ML-miRNA: Machine Learning-Based Prediction of Functional Effects of miRNA Variants in Cancer Cell Lines

Master's in Bioinformatics Thesis

Anik Saha

Supervisor:

Xavier Bofil De Ros

Co-Supervisor:

Juraj Bergman

Spring 2025

AARHUS UNIVERSITET
NATURAL SCIENCES

AARHUS UNIVERSITY

Aarhus University
http://www.au.dk

Thesis title:

ML-miRNA: Machine Learning-Based Prediction of Functional Effects of miRNA Variants in Cancer Cell Lines

Thesis Period:

Spring Semester 2025

Author:

Anik Saha

Supervisor:

Xavier Bofil De Ros

Co- Supervisor:

Juraj Bergman

Page Numbers: 54

Date of Completion:

June 15, 2025

Master’s of Science in Bioinformatics

Bioinformatics Research Center (BiRC)

Aarhus University, Denmark

Abstract

MicroRNAs (miRNAs) play crucial role as post-transcriptional regulators, however, the sequence variants of this small RNAs are capable of posing significant impact in mRNA functions by changing gene regulatory network in cancer. This study introduces a machine learning framework named ML-miRNA which helps to predict the functional effects of miRNA variants across 13 human cell lines (includes both immortalized and cancer cell lines). We analyzed a complete library consist of cell growth expression of those cell lines against 3,311 miRNA sequences (after filtration), which include 800 canonical miRNAs and five classes of positional and shift variation for each canonical sequence. A total of 123 sequence-based features, including nucleotide composition, seed/flank k-mers, and positional motifs, were extracted at first and then reduced to 31 essential features using a pipeline utilizing correlation filtering, a hierarchical feature-protection scheme and biological relevance. We evaluated the performance of Random Forest and XGBoost algorithms across three preprocessing methods (raw, z-score, $\log_2$-fold-change). Our findings suggests that a hyperparameter-tuned Random Forest using $\log_2$-fold-change data produces the highest performance, with a mean R^2 of approximately 0.46 ± 0.02, surpassing all other configurations. Analysis of feature importance indicated that seed-region G-content and overall GC percentage are the primary predictors of variant impact, which eventually supports the finding from previous studies on seed-centric toxicity. This framework is capable of identifying both universal and cell-specific sequence features of miRNA sequence, making it a useful tool for finding/selecting variants for further experimental validation. We believe that this study will provide some help to improve the understanding importance of different sequence features as it combines the domain knowledge with strong machine learning methodologies to predict the functional effects of miRNA variants in different cell line settings.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to everyone whom has supported me throughout the completion of this thesis.

I wish to express my gratitude to both of my supervisors, Professor Xavier Bofil De Ros and Professor Juraj Bergman, for their precious guidance and constant assistance and tolerance throughout my work. Their vast knowledge and expertise hence influential in keeping me motivated for this thesis as well.

Special thanks to the Bioinformatics Research Center (BiRC) at Aarhus University for providing a stimulating environment.

To my parents, thank you for your endless love and for believing in me. I am also grateful to my friends in Denmark for their support and for making this journey enjoyable.

Table of Contents

Table of Contents vi

List of Figures ix

List of Tables x

Chapter 1: Introduction 1

Chapter 2: Background 3

2.1 Introduction to MicroRNAs 3

2.2 MicroRNA Biogenesis Pathway 3

2.2.1 Canonical miRNA Biogenesis 4

2.2.2 Non-canonical Biogenesis Pathways 4

2.3 MicroRNA Target Recognition and Regulation 6

2.3.1 Seed Region and Target Specificity 6

2.3.2 Mechanisms of Gene Silencing 6

2.3.3 Seed-Based Toxicity and Off-Target Effects 7

2.4 MicroRNA Sequence Variants and Their Functional Impact 7

2.4.1 Types of miRNA Variants 7

2.4.2 Functional Consequences of Variants 8

2.4.3 Disease-Associated miRNA Variants 8

2.5 Computational Approaches for miRNA Analysis 8

2.5.1 Traditional Machine Learning Approaches for Target Prediction 8

2.5.2 Machine Learning for Variant Effect Prediction 9

2.5.3 Predicting Variant Effects 10

2.6 MicroRNAs in Cancer 10

2.6.1 Oncogenic and Tumor-Suppressive miRNAs 10

2.6.2 Clinical Applications and Challenges 10

2.7 Current Challenges and Research Gaps 11

2.7.1 Limitations in Variant Analysis 11

2.7.2 Computational Challenges 11

2.7.3 Need for Systematic Approaches 11

Chapter 3- Methodology 12

3.1 Dataset Description 12
3.2 Data Preprocessing and Normalization 13
3.3 Principal Component Analysis 14
3.4 Sequence Feature Extraction and Analysis 15
3.4.1 Feature Categories and Extraction 15
3.4.2 Positional and Terminal Features 16
3.4.3 Seed and Post-seed Region Characterization 16
3.4.4 K-mer Frequency Analysis and Filtering 16
3.4.5 Feature Integration and Dataset Construction 16
3.5 Feature Preprocessing and Sparsity Analysis 17
3.5.1 Pipeline Structure and Categorical Encoding 17
3.5.2 Sparsity-Based Filtering with Biological Protection 18
3.5.3 Variance-Based Filtering and Standardization 19
3.6 Correlation-Based Feature Selection 19
3.6.1 Correlation Analysis with Expression Targets 19
3.6.2 Multicollinearity Detection and Biologically Informed Resolution 20
3.7 Feature Selection Methods Comparison 20
3.8 Machine Learning Model Development 22
3.8.1 Random Forest Implementation Strategy 22
3.8.2 Multi-Variant Model Implementation Framework 22
3.8.3 Training and Validation Framework 23
3.9 XGBoost Implementation and Comparative Analysis 24
3.9.1 Optimization Strategy 25
3.9.2 Performance Metrics and Evaluation Framework 25
3.10 Model Performance Evaluation and Statistical Analysis 26
3.10.1 Data Aggregation and Metric Computation 26
3.10.2 Visualization Framework and Interpretability Analysis 26
3.10.3 Tools and Reproducibility 27
3.11 Advanced Model Evaluation and Statistical Comparison 27
3.11.1 Aggregation of Performance Data and Metric Consolidation 27

3.11.2 Visualization of Predictive Power with SEM 27

3.11.3 Statistical Significance Testing via Friedman-Nemenyi Analysis 28

3.11.4 Implementation and Reproducibility 29

3.12 Implementation Details and Computational Considerations 29

Chapter 4: Results 31

4.1 Data Preprocessing, Feature Engineering, and Exploratory Analysis 31

4.1.1 Expression Data Distribution Analysis 31

4.1.2 Principal Component Analysis 32

4.1.3 Feature Engineering and Selection 32

4.2 Models Selection and Comparative Performance Analysis 34

4.2.1 Overview of Model Performance Rankings 34

4.2.3 Effectiveness of Hyperparameter Tuning 37

4.2.4 Model Type Performance, Reliability, and Statistical Significance 38

4.2.5 Implications for Science and Technology 40

4.2.6 Recommendation for the Final Model 41

4.3 Analysis of Feature Importance 41

4.3.1 Main Feature Classes and Biological Significance 41

4.3.2 Cross-Model Consistency and Context Dependence 42

4.3.3 Model Specific Interpretations 43

4.3.4 Consequences for Predictive Analysis and Mechanistic Modeling 44

Chapter 5: Discussion, Limitation and Future Direction 44

5.1 Justification of Methods and Review of Key Studies 44

5.2 Biological Implications and Mechanistic Interpretation 45

5.3 Possible Integration with Modern AI/ML Framework in Future 45

5.4 Limitations 46

Chapter 6: Conclusion 47

References: 47

List of Figures

Figure 1: a, Canonical microRNA (miRNA) biogenesis. b, Non-canonical mechanisms of miRNA biogenesis, involving the generation of pre-miRNA hairpins independent of Microprocessor. c, Ago2 cleavage-dependent miRNAs. (Adapted from Shang et al., 2023) 5
Figure 2: Visual Representation of 3 Variants along with Canonical Sequence 12
Figure 3: Density Distribution of Three Datasets 31
Figure 4: PCA Component Analysis of Three Datasets 32
Figure 5: Flowchart of the Feature Engineering Framework 33
Figure 6: Model Performance Ranking Across 18 Configurations. The horizontal bar chart displays the R-scores of all Random Forest (RF) and XGBoost (XGB) models under various preprocessing methods and tuning conditions. 34
Figure 7: Raw vs. Z-score Normalization Performance Similarity. Scatter plot comparing R² scores between raw and z-score preprocessing for each model configuration, showing a mean absolute difference of only 0.50%. 35
Figure 8: Raw vs. Z-score Preprocessing Effects on Random Forest Models. Comparison of R², RMSE, and MAE across preprocessing methods, showing negligible R² differences but large RMSE and MAE scale reductions with z-score normalization. 36
Figure 9: Hyperparameter Tuning Effectiveness. Bar chart showing percentage improvements in R² scores after tuning for each model and dataset (no error bars). 37
Figure 10: Performance Distribution by Model Type. Boxplot summarizing the R² performance distribution for Basic, Tuned, and LASSO model variants. 38
Figure 11: Critical Difference Diagram (Nemenyi Test). Ranking of all model configurations by average performance, with statistically insignificant differences marked by horizontal connections. 39
Figure 12: Model Performance Summary with Standard Error. Horizontal bar chart illustrating mean R² values and standard errors for all model configurations. 40
Figure 13: Feature Importance Analysis – Log2FC Model. Top 20 features across cell lines including heatmap, frequency bar, and cumulative contribution. 42
Figure 14: Feature Importance Analysis – Z-score Model. Identical layout for comparison against Log2FC model. 43

List of Tables

Table 1: Sequence Feature Categories and Biological Significance 15
Table 2: Categorical Feature Encoding Specifications 17
Table 3: Biological Feature Protection Hierarchy 18
Table 4: Computational Performance Comparison 21
Table 5:Random Forest Advantages for IsomiR Analysis 22
Table 6: XGBoost Advantages for Genomic Applications 24
Table 7: Complete Software Environment Specification 29

Chapter 1: Introduction

MicroRNAs (miRNAs) are short RNA molecules, approximately 20–24 nucleotides in length, that regulate gene expression. The Argonaute-miRNA complex interacts with complementary sequences in messenger RNAs (mRNAs), resulting in either the inhibition of protein synthesis or the degradation of the mRNA. In humans, more than 2,600 mature miRNAs collectively regulate over 60% of protein-coding genes, therefore building dense regulatory networks essential for development, differentiation, stress responses, and disease homeostasis (Friedman et al., 2009; Treiber et al., 2019).

In oncology, miRNAs can act as either oncogenes ("oncomiRs") or tumor suppressors, depending on their expression levels and targets availability. Dysregulated miRNA profiles are associated with tumor initiation, progression, metastasis, and drug resistance (Rupaimoole & Slack, 2017; Condrat et al., 2020). To give just a few examples, the loss of the helpful miR-34a accelerates the growth of cancers that have mutated p53, and excessive levels of the bad miR-21 helps tumors to spread and become resistant to treatment in many cancers. This duality highlights the potential of miRNAs as biomarkers and therapeutic agents in oncology.

Even though a lot of work has been done to list changes in miRNA expression in cancer, it's still hard to predict how different sequence variants—like single-nucleotide polymorphisms, RNA editing, somatic mutations, and differences in isomiR processing—affect miRNA function. Variants in the important "seed" area (nucleotides 2–8) can change how miRNAs target other molecules, which can either enhance or reduce their function and lead to significant changes in traits. Early studies on siRNA showed that the frequency of matching hexamer seeds can predict unwanted effects (Anderson et al., 2008), and later tests of all 4,096 possible 6-mer seeds found that G-rich patterns can be harmful by targeting genes that help cells survive (DISE mechanism) (Gao et al., 2018). However, these studies mainly looked at how seeds affect siRNAs or single miRNAs, creating a gap in understanding how variants impact entire miRNA sequences and different cell types in a systematic and measurable way.

Current in silico tools mainly focus on predicting targets (like TargetScan, miRanda, PITA) by estimating how well they bind based on seed complementarity and site accessibility (Gebert & MacRae, 2019). While valuable, they do not model how sequence variants disturb global regulatory function, nor do they account for cell-type–specific expression landscapes. Machine-learning tools in miRNA biology have mostly focused on specific tasks, like scoring miRNA–mRNA interactions or classifying isomiRs, but they haven't created a broad system to predict how variants affect different cancer cell lines.

Objectives

This thesis presents ML-miRNA, a machine-learning framework aimed at predicting the functional implications of miRNA sequence variants in cancer cell lines. The primary contributions are as follows:

- **Systematic Variant Library:** A systematic variant library was developed, containing 5,574 unique miRNA sequences. This includes 800 canonical miRNAs and five classes of positional substitutions and shift variants, aimed at understanding the roles of seed and flanking regions.
- **Comparison of Preprocessing Strategies:** We assessed raw, z-score normalized, and $\log_2$-fold-change transformations to identify the most effective representation of expression data for modeling purposes.
- **Feature Engineering Overview**: We reduced an initial set of 123 sequence-based descriptors, which included nucleotide composition, seed/flank k-mers, and positional motifs, to 31 biol ogically relevant features. This was achieved through correlation filtering, followed by a feature-protection hierarchy to preserve essential features.
- **Algorithmic Evaluation and Interpretability:** We thoroughly compared the performance of Random Forest and XGBoost models using an algorithmic evaluation under several preprocessing settings. We also found major sequence determinants using feature-importance measures.
- **Cross-Cell-Line Validation:** Profiling variant effects across 13 different cancer cell lines helped to identify both universal and context-specific regulating patterns, hence improving the use of variant-effect prediction in cell line.

Structure of the Thesis

- Chapter 2 reviews miRNA biogenesis, variant mechanisms, and existing computational methods.
- Chapter 3 describes the engineered variant library, feature-engineering pipeline, data transformations, and modeling workflows.
- Chapter 4 presents model performance results, hyperparameter tuning outcomes, and feature-importance analyses.
- Chapter 5 discusses methodological rationale, biological implications, integration with AI/ML paradigms, limitations, and future directions.
- Chapter 6 concludes by summarizing contributions and outlining translational prospects for ML-miRNA in clinical and research applications.

Through this work, we hope to close the gap between sequence variation and functional outcome by offering a strong, interpretable method for classifying miRNA variants for experimental validation and cancerous treatment research.

Chapter 2: Background

2.1 Introduction to MicroRNAs

miRNAs (microRNAs) are short and evolutionarily conserved non-coding RNAs, which represent significant controllers of eukaryotic gene expression. This ~22 nucleotide-sized molecules usually act as vital post transcriptional regulators of gene expression in all metazoans through base-pairing with complementary sequences in target messenger RNAs (mRNAs) (Bartel, 2018; O'Brien et al., 2018). After they were first discovered in the Caenorhabditis elegans in the early 1990s (Lee et al., 1993), miRNAs have become fundamental in nearly every physiological and pathological process, including development, differentiation, immune regulation, metabolism, and disease progression (Treiber et al., 2019; Bartel, 2018).

According to miRBase v22 (Kozomara et al., 2019), the human genome encodes for more than 2,600 mature miRNAs, while, more exact estimates suggest that the actual count of real miRNAs could be less (Fromm et al., 2020). More than 60% of human protein-coding genes are expected to be influenced by these miRNAs, hence generating complex regulatory networks that maintain cellular homeostasis (Friedman et al., 2009; Hill & Tran, 2021). Whereas individual mRNAs may be controlled by various miRNAs, each miRNA can target many mRNAs, hence creating a complex layer of post-transcriptional control (Gebert & MacRae, 2019).

MicroRNAs primarily operate by partially base-pairing with target mRNAs, mainly in the 3' untranslated regions (3'UTRs), resulting in translational repression or mRNA degradation (Jonas & Izaurralde, 2015). Recent studies indicate that mRNA degradation is the primary mechanism of miRNA-mediated repression in mammalian cells, while translational repression is of secondary importance (Eichhorn et al., 2014; Gebert & MacRae, 2019). The specificity of miRNA targeting is largely determined by the "seed" region, which includes nucleotides at position 2-8 from the 5' end of the miRNA. However, recent findings indicate a more complex mechanism in target recognition (McGeary et al., 2019; Sheu-Gruttadauria et al., 2019).

In clinical contexts, the expression of miRNA profiles has been associated with a wide range of diseases, particularly in cancer, cardiovascular disease, and neurodegeneration. MicroRNAs can act as oncogenes or tumor suppressors, depending on their targets and the tissue from which they originate (Plotnikova et al., 2019). The dual role of miRNAs, along with their stability in biofluids and potential for therapeutic modulation, has generated significant interest in their application as diagnostic biomarkers and therapeutic agents (Rupaimoole & Slack, 2017; Condrat et al., 2020).

2.2 MicroRNA Biogenesis Pathway

MicroRNA (miRNA) biogenesis is a precisely controlled multistep process that controls the maturity of these about 22-nucleotide regulating molecules. This pathway can be divided into canonical and non-canonical routes, each with different purposes in miRNA diversity and activity.

2.2.1 Canonical miRNA Biogenesis

Starting with the transcription of pri-miRNAs by RNA polymerase II, the canonical miRNA pathway begins. Typically, pri-miRNAs are lengthy, capped, polyadenylated transcripts labeled by one or more hairpin structures. The Microprocessor Complex, comprising the RNase III enzyme Drosha and its cofactor DGCR8, cleaves the pri-miRNA in the nucleus, releasing a ~70-nucleotide precursor miRNA (pre-miRNA) hairpin (Kim et al., 2024; Bartel, 2018). Exportin-5 transfers the pre-miRNA to the cytoplasm in a Ran-GTP-dependent fashion. The RNase III enzyme Dicer further digests the pre-miRNA in the cytoplasm to a ~22 nucleotide miRNA duplex (Figure 2.1 a). One of the two strands of this duplex, the so-called guide strand, is specifically incorporated into an Argonaute (AGO) protein to constitute the RNA-induced silencing complex (RISC), whereas the other passenger strand is usually degraded. The guide strand subsequent leads the complex to complementary target mRNAs to repress or degrade them, mostly through seed region interactions (McGeary et al., 2019).

2.2.2 Non-canonical Biogenesis Pathways

As well as the canonical pathway, several non-canonical pathways have been identified, showing the flexibility of miRNA generation. One of these pathways is through mirtrons, which are short introns that form hairpin structures and avoid processing by Drosha. These mirtrons are then exported after splicing and debranching, and are cleaved by Dicer like canonical pre-miRNAs (Westholm & Lai, 2011) (Figure 2.1 b left).

Another notable exception is pre-mir-451, which bypasses Dicer processing owing to its unusually short stem of 17 base pairs (Cheloufi et al., 2010). Instead, pre-mir-451 is cleaved by Ago2's endonuclease activity, followed by trimming by the poly(A)-specific ribonuclease (PARN) to generate the mature miRNA (Yoda et al., 2013). This pathway (Figure 2.1 c) demonstrates remarkable flexibility in miRNA biogenesis and highlights the central role of Argonaute proteins.

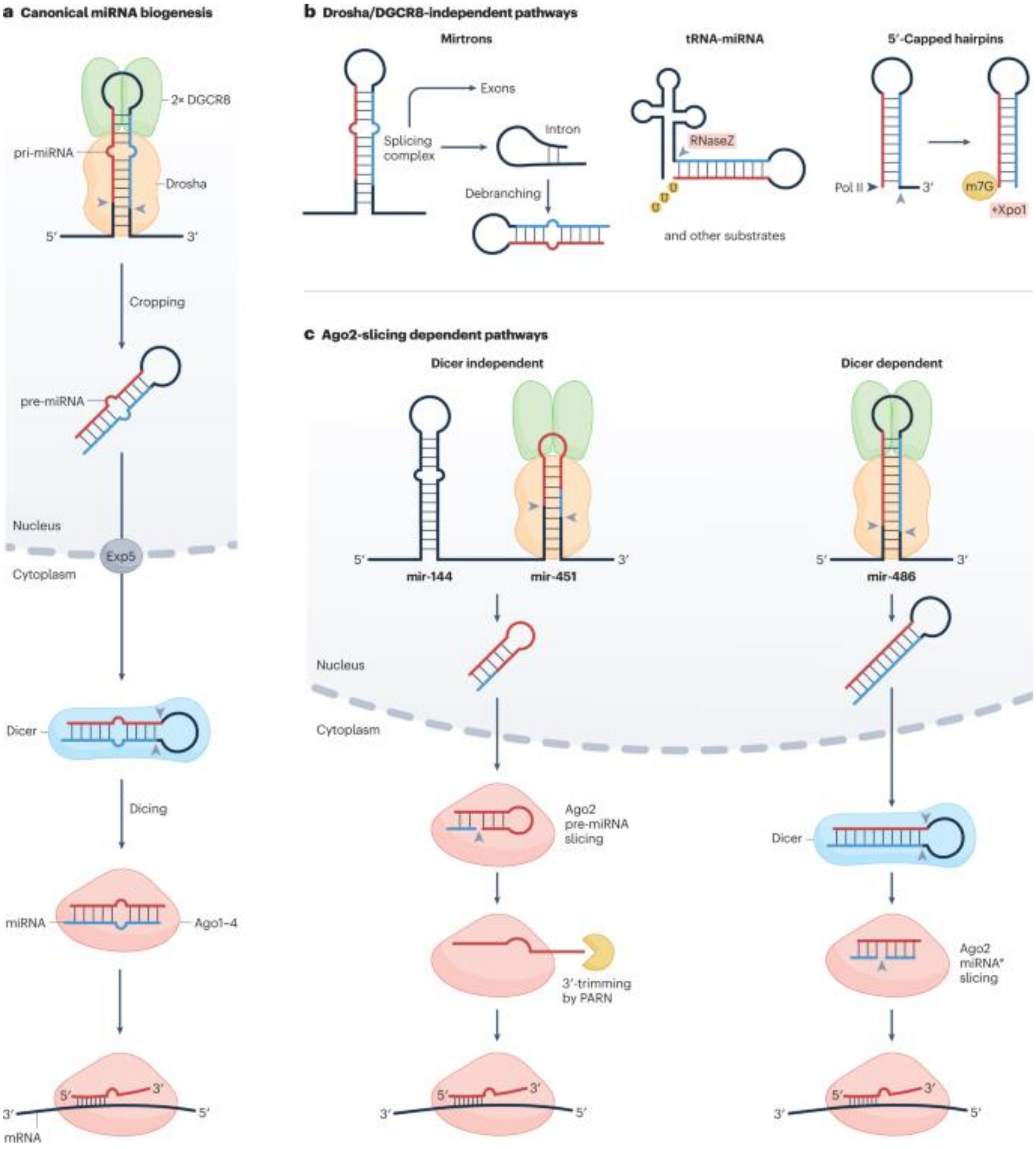


Figure 1: a, Canonical microRNA (miRNA) biogenesis. b, Non-canonical mechanisms of miRNA biogenesis, involving the generation of pre-miRNA hairpins independent of Microprocessor. c, Ago2 cleavage-dependent miRNAs. (Adapted from Shang et al., 2023)

In total, the diversity of miRNA biogenesis pathways and regulatory mechanisms highlight the cellular requirement of a versatile but strict regulation of gene expression. Both canonical and non-canonical pathways help to make the miRNAome and its dynamic responsiveness to developmental and environmental signals.

2.3 MicroRNA Target Recognition and Regulation

2.3.1 Seed Region and Target Specificity

The seed region, consisting of nucleotides 2-8 from the 5' end of the miRNA, is the main factor influencing miRNA target specificity (Bartel, 2018). Canonical seed matches are categorized into distinct types according to their complementarity patterns: 6mer (positions 2-7), 7mer-A1 (positions 2-7 with an A at position 1), 7mer-m8 (positions 2-8), and 8mer (positions 2-8 with an A at position 1) (Agarwal et al., 2015; McGeary et al., 2019). Recent high-throughput studies show that 8mer sites display the most significant repression, followed by 7mer-m8, 7mer-A1, and 6mer sites (McGeary et al., 2019). Apart from seed pairing, several auxiliary elements influence targeting efficiency including local AU content, target site accessibility, and cooperative binding of several miRNAs (Grimson et al., 2007; McGeary et al., 2019). Recent structural analyses of AGO-miRNA-target ternary complexes have explained the mechanisms by which these components enable target detection and repression (Sheu-Gruttadauria et al., 2019).

Experimental validation has shown that seed complement frequency (SCF) is crucial for miRNA specificity. Analysis of all possible hexamers revealed a nonuniform distribution across the 3' UTR transcriptome, with seed matches to highly expressed miRNAs showing evolutionary depletion, suggesting selective pressure to avoid targeting (Farh et al., 2005; Stark et al., 2005). Duplexes with low SCFs typically result in fewer off-targets effects compared to molecules with richer 3' UTR complements, which is relevant for the design of miRNA-based therapeutics (Anderson et al., 2008).

2.3.2 Mechanisms of Gene Silencing

MicroRNAs (miRNAs) promote gene silencing via two primary mechanisms: degradation of mRNA and repression of translation (Jonas & Izaurralde, 2015). The GW182 protein family functions as essential effectors, connecting AGO proteins to the cellular degradation machinery (Gebert & MacRae, 2019). GW182 proteins facilitate the recruitment of the CCR4-NOT deadenylase complex, resulting in the removal of the poly(A) tail, decapping, and subsequent mRNA degradation (Jonas & Izaurralde, 2015).

Recent studies using ribosome profiling and proteomics have found that the main mechanism of miRNA-mediated repression in mammalian cells is mRNA instability, therefore explaining 66–90% of protein downregulation (Eichhorn et al., 2014; Gebert & MacRae, 2019). For some targets and cellular settings, translational suppression is absolutely essential (Duchaine & Fabian, 2019)

2.3.3 Seed-Based Toxicity and Off-Target Effects

A significant discovery indicates that specific 6mer seed sequences possess intrinsic toxicity to cells by targeting crucial survival genes (Gao et al., 2018; Putzbach et al., 2017). The mechanism known as DISE (Death Induced by Survival gene Elimination) become obvious when siRNAs or miRNAs possessing specific G-rich seed sequences simultaneously target multiple survival genes (Murmann et al., 2018; Putzbach et al., 2018).

A systematic screening of all 4,096 possible 6mer seeds revealed that the most toxic seeds are G-rich, especially those with G nucleotides at positions 1-2, which preferentially target survival genes containing C-rich 3'UTRs (Gao et al., 2018). Many tumor-suppressive miRNAs, such as miR-34a-5p, contain toxic seed sequences, indicating a potential mechanism for inducing cancer cell death (Gao et al., 2018; Patel & Peter, 2018). In contrast, the majority of miRNAs have evolved to avoid these toxic sequences, suggesting a selective pressure reducing overall toxicity (Gao et al., 2018). This discovery has important implications for understanding miRNA evolution, function, and therapeutic applications. It also highlights the importance of considering off-target effects in RNAi-based therapeutics and the potential for designing super-toxic artificial miRNAs for cancer treatment (Murmann et al., 2018).

2.4 MicroRNA Sequence Variants and Their Functional Impact

2.4.1 Types of miRNA Variants

MicroRNA sequence variants include naturally occurring polymorphisms as well as somatic mutations, which can significantly influence miRNA function through various mechanisms (Ryan et al., 2010; Tomasello et al., 2021). Variations in single nucleotide polymorphisms (SNPs) within miRNA genes are observed at different frequencies among populations, with specific positions exhibiting greater evolutionary constraint compared to others (Zorc et al., 2012). The variants include SNPs that represent germline variations capable of affecting miRNA processing, stability, or target recognition (Moszynska et al., 2017).

IsomiRs, defined as length and sequence variants of canonical miRNAs that result from inaccurate processing or post-transcriptional modifications, represent a significant source of miRNA diversity (Tomasello et al., 2021). The functional consequences of miRNA variants are significantly influenced by their location within the mature miRNA sequence. Seed region variants (positions 2-8) generally have the most significant effects on function by modifying the total number of target genes (Hill et al., 2014). Single nucleotide variations in the seed region may divert miRNA towards different target sets, creating a novel regulatory molecule (Mencía et al., 2009). Variants located at the 3' end can influence miRNA stability and its incorporation into RISC, whereas changes in the pri-miRNA or pre-miRNA sequences may interfere with processing by Drosha or Dicer

(Fernandez et al., 2017). In addition to basic substitutions, miRNA variants include insertions and deletions that may modify the seed sequence or change the overall length of the mature miRNA (Bhattacharya et al., 2014). Recent deep sequencing studies have shown that isomiR expression is prevalent and regulated; different isomiRs could have diverse targets and functions (Tomasello et al., 2021; Guo & Chen, 2014).

2.4.2 Functional Consequences of Variants

The functional effect of the miRNA variants is significantly influenced by their location within the mature sequence (Tomasello et al., 2021). Target specificity (Bhattacharya & Cui, 2016) is substantially influenced by variants in the seed region between positions 2–8. Recent investigations have revealed that 5' isomiRs with altered seed sequences might display quite different target patterns than canonical miRNAs (Tan et al., 2014; Tomasello et al., 2021). On the other hand, variants at the 3' end influence miRNA stability and subcellular localization, whereas central variants can affect RISC loading and activity (Tomasello et al., 2021; Vickers et al., 2015). The careful analysis of miRNA variations has revealed that particular sites show more evolutionary constraint, implying their importance in function (Quang & Xie, 2016). This pattern of positional conservation provides insightful analysis for understanding miRNA evolution and future implications of variations.

2.4.3 Disease-Associated miRNA Variants

A number of studies have linked miRNA variations to human disorders, including cancer (Moszynska et al., 2017; Galka-Marciniak et al., 2019). Pathogenic variants can operate through a variety of pathways, including altered processing, whereby variants impacting pri- or pre-miRNA structure may lower mature miRNA levels (Ryan et al., 2010); seed disruption, which directly modifies targeting specificity (Hill et al., 2014); and the synthesis of new miRNAs, whereby mutations can generate de novo miRNA genes from previously non-functional sequences (Friedländer et al., 2014).

Recent cancer genomics research has found recurring miRNAs across several cancer types that point to possible driver involvement (Galka-Marciniak et al., 2019; Hrovatin et al., 2018). Without functional validation, it is difficult to separate driver from passenger mutations.

2.5 Computational Approaches for miRNA Analysis

2.5.1 Traditional Machine Learning Approaches for Target Prediction

Over the past 20 years, miRNAs' computational analysis has progressed significantly. Initial algorithms focused on target prediction using evolutionary conservation and seed complementarity (Lewis et al., 2005; Agarwal et al., 2015). These techniques did, however, show limited accuracy during experimental validation; false positive rates often exceeded 70% (Pinzón et al., 2017).

Modern target prediction systems combine several features outside seed pairing, including target site accessibility (Lorenz et al., 2011), local sequence context (Grimson et al., 2007), expression profiles (Liu & Wang, 2019), and cross-linking immunoprecipitation (CLIP) data (Karagkouni et al., 2020). They also use machine learning approaches. Recent benchmarking analyses show that ensemble approaches—which include several algorithms—outperform individual tools (Quillet et al., 2020; Riffo-Campos et al., 2016).

The evolution toward machine learning approaches began with the recognition that multiple features beyond seed pairing influence targeting efficacy (Grimson et al., 2007). miRanda incorporated thermodynamic stability calculations using Vienna RNA package algorithms, while PicTar introduced the concept of combinatorial targeting by co-expressed miRNAs (John et al., 2004; Krek et al., 2005). These developments laid the groundwork for more sophisticated machine learning implementations.

Support vector machines (SVMs) emerged as a popular choice for miRNA target prediction, as demonstrated by tools like MiRTarget2 and TargetMiner (Wang & El Naqa, 2008; Bandyopadhyay & Mitra, 2009). These methods could integrate diverse features including sequence composition, structural accessibility, and conservation patterns into unified prediction models (Betel et al., 2010).

2.5.2 Machine Learning for Variant Effect Prediction

Analysis of miRNA with the help of machine learning has benefited significantly due to the increased accessibility of high-throughput experimental data (Quang & Xie, 2016; Wen et al., 2019). Deep learning has been very promising, reflected in tools such as DeepMirTar that employs autoencoders to identify target site features, miRAW that employs convolutional neural networks to predict targets and DeepTarget that incorporates the various types of data to boost accuracy.

Ensemble methods have been shown to be quite powerful, Random Forest models appearing to give good results across the board in terms of predicting the effect sizes/variants on the outcome, XGBoost based methods being more robust to unbalanced dataset, and hybrid approaches using multiple algorithms together performing the best overall. Recent studies have identified novel predictive features such as RNA secondary structure dynamics (Lorenz et al., 2011), sequence motifs beyond seed regions (Briskin et al., 2020), tissue-specific expression patterns (Ludwig et al., 2016), and evolutionary signatures (Friedman et al., 2009). The innovations in feature engineering have markedly enhanced prediction accuracy.

2.5.3 Predicting Variant Effects

The prediction of miRNA variant effects presents unique difficulties in comparison to target prediction (Bhattacharya & Cui, 2016). Recent methodologies have focused on the creation of position-specific scoring matrices to reflect the varied significance of nucleotide positions (Quang & Xie, 2016), structural predictions that evaluate the impact of variants on RNA folding and processing (Lorenz et al., 2011), the integration of machine learning to amalgamate diverse features for holistic predictions (Salari et al., 2013), and cell-type specific models that consider context-dependent influences (Ludwig et al., 2016).

Regardless of developments, predicting the effects of variants continues to be difficult due to the limitations of training data and the context-dependent characteristics of miRNA function (Quang & Xie, 2016).

2.6 MicroRNAs in Cancer

2.6.1 Oncogenic and Tumor-Suppressive miRNAs

Depending on their targets and the cellular setting, microRNAs show two different roles in cancer: either tumor suppressors or oncogenes (oncomiRs). MicroRNAs exhibit dual roles in cancer, acting as either oncogenes (oncomiRs) or tumor suppressors, contingent upon their targets and the cellular context (Smolarz et al., 2022; Plotnikova et al., 2019). OncomiRs that are well-characterized include miR-21, which targets several tumor suppressors such as PTEN and PDCD4 (Bautista-Sánchez et al., 2020), miR-155, which facilitates proliferation and immune evasion (Bayraktar & Van Roosbroeck, 2018), and the miR-17~92 cluster, which promotes MYC-driven tumorigenesis (Fuziwara & Kimura, 2015).

However, tumor-suppressive miRNAs include the let-7 family, which inhibits several oncogenes including RAS and MYC (Chirshev et al., 2019), the miR-34 family, direct targets of p53 that promote cell cycle arrest and death (Hermeking, 2010; Rokavec et al., 2014), and the miR-200 family, known for its function in suppressing epithelial-mesenchymal transition (Title et al., 2018).

2.6.2 Clinical Applications and Challenges

As cancer cells can adapt through several pathways, the therapeutic targeting of miRNAs in cancer presents several challenges including the need of tissue-specific delivery and cellular adsorption, potential off-target effects associated with unintended targeting of toxic agents, stability concerns regarding the protection of RNA molecules from degradation, and the presence of resistance mechanisms (Rupaimoole & Slack, 2017; Gambari et al., 2019).

Recent studies show promise as miR-34a mimics (MRX34) were among the first candidates for clinical testing; but, the trial was stopped due to immune-related side events (Beg et al., 2017; Hong et al., 2020). Enhanced delivery systems and more focused approaches are underlined in current initiatives (Rupaimoole & Slack, 2017).

2.7 Current Challenges and Research Gaps

2.7.1 Limitations in Variant Analysis

Despite progress in miRNA research, significant gaps remain in our understanding of sequence variants (Tomasello et al., 2021; Bhattacharya & Cui, 2016). Most variant studies concentrate on individual miRNAs, which restricts our knowledge of fundamental concepts. Variant effects showcase significant variability across different cell types and conditions, complicating the prediction of outcomes in novel contexts. The technical challenges of high-throughput screening for variant effects persist, resulting in numerous isomiRs and variants remaining uncharacterized in existing databases.

2.7.2 Computational Challenges

Current computational approaches encounter several limitations, including a shortage of training data with experimentally validated variant effects, challenges in identifying the most informative features for prediction, a lack of biological interpretability in deep learning models, and the issue of models trained on specific cell types not transferring effectively to other contexts (Quang & Xie, 2016; Wen et al., 2019).

2.7.3 Need for Systematic Approaches

Systematic approaches are necessary to fully understand the effects of miRNA variants (Tomasello et al., 2021). This involves the development of standardized experimental protocols for variant characterization, the integration of seed toxicity concepts with traditional target prediction, the creation of cell line-specific models that account for contextual effects, and the construction of machine learning frameworks capable of managing sparse, high-dimensional data.

The challenges stated requirements for the advancement of novel computational and experimental methodologies to systematically characterize miRNA variants and predict their functional implications across various cellular contexts, which underpins the present study.

Chapter 3- Methodology

3.1 Dataset Description

The dataset used for the present study consists of a total of 5,574 unique miRNA sequences that were experimentally profiled to determine their effect on 13 different human cell lines. These are the lines of genetically modified and cancer-derived cells, which cover a wide variety of cellular contexts. In particular, the cell lines of the study are: Wild Type (WT), Triple Knockout (TKO), TUT2 Knockout (TUT2KO), TUT4 Knockout (TUT4 KO), HME1, G401, A549, HCT116, MCF7, SF268, H522, HOP92, and SKOV3. Each miRNA sequence was 22 nucleotides in length, representing the antisense strand designed to target specific mRNAs.

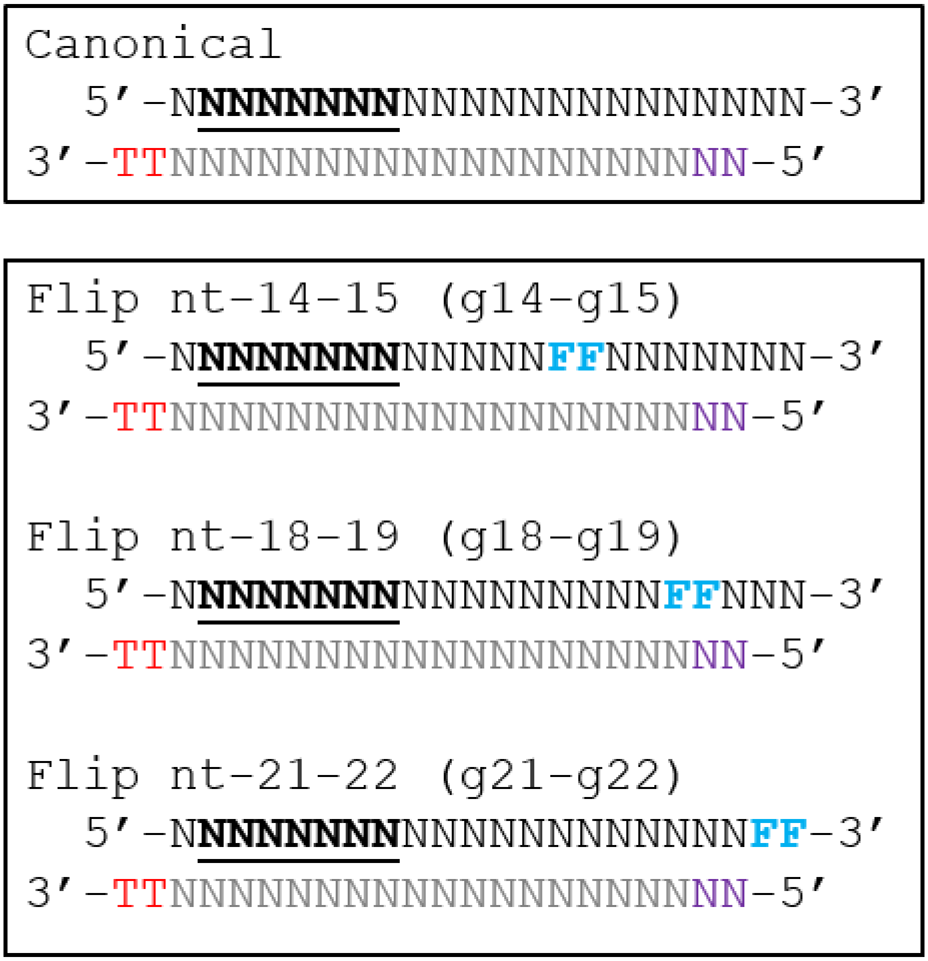

Canonical
5′-NNNNNNNNNNNNNNNNNNNNNN-3′
3′-TTNNNNNNNNNNNNNNNNNNNNNN-5′
Flip nt-14-15 (g14-g15)
5′-NNNNNNNNNNNNNFFNNNNNNN-3′
3′-TTNNNNNNNNNNNNNNNNNNNNNN-5′
Flip nt-18-19 (g18-g19)
5′-NNNNNNNNNNNNNNNNNFFNNN-3′
3′-TTNNNNNNNNNNNNNNNNNNNNNN-5′
Flip nt-21-22 (g21-g22)
5′-NNNNNNNNNNNNNNNNNNNNFF-3′
3′-TTNNNNNNNNNNNNNNNNNNNNNN-5′

Figure 2: Visual Representation of 3 Variants along with Canonical Sequence

Each data point in the dataset necessarily describes either a canonical miRNA or a designed variant of a canonical miRNA. Summarily, the dataset was obtained from 800 unique canonical miRNAs, and for each canonical miRNA, five forms of sequence variants were engineered to analyze the specific positional alteration effect. These were a result of the following systematic modification:

- **Point Substitution at Positions 14 and 15 (N1415)**: A pair of nucleotides at the 3′ end of the miRNA sequence was changed to examine the effect of such a change on gene regulatory potential.
- **Point Substitution at Positions 17 and 18 (N1819)**: Likewise, additional substitutions were brought downstream to look into positional effects towards the tail end.
- **Point Substitution at Positions 21 and 22 (N2122):** This improvement focuses on the most severe part of the miRNA sequence, i.e., the extreme 3′ end.
- **5′ End Shift – Deletion Variant (min1):** This variant was synthesized by deleting one nucleotide from the 5′-end of the miRNA and adding an extra nucleotide at the 3′-end of the miRNA, to shift the sequence downstream by one base pair.
- **5′ End Shift – Addition Variant (plus1):** Conversely, this variant involves the insertion of a second nucleotide at the 5′ end, and deletion of one nucleotide from the 3′ end, hence moving the sequence upstream.

3.2 Data Preprocessing and Normalization

Python 3.8+ was the main language used in the data processing pipeline, which relied on the scientific computing ecosystem for solid and repeatable results. The main libraries applied were Pandas (version 1.3 or above) for dealing with data (McKinney, 2010); NumPy (version 1.21 or higher) for performing efficient calculations (Harris et al., 2020); SciPy (version 1.7 or more) for using statistical methods and hypotheses (Virtanen et al., 2020); Matplotlib (version 3.5 or above) for developing good-looking visualizations (Hunter, 2007); and Seaborn (version 0.11+) for making improved statistical graphs and analyzing distributions (Waskom, 2021).

To handle the different analysis challenges for isomiR data and compare cell lines with variance in their expression levels and distributions, data were processed using two different methods. With these methods, the adaptability to various research topics and assumptions is still possible, as it can reveal many kinds of expression patterns. At the start of this analysis, we filtered out sequences denoted with symbol min1 and plus1 since they had only one changes in the whole sequence which eventually would not influence most of the sequence feature, such as features related to seed and post seed, left us with 3311 sequences in total.

For every cell line, Z-score normalizing was performed independently to standardize expression value distribution. This transformation standardized every cell line dataset to a mean of zero and a standard deviation of one, therefore enabling fair comparison between many biological systems. The equation describes the change:

$$\mathbf{z = (x - \mu) / \sigma}$$

where:

x = individual expression value

μ = cell line-specific mean expression

σ = cell line-specific standard deviation of expression

This normalizing technique lowers variability caused by variations in experimental circumstances or natural expression scaling among cell types.

Relative expression changes were assessed by converting data to the log2 fold change, which lets us observe changes with reference to a standard background. Since we set baseline_reference = 100, the transformation was calculated with the **log2FC = log2(expression_value / baseline_reference)** to compute log2FC. Experiments showed that the baseline of 100 allowed significant fold change interpretations, so that a change of +1 meant a two-fold increase and a change of -1 meant a two-fold decrease. If analysis of gene expression reveals values that are not what was anticipated for a particular biological condition, this method can be very valuable.

3.3 Principal Component Analysis

Principal Component Analysis was performed in all three transformed datasets (raw, z-score normalized, and log2 fold change) to understand how the different cell lines relate and find the hidden factors affecting miRNA sequence variations. The goal of this analysis was to reduce dimensionality while preserving as much variance as possible.

Considering the structure of the data, we transposed the matrix so that cell lines (n = 13) could be treated as an observation and miRNA sequences as a variable. This perspective allowed exploring the way cell lines group based on their overall expression signature and the specific miRNA sequence that generate the most variance between cell lines.

We performed PCA using the PCA class from the **sklearn.decomposition** package in scikit-learn. **StandardScaler** from sklearn. We used preprocessing to adjust the data before reducing its dimensions for datasets, such as raw and log2FC data, that had not yet undergone normalization. The **fit_transform()** technique was used for the transformation; the **components_ property** extracted the loadings—that is, individual miRNA sequence to every primary component. Using Matplotlib and Seaborn, we generated scatter plots of the top two main components that allow visually assessing the clustering of cell lines across various data manipulations. For interpretability, the top 20 miRNA sequences contributing to each of the first two principal components were identified based on absolute loading values. We then investigated these prominent characteristics to understand their likely biological relevance in elucidating the variations in expression among cell lines.

3.4 Sequence Feature Extraction and Analysis

3.4.1 Feature Categories and Extraction

We created a thorough feature engineering workflow to examine the molecular factors affecting cell line function and expression patterns. This pipeline classified sequence-derived traits based on recognized biological processes including stability, processing, targeted specificity, and motif enrichment.

Table1 summarizes the feature categories, including the number of features extracted, their biological relevance, and representative examples.

Table 1: Sequence Feature Categories and Biological Significance

Category	**Feature Count**	**Biological Rationale**	**Key Examples**
Basic Properties	4	Fundamental sequence characteristics affecting stability and processing	overall GC content, median folding energy
Nucleotide Composition	8	Base composition effects on structure and function	A/U/G/C counts and percentages
Positional Features	8	Terminal nucleotide effects on processing and stability	First and last nucleotide identity (one-hot encoded)
Seed Region Analysis	16	Target recognition and binding specificity determinants	Positions 2-8 composition, seed GC content
K-mer Patterns	64	Higher-order sequence motifs and structural preferences	Trinucleotide frequencies and patterns

Basic features like GC content give a thermodynamic stability whereas nucleotide composition (numbers and percentages) give additional information about the diversity of sequences. The percentage-based feature enables length-independent comparisons whereas the count-based features preserve absolute molecular composition which is biologically relevant in modeling.

3.4.2 Positional and Terminal Features

Terminal nucleotide similarity is essential for miRNA processing and silencing action especially at the first and last locations. The loading specificity of Argonaute proteins is influenced by the 5′ nucleotide, for example, uracil at the 5′ end is more frequently recognized by certain AGO family members. The 3′ nucleotide can influence subcellular localization and transcript stability (Kim et al., 2025). One-hot encoding helped to encode terminal positions in a machine-readable way. By converting every base identity into binary variables, this encoding lets machine learning models assess spatial impacts free from ordinal structure.

3.4.3 Seed and Post-seed Region Characterization

Since the bases at positions 2–8 pair with target mRNAs, the seed region is the main region in miRNAs that determines which mRNAs are targeted. The features were extracted to reflect nucleotide composition, the frequency of a base, and GC content in this particular region. The post-seed region (positions 9 to 22) was computed with a corresponding set of features to allow comparative analysis. The GC content in the seed area acts as an intermediary for hybridization strength, therefore affecting binding affinity and specificity. Comparative statistics between the seed and non-seed regions provide insight on functional difference and evolutionary boundaries between targeting domains.

3.4.4 K-mer Frequency Analysis and Filtering

The analysis of trinucleotide (3-mer) frequencies identified higher-order sequence patterns and motifs undetectable based on simple nucleotide composition analysis (Ghandi et al., 2014; Lee et al., 2015). A systematic examination of all possible trinucleotide combinations ($4^3 = 64$ total) makes it possible to identify a preference of sequences and themes that can potentially influence miRNA processing, stability or activity.

In order to identify more subtle sequence motifs beyond simple base composition, all 3-mers (trinucleotide compositions) were calculated with a sliding window along the length of the miRNA (window size = 3, stride = 1). These characteristics are of local characteristics which can affect the structure of miRNA or Dicer processing or interaction with the target. After generating all 3-mer features, we performed variance filtering to clean the data. This step helped eliminate features that were constant or nearly constant across all miRNAs (variance < 0.001) — since such features do not help the model learn anything meaningful. Removing them makes the model more efficient and prevents it from wasting effort on patterns that aren't biologically relevant.

3.4.5 Feature Integration and Dataset Construction

The completed sequence-derived features were integrated with the corresponding expression values (raw, Z-score normalized, and log2 fold change) to create the comprehensive feature matrices to be used in downstream machine learning. This design allowed the relative modeling of expression behavior among cell lines based on inherent sequence properties, appropriate to both predictive modeling and to the analysis of feature importance.

3.5 Feature Preprocessing and Sparsity Analysis

3.5.1 Pipeline Structure and Categorical Encoding

A modular preprocessing pipeline was used to prepare the dataset to be used in machine learning modeling but still be interpretable in a biological context. The design provides miRNA regulation biological knowledge to guide data cleaning and transformation. The preprocessing logic was implemented in a custom class, **miRNAPreprocessor75,** which made it reproducible and allowed its configurable use on different feature sets.

Categorical variables (indicating nucleotide identity at important positions: e.g. pos1_nt and last_pos_nt) were one-hot encoded to make them compatible with machine learning algorithms. This transformation retained biological meaning and transformed categorical input to binary features.

Algorithm 3.2: One-Hot Encoding Implementation

```
For each categorical feature f in {pos1_nt, last_pos_nt}:
  For each unique value v in f:
    Create binary feature f_v where:
      f_v = 1 if f == v, else 0
  Remove original categorical feature f
```

Table 2: Categorical Feature Encoding Specifications

Original Feature	**Encoded Features**	**Biological Rationale**	**Impact on Dataset**
pos1_nt	pos1_nt_A, pos1_nt_U, pos1_nt_G, pos1_nt_C	5' nucleotide effects on RISC loading and stability	+4 features, -1 categorical
last_pos_nt	last_pos_nt_A, last_pos_nt_U, last_pos_nt_G, last_pos_nt_C	3' nucleotide effects on processing and degradation	+4 features, -1 categorical

3.5.2 Sparsity-Based Filtering with Biological Protection

Sparsity Assessment

We evaluated each numerical feature for sparsity, which was considered as the ratio of zero values to the total sample number. The threshold of 75% sparsity is due to conventional bioinformatics usage (Saeys et al., 2007; Hira & Gillies, 2015), and it can be considered as a balance between keeping features and removing noise. Features with a sparsity level above 75% were identified as features to be eliminated:

sparsity_rate(feature) = (count_zeros(feature) / total_samples) * 100.

Features above the threshold might be removed, but this was done automatically by their biological relevance (sub-section 3.5.2.2 and Table 3.3) and thus ensured that none of the required functions were removed.

Biological Feature Protection Mechanism

After realizing that some features biologically necessary might be naturally sparse, a feature protection mechanism was introduced. In this system, features are given priority scores depending on their biological significance making sure that important features that are few in number are not lost after filtering.

Table 3: Biological Feature Protection Hierarchy

Priority Level	**Feature Category**	**Protection Rationale**	**Score Range**	**Examples**
1 (Highest)	Core properties	Fundamental miRNA characteristics essential for all analyses	85-95	GC_content, seq_length, median_energy
2	Position features	Terminal nucleotide effects critical for processing and function	76-82	pos1_nt_*, last_pos_nt_*
3	Seed region	Target binding specificity determinants with established functional importance	55-75	seed_GC_content, seed_*_percent

4	Composition	Overall sequence properties providing baseline molecular characteristics	45-50	A_percent, U_percent, G_percent, C_percent
5 (Lowest)	K-mer frequencies	Higher-order sequence patterns potentially relevant for specific contexts	20	AAA_freq, UUG_freq, etc.

This hierarchy was made by reviewing many literatures on miRNA biogenesis and regulation such as the recent review paper of Kim et al., 2025, which provide a complete overview on miRNA biogenesis, and Neph et al. highlighted the functional significance of certain motifs in their paper in 2012. This scoring system served to retain features of known biological importance despite statistical sparsity.

3.5.3 Variance-Based Filtering and Standardization

After sparsity filtering, variance cutoff of 0.001 was applied to remove ultra-low variance features which offer little discriminatory power. This step targeted the features with non-sparse distributions but limited variability across the samples, therefore, increasing model efficiency and interpretability. All the retained features were then standardized using z-score normalization. **scikit-learn StandardScaler** was used to compute feature-wise means and standard deviations using only training data. The training, validation, and test sets were subjected to these settings equally to eliminate data leakage and guarantee the model generalizability.

3.6 Correlation-Based Feature Selection

3.6.1 Correlation Analysis with Expression Targets

Pearson correlation was performed to determine the relationship between sequence features and the expression of multiple cell lines. It was decided to use this method because it is very clear, statistically rigorous, and widely acceptable in the processing of biological data although it is limited to linear connections (Hall, 1999; Guyon & Elisseeff, 2003).

The sequences were assessed regarding each characteristic correlating with the expression values across the 13 cell lines using Pearson product-moment correlation coefficient. Two-sided t-tests were used to test singularity at the alpha level of 0.01 with null hypothesis $H_{0:}$ ρ=0 and alternative hypothesis H_1: ρ≠0. With 52 traits and 13 targets (676 tests). We used uncorrected α = 0.01 to remain sensitive to physiologically meaningful, but moderate associations, but interpreted results cautiously.

3.6.2 Multicollinearity Detection and Biologically Informed Resolution

Multicollinearity was evaluated to ensure model stability and interpretability by identifying feature pairs with a correlation coefficient of $|r| \geq 0.8$ (Dormann et al., 2013). Features showing notable mutual correlation might cause redundancy, reduce generalizability, and maybe alter estimates of feature importance in predictive models.

Multicollinearity was solved through a hierarchy-based resolution system, which guaranteed a balance between statistical relevance and biological relevance.

Biological Priority: Features were ranked according to the biological importance scoring system outlined in Section 3.5, Table 3.3. This ensured the retention of essential features related to miRNA stability, targeting, and processing, despite potential statistical redundancy.

Predictive Breadth: Features showing stronger connections to expression targets were given higher priority within linked pairs.

Interpretability Preference: Features with greater interpretability from biological or computational angles were given higher priority in cases of tie breaking.

This biologically driven multicollinearity resolution reduced feature redundancy and kept domain-specific information, which allowed the development of efficient and interpretable machine learning models.

3.7 Feature Selection Methods Comparison

We performed comparison analysis with four popular techniques to investigate efficiency and practical significance of correlation-based feature selection method. Their respective performances needed to be benchmarked in terms of the number of characteristics selected, run time, and ease of biological interpretation. All the techniques were applied to the identical pre-processed dataset in controlled experimental conditions to ensure their fair comparison. The variability was minimized through this strategy via data management thus providing focused research on the choice behaviour and efficiency of each technique.

Feature Selection Methods Evaluated:

- **Mutual Information Regression:** Captures non-linear dependencies between features and targets based on information-theoretic measures (Kraskov et al., 2004). It offers a complementary perspective to linear correlation by identifying complex, non-linear relationships.

- **LASSO Regularization:** It selects a sparse set of predictive features through L1-penalised linear regression cross-validated. Feature selection and integrated model fitting is attained by retaining features whose coefficient is non-zero (Tibshirani, 1996).

- Principal Component Analysis: Though not a strict feature selector, PCA lowers dimensionality by converting features into uncorrelated components. We kept the minimum essential components to account for 95% of the dataset variance.

- **Univariate F-test Selection:** Ranks features based on their statistic correlation with target variables through individual F-tests. It is a fast baseline comparison standard filter method (Guyon & Elisseeff, 2003).

To assess practical performance, we recorded the number of selected features and total runtime for each method. These results are presented in Table 3.4.

Table 4: Computational Performance Comparison

Method	Features Selected	Runtime (s)
Correlation-Based	21	0.040
Mutual Information	30	1.257
LASSO Regularization	22	0.363
Univariate F-test	30	0.032
PCA (95% variance)	8 components	0.241

The univariate F -test was the fastest of all but does not control multicollinearity or provide interpretability beyond statistical association. Mutual Information and LASSO induced a moderate computational burden, but provided other modeling perspective, non-linear and embedded, respectively. PCA produced the smallest representation but rotated original features to components, which are not easy to interpret biologically.

The correlation-based approach offered a good compromise between computing performance and biology. It picked a smaller but understandable feature set (21 features) and accidentally retained competitive runtime performance, trailing only the F-test method. It also featured a domain-aware multicollinearity resolution mechanism unlike other approaches which makes it an attractive feature selection method on biologically complex systems like the modeling of miRNA expression.

3.8 Machine Learning Model Development

3.8.1 Random Forest Implementation Strategy

Random Forest was selected as the primary machine learning algorithm based on its demonstrated advantages for biological data analysis and specific suitability for cell line expression prediction tasks (Qi, 2012). The algorithm's ensemble nature provides robustness against overfitting, which is particularly important when working with high-dimensional biological data where the number of features may approach or exceed the number of samples.

Table 5:Random Forest Advantages for IsomiR Analysis

Advantage	**Biological Application**	**Technical Benefit**	**Implementation Impact**
Overfitting Resistance	Robust performance with noisy expression data	Stable predictions with limited sample sizes	Reliable model performance across cell lines
Non-linear Relationships	Captures complex sequence-expression interactions	No assumptions about linear relationships	Enhanced predictive accuracy for biological patterns
Feature Importance	Identifies key biological drivers	Interpretable results for biological insight	Direct biological interpretation of sequence determinants
Missing Value Handling	Robust to incomplete biological measurements	Minimal preprocessing requirements	Simplified data preparation pipeline
Computational Efficiency	Parallel processing capabilities	Scalable to larger datasets	Feasible training across multiple cell lines

The model was configured with the following hyperparameters: n_estimators=200, max_depth=15, min_samples_split=5, min_samples_leaf=2, max_features='sqrt', bootstrap=True, oob_score=True, random_state=42, and n_jobs=-1.

3.8.2 Multi-Variant Model Implementation Framework

Multiple model versions were methodically used and compared in order to fully assess Random Forest performance and find ideal settings for cell line expression prediction. While keeping constant evaluation procedures, this multi-variant technique allows examination of many optimization strategies and feature selection combinations.

Standard Random Forest: Baseline implementation using the correlation-selected feature set with empirically optimized hyperparameters derived from preliminary experimentation and biological data analysis best practices.

Hyperparameter-Tuned Models: Systematic optimization through grid search cross-validation across multiple parameter dimensions including tree count, depth limits, sampling requirements, and feature selection strategies. Such optimization procedure determines the set of parameters providing the best predictive results according to the specific characteristics of cell line expression data.

Parameter Grid Specification:

- n_estimators: [100, 200, 300] balancing prediction stability with computational cost
- max_depth: [10, 15, 20, None] exploring complexity vs. overfitting trade-offs
- min_samples_split: [2, 5, 10] investigating split robustness requirements
- min_samples_leaf: [1, 2, 4] assessing leaf node size impact on generalization
- max_features: ['sqrt', 'log2', 0.5] evaluating feature selection strategies

Additional Feature Selection Variants: The integration of secondary feature selection strategies (LASSO regularization and Recursive Feature Elimination) applied to the correlation-selected feature set to evaluate the influence of additional dimensionality reduction on model performance (Guyon et al., 2002).

3.8.3 Training and Validation Framework

A validation system was established to ensure a proper performance evaluation and the model generalizability in diverse biological settings. The validation procedure was designed to address the specific challenges related to the biological data analysis, including the possible batch effects, the sample size restrictions, and the need of the consistent performance across the different contexts of the cell lines.

5-Fold Stratified Cross-Validation: 5-fold cross-validation with stratified sampling is used as the main validation method to ensure that the training and validation folds have a balanced representation of the expression range. The approach provides powerful performance estimates and makes the best use of the available data to train a model.

Train-Test Split Protocol: The final model evaluation was done with a separate 80-20 train-test split, such that the test set was used only to report performance, and to avoid data leakage or optimization bias.

Performance Metrics Comprehensive Assessment: Multiple complementing measures were computed to provide an all-around assessment of model performance.

- **R² Score:** Primary metric indicating proportion of variance explained by sequence features
- **Root Mean Square Error (RMSE):** Absolute prediction error assessment in original expression units
- **Mean Absolute Error (MAE):** Robust error metric less sensitive to outliers than RMSE

3.9 XGBoost Implementation and Comparative Analysis

XGBoost (eXtreme Gradient Boosting) (Chen & Guestrin, 2016) was applied as an additional machine learning model to compare with Random Forest. XGBoost was chosen due to its good results on genomics tasks, and its advantages are regularization, gradient-based optimization, and the ability to deal with complex features interactions (Li et al., 2019). Its incorporation allows more thoroughly analyzing the predictive approaches of modeling cell line expression.

Table 6: XGBoost Advantages for Genomic Applications

Advantage	**Biological Relevance**	**Technical Implementation**	**Expected Benefit**
Sequential Error Correction	Learns from prediction mistakes to improve complex pattern recognition	Gradient boosting with residual learning	Enhanced accuracy on difficult-to-predict sequences
Built-in Regularization	Prevents overfitting in high-dimensional biological data	L1 and L2 penalty integration	Improved generalization with limited samples
Advanced Feature Interactions	Models complex sequence-expression relationships	Tree-based interaction detection	Captures non-additive sequence effects
Efficient Missing Value Handling	Robust performance with incomplete biological measurements	Native missing value processing	Simplified data preprocessing requirements

Early Stopping Mechanisms	Automatically determines optimal model complexity	Validation-based training termination	Prevents overfitting while maximizing performance

The base XGBoost model was configured with n_estimators=300, max_depth=6, learning_rate=0.1, subsample=0.8, colsample_bytree=0.8, reg_alpha=0.1, reg_lambda=1.0, random_state=42, n_jobs=-1.

3.9.1 Optimization Strategy

Randomized search cross-validation was used as hyperparameter optimization. It is a strategy that draws samples of defined parameter distributions; it is more effective than the exhaustive grid search in high-dimensional spaces.

Parameter Distribution Specifications:

- n_estimators: [200, 300, 500] investigating boosting round requirements
- max_depth: [4, 6, 8] exploring tree complexity vs. overfitting balance
- learning_rate: [0.05, 0.1, 0.15] assessing convergence speed vs. stability
- subsample: [0.8, 0.9] evaluating bootstrap sampling impact
- colsample_bytree: [0.8, 0.9] testing feature subsampling strategies
- reg_alpha: [0, 0.1, 0.5] examining L1 regularization effects
- reg_lambda: [1, 1.5, 2] investigating L2 regularization strength

A total of 50 random parameter combinations were evaluated using 3-fold cross-validation with R^2 as the scoring metric, balancing model accuracy and computational feasibility across 13 cell lines.

3.9.2 Performance Metrics and Evaluation Framework

Comprehensive Performance Assessment: Multiple complementary metrics were calculated for both algorithms to provide complete characterization of their predictive capabilities:

- **R^2 Score: Primary metric for assessing explained variance and biological relevance**
- **RMSE: Absolute prediction error in original expression units**
- **MAE: Robust error metric providing additional perspective on prediction accuracy**
- **Feature Importance Rankings: Comparative analysis of which sequence features each algorithm identifies as most predictive**

.

3.10 Model Performance Evaluation and Statistical Analysis

The post-modeling evaluation was done thoroughly to verify the predictive quality, generalization capacity, and statistical stability of all settings of machine learning which were applied. This was an important step towards interpreting model behaviour under transformations, algorithms and feature selection methods and gaining insight into the relative importance of model tuning and data normalization.

3.10.1 Data Aggregation and Metric Computation

A Python class (ModelPerformanceVisualizer) was developed specifically to read and process the output of 18 different model fits, including all combinations of two algorithms (Random Forest and XGBoost), three expression data transformations (Raw, Z-score, and log 2-fold change) and three model types (Basic, Tuned, LASSO-based feature subsets). Every setting generated a CSV file with performance statistics (R 2, RMSE, MAE) in 13 cell lines. The results were compiled into a single DataFrame to perform meta-analysis. Cross-cell line consistency and predictive robustness were assessed by calculating summary statistics such as mean, minimum, maximum, and standard deviation of R^2.

3.10.2 Visualization Framework and Interpretability Analysis

All figures were created using matplotlib and seaborn with modified aesthetic parameters (e.g., Times New Roman font, monochromatic color schemes, black-edged plots) suited for scientific presentation and publication, therefore guaranteeing academic-grade visual consistency. Important visual outputs were:

- **Overall Model Ranking:** Horizontal bar plot ranked average R^2 across all configurations
- **Transformation Effects:** Grouped bar plots showing how data normalization (Raw, Z-score, $\log_2$FC) impacted model accuracy across algorithms and model types.
- **Raw vs. Z-score Correlation:** Scatter plot comparing R^2 scores from Raw vs. Z-score transformations, annotated by model, with a perfect-correlation reference line and mean absolute difference reported.
- **Hyperparameter Tuning Impact:** Bar plots comparing baseline vs. tuned models, including percentage improvement labels for each configuration.
- **Model Type Distribution:** Boxplots visualizing R^2 score distributions across Basic, Tuned, and LASSO-based model variants.
- **XGBoost Preprocessing Independence:** Visual confirmation that XGBoost performance remained invariant across preprocessing strategies for each model type, verified by constant R^2 scores across input transformations.

3.10.3 Tools and Reproducibility

All scripts were developed in Python 3.8+ using pandas, numpy, matplotlib, and seaborn. The analysis pipeline was modularized and reproducible, with standardized directory input and academic figure export functionality.

3.11 Advanced Model Evaluation and Statistical Comparison

In this section of analysis, we designed a advance statistical model of post-hoc performance evaluation to gain better insight about the comparative model behaviour. This framework constituted strict statistical tests, visual diagnostics, and model ranking analyses with 18 machine learning models.

3.11.1 Aggregation of Performance Data and Metric Consolidation

The results of all models were gathered based on 18 different configurations that consisted of combinations of two algorithms (Random Forest, XGBoost), three types of data transformation (Raw, Z-score, Log 2 FC), and three types of models (Basic, Tuned, LASSO-based). All of the configurations were tested on 13 cancer cell lines, which provided the performance matrices of R^2, RMSE, and MAE.

For each model, we computed:

- Mean R^2, RMSE, and MAE
- Standard deviation (SD) and standard error of the mean (SEM)
- Raw R^2 and RMSE values for every cell line

A custom Python class, **FixedAdvancedModelAnalyzer**, was used to organize this data, which made the results reproducible and consistent between comparative assessments.

3.11.2 Visualization of Predictive Power with SEM

Mean R^2 scores were plotted with standard error bars at each of the configurations to illustrate central tendency and uncertainty in performance. A horizontal bar plot was used with coloring by algorithm and ranking in the descending order of mean R^2.

As recommended by Cumming et al. (2007), **Standard Error of the Mean (SEM)** was used to represent variability across replicates (cell lines), providing a biologically grounded view of model stability and generalizability.

3.11.3 Statistical Significance Testing via Friedman-Nemenyi Analysis

Although average R^2 scores are easy to interpret as an indication of model performance, they cannot alone tell whether differences between models are statistically significant. Accordingly, a non-parametric statistical system was adopted to formally evaluate variations in performance among model settings. This was necessary because the sample size was small (13 cell lines) and normality assumptions could be violated as is common with biological data.

Friedman Test for Repeated Measures

As the main inferential statistic, the Friedman test (non-parametric alternative to repeated-measures ANOVA) was utilized (Friedman, 1940). It checks the statistical hypothesis that all model settings work equally well on many datasets (here: cell lines), based only on the ranking of the model performances on each dataset. The test was quite appropriate in our evaluation design, in which each of the models was fitted to the 13 cell lines and resulted in paired observations under different conditions.

- For each of the N=13 cell lines, the 18 models were **ranked** based on their R^2 scores.
- Lower ranks corresponded to **better performance** (i.e., rank 1 for the highest R^2).
- The test statistic:

  $$\chi_F^2 = \frac{12n}{k(k+1)}\left[\sum_{j=1}^{k}\bar{R}_j^2\right] - 3n(k+1)$$

  where *n* is the number of datasets (cell lines), *k* is the number of models, and $R\bar{j}$ is the average rank of model *j*.
- The Friedman test indicated **statistically significant differences** among models ($p < 0.05$), justifying further pairwise comparisons.

Nemenyi Post-Hoc Test and Critical Difference

A Nemenyi post-hoc test (Demšar, 2006) was conducted to determine which pairs of models had significant differences. In this test the average ranks of each pair of models are compared with a value of Critical Difference (CD):

$$CD = q_\alpha \cdot \sqrt{\frac{k(k+1)}{6n}}$$

- q_α is the critical value from the Studentized range statistic at $\alpha = 0.05$

- *k* is the number of models (18)

- *n* is the number of datasets (13)

Model pairs with average rank differences that exceeds the CD threshold are considered significantly different at the given alpha level.

Critical Difference Diagram Visualization

To get an intuitive feel of the result of the Nemenyi test, a Critical Difference Diagram was created:

- The models are ranked on a horizontal axis based on average rank (low = good).
- The horizontal lines between models imply that the difference between them is not found to be significant.
- The Nemenyi test indicates significant difference between disconnected models (i.e. those that are not connected with a bar).

It is a high-level summary of performance groups of models, showing both the best-performing set of configurations as well as those that are statistically indistinguishable. It assists in preventing an over-interpretation of differences in marginal R^2, and it allows more evidence-based Choice of modelling strategies.

3.11.4 Implementation and Reproducibility

Python 3.8+ libraries including pandas, numpy, scipy.stats, and matplotlib were used across all statistical techniques. Visualizations used seaborn with scientific themes (Times New Roman font, black-edged bars, annotated legends).

3.12 Implementation Details and Computational Considerations

All analyses were performed in a version-controlled and well-defined software environment to enable reproducibility, computational efficiency and independence of the analysis platform. The entire analytical pipeline was written in Python and ordered a collection of scientific computing and machine learning libraries that are seeing extensive use in computational biology and bioinformatics.

Table 7: Complete Software Environment Specification

Component	Version Requirement	Primary Functions
Python	3.8+	Core programming language and ecosystem coordination
pandas	1.3+	Data manipulation, preprocessing, and analysis

NumPy	1.21+	Numerical computing foundation and array operations
scikit-learn	1.0+	Machine learning algorithms and evaluation frameworks
XGBoost	1.5+	Gradient boosting implementation with GPU support
SciPy	1.7+	Statistical functions and hypothesis testing
matplotlib	3.5+	Publication-quality visualization generation
seaborn	0.11+	Advanced statistical graphics and distribution analysis

Chapter 4: Results

4.1 Data Preprocessing, Feature Engineering, and Exploratory Analysis

The section reflects the findings of a thorough data preparation procedure, which began with an examination of the pattern of expressions through varying normalization techniques, proceeded to employ principal component analysis to visualize the relationship between cell lines, and lastly resulted in thoughtful feature engineering and selection. This reduced the original set of 3,311 miRNA sequences (123 features) through an analysis process to a more machine-learning-suitable set of 31 important sequence factors. Beginning with 3,311 miRNA sequences with 123 features, the analysis was reduced to 31 significant sequence factors more appropriate to use in machine learning.

4.1.1 Expression Data Distribution Analysis

Figure shows how preprocessing affects the way 'WT' data is spread out among different cancer cell lines, which is an important step for using machine learning to predict how miRNA variants work. The leftmost panel, 'Raw Data Distribution', shows that the data varies a lot between different cell lines, with different shapes, means, and deviations, including some that have two HCT116). This raw variability requires preprocessing to prevent biased model learning. The central panel, 'Z-Score Normalized Distribution', demonstrates how Z-score normalization effectively centers all distributions around zero and reduces variance, standardizing feature scales. This process is vital for distance-sensitive algorithms, ensuring that feature magnitudes do not disproportionately influence models. The rightmost panel, 'Log2 Fold Change Distribution', shows that Log2 fold-change transformation further compacts and symmetrizes distributions, emphasizing relative changes in expression. This transformation is particularly beneficial in biological contexts because it highlights biologically significant variations and is well-suited for the machine learning tasks in this thesis.

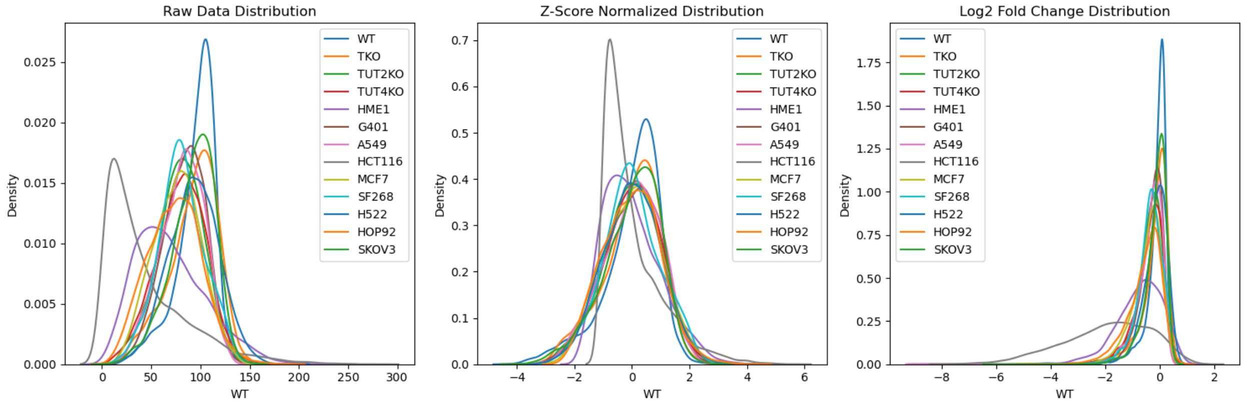


Figure 3: Density Distribution of Three Datasets

4.1.2 Principal Component Analysis

The figure below provides a comparison of Principal Component Analysis (PCA) of the data in three different preprocessing modes, namely raw, Z-score normalized, and Log2 fold-change (Log2FC) transformed. In the PCA of the raw data, the samples are dispersed across the two principal components (PC1 and PC2) with much variation and without any obvious clustering of the various cell lines. Following Z-score normalization, data points are likelier concentrated around the center, and the magnitude of the principal components is lesser, which indicates that variances of features have been made equal. Nevertheless, clear segregation or close grouping of certain cell lines is not very obvious. Conversely, the PCA performed on the Log2FC transformed data clearly exhibits a clustering of some of the cell lines more so those that have similar biological characteristics and also identifies a few outliers that have distinct expression profiles. This observation supports the idea that the Log2FC transformation is useful to magnify relative variations in feature expression and optimize the visualization of the underlying biological relationships and differences between the samples, which is essential to discern meaningful patterns in high-dimensional biological data. These differences observed help to prove the significant impact of data preprocessing on the interpretability of dimensionality reduction method to highlight the heterogeneity and sample relationships in a complex biological dataset.

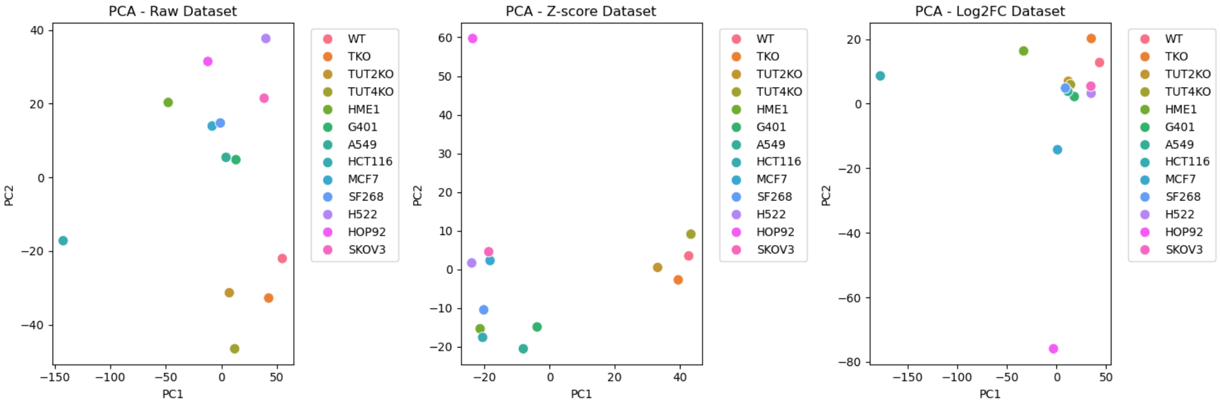
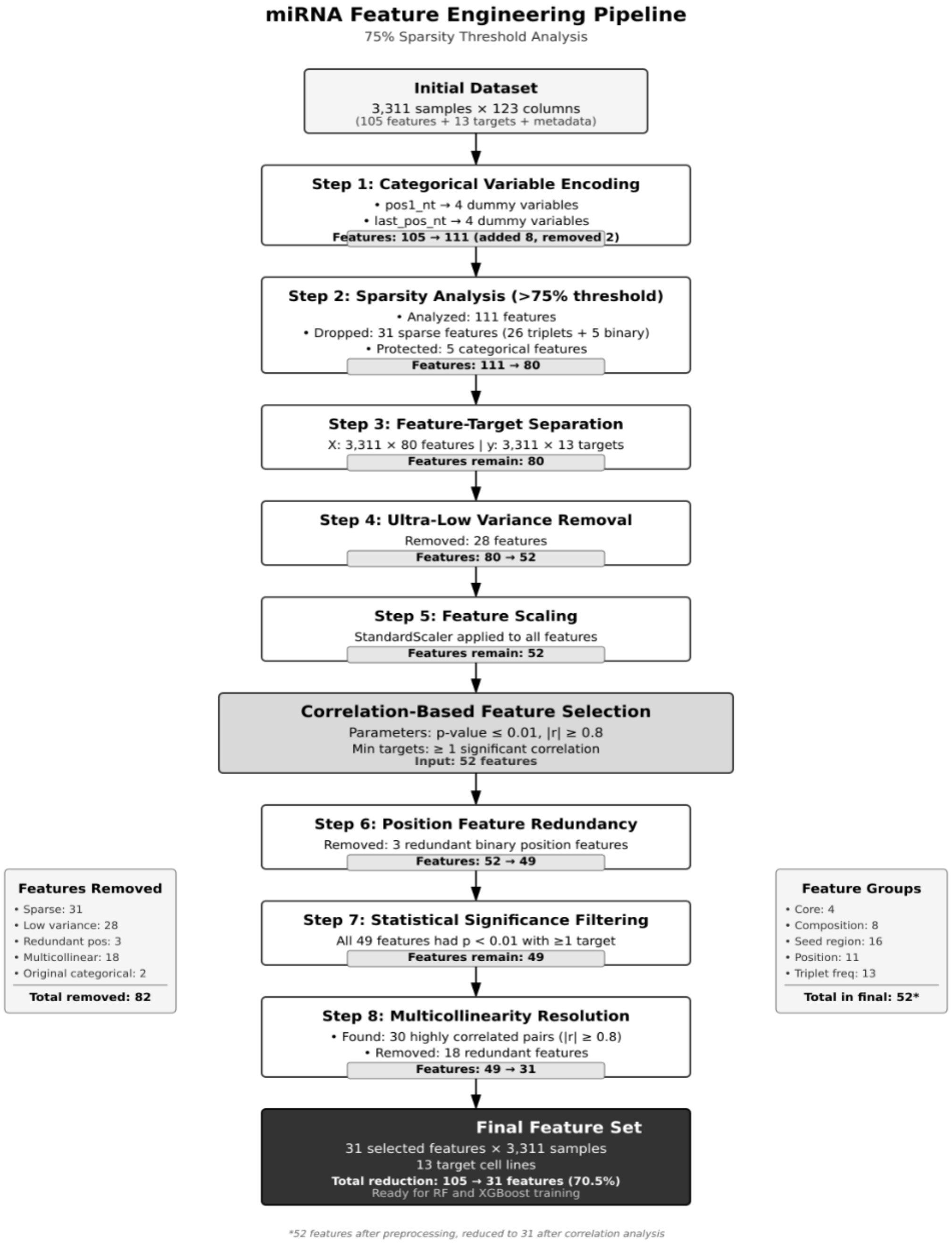


Figure 4: PCA Component Analysis of Three Datasets

4.1.3 Feature Engineering and Selection

This flowchart below represents a step wise method to enhance quality of miRNA expression data to predict multiple targets with 13 cell lines, to decrease features and raise accuracy. The preprocessing step starts with 3,311 miRNA sequences characterized by 105 features. It includes transforming nucleotide positions into 8 dummy variables, dropping 31 mostly empty features, retaining 5 valuable categorical features, and dropping 28 features with extremely minimal variation, and standardizing the data, producing a grand total of 52 features. The second step in the procedure eliminates redundancy among the positional features (3 features), ensures that all 49 features are statistically significant (with $p < 0.01$ in at least one target) and it also minimizes multicollinearity by dropping 18 features among 30 pairs that are strongly correlated (|r| >= 0.8) based on their biological significance. It shrinks the feature space by 70.5 per cent, (105 to 31 features) and preserves substantial predictors in five categories: core features (4), nucleotide composition (8), seed region characteristics (16), position encodings (11) and triplet frequencies (13). Such an approach is statistically significant, decreases computational burden, and minimizes the chances of overfitting in subsequent Random Forest and XGBoost models.



miRNA Feature Engineering Pipeline
75% Sparsity Threshold Analysis
Initial Dataset
3,311 samples × 123 columns
(105 features + 13 targets + metadata)
Step 1: Categorical Variable Encoding
• pos1_nt → 4 dummy variables
• last_pos_nt → 4 dummy variables
Features: 105 → 111 (added 8, removed 2)
Step 2: Sparsity Analysis (>75% threshold)
• Analyzed: 111 features
• Dropped: 31 sparse features (26 triplets + 5 binary)
• Protected: 5 categorical features
Features: 111 → 80
Step 3: Feature-Target Separation
X: 3,311 × 80 features | y: 3,311 × 13 targets
Features remain: 80
Step 4: Ultra-Low Variance Removal
Removed: 28 features
Features: 80 → 52
Step 5: Feature Scaling
StandardScaler applied to all features
Features remain: 52
Correlation-Based Feature Selection
Parameters: p-value ≤ 0.01, |r| ≥ 0.8
Min targets: ≥ 1 significant correlation
Input: 52 features
Step 6: Position Feature Redundancy
Removed: 3 redundant binary position features
Features: 52 → 49
Features Removed
• Sparse: 31
• Low variance: 28
• Redundant pos: 3
• Multicollinear: 18
• Original categorical: 2
Total removed: 82
Step 7: Statistical Significance Filtering
All 49 features had p < 0.01 with ≥1 target
Features remain: 49
Feature Groups
• Core: 4
• Composition: 8
• Seed region: 16
• Position: 11
• Triplet freq: 13
Total in final: 52*
Step 8: Multicollinearity Resolution
• Found: 30 highly correlated pairs (|r| ≥ 0.8)
• Removed: 18 redundant features
Features: 49 → 31
Final Feature Set
31 selected features × 3,311 samples
13 target cell lines
Total reduction: 105 → 31 features (70.5%)
Ready for RF and XGBoost training
*52 features after preprocessing, reduced to 31 after correlation analysis

Figure 5: Flowchart of the Feature Engineering Framework

4.2 Models Selection and Comparative Performance Analysis

All eighteen model setups were trained and tested using the same dataset of miRNA variant effects in cancer cell lines, but the data was presented in three different ways: raw expression values, z-score normalized values, and log2 fold change (Log2FC) transformed values. This setup allows for a precise comparison of the impacts of different preprocessing techniques on model performance, while controlling for variations in data content.

4.2.1 Overview of Model Performance Rankings

Using three main criteria—R^2, RMSE, and MAE—eighteen model configurations were assessed. The Random Forest model that was trained on Log2FC-transformed data and adjusted (RF_Log2FC_Tuned) always performed better than all the other models, as shown in Figure 4.1 and confirmed by the critical difference diagram (Figure 4.8). The model achieved the best predicted accuracy with $R^2 = 0.4614 \pm 0.0228$, RMSE = 0.4320, and MAE = 0.3208.

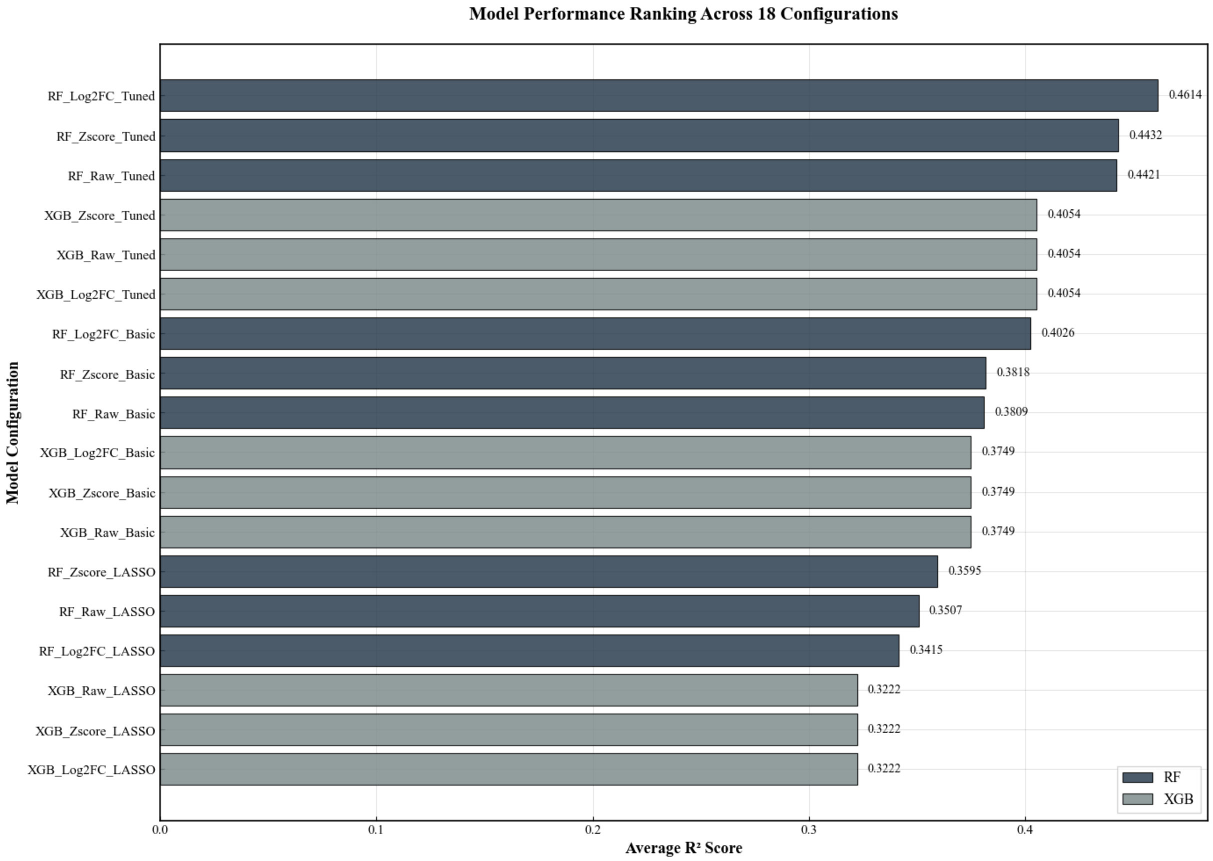


Figure 6: Model Performance Ranking Across 18 Configurations. The horizontal bar chart displays the R-scores of all Random Forest (RF) and XGBoost (XGB) models under various preprocessing methods and tuning conditions.

Although Random Forest models using z-score normalization (RF_Zscore_Tuned: R² = 0.4432 ± 0.0294) and raw data (RF_Raw_Tuned: R² = 0.4421 ± 0.0245) shown identical ability to describe the data, however, they had larger error which might be observed due to scaling. The XGBoost model using Log2FC-transformed data (XGB_Log2FC_Tuned) had R² of 0.4054 ± 0.00307, but it showed considerably higher RMSE (19.6392) and MAE (15.1949), which suggests that why it was less reliable in case of making overall predictions.

LASSO-regularized models had the lowest R² scores and the least variation in performance (for example, XGB_Log2FC_LASSO: R² = 0.3222 ± 0.0305), suggesting they may have missed important information because they were too strict in choosing which features to include.

4.2.2 Similarity of Raw Data and Z-score Normalization

Particularly for Random Forest models, figure 4.2 indicates that the performance of raw data and z-score adjusted data is identical across all measures. With an average difference of just 0.50% between the raw and z-score R² values, the scatter plot indicates most models to be close to the ideal correlation line. Still, the error metrics offer a more thorough grasp of the useful benefits of normalizing.

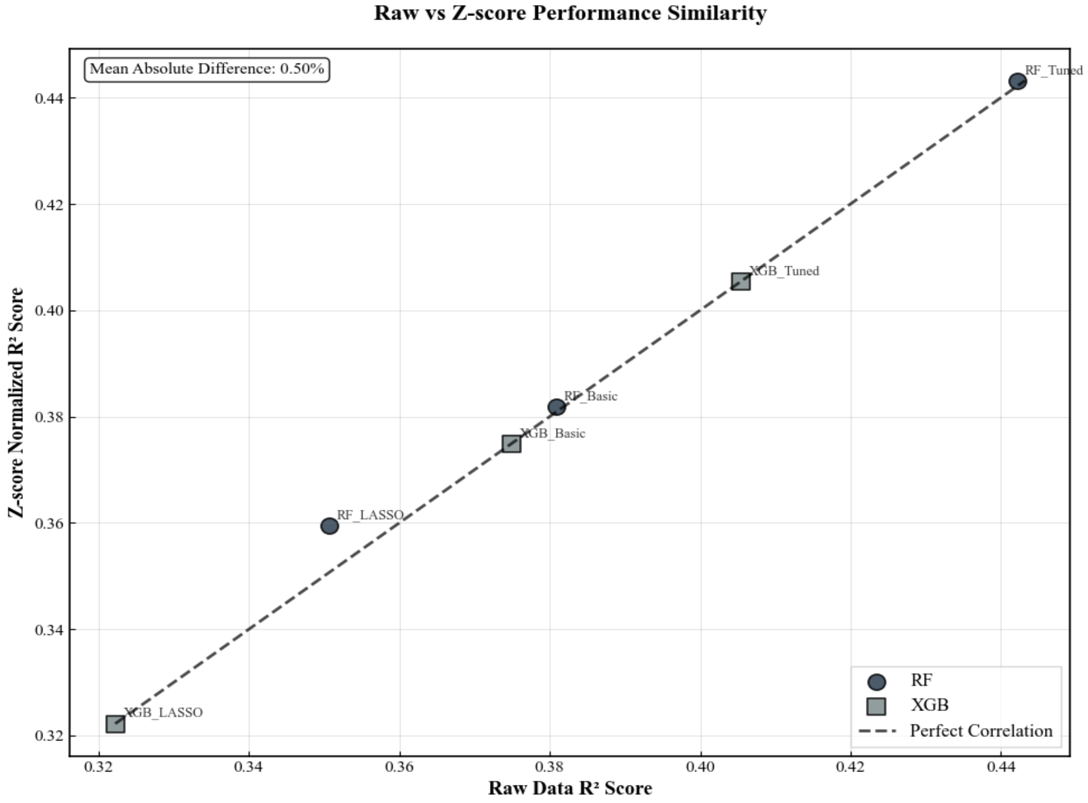
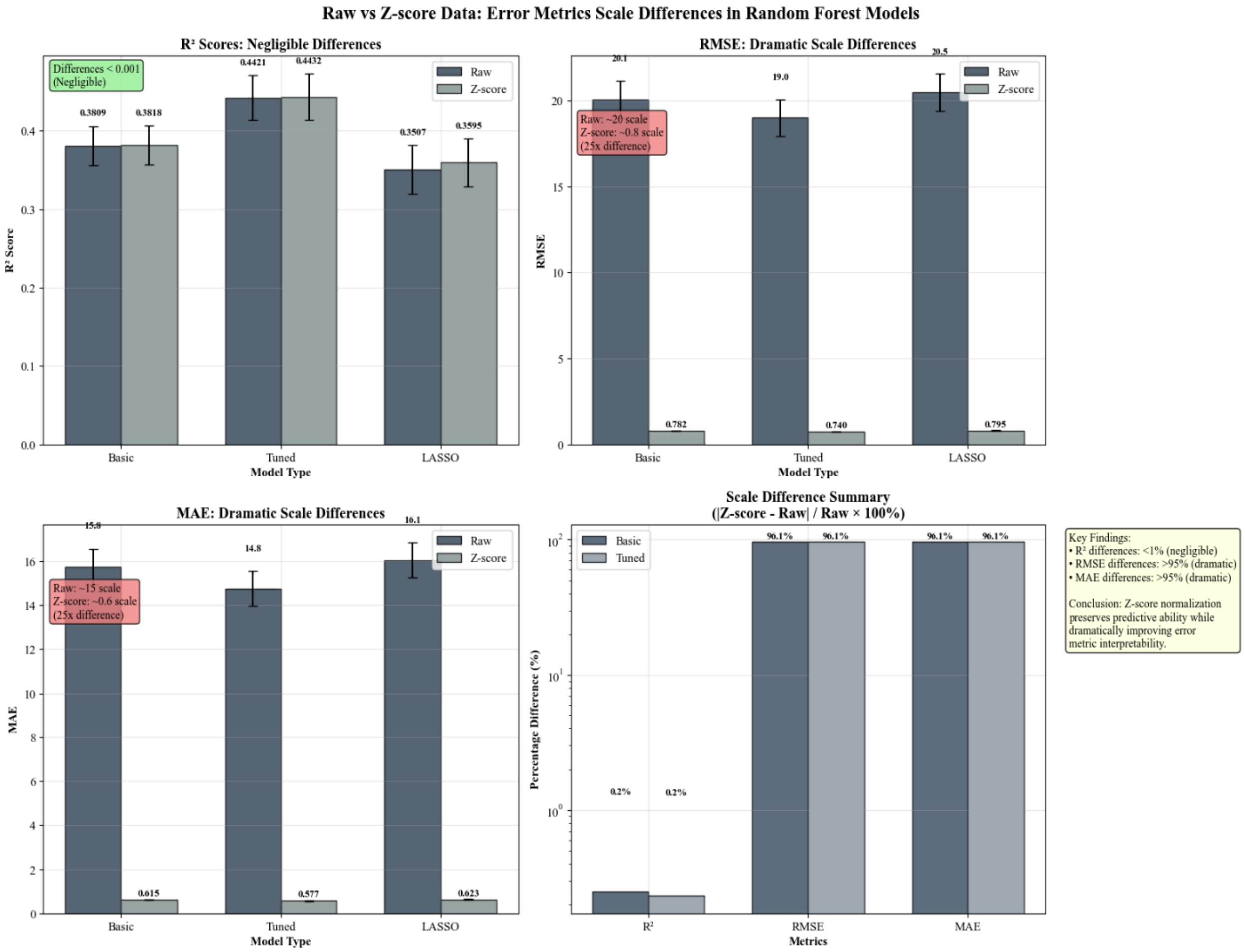


Figure 7: Raw vs. Z-score Normalization Performance Similarity. Scatter plot comparing R² scores between raw and z-score preprocessing for each model configuration, showing a mean absolute difference of only 0.50%.

While R² changes remained small across all variants (RF_Basic: 0.3809 vs 0.3818; RF_Tuned: 0.4421 vs 0.4432; RF_LASSO: 0.3507 vs 0.3595), the error metrics exhibited notable variations in scale among the Random Forest models. Unlike the models that used z-score normalization, which had RMSE values between 0.740 and 0.795 and MAE values from 0.577 to 0.621, the models using raw data had much higher error levels, with RMSE between 19.0 and 20.5 and MAE between 14.8 and 16.1. This variation suggests, over all Random Forest versions, a 25-fold scale change (figure 4.3). The results indicate that even though the predictive performance remains similar (with R² differences less than 1%), using z-score normalization makes it much easier to understand the error measurements.



Key Findings:
- R² differences: <1% (negligible)
- RMSE differences: >95% (dramatic)
- MAE differences: >95% (dramatic)

Conclusion: Z-score normalization preserves predictive ability while dramatically improving error metric interpretability.

Figure 8: Raw vs. Z-score Preprocessing Effects on Random Forest Models. Comparison of R², RMSE, and MAE across preprocessing methods, showing negligible R² differences but large RMSE and MAE scale reductions with z-score normalization.

This similarity suggests that z-score normalization mainly makes the scales of the features uniform while keeping the basic relationships that Random Forest algorithms rely on intact. Keeping the significant patterns discovered by correlation-based feature selection helps both approaches predict results similarly even if they explain findings differently.

4.2.3 Effectiveness of Hyperparameter Tuning

The noticeable differences seen in the bars for each model-dataset combination in Figure 4.4 show that hyperparameter tuning had different effects on each one, as it was used for all combinations. By adjusting the settings for all measures, Random Forest models showed clear improvements, leading to R² increases of 14.6–16.1% and significant drops in both RMSE and MAEs. RF_Log2FC, for example, rose from R2 = 0.4026 (RMSE = 0.4558, MAE = 0.3438) to R2 = 0.4614 (RMSE = 0.4320, MAE = 0.3208), therefore suggesting an increase in explained variance and prediction accuracy.

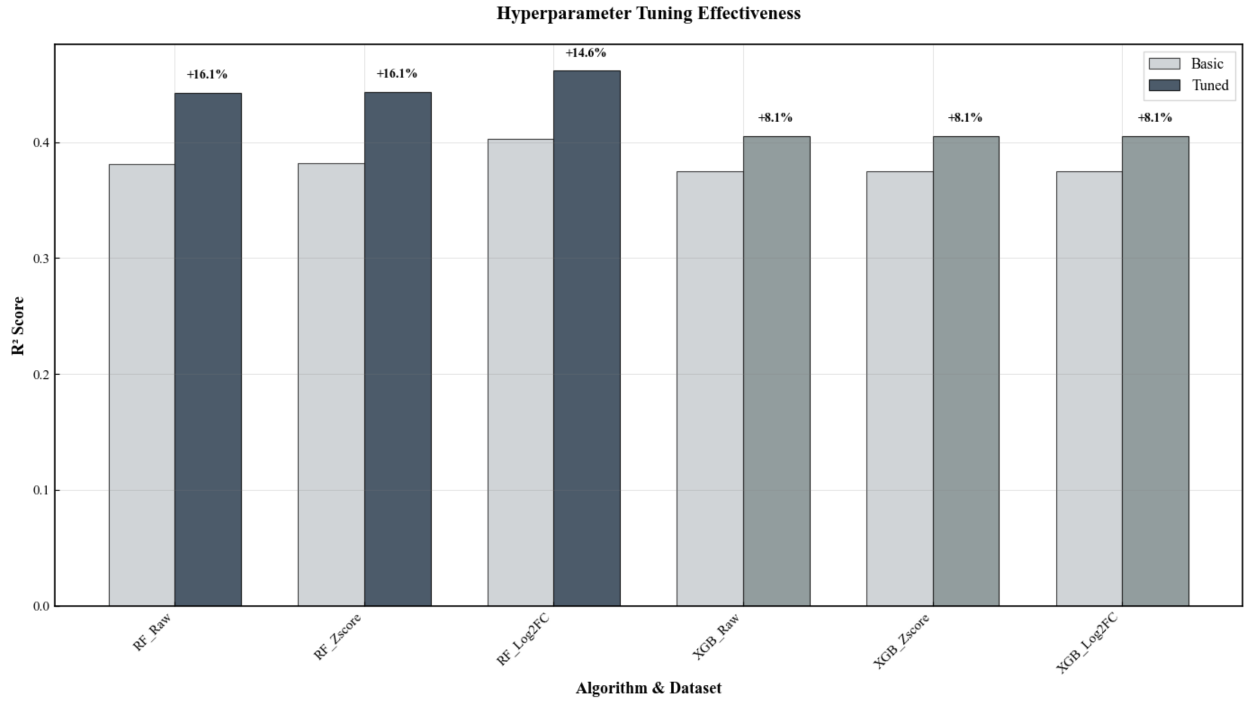


Figure 9: Hyperparameter Tuning Effectiveness. Bar chart showing percentage improvements in R² scores after tuning for each model and dataset (no error bars).

With adjustments across all preprocessing techniques, XGBoost models showed a consistent improvement of 8.1% in R2, paired with improvements in error measures. Together with consistent performance across preprocessing techniques, the steady improvement noted emphasizes XGBoost's preprocessing independence. On the other hand, it shows that Random Forest methods receive more benefits from systematic optimization in biological prediction applications and show increased sensitivity to hyperparameters.

The observed improvement trends across preprocessing techniques imply that the impacts of hyperparameter tweaking are mostly independent of data transformation methods, therefore permitting the construction of optimization algorithms without the necessity of substantial calibration, which is particular to preprocessing.

4.2.4 Model Type Performance, Reliability, and Statistical Significance

Figure 4.5 offers an in-depth look at model performance distributions using R² as the principal metric. Tuned models did much better than the others, with a median R² of about 0.424 and the smallest range of results, showing they are reliable and good at making predictions. Random Forest variants within this group also exhibited the lowest RMSE and MAE values across their respective preprocessing formats, reinforcing their overall superiority.

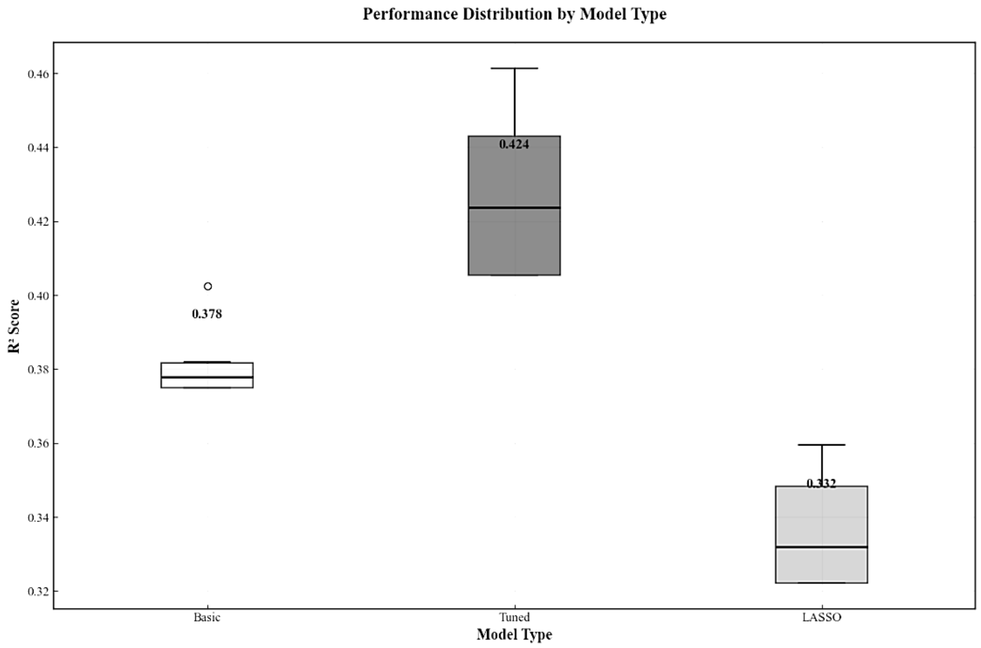


Figure 10: Performance Distribution by Model Type. Boxplot summarizing the R² performance distribution for Basic, Tuned, and LASSO model variants.

In comparison, basic models performed moderately well (median R² ≈ 0.378), while LASSO-regularized models did the worst, showing the lowest median R² (≈ 0.332) and little variation. The error metrics (RMSE and MAE) support this trend, indicating that using correlation-based feature selection kept important features, while strong LASSO filtering caused a loss of valuable information.

Statistical significance and reliability of these findings were confirmed through the Nemenyi test (Figure 4.6). RF_Log2FC_Tuned had the highest average rank (1.77), closely followed by RF_Raw_Tuned and RF_Zscore_Tuned (both 2.31), yet these differences were not statistically significant at $\alpha = 0.05$. This result indicates comparable predictive performance among the top three Random Forest configurations.

All tuned XGBoost variants clustered together with identical ranks (6.69), highlighting their insensitivity to preprocessing. Conversely, LASSO configurations, particularly XGBoost LASSO, ranked lowest (16.15), affirming the negative impact of excessive feature elimination.

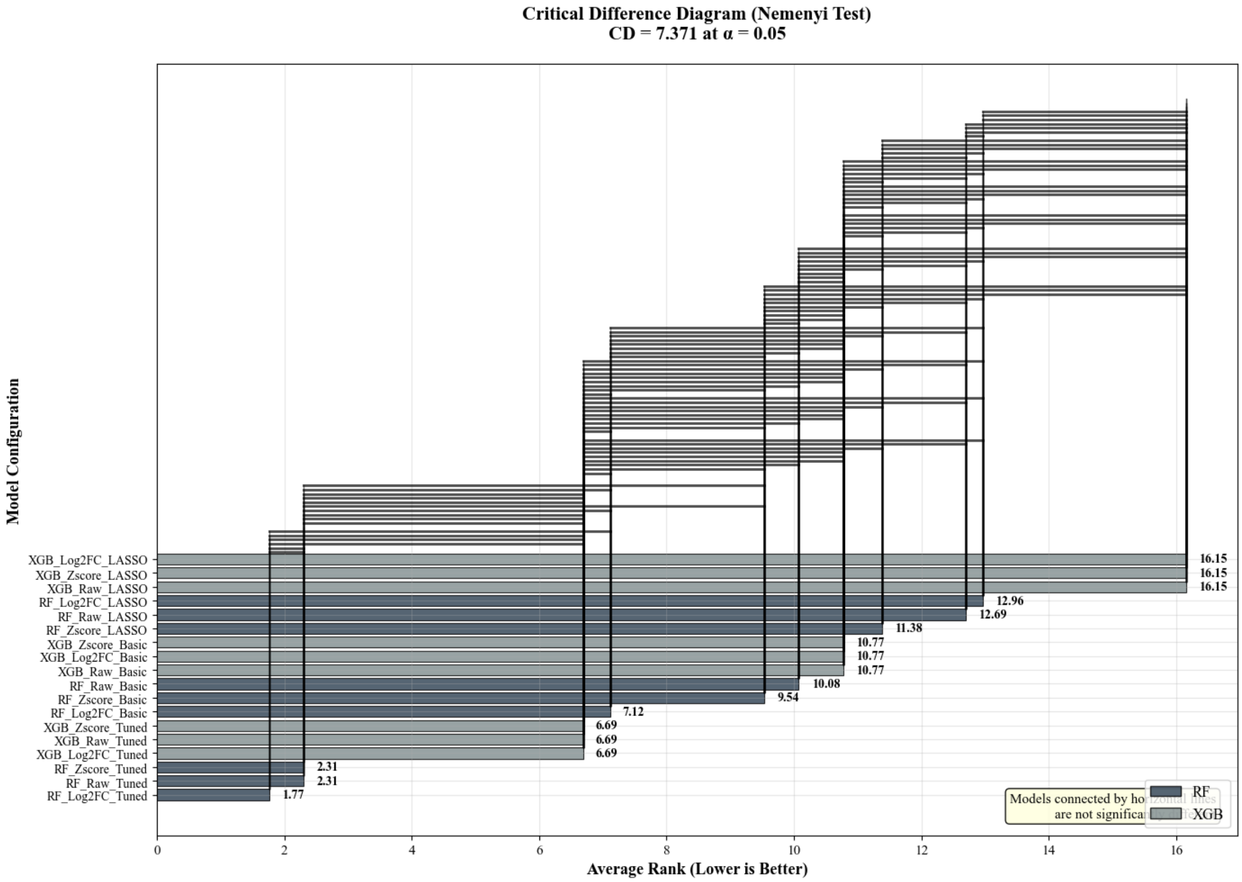


Figure 11: Critical Difference Diagram (Nemenyi Test). Ranking of all model configurations by average performance, with statistically insignificant differences marked by horizontal connections.

Figure 4.7 provides additional information about model reliability using standard error bars. RF_Log2FC_Tuned not only achieved the highest mean R² (0.4614) but also showed the lowest SEM (±0.0228), affirming both its accuracy and consistency. RF_Zscore_Tuned and RF_Raw_Tuned also maintained strong reliability (SEM ±0.0294 and ±0.0285, respectively). In contrast, XGBoost models displayed consistent SEM values (±0.0307), but with slightly greater variability. LASSO models—especially RF_LASSO—showed the highest SEM, reflecting unstable generalization across biological conditions.

Moreover, while R² values remained relatively stable across preprocessing approaches, RMSE and MAE varied drastically due to data scale effects. Models based on raw data had the biggest errors (for example, RF_Basic: RMSE = 20.0561, MAE = 15.7526), but using z-score normalization made the results much easier to understand without losing accuracy (for example, RF_Basic: RMSE = 0.7815, MAE = 0.6153).

Ultimately, Log2FC-transformed data produced the most precise and biologically relevant predictions, especially for Random Forest models. RF_Log2FC_Tuned had the best performance, with the lowest RMSE (0.4320) and MAE (0.3208), and a strong R² (0.4614), making it the best choice for predicting how miRNA variants affect cancer cell lines.

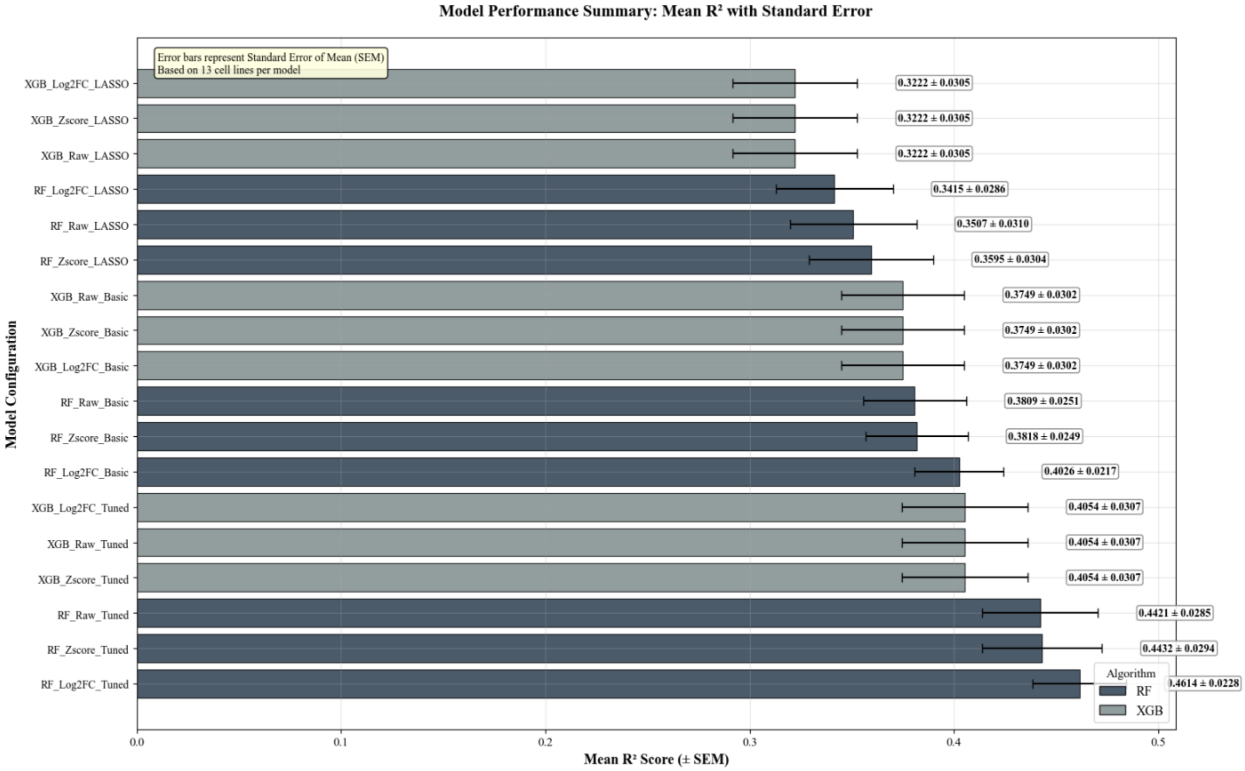


Figure 12: Model Performance Summary with Standard Error. Horizontal bar chart illustrating mean R² values and standard errors for all model configurations.

4.2.5 Implications for Science and Technology

The better performance of Log2FC transformation in Random Forest models shows that changes in gene expression are important signals that these models can effectively recognize. This highlights the necessity of aligning preprocessing techniques with the strengths of algorithms in biological modeling.

XGBoost's ability to handle data without needing much preprocessing is fast, but it might hide important biological information found in features that depend on scale. The poor results from LASSO-regularized models show that reducing the number of features too much can eliminate important biological information that correlation-based selection has kept.

These findings highlight the importance of careful feature selection, adjusting models to account for changes, and understanding results in a biological context. Random Forest models, especially RF_Log2FC_Tuned, are a strong choice for future research that values clear understanding and responsiveness to biological changes.

4.2.6 Recommendation for the Final Model

We have trained the Random Forest on Log2FC-transformed data and tuned it (RF_Log2FC_Tuned), making it the optimal model. It offers:

- The Random Forest model achieved the highest predictive accuracy among all evaluated configurations, with an R-value of 0.4614 to 0.0228.
- The analysis of feature importance enhances the interpretability of the results.
- Sensitivity to regulatory patterns based on fold change
- Consistent performance among biological replicates (minimal SEM)

We also suggest RF_Zscore_Tuned ($R^2 = 0.4432 \pm 0.0294$) as an additional model. Although it exhibits marginally lower performance compared to RF_Log2FC_Tuned, the difference lacks statistical significance and provides robustness via standardized scaling.

XGBoost models, while stable, are not ideal for biological interpretation because of their insensitivity to preprocessing and reduced interpretability. The suggested two-model approach offers better prediction results and helps explain how miRNA variant effects work.

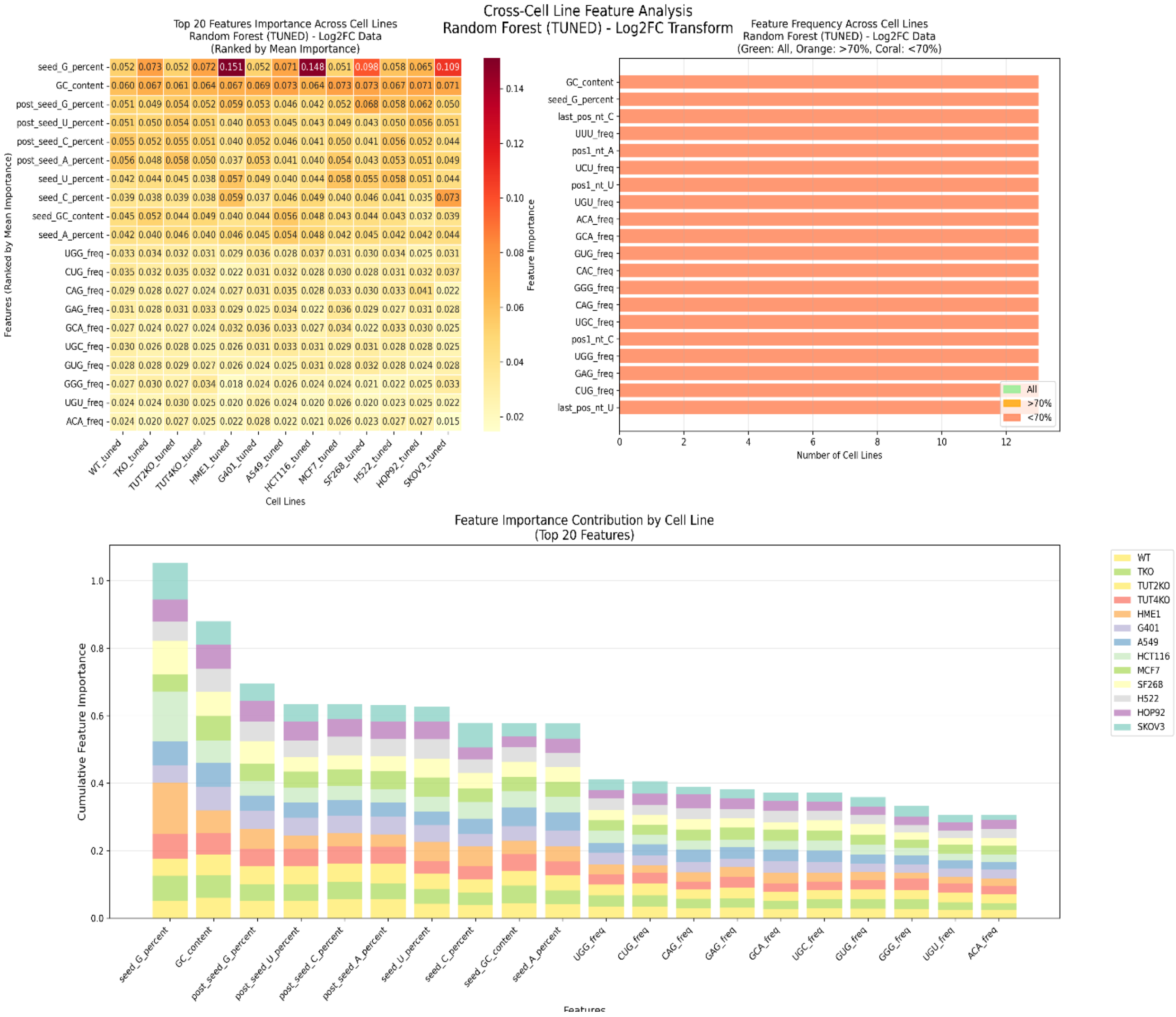
4.3 Analysis of Feature Importance

Following selection of RF_Log2FC_Tuned and RF_Zscore_Tuned as the best models, we investigated which features most significantly influence miRNA variant effect on function. Examining the top 20 traits across all 13 cell lines for both models, this study identified consistent, context-dependent regulating variables. Comparative feature importance profiles for the Log2FC and Z-score models are shown in Figures 4.10 and 4.11. Every figure consists of a stacked bar chart displaying cumulative feature contributions for every cell line, a frequency bar chart displaying the occurrence of features in top rankings, and a heatmap displaying feature importance scores across cell lines.

4.3.1 Main Feature Classes and Biological Significance

The primary influencing factor is seed region composition as it has been found at the top of the list for both models. With both features, seed_G_percent and GC_content, included in the top 20 across all cell lines,the earlier one showed the highest ranking among cell lines (0.052–0.109) in the Log2FC model closely followed by later one. Highlighting the biological relevance of guanine content in seed areas for target recognition and binding, the Z-score model showed a similar tendency.

Post-seed contributions: Features related to post-seed, such post-seed_G_percent and post-seed_U_percent, shown great relevance. Their relevance in both models indicates that miRNA binding and stability depend on nucleotides close to the seed region as well.



	WT_tuned	TKO_tuned	TUT2KO_tuned	TUT4KO_tuned	HME1_tuned	G401_tuned	A549_tuned	HCT116_tuned	MCF7_tuned	SF268_tuned	H522_tuned	HOP92_tuned	SKOV3_tuned
seed_G_percent	0.052	0.073	0.052	0.072	0.151	0.052	0.071	0.148	0.051	0.098	0.058	0.065	0.109
GC_content	0.060	0.067	0.061	0.064	0.067	0.069	0.073	0.064	0.073	0.073	0.067	0.071	0.071
post_seed_G_percent	0.051	0.049	0.054	0.052	0.059	0.053	0.046	0.042	0.052	0.068	0.058	0.062	0.050
post_seed_U_percent	0.051	0.050	0.054	0.051	0.040	0.053	0.045	0.043	0.049	0.043	0.050	0.056	0.051
post_seed_C_percent	0.055	0.052	0.055	0.051	0.040	0.052	0.046	0.041	0.050	0.041	0.056	0.052	0.044
post_seed_A_percent	0.056	0.048	0.058	0.050	0.037	0.053	0.041	0.040	0.054	0.043	0.053	0.051	0.049
seed_U_percent	0.042	0.044	0.045	0.038	0.057	0.049	0.040	0.044	0.058	0.055	0.058	0.051	0.044
seed_C_percent	0.039	0.038	0.039	0.038	0.059	0.037	0.046	0.049	0.040	0.046	0.041	0.035	0.073
seed_GC_content	0.045	0.052	0.044	0.049	0.040	0.044	0.056	0.048	0.043	0.044	0.043	0.032	0.039
seed_A_percent	0.042	0.040	0.046	0.040	0.046	0.045	0.054	0.048	0.042	0.045	0.042	0.042	0.044
UGG_freq	0.033	0.034	0.032	0.031	0.029	0.036	0.028	0.037	0.031	0.030	0.034	0.025	0.031
CUG_freq	0.035	0.032	0.035	0.032	0.022	0.031	0.032	0.028	0.030	0.028	0.031	0.032	0.037
CAG_freq	0.029	0.028	0.027	0.024	0.027	0.031	0.035	0.028	0.033	0.030	0.033	0.041	0.022
GAG_freq	0.031	0.028	0.031	0.033	0.029	0.025	0.034	0.022	0.036	0.029	0.027	0.031	0.028
GCA_freq	0.027	0.024	0.027	0.024	0.032	0.036	0.033	0.027	0.034	0.022	0.033	0.030	0.025
UGC_freq	0.030	0.026	0.028	0.025	0.026	0.031	0.033	0.031	0.029	0.031	0.028	0.028	0.025
GUG_freq	0.028	0.028	0.029	0.027	0.026	0.024	0.025	0.031	0.028	0.032	0.028	0.024	0.028
GGG_freq	0.027	0.030	0.027	0.034	0.018	0.024	0.026	0.024	0.024	0.021	0.022	0.025	0.033
UGU_freq	0.024	0.024	0.030	0.025	0.020	0.026	0.024	0.020	0.026	0.020	0.023	0.025	0.022
ACA_freq	0.024	0.020	0.027	0.025	0.022	0.028	0.022	0.021	0.026	0.023	0.027	0.027	0.015

Figure 13: Feature Importance Analysis – Log2FC Model. Top 20 features across cell lines including heatmap, frequency bar, and cumulative contribution.

4.3.2 Cross-Model Consistency and Context Dependence

In both models, a key group of features—seed_G_percent, GC_content, post_seed_G_percent, and motif-like features like UGU_freq—consistently showed high importance and were commonly found across different cell lines (Figure 4.10/4.11, top and middle panels). This justifies their essential role in miRNA function. Every cell line has unique variability. The stacked bar graphs in the lower panels show that their relative relevance differs across cell lines even if they have same properties. Features like last_pos_nt_C and ACA_freq showed relevance in particular settings, therefore showing the presence of interactions unique to the regulatory environment.

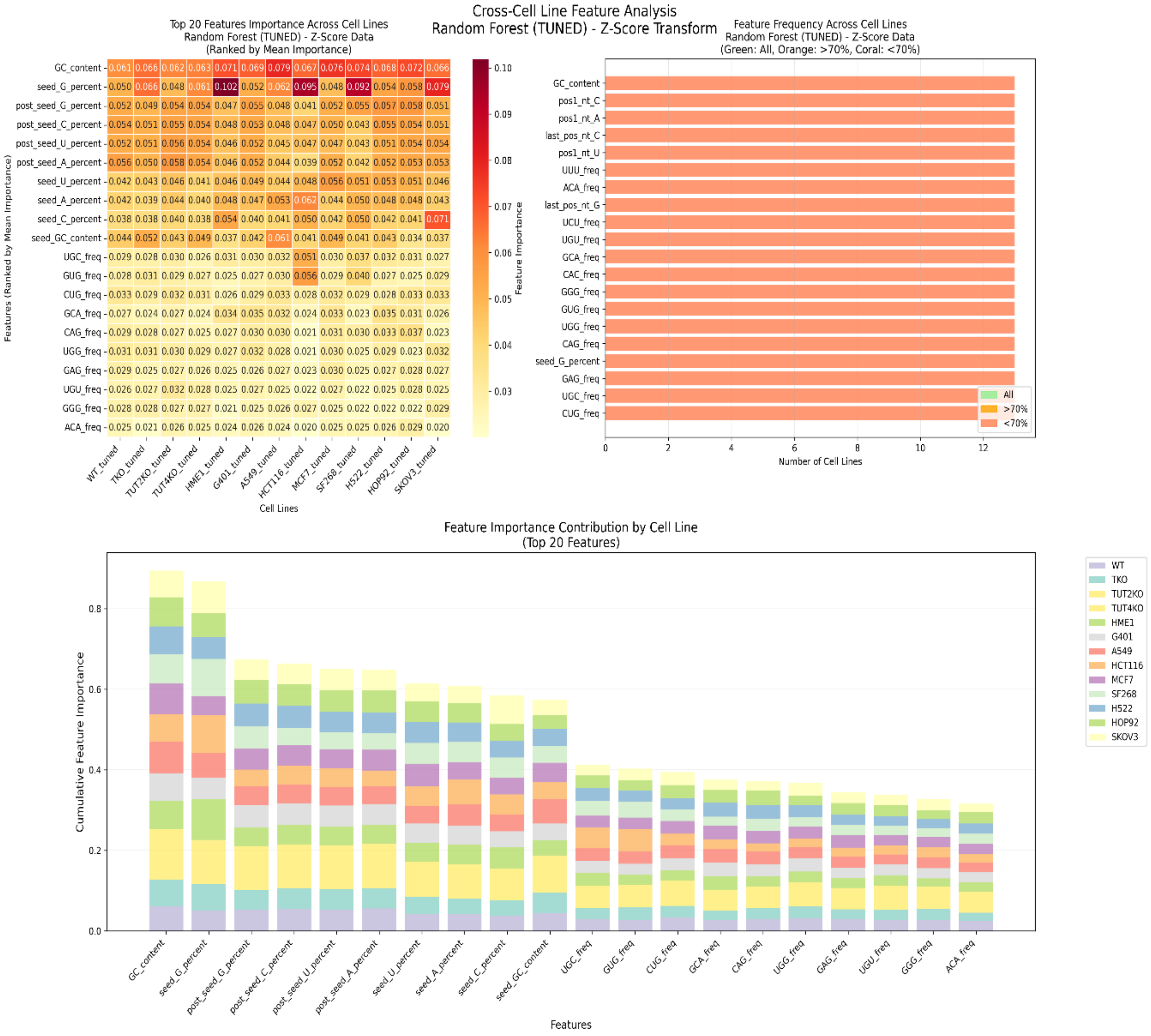


Figure 14: Feature Importance Analysis – Z-score Model. Identical layout for comparison against Log2FC model.

4.3.3 Model Specific Interpretations

The Log2FC model highlights the need of regulatory reactability. Focusing on elements of control that vary with time, the Log2FC-transformed model focused on elements related to stability and how expression levels vary. The frequency bar graph shows a strong degree of universality since over 10 of the top 20 traits were present in more than 70% of cell lines. The Z-Score methodology evaluates baseline expression and variability rather successfully. The Z-score model showed more range in ranks. While UUU_freq and ACA_freq gained context-specific importance, GC_content and seed_G_percent remained important. This data shows how responsively the model is to changes in baseline expression.

4.3.4 Consequences for Predictive Analysis and Mechanistic Modeling

This analysis supports two main conclusions:

- **Feature selection strategies** should focus on important biological measures, like G/C content and position-specific frequencies, instead of using general statistical filters.

- **Dual-model frameworks** that incorporate both Log2FC and Z-score preprocessing provide complementary insights. Log2FC effectively quantifies regulatory magnitude, whereas Z-score provides interpretability consistent with standardized metrics.

Chapter 5: Discussion, Limitation and Future Direction

5.1 Justification of Methods and Review of Key Studies

The advancements in ML-miRNA demonstrate a clear progression from early discoveries regarding how seed-based RNA regulates genes to the challenging task of predicting how different miRNA variations influence various cancer cell lines. The methodological framework was significantly influenced by two transformative studies that demonstrated the role of seed sequences in determining RNA regulatory specificity and toxicity. Anderson et al. (2008) showed that shRNAs with seeds sequences that are extremely abundant had a more toxic effect than those with a narrower set. Gao et al. (2018) advanced this concept by discovering that numerous tumor-suppressive microRNAs possess toxic 6mer seeds, which induce cancer cell death by targeting survival genes with C-rich 3'UTRs. They referred to this mechanism as DISE (Death Induced by Survival gene Elimination). Their detailed study of all 4,096 possible 6mer seeds showed that G-rich sequences, particularly those with guanines in the first two positions, were the most harmful. Importantly, tumor-suppressive miRNAs such as miR-34a-5p have evolved to incorporate these toxic seeds, whereas oncogenic miRNAs have largely avoided them. Gao et al. offered important information about how miRNAs can be toxic, but they mainly focused on the 6mer seed area and simply labeled sequences as toxic or non-toxic. We acknowledged that by expanding this framework to predict a range of variant effects, other than only focusing on seed region, across entire miRNA sequences and a subset of sequence permutations necessitated a more advanced computational methodology. Our findings encouraged the implementation of Random Forest as the primary algorithm for several reasons: its established capacity to capture non-linear relationships between sequence features and biological outcomes, its built-in feature importance calculations that support biological interpretability, and its resilience to overfitting in high-dimensional genomic data. We selected Random Forest instead of popular deep learning methods due to our dataset of 3,311 sequences, which, although considerable, was inadequate for training complex neural networks without the risk of overfitting. The complicated nature of deep learning would have made it hard for us to understand the biological processes we wanted to explain—being able to interpret the features was crucial for understanding how changes in sequences affect their function. We looked at three ways to change the data—raw, z-score normalized, and log2 fold change—because we knew that different methods might show us different biological signals. The effectiveness of the log2 fold change transformation ($R^2 = 0.4614$) backs up this approach, as fold changes truly show the scale of gene expression regulation and align with what biologists know about miRNA effects. The feature engineering process, which extracted 123 important features prior to correlation-based selection, was essential; if we had only used statistical filtering, we would have lost important but less common features. Using a biological feature protection hierarchy helped retain important features, such as terminal nucleotides and seed composition, even when there was not much data available for them in the dataset. The systematic engineering of positional variants (N1415, N1819, N2122, min1, plus1) rather than random mutagenesis enabled the isolation of position-specific effects while preserving statistical power. This approach revealed that, although seed toxicity is predominant, various regions within the miRNA play a role in overall function.

5.2 Biological Implications and Mechanistic Interpretation

Our findings fundamentally extend Gao et al.'s discovery of toxic seeds in tumor-suppressive miRNAs by showing that these ideas are widely applicable to different miRNA sequences and can be computed computationally with great accuracy. Our findings from machine learning closely match what Gao et al. discovered, which strongly supports each other: seed_G_percent was the most important feature in all cell lines (importance scores 0.052-0.109), confirming that G-rich seeds play a key role in toxicity. However, our detailed analysis showed that there is more to consider beyond the 6mer seed. The significant role of post-seed features (post_seed_G_percent, post_seed_U_percent) indicates that miRNA function involves sequences beyond positions 2-7, suggesting that while the seed starts the target recognition, the surrounding areas affect how well it binds and its regulatory strength. The identification of certain three-nucleotide patterns (UGU_freq, ACA_freq) as important features shows that more complex sequence arrangements can influence how miRNA is processed, loaded onto RISC, or how easily it can reach its target.

5.3 Possible Integration with Modern AI/ML Framework in Future

The ML-miRNA framework is in a good position to both guide and be supplemented by the recent developments in the AI-driven biological design. Transformer-based architectures like AlphaFold have brought a revolution in protein structure prediction, where the complicated sequence structure relationships are learned in a raw data manner (Jumper et al., 2021). Enformer is similarly able to predict highly detailed long DNA sequences corresponding to different cell types, suggesting that sequence-only models can learn highly detailed structural and functional grammars (Zhang et al., 2023) and TrASPr+BOS for tissue-specific splicing generation (Gupta et al., 2024).

These state-of-the-art models can be significantly initialized with our 31 features constructed by biological knowledge that work with much less data and are more interpretable. As an example, in a transformer model we may put attention heads to be more attentive to G-content and GC balance in the seed region, assisting the model to concentrate on the most significant aspects we observed. Alternatively, generative diffusion or variational-autoencoder based methods may synthesize novel miRNA variants optimized to achieve particular treatment objectives, such as maximizing damage to cancer cells whilst minimizing side effects in healthy tissues, by feedback of the predictions of ML-miRNA into a learning process.

In the future, single-cell transcriptomics, together with simple RNA language models, can be used to make predictions that take into account the precise context: we might generate families of models that make predictions about the effects of variants on different cell types in tumours, or we might use ML-miRNA with patient data to give rapid personalized predictions of the effects of variants.

5.4 Limitations

Although ML-miRNA's development is rigorous and offers new insights, it does not highlight the full potential of this study since there were still some analyses could not be done due to time limitations. Other than this, there are many limitations relating to this current study need recognition:

Insufficient Cohort for External Validation

Using the same collection of 3,311 miRNA variations investigated in 13 cancer cell lines, all model training and testing was conducted; cross-validation was used to assess performance. No alternative dataset, like an external miRNA screen or patient samples, was used to confirm whether the findings generalize more widely outside of this cohort. Future research should evaluate ML-miRNA on clinical data or held-out libraries to guarantee strong performance in practical environments.

Scope of Sequence-Only Feature Space

The predictive feature set comprises 31 sequence-based factors (such as nucleotide composition, seed/flanking k-mers and positional motifs) derived from an initial set of 123 by removing less relevant features. However, it excludes the other relevant parameters-such as RNA secondary structure, the accessibility of the target site, competing endogenous RNAs, as well as interactions with RNA-binding proteins-which have been shown to influence miRNA stability and targeting. Incorporation of computer-predicted structure (such as RNAfold) or CLIP-seq data may be able to incorporate these relevant terms and increase the current R^2 =0.46, resulting in more sensible results.

Chapter 6: Conclusion

In this thesis, we developed and tested ML-miRNA, a complete machine-learning system designed to predict how changes in miRNA sequences affect cancer cell lines. Using a carefully designed collection of 3,311 variants and a detailed set of biological features, we showed that how we prepare the data and which algorithm we use are very important for getting accurate predictions. Our improved Random Forest model, using transformed data, explained the most variation ($R^2 \approx 0.46$), setting a new standard for predicting the effects of miRNA variants. Importantly, the analysis of feature importance identified seed-region G-content and GC percentage as key predictors, offering a deeper understanding that supports and builds on important research about seed-related toxicity.

Besides looking at performance metrics, ML-miRNA helps researchers understand how predictions are made by linking them to specific sequence features, which aids in creating new hypotheses and designing experiments. The framework's cross-cell-line validation also highlights context-dependent regulatory patterns, indicating the importance of cell-type–aware modeling in oncology applications.

In the future, combining different types of data, like RNA secondary structure predictions, CLIP-seq binding profiles, and single-cell phenotypic readouts, will probably improve our understanding and reveal more detailed effects of variants. As deep-learning and generative AI models improve, the carefully selected features of ML-miRNA can be useful starting points for transformer-based models and guidance in designing sequences. Ultimately, ML-miRNA establishes a foundation for personalized evaluation and engineering of miRNA variants, expediting their transformation into diagnostic biomarkers and therapeutic agents.

References:

- Abdelfattah, N., Rajamanickam, S., Panneerdoss, S., Timilsina, S., Yadav, P., Onyeagucha, B. C., ... & Rao, M. K. (2014). MiR-584-5p potentiates vincristine and radiation response by inducing spindle defects and DNA damage in medulloblastoma. Nature Communications, 5(1), 4541.
- Agarwal, V., Bell, G. W., Nam, J. W., & Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. eLife, 4, e05005.
- Bartel, D. P. (2018). Metazoan microRNAs. Cell, 173(1), 20-51.
- Bautista-Sánchez, D., Arriaga-Canon, C., Pedroza-Torres, A., De La Rosa-Velázquez, I. A., González-Barrios, R., Contreras-Espinosa, L., ... & Herrera, L. A. (2020). The promising

role of miR-21 as a cancer biomarker and its importance in RNA-based therapeutics. Molecular Therapy-Nucleic Acids, 20, 409-420.

- Bayraktar, R., & Van Roosbroeck, K. (2018). miR-155 in cancer drug resistance and as target for miRNA-based therapeutics. Cancer and Metastasis Reviews, 37(1), 33-44.
- Beg, M. S., Brenner, A. J., Sachdev, J., Borad, M., Kang, Y. K., Stoudemire, J., ... & Hong, D. S. (2017). Phase I study of MRX34, a liposomal miR-34a mimic, administered twice weekly in patients with advanced solid tumors. Investigational New Drugs, 35(2), 180-188.
- Bhattacharya, A., & Cui, Y. (2016). SomamiR 2.0: a database of cancer somatic mutations altering microRNA–ceRNA interactions. Nucleic Acids Research, 44(D1), D1005-D1010.
- Briskin, D., Wang, P. Y., & Bartel, D. P. (2020). The biochemical basis for the cooperative action of microRNAs. Proceedings of the National Academy of Sciences, 117(30), 17764-17774.
- Cheloufi, S., Dos Santos, C. O., Chong, M. M., & Hannon, G. J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. Nature, 465(7298), 584-589.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
- Chirshev, E., Oberg, K. C., Ioffe, Y. J., & Unternaehrer, J. J. (2019). Let-7 as biomarker, prognostic indicator, and therapy for precision medicine in cancer. Clinical and Translational Medicine, 8(1), 24.
- Condrat, C. E., Thompson, D. C., Barbu, M. G., Bugnar, O. L., Boboc, A., Cretoiu, D., ... & Voinea, S. C. (2020). miRNAs as biomarkers in disease: latest findings regarding their role in diagnosis and prognosis. Cells, 9(2), 276.
- Duchaine, T. F., & Fabian, M. R. (2019). Mechanistic insights into microRNA-mediated gene silencing. Cold Spring Harbor Perspectives in Biology, 11(3), a032771.
- Eichhorn, S. W., Guo, H., McGeary, S. E., Rodriguez-Mias, R. A., Shin, C., Baek, D., ... & Bartel, D. P. (2014). mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. Molecular Cell, 56(1), 104-115.
- Friedländer, M. R., Lizano, E., Houben, A. J., Bezdan, D., Báñez-Coronel, M., Kudla, G., ... & Estivill, X. (2014). Evidence for the biogenesis of more than 1,000 novel human microRNAs. Genome Biology, 15(4), R57.
- Friedman, R. C., Farh, K. K. H., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Research, 19(1), 92-105.
- Fromm, B., Domanska, D., Høye, E., Ovchinnikov, V., Kang, W., Aparicio-Puerta, E., ... & Peterson, K. J. (2020). MirGeneDB 2.0: the metazoan microRNA complement. Nucleic Acids Research, 48(D1), D132-D141.
- Fuziwara, C. S., & Kimura, E. T. (2015). Insights into regulation of the miR-17-92 cluster of miRNAs in cancer. Frontiers in Medicine, 2, 64.

- Galka-Marciniak, P., Urbanek-Trzeciak, M. O., Nawrocka, P. M., Dutkiewicz, A., Giefing, M., Lewandowska, M. A., & Kozlowski, P. (2019). Somatic mutations in miRNA genes in lung cancer—potential functional consequences of non-coding sequence variants. Cancers, 11(6), 793.
- Gambari, R., Brognara, E., Spandidos, D. A., & Fabbri, E. (2019). Targeting oncomiRNAs and mimicking tumor suppressor miRNAs: New trends in the development of miRNA therapeutic strategies in oncology. International Journal of Oncology, 49(1), 5-32.
- Gao, Q. Q., Putzbach, W. E., Murmann, A. E., Chen, S., Sarshad, A. A., Peter, J. M., ... & Peter, M. E. (2018). 6mer seed toxicity in tumor suppressive microRNAs. Nature Communications, 9(1), 4504.
- Gebert, L. F., & MacRae, I. J. (2019). Regulation of microRNA function in animals. Nature Reviews Molecular Cell Biology, 20(1), 21-37.
- Grimson, A., Farh, K. K. H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., & Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Molecular Cell, 27(1), 91-105.
- Guo, L., & Chen, F. (2014). A challenge for miRNA: multiple isomiRs in miRNAomics. Gene, 544(1), 1-7.
- Hermeking, H. (2010). The miR-34 family in cancer and apoptosis. Cell Death & Differentiation, 17(2), 193-199.
- Hill, C. G., Jabbari, N., Matyunina, L. V., & McDonald, J. F. (2014). Functional and evolutionary significance of human microRNA seed region mutations. PLoS One, 9(12), e115241.
- Hill, M., & Tran, N. (2021). miRNA interplay: mechanisms and consequences in cancer. Disease Models & Mechanisms, 14(4), dmm047662.
- Hong, D. S., Kang, Y. K., Borad, M., Sachdev, J., Ejadi, S., Lim, H. Y., ... & Beg, M. S. (2020). Phase 1 study of MRX34, a liposomal miR-34a mimic, in patients with advanced solid tumours. British Journal of Cancer, 122(11), 1630-1637.
- Hrovatin, K., Kunej, T., & Glavač, D. (2018). The role of microRNA genetic variants in cancer. Human Genetics, 137(5), 335-342.
- Jonas, S., & Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. Nature Reviews Genetics, 16(7), 421-433.
- Karagkouni, D., Paraskevopoulou, M. D., Tastsoglou, S., Skoufos, G., Karavangeli, A., Pierros, V., ... & Hatzigeorgiou, A. G. (2020). DIANA-LncBase v3: indexing experimentally supported miRNA targets on non-coding transcripts. Nucleic Acids Research, 48(D1), D101-D110.
- Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Research, 47(D1), D155-D162.
- Kwon, S. C., Nguyen, T. A., Choi, Y. G., Jo, M. H., Hohng, S., Kim, V. N., & Woo, J. S. (2019). Structure of human DROSHA. Cell, 164(1-2), 81-90.

- Lee, B., Baek, J., Park, S., & Yoon, S. (2016). deepTarget: end-to-end learning framework for microRNA target prediction using deep recurrent neural networks. In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 434-442).
- 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.
- 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(1), 15-20.
- Liu, W., & Wang, X. (2019). Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. Genome Biology, 20(1), 18.
- Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. Algorithms for Molecular Biology, 6(1), 26.
- Ludwig, N., Leidinger, P., Becker, K., Backes, C., Fehlmann, T., Pallasch, C., ... & Meese, E. (2016). Distribution of miRNA expression across human tissues. Nucleic Acids Research, 44(8), 3865-3877.
- 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), eaav1741.
- Medley, J. C., Panzade, G., & Zinovyeva, A. Y. (2021). microRNA strand selection: Unwinding the rules. Wiley Interdisciplinary Reviews: RNA, 12(3), e1627.
- Michlewski, G., & Cáceres, J. F. (2019). Post-transcriptional control of miRNA biogenesis. RNA, 25(1), 1-16.
- Moszynska, A., Gebert, M., Collawn, J. F., & Bartoszewski, R. (2017). SNPs in microRNA target sites and their potential role in human disease. Open Biology, 7(4), 170019.
- Murmann, A. E., Gao, Q. Q., Putzbach, W. E., Patel, M., Bartom, E. T., Law, M., ... & Peter, M. E. (2018). Small interfering RNAs based on huntingtin trinucleotide repeats are highly toxic to cancer cells. EMBO Reports, 19(3), e45336.
- Nguyen, T. A., Jo, M. H., Choi, Y. G., Park, J., Kwon, S. C., Hohng, S., ... & Woo, J. S. (2015). Functional anatomy of the human microprocessor. Cell, 161(6), 1374-1387.
- O'Brien, J., Hayder, H., Zayed, Y., & Peng, C. (2018). Overview of microRNA biogenesis, mechanisms of actions, and circulation. Frontiers in Endocrinology, 9, 402.
- Okada, C., Yamashita, E., Lee, S. J., Shibata, S., Katahira, J., Nakagawa, A., ... & Tsukihara, T. (2009). A high-resolution structure of the pre-microRNA nuclear export machinery. Science, 326(5957), 1275-1279.
- Patel, M., & Peter, M. E. (2018). Identification of DISE-inducing shRNAs by monitoring cellular responses. Cell Cycle, 17(4), 506-514.

- Pinzón, N., Li, B., Martinez, L., Sergeeva, A., Presumey, J., Apparailly, F., & Seitz, H. (2017). microRNA target prediction programs predict many false positives. Genome Research, 27(2), 234-245.
- Pla, A., Zhong, X., & Rayner, S. (2018). miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts. PLoS Computational Biology, 14(7), e1006185.
- Plotnikova, O., Baranova, A., & Skoblov, M. (2019). Comprehensive analysis of human microRNA–mRNA interactome. Frontiers in Genetics, 10, 933.
- Putzbach, W., Gao, Q. Q., Patel, M., van Dongen, S., Haluck-Kangas, A., Sarshad, A. A., ... & Peter, M. E. (2017). Many si/shRNAs can kill cancer cells by targeting multiple survival genes through an off-target mechanism. eLife, 6, e29702.
- Putzbach, W., Gao, Q. Q., Patel, M., Haluck-Kangas, A., Murmann, A. E., & Peter, M. E. (2018). DISE: A seed-dependent RNAi off-target effect that kills cancer cells. Trends in Cancer, 4(1), 10-19.
- Quang, D., & Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Research, 44(11), e107.
- Quillet, A., Anouar, Y., Lecroq, T., & Dubessy, C. (2020). MiRiFIC: a method for prediction of plant miRNAs using integrated data from GEO and SRA databases. Scientific Reports, 10(1), 1-16.
- Riffo-Campos, Á. L., Riquelme, I., & Brebi-Mieville, P. (2016). Tools for sequence-based miRNA target prediction: what to choose?. International Journal of Molecular Sciences, 17(12), 1987.
- Rokavec, M., Li, H., Jiang, L., & Hermeking, H. (2014). The p53/miR-34 axis in development and disease. Journal of Molecular Cell Biology, 6(3), 214-230.
- 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(3), 203-222.
- Ryan, B. M., Robles, A. I., & Harris, C. C. (2010). Genetic variation in microRNA networks: the implications for cancer research. Nature Reviews Cancer, 10(6), 389-402.
- Salari, R., Wojtowicz, D., Zheng, J., Levens, D., Pilpel, Y., & Przytycka, T. M. (2013). Teasing apart translational and transcriptional components of stochastic variations in eukaryotic gene expression. PLoS Computational Biology, 9(8), e1003161.
- Sheu-Gruttadauria, J., Pawlica, P., Klum, S. M., Wang, S., Yario, T. A., Schirle Oakdale, N. T., ... & MacRae, I. J. (2019). Structural basis for target-directed microRNA degradation. Molecular Cell, 75(6), 1243-1255.
- Smolarz, B., Durczyński, A., Romanowicz, H., Szyłło, K., & Hogendorf, P. (2022). miRNAs in cancer (review of literature). International Journal of Molecular Sciences, 23(5), 2805.

- Stavast, C. J., & Erkeland, S. J. (2019). The non-canonical aspects of microRNAs: many roads to gene regulation. Cells, 8(11), 1465.
- Tan, G. C., Chan, E., Molnar, A., Sarkar, R., Alexieva, D., Isa, I. M., ... & Dibb, N. J. (2014). 5′ isomiR variation is of functional and evolutionary importance. Nucleic Acids Research, 42(14), 9424-9435.
- Title, A. C., Hong, S. J., Pires, N. D., Hasenöhrl, L., Godbersen, S., Stokar-Regenscheit, N., ... & Stoffel, M. (2018). Genetic dissection of the miR-200–Zeb1 axis reveals its importance in tumor differentiation and invasion. Nature Communications, 9(1), 4671.
- Tomasello, L., Distefano, R., Nigita, G., & Croce, C. M. (2021). The microRNA family gets wider: The isomiRs classification and role. Frontiers in Cell and Developmental Biology, 9, 668648.
- Treiber, T., Treiber, N., & Meister, G. (2019). Regulation of microRNA biogenesis and its crosstalk with other cellular pathways. Nature Reviews Molecular Cell Biology, 20(1), 5-20.
- Vickers, K. C., Roteta, L. A., Hucheson-Dilks, H., Han, L., & Guo, Y. (2015). Mining diverse small RNA species in the deep transcriptome. Trends in Biochemical Sciences, 40(1), 4-6.
- Wang, Y., Liang, H., & Jin, M. (2019). miRNA editing: New insights into the fast and dynamic control of miRNA expression. Signal Transduction and Targeted Therapy, 4(1), 1-3.
- Wen, M., Cong, P., Zhang, Z., Lu, H., & Li, T. (2019). DeepMirTar: a deep-learning approach for predicting human miRNA targets. Bioinformatics, 35(18), 3348-3354.
- Westholm, J. O., & Lai, E. C. (2011). Mirtrons: microRNA biogenesis via splicing. Biochimie, 93(11), 1897-1904.
- Wilson, R. C., Tambe, A., Kidwell, M. A., Noland, C. L., Schneider, C. P., & Doudna, J. A. (2015). Dicer-TRBP complex formation ensures accurate mammalian microRNA biogenesis. Molecular Cell, 57(3), 397-407.
- Waskom, M. L. (2021). Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.
- McKinney, W. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51-56).
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods, 17(3), 261–272.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95.

- Kim, H., Lee, Y. Y., & Kim, V. N. (2025). The biogenesis and regulation of animal microRNAs. Nature Reviews Molecular Cell Biology, 26(4), 276-296.
- Ghandi, M., Lee, D., Mohammad-Noori, M., & Beer, M. A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. PLoS computational biology, 10(7), e1003711.
- Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., & Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. Nature genetics, 47(8), 955-961.
- Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507-2517.
- Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics, 2015.
- Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., ... & Stamatoyannopoulos, J. A. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature, 489(7414), 83-90.
- Hall, M. A. (1999). Correlation-based feature selection for machine learning. PhD thesis, The University of Waikato.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.
- Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27-46.
- Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. Physical review E, 69(6), 066138.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal statistical Society: Series B (Methodological), 58(1), 267-288.
- Qi, Y. (2012). Random forest for bioinformatics. In Ensemble machine learning (pp. 307-323). Springer, Boston, MA.
- Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1), 389-422.
- Li, Y., Xu, J., Chen, L., & Guo, Y. (2019). Prediction of microRNA-disease associations using XGBoost with high accuracy. IEEE Access, 7, 136735–136744. https://doi.org/10.1109/ACCESS.2019.2942591
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
- Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. The Journal of Cell Biology, 177(1), 7-11.(5986), 1694-1698.

- Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. The Annals of Mathematical Statistics, 11(1), 86–92. https://doi.org/10.1214/aoms/1177731944
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1-30.
- Yoda, M., Cifuentes, D., Izumi, N., Sakaguchi, Y., Suzuki, T., Giraldez, A. J., & Tomari, Y. (2013). Poly (A)-specific ribonuclease mediates 3′-end trimming of Argonaute2-cleaved precursor microRNAs. *Cell reports*, *5*(3), 715-726.
- Shang, R., Lee, S., Senavirathne, G., & Lai, E. C. (2023). microRNAs in action: biogenesis, function and regulation. Nature Reviews Genetics, 24(12), 816-833.