MASARYK
UNIVERSITY

FACULTY OF INFORMATICS

# Modelling small RNA binding using Convolutional Neural Networks

Master's Thesis

EVA KLIMENTOVÁ

Brno, Spring 2022

MASARYK UNIVERSITY

FACULTY OF INFORMATICS

# Modelling small RNA binding using Convolutional Neural Networks

Master's Thesis

EVA KLIMENTOVÁ

Advisor: Panagiotis Alexiou, PhD

Department of Machine Learning and Data Processing

Brno, Spring 2022


FACULTAS ARTIS INFORMATICÆ
UNIVERSITAS MASARYKIANA

## Declaration

Hereby I declare that this paper is my original authorial work, which I have worked out on my own. All sources, references, and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Eva Klimentová

**Advisor:** Panagiotis Alexiou, PhD

## Acknowledgements

I would like to thank my supervisor Panagiotis Alexiou for his valuable advice during the project realization and thesis writing. Many thanks to all colleagues from Panagiotis Alexiou Research Group and Bioinformatics Core Facility who cooperated with me on this project, namely Ilektra - Chara Giassa, Václav Hejret, Katarína Grešová and Ján Krčmář. The last but most important thanks go to my boyfriend who stood by me, supported me and cooked for me during the hard times.

## Abstract

Gene expression regulation by Ago-loaded small RNAs is a complex but essential process across species. Prediction of small RNA – target site binding is an important first step in all small RNA target prediction programs. To date, there are two widely used techniques for small RNA – target site prediction: seed and cofold. Limitations of both these techniques have presented target prediction tools selectively focusing on targets with "canonical seed", although unbiased experiments have shown that less than 50 % of the small RNA targets are "canonical".

In this thesis, we present a machine learning method for the prediction of potential small RNA – target site binding. It is trained on seed-unbiased experimental data and we show that our method outperforms state-of-the-art approaches. The code, data and a web server for two projects included in this thesis are available at `https://github.com/ML-Bioinfo-CEITEC/miRBind` and `https://github.com/evaklimentova/smallRNA_binding`.



## Keywords

bioinformatics, Convolutional Neural Network, small RNA binding, small RNA – target prediction, CLASH

# Contents

# List of Tables

# List of Figures

# Abbreviations

**ago** argonaute

**AU PRC** area under the PR curve

**bp** base pairs

**CLASH** Crosslinking, Ligation, And Sequencing of Hybrids

**CLIP** CrossLinking ImmunoPrecipitation

**CNN** Convolutional Neural Network

**miRNA** microRNA

**ML** machine learning

**mRNA** messenger RNA

**NN** Neural Network

**PR curve** precision recall curve

**snoRNA** small nucleolar RNA

**tRF** tRNA fragment

**tRNA** transfer RNA

# Introduction

MicroRNAs and potentially other types of small RNAs post-transcriptionally regulate gene expression across species. The regulation works on the principle of Ago protein-mediated small RNA binding to its mRNA target. Gene expression regulation by small RNAs plays a role in multiple essential processes and the effects of its dysfunction may be important in many diseases such as schizophrenia, Alzheimer's disease or cancer. Identification of small RNA – mRNA target interactions is thus crucial for the exploration of the regulation network. The easiest way is usually using some computational prediction tool followed by experimental validation of predicted pairs.

An important part of the prediction of new small RNA and targets they are able to regulate is predicting the first step of this machinery – small RNA target Ago-mediated binding. There exist plenty of target prediction tools but the majority of them use in their first filtering step to determine the potential small RNA target binding place either co-folding free energy measures or approaches based on the identification of small RNA seed region binding. Limitations of both these techniques have presented target prediction tools selectively focusing on targets with "canonical seed", although unbiased experiments have shown that less than 50 % of the small RNA targets are "canonical".

In this thesis, we present a machine learning method for the prediction of potential small RNA – target site binding. The method is based on two experimental high throughput CLASH datasets, one focusing on Ago1 interactions, the second on Ago2. We show that our method outperforms state-of-the-art tools for small RNA target binding site recognition. To open up our method to the biology researchers we develop a user-friendly standalone tool as well as a web server where they can evaluate their potential small RNA – target pairs.

The thesis is divided into five chapters. Chapter 1 introduces the biological background necessary to understand the concept and goals of this work. Chapter 2 focuses on an overview of the fields of deep learning and its relation to genomics. In chapter 3, methods used to build the small RNA target binding site prediction tool are summarized and chapter 4 follows by presenting the achieved results in a state-of-

the-art context. Finally, the chapter 5 summarizes the findings of this thesis and discusses further potential extensions and improvements.

# 1 Biological background

This chapter will briefly introduce multiple terms and concepts from biology to help readers properly understand the context and goals of this thesis. The first section 1.1 presents small RNAs and their different types, section 1.2 deals with Ago proteins and their function and introduces the problem of small RNA target binding prediction.

## 1.1 Small RNAs

The discovery of small RNAs can be traced back to the 1950s when transfer RNA and ribosomal RNA were discovered [1, 2]. Over the years, a number of similar small RNA molecules that do not function as messenger RNAs (mRNAs) were discovered and the term small RNA was established. It is used for RNA molecules that are shorter than 200 nucleotides long and are usually non-coding. The small RNAs have an indispensable role in a wide range of cell functions such as gene silencing, RNA processing and modification, or gene expression regulation [3].

The following paragraphs briefly present some types of small RNAs.

**MicroRNAs (miRNAs)** are approximately 22 nucleotide long small RNAs that negatively regulate gene expression at the level of mRNA. MiRNAs are processed from longer RNA sequences called pri-miRNAs which contain a region that folds back on itself and forms a hairpin structure. These folded molecules are subsequently processed and cleaved into small double-stranded RNAs. One strand (or both) of the miRNA duplex is then called the mature miRNA which then plays its role in mRNA repression [4].

**Small nucleolar RNAs (snoRNAs)** are about 60 – 300 nucleotides long molecules that are mainly found in the nucleolus. Their primary function is the posttranscriptional modification of ribosomal RNA or transfer RNA [5].

**Transfer RNAs (tRNAs)** are typically between 70 and 100 nucleotides long molecules folding into the cloverleaf structure. They have a key role in RNA translation into protein by recognizing specific tri-nucleotide codon and attaching appropriate amino acid to the growing polypeptide chain [6].

**Vault RNAs** were firstly discovered as part of the largest ribonucleoprotein complexes named "vault" and the group consists of only four members in human. It has been shown that they are involved in central signalling pathways and cell to cell communication but most of their function is still unknown and their role is still under investigation [7].

**YRNAs** are about 100 nucleotides long with a stem-loop structure. Even though they were discovered more than 40 years ago, their role is still under investigation. They are for example reported to be important for DNA replication. In total there exist only four different YRNAs [8].

## 1.2 Ago proteins

The Argonaute (ago) protein family is a widely conserved set of proteins found in multiple species ranging from yeast to humans and plants. In mammals, the Ago protein family is divided into two clades, Ago and PIWI, each associated with small RNA driver sequences that can target RNAs using some form of sequence complementarity. There are four Ago proteins (Ago1 – 4) known in mammals that primarily interact with miRNAs and form a complex called the RNA-induced silencing complex (RISC) (Figure 1.1). Loaded miRNA then guides Ago through base pairing to target mRNAs and regulate translation. Of the four Ago proteins found in mammalian cells, Ago2 is unique in its "slicer" ability which allows it to cleave highly complementary targets [9]. Experiments performed on mice [9, 10] show that Ago2 is the most important member of a possibly partially redundant family of Ago proteins.

Primarily associated drivers of Ago proteins are miRNAs. However, it is becoming more and more apparent that other small RNA molecules can also be loaded to Ago proteins and may act as targeting drivers similarly to miRNAs [11].

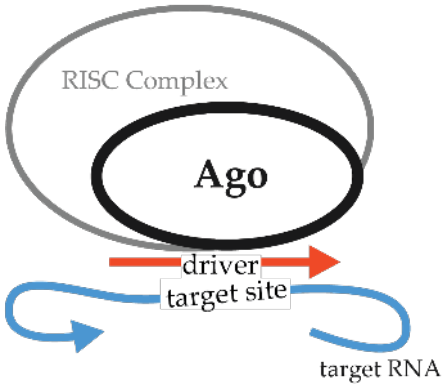


**Figure 1.1:** Interaction between small RNA and target RNA

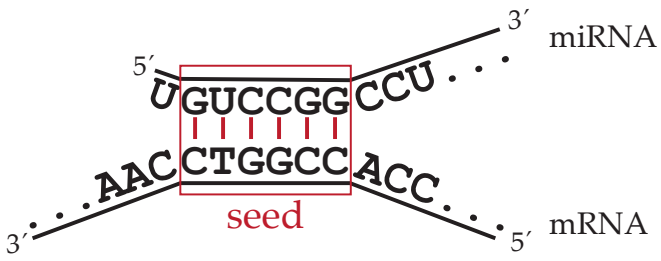


**Figure 1.2:** miRNA – mRNA "canonical" seed binding

The exact process of how small RNA sequences loaded to Ago protein bind to their mRNA targets has been inspected by many studies mostly working with miRNA-target interactions. The early identified pattern important in miRNA target recognition was the seed region pairing. The "canonical" seed is a stretch of six to eight nucleotides starting from position 2 from the 5′ end of the miRNA molecule which binds to the target sequence by perfect nucleotide complementarity (Figure 1.2), while it is known that "non-canonical" seeds allowing a small number of mismatches or bulges are also functional [12]. However, also non-seed interactions have been known for a long time [13].

### 1.2.1 Prediction of miRNA targets

Early computational approaches to miRNA target prediction were heavily based on seed complemented with additional features such as evolutionary conservation of targets, the position of the target sites on 3'-UTRs, nucleotide content and others [12]. Other types of miRNA target prediction methods utilized alignment or co-fold methodologies ignoring the seed region. In these approaches, an ideal structure is calculated based on the affinity of the miRNA sequence to its target sequence, and then measures such as alignment score or free energy of binding of the two molecules are used to score binding probability [14].

When the first high-throughput miRNA targeting datasets became available [15, 16], it showed up that seed-based approaches outperform "cofold" based methods [17]. The following years produced a wealth of high-throughput miRNA targeting data utilizing methods such as CrossLinking ImmunoPrecipitation (CLIP) sequencing that identified thousands of miRNAs and their targets [18]. An important limitation of these techniques is that they do not produce specific miRNA – target site pairs. Instead, they produce peaks of Ago protein binding, to which miRNAs need to be assigned. This is often done by using a miRNA target prediction program, which usually utilises the seed heuristic [19]. This creates a feedback loop of "seed bias" in which putative target sites with "canonical" seeds score high for miRNA target prediction programs based on a seed heuristic, and then in turn are prioritized for experimental validation. These validated targets are in turn used to train the new generations of miRNA target prediction programs. Even though more functional "non-canonical" seed binding sites are being continuously discovered, they remain underrepresented by all miRNA prediction programs and databases of validated miRNA targets.

Until 2013 when Crosslinking, Ligation, And Sequencing of Hybrids (CLASH) protocol was presented [20], no unbiased high-throughput experiment was performed to directly identify RNA – RNA Ago mediated interactions. In the CLASH method, RNA – protein complexes are stabilized and the Ago-loaded small RNA and RNA interacting molecules are intermolecularly ligated. RNA – protein complexes are then pulled out using immunoprecipitation and RNA se-

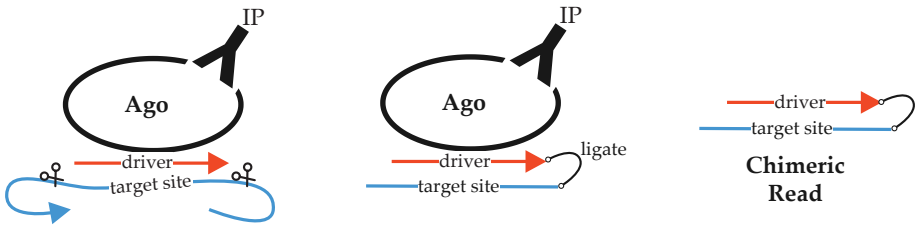


**Figure 1.3:** Schematic of the CLASH experiment

quences are sequenced (Figure 1.3). CLASH sequencing results offer three types of information: precise Ago-binding sites on RNAs (similar to CLIP methods), Ago-loaded driver sequences and also RNA driver – RNA target pairs (called chimeric reads) mediated by Ago. This new CLASH experiment presented that only approximately 20 % of identified miRNA – target chimeras have "canonical" seed and about 20 % chimeras do not show any type of seed binding at all [20]. This strengthens the need of using different algorithms than the seed heuristic as a first filtering step in the miRNA target prediction programs.

### 1.2.2 Non-miRNA Ago drivers

The small RNAs primarily associated with Ago proteins are miRNAs. However, it has been demonstrated that other small RNA molecules can be loaded to Ago too. It has been shown that for example tRNA fragments (tRFs) can associate with Ago proteins [21] and silence their targets in a similar manner to miRNAs [22]. When data from Ago1 CLASH experiment [20] were reanalyzed, it showed out that miRNAs are not the absolute majority in the chimeric reads [23]. In a CLASH Ago2 experiment performed in CEITEC (see section 3.1.2), approximately 14,000 chimeric target sites were found. Less than 50 % of these target sites were associated with miRNAs, the rest was associated with tRNA, snoRNA, vault RNA and YRNA fragments.

Until recently, non-miRNA target prediction tools were limited to miRNA target prediction tools. In the past couple of years, several tRF prediction tools were introduced [24, 25, 26]. They work on the same

ideas as miRNA prediction tools and use cofold or seed match. So far any other non-miRNA prediction tools have been developed.

# 2 Bioinformatics and Machine Learning Approaches

This chapter serves as a brief introduction to the terms such as machine learning and Neural Network. It will also provide a number of practical examples that highlight the applications of machine learning on biological data.

## 2.1 Machine learning

*Machine learning* (*ML*) is a set of algorithms that are able to learn concepts and extract patterns from raw data. The ML algorithms can help with many real-world problems however the behaviour of these algorithms heavily depends on the input data representation. Many tasks can be easily solved if appropriate features are extracted from the input data and provided to a simple ML algorithm. Nonetheless, for some problems, it might be challenging to construct such features [27].

A subset of ML called *Deep learning* solves the problem of representation by its ability to obtain complex concepts expressed using simpler ones. This eliminates the issue of handcrafted features by incorporating the computation of these features into the deep learning model itself [28].

## 2.2 Neural Network

One of the earliest algorithms in deep learning that has survived until today was based on the idea of how learning in the brain could work. From this concept emerged *Artificial Neural Networks*, usually called simply Neural Network (NN). They consist of multiple layers where the first layer is the input layer, the final layer is named the output layer and all the remaining layers between them are hidden layers. Each of these layers is then composed of individual nodes termed *artificial neurons* which are interconnected between individual subsequent layers. To go from one layer to the next one, the neurons calculate a weighted sum of their inputs and pass the result through a non-linear function [28].

## 2.3 Convolutional Neural Networks

*Convolutional Neural Networks* (*CNNs*) are a special type of NNs that are designed to process data composed of multiple arrays. They are typically used for data such as 2D images or 3D video. The classical CNNs are contain two special types of layers: convolutional and pooling. Neurons in convolutional layers are connected to the local regions of the previous layer and compute the scalar product between their weights and the connected region. The result is then passed through a non-linear function. Neurons in the convolutional layer are organized to feature maps where the neurons from one feature map use the same set of weights called the kernel. The idea behind this architecture is that each set of neurons in the same feature map searches for one simple pattern independently of its position in the input layer. The role of the pooling layer is to downsample the convolutional layer output and merge similar features into one. A pooling neuron typically computes the maximum of a local area of neurons in one or multiple feature maps. Typical CNN consists of multiple stacked convolutional and pooling layers, followed by classical fully-connected layers [27].

Thanks to their architecture CNNs are able to extract low-level features and compose them into higher-level features. An example in images may be the detection of simple patterns like edges which can be then merged into simpler objects or their parts which can form bigger objects [28]. The same idea can be applied to DNA sequences. In them, simple motifs and motif interactions can be found, from which a function may emerge [29].

## 2.4 Deep learning in the context of genomics

In the last couple of years, there has been an explosion of publications presenting deep learning approaches to study the genome. This could have happened thanks to the fast development of high-throughput methods for analyzing genome structures and functions and thus the availability of large datasets [30]. It has been demonstrated that deep learning suits well tasks related to genomics as the multiple layers can capture complex multi-level information. In comparison to the older tools, deep learning methods do not need any handcrafted features

as they can extract the features straight from the raw data and they are able to catch not only individual motifs but also their higher-level interactions [31]. Deep learning techniques can be nowadays found in areas such as prediction of splicing [32], methylation status [33] or sequence specificity of binding proteins [34]. Another interesting example of deep learning usage is the application of the natural language processing model on DNA sequences [35].

# 3 Methods

This chapter outlines the datasets and methods used in the CNN construction. Section 3.1 describes all datasets used, their origin and preprocessing steps. Section 3.2 covers the NN technicalities such as architecture or evaluation metrics.

## 3.1 Datasets

The major problem in the field of small RNA target prediction is the lack of available data. These days there is only one published unbiased high-throughput dataset focusing on Ago1 miRNA drivers and their targets, namely the CLASH experiment from 2013 [20] (called CLASH Ago1 dataset in this thesis). The second dataset which will be used in this thesis is not yet published and comes from an experiment done in CEITEC (called CLASH Ago2 dataset). It is similar to the CLASH 2013 dataset but is done not with Ago1 protein but Ago2 and brought in addition to miRNA targets a lot of other non-miRNA small RNA targets. Yet another dataset can potentially arise from a new eCLIP experiment [36] when released.

The following subsections describe in detail the methods used for obtaining the experimental CLASH Ago1 and Ago2 data and their processing.

### 3.1.1 CLASH Ago1 dataset

Supplementary text file S1 with information about miRNAs and their targets was downloaded from Helwak et al. [20] supplement. To fit the NN input shape, obtained miRNA sequences were cut to contain only the first 20 bases, whereas shorter sequences were left untouched. Target sequences needed to be reshaped to the length of 50 thus their coordinates were centred around the original target middle point and resized to the window length of 50 base pairs (bp). A custom database of human transcripts was downloaded from *hyb* pipeline [1] [37] which was originally used to process experimental CLASH Ago1 data. Adjusted target sequences were obtained from the modified coordinates

1. `https://github.com/gkudla/hyb/blob/master/data/db/hOH7.fasta.gz`

using bedtools [38] and *hyb* transcript database. The processed dataset with 20 bp long miRNAs and 50 bp long targets was called the positive dataset of the classification problem and contained 18,392 pairs.

The dataset was further divided into training, validation and testing set composed of 15,392, 2,000 and 1,000 miRNA – target pairs. The splitting could not be done based on chromosome number, which is usual when working with DNA sequences, because the original dataset did not contain this information, thus the dataset was split randomly.

The negative sets were formed by matching randomly selected miRNA from the positive set and randomly selected target from the positive set, taking care that the selected pair is not present in the positive set. This approach was selected because it reflects the real situation – for example in Ago CLIP-seq experiments when it is needed to match sequenced miRNAs to their targets on Ago-CLIP peaks. When CLIP-seq experiments are nowadays done, peaks are often assigned to the miRNAs using seed or cofold methods.

From each positive dataset (training, evaluation and testing) three final datasets with different positive : negative ratios were composed – 1:1, 1:10 and 1:100. The 1:1 dataset represents the classically balanced classification task whereas the 1:10 and 1:100 datasets reflect better the realistic scenario since non-target sequences usually outnumber the target sequences.

### 3.1.2 CLASH Ago2 dataset

The sequencing data for this dataset were obtained by performing CLASH experiment on Ago2 which was done in CEITEC [2]. This section will explain a bit deeper how to get from the raw sequencing data to the dataset of chimeras and what problems and difficulties may arise in this process.

The processing starts with high throughput Illumina sequencing reads on the input. At first, the unique molecular identifiers are extracted and the quality of reads is checked. If any adapters are detected in the quality control step, they are trimmed as well as low-quality reads or bases. The preprocessed reads are mapped to the human

2. Thanks belong to Nandan Mysore Varadarajan and prof. Štěpánka Vaňáčová

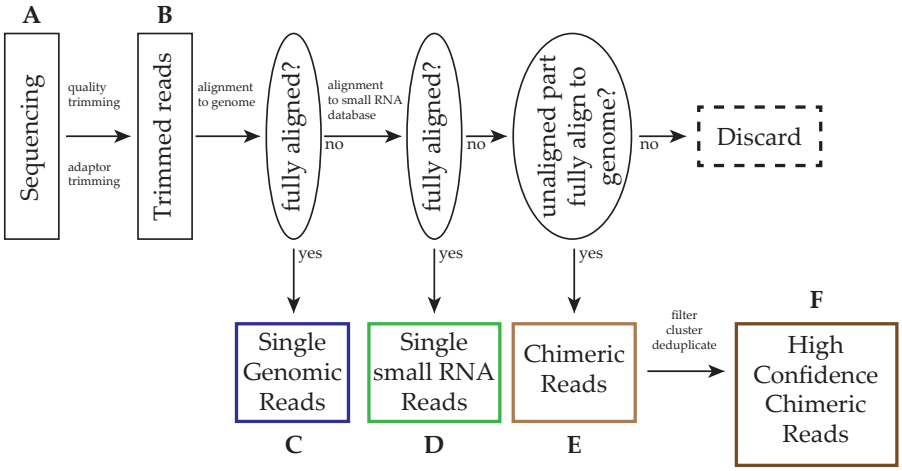


**Figure 3.1:** Visualization of individual CLASH Ago2 pipeline steps

GRCh38 reference genome with very strict settings to detect only precisely aligned target genomic reads. All remaining reads from the pure genomic read separation step were aligned to a custom database of small RNAs containing miRNAs, tRNAs, snoRNAs, vault RNAs and YRNAs. Reads fully aligned to the small RNA database were separated and saved for future usage. Reads aligned to a small RNA database with flanking part longer than 16 bp were considered as potential chimeric reads and further examined. The rest of the reads was trashed. From potential chimeric reads unaligned part was extracted and mapped to the human genome. Only reads with properly mapped target genomic part were kept. To obtain only high confidence chimeras, the chimeric reads were further collapsed, clustered and filtered and the genomic targets were standardized to the length of 50 bp. Schema of the whole pipeline [3] is outlined in Figure 3.1.

To illustrate the yield of the reads in individual steps of the pipeline in comparison to the input raw reads count, see Table 3.1. What can be immediately seen is that out of the initial more than 800 million reads there are only 14,205 unique chimeras, which makes the yield 0.0002 %. This low number makes the CLASH experiment very demanding. The

3. Credits for the whole pipeline design and development goes to Václav Hejret

**Table 3.1:** The number of reads in individual steps of the CLASH Ago2 pipeline. Labels of the pipeline steps correspond to the labels in Figure 3.1.

| Pipeline step | Reads number |
|---|---|
| (A) Sequenced reads | 812,717,759 |
| (B) Trimmed reads | 730,731,281 |
| (C) Single genomic target | 194,551,480 |
| (D) Single small RNA driver | 119,640,814 |
| (E) Chimeras partially mapped to genomic targets | 16,817,797 |
| (F) Chimeras after deduplication and filtering | 14,205 |

issue of losing a large number of reads occurs in multiple steps. Due to the low efficiency of the intermolecular ligation step of the CLASH protocol, there is a much higher number of only single genomic target reads and single small RNA reads than the potential chimeric reads. The second major issue is in the final filtering step, where small RNA chimeric parts, that align with the same confidence to more than one type of small RNA are trashed.

For the ML part, the chimeric reads were sorted into six sets based on the small RNA driver type. Only half of them (miRNAs, tRNAs and YRNAs) could be used as a positive set for training because the rest (snoRNAs, and vault RNAs) contained less than 1,000 chimeras. From the miRNA set, two different datasets were prepared: one with the whole mature miRNA sequence taken from the miRBase database [39] and the second one with the truly observed part of the miRNA sequence. As the classical whole tRNA and YRNA sequences are much longer than the observed fragments loaded to Ago, only datasets with truly observed sequences could be constructed for tRNAs and YRNAs. The disadvantage of using the truly observed sequences is that due to the ligation and sequencing process the Ago-loaded sequence may be incomplete, which may for example shift the seed region.

All positive datasets were further processed by cutting small RNA sequences from the beginning to the length of 20 bp and splitting them into train and test sets based on the target chromosome number.

**Table 3.2:** Number of small RNA – target pairs in individual positive datasets

| Small RNA type | Train set samples | Test set samples |
|---|---|---|
| miRNA | 3,860 | 719 |
| tRNA | 6,638 | 1,195 |
| YRNA | 956 | 163 |

Table 3.2 lists the number of small RNA – target tuples in each of the datasets.

The negative sets were formed the same way as in the CLASH Ago1 dataset. The final training sets were composed of positive : negative ratio 1:10, the test sets had all 1:1, 1:10 and 1:100 ratios.

### 3.1.3 Data representation

As the problem we are dealing with is a binding of two RNA molecules, it is indisputable that Watson-Crick base pairing is important in this task. We decided to help the model with understanding the binding rules and thus elected a sequence agnostic approach where we completely hide the small RNA and target sequences from the input. Instead of using standard one hot encoded sequences, we build a 20 (small RNA size) x 50 (target size) matrix in which any Watson-Crick binding nucleotide pair (A – T, C – G) is represented by 1, and any non-binding pair or empty space in shorter sequences by 0. An example of an encoded small RNA and target pair is shown in Figure 3.2.

## 3.2 Neural Network architecture

For the building of the classifier recognizing small RNA – target pairs, Convolutional Neural Networks were chosen. Due to the nature of 2D image-like input data CNNs were chosen as they were shown to perform well in tasks with image-like inputs [28]. The architecture similar to PENGUINN [40] was chosen as the architecture on which all attempts were based. The architecture consists of multiple layered blocks composed of a convolutional layer, leaky ReLU, batch normalization, pooling and dropout layer. The output of the last dropout

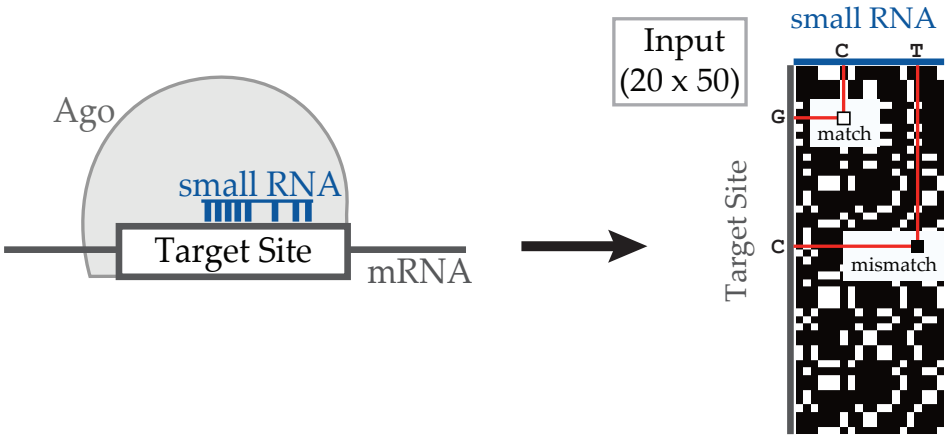


**Figure 3.2:** Example of encoded small RNA and target

layer is flattened and connected to the layered blocks of dense, leaky ReLU, batch normalization and dropout layer. The last layer is formed of a single neuron with a sigmoid activation function, which outputs the probability of input small RNA : target site binding. A schematic picture of the network architecture can be found in Figure 3.3. The network was compiled with Adam optimizer, binary cross-entropy loss function was used.

To find the best set of parameters to use, hyperparameter search was performed. It was done on the CLASH Ago1 dataset separately for all three positive : negative ratios (1:1, 1:10 and 1:100) using the train set for model training and evaluation set for comparison. Bayesian optimization implemented in Keras Tuner was used to perform the hyperparameters search. The optimized parameters were number of blocks with convolutional layer (2–6) and number of blocks with dense layer (2–6) ($\alpha$ and $\beta$ in Figure 3.3), convolutional layer kernel size (3–6), pool size of the pooling layer (2–5), dropout rate (0–0.6) and learning rate (0.0001–0.01). The total number of model configuration trials was set to 100. All the models were trained over 10 epochs with batch size 32.

For CLASH Ago2 datasets, the best performing models architecture with training dataset 1:10 ratio (see section 4.1.1) was selected for the training of all models.

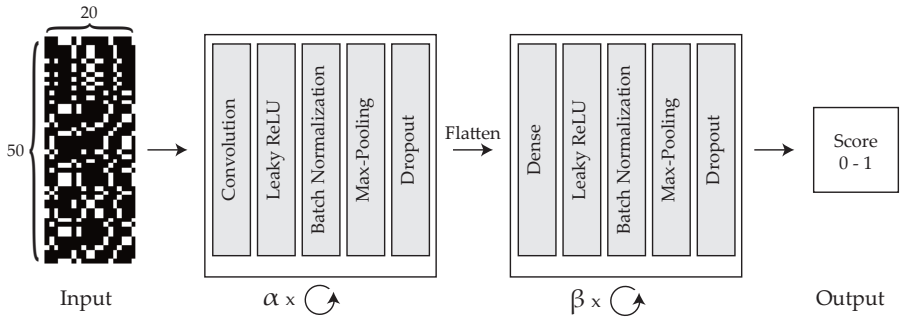


**Figure 3.3:** Compact representation of network architecture, $\alpha$ and $\beta$ are hyperparameters

The code for training and bayesian optimization can be found on CLASH Ago1 github and CLASH Ago2 github.

Another two architectures were tried for the CLASH Ago1 dataset by two different people. Katarína Grešová was working on an architecture based on a recently published transformer-based model named DNABERT [35]. DNABERT uses tokenized k-mer sequences as input and it can be fine-tuned for multiple tasks. As the input to DNABERT is a set of sequences, miRNA – target pair were converted into a single sequence, in which miRNA and target sequences are interlaid with four N nucleotides, as shown in Figure 3.4. The DNABERT model was fine-tuned on the 1:1 training set.

The second model was recently presented by Ján Krčmář [41]. The architecture used in his approach is based on the ResNet architecture [42]. Another concept used in his work is label smoothing, where the labels of samples which are hard to classify are changed to prevent the model from being too confident with its predictions. This is achieved by first training multiple small models and then evaluating them. The hard samples are then chosen and their labels are smoothed. The new regular model is then trained on the modified dataset. Jan created two final models – one "classical" ResNet and one ensemble of multiple ResNets.

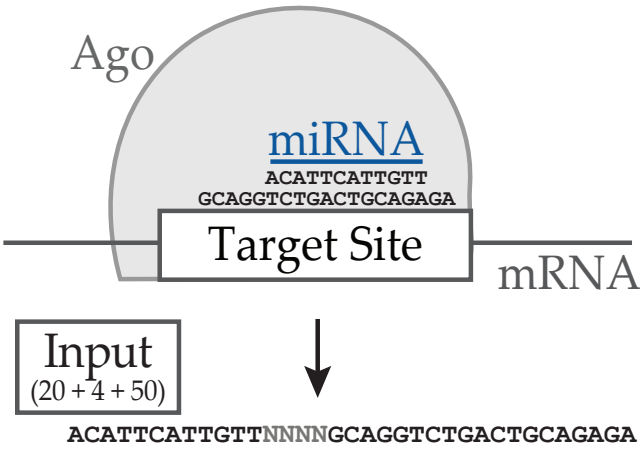


**Figure 3.4:** Representation of DNABERT model architecture input encoding

### 3.2.1 Evaluation metrics

As the problem of finding small RNA targets is imbalanced, it is important to choose appropriate metrics because, for example, a widely used accuracy metric is unsuitable for this problem. The metrics advisable for detecting rare events is precision and recall [27]. They are defined as

$$precision = \frac{TP}{TP + FP} \tag{3.1}$$

$$recall = \frac{TP}{TP + FN} \tag{3.2}$$

where TP are true positives, FP false positives and FN false negatives. For the visual idea of the model performance, precision recall curve (PR curve) is used, for direct comparison area under the PR curve (AU PRC). The AU PRC metric was also applied in the hyperparameter tuning task.

### 3.2.2 Prediction of new targets

To enable potential users to make a new prediction of small RNA targets, a standalone python script and a web server were created.

The python script allows the user to upload tsv file with small RNA and target sequences and outputs a tsv where for each small RNA – target pair a score is predicted by the loaded selected model. The web interface is a more user-friendly and intuitive way of predicting user-submitted small RNA and target binding score. It works on a similar principle as the python script using trained keras models converted to json format. The web page for miRNA – target prediction based on CLASH Ago1 dataset was developed by Ilektra - Chara Giassa, the other web page based on CLASH Ago2 was developed by the author of this thesis. All the tools described in this paragraph are available on github [4] [5].

## 3.3 Small RNA target binding state of the art

The problem with small RNA target binding prediction is that most of the available tools focus not only on the binding problem itself but go further and consider also the functionality of the binding. Classical target prediction state-of-the-art tools are thus not directly comparable to the ML methods developed in this thesis. Another problem with the comparison arises because target prediction methods usually use as the input not only small RNA and potential target sequence, but also many additional features such as conservation score of the target, upstream or downstream nucleotide content or minimum free energy. The methods with which the developed ML tool can be directly compared are thus very limited as the tool is designed more as a potential first filtering step in the classical target prediction programs, it can be used for example instead of seed.

This section describes the chosen external tools and methods used in this thesis as a state of the art comparison in target binding prediction. For every tool, there is a brief description and explanation of how it was run.

4. CLASH Ago1: `https://github.com/ML-Bioinfo-CEITEC/miRBind`
5. CLASH Ago2: `https://github.com/evaklimentova/smallRNA_binding`

### 3.3.1 Seed

"Canonical" seed consisting of a stretch of perfectly binding six nucleotides starting from position 2 at the small RNAs 5′ end was used. A custom implementation of this method was done in python with a small RNA driver and mRNA target sequences at the input. The seed section from second to seventh (including) nucleotide from the beginning of the small RNA sequence is extracted and reverse complemented. The target sequence is then searched with the reverse complemented seed for an exact match. If there is such a match, a score of 1 goes to the output, in case of not finding a match, the output is 0.

### 3.3.2 Cofold

The idea of folding methods is represented by the *RNAcofold* tool from *ViennaRNA Package* [14]. *RNAcofold* computes the hybridization energy and base-pairing pattern of an input pair of interacting RNA molecule sequences. The computed minimum free energy [43] of the folding represents how well the two molecules hold together. The lower the free energy, the stronger the binding.

Input to *RNAcofold* are small RNA and target sequences concatenated with "&" symbol, formatted to fasta file. Their minimum free energy was computed using the following command:

```
RNAcofold --noPS input.fasta > output.fasta
```

To simplify direct comparison with other tools, minimum free energy scores were normalized to the range from 0 to 1 where 1 represents the strongest binding.

### 3.3.3 RNA22

RNA22 is a method for identifying miRNA binding sites and is one of the few methods that does not rely on any additional features except miRNA and target sequences. The algorithm is based on the Markov chain which helps to find recurring patterns in miRNA sequences. Potential targets are then searched with the identified patterns and areas with accumulated hits are paired with miRNAs based on the nucleotide pairing and free energy.

The standalone version of the RNA22 program was used [6] and run with default parameters apart from "minenergy" which was set to -5. On the input, there were individual miRNA and target sequences in fasta format. The program can output either an empty file, which means that the input miRNA and target do not bind or potentially multiple exact positions of the miRNA binding to the target accompanied by a p-value representing the likelihood that the target site loci is random. Processing of the RNA22 output was done by putting score 1 to the pairs that were not recognized as binding and reporting the lowest score for pairs with multiple recognized target binding places.

6. https://cm.jefferson.edu/rna22/Interactive/remoteRNA22v2.zip

# 4 Results

## 4.1 CLASH Ago1 dataset

This section provides a summary of models trained on CLASH Ago1 dataset and evaluates their performance in the state-of-the-art context.

### 4.1.1 Hyperparameter tuning results

As described in section 3.2, hyperparameter tuning was performed for models trained on all three datasets with ratios 1:1, 1:10 and 1:100. The best hyperparameters for all three models are summarised in Table 4.1. To pick one best model for the whole CLASH Ago1 dataset, all three tuned models were compared on all three evaluation datasets (Table 4.2). The best performing model on all three datasets was the one trained on the dataset with 1:10 ratio, which was named miRBind. The rest of this thesis will use for all other comparisons only this one miRBind model.

**Table 4.1:** Summary of the best hyperparameters for models trained on datasets with 1:1, 1:10 and 1:100 ratios

| Hyperparameter | Dataset ratio | | |
|---|---|---|---|
| | 1:1 | 1:10 | 1:100 |
| Number of convolutional layer blocks | 6 | 6 | 6 |
| Number of dense layer blocks | 3 | 2 | 3 |
| Convolutional layer kernel size | 4 | 5 | 5 |
| Pool size of the pooling layer | 2 | 2 | 2 |
| Dropout rate | 0.2 | 0.3 | 0.1 |
| Learning rate | 0.01 | 0.00152 | 0.00027 |

### 4.1.2 Comparison with state-of-the-art

miRBind method was compared with three other state-of-the-art methods on all three left out test sets (Figure 4.1). The seed approach recognizes a perfectly complementary match of the 2–7 miRNA hexamer on

**Table 4.2:** AU PRC for three models trained on train sets with different pos : neg ratios evaluated on three evaluation sets

| Evaluation set | Model trained on | | |
|---|---|---|---|
| | 1:1 | 1:10 | 1:100 |
| 1:1 | 0.9660 | 0.9670 | 0.9645 |
| 1:10 | 0.7986 | 0.8140 | 0.8001 |
| 1:100 | 0.4211 | 0.4629 | 0.4512 |

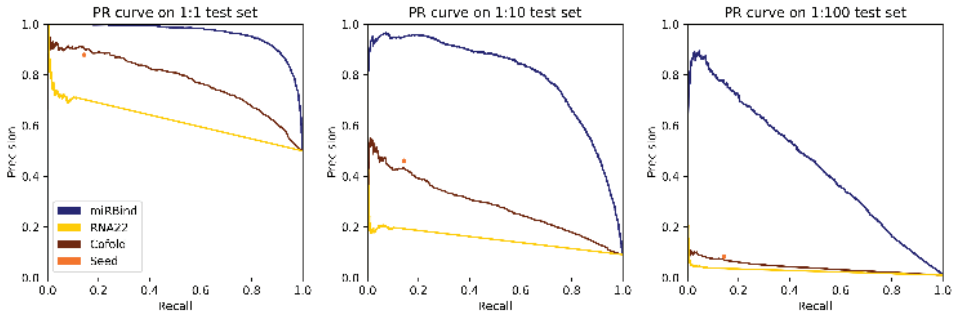


**Figure 4.1:** PR curve for miRBind, RNA22, cofold and seed tested on 1:1, 1:10, and 1:100 test sets

the target sequence. Since the match is a binary decision, no AU PRC may be calculated. The cofold method evaluates potential binding pairs based on the free energy of folding the sequences. Seed and cofold represent widely used approaches to quickly identify potential miRNA (or small RNA) targets and are commonly plugged in to more complex target prediction programs. The third method RNA22 represent a lightweight class of target prediction programs but in comparison to the standard target prediction program, it does not predict only functional targets but all putative miRNA binding sites.

In the 1:1 dataset, miRBind outperforms both cofold and RNA22 with the AU PRC of 0.9634 versus 0.7784 for cofold and 0.6203 for RNA22. The difference is even more highlighted in the more realistic 1:10 and 1:100 datasets. The seed method performs similarly to cofold, it is very precise (precision 0.8796, recall 0.1425) in the balanced task,

which made it promising as the first step in target prediction programs as well as for assigning miRNAs to CLIP-Seq peaks. However, when going to the imbalanced datasets, the precision drops again very quickly. Assigning targets to miRNAs using the standard cofold or seed method is thus unreliable in realistic scenarios. In contrast, miRBind shows an almost perfect precision up to 50 % recall in the 1:1 dataset and is more robust in comparison to other methods in the 1:10 and 1:100 datasets.

### 4.1.3 Comparison with different architectures

Two other architectures used for training on CLASH Ago1 dataset were presented in this thesis (see section 3.2). Figure 4.2 shows their comparison with our miRBind method on the test sets. The ResNet method with label smoothing developed by Ján Krčmář outperforms the miRBind model on all sets. While the difference on the 1:1 set is minimal (0.9634 for miRBind versus 0.9689 for ResNet), the biggest difference can be observed on the hardest 1:100 set (0.4464 for miRBind versus 0.5372 for ResNet). DNABERT performs the worse of all models. As DNABERT was pretrained on single DNA sequences and the task we are dealing with is the pairing of two RNA sequences, this shift might be too hard for the model to perform well. Also, the input to the DNABERT model is different than for the rest of the models. Revealing Watson-Crick base pairing from the one hot encoded sequences on the input might be too hard for the model but it is crucial for the miRNA – target binding.

### 4.1.4 Usage

The expected target group of users are biologists or bioinformaticians interested in miRNA binding, for example, to allocate miRNAs to CLIP-Seq peaks. To run a large number of predictions or to plug miRBind into a custom pipeline, a standalone python script is provided. For the less experienced group of users, web server may be a user friendly and easy way to use miRBind method for custom miRNA target binding prediction (Figure 4.3).

All data concerning the miRBind project are available on miRBind github.

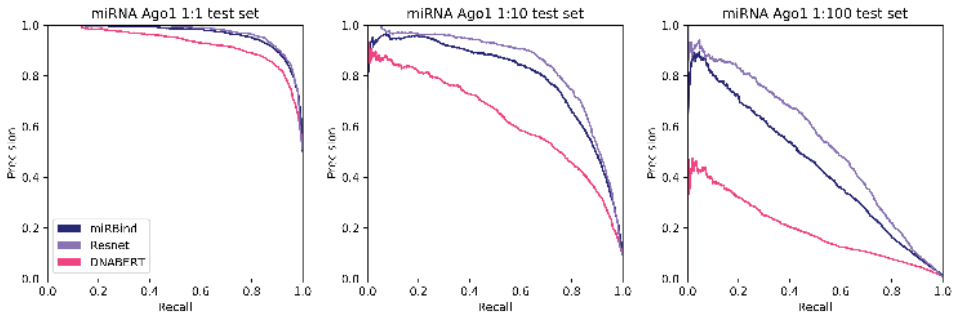


**Figure 4.2:** PR curve for miRBind, fine-tuned DNABERT and Jan's ResNet model tested on 1:1, 1:10, and 1:100 test sets

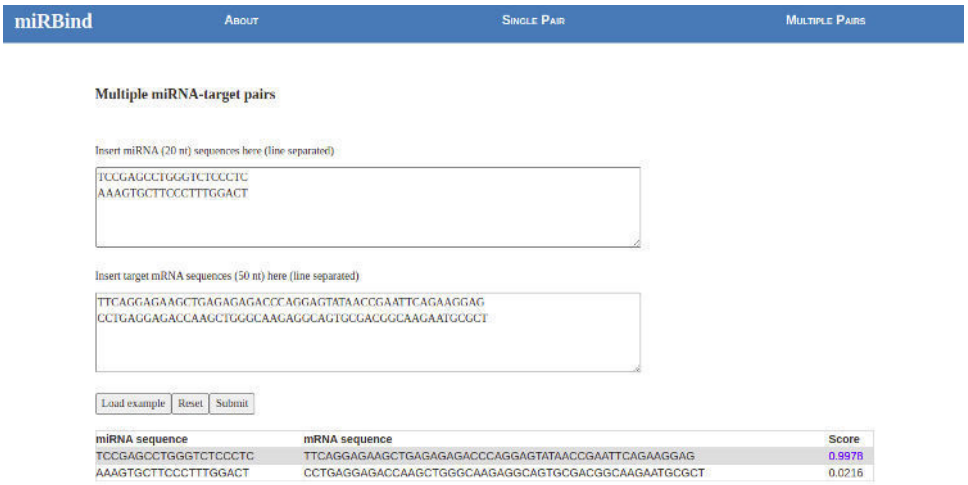


**Figure 4.3:** Screenshot from the miRBind web page

## 4.2 CLASH Ago2 dataset

This section provides a summary of all four models trained on CLASH Ago2 datasets (see section 3.1.2) and compares their performance with the state-of-the-art as well as individual models with each other on the left-out test sets.

### 4.2.1 Comparison with state-of-the-art

Both models trained on miRNA dataset with database (named miRNA model) or truly observed miRNA sequences (named miRNA real sequence model) were compared with RNA22, cofold and seed methods like in the CLASH Ago1 section 4.1.2. For comparison with models trained on tRNA and YRNA dataset, only cofold and seed were used. As RNA22 is a tool developed and tested only on miRNAs, we omitted it from this comparison. Specific tools for the prediction of non-miRNA target binding have not been developed yet, but there exist a couple of target prediction programs for tRNA drivers which use in their first filtering step seed or cofold.

Both miRNA models (Figure 4.4a and 4.4b) show the same trends as miRBind. The models outperform state-of-the-art and the differences are more pronounced in the datasets with higher negative ratios.

The tRNA model again outperforms both cofold and seed methods (Figure 4.4c) across all datasets. In the balanced 1:1 dataset, the tRNA model outperforms cofold with AU PRC 0.8387 against 0.6990. When adding more negatives, the difference deepens. Seed keeps its high precision (0.8502) in the balanced dataset but drops with recall (0.1473) in comparison to the CLASH Ago2 miRNA datasets (around 0.24). This drop might be caused by tRNAs behaving differently than miRNA drivers when targeting the mRNA as the seed region might not be the most important binding part.

The YRNA model performs very poorly as well as the other methods (Figure 4.4d). The problem for the YRNA model might be in the size of the training set, as it is relatively small – it contains only less than a thousand positive samples. The model thus probably was not able to properly train. Yet another problem may be in the dataset itself as some parts of the data can be artefacts from the experiment preparation.

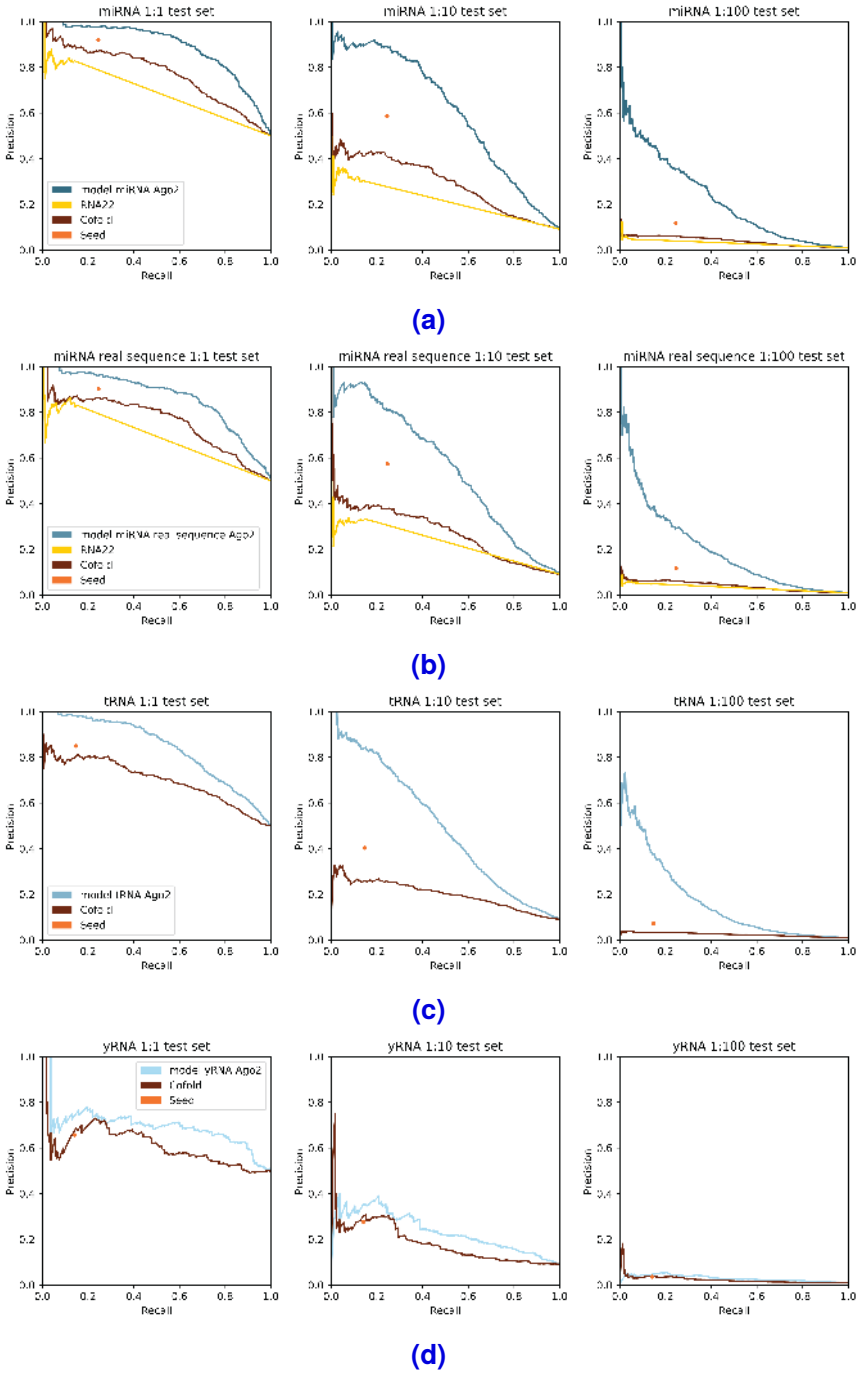


**Figure 4.4:** PR curves for methods evaluated on individual CLASH Ago2 test datasets with different positive : negative ratios

### 4.2.2 Cross comparison between CLASH Ago2 datasets

Two datasets were created for CLASH Ago2 miRNAs – dataset with database miRNA sequences and dataset with experimentally observed miRNAs. Based on these two datasets, two different models were trained. Next, the trained models were evaluated on both database-based and real miRNA test datasets. On the miRNA database test set both models performed very similarly across all ratios (Table A.2) but on the miRNA real sequence dataset the model trained on this dataset performed a bit better than the model trained on database miRNA sequences (Table A.3). The results show, that there is enough information also in the truly observed miRNA sequences even though the sequences may be noisier (couple of nucleotides shorter or contain mismatches). This supported the idea that training on observed tRNA fragments may bring reasonable results. The possible reason why the miRNA real sequence model outperforms the database sequence model on miRNA real sequence dataset but performs just as well as the database sequence model on the miRNA database sequence dataset is that the real sequence model orients better in the noisier dataset and when evaluated on the clean dataset it can spot the same patterns. However, the database model might not be able to deal with the noisier sequences. Another possible explanation is that the real miRNA sequences contain some extra information which is lost when using database sequences.

An interesting fact is, that even the tRNA model performs pretty well on both miRNA datasets and similarly, both miRNA models perform well on tRNA datasets and outperform cofold and seed (see supplement A). These results support the idea of Ago loaded miRNAs and tRNAs binding to their targets in a similar way.

When evaluating the YRNA model on any other small RNA dataset, the model performs very poorly. On the other hand, all models evaluated on the YRNA datasets perform badly and close to random (see supplement A). This could mean that either YRNA targeting works in a completely different way than miRNA or tRNA targeting or that the dataset is very noisy, which is the more probable version.

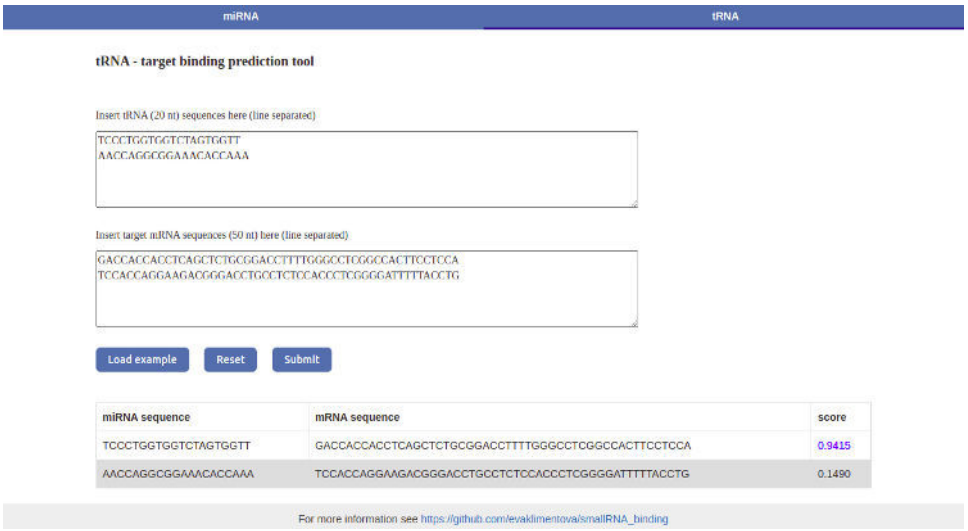


**Figure 4.5:** Screenshot from the web page for CLASH Ago2 miRNA and tRNA target predictions

### 4.2.3 Usage

For evaluation of new potential small RNA and target tuples, similar interfaces as for miRBind were created (see section 4.1.4). The standalone python script with the possibility to load different small RNA models serves for fast evaluation of bigger datasets. For simple requests, there is a webpage (Figure 4.5) with the possibility to evaluate miRNAs or tRNAs and their targets. The model evaluating miRNAs is the one trained with database sequences. Due to the YRNA model's poor performance, YRNAs are not included on the webpage.

All models and the web server from the small RNA CLASH Ago2 project can be found on github.

## 4.3 Comparison between CLASH Ago1 and Ago2

To prove that the trained models are not overfitted and able to generalize, CLASH Ago1 dataset was used for the evaluation of models trained on miRNA CLASH Ago2 and vice versa (Figures 4.6a and 4.6b). In both datasets, native models perform the best, however, the

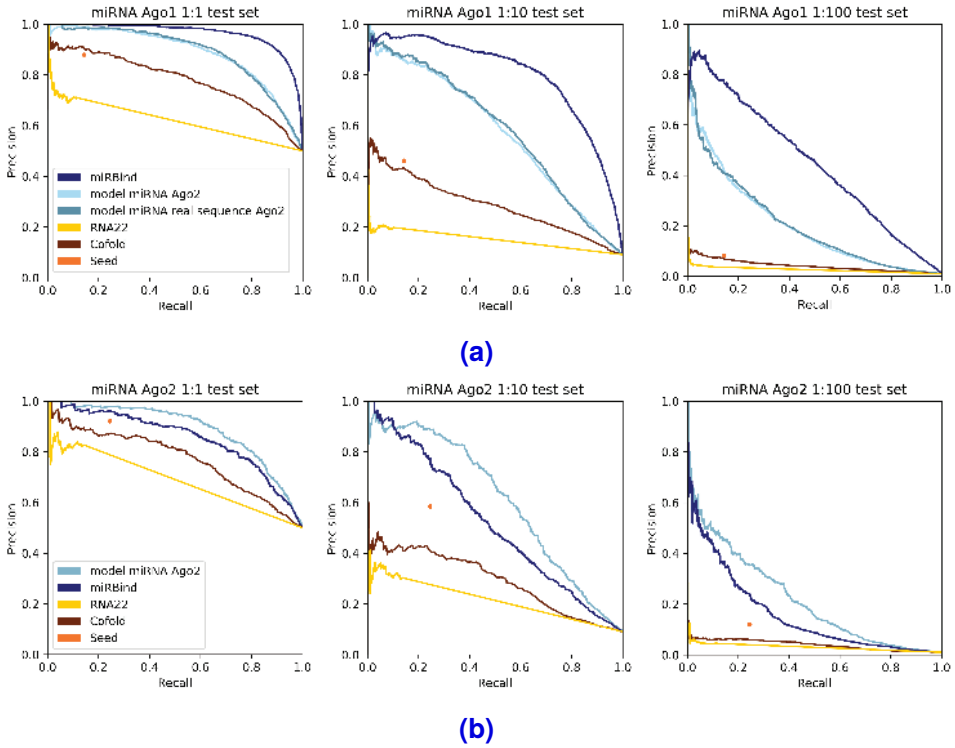


**Figure 4.6:** PR curves for methods evaluated on (a) CLASH Ago1 and (b) CLASH Ago2 test datasets

model trained on the other dataset still performs pretty well and outperforms other state-of-the-art methods. From the biological point of view, Ago1 and Ago2 work generally similarly, nonetheless Ago2 is known to have an extra "slicer" activity. The CLASH Ago1 and Ago2 datasets can thus have mostly similar but partially different binding rules. The good performance of models on different dataset confirms the common binding rules for both datasets and shows that the models are able to apply the learned rules even for a bit different dataset.

# 5 Conclusion

In this thesis, we presented multiple machine learning models that can be used to predict the pairing of Ago1 or Ago2 loaded miRNAs or tRFs to their mRNA targets. We show that our method performs better in comparison to older state-of-the-art methods such as seed and cofold. An important fact is that our method is trained on unbiased experimental data which are not overfilled with "canonical seed" targets and offers thus a more realistic view of the small RNA binding problem. Thanks to this, our method may be able to replace existing widely used tools and help with revealing non-canonical binding sites on top of an increasing number of experimentally identified ones. To open up the presented method to wider biological society, we prepared a python script for more complex requests supplemented by easy to use web application. The findings introduced in this thesis will be presumably published in two separate papers – one dealing with miRBind method trained on CLASH Ago1 dataset and one dedicated to the CLASH Ago2 project including the experimental part, bioinformatical processing and machine learning model.

We hope our method will be used for example as a way to allocate miRNAs to CLIP-Seq peaks instead of the currently used seed. Another very interesting application may be to plug our method to some target prediction program as the first filtering step to obtain small RNA binding sites which will be then filtered to get only functional targets.

Up to this date, there are only two unbiased high throughput Ago CLASH datasets, but another can be built from the new eCLIP method when the data are released. The already presented models can be evaluated on this dataset and even a new model may be trained.

There exist studies trying to sort different binding rules into multiple categories, for example, canonical seed binding, 3'-end binding or centred miRNA pairing. It would be interesting as a future plan to look into the machine learning model features and try to interpret what the model has learned and match it with the known binding categories.

# Bibliography


1. HOAGLAND, Mahlon B.; KELLER, Elizabeth B.; ZAMECNIK, Paul C. Enzymatic Carboxyl Activation of Amino Aicds. *The Journal of biological chemistry*. 1956, vol. 218, pp. 345–58. Available from DOI: 10.1016/S0021-9258(18)65898-3.
2. PALADE, George E. A small particulate component of the cytoplasm. *The Journal of biophysical and biochemical cytology*. 1955, vol. 1, pp. 59–68. Available from DOI: 10.1083/jcb.1.1.59.
3. STORZ, Gisela. An Expanding Universe of Noncoding RNAs. *Science*. 2002, vol. 296, pp. 1260–1263. Available from DOI: https://www.science.org/doi/10.1126/science.1072249.
4. BARTEL, David P. Metazoan MicroRNAs. *Cell*. 2018, vol. 173, pp. 20–51. Available from DOI: https://doi.org/10.1016/j.cell.2018.03.006.
5. KONDETIMMANAHALLI, Ragini; GHARPURE, Kshipra M.; WU, Sherry Y.; LOPEZ-BERESTEIN, Gabriel; SOOD, Anil K. Chapter 24 - Noncoding RNAs: Novel Targets in Anticancer Drug Development. In: CHAKRABARTI, Dr.Jayprokas; MITRA, Dr. Sanga (eds.). *Cancer and Noncoding RNAs*. Boston: Academic Press, 2018, pp. 447–459. Translational Epigenetics. ISSN 25425358. Available from DOI: https://doi.org/10.1016/B978-0-12-811022-5.00024-3.
6. KRAHN, Natalie; FISCHER, Jonathan T.; SÖLL, Dieter. Naturally Occurring tRNAs With Non-canonical Structures. *Frontiers in Microbiology*. 2020, vol. 11. Available from DOI: 10.3389/fmicb.2020.596914.
7. HAHNE, Jens Claus; LAMPIS, Andrea; VALERI, Nicola. Vault RNAs: hidden gems in RNA and protein regulation. *Cellular and Molecular Life Sciences*. 2021, vol. 78, pp. 1487–1499. Available from DOI: https://doi.org/10.1007/s00018-020-03675-9.
8. GUGLAS, Kacper; KOŁODZIEJCZAK, Iga; KOLENDA, Tomasz; KOPCZYŃSKA, Magda; TERESIAK, Anna; SOBOCIŃSKA, Joanna; BLIŹNIAK, Renata; LAMPERSKA, Katarzyna. YRNAs and YRNA-Derived Fragments as New Players in Cancer Research and Their

Potential Role in Diagnostics. *International Journal of Molecular Sciences*. 2020, vol. 21. Available from DOI: 10.3390/ijms21165682.

9. LIU, Jidong; CARMELL, Michelle A.; RIVAS, Fabiola V.; MARSDEN, Carolyn G.; THOMSON, J. Michael; SONG, Ji-Joon; HAMMOND, Scott M.; JOSHUA-TOR, Leemor; HANNON, Gregory J. Argonaute2 Is the Catalytic Engine of Mammalian RNAi. *Science*. 2004, vol. 305, pp. 1437–1441. Available from DOI: 10.1126/science.1102513.

10. MORITA, Sumiyo; HORII, Takuro; KIMURA, Mika; GOTO, Yuji; OCHIYA, Takahiro; HATADA, Izuho. One Argonaute family member, Eif2c2 (Ago2), is essential for development and appears not to be involved in DNA methylation. *Genomics*. 2007, vol. 89, pp. 687–696. Available from DOI: https://doi.org/10.1016/j.ygeno.2007.01.004.

11. B., Alexander Maxwell; ANDO, Yoshinari; DE HOON, Michiel L.; TOMARU, Yasuhiro; SUZUKI, Harukazu; HAYASHIZAKI, Yoshihide; DAUB, Carsten O. Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin. *RNA Biology*. 2011, vol. 8, pp. 158–177. Available from DOI: 10.4161/rna.8.1.14300.

12. BARTEL, David P. MicroRNAs: Target Recognition and Regulatory Functions. *Cell*. 2009, vol. 136, pp. 215–233. Available from DOI: https://doi.org/10.1016/j.cell.2009.01.002.

13. LAL, Ashish; NAVARRO, Francisco; MAHER, Christopher A.; MALISZEWSKI, Laura E.; YAN, Nan; O'DAY, Elizabeth; CHOWDHURY, Dipanjan; DYKXHOORN, Derek M.; TSAI, Perry; HOFMANN, Oliver; BECKER, Kevin G.; GOROSPE, Myriam; HIDE, Winston; LIEBERMAN, Judy. miR-24 Inhibits Cell Proliferation by Targeting E2F2, MYC, and Other Cell-Cycle Genes via Binding to "Seedless" 3'UTR MicroRNA Recognition Elements. *Molecular Cell*. 2009, vol. 35, pp. 610–625. Available from DOI: https://doi.org/10.1016/j.molcel.2009.08.020.

14. LORENZ, Ronny; BERNHART, Stephan H.; HÖNER ZU SIEDERDISSEN, Christian; TAFER, Hakim; FLAMM, Christoph; STADLER, Peter F.; HOFACKER, Ivo L. ViennaRNA Package 2.0. *Algorithms*

*for Molecular Biology*. 2011, vol. 6. Available from DOI: https://doi.org/10.1186/1748-7188-6-26.

15. BAEK, Daehyun; VILLÉN, Judit; SHIN, Chanseok; CAMARGO, Fernando D.; GYGI, Steven P.; BARTEL, David P. The impact of microRNAs on protein output. *Nature*. 2008, vol. 455, pp. 64–71. Available from DOI: 10.1038/nature07242.

16. SELBACH, Matthias; SCHWANHÄUSSER, Björn; THIERFELDER, Nadine; FANG, Zhuo; KHANIN, Raya; RAJEWSKY, Nikolaus. Widespread changes in protein synthesis induced by microRNAs. *Nature*. 2008, vol. 455, pp. 58–63. Available from DOI: 10.1038/nature07228.

17. ALEXIOU, Panagiotis; MARAGKAKIS, Manolis; PAPADOPOULOS, Giorgos L.; RECZKO, Martin; HATZIGEORGIOU, Artemis G. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*. 2009, vol. 25, pp. 3049–3055. Available from DOI: 10.1093/bioinformatics/btp565.

18. KARAGKOUNI, Dimitra; PARASKEVOPOULOU, Maria D.; CHATZOPOULOS, Serafeim; VLACHOS, Ioannis S.; TASTSOGLOU, Spyros; KANELLOS, Ilias; PAPADIMITRIOU, Dimitris; KAVAKIOTIS, Ioannis; MANIOU, Sofia; SKOUFOS, Giorgos; VERGOULIS, Thanasis; DALAMAGAS, Theodore; HATZIGEORGIOU, Artemis G. DIANA-TarBase v8: A decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic acids research*. 2017, vol. 46, pp. 239/245. Available from DOI: 10.1093/nar/gkx1141.

19. RIOLO, Giulia; CANTARA, Silvia; MARZOCCHI, Carlotta; RICCI, Claudia. miRNA Targets: From Prediction Tools to Experimental Validation. *Methods and Protocols*. 2021, vol. 4. Available from DOI: 10.3390/mps4010001.

20. HELWAK, Aleksandra; KUDLA, Grzegorz; DUDNAKOVA, Tatiana; TOLLERVEY, David. Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. *Cell*. 2013, vol. 153, pp. 654–665. Available from DOI: https://doi.org/10.1016/j.cell.2013.03.043.

21. HAUSSECKER, Dirk; HUANG, Yong; LAU, Ashley; PARAMESWARAN, Poornima; MARK A. KAY, Andrew Z. Fire nad. Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA*. 2010, vol. 16, pp. 673–695. Available from DOI: 10.1261/rna.2000810.

22. KUSCU, Canan; KUMAR, Pankaj; KIRAN, Manjari; SU, Zhangli; MALIK, Asrar; DUTTA, Anindya. tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer independent manner. *RNA*. 2018, vol. 24, pp. 1093–1105. Available from DOI: 10.1261/rna.066126.118.

23. KUMAR, Pankaj; ANAYA, Jordan; MUDUNURI, Suresh B; DUTTA, Anindya. Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Biology*. 2014, vol. 12. Available from DOI: https://doi.org/10.1186/s12915-014-0078-0.

24. XIAO, Qiong; GAO, Peng; HUANG, Xuanzhang; CHEN, Xiaowan; CHEN, Quan; LV, Xinger; FU, Yu; SONG, Yongxi; WANG, Zhenning. tRFTars: predicting the targets of tRNA-derived fragments. *Journal of Translational Medicine*. 2021, vol. 19. Available from DOI: https://doi.org/10.1186/s12967-021-02731-7.

25. ZHOU, Yiran; PENG, Haoran; CUI, Qinghua; ZHOU, Yuan. tRFTar: Prediction of tRF-target gene interactions via systemic reanalysis of Argonaute CLIP-seq datasets. *Methods*. 2021, vol. 187, pp. 57–67. Available from DOI: https://doi.org/10.1016/j.ymeth.2020.10.006.

26. PARIKH, Rohan; WILSON, Briana; MARRAH, Laine; SU, Zhangli; SAHA, Shekhar; KUMAR, Pankaj; HUANG, Fenix; DUTTA, Anindya. tRForest: a novel random forest-based algorithm for tRNA-derived fragment target prediction. *bioRxiv*. 2021. Available from DOI: 10.1101/2021.12.13.472430.

27. GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

28. LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. *Nature*. 2015, vol. 521, pp. 436–444. Available from DOI: https://doi.org/10.1038/nature14539.

29. ZENG, Haoyang; EDWARDS, Matthew D.; LIU, Ge; GIFFORD, David K. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics*. 2016, vol. 32, pp. i121–i127. Available from DOI: https://doi.org/10.1093/bioinformatics/btw255.

30. ZOU, James; HUSS, Mikael; ABID, Abubakar; MOHAMMADI, Pejman; TORKAMANI, Ali; TELENTI, Amalio. A primer on deep learning in genomics. *Nature Genetics*. 2019, vol. 51, pp. 12–18. Available from DOI: https://doi.org/10.1038/s41588-018-0295-5.

31. PARK, Yongjin; KELLIS, Manolis. Deep learning for regulatory genomics. *Nature biotechnology*. 2015, vol. 33, pp. 825–826. Available from DOI: 10.1038/nbt.3313.

32. JHA, Anupama; GAZZARA, Matthew R; BARASH, Yoseph. Integrative deep models for alternative splicing. *Bioinformatics*. 2017, vol. 33, pp. i274–i282. Available from DOI: https://doi.org/10.1093/bioinformatics/btx268.

33. ANGERMUELLER, Christof; LEE, Heather J; REIK, Wolf; STEGLE, Oliver. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome biology*. 2017, vol. 18, pp. 1–13. Available from DOI: https://doi.org/10.1186/s13059-017-1189-z.

34. ALIPANAHI, Babak; DELONG, Andrew; WEIRAUCH, Matthew T; FREY, Brendan J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature biotechnology*. 2015, vol. 33, pp. 831–838. Available from DOI: https://doi.org/10.1038/nbt.3300.

35. JI, Yanrong; ZHOU, Zhihan; LIU, Han; DAVULURI, Ramana V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*. 2021, vol. 37, pp. 2112–2120. ISSN 1367-4803. Available from DOI: https://doi.org/10.1093/bioinformatics/btab083.

36. MANAKOV, Sergei A; SHISHKIN, Alexander A; YEE, Brian A; SHEN, Kylie A; COX, Diana C; PARK, Samuel S; FOSTER, Heather M; CHAPMAN, Karen B; YEO, Gene W; VAN NOSTRAND, Eric L. Scalable and deep profiling of mRNA targets for individual microRNAs with chimeric eCLIP. *bioRxiv*. 2022. Available from DOI: 10.1101/2022.02.13.480296.

37. TRAVIS, Anthony J.; MOODY, Jonathan; HELWAK, Aleksandra; TOLLERVEY, David; KUDLA, Grzegorz. Hyb: A bioinformatics pipeline for the analysis of CLASH (crosslinking, ligation and sequencing of hybrids) data. *Methods*. 2014, vol. 65, pp. 263–273. Available from DOI: https://doi.org/10.1016/j.ymeth.2013.10.015.

38. QUINLAN, Aaron R.; HALL, Ira M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010, vol. 26, pp. 841–842. Available from DOI: 10.1093/bioinformatics/btq033.

39. KOZOMARA, Ana; BIRGAOANU, Maria; GRIFFITHS-JONES, Sam. miRBase: from microRNA sequences to function. *Nucleic Acids Research*. 2018, vol. 47, pp. D155–D162. Available from DOI: https://doi.org/10.1093/nar/gky1141.

40. KLIMENTOVA, Eva; POLACEK, Jakub; SIMECEK, Petr; ALEXIOU, Panagiotis. PENGUINN: Precise Exploration of Nuclear G-Quadruplexes Using Interpretable Neural Networks. *Frontiers in Genetics*. 2020, vol. 11. Available from DOI: 10.3389/fgene.2020.568546.

41. KRČMÁŘ, Ján. *Deep Learning for small RNA mediated targeting*. 2022. MA thesis. Masaryk University.

42. HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

43. ZUKER, Michael; STIEGLER, Patrick. Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information. *Nucleic acids research*. 1981, vol. 9, pp. 133–48. Available from DOI: 10.1093/nar/9.1.133.

# A AU PRC for different methods evaluated across all test datasets

**Table A.1:** AU PRC for listed methods evaluated on miRNA CLASH Ago1 test datasets. For seed method, sensitivity and precision are given.

| | 1:1 test set | 1:10 test set | 1:100 test set |
|---|---|---|---|
| miRBind | 0.9634 | 0.7969 | 0.4464 |
| miRNA Ago2 | 0.8863 | 0.5737 | 0.2015 |
| miRNA real seq Ago2 | 0.8890 | 0.5831 | 0.2055 |
| Cofold | 0.7784 | 0.2842 | 0.0413 |
| RNA22 | 0.6203 | 0.1507 | 0.0265 |
| Seed | sens: 0.1425<br>prec: 0.8796 | sens: 0.1425<br>prec: 0.46117 | sens: 0.1425<br>prec: 0.0824 |

**Table A.2:** AU PRC for listed methods evaluated on miRNA CLASH Ago2 test datasets. For seed method, sensitivity and precision are given.

| | 1:1 test set | 1:10 test set | 1:100 test set |
|---|---|---|---|
| miRBind | 0.8535 | 0.5167 | 0.1564 |
| miRNA Ago2 | 0.8891 | 0.6058 | 0.2178 |
| miRNA real seq Ago2 | 0.8902 | 0.6169 | 0.2384 |
| tRNA Ago2 | 0.8338 | 0.4871 | 0.1496 |
| YRNA Ago2 | 0.6980 | 0.2088 | 0.0265 |
| Cofold | 0.7709 | 0.2896 | 0.0394 |
| RNA22 | 0.6884 | 0.2151 | 0.0311 |
| Seed | sens: 0.2448<br>prec: 0.9215 | sens: 0.2448<br>prec: 0.5847 | sens: 0.2448<br>prec: 0.1193 |

**Table A.3:** AU PRC for listed methods evaluated on miRNA real sequence CLASH Ago2 test datasets. For seed method, sensitivity and precision are given.

| | 1:1 test set | 1:10 test set | 1:100 test set |
|---|---|---|---|
| miRBind | 0.8128 | 0.4079 | 0.1122 |
| miRNA Ago2 | 0.8366 | 0.4906 | 0.1486 |
| miRNA real seq Ago2 | 0.8638 | 0.5564 | 0.1959 |
| tRNA Ago2 | 0.8069 | 0.4569 | 0.1292 |
| YRNA Ago2 | 0.6916 | 0.2114 | 0.0266 |
| Cofold | 0.7577 | 0.2737 | 0.0399 |
| RNA22 | 0.6897 | 0.2289 | 0.0334 |
| Seed | sens: 0.2462<br>prec: 0.9031 | sens: 0.2462<br>prec: 0.5747 | sens: 0.2462<br>prec: 0.1164 |

**Table A.4:** AU PRC for listed methods evaluated on tRNA CLASH Ago2 test datasets. For seed method, sensitivity and precision are given.

| | 1:1 test set | 1:10 test set | 1:100 test set |
|---|---|---|---|
| miRBind | 0.6912 | 0.2293 | 0.0339 |
| miRNA Ago2 | 0.7281 | 0.2895 | 0.0575 |
| miRNA real seq Ago2 | 0.7443 | 0.3025 | 0.0601 |
| tRNA Ago2 | 0.8387 | 0.4947 | 0.1643 |
| YRNA Ago2 | 0.6301 | 0.1472 | 0.0164 |
| Cofold | 0.6990 | 0.1983 | 0.0239 |
| Seed | sens: 0.1473<br>prec: 0.8502 | sens: 0.1473<br>prec: 0.4037 | sens: 0.1473<br>prec: 0.0719 |

**Table A.5:** AU PRC for listed methods evaluated on YRNA CLASH Ago2 test datasets. For seed method, sensitivity and precision are given.

| | 1:1 test set | 1:10 test set | 1:100 test set |
|---|---|---|---|
| miRBind | 0.6179 | 0.1850 | 0.0237 |
| miRNA Ago2 | 0.5845 | 0.1568 | 0.0198 |
| miRNA real seq Ago2 | 0.6451 | 0.1758 | 0.02311 |
| tRNA Ago2 | 0.6389 | 0.1879 | 0.0236 |
| YRNA Ago2 | 0.6934 | 0.2331 | 0.0292 |
| Cofold | 0.6084 | 0.1871 | 0.0230 |
| Seed | sens: 0.1411<br>prec: 0.6571 | sens: 0.1411<br>prec: 0.2771 | sens: 0.1411<br>prec: 0.0362 |