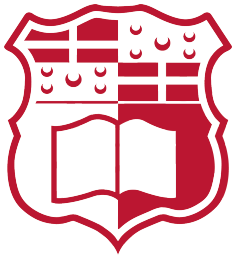

Advancing microRNA Target Site Identification via Bias-Corrected Chimeric Datasets for Machine Learning Approaches

Stephanie Sammut

Supervised by Dr Panagiotis Alexiou

Centre for Molecular Medicine and Biobanking

University of Malta

October, 2025

A dissertation submitted in partial fulfilment of the requirements for the degree of M.Sc. in Molecular Medicine (By Research).

WWW.UM.EDU.MT

First edition, Tuesday 14th October, 2025

To my parents

For investing in my education and supporting me with unwavering faith and love.

Acknowledgements

I would like to express my sincerest gratitude to my supervisor, Dr Panagiotis Alexiou, for his invaluable mentorship and guidance throughout this journey. I am truly grateful for the opportunity to work under his supervision and for the personal and professional growth that this experience has provided.

Thanks to his efforts, I was able to work closely with researchers at the Bioinformatics Core Facility at CEITEC in Brno, Czech Republic, who have since become both colleagues and friends.

I am especially grateful to Dr Katarina Gresova for her patient support and technical guidance, particularly during the early stages of my transition into this new field. I would also like to thank Dimosthenis Tzimotoudis, Eva Marsalkova, and David Cechak. Their collaboration has greatly enriched this experience.

Finally, I am deeply grateful to my boyfriend for his patience, continued support, and encouragement, especially during the busiest times. I would also like to thank my parents, sisters, in-laws, and friends, for helping to carry life's burdens throughout this journey, and for believing in me and always cheering me on.

Funding

This work was 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.

Computational resources

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).

Abstract

microRNAs (miRNAs) are small, non-coding RNA molecules that regulate gene expression post-transcriptionally. These ∼22-nucleotide long RNA molecules are loaded onto a protein of the Argonaute (AGO) family guiding it to specific RNA target transcripts, which are consequently inhibited from translation to protein. Despite years of research, the precise mechanisms that determine miRNA target recognition remain unclear.

Given that a single miRNA can potentially target any RNA transcript, experimental validation of all possible interactions is impractical. To this end, with the recent availability of volumes of data from high-throughput CLASH experiments that capture interacting RNA molecules mediated by a protein, many miRNA target site prediction methods employing data-driven approaches have been developed.

However, despite substantial efforts to produce more accurate models, currently, no standardised framework for benchmarking miRNA target site prediction methods exists, and various strategies have been adopted, hindering fair and reliable comparison. Consequently, this undermines the validity of performance claims.

Recently, another high-throughput experimental technique, called chimeric eCLIP, was developed, leading to a 70-fold increase in the recovery of miRNA–target site interactions. Here, we identify a unique opportunity to leverage this immense resource of publicly available raw data, to curate benchmark datasets for miRNA target site prediction.

Throughout this work, we additionally uncover a miRNA frequency class bias that arises as a result of the method used to generate negative examples in silico. Such methods are required when modeling miRNA target site prediction as a supervised binary classification problem, due to the absence of experimentally confirmed non-binding miRNA–target pairs. To this end, we develop a novel method for generating negatives that mitigates the identified bias.

Leveraging data from these high-throughput experiments and applying this new method for generating negatives, we curate three collections of datasets: one novel dataset containing almost three million examples, and bias-corrected versions of two smaller, published datasets. We contribute these datasets to the publicly available and easy-to-use `miRBench` Python package, providing a framework for benchmarking miRNA target site prediction methods.

We benchmark six state-of-the-art deep learning models on these datasets and train simple models to establish a baseline. Retraining a convolutional neural network, that was originally trained on a biased dataset (Average Precision Score (APS): 0.71), on a larger, bias-corrected dataset improves its performance (APS: 0.81), surpassing the previous state of the art (TargetScanCnn, APS: 0.76).

This highlights the advantages of well-curated, unbiased benchmarks, facilitating the development of more accurate miRNA target site prediction models thus enabling more reliable downstream analyses.

Contents

List of Figures

List of Tables

List of Abbreviations

AUC Area Under Curve

APS Average Precision Score

CNN Convolutional Neural Network

CLASH Crosslinking, Ligation, and Sequencing of Hybrids

CLIP Crosslinking and ImmunoPrecipitation

DL Deep Learning

DOME Data, Optimisation, Model, and Evaluation

eCLIP enhanced CLIP

FAIR Findable, Accessible, Interoperable, and Reusable

FN False Negative

FP False Positive

FPR False Positive Rate

ML Machine Learning

MFE Minimum Free Energy

NN Neural Network

PR Precision-Recall

ResNet Residual Neural Network

ROC Receiver Operating Characteristic

TP True Positive

TN True Negative

TPR True Positive Rate

UMI Unique Molecular Identifier

UTR Untranslated Region

1

Introduction

microRNAs (miRNAs) are small, non-coding RNA molecules that are involved in the post-transcriptional regulation of gene expression. These ∼22-nucleotide long RNAs are loaded onto a protein of the Argonaute (AGO) family forming the RNA-induced silencing complex (RISC). The miRNA serves as a guide, directing the RISC to its target RNA, including messenger RNAs (mRNAs), by binding specific target sites on the target transcript. Consequently, the mRNA is degraded or suppressed from being translated into the respective protein.

Since their discovery just over 30 years ago, miRNAs have been identified as master regulators of gene expression in fundamental biological processes such as embryogenesis, cell differentiation and proliferation, and aging (Bernstein et al., 2003; Ivey & Srivastava, 2010; Smith-Vikos & Slack, 2012; Zhao et al., 2005). Consequently, deregulation of miRNAs has been linked to various diseases, including cancer, cardiovascular disease, respiratory disease, neurological disorders, and immunoregulatory and inflammatory conditions (Calin & Croce, 2006; Dai & Ahmed, 2011; Hébert & Strooper, 2009; Maltby et al., 2016; Sonkoly & Pivarcsi, 2009; Thum & Mayr, 2012). Their involvement in key regulatory networks and disease pathogenesis has positioned miRNAs as promising candidates for both molecular diagnostics and therapeutic development (Condrat et al., 2020; Hanna et al., 2019). The 2024 Nobel Prize in Physiology or Medicine, awarded for the discovery of miRNAs, underscores their pivotal role in biology and biomedicine.

The identification and validation of specific miRNA–target interactions is essential for understanding the regulatory roles of miRNAs in complex biological systems. Given that a single miRNA can potentially target any of the ∼20,000 annotated human protein-coding RNA transcripts, as well as any other transcripts, experimental validation of all possible interactions is unrealistic, as it is expensive and time-consuming. Computational approaches have been leveraged to address this problem, enabling large-scale screening to prioritise candidate targets for downstream experimental validation. Consequently, a variety of miRNA target prediction programs have been developed.

Traditionally, miRNA target prediction tools employ rule-based approaches. An initial step in many such miRNA *target gene* prediction methods is the identification of a miRNA *target site*. For a single miRNA, target sites are scored based on their binding potential, facilitating the identification of target genes of that miRNA. Commonly employed rules involve seed pairing (described next), miRNA–target duplex energy, evolutionary conservation of the target site, and structural accessibility of the 3′ untranslated region (UTR) of the target mRNA.

The primary determinant of target specificity is the hexamer spanning nucleotide positions 2 to 7 at the 5′ end of the miRNA, known as the seed region. Canonical binding is defined by perfect Watson–Crick base pairing between this region and complementary sites in the 3′ UTR of the target mRNA (Bartel, 2009). However, several non-canonical modes of binding have also been described, adding complexity to miRNA target recognition. These include 3′ supplementary binding (Chipman & Pasquinelli, 2019; Grimson et al., 2007; Moore et al., 2015; Sheu-Gruttadauria et al., 2019), 3′ compensatory binding that allows imperfect pairing in the seed region (Bartel, 2009; Chi et al., 2012), and central pairing in the miRNA (Shin et al., 2010). Functional interactions lacking seed pairing entirely have also been reported (Helwak et al., 2013; Lal et al., 2009). Notably, Helwak et al. (2013) found that 18% of miRNA–target interactions detected via crosslinking assays lacked pairing in the seed region, showing binding only at the miRNA 3′ end. Such findings illustrate that the full spectrum of miRNA target recognition mechanisms remains incompletely understood. This diversity of binding modes challenges the assumptions of rule-based models, which remain limited by high false positive and false negative rates (Alexiou et al., 2009; Pinzón et al., 2017).

Just over ten years ago, a novel experimental method called CLASH, for Crosslinking, Ligation, and Sequencing of Hybrids, marked a pivotal advancement in the identification of miRNA–target site interactions by enabling the direct detection of the RNA–RNA hybrids bound to the AGO protein. Introduced by Helwak et al. (2013), CLASH involves in vitro UV crosslinking of AGO–RNA complexes, followed by a key ligation step that covalently binds the interacting RNA strands. This ligation step allows for the recovery of chimeric reads. Chimeric reads are sequencing reads comprising both the miRNA sequence and its ligated target site sequence, thereby providing direct evidence of AGO-mediated miRNA–target site interactions.

Since chimeric reads span two joined RNA sequences, aligning them to the reference genome using standard tools is not straightforward. Each read originates from two distinct transcripts, requiring specialised processing. To address this, chimeric read data must be handled by dedicated bioinformatics pipelines that identify chimeric reads among background sequences, split them into their miRNA and target site components,

and reconstruct the corresponding interactions. Several such pipelines have been developed. These tools typically output structured sets of miRNA–target site sequence pairs, which serve as valuable resources for downstream analysis. One such tool is HybriDetector (Hejret et al., 2023), which is employed in this work and adapted to handle larger input files.

The availability of high-throughput datasets of chimeric miRNA–target site interactions from CLASH and similar experiments has since motivated the training of machine learning, specifically deep learning, models on the task of miRNA target site prediction. Such models are able to capture binding information across the interacting molecules *derived from data*, which has demonstrated improvements in predictive performance compared to traditional rule-based methods (Parveen et al., 2019).

The task of miRNA target site prediction is generally modeled as a supervised binary classification problem, whereby a machine learning (ML) model is trained on labeled examples of binding and non-binding miRNA–target site sequence pairs, and is then used to classify previously unseen examples. The interactions must be represented by features that are manually extracted from the data, encoding relevant biological information into a numeric form compatible with the expected input for ML models. However, hand-crafting features is a laborious process that requires deep domain expertise and is therefore guided by prior knowledge in the field.

Conversely, the multiple hidden layers in deep learning (DL) models allow them to learn relevant features directly from raw data (LeCun et al., 2015). Therefore, since miRNA targeting remains an open question, DL approaches may capture previously overlooked features relevant to the interaction into miRNA target site prediction models (Hwang et al., 2023). To this end, several data representations have been explored to represent the raw interacting sequences.

In this supervised learning setup, each miRNA–target site pair is labeled as either binding or non-binding. For this task, the miRNA–target site interactions derived from chimeric experiments serve as the positive (binding) class. Publicly available experimentally validated negative (non-binding) examples are scarce. In the absence of sufficient negative examples, a variety of computational strategies have been proposed to simulate them.

Once a final dataset is curated, ideally with a balanced ratio of positive to negative examples, it is typically split into the training and test sets. The training set is further split into training and validation subsets. The model is trained on the training subset, while the validation subset is used to evaluate the model's performance throughout model development. When a final model is obtained, it is evaluated on the held-out test set, which contains examples that are previously unseen by the model. This final eval-

uation provides an unbiased measure of the model's performance and generalisability, and should be compared to the performance of other methods on the same dataset, for the same ML task.

Benchmarking refers to the structured evaluation and comparison of different methods against defined datasets. In ML and increasingly in bioinformatics, benchmarks are key drivers of progress through rigorous, reproducible comparisons (Deng et al., 2009; Jumper et al., 2021).

1.1 | Motivation

Despite substantial efforts to develop more accurate models for miRNA target site prediction, no standardised framework currently exists for benchmarking such methods. Instead, a wide range of data sources are used to train and benchmark miRNA target site prediction models. In addition to source diversity, there is considerable variation in dataset size, ranging from hundreds to tens of thousands of examples, as well as in strategies for generating negative examples, the class balance, and the choice of methods compared during benchmarking.

These factors make benchmarking in the field challenging, and undermine the validity of performance claims making it difficult to assess genuine improvements. Consequently, downstream research is affected, including the development of miRNA target gene prediction tools and subsequent experimental validation seeking to advance miRNA-mediated regulatory networks. These limitations highlight the need for a standardised benchmarking framework for miRNA target site prediction methods, to enable fair, reliable model evaluations.

Recently, another high-throughput experimental technique called chimeric eCLIP, for enhanced Crosslinking and Immunoprecipitation, was developed by Manakov et al. (2022). Similar to CLASH, the method integrates a chimeric ligation step into the eCLIP method, which itself has an improved library generation efficiency relative to earlier CLIP methods, while maintaining single-nucleotide binding resolution (Van Nostrand et al., 2016). Additionally, two variations of the protocol were described: one including and one excluding the SDS-PAGE separation and nitrocellulose membrane transfer step, a technically demanding purification step that is standard in CLIP-based protocols. Comparison of the two variants revealed that omission of this laborious step led to a 70-fold increase in the recovery of miRNA–target RNA interactions (Manakov et al., 2022).

Additionally, the chimeric eCLIP experiment was carried out on the AGO2 protein, the only AGO protein, out of the four in mammals, whose disruption was lethal to embryonic mice, suggesting a superior, non-redundant regulatory role relative to its counterparts (J. Liu et al., 2004; Morita et al., 2007). Furthermore, in contrast with the CLASH technique, which involves AGO protein induction, chimeric eCLIP leverages endogenous concentrations of AGO protein, which may yield a more accurate representation of native miRNA targeting, at least in the HEK293T cell line, on which both the CLASH and chimeric eCLIP experiments were performed.

Manakov et al. (2022) have made this vast resource of raw reads publicly available and to the best of our knowledge, it has not yet been leveraged for any ML approaches in the field. As such, we identify a valuable opportunity to curate a large, easy-to-use benchmarking dataset for training and evaluating deep learning models on the task of miRNA–target site prediction.

Throughout this study, we additionally uncover what we term a miRNA frequency class bias in existing chimeric datasets, used for the training and evaluation of several deep learning models in the field (Hejret et al., 2023; Klimentová et al., 2022; Pla et al., 2018). The bias occurs when the miRNA frequency distribution in the negative class differs from that in the positive class, arising from the method used to generate negative examples in silico. Further analyses show that it is a pervasive problem in the field yielding models that struggle to generalise to unseen examples, as they learn the underlying biases of the data rather than biologically significant patterns that separate binding from non-binding sequence pairs (Hejret et al., 2023; Klimentová et al., 2022; Min et al., 2022; Zheng et al., 2020).

This provides another opportunity to develop a novel method for generating negative examples. The method will be used to curate the negative class of the new benchmarking dataset, but also to recreate the negative class for some published biased datasets (Hejret et al., 2023; Klimentová et al., 2022) to produce bias-corrected versions.

Overall, we hope that this work strengthens current benchmarking strategies for miRNA target site prediction methods, facilitating fair comparison that drives the development of more accurate models, and enabling more reliable downstream analyses.

1.2 | Aims and Objectives

The aim of this work is to standardise benchmarking in the field of miRNA target site prediction. We define the following objectives:

Objective 1. To produce carefully-curated benchmarking datasets and make them FAIR (Findable, Accessible, Interoperable, and Reusable) which lowers the barrier of entry to the field for machine learning experts.

Objective 2. To review and evaluate the state of the art in the field of miRNA target site prediction using the curated datasets.

Objective 3. To produce baseline models by training and benchmarking classical machine learning and neural network models on the curated datasets.

1.3 | Approach

We first review the literature to understand the history of the field and related topics, and to survey key developments in the following areas: chimeric experimental techniques and the availability of chimeric datasets; computational pipelines for detecting chimeric interactions from raw reads; machine learning approaches to miRNA target site prediction; in silico methods for generating the negative class; and benchmarking strategies in the field.

Our approach for the first objective involves several steps and analyses. Using an adaptation of the HybriDetector pipeline for detecting chimeric interactions (Hejret et al., 2023), we process publicly available raw read data from the high-throughput chimeric eCLIP experiment (Manakov et al., 2022) to curate a structured dataset of miRNA–target site interactions that defines the positive class of the Manakov2022 dataset collection.

For the negative class, we first reproduce the in silico method for generating negative examples described by Hejret et al. (2023) and Klimentová et al. (2022). Initial exploratory data analysis reveals a bias in the miRNA frequencies across the classes. We identify the same bias in other published datasets: the Hejret, Klimentova and miRAW datasets (Hejret et al., 2023; Klimentová et al., 2022; Pla et al., 2018).

To correct this, we develop a new method for generating negatives. We use this method to generate negatives for the new Manakov2022 dataset and for the positive class of the published Hejret and Klimentova datasets, producing three final dataset collections: the Manakov2022, Hejret2023, and Klimentova2022 datasets. After further

processing to split the datasets for use in machine learning, we produce a total of six sets: two train sets (Manakov2022 train set, Hejret2023 train set) and four evaluation sets (Manakov2022 test and left-out sets, Hejret2023 test set, Klimentova2022 test set). We contribute these datasets to the publicly available and easy-to-use `miRBench` Python package for miRNA target site prediction benchmarks, which also includes implementations of state-of-the-art models and their corresponding data encoders (Sammut et al., 2025b).

To achieve our second objective, we review the state of the art and identify six relevant deep learning models, including three convolutional neural networks (CNNs) (Hejret et al., 2023; McGeary et al., 2019; Zheng et al., 2020), two residual neural networks (ResNets) (Klimentová et al., 2022; Min et al., 2022), and one attention-based network (T.-H. Yang et al., 2024). Leveraging `miRBench`, we encode each of the four curated evaluation sets into the input representation originally used to train each of the models. We benchmark all six models on all four evaluation sets.

For our third objective, we first design a feature set based on interacting miRNA–target site sequence pairs in our newly curated dataset collection (Manakov2022 train, test and left-out sets). We train and evaluate three tree-based machine learning models (Decision Tree, Random Forest, and XGBoost), establishing a baseline for simpler models.

Secondly, we implement the data representation described by Hejret et al. (2023) and Klimentová et al. (2022), which uses the raw miRNA–target site sequences and Watson–Crick base pairing, to encode all curated dataset collections. We adopt the architecture of a published CNN model, originally trained on the biased Hejret dataset (Hejret et al., 2023), and retrain it on each of the two new unbiased train sets. We evaluate each retrained model on all four evaluation sets to establish new performance baselines for deep learning models. Additionally, we explore an enhanced data representation that incorporates intermolecular binding derived from co-folding methods.

1.4 | Document Structure

This dissertation comprises a total of five chapters. The four chapters that follow are summarised next.

Chapter 2: Background and Literature Review We provide an overview of the field of miRNA target site prediction, highlighting key developments that have shaped its progression, with emphasis on topics most relevant to this study. We review state-of-the-art miRNA target site prediction methods, focusing on model types, data representations, methods for generating negative examples in silico, and benchmarking strategies.

Chapter 3: Methodology We describe in detail all steps and analyses carried out to achieve the objectives of this study. The chapter defines all data sources and software used, including their versions, and includes descriptions of key concepts where necessary. All methodological choices are supported by rationale and aligned with the specific goals of this work.

Chapter 4: Results & Discussion We present the results of all analyses carried out in this study, compare and contrast them where relevant, and discuss potential inferences, including biological insights. Where applicable, results are interpreted with reference to existing literature.

Chapter 5: Conclusion We summarise the work carried out and highlight our contributions to the field, while revisiting the aims and objectives. We conclude by providing critique and acknowledging the limitations of our work, and propose potential future directions that build on this study.

2

Background & Literature Overview

In this chapter, we introduce the key concepts necessary to understand and contextualise this study, which focuses on the computational prediction of interacting miRNA–target site sequences. The field has evolved from the discovery of miRNAs and their peculiar, still poorly understood targeting behaviour. miRNAs can potentially interact with any RNA transcript, making experimental validation of targets impractical at scale. Computational methods thus offer a way to prioritise candidate targets. In this study, we specifically focus on miRNA target *site* prediction which is the first step in many miRNA target gene prediction tools. This field has seen a range of approaches, from traditional rule-based strategies to more recent data-driven methods, enabled by emerging experimental techniques that capture direct evidence of protein-mediated RNA–RNA interactions. We provide an overview of the relevant literature and the evolution of the field, and identify current gaps that motivate our work.

2.1 | Biological Background

RNA is broadly categorised into coding messenger RNAs (mRNAs), that transmit genetic instructions to make proteins, and non-coding RNAs (ncRNAs), which do not translate into proteins. The latter encompass housekeeping RNAs, like small nuclear RNAs (snRNAs), ribosomal RNAs (rRNAs), and transfer RNAs (tRNAs), as well as regulatory RNAs, such as long non-coding RNAs (lncRNAs) and small RNAs (L.-L. Chen & Kim, 2024). LncRNAs are over 200 nucleotides (nt) in length, varying from several hundreds to over 100,000 nt. Small RNAs are defined by their size and are typically less than 200 nt in length. These mainly include microRNAs (miRNAs), small interfering RNAs (siRNAs), and PIWI-interacting RNAs (piRNAs), each involved in different regulatory functions within the cell (L.-L. Chen & Kim, 2024). This work focuses on the small non-coding RNAs—*microRNAs*.

miRNAs are small RNA molecules, approximately 22 nt long, that are involved in the post-transcriptional regulation of gene expression, by guiding Argonaute (AGO) proteins to target mRNA and suppressing its translation to protein.

The first miRNA, *lin-4*, was discovered in the invertebrate model organism *C. elegans*, where it was shown to regulate developmental timing by producing this small, 22-nt non-coding RNA instead of a protein-coding transcript (Lee et al., 1993). This RNA was found to repress the protein levels of its target, *lin-14*, through imperfect base pairing in the 3′ untranslated region (3′ UTR) of the mRNA (Wightman et al., 1993).

A second small RNA, *let-7*, was later identified and shown to act similarly (Reinhart et al., 2000), and its temporal expression pattern was found to be conserved across bilaterian animals, including humans (Pasquinelli et al., 2000).

These discoveries suggested that small RNAs might represent a broader class of post-transcriptional regulators (Bartel, 2018). Subsequent large-scale screens confirmed this, identifying many similar RNAs processed from hairpin-like precursors, now known as microRNAs (miRNAs) (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee & Ambros, 2001). Since then, a total of 48,860 mature miRNAs from 271 species have been annotated, including 2,654 mature miRNAs in humans (Kozomara et al., 2019).

2.1.1 | microRNA Biogenesis and Function

miRNAs are initially transcribed as long transcripts, which can be several kilobases in length, that undergo a series of enzymatic processes to produce the mature approximately 22-nucleotide long sequences termed mature miRNAs.

As reviewed by Bartel (2018), the canonical biogenesis of a mature miRNA (Figure 2.1) begins in the nucleus with the transcription of miRNA genes by RNA polymerase II (Pol II). The resulting primary transcripts, known as pri-miRNAs, are typically capped and polyadenylated. A key characteristic of pri-miRNAs is their ability to fold back on themselves to form one or more distinctive hairpin structures.

These hairpin-like pri-miRNAs are then processed by a nuclear complex called the Microprocessor. This complex is composed of one Drosha endonuclease molecule and two DGCR8 RNA-binding protein (RBP) molecules. Drosha includes two Ribonuclease III (RNase III) enzymes that each cleave one strand of the pri-miRNA within the hairpin stem with an offset of two nucleotides. DGCR8 aids in the recognition and positioning of Drosha for precise cleavage. This cleavage event releases a shorter hairpin RNA of approximately 60 nucleotides, known as the precursor miRNA (pre-miRNA), which possesses a 5′ phosphate and a 2-nucleotide 3′ overhang.

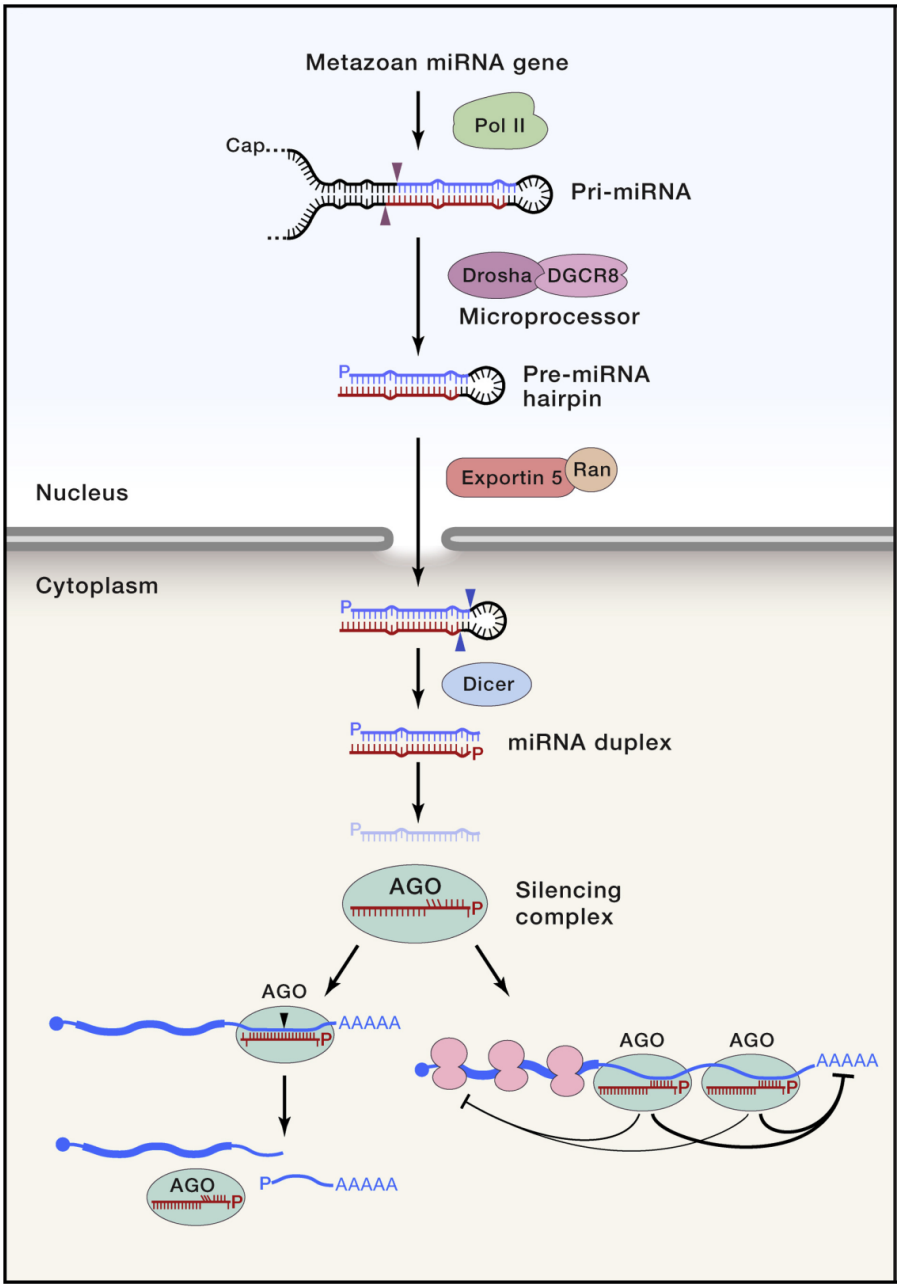

Figure 2.1: An illustration of miRNA biogenesis and function. MicroRNA genes are transcribed into pri-miRNA in the nucleus. Pri-miRNA are cleaved into the pre-miRNA and exported into the cytoplasm. Pre-miRNA are further cleaved to form a miRNA duplex. One of the strands is loaded on the RISC (Silencing complex). The loaded miRNA guides the RISC to its target transcripts, inhibiting translation to protein via slicing or other repression modes. Reproduced from Bartel (2018).

The pre-miRNA is then transported out of the nucleus into the cytoplasm by Exportin 5 and RAN-GTP. In the cytoplasm, the pre-miRNA undergoes a second processing step mediated by another RNase III enzyme called Dicer. Dicer recognises the pre-miRNA hairpin and cleaves it near the loop to produce a miRNA duplex, which is approximately 20 nucleotides long and also features 2-nucleotide 3′ overhangs.

Finally, the miRNA duplex can interact with an Argonaute (AGO) protein to form the RNA-induced silencing complex (RISC). One strand from the duplex, known as the guide strand or mature miRNA, is preferentially incorporated into the RISC, while the other strand, the passenger strand (miRNA*), is typically degraded. The selection of the guide strand depends on features at the 5′ end of each strand. The 5′-terminal nucleotide is anchored in a pocket within one of the four structural domains composing the AGO protein—the MID domain—influencing how the duplex is loaded into AGO (Bartel, 2018).

Four AGO proteins have been identified in mammals (AGO1 to AGO4). AGO2 is particularly critical; unlike AGO1, AGO3, and AGO4, its disruption proved lethal to embryonic mice, highlighting its superior, non-redundant regulatory role relative to its counterparts (J. Liu et al., 2004; Morita et al., 2007).

Once loaded onto the AGO protein, the mature miRNA can then guide the complex to specific *target sites* on its *target mRNA transcripts*, mediating post-transcriptional repression through one of two distinct mechanisms. One mode involves extensive complementarity between the miRNA and its target, enabling the AGO protein to cleave the mRNA through its endonucleolytic activity (Bartel, 2018). This function is restricted to AGO2, which has retained the ancestral RNAi slicing capability (J. Liu et al., 2004; Meister et al., 2004). This slicing mechanism is rare in humans. The dominant repression mode in humans operates independently of slicing and does not require extensive base-pairing (Bartel, 2018). Instead, it depends on the recruitment of the TNRC6 adaptor protein, which binds to AGO and interacts with the poly(A)-binding protein (PABPC) associated with the poly(A) tail of the mRNA. TNRC6 also interacts with deadenylase complexes like CCR4–NOT, which shorten the poly(A) tail, leading to mRNA destabilisation and degradation. CCR4–NOT also recruits the helicase DDX6 and the 4E-T protein, which together enhance mRNA decay and also interfere with translation initiation (Bartel, 2018).

2.1.2 | Regulatory Role and Clinical Applications

miRNAs play crucial roles in fundamental biological processes, such as embryogenesis (Bernstein et al., 2003), cell differentiation and proliferation (Ivey & Srivastava, 2010; Zhao et al., 2005), and aging (Smith-Vikos & Slack, 2012), and are thus regarded as *master regulators* of diverse pathways. Consequently, deregulation of miRNAs is linked to a host of diseases, including cancer (Calin & Croce, 2006; Esquela-Kerscher & Slack, 2006), cardiovascular disease (Thum & Mayr, 2012), respiratory disease (Maltby et al., 2016), neurological disorders (Hébert & Strooper, 2009; kumar Karnati et al., 2015), and immunoregulatory and inflammatory conditions (Dai & Ahmed, 2011; Sonkoly & Pivarcsi, 2009).

Reflecting their clinical potential, miRNAs hold promise in molecular diagnostics as biomarkers (Condrat et al., 2020), but also in therapeutics, via miRNA mimics and antagomirs (Chakraborty et al., 2017; Hanna et al., 2019; Rupaimoole & Slack, 2017; van Rooij & Kauppinen, 2014).

2.2 | miRNA Target Site Prediction

The identification and validation of miRNA–target interactions are critical tasks for understanding the regulatory functions of miRNAs in complex biological systems. Given that a single miRNA can potentially target hundreds of transcripts, experimental validation of all interactions is impractical due to cost and time constraints. Computational approaches address this challenge by enabling large-scale screening to prioritise candidate targets for downstream experimental validation. Consequently, a variety of miRNA target prediction programs have been developed.

The early generation of target prediction tools relied on a set of broad, *rule-based* approaches, which considered features such as canonical seed pairing, miRNA–target duplex energy, conservation of target sites, and structural accessibility of the target mRNA 3′-UTR. Examples of such tools include RNAhybrid (Krüger & Rehmsmeier, 2006), miRanda (Enright et al., 2003), TargetScan (Lewis et al., 2005), and PITA (Kertesz et al., 2007), which rely on one or a combination of these features. A common initial step in these miRNA target *gene* prediction methods involves the transcriptome-wide scanning for candidate miRNA target *sites*. These sites are scored based on their binding potential, which helps identify target genes that are more likely regulated by the miRNA.

Traditionally, miRNA binding to a target site has been thought to depend primarily on the so-called *seed* region: the hexamer spanning nucleotide positions 2 to 7 at the 5′ end of the miRNA. Watson–Crick base pairing between this seed region and com-

plementary sites, typically located in the 3′ untranslated region (UTR) of mRNA transcripts, enables miRNA–RNA interactions.

In his review, Bartel (2018) describes various *canonical* binding sites (see Figure 2.2A). *6-mer* sites match miRNA nucleotides 2 to 7; *7-mer-m8* sites extend the 6-mer with a match at nucleotide 8; *7-mer-A1* sites pair the 6-mer with an adenine opposite miRNA nucleotide position 1; *8-mer* sites combine both the m8 match and the A1 feature; and *offset 6-mer* sites are shifted by one nucleotide in either the 5′ or 3′ direction.

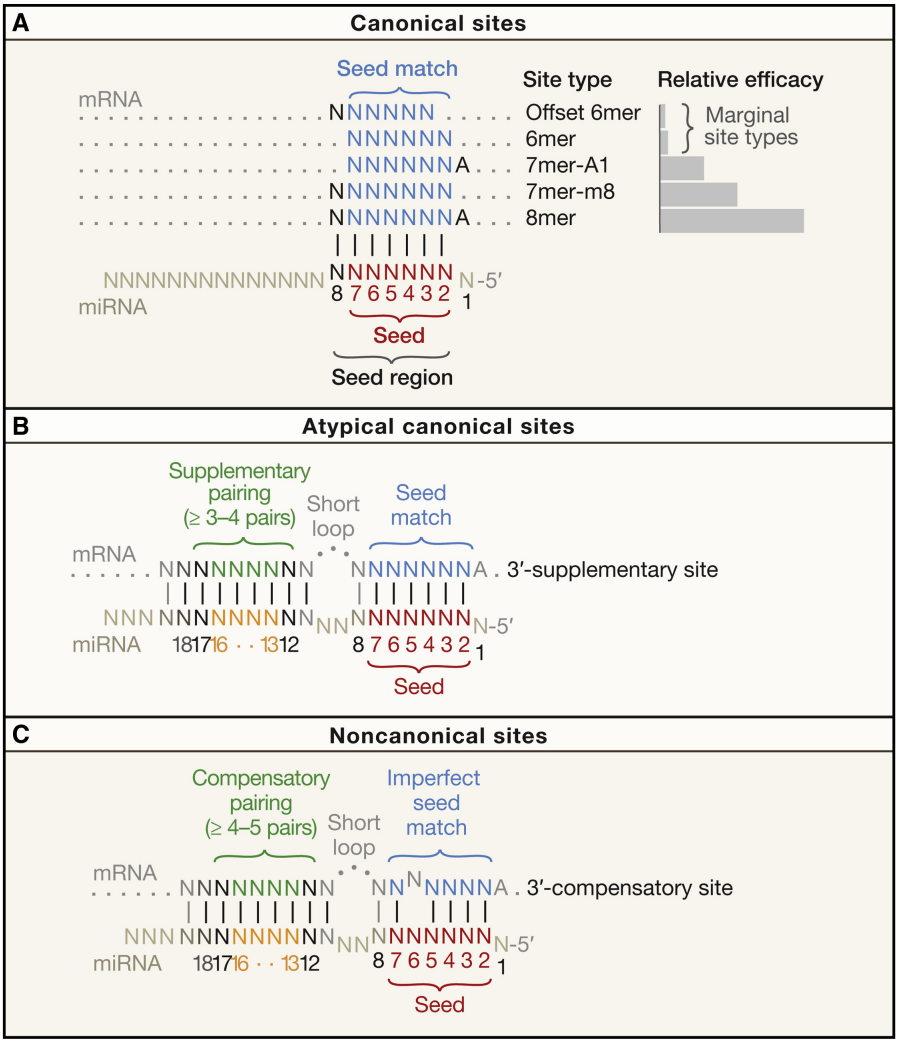

Figure 2.2: An illustration of miRNA target site types, as described by Bartel (2018). Traditionally, seed binding modes are what determines miRNA–target interaction. Reproduced from Bartel (2018).

However, seed pairing alone is not always sufficient to mediate effective target repression. Additional interactions involving the miRNA 3′ end, known as *3′ supplementary pairing*, may be necessary for regulatory efficiency on certain targets (Chipman & Pasquinelli, 2019; Grimson et al., 2007; Moore et al., 2015; Sheu-Gruttadauria et al., 2019). Bartel (2018) describes these binding sites as *atypical canonical* sites (see Figure 2.2B).

Non-canonical sites, by contrast, do not exhibit perfect seed pairing (see Figure 2.2C). These may accommodate bulges or mismatches within the seed region (Chi et al., 2012), or rely on *3′ compensatory pairing* (Bartel, 2009). Additional interaction modes such as *centered binding*, involving contiguous complementarity in the central region of the miRNA, have also been described (Shin et al., 2010).

Furthermore, target interactions that completely lack binding in the seed region have been documented (Helwak et al., 2013; Lal et al., 2009). Notably, Helwak et al. (2013) found that 60% of AGO1-bound human miRNA interactions detected by high-throughput cross-linking assays (discussed further in Section 2.3) were non-canonical, and 18% were non-seed interactions with binding limited solely to the 3′ end of the miRNA. Upon experimental validation some of these interactions even showed functional repression.

Unsurprisingly, the reliance of miRNA target prediction tools on these general rules, primarily canonical seed pairing and its known variants, combined with a limited understanding of non-canonical and non-seed binding patterns, has limited their accuracy. As a result, these tools often produce high rates of both false positive and false negative predictions (Alexiou et al., 2009; Pinzón et al., 2017).

Moreover, miRNA target sites are not limited to the 3′ UTR; they can also occur in the 5′ UTR and coding regions of target transcripts (Broughton et al., 2016; Lytle et al., 2007). In addition to the diversity of binding locations, the regulatory relationships themselves can be complex. Regulation can be combinatorial: a single miRNA may coordinately repress multiple transcripts within a biological pathway, while individual mRNAs can be co-regulated by several miRNAs (Wilczyńska & Bushell, 2014). These features add to the complexity of miRNA targeting, that remains an open question and is still far from being fully understood.

2.3 | Experimental Capture of AGO-bound RNA Hybrids

Cross-linking immunoprecipitation (CLIP) techniques are widely used to study RNA–protein interactions. They rely on in vitro ultraviolet (UV) irradiation to induce covalent bonds between RNAs and adjacent proteins, stabilising RNA–protein complexes for purification and sequencing. This approach enables transcriptome-wide identification of RNA binding sites for a given RNA-binding protein (RBP).

Despite methodological differences, CLIP-based protocols share a common workflow (Hafner et al., 2021). Following UV cross-linking of RNA and associated proteins, RNase digestion reduces RNA to shorter fragments, and the target RNA–protein complexes are isolated, typically via immunoprecipitation. These are further purified using SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel electrophoresis) and transferred to a nitrocellulose membrane, which selectively retains cross-linked complexes. Proteolytic digestion then releases the RNA, which is converted into cDNA for high-throughput sequencing. Finally, reads are aligned to the genome and computational methods are applied to distinguish possible binding sites, represented by clusters of overlapping reads, from background signals.

The adaptation of CLIP for high-throughput sequencing led to the development of HITS-CLIP (Chi et al., 2009). Subsequent variants such as individual-nucleotide resolution CLIP (iCLIP), infrared CLIP (irCLIP), and enhanced CLIP (eCLIP) introduced refinements in the purification steps and cDNA library preparation protocols allowing the precise identification of protein–RNA contact points at single-nucleotide resolution (Hafner et al., 2021).

Therefore, conventional CLIP-based approaches primarily capture *single* RNA strands that are cross-linked to RBPs, which limits their ability to directly resolve RNA–RNA interactions mediated by these proteins. Variants such as AGO HITS-CLIP (Chi et al., 2009) and PAR-CLIP (Hafner et al., 2010) specifically cross-link AGO proteins to bound RNAs, thereby enabling the identification of both AGO-bound miRNAs and AGO-bound target transcripts. These two datasets are then computationally integrated to infer miRNA–target interactions. However, the pairing is guided by sequence complementarity to the miRNA seed region, inherently biasing the analysis towards canonical seed-based interactions. As a result, these approaches provide only indirect evidence of interaction and operate under assumptions that may limit their ability to capture the full range of miRNA binding events.

In 2013, CLASH (Crosslinking, Ligation, and Sequencing of Hybrids) marked a pivotal advancement in the identification of miRNA–target interactions by enabling the direct detection of RNA–RNA hybrids bound to AGO1. Introduced by Helwak et al.

(2013), CLASH uses in vitro UV crosslinking of AGO–RNA complexes, followed by a key *ligation step* that covalently joins the interacting RNA strands. This ligation allows for the recovery of *chimeric reads*. These are sequencing reads comprising the miRNA and its target site joined together, thereby providing direct evidence of RNA–RNA interactions without relying on prior assumptions about binding patterns.

In their study, Helwak et al. (2013) reported that approximately 2% of sequencing reads were chimeric, yielding 18,514 miRNA–RNA interactions. As previously noted, 60% of these interactions were non-canonical, including bulges or mismatches in the seed region, and 18% were non-seed interactions. The availability of this resource highlights the potential to study miRNA targeting beyond conventional seed-based models (Helwak et al., 2013).

Recently, Hejret et al. (2023) produced a valuable resource of chimeric interactions by applying the CLASH method to the functionally essential and non-redundant AGO2 protein in human HEK293T cells. This experiment yielded 14,205 high-confidence chimeric interactions.

Although highly informative, CLASH involves multiple purification and ligation steps, including SDS-PAGE separation and nitrocellulose membrane transfer, that are standard in CLIP-based protocols but are technically demanding and difficult to scale. These constraints may limit RNA throughput and reduce final recovery.

Another notable recent development, described by Manakov et al. (2022), is chimeric eCLIP, also performed in the HEK293T cell line. This method incorporates a ligation step into the AGO2 eCLIP protocol, enabling the recovery of AGO2-mediated chimeric reads at single-nucleotide binding resolution (see Figure 2.3). Importantly, the protocol was described both with and without the SDS-PAGE separation and nitrocellulose membrane transfer step. Comparison of the two variants revealed that omitting this laborious step led to a 70-fold increase in the recovery of miRNA–RNA interactions (Manakov et al., 2022). This data provides a valuable opportunity to further investigate miRNA regulatory networks in the well-characterised HEK293T cell line.

A key distinction between CLASH and chimeric eCLIP is that CLASH involves artificial overexpression of AGO, which may expand the pool of detectable miRNA–target interactions by including those with weaker binding affinity. In contrast, chimeric eCLIP captures interactions mediated by endogenously expressed AGO, offering a snapshot of miRNA targeting under physiological conditions. This difference has led to the suggestion that CLASH may recover a broader, but potentially less biologically representative, set of interactions (Hejret et al., 2023), whereas chimeric eCLIP may provide a more accurate view of native miRNA targeting.

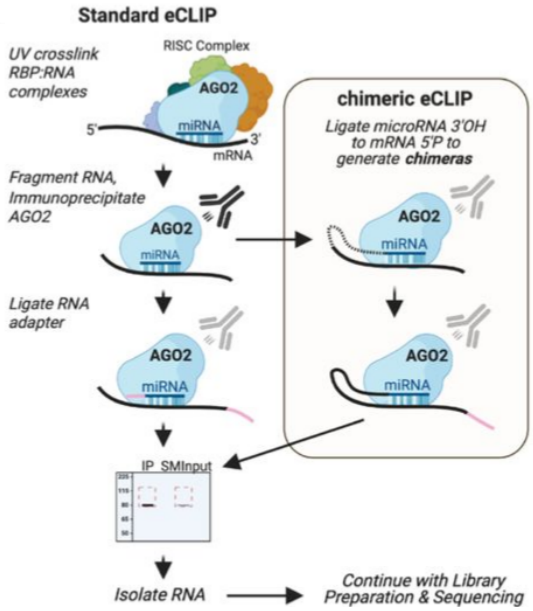

Figure 2.3: Schematic of the chimeric eCLIP experiment. The ligation step joins the AGO-mediated interacting miRNA and target site sequences, which are then recovered as chimeric reads. Reproduced from Manakov et al. (2022).

2.4 | Computational Detection of Chimeric Reads

Since chimeric reads consist of two RNA sequences joined together, it is not straightforward to directly align them to the reference genome using standard alignment tools such as STAR, HISAT2, or Bowtie2. This is because each read spans two distinct transcripts: one part maps to a miRNA and the other to a target RNA.

To address this, both CLASH and chimeric eCLIP require specialised bioinformatics pipelines capable of identifying chimeric reads among sequenced reads, splitting them into two alignable parts, and reconstructing the corresponding miRNA–target interactions. A number of tools have been developed for this purpose, including Hyb (Travis et al., 2013), CLAN (Zhong & Zhang, 2017), ChiRA (Videm et al., 2021), SCRAP (Mills et al., 2023), and HybriDetector (Hejret et al., 2023), among others. These pipelines typically output a set of miRNA–target sequence pairs, annotated with metadata such as miRNA name and family names, and genomic coordinates and feature annotation of the target transcript.

Notably, besides miRNAs, AGO proteins can bind various other small RNA biotypes. To the best of our knowledge, HybriDetector is the only tool that is able to account for and annotate such non-miRNA guide sequences. Specifically, the authors cu-

rated non-coding RNA databases for each of tRNAs, rRNAs, snoRNAs, YRNAs, and vaultRNAs, from multiple sources in an attempt to produce comprehensive and complete sources of annotations for known non-coding RNA biotypes (Hejret et al., 2023). This reduces the risk of false alignments. For this reason, in this study we employ an adaptation of the HybriDetector pipeline to detect chimeric interactions in the raw read data produced by the chimeric eCLIP experiment by Manakov et al. (2022). Since it is particularly relevant to our work, an overview of the pipeline is outlined next, and briefly in Figure 2.4. Bold italicised titles and italicised sub-titles are used to guide the description for clarity where needed.

2.4.1 | The HybriDetector Pipeline

The pipeline begins with optional read pre-processing steps, including Unique Molecular Identifier (UMI) extraction to the read name, adapter trimming, and quality filtering.

RNA fragments recovered from methods such as CLASH and chimeric eCLIP, may include a number of molecules resulting from failed intermolecular ligation, including *single non-coding RNA reads* and *single genomic reads*. These interfere with the capture of chimeric reads and are therefore accounted for in chimeric read detection pipelines such as HybriDetector.

Filtering Single Genomic Reads

Reads are first aligned to the genome to separate fully mapped reads as single genomic reads. A minimum of 85% of the read length must be aligned, allowing for a maximum mismatch of only 10% of the read length.

Filtering Single Non-Coding RNA Reads

The remaining reads are aligned to the several individual non-coding RNA databases curated by the authors. Reads with a minimum of 25% of the read length aligned, allowing for a maximum mismatch of only 10% of the read length, are retrieved.

Files resulting from alignment to all individual non-coding RNA databases are pooled and sorted by read name, producing blocks of per read possible alignments from individual reference databases. Further processing is carried out per read block. A read is considered a single non-coding RNA read if it has only one alignment, or more alignments of the same sequence to multiple references, and does not contain more than 6 consecutive unmapped nucleotides at each end.

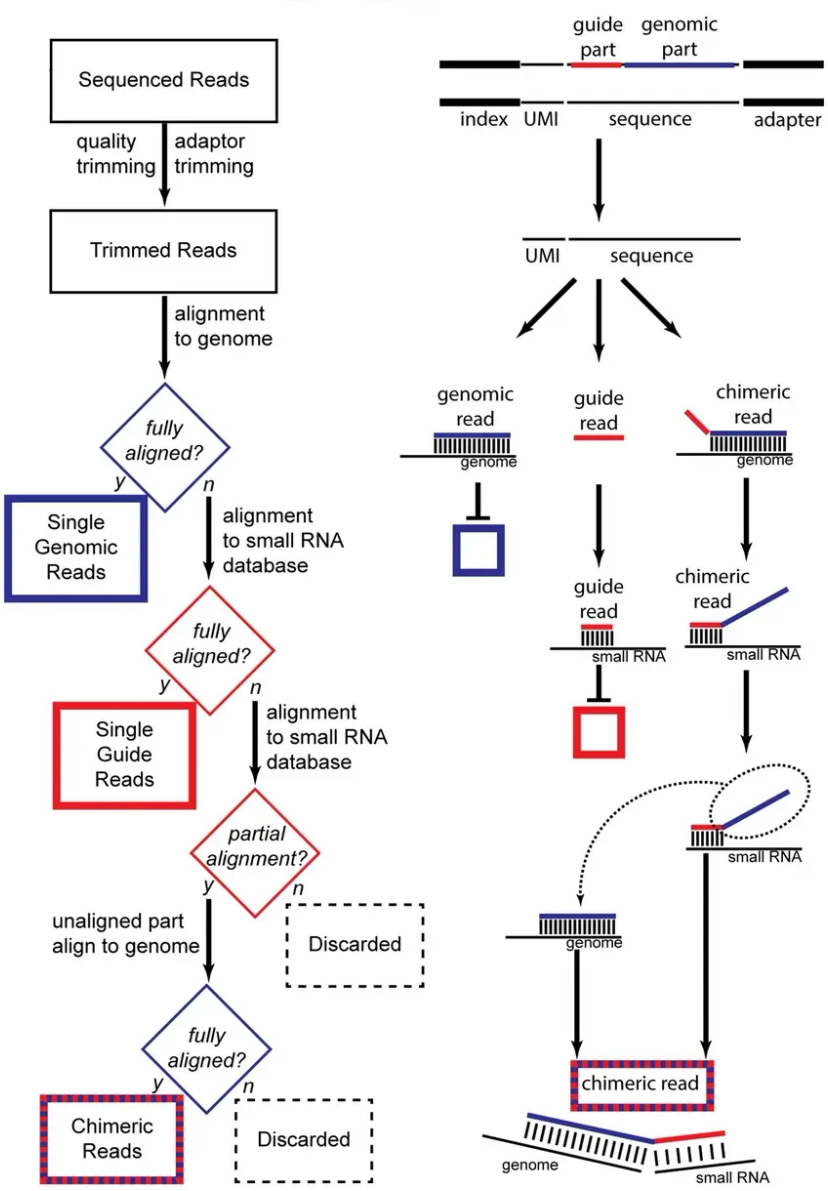

Figure 2.4: Workflow of the HybriDetector pipeline: from sequenced reads (top) to chimeric interactions (bottom). Reproduced from Hejret et al. (2023).

Filtering Potential Chimeric Reads

The next steps identify potential chimeric reads. First, the non-coding RNA part in the chimera is defined. To do this, for the remaining reads containing more than one alignment, the aligned parts (matched positions) of the reads are first identified and compared as follows.

Identifying the Non-Coding RNA in Potential Chimeric Reads

First, for each read block, the length of intersecting match positions, and the length of the entire spanned region of match positions (union), are computed. Then, if the intersection length is greater than 80% of the union length, the reads are considered sufficiently similar, and the read with the longest unmapped or *soft-clipped* part is selected as the representative read. Furthermore, if the read bears 15 or more soft-clipped bases at one end, and 6 or less at the other end, it is marked 'potentially chimeric'. Altogether, this effectively identifies the non-coding RNA parts in potential chimeric reads.

Identifying the Genomic Target in Potential Chimeric Reads

Next, the target RNA part in the chimera is defined. The longer soft-clipped part of the potential chimeric read is extracted and subject to another round of alignment against the entire human genome. Stricter settings are employed, whereby a minimum of 75% of the read length must be aligned, and only uniquely mapped soft-clipped parts are considered, to ensure more precise identification of genomic targets.

Connecting Non-Coding RNA to Genomic Target, Deduplication, and Annotation

The identified non-coding RNA part and the genomic target part of the potential chimeric reads are then assigned to one another based on the read name. These are deduplicated based on the Unique Molecular Identifier (UMI) sequence in the read name, and the sequence of the read. Annotations are added, including the gene name, genomic feature and coordinates, strand, single genomic read support, non-coding RNA sequence complexity, and miRNA family (if the non-coding RNA is a miRNA), among others.

Filtering and Collapsing Potential Chimeric Reads

A final step is the filtering and collapsing of potential chimeric reads to produce a final set of chimeric reads supported by a higher level of certainty. A number of cut-offs are first applied. The genomic target part must be at least 20 nucleotides long and supported by at least one single genomic read. The non-coding RNA part must not be longer than 30 nucleotides so it is compatible with AGO loading and must be sufficiently complex; each nucleotide type must not dominate more than 50% of the sequence.

Next, interactions are first collapsed to resolve redundancy. The data is grouped by various features related to non-coding RNA identity, target genomic position, and read mapping and quality. Strict grouping (i.e. by many features) ensures only identical interactions are subsequently merged. Within each group, corresponding to a unique non-coding RNA bound at an exact genomic site, the interactions are collapsed based

on the longest target sequence, and cumulative read support. This produces a single row per unique non-coding RNA–target pair, representing a distinct interaction.

These interactions are then collapsed a second time, whereby for the same non-coding RNA, nearby target sites, starting within ± 20 nucleotides, are considered functionally redundant due to shifted or fragmented alignments, and are merged into a single broader chimeric interaction. The final target region is defined as the union of such overlapping sites.

Single nucleotide polymorphisms, commonly resulting from the technical shortcomings in the library preparation and sequencing steps of CLASH and similar experiments, are then resolved via hierarchical clustering of the non-coding RNA sequences, producing a unique list of chimeric interactions. This is done by first computing a distance matrix, corresponding to the edit or Levenshtein distance between all possible pairs of non-coding RNA sequences. This distance reflects the number of insertions, deletions, or substitutions needed to transform one sequence into another. Each value in the matrix is then normalised by the average length of the corresponding sequence pair. Hierarchical clustering is subsequently performed on the normalised distances, and the resulting tree is cut at a defined height to assign cluster memberships. Chimeric interactions are then collapsed on the cluster ID, and the most supported—the one with the highest read count—is selected to represent the cluster.

Unifying Length of Genomic Targets

Target sites are normalised to a fixed length of 50 nucleotides. This is achieved by mapping each site to its genomic interval, computing the midpoint, and then generating a new genomic interval centred at this position, extending 24 nucleotides in one direction and 25 in the other. This step ensures that the entire target site is included within each chimeric interaction.

Final Output of HybriDetector

Finally, two main output files are produced, one containing all chimeric interactions detected, and another filtered 'high confidence' version containing chimeric interactions whose alignment of the small RNA sequence contains no mismatches. The output annotated interactions include details on miRNA–target site pairing, supporting evidence, secondary structure prediction, and genomic context, offering a high-confidence graph of miRNA–target interactions. Examples of such structured datasets produced using the HybriDetector pipeline include those presented by Hejret et al. (2023) derived from an AGO2 CLASH experiment, and those presented by Klimentová et al. (2022) derived from a small-scale AGO2 chimeric eCLIP experiment.

Such datasets serve an important role as direct evidence of true miRNA–target site binding interactions and can be used to model such interactions. In particular, they have been used as positive examples in machine learning, especially deep learning approaches for miRNA target site prediction. These models can capture features of the interacting molecules *directly from data*, and have shown improved predictive performance compared to traditional rule-based methods (Parveen et al., 2019). These approaches are described in the sections that follow.

2.5 | Machine Learning Approaches

Machine learning (ML) refers to a subset of artificial intelligence that involves the development of algorithms that enable systems to learn from data and improve their performance without explicit programming (Mitchell et al., 1997; Samuel, 1959). Most ML methods require feature engineering, a laborious process that requires domain expertise.

The task of miRNA target site prediction is generally modeled as a supervised binary classification problem, whereby a ML model is trained on labeled examples of binding (positive), and non-binding (negative), miRNA–target site interactions and is then deployed to classify previously unseen examples into one of the two categories. The negative class is described further in Section 2.5.3. The interactions must be represented by features extracted from the data that encode relevant biological information into a numeric form compatible with the expected input for ML models.

Examples of miRNA target site prediction tools that employ a ML approach using features extracted from chimeric datasets include `chimiRic` (Lu & Leslie, 2016) and `miRTarget` (Wang, 2016) that are based on support vector machine (SVM) classifiers, and `TarPmiR` (Ding et al., 2016) that is based on the random forest ensemble method.

The features used to train `chimiRic` describe the predicted duplex structures of miRNA–target site interactions. Three types of features are included: (i) Watson–Crick base pairing, and G–U and U–G wobbles, thermodynamically stable base pairs common in RNA secondary structure (Varani & McClain, 2000), at each position of the aligned pair of sequences; (ii) position of loop opening and extension in the duplex structure; and (iii) miRNA position-wise binary variables indicating pairing to a target site base.

In `miRTarget`, 50 features are used describing nucleotide composition, accessibility of the target site, seed conservation, seed base-pairing stability, and target site location. Specifically, for example, the secondary structure of the target sites is computed using `RNAfold` (Hofacker, 2003) and the folded structure and its associated free energy are

used as a proxy for the accessibility of the target site, while the number of conserved species for the target site is used as a proxy for seed conservation.

`TarPmiR` is trained on 13 features, selected from an initial pool of 18 features via four computational feature selection methods (Ding et al., 2016). These include conventional features such as seed match, folding energy, AU content, accessibility, and measures of conservation, the latter computed using average PhyloP scores (Pollard et al., 2009), across different regions of the miRNA–target site duplex. Additional features capture pairing characteristics, including per-position miRNA pairing probabilities (based on folded secondary structure), the total number of paired positions (overall and at the miRNA 3′ end), and the length and position of the longest consecutive base-pairing stretch.

This overview illustrates that feature extraction is a laborious process that requires deep domain expertise and is guided by prior knowledge about miRNA–target site binding in the field.

2.5.1 | Deep Learning Approaches

A neural network is a machine learning model inspired by the way biological neural networks in the brain process information. It typically consists of an input layer, one or more hidden layers, and an output layer. Each layer is composed of nodes, or artificial neurons, connected to those in the next layer. As data flows through the network, each neuron computes a weighted sum of its inputs and applies a nonlinear activation function, enabling the model to learn complex input–output relationships (LeCun et al., 2015).

Deep learning (DL) is a subset of neural networks, distinguished by the use of *multiple* hidden layers. These deep architectures allow models to learn data representations at increasing levels of abstraction, enabling the capture of intricate patterns in large datasets (LeCun et al., 2015). This means that DL allows automated feature extraction from raw data. The 2024 Nobel Prize in Physics, awarded for foundational discoveries and inventions enabling machine learning with artificial neural networks, underscores the central importance of neural networks in modern artificial intelligence and its applications.

Since miRNA targeting remains an open question, provided sufficient amounts of data, DL approaches may therefore capture previously overlooked features into miRNA target site prediction tools (Hwang et al., 2023). The availability of the unprecedented volume of data produced by recent high-throughput chimeric experiments (described

in Section 2.3), drove the training of deep learning models on the raw miRNA–target site sequences derived from chimeric reads.

Examples of prediction tools that adopt a DL approach using raw sequence-level data representations of chimeric interactions for miRNA target site prediction include: `miRAW` (Pla et al., 2018), based on a multi-layer perceptron; `mirLSTM` (Paker & Ogul, 2019), using a long short-term memory (LSTM) model; `DeepMirTar` (Wen et al., 2018), based on a stacked de-noising auto-encoder (SdA); `miRBind` (Klimentová et al., 2022) and `TargetNet` (Min et al., 2022), both using residual neural networks (ResNet); several convolutional neural network (CNN) models including `TargetScanCnn` (McGeary et al., 2019), `CnnMirTarget` (Zheng et al., 2020), and another unnamed CNN model reported by Hejret et al. (2023) (referred to as the `miRNA_CNN` throughout this work); and more recently, the `InteractionAwareModel` (T.-H. Yang et al., 2024) which uses a deep multi-head mutual attention network.

`DeepMirTar`, similar to conventional ML-based tools, uses 750 hand-crafted features that represent seed match type, miRNA pairing position and count, sequence composition, free energy, site accessibility, and conservation. In addition to these, it incorporates raw sequence-level features, including one–hot encoding. This method converts each of the four nucleotides into a 4-D binary vector. This representation preserves both nucleotide identity and position in a format suitable for machine learning models.

Other methods, including `miRAW`, `InteractionAwareModel`, and `CnnMirTarget`, are trained exclusively on representations derived from raw interacting sequences. In `miRAW` and the `InteractionAwareModel`, the miRNA and target site sequences are independently one–hot encoded into binary vectors and concatenated. In `CnnMirTarget`, intact miRNA–target site chimeric sequences are padded to a fixed length of 110 using randomly generated bases that exclude any 4-mers complementary to the miRNA. The resulting sequence is then one–hot encoded.

Some methods leverage low-level features such as alignment and/or Watson–Crick base pairing to strengthen pairing signal representations. `miRLSTM` and `TargetNet` involve entire or partial alignment of the miRNA and target site sequences. In `miRLSTM`, the miRNA and target site sequences are aligned, using a global complementary alignment algorithm, then Watson–Crick base pairs and any other base pair combination are represented by a, b, c, d and q, respectively, producing a single duplex sequence, which is then converted to alphabetical integer indices and used to produce an embedded vector. In `TargetNet`, extended seed regions in the miRNA are aligned to the target site. These aligned subsequences are substituted into the original sequences, which are then one–hot encoded into 5-D $(A, U, C, G, -)$ vectors. The two vectors are position-wise

concatenated and padded to a final 10×50 matrix.

`TargetScanCnn`, `miRBind`, and `miRNA_CNN` explicitly model Watson–Crick base-pairing. In `TargetScanCnn`, the first 10 nucleotides of the miRNA and 12-nucleotide target site sequences are encoded into a $10 \times 12 \times 16$ matrix, where the final dimension encodes all possible dinucleotide combinations (4^2) across the two sequences.

`miRBind` and `miRNA_CNN` are trained on a common data representation first introduced by Klimentová et al. (2022). Each miRNA–target site interaction is encoded as a 2-D matrix representing potential Watson–Crick base-pairing between the sequences. miRNA and target site sequences are truncated or padded to fixed lengths of 20 and 50 nucleotides, respectively, to ensure consistent input dimensions. Each matrix element indicates the presence (1) or absence (0) of base pairing between corresponding nucleotide positions.

`TargetScanCnn`, `CnnMirTarget`, `TargetNet`, `miRBind`, `miRNA_CNN`, and the `InteractionAwareModel` are included in the benchmarking effort presented in this study, described further in Chapter 3, Section 3.3. Benchmarking of miRNA target site prediction methods is discussed further in Section 2.6.1.

These approaches highlight the diversity of deep learning architectures and representations used for miRNA target site prediction, with convolutional neural networks (CNNs) emerging as a particularly effective class.

2.5.2 | Convolutional Neural Networks

Since the early 2000s, CNNs have excelled at image detection, segmentation, and recognition tasks, largely due to the availability of vast amounts of labeled data (LeCun et al., 2015). The 2012 ImageNet competition (Krizhevsky et al., 2012) reignited interest by showing that *deep CNNs*, trained on a million web-sourced images spanning 1,000 classes, reduced error rates by nearly 50% compared to the previous state of the art, and lead to the rapid adoption of deep learning by the computer vision community (LeCun et al., 2015).

CNNs transform raw inputs into informative representations using four core components: convolution, padding, stride, and pooling (see Figure 2.5). Convolution applies a learnable kernel across the input array, computing local dot-products to produce feature maps, analogous to automated feature extraction. Padding (typically with zeros) preserves border information, while the stride controls the kernel's step size, balancing detection detail against computational cost. Pooling, such as max or average pooling, aggregates values over small windows to downsample feature maps, reduce redundancy, and mitigate overfitting (Li et al., 2022).

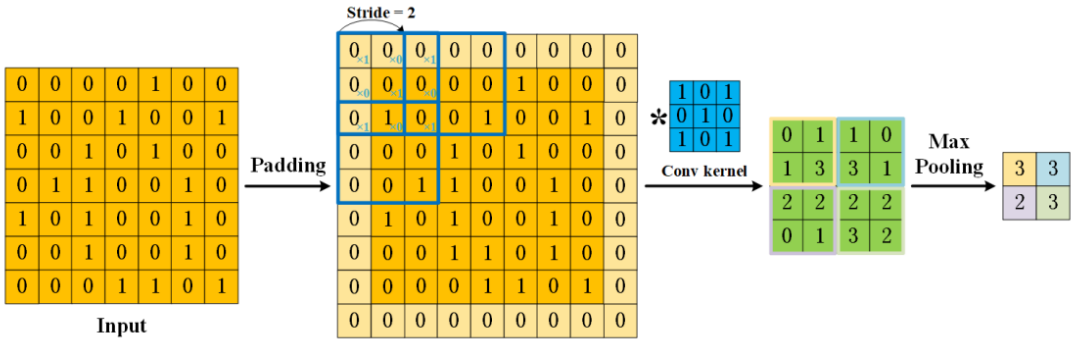

Figure 2.5: Illustration of the four main components of a convolutional neural network (CNN): convolution, padding, stride and pooling operations for a 2-D input. Reproduced from Li et al. (2022), © 2022 IEEE.

Many have adopted CNNs for miRNA target site prediction, applying image-like representations to encode miRNA–target site interactions, as described in Section 2.5.1.

The `miRNA_CNN` developed by Hejret et al. (2023) is particularly relevant to this work. It was trained on AGO2 CLASH miRNA–target site interactions generated by the authors. As outlined in Section 2.5.1, each interaction is encoded into a two-dimensional binary matrix indicating the presence or absence of Watson–Crick base pairing across miRNA and target site sequences, effectively forming a black-and-white image (Figure 2.6, *INPUT*).

The architecture, illustrated in Figure 2.6, consists of an input layer; followed by six convolutional blocks, each comprising a 2D convolutional layer used to detect local spatial patterns in the input, employing filters of size 5×5 with the number of filters linearly increasing across the six convolutional blocks; leaky rectified linear unit (Leaky ReLU) activation which is a non-linear function that allows small gradients when inactive to avoid dead neurons; batch normalisation which stabilises and accelerates training by normalising activations; max pooling which downsamples feature maps by retaining only the most prominent signals; and a dropout layer with a dropout rate of 0.3 to reduce overfitting by randomly deactivating a subset of neurons during training.

After the convolutional blocks, the output is flattened into a one-dimensional vector to transition from spatial to vector-based processing, and passed through two fully connected (dense) layers, which integrate information across the entire input. Each dense layer is coupled with Leaky ReLU activation, batch normalisation, and dropout. The final output layer consists of a single neuron with a sigmoid activation function, which maps the output to a probability score indicating the likelihood of miRNA–target site binding for each input example.

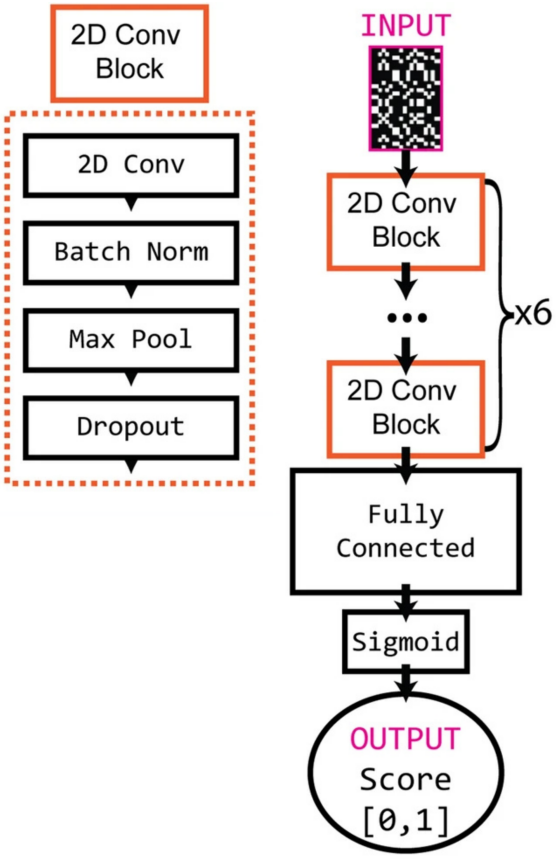

Figure 2.6: An illustration of the architecture of the CNN implemented by Hejret et al. (2023). In this study, the same configuration is retrained on a bias-corrected version of the original Hejret dataset and on the newly curated unbiased Manakov2022 dataset, presented in this work. Adapted from Hejret et al. (2023).

2.5.3 | The Negative Class

Binary classification is a supervised learning task. This means that the model is trained on labeled examples. In binary classification, examples are assigned one of two possible labels. In miRNA target site prediction, the model is trained to classify whether a given pair of miRNA–target site sequences represents a binding or non-binding interaction. Binding and non-binding interactions, or positive and negative examples respectively, correspond to the binary categories or labels. For this task, miRNA–target site interactions derived from experiment serve as the positive class.

The negative class plays a critical role in enabling ML models to learn meaningful discriminative patterns that define the decision boundary between the classes. It has been shown that one-class classification models trained solely on positive interac-

tions perform poorly, tending to classify most negative interactions as positives (Cohen-Davidi & Veksler-Lublinsky, 2024). This highlights the necessity to include negative examples for effective miRNA target site prediction.

Another important consideration in this context is class imbalance. Given the vast number of non-binding sites across the transcriptome compared to the relatively limited number of true binding sites for a specific miRNA, the underlying biological reality is inherently imbalanced. Accordingly, adopting training strategies that reflect this imbalance, such as the use of datasets that are highly imbalanced towards the negative class, can better approximate realistic biological conditions and improve model generalisability.

However, publicly available experimentally validated negative examples are scarce. To enable effective training, negative examples must be computationally generated to maintain at least a balanced ratio of positive to negative examples.

2.5.3.1 | In silico Approaches to Generate the Negative Class

In the absence of sufficient experimentally validated negative examples, a variety of strategies have been proposed to simulate them.

One simple approach is the use of mock miRNAs which primarily involves shuffling the real miRNA sequence to create an artificial miRNA. This approach is implemented by Paker and Ogul (2019) and Wen et al. (2018) for `miRLSTM` and `DeepMirTar`, respectively.

For `miRAW`, Pla et al. (2018) leverage the negative experimentally validated miRNA–mRNA interactions from the DIANA-TarBase database (Vlachos et al., 2014). The authors consider all possible 30-nucleotide subsequences from the 3′ UTR of the mRNA sequences of these negative interactions. They model the secondary structure of the resulting RNA duplexes with RNACofold (Lorenz et al., 2011), and if the duplex has a negative binding energy, an indicator of stability, it is accepted as a negative example. Since Min et al. (2022) leverage the miRAW datasets for `TargetNet`, they therefore also rely on this approach for their negative examples.

For the `InteractionAwareModel` and `CnnMirTarget`, the authors leverage target sites from transcripts sampled from the genome, that are different to those associated to the respective miRNA in the positive class (T.-H. Yang et al., 2024; Zheng et al., 2020). This method may introduce bias due to the underlying sequence composition of different genomic features such as introns, exons, untranslated regions, repeat elements, *etc*, which are difficult to control for.

Conversely, for `TarPmiR`, Ding et al. (2016) search for alternative, non-overlapping binding regions on the same transcript, referred to as non-positive sites, with selection based on thermodynamic stability or random sampling.

For `chimiRic`, Lu and Leslie (2016) pair miRNAs to AGO-bound RNA fragments recovered from independent CLIP experiments, excluding combinations recovered as chimeras in the first AGO1 CLASH experiment. Such a strategy dismisses any potential affects that may derive from experiment- or dataset- specific characteristics.

On the other hand, Hejret et al. (2023), Klimentová et al. (2022), and Wang (2016) generate new pairs of miRNA and target site sequences recovered from the same CLIP- or CLASH- experiment, excluding the recovered positive interaction combinations, for `miRNA_CNN`, `miRBind`, and `miRTarget`, respectively.

Notably, McGeary et al. (2019) model the task of miRNA target site prediction as a supervised *regression* problem in `TargetScanCnn`. The examples are each labeled with a *negative* continuous value for the dissociation constant (K_d), which serves as a proxy for the binding affinity of the RNA duplex. A low $-\log(K_d)$ value indicates low stability. Examples associated with a low $-\log(K_d)$ value represent low affinity miRNA target sites. These are analogous to the negative examples in a binary classification task. For this reason, negative examples did not need to be generated to train the `TargetScanCnn` supervised regression model.

While various strategies exist for generating negative data, each introduces specific biases that influence model behaviour. As shown by Cohen-Davidi and Veksler-Lublinsky (2024), these biases can limit generalisation, with models often performing poorly when evaluated on negatives constructed using alternative approaches.

This underscores the critical importance of carefully considering how negative data is generated when designing, evaluating, and comparing miRNA target site prediction models. Exploratory data analysis of the final dataset is therefore an important step to ensure no critical biases are present, that could lead to models learning artifacts of the data rather than biologically meaningful patterns.

2.6 | Benchmarking

Benchmarking is a structured approach that involves the evaluation and comparison of the performance of different methods or tools against a defined set of criteria or datasets. By establishing a standardised framework for comparison, benchmarking enables researchers to identify the strengths and weaknesses of various approaches, and pinpoint areas that need improvement, to guide the advancement of more effective and reliable methods.

In machine learning and, increasingly, bioinformatics, benchmarks have become key drivers for advancement based on rigorous, reproducible comparisons. For instance, in computer vision, benchmarks such as ImageNet have been pivotal in advancing the field by providing a large-scale, annotated dataset that serves as a standardised reference for model evaluation (Deng et al., 2009). Likewise, the Critical Assessment of Protein Structure Prediction (CASP) has set a benchmark for evaluating protein structure prediction methods, leading to significant improvements in computational model accuracy, including the development of AlphaFold, which achieved near-experimental precision (Jumper et al., 2021), contributing to the Nobel Prize in Chemistry 2024 for advancements in protein structure prediction. Several other benchmarking initiatives have emerged in bioinformatics, including Genomic Benchmarks for benchmarking genomic sequence classification methods (Grešová, Martinek, et al., 2023), GUANinE for genomic sequence-to-function models (Robson & Ioannidis, 2024), BEND for DNA language models (Marin et al., 2023), BEACON for RNA tasks and language models (Ren et al., 2024), and GenBench for genomic foundation models (Z. Liu et al., 2024).

Creating a high-quality benchmark requires: (i) extensive domain expertise to minimise biases; (ii) adherence to the well-known FAIR principles that provide guidance on data findability, accessibility, interoperability, and reusability (Wilkinson et al., 2016), and; (iii) compliance with the DOME recommendations for data, optimisation, model, and evaluation in supervised machine learning validation in biology (Walsh et al., 2021). As machine learning (ML) has become a common approach in computational biology, the DOME recommendations were developed to help non-specialists apply ML methods more effectively by highlighting common pitfalls, their potential consequences, and practical strategies to avoid them.

One of the key components of a benchmark for ML approaches is the dataset, which consists of training, validation, and testing subsets. The creation of these independent subsets, and their deposition in long-term repositories for community use and validation, are among the DOME recommendations most relevant to benchmarking. During the model development process, the dataset is divided into these subsets: the training

set is used for model training, the validation set is used for evaluation throughout the model development process such as hyperparameter tuning, and the testing dataset, which remains unseen during training and validation, is used for evaluation of the final model. By evaluating the model on an unseen test set, benchmarking provides an unbiased measure of the model's performance and generalisability. Comparing this estimate to that of other published models, as well as to simple baseline models evaluated on the same unseen test set, also aligns with the DOME recommendations related to benchmarking.

2.6.1 | Benchmarking miRNA Target Site Prediction Methods

No standardised framework currently exists for benchmarking miRNA target site prediction methods. Popular databases in the field include DIANA-TarBase (Skoufos et al., 2024), and MirTarBase (Cui et al., 2025). These are miRNA–target *gene* interaction databases that include experimentally validated interactions, curated either manually or semi-automatedly from literature of studies performing experimental validation. Multiple versions of each database are available.

Generally, miRNA target genes are first computationally predicted to narrow down candidate interactions, which are then subjected to experimental validation. As a result, any bias present in the prediction method is propagated into the experimentally validated interactions, underscoring the need for accurate and unbiased prediction models.

Since miRNA target *site* prediction is typically the first step in many gene-level prediction programs, reliance on rules or features derived from prior knowledge leads to experimentally validated targets that are skewed towards well-established miRNA binding rules, limiting the capture of the full spectrum of targets. In turn, the use of these experimentally validated interaction datasets to train miRNA target site prediction models creates a feedback loop, restricting the advancement of new biological knowledge about miRNA binding rules. A prominent example of such rules or features is the overreliance on seed measures, which has led to a strong bias in experimentally validated targets towards canonical seed-based interactions.

It is therefore essential to train and benchmark unbiased models for miRNA target site prediction. As discussed in Sections 2.3 and 2.5.1, the availability of high throughput chimeric data has enabled the development of deep learning models trained directly on *raw* miRNA–target site sequences, avoiding introduction of any preconceptions about miRNA–target site binding. Several datasets have been curated for this purpose. In what follows, we review the datasets and benchmarking strategies employed by the models most relevant to this work.

Several models are trained on data derived from a single experiment. `miRBind` is trained on the AGO1 CLASH data from Helwak et al. (2013), the first reported chimeric experiment that introduced a ligation step to capture RNA–RNA interactions mediated by the AGO1 protein. The chimeric interactions were detected using the HybriDetector pipeline described in Section 2.4.1. The model is evaluated on a randomly held-out subset of the AGO1 CLASH data, and on a smaller chimeric eCLIP test set (Klimentová et al., 2022).

The `InteractionAwareModel` (T.-H. Yang et al., 2024) is trained on a re-analysed version of the AGO1 CLASH raw data using a different chimeric interactions detection pipeline. Evaluation is performed on a held-out test set containing interactions whose target sites are located on chromosomes 12 and X.

`miRNA_CNN` is trained on newly generated CLASH data using the AGO2 protein (Hejret et al., 2023), and applies the same chimeric interaction detection pipeline as `miRBind`. It is evaluated on a held-out test set containing interactions whose target sites are located on chromosome 1.

`TargetScanCnn` is trained on data obtained using a novel experimental technique called RNA Bind-n-Seq (RBNS) (McGeary et al., 2019). This technique measures relative binding affinities between 10-nucleotide *k*-mers derived from specific miRNAs and all possible 12-mer target site sequences. This experiment is biased towards short-range interactions and thus may favour binding in the seed region. Evaluation is performed across different miRNAs.

Other models are trained on datasets compiled from multiple sources. `CnnMirTarget` is trained on a composite dataset derived using the AGO1 CLASH data (Helwak et al., 2013) merged with *C. elegans* and mammalian iPAR-CLIP data (Grosswendt et al., 2014), and high-confidence experimentally validated interactions from MirTarBase (Hsu et al., 2010). Evaluation is performed on a randomly held-out subset of the combined dataset.

`miRAW` is trained using the miRNA–mRNA interactions derived from DIANA-TarBase (Vlachos et al., 2014) and MirTarBase (Chou et al., 2015), that are cross-referenced with the miRNA database, miRBase version 21 (Kozomara & Griffiths-Jones, 2014), and Ensembl version 87 (Aken et al., 2016). Candidate binding sites are identified using iPAR-CLIP and AGO1 CLASH datasets, with additional conserved sites incorporated from TargetScanHuman 7.1 (Agarwal et al., 2015). The test set is a held-out subset that is heavily imbalanced (97% positive class); a balanced test set is obtained through random subsampling. `TargetNet` uses the same `miRAW` datasets for training and testing. Additionally, they produce a test set that is composed of interactions involving miRNAs that are exclusive to the test set and unseen during training.

These examples demonstrate the wide range of data sources used to train and benchmark miRNA target site prediction models. In addition to source diversity, there is considerable variation in dataset size, ranging from hundreds to tens of thousands of examples; as well as in strategies for generating negative examples and class balance (as discussed in Section 2.5.3); and the choice of methods compared during benchmarking. These factors make model performance comparisons challenging.

Despite substantial efforts to develop more accurate models, benchmarking practices remain inconsistent and insufficiently standardised. This undermines the validity of performance claims and makes it difficult to assess genuine improvements. As a result, researchers developing miRNA target gene prediction models face uncertainty when selecting a suitable site-level predictor. In turn, researchers relying on these gene-level models to guide experimental validation and investigate miRNA regulatory functions are also impacted. These limitations underscore the need for a standardised benchmarking framework to enable fair, reliable model evaluations and to support downstream research that advances understanding of miRNA-mediated regulation.

2.7 | Summary

In this chapter, we provide an overview of the topics and theoretical concepts relevant to this work. We introduce the biology of miRNAs, focusing on their biogenesis, regulatory role, and emerging clinical relevance. We then present the field of computational miRNA target prediction, which offers an alternative to the experimental validation of all potential targets, and we emphasise its reliance on accurate miRNA target *site* prediction models. We outline traditional rule-based methods and their limitations. We then describe how evolving high-throughput experimental techniques, based on ligation of AGO-mediated interacting RNA molecules, have enabled the direct detection of in vitro miRNA–target site interactions. We explain how these techniques motivate the development of chimeric interaction detection pipelines, such as HybriDetector, which produce structured datasets of interacting sequence pairs. We note that such datasets facilitate data-driven deep learning approaches in miRNA target site prediction. We summarise the interaction representation strategies used in such models to represent the interacting raw sequences, and we highlight a CNN model particularly relevant to this study. We also describe various strategies for generating negative examples, necessary due to the scarcity of experimentally confirmed non-binding pairs. Finally, we discuss benchmarking practices in the field and identify their lack of standardisation as a key limitation that motivates this work.

3

Methodology

In this chapter, we present a comprehensive account of the methodology adopted throughout this study to fulfill our aims and objectives. We begin with a schematic overview of the key components of the workflow, followed by detailed descriptions of each step. All processes are documented with sufficient detail to ensure full reproducibility, including the software tools and version numbers used. Where relevant, we refer to other chapters for contextual information or supporting results, and introduce theoretical concepts when necessary. We also include a final section outlining the evaluation metrics used to assess model performance across this study. The chapter concludes with notes on code and data availability, followed by a concise summary of the work carried out.

3.1 | Overview

This work comprises four main components, as illustrated in Figure 3.1. Each is briefly outlined below and described in detail in the subsequent sections.

1. **Dataset Curation:** First, we leverage raw read files from several chimeric eCLIP experiments (Manakov et al., 2022), and via a series of processing steps, we produce the positive class for a benchmarking dataset. We then reproduce an established in silico method to generate the negative class and identify an associated bias (Hejret et al., 2023; Klimentová et al., 2022). Accordingly, we develop a new method for generating negatives that corrects this bias, and apply it to construct the negative class of the new benchmarking dataset and to correct existing published datasets. Finally, we split the datasets into train and test sets, and ensure all datasets comply with FAIR (Findable, Accessible, Interoperable, and Reusable) principles. The final datasets are contributed to the `miRBench` Python package for miRNA target site prediction benchmarks (Sammut et al., 2025b).

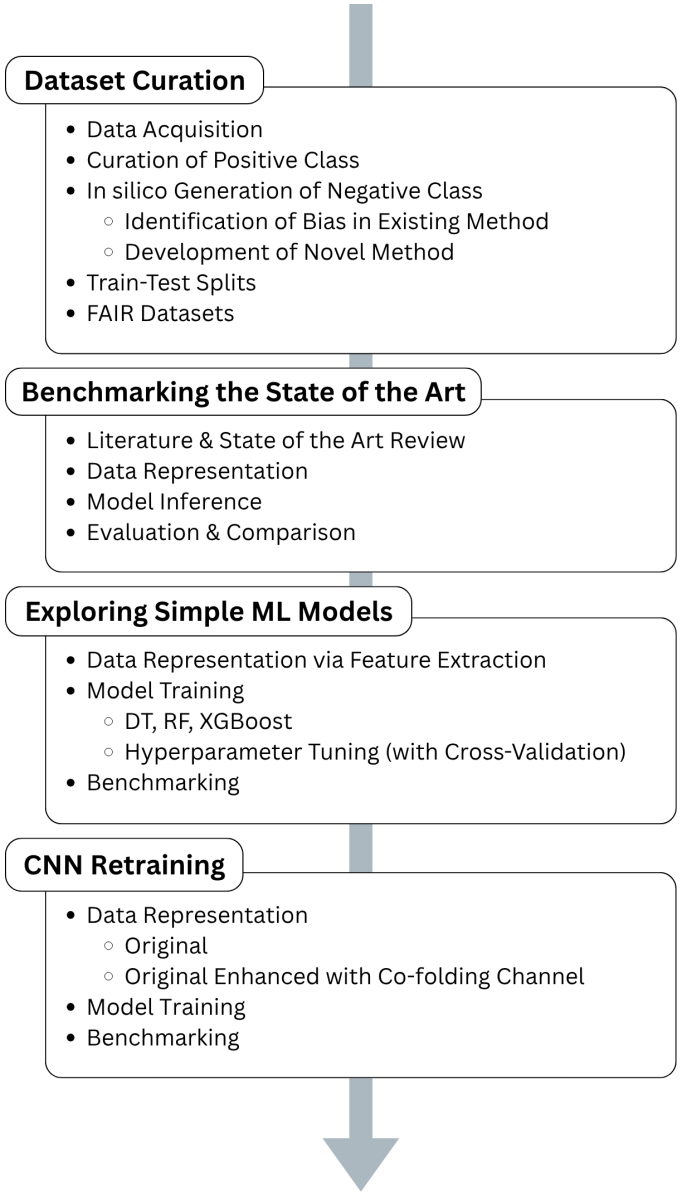

Figure 3.1: A flowchart of the workflow implemented in this study, comprising four main components.

2. **Benchmarking the State of the Art:** We first review the literature and identify six relevant state-of-the-art deep learning models (Hejret et al., 2023; Klimentová et al., 2022; McGeary et al., 2019; Min et al., 2022; T.-H. Yang et al., 2024; Zheng et al., 2020). We implement the corresponding input data representations to encode our benchmarking datasets. We run inference on these datasets, and evaluate the predictions to produce comparable performance metrics.

3. **Exploring Simple Machine Learning (ML) Models:** First, we design a feature set based on the miRNA–target site sequence pairs in our dataset, and we explore tree-based classifiers. We perform hyperparameter tuning with cross-validation and select the best configuration for each model type. Finally, we train the final model configurations on the full train set, and benchmark the models.

4. **Retraining of State-of-the-Art Convolutional Neural Network (CNN):** Using our bias-corrected datasets, we retrain a published CNN architecture that was originally trained on a biased dataset (Hejret et al., 2023). To do so, we first encode the data into the input representation expected by the model. We then train and benchmark the model. Additionally, we explore a data representation that incorporates co-folding information.

3.2 | Curation of the Datasets

Most relevant to this work are the published Hejret et al. (2023) and Klimentová et al. (2022) datasets, summarised in Table 3.1. Building on these, we curated comparable datasets by leveraging newly available raw miRNA–target site interaction data from Manakov et al. (2022).

Table 3.1: Summary of Original Published Datasets.

Dataset & Source	Split	No. of Pos.	No. of Neg.	Total	Pos:Neg Ratio
Hejret (Hejret et al., 2023)	Train	3,860	38,600	42,460	1 : 10.0
	Test	719	719	1,438	1 : 1.0
	Test	719	7,190	7,909	1 : 10.0
	Test	719	71,900	72,619	1 : 100.0
Klimentova (Klimentová et al., 2022)	Test	477	477	954	1 : 1.0
	Test	477	4,770	5,247	1 : 10.0
	Test	477	47,676	48,153	1 : 99.9

As described in Chapter 2, Section 2.3, the chimeric eCLIP method produces chimeric reads, among other reads. These are sequencing reads that contain both the miRNA and its target site joined together owing to the ligation step in the experimental method. Vast raw data produced via this novel technique provides an opportunity to construct a structured dataset of miRNA–target site interactions, that is useful for training and benchmarking machine learning and deep learning models on the binary classification task of miRNA target site prediction.

Through a series of processing steps (see Figure 3.2), the data was transformed into a structured format comprising multiple columns, including the miRNA sequence, the target site sequence, and the corresponding binary interaction label.

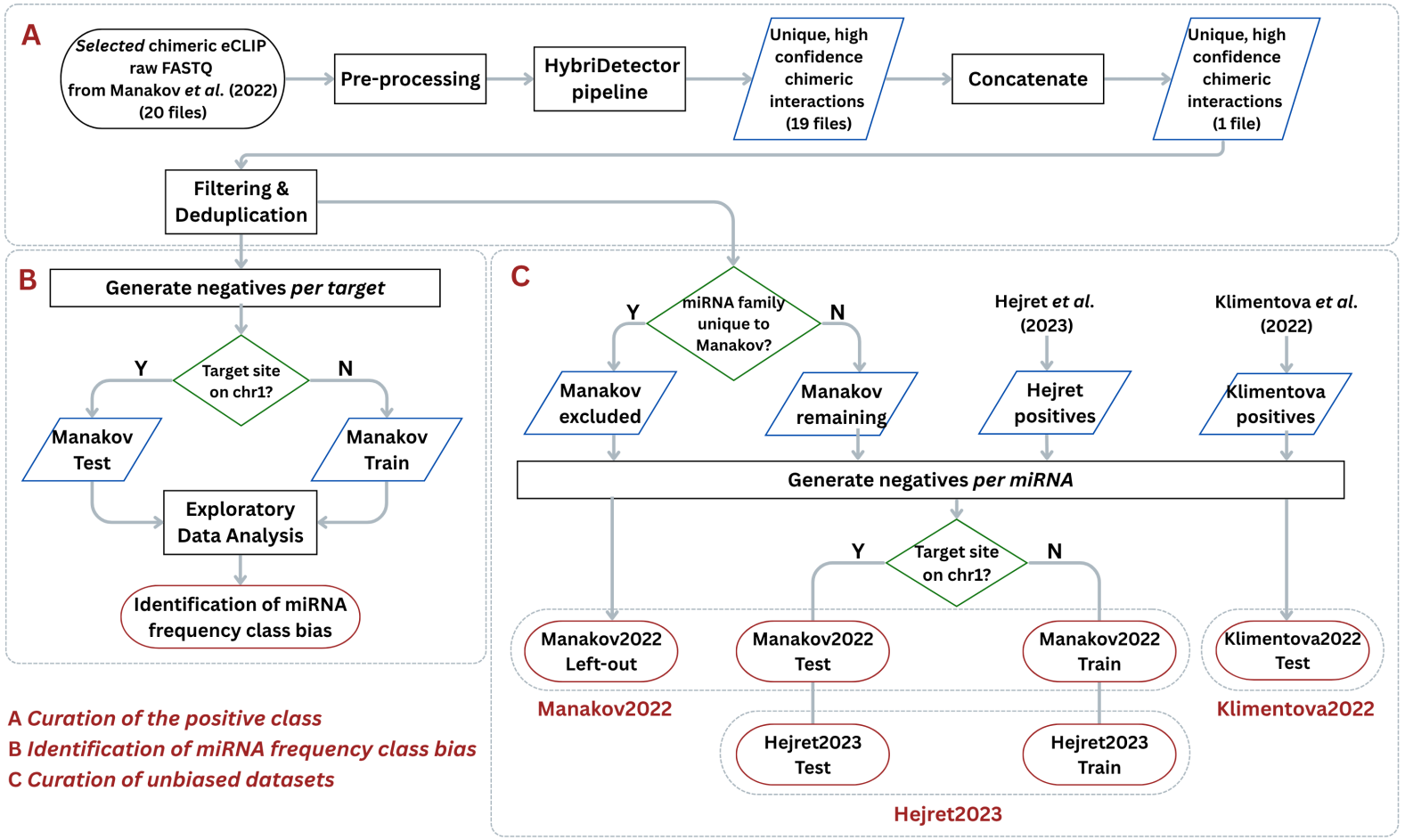

Figure 3.2: A flowchart of the entire dataset curation process: from the raw FASTQ files from the chimeric eCLIP experiment, to the three final dataset collections presented in this work.

3.2.1 | Data Acquisition

We identified a resource of publicly available chimeric eCLIP data produced from experiments performed by Manakov et al. (2022). The data is deposited on the Gene Expression Omnibus (GEO) on the National Center for Biotechnology Information (NCBI) platform, under GEO ID GSE198251, SubSeries GSE198250. 188 samples are available for download.

The data originates from various experiments performed in Human Embryonic Kidney (HEK) 293T cells, rat C9 cells, mouse liver tissue, and a mixture of HEK293T and C9 cells. We selected samples from the human cell line only, since our study aims to elucidate miRNA–target site interactions in a human cellular context, minimising variability from interspecies differences.

Some of the experiments involve enrichment of a miRNA or gene of interest, whereas others capture the total chimeric interactions. We selected samples from the total chimeric interactions experiments only, since these provide an unbiased overview of the full spectrum of miRNA–target interactions. This approach enables us to derive general binding patterns that apply across all miRNAs and target sites.

Finally, Manakov et al. (2022) performed the chimeric eCLIP methodology in two ways; with and without the polyacrylamide gel and nitrocellulose membrane transfer step. This step is a standard feature of CLIP technology used for size-selective purification of RBP–RNA complexes and for reducing the co-purification of non-crosslinked RNA, as the latter binds weakly to nitrocellulose and is therefore largely washed away during the membrane purification steps (Hafner et al., 2021). However, Manakov et al. (2022) have shown that the elimination of this cumbersome step improves the recovery of miRNA to mRNA interactions. We opted to include both types of experiments.

These criteria narrowed down the selection to a total of 20 samples. These samples were selected using the GEOparse Python package (version 2.0.4), and downloaded using the enaBrowserTools GitHub repository (release v1.7.1)[1].

Some experiments used a single-end library layout, where only one end of each fragment is sequenced, while others employed a paired-end layout, which sequences both ends of each fragment. For the paired-end samples, 151 bp were sequenced from each end. Since this read length is sufficient to capture the entire chimeric read, which consists of a ∼22-nucleotide (nt) miRNA and a target site that is assumed to be no longer than 50 nucleotides throughout this study, only the R1 read files were used. The sequenced reads from each sample were stored in a FASTQ file, which is a text-based format that includes both the nucleotide sequences and their corresponding quality scores.

[1] `https://github.com/enasequence/enaBrowserTools`. Last accessed May 2024.

This process resulted in one FASTQ file per sample for a total of 20 samples, for subsequent processing.

3.2.2 | Curation of the Positive Class

Once the chimeric eCLIP raw FASTQ files were chosen and downloaded, they required processing to extract interacting miRNA–target site pairs of sequences that would serve as the positive class for the task of miRNA target site prediction. This involved multiple steps (see Figure 3.2A), including UMI extraction, adapter trimming, and downstream processing using an adaptation of HybriDetector to handle larger files. Additional steps such as concatenation, filtering for miRNA interactions, and deduplication followed. These steps are described in detail in the subsections that follow.

3.2.2.1 | UMI Extraction and Adapter Trimming

In this section, we mirror the processing suggested by Manakov et al. (2022), specifically the first part of the workflow for total chimeric eCLIP available on the chim-eCLIP GitHub repository from YeoLab[2].

For each of the downloaded files, the 10-nt 5′ unique molecular identifiers (UMIs) were extracted and appended to the read names, then trimmed from the sequences using umi-tools (version 1.1.4). UMIs are short, random nucleotide sequences that serve as molecular barcodes to uniquely label each original RNA molecule. This enables us to distinguish true biological reads from duplicates. PCR amplification is a method used to generate multiple copies of nucleic acid sequences for sequencing. This can artificially inflate read counts. Therefore, umi-tools can identify and remove these PCR-generated duplicates, ensuring more accurate downstream analyses.

Next, the 3′ adapters were trimmed with Cutadapt (version 2.8) to remove any synthetic sequences introduced during library preparation. Adapters are short, predetermined sequences ligated to RNA fragments to facilitate sequencing. However, they must be removed before downstream analysis to prevent interference with the biological signals. Cutadapt is a tool specifically designed to identify and remove such adapter sequences from high-throughput sequencing data. It was also used to trim low-quality 3′ ends and discard reads shorter than 18 nucleotides, ensuring that only high-quality relevant reads were retained for further processing.

Finally, another 10 nucleotides were trimmed from the 3′ end, again using Cutadapt (version 2.8), to remove any residual random sequences specific to the chimeric eCLIP

[2] `https://github.com/YeoLab/chim-eCLIP`. Last accessed June 2024.

experimental design, thereby ensuring accurate representation of the insert sequences in the final dataset.

3.2.2.2 | HybriDetector Processing and Adaptations for Larger Files

HybriDetector[3] is a pipeline for detecting AGO-mediated chimeric interactions between small non-coding RNAs and their target sites from CLASH and similar experiments, described by Hejret et al. (2023).

The input to the HybriDetector pipeline is a file in FASTQ format containing reads from sequencing libraries produced from such experiments. Through a series of alignment steps against the human genome and several small RNA databases, sequencing reads are classified into three distinct categories: (i) reads that align exclusively to the genome, (ii) reads that align to a small RNA database, including miRNAs, tRNAs, rRNAs, snoRNAs, YRNAs, or vaultRNAs, and (iii) reads that exhibit partial alignment to a small RNA database, with the remainder of the read aligning to the genome (see Figure 2.4).

Category (iii), referred to as chimeric reads, represent instances where a small RNA sequence is directly linked to a genomic target site. These chimeric reads serve as the basis for extracting small non-coding RNA–target site sequence pairs.

These reads are deduplicated based on their UMIs and subjected to filtering using multiple criteria, including stringent alignment requirements that permit little to no mismatches (Hejret et al., 2023). Chimeric reads that align to the same small RNA sequence and the same target sequence, allowing for minor discrepancies, are subsequently collapsed into a representative chimeric interaction. This process results in the final dataset consisting of a curated list of small non-coding RNA–target site sequence pairs. The HybriDetector pipeline is described in detail in the Background and Literature Overview in Chapter 2, Section 2.4.1.

Following UMI extraction and adapter trimming as described in Section 3.2.2.1, the sample files were further processed by the HybriDetector pipeline. It was run with default parameters, except for `--read_length`, which was set to 151 to match the sequencing read length, and `--is_umi`, which was set to 'True' to indicate that the library already contains UMIs extracted to the read headers (since HybriDetector also includes an optional utility for UMI sequence extraction).

As we attempted to process the FASTQ files, adaptations needed to be made to the pipeline to be able to handle some of these larger files (ranging 243 MB to 6 GB). In

[3] `https://github.com/ML-Bioinfo-CEITEC/HybriDetector`. Last accessed December 2024.

the final steps of the pipeline, when potential chimeric reads are filtered and collapsed, sequence similarity clustering is performed on the small RNA sequences before they are collapsed by the identified clusters (Hejret et al., 2023). This is done by first computing a distance matrix, corresponding to the edit or Levenshtein distance between all possible pairs of sequences. This distance corresponds to the number of insertions, deletions, or substitutions needed to transform one sequence into another. Each of the distances in the matrix are then normalised by corresponding average length of the pair of sequences compared. Finally hierarchical clustering is performed on the normalised distances and the tree is cut at a specified height to define the clusters, which are assigned to each sequence for further processing.

For large files, the first step may compute too large a distance matrix that exceeds the allocated memory usage limit. To mitigate this, an additional clustering step was performed prior to the distance matrix computation, in order to produce smaller subsets of the dataset that can be independently processed more efficiently.

A k-mer frequency matrix for $k = 4$ was computed for the small RNA sequences. A k-mer frequency matrix is a numeric array in which each row corresponds to an individual RNA sequence and each column represents one of the 4^4 possible k-mers (i.e., nucleotide strings of length 4). The entries in the matrix indicate the frequency with which each specific 4-mer occurs in the sequence. This computation was performed in R, consistent with the programming language used for this part of the HybriDetector pipeline, using the `DNAString` and `oligonucleotideFrequency` functions from the Biostrings package, which is already in use by HybriDetector.

K-means clustering was then performed on these features using 100 centers. K-means is an unsupervised machine learning algorithm that partitions data into a predefined number of clusters by iteratively assigning each data point to the nearest cluster center and updating the centers to minimise within-cluster variance. In our context, setting the number of centers to 100 grouped similar small RNA sequences based on their k-mer frequency profiles, thereby reducing the complexity of subsequent distance matrix computations.

This process assigned a cluster ID to each small RNA (or row), generating subsets of similar sequences. The original sequence similarity clustering was then applied to each subset individually, and the outputs of these clustering operations were vertically concatenated to produce a single data table. This data table was then used to continue with the remainder of the HybriDetector pipeline.

The described modifications were added to a branch on the public HybriDetector

GitHub repository[4]. Despite these modifications, one of the 20 sample files remained too large to process even with up to 2,000 centres for *k*-means clustering, and was finally abandoned, such that only 19 of the 20 selected sample files were successfully processed by HybriDetector. Further improvement of the HybriDetector pipeline was deemed as out of scope of this work.

3.2.2.3 | Post-processing of Chimeras Detected by HybriDetector

The HybriDetector pipeline produces tab-separated values (TSV) output files containing a 'noncodingRNA_seq' column, corresponding to various types of small RNA sequences, and a 'seq.g' column, corresponding to a target site sequence, among other descriptive columns.

Two branches of output files are produced based on the uniqueness of the alignment of the small RNA part of the chimeric read: the 'unique' and 'ambiguous' branches. The 'unique' branch contains chimeric interactions with small RNAs that have been mapped to a single small RNA. Conversely, the 'ambiguous' branch contains the rest of chimeric interactions. We opted to use the more accurate 'unique' branch of output files.

Additionally, for each branch, two main output files were produced, one containing all chimeric interactions detected, and another filtered 'high confidence' version containing chimeric interactions whose alignment of the small RNA sequence contains no mismatches. We again opted for the latter, more accurate output files. These were deposited on the data sharing platform, Zenodo[5].

The 'unique', 'high confidence' output files of each of the 19 successfully processed samples were vertically concatenated to produce a single file with binding small RNAs–target site sequences, corresponding to the previously published HybriDetector output datasets from Klimentová et al. (2022) and Hejret et al. (2023)[6,7]. This new dataset was also made publicly available on Zenodo[8].

Furthermore, the concatenated file was filtered for miRNAs via the column containing annotations for small RNA type, excluding all other non-coding RNA biotypes: tRNA, rRNA, snoRNA, vaultRNA and YRNA, from the dataset.

[4] `https://github.com/ML-Bioinfo-CEITEC/HybriDetector/tree/fix_clustering`. Last accessed May 2024.

[5] `https://zenodo.org/records/14730307`. Last accessed July 2025.

[6] `https://github.com/ML-Bioinfo-CEITEC/miRBind/blob/main/Datasets/AGO2_eCLIP_Klimentova22_full_dataset.tsv`. Last accessed March 2025.

[7] `https://github.com/ML-Bioinfo-CEITEC/HybriDetector/blob/main/ML/Datasets/AGO2_CLASH_Hejret2023_full_dataset.tsv`. Last accessed March 2025.

[8] `https://zenodo.org/records/14501607/files/AGO2_eCLIP_Manakov2022_full_dataset.tsv.gz`. Last accessed March 2025.

The file was then restructured by dropping some less useful columns and renaming existing ones. Importantly, the 'noncodingRNA_seq' column containing the miRNA sequences was renamed to 'noncodingRNA', and the 'seq.g' column containing the corresponding target site sequences was renamed to 'gene'.

A new column called 'label' was added and all rows were assigned the value '1', to denote that they are positive examples. Another new column called 'test' was added containing a boolean indicator of whether the target site sequence was located on chromosome 1, based on the column containing the chromosome annotation of the target site sequences. This new 'test' column was to be used further downstream for the train–test splitting strategy, as described in Section 3.2.4.

Finally, the rows were deduplicated based on the combination of the 'noncodingRNA' and 'gene' columns, removing redundancies hence ensuring unique positive examples. This is the final set of positive examples for what we term the *Manakov2022* dataset, containing 1,431,689 examples.

3.2.3 | In silico Generation of the Negative Class

Establishing a well-defined negative class for miRNA–target site interactions presents a significant challenge due to the scarcity of datasets containing experimentally validated non-binding sequence pairs. Due to this limitation, the negative class is simulated and derived from the positive class based on some considerations.

One key assumption is that within the set of identified miRNA sequences and the set of target site sequences, those that do not appear paired together as chimeric interactions in the dataset are presumed to exist within the same cellular environment but do not interact. We therefore assumed a strategy of mismatching the pairs of interactions in the positive class. This systematic pairing approach offers a structured and reproducible method for constructing the negative class while leveraging available data.

Another important consideration is class balance. In the cellular environment, a miRNA encounters numerous non-binding sequences before binding a target site, reflecting its selective and finely tuned regulatory function. To better approximate this biological reality, class-imbalanced datasets were constructed with varying positive-to-negative ratios: 1:1, 1:10, and 1:100. Among these, the 1:100 ratio is the most representative of the natural scenario, as it best captures the prevalence of non-binding interactions in relation to binding events.

The following steps outline the procedure for generating negative examples at different class imbalance ratios reproduced from the method implemented by Hejret et al.

(2023) and Klimentová et al. (2022). This method was later found to produce biased datasets, as described further in Section 3.2.5.

Prior to generating negative examples, the positive examples were alphabetically sorted by target site sequence to group rows with the same target site sequence into 'target site blocks'. The dataset was then processed one block at a time.

The criteria for pairing a miRNA sequence with a target site sequence to generate a negative example are as follows:

i The interaction does not already exist as a positive example.

ii The interaction is sufficiently different from interactions in the positive examples.

iii The interaction is unique in the negative examples.

For each unique miRNA sequence in the dataset, a set of sufficiently dissimilar miRNA sequences was precomputed. Levenshtein distance was used to quantify the sequence dissimilarity, ensuring that only miRNAs with an edit distance greater than 3 from the given miRNA were included in the set. This approach allowed for the systematic selection of alternative miRNAs while maintaining a minimum sequence dissimilarity threshold, thereby supporting the selection criteria (i) and (ii).

For each 'target site block' in the positive examples, the intersection of the sets of alternative miRNAs for the miRNAs in that block was computed. This approach allowed for the systematic pooling of miRNAs that may be paired to the common target site in the block, producing unique negative examples that are sufficiently distinct from the corresponding positive examples, thereby supporting selection criteria (ii) and (iii).

The number of negative examples generated for each block was determined by the number of positive examples within that block, the desired class imbalance ratio, and, if applicable, the unmet negative example quota from previous blocks due to criterion constraints. This number of miRNAs was sampled from the pool of suitable miRNAs computed in the previous step. Any remaining negative examples that could not be generated due to criterion constraints across all blocks (edge cases) were recorded in a log file, if applicable.

To mitigate potential inconsistencies, positive examples with discrepancies in feature annotation or chromosome number of the target site within a 'target site block' were excluded from the final dataset, and no negative examples were generated for them. If applicable, these exclusions were also documented in a log file.

The final rows for the negative examples were constructed to match the format of the positive examples. The 'label' column was assigned a boolean value of '0'. The neg-

ative examples were then vertically concatenated with the positive examples, yielding a dataset containing both classes in the selected class imbalance ratio.

3.2.4 | Train-Test Splits

In this work, the datasets were split into the train and test sets based on whether the target site sequence was mapped to chromosome 1. Target site sequences mapped to chromosome 1 were assigned to the test set, while those mapped to other chromosomes were assigned to the train set.

Compared to random splitting, this approach helps to ensure that models are trained and evaluated on sufficiently distinct miRNA–target site interactions, enabling assessment of the model's generisability to new, unseen genomic regions rather than memorisation of patterns from the same chromosomal context. Additionally, this chromosome-based splitting strategy consistently yielded an approximate 90:10 train–test ratio, which aligns with common best practices in machine learning.

This was carried out via the boolean value in the 'test' column. As a final step, the 'test' column was removed to eliminate redundant information.

3.2.5 | Identification of a miRNA Frequency Class Bias

Since negative examples were generated in silico, it is important to verify that no unintended biases were introduced between the two classes. To assess this, we examined several properties, including the distribution of miRNA sequences across the positive and negative classes. Ideally, these distributions should be similar and not exhibit obvious differences.

Despite this expectation, we observed differences in the frequency of miRNA sequences between the two classes, indicating a dataset artifact that could influence model training. In particular, models may learn to distinguish positive from negative examples based primarily on the miRNA sequence, rather than the underlying interaction features. This risks inflating performance metrics and misrepresenting the model's generalisability. We refer to this artifact as the *miRNA frequency class bias*.

To investigate this further, we hypothesised that the distribution of miRNA sequences alone could be sufficient to distinguish between positive and negative examples. Specifically, we tested whether a model trained only on miRNA-derived features, excluding any target site information, could classify interactions accurately.

We computed all possible k-mers for $k = 3$ for each miRNA sequence, generating a 3-mer frequency matrix, where each example was represented by a 64-dimensional

feature vector ($4^3 = 64$). This served as the input feature matrix for training a simple machine learning model aimed at isolating the effect of the bias.

We used a decision tree classifier trained on the input feature matrix and class labels from the train set. A decision tree classifier is an algorithm that uses a tree-like structure to split the feature space based on threshold values at each node, leading to a series of decisions that ultimately assign a class label to each example. The classifier was implemented in Python using the default parameters of the `DecisionTreeClassifier` function from the `sklearn.tree` module in the scikit-learn library (version 1.5.1). The trained model was evaluated on the test set using the corresponding 3-mer feature matrix to assess predictive performance based solely on miRNA sequence.

As a baseline comparison, a random classifier was implemented, assigning uniformly distributed values between 0 and 1 to each example, simulating random guessing.

This analysis was conducted using the following datasets:

- **Manakov datasets**, newly produced following Section 3.2.3 and Section 3.2.4.
  - 1:1 train set
  - 1:1 test set
- **Hejret datasets** from Hejret et al. (2023), including:
  - 1:10 train set[9], used for training the convolutional neural network (`miRNA_CNN`) in Hejret et al. (2023)
  - 1:1 test set[10]
- **miRAW datasets** from Pla et al. (2018)[11], including:
  - 1:1 train set[12], used for training the neural network (`miRAW`) in Pla et al. (2018) and the residual neural network (`TargetNet`) in Min et al. (2022)
  - 1:1 test set[13]

[9] `https://github.com/ML-Bioinfo-CEITEC/HybriDetector/blob/main/ML/Datasets/miRNA_train_set.tsv`. Last accessed January 2025.

[10] `https://github.com/ML-Bioinfo-CEITEC/HybriDetector/blob/main/ML/Datasets/miRNA_test_set_1.tsv`. Last accessed January 2025.

[11] All data is available in the repository `https://bitbucket.org/bipous/miraw_data/src/master/`. Last accessed January 2025.

[12] `miraw_data/PLOSComb/Data/ValidTargetSites/allTrainingSites.txt`. Last accessed January 2025.

[13] `miraw_data/PLOSComb/Data/TestData/balanced10/randomLeveragedTestSplit_0.csv`. Last accessed January 2025.

- **Yang datasets**[14] from T.-H. Yang et al. (2024), who claim to have mitigated the same bias, including:
  - 1:1 train set, used for training the attention-based model (`InteractionAwareModel`) in T.-H. Yang et al. (2024)
  - 1:1 test set

In this analysis, we identified the bias in all the analysed datasets except the Yang datasets. We confirmed that there is a difference between the frequencies of unique miRNAs occurring in each class, and that machine learning models can leverage this feature during inference. We termed the bias as the *miRNA frequency class bias*. This result is reported and further discussed in Chapter 4, Section 4.2.

To address this issue, we developed an alternative approach for generating negative examples aimed at mitigating the miRNA frequency class bias; this method is described in Section 3.2.7. To further enable robust evaluation of model generalisability, we also curated a completely unseen test set, designed to prevent reliance on such dataset-specific biases. The construction of this evaluation set precedes the negative example generation step to mitigate the bias, and is therefore described in the following section.

3.2.6 | Curating a Completely Unseen Evaluation Set

A key characteristic of miRNA–target site interaction datasets is that each example consists of a pair of sequences: the miRNA sequence and its corresponding target site sequence. It is common for a miRNA sequence to appear several times in the dataset and to be paired with multiple target sites, since only a limited number of mature miRNAs are known and each can target and repress the expression of many genes. Conversely, a target site sequence may also be appear several times in the dataset associated with various miRNAs, since a single target site can be targeted by different miRNAs.

Because each example consists of a pair of sequences, it is likely that one or both will *independently* appear in both the train and test sets. This introduces an element of data leakage, where the model encounters part of the test data during training. Such leakage can inflate prediction scores and lead to an overestimation of the model's true generalisability.

Information overlap between train and test sets at the *target site* sequence level is relatively minor. This is because the large number of potential target sites across the

[14] All data is available at `http://cosbi2.ee.ncku.edu.tw/mirna_binding/download`. Last accessed March 2025.

transcriptome lead to a vast diversity of target site sequences. This makes exact duplication of target sites uncommon. To further minimise this effect, we split the data based on the chromosome number associated with the target site, as outlined in Section 3.2.4. This ensures that train and test sets contain miRNA–target site interactions from distinct genomic regions, improving the independence of evaluation.

In contrast, overlap involving *miRNA* sequences is more substantial due to the limited number of distinct miRNAs, making sequence repetition across splits more likely. To address this, we curated an evaluation set in which all miRNAs are entirely absent from the train sets.

This evaluation set was constructed from the positive examples of the full Manakov dataset, curated as described in Section 3.2.2, prior to the negative examples generation step and the train–test splits. We also leveraged the positive class of the Hejret dataset (combined train and test sets) from Hejret et al. (2023), and the positive class of the Klimentova dataset from Klimentová et al. (2022) that only includes a small test set[15]. All three of these datasets were produced using the HybriDetector pipeline. As such, they share a similar structure and annotations.

We used the propagated miRNA family annotations retrieved from miRBase (version 22) and assigned to each example as part of the HybriDetector pipeline. We identified miRNA families unique to the positive examples of the Manakov dataset relative to the positive examples of the Hejret and the Klimentova datasets. The examples in the Manakov dataset comprising miRNAs belonging to such miRNA families were selected to form the positive class of this held-out evaluation set, thereafter referred to as the Manakov *left-out* set.

Although relatively small, comprising 10,027 positive examples including 328 unique miRNAs from 235 unique miRNA families, the Manakov left-out set plays a crucial role in this benchmarking effort. It enables the assessment of the true generalisability of models trained on any of the final train sets produced in this work, to interactions involving entirely unseen miRNAs.

The Manakov examples remaining following extraction of the left-out set, referred to as the Manakov *remaining* dataset, is the parent dataset of the final Manakov2022 train and test sets described downstream: after negative example generation that mitigates the identified bias described in the next section, followed by the same train–test splitting strategy described earlier in Section 3.2.4.

[15] `https://github.com/ML-Bioinfo-CEITEC/miRBind/blob/main/Datasets/AGO2_eCLIP_Klimentova2022_1_test.tsv`. Last accessed January 2025.

3.2.7 | Generating Negatives to Mitigate miRNA Frequency Class Bias

To mitigate the miRNA frequency class bias that was identified, which arises from the negative example generation method described in Section 3.2.3, an alternative approach was developed to preserve the miRNA frequency distribution across the positive and negative classes. This was achieved by generating negative examples *per miRNA* sequence instead of *per target site* sequence (see Figure 3.3).

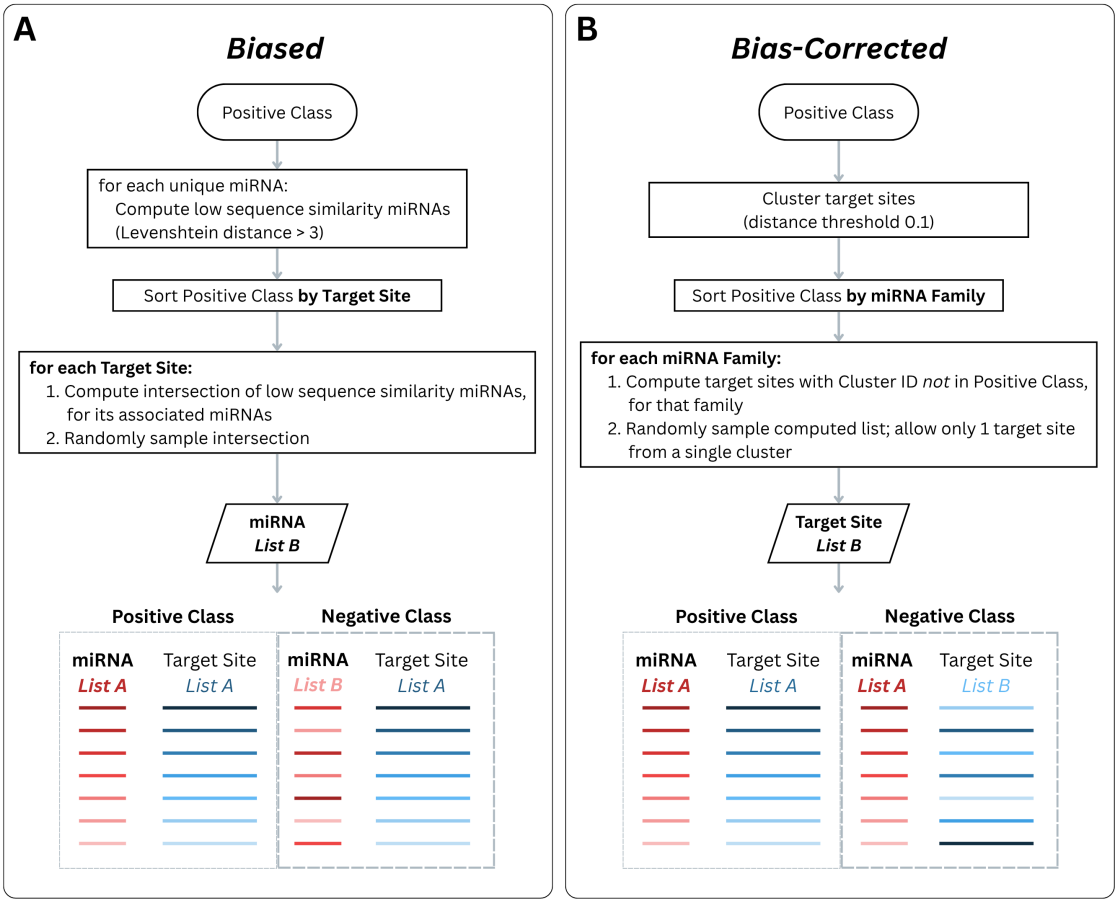

Figure 3.3: Comparison of the biased and bias-corrected negative example generation methods. Method A is reproduced from Hejret et al. (2023) and Klimentová et al. (2022) and produces datasets with the identified miRNA frequency class bias. Method B is the new method proposed and implemented in this study that mitigates this bias.

In accordance with the criteria for pairing a miRNA sequence with a target site sequence to generate a negative example, outlined in Section 3.2.3, the following steps outline the procedure for generating negative examples while mitigating the identified miRNA frequency class bias.

Prior to generating negative examples, the positive examples were alphabetically sorted by miRNA family name to group rows with the same miRNA family name into 'miRNA family blocks'. miRNA families effectively group together miRNAs that share sequence similarity. The dataset was then processed one block at a time.

In R, the `Clusterize` function from the DECIPHER package (version 2.30.0) in Bioconductor (version 3.18) was used to accurately group similar gene sequences together in linear time (Wright, 2024). The function assigns a cluster ID to each sequence based on a specified distance threshold, effectively clustering those that are sufficiently similar. A distance threshold of 0.1 was applied, grouping target site sequences that are less than 10% distant from one another (or, equivalently, more than 90% similar) in the same cluster.

This approach for generating negatives was applied to the Manakov left-out and remaining datasets, produced from the split described in Section 3.2.6. We also used it to generate new unbiased versions of the datasets previously published by Hejret et al. (2023) and Klimentová et al. (2022), by applying it to their respective positive examples. The clustering step produced a majority of singleton clusters across all datasets: approximately 60% of total clusters for Manakov remaining, and 90% for Manakov left-out, for the Hejret dataset, and for the Klimentova dataset, with the largest cluster occurring in the Manakov remaining dataset, comprising 253 sequences, corresponding to 0.02% of target sequences.

For each miRNA family block in the positive (binding) examples, candidate non-binding sites were identified as target site sequences belonging to clusters distinct from those in the miRNA family block. This method for selecting alternative non-binding sites provided support for Section 3.2.3 criteria (i) and (ii), generating negative interactions that do not already exist as positive examples, and that are sufficiently different from those in the positive examples. These candidate sites were then randomly sampled, ensuring that only a single binding site was selected from each candidate cluster. This supports criterion (iii), generating unique negative examples. The number of sampled candidates was determined based on the number of positive examples within the miRNA family block, maintaining a balanced 1:1 positive-to-negative class ratio. It was not feasible to generate higher class imbalance ratios with this method as there were insufficient candidate target sites to sample from for the most abundant miRNAs.

This method for generating negatives was applied to the positive examples comprising the Manakov remaining and left-out datasets, and the Hejret and Klimentova datasets. For the Manakov remaining and Hejret datasets, the negative example generation was followed by the train–test splitting strategy described in Section 3.2.4. The Klimentova dataset, comprising only 477 positive examples, was too small to split and was therefore entirely assigned as a test set. This produced three dataset collections, hereafter referred to as:

- **the Manakov2022 dataset**: train, test, left-out sets
- **the Hejret2023 dataset**: train, test sets
- **the Klimentova2022 dataset**: test set

3.2.7.1 | Sanity Checks: Evaluating Bias Removal in the Newly Developed Negative Generation Method

To verify the elimination of the miRNA frequency class bias and ensure that models trained on these miRNA interaction datasets are no longer able to classify binding and non-binding sequences from the miRNA sequence alone, another decision tree classifier was trained following the approach described in Section 3.2.5. *K*-mer frequency matrices were computed from miRNA sequences in the newly generated class-balanced datasets, which are supposedly free of the previously identified bias artifact.

Since there are a significantly larger number of unique target site sequences than miRNA sequences, the likelihood of a target site frequency imbalance between the positive and negative classes is reduced. However, to ensure that the new method for generating negative examples does not inadvertently introduce this bias artifact, an additional test was conducted. Another decision tree classifier was trained following the approach described in Section 3.2.5, but this time using *k*-mer frequency matrices computed from target site sequences instead of miRNA sequences.

To establish a baseline for comparison, random classifiers were implemented again. These analyses were conducted on the newly curated Manakov2022 train and test sets, and on the newly generated versions of the Hejret2023 train and test sets.

3.2.8 | FAIR data

In this study, three dataset collections were produced, comprising a total of six datasets, as follows:

1. **Manakov2022**: newly curated train, test, and left-out sets
2. **Hejret2023**: new versions of the train and test sets
3. **Klimentova2022**: new version of the test set

To ensure that these datasets adhere to the FAIR (Findable, Accessible, Interoperable, Reusable) principles for (Wilkinson et al., 2016), we have made them publicly available through the Zenodo data-sharing platform, assigning them a Digital Object Identifier (DOI) and maintaining a versioned release (version 5)[16].

To enhance accessibility and usability, we have also contributed these datasets to the publicly available `miRBench` Python package[17]. `miRBench` is a novel benchmarking framework designed for assessing miRNA target site prediction methods. It offers functionalities for dataset retrieval, inference from state-of-the-art models, and implementation of their respective input sequence interaction encoding. Further details are provided in an associated publication by Sammut et al. (2025b) (refer to Appendix A).

3.3 | Benchmarking the State of the Art

When selecting state-of-the-art models to benchmark on the task of miRNA target site prediction, a number of criteria should be considered. First is the ability to produce a score reflecting binding site affinity. Second is the direct use of miRNA and target site sequences, and only these sequences, without requiring additional features. Third is the availability of a functional and accessible implementation. Consequently, this benchmark excludes gene- (transcript-) level target prediction tools that do not focus on binding sites, models that require supplementary input data, and methods lacking available or complete implementations, as re-implementation is not feasible.

In accordance to these criteria, we selected six state-of-the-art deep learning models that have also been reviewed in the Background and Literature Review in Chapter 2, and are summarised in Table 3.2.

[16] `https://zenodo.org/records/14501607`. Last accessed June 2025.
[17] `https://github.com/katarinagresova/miRBench/releases/tag/v1.0.1`. Last accessed June 2025.

Table 3.2: State-of-the-art deep learning models, available on the `miRBench` Python package, and benchmarked in this work.

Model	Architecture	Training Data	Model Output	Reported Performance
`TargetScanCnn` (McGeary et al., 2019)	CNN	Dissociation constants (K_d) from novel RNA Bind-n-Seq (RBNS) experiment (developed by authors)	$-\log(K_d)$, serving as a proxy for binding affinity	–
`CnnMirTarget` (Zheng et al., 2020)	CNN	Human AGO1 CLASH data (Helwak et al., 2013), *C. elegans* and mammalian iPAR-CLIP datasets (Grosswendt et al., 2014), MirTarBase data with strong experimental evidence (Hsu et al., 2010)	Probability of binding	0.95*
`TargetNet` (Min et al., 2022)	ResNet	miRAW dataset (Pla et al., 2018)	Probability of binding	0.77*
`miRBind` (Klimentová et al., 2022)	ResNet	Human AGO1 CLASH data (Helwak et al., 2013)	Probability of binding	0.95†
`miRNA_CNN` (Hejret et al., 2023)	CNN	Human AGO2 CLASH data (produced by authors)	Probability of binding	0.89†
`InteractionAwareModel` (T.-H. Yang et al., 2024)	Deep multi-head mutual attention network	Re-analysed data from human AGO1 CLASH data (Helwak et al., 2013)	Probability of binding	0.97‡

* F1 score † PR–AUC ‡ ROC–AUC – No reported performance metric

In this study, we leveraged the `miRBench` package (version 1.0.1), to programmatically access these models and their respective input data encoders. We encoded each of the four evaluation sets curated in this work (Manakov2022 test and left-out, Hejret2023 test, Klimentova2022 test) using the same respective encoder employed during training of each of the six models. This ensures compatibility with the data input format expected by each model. Predictions were obtained from each model on each dataset, which were then evaluated. Detailed notes on evaluation are found at the end of this chapter in Section 3.6. This workflow describes the benchmarking process. Additionally, we also benchmarked all six models on the original version of the Hejret 1:1 test set (Hejret et al., 2023), for comparison with the new unbiased version presented here.

Beyond these deep learning models, that have been specifically trained on the task of miRNA target site prediction, we also implemented a random classifier, and two traditional methods based on seed complementarity and co-folding of the miRNA and target site sequences. These are described in the sections that follow.

3.3.1 | Random Classifier

Following negative example generation as per the method described in Section 3.2.7, that corrects for the identified miRNA frequency class bias, upon splitting the produced 1:1 class balanced dataset, as per the technique described in Section 3.2.4, based on chromosome number annotation, a slight class imbalance is introduced in each of the final train and test splits. This is because in the method used for generating negatives, only the list of miRNA sequences is propagated from the positive to the negative class. The target sites are sampled from a candidate pool of target site sequences. As such the number of target sites occurring on chromosome 1 is not consistent across the classes. Therefore, upon splitting, the number of positive and negative examples are redistributed between the train and test sets, yielding a slight class imbalance. As a result, the probability of selecting examples from each class becomes unequal, leading to deviations in performance from the expected random baseline of 0.5. This phenomenon is further discussed in Chapter 4, Section 4.3.

In this context, the inclusion of a random classifier provides a meaningful reference point for evaluating model performance on each of the different test sets.

3.3.2 | Seed Complementarity

The `miRBench` package also includes encoders and predictors for predefined seed types (see Figure 3.4). Seed types are categorised based on complementarity patterns between the miRNA and its target site. In this work, we defined a number of seed types similar to those described by Bartel (2018), as outlined in Chapter 2, Section 2.2. Our seed definitions are depicted in Figure 3.4, and defined as follows:

- **`Seed8mer`** exhibits full complementarity at positions 2–8, with an adenine (A) at position 1.
- **`Seed7mer`** maintains full complementarity at positions 2–8, or at 2–7 with an A at position 1.
- **`Seed6mer`** includes full complementarity at positions 2–7, or 3–8, or 2–6 with an A at position 1.
- **`Seed6merBulgeOrMismatch`** follows the same criteria as those described for the `Seed6mer` but additionally allows for a single bulge or mismatch within the seed region.

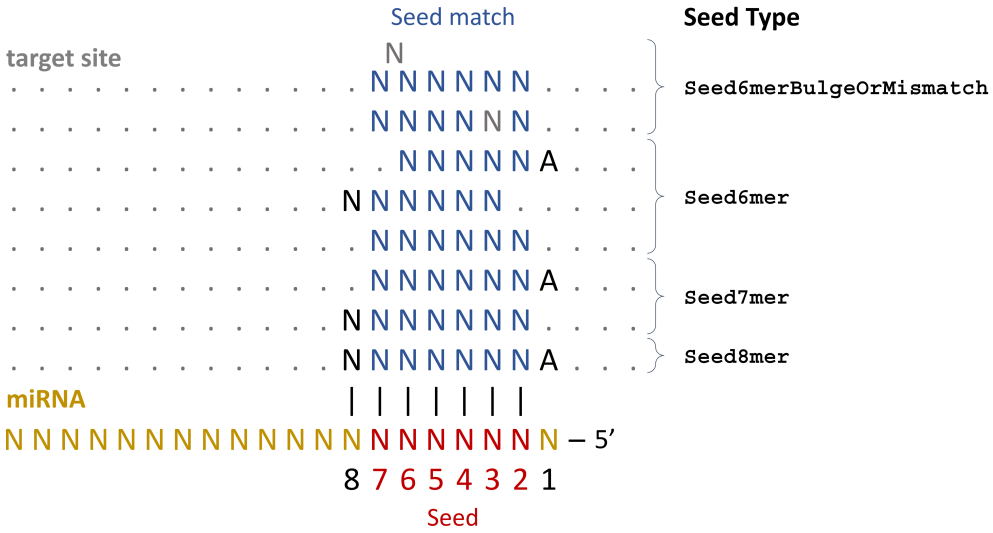

Figure 3.4: An illustration of the target site seed types as defined in the `miRBench` package. These seed classifiers provide a performance baseline.

These seed classifiers are therefore designed in a hierarchy, where the looser definitions include the stricter ones. The loosest seed type is the `Seed6merBulgeOrMismatch`, which captures all seed types in addition to seeds with a bulge or mismatch. This is followed by the `Seed6mer`, which also captures the `Seed7mer` and `Seed8mer`, and then the `Seed7mer`, which only additionally captures any `Seed8mer`. The strictest seed type is the longest, the `Seed8mer`, which captures only this seed type.

The seed encoder in `miRBench` returns an array containing the miRNA sequence and the reverse complement of the target site, since both sequences run in the 5′ to 3′ direction. The predictor then extracts the defined seed from the miRNA sequence and determines whether it appears anywhere within the reverse complement of the target site sequence, outputting a boolean value indicating presence (1) or absence (0).

This provides an additional baseline for evaluation. Given that approximately 20% of interactions occur outside the seed region (Helwak et al., 2013), by comparing the more complex state-of-the-art models summarised in Table 3.2 against this baseline, we can assess whether these models capture binding patterns beyond seed-based complementarity, offering insight into the extent of their learning.

3.3.3 | RNA–RNA Folding

Another tool integrated into the `miRBench` package is `RNACofold` (Bernhart et al., 2006), that is part of the `ViennaRNA` Package (Lorenz et al., 2011). `RNACofold` predicts the binding energy and base-pairing pattern of interacting RNA sequences. The program outputs the minimum folding energy (MFE) of the folded structure, among others. In the `miRBench` package, the MFE is leveraged and multiplied by -1 to serve as a binding affinity score. This assumes that lower `RNACofold` energy corresponds to a more stable interaction and, consequently, a higher likelihood of binding, as molecular interactions generally favor states of lower energy.

Including `RNACofold` in benchmarking provides an additional baseline using a more traditional approach for approximating RNA–RNA binding. This approach leverages conventional dynamic programming algorithms. These are computational methods that break down a complex problem into simpler, repeated sub-problems and store intermediate results for reuse to efficiently obtain a solution. This approach is common in RNA secondary structure prediction. The `RNACofold` baseline allows for an assessment of whether data-driven models capture additional interaction complexities beyond those predicted by these traditional methods.

3.4 | Exploring Simple Machine Learning Models Using Engineered Features

In this section, we present a proof of concept demonstrating how the curated datasets may be used to train classical machine learning models on the task of miRNA target site prediction, and provide a baseline for such models.

3.4.1 | Data Representation

To explore the utility of classical machine learning models for miRNA target site prediction, a set of features were designed to numerically represent the interaction between the miRNA and target site sequences.

Since examples in the dataset occur as *pairs* of miRNA–target site sequences, they must therefore be represented as such. Commonly used encodings to represent an RNA or DNA sequence, such as simple k-mer counts, are therefore not sufficient. A data representation that captures information across each interaction is required.

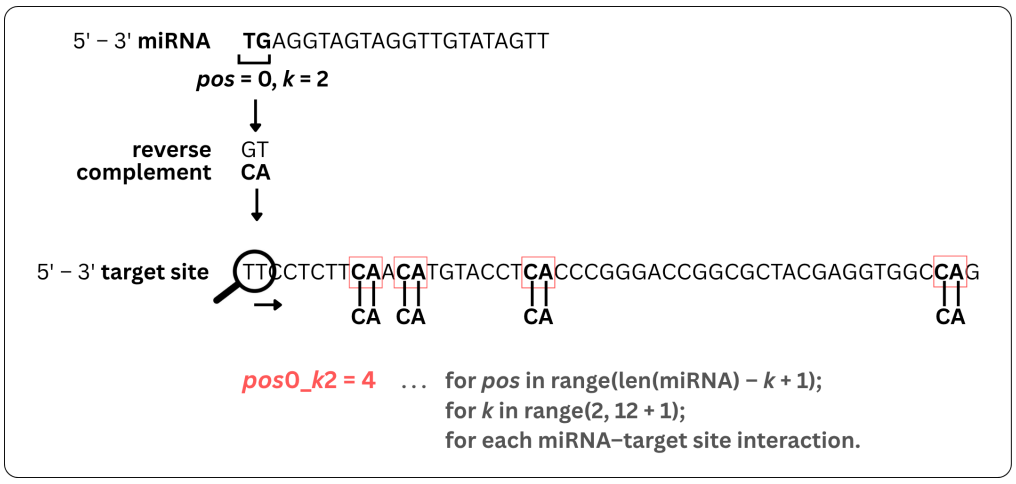

Figure 3.5: An illustration of the feature extraction method used to capture information about the interactions between the miRNA and target site sequences.

We designed and extracted features that record the position and length of k-mers from the miRNA, and quantified the occurrence of the reverse complement of these k-mers within the corresponding target site sequence. We used the *reverse* complement since both sequences run in the 5′ to 3′ direction. These features quantify the frequency of contiguous complementary stretches of varying k-mer lengths between the miRNA and its target site.

Each miRNA sequence was first truncated or padded to a uniform length of 20 nucleotides, since miRNA sequences vary between 16 and 28 nucleotides in length in our datasets. For values of k ranging from 2 to 12, all possible k-mers were extracted from the miRNA. For each k-mer, its reverse complement was computed and the number of times this appeared in the target site sequence was recorded. A separate feature named `pos<i>_k<k>` was defined for each k-mer starting at position i, recording the frequency of its reverse complement in the target site. This process was repeated for all k and all relevant positions in the miRNA sequence, resulting in a fixed-length feature vector of 154 features per example, with no missing values.

The Manakov2022 train, test and left-out datasets were encoded in this way to be used throughout this experiment.

3.4.2 | Model Training

The goal of model training is to learn a function that captures the relationship between input features and labels, and generalises effectively to unseen examples. As a proof of concept for dataset usage and to set a baseline for ML models on the task of miRNA target site prediction, we chose to train tree-based, inherently interpretable binary classifiers that also facilitate insights into feature contributions, an important downstream objective for deciphering the rules governing miRNA–target interactions. Although support vector machines (SVMs) have shown good performance on this task (Lu & Leslie, 2016; Wang, 2016), they do not scale well to large datasets and require long training times. Given the size of our dataset and the scope of this experiment, we chose not to include them.

We selected three widely used tree-based algorithms: Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost). These models inherently perform automatic feature selection by identifying the features that most effectively split the data at each decision point, to maximise information gain or minimise leaf impurity, eliminating the need for manual feature selection prior to training. Notably, tree-based models are also invariant to feature scaling and normalisation, as they split data by comparing individual features to threshold values. These models rely on the relative order of feature values, not their absolute magnitudes. As a result, such transformations were not required.

We implemented the `DecisionTreeClassifier` and `RandomForestClassifier` classes from the scikit-learn library version 1.5.1, and the `XGBClassifier` from the xgboost library (version 3.0.0), in Python. A set of hyperparameters was selected for tuning, and a corresponding search space was defined, for each model (see Table 3.3).

Table 3.3: Hyperparameter search spaces for the tree-based models trained in this study.

Model	Hyperparameter	Search Space
Decision Tree	`max_depth`	Integer(10, 25)
	`min_samples_split`	Real(0.0002, 0.02, prior=*log-uniform*)
	`min_samples_leaf`	Real(0.0001, 0.01, prior=*log-uniform*)
	`criterion`	Categorical([*gini*, *entropy*])
	`max_features`	Categorical([*sqrt*, *log2*, 0.25, 0.5])
Random Forest	`max_depth`	Integer(10, 25)
	`min_samples_split`	Real(0.0002, 0.02, prior=*log-uniform*)
	`min_samples_leaf`	Real(0.0001, 0.01, prior=*log-uniform*)
	`criterion`	Categorical([*gini*, *entropy*])
	`max_features`	Categorical([*sqrt*, *log2*, 0.25, 0.5])
	`n_estimators`	Integer(100, 300)
XGBoost	`max_depth`	Integer(3, 10)
	`n_estimators`	Integer(100, 300)
	`eta`	Real(0.01, 0.5, prior=*log-uniform*)

3.4.2.1 | Decision Tree

Decision Trees are supervised models that recursively split the feature space to achieve class separation in as few splits as possible. At each node, the algorithm selects the feature and threshold that maximise the reduction in node impurity (such as Gini impurity or entropy), which is a measure of how heterogeneous the class labels are within a node. This process continues until the leaves of the tree represent pure, or nearly pure, class labels, or until other stopping criteria, such as a minimum number of samples per leaf, or maximum depth, are met. The leaf nodes are then assigned an output value based on the majority class, and unseen examples are classified by traversing the tree from the root down to a leaf, following the decision paths.

For the Decision Tree, we tuned standard tree-complexity hyperparameters with search ranges chosen to suit the ∼2.5 M example Manakov2022 train set. We varied the maximum depth (`max_depth`) to explore both shallow (high-bias, low-variance) and deeper (low-bias, high-variance) trees without under- or over- fitting. The minimum examples required to split a node (`min_samples_split`) and per leaf (`min_samples_leaf`) were defined as fractions of the train set so that splits and leaves span from a few hundred to tens of thousands of examples—small enough to capture real patterns yet large enough to ignore random noise. The log-uniform prior ensures even chance for exploration across orders of magnitude. We varied the number of fea-

tures considered at each split (`max_features`). We also varied the function used to measure the quality of a split (`criterion`).

3.4.2.2 | Random Forest

Random forest is an ensemble technique that builds on decision trees by combining many trees, each constructed on a different bootstrapped sample of the original dataset. For each tree, a bootstrapped dataset is created by sampling a subset with replacement. Additionally, as each decision tree is built, at each splitting node, only a random subset of features is considered, producing diverse trees thus enhancing robustness. The final prediction is derived by aggregating the outputs of the individual trees and selecting the most frequent class as the final prediction. The process of bootstrapping the data to train multiple models and using the aggregate of model outputs to decide on a final output is known as bagging, and is known to reduce variance. For the Random Forest, the same hyperparameters and search spaces as the Decision Tree were used, along with the number of trees in the ensemble (`n_estimators`).

3.4.2.3 | XGBoost

Boosting is another ensemble method where models are trained sequentially, with each new model focusing on reducing the error made by the previous. Gradient boosting is one such technique which builds a strong predictive model by sequentially adding weak learners, typically decision trees, each trained to correct the residual errors of the combined ensemble. A scalable implementation of gradient boosting is Extreme Gradient Boosting known as XGBoost (T. Chen & Guestrin, 2016), which was designed to be used with large, more complex datasets. For XGBoost, we used the `binary:logistic` objective to model the probability of the positive class. We tuned the maximum tree depth and the number of trees, as well as the learning rate (`eta`), which controls the contribution of each tree to the final model.

3.4.2.4 | Hyperparameter Optimisation with Cross-Validation

While hyperparameter optimisation is typically a step-wise and iterative process, involving extensive experimentation and evaluation, the final selections on hyperparameter search space used here were considered sufficient for the scope of this work.

We implemented Bayesian optimisation to identify suitable hyperparameter configurations. Instead of blindly exploring the hyperparameter space, as in Random or Grid search, Bayesian optimisation uses past evaluations, each corresponding to a set

of hyperparameters and its associated performance, to guide the future search. This means that each new evaluation is informed by the results of previous experiments, which in turn directs the search towards more promising areas of the hyperparameter space. It therefore improves the search speed for the optimal hyperparameters. We ran the search for 30 iterations. This was done using the `BayesSearchCV` class from the scikit-optimize library (version 0.10.2), which integrates hyperparameter tuning with cross-validation for each configuration.

Cross-validation involves dividing the available train set into k subsets, or *folds*, and using $k-1$ folds for training while the remaining fold is reserved for validation. This process is repeated k times so that each fold serves as the validation set exactly once. We employed the default stratified k-fold cross-validation with $k = 5$, resulting in an 80:20 train–validation split for each fold. *Stratified* k-fold cross-validation ensures that each fold preserves the class proportions of the full dataset.

During hyperparameter tuning, each hyperparameter configuration was evaluated on all five folds, producing a distribution of performance scores. The mean and standard deviation of the chosen scoring metric were computed across the five folds to gauge performance stability and model generalisability. The main advantage of this approach is that it aggregates performance over multiple validation sets, leading to a more robust estimate of model fitness.

We used the Average Precision Score (APS) to consistently evaluate model performance across all possible classification thresholds, as described in more detail in Section 3.6. Model performance was evaluated on both train and validation splits per fold during cross-validation to monitor for model over- or under- fitting.

Following cross-validation, the hyperparameter configuration with the highest mean APS for each model type was selected, and a final model was trained on the entire Manakov2022 train set, resulting in three final models, one for each algorithm[18]. Final evaluation of each of these three models was performed on the unseen Manakov2022 test and left-out splits, to test model generalisability.

[18]Deposited on Zenodo at `https://zenodo.org/records/16307664`. Last accessed July 2025.

3.5 | State-of-the-Art Neural Network Retrained on Unbiased Datasets

Building upon the baseline established by simpler machine learning models relying on explicit feature engineering, we now investigate deep learning architectures that inherently alleviate the need for extensive manual feature construction. Given the complexity and largely unknown binding patterns within miRNA–target site interactions, neural networks offer the advantage of bypassing the explicit feature engineering step. Although a meaningful data representation remains necessary as input, a neural network can infer relevant features itself directly from the data.

In this work, we retrained the state-of-the-art convolutional neural network (CNN) architecture introduced by Hejret et al. (2023), `miRNA_CNN`, included in our benchmarking effort described earlier in Section 3.3. This CNN was retrained on the revised Hejret2022 train set, corrected for the miRNA frequency class bias identified in this work, thus yielding an unbiased variant of the original CNN model. Additionally, we also independently retrained the model using the newly curated and unbiased Manakov2022 train set, providing a significantly larger data resource.

3.5.1 | Data Representation using Watson–Crick Base-Pairing

We leveraged the data representation initially described and implemented for `miRBind` by Klimentová et al. (2022), and later re-implemented for the `miRNA_CNN` described by Hejret et al. (2023), both benchmarked in this work. An implementation of this encoding is available through the `miRBench` package.

Each miRNA–target pair in the dataset, corresponding to a single example, was encoded into a two-dimensional (2D) matrix (or array) representing possible Watson–Crick base pairing, specifically: AT, TA, AU, UA, GC, and CG. The miRNA and target sequences were truncated or padded with 'N' to a constant length of 20 and 50 nucleotides, respectively. This allowed for consistent dimensions of the 2D matrix across all examples, with the length of the miRNA sequence as the width of the matrix (inner array length) and the length of the target sequence as the height of the matrix (outer array length). Each position within the matrix indicated the presence (1) or absence (0) of Watson–Crick base pairing between the respective nucleotides on the miRNA and target site sequences. This encoding is hereafter referred to as the *Sequence-only* data representation. An example is shown in Figure 3.6.

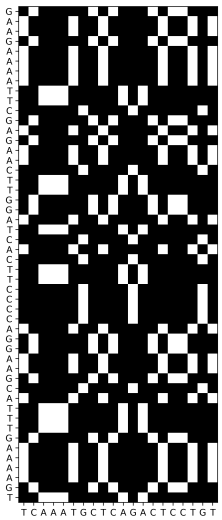

Figure 3.6: An illustration of an example of the *Sequence-only* $50 \times 20 \times 1$ data representation, encoding Watson–Crick base-pairing between a miRNA and target site sequence. This representation was used to retrain the `miRNA_CNN` architecture on the bias-corrected datasets.

To train the CNN, an input tensor with dimensions of $N \times 50 \times 20 \times 1$ was used, where N was the batch size representing the number of examples processed at a time in a batch, 50 was the height of the matrix representing the length of the target gene sequence, 20 was the width of the matrix representing the length of the miRNA sequence, and 1 indicated a single channel, analogous to a binary image, with pixel intensities either 0 or 1 (black-and-white representation). Thus, the binary matrix captured positional interaction information without explicitly revealing the nucleotide identities of either sequence. This encoding procedure was consistently applied across batches of examples from the dataset. Furthermore, this tensor shape matched the expected input dimensions of the first convolutional layer (Conv2D) in the CNN architecture.

The Hejret2023 train and test sets, and the Manakov2022 train, test and left-out sets were encoded in this way to be used throughout this experiment.

3.5.2 | Data Representation Enhanced with Co-folded Structure

To further enrich the data with structural information, we incorporated a new representation derived from the dot–bracket notation of the minimum free energy (MFE) conformation predicted by `RNACofold` from the `ViennaRNA` Package (version 2.7.0) (Lorenz et al., 2011). We retrieved the dot–bracket notation for each example in the dataset. In the dot–bracket structure, each character reflects the pairing status of a nucleotide: a

dot (".") indicates an unpaired nucleotide, while an opening "(" and a closing ")" bracket denote paired nucleotides. The dot–bracket string is a concatenation of two parts: the miRNA followed by the target sequence, so its length equals the sum of the individual sequence lengths.

With this understanding, the miRNA and target sequences were again truncated or padded with 'N' to a constant length of 20 and 50 nucleotides, respectively. For each example, a zero matrix with dimensions of $50 \times 20 \times 1$ was initialised. A stack-based algorithm was used to parse the dot–bracket string. As the function iterates over the string, each opening bracket is pushed onto a stack; when a closing bracket is encountered, the corresponding opening bracket is popped to establish a base pair. The algorithm checks if this pair represents an intermolecular interaction, i.e. whether one nucleotide is part of the miRNA region (first 20 nucleotides) and the other is part of the target region (last 50 nucleotides). For each valid intermolecular interaction, the corresponding element in the binding matrix is set to 1. Figure 3.7 illustrates an example encoding. The dot–bracket structure for this example is:

```
(((((((((.....(((((.((((...((((((....)))))).))))).))))))))))).....
```

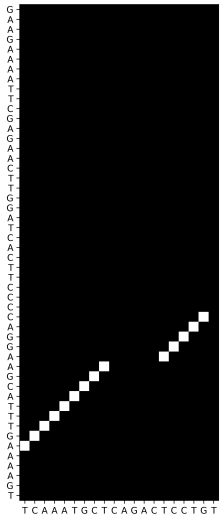

Figure 3.7: An illustration of an example of the second channel for the *Sequence & Co-folding* $50 \times 20 \times 2$ data representation, encoding intermolecular binding of a miRNA and target sequence derived from the dot–bracket MFE structure from `RNACofold`.

Once all structures were processed, the resulting matrices were combined into a four-dimensional tensor with dimensions of $N \times 50 \times 20 \times 1$. This tensor was concatenated with the $N \times 50 \times 20 \times 1$ tensor, described in Section 3.5.1, along the last channel dimension, producing a tensor with dimensions of $N \times 50 \times 20 \times 2$. This enhanced representation was intended to provide additional insight into intermolecular binding, complementing the information on positional complementarity, and thereby offering the neural network a richer data representation from which to infer binding patterns. This encoding is referred to as the *Sequence & Co-folding* data representation throughout this study.

All six subsets from the three dataset collections produced in this work were encoded in this way to be used throughout this experiment.

3.5.3 | Model Training

The CNN model was independently trained on two datasets. Initially, it was retrained on the bias-corrected Hejret2023 train set to generate an unbiased variant of the original `miRNA_CNN` model. Subsequently, the model was trained on the larger, newly curated Manakov2022 train set to assess if increased dataset size enhances classification performance and captures additional relevant binding information. The CNN was trained on each of the data representations described in Sections 3.5.1 and 3.5.2, adjusting the number of channels in the input layer of the CNN accordingly.

We employed the Keras high-level API from the TensorFlow library (version 2.13.1), following the CNN architecture described by Hejret et al. (2023), that is visualised in Chapter 2, Figure 2.6. Specifically, the network consisted of an input layer, followed by six sequential convolutional blocks, each comprising a 2D convolutional layer, employing filters of size 5×5 with the number of filters incrementally increasing across the six convolutional blocks, leaky rectified linear unit (Leaky ReLU) activation, batch normalisation, max pooling, and a dropout layer with a dropout rate of 0.3. Following the convolutional blocks, the network output is flattened and passed through two fully connected (dense) layers, each coupled with Leaky ReLU activation, batch normalisation, and a dropout layer with a dropout rate of 0.3. The final output layer consists of a single neuron using a sigmoid activation function, producing probability scores indicating the likelihood of miRNA-target binding for each input example.

The CNN was compiled using the Adam optimiser with a learning rate of 0.00152 and binary cross-entropy loss as the objective function, as per the original implementation by Hejret et al. (2023). The model was trained using a randomised hold-out validation strategy, splitting the dataset into 90% training and 10% validation subsets. The

training subset was shuffled at the end of each epoch. This ensures that batches contain different examples in different epochs, which helps the model avoid memorising the order thus promoting robustness, and also preventing biases from the ordering of the dataset, ensuring that every training epoch sees varied data combinations. The validation subset was maintained consistent, ensuring meaningful performance comparisons across epochs over training. The training process was monitored on both the train and validation sets using the binary cross-entropy loss and accuracy evaluation metrics. The model was trained over 10 epochs, with a batch size of 32.

The final models[19] were evaluated on all four evaluation sets presented in this work, using relevant evaluation metrics as described in the next section.

3.6 | Evaluation Metrics

Evaluation enables an objective assessment of a model's performance, providing insights into its capability to generalise to unseen data. In machine learning approaches, evaluation is carried out multiple times as part of the training process, and once more following training of the final model. Throughout the training process, only the train set is used. For evaluation, the train set is therefore further split into a train set and a validation set, typically in an 80:20 or 90:10 ratio, and the validation set is used for periodic evaluation to monitor the training process. For evaluation of the final model, a separate held-out test set is used.

In all cases, the first step of evaluation involves inference, i.e. obtaining predictions by the model on each of the examples in the respective dataset. The second step of evaluation involves selecting the appropriate metric that summarises the performance of the model into a single, comparable figure. Various such metrics exist, each suited for specific tasks and dataset characteristics.

Most metrics are calculated using terms from the confusion matrix (see Table 3.4), which summarises predictions into the following terms:

- **True Positive (TP)** is the number of positive examples correctly predicted as positive by the model.
- **True Negative (TN)** is the number of negative examples correctly predicted as negative by the model.

[19] Deposited on Zenodo at `https://zenodo.org/records/16307664`. Last accessed July 2025.

Table 3.4: The components of a confusion matrix for binary classification.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

- **False Positive (FP)** is the number of negative examples incorrectly predicted as positive by the model.
- **False Negative (FN)** is the number of positive examples incorrectly predicted as negative by the model.

These terms are calculated at a single classification threshold value, typically 0.5, where predicted probabilities equal to or greater than this threshold are classified as positives, and those below this threshold are classified as negatives. These predictions are compared to their actual class label and summarised in a confusion matrix (see Table 3.4), from which further evaluation metrics may be computed.

One commonly used metric is accuracy, defined as the number of correct classifications out of the total classifications, and given by the Equation 3.1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

In this work, while the negative example generation process ensured class balance (Section 3.2.7), the subsequent dataset splitting step into the train and test sets (Section 3.2.4) reintroduced some degree of class imbalance. Since accuracy does not differentiate between the correctly classified examples of each class (TP and TN are summed in the numerator in Equation 3.1), precision and recall were prioritised. Precision is defined as the number of correctly classified actual positives out of all the examples classified as positives, given by Equation 3.2, and recall, or True Positive Rate Rate (TPR), is defined as the number of correctly classified actual positives out of all positives, given by Equation 3.3. This ensures reliable evaluation.

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

However, it is useful to evaluate model performance across all possible classification thresholds because it reveals the trade-off between different types of errors (FP and FN) at various decision points, as opposed to relying on a single arbitrary threshold which can be misleading. Metrics such as the area under the Receiver-Operating Characteristic (ROC) curve or the Precision-Recall (PR) curve, described next, summarise model performance across all thresholds.

The ROC curve is constructed by plotting the recall or TPR (Equation 3.3) against the False Positive Rate (FPR) (Equation 3.4) across all possible classification thresholds. Ideally, the TPR is high and the FPR is low, therefore good performance is indicated by a higher top-left curve, hence a larger Area Under Curve (AUC). The AUC of a ROC curve quantifies the likelihood that a model scores a randomly selected positive example higher than a randomly selected negative example. ROC–AUC is particularly useful for comparing model performance when the classes in the dataset are roughly balanced, and detecting TPs and TNs is equally important.

$$\textit{False Positive Rate (FPR)} = \frac{FP}{FP + TN} \tag{3.4}$$

For heavily imbalanced datasets, PR curves may be more informative. A PR curve illustrates precision against recall across varying thresholds. Ideally, both precision and recall are high, therefore good performance is indicated by a higher top-right curve, hence a larger AUC. The PR–AUC quantifies the model's ability to distinguish between classes across all thresholds. This metric is also more useful in needle-in-a-haystack type problems, where detecting TPs is more important or interesting than detecting TNs, as is the scenario in this study.

Another useful metric is the Average Precision Score (APS), given by Equation 3.5 where P_n is precision at the n^{th} threshold, given by the proportion of positive predictions that are correct at this threshold, and R_n is recall at the n^{th} threshold, given by the proportion of actual positives correctly predicted at or above this threshold. Therefore, APS is the weighted sum of precision values at various prediction thresholds, where the weight for each precision value corresponds to the change in recall between consecutive thresholds.

$$\textit{Average Precision Score (APS)} = \sum_{n} (R_n - R_{n-1})\, P_n \tag{3.5}$$

Although commonly used, the PR–AUC may sometimes be misleading as it relies on linear interpolation between PR points on the PR curve. During the analysis outlined in Section 3.2.5, during evaluation of decision tree classifiers, we observed an abrupt drop in precision in some PR curves. This likely indicates the presence of a single highly scored prediction correctly identified as positive, immediately followed by many predictions with lower but identical scores that are predominantly false positives. Once these subsequent predictions are included, the number of FP sharply increases, causing the precision to rapidly decline (see Equation 3.2). As PR–AUC is computed by linear interpolation of these points, this inflates the PR–AUC metric, which may wrongly infer higher performance.

APS offers a more robust evaluation as it avoids this issue by computing precision and recall across multiple thresholds via weighted sum (see Equation 3.5), thus reflecting performance more accurately. Therefore, to achieve reliable comparisons across all models, APS was selected as the primary performance metric throughout this study. The scikit-learn Python library (version 1.5.1) was used.

3.7 | Code and Data Availability

All code implemented to produce this work is available on GitHub[20].
The final dataset collections produced in this study are available on Zenodo[21].

3.8 | Summary

In this chapter we thoroughly explain the methodology applied in this study. We first provide an overview of the four components that comprise this work to achieve our defined aims and objectives. This is followed by a detailed description of the analyses and experiments comprising each component.

We first describe the dataset curation process, which includes several parts. Notably, we identify a bias derived from a negative generation method reproduced from the literature. To address this, we develop a new method for generating negatives and use it to curate the new benchmarking dataset and to correct two existing published datasets. We contribute a total of three dataset collections, which include various train and evaluation sets, to the `miRBench` Python package for miRNA–target site prediction benchmarks, while ensuring compliance with FAIR principles.

[20] `https://github.com/stephaniesamm/miRBench-BiasCorrectedChimericDatasets-ML`
[21] `https://zenodo.org/records/14501607`

We then explain the benchmarking process. This includes leveraging `miRBench` to access six state-of-the-art deep learning models and their corresponding data encoders, and running inference and evaluation to obtain a single comparable performance metric.

Next, we describe a proof of concept demonstrating the utility of the curated datasets by designing a feature set based on interacting sequence pairs, and training decision tree, random forest, and XGBoost classifiers to establish a baseline for simpler models.

Finally, we describe the retraining of a published CNN architecture, originally trained on a biased dataset, using our bias-corrected datasets to assess the effect of bias removal and establish a new baseline for deep learning models.

At the end of the chapter, we provide notes on evaluation and performance metrics. Throughout this work, we use the APS for evaluation, accompanied by PR curves where necessary to aid visualisation of results. All results derived from the work described in this chapter are presented and discussed in the Results & Discussion in the chapter that follows.

4

Results & Discussion

The aim of this study was to standardise benchmarking in the field of miRNA target site prediction by (i) producing carefully-curated datasets and making them accessible and easy to use which lowers the barrier of entry to the field for deep learning experts; (ii) reviewing and evaluating the state of the art in the field using the curated datasets; and (iii) producing baseline models by training and benchmarking classical machine learning (ML) and neural networks models on the curated datasets.

In this chapter we present and discuss the main findings of the analyses and experiments carried out in this work. Specifically we first present the newly curated positive class of the Manakov2022 dataset accompanied by biologically relevant descriptive metrics comparing it to the positive class of existing datasets (Hejret et al., 2023; Klimentová et al., 2022). We then present the identification of a miRNA frequency class bias following in silico generation of the negative class for this dataset using a published method (Hejret et al., 2023; Klimentová et al., 2022). We present results from an experiment designed to isolate the effect of this bias on model performance in various datasets. To mitigate this bias we implement a novel method for generating negatives and repeat the experiment on the bias-corrected datasets. We adapt the experiment to ensure we do not inadvertently introduce a bias in the target site frequencies. We present and discuss the results of all these experiments. Finally, in accordance with our first objective, we present the structure of the final three bias-corrected dataset collections (Manakov2022, Hejret2023, Klimentova2022), and demonstrate their accessibility and ease of use.

For our second objective, we first review the state of the art and provide a detailed comparison throughout the Background and Literature Overview in Chapter 2. In this chapter, we present a comparison of benchmarking six state-of-the-art miRNA target site prediction tools on the original biased version of the Hejret dataset (Hejret et al., 2023) and on the bias-corrected Hejret2023 dataset presented in this work. We then present and discuss results from benchmarking the same tools on all four bias-corrected evaluation sets presented here (Manakov2022 test and left-out sets, Hejret2023 test set, Klimentova2022 test set).

For our final objective, we present the results from training and benchmarking three tree–based models (Decision Tree, Random Forest, XGBoost), and demonstrate the effect of retraining the biased convolutional neural network (CNN) from Hejret et al. (2023) on the bias-corrected Hejret2023 dataset and on the larger Manakov2022 dataset. Furthermore, in one final experiment, we explore the effect of enhancing raw read data representations with intermolecular binding information derived from co-folding methods.

4.1 | A Novel miRNA–Target Site Dataset Comprising 1.4 M Examples

The recent availability of vast raw read data produced by the chimeric eCLIP experiment (Manakov et al., 2022), provided a unique opportunity to curate a new larger miRNA–target site interaction dataset, that is useful for training machine learning models on the task of miRNA target site prediction.

Following extensive processing of raw read data downloaded from selected experiments as described in Chapter 3, Sections 3.2.1 and 3.2.2, we successfully curate the positive class for the Manakov2022 dataset that includes 1.4 M examples; a substantial increase over previous chimeric datasets, which typically contain only tens of thousands of examples, representing an improvement by approximately two orders of magnitude. The dataset captures miRNA–target site interactions across 1,230 unique human mature miRNA sequences, from the current total known 2,654 sequences (Kozomara et al., 2019).

Preliminary descriptive analyses show that the target sites are found across various locations on the genome, and that interactions bear a variety of seed-types or no seed at all (see Figure 4.1). For comparison, these analyses include the *positive class* of the publicly available Hejret2023 and Klimentova2022 datasets.

In Klimentova2022, only 23% of the target sites fall within regions annotated as 3′ UTRs (Figure 4.1A). By contrast, Hejret2023 and Manakov2022 report higher proportions of 3′ UTR interactions, at 37% and 41%, respectively. Overall, our observations suggest that chimeric eCLIP data (Klimentova2022, and Manakov2022) tend to yield more intronic target sites than CLASH (Hejret2023), although this conclusion is limited by dataset size.

Notably, most existing miRNA target prediction tools restrict their analysis to 3′ UTR regions, occasionally including exonic sites but typically excluding introns, owing to existing preconceptions in the field. This contrasts with the binding site distributions

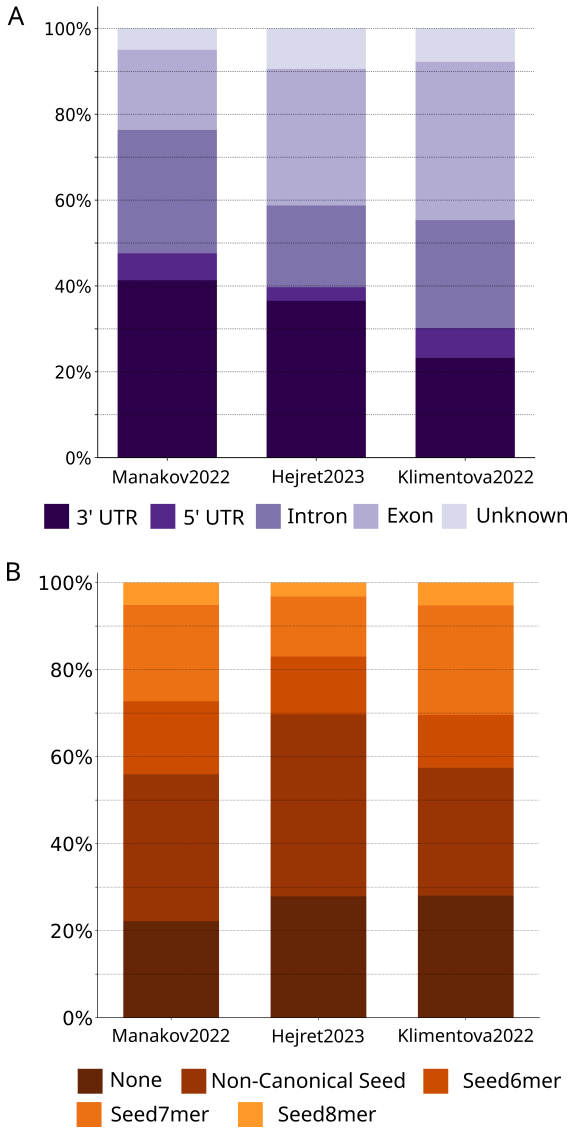

Figure 4.1: Barplots for (A) the distribution of target site genomic feature annotation, and (B) the distribution of miRNA–target site seed types, for the recovered chimeric interactions across the Manakov2022, Hejret2023, and Klimentova2022 datasets.

reported here, indicating a possible blind spot in current prediction approaches that may overlook a substantial fraction of regulatory interactions.

Since miRNA target prediction often relies heavily on seed-based criteria, we examine how frequently canonical seed matches occur within binding site sequences across the datasets analysed (Figure 4.1B). With reference to the seed definitions described in Chapter 3, Figure 3.4, canonical seed interactions account for less than half of the inter-

actions in each dataset: 44% in Manakov2022, 30% in Hejret2023, and 43% in Klimentova2022. Among these, Seed7mer interactions are the most common canonical type, representing 22%, 14%, and 25% of the total in Manakov2022, Hejret2023, and Klimentova2022, respectively.

Notably, interactions with no exact match to any canonical seed (*None*), make up around 22% of Manakov2022, and 28% of each of Hejret2023 and Klimentova2022. It is also noted that the chimeric eCLIP-based protocols employed in Manakov2022 and Klimentova2022 may have a greater propensity to capture canonical seed interactions than the CLASH approach used in Hejret2023, though this interpretation is limited by the available dataset sizes.

4.2 | Identification of a Prevalent microRNA Frequency Class Bias

The task of miRNA target site prediction is modeled as a binary classification problem when implementing a machine learning (ML) approach. This means that examples representing two categories or classes are required. Ligation-based CLIP experiments produce chimeric reads that are leveraged to computationally derive miRNA–target site interactions that would then serve as examples in the positive class. Publicly available experimentally validated negative examples are scarce, and strongly affected by preconceptions of miRNA targeting in the field. ML models for miRNA target site prediction must therefore rely on in silico generated negative examples.

In this work, we first generated negatives by reproducing the method implemented by Hejret et al. (2023) and Klimentová et al. (2022). For each target site in the positive class, the authors assigned a miRNA sequence that is not already associated with that target site in the positive class (see Chapter 3, Section 3.2.3, and Figure 3.3A). Using this method, we were able to produce new Manakov datasets at three different class imbalance ratios: 1:1, 1:10, and 1:100 positive-to-negative examples[1].

Preliminary standard exploratory data analysis reveals severe bias in the miRNA frequency between the positive and negative classes (see Figure 4.2, *Biased Dataset* in red). We identify a difference between the miRNA frequencies occurring in each class. We attribute the bias to the implemented negative example generation method reproduced from Hejret et al. (2023) and Klimentová et al. (2022). We term this artifact the

[1] `https://zenodo.org/records/13909173`. Last accessed June 2025.

miRNA frequency class bias. This is solely observed for the miRNA sequences. The target sequences did not show any critical bias.

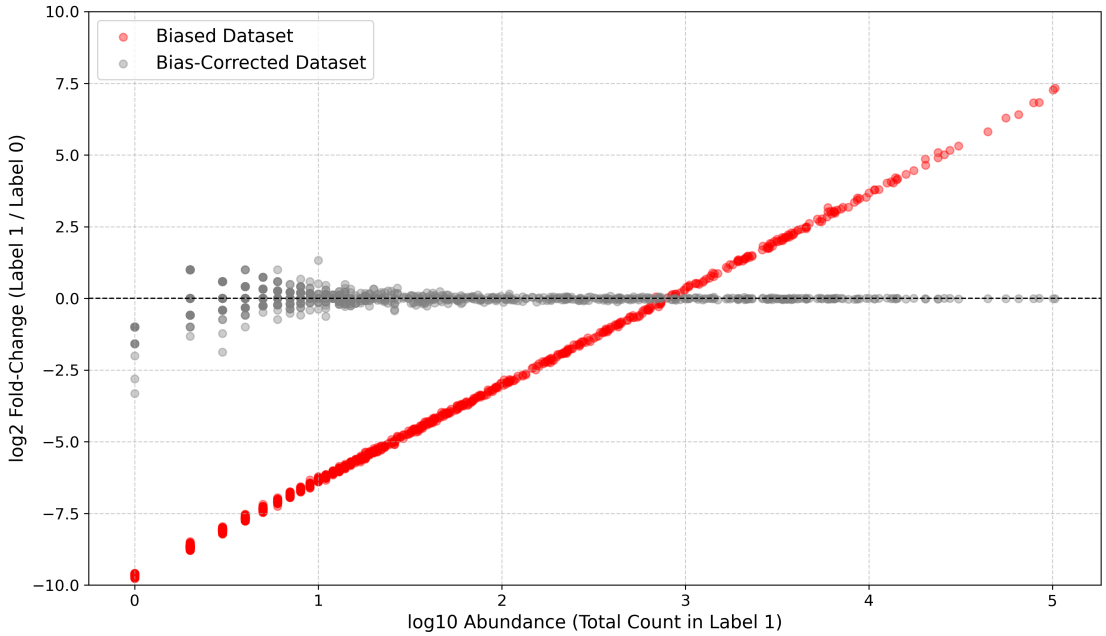

Figure 4.2: A scatter plot showing per miRNA log2 Fold-Change (FC) of the ratio of its frequency in the positive class (Label 1) to that in the negative class (Label 0), against the log10 Abundance or Total Count of the miRNA in the positive class: (A) for the biased Manakov 1:1 train set (red), and (B) for the bias-corrected Manakov2022 train set (grey). Each mark is a single miRNA. Log2FC deviations from 0 infer over-representation in one class, depending on the direction. Such deviations in the more abundant miRNAs (high log10 Abundance) are the most influential on model training. The severe miRNA frequency class bias (red) is a by-product of the negative generation method that produces negatives *per target site* in the positive class.

To investigate the impact of the identified bias, decision tree classifiers were trained and evaluated on k-mer counts computed from miRNA sequences alone, as described in Section 3.2.5. Since this input excludes any information regarding the interaction between the miRNA–target site sequences, we expect that any models trained on these features alone, do not perform any better than random chance. The experiment practically isolates the effect of any underlying biases related to the miRNA sequences.

This was carried out on the newly curated Manakov dataset, as well as on publicly available datasets from Hejret et al. (2023); the miRAW datasets from Pla et al. (2018) that are widely used in the field; and datasets from T.-H. Yang et al. (2024) who claim

to have curated information-balanced miRNA–target site datasets. A random classifier was also implemented and evaluated on each test set.

The average precision scores (APS) for the decision tree classifiers trained on datasets from Hejret et al. (2023), the miRAW datasets, and the newly curated Manakov datasets, are notably higher than that of the random classifier (see Table 4.1). These results confirm that even simple ML models are capable of exploiting underlying data biases, such as the miRNA frequency class bias identified here, producing artificially high performance metrics.

Table 4.1: Per-dataset evaluation of Decision Tree (DT) models trained on miRNA sequences only, from biased datasets. For each dataset, a model was trained on its own train set and evaluated on the corresponding test set. Average Precision Score (APS) is reported for both the Random Classifier baseline and the independently trained DT model per dataset.

Model	Hejret Test	miRAW Test	Manakov Test	Yang Test
Random Classifier	0.489	0.482	0.500	0.505
Trained DT Model	0.749	0.853	0.976	0.428

Models trained on miRNA interaction datasets featuring such biases risk learning superficial correlations in miRNA frequency rather than biologically meaningful features such as base-pair complementarity, structural features, or true binding patterns. This undermines the model's ability to generalise, as performance becomes dependent on the specific distribution of miRNAs in the training data. Consequently, a model that learns from these spurious patterns may perform poorly when evaluated on datasets with different miRNA distributions.

The decision tree classifier trained and evaluated on the Yang datasets exhibits close to random performance ($APS = 0.428$). This indicates that the method for selecting negatives presented by T.-H. Yang et al. (2024), and used to train the `InteractionAwareModel`, is an effective method at eliminating the miRNA frequency class bias.

To generate negative examples, T.-H. Yang et al. (2024) leverage the human transcriptome (GENCODE version 38). They first exclude any transcripts that are suggested to bind with a given miRNA according to AGO1 CLASH chimeric interactions from Helwak et al. (2013) or from the literature. From the remaining transcripts, they randomly select 40-nucleotide fragments. These are further filtered to remove sequences overlapping with positive regions, and those with a negative binding energy derived from `RNAup` for computing RNA–RNA interaction thermodynamics. The latter helps

to avoid choosing obvious negatives. They perform this for each of the 567 miRNAs in the positive class, until they generate roughly the same number of negative examples as positives, while ensuring that the miRNA distribution in the negative class mirrors that in the positive class.

The drawback of this method is that drawing negative examples from the full transcriptome may introduce sequence-level artifacts or distribution shifts. These differences in background sequence characteristics could lead models to distinguish classes based on transcript-specific features, such as sequence composition or structure, rather than genuine miRNA–target site interaction pattern. For this reason we chose to avoid sampling the transcriptome for alternative non-binding sites in our negative example generation strategy, and instead leverage transcripts from the same dataset derived from a common cellular context, but show no evidence of chimeric interactions with the respective miRNAs.

4.3 | Novel Method for Generating Negatives Produces Unbiased Datasets

To address the limitations of existing strategies for generating negatives, a novel method was developed that preserves the same miRNA frequencies across the positive and negative classes. As described in Chapter 3, Section 3.2.7, and illustrated in Figure 3.3B, instead of assigning an alternative miRNA to each positive target site, negatives were generated by assigning an alternative target site to each miRNA in the positive class, selected from within the same dataset. This strategy propagates the list of miRNAs in the positive class to the negative class, thereby maintaining an identical miRNA frequency distribution (see Figure 4.2, *Bias-Corrected Dataset* in grey). Additionally, this approach ensures that both classes are derived from the same transcript pool, maintaining a consistent cellular and sequence context, and reducing the risk of introducing artifacts that models could exploit.

Decision tree classifiers were again trained on *k*-mer frequency matrices derived from miRNA sequences, using the bias-corrected Hejret2023 and Manakov2022 datasets. Upon evaluation, all models exhibited performance comparable to that of a random classifier (Table 4.2, *Input: miRNA*). Notably, the model trained on the corrected Hejret2023 train set and evaluated on its corresponding test set ($APS = 0.387$) performed slightly below random ($APS = 0.510$). This is due to the chromosome-based strategy used for the train–test split.

Table 4.2: Per-dataset evaluation of Decision Tree (DT) models trained on bias-corrected datasets. For each dataset, a model was trained on its own train set and evaluated on the corresponding test set. Average Precision Score (APS) is reported for both the Random Classifier baseline and the independently trained DT model per dataset. Results are shown for two input settings: using only miRNA sequences, and using only target site sequences.

Model	Hejret Test	Manakov Test	Manakov Left-out
Input: miRNA			
Random Classifier	0.510	0.515	0.499
Trained DT Model	0.387	0.498	0.500
Input: Target site			
Random Classifier	0.508	0.515	0.513
Trained DT Model	0.498	0.524	0.505

Since the per-miRNA positives and negatives are balanced prior to splitting, and the split is based on a feature of the target site sequence (chromosome) that is independent of the miRNA, certain dataset characteristics emerge. The class balance is no longer perfect, and the miRNA frequency class bias is slightly reintroduced, although it does not appear to affect classification performance overall (Table 4.2). Finally, the positives and negatives are balanced across the train and test sets, meaning that any positive or negative examples for a given miRNA assigned to one set, are missing in the other. As a result, the miRNA frequency class bias is reintroduced with equal magnitude but in opposite directions in the two splits relative to each other. If a decision tree trained on miRNA *k*-mer frequencies learns that a particular miRNA is more often associated with positive labels in the training set, it will tend to classify that miRNA as positive in the test set as well. However, in the test set, that same miRNA will predominantly occur in negative examples. This leads to a number of false positives (FP). The same applies for miRNAs that are more frequent in the negative class of the training set, leading to an increase in false negatives (FN) when these are misclassified in the test set. Both effects inflate the FP and FN terms in the precision and recall equations (see Equations 3.2 and 3.3 in Chapter 3, Section 3.6) respectively, resulting in the observed deflated average precision score (APS).

To ensure that the new negative example generation strategy did not inadvertently introduce a similar bias in the target site sequences, additional decision tree classifiers were trained using *k*-mer counts derived from the target site sequences, as described in

Section 3.2.7.1. These also achieved random performance across datasets (see Table 4.2, *Input: Target site*), supporting the conclusion that the new method for generating negatives does not itself introduce class-distinguishing artifacts in either miRNA or target site sequence.

We compared the performance of several state-of-the-art models from the `miRBench` Python package on the original, biased test set from Hejret et al. (2023), and on the bias-corrected Hejret2023 test set, presented in this study. We also included a random classifier as a baseline for performance on each dataset. We used the average precision score (APS) metric, and noticed an overall decrease in performance across all models (see Table 4.3).

Table 4.3: Comparison of performance of state-of-the-art models on the original biased Hejret test set versus on the new bias-corrected Hejret2023 test set. A random classifier is included as a baseline. The evaluation metric is the Average Precision Score (APS).

Tool	*Original* Hejret2023 Test	*Corrected* Hejret2023 Test
`TargetScanCnn`	0.737	0.711
`CnnMirTarget`	0.630	0.530
`TargetNet`	0.661	0.575
`miRBind`	0.870	0.796
`miRNA_CNN`	0.889	0.773
`InteractionAwareModel`	0.775	0.741
`RNACofold`	0.770	0.739
Random Classifier	0.489	0.506

The evaluation of `miRNA_CNN` on the original biased Hejret2023 test set (APS = 0.889) validates our implementation of the method, as it was evaluated on the same test set by Hejret et al. (2023) with a comparable performance (PR–AUC = 0.89), as shown in Chapter 3, Table 3.2. As expected, this CNN model, originally trained on the biased dataset, showed a notable drop in performance when evaluated on the corrected dataset ($APS = 0.889$ to 0.773), reflecting its reliance on the miRNA frequency class bias present in the original training data, to accurately classify examples.

The `InteractionAwareModel` only showed minor decrease in performance. As described in the previous section, this model was trained on the Yang datasets which do not bear the miRNA frequency class bias and should therefore not be affected by the bias-correction. `RNACofold` and `TargetScanCnn` also showed minor decreases in

performance. These models were not trained on such miRNA–target site interaction datasets, and therefore should also not show any effect. A completely random model also displayed a negligible shift in APS, suggesting that the performance shifts for these models are within the range of expected noise.

Interestingly, other models such as `miRBind`, `TargetNet`, and `CnnMirTarget` also showed notable performance drops. These models were all fully or partially trained on data from the older AGO1 CLASH experiment (Helwak et al., 2013), which, like the experiments performed by Hejret et al. (2023) and Manakov et al. (2022), was conducted using the HEK293 cell line. As a result, they may share a similar miRNA frequency distribution owing to cellular context. Moreover, `TargetNet` was trained using the biased miRAW datasets (Min et al., 2022), and the negative examples for `miRBind` and `CnnMirTarget` were also generated by selecting alternative miRNAs *per target site* (Klimentová et al., 2022; Zheng et al., 2020), as in the initial method reproduced in this work, therefore leading to a similar bias. These observations suggest that the miRNA frequency class bias is not limited to a single dataset, but is instead a recurring issue across multiple datasets derived from similar experimental conditions and using similar negative example generation strategies.

Together, these results emphasise the importance of evaluating models on unbiased datasets. The drop in performance across diverse models indicates that the miRNA frequency class bias is a pervasive problem in the field. All this highlights that reliable assessment of model generalisability requires the use of unbiased benchmarks, such as those introduced in this work.

4.4 | Novel Unbiased Datasets made FAIR

We produced three dataset collections comprising six datasets: bias-corrected versions of the Klimentova2022 test set, and the Hejret2023 train and test sets, and the newly curated Manakov2022 train, test, and left-out sets.

All datasets follow a consistent structure with the following column order and definitions:

1. **`gene`**: A string of length 50 indicating the binding site sequence in the 5′ to 3′ direction.
2. **`noncodingRNA`**: A string of variable length (16–28) indicating the mature miRNA sequence in the 5′ to 3′ direction.
3. **`noncodingRNA_name`**: A string indicating the name of the miRNA.

4. **noncodingRNA_fam**: A string indicating the name of the miRNA family the miRNA belongs to.
5. **feature**: A string indicating the feature annotation on the genome where the binding site occurs.
6. **label**: A boolean value indicating whether the example belongs to the positive or negative class.
7. **chr**: A string indicating the chromosome number on the genome where the binding site occurs.
8. **start**: An integer indicating the 1-based start position of the binding site on the genome.
9. **end**: An integer indicating the 1-based end position of the binding site on the genome.
10. **strand**: A string indicating whether the binding site occurs on the '+' or '-' strand on the genome.
11. **gene_cluster_ID**: An integer indicating the cluster ID of the binding site sequence used to generate the negative class.

The datasets are approximately class-balanced (Table 4.4). Using the new method for generating negatives, it was not possible to exceed a 1:1 positive-to-negative ratio due to constraints imposed by the stricter criteria. In particular, for highly abundant miRNAs, there were not enough eligible non-binding target sites remaining to sample from, after excluding all targets from clusters containing binding sites for the given miRNA, and allowing only one target per cluster from the remaining clusters, to be selected per miRNA family. While necessary for removing bias, these criteria limit the number of valid negative examples and consequently constrain the achievable class ratio. We recognise this as a limitation of our method for generating negatives since highly imbalanced datasets, with non-binding sites (negatives) far outnumbering binding sites (positives), are more representative of the true scenario in the cellular environment.

As discussed in Section 4.3, although the new method for generating negatives mitigates the miRNA frequency class bias, the chromosome-based train-test split can reintroduce it to a small degree. This is a known and acceptable limitation of the curation process. To further mitigate this issue and better assess model generalisability, we curated a dedicated evaluation set, referred to as the Manakov2022 left-out set (described

Table 4.4: Summary of the New Dataset Collections.

Dataset	Split	No. of Positives	No. of Negatives	Total Examples	Pos:Neg Ratio
Klimentova2022	Test	477	477	954	1 : 1.0
Hejret2023	Test	495	470	965	1 : 0.9
	Train	4,084	4,109	8,193	1 : 1.0
Manakov2022	Left-out	10,027	10,027	20,054	1 : 1.0
	Test	168,342	158,787	327,129	1 : 0.9
	Train	1,253,320	1,262,875	2,516,195	1 : 1.0

in Chapter 3, Section 3.2.6). This dataset comprises interactions involving miRNAs that are exclusive to the Manakov2022 left-out set, thus excluding miRNAs that comprise all other sets in the three data collections (Manakov2022, Hejret2023, Klimentova2022) presented in this work. It therefore provides a stricter benchmark for evaluating whether models trained on the Manakov2022 or Hejret2023 train sets can generalise to truly unseen miRNA interactions.

The three dataset collections were released alongside detailed metadata on the data sharing platform, Zenodo, under record number 14501607[2], and contributed to the `miRBench` Python package[3] for miRNA target site prediction benchmarks, such that the datasets are findable, accessible, interoperable, and reusable, in accordance with the FAIR principles (Wilkinson et al., 2016).

4.5 | Benchmarking the State of the Art in miRNA Target Site Prediction

We leverage `miRBench` to access the datasets, six state-of-the-art deep learning models, and implementations of their corresponding input data encoders. We benchmarked all six models on all four evaluation sets presented in this work. To contextualise model performance, we also included several baselines: seed-based measures, a co-folding approach, and a random classifier. All evaluations were conducted using the Average Precision Score (APS) metric.

The set of benchmarked models includes convolutional neural networks (`TargetScanCnn`, `CnnMirTarget`, `miRNA_CNN`), ResNet-based models (`TargetNet`, `miRBind`), and an attention-based model (`InteractionAwareModel`). These models

[2] `https://zenodo.org/records/14501607`. Last accessed June 2025.
[3] `https://github.com/katarinagresova/miRBench`. Last accessed June 2025.

were evaluated on the bias-corrected Hejret2023 and Klimentova2022 test sets, and on the new Manakov2022 test and left-out sets.

Performance varies across models and datasets (see Figure 4.3 and Table 4.5). Within individual test sets, `miRBind` achieves the highest APS on the corrected Klimentova2022 and Hejret2023 test sets (0.753 and 0.796, respectively), while `TargetScanCnn` is the best-performing model on the Manakov2022 test and left-out sets (0.774 and 0.757).

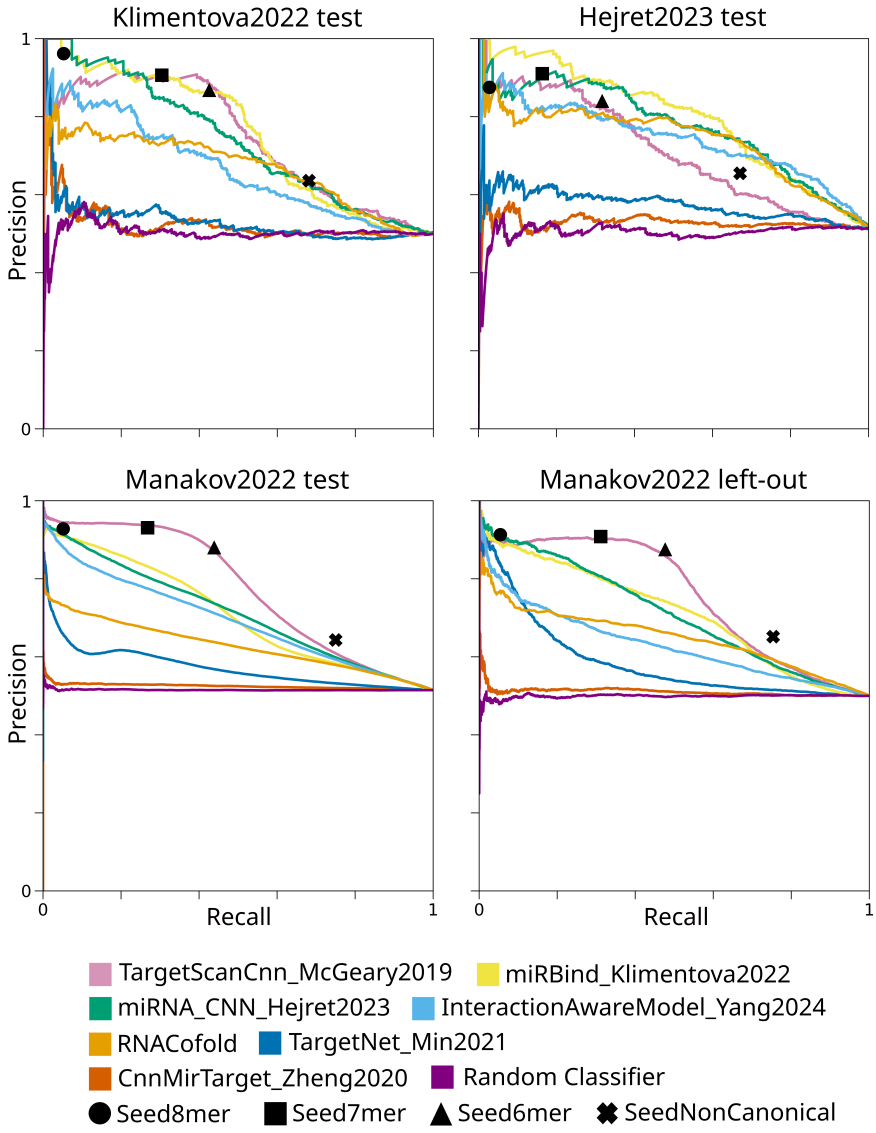

Figure 4.3: Precision–Recall curves produced from benchmarking the six state-of-the-art deep learning models, an RNA–RNA folding method, four seed classifiers, and a random baseline, on all unbiased evaluation sets presented in this work.

Table 4.5: Benchmarking of State-of-the-Art models on all unbiased evaluation sets. A random classifier is included as a baseline. The evaluation metric is the Average Precision Score (APS). The *Overall* column reports the mean ± standard deviation of APS across all evaluation sets.

Model	Klimentova2022	Hejret2023	Manakov2022		*Overall*
	Test	Test	Test	Left-out	
`TargetScanCnn`	0.745	0.711	**0.774**	**0.757**	0.747 ± 0.027
`CnnMirTarget`	0.524	0.530	0.525	0.511	0.523 ± 0.008
`TargetNet`	0.533	0.575	0.572	0.585	0.566 ± 0.023
`miRBind`	**0.753**	**0.796**	0.709	0.711	0.742 ± 0.041
`miRNA_CNN`	0.739	0.773	0.711	0.710	0.733 ± 0.030
`InteractionAwareModel`	0.664	0.741	0.692	0.631	0.682 ± 0.047
`RNACofold`	0.669	0.739	0.628	0.648	0.671 ± 0.048
Random Classifier	0.506	0.506	0.515	0.499	0.507 ± 0.007

In terms of cross-dataset consistency (Table 4.5, *Overall*), `TargetScanCnn` and `miRNA_CNN` exhibit similarly high and stable performance. `miRBind` performs well but shows higher variation across datasets. In contrast, `CnnMirTarget` and `TargetNet` demonstrate consistently lower APS scores with the lowest variation across datasets, while the `InteractionAwareModel` achieves moderate scores but among the highest variation across datasets. These results show that no individual model consistently outperforms all others across the evaluation sets, which suggests that the field remains open for the development of better models.

The inclusion of seed-based predictors and `RNACofold` provide additional baselines. These simpler methods help assess whether complex models are able to learn beyond canonical seed complementarity or traditional energy-based interaction scoring. Figure 4.3 shows that seed measures consistently outperform `RNACofold`, which is consistent with earlier benchmarks (Alexiou et al., 2009), indicating that seed complementarity is a better feature for classifying miRNA target sites than energy-based measures. As expected, the strictest seed-based predictor (`Seed8mer`) has high precision (> 0.8) but low recall (< 0.5), while the loosest predictor (`SeedNonCanonical`) drops in precision (0.6–0.8) but increases in recall (0.6–0.8), across datasets. Interestingly, none of the deep learning models consistently outperforms seed-based predictors across all evaluation sets. Since it has been repeatedly shown that non-seed interactions between some miRNAs and their target sites exist (Helwak et al., 2013; Lal et al., 2009), the field therefore remains open for the development of improved models capable of capturing the full spectrum of miRNA interactions.

Finally, the random classifier serves as a lower-bound reference to account for minor variations due to any residual class imbalance or dataset artifacts. All models exceed the performance of the random classifier, although `CnnMirTarget` showed near-random performance, suggesting that it is not able to generalise. We suspect the miRNA frequency class bias is one reason, as the negatives were generated in a way that does not preserve the miRNA distribution across the classes. This is further supported by the inflated performance of this model when evaluated on the original biased Hejret dataset (see Table 4.3).

4.6 | A Baseline for Tree-based Machine Learning Models

As proof of concept for demonstrating the utility of the curated Manakov2022 dataset, as described in Chapter 3, Section 3.4, we trained three tree-based models: a simple decision tree, a bagging ensemble random forest, and a boosting ensemble XGBoost classifier. To select suitable hyperparameters within the defined search spaces, we used Bayesian optimisation with 30 iterations and 5-fold cross-validation. Throughout the training process, the models were evaluated on each of the train and validation split per fold to assess model fit. The results obtained from the hyperparameter tuning experiment are presented in Table 4.6.

The cross-validation results show respectable performance across all models (*Val. APS* $=$ 0.795 to 0.801). Comparison of the mean evaluation score on the train versus validation splits for each model shows that the models fit the data well, with no critical signs of severe overfitting ($\max(\overline{\text{APS}}_{\text{train}} - \overline{\text{APS}}_{\text{val}}) = 0.020$).

Favourably, optimal hyperparameters landed well within the defined ranges of the search space for the Decision Tree and XGBoost models. The Random Forest model includes optimal hyperparameters identified at the very limits of the specified ranges (`min_samples_split` and `min_samples_leaf` at their lower limit corresponding to approximately 400 and 200 samples, respectively, and `n_estimators` at its higher limit) indicating that the search space may be too constrained and that it may be beneficial to explore broader ranges. While hyperparameter optimisation is typically a step-wise and iterative process, involving extensive experimentation and evaluation, the results obtained from the selected hyperparameter search spaces were considered sufficient for the scope of this work.

Table 4.6: Results from hyperparameter tuning via Bayesian optimisation (30 iterations) with 5-fold cross-validation for tree-based models on the Manakov2022 train set. Mean train and mean validation (Val.) Average Precision Scores (APS) for the optimal model hyperparameter configuration are reported.

Hyperparameter	Search space	Opt. value	Train APS	Val. APS
Decision Tree				
`max_depth`	Integer(10, 25)	21		
`min_samples_split`	Real(0.0002, 0.02, prior=*log-uniform*)	0.001		
`min_samples_leaf`	Real(0.0001, 0.01, prior=*log-uniform*)	0.0007	0.805 ±0.005	0.795 ±0.020
`criterion`	Categorical([*gini*, *entropy*])	*entropy*		
`max_features`	Categorical([*sqrt*, *log2*, 0.25, 0.5])	0.5		
Random Forest				
`max_depth`	Integer(10, 25)	21		
`min_samples_split`	Real(0.0002, 0.02, prior=*log-uniform*)	0.0002		
`min_samples_leaf`	Real(0.0001, 0.01, prior=*log-uniform*)	0.0001	0.819	0.801
`criterion`	Categorical([*gini*, *entropy*])	*entropy*	±0.005	±0.020
`max_features`	Categorical([*sqrt*, *log2*, 0.25, 0.5])	0.25		
`n_estimators`	Integer(100, 300)	300		
XGBoost				
`max_depth`	Integer(3, 10)	8		
`n_estimators`	Integer(100, 300)	287	0.821 ±0.005	0.801 ±0.021
`eta`	Real(0.01, 0.5)	0.05		

Following identification of a suitable hyperparameter configuration per model, we trained each model on the entire Manakov2022 train set, leveraging all ∼2.5 M examples, and evaluated each model's performance on the Manakov2022 test and left-out datasets; both held-out during training therefore containing previously unseen examples. This serves to evaluate the models' generalisability to new data. The results obtained are presented in Table 4.7.

Table 4.7: Evaluation of the final tree-based models following hyperparameter optimisation, trained on the full Manakov2022 train set, and evaluated on the Manakov2022 test and left-out sets. Evaluation of a state-of-the-art (SotA) model is included for comparison. The evaluation metric is the Average Precision Score (APS).

Model	Test	Left-out
`TargetScanCnn` (SotA)	0.774	0.757
Decision Tree	0.812	0.798
Random Forest	0.818	0.802
XGBoost	**0.820**	**0.803**

The final evaluation shows that all three tree-based models trained on the new Manakov2022 dataset exceed the performance of the state of the art, `TargetScanCnn` model, established from earlier benchmarking on the Manakov2022 test and left-out datasets, by at least 4%. Out of the three tree-based models, it is unsurprising that the scalable XGBoost model is the best performing model ($APS = 0.803$ on the left-out set), albeit by a small margin (< 1%). The Random Forest performs similarly ($APS = 0.802$ on the left-out set). The simpler Decision Tree also performs surprisingly well ($APS = 0.798$ on the left-out set). This indicates that the splits of a single tree capture the dominant decision boundary, and that bootstrapping with aggregation in Random Forest (designed to reduce variance) and sequential residual-correction in XGBoost (designed to reduce bias) both yield minimal additional benefit. The lack of a notable discrepancy between the performance of the simple Decision Tree and that of the ensemble methods may indicate a plateau for tree-based modeling, and the need to explore other data representations or model architectures.

Nevertheless, these tree-based machine learning models serve as a strong first baseline for models trained on the new, unbiased, much larger Manakov2022 dataset.

4.7 | CNN Retraining Surpasses State-of-the-Art Performance

To assess the impact of dataset bias on model performance, we retrained the original `miRNA_CNN` model that was trained on the biased Hejret train set, using the corrected Hejret2023 train set that is free from the miRNA frequency class bias. The original architecture and hyperparameters were left unchanged to isolate the effect of training data quality. Although this may not represent the optimal model configuration for these datasets, the retraining experiment enables a comparison solely focused on training data quality.

Retraining on the corrected version of the dataset led to a clear performance improvement ($APS = 0.865$ vs 0.773 on the Hejret2023 test set in Table 4.8, *Sequence-only*), confirming that bias in the training data was indeed limiting the model's learning potential. Furthermore, it outperformed all previously benchmarked models across all evaluation sets (Table 4.5), demonstrating that data quality alone can drive significant improvements. The performance improvement is also reflected by the higher top-right precision-recall curves, shown in Figure 4.4.

Table 4.8: Evaluation of published `miRNA_CNN` architecture retrained on bias-corrected Hejret2023 train set and on the larger, unbiased Manakov2022 train set, using the *Sequence-only* and *Sequence & Co-folding* data representations. Evaluation of the original CNN model is shown for comparison. All evaluations were carried out using the unbiased evaluation sets. The evaluation metric is the Average Precision Score (APS).

Train set	Klimentova2022	Hejret2023	Manakov2022	
	Test	Test	Test	Left-out
Sequence-only				
Original Hejret (Biased)	0.739	0.773	0.711	0.710
Hejret2023	0.774	0.865	0.785	0.783
Manakov2022	0.842	0.842	0.841	0.806
Sequence & Co-folding				
Hejret2023	0.777	0.866	0.782	0.784
Manakov2022	0.846	0.847	0.842	0.806

Notably, the retrained model performs best on the Hejret2023 test set ($APS = 0.865$), with lower performance on other test sets ($APS = 0.774$ to 0.785). This suggests that experiment- or dataset- specific features may have been learned, limiting generalisability. The Hejret2023 dataset was generated using the CLASH technique, while the Klimentova2022 and Manakov2022 datasets were derived from chimeric eCLIP experiments. In addition, the Hejret2023 dataset contains a higher proportion of non-seed and non-canonical seed interactions (see Figure 4.1B), which may bias the model towards these less common patterns, reducing its ability to generalise across datasets with different interaction profiles. Given these discrepancies, reporting performance on each dataset independently in future benchmarking efforts is advisable to better account for dataset-specific biases. Additional experimental datasets may help expose experiment-specific biases that affect model performance.

Furthermore, we retrained the same model on the larger Manakov2022 train set. Performance improved across all test sets; with average precision scores ranging from 0.806 to 0.842. The lower APS observed on the Manakov2022 left-out set, relative to the test set, may suggest the existence of miRNA-specific binding rules, as this evaluation set was specifically curated to contain a completely unseen subset of interactions involving miRNAs unique to that set. The concept of miRNA-specific binding rules remains underexplored in miRNA target site prediction methods.

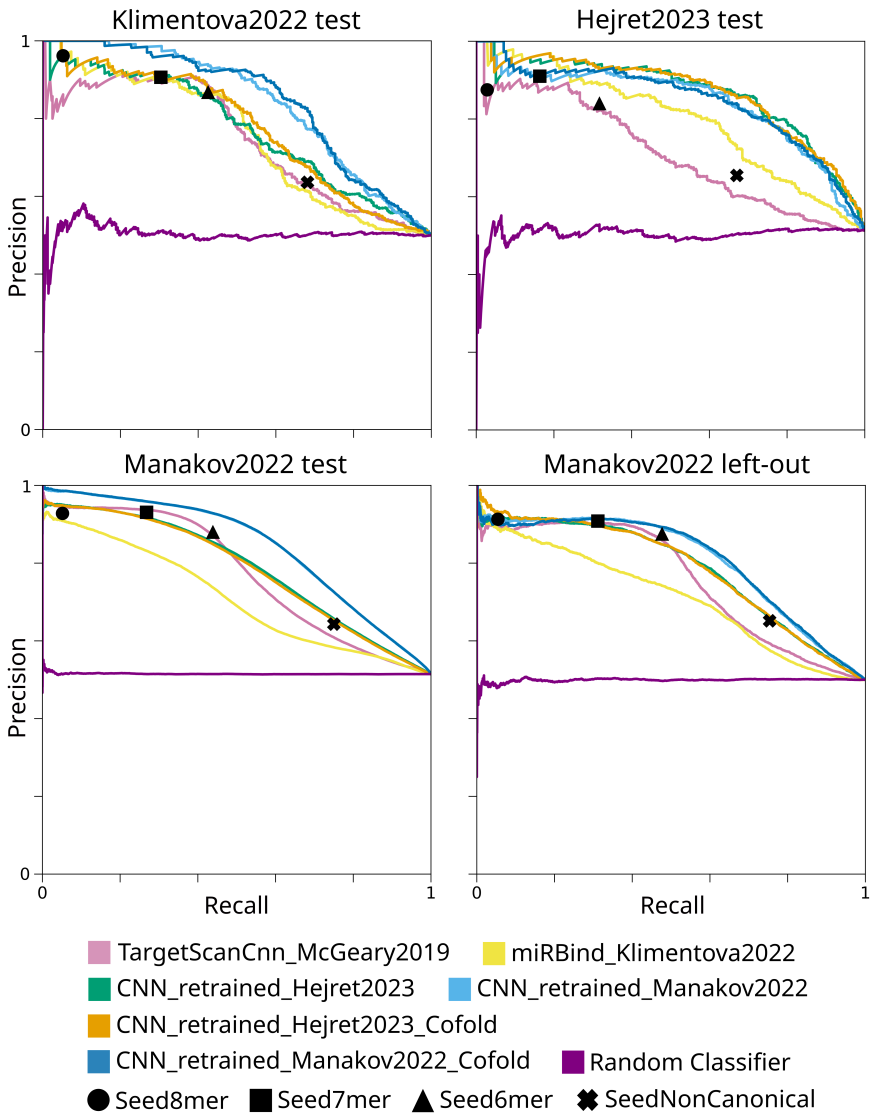

Figure 4.4: Precision–Recall curves produced from evaluating the same CNN architecture retrained on the unbiased train sets and two variations of the data representation (*Sequence-only* and *Sequence & Cofolding*) on all unbiased evaluation sets presented in this work. A state-of-the-art deep learning model, an RNA–RNA folding method, four seed classifiers, and a random baseline are also included.

We also note that the Precision–Recall curves for this model consistently exceed the performance of the rule-based seed classifiers implemented as a baseline in this study (see Figure 4.4). This observation suggests that the model learns class-distinguishing features beyond traditional seed-based measures, highlighting the importance of con-

sidering the full spectrum of interactions rather than relying solely on seed-based features when prioritising candidate target transcripts.

Overall, the improved performance demonstrates that even a relatively simple CNN architecture benefits from training on a well-curated, larger dataset. The increased data diversity and scale enables the model to generalise more effectively, highlighting the value of comprehensive dataset curation.

In addition, we explored whether richer input representations could enhance model performance. The CNN was retrained once more on the Hejret2023 train set and another time on the Manakov2022 train set, using the $50 \times 20 \times 2$ *Sequence & Co-folding* data representation described in Chapter 3, Section 3.5.2, that integrates co-folding information derived from the `RNACofold`-predicted dot–bracket structure as an additional channel to the original Watson–Crick complementarity matrix. These runs only show improvement in the 3rd decimal of the APS, if any, relative to the original 1-channel *Sequence-only* representation (see Table 4.8, *Sequence & Co-folding*). This may warrant further exploration in future model development.

These results were achieved without any hyperparameter optimisation or architectural tuning, indicating that, given the purpose of the experiment here, the existing configuration was sufficiently suited for the scale and complexity of the datasets. However, further experimentation on the model architecture and training pipeline is highly recommended in future works.

Together, these results reinforce that unbiased and well-structured datasets are critical for developing accurate and generalisable models, and that enhancements to the data representation warrants further exploration and may offer performance improvements.

4.8 | Summary

In this chapter, we first revisit the aims and objectives of the study, along with the analyses and experiments designed to address them. This is followed by the presentation and discussion of all results produced in this work. In this section, we summarise the main findings.

We successfully curate a new benchmarking dataset: the Manakov2022 dataset, which includes $\sim$1.3 M positive examples, equivalent to two orders of magnitude larger than previous datasets. Preliminary descriptive analysis of this dataset, and the existing Hejret and Klimentova datasets, reveals that target sites may occur in genomic regions beyond the 3′ UTR suggesting such regions should be considered in miRNA target site

prediction methods. Further analysis reveals that although canonical seed binding is the primary determinant of target recognition in miRNA target prediction methods, it accounts for less than half of interactions, and non-seed binding comprises 22–28% of interactions across our datasets.

We generate negative examples by reproducing an existing method from the literature, and identify a miRNA frequency class bias that derives from it. The bias inflates performance metrics and is prevalent in the field, such as in the Hejret (Hejret et al., 2023), Klimentova (Klimentová et al., 2022), and miRAW datasets (Pla et al., 2018). We observe that the Yang datasets (T.-H. Yang et al., 2024) are not affected by this bias and note that the negative example generation method used for these datasets ensures consistent miRNA frequency distribution across the classes. However, target sites are drawn from the human transcriptome which may introduce other sequence-level biases.

We therefore develop a new method for generating negatives and produce bias-corrected datasets. We test our method by training independent simple machine learning models on miRNA sequences only and on target site sequences only, and show that they exhibit random performance suggesting that there are no artifacts in either sequence that distinguish the classes in the new bias-corrected datasets.

When we evaluate state-of-the-art models on these datasets, we notice an overall performance drop relative to evaluation carried out on the biased datasets. This suggests that the identified bias is a pervasive problem in the field and emphasises the importance of benchmarking models on well-curated, unbiased datasets for a reliable assessment of model generalisability.

We finally present three dataset collections: Manakov2022, Hejret2023, and Klimentova2022; including two trains sets and four evaluation sets. Importantly, the Manakov2022 left-out set comprises interactions involving completely unseen miRNAs which is useful for testing true generalisability. We contribute all these datasets to the `miRBench` Python package which also provides access to state-of-the-art models and implementations of the corresponding input data representations. Collectively, the framework facilitates benchmarking for miRNA target site prediction methods.

We leverage the package to benchmark six state-of-the-art deep learning models on all four evaluation sets curated in this study. Results show that no single model outperforms all others across all datasets. The field therefore remains open for the development of better models.

We present results from training a decision tree, a random forest, and an XGBoost classifier as proof of concept for dataset usage, while establishing a baseline for simpler machine learning models. We show that ensemble methods yield minimal additional benefit over a single decision tree, which may indicate a plateau for tree-based mod-

els and the need to explore other data representation or model types. We then present results from retraining a published CNN configuration, that was originally trained on a biased dataset, on a bias-corrected version of the same dataset and observe overall performance improvements. This suggests that the bias was limiting the model's learning potential, and that data quality alone can drive marked improvement. We also explore the effect of retraining the CNN on the larger Manakov2022 train set. The overall higher performance shows that increased data diversity and scale can help models generalise. Notably, performance varies across evaluation sets which suggests the presence of experiment- or dataset- specific features. It is therefore advisable to benchmark on all datasets. We acknowledge that additional experimental datasets may help expose these biases. A lower APS on the left-out set may suggest miRNA-specific binding patterns which remains underexplored in the field. In one final experiment, we show that data representation enhanced by structural features may warrant further exploration in model development. A comparison of the performance of all models trained in this study against the previous state of the art, is shown in Figure 4.5.

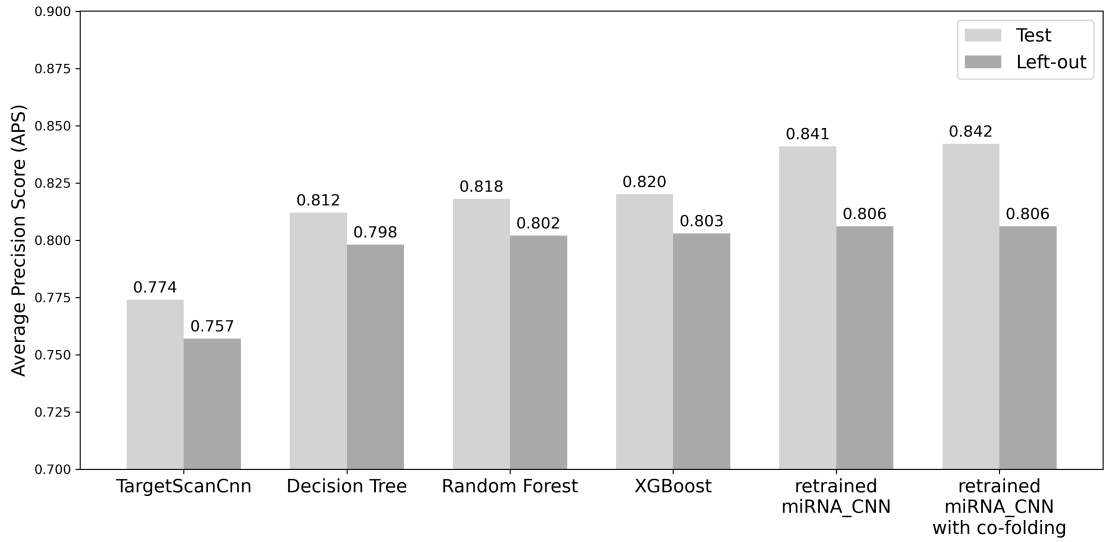

Figure 4.5: A barplot of the Average Precision Score (APS) of all models trained on the Manakov2022 train set, and evaluated on the Manakov2022 test and left-out sets, throughout this study. The previous state of the art (`TargetScanCnn`) is also included.

5

Conclusions

MicroRNAs are small, non-coding RNA molecules that regulate gene expression post-transcriptionally by targeting specific messenger RNA transcripts and repressing their translation into protein. They have been described as *master regulators* in a variety of fundamental biological processes, such as cell differentiation and proliferation (Ivey & Srivastava, 2010). Consequently, when deregulated, miRNAs have been linked to various diseases, including cancer (Calin & Croce, 2006). Given their key role in regulatory networks, miRNAs have emerged as promising tools for both molecular diagnostics and therapeutic intervention (Condrat et al., 2020; Hanna et al., 2019).

Identifying miRNA–target interactions is therefore important for understanding the regulatory role of miRNAs in complex biological systems. However, because a miRNA can, in principle, target any RNA transcript, experimentally validating all possible interactions is unfeasible due to the associated cost and time. Computational methods have thus been developed to prioritise candidate targets.

The first step in most miRNA target prediction tools is the identification of miRNA target sites on potential RNA target transcripts. This is traditionally based on the presence of Watson–Crick base pairing between the miRNA 5′ seed region and the 3′ UTR of the target transcript. Several such canonical and non-canonical configurations have been described (Bartel, 2018). However, functional miRNA–target interactions lacking base pairing in the seed region have been recognised since the early days of miRNA research (Helwak et al., 2013; Lal et al., 2009). Consistent with earlier observations, in this study we find that non-seed binding comprise 22–28% of positive interactions across all experimentally derived datasets presented here. Additionally, a CNN model we retrain on one such dataset shows higher precision and recall than simpler rule-based seed classifiers, suggesting the presence of class-distinguishing features beyond seed-based measures. As a result, rule-based approaches have been limited by high false positive and false negative rates (Alexiou et al., 2009; Pinzón et al., 2017).

The development of experimental methods that capture interacting RNAs mediated by a protein has been instrumental in addressing this limitation. Techniques such as

CLASH introduce a ligation step that joins the miRNA to its target site sequence while loaded onto the AGO protein (Helwak et al., 2013). The protein is then immunoprecipitated and the RNA is isolated. This is followed by standard library preparation and sequencing steps, enabling the recovery of chimeric reads that contain both interacting sequences. Effectively, such methods provide a resource of experimentally derived miRNA–target site interactions.

Specialised bioinformatics pipelines have been developed to detect these interacting sequences from chimeric reads. In this study we employ and adapt the HybriDetector pipeline (Hejret et al., 2023), which produces structured datasets containing pairs of miRNA–target site sequences that can be used in machine learning (ML) approaches to miRNA target site prediction. In this work, we leverage such datasets from Hejret et al. (2023) and Klimentová et al. (2022). The ML task is specifically framed as binary classification, where experimentally derived interactions form the positive (binding) class. Experimentally confirmed non-binding interactions are scarce; thus, negative examples are generated in silico using various strategies. In this work, we initially reproduce the method implemented by Hejret et al. (2023) and Klimentová et al. (2022), however we discover a miRNA frequency class bias in the datasets produced, that is a by-product of the reproduced method. We identify the bias in various datasets (Hejret et al., 2023; Klimentová et al., 2022; Pla et al., 2018) that have been used to train several state-of-the-art models (Hejret et al., 2023; Klimentová et al., 2022; Min et al., 2022; Pla et al., 2018), suggesting it is a pervasive problem in the field. We therefore propose a strategy that corrects the identified bias, and implement it to correct the bias in the datasets from Hejret et al. (2023) and Klimentová et al. (2022). The resulting binary-labeled datasets are then used to train models to predict the probability of binding.

Most machine learning models rely on engineered features informed by prior biological knowledge and assumptions in the field. Deep learning models, however, automate this process and can learn relevant patterns directly from raw sequence-level data, potentially capturing signals relevant to the interaction that earlier machine learning approaches may overlook. In this study, we review the literature and identify several state-of-the-art deep learning models, including CNNs, ResNets, and an attention-based model (Hejret et al., 2023; Klimentová et al., 2022; McGeary et al., 2019; Min et al., 2022; T.-H. Yang et al., 2024; Zheng et al., 2020).

However, the validity of their reported performance is often unclear. A variety of evaluation strategies are employed, and no standardised framework exists to support fair benchmarking in this fast-moving field. This gap motivates our work. Additionally, the recent availability of vast raw data from a new chimeric eCLIP method (Manakov et al., 2022), which yields a higher recovery of chimeric reads presents an opportunity.

This data had not yet been used for machine learning approaches to miRNA target site prediction. In this study, we curate structured benchmarking datasets from this data and contribute them to an open-source, easy-to-use Python package for benchmarking miRNA target site prediction tools (Sammut et al., 2025b). We benchmark six state-of-the-art deep learning models identified from the literature, and train and benchmark several models using the new datasets, establishing a baseline for future model development.

5.1 | Revisiting the Aims and Objectives

The aim of this study was to standardise benchmarking in the field of miRNA target site prediction. With the work presented here, we believe that we have achieved our aim by:

- presenting a novel way for generating negative examples for miRNA target site interaction datasets that mitigates the identified prevalent *miRNA frequency class bias* in the field,
- producing new, easy-to-use, significantly larger, bias-corrected datasets, that have been made publicly available, to train and benchmark deep learning models for miRNA target site prediction, and
- benchmarking and retraining state-of-the-art models on the new datasets, thereby establishing a baseline for the field.

Specifically, we have achieved our predefined objectives as follows:

Objective 1. To produce carefully-curated benchmarking datasets and make them FAIR, which lowers the barrier of entry to the field for machine learning experts.

We have produced three collections of class-balanced, bias-corrected datasets as follows (approximate total number of examples is indicated in brackets):

- new versions of:
  - the Klimentova2022 Test ($\sim 1K$) set
  - the Hejret2023 Train ($\sim 8K$) and Test ($\sim 1K$) sets, and
- the newly curated Manakov2022 dataset, including Train ($\sim 2.5M$), Test ($\sim 300K$), and Left-out ($\sim 20K$) sets

We have uploaded the datasets to the data sharing platform, Zenodo, under record number 14501607[1], accompanied by detailed metadata. We have contributed the datasets to the `miRBench` Python package[2], for miRNA target site prediction benchmarks. In this way, **the datasets are findable, accessible, interoperable, and reusable, in accordance with the FAIR principles** for data (Wilkinson et al., 2016).

Objective 2. To review and evaluate the state of the art using the curated datasets.

We have extensively reviewed the current state of the art in the field throughout this entire study. **We have benchmarked a total of six deep learning miRNA target site prediction models on the four new evaluation sets** presented in this work.

Objective 3. To produce baseline models by training and benchmarking classical machine learning and neural network models on the curated datasets.

We have trained simple tree-based machine learning models and retrained the biased CNN model from Hejret et al. (2023) on the new, unbiased version of the Hejret2022 train set, and on the new, much larger Manakov2022 train set. **We have set a new baseline for benchmarking miRNA binding site prediction models**, with the published CNN architecture from Hejret et al. (2023) retrained on the Manakov2022 train set.

Beyond the objectives, we have also identified a miRNA frequency class bias that is prevalent in the field and that affects the performance of state-of-the-art models. The bias results from the in silico method used to generate negative examples, whereby the target site sequences from all the examples in the positive class are propagated to the negative class and paired to alternative miRNA sequences. This results in different miRNA frequencies across the positive and negative classes, producing a bias in the dataset. Models trained on such datasets would learn the bias and classify unseen examples based on the miRNA frequency in the classes.

A new method for generating negatives that preserves the miRNA frequencies across the positive and negative classes has been proposed. In the new method, instead of the target site sequences, the miRNA sequences from all the examples in the positive class are propagated to the negative class and paired to alternative target site sequences, Using this method, we have corrected the previously published datasets: Klimentova2022 and Hejret2023, and produced the new, larger unbiased Manakov2022 dataset, all of

[1] `https://zenodo.org/records/14501607`. Last accessed June 2025.
[2] `https://github.com/katarinagresova/miRBench`. Last accessed June 2025.

which can be used to train and evaluate models on the task of miRNA target site prediction.

5.2 | Critique and Limitations

Having achieved our aims and objectives, in this section we acknowledge and discuss the limitations of our work.

In this study, we leveraged publicly available raw data produced by the chimeric eCLIP experimental method described by Manakov et al. (2022) in their pre-print publication on bioRxiv. To the best of our knowledge, this study has not yet undergone peer review or been accepted in a scientific journal. This introduces potential risks of undetected errors, misinterpretations, or unsupported claims, and raises the possibility of future retractions or revisions that could render aspects of our work outdated or invalid. Nonetheless, Manakov et al. (2022) has been cited 19 times, according to Google Scholar (last accessed 2 June 2025), including by peer-reviewed articles. While this may reflect some level of community validation, we acknowledge that citation alone cannot substitute for formal peer review, nor does it ensure the accuracy or quality of the underlying data.

The chimeric eCLIP experiments were performed in vitro using the HEK293T cell line. Relying on this data to produce miRNA target site interaction datasets assumes that miRNA targeting behaves similarly in vivo and across different cellular contexts. However, miRNA–target interactions may be context-specific. For instance, McGeary et al. (2019) demonstrated that the concentration of AGO–miRNA complexes influences target site occupancy and repression efficacy. These factors are not accounted for in this work.

To extract chimeric reads and identify the miRNA and target site sequences from the raw FASTQ files, we used the publicly available HybriDetector pipeline (Hejret et al., 2023). As part of this pipeline, target sites are normalised to a fixed length of 50 nucleotides. This is achieved by mapping each site to its genomic interval, computing the midpoint, and then generating a new genomic interval centred at this position, extending 24 nucleotides in one direction and 25 in the other. This approach supports comparability and facilitates downstream analyses. We assume that the miRNA target site is located within this window. However, because the exact position is not precisely known, this normalisation step may shift or distort the true binding site.

Moreover, the target site sequence in the final datasets is not taken directly from the sequenced read, but rather from the reference genome at the mapped position. The

same applies to the miRNA sequence, which corresponds to the annotated reference sequence of the noncoding RNA. However, the actual sequenced reads may differ from these reference sequences due to biological variation, sequencing errors, or mapping artifacts. As a result, some binding interactions may be inaccurately represented.

No functional validation was performed on the identified interactions. As a result, models trained on this data may predict miRNA binding sites that are not functionally relevant. Such predictions are inherently difficult to validate. Moreover, functional assessment is beyond the scope of miRNA target site prediction tools, which are intended only as an initial filtering step for potential target transcripts.

Due to the scarcity of experimentally validated negative interactions, we generated negative examples in silico, which is a common approach in the field. This process relies on several assumptions. Specifically in our approach, if a miRNA and target site are not recovered as a chimeric read, it is assumed that no binding or interaction occurs. Using additional filtering criteria, these pairs were used to construct the negative class. However, this assumption is not guaranteed, since miRNAs and target site sequences that do not form chimeras under the sampled conditions may still interact in other contexts.

The final method used to generate negatives involves keeping the same list of miRNAs from the positive class and sampling alternative target sites from the positive class to assign to these miRNAs, ensuring that the resulting miRNA–target site pairs do not match any interactions in the positive class. This approach limits the number of negatives that can be produced, as the available pool of target sites to sample from is finite. As a result, class imbalance ratios greater than 1:1 could not be achieved. This restricts our ability to reflect the true biological scenario, whereby binding sites are rare among many non-binding sites.

To understand the effect of the identified miRNA frequency class bias on model generalisabilty and the effect of training on larger datasets, we retrained the published biased CNN Hejret et al. (2023) on a bias-corrected version of the original dataset, and on the new, larger Manakov2022 dataset. We adopted the same data representation: a 2D black-and-white image encoding Watson–Crick base-pairing between miRNA and target site sequences, originally described by Klimentová et al. (2022). Following their approach, miRNA sequences were truncated to the first 20 nucleotides, to ensure fixed-size input arrays. However, mature miRNAs are typically 21–25 nucleotides long. This truncation results in data loss and may exclude biologically relevant information, potentially impairing the model's ability to learn meaningful patterns and generalise accurately. Additionally, the CNN architecture includes multiple pooling layers, which result in extensive padding of the 2D representation. Consequently, the informative signal is diluted potentially reducing the model's ability to learn relevant features.

A further concern with training deep learning models on the curated miRNA interaction datasets is the potential of models to learn the pattern resulting from the method used to generate negative examples, rather than features associated with true binding versus non-binding interactions. Additionally, deep learning models are black boxes. Extracting biological insight requires complex model explainability techniques. This, however, is beyond the scope of this work.

Finally, since the curation of the then-current state-of-the-art miRNA target site prediction tools and the release of the `miRBench` package, several new tools have been published (Bi et al., 2024; Wu et al., 2025; T. Yang et al., 2024). These may offer competitive performance and warrant consideration in future benchmarking efforts.

5.3 | Future Work

In this section, we outline potential directions for extending and improving the present work.

The HybriDetector pipeline for detecting miRNA–target site interactions from chimeric reads could be further refined. One possible improvement is extending the target site window beyond 50 nucleotides to better ensure complete capture of the binding site. Additionally, replacing miRNA family annotations from miRBase with those from the more recent MirGeneDB database (Clarke et al., 2025), may enhance annotation accuracy. MirGeneDB also incorporates evolutionary context.

Future work on data representation could explore the use of padding instead of truncation for miRNA sequences, allowing retention of full-length sequence information and potentially improving model performance. Additional features could also be encoded to enrich the input, such as base-pair identity, evolutionary conservation, and structural context.

The notion of miRNA-specific rules of binding remains underexplored in the field. Future work could therefore include training miRNA-specific models, using a single miRNA and its corresponding interaction data from the presented datasets. Similarly, models could be trained on subsets of data defined by genomic context, such as the annotated genomic feature of the target site.

To improve predictive performance, hyperparameter tuning of the CNN could be conducted using optimisation tools such as Optuna (Akiba et al., 2019). Alternative architectures may also be worth exploring. For example, transformer-based models (Vaswani et al., 2017), which underpin modern large language models, have recently been applied to miRNA target site prediction (Bi et al., 2024; T. Yang et al., 2024; T.-H.

Yang et al., 2024).

Finally, once a satisfactory model is obtained, explainability methods may be applied to derive biological insights. For instance, Grešová, Vaculík, and Alexiou (2023) describe a technique applied to models that use the 2D-binding representation introduced by Klimentová et al. (2022) and adopted in this work as input. They first compute attribution scores using implementations of the widely used SHAP (SHapley Additive exPlanations) method that assigns feature-level attribution scores for individual predictions of any ML model. Furthermore Grešová, Vaculík, and Alexiou (2023) describe an attribution sequence alignment algorithm based on dynamic programming principles for semi-global alignment of the miRNA to its target site using the attribution score matrix. The output includes a sequence of attribution scores describing the importance of each position for the interaction. This facilitates visualisation of the internal representations learned by the model, offering insight into its decision-making process and a potential route to interpreting underlying biological patterns captured during training.

5.4 | Final Remarks

In this study, we have produced FAIR datasets, benchmarked the state of the art, and established a baseline for future model development in the task of miRNA target site prediction. These efforts support our aim of standardising benchmarking in the field, which we hope will catalyse progress by lowering the barrier to entry for machine learning experts. Moreover, we identified a prevalent *miRNA frequency class bias* in commonly used datasets and mitigated it by developing a novel method for generating negative examples in silico. All data and code have been made publicly available to encourage transparency and reusability.

This work has been published as a pre-print on *bioRxiv* (Sammut et al., 2025a), and following peer review, it has been accepted for presentation at ISMB/ECCB 2025 (presented by Dr Panagiotis Alexiou), the world's largest bioinformatics and computational biology conference, jointly organised by the Intelligent Systems for Molecular Biology (ISMB) and the European Conference on Computational Biology (ECCB). The work has also been accepted for inclusion in the conference proceedings, which has been published online in the leading journal in its field, *Bioinformatics* (Sammut et al., 2025b) (see Appendix A).

This work has also been presented as a poster or talk, with the author contributing as either presenting author or co-author, at the following national or international scientific venues:

- **Weekly Scientific Seminars** of the BioGeMT and TargetMI projects at UM, in March 2024
- **Microsymposium on RNA Biology** at the Vienna BioCenter, in Austria, in April 2024
- **ECCB 2024** at Logomo in Turku, Finland, in September 2024
- **18th Conference of the Hellenic Society for Computational Biology and Bioinformatics (HSCBB24)** in Lamia, Greece in October 2024
- **EMBO AI, Bioinformatics, Computation & Data (ABCD) Sectoral Meeting** in Heidelberg, Germany, in March 2025
- **Bioinformatics Seminar Series** at the Central European Institute of Technology (CEITEC) in Brno, Czech Republic, in April 2025
- **Microsymposium on RNA Biology** at the Vienna BioCenter, in Austria, in May 2025
- **University of Malta Research Expo** at UM, in May 2025
- **SCISEM Life Science Research Seminar Series** at UM, in July 2025
- **ISMB/ECCB 2025** at the ACC Liverpool in Liverpool, UK, in July 2025

This dissemination has facilitated valuable feedback and engagement across diverse research environments. These presentations reflect the sustained interest and relevance of this work within the international RNA and bioinformatics communities. By openly sharing all resources, we aim to support continued research and promote reproducible benchmarking in miRNA target site prediction. We hope this work will serve as a foundation for future efforts in the development and evaluation of machine learning models in this domain.

A

Research Paper

Most of the work presented in this dissertation has been included in an associated research paper that is cited as Sammut et al. (2025b) in the main text, and can be accessed at: https://doi.org/10.1093/bioinformatics/btaf233.

References

Agarwal, V., Bell, G. W., Nam, J.-W., & Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, *4*, e05005. https://doi.org/10.7554/elife.05005

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., Howe, K., Kähäri, A., Kokocinski, F., Martin, F. J., Murphy, D. N., Nag, R., Ruffier, M., Schuster, M., Tang, Y. A., … Searle, S. M. J. (2016). The Ensembl gene annotation system. *Database*, baw093. https://doi.org/10.1093/database/baw093

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Alexiou, P., Maragkakis, M., Papadopoulos, G. L., Reczko, M., & Hatzigeorgiou, A. G. (2009). Lost in translation: An assessment and perspective for computational microRNA target identification. *Bioinformatics*, *25*(23), 3049–3055. https://doi.org/10.1093/bioinformatics/btp565

Bartel, D. P. (2009). MicroRNAs: Target recognition and regulatory functions. *Cell*, *136*(2), 215–233. https://doi.org/10.1016/j.cell.2009.01.002

Bartel, D. P. (2018). Metazoan microRNAs. *Cell*, *173*(1), 20–51. https://doi.org/10.1016/j.cell.2018.03.006

Bernhart, S. H., Tafer, H., Mückstein, U., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2006). Partition function and base pairing probabilities of RNA heterodimers.

Algorithms for Molecular Biology, *1*(3). https://doi.org/10.1186/1748-7188-1-3

Bernstein, E., Kim, S. Y., Carmell, M. A., Murchison, E. P., Alcorn, H. L., Li, M. Z., Mills, A. A., Elledge, S. J., Anderson, K. V., & Hannon, G. J. (2003). Dicer is essential for mouse development. *Nature Genetics*, *35*(3), 215–217. https://doi.org/10.1038/ng1253

Bi, Y., Li, F., Wang, C., Pan, T., Davidovich, C., Webb, G. I., & Song, J. (2024). Advancing microRNA target site prediction with transformer and base-pairing patterns. *Nucleic Acids Research*, *52*(19), 11455–11465. https://doi.org/10.1093/nar/gkae782

Broughton, J. P., Lovci, M. T., Huang, J. L., Yeo, G. W., & Pasquinelli, A. E. (2016). Pairing beyond the seed supports microRNA targeting specificity. *Molecular cell*, *64*(2), 320–333. https://doi.org/10.1016/j.molcel.2016.09.004

Calin, G. A., & Croce, C. M. (2006). MicroRNA signatures in human cancers. *Nature Reviews Cancer*, *6*, 857–866. https://doi.org/10.1038/nrc1997

Chakraborty, C., Sharma, A. R., Sharma, G., Doss, C. G. P., & Lee, S.-S. (2017). Therapeutic miRNA and siRNA: Moving from bench to clinic as next generation medicine. *Molecular Therapy. Nucleic Acids*, *8*, 132–143. https://doi.org/10.1016/j.omtn.2017.06.005

Chen, L.-L., & Kim, V. N. (2024). Small and long non-coding RNAs: Past, present, and future. *Cell*, *187*(23), 6451–6485. https://doi.org/10.1016/j.cell.2024.10.024

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chi, S. W., Hannon, G. J., & Darnell, R. B. (2012). An alternative mode of microRNA target recognition. *Nature structural & molecular biology*, *19*, 321–327. https://doi.org/10.1038/nsmb.2230

Chi, S. W., Zang, J. B., Mele, A., & Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature*, *460*(7254), 479–486. https://doi.org/10.1038/nature08170

Chipman, L. B., & Pasquinelli, A. E. (2019). miRNA targeting: Growing beyond the seed. *Trends in genetics : TIG*, *35*(3), 215–222. https://doi.org/10.1016/j.tig.2018.12.005

Chou, C.-H., Chang, N.-W., Shrestha, S., Hsu, S.-D., Lin, Y.-L., Lee, W.-H., Yang, C.-D., Hong, H.-C., Wei, T.-Y., Tu, S.-J., Tsai, T.-R., Ho, S.-Y., Jian, T.-Y., Wu, H.-Y., Chen, P.-R., Lin, N.-C., Huang, H.-T., Yang, T.-L., Pai, C.-Y., … Huang, H.-D. (2015). miRTarBase 2016: Updates to the experimentally validated miRNA–target interactions database. *Nucleic Acids Res*, *44*(D1), D239–D247. https://doi.org/10.1093/nar/gkv1258

Clarke, A. W., Høye, E., Hembrom, A. A., Paynter, V. M., Vinther, J., Wyrożemski, Ł., Biryukova, I., Formaggioni, A., Ovchinnikov, V., Herlyn, H., Pierce, A., Wu, C., Aslanzadeh, M., Cheneby, J., Martinez, P., Friedländer, M. R., Hovig, E., Hackenberg, M., Umu, S. U., … Fromm, B. (2025). MirGeneDB 3.0: Improved taxonomic sampling, uniform nomenclature of novel conserved microRNA families and updated covariance models. *Nucleic Acids Research*, *53*(D1), D116–D128. https://doi.org/10.1093/nar/gkae1094

Cohen-Davidi, E., & Veksler-Lublinsky, I. (2024). Benchmarking the negatives: Effect of negative data generation on the classification of miRNA–mRNA interactions. *PLOS Computational Biology*, *20*(8), 1–25. https://doi.org/10.1371/journal.pcbi.1012385

Condrat, C. E., Thompson, D. C., Barbu, M. G., Bugnar, O. L., Boboc, A. E., Crețoiu, D., Suciu, N., Crețoiu, S. M., & Voinea, S. C. (2020). miRNAs as biomarkers in disease: Latest findings regarding their role in diagnosis and prognosis. *Cells*, *9*(2). https://doi.org/10.3390/cells9020276

Cui, S., Yu, S., Huang, H.-Y., Lin, Y.-C.-D., Huang, Y., Zhang, B., Xiao, J., Zuo, H., Wang, J., Li, Z., Li, G., Ma, J., Chen, B., Zhang, H., Fu, J., Wang, L., & Huang, H.-D. (2025). miRTarBase 2025: Updates to the collection of experimentally validated microRNA–target interactions. *Nucleic Acids Research*, *53*(D1), D147–D156. https://doi.org/10.1093/nar/gkae1072

Dai, R., & Ahmed, S. A. (2011). MicroRNA, a new paradigm for understanding immunoregulation, inflammation, and autoimmune diseases. *Translational research: the journal of laboratory and clinical medicine*, *157*(4), 163–179. https://doi.org/10.1016/j.trsl.2011.01.007

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/cvpr.2009.5206848

Ding, J., Li, X., & Hu, H. (2016). TarPmiR: A new approach for microRNA target site prediction. *Bioinformatics*, *32*(18), 2768–2775. https://doi.org/10.1093/bioinformatics/btw318

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., & Marks, D. S. (2003). MicroRNA targets in Drosophila. *Genome Biology*, *5*(1), 1–27. https://doi.org/10.1186/gb-2003-5-1-r1

Esquela-Kerscher, A., & Slack, F. J. (2006). Oncomirs — microRNAs with a role in cancer. *Nature Reviews Cancer*, *6*, 259–269. https://doi.org/10.1038/nrc1840

Grešová, K., Martinek, V., Čechák, D., Šimeček, P., & Alexiou, P. (2023). Genomic benchmarks: A collection of datasets for genomic sequence classification. *BMC Genom Data*, *24*(1), 25. https://doi.org/10.1186/s12863-023-01123-8

Grešová, K., Vaculík, O., & Alexiou, P. (2023). Using attribution sequence alignment to interpret deep learning models for miRNA binding site prediction. *Biology*, *12*(3). https://doi.org/10.3390/biology12030369

Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-engele, P. W., Lim, L. P., & Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. *Molecular cell*, *27*(1), 91–105. https://doi.org/10.1016/j.molcel.2007.06.017

Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Gottwein, E., & Rajewsky, N. (2014). Unambiguous identification of miRNA:Target site interactions by different types of ligation reactions. *Molecular Cell*, *54*(6), 1042–1054. https://doi.org/10.1016/j.molcel.2014.03.049

Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., & Zavolan, M. (2021). CLIP and complementary methods. *Nature Reviews Methods Primers*, *1*. https://doi.org/10.1038/s43586-021-00018-1

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr, Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., & Tuschl, T. (2010). Transcriptome-wide identification of RNA-Binding protein and MicroRNA target sites by PAR-CLIP. *Cell*, *141*(1), 129–141. https://doi.org/10.1016/j.cell.2010.03.009

Hanna, J., Hossain, G. S., & Kocerha, J. (2019). The potential for microRNA therapeutics and clinical research. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00478

Hébert, S. S., & Strooper, B. D. (2009). Alterations of the microRNA network cause neurodegenerative disease. *Trends in Neurosciences*, *32*(4), 199–206. https://doi.org/10.1016/j.tins.2008.12.003

Hejret, V., Varadarajan, N. M., Klimentova, E., Gresova, K., Giassa, I.-C., Vanacova, S., & Alexiou, P. (2023). Analysis of chimeric reads characterises the diverse targetome of AGO2-mediated regulation. *Scientific Reports*, *13*. https://doi.org/10.1038/s41598-023-49757-z

Helwak, A., Kudla, G., Dudnakova, T., & Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, *153*(3), 654–665. https://doi.org/10.1016/j.cell.2013.03.043

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res*, *31*(13), 3429–3431. https://doi.org/10.1093/nar/gkg599

Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., Tsai, W.-T., Chen, G.-Z., Lee, C.-J., Chiu, C.-M., Chien, C.-H., Wu, M.-C., Huang, C.-Y., Tsou, A.-P., & Huang, H.-D. (2010). miRTarBase: A database curates experimentally validated microRNA–target interactions. *Nucleic Acids Res*, *39*, D163–D169. https://doi.org/10.1093/nar/gkq1107

Hwang, H., Chang, H. R., & Baek, D. (2023). Determinants of functional MicroRNA targeting. *Molecules and Cells*, *46*(1), 21–32. https://doi.org/10.14348/molcells.2023.2157

Ivey, K. N., & Srivastava, D. (2010). MicroRNAs as regulators of differentiation and cell fate decisions. *Cell stem cell*, *7*(1), 36–41. https://doi.org/10.1016/j.stem.2010.06.012

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., & Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nature Genetics*, *39*, 1278–1284. https://doi.org/10.1038/ng2135

Klimentová, E., Hejret, V., Krčmář, J., Grešová, K., Giassa, I.-C., & Alexiou, P. (2022). miRBind: A deep learning method for miRNA binding classification. *Genes*, *13*(12). https://doi.org/10.3390/genes13122323

Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). miRBase: From microRNA sequences to function. *Nucleic Acids Research*, *47*(D1), D155–D162. https://doi.org/10.1093/nar/gky1141

Kozomara, A., & Griffiths-Jones, S. (2014). miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, *42*(D1), D68–D73. https://doi.org/10.1093/nar/gkt1181

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Krüger, J., & Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Research*, *34*, W451–W454. https://doi.org/10.1093/nar/gkl243

kumar Karnati, H., Panigrahi, M. K., Gutti, R. K., Greig, N. H., & Tamargo, I. A. (2015). miRNAs: Key players in neurodegenerative disorders and epilepsy. *Journal of Alzheimer's disease : JAD*, *48*(3), 563–580. https://doi.org/10.3233/jad-150395

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., & Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science*, *294*(5543), 853–858. https://doi.org/10.1126/science.1064921

Lal, A., Navarro, F., Maher, C. A., Maliszewski, L. E., Yan, N., O'Day, E. M., Chowdhury, D., Dykxhoorn, D. M., Tsai, P., Hofmann, O. M., Becker, K. G., Gorospe, M., Hide, W. A., & Lieberman, J. (2009). 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*, *35*(5), 610–625. https://doi.org/10.1016/j.molcel.2009.08.020

Lau, N. C., Lim, L. P., Weinstein, E. G., & Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in caenorhabditis elegans. *Science*, *294*(5543), 858–862. https://doi.org/10.1126/science.1065062

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, R. C., & Ambros, V. (2001). An extensive class of small RNAs in caenorhabditis elegans. *Science*, *294*(5543), 862–864. https://doi.org/10.1126/science.1065329

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The c. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, *75*(5), 843–854. https://doi.org/10.1016/0092-8674(93)90529-y

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are MicroRNA targets. *Cell*, *120*, 15–20. https://doi.org/10.1016/j.cell.2004.12.035

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural*

Networks and Learning Systems, *33*(12), 6999–7019. https://doi.org/10.1109/tnnls.2021.3084827

Liu, J., Carmell, M. A., Rivas, F. V., Marsden, C. G., Thomson, J. M., Song, J.-J., Hammond, S. M., Joshua-Tor, L., & Hannon, G. J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. *Science*, *305*(5689), 1437–1441. https://doi.org/10.1126/science.1102513

Liu, Z., Li, J., Li, S., Zang, Z., Tan, C., Huang, Y., Bai, Y., & Li, S. Z. (2024). GenBench: A benchmarking suite for systematic evaluation of genomic foundation models. *ArXiv*, *abs/2406.01627*. https://doi.org/10.48550/arXiv.2406.01627

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26. https://doi.org/10.1186/1748-7188-6-26

Lu, Y., & Leslie, C. S. (2016). Learning to predict miRNA-mRNA interactions from AGO CLIP sequencing and CLASH data. *PLOS Computational Biology*, *12*(7), 1–18. https://doi.org/10.1371/journal.pcbi.1005026

Lytle, J. R., Yario, T. A., & Steitz, J. A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. *Proceedings of the National Academy of Sciences*, *104*, 9667–9672. https://doi.org/10.1073/pnas.0703820104

Maltby, S., Plank, M. W., Tay, H. L., Collison, A. M., & Foster, P. S. (2016). Targeting MicroRNA function in respiratory diseases: Mini-review. *Frontiers in Physiology*, *7*. https://doi.org/10.3389/fphys.2016.00021

Manakov, S. A., Shishkin, A. A., Yee, B. A., Shen, K. A., Cox, D. C., Park, S. S., Foster, H. M., Chapman, K. B., Yeo, G. W., & Van Nostrand, E. L. (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, F. I., Teufel, F., Horlacher, M., Madsen, D., Pultz, D., Winther, O., & Boomsma, W. (2023). BEND: Benchmarking DNA language models on biologically meaningful tasks. *ArXiv*, *abs/2311.12570*. https://doi.org/10.48550/arXiv.2311.12570

McGeary, S. E., Lin, K. S., Shi, C. Y., Pham, T. M., Bisaria, N., Kelley, G. M., & Bartel, D. P. (2019). The biochemical basis of microRNA targeting efficacy. *Science*, *366*(6472). https://doi.org/10.1126/science.aav1741

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., & Tuschl, T. (2004). Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell*, *15*(2), 185–197. https://doi.org/10.1016/j.molcel.2004.07.007

Mills, W. T., Eadara, S., Jaffe, A. E., & Meffert, M. K. (2023). SCRAP: A bioinformatic pipeline for the analysis of small chimeric RNA-seq data. *RNA*, *29*(1), 1–17. https://doi.org/10.1261/rna.079240.122

Min, S., Lee, B., & Yoon, S. (2022). TargetNet: Functional microRNA target prediction with deep neural networks. *Bioinformatics*, *38*(3), 671–677. https://doi.org/10.1093/bioinformatics/btab733

Mitchell, T. M., et al. (1997). *Machine learning*. McGraw-hill New York.

Moore, M. J., Scheel, T. K. H., Luna, J. M., Park, C. Y., Fak, J., Nishiuchi, E., Rice, C. M., & Darnell, R. B. (2015). miRNA–target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms9864

Morita, S., Horii, T., Kimura, M., Goto, Y., Ochiya, T., & Hatada, I. (2007). One argonaute family member, Eif2c2 (Ago2), is essential for development and appears not to be involved in DNA methylation. *Genomics*, *89*(6), 687–696. https://doi.org/10.1016/j.ygeno.2007.01.004

Paker, A., & Ogul, H. (2019). mirLSTM: A deep sequential approach to MicroRNA target binding site prediction. *DEXA Workshops*. https://doi.org/10.1007/978-3-030-27684-3_6

Parveen, A., Mustafa, S. H., Yadav, P., & Kumar, A. (2019). Applications of machine learning in miRNA discovery and target prediction. *Curr Genomics*, *20*(8), 537–544. https://doi.org/10.2174/1389202921666200106111813

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Müller, P., Spring, J., Srinivasan, A., Fish-

man, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., & Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, *408*(6808), 86–89. https://doi.org/10.1038/35040556

Pinzón, N. E., Li, B., Martinez, L. D., Sergeeva, A., Presumey, J., Apparailly, F., & Seitz, H. (2017). microRNA target prediction programs predict many false positives. *Genome Research*, *27*, 234–245. https://doi.org/10.1101/gr.205146.116

Pla, A., Zhong, X., & Rayner, S. (2018). miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts. *PLoS Comput Biol*, *14*(7), e1006185. https://doi.org/10.1371/journal.pcbi.1006185

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2009). Detection of non-neutral substitution rates on mammalian phylogenies. *Genome Res*, *20*(1), 110–121. https://doi.org/10.1101/gr.097857.109

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R., & Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in caenorhabditis elegans. *Nature*, *403*(6772), 901–906. https://doi.org/10.1038/35002607

Ren, Y., Chen, Z., Qiao, L., Jing, H., Cai, Y., Xu, S., Ye, P., Ma, X., Sun, S., Yan, H., Yuan, D., Ouyang, W., & Liu, X. (2024, June). *BEACON: Benchmark for comprehensive RNA tasks and language models*. https://doi.org/10.1101/2024.06.22.600190

Robson, E. S., & Ioannidis, N. M. (2024). GUANinE v1.0: Benchmark datasets for genomic AI sequence-to-function models. *bioRxiv*. https://doi.org/10.1101/2023.10.12.562113

Rupaimoole, R., & Slack, F. J. (2017). MicroRNA therapeutics: Towards a new era for the management of cancer and other diseases. *Nature Reviews Drug Discovery*, *16*, 203–222. https://doi.org/10.1038/nrd.2016.246

Sammut, S., Gresova, K., Tzimotoudis, D., Marsalkova, E., Cechak, D., & Alexiou, P. (2025a). miRBench: Novel benchmark datasets for microRNA binding site prediction that mitigate against prevalent microRNA frequency class bias. *bioRxiv*. https://doi.org/10.1101/2024.11.14.623628

Sammut, S., Gresova, K., Tzimotoudis, D., Marsalkova, E., Cechak, D., & Alexiou, P. (2025b). miRBench: Novel benchmark datasets for microRNA binding site prediction that mitigate against prevalent microRNA frequency class bias. *Bioinformatics*, *41*, i542–i551. https://doi.org/10.1093/bioinformatics/btaf233

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Sheu-Gruttadauria, J., Xiao, Y., Gebert, L. F. R., & MacRae, I. J. (2019). Beyond the seed: Structural basis for supplementary microRNA targeting by human Argonaute2. *The EMBO Journal*, *38*. https://doi.org/10.15252/embj.2018101153

Shin, C., Nam, J.-W., Farh, K. K.-H., Chiang, H. R., Shkumatava, A., & Bartel, D. P. (2010). Expanding the microRNA targeting code: Functional sites with centered pairing. *Molecular cell*, *38*(6), 789–802. https://doi.org/10.1016/j.molcel.2010.06.005

Skoufos, G., Kakoulidis, P., Tastsoglou, S., Zacharopoulou, E., Kotsira, V., Miliotis, M., Mavromati, G., Grigoriadis, D., Zioga, M., Velli, A., Koutou, I., Karagkouni, D., Stavropoulos, S., Kardaras, F. S., Lifousi, A., Vavalou, E., Ovsepian, A., Skoulakis, A., Tasoulis, S. K., … Hatzigeorgiou, A. G. (2024). TarBase-v9.0 extends experimentally supported miRNA–gene interactions to cell-types and virally encoded miRNAs. *Nucleic Acids Research*, *52*(D1), D304–D310. https://doi.org/10.1093/nar/gkad1071

Smith-Vikos, T., & Slack, F. J. (2012). MicroRNAs and their roles in aging. *Journal of Cell Science*, *125*, 7–17. https://doi.org/10.1242/jcs.099200

Sonkoly, E., & Pivarcsi, A. (2009). Advances in microRNAs: Implications for immunity and inflammatory diseases. *Journal of Cellular and Molecular Medicine*, *13*(1), 24–38. https://doi.org/10.1111/j.1582-4934.2008.00534.x

Thum, T., & Mayr, M. (2012). Review focus on the role of microRNA in cardiovascular biology and disease. *Cardiovascular research*, *93*(4), 543–544. https://doi.org/10.1093/cvr/cvs085

Travis, A. J., Moody, J., Helwak, A., Tollervey, D., & Kudla, G. (2013). Hyb: A bioinformatics pipeline for the analysis of CLASH (crosslinking, ligation and sequencing of hybrids) data. *Methods*, *65*(3), 263–273. https://doi.org/10.1016/j.ymeth.2013.10.015

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M., & Yeo, G. W. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, *13*(6), 508–514. https://doi.org/10.1038/nmeth.3810

van Rooij, E., & Kauppinen, S. (2014). Development of microRNA therapeutics is coming of age. *EMBO Molecular Medicine*, *6*, 851–864. https://doi.org/10.15252/emmm.201100899

Varani, G., & McClain, W. H. (2000). The G x U wobble base pair. a fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep*, *1*(1), 18–23. https://doi.org/10.1093/embo-reports/kvd001

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1706.03762

Videm, P., Kumar, A., Zharkov, O., Grüning, B. A., & Backofen, R. (2021). ChiRA: An integrated framework for chimeric read analysis from RNA-RNA interactome and RNA structurome data. *GigaScience*, *10*(2), giaa158. https://doi.org/10.1093/gigascience/giaa158

Vlachos, I. S., Paraskevopoulou, M. D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I.-L., Maniou, S., Karathanou, K., Kalfakakou, D., Fevgas, A., Dalamagas, T., & Hatzigeorgiou, A. G. (2014). DIANA-TarBase v7.0: Indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res*, *43*(Database issue), D153–D159. https://doi.org/10.1093/nar/gku1215

Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., Capriotti, E., Casadio, R., Capella-Gutierrez, S., Cirillo, D., Del Conte, A., Dimopoulos, A. C., Del Angel, V. D., Dopazo, J., Fariselli, P., Fernández, J. M., Huber, F., Kreshuk, A.,

Lenaerts, T., Martelli, P. L., … ELIXIR Machine Learning Focus Group. (2021). DOME: Recommendations for supervised machine learning validation in biology. *Nature Methods*, *18*(10), 1122–1127. https://doi.org/10.1038/s41592-021-01205-4

Wang, X. (2016). Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics*, *32*(9), 1316–1322. https://doi.org/10.1093/bioinformatics/btw002

Wen, M., Cong, P., Zhang, Z., Lu, H., & Li, T. (2018). DeepMirTar: A deep-learning approach for predicting human miRNA targets. *Bioinformatics*, *34*(22), 3781–3787. https://doi.org/10.1093/bioinformatics/bty424

Wightman, B., Ha, I., & Ruvkun, G. (1993). Post–transcriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, *75*(5), 855–862. https://doi.org/10.1016/0092-8674(93)90530-4

Wilczyńska, A., & Bushell, M. (2014). The complexity of miRNA-mediated repression. *Cell Death and Differentiation*, *22*, 22–33. https://doi.org/10.1038/cdd.2014.112

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wright, E. (2024). Accurately clustering biological sequences in linear time by relatedness sorting. *Nature Communications*, *15*(1), 3047. https://doi.org/10.1038/s41467-024-47371-9

Wu, X., Zhang, L., Tong, X., Wang, Y., Zhang, Z., Kong, X., Ni, S., Luo, X., Zheng, M., Tang, Y., & Li, X. (2025). miCGR: Interpretable deep neural network for predicting both site-level and gene-level functional targets of microRNA. *Briefings in Bioinformatics*, *26*(1), bbae616. https://doi.org/10.1093/bib/bbae616

Yang, T., Wang, Y., & He, Y. (2024). TEC-miTarget: Enhancing microRNA target prediction based on deep learning of ribonucleic acid sequences. *BMC Bioinformatics*, *25*(1), 159. https://doi.org/10.1186/s12859-024-05780-z

Yang, T.-H., Chen, J.-C., Lee, Y.-H., Lu, S.-Y., Wu, S.-H., Chang, F.-Y., Huang, Y.-C., Lee, M.-H., Tseng, Y.-Y., & Wu, W.-S. (2024). Identifying human miRNA target sites via learning the interaction patterns between miRNA and mRNA segments. *J. Chem. Inf. Model.*, *64*(7), 2445–2453. https://doi.org/10.1021/acs.jcim.3c01150

Zhao, Y., Samal, E., & Srivastava, D. (2005). Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature*, *436*, 214–220. https://doi.org/10.1038/nature03817

Zheng, X., Chen, L., Li, X., Zhang, Y., Xu, S., & Huang, X. (2020). Prediction of miRNA targets by learning from interaction sequences. *PLoS One*, *15*(5), e0232578. https://doi.org/10.1371/journal.pone.0232578

Zhong, C., & Zhang, S. (2017). CLAN: The CrossLinked reads ANalyais tool. *bioRxiv*. https://doi.org/10.1101/233841