miRBench: A Comprehensive microRNA Binding Site Prediction Training and Benchmarking Dataset

Stephanie Sammut[*,1,2], Katarina Gresova*[*,1,2,3], Dimosthenis Tzimotoudis[1,2], Eva Marsalkova[3,4] , David Cechak[3,4] , Panagiotis Alexiou [§,1,2]*

[1] Centre for Molecular Medicine and Biobanking, University of Malta, Msida, Malta
[2] Department of Applied Blomedical Science, Faculty of Health Sciences, University of Malta, Msida, Malta
[3] National Centre for Biomolecular Research, Faculty of Science, Masaryk University, 61137 Brno, Czech Republic
[4] Central European Institute of Technology, Masaryk University, Brno, Czech Republic
[*] these authors have contributed equally to this work
[§] to whom correspondence should be addressed

Abstract

Benchmarking drives innovation by providing a standardised framework that challenges existing methods and encourages the development of more advanced and effective solutions. Having well-curated benchmarking datasets has allowed machine learning experts to enter and revolutionise various fields of bioinformatics, such as protein structure prediction. Here, we present a set of benchmarking datasets for the yet unsolved task of accurate microRNA (miRNA) binding site prediction.

miRNAs are small, non-coding RNAs that regulate gene expression post-transcriptionally. miRNAs are loaded onto Argonaute family proteins (AGO) and serve as guides to target specific binding sites on messenger RNA through partial sequence complementarity, the exact rules of which remain to be fully understood. The precise identification of miRNA binding sites is crucial for the further understanding of these master regulators of gene expression.

Despite decades of attempts to accurately predict miRNA binding sites, the field has been unable to conclusively determine the exact rules of miRNA binding. An important issue is the relative paucity of unbiased experimental datasets that could be used to train more accurate machine learning models. In recent years, novel experimental approaches have revolutionised the field, producing thousands of unbiased miRNA-binding site sequence pairs. Recently, over one million such pairs were produced and made publicly available, posing the potential to revolutionise the field.

In this manuscript, we process and produce several comprehensive, well-annotated, and easy-to-use benchmark datasets from these experimental approaches. We use these datasets to benchmark state-of-the-art methods. We make the benchmarking datasets easy to access programmatically with a few lines of Python code, aiming to stimulate further machine learning research in the field by lowering the barrier to entry for machine learning experts into the field of miRNA binding site prediction. All code, datasets, and prediction tools are accessible at https://github.com/katarinagresova/miRBench/releases/tag/v1.0.0.

Introduction

Benchmarking is a systematic process that involves evaluating and comparing the performance of various methods or tools against a standard set of criteria or datasets. By establishing a common framework for comparison, benchmarking allows researchers to identify the strengths and weaknesses of different approaches, highlight areas for improvement, and drive the development of more advanced and reliable methods. In the field of machine learning, and increasingly bioinformatics, benchmarks have become strong drivers for advancement grounded in rigorous, reproducible comparisons. For example, in computer vision, benchmarks like ImageNet have been instrumental in driving progress by providing a large-scale, annotated dataset that serves as a common reference point for model evaluation (Deng et al. 2009). Similarly, the Critical Assessment of protein Structure Prediction (CASP) has provided a standardised benchmark for evaluating protein structure prediction methods, leading to significant improvements in the accuracy of computational models, including the development of AlphaFold, which achieved near-experimental accuracy (Jumper et al. 2021).

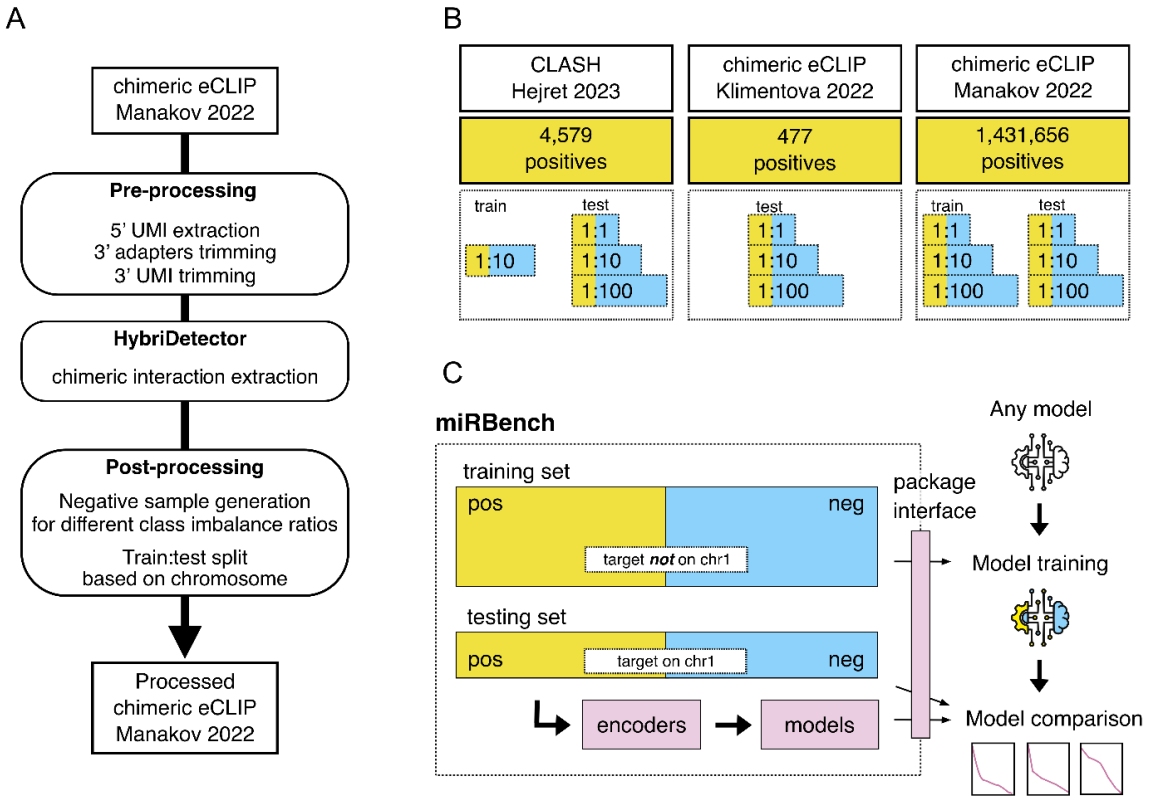
We have previously developed a collection of general Genomic Benchmarks with the objective of making training machine learning models on various biological tasks straightforward for non-biologists (Grešová et al. 2023). Following our initiative, several other general benchmarking studies were produced, such as GUANinE (Robson and Ioannidis 2024), BEND (Marin et al. 2023), BEACON (Ren et al. 2024), and GenBench (Z. Liu et al. 2024). Producing a high quality benchmark needs deep domain knowledge so as to avoid pitfalls and biases; adherence to the findability, accessibility, interoperability, and reusability (FAIR) principles (Wilkinson et al. 2016); as well as to the Data, Optimization, Model, and Evaluation (DOME) principles recently proposed for Machine Learning evaluation in Bioinformatics (Walsh et al. 2021). Following these directions, we have decided to produce a comprehensive benchmarking dataset for the pressing biological question of microRNA (miRNA) binding site prediction, a task that has stumped researchers for the better part of the past three decades.

Over thirty years ago, the regulatory map of the animal cell was fundamentally changed when some previously considered non-functional fragments of cellular RNA were identified as important regulators of the post-transcriptional life of RNA (Lee, Feinbaum, and Ambros 1993). These small regulatory RNAs, termed miRNAs, were quickly shown to play a wide variety of roles, including as master regulators of the degradation of aberrant mRNAs during embryogenesis and homeostasis. For instance, during embryogenesis, specific miRNAs orchestrate the temporal and spatial expression of genes necessary for tissue formation and organ development (Bernstein et al. 2003; Zhao, Samal, and Srivastava 2005; Ivey and Srivastava 2010). In adults, miRNAs maintain cellular equilibrium and adapt to environmental changes (Ambros 2004). Dysregulation of miRNAs is associated with various pathologies, including cancer (L. He and Hannon 2004; Calin and Croce 2006; Esquela-Kerscher and Slack 2006), cardiovascular diseases (van Rooij et al. 2006; Ikeda et al. 2007; Thum and Condorelli 2015), neurological disorders (Hébert and De Strooper 2009), and immune-related conditions (O'Connell et al. 2007; Sonkoly and Pivarcsi 2009; Dai and Ahmed 2011) among many others. Beyond their physiological function, miRNAs have shown great promise as biomarkers (Condrat et al. 2020), therapeutics (van Rooij and Olson 2012; Rupaimoole and Slack 2017; Chakraborty et al. 2017), and even combined with novel therapeutic methods such as CAR-T therapy (Rad et al. 2022; Shen et al. 2024). Fully

understanding the rules by which miRNAs bind their target binding sites is a crucial step in these applications.

miRNAs are initially transcribed as long transcripts (primary microRNAs) that undergo a series of enzymatic processes to produce the mature ~23 nucleotide long sequences termed mature miRNAs. These mature miRNAs are then loaded onto proteins of the Argonaute (AGO) family, which are fundamental components of the RNA-induced silencing complex (RISC). In mammals, the four AGO proteins (AGO1-4) specifically interact with miRNAs to form ribonucleoprotein complexes central to RNA silencing. Among them, AGO2 is particularly critical; unlike AGO1, AGO3, and AGO4, its loss leads to embryonic lethality in mice, underscoring its essential role (J. Liu et al. 2004; Morita et al. 2007). After AGO2 is loaded with a miRNA 'guide' (AGO2:miRNA), it can use partial complementarity between the 'guide' to identify 'target' sequences on other RNA molecules (Bartel 2004). We define the sites on these target RNAs where the miRNA binds as 'miRNA binding sites'. Unlike plant miRNAs that have fully complementary binding sites, in mammals most known AGO2:miRNA binding sites show partial complementarity, often focused on a 'seed' region located at the 5' end of the 'guide' sequence. A 'canonical seed' sequence denotes a fully Watson–Crick complementary stretch of at least six nucleotides starting at the second position from the 5' end of the miRNA 'guide'. Further binding outside the seed area can stabilise the interaction and is commonly called 3' complementary binding. However, functional interactions not mediated by a 'canonical seed' have been known since the early days of miRNA targeting research in worms and mammals (Didiano and Hobert 2006; Broughton et al. 2016). We, and others, have previously shown that less than 50%, of experimentally identified miRNA binding sites, are mediated by a canonical seed (Helwak et al. 2013; Klimentová et al. 2022; Hejret et al. 2023).

As the exact rules of AGO2:miRNA target recognition remain largely unknown, several approximations have been produced to date as crucial parts of miRNA target gene prediction programs (Alexiou et al. 2009; Grešová, Alexiou, and Giassa 2022). Most such methods use a two-step approach. In the first step, putative binding sites are identified using either a seed-based or a 'cofolding'-based approach. The seed-based approach tries to identify canonical or almost canonical seed sequences and weigh them based on categories of binding. Moreover, these seed-based rules of early miRNA target prediction programs were only based on a handful of known and experimentally validated interactions. The cofolding approach uses methods that calculate the minimum free energy of the structure produced from dimerization of the miRNA sequence and the putative target sequence, using this as a prioritisation and weighing technique. As a second step, various features of the identified putative binding sites are combined into an overall prediction of the probability of a specific AGO2:miRNA complex to repress a target RNA as a whole.



D

```
> from miRBench.dataset import get_dataset_df
> from miRBench.encoder import get_encoder
> from miRBench.predictor import get_predictor

> df = get_dataset_df("AGO2_CLASH_Hejret2023", split="test", ratio="1")
> tool = "miRBind_Klimentova2023"
> encoder = get_encoder(tool)
> predictor = get_predictor(tool)
> predictions = predictor(encoder(df))
```

Figure 1: (A) Processing workflow of the Manakov2022 chimeric eCLIP dataset. The dataset is processed using HybriDetector to get miRNA:binding site pairs as positive examples. Negative examples are generated from mismatching miRNA:binding site pairs from the positive examples. (B) The miRBench package contains three collections of datasets of various sizes, each with multiple positive to negative class imbalance ratios. The raw Manakov2022 dataset was first processed in this work, whereas the processed Hejret2023 and Klimentova2022 datasets were obtained from earlier publications. All datasets were processed by HybriDetector. (C) Datasets are divided into training and testing sets. Examples with the binding site sequence located on chromosome 1 are assigned to the testing set. The remaining examples are allocated to the training set. Various models are evaluated on the testing datasets. All datasets are made publicly available and easily accessible by the miRBench Python package. The miRBench package can be easily used to train and benchmark user models against the current SotA. (D) Python code snippet for loading a dataset, encoder, and predictor and producing predictions with miRBench.

In 2013, a new experimental method, termed CLASH (Crosslinking, Ligation, and Sequencing of Hybrids), was developed (Helwak et al. 2013). CLASH utilises a ligation step between the 'guide' and 'target' sequences, producing 'chimeric' reads that contain both sequences of the AGO-bound guide:target interaction. This technique was first applied to the AGO1 protein, revealing that only 40% of the identified miRNA binding sites contained a canonical seed in human cells (Helwak et al. 2013). More recently, in 2022, a newer high-throughput technique, termed chimeric eCLIP (chimeric enhanced Crosslinking and Immunoprecipitation) has been developed (Manakov et al. 2022). The method integrates a chimeric ligation step into the eCLIP method, a highly reproducible, robust method for identifying RBP (RNA-binding protein) binding sites, at single-nucleotide resolution, and an improved library generation efficiency relative to earlier CLIP methods (Van Nostrand et al. 2016). Chimeric eCLIP is therefore also able to capture AGO-bound chimeric reads of guide and target RNA sequences, at a higher resolution and sensitivity, providing a comprehensive map of miRNA binding sites. An important distinction between CLASH and chimeric eCLIP is the fact that CLASH uses induced AGO protein, while chimeric eCLIP exploits the endogenous AGO protein of the cells. We have also previously suggested (Hejret et al. 2023) that CLASH may be retrieving more interactions of lower affinity due to the higher abundance of AGO in CLASH compared to chimeric eCLIP.

In this study, we analyse three data sources produced by AGO2 CLASH or chimeric eCLIP methods (Klimentová et al. 2022; Manakov et al. 2022; Hejret et al. 2023). We extract chimeric interactions between human AGO2:miRNAs and their binding sites, create negative examples at three class imbalance levels, split the data to training and testing (Fig 1B), and evaluate several existing miRNA binding site prediction models on the testing datasets (Fig 1C). We have benchmarked eight recently published models for miRNA binding site prediction, all of which are able to score miRNA:binding site pairs based on sequence information alone. We have also made the training datasets easily accessible programmatically to stimulate further development. All presented datasets and benchmarked models are easily accessible with a few lines of Python code using the miRBench Python package (Fig 1D).

Results

Novel miRNA:binding-site datasets curated

We present here three curated collections of datasets (Table 1). The first dataset, Hejret2023, is based on data produced using the CLASH technique on AGO2, containing 4,579 positive examples (Hejret et al. 2023). The second dataset, Klimentova2022, was generated using the chimeric eCLIP method and includes 477 positive examples (Klimentová et al. 2022). The third dataset, Manakov2022, employs the chimeric eCLIP technique and comprises 1,431,656 positive examples (Manakov et al. 2022), approximately 170 times the number of positive examples of the other two datasets put together. Hejret2023 and Klimentova2022 were downloaded from the repositories of the corresponding papers. Manakov2022 was processed from raw reads using the pipeline originally employed in the other two dataset studies for consistency.

Dataset Name	Publication Year	Publication DOI	Experiment Type	N positive examples
Hejret2023	2023	10.1038/s41598-023-49757-z	CLASH	4,579
Klimentova2022	2022	10.3390/genes13122323	Chimeric eCLIP	477
Manakov2022	2022 (preprint)	10.1101/2022.02.13.480296	Chimeric eCLIP	1,431,656

Table 1: Summary statistics of miRNA:binding-site benchmark datasets.

Adhering to FAIR principles, and in order to facilitate the availability and interoperability of the novel datasets presented here, we have developed a Python package called miRBench that allows easy access to all produced datasets. The latest version of the package is freely available on GitHub (https://github.com/katarinagresova/miRBench/releases/tag/v1.0.0) and has a permanent version deposited at Zenodo ((https://doi.org/10.5281/zenodo.13934721). The Python package is also distributed through the Python Package Index (PyPI) (https://pypi.org/project/miRBench/1.0.0/) so that it can be installed using a single-line command, and can provide access to the datasets using a single line of Python code (Fig 1D). Additionally, the package provides access to all models benchmarked in this work, described in the Section 'Description of Benchmarked Models', under 'Materials & Methods', as well as access to methods for preprocessing input data so that it is compatible with each of the models. This thus ensures reproducibility of the results, and allows for future use of the models for benchmarking in novel datasets as they become available.

Binding sites in the 3'UTR are most prevalent but not an absolute majority

As part of the curation process of the three collections of datasets described, the target binding site sequences were aligned to the human genome via HybriDetector (Hejret et al. 2023), enabling assignment of genic feature annotations (Fig 2A). The Klimentova2022 chimeric eCLIP dataset appears to have a prevalence of binding sites mapping to regions annotated as transcript introns. However, the larger datasets (Hejret2023 and Manakov2022) show consistent prevalence of 3'UTR binding sites at 37% and 41% of

annotated interactions, respectively. Hejret2023 shows binding within exons at 32% and introns at 19%, while Manakov2022 shows binding within exons at 19% and introns at 29%. We observe that, overall, chimeric eCLIP appears to have more intronic targets than CLASH, at least for our limited sample size. We also observe that the number of binding sites on 3'UTRs is at a maximum of only 41%, in the Manakov2022 dataset. Interestingly, most miRNA target prediction methods only take into consideration targets within the 3'UTR, rarely exons, and never introns, despite the binding site distribution seen here. This is a potential blind spot that ignores a large number, potentially a majority, of binding sites from consideration of their effect on the overall regulation of the target messenger RNA.

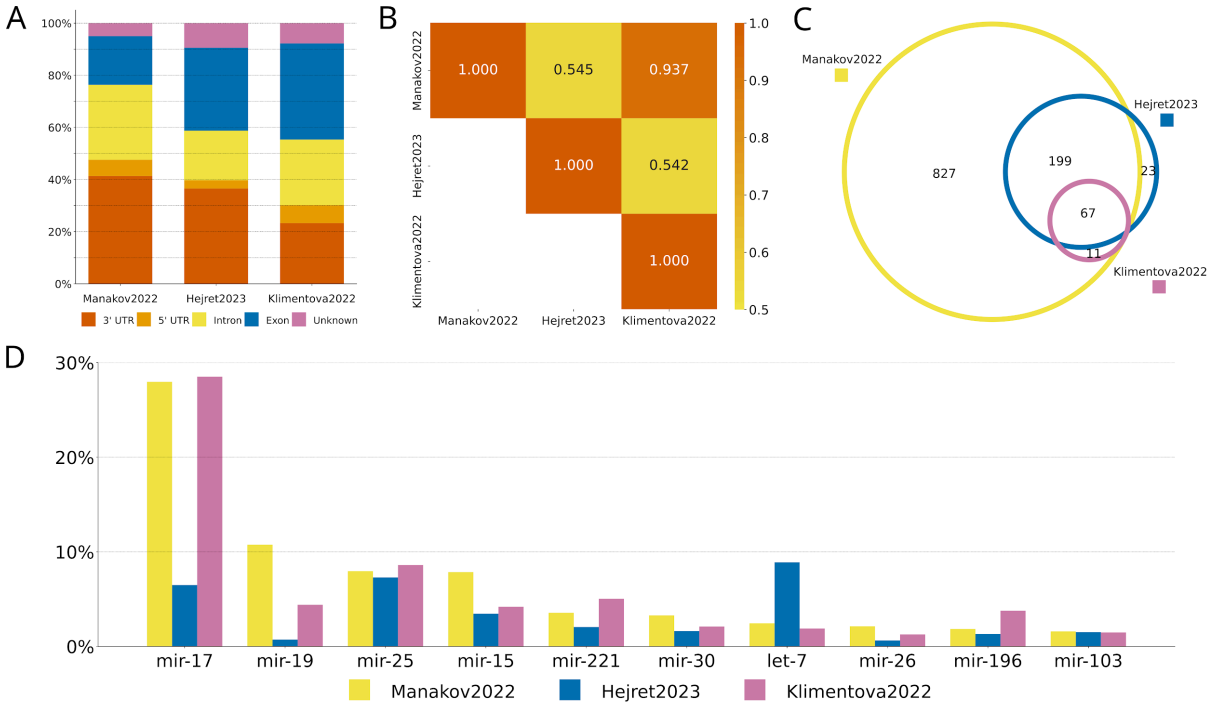


Figure 2: Composition of datasets. A) Distribution of binding site percentage overlap to genomic element annotations (5' UTR, 3' UTR, Intron, Exon, Unknown). B) Correlation heatmap of datasets based on similarity of miRNA families composition. C) Dataset overlap for miRNA families, including counts for unique families. D) Prevalence of the top 10 miRNA families by counts across the three datasets.

Detected microRNAs differ between experimental techniques

The three experimental datasets produced using the same cell line are expected to retrieve similar ratios of miRNA targets per miRNA family. We have correlated the fraction of chimeric reads per dataset for all detected microRNA families across samples (Fig 2B). We observe that Manakov2022 and Klimentova2022 display high correlation (R^2=0.937), while Hejret2023 differs substantially from both Manakov2022 (R^2=0.545) and Klimentova2022 (R^2=0.542). Additionally, we observe that the much larger Manakov2022 dataset identified chimeric reads for 827 miRNA families that none of the other methods identified (Fig 2C). Conversely, there were 23 miRNA families unique to Hejret2023, and none unique to Klimentova2022, while 67 are common among all three datasets. We find that the most prevalent miRNA family is mir-17 at 28.0% for Manakov2022 and 28.5% for Klimentova2022, and let-7 at 9.0% for Hejret2023. (Fig 2D). The discrepancy between the two methods could

be attributed to the fact that the CLASH method (Hejret2023) induces the expression of the AGO2 protein, while chimeric eCLIP (Manakov2022 and Klimentova2022) uses an antibody against endogenous AGO2. It is possible that induced AGO2 expression reduces competition between miRNA 'guides' and thus changes the composition of detected miRNAs in chimeras.

The majority of binding sites do not contain a canonical seed

Given the predominance of seed-like measures for miRNA binding site prioritisation, we explored the prevalence of canonical seed-like sequences in the produced datasets (Fig. 3A). Interactions lacking exact match for any of the defined seeds comprised approximately 22% of Manakov2022, and 28% for Hejret2023 and for Klimentova2022. Additionally, non-canonical seed interactions, that is, interactions that bear a 6mer seed match with a single bulge or mismatch, comprised 34% of Manakov2022, 42% of Hejret 2023, and 29% of Klimentova2022 datasets. The remaining interactions involved a canonical seed; Manakov2022 (44%), Hejret2023 (30%), and Klimentova2022 (43%). For all three datasets, the 7mer seed was the most prevalent type of canonical seed interaction, at 22%, 14%, and 25% of interactions in Manakov2022, Hejret2023, and Klimentova2022, respectively.

Exploring the distribution of the prevalence of canonical seed type interactions across miRNA families (Fig 3B), we find that the canonical seed type interactions in the Manakov2022 dataset have a range of 7.5% to 67.5% prevalence. That range shrinks for the Hejret2023 dataset at around 2.5% to 62.5%, and stretches for the Klimentova2022 dataset at 2.5% to 97.5%. Interestingly, both the Manakov2022 and Klimentova2022 datasets displayed high median prevalence of canonical seed type categories at 40% or more, in contrast to 33% in Hejret2023. It appears that the chimeric eCLIP methodology is more effective in capturing canonical seed-type interactions compared to the CLASH method, at least for the limited samples produced to date.

We examined the seed type distribution of the top 10 miRNA families that had the most identified targets. Pooling targets from all datasets (Fig 3C), 24% to 39% of interactions of the top miRNA families have a non-canonical seed. Generally, canonical Seed7mer is the most prevalent canonical seed type, ranging between 16% and 31% of total interactions, with some exceptions; the canonical Seed6mer is the most prevalent seed type for miR-221 (15%), and miR-26 (20%).

These findings taken together, show that no-seed, non-canonical, and canonical seed interactions of various types are found across different miRNAs and across experiments. The prevalence of different types of interactions could have a miRNA-specific component. This variation among different miRNA families currently remains underexplored in miRNA binding site prediction methods.

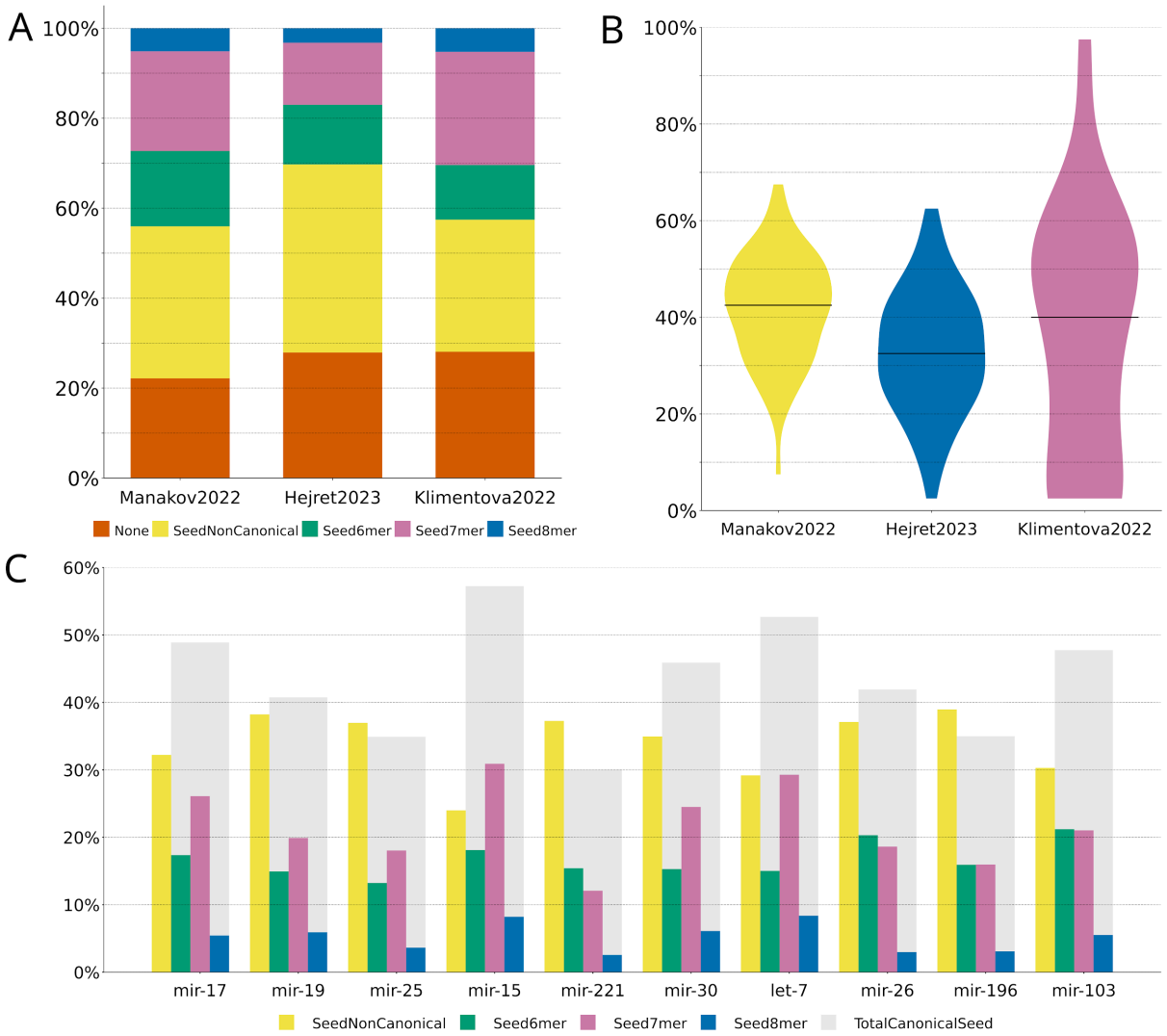


Figure 3: Seed type interactions analysis. A) Distribution of miRNA seed types across the three datasets. B) Distribution of miRNA canonical seed type prevalence in unique miRNA families across the datasets. C) Prevalence of miRNA seed types within the top 10 miRNA families by counts across the datasets.

Benchmarking of miRNA binding site prediction tools

We have identified 6 published machine learning models (further described in Section 'Description of Benchmarked Models', under 'Materials & Methods') that can be used to classify miRNA:target pairs as potential binding interactions. Additionally, we are benchmarking seed-based measures per seed category, as well as a binding affinity measure from RNACofold from the Vienna RNA Package (Lorenz et al. 2011). We have benchmarked each of these methods on the testing subsets of our datasets, for 1:1, 1:10, and 1:100 positive to negative example ratios. It is crucial to consider imbalanced datasets for the task of miRNA binding site prediction because for any given miRNA, only a few sites, on the entire transcriptome, serve as binding sites. The positive class should therefore be the minority class. This is represented in our data by considering larger ratios of negative examples, relative to the positive class. Evaluation of classification performance on highly imbalanced datasets is carried out using metrics that consider the class distribution, namely Precision and Recall. Accuracy is not a good metric here, as it does not distinguish between

the number of correctly classified examples of the different classes and could, therefore, be biassed towards one class. We have also calculated the Area Under the Precision Recall Curve (AU-PRC) for all methods on all datasets (Fig.4, Table 2). We notice that no single method consistently outperformed others across all datasets.

	Dataset								
	Hejret2023			Klimentova2022			Manakov2022		
Tool	**1:1**	**1:10**	**1:100**	**1:1**	**1:10**	**1:100**	**1:1**	**1:10**	**1:100**
TargetScanCnn_McGeary2019	0.74	0.33	0.09	0.78	0.41	0.13	0.79	0.46	0.18
CnnMirTarget_Zheng2020	0.63	0.15	0.02	0.63	0.15	0.02	0.91	0.53	0.11
TargetNet_Min2021	0.66	0.18	0.02	0.57	0.12	0.01	0.64	0.16	0.02
miRBind_Klimentova2022	0.87	0.58	0.22	0.84	0.47	0.15	0.89	0.57	0.19
miRNA_CNN_Hejret2023	0.89	0.61	0.22	0.80	0.39	0.11	0.90	0.62	0.26
InteractionAwareModel_Yang2024	0.77	0.31	0.05	0.70	0.23	0.04	0.65	0.21	0.04
RNACofold	0.77	0.29	0.04	0.68	0.19	0.02	0.62	0.14	0.02

Table 2: AU-PRC metrics. Highlighting the three top performing models per benchmarking dataset and ratio.

For Hejret2023 1:1, miRNA_CNN_Hejret2023 outperformed others at 0.89 AU-PRC, with miRBind_Klimentova2022 closeby at 0.87, and other methods such as TargetScanCnn_McGeary2019 (0.74 AU-PRC) and InteractionAwareModel_Yang2024 (0.77 AU-PRC) being very close to the baseline for RNACofold (0.77 AU-PRC). For Hejret2023 1:10 and 1:100, the same two models outperform others as well. It should be noted that the model miRNA_CNN_Hejret2023 was trained on the Hejret2023 1:10 dataset also made available in the miRBench Python Package.

For the Klimentova2022_1:1 dataset, the model miRBind_Klimentova2022 outperforms other models with AU-PRC 0.84, followed by miRNA_CNN_Hejret2023 at 0.80 and TargetScanCnn_McGeary2019 at 0.78. Despite the fact that the Klimentova2022 dataset was published in the same manuscript as the miRBind_Klimentova2022 model, the model was only evaluated on this dataset, and was not trained on it but on an older AGO1 CLASH dataset (Helwak et al. 2013). In the 1:10 and 1:100 datasets a similar pattern was observed with miRBind_Klimentova2022, miRNA_CNN_Hejret2023, and TargetScanCnn_McGeary2019 being the top performing models.

For the Manakov2022 1:1 dataset, CnnMirTarget_Zheng2020 has the best AU-PRC at 0.91, closely followed by miRNA_CNN_Hejret2023 at 0.90 and miRBind_Klimentova2022 at 0.89. TargetScanCnn_McGeary2019 (0.79 AU-PRC) is the only other method that shows a clear separation from the RNACofold baseline (0.62 AU-PRC). Interestingly, when we examine the Manakov2022 1:10 dataset, the top method of the 1:1 datasets loses performance faster than the following methods, bringing CnnMirTarget down to third place, below miRNA_CNN_Hejret2023 and miRBind_Klimentova2022.

Examining the largest benchmark Manakov2022_1:100, we see that the top three methods (miRNA_CNN_Hejret2023, miRBind_Klimentova2022, and TargetScanCnn_McGeary2019) outperform CnnMirTarget_Zheng2020. It is important to note that even the best performing method in this dataset (miRNA_CNN_Hejret2023) has a precision below 0.2 at the 0.5 recall mark (Fig 4l), meaning that retrieving half of all *bona fide* miRNA binding sites from this dataset would bring along four false positives for each true positive recovered. For the third method (TargetScan_CNN_McGeary2019), the precision at

0.5 recall is around 5%, meaning that 9 false positives would be retrieved for each true positive at that recall mark.

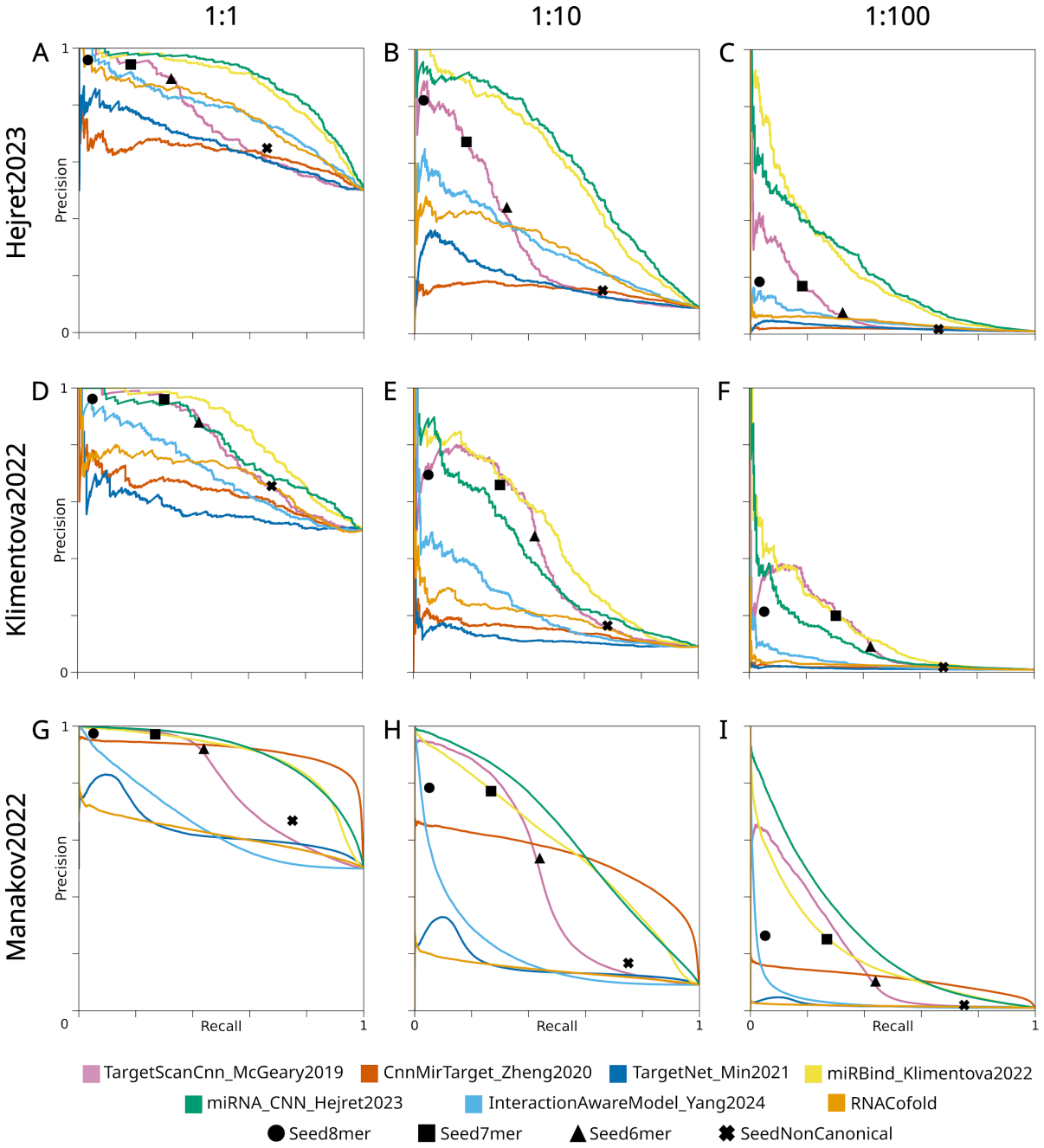


Figure 4: Precision-recall curves. Various miRNA binding site prediction tools evaluated on (A) Hejret2023 1:1 (B) Hejret2023 1:10 (C) Hejret2023 1:100 (D) Klimentova2022 1:1 (E) Klimentova2022 1:10 (F) Klimentova2022 1:100 (G) Manakov2022 1:1 (H) Manakov2022 1:10 (I) Manakov2022 1:100

Dataset Size improves CNN performance with diminishing results

The Manakov2022 has produced 1,262,123 positive examples in its 1:1 training set, in contrast to 3,860 positive examples in the previously largest 1:10 training dataset; Hejret2023. In light of this major increase in dataset size, we explore the potential improvement in CNN performance. We have decided to retrain the CNN model miRNA_CNN_Hejret2023 (Hejret et al. 2023) using various training set sizes, ranging from 200 examples, up to the full 2,524,246 examples in the Manakov2022 1:1 dataset. We subsampled the Manakov2022 1:1 training dataset and trained models based on the miRNA_CNN_Hejret2023 architecture and hyperparameters at various sample sizes. The performance of the retrained models was then evaluated on the Manakov2022 1:100 testing set.

We observe a performance improvement across the models trained on Manakov2022 subsets with increasing training set size (Fig 5). A model trained on the 200 examples shows similar performance to a completely random classifier, whereas the model trained on the subset of size 7,720 (the number of positive examples in this dataset correspond to the number of positives in Hejret2023 training set) already shows improved performance, with 0.43 AU-PRC. The model retrained on the entire Manakov2022 1:1 train set, with 2,524,246 examples, shows the best performance of 0.66 AU-PRC, closely followed by the model trained on the second largest set, with 200,000 examples, which displays very similar metrics of 0.62 AU-PRC (Fig. 5A). Considering the improvement in AU-PRC against training sample size (Fig. 5B), we observe diminishing improvement in AU-PRC as the training set becomes larger. At a recall level of 0.5, in which 50% of positive test examples are correctly retrieved, we see an increase in precision from 0.01 to 0.25 between 200 and 2,000 samples (+0.24), from 0.25 to 0.48 between 2,000 and 20,000 samples (+0.23), from 0.48 to 0.67 between 20,000 and 200,000 samples (+0.19), and finally from 0.67 to 0.71 from 200,000 to the entire training set ~2.5M samples (+0.04). (Fig. 5C).

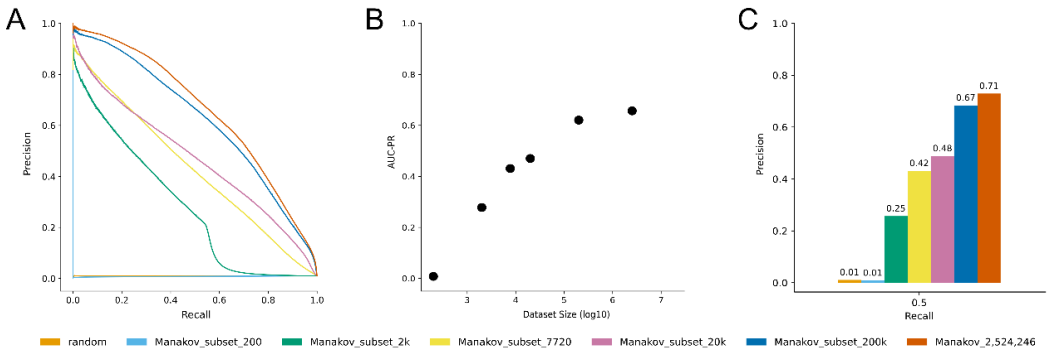


Figure 5: (A) Precision-recall curves for miRNA_CNN_Hejret2023 model retrained on various subsets of the Manakov2022 1:1 training dataset and evaluated on the Manakov2022 1:100 test dataset. The models' AU-PRC values considerably decrease with the size of the training subset:. (B) Scatter plot illustrating the relationship between training set size and AU-PRC for the miRNA_CNN_Hejret2023 model retrained on Manakov2022 1:1 train set subsets and tested on the Manakov2022 1:100 test set. (C) The precision of the retrained models at recall 0.5 when evaluated on the Manakov2022 1:100 test dataset.

Discussion and Conclusions

In this study, we have presented a set of miRNA:binding site training and benchmarking datasets called miRBench. We have benchmarked state-of-the-art methods at three levels of negative class imbalance. We have developed the miRBench Python package to facilitate use of the datasets by other researchers. Finally, we have showcased the potential of the new larger datasets in retraining existing models or training new models with novel machine learning architectures. Our primary goal was to lower the barrier for entry into this field and drive further innovation in miRNA binding site prediction, emphasizing progress and encouraging new developments.

The results from benchmarking eight commonly used miRNA binding site prediction tools across different datasets reveal a few key insights. First, no single tool consistently outperforms others across all datasets and conditions. This is particularly important when evaluating performance under different levels of class imbalance, where the positive class, representing bona fide miRNA binding sites, is often the minority class. The fact that tools perform variably depending on the dataset and imbalance levels indicates that there is no one-size-fits-all solution for miRNA binding site prediction. This highlights the need for continuous improvement and refinement in these prediction algorithms. Tools such as miRNA_CNN_Hejret2023 and miRBind_Klimentova2022 performed best on most benchmarks, particularly in imbalanced conditions. Best-performing models achieved precision below 0.2 at 0.5 recall in large datasets like Manakov2022, highlighting the need for advancements in prediction accuracy.

One important observation from our dataset analysis is the prevalence of non-canonical seed interactions. Across the three datasets, we found that less than 50% of binding sites contained canonical seed matches. Non-canonical interactions, including those with bulges or mismatches in the seed region, or those with no seed at all, were prevalent across miRNA families and experimental conditions. This finding suggests that many current prediction models, which primarily focus on canonical seed matches, may overlook a significant portion of true miRNA binding events. These results point to an under-explored area in miRNA binding site prediction, suggesting that modelling non-canonical interactions could improve model accuracy.

Large, curated training datasets are crucial for improving model performance. Retraining the miRNA_CNN_Hejret2023 model with the larger Manakov2022 1:1 dataset and its subsets led to significant improvements. However, these improvements followed a logarithmic scale, with diminishing returns as dataset size increased. Optimised models trained on large datasets can enhance predictive power and understanding of miRNA binding. While optimising model performance is beyond the scope of this study, demonstrating the potential of larger datasets highlights their role in improving model generalizability and robustness. Larger datasets also enable training of more complex architectures that can capture intricate miRNA binding patterns which smaller datasets may miss.

In conclusion, our study underscores the critical need for continued benchmarking and standardisation in the field of miRNA binding site prediction. By providing easily accessible and curated datasets through miRBench, we aim to establish a new standard for evaluating prediction tools, addressing discrepancies in data quality and inconsistency in evaluation metrics. While significant advancements have been made, current models still face challenges in terms of precision and recall, particularly in large, imbalanced datasets.

The prevalence of non-canonical interactions, the broader distribution of binding sites across different genomic regions, and the importance of large, unbiased datasets all represent areas that require further attention. We hope that miRBench will serve as a valuable resource for future research, fostering reproducibility, healthy competition, and the development of more accurate and generalizable miRNA binding site prediction tools, ultimately advancing our understanding of miRNA-mediated gene regulation.

Materials & Methods

Accessing Klimentova2022 and Hejret2023

The Klimentova2022 datasets are publicly available for download on GitHub at https://github.com/ML-Bioinfo-CEITEC/miRBind/tree/main/Datasets. Three test sets with positive to negative example ratios of 1:1, 1:10, and 1:100, are available as `AGO2_eCLIP_Klimentova_1_test.tsv`, `AGO2_eCLIP_Klimentova_10_test.tsv`, and `AGO2_eCLIP_Klimentova_100_test.tsv`, respectively. These datasets were downloaded and deposited on Zenodo, for easy accessing through the miRBench Python package. These datasets were produced using the chimeric read annotation pipeline, HybriDetector (Hejret et al. 2023), to separate different types of reads, and discern the miRNA and binding site sequences from chimeric reads. Since only a few hundreds of miRNA to binding site pairs were identified in the Klimentova2022 datasets, the 477 pairs were entirely assigned to a testing split and no training split was produced. Unlike other datasets presented here, this produced a testing split with binding site sequences mapping to any chromosome for the Klimentova2022 datasets. Following the split, negative examples were generated by matching binding site sequences to random miRNA sequences from the same experiment, excluding the miRNA:binding site pairs detected by HybriDetector that serve as positive examples ((Helwak et al. 2013; Klimentová et al. 2022; Hejret et al. 2023)).

Similarly, the Hejret2023 datasets are publicly available at https://github.com/ML-Bioinfo-CEITEC/HybriDetector/tree/main/ML/Datasets as `miRNA_test_set_1.tsv`, `miRNA_test_set_10.tsv`, and `miRNA_test_set_100.tsv`, for positive to negative example ratios 1:1, 1:10, and 1:100, respectively. A larger train set with a class imbalance ratio of 1:10, is also available as, `miRNA_train_set.tsv` (Hejret et al. 2023). Train-test split was carried out based on which chromosome the binding site sequence maps to. Unlike the Klimentova2022 datasets, binding site sequences on chromosome 1 were assigned to the test set, whereas those on any other chromosome were assigned to the training set. This produced a training-to-testing split ratio of 84/16 (3,860 / 719). As per the Klimentova2022 datasets, the negative examples were then created by matching randomly sampled miRNA sequences to randomly sampled binding site sequences, whilst ensuring that the match is not already present in the positive examples (Hejret et al. 2023). These datasets were also downloaded and deposited on Zenodo, to be easily accessed by the miRBench Python package.

Curating Manakov2022

Downloading the data

Publicly available data from high depth profiling of miRNA targets with the new chimeric eCLIP experimental method described by (Manakov et al. 2022), was downloaded from the Gene Expression Omnibus under SubSeries GEO ID GSE198250 (on 14 May 16:01:35 UTC 2024). The samples selected for download were to include only chimeric eCLIP data from the human cell line HEK293T. Only files for the R1 read sequences were used, yielding one file per sample to be processed. The files were selected using the GEOparse Python package (version 2.0.4) and downloaded using enaBrowserTools available at https://github.com/enasequence/enaBrowserTools (accessed in May 2024).

Pre-processing the data

The downloaded samples were pre-processed as per part of the pipeline for the analysis of total chimeric eCLIP datasets, described by (Manakov et al. 2022), available at https://github.com/YeoLab/chim-eCLIP. The pre-processing step extracts the 5' UMI to the read name, trims the 3' adapters, and trims 10 nucleotides from the 3' end to ensure no random sequences from the 3' UMI remain. Downstream processing of this pre-processed data is inspired by the data preparation methodology adopted in (Helwak et al. 2013; Klimentová et al. 2022; Hejret et al. 2023), in an attempt to replicate the data processing on the Hejret2023 and Klimentova2022 datasets (Fig. 1A).

Processing the data with HybriDetector

The sample files were then processed by the HybriDetector pipeline for detecting chimeric interactions, described in (Hejret et al. 2023), available at https://github.com/ML-Bioinfo-CEITEC/HybriDetector. The pipeline outputs high confidence pairs of miRNA sequence ('driver') to binding site sequence ('target'). Some modifications were applied to the HybriDetector pipeline for it to be able to process these larger files from (Manakov et al. 2022). The modifications are described further in Supplementary Methods and are available on the `fix_clustering` branch at https://github.com/ML-Bioinfo-CEITEC/HybriDetector/tree/fix_clustering. A total of 19 sample files were successfully processed with HybriDetector.

Post-Processing the HybriDetector Output

The driver-to-target sequence pairs from the HybriDetector pipeline were processed to curate a new miRNA benchmarking dataset. The 19 output files produced from HybriDetector were vertically concatenated into one dataset. The dataset was first filtered to retain only sequence pairs for which the non-coding RNA is a miRNA, eliminating other annotated non-coding RNA types, such as tRNAs and yRNAs. Next, the dataset was deduplicated based on the combination of miRNA to gene sequence, to ensure unique positive entries. To facilitate the exploratory data analysis presented here, it was ensured that all miRNAs were assigned their respective accurate miRNA family name, using mature sequence data from miRBase (Release 22.1) (Kozomara, Birgaoanu, and Griffiths-Jones 2019). The dataset was then alphabetically sorted by the gene sequence, to facilitate the next step comprising negative example generation. The negative examples were generated

by matching genes to miRNAs in the same dataset, excluding those miRNAs the genes are already matched to in the positive examples. Negative examples were produced in such a way that ensures they are unique, and sufficiently dissimilar (having a minimum Levenshtein distance of 3) to all the positive counterparts of each respective gene (described in more detail in Supplementary Methods). Negative examples were generated at ratios 1:1, 1:10, and 1:100. The three datasets were then each split into train and test sets according to which chromosome the binding site sequence originates from. In accordance with the Hejret2023 datasets, for the Manakov2022 datasets, binding site sequences from chromosome 1 were assigned to the testing set and the rest were assigned to the training set. This produced a training-to-testing split ratio of around 88/12 (1,262,123 / 169,533).

This pipeline generated a total of 6 comprehensive datasets (Fig. 1B) from the data presented by Manakov et al (Manakov et al. 2022). A detailed technical overview of the pipeline can be found in Supplementary Methods. The pipeline itself is freely available at: https://github.com/BioGeMT/miRBench_paper/tree/v0.1.1/code/post_process.

Data availability

The six Manakov2022 datasets produced were deposited on Zenodo at https://zenodo.org/records/13909173, alongside the Hejret2023 and Klimentova2022 datasets, so they are all easily accessible by the miRBench Python package presented here.

Description of Benchmarked Models

`TargetScanCnn_McGeary2019` (McGeary et al. 2019) is a convolutional neural network with two convolutional layers and two fully connected layers. It uses an outer product of a one-hot-encoded representation of the first 10 nucleotides of a miRNA and 12 nucleotides of its putative target. All possible 12-nucleotide subsequences of the putative target are scored and the highest score is used as a final prediction. TargetScanCnn was trained using RNA Bind-n-Seq (RBNS) -derived dissociation constants (K_d) and mRNA-transfection fold-change measurements produced by the authors. The model predicts -log(K_d) between miRNA and putative target, which can be used as a proxy for a binding affinity score.

`CnnMirTarget_Zheng2022` (Zheng et al. 2020) uses a convolutional neural network consisting of four convolutional layers with max-pooling followed by two dense layers with dropout. It uses a one-hot-encoded representation of miRNA concatenated with a putative target, padded to 110 nucleotides. CnnMirTarget was trained on data from three sources: 1) human AGO1 CLASH Helwak 2013 data (Helwak et al. 2013); 2) *C. elegans* and mammalian iPAR-CLIP data (Grosswendt et al. 2014); and 3) MirTarBase data with strong experimental evidence (Hsu et al. 2011). The model's predictions may be directly used to measure the probability of binding.

`TargetNet_Min2021` (Min, Lee, and Yoon 2022) is a deep learning model based on the ResNet architecture (K. He et al. 2015). It uses an one-hot-encoded representation of a miRNA and its putative target after extended seed alignment. TargetNet was trained on the miRAW dataset (Pla, Zhong, and Rayner 2018). This dataset was constructed from four sources: 1) Diana TarBase (Vlachos et al. 2015); 2) MirTarBase (Chou et al. 2016); 3) human AGO1 CLASH Helwak 2013 data (Helwak et al. 2013); and 4) *C. elegans* iPAR-CLIP data (Grosswendt et al. 2014). The model's predictions may be directly used to measure the probability of binding.

miRBind_Klimentova2022 (Klimentová et al. 2022) is a deep learning model based on the ResNet architecture (K. He et al. 2015). It uses a two-dimensional representation of miRNA and putative target sequence. To address the imbalance between positive and negative classes, miRBind uses an instance hardness-based label smoothing approach, which keeps important, discriminative examples in the final training dataset and discards easily classifiable examples to rebalance the skewed ratio (Guo and Viktor 2004). miRBind was trained on AGO1 CLASH Helwak 2013 data (Helwak et al. 2013). The model's predictions are directly used to measure the probability of binding.

miRNA_CNN_Hejret2023 (Hejret et al. 2023) is a convolutional neural network consisting of six convolutional layers followed by dense layers. It uses a two-dimensional representation of miRNA and putative binding site interaction introduced in miRBind (Klimentová et al. 2022). miRNA_CNN_Hejret2023 was trained on the AGO2 CLASH Hejret 2023 training set that is also presented in this manuscript. The model's predictions are directly used to measure the probability of binding.

InteractionAwareModel_Yang2024 (Yang et al. 2024) is a deep multi-head attention network consisting of three parts: 1) sequence feature extraction; 2) interaction pattern identification; and 3) classification. It uses a one-hot-encoded representation of miRNA sequence padded to 30 nucleotides and the first 40 nucleotides of a putative target. The Interaction-Aware Model was trained on re-analyzed data from human AGO1 CLASH Helwak 2013 (Helwak et al. 2013). This model's predictions may also be used directly to measure the probability of binding.

Beyond these ML methods that have been specifically trained for the task, we also implemented two simple methods based on co-folding and seed. **RNACofold** is a tool offered by the ViennaRNA Package (Lorenz et al. 2011). It computes the binding energy and base-pairing pattern of an input pair of interacting RNA molecule sequences. The model's output predicted minimum free energy of the folded structure is multiplied by -1 to be used as a measure of binding affinity. For the seed measure, we followed previous studies (McGeary et al. 2022) and four different definitions of seed were used: 1) **Seed8mer** - full complementarity on positions 2-8 and A on position 1; 2) **Seed7mer** - full complementarity on positions 2-8, or full complementarity on positions 2-7 and A on the position 1; 3) **Seed6mer** - full complementarity on positions 2-7, or full complementarity on positions 3-8 or full complementarity on positions 2-6 and A on the position 1; and 4) **SeedNonCanonical** - the same as Seed6mer, but allowing for a single bulge or mismatch.

miRBench Python Package

The package is structured into three modules: dataset, encoder, and predictor. The dataset module is responsible for access to the benchmark datasets described in this paper. The main function of the module is `get_dataset_path(dataset_name, ratio, split, force_download)` which downloads the selected dataset of a specified ratio and split, from Zenodo, into the local cache, and returns the local path to the dataset. A complementary function is `get_dataset_df(dataset_name, ratio, split, force_download)` which instead of returning the path to the dataset, returns the dataset loaded into a Pandas DataFrame.

The encoder module is responsible for encoding data into the format expected by a predictor module. The main function of the module is `get_encoder(predictor_name)` which returns an instance of an encoder object implemented for a specified predictor. The

encoder expects data as a Pandas DataFrame with columns named `noncodingRNA` and `gene`. Specifying custom column names is possible when calling the encoder. The returned data format differs for every encoder and is specific to the predictor.

The predictor module is responsible for predicting miRNA-binding site interaction. The main function of the module is `get_predictor(predictor_name)` which returns an instance of a specified predictor object. The predictor object expects data encoded by a corresponding encoder and returns an array of predictions. Supported encoders and predictors in miRBench version 1.0.0 include `TargetScanCnn_McGeary2019`, `CnnMirTarget_Zheng2020`, `TargetNet_Min2021`, `miRBind_Klimentova2022`, `miRNA_CNN_Hejret2023`, `InteractionAwareModel_Yang2024`, `RNACofold`, `Seed8mer`, `Seed7mer`, `Seed6mer`, and `Seed6merBulgeOrMismatch`, described under Materials & Methods, in the Section 'Description of Benchmarked Models'.

Running Benchmark Comparison of Selected Computational Methods on Selected Datasets

The miRBench Python package v 1.0.0 was installed with pip, and all the available predictor models were benchmarked on all 1:1, 1:10, and 1:100 positive to negative class imbalance ratios of each of the Hejret2023, Klimentova2022, and Manakov2022 datasets. The resulting predictions were evaluated against the true labels, to produce the Precision-Recall curves presented in Figure 4 and their respective Area-Under-Curve metrics presented in Table 2.

Retraining `miRNA_CNN_Hejret2023` Model on Manakov2022 1:1 Train Set and Subsets

The miRNA_CNN_Hejret2023 model was retrained on 6 datasets of increasing sizes subsampled from the Manakov2022 1:1 train set. The sizes of the datasets, including both positive and negative examples, are: 200, 2,000, 7,720 (the number of positives corresponds to the number of positives in the Hejret2023 1:10 train set), 20,000, 200,000, and 2,524,246 examples (full Manakov2022 1:1 train set). The subsampling was done randomly while keeping the positive:negative ratio exactly 1:1. Retraining of the miRNA_CNN_Hejret2023 model for all dataset sizes was carried out in the same way as in the original paper including all the hyperparameters settings (Hejret et al. 2023). The performance of the retrained models was then evaluated on the Manakov2022 1:100 testing set.

Acknowledgements

This work has been supported by funding from the projects BioGeMT (HORIZON-WIDERA-2022 Grant ID: 101086768) at the University of Malta and miRBench RNS-2024-022 for "Collaboration for microRNA Benchmarking" from Xjenza Malta awarded to Panagiotis Alexiou. Scientific data presented in this paper was obtained with the help of the Bioinformatics Core Facility of CEITEC Masaryk University supported by the NCMG research infrastructure (LM2023067 funded by MEYS CR). This work was supported by computational resources provided by the TargetID project, Novel Drug Targets for Infectious Diseases, funded by the Malta Council for Science and Technology (MCST) COVID-19 R&D Fund 2020 (Grant COV.RD.2020-11), and the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. The authors would like to thank E. Maragkakis and V. Martínek for their constructive criticism of the manuscript.

References

Alexiou, Panagiotis, Manolis Maragkakis, Giorgos L. Papadopoulos, Martin Reczko, and Artemis G. Hatzigeorgiou. 2009. “Lost in Translation: An Assessment and Perspective for Computational microRNA Target Identification.” *Bioinformatics* 25 (23): 3049–55.

Ambros, Victor. 2004. “The Functions of Animal microRNAs.” *Nature* 431 (7006): 350–55.

Bartel, David P. 2004. “MicroRNAs: Genomics, Biogenesis, Mechanism, and Function.” *Cell* 116 (2): 281–97.

Bernstein, Emily, Sang Yong Kim, Michelle A. Carmell, Elizabeth P. Murchison, Heather Alcorn, Mamie Z. Li, Alea A. Mills, Stephen J. Elledge, Kathryn V. Anderson, and Gregory J. Hannon. 2003. “Dicer Is Essential for Mouse Development.” *Nature Genetics* 35 (3): 215–17.

Broughton, James P., Michael T. Lovci, Jessica L. Huang, Gene W. Yeo, and Amy E. Pasquinelli. 2016. “Pairing beyond the Seed Supports MicroRNA Targeting Specificity.” *Molecular Cell* 64 (2): 320–33.

Calin, George A., and Carlo M. Croce. 2006. “MicroRNA Signatures in Human Cancers.” *Nature Reviews. Cancer* 6 (11): 857–66.

Chakraborty, Chiranjib, Ashish Ranjan Sharma, Garima Sharma, C. George Priya Doss, and Sang-Soo Lee. 2017. “Therapeutic miRNA and siRNA: Moving from Bench to Clinic as Next Generation Medicine.” *Molecular Therapy. Nucleic Acids* 8 (September):132–43.

Chou, Chih-Hung, Nai-Wen Chang, Sirjana Shrestha, Sheng-Da Hsu, Yu-Ling Lin, Wei-Hsiang Lee, Chi-Dung Yang, et al. 2016. “miRTarBase 2016: Updates to the Experimentally Validated miRNA-Target Interactions Database.” *Nucleic Acids Research* 44 (D1): D239–47.

Condrat, Carmen Elena, Dana Claudia Thompson, Madalina Gabriela Barbu, Oana Larisa Bugnar, Andreea Boboc, Dragos Cretoiu, Nicolae Suciu, Sanda Maria Cretoiu, and Silviu Cristian Voinea. 2020. “miRNAs as Biomarkers in Disease: Latest Findings Regarding Their Role in Diagnosis and Prognosis.” *Cells* 9 (2). https://doi.org/10.3390/cells9020276.

Dai, Rujuan, and S. Ansar Ahmed. 2011. “MicroRNA, a New Paradigm for Understanding Immunoregulation, Inflammation, and Autoimmune Diseases.” *Translational Research: The Journal of Laboratory and Clinical Medicine* 157 (4): 163–79.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. “ImageNet: A Large-Scale Hierarchical Image Database.” In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–55. IEEE.

Didiano, Dominic, and Oliver Hobert. 2006. “Perfect Seed Pairing Is Not a Generally Reliable Predictor for miRNA-Target Interactions.” *Nature Structural & Molecular Biology* 13 (9): 849–51.

Esquela-Kerscher, Aurora, and Frank J. Slack. 2006. “Oncomirs - microRNAs with a Role in Cancer.” *Nature Reviews. Cancer* 6 (4): 259–69.

Grešová, Katarína, Panagiotis Alexiou, and Ilektra-Chara Giassa. 2022. “Small RNA Targets: Advances in Prediction Tools and High-Throughput Profiling.” *Biology* 11 (12): 1798.

Grešová, Katarína, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. 2023. “Genomic Benchmarks: A Collection of Datasets for Genomic Sequence Classification.” *BMC Genomic Data* 24 (1): 25.

Grosswendt, Stefanie, Andrei Filipchyk, Mark Manzano, Filippos Klironomos, Marcel Schilling, Margareta Herzog, Eva Gottwein, and Nikolaus Rajewsky. 2014. “Unambiguous Identification of miRNA:target Site Interactions by Different Types of Ligation Reactions.” *Molecular Cell* 54 (6): 1042–54.

Guo, Hongyu, and Herna L. Viktor. 2004. “Learning from Imbalanced Data Sets with Boosting and Data Generation.” *SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining* 6 (1): 30–39.

Hébert, Sébastien S., and Bart De Strooper. 2009. “Alterations of the microRNA Network Cause Neurodegenerative Disease.” *Trends in Neurosciences* 32 (4): 199–206.

Hejret, Vaclav, Nandan Mysore Varadarajan, Eva Klimentova, Katarina Gresova, Ilektra-Chara Giassa, Stepanka Vanacova, and Panagiotis Alexiou. 2023. “Analysis of Chimeric Reads Characterises the Diverse Targetome of AGO2-Mediated Regulation.” *Scientific Reports* 13 (1): 22895.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. “Deep Residual Learning for Image Recognition.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1512.03385.

He, Lin, and Gregory J. Hannon. 2004. “MicroRNAs: Small RNAs with a Big Role in Gene Regulation.” *Nature Reviews. Genetics* 5 (7): 522–31.

Helwak, Aleksandra, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. 2013. “Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding.” *Cell* 153 (3): 654–65.

Hsu, Sheng-Da, Feng-Mao Lin, Wei-Yun Wu, Chao Liang, Wei-Chih Huang, Wen-Ling Chan, Wen-Ting Tsai, et al. 2011. “miRTarBase: A Database Curates Experimentally Validated microRNA-Target Interactions.” *Nucleic Acids Research* 39 (Database issue): D163–69.

Ikeda, Sadakatsu, Sek Won Kong, Jun Lu, Egbert Bisping, Hao Zhang, Paul D. Allen, Todd R. Golub, Burkert Pieske, and William T. Pu. 2007. “Altered microRNA Expression in Human Heart Disease.” *Physiological Genomics* 31 (3): 367–73.

Ivey, Kathryn N., and Deepak Srivastava. 2010. “MicroRNAs as Regulators of Differentiation and Cell Fate Decisions.” *Cell Stem Cell* 7 (1): 36–41.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.

Klimentová, Eva, Václav Hejret, Ján Krčmář, Katarína Grešová, Ilektra-Chara Giassa, and Panagiotis Alexiou. 2022. “miRBind: A Deep Learning Method for miRNA Binding Classification.” *Genes* 13 (12): 2323.

Kozomara, Ana, Maria Birgaoanu, and Sam Griffiths-Jones. 2019. “miRBase: From microRNA Sequences to Function.” *Nucleic Acids Research* 47 (D1): D155–62.

Lee, R. C., R. L. Feinbaum, and V. Ambros. 1993. “The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14.” *Cell* 75 (5): 843–54.

Liu, Jidong, Michelle A. Carmell, Fabiola V. Rivas, Carolyn G. Marsden, J. Michael Thomson, Ji-Joon Song, Scott M. Hammond, Leemor Joshua-Tor, and Gregory J. Hannon. 2004. “Argonaute2 Is the Catalytic Engine of Mammalian RNAi.” *Science* 305 (5689): 1437–41.

Liu, Zicheng, Jiahui Li, Siyuan Li, Zelin Zang, Cheng Tan, Yufei Huang, Yajing Bai, and Stan Z. Li. 2024. “GenBench: A Benchmarking Suite for Systematic Evaluation of Genomic Foundation Models.” *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/2406.01627.

Lorenz, Ronny, Stephan H. Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. 2011. “ViennaRNA Package 2.0.” *Algorithms for Molecular Biology: AMB* 6 (November):26.

Manakov, Sergei A., Alexander A. Shishkin, Brian A. Yee, Kylie A. Shen, Diana C. Cox, Samuel S. Park, Heather M. Foster,

Karen B. Chapman, Gene W. Yeo, and Eric L. Van Nostrand. 2022. “Scalable and Deep Profiling of mRNA Targets for Individual microRNAs with Chimeric eCLIP.” *bioRxiv*. https://doi.org/10.1101/2022.02.13.480296.

Marin, Frederikke Isa, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. 2023. “BEND: Benchmarking DNA Language Models on Biologically Meaningful Tasks.” *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/2311.12570.

McGeary, Sean E., Namita Bisaria, Thy M. Pham, Peter Y. Wang, and David P. Bartel. 2022. “MicroRNA 3’-Compensatory Pairing Occurs through Two Binding Modes, with Affinity Shaped by Nucleotide Identity and Position.” *eLife* 11 (February). https://doi.org/10.7554/eLife.69803.

McGeary, Sean E., Kathy S. Lin, Charlie Y. Shi, Thy M. Pham, Namita Bisaria, Gina M. Kelley, and David P. Bartel. 2019. “The Biochemical Basis of microRNA Targeting Efficacy.” *Science* 366 (6472). https://doi.org/10.1126/science.aav1741.

Min, Seonwoo, Byunghan Lee, and Sungroh Yoon. 2022. “TargetNet: Functional microRNA Target Prediction with Deep Neural Networks.” *Bioinformatics* 38 (3): 671–77.

Morita, Sumiyo, Takuro Horii, Mika Kimura, Yuji Goto, Takahiro Ochiya, and Izuho Hatada. 2007. “One Argonaute Family Member, Eif2c2 (Ago2), Is Essential for Development and Appears Not to Be Involved in DNA Methylation.” *Genomics* 89 (6): 687–96.

O’Connell, Ryan M., Konstantin D. Taganov, Mark P. Boldin, Genhong Cheng, and David Baltimore. 2007. “MicroRNA-155 Is Induced during the Macrophage Inflammatory Response.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (5): 1604–9.

Pla, Albert, Xiangfu Zhong, and Simon Rayner. 2018. “miRAW: A Deep Learning-Based Approach to Predict microRNA Targets by Analyzing Whole microRNA Transcripts.” *PLoS Computational Biology* 14 (7): e1006185.

Rad, Seyed Mohammad Ali Hosseini, Joshua Colin Halpin, Supannikar Tawinwung, Koramit Suppipat, Nattiya Hirankarn, and Alexander D. McLellan. 2022. “MicroRNA-Mediated Metabolic Reprogramming of Chimeric Antigen Receptor T Cells.” *Immunology and Cell Biology* 100 (6): 424–39.

Ren, Yuchen, Zhiyuan Chen, Lifeng Qiao, Hongtai Jing, Yuchen Cai, Sheng Xu, Peng Ye, et al. 2024. “BEACON: Benchmark for Comprehensive RNA Tasks and Language Models.” *arXiv [q-bio.QM]*. arXiv. http://arxiv.org/abs/2406.10391.

Robson, Eyes S., and Nilah M. Ioannidis. 2024. “GUANinE v1.0: Benchmark Datasets for Genomic AI Sequence-to-Function Models.” *bioRxiv : The Preprint Server for Biology*, March. https://doi.org/10.1101/2023.10.12.562113.

Rooij, Eva van, and Eric N. Olson. 2012. “MicroRNA Therapeutics for Cardiovascular Disease: Opportunities and Obstacles.” *Nature Reviews. Drug Discovery* 11 (11): 860–72.

Rooij, Eva van, Lillian B. Sutherland, Ning Liu, Andrew H. Williams, John McAnally, Robert D. Gerard, James A. Richardson, and Eric N. Olson. 2006. “A Signature Pattern of Stress-Responsive microRNAs That Can Evoke Cardiac Hypertrophy and Heart Failure.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (48): 18255–60.

Rupaimoole, Rajesha, and Frank J. Slack. 2017. “MicroRNA Therapeutics: Towards a New Era for the Management of Cancer and Other Diseases.” *Nature Reviews. Drug Discovery* 16 (3): 203–22.

Shen, Lianghua, Jiali Yang, Chijian Zuo, Jian Xu, Ling Ma, Qiaomei He, Xiao Zhou, et al. 2024. “Circular mRNA-Based TCR-T Offers a Safe and Effective Therapeutic Strategy for Treatment of Cytomegalovirus Infection.” *Molecular Therapy: The Journal of the American Society of Gene Therapy* 32 (1): 168–84.

Sonkoly, E., and A. Pivarcsi. 2009. “Advances in microRNAs: Implications for Immunity and Inflammatory Diseases.” *Journal of Cellular and Molecular Medicine* 13 (1): 24–38.

Thum, Thomas, and Gianluigi Condorelli. 2015. “Long Noncoding RNAs and microRNAs in Cardiovascular Pathophysiology.” *Circulation Research* 116 (4): 751–62.

Van Nostrand, Eric L., Gabriel A. Pratt, Alexander A. Shishkin, Chelsea Gelboin-Burkhart, Mark Y. Fang, Balaji Sundararaman, Steven M. Blue, et al. 2016. “Robust Transcriptome-Wide Discovery of RNA-Binding Protein Binding Sites with Enhanced CLIP (eCLIP).” *Nature Methods* 13 (6): 508–14.

Vlachos, Ioannis S., Maria D. Paraskevopoulou, Dimitra Karagkouni, Georgios Georgakilas, Thanasis Vergoulis, Ilias Kanellos, Ioannis-Laertis Anastasopoulos, et al. 2015. “DIANA-TarBase v7.0: Indexing More than Half a Million Experimentally Supported miRNA:mRNA Interactions.” *Nucleic Acids Research* 43 (Database issue): D153–59.

Walsh, Ian, Dmytro Fishman, Dario Garcia-Gasulla, Tiina Titma, Gianluca Pollastri, ELIXIR Machine Learning Focus Group, Jennifer Harrow, Fotis E. Psomopoulos, and Silvio C. E. Tosatto. 2021. “DOME: Recommendations for Supervised Machine Learning Validation in Biology.” *Nature Methods* 18 (10): 1122–27.

Wilkinson, Mark D., Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (March):160018.

Yang, Tzu-Hsien, Jhih-Cheng Chen, Yuan-Han Lee, Shang-Yi Lu, Sheng-Hang Wu, Fang-Yuan Chang, Yan-Cheng Huang, Mei-Hsien Lee, Yan-Yuan Tseng, and Wei-Sheng Wu. 2024. “Identifying Human miRNA Target Sites via Learning the Interaction Patterns between miRNA and mRNA Segments.” *Journal of Chemical Information and Modeling* 64 (7): 2445–53.

Zhao, Yong, Eva Samal, and Deepak Srivastava. 2005. “Serum Response Factor Regulates a Muscle-Specific microRNA That Targets Hand2 during Cardiogenesis.” *Nature* 436 (7048): 214–20.

Zheng, Xueming, Long Chen, Xiuming Li, Ying Zhang, Shungao Xu, and Xinxiang Huang. 2020. “Prediction of miRNA Targets by Learning from Interaction Sequences.” *PloS One* 15 (5): e0232578.