 modeling of functional gene regulation through machine learning and deep learning methods

Yuhao Chen
Washington University – McKelvey School of Engineering

Follow this and additional works at: https://openscholarship.wustl.edu/eng_etds

Recommended Citation
https://openscholarship.wustl.edu/eng_etds/961

This Dissertation is brought to you for free and open access by the McKelvey School of Engineering at Washington University Open Scholarship. It has been accepted for inclusion in McKelvey School of Engineering Theses & Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
Modeling of Functional Gene Regulation
Through Machine Learning and Deep Learning Methods

by

Yuhao Chen

A dissertation presented to
the McKelvey School of Engineering
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2023
St. Louis, Missouri
Table of Contents

List of Figures ................................................................................................................................. v
List of Tables .................................................................................................................................. vii
Acknowledgments ........................................................................................................................... viii
Abstract ........................................................................................................................................ xi

Chapter 1: Introduction .................................................................................................................... 1
  1.1 MicroRNAs: History and Functions in Biological Processes .................................................. 1
    1.1.1 MicroRNA Discovery and Function .................................................................................. 1
    1.1.2 MicroRNA-Gene Regulatory Analysis .............................................................................. 4
    1.1.3 Prediction of MicroRNA Targets ...................................................................................... 6
  1.2 CRISPR/Cas Genome Editing System ...................................................................................... 8
    1.2.1 History of CRISPR Technology Development ................................................................. 8
    1.2.2 Target Specificity and Cleavage Efficiency ..................................................................... 11
    1.2.3 Computational Research for Guide RNA Design of CRISPR/Cas9 system ..................... 14
  1.3 Machine Learning and Deep Learning Applications ............................................................... 17
    1.3.1 Backgrounds of Machine Learning .................................................................................. 17
    1.3.2 Machine Learning and Deep Learning Application for CRISPR Cleavage Efficiency
         Prediction .............................................................................................................................. 19
    1.3.3 Machine Learning and Deep Learning Application for MicroRNA Targets Prediction ... 24
  1.4 Project Aims ............................................................................................................................. 28
  1.5 References ............................................................................................................................... 30

Chapter 2: Generalizable sgRNA Design for Improved CRISPR/Cas9 Editing Efficiency ........... 42
  2.1 Abstract ..................................................................................................................................... 42
  2.2 Introduction ............................................................................................................................... 43
  2.3 Materials and Methods ............................................................................................................. 45
    2.3.1 Cloning of sgRNA Plasmid Library ................................................................................... 45
    2.3.2 Lentivirus Preparation ....................................................................................................... 47
    2.3.3 Plasmid Library Delivery into HeLa/Cas9 Cells ................................................................. 48
    2.3.4 Sequencing Library Construction ..................................................................................... 48
    2.3.5 Quantification of sgRNA Efficiency .................................................................................. 48
Chapter 2: Computational Tools for CRISPR/Cas9 gRNA Selection

2.3.6 Computational Tools ............................................................ 49
2.3.7 Independent Testing Datasets .................................................. 49
2.3.8 Model Performance Evaluation ................................................ 50
2.3.9 Availability of Data ............................................................... 51

2.4 Results .................................................................................... 51
2.4.1 An sgRNA Library for Quantifying CRISPR/Cas9 Editing Efficiency ........................................ 51
2.4.2 sgRNA/Target Features .......................................................... 53
2.4.3 Assessment of Modeling Methods .............................................. 56
2.4.4 Validation of sgDesigner ........................................................ 59
2.4.5 Genome-wide sgRNA Design Database ...................................... 60

Chapter 3: Evaluation of Efficiency Prediction Algorithms and Development of Ensemble Model for CRISPR/Cas9 gRNA Selection

3.1 Abstract .................................................................................. 71
3.2 Introduction ............................................................................ 72
3.3 Materials and Methods .............................................................. 73
3.3.1 Computational Algorithms for Predicting CRISPR/Cas9 Efficacy ................................................. 73
3.3.2 CRISPR gRNA Datasets for Evaluation of Editing Efficiency ......................................................... 76
3.3.3 Assessment of Model Performance ........................................... 79

3.4 Results .................................................................................... 79
3.4.1 Evaluating the Performance of Public Algorithms for Prediction of CRISPR/Cas9 Editing Efficiency ................................................. 79
3.4.2 Developing an Ensemble Model for Improved Prediction of CRISPR/Cas9 Efficiency ... 84
3.4.3 Validation of the Ensemble Models with Independent Datasets ......................................................... 89
3.4.4 Re-training the Ensemble Model for the T7 Promoter System .............................................................. 93
3.4.5 Online Web Server for gRNA Design ........................................ 94

3.5 Discussion ............................................................................... 95
3.6 References ............................................................................. 97
3.7 Supplementary Figures ................................................................ 101

Chapter 4: Optimized Prediction of Functional MicroRNA Targets by Deep Learning

4.1 Abstract ................................................................................. 104
4.2 Introduction ........................................................................................................... 105
4.3 Materials and Methods ......................................................................................... 107
  4.3.1 Public CLIP-seq Data Retrieval ......................................................................... 107
  4.3.2 RNA-seq to Identify Transcripts Downregulated by MicroRNA Overexpression ........ 108
  4.3.3 Deep Learning Model Development .................................................................. 110
  4.3.4 Model Training and Parameter Tuning ............................................................. 114
4.4 Results .................................................................................................................... 115
  4.4.1 miRDB Web Portal Improvement ...................................................................... 115
  4.4.2 Combining Target Downregulation and CLIP Binding Data to Identify Common Targeting Features 121
  4.4.3 Performance Comparison of Various Modeling Methods .................................. 125
  4.4.4 Algorithm Evaluation with Independent Experimental Data ............................. 128
4.5 Discussion .............................................................................................................. 133
4.6 References ............................................................................................................ 134
Chapter 5: Conclusions .............................................................................................. 139
List of Figures

Figure 1.1: miRNA biogenesis and function.................................................................3
Figure 1.2: MicroRNAs can affect gene expression directly or indirectly......................5
Figure 1.3: CRISPR/Cas9 genome editing system.......................................................9
Figure 1.4: Target specificity and cleavage efficiency of CRISPR/Cas9 system.............12
Figure 1.5: The architecture of the generally used convolutional neural network..........22
Figure 1.6: The architecture of basic recurrent neural network (RNN) .......................23
Figure 1.7: Basic architecture of a single layer autoencoder .....................................27
Figure 2.1: Determining sgRNA efficiencies using an in silico designed sgRNA library....52
Figure 2.2: Feature analysis of efficient sgRNAs.........................................................54
Figure 2.3: Stacking model framework........................................................................57
Figure 2.4: Comparison of different computational models.......................................58
Figure 2.5: Comparison of sgDesigner with public sgRNA design tools......................60
Supplementary Figure 2.1: Additional flanking-region features did not improve predictions....67
Supplementary Figure 2.2: ROC comparison of sgRNA design tools in five additional datasets...............................................................68
Supplementary Figure 2.3: PR comparison of sgRNA design tools in five additional datasets....69
Supplementary Figure 2.4: Spearman correlation comparison of sgRNA design tools in five additional datasets...............................................................70
Figure 3.1: Performance comparison of 15 public algorithms with Spearman correlation analysis.........................................................................................81
Figure 3.2: Performance comparison of 15 public algorithms with ROC analysis...........83
Figure 3.3: The stacking ensemble modeling strategy................................................85
Figure 3.4: Contribution of individual algorithms and optimal order of the ensemble model......87
Figure 3.5: Validation of the ensemble models by Spearman correlation analysis.......................90
Figure 3.6: Validation of the ensemble models by ROC analysis..............................................91
Supplementary Figure 3.1: Spearman correlation comparison of ensemble models assembling all
individual algorithms using different meta models on 11 independent validation datasets.......101
Supplementary Figure 3.2: The overall Spearman correlation performance.............................102
Supplementary Figure 3.3: Performance comparison of T7 promoter-based algorithms...........103
Figure 4.1: LSTM cell architecture.................................................................112
Figure 4.2: miRDB Target Search page................................................................................115
Figure 4.3: miRDB target expression analysis.................................................................119
Figure 4.4: miRDB target ontology analysis...............................................................120
Figure 4.5: Performance of different algorithms for miRNA target prediction....................126
Figure 4.6: The workflow of LSTM+Biofeat model.......................................................128
Figure 4.7: ROC analyses of algorithm performance on validation datasets.....................130
List of Tables

Table 2.1. List of Primers Used in the Study .................................................................46
Table 2.2. Summary of validation datasets ........................................................................50
Table 3.1. The collection of scoring algorithms for predicting gRNA cleavage efficiency ......74
Table 3.2. The collection of experimental datasets for evaluating gRNA efficacy .................77
Table 4.1. Twenty-five miRNAs analyzed in the RNA-seq experiments ..............................108
Table 4.2. Summary of biological features of miRNA targeting ........................................122
Table 4.3. Number of validated targets for five miRNA target prediction algorithms ........132
Acknowledgments

First and foremost, I would like to express my gratitude to Dr. Xiaowei Wang, my mentor, for leading me through this extraordinary journey. His exemplary guidance, teaching, and role modeling have been a source of inspiration as I strive to become a researcher of similar caliber.

I want to give special thanks to my thesis committee members, Drs. Lan Yang, Ulugbek Kamilov, Hong Chen, and Tiezhi Zhang, for guiding me through this research process and offering invaluable insights and direction.

I also want to thank everyone in the Wang lab, both current and former members. Special thanks go to Weijun, Ping, Arlise, Tony, Xinyi, Gongyu, Minsu, and Yuanxiang in no specific order, for contributing to a stimulating and pleasant working environment in the lab.

I would like to extend my thanks to the Department of Electrical and Systems Engineering for admitting me into both the master's and Ph.D. programs. These opportunities have significantly altered the course of my life, and I am truly grateful for the department's support, which has played a crucial role in shaping the person I am today.

Many big hugs to my parents Hong Ye and Shunzhi Chen. Thank you so much for providing a nurturing environment, endorsing my decisions, and guiding me when I was lost. Your influence has motivated me to not only excel academically but also to grow as an individual.

Special thanks to my mother-in-law Dongmei Sun, for her unwavering support during challenging times. Your encouragement and insightful suggestions have been invaluable to me.
and my family. I cannot envision my life without your kind guidance. Your life wisdom serves as a great source of inspiration, and I am committed to continually improving myself.

Above all, I want to convey my deep love and gratitude to my wife, Yunfei Ta, for sharing this incredible yet challenging journey with me. The sacrifices you've made for my career and our family have been invaluable. It’s the strength drawn from you and our children that propels me to work tirelessly toward my goals. Words cannot adequately express how your support and inspiration have propelled me to this point. I will forever cherish your love and kindness, striving to be a better husband and father each day. As you embark on your Ph.D. journey, I extend my heartfelt wishes for your success and eagerly anticipate witnessing your remarkable achievements.

Yuhao Chen

Washington University in St. Louis

December 2023
Dedicated to my wife, Yunfei, and my children, Nathan and Noah.
ABSTRACT OF THE DISSERTATION

Modeling of Functional Gene Regulation
Through Machine Learning and Deep Learning Methods

by

Yuhao Chen

Doctor of Philosophy in Electrical Engineering

Washington University in St. Louis, 2023

Professor Lan Yang, Chair
Professor Xiaowei Wang, Co-Chair

Genes are specific sequences of nucleotides within the genome that undergo transcription to generate products that can diffuse throughout the cell. These gene products can take the form of either proteins or RNAs and play crucial roles in virtually all cellular functions, encompassing structural support, enzymatic activities, and signaling functions. An essential focus of contemporary biological research revolves around unraveling the functions of genes and their resulting products across different organisms. One primary means to conduct this research is through the regulation of gene expression.

MicroRNAs (miRNAs), short single-stranded non-coding RNAs, play vital roles in gene expression regulation. Both computational and experimental analyses indicate that most human protein-coding genes are regulated by one or more miRNAs. For functional miRNA analysis, one critical first step is to identify genes targeted by the miRNA. We have identified key characteristics of functional miRNA-target pairs and developed a machine learning prediction algorithm for genome-wide miRNA target prediction accordingly. This tool has gained significant popularity in the field, with hundreds of daily usages.
Additionally, the CRISPR (clustered regularly interspaced short palindromic repeats) Cas system is a rapidly advancing technology associated with gene regulation. It enables gene expression activation, repression, or knockout in mammalian systems. Previous studies indicate the efficiency of CRISPR/Cas system is dependent on one of its core components, guide RNA (gRNA). Leveraging innovative experimental designs and an ensemble learning framework, we have created a state-of-the-art gRNA design algorithm, offering superior performance for predicting CRISPR efficiency.

Importantly, these algorithms have been implemented as web-accessible applications (miRDB.org and CRISPRDB.org) to enable flexible bioinformatics analyses of miRNA target regulation and CRISPR design by the research community.
Chapter 1: Introduction

Gene regulatory networks play a vital role in various life processes, such as cell differentiation, metabolism, the cell cycle, and signal transduction. Understanding how different genes function provides insights into the mechanisms of diseases resulting from disruptions in these cellular processes. One direct approach to investigate a gene's role in cellular processes is by regulating its expression, achieved through technologies like RNA interference (RNAi) and CRISPR. Predicting the behavior of critical components in RNAi and CRISPR research accelerates biotechnological projects, offering faster and cost-effective alternatives to lab experiments. In this context, we will outline the application of bioinformatics pipelines and machine learning techniques in advancing research related to microRNA (miRNA) and CRISPR.

1.1 MicroRNAs: History and Functions in Biological Processes

1.1.1 MicroRNA Discovery and Function

Noncoding or non-messenger RNAs are molecules that serve various roles, including structural, enzymatic, and regulatory functions. MicroRNAs (miRNAs) are noncoding RNAs with regulatory activity, which typically consist of approximately 22 nucleotides and have been found in every studied metazoan organism (1). MiRNAs function as part of the post-transcriptional regulatory mechanism known as RNA interference (RNAi). The first miRNA, lin-4 in Caenorhabditis elegans, was discovered in 1993 through loss-of-function mutations that disrupted developmental timing in worm larvae (2). However, it was in the early 2000s that miRNAs were recognized as a distinct class of biological regulators. In 2000, a second small
RNA, *let-7*, was characterized in *C. elegans*. It was later identified as conserved in various species, suggesting its role in regulating developmental timing across different animals, including humans (3, 4). Over the years, miRNAs have been discovered in 223 species, with humans alone possessing 2588 high-confidence mature miRNA sequences (5).

Ha and Kim (6) have elucidated the role of miRNAs in gene expression regulation. Initially, the miRNA gene is transcribed into pri-miRNAs by RNA-polymerase II. Next, the pri-miRNA is processed within the nucleus by the Drosa-DGCR8 complex into the hairpin-shaped stem-loop pre-miRNA. The pre-miRNA is then exported to the cytoplasm through a transport complex comprised of exportin 5 and RAN-GTP. Subsequently, Dicer cleaves the pre-miRNA near the terminal loop, yielding a small RNA duplex. Meanwhile, one of the strands combines with a group of proteins, specifically the Argonaute (Ago) protein within the silencing complex, forming a miRNA-protein complex termed RNA-induced silencing complex (RISC). The other strand, the passenger strand, is typically discarded (7). The miRNA within the RISC complex then pairs with messenger RNA (mRNA) to facilitate post-transcriptional repression. In instances of extensive pairing complementarity, miRNAs can catalyze mRNA cleavage via Ago protein (8,9). More frequently, however, miRNAs induce translational repression, mRNA destabilization, or a combination of both (10,11) (Figure 1.1).
Figure 1.1: miRNA biogenesis and function. The miRNA gene is transcribed into the pri-miRNA, cleaved by the Drosha/DGCR8 complex to form the pre-miRNA hairpin structure. The pre-miRNA is transported from the nucleus to the cytoplasm by Exportin 5. The hairpin is cleaved by Dicer before the paired miRNA strands are loaded into the Argonaute proteins. The passenger strand of miRNA-miRNA duplex is then removed, resulting in the mature miRNA silencing complex, which can bind to the 3'UTR of target sequences through perfect complementarity or partial complementarity and induce mRNA cleavage or translational repression. The figure is from Salim, Hiba, et al. (128).
1.1.2 MicroRNA-Gene Regulatory Analysis

Since miRNAs function mainly by downregulating the expression of their target genes, the primary objective in miRNA-gene regulatory analysis is to elucidate the connection between miRNAs and their target genes, while quantifying the impact of miRNAs on the expression levels of these genes. Previous computational studies have employed diverse high-throughput profiling datasets for conducting miRNA-gene regulatory analysis, including microarray profiling data (12, 13) and, more recently, crosslinking and immunoprecipitation (CLIP) sequencing data (14, 15, 16).

CLIP sequencing data effectively identifies transcript targets associated with the functional miRNA-RNA-induced silencing complex (RISC) complex (17, 18, 19). In a typical CLIP experiment, short transcript sequences bound to the Ago protein are identified by crosslinking the target RNA to the RISC protein complex, followed by immunoprecipitation and high-throughput RNA-seq analysis (18). Recent advancements in CLIP studies have allowed for unambiguously identifying paired miRNA-target transcripts residing in the same RISC complex through direct ligation of the miRNA and its corresponding target transcript (20, 21). However, it's essential to note that while CLIP data has been widely used in miRNA-gene regulatory analysis, a significant concern is that miRNA target binding, as revealed by CLIP, may not necessarily result in functional target suppression (13, 22). Consequently, many miRNA-gene relationships identified through CLIP data may not be functionally relevant in gene expression regulation.

In addition to utilizing CLIP data, another widely adopted strategy for miRNA-gene regulatory analysis involves identifying downregulated transcripts resulting from miRNA overexpression (11, 23, 24). This approach is favored because target genes identified by this
Figure 1.2: MicroRNAs can affect gene expression directly or indirectly. The regulatory impact of a miRNA extends beyond direct RISC-dependent targeting. When a miRNA is over-expressed, its direct targets typically experience down-regulation. Suppose these direct targets act as repressors of downstream genes. In that case, miRNA regulation can lead to the de-repression of these genes, resulting in elevated levels (referred to as Up-regulated differentially expressed genes or DEGs). Conversely, genes downstream of activators and transcription factors will exhibit decreased expression following the changes in the direct targets (referred to as Down-regulated DEGs). Additionally, proteins that physically interact with the direct targets to form functional complexes may also be influenced by this regulatory process. The figure is from Yang et al. (129).

method are more likely to have functional significance, indicated by significant expression downregulation. However, challenges arise when employing miRNA overexpression strategies, due to the difficulty distinguishing direct miRNA target genes from indirect ones, primarily because of the involvement of transcription factors (TFs) (Figure 1.2). In molecular biology, TFs are proteins that govern the transcription rate of genetic information from DNA to messenger RNA by binding to specific DNA sequences (25, 26). When a particular gene encodes a TF, the regulatory effects induced by the direct binding of miRNAs can result in differential TF expression levels, subsequently affecting the transcription rates of multiple target genes controlled by that TF. In this scenario, miRNAs exert regulatory effects on several genes they
did not directly bind to, referred to as indirect target genes. Another concern is the physiological relevance of target genes identified under miRNA overexpression in cell culture. Furthermore, the analysis of miRNA overexpression is significantly hampered by the need for high-quality transcriptome-wide profiling data. Specifically, existing datasets are often of limited scope, focusing on only a small number of miRNAs in each study, rendering them unsuitable for comprehensive miRNA-gene regulatory analysis. While it is feasible to combine data from multiple small-scale studies, substantial heterogeneity among various experiments poses a significant challenge to accurate analysis.

### 1.1.3 Prediction of MicroRNA Targets

A single miRNA can potentially target hundreds of different mRNA molecules, while a particular mRNA molecule can be regulated by multiple miRNAs (27, 28). Research conducted in the 2000s revealed that mammalian miRNAs exhibit a remarkable capacity for targeting numerous distinct genes. For instance, an analysis of highly conserved miRNAs in vertebrates revealed that, on average, each of these miRNAs targets around 400 conserved genes (27). Similarly, experimental studies have demonstrated that a single miRNA species can negatively impact the stability of hundreds of unique messenger RNAs (11). Given the sheer number of miRNAs, relying solely on experimental methods for identifying miRNA targets becomes exceedingly challenging. Consequently, computational algorithms have proven to be invaluable tools for miRNA target prediction.

Animal miRNAs can recognize their target mRNAs using as few as 6–8 nucleotides at the 5’ end, known as the seed region. This characteristic has led to the development of research tools like TargetScan, which first emerged in 2005 (29). Initially created by the Bartel group, TargetScan relied on seed matching as the primary factor in defining miRNA targets and Gibbs
free energy of the binding site to score potential miRNA-target interactions. Subsequent score ranking defined a cutoff that maximized the signal-to-noise ratio (30). TargetScan7, introduced in 2015, identifies miRNA targets by seeking 8-mer, 7-mer, and 6-mer sites that match the seed region. Users can choose to consider only conserved sites. The algorithm evaluates additional factors like seed-pairing stability, 3' compensatory pairing, and target site abundance. The putative targets are then ranked based on the predicted efficacy of targeting, calculated using cumulative scores considering all the features of the sites (13). The most recent version, TargetScan8 (2021), takes a novel approach by measuring binding affinities for six miRNAs and synthetic targets. It constructs a biochemical model for miRNA-mediated repression and extends it to all miRNAs using a convolutional neural network. This advancement offers enhanced insights into miRNA targeting and improves predictions of intracellular miRNA repression efficacy compared to previous algorithms (31).

In addition to TargetScan, a well-known miRNA prediction algorithm, several other computational algorithms have been developed and widely adopted in recent years within the research community. Notable algorithms such as miRanda, DIANA-microT, PITA, and MirTarget employ various bioinformatics techniques to assess essential and supplementary factors in microRNA targeting (32-35). MiRanda, for instance, evaluates complementarity between miRNAs and 3' untranslated regions (3’-UTRs), the thermodynamic stability of the duplex structure, the evolutionary conservation of the entire binding site, and its position within 3’-UTR as a final filtering step (32). On the other hand, DIANA-microT uses stepwise regression based on the Akaike information criterion (AIC) to identify features used in non-canonical binding. In its recent update, DIANA-microT 2023 also predicts interactions between miRNAs encoded by 20 viruses and host transcripts from human, mouse, and chicken species (36). PITA,
taking a different approach, primarily focuses on the accessibility of the target site. It assesses the free energy associated with miRNA-mRNA pairing and the energy required to make the target accessible to the miRNA. Then, PITA would calculate the difference between these two parameters. It also considers "flank sites" surrounding the seed region, contributing to site accessibility (34). MirTarget, which shares its results through miRDB, utilizes public miRNA target binding data and original miRNA overexpression data to systematically identify miRNA targeting features characteristic of both miRNA binding and target downregulation. By integrating these common features into a Support Vector Machine (SVM) model, MirTarget was developed and validated as an enhanced computational model for genome-wide miRNA target prediction (35).

1.2 CRISPR/Cas Genome Editing System

1.2.1 History of CRISPR Technology Development

The discovery of clustered DNA repeats occurred in 1987 when a team of Japanese researchers accidentally cloned a portion of a sequence from the clustered regularly interspaced short palindromic repeats (CRISPR) found in the genome of Escherichia coli (37, 38). Subsequently, an increasing number of scientists initiated investigations into enigmatic DNA sequence arrays known as CRISPR, which are frequently observed in microbial genomes alongside genes encoding CRISPR-associated (Cas) proteins. The presence of short DNA sequences within CRISPRs, matching those in viruses, hinted at the role of these systems in acting as adaptive immunity pathways, preventing viral infections (39). Furthermore, various studies demonstrated that CRISPR systems employ RNA molecules transcribed from these sequence arrays to guide Cas proteins in cleaving and thereby eliminating viral DNA or RNA
Notably, researchers also illustrated how CRISPR's RNA-programmed cleaving mechanism could be leveraged to modify DNA sequences in any cell with unparalleled ease, aligning with the utility of the CRISPR system for genome editing.

**Figure 1.3**: CRISPR/Cas9 genome editing system. The CRISPR/Cas9 system comprises a single Cas9 enzyme and a guide RNA (gRNA) that forms a complex. The gRNA consists of trans-activating crispr RNA (tracer RNA) and crispr RNA (crRNA). The protospacer adjacent motif (PAM) is a small sequence of 2–6 bp (NGG) required for Cas9 nuclease activity. This PAM sequence is located three nucleotides downstream from the cleavage site, and the seed sequence is positioned five nucleotides upstream of the PAM. The figure is from Kashtwari et al. (130).

CRISPR/Cas systems can be categorized into two classes: Class 1 systems employ a complex comprising multiple Cas protein subunits to degrade foreign nucleic acids, whereas Class 2 systems use a single large Cas protein for the same purpose. Among these, the CRISPR/CRISPR-associated protein 9 (Cas9) system, which belongs to the Class 2 systems, stands out as the most commonly utilized genome editing system due to its simplicity, high
efficiency, and cost-effectiveness (45). The fundamental capability of CRISPR as a genome editing technology arises from its chemical mechanism of DNA cleavage guided by RNA-determined sequence recognition. Because Cas proteins rely on RNA-DNA base pairings for DNA identification, a single protein like Cas9 can target a diverse array of DNA sequences by merely swapping out guide RNAs. In genome editing experiments, the RNA responsible for sequence recognition is engineered into a single guide RNA (sgRNA or gRNA) with a customized guide sequence matching the desired target (42). Functioning in a manner akin to the natural system, the sgRNA directs Cas9 to cleave the DNA at a specific genomic location based on sequence complementarity, resulting in a precise double-stranded DNA break. This break occurs exactly three nucleotides upstream of an NGG protospacer adjacent motif (PAM) sequence (46) (Figure 1.3). In mammalian cells, the DNA repair process often introduces insertions or deletions (indels), leading to frameshift mutations and the consequent knockout of the functional gene. Building upon this editing framework, advanced strategies have been devised, including paired nicking to enhance specificity (47) and inserting nucleotide sequences during double-strand break repair to generate knock-ins (48). Additionally, other Cas proteins, including RNA-targeting proteins, have been explored as tools for genome modification, facilitated by extensive discovery efforts and biochemical and structural characterization (49-52). Some of these enzymes have also found applications in developing imaging techniques (53, 54) and diagnostic approaches (55, 56).

CRISPR technology applications have played a pivotal role in advancing clinical trials for various medical conditions, including sickle cell disease, beta-thalassemia, transthyretin (TTR) amyloidosis, congenital eye disease, and upcoming trials for both rare conditions such as progeria, severe combined immunodeficiency, and familial hypercholesterolemia, as well as
common ailments like cancer and HIV infection. Additionally, CRISPR-associated nucleases have proven highly valuable as molecular testing tools due to their precise targeting capabilities amidst a complex backdrop of non-target sequences (57). For example, they have been employed in the direct diagnosis of nucleic acids, achieving sensitivity down to the level of single molecules (56, 58). Furthermore, through integration with additional enzymatic processes, CRISPR-based diagnostics can expand their scope to detect molecules beyond nucleic acids. For instance, CRISPR-Cas platforms are also being explored for detection (59-63) and inactivation of SARS-CoV-2 (64), the virus responsible for COVID-19. This dynamic field of CRISPR research is rapidly evolving, offering an increasingly comprehensive toolkit for genetic manipulation.

1.2.2 Target Specificity and Cleavage Efficiency

The advancement of CRISPR technology has brought forth a set of crucial challenges necessitating creative solutions. Among these challenges, two stand out as primary concerns: target specificity (accuracy) and cleavage efficiency (precision) (Figure 1.4). In the context of CRISPR genome editing experiments, the issue of off-target Cas9 activity arises when genomic sequences resembling the guide RNA (gRNA) are encountered elsewhere, potentially resulting in unintended knockout effects (65). Such off-target editing can lead to permanent sequence alterations, causing genome instability or disruption of normal gene function. Besides target specificity, cleavage efficiency is equally paramount and warrants thorough research. In traditional CRISPR-Cas9 editing within eukaryotic cells, complete control over the editing outcome following the introduction of a double-stranded DNA break (DSB) remains elusive for scientists. Over the past decade, advancements have been made in understanding the cleavage efficiency of single-guide RNAs (sgRNAs), with several studies identifying various sequence
and structural features influencing on-target cleavage efficiency. Nevertheless, the scientific community still lacks a universally applicable computational tool to enhance cleavage efficiency across diverse biological conditions.

**Figure 1.4**: Target specificity and cleavage efficiency of CRISPR/Cas9 system. Target specificity (accuracy) involves specificity for the target location, which is challenged by off-target binding by the Cas9 RNP. Cleavage efficiency (precision) involves producing the correct intended edit with no unintended edits and is challenged by undesired indels and bystander edits. The figure is from Wang et al. (131).
To mitigate off-target effects, recent advances in experimental techniques have focused on enhancing the precision of the Cas9 nuclease activity. This focus has led to the development of high-fidelity Cas9 variants like SpCas9-HF1 (66), evoCas9 (67), HiFiCas9 (68), and Cas9_R63A/Q768A (69), resulting in significantly improved specificity for CRISPR/Cas9 targeting (70). Notably, the sgRNAs used in clinical applications by CRISPR Therapeutics/Vertex and Intellia have shown no detectable off-target sites when subjected to rigorous US Food and Drug Administration (FDA)–grade assays (71, 72). Furthermore, recent research has introduced bioinformatics tools for designing sgRNA sequences that minimize off-target effects (65, 73, 74). However, it's worth noting that off-target editing inaccuracies can also arise due to the Cas9-independent behavior of effector domains like deaminases, reverse transcriptases, and transcriptional regulators. This phenomenon is particularly evident in the analysis of base editing outcomes (75, 76). To address this issue, ongoing efforts involve using high-fidelity Cas variants and strategic modifications to the deaminase domain. These modifications aim to reduce the binding of nucleic acids without requiring Cas assistance (75-79). Encouragingly, early-stage clinical trials of base editing provide optimism for the continued advancement of this research direction (80, 81).

While various experimental innovations have made notable strides in minimizing off-target effects, the CRISPR/Cas9 system still faces potential variations in cleavage efficiency. In response to this challenge, more studies have sought to address the issue of sgRNA efficiency prediction, identifying correlations between sgRNA and target characteristics and Cas9 cleavage efficiency (82-87). Given the complexity of these features, machine learning techniques are frequently employed for data modeling. Constructing such computational models necessitates the experimental testing of a substantial number of sgRNAs to establish a robust training dataset for
efficiency prediction. Traditionally, existing studies have often adopted biological enrichment strategies, wherein gene editing events impact cell survival or other observable biological phenotypes, to generate training data. While these strategies alleviate the labor intensiveness associated with experimentation, indirect biological readouts can introduce artifacts into the training data. This is because equally efficient Cas9 cleavage sites may not necessarily lead to equivalent phenotypic changes or survival pressures. Moreover, existing experimental studies have often concentrated on a limited subset of genes or a single cell line, which restricts the applicability of the training data for broader predictions. To address these limitations, several representative studies have employed the plasmid target library strategy for the experimental quantification of sgRNA efficiency within the CRISPR/Cas9 system. Utilizing in silico-designed target sites obtained from an extensive plasmid library, the resulting large-scale training dataset minimizes potential biases associated with specific experimental setups and can be applied more broadly across various studies (88-91).

1.2.3 Computational Research for Guide RNA Design of CRISPR/Cas9 system

Research has demonstrated that the genetic sequence of the guide RNA (gRNA) plays a crucial role in both targeting precision and cleavage efficiency (65, 92). While these studies have offered qualitative insights into the factors affecting specificity and efficiency, they still need to provide a comprehensive understanding. Identifying generalizable patterns in this context is a formidable task, due to the vast number of potential imperfect interactions between gRNAs and DNA sequences. It requires substantial datasets to uncover predictive sequence features. Relying solely on experimental methods would entail significant time and resource expenses. Consequently, computational approaches, particularly leveraging machine learning and deep
learning technologies, have emerged as valuable tools for extracting insights from extensive biological datasets related to gRNAs and their target sequences. These computational methods are capable of predicting CRISPR activities, under diverse experimental conditions. It's important to note that target specificity and cleavage efficiency represent distinct aspects of the CRISPR system, each with unique characteristics. As a result, computational modeling typically focuses on one of these aspects to enhance the performance of the CRISPR system.

In the initial phases of CRISPR off-target analysis, numerous bioinformatics tools were employed to search for off-target sites. These tools included conventional alignment algorithms like bowtie (93), bowtie2 (94), bwa (95), TagScan (96), and GPGPU-enabled CUSHAW (97). However, these methods shared two common limitations: they were constrained by the number of allowable mismatches and relied on fixed PAM sequences. Consequently, new algorithms emerged that departed from alignment-based approaches. Initially, these algorithms adopted hypothesis-driven methods, which evaluated off-target activity using specific formulas. Over time, they evolved into learning-based methods. For example, a web-based software tool developed by Feng Zhang's group assessed off-target scores using a formula based on the count of mismatched nucleotides and their distances (65). This formula was then used to classify whether a given gRNA had a significant off-target effect by applying a predefined threshold (98). Subsequently, Doench J.G. et al. (74) introduced a method called CFD (cutting frequency determination), which predicted off-target scores by considering the frequency of bases at each position within the gRNA spacer sequence (74). Haeussler et al. (99) conducted an extensive evaluation of various machine learning methods and integrated them into the gRNA design tool CRISPOR (99). They recommended CFD as an off-target reference because it provided an aggregated score for a single gRNA that encapsulated its impact on off-target sequences,
demonstrating high accuracy. In 2018, deep learning-based methods entered the scene for CRISPR off-target scoring. Two models, CNN_std (100) and DeepCRISPR (73), employed convolutional neural networks (CNNs) to predict gRNA specificity scores within the CRISPR-Cas9 system.

Thanks to the extensive library generation, many machine learning-based methods have emerged for predicting CRISPR on-target activity. As this field has progressed, model performance has significantly improved, but with the cost of increased model complexity. For instance, CRISPRater (84), CRISPRScan (101), and SSC (87) employ straightforward linear models, while Azimuth1.0 (83) is trained using generalized linear models. These linear models are easy to interpret and train, allowing users to run them quickly. However, linear models have limitations in handling non-linear feature relationships. Consequently, non-linear models like WU-CRISPR (86), Azimuth2.0 (74), and TUSCAN (102) have emerged, offering enhanced prediction capabilities while posing challenges in model interpretation. Despite benefiting from large-scale library generation, most models show modest performance in individual gRNA/target design (84), leading to further research on modeling strategies. In response, two advanced models, TSAM (85) and sgDesigner (88), employ ensemble-learning methods to integrate multiple machine learning models, surpassing single-model algorithms in overall prediction performance. Similar to the model development of off-target analysis, several deep learning-based algorithms, including DeepCas9 (103), DeepCRISPR (73), DeepCpf1 (104), and CRISPRCpf1 (105), utilize convolutional neural networks (CNNs) to predict sgRNA activity by automatically recognizing sequence features. The primary strength of deep learning lies in its ability to automatically identify crucial features, due to its complex neural network structure. However, the feature extraction process can be likened to a black box, making it challenging to
validate these features functionally (106). Furthermore, existing public datasets contain only tens of thousands of human cells. While data augmentation techniques can artificially expand the dataset, this refinement can mask the actual information in the data, making it difficult to accumulate the millions of data points required to build a powerful deep learning model.

1.3 Machine Learning and Deep Learning Applications

1.3.1 Backgrounds of Machine Learning

Machine learning is a field that revolves around two core inquiries: firstly, the creation of computer systems that enhance their performance autonomously through accumulated experience. Secondly, the examination of the underlying statistical, computational, and information-theoretic principles that govern learning systems, encompassing machines, humans, and organizations. The study of machine learning holds significance in tackling these fundamental scientific and engineering challenges, as well as in harnessing the practical computer software it has generated and implemented across various domains.

Over the past two decades, machine learning has advanced significantly, evolving from a mere laboratory curiosity to a highly practical technology now extensively used in commercial applications. In artificial intelligence (AI), machine learning has become the preferred approach for creating functional software across various domains, including computer vision, speech recognition, natural language processing, robot control, and more. Many AI developers now recognize the efficiency of training systems by providing examples of desired input-output behavior, as opposed to manual programming that requires anticipating responses for all possible inputs. The impact of machine learning extends beyond AI and has had a profound influence on computer science and numerous data-intensive industries, such as consumer services, complex
system fault diagnosis, and logistics chain management. Its effects have also reverberated across empirical sciences, such as biology, where machine-learning methods have been innovatively applied to analyze high-throughput experimental data.

A learning problem involves enhancing performance in task execution through training. For instance, in the context of identifying specific molecular activities in biological processes, the task is to label processes as "occurring" or "not occurring." The performance metric we seek to improve is typically the accuracy of this molecular activity classifier. To achieve this, we use training experiences, which consist of historical data from experiments on related processes. These historical data are retrospectively labeled as either the activity of interest occurring or not. Alternatively, one can define a performance metric that penalizes labeling "occurring" as "not occurring" more than the reverse error. Moreover, the training experience can be adapted, such as incorporating unlabeled data alongside labeled examples from biological processes.

Deep learning is a subset of machine learning algorithms that employ multiple layers to extract higher-level features from raw input data progressively. It aims to replicate and automate human learning processes, such as recognizing objects like dogs from visual data like images. The foundation of most modern deep learning models lies in multi-layered artificial neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). In these neural network models, each layer is responsible for learning to transform input data into a more abstract and composite representation. For instance, in an image recognition task, the initial input could be a pixel matrix, with the first layer abstracting and encoding pixel-level information, the subsequent layer capturing edge patterns, the following layer identifying facial features like a nose and eyes, and the final layer recognizing the presence of a face. Notably, a deep learning process autonomously determines which features to place at each level optimally.
In supervised learning tasks, deep learning methods eliminate the need for manual feature engineering by converting data into compact, intermediate representations similar to principal components. They also establish layered structures that reduce redundancy in data representation. These approaches are also applicable to unsupervised learning tasks. Compared to traditional machine learning models, deep learning models benefit significantly from large-scale datasets, demand less human intervention, and deliver superior predictive performance. However, it's essential to acknowledge that setting up deep learning models can be more intricate and necessitate more powerful hardware and computational resources.

Machine learning, as an academic field, resides at the intersection of computer science, statistics, and various disciplines that focus on automatic improvement over time, as well as making inferences and decisions in uncertain situations. While there has been more collaboration with other fields in the past decade, we are only starting to explore the potential synergies and the vast array of formalisms and experimental approaches employed in these diverse disciplines to study systems that enhance their performance through experience.

1.3.2 Machine Learning and Deep Learning Application for CRISPR

Cleavage Efficiency Prediction

As introduced previously, machine learning and deep learning technologies were implemented to build prediction algorithms for both target specificity and cleavage efficiency of the CRISPR system. However, because greater Cas9 cleavage efficiency can potentially result in more pronounced knockout phenotypic alterations, the relevance of Cas9 cleavage efficiency gains particular prominence in scenarios involving extensive screening assays, where multiple genes are targeted using a comprehensive genome-wide CRISPR/Cas9 sgRNA library. In light of
this, the forthcoming section will concentrate on the utility of machine learning and deep learning in predicting CRISPR on-target activity.

In the context of applying machine learning to biological problems, linear models serve as an initial approach to gain preliminary insights. This is because linear models are adept at representing the relationship between a single output variable (known as the response) and one or more input variables (referred to as explanatory variables) through linear predictor functions. For instance, the sgRNA Designer model, introduced by Doench et al. (83), stands as an early prediction algorithm for estimating the cleavage efficiency of gRNAs. The authors analyzed potential target sites within six mouse genes and three human genes to extract relevant features. Subsequently, they employed Logistic Regression (LR), a statistical model, to train a classification model for gRNA cleavage efficiency prediction. In LR, the probability of an event occurring is modeled by expressing the logarithm of the odds of the event as a linear combination of one or more independent variables. The output of this model provides the probability that a given gRNA will be efficient in CRISPR experiments.

Linear models alone may not suffice for addressing scientific problems, as they cannot inherently capture intricate non-linear aspects. Therefore, it becomes essential to incorporate non-linear models to delve deeper into the relationship between training features and the predicted variable. To illustrate, the Doench group enhanced their prediction model (74) by expanding their training dataset with new experimental data. They also integrated various new features, including melting temperature, counts of position-independent nucleotides, and the target site's location within the corresponding gene. Moreover, they harnessed the capabilities of the Gradient Boosted Decision Tree (GBDT), a potent non-linear machine learning model. Gradient Boosting operates by constructing sequential weak prediction models, each aiming to
predict the residual error left behind by its predecessor. GBDT employs Decision Tree (DT) models as its weak learners and adopts a stage-wise approach similar to other boosting techniques. Notably, it generalizes the other methods by enabling the optimization of arbitrary differentiable loss functions, allowing it to estimate all types of relationships between the target-dependent variable and the independent variables. Furthermore, GBDT frequently outperforms other tree-based models like Random Forest (RF).

Due to the growing need for enhanced gRNA cleavage efficiency and more accurate predictions, researchers sought to improve prediction capabilities beyond single machine learning models. Several studies have explored the combination of multiple models to elevate predictive performance. For example, in the study by Peng et al. (85), the TSAM algorithm was developed by aggregating predictions from diverse models. The process involved creating an initial feature set, generating first-step scores, and assessing feature importance using a pre-trained Gradient Boosting Regressor. Subsequently, the most critical features were combined with profiled Hidden Markov Model (pHMM) features, and second-step scores were computed using an SVM regressor. The final regression scores were derived as the average of the two-step scores. The authors emphasized that their two-step averaging strategy highlights the complementary nature of the boosting regression and SVM regression approaches, and the integration of these two regression results led to an improved prediction performance in the context of sgRNA cleaving efficiencies.

With the rapid advancement of deep learning, researchers have gained a more potent tool to analyze gRNA activities more comprehensively. One commonly employed neural network for predicting gRNA cleavage efficiency is the Convolutional Neural Network (CNN) (73, 90, 104). CNN is a feed-forward neural network that employs regularization and automatically learns
Figure 1.5: The architecture of the generally used convolutional neural network (CNN). The figure is from Chang-Cheng et al. (132).

Feature engineering through filter optimization. Specifically, CNN utilizes convolutional layers to convolve input data and reduce the number of free parameters, allowing for deeper network architectures (107). Moreover, it employs regularized weights over fewer parameters to mitigate issues like vanishing and exploding gradients encountered during backpropagation in earlier neural networks (108, 109). To further reduce data dimensions, local or global pooling layers are incorporated, which consolidate the outputs of neuron clusters in one layer into a single neuron in the subsequent layer. Fully connected layers, linking every neuron in one layer to every neuron in another, can be introduced to handle flattened matrix data and yield classification or regression results (Figure 1.5). For instance, Kim et al. (90) developed a neural network, which comprised one convolutional layer and three fully connected layers, for predicting gRNA efficiency using a CNN architecture. They converted input gRNA sequences into a four-dimensional binary matrix using one-hot encoding (representing A, T, C, and G nitrogenous bases at each position in the DNA molecule). The authors introduced 210 filters of various sizes (100 filters of 3 nucleotides, 70 filters of 5 nucleotides, and 40 filters of 7 nucleotides) to enhance CNN performance (110). They reported that the final model, "DeepSpCas9", exhibited high generalization performance in predicting CRISPR/Cas9 on-target activity.
Figure 1.6: The architecture of basic recurrent neural network (RNN). RNN does not assume that the data points are intensive. Instead, it carries out a specific task based on the preceding data in a sequence, essentially leveraging a form of memory. RNN cannot remember from longer sequences or time. It is unfolded during the training process. The figure is from Shanmugamani et al. (133).

In addition to Convolutional Neural Network (CNN), recent studies have demonstrated the effectiveness of Recurrent Neural Network (RNN), another deep learning architecture, for analyzing DNA and protein sequences (111, 112). RNN can automatically extract valuable features from raw DNA and protein sequences without manual feature engineering, similar to CNN. However, there is a fundamental difference between them: while CNN treats DNA input as grayscale images with binary pixel values, RNN is explicitly designed for ordered sequence problems (113). RNN is more naturally suited to treating DNA sequences as sentences composed of four character types: A, C, G, and T, rather than as images. Thus, this perspective could benefit from insights gained from natural language processing research (114). Notably, RNN incorporate loops in their connections, introducing feedback mechanisms and memory to the networks over time or spatial sequences. This memory capacity allows RNN to learn and generalize across sequences of inputs rather than individual patterns (Figure 1.6). Additionally,
the subclass of RNN, known as Long Short-Term Memory (LSTM), uses "gates" to control the influence of previous sequence features, enhancing flexibility in memory management. To illustrate the superior performance of RNN, Wang et al. (89) conducted a study comparing four conventional models (Linear regression, Ridge regression, XGBoost, and multilayer perceptron), as well as two deep learning models, CNN and RNN, for predicting cleavage efficiency. Their results indicated that RNN outperformed all other models. Furthermore, because deep learning models can benefit from adding external features not automatically extracted by the models (115), the authors combined indirect biological features with the RNN architecture to enhance prediction performance. The resulting algorithm was validated and demonstrated superior performance when compared to other published algorithms using eight human CRISPR datasets.

### 1.3.3 Machine Learning and Deep Learning Application for MicroRNA Targets Prediction

In miRNA target identification, machine learning (ML) methods differ from traditional approaches by focusing on recognizing potential miRNA targets based on established miRNA-mRNA interactions with known biological significance. These algorithms undergo a training process using experimentally confirmed miRNA-mRNA interactions as positive instances and artificially generated negative instances. During this training, the ML model strives to identify patterns that can distinguish between real and false targets. When presented with previously unseen datasets, these learned patterns are employed to predict whether a given target is real accurately. Notably, ML doesn't rigidly rely on exact seed matches in the 3' untranslated regions (3'-UTRs) but instead learns from provided examples that hold biological relevance. This flexibility allows ML to identify non-canonical binding sites, even those within coding regions.
However, it's crucial to recognize that ML solely learns from the provided examples and can therefore only produce results resembling those examples (116).

As mentioned earlier, TargetScan is one of the earliest models for predicting miRNA targets. In its updated 2015 version, the authors adopted a straightforward linear regression approach. To build this model, they leveraged a comprehensive dataset comprising 74 microarray datasets involving HeLa cells transfected with individual miRNAs. Through stepwise regression, this linear model efficiently identified and utilized 14 out of the available 26 pertinent features for training, resulting in a highly interpretable model (13). However, it's important to note that the simplicity of this model comes with certain limitations. Specifically, it primarily captures linear relationships within the data. Additionally, the stepwise regression technique, while facilitating model interpretability, has been criticized for potentially oversimplifying models and introducing biases toward the training data (117, 118). In the context of functionality, TargetScan predicts the repressive potential of individual target sites and then combines these predictions using a mathematical model to generate scores for miRNA-mRNA pairs. It's worth noting that this approach does not account for the potential synergistic effects of binding sites on the same mRNA based on experimentally derived training data. This omission has the potential to introduce inaccuracies in the predictions.

In contrast to TargetScan, the miSTAR method introduced in 2016 employs a more intricate approach to tackle a significant challenge (119). Specifically, the authors devised a two-layered stacked model. The initial layer uses a Random Forest (RF) classification to gauge whether a single miRNA binding site triggers repression. The subsequent layer employs these assessments per-transcript basis and forecasts repression for the miRNA-transcript pair using Logistic Regression (LR). miSTAR's foundation rests upon 3’UTR reporter assays conducted for
470 miRNAs and 3' UTRs sourced from 17 genes. This design offers an advantage in light of recent findings that reveal substantial variations in binding patterns across different miRNAs (120), enhancing its applicability to unobserved miRNAs. When evaluated using the AUC-ROC score, miSTAR surpasses other widely used models in performance. It is worth noting, however, that the authors' performance assessment employed a rigorous cross-validation method but was limited to the dataset used for model optimization. Additionally, the model was trained on a relatively small number of interactions, the majority of which were negative in nature.

Recent advances in deep learning technology have significantly improved computational modeling research in miRNA target prediction. In the latest version of TargetScan, McGeary et al. harnessed the power of CNN to enhance their miRNA regulatory analysis (31). They conducted in vitro binding affinity experiments, demonstrating a strong correlation between the affinities they measured and the mRNA repressions observed in miRNA-overexpression experiments. To extend the applicability of their experimental findings from a limited number of individual miRNAs to the entire set of annotated miRNAs, they employed a CNN, which predicts binding affinities based on provided miRNA-target pair sequences. Their measurements and predictions exhibited greater accuracy than previous approximations of binding affinity. Notably, they achieved a 31% improvement in predictive performance for miRNA-mediated target repression compared to the previous TargetScan model (13). This enhancement resulted from integrating the CNN-predicted binding affinity into a regression model and training both models simultaneously.
Alongside commonly employed deep learning architectures like CNN and RNN, the autoencoder has emerged as a precious tool in this context (121-123). An autoencoder, a type of artificial neural network, excels at learning efficient representations of unlabeled data through unsupervised learning (124, 125). It comprises two primary functions: an encoding function that transforms input data and a decoding function that reconstructs the input data from the encoded representation. A basic autoencoder consists of three layers: an input layer for data reception, a hidden layer to host the encoded data representation, and an output layer for generating reconstructed results from this encoded representation (Figure 1.7). Additionally, employing multi-layered (deep) encoders and decoders offers several advantages (126). Specifically, increased depth can significantly reduce the computational cost of representing certain functions, decrease the required amount of training data for learning these functions, and empirically, deep autoencoders outperform shallow or linear autoencoders in compression capabilities (127). For instance, Pla et al. incorporated a deep autoencoder architecture into their miRAW model (122).
This approach utilized a pre-trained autoencoder with five layers: the initial hidden sparse layer enhanced the dimensionality of the data, enabling representation in a more complex dimension (over-completion). Hidden layers two through five were dedicated to identifying relevant features representing the data. These layers were pre-trained independently as isolated autoencoders to capture the most representative features of miRNA-mRNA duplexes. Furthermore, the architecture incorporated three fully connected layers to classify the features learned by the autoencoder. The authors subsequently validated miRAW on two independent test sets, achieving a 23% performance improvement over the second-best model.

1.4 Project Aims

The goals of this project were twofold:

1. To create an advanced CRISPR/Cas9 gRNA design tool that excels in predicting cleavage efficiency and establish a robust algorithm comparison approach to demonstrate the superiority of our tool in comparison to existing published algorithms.

2. To develop a state-of-the-art miRNA target prediction algorithm utilizing deep learning technologies and enhance our miRDB web portal by incorporating this new algorithm along with significant analysis features.

To achieve our first objective, we initiated the process by creating a plasmid target library. This library served as the foundation for quantifying the effectiveness of gRNAs in the CRISPR/Cas9 system through experimentation. We leveraged a meticulously designed dataset of gRNAs targeting plasmid sites generated through in-silico methods. Subsequently, we conducted a thorough analysis of the features within our gRNA library and harnessed these features to train a machine learning model tailored explicitly for gRNA design. Our ultimate model, sgDesigner, was crafted using a stacked generalization framework that integrated distinct models, resulting in
heightened prediction robustness. Notably, sgDesigner surpasses existing gRNA design algorithms in predicting gRNA potency. However, it's crucial to acknowledge that individual algorithms were constructed using diverse computational strategies, relying on distinct training datasets and feature designs. Consequently, their performance can exhibit substantial variations when applied in various experimental scenarios. To address this variability, we further conducted an exhaustive comparison of established gRNA design algorithms across an extensive range of published datasets. We identified the top-performing algorithms and improved our predictive model by employing the Stacking ensemble method, which combines these top algorithms. Our validation analysis conclusively demonstrates that this new ensemble model outperforms any individual algorithm in our study, marking it as the superior choice regarding overall performance. Finally, the new model is accessible as a web application at https://crisprdb.org.

To accomplish the second goal of this project, we first conducted a comprehensive analysis utilizing both CLIP binding data and miRNA overexpression data. We aimed to identify common characteristics shared between miRNA binding and the downregulation of target genes. Moreover, we harnessed the power of a Long Short-Term Memory (LSTM) deep learning architecture to automatically extract sequence-related features from experimentally validated miRNA-mRNA interactions. These biological attributes, derived from both CLIP binding and miRNA overexpression datasets, were then integrated with the LSTM architecture's outputs through several fully connected layers to yield our final prediction results. Our model demonstrated significant improvements in prediction performance, as validated through various analyses. Furthermore, we have implemented this new prediction algorithm into our web application, miRDB, enabling users to perform custom target predictions using their provided sequences. Another noteworthy addition to our web application is the capability to predict
miRNA targets specific to different cell types. MiRDB now encompasses expression profiles from over 1000 cell lines, tailoring target predictions to specific cellular models. Lastly, we have introduced a new web query interface on miRDB for predicting miRNA functions through the integrative analysis of target prediction and Gene Ontology data. All these valuable resources are readily accessible through our website at https://mirdb.org.

1.5 References


Chapter 2: Generalizable sgRNA Design for Improved CRISPR/Cas9 Editing Efficiency

This chapter is adapted from and expanded upon the following publication (1):


* The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

2.1 Abstract

**MOTIVATION:** The development of clustered regularly interspaced short palindromic repeat (CRISPR)/CRISPR-associated protein 9 (Cas9) technology has provided a simple yet powerful system for targeted genome editing. In recent years, this system has been widely used for various gene editing applications. The CRISPR editing efficacy is mainly dependent on the single guide RNA (sgRNA), which guides Cas9 for genome cleavage. While there have been multiple attempts at improving sgRNA design, there is a pressing need for greater sgRNA potency and generalizability across various experimental conditions.

**RESULTS:** We employed a unique plasmid library expressed in human cells to quantify the potency of thousands of CRISPR/Cas9 sgRNAs. Differential sequence and structural features among the most and least potent sgRNAs were then used to train a machine learning algorithm for assay design. Comparative analysis indicates that our new algorithm outperforms existing CRISPR/Cas9 sgRNA design tools.

2.2 Introduction

The clustered regularly interspaced short palindromic repeat (CRISPR)/Cas systems have provided an unprecedented opportunity for performing site-specific editing of a variety of genomes. In prokaryotes, CRISPRs are virus-derived DNA fragments which encode CRISPR RNA (crRNA) (2). The CRISPR/Cas9 system is the most widely used editing system due to its simplicity, high efficiency and low cost (3). In conjunction with trans-activating crRNA (tracrRNA), the crRNA serves as a guide for Cas9 to bind and cleave foreign DNA (4). In genome editing experiments, tracrRNA and crRNA are engineered into a combined single guide RNA (sgRNA) with a designed guide sequence complementary to the desired target (5). Similar to the natural system, the sgRNA guides Cas9 to cleave the DNA at a specific genomic locus based on sequence match, resulting in a double-stranded DNA break. This break occurs precisely 3 nt upstream of an NGG protospacer adjacent motif (PAM) sequence (6). In mammalian cells, the DNA repair process often introduces indels, causing frameshift mutations and resulting in functional gene knockout. From this editing framework, advanced strategies have also been developed such as paired nicking for increased specificity (7) or inserting nucleotide sequences during the repair of double-strand break to generate knock-ins (8). The broad applicability of the CRISPR/Cas9 system stems from its ability to target DNA based on a synthetic sgRNA sequence, specifically the 20 nt guide sequence (gRNA) at the 5’ end of the sgRNA sequence.
It has been shown that the gRNA sequence is important for both targeting specificity and cleavage efficiency (9, 10). Off-target Cas9 activity occurs when sequences similar to the gRNA occur elsewhere in the genome, potentially resulting in unintended knockout effects (9). To address the off-target effects, various experimental approaches, mainly by altering the nuclease activity of Cas9 (resulting in high-fidelity Cas9) or gRNA design, have been established in recent years, resulting in significantly improved specificity for CRISPR/Cas9 targeting (7, 11-14). Further, recent studies have also developed bioinformatics methods to design sgRNA sequences with reduced off-target effects (9, 15, 16). However, these experimental innovations could still suffer from potential cleavage efficiency variations. More efficient Cas9 cleavage can potentially result in stronger knockout phenotypic changes. The importance of Cas9 cleavage efficiency is amplified when considering large-scale screening assays where many genes are to be knocked out using a genome-wide CRISPR/Cas9 sgRNA library.

Several studies have approached the problem of sgRNA efficiency prediction, revealing sgRNA and target features that correlate with Cas9 cleavage efficiency (15-22). Given the large number of features involved, machine learning methods are commonly employed for data modeling. To construct such computational models, a large number of sgRNAs need to be experimentally tested to build a robust training dataset for efficiency prediction. In order to do so, existing studies typically adopted biological enrichment schemes, in which gene editing events impact cell survival or other observable biological phenotypes. While these strategies avoid labor intensiveness on the experimental side, such indirect biological readouts could produce artifacts in the training data, as equally efficient Cas9 cleavage sites may not result in equal phenotypic changes or survival pressure. Furthermore, existing experimental studies were often focused on a small subset of genes and/or a single cell line, which limits the usefulness of
the training data for general predictions. In our study, we generated a plasmid target library for experimental quantification of sgRNA efficiency in the CRISPR/Cas9 system. Using in silico designed target sites as presented from a large plasmid library, our large-scale training dataset reduces potential bias from specific experimental systems and is generalizable across other datasets. We performed comprehensive feature analysis of our sgRNA library and used the extracted features to train a machine learning model for sgRNA design.

Our final model, which we named sgDesigner, was developed by utilizing a stacked generalization framework to combine distinct models, resulting in more robust predictions (23). sgDesigner outperforms existing sgRNA design algorithms for sgRNA potency prediction and is publically accessible as a web application via http://crispr.wustl.edu.

2.3 Materials and Methods

2.3.1 Cloning of sgRNA Plasmid Library

A pool of 12 472 oligonucleotides were synthesized by CustomArray, Inc. (Bothell, WA, USA). Each oligonucleotide contains a 20 nt gRNA sequence and paired 53 nt target sequence (including a NGG PAM). Among these oligos, 11 472 gRNA sequences were randomly selected from coding exons in humans. Most of these gRNAs (93%) cannot target the endogenous exon sites due to the lack of adjacent PAM domains in the genomic sequence. In these cases, a PAM domain was added next to the gRNA sequence in the plasmid to make the site targetable by Cas9. In addition, 1000 randomly shuffled gRNA sequences were included in the oligo pool to serve as negative control. Between the gRNA and target sequence, two BsmBI sites for Cas9 sgRNA scaffold cloning and a 12 nt unique molecular index (UMI) sequence for bioinformatics analysis were inserted. Two constant regions (20 nt each) at the 5’ and 3’ ends were added for
PCR amplification of the oligonucleotides. The oligo pool was amplified by PCR with Phusion DNA polymerase (ThermoFisher) using primers ‘Cas9Lib_FP’ and ‘Cas9Lib_RP’ (Table 2.1). Amplified DNA oligos were then gel purified using the QIAquick Gel Extraction kit (QIAGEN). Next, purified PCR products were assembled into the BsmBI-digested plasmid Lenti-gRNA-Puro (Addgene #84752) using the NEBuilder HiFi DNA Assembly kit. This plasmid was referred as the Library-1st plasmids in our study.

Cas9 sgRNA scaffold sequence was amplified from Lenti-CRISPR V2 (Addgene #52961) using the primers ‘scaffold RNA FP’ and ‘scaffold RNA RP’ (Table 2.1). The Cas9 sgRNA scaffold PCR products were then gel purified. After BsmBI digestion of both library-1st plasmids and Cas9 sgRNA scaffold, the two fragments were ligated by T4 DNA ligase (Intact Genomics) to get the final library plasmids.

**Table 2.1. List of Primers Used in the Study**

<table>
<thead>
<tr>
<th>Primers</th>
<th>Sequences (5’ -&gt; 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target primer FP</td>
<td>acactctttccctacacgacgctcttcggatctcttgtggaaaggacgaaacacc</td>
</tr>
<tr>
<td>Target primer RP</td>
<td>gtgactggagtccagctgtgtctctttccgatctttgtggatgaataactgccatattgct</td>
</tr>
<tr>
<td>Illumina adaptor primer</td>
<td>atatgatacggegccagccgccgatctacacactctttccctacagac</td>
</tr>
<tr>
<td>SIC-index primer</td>
<td>caacgagaagccgacgggataacgagttctagctgtgcaggtggagtgcagctgttcgtccga</td>
</tr>
<tr>
<td>cas9Lib_FP</td>
<td>agtttttccgatctttgtttatatatatctttgtgtggaaaggacgaaacacc</td>
</tr>
<tr>
<td>cas9Lib_RP</td>
<td>tgcagtttaaggagtagaagctgacgctgaagtaacagacgtgagtgcactggctctctttag</td>
</tr>
<tr>
<td>scaffold RNA FP</td>
<td>gctttgctttttgtacgtctctcttttag</td>
</tr>
<tr>
<td>scaffold RNA RP</td>
<td>gtaggacagtccggtggaaaaagcaccgactgtgcagacgtgcactggctctctttag</td>
</tr>
</tbody>
</table>
Following Gibson assembly or T4 DNA ligation, 2 µl of the reaction was transformed into 25 µl of ig™ 10B ElectroCompetent cells (Intact Genomics) by electroporation. To maximize library coverage, two electroporation reactions were performed. After transformation, cells were pooled and spread onto LB agar plates supplemented with 100 µg/ml ampicillin. All clones were harvested for plasmid DNA extraction by the PureYield Plasmid Midiprep kit (Promega). Throughout the cloning process, the transformation efficiency and library coverage were evaluated according to previously published guidelines (24). On average, there are about 300 colonies per sgRNA oligo in the plasmid library.

2.3.2 Lentivirus Preparation

The infectious lentivirus particles were generated and packaged using 293 T cells (ATCC). In a 60 mm dish, 2 × 10^6 cells were seeded in 2.5 ml Dulbecco’s modified Eagle’s medium (DMEM) supplemented with 10% fetal bovine serum (FBS) (Gibco). About 2 µg of library plasmids or Cas9 expressing plasmids were mixed with 1.8 µg of psPAX2 (Addgene #12260) and 0.2 µg of pCMV-VSVG (Addgene #8454) in 250 µl of OPTI-MEM medium (ThermoFisher), while 12 µl of Lipofectamine 2000 (ThermoFisher) was diluted in 250 µl of OPTI-MEM medium. After 5 min of incubation at room temperature, the plasmid mixture and the diluted Lipofectamine 2000 were combined and incubated for 20 min at room temperature. After incubation, the 500 µl plasmid–lipofectamine mixture was dropped onto the 293 T cells. The transfection medium was replaced with regular cell culture medium 6 h post transfection. Virus was harvested at 40 h post transfection and filtered through a 0.45 µm Millex-HV membrane (Millipore).
2.3.3 Plasmid Library Delivery into HeLa/Cas9 Cells

Adherent HeLa cells (ATCC) were cultured in DMEM medium supplemented with 10% FBS. Cells were cultured in a 37°C incubator supplied with 5% CO2. To establish the Cas9 stable expressing cells, HeLa cells (ATCC) were transduced with lentivirus containing Cas9 expressing transcripts from LentiCas9-Blast (Addgene #52962) at an MOI of 0.7. Two days after transduction, cells were selected with 10 µg/ml blasticidin for 4 days. The blasticidin-resistant cells were pooled and maintained in the presence of 10 µg/ml blasticidin. Cas9-expressing cells were transduced with the lentivirus expressing the plasmid library at an MOI of 0.3. Two days after transduction, cells were treated with 2 µg/ml puromycin for 3 days. Survived cells were harvested and genomic DNA was isolated for sequencing library construction.

2.3.4 Sequencing Library Construction

Genomic DNA was isolated using the GenElute Mammalian Genomic DNA Purification kit (SigmaMillipore). The sequencing library was constructed according to the methods described previously (25). In brief, the target sequence was first amplified and then the Illumina adaptor and barcode sequences were introduced by a second PCR. All primers used in these two PCRs were listed in Table 2.1. The final PCR products were purified with AmpureXP beads (Beckman Coulter), quantified with the Quantifluor system (Promega) and then sequenced with MiSeq (Illumina).

2.3.5 Quantification of sgRNA Efficiency

FASTQ raw sequencing data were de-multiplexed and ambiguous reads were filtered out. Each sequencing read was identified using its gRNA sequence and UMI and subsequently aligned to its reference sequence using Smith–Waterman alignment with affine gap penalty to
detect editing (with parameters for match = 3, mismatch = −2, gap opening = −10 and gap extension = −1). In this step, plasmids that exhibited indels prior to exposure to Cas9 were excluded from further analysis. Proportion of reads edited was used to quantify sgRNA efficiency. sgRNAs with 100% of associated reads edited were considered of high efficiency, while those with \( \leq 50\% \) of associated reads edited were considered of low efficiency. A minimum read count of 10 per sgRNA was required for sgRNA inclusion in the analysis, and an additional criterion of at least two UMIs were required for each sgRNA included the high-efficiency group in order to maximize training data quality.

### 2.3.6 Computational Tools

Data processing, sequence alignment and feature extraction were performed using custom Perl scripts. RNAfold (26) was used to compute sgRNA structural features. Features analysis was performed using MATLAB. Significance levels (P-values) were calculated using Student’s t-test for numerical features and \( \chi^2 \) test for binary features. Feature enrichment was determined by comparing functional sgRNAs with non-functional sgRNAs. Computational modeling and performance evaluation were performed using Python.

### 2.3.7 Independent Testing Datasets

A total of six testing datasets were gathered from published studies, namely: Wang, Koike-Yusa, FC, RES, Shalem and Chari (16-18, 27-29). The Wang and Koike-Yusa datasets were downloaded from supplemental tables provided by Xu et al. (22). We used the negative log2 fold change values for correlative analysis of sgRNA prediction. The FC and Shalem datasets were provided by Doench et al. (18). The RES dataset was downloaded from the Azimuth website, which was implemented by Microsoft (16). The Chari dataset was directly
retrieved from supplemental tables at the journal’s website (17). Additional details about these datasets can be found in Table 2.2.

**Table 2.2. Summary of validation datasets**

<table>
<thead>
<tr>
<th>Name</th>
<th>sgRNAs</th>
<th>High Efficiency gRNAs</th>
<th>Low Efficiency gRNAs</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang Dataset</td>
<td>2077</td>
<td>1402</td>
<td>675</td>
<td>Wang et al. (29) (a)</td>
</tr>
<tr>
<td>Shalem Dataset*</td>
<td>1278</td>
<td>256</td>
<td>256</td>
<td>Shalem et al. (28) (b)</td>
</tr>
<tr>
<td>Koike-Yusa Dataset</td>
<td>1064</td>
<td>830</td>
<td>234</td>
<td>Koike-Yusa et al. (27) (a)</td>
</tr>
<tr>
<td>FC Dataset*</td>
<td>1841</td>
<td>368</td>
<td>368</td>
<td>Doench et al. (18)</td>
</tr>
<tr>
<td>Chari Dataset</td>
<td>1234</td>
<td>133</td>
<td>146</td>
<td>Chari et al. (17)</td>
</tr>
<tr>
<td>RES Dataset*</td>
<td>2549</td>
<td>509</td>
<td>509</td>
<td>Doench et al. (16)</td>
</tr>
</tbody>
</table>

*For datasets without high/low gRNA classification provided, the top 20% and bottom 20% gRNAs were labeled high and low efficiency, respectively

a. Provided by Xu et al. (22)  
b. Provided by Doench et al. (18)

**2.3.8 Model Performance Evaluation**

We compared sgDesigner with three existing sgRNA design tools included RS2, Sequence Scan for CRISPR (SSC) and DeepCRISPR. RS2 prediction results were retrieved from Microsoft’s Azimuth 2.0 website, using predefined in vitro parameters (16). SSC prediction
results were computed using the authors’ web-based implementation (22). DeepCRISPR predictions were generated using the command-line version with sequence features only (15). Receiver operating characteristic (ROC) and Spearman correlation analyses were performed using the testing datasets to assess the consistency between experimentally determined sgRNA efficiencies and predicted efficiencies.

2.3.9 Availability of Data

Our sgRNA design tool, sgDesigner, is freely accessible as a web application via http://crispr.wustl.edu. In addition, the source code and stand-alone version of sgDesigner are freely accessible at GitHub (https://github.com/wang-lab/sgDesigner). Additional supplementary data can be downloaded from journal’s website and Zenodo.org (http://doi.org/10.5281/zenodo.3572803).

2.4 Results

2.4.1 An sgRNA Library for Quantifying CRISPR/Cas9 Editing Efficiency

The overall experimental procedure is summarized in Figure 2.1a. In summary, we synthesized an sgRNA library and used it for cleavage efficiency quantification by high-throughput sequencing. Specifically, we designed a pool of oligonucleotides each containing a U6 promoter sequence, a gRNA sequence and a corresponding target sequence. The target sequence included a 20 nt gRNA-matching region, followed by an NGG PAM domain. In addition, to examine the potential impact of target flanking regions, we also included 30 distinct nt surrounding the target site, forming an extended target site of 53 nt. The oligos were cloned into plasmids by Gibson assembly, and then the sgRNA scaffold was inserted downstream of the gRNA sequence. In this way, each plasmid contains both an sgRNA expression cassette and a
paired target sequence. This plasmid library was transduced into Cas9-expressing cells, and the editing of in silico designed target sequences was determined by sequencing. In this way, the

Figure 2.1: Determining sgRNA efficiencies using an in silico designed sgRNA library. (a) Experimental outline for construction of the sgRNA library and generation of training data for computational modeling. (b) Stratification of sgRNAs based on editing efficiency.

potency of 7407 sgRNAs were quantified after filtering out low-quality reads. Overall, Cas9 activity was high, with 81.1% reads edited. We stratified the sgRNAs into high-efficiency (100%
editing), medium-efficiency (51–99% editing) and low-efficiency (0–50% editing) groups and selected high-confidence sgRNAs for training (Figure 2.1b). To emphasize the most predictive features affecting sgRNA efficiency, our strategy was to only consider the high- and low-efficiency groups for data modeling, resulting in a training set comprised of 746 functional sgRNAs and 563 non-functional sgRNAs.

2.4.2 sgRNA/Target Features

Previous studies have identified multiple sgRNA and target features contributing toward Cas9 activity, such as position-specific nucleotide composition and GC content (15-22). However, feature comparison between published datasets reveals considerable discordance and further study is warranted to identify generalizable features affecting Cas9 efficiency. For example, Xu et al. (22) demonstrated that guanines are preferred at positions −14 to −17 of the 20-mer gRNA sequence, whereas this was not observed by Doench et al. (18). Our new dataset, with its direct cleavage quantification by employing a plasmid-based system for generation of target sites, provides a unique opportunity to isolate and quantify features that are intrinsically associated with Cas9 activity. We included a total of 26 nt for analysis, including the 20 nt gRNA-matching sequence, the NGG PAM and 3 nt downstream of the PAM (Figure 2.2a). In total, 302 features were chosen as machine learning input. These features were extracted by a combination of sequence and structural analyses of the gRNA and target sites, as described in detail below. We also explored additional features of the 53 nt extended target sequence, including 15 nt and 12 nt flanking the gRNA-matching/PAM sequence at the 5’ end and 3’ end, respectively. However, none of the additional extended positions were statistically significant after correcting for multiple comparison, nor did they improve model performance.
(Supplementary Figure 2.1). Thus, we excluded these additional nucleotide positions from further analysis.

Figure 2.2: Feature analysis of efficient sgRNAs. (a) Position-specific nucleotide composition. A positive or negative value represents enrichment or depletion of the nucleotide, respectively. Statistically significant nucleotides are depicted in blue. (b) GC content of the gRNA. (c) Significance of nucleotide accessibility at each position in the sgRNA. The significance level (P = 0.05) is depicted with red dashes. (d) Percent enrichment of nucleotide accessibility at each position in the gRNA. Significant positions are depicted in blue.

Nucleotide composition

Nucleotide composition at each target position is summarized in Figure 2.2a. Nucleotides at positions 1–20 are identical to those in the gRNA [with a thymine (T) to uracil (U)]
conversion], followed by the NGG PAM, which is a requirement for Cas9 targeting (30). Positions +1, +2 and +3 represent the genomic context of the target sequence. Functional gRNAs were depleted in T throughout 19 of the 23 positions (P-values in the range of 2.4E−27–1.8E−02; average depletion of 38%), while position +1 was significantly enriched in T (P = 1.3E−02; enrichment of 28%; Figure 2.2a). The overall depletion is in part due to transcriptional efficiency as opposed to an interaction involving Cas9. The U6 promotor used in the study recruits RNA polymerase III which recognizes a poly-T sequence as a termination signal (31). Consistent with this mechanism, none of the gRNAs in the high efficiency group contained a sequence of five or more contiguous T bases. However, this phenomenon does not account for the entirety of the depletion in T, since we still observed significant overall T depletion after excluding gRNAs with four or more contiguous Ts.

Nucleotides proximal to the PAM were the most predictive of Cas9 activity. Most significantly, functional gRNAs had strong enrichment in guanine (G) at positions 19, 20, and the N position of the PAM (P = 4.7E−19, 1.4E−36 and 5.9E−08; enrichment of 165, 392 and 71%, respectively). Adenines (A) were, however, more enriched toward the middle of the gRNA specifically at positions 9–12 and 14–16 (P-values in the range of 2.0E−08–1.8E−02; average enrichment of 53%). Cytosines (C) were most enriched at positions 17 and 18 (P = 1.2E−06 and 1.2E−07, enrichment of 60 and 73%, respectively).

**GC content**

We found decreased activity in gRNAs with extreme overall GC content. As shown in Figure 2.2b, the vast majority of gRNAs with GC content >80% and <30% were non-functional (depletion of 81 and 93%, respectively). These two features were significant and improved
overall predictions (P = 1.8E−02 and 1.1E−09, respectively). In contrast, we did not observe model improvement using absolute GC content values.

**Structural features**

RNA molecules commonly form secondary structures through intramolecular interactions, resulting in differential accessibility for the nucleotides within the folded structure. This phenomenon can potentially result in unfavorable sgRNA structures affecting Cas9 efficiency. However, most sgRNA design tools did not consider sgRNA nucleotide accessibility for Cas9 editing prediction. Here, we investigated these structural features using RNAfold (26) for structure prediction. The present dataset showed that sgRNA nucleotide accessibility at positions 18–20 of the gRNA domain are crucial for efficient editing (P = 2.1E−3, 5.8E−13 and 9.9E−23, respectively; Figure 2.2c). Functional sgRNAs tend to be accessible at these positions with enrichment values of 20, 41 and 68%, respectively (Figure 2.2d). In the predicted sgRNA secondary structure, these three nucleotides proximal to the PAM align with the nucleotides in the scaffold at positions 51–53 due to a stem-loop formation at positions 21–50. Thus, interestingly, increased accessibility at 51–53 is also significantly correlated with high Cas9 efficiency (Figure 2.2c). The sgRNA sequence at positions 51–53 is AAG, which would ideally bind to a CUU sequence at positions 18–20, or a UUU sequence when considering wobble base pairing. This may explain the observed depletion of U nucleotides at the 3’ end of the 20 nt gRNA. Our results suggest that there are more complex intramolecular interactions which may have been missed in other algorithms that do not consider structural features.

### 2.4.3 Assessment of Modeling Methods

Recently, several machine learning algorithms have been used to predict sgRNA efficiency. We summarize these algorithms into three categories: (i) regression models such as
gradient boosting regression tree (16) and extreme gradient boost (XGBoost) (20), (ii) classification models, such as support vector machines (SVM) (17, 21) and logistic regression (18) and (iii) emerging technologies or hybrid algorithms, such as deep learning technology (15) and simple average of multiple models (20). Given the variety of potentially useful models, our strategy was to use a stacking framework in order to capture the advantages of multiple models. Our Stacking model was designed by stacking SVM and XGBoost using a logistic regression model as the combiner (Figure 2.3). We compared Stacking performance against commonly used machine learning algorithms as well as models proposed in other sgRNA design studies.

**Figure 2.3:** Stacking model framework. The features were first used as input to train first layer models (SVM and XGBoost). Fivefold cross-validation was performed for each individual model in the first layer and the predictions from each model were merged into a two-column feature set. The resulting feature set was then used to train the second layer model (logistic regression).

Using our training dataset comprised of 746 high-efficiency sgRNAs and 563 low-efficiency sgRNAs, we employed 7 different classification-based approaches and tested the performance of each model. The models tested include Stacking, XGBoost, L2-regularization logistic regression, SVM, adaboost, random forests and decision trees. Each of the model parameters were tuned based on the training dataset. We used ROC curve, Spearman correlation...
and predicted mean accuracy from cross-validation analysis to evaluate the performance of each model. Based on these assessment metrics, we conclude that the Stacking model had the best performance among all models included in the analysis (Figure 2.4). The Stacking model developed using the training dataset was thus chosen to be the final model, which we named sgDesigner.

**Figure 2.4**: Comparison of different computational models. (a) ROC curve analysis. AUC values for individual models are shown in the legend. (b) Spearman correlation between experimentally determined efficiency and predicted efficiency score. Error bars indicate the SD. (c) Mean accuracy of sgRNA classification (high or low efficiency).
2.4.4 Validation of sgDesigner

To evaluate the general applicability of sgDesigner at predicting sgRNA efficiency, we curated six CRISPR/Cas9 sgRNA datasets from various cell lines (see Materials and methods for details; Supplementary Table 2.4). With these datasets, we compared sgDesigner with three existing state-of-the-art sgRNA design tools, including Doench Rule Set 2 (RS2) (16), SSC (22) and DeepCRISPR (15). These existing tools were selected for comparison because they are currently widely used and freely accessible to the public. To avoid training bias, we only considered independent datasets that were not used to train respective models. The prediction results for these independent datasets were separately generated using each design tool. For each tool, prediction results for all independent datasets were combined for subsequent performance evaluation. Specifically, we performed ROC and Spearman correlation analyses and used true data labels to evaluate the performance of the design tools. The area under the curve (AUC)–ROC and correlation coefficient analysis results are summarized in Figure 2.5(a, b), with more detailed results for each dataset presented in Supplementary Figures 2.2 and 2.4. Further, we present detailed performance evaluation of sgDesigner across six individual datasets in Figure 2.5(c, d). Compared with other tools, sgDesigner had the best performance, as evaluated by ROC, precision-recall and correlation analyses. Specifically, sgDesigner consistently outperformed all competing tools across all six independent datasets (Supplementary Figures 2.2-2.4). Overall, sgDesigner had consistently high performance across all testing datasets, with average ROC–AUC of 0.833 and a range of 0.765–0.899 (Figure 2.5c). From these validation results, we conclude that sgDesigner has robust performance and consistently performs well across various experimental settings.
Figure 2.5: Comparison of sgDesigner with public sgRNA design tools. sgDesigner and three other algorithms were included in this analysis. Validation analysis was performed using six independent datasets, and the combined results are summarized here. Detailed results on each testing dataset are presented in Supplementary Figures 2.2 and 2.4. (a) ROC curve analysis. (b) Spearman correlations between experimentally determined efficiency and prediction score. (c) Summary of ROC–AUC values for sgDesigner on six independent datasets. (d) Summary of Spearman correlation coefficient values for sgDesigner on six independent datasets.

2.4.5 Genome-wide sgRNA Design Database

Using the sgDesigner algorithm, we computed cleavage efficacy for CRISPR/Cas9 sgRNAs to target all human and mouse genes annotated in the NCBI RefSeq database. To reduce potential off-target editing, we also computed off-target scores for the sgRNAs and select those
with greater specificity using our previously published algorithm (21). In brief, we performed both gRNA seed search and BLAST alignment to identify potential off-targets that share identical 13-mer seed sequence or with at least 85% overall sequence homology to the gRNA sequence. Of note, we focused on identifying off-targets from all known exons (for both coding and non-coding genes) instead of the entire genome space which contains other potentially important non-coding regions.

2.5 Discussion

As the CRISPR/Cas9 system has quickly become a ubiquitous gene editing tool in biological research, an increasingly pressing challenge is the design of efficient sgRNAs. Various bioinformatics tools have been developed to address this important issue. However, one major limitation of previous studies is related to the quality of the datasets used to train such tools. Most experimental methods are based on phenotypic screening and are not ideal at quantifying CRISPR/Cas9 editing efficiency. Successful gene editing is unlikely to produce consistent and precise phenotypic changes across all genes and target sites tested. Thus, such indirect methods introduce undesired noises in the datasets used to train machine learning algorithms, which could mask true features that are characteristic of sgRNA-guided Cas9 cleavage. Furthermore, sgRNAs tested in functional screens are typically designed for a subset of genes and tested in a single cell line. These restrictions may introduce biases specific to each experimental setting, such as those related to different levels of genomic accessibility, or different responses to DNA cleavage in a cell line or gene specific manner. All these factors may potentially reduce model generalizability. In the present study, we address these issues by using a new in silico designed plasmid library for sgRNA expression and target site presentation. We
produced a new training dataset with precise and direct quantification of sgRNA efficiency, which was used to characterize general sgRNA features that are intrinsically associated with CRISPR/Cas9 cleavage. This new strategy was feasible due to a unique experimental design in which oligonucleotides were synthesized with both an sgRNA expression cassette and a corresponding target sequence in the same construct. Similar strategies were recently used to generate large-scale datasets for analysis of CRISPR/Cpf1 efficiency as well as for the analysis of CRISPR/Cas9 editing patterns and specificity (32-35). Here, we demonstrate that an in silico designed Cas9 targeting system is useful at generating large-scale training data to characterize CRISPR/Cas9 cleavage efficiency. We were able to precisely quantify the efficiencies of a large number of sgRNAs within a single experiment, thus avoiding inconsistencies when merging datasets from heterogeneous experiments. Our final model, sgDesigner, had stable, high-quality performance across vastly different independent testing datasets in human and mouse experimental systems. However, it remains to be tested whether sgDesigner can be robustly applied to other biological systems, as the rules for CRISPR/Cas9 targeting could be different in other organisms not assessed in our study.

Equally important to training data quality, the choice of machine learning modeling methods also has great impact on the quality of predictions. Previous studies have not reached a consensus on the best modeling approach as seen in the variety of distinct frameworks proposed in sgRNA design studies. Most studies tested a single model or a small number of similar models, limiting the potential for model improvement. Thus, in the present study, we explored multiple vastly different frameworks to identify the best one at sgRNA efficiency prediction. Our final stacking model combined the advantages of multiple models and exhibited greater performance than individual models alone. In summary, through improvements in experimental
design, data quality and computational modeling, we developed a new sgRNA design tool, which consistently outperformed competing tools under various experimental settings. Our tool is freely accessible as a web application via http://crispr.wustl.edu.

2.6 References


2.7 Supplementary Figures

Supplementary Figure 2.1: Additional flanking-region features did not improve predictions. (a) ROC curves from 10-fold cross-validation of each model. (b) Spearman correlations between predicted efficiency scores and experimental measurements. Stacking models were trained either using features of the 26 n.t. sequence (red) or the 53 n.t. extended target sequence (blue). The 26 n.t. sequence consisted of 20 n.t. gRNA-matching sequence, 3 n.t. PAM and 3 n.t. downstream of the PAM. The 53 n.t. sequence had an additional 15 n.t. on the 5’ end and 12 n.t. on the 3’ end.
**Supplementary Figure 2.2:** ROC comparison of sgRNA design tools in five additional datasets. (a) The Chari dataset. (b) The FC dataset was used to train RS2. (c) The Wang dataset was used to train SSC and DeepCRISPR. (d) The RES dataset was used to train RS2 and DeepCRISPR. (e) The Koike-Yusa dataset was used to train SSC. (f) The Shalem dataset.
Supplementary Figure 2.3: PR comparison of sgRNA design tools in five additional datasets. (a) The Chari dataset. (b) The FC dataset was used to train RS2. (c) The Wang dataset was used to train SSC and DeepCRISPR. (d) The RES dataset was used to train RS2 and DeepCRISPR. (e) The Koike-Yusa dataset was used to train SSC. (f) The Shalem dataset.
Supplementary Figure 2.4: Spearman correlation comparison of sgRNA design tools in five additional datasets. (a) The Chari dataset. (b) The FC dataset was used to train RS2. (c) The Wang dataset was used to train SSC and DeepCRISPR. (d) The RES dataset was used to train RS2 and DeepCRISPR. (e) The Koike-Yusa dataset was used to train SSC. (f) The Shalem dataset.
Chapter 3: Evaluation of Efficiency Prediction Algorithms and Development of Ensemble Model for CRISPR/Cas9 gRNA Selection

This chapter is adapted from and expanded upon the following publication (1):


3.1 Abstract

MOTIVATION: The CRISPR/Cas9 system is widely used for genome editing. The editing efficiency of CRISPR/Cas9 is mainly determined by the guide RNA (gRNA). Although many computational algorithms have been developed in recent years, it is still a challenge to select optimal bioinformatics tools for gRNA design in different experimental settings.

RESULTS: We performed a comprehensive comparison analysis of 15 public algorithms for gRNA design, using 16 experimental gRNA datasets. Based on this analysis, we identified the top-performing algorithms, with which we further implemented various computational strategies to build ensemble models for performance improvement. Validation analysis indicates that the new ensemble model had improved performance over any individual algorithm alone at predicting gRNA efficacy under various experimental conditions.

AVAILABILITY AND IMPLEMENTATION: The new sgRNA design tool is freely accessible as a web application via https://crisprdb.org. The source code and stand-alone version
3.2 Introduction

The CRISPR system has quickly become the first choice for performing site-specific genome editing. It has been used in a variety of organisms and cell types (2). Among all the systems with various choices of Cas proteins, the CRISPR/Cas9 system is the most widely used because of its simplicity, high efficiency and low cost (2). This system consists of two core components, the Cas9 protein and the guide RNA (gRNA). The Cas9/gRNA complex is guided by the gRNA to its target site by recognizing a 20-nucleotide target sequence proximal to an NGG protospacer adjacent motif (PAM). Then, Cas9 induces a site-specific double-strand break (DSB) at the target site. As for the editing performance of the CRISPR/Cas9 system, two major considerations are targeting specificity and cleavage efficiency. Off-target effects occur when the Cas9/gRNA complex binds and cleaves unintended genomic loci (3). This is potentially a serious concern especially in clinical applications. Thus, many studies have been reported to address this issue using both experimental (4-6) and computational (3, 7, 8) methods. Besides targeting specificity, there is also a high demand for high cleavage efficiency, as inefficient gene editing often obscures real functional changes in experiments.

To predict the cleavage efficiency of gRNAs, individual studies have utilized various experimental datasets to train prediction algorithms. Besides the differences in training data, individual algorithms employed different feature sets based on various sequence and structural attributes related to cleavage efficiency, such as nucleotide composition and gRNA secondary structure, as recently reviewed by Konstantakos et al. (9). Moreover, a variety of traditional
machine learning (8, 10-17) and deep learning (7, 18-20) models have been employed by individual prediction algorithms. Among them, we recently reported an ensemble learning-based computational model, sgDesigner, which utilized the Stacking ensemble method to combine multiple machine learning models (11). As individual algorithms were developed using different computational strategies based on different training datasets and feature designs, their performance could vary significantly when applied in different experimental settings (9). Thus, there is a strong demand in the CRISPR field for comprehensive guidance on selecting optimal tools for gRNA design. To this end, we performed a comprehensive comparison of existing gRNA design algorithms using a large collection of published datasets. The top-performing algorithms were then combined using the Stacking ensemble method to further enhance the power of the predictive model. Our validation analysis showed the new ensemble model has the best overall performance compared with any individual algorithm included in this study.

3.3 Materials and Methods

3.3.1 Computational Algorithms for Predicting CRISPR/Cas9 Efficacy

We assessed seventeen published scoring algorithms for gRNA cleavage efficiency: CRISPRon (20), DeepSpCas9 (18), DeepHF (19), sgDesigner (11), uCRISPR (21), TSAM (14), RuleSet2 (8), SSC (17), predictSGRNA (13), E-crisp (22), TUSCAN (16), ge-CRISPR (12), DeepCRISPR (7), sgRNAScorer2.0 (10), CRISPRpred (15), CRISPRscan (23) and CRISPRater (24). The prediction results from most algorithms were generated using the respective stand-alone packages. Among them, CRISPRon, DeepSpCas9, sgDesigner, SSC, predictSGRNA and CRISPRpred were run with default settings; DeepHF was run with `wt_u6` as the `enzyme` option; uCRISPR scores were generated using the ‘on-target’ module; TSAM was run with
additional ‘pHMM’ features as instructed by the authors; TUSCAN scores were computed using
the regression version; DeepCRISPR predictions were generated using the sequence features
only; sgRNA Scorer 2.0 model parameters were set to 20-bp as the seed length and ‘NGG’ as the
PAM sequence. As for the remaining algorithms, the prediction results of RuleSet2, ge-CRISPR
and E-crisp were obtained via the respective web servers; CRISPRscan was not included in our
comparative analysis because the website does not support batch prediction for multiple gRNA
sequences; CRISPRater was excluded from our analysis as we could not finish processing all
queries due to usually long waiting time at the web server. More details about these scoring
algorithms are summarized in Table 3.1.

Table 3.1. The collection of scoring algorithms for predicting gRNA cleavage efficiency

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Machine learning function</th>
<th>Models</th>
<th>Input seq format</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRISPRon</td>
<td>Deep Learning regression</td>
<td>CNN + biological features</td>
<td>30mer (4mer + 20mer + PAM + 3mer)</td>
<td>Xiang et al. (20)</td>
</tr>
<tr>
<td>sgDesigner</td>
<td>classification</td>
<td>Stacking (SVM, XGBoost)</td>
<td>26mer (20mer + PAM + 3mer)</td>
<td>Hiranniramol et al.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(11)</td>
</tr>
<tr>
<td>DeepHF</td>
<td>Deep Learning regression</td>
<td>LSTM (RNN) + biological features</td>
<td>23mer (20mer + 3mer)</td>
<td>Wang et al. (19)</td>
</tr>
<tr>
<td>DeepSpCas9</td>
<td>Deep Learning regression</td>
<td>CNN</td>
<td>30mer (4mer + 20mer + PAM + 3mer)</td>
<td>Kim et al (18)</td>
</tr>
<tr>
<td>TSAM</td>
<td>regression</td>
<td>Average (SVM, GBM)</td>
<td>30mer (4mer + 20mer + PAM + 3mer)</td>
<td>Peng et al. (14)</td>
</tr>
<tr>
<td>RuleSet2</td>
<td>regression</td>
<td>GBDT</td>
<td>30mer (4mer + 20mer + PAM + 3mer)</td>
<td>Doench et al. (8)</td>
</tr>
<tr>
<td>Method</td>
<td>Type</td>
<td>Model</td>
<td>Seq Format</td>
<td>Ref</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------------</td>
<td>----------------</td>
<td>-----------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>SSC</td>
<td>regression</td>
<td>Elastic Net</td>
<td>30mer (20mer + PAM + 7mer)</td>
<td>Xu et al. (17)</td>
</tr>
<tr>
<td>DeepCRISPR</td>
<td>classification or regression</td>
<td>CNN</td>
<td>23mer (20mer + PAM)</td>
<td>Chuai et al. (7)</td>
</tr>
<tr>
<td>ge-CRISPR</td>
<td>classification or regression</td>
<td>SVM</td>
<td>23mer (20mer + PAM)</td>
<td>Kaur et al. (12)</td>
</tr>
<tr>
<td>predictSGRNA</td>
<td>regression</td>
<td>Elastic Net</td>
<td>40mer (10mer + 20mer + PAM + 7mer)</td>
<td>Kuan et al. (13)</td>
</tr>
<tr>
<td>CRISPRpred</td>
<td>classification</td>
<td>SVM</td>
<td>30mer (4mer + 20mer + PAM + 3mer)</td>
<td>Rahman et al. (15)</td>
</tr>
<tr>
<td>TUSCAN</td>
<td>classification or regression</td>
<td>Random Forest</td>
<td>30mer (4mer + 20mer + PAM + 3mer)</td>
<td>Wilson et al. (16)</td>
</tr>
<tr>
<td>uCRISPR</td>
<td>N/A</td>
<td>Physical Scoring Function</td>
<td>23mer (20mer + PAM)</td>
<td>Zhang et al. (21)</td>
</tr>
<tr>
<td>sgRNA Scorer 2.0</td>
<td>classification</td>
<td>SVM</td>
<td>23mer (20mer + PAM)</td>
<td>Chari et al. (10)</td>
</tr>
<tr>
<td>E-crisp</td>
<td>N/A</td>
<td>Physical Scoring Function</td>
<td>20mer</td>
<td>Heigwer et al. (22)</td>
</tr>
<tr>
<td>CRISPRscan</td>
<td>regression</td>
<td>Linear Regression</td>
<td>23mer (20mer + PAM)</td>
<td>Moreno-Mateos et al. (23)</td>
</tr>
<tr>
<td>CRISPRater</td>
<td>regression</td>
<td>LARS</td>
<td>23mer (20mer + PAM)</td>
<td>Labuhn et al. (24)</td>
</tr>
</tbody>
</table>

PS: The 20mer sequence structure shown in 'input_seq_format' column is the seed region of gRNAs
3.3.2 CRISPR gRNA Datasets for Evaluation of Editing Efficiency

Relevant experimental studies for gRNA potency analysis are summarized in Supplementary Table S2. We downloaded the Chari_293T dataset directly from the Supplementary Tables at the journal’s website (25). The Shalem dataset was originally generated by Shalem et al. (26) and further processed by Doench et al. (27) to calculate ‘Normalized sgRNA Activity’. The gene percent rank from the Doench2014_hs and Doench2014_mm datasets were used to assess gRNA efficiency. The Shalem, Doench2014_hs and Doench2014_mm datasets were downloaded from the Supplementary Tables of Doench et al. (27). We downloaded the Doench2016 dataset from the Azimuth website (8). The XuHL60 and XuKBM7 datasets were originally generated by Wang et al. (28) and further curated by Xu et al. (17). We merged the gRNAs targeting both ‘ribosomal’ and ‘non-ribosomal’ genes and used the negative log2-fold change values with inverted sign to represent gRNA cleavage efficiency. As for the data from Hart et al. (29), we included the representative Hct116-2 Lib1 dataset as curated by Haeussler et al. (30). The CRISPRon_train, DeepSpCas9_train, DeepHF_train and sgDesigner_train datasets were respective training data of four individual scoring algorithms and downloaded from the journals’ websites (11, 18-20). The Cheruiyot dataset was obtained from the authors (31). The CHChen and Fiona_Breast datasets were downloaded from the journals’ websites (32, 33). The Achilles dataset was originally generated in Project Achilles and hosted on the Cancer Dependency Map Portal (DepMap). It is publicly available on Figshare (34). The Gagnon_2014, Varshney_2015 and Moreno-Mateos_2015 datasets from T7 expression systems were curated by Haeussler et al. (30). For the datasets without originally provided efficiency labels, the gRNAs were first ranked by the corresponding experimental values. Then, we
assigned ‘high’ and ‘low’ efficiency labels to the top and bottom 20% gRNAs, respectively. Details of all included validation datasets are described in Table 3.2.

Table 3.2. The collection of experimental datasets for evaluating gRNA efficacy

<table>
<thead>
<tr>
<th>Datasets</th>
<th>gRNA counts</th>
<th>Cell line</th>
<th>Species</th>
<th>Experiment strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chari_293T</td>
<td>1234</td>
<td>HEK293T</td>
<td>Human</td>
<td>Cells are first transduced by lentivirus target library and performed puromycin selection, then transfected with Cas9, corresponding gRNA library and end-processing enzyme/empty vector, finally DNA from cells were harvested and indel frequency are determined by high-throughput sequencing.</td>
</tr>
<tr>
<td>Shalem</td>
<td>1278</td>
<td>A375</td>
<td>Human</td>
<td>Conduct a negative selection screen by profiling the depletion of sgRNAs targeting essential survival genes.</td>
</tr>
<tr>
<td>Hart_Hct1162Lib1</td>
<td>4239</td>
<td>HCT116</td>
<td>Human</td>
<td></td>
</tr>
<tr>
<td>XuHL60</td>
<td>2076</td>
<td>HL60</td>
<td>Human</td>
<td>Infect cells with gRNAs targeting essential genes for cell survival and use deep sequencing to monitor the change in abundance of each sgRNA between the initial seeding and a final population.</td>
</tr>
<tr>
<td>XuKBM7</td>
<td>2076</td>
<td>KBM7</td>
<td>Human</td>
<td></td>
</tr>
<tr>
<td>Doench2014_hs</td>
<td>881</td>
<td>MOLM13, NB4, TF1</td>
<td>Human</td>
<td>Design gRNAs that targeting cell surface markers, and then isolating complete knockout cells by fluorescence-activated cell sorting, to find most active gRNAs.</td>
</tr>
<tr>
<td>Doench2014_mm</td>
<td>951</td>
<td>EL4</td>
<td>Mouse</td>
<td></td>
</tr>
<tr>
<td>Doench2016</td>
<td>2333</td>
<td>A375</td>
<td>Human</td>
<td>Design gRNAs targeting genes known to confer resistance to 4 drugs, then screening the infected cells with corresponding drugs. The log2 fold change of gRNAs are determined relative to control cells.</td>
</tr>
<tr>
<td>Dataset</td>
<td>Cell Line</td>
<td>Condition</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>CHChen (a)</td>
<td>LNCaP-abl, T47D</td>
<td>Human</td>
<td>Cells are infected by gRNAs targeting nonlethal region and lethal genes. Then after puromycin selection, survival cells are divided into control group and another group to be cultured and DNA extraction.</td>
<td></td>
</tr>
<tr>
<td>Cheruiyot</td>
<td>U2OS</td>
<td>Human</td>
<td>Cells are first infected with lentiviruses expressing Cas9 and selected with blasticidin, then U2OS reporter Cas9 cells were infected with lentiviruses expressing the GeCKOv2 library, finally DNA from cells were harvested and indel frequency are determined by high-throughput sequencing.</td>
<td></td>
</tr>
<tr>
<td>Fiona_Breast (b)</td>
<td>25 Breast Cancer cell lines</td>
<td>Human</td>
<td>Conduct a negative selection screen by profiling the depletion of sgRNAs targeting essential survival genes.</td>
<td></td>
</tr>
<tr>
<td>Achilles</td>
<td>769 screens of various cancer cell lines</td>
<td>Human</td>
<td>Performed genome-scale CRISPR-Cas9 loss-of-function screens and assessed the cell-proliferation effects of introducing each gRNA.</td>
<td></td>
</tr>
<tr>
<td>DeepHF_train</td>
<td>HEK293T</td>
<td>Human</td>
<td>Generate gRNA expressing plasmids using synthesized oligonucleotides, transduced into Cas9-expressing cells and determine the indel frequency using deep sequencing.</td>
<td></td>
</tr>
<tr>
<td>DeepSpCas9_train</td>
<td>HEK293T</td>
<td>Human</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sgDesigner_train</td>
<td>ATCC</td>
<td>Human</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRISPRon_train</td>
<td>HEK293T</td>
<td>Human</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gagnon_2014</td>
<td>in vitro</td>
<td>Zebrafish</td>
<td>Guides are made by in vitro transcription with the T7 promoter and injected into eggs of zebrafish</td>
<td></td>
</tr>
<tr>
<td>Varshney_2015</td>
<td>in vitro</td>
<td>Zebrafish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moreno-Mateos_2015</td>
<td>in vitro</td>
<td>Zebrafish</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a): For datasets originally from CRISPR screening studies, we first conducted median normalization, then calculated the log2 fold change of each gRNA between treatment and control samples. The datasets only included gRNAs that are targeting on the essential genes.

(b): For the Fiona_Breast dataset, after the same processing described in (a) to calculate the log2 fold change of each gRNA, the top and bottom 10% of the gRNAs based on the log2 fold change were selected for the comparative analysis.
3.3.3 Assessment of Model Performance

In this study, we employed multiple computational methods to evaluate the performance of individual algorithms. Specifically, Spearman correlation analysis was performed to assess non-parametric correlation between the prediction scores and experimental values. Receiver operating characteristic (ROC) curve analysis was performed to assess the diagnostic ability of a prediction model based on both prediction sensitivity and specificity. Random forest (RF)-based feature analysis was performed with the Python scikit-learn package to determine the relative contribution of each algorithm in the final ensemble model. During the training of the RF model, a specified number (n = 5000 assigned to ‘n_estimators’) of independent decision trees were trained simultaneously. Every decision tree was trained with random collections of the features, and then the importance of a feature was collectively quantified.

3.4 Results

3.4.1 Evaluating the Performance of Public Algorithms for Prediction of CRISPR/Cas9 Editing Efficiency

We selected fifteen published algorithms that were developed for predicting gRNA efficiency (Table 3.1). These algorithms employed various computational techniques as well as diverse training datasets to predict the efficiency of gRNAs. Specifically, CRISPRon, DeepSpCas9, DeepHF and DeepCRISPR applied deep learning neural networks to their prediction models; RuleSet2, predictSGRNA, SSC, E-crisp, TUSCAN, ge-CRISPR, sgRNA Scorer 2.0 and CRISPRpred employed single machine learning models, whereas sgDesigner and TSAM combined multiple machine learning models into assemble models; uCRISPR relied on an empirical scoring function based on experimental observations. As for the experimental
strategies for generating model training data, CRISPRon, DeepSpCas9, DeepHF and sgDesigner relied on DNA sequencing of edited target plasmid libraries for direct quantitation of CRISPR/Cas9 efficiency; in contrast, the other algorithms relied on DNA sequencing data from genome-wide functional screens with CRISPR/Cas9 libraries. To comprehensively evaluate the performance of individual algorithms, we collected 16 independent datasets, encompassing more than 90,000 gRNAs, from various experimental sources. This collection of data included the training datasets used to build gRNA prediction algorithms (8, 11, 17-20, 25, 27, 29) as well as other datasets from high-throughput screening studies (26, 31-34) (Table 3.2) that are independent to all included algorithms.

We first conducted correlation analysis to evaluate the prediction performance of individual algorithms on the testing datasets. To this end, Spearman correlation analysis was performed to assess the predictive power of individual algorithms. Specifically, for each pair of the algorithm and testing dataset, we computed Spearman correlation coefficient between the prediction scores and the experimental values of gRNA activities (Figure 3.1). With each testing dataset, individual algorithms were ranked based on their correlation coefficients. Then, the average rank from all datasets was calculated for each algorithm. Overfitting is a common concern in data modeling as the model may perform well on its training data, but not on independent testing data. Thus, in this ranking analysis, we only included the rank for an algorithm/dataset pair if the dataset had not been used to train the algorithm.
Figure 3.1: Performance comparison of 15 public algorithms with Spearman correlation analysis. The heatmap matrix table presents the Spearman correlation coefficients between algorithm prediction scores and corresponding experimental values in individual validation datasets. In this table, each row represents one prediction algorithm, and each column represents one testing dataset. The correlation between an algorithm and its own training dataset is marked in gray. The average rank for each algorithm was computed by averaging individual performance ranks with all independent datasets (i.e., algorithm self-training datasets, as marked in gray, were excluded from the analysis).

From the correlation matrix, most deep learning models showed better predictive power than conventional models, with the top-performing algorithms being CRISPRon, DeepSpCas9 and DeepHF. Besides the adoption of deep learning techniques, these models also took advantage of high-quality training datasets generated from large-scale plasmid libraries for direct quantification of gRNA efficiency. In our analysis, CRISPRon ranked first when applied to 11 out of 14 validation datasets (CRISPRon_train and DeepSpCas9_train were excluded from analysis as they were used to train CRISPRon). The average rank of CRISPRon was 1.36 when
all validation results were combined. In the same way, the average ranks of DeepHF and DeepSpCas9 were computed to be 2.69 and 2.93, respectively.

Among other models, one promising computational strategy is to combine multiple single models into assemble models. The advantages of this assemble approach were exemplified by sgDesigner and TSAM, which outperformed other single-model algorithms for overall prediction performance. In combined analysis of all validation results, the average ranks of sgDesigner and TSAM were 3.8 and 4.54, respectively. As for other single-model algorithms, predictSGRNA and RuleSet2 had comparable overall performance, with the respective average ranks being 6.57 and 6.77. Both algorithms outperformed the remaining ones trained using single machine learning methods. Although most algorithms were trained using machine learning methods, one exception is uCRISPR, which was based on an empirical scoring scheme. uCRISPR had relatively robust performance, with the average rank being 5.8. By correlating prediction results to experimental testing data, we demonstrated that CRISPRon, DeepSpCas9, DeepHF, sgDesigner and TSAM were the top-performing algorithms according to average correlation rank.

We further performed receiver operating characteristic (ROC) curve analysis to evaluate the overall sensitivity and specificity of individual algorithms. Fifteen published datasets were directly used to construct ROC curves for algorithm evaluation, and the area-under-the-curve (AUC) values of these ROC curves were determined to quantify the performance of individual algorithms (Figure 3.2). As similarly described in the correlation analyses, we also calculated the average AUC rank of each algorithm when applied to all testing datasets. For each algorithm, the self-training datasets were excluded from the performance ranking.
Table 3.2: Performance comparison of 15 public algorithms with ROC analysis. The heat map matrix table presents the AUC values of the ROC curves generated by comparing the algorithm prediction scores with the corresponding efficiency labels (high versus low) from each validation dataset. In this table, each row represents one prediction algorithm, and each column represents one testing dataset. The AUC between an algorithm and its own training data is marked in gray. The average rank for each algorithm was computed by averaging individual performance ranks with all independent datasets.

According to the AUC matrix, we identified the same top-performing deep learning algorithms as previously identified in the correlation analysis. Specifically, CRISPRon had the AUC values in the range of 0.792–0.963 for independent testing datasets, with the best average rank of 1.5 among all individual algorithms. Similarly, DeepSpCas9 and DeepHF were also top-performing algorithms, with the average ranks of 2.8 and 2.92, respectively. For sgDesigner and TSAM that adopted ensemble modeling strategies, they had the average ranks of 4.13 and 4.62, respectively. They performed worse than the top-performing deep learning algorithms but better than other single-model algorithms, as similarly described in previous correlation analysis.
Among the remaining algorithms, we observed various algorithm rankings when different validation datasets were tested. For instance, ge-CRISPR performed worse than E-crisp on the XuHL60 dataset but better on the Chari_293T dataset. From the ROC analysis, it was clear that CRISPRon, DeepSpCas9, DeepHF, sgDesigner and TSAM consistently outperformed the other algorithms when generally applied to various testing datasets.

In summary, from both Spearman correlation analysis and ROC analysis, we demonstrated that CRISPRon, DeepSpCas9, DeepHF, sgDesigner and TSAM were consistently identified as the top-performing algorithms for predicting the efficacy of gRNAs.

3.4.2 Developing an Ensemble Model for Improved Prediction of CRISPR/Cas9 Efficiency

Although the top-performing algorithms produced robust prediction results, there was significant room for improvement based on both ROC and correlation analyses. In our previous experience, assemble models (e.g., sgDesigner) that integrated multiple algorithms had superior performance when compared with any individual algorithm included in the assemble model. Based on that observation, we hypothesized that, by assembling top-performing algorithms, it would be possible to generate an integrated model with improved performance over any individual top-performing algorithm.

To achieve this goal, one computational strategy is the stacking ensemble method. The stacking ensemble models have two model layers. The first layer contains multiple individual models to be assembled, with each model generating prediction results independently. The second layer is a meta-model that collects the predictions from the first layer as input features to generate final integrated prediction results. The stacking method not only can be used to
assemble various internal models but also can be adopted as a framework to integrate various published algorithms (Figure 3.3).

![Stacking Ensemble Modeling Strategy](image)

**Figure 3.3:** The stacking ensemble modeling strategy. In the first layer, the gRNA sequences were input to individual algorithms to generate prediction scores. Then, these scores were merged as input features for second layer modeling. Specifically, a meta-model was used in the second layer to predict the final score for gRNA efficiency.

Selection of training data to build ensemble models

For machine learning and deep learning models, high-quality training data are crucial for their robust performance. From the Spearman correlation analysis on public algorithms and datasets, we found the four algorithms with top performance implemented similar strategies to generate their training data (11, 18-20). Specifically, a pool of oligonucleotides was first designed and synthesized, which contain the gRNA sequences and corresponding target regions. A plasmid library containing these oligonucleotides was transduced into Cas9-expressing cells, which led to indel formation at the target region with frequencies dependent on the gRNA on-target activity. The final products were amplified and the indel frequency of every design was determined by deep sequencing. In another word, the efficacy of each gRNA was determined by observed rate of mutagenesis of the target site. In this way, we focused on CRISPR/Cas9
cleavage of double-stranded DNA that led to detectable gene mutations. The high quality of these datasets was clearly demonstrated by the high performance of the respective algorithms being trained, using either machine learning or deep learning methods. In addition, these experimental datasets also had relatively high Spearman correlation coefficients with the prediction results from other public algorithms. In our analysis, we combined the high-quality training data for CRISPRon, DeepSpCas9, DeepHF and sgDesigner (Table 3.2) to train new ensemble models.

**Ranking the relative contribution of individual algorithms in the ensemble model**

One crucial step in developing ensemble models is to decide which individual algorithms should be included to further boost the performance of the final model. To this end, the relative contribution of individual algorithms was evaluated by feature importance analysis in random forest (RF) modeling. Specifically, RF models were trained with the prediction scores of individual algorithms on each non-self training data (CRISPRon_train, DeepSpCas9_train, DeepHF_train or sgDesigner_train). In another word, algorithm prediction scores on self-training datasets were already excluded from the analysis to reduce overfitting concerns. Prediction results from selected algorithms were treated as input features for the second-layer RF model (Figure 3.3), and the Mean Decrease in Impurity (MDI) score computed by the model was used to assess the contribution of each algorithm. As we were interested in both classification and regression-based ensemble models, the RF classifier and regressor were separately applied for algorithm evaluation. Then, combined MDI scores from all subsets of the training data were used to determine the overall feature ranking (Figure 3.4a and b for the regression and classification models, respectively). From this RF analysis, the algorithms with top performance in Spearman correlation analysis (i.e., CRISPRon, DeepSpCas9, DeepHF, sgDesigner in Figure 3.1)
maintained their top ranks. In contrast, for other algorithms, we observed inconsistent ranking between RF analysis and Spearman correlation analysis, suggesting that correlative results may not fully reflect the independent contribution of each algorithm in the assemble model.

Figure 3.4: Contribution of individual algorithms and optimal order of the ensemble model. The feature importance score (MDI) of each algorithm was computed using either (a) random forest regressor or (b) random forest classifier, as presented in the boxplots. Further, the optimal order of the ensemble model was determined by stepwise addition of individual algorithms as ranked by either (c) the random forest regressor or (d) the random forest classifier. The optimal order of the ensemble model is indicated by the peak in each curve.

**Developing ensemble models by integrating top-ranking individual algorithms**

We tested two common ensemble strategies that were based on either regression or classification models. For both strategies, we implemented different machine learning meta-
models in the Stacking framework (Figure 3.3). After assembling all the 15 algorithms into meta-models, we performed Spearman correlation analysis of these ensemble models with 11 independent validation datasets. Specifically, we computed the Spearman correlation coefficients to evaluate the prediction performance of each ensemble model. For regression-based ensemble modeling, we evaluated ridge regression, lasso regression, XGBoost regression and random forest regression models. Among them, ridge regression was ranked as the top ensemble model for 7 out of 11 validation datasets (Supplementary Figure 3.1a). For classification-based analysis, we compared ensemble models based on logistic regression, XGBoost classification and random forest classification. Logistic regression outperformed other ensemble models for 9 out of 11 validation datasets (Supplementary Figure 3.1b). Thus, we implemented ridge regression and logistic regression as the meta-models for further regression and classification modeling analysis, respectively.

Then, we tested various collections of individual algorithms as input to maximize the performance of the ensemble model. Specifically, we started the process by assembling the two best-performing algorithms based on relative feature importance (Figure 3.4a and b). Then, the ensemble model was expanded by stepwise inclusion of the next best-performing algorithm. This process was repeated with one new algorithm added to the ensemble model at each step based on the algorithm ranking until all 15 individual algorithms were assembled. We further tested these models by performing Spearman correlation analysis using leave-one-dataset-out cross-validation with the four training datasets. Specifically, for each iteration in the cross-validation analysis, we trained the ensemble model using three training datasets and tested it on the remaining fourth dataset by Spearman correlation analysis. This process was repeated until all four training datasets had been used as testing data. Then, the performance of the ensemble
model was evaluated by the average value of the Spearman correlation coefficients. The performance of the regression and classification models built on various number of individual algorithms is summarized in Figure 3.4c and d, respectively. Interestingly, performance of the regression model was peaked when the top five algorithms were assembled (Figure 3.4c); similarly, maximal performance of the classification model was observed when the top six algorithms were assembled (Figure 3.4d). However, after more algorithms were assembled into the models, the prediction performance was decreased as assessed by Spearman correlation coefficient. One possible explanation for this observation was that assembling the algorithms with relatively poor performance would worsen the final prediction results, reflecting the nature of the Stacking ensemble method. Thus, five and six top-ranking algorithms were included as input to build the final regression and classification ensemble models, respectively.

### 3.4.3 Validation of the Ensemble Models with Independent Datasets

Among the 16 collected gRNA datasets, our ensemble models were trained with CRISPRon_train, DeepSpCas9_train, DeepHF_train and sgDesigner_train. Thus, we used the other 12 datasets for independent model validation. The performance of the ensemble models on individual datasets was evaluated by both Spearman correlation analysis (Figure 3.5a) and ROC analysis (Figure 3.6a). Both regression and classification ensemble models generated robust prediction results, with the regression model outperforming the classification model in 10 out of 12 validation datasets by correlation analysis, and 9 out of 12 datasets by ROC analysis. Moreover, compared with CRISPRon (the top-performing individual algorithm), the ensemble regression model produced improved Spearman correlation results for ten out of twelve validation datasets and tied for the remaining two datasets. Similarly, the ensemble regression model had higher AUC-ROC values than CRISPRon for ten out of twelve validation datasets. To
Figure 3.5: Validation of the ensemble models by Spearman correlation analysis. (a) The Spearman correlation coefficients between the prediction scores and the experimental values were calculated using 12 independent validation datasets. (b) Performance comparison of the ensemble models with included public algorithms, ordered by the average rank of the correlation coefficients computed with individual validation datasets.

summarize the performance of the ensemble models, the average values of the Spearman correlation coefficients and the AUC-ROC for all included datasets were calculated. In this analysis, the training datasets of respective ensemble models were excluded to reduce overfitting bias. From Spearman correlation analysis, the ensemble_ridge and ensemble_logistic models had average correlation coefficients of 0.54 and 0.52, respectively, both of which were higher than the average correlation coefficients from all individual algorithms (Figure 3.1). We also observed
similar results from the ROC analysis, where the average AUC values of the ensemble models (0.89 and 0.88 for ensemble_ridge and ensemble_logistic, respectively) showed improved performance over all individual algorithms (Figure 3.2).

![Ensemble AUC values](image)

**Figure 3.6**: Validation of the ensemble models by ROC analysis. (a) The AUC-ROC values were determined using 12 independent validation datasets. (b) Performance comparison of the ensemble models with included public algorithms, ordered by the average rank of the AUC-ROC values computed with individual validation datasets.

Next, the rank distributions of the models among all validation datasets were determined to assess general model performance across experimental conditions. Specifically, for each model, we computed its performance rank for every validation dataset as determined either by
correlation coefficient or AUC-ROC. To alleviate overfitting concerns, training datasets for respective models were not considered in this analysis. In this way, the rank distributions for each model were determined, and presented in Figure 3.5b for correlation rank and Figure 3.6b for AUC rank, respectively. From these results, the mean correlation ranks of the two ensemble models were 1.25 for ensemble_ridge and 2.75 for ensemble_logistic, both of which were higher than any individual algorithm (Figure 3.5b). Similarly, the mean AUC ranks of the two ensemble models were the highest, with 1.42 for ensemble_ridge and 2.5 for ensemble_logistic, respectively (Figure 3.6b). In summary, the two ensemble models outperformed all individual algorithms as evaluated by both correlation and AUC ranks.

Based on these validation analyses, we concluded that the two ensemble models had improved performance over all included individual algorithms, as shown by consistent results across various experimental datasets. In particular, the regression ensemble model, ensemble_ridge, demonstrated the best performance. To further demonstrate the superior performance of ensemble_ridge, we conducted overall Spearman correlation analysis (Supplementary Figure 3.2), which correlated algorithm prediction scores to experimental data. In this analysis, we determined the overall Spearman correlation performance of ensemble_ridge, as well as other public algorithms on the combined dataset consisting of all twelve independent validation datasets (CRISPRon_train, DeepSpCas9_train, DeepHF_train and sgDesigner_train were excluded). Specifically, for each algorithm, the prediction scores of the twelve validation datasets were obtained. Then, only prediction results of independent validation datasets were combined for Spearman correlation analysis. Among all individual algorithms, CRISPRon had the best performance (r = 0.445), followed by DeepHF (r = 0.416), DeepSpCas9 (r = 0.399) and
sgDesigner ($r = 0.396$). The ensemble_ridge model outperformed all individual algorithms in this correlative analysis ($r = 0.471$).

From the aforementioned comparative analyses, CRISPRon was the best single algorithm among the 15 public algorithms, while ensemble_ridge outperformed CRISPRon consistently across various validation datasets. To examine the cases where the prediction results of ensemble_ridge and CRISPRon disagreed, we conducted a case study using the Achilles dataset. To this end, we normalized the original efficacy and prediction scores from individual algorithms to percentage rank values. Then, the performance difference between ensemble_ridge and CRISPRon was assessed. To focus on the most discrepant cases, 50 gRNAs with the largest percentage rank differences between ensemble_ridge and CRISPRon were selected for further examination. Individual algorithms including CRISPRon had variable performance on predicting the efficacy of these gRNAs. Specifically, CRISPRon, DeepSpCas9, DeepHF, sgDesigner and TSAM had the best performance on 1, 6, 20, 18 and 5 gRNAs, respectively. With the adoption of the ensemble strategy, the ensemble model balanced the performance of individual algorithms, leading to more stable and robust performance than any single algorithm in general. As a result, ensemble_ridge correlated better to experimental data than CRISPRon for 86% of the 50 selected gRNAs.

### 3.4.4 Re-training the Ensemble Model for the T7 Promoter System

Previous studies reported that the prediction algorithms trained with U6 promoter-based datasets may not be ideal to predict the efficacy of gRNAs designed for the T7 promoter system. To evaluate the performance of the new ensemble model as well as five other top-performing algorithms on T7 promoter-based datasets, we collected three datasets from zebrafish screens (Gagnon_2014, Varshney_2015 and Moreno-Mateos_2015). Then, we conducted Spearman
correlation analysis between the prediction scores from each algorithm and experimental gRNA efficiencies as presented in the T7 datasets. Comparative analysis indicates that all these algorithms (trained with the U6 datasets) performed relatively poorly on the T7 datasets (Supplementary Figure 3.3a), with the Spearman correlation in the range of 0.074–0.321.

To address data discrepancy between the U6 and T7 systems, in the DeepHF and TSAM studies, the authors generated T7-focused algorithms (named as DeepHF_T7 and TSAM_T7, respectively) by training with the Moreno-Mateos_2015 dataset. Comparative analysis indicates these T7-tailored algorithms had improved performance over the corresponding U6-based algorithms (Supplementary Figure 3.3b). Accordingly, we re-trained our ensemble_ridge model using the Moreno-Mateos_2015 dataset by assembling these two T7 algorithms (DeepHF_T7 and TSAM_T7). As shown in Supplementary Figure 3.3b, ensemble_ridge_T7 has the best performance on two independent validation datasets that were generated from the T7 expression system (Spearman correlation in the range of 0.385–0.513).

### 3.4.5 Online Web Server for gRNA Design

In our study, with the final ensemble model ‘ensemble_ridge’, we performed genome-wide design of gRNAs that were predicted to have high efficacy to target human and mouse genes. Specifically, for each gene, we searched for potential target sites within the 5’-end exons (70% portion of the coding sequence). The focus on 5’-end exons was to maximize functional disruption of the gene resulting from frameshift mutations. Besides on-target activity, we also predicted the specificity of the gRNA using our previously published algorithm (35). Specifically, we performed gRNA seed search to identify potential off-target sites that had any 13-mer seed match to the gRNA sequence. In parallel, BLAST alignment was conducted to search for potential off-target sequences with at least 85% overall homology to the gRNAs. By
implementing these specificity filters, we selected gRNAs with reduced off-target effects. Of note, previous studies indicate that limited off-target editing (i.e., involving <20 nucleotides) had little functional consequence when the target sites are within intergenic regions (27, 36). Therefore, we focused our off-target analysis on all known exon regions, from both protein-coding genes and non-coding genes. An online database resource, CRISPRDB, which contains all pre-designed gRNAs, is freely accessible at https://crisprdb.org. CRISPRDB ranks all gRNAs based on predicted efficacy, and then the website presents up to twenty gRNAs with the highest scores. This online resource also presents a web server interface for custom design of gRNAs with user-provided sequences. Specifically, users can input target site sequences and use either the ‘ensemble_ridge’ or ‘ensemble_ridge_T7’ model to design gRNAs for the U6 or T7 expression system, respectively.

3.5 Discussion

Although the CRISPR/Cas9 system has been widely used, it is still a challenge to consistently design gRNAs with high cleavage efficiency. Previous studies described numerous computational algorithms on predicting the gRNA on-target activities based on various experimental datasets and modeling methods. For experimental datasets, many studies adopted biological enrichment schemes, in which gene editing events impact cell survival or other observable biological phenotypes. Although these strategies have been widely used, such indirect biological readouts could potentially introduce undesired noises in the training data for computational modeling, as equally efficient Cas9 cleavage sites may not result in equal phenotypic changes or survival pressure. Furthermore, gRNAs tested in experimental screens are typically designed for a subset of genes and tested in a single cell line. These restrictions could introduce biases specific to each experimental setting, such as those related to different levels of
genomic accessibility, or different responses to DNA cleavage in a cell line or gene-specific manner. All these factors may potentially reduce model generalizability. As a result, these algorithms often had inconsistent performance when predicting gRNA efficiency under different experimental settings. That being said, we envision that common biological features learned from multiple experimental screening datasets would be valuable to further enhance the performance of the prediction model, as they are currently missing from the artificial plasmid expression systems. To this end, our ensemble model is a flexible framework that could easily incorporate new advancement in this field for performance improvement.

One major issue in comparing gRNA design algorithms is that researchers usually used different independent datasets and computational analysis strategies to validate their algorithms; this presents a major challenge to directly compare the performance of reported algorithms from different studies. To fill in this gap, we performed comprehensive comparison of 15 most widely used algorithms for gRNA design, using a variety of independent experimental datasets for performance validation. In this way, we ranked these algorithms based on their overall performance across many independent datasets. This provides a useful guidance to CRISPR researchers for selecting the most appropriate gRNA design tools for their studies. Moreover, our algorithm comparison analysis revealed interesting features that are characteristic of top-performing prediction models. For example, the same experimental strategy for training data generation was applied by all four top-performing algorithms. Although these four algorithms adopted various modeling strategies, they consistently outperformed other algorithms, which highlight the paramount importance of high-quality data for model training.

As described above, we identified important characteristics that are common among top performing models. Thus, it would be interesting to develop a new prediction algorithm that
could ideally combine the advantages of the top-performing algorithms while avoiding their shortfalls. In computational modeling analysis, the stacking ensemble method is commonly used to linearly combine prediction results from multiple models to gain further performance improvement (37). In the present study, we extended this stacking strategy to assembling top-performing published algorithms. Through validation analysis, the final ensemble model clearly demonstrated performance gains over individual algorithms in the assembly. We hope this ensemble-learning method would further contribute to future improvement of gRNA design when more robust algorithms, based on high-quality training data and enhanced modeling strategies, become available.

3.6 References


target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol*, 17, 148.


3.7 Supplementary Figures

**Supplementary Figure 3.1**: Spearman correlation comparison of ensemble models assembling all individual algorithms using different meta models on 11 independent validation datasets. (a) Regression strategy performance comparison, included Ridge Regression, Lasso Regression, XGBoost and Random Forest regressors to be meta models. (b) Classification strategy performance comparison, included Logistic Regression, XGBoost and Random Forest classifiers to be meta models.
Supplementary Figure 3.2: The overall Spearman correlation performance. Including the performance of the final ensemble model, 'ensemble_ridge', and other public algorithms on independent validation datasets (for each algorithm, only prediction results from independent validation datasets were included).
**Supplementary Figure 3.3**: Performance comparison of T7 promoter-based algorithms. (a) Spearman correlation comparison of U6 and T7 versions of DeepHF and TSAM on three T7 promoter-based datasets. (b) Spearman correlation comparison of the T7 version of ensemble_ridge, DeepHF, and TSAM on three T7 promoter-based datasets. Algorithm self-training datasets were marked in gray.

<table>
<thead>
<tr>
<th></th>
<th>Gagnon 2014</th>
<th>Varshney 2015</th>
<th>Moreno-Mateos 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DeepHF_U6</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DeepHF_T7</strong></td>
<td>0.504</td>
<td>0.369</td>
<td>0.581</td>
</tr>
<tr>
<td><strong>TSAM_U6</strong></td>
<td>0.166</td>
<td>0.233</td>
<td>0.074</td>
</tr>
<tr>
<td><strong>TSAM_T7</strong></td>
<td>0.409</td>
<td>0.363</td>
<td>0.923</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Gagnon 2014</th>
<th>Varshney 2015</th>
<th>Moreno-Mateos 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ensemble_ridge_T7</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DeepHF_T7</strong></td>
<td>0.504</td>
<td>0.369</td>
<td>0.581</td>
</tr>
<tr>
<td><strong>TSAM_T7</strong></td>
<td>0.409</td>
<td>0.363</td>
<td>0.923</td>
</tr>
</tbody>
</table>
Chapter 4: Optimized Prediction of Functional MicroRNA Targets by Deep Learning

This chapter is adapted from and expanded upon the following publication (1):


4.1 Abstract

MOTIVATION: MicroRNAs (miRNAs) are tiny noncoding RNA molecules that serve as primary controllers in numerous biological processes. Their primary role involves reducing the expression of specific target genes. Consequently, precise prediction of miRNA target genes holds paramount importance in comprehending the functions of miRNAs.

RESULTS: We conducted a large-scale RNA sequencing study aimed at experimentally pinpointing genes that are downregulated by 25 miRNAs. This RNA-seq dataset was merged with publicly available miRNA target binding information to systematically detect unique characteristics associated with miRNA binding and subsequent target downregulation. By incorporating these shared attributes into a deep learning framework, we created and verified an enhanced computational model for predicting miRNA targets on a genome-wide scale.

AVAILABILITY AND IMPLEMENTATION: The new miRNA target prediction tool is freely accessible as a web application via https://mirdb.org.
4.2 Introduction

MicroRNAs (miRNAs) are short, noncoding RNA molecules that play a crucial role in regulating various biological processes (2, 3). MiRBase has documented approximately 2000 miRNAs in the human genome (4). Both computational and experimental investigations have revealed that most protein-coding genes are subject to regulation by miRNAs, impacting gene expression at post-transcriptional and translational levels (5-7). MiRNAs primarily function by downregulating the expression of their target genes, making the accurate prediction of miRNA targets essential for characterizing miRNA functions. However, accurately identifying miRNA targets remains a significant challenge in current research. Many researchers employ computational tools for the initial prediction of candidate gene targets, followed by experimental validation. Over the past decade, numerous computational tools have been developed, gradually improving their performance in target prediction. Recent advancements in experimental techniques and novel insights into the mechanisms of miRNA target regulation have led to the creation of new models (8). Several standard features have been identified in miRNA target regulation, such as the requirement for a precise pairing between the miRNA's 5'-end (known as the seed region) and the target site, as well as the relatively low GC content of the target site, which results in increased site accessibility for miRNA binding (9-15). Despite the ongoing progress in miRNA target prediction, existing algorithms still have suboptimal performance, resulting in frequent false predictions that necessitate costly experimental validation. Consequently, further enhancement in computational target prediction methods are crucial for practical applications.

Indeed, commonly utilized target prediction algorithms have been trained using various high-throughput profiling data. Notably, crosslinking and immunoprecipitation (CLIP)
sequencing data and miRNA overexpression data have been particularly prominent in these datasets. CLIP studies enable the unequivocal identification of paired miRNA-target transcripts residing within the same RISC complex. This identification is achieved through the direct ligation of the miRNA to its corresponding target transcript. Additionally, miRNA overexpression experiments can discern functionally relevant targets, as evidenced by their significant downregulation in expression. In the present study, we conducted an analysis employing both CLIP binding data and miRNA overexpression data to pinpoint common attributes that characterize both miRNA binding and target downregulation. Our initial step involved analyzing an extensive miRNA overexpression dataset originally generated by our lab in a prior study (16), which, to the best of our knowledge, stands as the most extensive RNA-seq study of its kind for miRNA target identification. This comprehensive dataset provided the foundation for a quantitative re-evaluation of previously reported features within the context of transcriptome-level target downregulation. To enhance the predictive power of our miRNA target identification, we compared the miRNA targeting characteristics extracted from overexpression data with those derived from publicly available CLIP binding data. This comparison allowed us to consolidate significant biological features. Subsequently, we integrated these biological features into a deep learning-based architecture to develop a predictive algorithm for the functional prediction of miRNA targets. Finally, we established a web portal to host both the algorithm and the database housing prediction results, making them readily accessible to the research community.
4.3 Materials and Methods

4.3.1 Public CLIP-seq Data Retrieval

The CLIP-ligation data retrieval process is detailed in our previous study (17). In summary, we amalgamated data from two studies, namely the Helwak study (18) and the Grosswendt study (19). Raw RNA-seq data from the Helwak study were acquired from the NCBI GEO database under accession number GSE50452 (20). Curated lists of miRNA/target pairs were obtained from the respective journals' websites (18, 19). We retrieved HITS-CLIP data (21) from http://ago.rockefeller.edu and aligned the raw sequence tags to the transcriptome using BLAT (22).

miRNA sequences were sourced from miRBase (4), and RefSeq transcript sequences, along with related gene mapping index files, were downloaded from NCBI (23). To establish orthologous gene relationships across species, we utilized the NCBI HomoloGene database (23). Basewise conservation was determined by calculating phyloP conservation scores with PHAST (24) and downloading them from the UCSC Genome Browser (https://genome.ucsc.edu/). For miRNA target prediction data, we gathered information from various public tools available on their respective websites: TargetScan 8.0 (25) (http://targetscan.org), DIANA-MicroT (26) (http://diana.imis.athena-innovation.gr), miRanda-mirSVR (27) (http://microrna.org), and PITA (28) (https://genie.weizmann.ac.il/pubs/mir07/). To facilitate direct comparison, we mapped the target transcript IDs from all algorithms to NCBI Gene IDs.
4.3.2 RNA-seq to Identify Transcripts Downregulated by MicroRNA

Overexpression

For our miRNA overexpression experiments, we individually introduced 25 miRNAs and a negative control RNA into HeLa cells through transfection. These 25 miRNAs are listed in Table 4.1. To assess the impact of miRNA overexpression on the cell's transcriptional activity, we conducted RNA-seq experiments at the transcriptome level. To ensure the reliability of our results and control for any experimental variations, we performed each miRNA transfection in duplicate on different days. Additionally, we constructed RNA-seq libraries and conducted sequencing runs in duplicate on separate days. In total, we generated 1.5 billion reads to analyze the expression patterns across 52 RNA samples. To identify genes that were downregulated due to miRNA overexpression, we combined all the sequencing data. Our analysis considered transcripts that contained at least one miRNA seed binding site and exhibited a reduction of at least 40% in expression in both duplicated experiments as miRNA targets. Conversely, transcripts that contained at least one seed site but showed no change in expression were classified as non-target controls. In this manner, we identified 2,240 miRNA targets and 4,127 non-target controls through RNA-seq. On average, there were 90 targets for each miRNA, with considerable variation in the number of targets among individual miRNAs, as outlined in Table 4.1 (ranging from 11 to 206 targets).

Table 4.1. Twenty-five miRNAs analyzed in the RNA-seq experiments

<table>
<thead>
<tr>
<th>miRNA name</th>
<th>miRNA sequence</th>
<th>Identified targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-let-7c-5p</td>
<td>UGAGGUAGGUAGGUUGUAUGGUU</td>
<td>31</td>
</tr>
<tr>
<td>hsa-miR-107</td>
<td>AGCAGCAUUGUACAGGGCUAUCA</td>
<td>35</td>
</tr>
<tr>
<td>miRNA</td>
<td>Sequence</td>
<td>Length</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>hsa-miR-10a-5p</td>
<td>UACCCUGUAGAUCCGAAUUUGUG</td>
<td>32</td>
</tr>
<tr>
<td>hsa-miR-124-3p</td>
<td>UAAGGCACGCGGUGAAUGCC</td>
<td>151</td>
</tr>
<tr>
<td>hsa-miR-126-3p</td>
<td>UCGUACCGUGAGUAUAUGCG</td>
<td>11</td>
</tr>
<tr>
<td>hsa-miR-126-5p</td>
<td>CAUUAAUACUUUUGGUACGCG</td>
<td>48</td>
</tr>
<tr>
<td>hsa-miR-133b</td>
<td>UUUGGUCCCCUUCAACCAGCUA</td>
<td>108</td>
</tr>
<tr>
<td>hsa-miR-142-3p</td>
<td>UGUAGUGUUUCCUACUUUAUGGA</td>
<td>108</td>
</tr>
<tr>
<td>hsa-miR-145-5p</td>
<td>GUCCAGUUUUCCCAGGAAUCCCU</td>
<td>82</td>
</tr>
<tr>
<td>hsa-miR-146a-5p</td>
<td>UGAGAACUGAAUUCUAGGGUU</td>
<td>42</td>
</tr>
<tr>
<td>hsa-miR-155-5p</td>
<td>UUAAUGCUAAUCGUGAUAGGGGU</td>
<td>154</td>
</tr>
<tr>
<td>hsa-miR-15a-5p</td>
<td>UAGCAGCACAUAUUGGUUUUGUG</td>
<td>108</td>
</tr>
<tr>
<td>hsa-miR-16-5p</td>
<td>UAGCAGCAGUAAUAUUGGCCG</td>
<td>122</td>
</tr>
<tr>
<td>hsa-miR-17-5p</td>
<td>CAAAGUGCUUACAGUGCAGGUAG</td>
<td>74</td>
</tr>
<tr>
<td>hsa-miR-193b-3p</td>
<td>AACUGGCCCCUAAAGUCCCGCU</td>
<td>102</td>
</tr>
<tr>
<td>hsa-miR-200a-3p</td>
<td>UAACACUGUCUGGUAACGAUGU</td>
<td>35</td>
</tr>
<tr>
<td>hsa-miR-200b-3p</td>
<td>UAAUACUGCCUGGUAAAGUAAGA</td>
<td>126</td>
</tr>
<tr>
<td>hsa-miR-200c-3p</td>
<td>UAAUACUGCCGGGUAAAGUAGGA</td>
<td>93</td>
</tr>
<tr>
<td>hsa-miR-206</td>
<td>UGGAAUGUAAGGAAGUGUGUGG</td>
<td>206</td>
</tr>
<tr>
<td>hsa-miR-210-3p</td>
<td>CUGUGCGUGUGACAGCGGCUGA</td>
<td>43</td>
</tr>
<tr>
<td>hsa-miR-21-5p</td>
<td>UAGCUUAUCAGACUGAUGUUGA</td>
<td>11</td>
</tr>
<tr>
<td>hsa-miR-31-5p</td>
<td>AGGCAAGAUGCUGGCAUAGCU</td>
<td>85</td>
</tr>
<tr>
<td>hsa-miR-34a-5p</td>
<td>UGGCAGUCCUUAGCGGUUGU</td>
<td>155</td>
</tr>
<tr>
<td>hsa-miR-9-3p</td>
<td>AUAAAGCUAGAUACCAGAAAGU</td>
<td>182</td>
</tr>
<tr>
<td>hsa-miR-9-5p</td>
<td>UCUUUGGUUACUCUGCUGUAGA</td>
<td>106</td>
</tr>
</tbody>
</table>
4.3.3 Deep Learning Model Development

The task of predicting miRNA targets can be framed as a regression problem. We developed a mapping function that takes the representations of miRNAs and potential target mRNAs as input and generates a score indicating whether the given miRNA-mRNA pairs are functional direct targets. This mapping function is implemented using a Bidirectional Long Short-Term Memory neural network (BiLSTM), which is a specific type of Recurrent Neural Network (RNN). In contrast to Convolutional Neural Networks (CNNs), which treat DNA input as grayscale images with binary pixel values, the RNN approach treats a DNA sequence more naturally, considering it as a sentence composed of four character types: A, C, G, and T, rather than an image. This perspective draws from related research in natural language processing, which has provided valuable insights into DNA sequence modeling (29). RNNs have loops in their connections, enabling them to incorporate feedback and memory when processing sequences over time or space. The Long Short-Term Memory (LSTM) variant of RNNs further enhances this capability by utilizing "gates" to control the influence of previous sequence features, making it more adaptable for memory control. A BiLSTM leverages the order sensitivity inherent in RNNs by employing two LSTM layers, each processing the input sequence in opposite directions (chronologically and anti-chronologically) and then combining their representations. This bidirectional processing allows a BiLSTM to capture patterns that a unidirectional RNN might miss. Borrowed from the natural language processing applications (30), the miRNA sequence input was transformed to a matrix \( x = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) \), which was a \( 8 \times 4 \) matrix (the length of the miRNA seed region is 8 here, and \( x_i \) means binary vector of four A, T, C, G nucleotides).
Embedding

Embedding plays a crucial role in our study by transforming a sparse matrix, derived from input vectors, into a dense, high-dimensional real-valued space. This transformation greatly aids the training process. In our research, we treat nucleotides in the miRNA/mRNA sequence as individual words (each essentially a character), and the entire sequence is considered a sentence. Specifically, we take the input matrix $x \in \mathbb{R}^{8 \times 4}$ (with 8 being the length of the miRNA seed region), and project it into a dense real-valued space denoted as $E \in \mathbb{R}^{8 \times m}$. Here, $E$ represents the embedding matrix, and $m$ is a hyperparameter corresponding to the embedding dimension. This projection is achieved using the lookup matrix, also known as the embedding weight matrix, $W_m$ (i.e., $E = xW_m$). Subsequently, this embedding matrix $E$ serves as the input matrix for the BiLSTM model.

BiLSTM

In a vanilla RNN, when provided with an input matrix $E \in \mathbb{R}^{8 \times m}$, it generates a matrix $H$ of dimensions $8 \times n$, where $n$ represents the number of RNN units. At each time step (or sequence position) $p$, we have $e_p \in \mathbb{R}^m$ as the input column vector, $h_{p-1} \in \mathbb{R}^n$ as the previous hidden state vector, and we compute the current state $h_p$ as follows:

$$h_p = \sigma(e_pW + h_{p-1}U + b)$$

(1)

where $W$, $U$, and $b$ are the trainable parameters, and $\sigma$ is the nonlinear activation function. However, LSTM incorporates a meticulously crafted gating mechanism that can be trained to regulate the flow of information within hidden neurons. In this study, the LSTM unit is implemented by substituting the aforementioned Equation (1), which takes $e_p$, $h_{p-1}$, $c_{p-1}$ as inputs and generates $h_p$ and $c_p$ (where $c_p$ represents the candidate state):

$$i_p = \sigma(e_pW^i + h_{p-1}U^i + b^i)$$

(2)
\[ f_p = \sigma(e_p W^f + h_{p-1} U^f + b^f) \]  
\[ o_p = \sigma(e_p W^o + h_{p-1} U^o + b^o) \]  
\[ \tilde{c}_p = \tanh(e_p W^c + h_{p-1} U^c + b^c) \]  
\[ c_p = f_p \odot c_{p-1} + i_p \odot \tilde{c}_p \]  
\[ h_p = o_p \odot \tanh(c_p) \]

where \( W, U, \) and \( b \) are the trainable parameters, \( \sigma(\cdot), \tanh(\cdot), \) and \( \odot \) are element-wise sigmoid, hyperbolic tangent, and multiplication functions, respectively. \( i_p, f_p, \) and \( o_p \) are the input, forget, and output gates (Figure 4.1).

**Figure 4.1**: LSTM cell architecture.

Due to the bidirectional reason, the model processes input data in both forward and backward directions, enabling the integration of information from both the 5' and 3' miRNA sequences at each time step. In this process, it produces two outputs: one in the forward direction
labeled as $\vec{h}_p$ and another in the reverse direction labeled as $\vec{h}_p$. These two output states are then combined to create an output vector referred to as $h_{bi,p}$:

$$h_{bi,p} = \text{Bidirectional}(\vec{h}_p, \vec{h}_p)$$

where Bidirectional (⋅) is a function used to combine the two output sequences. In this research, a concatenating function was adopted and $h_{bi,p}$ is an output vector with dimension $8 \times 2n$ ($n$ represents the number of RNN units).

**Significant biological features**

Numerous studies have demonstrated that incorporating hand-crafted features alongside CNN or RNN can yield better results compared to relying solely on automatically generated features from representation learning. In line with this, we integrated seed conservation features, GC content features, free energy of seed binding features, UTR related position features, and phyloP score (a measure of evolutilional constraint) based features with feature representations obtained from the LSTM to enhance its predictive capacity. Specifically, we concatenated the biological feature vector $h_{bio}$ with $h_{bi,p}$ from the preceding Bidirectional LSTM stage, resulting in $h_{new} = \text{concat}([h_{bio}, h_{bi,p}])$, which serves as the final representation of the functional miRNA-mRNA target pair.

**Fully connected layers**

The feature representation $h_{new}$, obtained through concatenation, serves as the input for the fully connected layers. The specifications for the number of layers, hidden units, activation functions, and dropout rates for each layer were determined through hyperparameter tuning. These layers are essential for producing nonlinear combinations of features and enabling the network to generate predictions.
4.3.4 Model Training and Parameter Tuning

To facilitate the development of deep learning models, we initially divided the original dataset into three distinct subsets: 76.5% (9837 data points) for training, 8.5% (1093 data points) for validation, and 15% (1928 data points) for testing. The training subset was utilized to fine-tune model parameters, the validation subset played a role in preventing overfitting, and the testing subset served as the basis for assessing model performance. The performance of the trained model was evaluated using two key metrics: the Area Under the Curve (AUC) value obtained from ROC analysis and the Spearman correlation score. The search for optimal hyperparameters involved exploring a space comprising 12 different hyperparameters for RNN and 10 different hyperparameters for CNN. This resulted in a staggering 200 billion possible parameter combinations, far too numerous for efficient exploration using conventional grid search methods. To address this challenge, we implemented a Bayesian optimization approach using GpyOpt (31). This method significantly reduced the time required for hyperparameter optimization and has demonstrated greater reliability compared to randomized grid search.

Regarding the parameter tuning of traditional machine learning models, we applied a consistent approach to divide the initial dataset. We allocated 85% (10930 data points) for training and 15% (1928 data points) for testing. We evaluated the model performance using the same metrics as those used for deep learning models. In the case of the logistic regression model, we employed grid search to identify the best parameters. For XGBoost and random forest models, we utilized Hyperopt, a hyperparameter tuning tool that uses the Tree of Parzen Estimators (TPE) algorithm, to enhance the optimization process (32).
4.4 Results

4.4.1 miRDB Web Portal Improvement

miRDB presents a flexible web server interface for miRNA target retrieval. The default query form, the Target Search page, allows the users to retrieve target prediction data for one specific miRNA or gene target at a time (Figure 4.2). In addition, an advanced query form, the Target Mining page, enables the search for multiple miRNAs or gene targets simultaneously. The Target Mining page also presents additional search filters to allow various combinations of search strategies based on user preference for miRNA and target selection.

miRDB is an online database for miRNA target prediction and functional annotations. All the targets in miRDB were predicted by a bioinformatics tool, MiTarget, which was developed by analyzing thousands of miRNA-target interactions from high-throughput sequencing experiments. Common features associated with miRNA binding and target downregulation have been identified and used to predict miRNA targets with machine learning methods. miRDB hosts predicted miRNA targets in five species: human, mouse, rat, dog and chicken. Users may also provide their own sequences for custom target prediction using the updated prediction algorithms. In addition, through combinatorial computational analysis and literature mining, functionally active miRNAs in humans and rats were identified. These miRNAs, as well as associated functional enrichments presented in the FuncIR Collection in miRDB. As a recent update, miRDB presents the expression profiles of hundreds of cell lines and the user may limit their search for miRNA targets that are expressed in a cell line of interest. To facilitate the prediction of miRNA functions, miRDB presents a new web interface for integrative analysis of target prediction and Gene Ontology data.

Figure 4.2: miRDB Target Search page.

Presentation of updated target prediction data

We have recently developed an improved computational model for miRNA target prediction. Details of this prediction model have been described in our recent publication (16). The final support vector machine (SVM) model, MirTarget, was trained with the identified features for miRNA target prediction. Comparative analysis using independent datasets indicates that MirTarget has improved performance over other existing prediction algorithms (16).
With MirTarget, we performed transcriptome-wide miRNA target prediction for five species: human, mouse, rat, dog and chicken. Specifically, the miRNA sequences were downloaded from miRBase version 22 (4); target transcript sequences were retrieved from the NCBI RefSeq database and further parsed with BioPerl to extract the 3’-UTR sequences. Unlike many existing algorithms, evolutionary conservation of the target binding site is not a required feature for miRNA target prediction with MirTarget. In this way, both conserved and nonconserved targets can be predicted by MirTarget. For each candidate target site, MirTarget generates a probability score as computed by the underlying SVM modeling tool. This score reflects statistical confidence of the prediction results. If a transcript contains multiple candidate target sites, individual site scores are combined to compute a final score for the entire transcript, as described in detail in (16). MirTarget prediction scores are in the range of 0–100, and candidate transcripts with scores \( \geq 50 \) are presented as predicted miRNA targets in miRDB. In total, 3.5 million gene targets were predicted to be regulated by 7000 miRNAs across five species in the current version of miRDB (Version 6.0). In comparison, 2.1 million gene targets and 6700 miRNAs were included in the previous version (Version 5.0). On average, there are 497 gene targets per miRNA across the five species, an increase of 58% from the previous version. The significant increase in gene target number is mainly a result of newly implemented MirTarget features such as integrative analysis of multiple miRNA seed types in a single prediction model and more comprehensive assessment of cross-species conservation of the seed binding sites. Specifically for humans, the number of predicted targets per miRNA is 606, which is significantly higher than other species. This likely reflects the relatively rich annotations of the human transcriptome as compared to other transcriptomes.
Custom target prediction by implementing MirTarget into a web server

We performed a significant update on the custom prediction function of miRDB by implementing the new MirTarget algorithm. miRDB allows the users to provide custom miRNA or gene target sequences for transcriptome-wide prediction of gene targets or miRNA regulators in one of the five species: human, mouse, rat, dog, or chicken. The custom sequence length is in the range of 17–30 nt for miRNA and 100–30 000 nt for gene target. The users should select the species and search type (miRNA or gene target) and then input their custom sequence. Then, the Perl script implementing the MirTarget algorithm takes the web form inputs and starts the target prediction process. Two precompiled sequence files are used to predict potential miRNA/target pairs: one contains the 3’-UTR sequences from all known genes in the five species, and the other one includes the sequences of species-specific miRNAs, as collated from miRBase version 22.

The target prediction process is as follows. First, the web server script collects the miRNA or candidate sequence from the web form. If the users input a custom miRNA sequence, the server script will import all 3’-UTR sequences from the selected species for target prediction; on the other hand, if the users input a candidate target sequence, all miRNA sequences from the selected species will be imported. Next, for every miRNA/candidate target pair, the server script scans for miRNA seed binding sites and generates targeting features for MirTarget prediction. The prediction data are presented as an annotation table for all miRNA/candidate target pairs, including target prediction scores and the miRNA/target sequences.

The prediction results are sorted in descending order as ranked by the target score. Then, the web server script imports the sorted results for web presentation, including target rank, target prediction score, miRNA name, target gene symbol, and description. Additionally, the users can review the details of the prediction result for every miRNA/target pair, including miRNA sequence and target sequence with highlighted miRNA seed binding positions. Depending on the
number of predicted targets for the input miRNA, the whole prediction process typically completes in about 30-60 s.

*Target expression profiles in specific cell models*

By default, miRDB presents miRNA target prediction data for all known genes in the genome. However, not all potential miRNA targets are functionally relevant in any given cell. Thus, researchers often need to perform target analysis in the context of specific cell models. To facilitate the selection of cell-specific miRNA targets, miRDB presents a Target Expression page, enabling users to combine target prediction data with target expression profiles from over 1000 cell lines (Figure 4.3A). Specifically, we downloaded RNA-seq gene expression profiling data from two large-scale transcriptome studies (42, 43) deposited in Expression Atlas (44). Combined together, these studies have profiled RNA expression in 1178 cell lines by RNA-seq analysis. Gene expression levels are represented as normalized RPKM read counts (Reads Per Kilobase of transcript per Million mapped reads). Based on the expression values, we defined four gene groups: high expression (RPKM > 20), moderate expression (RPKM 5–20), low expression (RPKM 1–5), and no detectable expression (RPKM < 1). On average, there are about 11 000 genes with detectable expression per cell line. Gene targets with high or moderate expression in a specific cell model are more likely to be functionally impacted by miRNA regulation. As shown in Figure 4.3B, the users may further limit target selection by defining a desired gene expression threshold. The expression level of each gene target is presented together with the MirTarget prediction score. By integrating target prediction and expression data, the users can quickly identify cell-specific targets for further experimental validation.
Figure 4.3: miRDB target expression analysis. miRDB hosts the expression profiles of over 1000 cell lines. (A) A screenshot for selection of specific cell models. (B) A screenshot for integrative presentation of both target prediction and expression data for the selected cell model.

Prediction of miRNA functions by target ontology analysis

The function of a miRNA is defined by its gene targets. Thus, biological pathways regulated by miRNAs can be inferred by target analysis. However, as one miRNA can potentially regulate hundreds of gene targets, it is a challenge to identify significant pathways impacted by miRNA regulation reliably. One popular approach for miRNA functional prediction is to perform target enrichment analysis, i.e., identifying pathways or functional categories that
are statistically enriched in miRNA targets. To this end, we have implemented a new web interface for target ontology analysis. As presented in the Target Ontology page, the users may first retrieve all predicted targets for a specific miRNA of interest and then directly submit the target list for GO enrichment analysis (Figure 4.4). The GO enrichment analysis employs the PANTHER web server engine (45) using up-to-date GO terms (46). By providing a new web query interface, miRDB integrates both target prediction and GO enrichment analyses and presents a streamlined pipeline for the prediction of miRNA functions.

![Figure 4.4: miRDB target ontology analysis. miRNA target prediction data and GO data were integratively analyzed to predict miRNA functions.](image)

In summary, we have performed major updates on miRDB, including the implementation of a new miRNA target prediction algorithm, as well as the presentation of new database features for the prediction of miRNA functions. By combining miRNA target prediction data with other biological data, such as cell-specific expression profiles or GO annotations, we expect these new miRDB features to be helpful for researchers to identify relevant miRNA functions in specific experimental systems quickly. In the next step, we will continue to make improvements to the target prediction algorithm as well as integrate more heterogeneous types of data in miRDB for flexible analysis of miRNA functions in various experimental settings.
4.4.2 Combining Target Downregulation and CLIP Binding Data to Identify Common Targeting Features

A common challenge in miRNA overexpression studies is precisely pinpointing the miRNA binding site within the target transcript. To address this concern, we sought out potential target sites using canonical 7-mer or 8-mer seed sites as markers. In contrast, CLIP-ligation studies offer a definitive method for identifying miRNA binding sites in the target transcript by creating a crosslink between the miRNA and its matching target site in the same RISC complex. However, it is not straightforward to determine the functional impact of miRNA target binding, as revealed by CLIP. Therefore, both CLIP binding and miRNA overexpression methods have advantages and disadvantages. Each method focuses on a distinct aspect of miRNA target regulation, namely, target binding or functional suppression. In this study, we merged the CLIP binding dataset with new miRNA overexpression data to pinpoint shared characteristics of both miRNA binding and target suppression. We subsequently analyzed 4,774 target sites and 8,081 non-target sites, which were derived from both CLIP and miRNA overexpression studies for further feature analysis.

In the combined dataset, we compared target sites and non-target sites to identify shared characteristics related to miRNA target regulation, which are often evolutionarily conserved (33, 34). To assess target conservation in our study, we employed two complementary methods. Firstly, we calculated the difference in conservation scores between the seed binding and flanking positions using phyloP scores from a 100-way multi-genome alignment (35). Secondly, we conducted a word search to determine if the entire seed site (7-mer or 8-mer) appeared across multiple species. Both conservation analyses consistently demonstrated that target sites exhibited significantly higher conservation than non-target sites. Notably, seed conservation emerged as
one of the most pronounced features, whether we analyzed miRNA overexpression and CLIP binding data individually or in combination. In particular, the conserved seed8A1 type stood out as the most enriched in target sites (p = 2.8E-245 by cross-species seed match and p = 7.3E-218 by phyloP score, respectively). Conversely, the non-conserved seed7A1 type was the most depleted seed variant (9.5E-134 by seed match and p = 1.3E-138 by phyloP score, respectively).

In addition to seed conservation, several other features were commonly observed in both datasets. For instance, miRNA target sites were notably linked to shorter 3'-UTR sequences (p = 1.5E-144), tended to be positioned toward the end of the 3'-UTR sequence (p = 5.4E-66), and were less likely to be found in the center of long transcripts (p = 2.5E-87).

**Table 4.2.** Summary of biological features of miRNA targeting.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Combined Data</th>
<th>Overexpression Data</th>
<th>CLIP Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Targets</td>
<td>Non-targets</td>
<td>P-value</td>
</tr>
<tr>
<td>Seed 8A1, conserved</td>
<td>0.184</td>
<td>0.018</td>
<td>2.8E-245</td>
</tr>
<tr>
<td>Seed7b, conserved</td>
<td>0.124</td>
<td>0.048</td>
<td>8.7E-55</td>
</tr>
<tr>
<td>Seed7A1, conserved</td>
<td>0.052</td>
<td>0.028</td>
<td>6.7E-12</td>
</tr>
<tr>
<td>Seed8A1, non-conserved</td>
<td>0.210</td>
<td>0.123</td>
<td>8.5E-39</td>
</tr>
<tr>
<td>Seed7b, non-conserved</td>
<td>0.289</td>
<td>0.441</td>
<td>1.3E-65</td>
</tr>
<tr>
<td>Seed7A1, non-conserved</td>
<td>0.142</td>
<td>0.341</td>
<td>9.5E-134</td>
</tr>
<tr>
<td>Feature</td>
<td>Terminal A-U match</td>
<td>Terminal C-G match</td>
<td>Distance to UTR end (log2)</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>--------------------</td>
<td>--------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.360</td>
<td>0.264</td>
<td>1.3E-30</td>
</tr>
<tr>
<td></td>
<td>0.014</td>
<td>0.042</td>
<td>4.4E-18</td>
</tr>
<tr>
<td></td>
<td>0.299</td>
<td>0.170</td>
<td>5.4E-66</td>
</tr>
<tr>
<td></td>
<td>0.248</td>
<td>0.421</td>
<td>2.5E-87</td>
</tr>
<tr>
<td></td>
<td>10.960</td>
<td>11.430</td>
<td>1.5E-144</td>
</tr>
<tr>
<td></td>
<td>-2.583</td>
<td>-2.596</td>
<td>2.6E-01</td>
</tr>
<tr>
<td></td>
<td>1.554</td>
<td>1.901</td>
<td>4.9E-177</td>
</tr>
<tr>
<td></td>
<td>1.595</td>
<td>1.926</td>
<td>3.0E-89</td>
</tr>
<tr>
<td></td>
<td>1.558</td>
<td>1.936</td>
<td>2.4E-96</td>
</tr>
<tr>
<td></td>
<td>0.052</td>
<td>0.022</td>
<td>2.2E-19</td>
</tr>
<tr>
<td></td>
<td>0.146</td>
<td>0.009</td>
<td>7.3E-218</td>
</tr>
<tr>
<td></td>
<td>0.093</td>
<td>0.022</td>
<td>8.7E-74</td>
</tr>
<tr>
<td>Seed7A1, high phyloP score</td>
<td>0.036</td>
<td>0.014</td>
<td>1.0E-15</td>
</tr>
<tr>
<td>Seed8A1, moderate phyloP score</td>
<td>0.047</td>
<td>0.006</td>
<td>7.5E-53</td>
</tr>
<tr>
<td>Seed7b, moderate phyloP score</td>
<td>0.047</td>
<td>0.022</td>
<td>3.0E-15</td>
</tr>
<tr>
<td>Seed7A1, moderate phyloP score</td>
<td>0.021</td>
<td>0.016</td>
<td>5.8E-02</td>
</tr>
<tr>
<td>Seed8A1, low phyloP score</td>
<td>0.200</td>
<td>0.126</td>
<td>1.5E-29</td>
</tr>
<tr>
<td>Seed7b, low phyloP score</td>
<td>0.273</td>
<td>0.445</td>
<td>3.2E-84</td>
</tr>
<tr>
<td>Seed7A1, low phyloP score</td>
<td>0.137</td>
<td>0.339</td>
<td>1.3E-138</td>
</tr>
</tbody>
</table>

While miRNA overexpression and CLIP binding data share many similarities, they also exhibit distinct differences. Notably, one significant distinction pertains to the GC content of the target sites. In CLIP binding data, the GC content at target sites was markedly lower compared to non-target sites (p = 1.9E−146), while in miRNA overexpression data, this difference was only moderately lower (p = 2.1E−10). This variation might be attributed to the use of RNase T1 in CLIP studies, which prefers cutting at G nucleotides, consequently leading to a reduction in internal G within sequencing reads. Nevertheless, it's also plausible that the enrichment of G hampers the binding of the miRISC complex to target sites, as G was similarly depleted in miRNA overexpression data, albeit to a lesser extent. Another noteworthy feature is the stability of seed binding, as indicated by the free energy of the seed/target duplex. In miRNA
overexpression data, seed binding stability was favored \((p = 2.5E^{-12})\), whereas in CLIP binding data, it was disfavored \((p = 5.4E^{-26})\). However, when the two datasets were combined, this feature ceased to be statistically significant \((p = 0.26)\).

### 4.4.3 Performance Comparison of Various Modeling Methods

We aim to create an advanced prediction algorithm by leveraging key biological characteristics for functional miRNA target pairs. To select the most effective modeling approach, we assessed the training performance of three conventional machine learning models, two deep learning models, and an integrated model. We measured model performance by calculating AUC values through ROC analysis (Figure 4.5).

**Conventional machine learning models**

Our primary goal of this project is to create a classification model for identifying functional miRNA targets. Initially, we evaluated the performance of three conventional classification models: Logistic regression, Random Forest classifier, and XGBoost classifier. We used the combined dataset we generated in the current study for this evaluation. To prevent overfitting, we divided the dataset randomly into two subsets. 85% of the data was utilized for training the models, while the remaining 15% was set aside to evaluate the models' ability to generalize. We fine-tuned the model parameters for optimal performance using a hyperparameter optimization tool (32).

Among the three models examined, the XGBoost classifier demonstrates the highest predictiveness, boasting a ROC-AUC value of 0.683. Logistic regression is the second most predictive, achieving an ROC-AUC value of 0.674. Random forest classifier also performs well but yields a comparatively lower AUC score (Figure 4.5).
Figure 4.5: Performance of different algorithms for miRNA target prediction. The performance of various algorithms in predicting miRNA targets was assessed through a 10-fold cross-validation ROC analysis. The bar plot illustrates the average AUC value of ROC curves generated from predicted values and experimentally validated miRNA-target labels, presented as mean ± standard deviation.

Deep learning-based models

Recent research has demonstrated the effectiveness of two deep learning algorithms, namely the Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), in analyzing DNA, RNA, and protein sequences (33-37). These algorithms can automatically extract valuable features from the raw sequences, eliminating the need for manual feature engineering. Consequently, we employed both CNN and LSTM (a subclass of RNN) for predicting functional miRNA-target pairs. To mitigate overfitting, we randomly divided the
dataset into three subsets: 76.5% for model training, 8.5% for validation, and the remaining 15% for evaluating the models' generalization performance.

The LSTM algorithm demonstrated superior performance in miRNA target prediction, achieving an AUC-ROC score of 0.698, outperforming both CNN and other algorithms. In comparison, CNN showed performance comparable to logistic regression, achieving an AUC score of 0.674.

An integrated model with additional biological features

The core of deep learning algorithms, such as CNN or RNN, primarily leverages k-mer composition or its dependencies (33, 38). Recent research has demonstrated that enhancing the predictive capabilities of deep learning models is achievable by incorporating additional biological features that are not automatically accessible to these models (39). In our current study, we extracted significant biological features from a comparison between target and non-target miRNA-gene pairs. These features, for example, seed conservation, GC content, and relative positioning within the entire UTR sequence are not inherently accessible through deep learning methods. Notably, our LSTM model demonstrated superior performance among all algorithms. As a result, we integrated these biological features with the LSTM model to further improve its predictive power (Figure 4.6). The inclusion of these features significantly boosted prediction performance, achieving an AUC-ROC value of 0.817. Consequently, we adopted the LSTM model combined with these biological features, referred to as LSTM + Biofeat, as our final model for miRNA target prediction.
Figure 4.6: The workflow of LSTM+Biofeat model. The miRNA-gene pairs from the training dataset were first used to extract sequence-independent features (i.e., biological features). Simultaneously, the original miRNA/mRNA sequence is encoded and embedded to get sequence-dependent features, including information like nucleotide counts and position-specific nucleotides. A BiLSTM network then processed these feature sets to generate the final representation, which was subsequently combined with the biological features and used as input for fully connected layers. The fully connected layers conducted non-linear transformations and provided the final prediction score.

4.4.4 Algorithm Evaluation with Independent Experimental Data

A common concern when developing algorithms is that a model may perform well during training but struggle with unseen data. Therefore, to assess the effectiveness of LSTM+Biofeat, it is essential to test it with independent experimental data. In our current research, we evaluated the algorithm using a diverse set of experimental data sources, which included data from CLIP binding experiments, miRNA overexpression experiments, and an integrated database of experimentally confirmed miRNA-target interactions. Additionally, we compared the performance of LSTM+Biofeat with four other widely recognized algorithms: TargetScan8,
DIANA-MicroT, miRanda (mirSVR), and PITA. These algorithms are well-established in the field of miRNA target prediction, and transcriptome-wide prediction data are presented from their respective websites.

**Validation analysis using CLIP and miRNA overexpression data**

Chi et al. introduced the HITS-CLIP method to experimentally identify miRNA target transcripts (40). In this approach, they used crosslinking immunoprecipitation to isolate mRNA transcripts associated with miRISC in the mouse brain. High-throughput sequencing was employed to pinpoint these mRNA transcript tags, which are short RNA fragments protected by Ago from RNase digestion. Chi et al. demonstrated that, typically, these transcript tags are centered on the seed binding sites (40). We leveraged this HITS-CLIP dataset in our research to identify potential miRNA target sites. Specifically, we identified 886 potential target sites based on seed-matching sequences for the six most abundantly expressed miRNAs. As part of negative controls, we selected a set of possible non-target sequences according to two criteria: (1) they had no overlap with any sequence tags from the HITS-CLIP experiment, and (2) they originated from transcripts with detectable expression levels as determined by microarrays. We designated 889 of these non-target sites with seed-matching sequences as negative controls.

We also assessed target prediction algorithms in the context of changes in target gene expression. In this comparative analysis, we used miRNA overexpression data from our current study to evaluate these algorithms. Specifically, we considered transcripts that showed a decrease of at least 40% in expression in one of the duplicated experiments as potential miRNA targets. Transcripts with expression reductions of less than 5% or even those with an increase in expression of up to 10% were classified as non-targets. It's important to note that all transcripts, both miRNA targets and non-targets, contained at least one miRNA seed binding site. As a
result, we identified 1107 target sites and 2214 non-target sites from the miRNA overexpression data.

**Figure 4.7**: ROC analyses of algorithm performance on validation datasets. (a) ROC analysis of five algorithms using HITS-CLIP validation dataset. (b) ROC analysis of five algorithms using miRNA overexpression validation dataset.

In our analysis, we assessed the performance of five computational algorithms—namely LSTM+Biofeat, TargetScan8, DIANA-MicroT, miRanda, and PITA. We evaluated these algorithms by measuring their capacity to distinguish between target and non-target of miRNAs, as indicated by HITS-CLIP and miRNA overexpression datasets. To achieve this, we employed ROC analysis to gauge the overall sensitivity and specificity of these prediction algorithms. As presented in Figure 4.7, LSTM+Biofeat demonstrated the highest performance when validated with both HITS-CLIP and miRNA overexpression datasets. It achieved an area under the ROC curve (AUC) of 0.784 and 0.750, respectively. Notably, DIANA-MicroT exhibited the second-best performance (AUC = 0.729) in the ROC analysis using the HITS-CLIP validation dataset. This result is unsurprising, given that DIANA-MicroT was developed based on training with...
CLIP binding data, making it well-suited for CLIP testing data. Conversely, TargetScan8, trained using miRNA overexpression data and miRNA binding data, demonstrated strong performance in the initial part of the ROC curve (with a false positive rate $\leq 0.2$) when assessed with the miRNA overexpression validation dataset. However, it primarily focuses on predicting conserved target sites, leading to a reduced true positive rate in the region where the false positive rate exceeds 0.2. Consequently, TargetScan8 attained a relatively lower AUC value of 0.643.

A case study using experimentally validated miRNA-target pairs from miRTarBase

The LSTM+Biofeat algorithm exhibited exceptional performance during validation with high-throughput miRNA datasets such as CLIP binding and miRNA overexpression data. We further substantiated its effectiveness by validating with experimentally confirmed miRNA-target pairs. To accomplish this, we gathered verified miRNA-target pairs from the miRTarBase database (41), a meticulously curated repository of MicroRNA-Target Interactions (MTIs). This database has compiled over two million confirmed MTIs through the mining of 13,389 relevant research articles. These articles were systematically analyzed to extract studies on the functional examination of miRNAs. Typically, the collected MTIs have been experimentally validated through reporter assays, western blots, microarrays, and next-generation sequencing experiments. miRTarBase stands out as the largest repository of validated MTIs and remains the most up-to-date collection compared to other previously developed databases (41).

In this study, we obtained 8157 confirmed human functional MTIs from miRTarBase, validated through low-throughput biological experiments such as luciferase reporter assays, Western blots, Northern blots, and qRT-PCR. This dataset encompasses 200 distinct human miRNAs. Five miRNA target prediction algorithms were used to predict potential targets for
these 200 miRNAs. The top-scored predicted targets for each miRNA were chosen, ensuring an equal number of predicted targets across different algorithms for these 200 miRNAs. By cross-referencing these predicted targets with the 8157 validated functional MTIs, we assessed the effectiveness of the miRNA target prediction algorithms. Table 4.3 illustrates the results for four sets of predicted targets: the top-10, top-5, top-3, and top-1 predicted targets for each miRNA. Notably, the LSTM+Biofeat algorithm produced the highest number of validated predicted targets, with 219, 143, 96, and 41 overlapping targets with miRTarBase for the four collections, respectively. Compared to the second-best algorithm, LSTM+Biofeat identified 37.7.8%, 60.7%, 65.5%, and 78.3% more validated targets for the top-10, top-5, top-3, and top-1 collections of predicted targets, respectively. It's worth noting that when stricter criteria are applied to select predicted targets, LSTM+Biofeat is expected to show even greater improvement. These results clearly demonstrate that LSTM+Biofeat is the most informative prediction algorithm in this comparative analysis.

Table 4.3. Number of validated targets for five miRNA target prediction algorithms.

<table>
<thead>
<tr>
<th>Prediction Algorithms</th>
<th>Top-10 predicted targets (total: 2000)</th>
<th>Top-5 predicted targets (total: 1000)</th>
<th>Top-3 predicted targets (total: 600)</th>
<th>Top-1 predicted targets (total: 200)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM + Biofeat</td>
<td>219</td>
<td>143</td>
<td>96</td>
<td>41</td>
</tr>
<tr>
<td>TargetScan8</td>
<td>147</td>
<td>89</td>
<td>58</td>
<td>23</td>
</tr>
<tr>
<td>DIANA-MicroT</td>
<td>159</td>
<td>78</td>
<td>56</td>
<td>18</td>
</tr>
<tr>
<td>miRanda</td>
<td>77</td>
<td>46</td>
<td>32</td>
<td>10</td>
</tr>
<tr>
<td>PITA</td>
<td>124</td>
<td>83</td>
<td>49</td>
<td>15</td>
</tr>
</tbody>
</table>
4.5 Discussion

Advancements in miRNA target prediction heavily rely on the availability of high-quality training datasets. Recently, innovative CLIP-seq techniques have enabled the direct identification of target transcripts bound to the miRISC complex. While this approach is valuable, it raises concerns when applied to training target prediction algorithms. One significant concern is that most binding events observed in CLIP experiments appear to have limited functional impact, as indicated by minimal changes in target expression (47, 48). Likely, many transcripts identified by CLIP are only temporarily recognized by miRISC, quickly dissociating without resulting in expression changes. There is also the possibility that miRISC binding has functional relevance beyond target downregulation, such as influencing the cytoplasmic distribution of miRNAs.

In most miRNA studies, researchers aim to identify target transcripts downregulated by the specific miRNA of interest. Consequently, our present study combines CLIP binding data with miRNA overexpression data to systematically pinpoint functional miRNA targets. However, it's important to note that miRNA overexpression may distort target regulation under normal physiological conditions, introducing a caveat.

Both CLIP and miRNA overexpression methods have their distinct advantages and disadvantages when it comes to analyzing miRNA targets. Our analysis reveals that CLIP binding and miRNA overexpression data share several common features, particularly those related to seed conservation. Nonetheless, we also observed notable differences in certain features, suggesting that these two methods capture different facets of miRNA target regulation. By incorporating both types of data into the prediction algorithm, it can be more broadly applicable across various experimental settings.
In terms of the modeling approach, recent years have seen the emergence of deep learning and related technologies as a superior choice for tackling complex problems. These problems include automatic speech recognition, image recognition, and natural language processing, among others. Deep learning models have the unique ability to automatically extract hidden features from input data, making them more potent than conventional machine learning models that rely on manual feature engineering. Furthermore, deep learning models can effectively utilize large datasets and leverage multi-layer neural networks to enhance prediction performance. In contrast, conventional machine learning models face limitations when dealing with extensive training datasets.

We successfully developed an integrated model based on the LSTM architecture by combining datasets from CLIP binding and miRNA overexpression. Through extensive validation analyses, we demonstrated the effectiveness and superior predictive capabilities of our LSTM+Biofeat model. As high-throughput miRNA experimental technologies continue to advance, we anticipate the emergence of more valuable miRNA data. This, in turn, will enable deep learning technologies to enhance further model performance and better support research on miRNA regulation.

4.6 References


Chapter 5: Conclusions

In this dissertation, we aimed to employ computational methods to identify critical biological characteristics related to gene regulation mechanisms and technologies and develop engineering tools accordingly. Gene regulation plays a pivotal role in understanding the function of genes and the products they produce in various organisms. We explicitly focus on two main areas of gene regulation: the miRNA regulatory system and the CRISPR genome editing system. Because of the growing need for additional computational tools to aid researchers in their biological analyses and experimental design, we leveraged machine learning and deep learning technologies to create a suite of bioinformatics tools for advanced CRISPR/Cas9 guide RNA design and miRNA target prediction.

To uncover the key qualities of effective guide RNAs within the CRISPR/Cas9 system, we initiated a novel experimental approach. This method used in silico-designed target sites derived from an extensive plasmid library. Subsequently, we directly determined target editing through sequencing, minimizing potential biases associated with specific experimental setups. With the aid of this innovative training dataset, we conducted a thorough feature analysis, leading to the discovery of several new attributes linked to efficient guide RNAs, such as GC content and secondary structural features. Consequently, we constructed a predictive model called sgDesigner, using an ensemble approach that combined Support Vector Machine and XGBoost models. This model outperformed several existing guide RNA design tools across multiple validation datasets.

Leveraging our prior expertise in CRISPR guide RNA design, we embarked on a more intricate challenge in this field. Specifically, predictive algorithms for guide RNA efficiency exhibit varying performance across different experimental settings, attributable to disparities in
modeling approaches and training datasets. To comprehensively assess these algorithms, we conducted a comparative analysis involving 15 established guide RNA prediction models, encompassing a range of modeling strategies, including traditional machine learning, ensemble learning, deep learning, and physical scoring functions. To assess each algorithm, we employed both ROC analysis and Spearman correlation analysis and validated these models using 16 independent experimental datasets. This extensive evaluation represents the most comprehensive analysis in this field, to the best of our knowledge. Building on the insights gained from this comparative analysis, we developed an integrated model named CRISPRDB, which combines the top-performing algorithms. Our final model demonstrated superior performance to any individual algorithm from the comparative analysis. In the last phase, we created a guide RNA library using CRISPRDB and made it available, along with the algorithm, through the website https://crisprdb.org.

We then shifted our focus to predicting miRNA targets. To make the most of the strengths of CLIP binding data and miRNA overexpression data, we decided to use the shared features that emerge from both data types, representing the characteristics of direct functional interactions between miRNAs and their targets. We conducted statistical analyses for each potential feature in three datasets: CLIP binding data, miRNA overexpression data, and their combined data. Features that showed statistical significance in all three datasets were considered common features. Subsequently, we assessed the performance of various machine learning and deep learning models for training. Among them, the Long Short-Term Memory (LSTM), a subtype of recurrent neural networks (RNN), emerged as the top-performing model. Furthermore, we improved the LSTM model by incorporating additional biological features that it couldn't generate automatically. This enhanced model, named LSTM+Biofeat, outperformed
other algorithms for miRNA target prediction when tested with both independent CLIP binding data and miRNA overexpression data. This testing performance confirmed the LSTM+Biofeat model’s effectiveness in identifying direct functional miRNA-target interactions. With the support of this newly developed algorithm, we made updates to our miRDB web portal. We introduced several new functions on the website, including custom target prediction, target expression analysis in specific cell models, and predicted target ontology analysis. The miRDB web portal can be accessed at https://mirdb.org and receives hundreds of unique visitors daily, demonstrating its significant impact on the research community.

This dissertation focuses on utilizing machine learning and deep learning technologies in the realms of CRISPR guide RNA design and miRNA target prediction. These advanced technologies enable us to gain valuable insights into complex biological challenges that are traditionally difficult to address through experimental methods. Our research showcases the effectiveness of our bioinformatics tools in forecasting functional miRNA targets and crafting effective guide RNAs for the CRISPR system. This, in turn, facilitates significant time and resource savings for researchers. In summary, our bioinformatics tools, tailored for gene regulation-related domains, have found widespread adoption and have made a substantial impact in the research community.