DiCleave-Plus occur between negative patterns and positive patterns with cleavage sites located near the boundary intervals of the extended Cleavage Pattern. This observation suggests that the current input features are insufficient to effectively distinguish subtle differences in these cases. Incorporating more biologically meaningful structural or thermodynamic features may help alleviate this issue. Addressing these limitations of DiCleavePlus is therefore left for future investigation.

Another important future direction involves extending the framework to include data from additional species. At present, most Dicer cleavage site prediction models are trained exclusively on human data, despite the fact that miRBase contains miRNA data spanning a wide range of taxa, from Alveolata to viruses. Given that many species share conserved miRNA processing mechanisms, training a unified model on multi-species datasets may enhance generalization performance. Such a framework could serve as a universal Dicer cleavage site predictor applicable across species, particularly for organisms with limited experimental annotations.

Finally, a promising direction is the development of an end-to-end framework capable of directly predicting mature miRNA sequences or functional outcomes from pre-miRNAs. Current cleavage site prediction approaches depend heavily on fully annotated pre-miRNA hairpin sequences; however, only a fraction of entries in miRBase are comprehensively annotated. A robust end-to-end model that directly outputs mature miRNAs could function as an automatic annotation tool, supporting database refinement and facilitating novel miRNA discovery.

Through continued advances in model design, data integration, and biological validation, future work building upon this dissertation has the potential to further deepen our understanding of miRNA biogenesis and enhance the practical utility of computational frameworks in RNA biology.

List of Publications

- Chapter 3 is based on a paper in *BMC Bioinformatics*:

Mu, L., Song, J., Akutsu, T., & Mori, T. (2024). DiCleave: a deep learning model for predicting human Dicer cleavage sites. *BMC Bioinformatics*, 25(1), 13.

- Chapter 4 is based on a paper in *Genes to Cells*:

Mu, L. & Akutsu, T. (2026). DiCleavePlus: a Transformer-based model to detect human Dicer cleavage sites within Cleavage Patterns. *Genes to Cells*, 31(1), e70074.

References

- [1] C. Xue, F. Li, T. He, G.-P. Liu, Y. Li, and X. Zhang, “Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine,” *BMC Bioinformatics*, vol. 6, no. 1, p. 310, 2005. [Cited on page 2]
- [2] K. L. S. Ng and S. K. Mishra, “De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures,” *Bioinformatics*, vol. 23, no. 11, pp. 1321–1330, 2007. [Cited on page 2]
- [3] G. Terai, T. Komori, K. Asai, and T. Kin, “miRRim: a novel system to find conserved miRNAs with high sensitivity and specificity,” *RNA*, vol. 13, no. 12, pp. 2081–2090, 2007. [Cited on page 2]
- [4] D. T.-H. Chang, C.-C. Wang, and J.-W. Chen, “Using a kernel density estimation based classifier to predict species-specific microRNA precursors,” *BMC Bioinformatics*, vol. 9, no. Suppl 12, p. S2, 2008. [Cited on page 2]
- [5] S. Park, S. Min, H.-S. Choi, and S. Yoon, “Deep recurrent neural network-based identification of precursor microRNAs,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017. [Cited on page 3]
- [6] J. Raad, L. A. Bugnon, D. H. Milone, and G. Stegmayer, “miRe2e: a full end-to-end deep model based on transformers for prediction of pre-miRNAs,” *Bioinformatics*, vol. 38, no. 5, pp. 1191–1197, 2022. [Cited on page 3]
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances*

- in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017. [Cited on pages 3, 10, 11, 34, 36, and 43]
- [8] B. P. Lewis, I.-h. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge, “Prediction of mammalian microRNA targets,” *Cell*, vol. 115, no. 7, pp. 787–798, 2003. [Cited on pages 3 and 14]
- [9] J. Ding, X. Li, and H. Hu, “TarPmiR: a new approach for microRNA target site prediction,” *Bioinformatics*, vol. 32, no. 18, pp. 2768–2775, 2016. [Cited on page 3]
- [10] M. Wen, P. Cong, Z. Zhang, H. Lu, and T. Li, “DeepMirTar: a deep-learning approach for predicting human miRNA targets,” *Bioinformatics*, vol. 34, no. 22, pp. 3781–3787, 2018. [Cited on page 3]
- [11] S. Cheng, M. Guo, C. Wang, X. Liu, Y. Liu, and X. Wu, “MiRTDL: a deep learning approach for miRNA target prediction,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 6, pp. 1161–1169, 2015. [Cited on page 3]
- [12] F. Ahmed, R. Kaundal, and G. P. Raghava, “PHDcleav: a SVM based method for predicting human Dicer cleavage sites using sequence and secondary structure of miRNA precursors,” *BMC Bioinformatics*, vol. 14, no. Suppl 14, p. S9, 2013. [Cited on pages 3, 14, 26, and 35]
- [13] Y. Bao, M. Hayashida, and T. Akutsu, “LBSizeCleav: improved support vector machine (SVM)-based prediction of Dicer cleavage sites using loop/bulge length,” *BMC Bioinformatics*, vol. 17, no. 1, p. 487, 2016. [Cited on pages 3, 4, 14, 26, and 35]
- [14] P. Liu, J. Song, C.-Y. Lin, and T. Akutsu, “ReCGBM: a gradient boosting-based method for predicting human dicer cleavage sites,” *BMC Bioinformatics*, vol. 22, no. 1, p. 63, 2021. [Cited on pages 3, 4, 14, 16, 34, 35, and 58]
- [15] J. Bell and D. A. Hendrix, “Predicting Drosha and Dicer cleavage sites with DeepMirCut,” *Frontiers in Molecular Biosciences*, vol. 8, p. 799056, 2022. [Cited on pages 3, 4, 35, and 50]

- [16] S. Gu, L. Jin, Y. Zhang, Y. Huang, F. Zhang, P. N. Valdmánis, and M. A. Kay, “The loop position of shRNAs and pre-miRNAs is critical for the accuracy of Dicer processing in vivo,” *Cell*, vol. 151, no. 4, pp. 900–911, 2012. [Cited on pages 3 and 63]
- [17] B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” *Science*, vol. 315, no. 5814, pp. 972–976, 2007. [Cited on page 4]
- [18] G. Fu, J. Brkić, H. Hayder, and C. Peng, “MicroRNAs in human placental development and pregnancy complications,” *International Journal of Molecular Sciences*, vol. 14, no. 3, pp. 5519–5544, 2013. [Cited on page 8]
- [19] D. C. Swarts, K. Makarova, Y. Wang, K. Nakanishi, R. F. Ketting, E. V. Koonin, D. J. Patel, and J. Van Der Oost, “The evolutionary journey of Argonaute proteins,” *Nature Structural & Molecular Biology*, vol. 21, no. 9, pp. 743–753, 2014. [Cited on page 8]
- [20] S. Jonas and E. Izaurralde, “Towards a molecular understanding of microRNA-mediated gene silencing,” *Nature Reviews Genetics*, vol. 16, no. 7, pp. 421–433, 2015. [Cited on page 8]
- [21] A. M. Denli, B. B. Tops, R. H. Plasterk, R. F. Ketting, and G. J. Hannon, “Processing of primary microRNAs by the Microprocessor complex,” *Nature*, vol. 432, no. 7014, pp. 231–235, 2004. [Cited on page 8]
- [22] C. Okada, E. Yamashita, S. J. Lee, S. Shibata, J. Katahira, A. Nakagawa, Y. Yoneda, and T. Tsukihara, “A high-resolution structure of the pre-microRNA nuclear export machinery,” *Science*, vol. 326, no. 5957, pp. 1275–1279, 2009. [Cited on page 8]
- [23] H. Zhang, F. A. Kolb, L. Jaskiewicz, E. Westhof, and W. Filipowicz, “Single processing center models for human Dicer and bacterial RNase III,” *Cell*, vol. 118, no. 1, pp. 57–68, 2004. [Cited on pages 8 and 14]
- [24] H. A. Meijer, E. M. Smith, and M. Bushell, “Regulation of miRNA strand selection: follow the leader?,” *Biochemical Society Transactions*, vol. 42, no. 4, pp. 1135–1140, 2014. [Cited on page 9]
- [25] M. Ha and V. N. Kim, “Regulation of microRNA biogenesis,” *Nature Reviews Molecular Cell Biology*, vol. 15, no. 8, pp. 509–524, 2014. [Cited on pages 9, 13, and 14]

-
- [26] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, “Efficient processing of deep neural networks: a tutorial and survey,” *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017. [Cited on page 9]
- [27] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. [Cited on page 10]
- [28] D. H. Hubel and T. N. Wiesel, “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex,” *The Journal of Physiology*, vol. 160, no. 1, p. 106, 1962. [Cited on page 10]
- [29] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *arXiv preprint arXiv:1207.0580*, 2012. [Cited on pages 10 and 21]
- [30] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, (Cambridge, MA, USA), pp. 3104–3112, MIT Press, 2014. [Cited on page 10]
- [31] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (A. Moschitti, B. Pang, and W. Daelemans, eds.), (Doha, Qatar), pp. 1724–1734, Association for Computational Linguistics, oct 2014. [Cited on page 10]
- [32] J. O’Brien, H. Hayder, Y. Zayed, and C. Peng, “Overview of microRNA biogenesis, mechanisms of actions, and circulation,” *Frontiers in Endocrinology*, vol. 9, p. 402, 2018. [Cited on pages 13 and 14]
- [33] A. Shimomura, S. Shiino, J. Kawauchi, S. Takizawa, H. Sakamoto, J. Matsuzaki, M. Ono, F. Takeshita, S. Niida, C. Shimizu, *et al.*, “Novel combination of serum microRNA for detecting breast cancer in the early stage,” *Cancer Science*, vol. 107, no. 3, pp. 326–334, 2016. [Cited on page 13]

- [34] W. Usuba, F. Urabe, Y. Yamamoto, J. Matsuzaki, H. Sasaki, M. Ichikawa, S. Takizawa, Y. Aoki, S. Niida, K. Kato, *et al.*, “Circulating miRNA panels for specific and early detection in bladder cancer,” *Cancer Science*, vol. 110, no. 1, pp. 408–419, 2019. [Cited on page 13]
- [35] R. Liu, X. Chen, Y. Du, W. Yao, L. Shen, C. Wang, Z. Hu, R. Zhuang, G. Ning, C. Zhang, *et al.*, “Serum microRNA expression profile as a biomarker in the diagnosis and prognosis of pancreatic cancer,” *Clinical Chemistry*, vol. 58, no. 3, pp. 610–618, 2012. [Cited on page 13]
- [36] F. Jansen, X. Yang, S. Proebsting, M. Hoelscher, D. Przybilla, K. Baumann, T. Schmitz, A. Dolf, E. Endl, B. S. Franklin, *et al.*, “MicroRNA expression in circulating microvesicles predicts cardiovascular events in patients with coronary artery disease,” *Journal of the American Heart Association*, vol. 3, no. 6, p. e001249, 2014. [Cited on page 13]
- [37] S. Ikeda, S. W. Kong, J. Lu, E. Bisping, H. Zhang, P. D. Allen, T. R. Golub, B. Pieske, and W. T. Pu, “Altered microRNA expression in human heart disease,” *Physiological Genomics*, vol. 31, no. 3, pp. 367–373, 2007. [Cited on page 13]
- [38] V. Neudecker, M. Haneklaus, O. Jensen, L. Khailova, J. C. Masterson, H. Tye, K. Biette, P. Jedlicka, K. S. Brodsky, M. E. Gerich, *et al.*, “Myeloid-derived miR-223 regulates intestinal inflammation via repression of the NLRP3 inflammasome,” *Journal of Experimental Medicine*, vol. 214, no. 6, pp. 1737–1752, 2017. [Cited on page 13]
- [39] L.-L. Lv, Y. Feng, M. Wu, B. Wang, Z.-L. Li, X. Zhong, W.-J. Wu, J. Chen, H.-F. Ni, T.-T. Tang, *et al.*, “Exosomal miRNA-19b-3p of tubular epithelial cells promotes M1 macrophage activation in kidney injury,” *Cell Death & Differentiation*, vol. 27, no. 1, pp. 210–226, 2020. [Cited on page 13]
- [40] C. Wang, C. Zhang, L. Liu, B. Chen, Y. Li, J. Du, *et al.*, “Macrophage-derived mir-155-containing exosomes suppress fibroblast proliferation and promote fibroblast inflammation during cardiac injury,” *Molecular Therapy*, vol. 25, no. 1, pp. 192–204, 2017. [Cited on page 13]
- [41] D. M. Santarelli, A. P. Carroll, H. M. Cairns, P. A. Tooney, and M. J. Cairns, “Schizophrenia-associated microRNA-gene interactions in the

- dorsolateral prefrontal cortex,” *Genomics, Proteomics & Bioinformatics*, vol. 17, no. 6, pp. 623–634, 2019. [Cited on page 13]
- [42] C. B. Brum, V. R. Paixao-Cortes, A. M. Carvalho, T. Martins-Silva, M. X. Carpena, K. F. Ulguim, K. Y. S. Luquez, A. Salatino-Oliveira, and L. Tovo-Rodrigues, “Genetic variants in miRNAs differentially expressed during brain development and their relevance to psychiatric disorders susceptibility,” *The World Journal of Biological Psychiatry*, vol. 22, no. 6, pp. 456–467, 2021. [Cited on page 13]
- [43] D. M. Santarelli, N. J. Beveridge, P. A. Tooney, and M. J. Cairns, “Up-regulation of Dicer and microRNA expression in the dorsolateral prefrontal cortex brodmann area 46 in schizophrenia,” *Biological Psychiatry*, vol. 69, no. 2, pp. 180–187, 2011. [Cited on page 13]
- [44] A. M. Ardekani and M. M. Naeini, “The role of microRNAs in human diseases,” *Avicenna Journal of Medical Biotechnology*, vol. 2, no. 4, p. 161, 2010. [Cited on page 13]
- [45] T. Conrad, E. Ntini, B. Lang, L. Cozzuto, J. B. Andersen, J. U. Marquardt, J. Ponomarenko, G. G. Tartaglia, and U. A. V. Ørom, “Determination of primary microRNA processing in clinical samples by targeted pri-miR-sequencing,” *RNA*, vol. 26, no. 11, pp. 1726–1730, 2020. [Cited on page 13]
- [46] K. J. Beezhold, V. Castranova, and F. Chen, “Microprocessor of microRNAs: regulation and potential for therapeutic intervention,” *Molecular Cancer*, vol. 9, no. 1, p. 134, 2010. [Cited on page 14]
- [47] B. Czech and G. J. Hannon, “Small RNA sorting: matchmaking for Argonautes,” *Nature Reviews Genetics*, vol. 12, no. 1, pp. 19–31, 2011. [Cited on page 14]
- [48] V. N. Kim, J. Han, and M. C. Siomi, “Biogenesis of small RNAs in animals,” *Nature Reviews Molecular Cell Biology*, vol. 10, no. 2, pp. 126–139, 2009. [Cited on page 14]
- [49] T. N.-Y. Le, C. T. Le, and T. A. Nguyen, “Determinants of selectivity in the dicing mechanism,” *Nature Communications*, vol. 15, no. 1, p. 8989, 2024. [Cited on pages 14 and 63]

- [50] P.-W. Lau, K. Z. Guiley, N. De, C. S. Potter, B. Carragher, and I. J. MacRae, “The molecular architecture of human Dicer,” *Nature Structural & Molecular Biology*, vol. 19, no. 4, pp. 436–440, 2012. [Cited on page 14]
- [51] T. D. Nguyen, T. A. Trinh, S. Bao, and T. A. Nguyen, “Secondary structure RNA elements control the cleavage activity of DICER,” *Nature Communications*, vol. 13, no. 1, p. 2138, 2022. [Cited on pages 14, 15, and 63]
- [52] G. Hutvagner, J. McLachlan, A. E. Pasquinelli, É. Bálint, T. Tuschl, and P. D. Zamore, “A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA,” *Science*, vol. 293, no. 5531, pp. 834–838, 2001. [Cited on page 14]
- [53] A. Vermeulen, L. Behlen, A. Reynolds, A. Wolfson, W. S. Marshall, J. Karpilow, and A. Khvorova, “The contributions of dsRNA structure to Dicer specificity and efficiency,” *RNA*, vol. 11, no. 5, pp. 674–682, 2005. [Cited on page 14]
- [54] J. Starega-Roslan, P. Galka-Marciniak, and W. J. Krzyzosiak, “Nucleotide sequence of miRNA precursor contributes to cleavage site selection by Dicer,” *Nucleic Acids Research*, vol. 43, no. 22, pp. 10939–10951, 2015. [Cited on page 14]
- [55] Q.-J. Luo, J. Zhang, P. Li, Q. Wang, Y. Zhang, B. Roy-Chaudhuri, J. Xu, M. A. Kay, and Q. C. Zhang, “RNA structure probing reveals the structural basis of Dicer binding and cleavage,” *Nature Communications*, vol. 12, no. 1, p. 3397, 2021. [Cited on page 14]
- [56] L. J. Wee, T. W. Tan, and S. Ranganathan, “SVM-based prediction of caspase substrate cleavage sites,” *BMC Bioinformatics*, vol. 7, no. Suppl 5, p. S14, 2006. [Cited on page 14]
- [57] L. J. Wee, T. W. Tan, and S. Ranganathan, “CASVM: web server for SVM-based prediction of caspase substrates cleavage sites,” *Bioinformatics*, vol. 23, no. 23, pp. 3241–3243, 2007. [Cited on page 14]
- [58] D. A. DuVerle, Y. Ono, H. Sorimachi, and H. Mamitsuka, “Calpain cleavage prediction using multiple kernel learning,” *PLOS ONE*, vol. 6, no. 5, p. e19035, 2011. [Cited on page 14]

- [59] M. Piippo, N. Lietzén, O. S. Nevalainen, J. Salmi, and T. A. Nyman, “Pripper: prediction of caspase cleavage sites from whole proteomes,” *BMC Bioinformatics*, vol. 11, no. 1, p. 320, 2010. [Cited on page 14]
- [60] J. Song, H. Tan, H. Shen, K. Mahmood, S. E. Boyd, G. I. Webb, T. Akutsu, and J. C. Whisstock, “Cascleave: towards more accurate prediction of caspase substrate cleavage sites,” *Bioinformatics*, vol. 26, no. 6, pp. 752–760, 2010. [Cited on page 14]
- [61] J. Song, H. Tan, A. J. Perry, T. Akutsu, G. I. Webb, J. C. Whisstock, and R. N. Pike, “PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites,” *PLOS ONE*, vol. 7, no. 11, p. e50300, 2012. [Cited on page 14]
- [62] M. Wang, X.-M. Zhao, H. Tan, T. Akutsu, J. C. Whisstock, and J. Song, “Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets,” *Bioinformatics*, vol. 30, no. 1, pp. 71–80, 2014. [Cited on page 14]
- [63] O. Singh and E. C.-Y. Su, “Prediction of HIV-1 protease cleavage site using a combination of sequence, structural, and physicochemical features,” *BMC Bioinformatics*, vol. 17, no. Suppl 17, p. 478, 2016. [Cited on page 14]
- [64] Z.-X. Liu, K. Yu, J. Dong, L. Zhao, Z. Liu, Q. Zhang, S. Li, Y. Du, and H. Cheng, “Precise prediction of calpain cleavage sites and their aberrance caused by mutations in cancer,” *Frontiers in Genetics*, vol. 10, p. 715, 2019. [Cited on page 14]
- [65] Y.-X. Fan, Y. Zhang, and H.-B. Shen, “LabCaS: labeling calpain substrate cleavage sites from amino acid sequence using conditional random fields,” *Proteins: Structure, Function, and Bioinformatics*, vol. 81, no. 4, pp. 622–634, 2013. [Cited on page 14]
- [66] Y. Jin, Y. Yang, and P. Zhang, “New insights into RNA secondary structure in the alternative splicing of pre-mRNAs,” *RNA Biology*, vol. 8, no. 3, pp. 450–457, 2011. [Cited on page 15]
- [67] J. Singh, J. Hanson, K. Paliwal, and Y. Zhou, “RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks

- and transfer learning,” *Nature Communications*, vol. 10, no. 1, p. 5407, 2019. [Cited on page 15]
- [68] K. Sato, M. Akiyama, and Y. Sakakibara, “RNA secondary structure prediction using deep learning with thermodynamic integration,” *Nature Communications*, vol. 12, no. 1, p. 941, 2021. [Cited on page 15]
- [69] J. Y. Ryu, H. U. Kim, and S. Y. Lee, “Deep learning improves prediction of drug-drug and drug-food interactions,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 18, pp. E4304–E4311, 2018. [Cited on page 15]
- [70] P. Kumar Shukla, P. Kumar Shukla, P. Sharma, P. Rawat, J. Samar, R. Moriwal, and M. Kaur, “Efficient prediction of drug-drug interaction using deep learning models,” *IET Systems Biology*, vol. 14, no. 4, pp. 211–216, 2020. [Cited on page 15]
- [71] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, “Deep learning for computational biology,” *Molecular Systems Biology*, vol. 12, no. 7, p. 878, 2016. [Cited on page 15]
- [72] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, “Deep learning: new computational modelling techniques for genomics,” *Nature Reviews Genetics*, vol. 20, no. 7, pp. 389–403, 2019. [Cited on page 15]
- [73] S. Jin, X. Zeng, F. Xia, W. Huang, and X. Liu, “Application of deep learning methods in biological networks,” *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1902–1917, 2021. [Cited on page 15]
- [74] S. Griffiths-Jones, H. K. Saini, S. Van Dongen, and A. J. Enright, “miR-Base: tools for microRNA genomics,” *Nucleic Acids Research*, vol. 36, no. suppl_1, pp. D154–D158, 2007. [Cited on page 16]
- [75] I. L. Hofacker, “Vienna RNA secondary structure server,” *Nucleic Acids Research*, vol. 31, no. 13, pp. 3429–3431, 2003. [Cited on pages 16, 42, and 66]
- [76] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014. [Cited on page 20]
- [77] S. Ioffe and C. Szegedy, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the 32nd*

- International Conference on Machine Learning*, pp. 448–456, PMLR, 2015. [Cited on page 21]
- [78] B. W. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975. [Cited on page 23]
- [79] J. Gorodkin, “Comparing two K-category assignments by a K-category correlation coefficient,” *Computational Biology and Chemistry*, vol. 28, no. 5-6, pp. 367–374, 2004. [Cited on page 24]
- [80] W. Li and A. Godzik, “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences,” *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006. [Cited on pages 32 and 38]
- [81] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, “CD-HIT: accelerated for clustering the next-generation sequencing data,” *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012. [Cited on pages 32 and 38]
- [82] S. Zhang, R. Fan, Y. Liu, S. Chen, Q. Liu, and W. Zeng, “Applications of transformer-based language models in bioinformatics: a survey,” *Bioinformatics Advances*, vol. 3, no. 1, p. vbad001, 2023. [Cited on page 36]
- [83] D. Yao, B. Li, X. Zhan, X. Zhan, and L. Yu, “GCNFORMER: graph convolutional network and transformer for predicting lncRNA-disease associations,” *BMC Bioinformatics*, vol. 25, no. 1, p. 5, 2024. [Cited on page 36]
- [84] Z. Li, P. Ren, H. Yang, J. Zheng, and F. Bai, “TEFDTA: a transformer encoder and fingerprint representation combined prediction method for bonded and non-bonded drug-target affinities,” *Bioinformatics*, vol. 40, no. 1, p. btad778, 2024. [Cited on page 36]
- [85] S. Zhang, Y. Xu, and Y. Liang, “TMSC-m7G: a transformer architecture based on multi-sense-scaled embedding features and convolutional neural network to identify RNA N7-methylguanosine sites,” *Computational and Structural Biotechnology Journal*, vol. 23, pp. 129–139, 2024. [Cited on pages 36 and 66]
- [86] A. L. Maas, A. Y. Hannun, A. Y. Ng, *et al.*, “Rectifier nonlinearities improve neural network acoustic models,” in *Proceedings of the 30th*

- International Conference on Machine Learning*, vol. 30, p. 3, 2013. [Cited on page 43]
- [87] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. [Cited on page 43]
- [88] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, “Attentional feature fusion,” in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3560–3569, 2021. [Cited on pages 43, 45, and 66]
- [89] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017. [Cited on page 48]
- [90] I. J. MacRae, K. Zhou, F. Li, A. Repic, A. N. Brooks, W. Z. Cande, P. D. Adams, and J. A. Doudna, “Structural basis for double-stranded RNA processing by Dicer,” *Science*, vol. 311, no. 5758, pp. 195–198, 2006. [Cited on page 63]
- [91] I. J. MacRae, K. Zhou, and J. A. Doudna, “Structural determinants of RNA recognition and cleavage by Dicer,” *Nature Structural & Molecular Biology*, vol. 14, no. 10, pp. 934–940, 2007. [Cited on pages 63 and 66]
- [92] J.-E. Park, I. Heo, Y. Tian, D. K. Simanshu, H. Chang, D. Jee, D. J. Patel, and V. N. Kim, “Dicer recognizes the 5’ end of RNA for efficient and accurate processing,” *Nature*, vol. 475, no. 7355, pp. 201–205, 2011. [Cited on page 63]
- [93] S. M. Elbashir, J. Harborth, W. Lendeckel, A. Yalcin, K. Weber, and T. Tuschl, “Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells,” *Nature*, vol. 411, no. 6836, pp. 494–498, 2001. [Cited on page 63]
- [94] E. Bernstein, A. A. Caudy, S. M. Hammond, and G. J. Hannon, “Role for a bidentate ribonuclease in the initiation step of RNA interference,” *Nature*, vol. 409, no. 6818, pp. 363–366, 2001. [Cited on page 63]
- [95] R. F. Ketting, S. E. Fischer, E. Bernstein, T. Sijen, G. J. Hannon, and R. H. Plasterk, “Dicer functions in RNA interference and in synthesis of

- small RNA involved in developmental timing in *C. elegans*,” *Genes & Development*, vol. 15, no. 20, pp. 2654–2659, 2001. [Cited on page 63]
- [96] S. W. Knight and B. L. Bass, “A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in *Caenorhabditis elegans*,” *Science*, vol. 293, no. 5538, pp. 2269–2271, 2001. [Cited on page 63]
- [97] Y. Weng, H. Xiao, J. Zhang, X.-J. Liang, and Y. Huang, “RNAi therapeutic and its innovative biotechnological evolution,” *Biotechnology Advances*, vol. 37, no. 5, pp. 801–825, 2019. [Cited on page 64]
- [98] E. Sardh, P. Harper, M. Balwani, P. Stein, D. Rees, D. M. Bissell, R. Desnick, C. Parker, J. Phillips, H. L. Bonkovsky, *et al.*, “Phase 1 trial of an RNA interference therapy for acute intermittent porphyria,” *New England Journal of Medicine*, vol. 380, no. 6, pp. 549–558, 2019. [Cited on page 64]
- [99] S. Agarwal, A. R. Simon, V. Goel, B. A. Habtemariam, V. A. Clausen, J. B. Kim, and G. J. Robbie, “Pharmacokinetics and pharmacodynamics of the small interfering ribonucleic acid, givosiran, in patients with acute hepatic porphyria,” *Clinical Pharmacology & Therapeutics*, vol. 108, no. 1, pp. 63–72, 2020. [Cited on page 64]
- [100] B. Hu, L. Zhong, Y. Weng, L. Peng, Y. Huang, Y. Zhao, and X.-J. Liang, “Therapeutic siRNA: state of the art,” *Signal Transduction and Targeted Therapy*, vol. 5, no. 1, p. 101, 2020. [Cited on page 64]
- [101] S. Yao, B. Peng, C. Papadimitriou, and K. Narasimhan, “Self-attention networks can process bounded hierarchical languages,” *arXiv preprint arXiv:2105.11115*, 2021. [Cited on page 65]
- [102] Y. Feng, X. Zhang, P. Graves, and Y. Zeng, “A comprehensive analysis of precursor microRNA cleavage by human Dicer,” *RNA*, vol. 18, no. 11, pp. 2083–2092, 2012. [Cited on page 66]
- [103] W. He, Y. Wang, L. Cui, R. Su, and L. Wei, “Learning embedding features based on multisense-scaled attention architecture to improve the predictive performance of anticancer peptides,” *Bioinformatics*, vol. 37, no. 24, pp. 4684–4693, 2021. [Cited on page 66]