Computational and Structural Biotechnology Journal 23 (2024) 1740–1754

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Advanced computational predictive models of miRNA-mRNA interaction efficiency

Sharon Bader [a], Tamir Tuller [a,b,*]

[a] Department of Biomedical Engineering, Tel-Aviv University, Tel Aviv, Israel
[b] The Segol School of Neuroscience, Tel-Aviv University, Tel Aviv, Israel

ARTICLE INFO

Keywords:
MiRNAs
MRNA
Biophysics
Competition
Machine Learning

ABSTRACT

The modeling of miRNA-mRNA interactions holds significant implications for synthetic biology and human health. However, this research area presents specific challenges due to the multifaceted nature of mRNA downregulation by miRNAs, influenced by numerous factors including competition or synergism among miRNAs and mRNAs. In this study, we present an improved computational model for predicting miRNA-mRNA interactions, addressing aspects not previously modeled. Firstly, we integrated a novel set of features that significantly enhanced the predictor's performance. Secondly, we demonstrated the cell-specific nature of certain aspects of miRNA-mRNA interactions, highlighting the importance of designing models tailored to specific cell types for improved accuracy. Moreover, we introduce a miRNA binding site interaction model (miBSIM) that, for the first time, accounts for both the distribution of miRNA binding sites along the mRNA and their respective strengths in regulating mRNA stability. Our analysis suggests that distant miRNA sites often compete with each other, revealing the intricate interplay of binding site interactions. Overall, our new predictive model shows a significant improvement of up to 6.43% over previous models in the field.

The code of our model is available at https://www.cs.tau.ac.il/~tamirtul/miBSIM

1. Introduction

MicroRNAs (miRNAs) are short strands of non-coding RNA, typically around 21–24 nucleotides in length, tasked with the important role of regulating gene expression [1]. Despite their short length, miRNAs play a pivotal role in orchestrating various cellular processes by regulating protein production, particularly the translation of mRNA into protein. Operating within the cytoplasm, miRNAs interact with mRNA molecules by attaching to binding sites, leading to the mRNA's destabilization and degradation, thereby actively repressing mRNA levels within the cell [2]. This straightforward mechanism enables miRNAs to participate in a multitude of cellular functions, ranging from pluripotency and developmental processes to metabolic and cellular pathways, as well as influencing the cell cycle. Consequently, dysregulation of miRNAs is associated with pathological conditions and has been implicated in oncogenesis [3–8].

Transcribed in the nucleus, miRNAs initially are in a hairpin structure and are subsequently processed by the Dicer enzyme in the cytoplasm, resulting in the formation of mature miRNA strands. RNA repression is facilitated by the recruitment of an RNA-induced silencing complex (RISC), which consists of Argonaute (AGO) nucleases responsible for cleaving or destabilizing mRNA molecules [9]. The miRNA-RISC complex interacts with mRNA through complementary base pairing. Canonical interactions involve a perfect match to the 2–8 nucleotides at the 5' end of the miRNA, known as the seed region. These interactions are classified into four types (6mer, 7mer-A1, 7mer-m8, 8mer), each characterized by the length of base pairing and associated with the affinity and efficiency of the interaction. Although potential binding sites can be found throughout the mRNA strand, interactions within the 3'UTR region are particularly effective for miRNA-mediated repression [10,11].

Due to miRNAs' pivotal regulatory role, investigating miRNA-mRNA interactions is fundamental in understanding their involvement in numerous cellular processes, including cellular differentiation, proliferation, apoptosis, oncogenesis, pathogenesis, and defense against viruses [12–19]. Deciphering these interactions holds significant value, as they can serve as a simple and effective tool in various biotechnological and medical applications. Researchers utilize them as biomarkers, integrate them into medical drugs, and employ miRNAs in cancer treatment [20,21]. Many of these studies require the engineering of

* Corresponding author at: Department of Biomedical Engineering, Tel-Aviv University, Tel Aviv, Israel.
E-mail address: tamirtul@tauex.tau.ac.il (T. Tuller).

https://doi.org/10.1016/j.csbj.2024.04.015
Received 10 January 2024; Received in revised form 6 April 2024; Accepted 7 April 2024
Available online 19 April 2024

effective binding sites for endogenous miRNAs, balancing desired interactions while avoiding interference with other signals [22]. Previous studies have developed computational models aiming to predict functional targets of miRNAs and quantify miRNA-mediated repression, albeit with varying success [23–31]. However, capturing these interactions and their effects presents challenges, given the multitude of known and unknown variables involved. While sequencing transfection experiments is commonly used to observe changes in mRNA levels due to specific miRNAs, it lacks detailed information on miRNA-mRNA interactions. Conversely, techniques such as CLASH mapping (cross--linking, ligation, and sequencing of hybrids) have primarily identified non-canonical miRNA-mRNA interactions with insignificant effects on mRNA repression [32]. Furthermore, highly-performing models have often been overfitted and case-specific, trained on small-scale experimental data. Finally, repression by individual miRNAs tends to be modest, resulting in relatively weak signals prone to noise and experimental bias [33]. These challenges hinder prediction accuracy and the relevance of models.

Most existing models have simplified the complexity of miRNA-mediated repression by employing linear algorithms that incorporate various interaction features such as thermodynamics, conservation, and sequence context. These models treat binding sites as independent entities, assuming that interactions between adjacent sites are insignificant. However, emerging evidence suggests that miRNAs may regulate their targets in a cooperative or competitive manner. Studies have shown that mRNAs strongly regulated by miRNAs often harbor multiple binding sites for the same or different miRNAs [33,34]. Developing a model that incorporates binding interactions could lead to a more accurate representation of the repression mechanism and improve predictive capabilities. By considering the interplay between binding sites, such a model could better capture the complexity of miRNA-mediated regulation.

Each cell type expresses a unique set of miRNAs allowing for cell-specific pathway regulation [26,35,36]. Investigating cell specificity in the context of miRNA-mediated repression holds promise for enhancing our understanding of miRNA-mRNA interactions. Tools like PUMA [37] and miTAlos [26] have been developed to categorize subgroups of miRNAs and their roles in different cell-specific processes. Similarly, researchers have worked on comparing cell-specific miRNA expression profiles in both healthy and pathologic tissues, such as cancerous tumors, heart disease and many more, in the aim of characterizing diseased tissue. However, while efforts have been made to identify cell-specific miRNA expression patterns, predicting miRNA-mediated repression quantitatively in a cell specific context has not been thoroughly addressed until now. Since different cells harbor distinct sets of active miRNAs, they may possess unique mechanisms or signals that allow different subsets of miRNAs to function optimally. These hidden signals could influence the importance of features and lead to biased models trained on specific cell types. Consequently, the development of cell-specific repression models has the potential to enhance prediction accuracy and shed light on the cell-specific attributes of miRNA-mRNA interaction.

Here, we investigate the importance of cell-specific repression models and present a novel miRNA-mediated repression model that integrates miRNA binding site interactions for the first time. Our findings support the importance of considering binding site distribution, resulting in an improved predictive model.

2. Results

mRNA levels within a cell dictate protein translation, thereby influencing cellular processes and functionality. As miRNAs contribute to mRNA regulation through silencing mechanisms, investigating their interactions holds significant value. Here, we aim to develop a computational model that predicts miRNA-mediated repression and improves upon the performance of existing models. We approach this objective in three ways, as illustrated in Fig. 1. First, we establish a basic model that incorporates elements from previously published models and tools. Recognizing the ongoing emergence of new studies and data, it is essential to continuously update computational models to attain a better understanding of miRNA activity and to enhance model performance. While individual studies may focus on specific aspects of miRNA-mRNA interaction, our goal is to integrate these elements into a comprehensive model, capturing as many variables involved in the mRNA silencing process as possible.

Second, we investigate whether these models exhibit cell-specific attributes by comparing the performance of models trained on the same cell type to those trained on different cell types. Since miRNA profiles and activities vary among different tissues, we anticipate that transcripts may encode cell-specific information regarding miRNA-mRNA interactions. As feature selection and model training heavily rely on the training dataset, if indeed features encompass cell-specific signals we expect models trained on the same cell type to better predict miRNA-mediated repression compared to models trained on different cell types.

Third, we develop a binding site interaction model, which extends the basic model with a correction step incorporating target site cooperativity or competition. Here, we aim to further improve our model by proposing that interactions between binding sites may augment repression. Sites may act cooperatively to facilitate more efficient binding and destabilization of the mRNA [38,39].

2.1. The experimental data of miRNA based mRNA repression is noisy

Transfection experiments measuring differential gene expression currently represent the best method for recording mRNA silencing due to miRNA, yet they come with several disadvantages. Foremost among these is the limited information they provide, as they only measure changes in mRNA levels. However, this simplistic measurement overlooks the complex interplay of factors contributing to mRNA regulation. Notably, changes in mRNA levels observed in these experiments cannot always be attributed solely to the transfected miRNAs. Studies have indicated that siRNA-independent effects may also influence mRNA levels, particularly in miRNA transfection experiments. It has been shown that when clustering differential mRNA expression, sets from the same group of experiments or the same transfection protocol clustered strongly together [30,40]. Further complicating matters, transcriptome-wide responses are strongly correlated with aspects such as AU content and 3'UTR length, leading to global expression changes that are dependent on experimental context and resulting in batch effects [23,41–44]. Additionaly, it has been suggested that the transfected miRNAs may cause derepression of mRNAs by competing with active endogenous miRNAs on the finite pool of silencing complex elements [42,45]. These effects are particularly pronounced when looking at fold changes, as these tend to have a low signal, resulting in low correlations between fold changes across different experiments and batches. Analyzing transfection dataset performed by the same experimental protocol and group, we see a mean correlation between repetitions of $r = 0.582$ overall (Fig. 2; $r = 0.672$ for the HeLa cells, $r = 0.455$ for the HEK293FT cells [46]) which range between 0.123 and 0.909. Even when experimental biases are attempted to be corrected, the limitations of training and assessing model performance on this type of data must be acknowledged. Moreover, miRNA-mediated repression is highly dependent on cellular context, including factors such as the number of mRNAs, miRNA concentrations, and the pool size of available AGO proteins [47]. As these parameters are dynamic, it is important to recognize the inherent limitations of computational models in capturing such complexities. The correlations reported in this study thus represent an upper bound on the correlations achievable with experimental data.

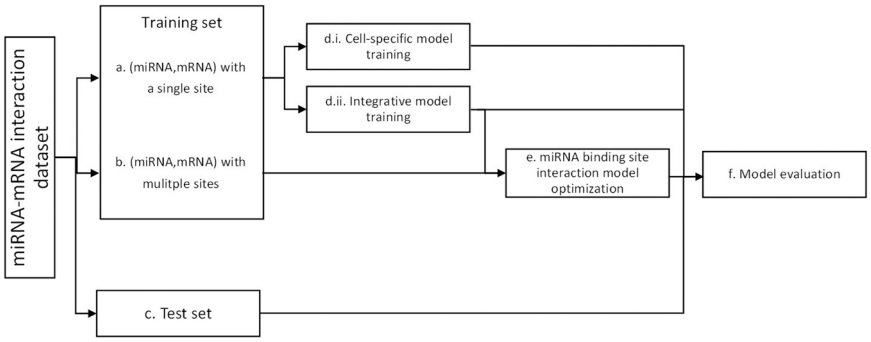

Fig. 1. Study Schematic. Starting with a miRNA-transfection dataset we partition our data into 2 groups: training set and test set, each with distinct miRNAs and mRNAs. The training set is then split to (a.) (miRNA,mRNA) couples that have a single potential target site, and (b.) (miRNA,mRNA) couples with multiple potential target sites. The first set (a.) is used to train the (d.i.) cell-specific models and (d.ii.) Integrative model (similar models with slight modifications to remove bias). The second set (b.) is then used together with the trained Integrative model (d.ii) to optimize the miBSIM. Finally, all models are evaluated (f.) based on their performance on the test set (c.).

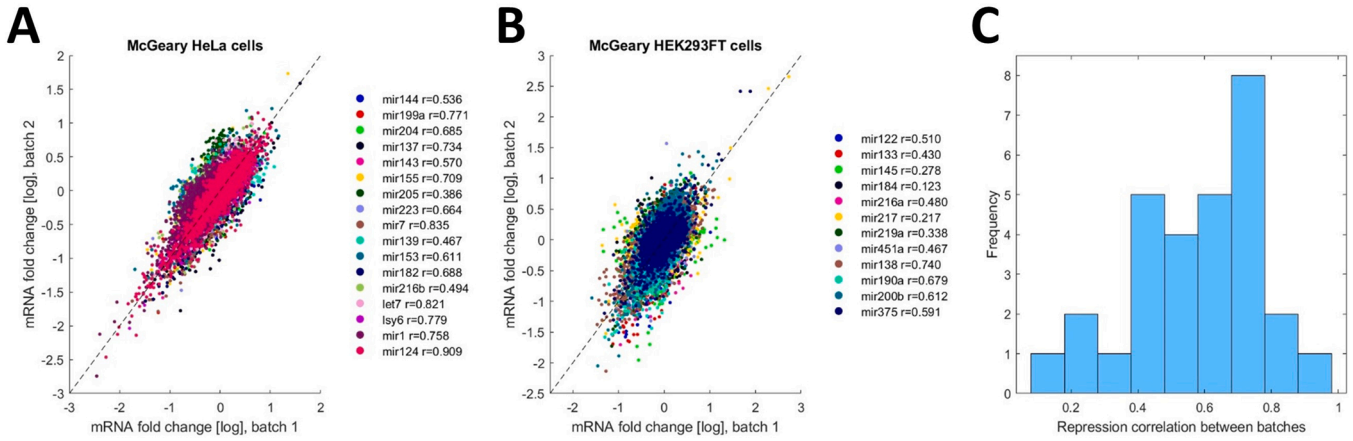

Fig. 2. Experimental Bias When Measuring mRNA Levels. Repression variance between repetition of transfection experiment, for each miRNA transfected (Eq. 1). (A) HeLa cells mRNA expression fold change following Transfection of 14 miRNAs. (B) HEK293FT cells mRNA expression fold change following transfection of 12 miRNAs. (C) Histogram of the repression correlations for both HeLa and HEK293FT cells. Both transfection experiments performed and published by McGeary et al. [46].

2.2. Integrative model outperforms previous models

miRNA-mediated repression holds an important regulatory role, as such we aim to investigate the relationship between miRNA and potential binding sites upon the mRNA. To that end, we used published transfection datasets to train a linear regression model that predicts miRNA-mediated repression given a miRNA and a mRNA.

Our study utilized two main sources of data. The first source consisted of transfection data from HeLa and HCT116 cells published by Agarwal et al. [30]. These datasets encompassed various experiments and publications, which were processed to mitigate bias introduced by batch effects and different experimental contexts. The second data source comprised transfection data from HeLa and HEK239FT cells published by McGeary et al. [46]. Unlike the previous dataset, this batch was smaller in size and derived from a single experimental protocol, purportedly offering a higher signal-to-noise ratio (SNR) compared to previous experiments. Each cell expression data was divided into training and test sets, enabling performance analysis of all models on both the same cell type and data source, as well as on different cell types and data sources, as depicted in Fig. 1.

First, for each (miRNA, mRNA) pair, we conducted a search for potential binding sites by examining complementary base pairing to the seed region of the miRNA, adhering to the pairing rules of canonical interactions. Subsequently, we filtered the pool of potential sites by retaining only those located in the 3'UTR and the last third of the ORF, as previous studies have demonstrated that sites in other regions have minimal impact on mRNA repression [10,11,46]. To isolate the contribution of each site to the mRNAs' repression, the training data consisted solely of (miRNA, mRNA) pairs with a single potential canonical binding site, following the aforementioned guidelines [31].

A feature table was then generated given the site information (see model features in methods section). Here, we combined features from several published models to create a vastly integrative model, in the aim of adding information that most accurately depicts the in-cell repression mechanism (Table 1). Many of these features are considered traditional in computational biology, including thermodynamic features, sequence features, and evolutionary features used in models such as TargetScan [30], miRmap [27], TarPmiR [24] and MIRZA-G [23]. In an effort to enhance prediction sensitivity, we incorporated features proposed by Bergman et al. [31], which emphasize the importance of ORF binding sites and potential interactions between the miRNA-AGO complex and the active ribosome. Additionally, we expanded our feature list by including biochemical features calculated using predictive tools developed by McGeary et al. [46]. One of these features, Kd, represents the affinity between the mRNA binding site and flanking nucleotides and the miRNA, estimated using a CNN algorithm. This feature has been shown to exhibit a high correlation with miRNA-mediated repression.

Finally, the model was trained on each training set using elastic net regularization (see model training in methods section). Given that both the type of miRNA-mRNA canonical interaction and the region of

Table 1
Features Used in Integrative Model.

Category	Feature	Previous Model
Thermodynamics	ΔG Binding/ΔG Binding Seed	miRmap
	ΔG Duplex/ΔG Duplex Seed	
	ΔG Open	
	Seed Pairing Stability	TargetScan
	Structural Accessibility	
	MicroRNA Recognition Elements	Bergman et al.
Biochemical	Kd miRNA-mRNA Affinity	Biochemical+
	Steady State Occupancy	
Sequence	AU Content	TargetScan
	3' Pairing	miRmap, TarPmiR
	Distance Score	TargetScan
	Relative Distance	TarPmiR
	Nucleotides in miRNA and target site	TargetScan
	Regions Length	
	Nucleotide Frequency	MIRZA-G
	Non-Canonical Sites	TargetScan
	Target Abundance	
	PUM Sites	Bergman et al.
	RNA Binding Proteins	
Evolution	phastCons/phyloP: Seed/Site/Flank	miRmap
	phyloP 2-6	
	Binomial/Exact Probability	
Translation	CAI	Bergman et al.
	Amino acid charge	
	Slow amino acids	
	Typical Decoding Rate	
	tAI	
	Ribosomal density	

interaction have been shown to impact repression, each model is specific to these parameters, resulting in eight formulas (2 regions: {ORF, 3'UTR}, and 4 seed types: {6mer, 7mer-A1, 7mer-m8, 8mer}). Therefore, each submodel was trained to target a specific site type based on the region of the site and its canonical sequence. In the final model, potential canonical binding sites are searched for each (miRNA, mRNA) pair. The model then calculates the predicted repression of each site based on these characteristics and sums the contribution of each site to determine the total effect on the mRNA.

To assess the performance of the different models, a test set was extracted from each cell type. Each test set comprised a unique set of mRNAs and miRNAs, not shared with the training set. This process involved Bootstrap resampling, repeated 100 times, resulting in 100 different models per cell type. Each model was trained on a unique random subgroup of miRNAs and mRNAs, and then used to predict miRNA repression on its corresponding test set (only (mRNA, miRNA) pairs with at least one 7–8nt site in the 3'UTR were included [30,31]). The performance of each model was assessed using the Pearson correlation coefficient (r^2) between the measured repression and the repression predicted by each model on the test set (Fig. 3). To determine whether our integrative model improved upon our previous model, we compared the performance (Pearson r^2) of each integrative model to the Bergman et al. [31] model trained on the same subset. Given that different experimental contexts introduce different dataset biases [30], models were tested on the same type of cell sourced from the same dataset.

Integrating models and tools created by different research groups and data sets provides a more comprehensive and accurate representation of the miRNA silencing process. As such, adding even a small number of features has the potential to improve the performance of the model. Upon comparing performances of previous models for each dataset, we found that our models exhibited comparable performance or improved upon existing models by up to ~10% on the model trained on McGeary HEK293FT cells (Fig. 3; $p = 0.0287$ $p = 3.46e\text{-}13$ $p = 0.0603$ for the McGeary HEK293FT set, the Agarwal HeLa set and the Agarwal HCT116 set respectively).

To assess our model on a foreign dataset, each of the 100 submodels was tested on the remaining datasets, with the exception of the McGeary HeLa dataset to prevent bias due to same cell type as well as bias toward the biochemical features which were trained on said set (Fig. 3C). Comparing the performances of the Bergman et al. models to the Integrative model (Pearson r^2), we found that our models were either comparable to previous models with minimal improvement and insignificant differences (less than 2% improvement for the models trained on Agarwal HeLa set and tested on the Agarwal HCT116 set $p = 0.997$, and the models trained on Agarwal HCT116 set and tested on the McGeary HEK293FT set $p = 0.100$) or significantly improving previous models by up to ~40% ($p = 1.12e\text{-}15$ $p = 4.27e\text{-}8$ for the models trained on the McGeary HEK293FT sets and tested on the Agarwal HeLa set and Agarwal HCT116 set respectively, $p = 3.20e\text{-}18$ for the models trained on the Agarwal HeLa set and tested on the McGeary HEK293FT set, $p = 5.79e\text{-}9$ for the models trained on the Agarwal HCT116 set and tested on the Agarwal HeLa set).

Finally, we trained a model on the complete Agarwal HeLa dataset (i.e. 100% of the Agarwal HeLa dataset was used to train model) and tested its performance on the Agarwal HCT116 cells and McGeary HEK293FT cells (Fig. 3D). The integrative model exhibited the most predictive performance, yielding results comparable to or higher than those of the Bergman et al. model, achieving a correlations of 0.384 and 0.437 for the Agarwal HCT116 cells and McGeary HEK293FT cells, respectively. This marks a 5.42% and 0.31% improvement over the Bergman et al. model. Given that the quantity and quality of data define the upper limit of machine learning model's performance, and considering the overlap between the current Integrative model and the previous Bergman et al. model, even a small improvement may aid in exposing new intricate relevant biological aspects contributing to the repression mechanism.

When assesing the contribution of the novel biochemical features: Kd, and occupancy (occ), we obsereved that these features were robustly selected across 3'UTR sites. In many casesn they ranked among the 20 highest selected features for the model (based upon the frequency chosen in the cross-validation sets, Fig. 4, Table 2). Reasoning that features robuslty selected would be the most predictive, their placement among the more classic features emphasizes their contribution. It's noteworthy that although features related to affinity already exist in the feature set we trained, the Kd and occupancy features are based on a unique set of biochemical experiments, thus adding novel information to our feature set. Additionally, it's important to highlight that the occupancy feature is predicted using the Kd feature, making them dependent on each other. Therefore, in most cases, only one of the two features was selected for inclusion in the model.

2.3. *Cell specific models improve predictability*

Creating cell specific models has the potential to improve prediction power, as different subgroups of miRNAs are active in different tissues. To evaluate the performance of cell specific models for each validation set, we compared models trained on the same cell type as the test to those trained on a different cell type. Model performance was assessed based on the Pearson correlation between the measured repression and the repression predicted by the model, and compared using a right-tailed Wilcoxon signed-rank test.

It is well known that models trained on a comprehensive number of features may be prone to overfitting. Managing this through feature selection may help create models that are not overfitted to the data yet are more relevant to certain cell types. If our features include miRNA-mRNA interaction signals that are cell-specific, we may develop more robust models that dipict miRNA interaction more accurately. To minimize experimental bias, this analysis was conducted solely on the McGeary datasets, as they were generated using the same process and by same group. Features were then processed by filtering out any features that have a bias towards a certain cell type (Supplementary Table S1). Finally, since the four sets differ in the number of transfected miRNAs, the number of potential target sites varied greatly as well, ranging

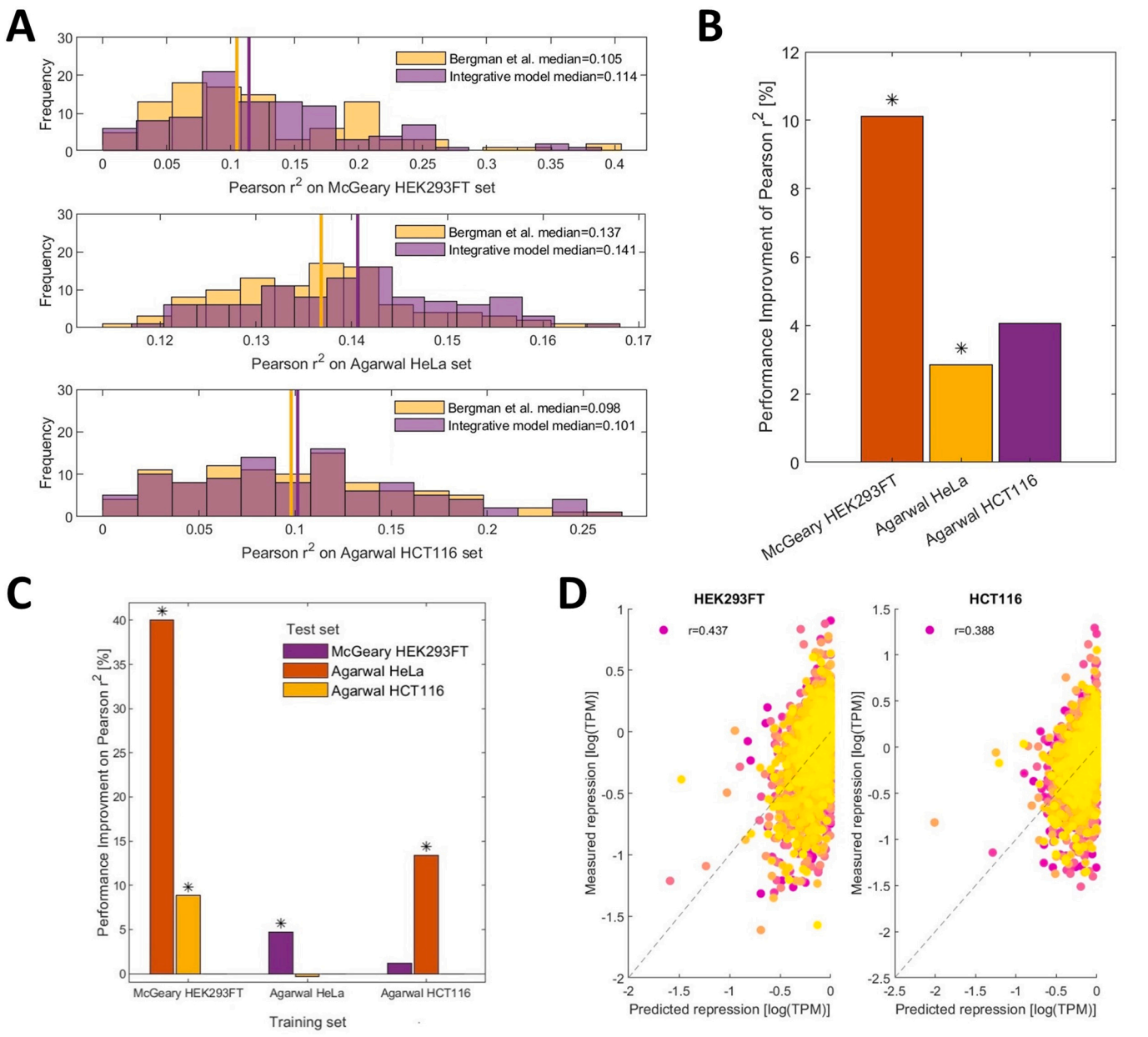

Fig. 3. Integrative model performance on test set and foreign data sets. Model performance evaluation. (A) Pearson correlation distribution of submodels comparing predicted repression and measured repression on the test set. For each dataset a random 80% of the corresponding mRNAs and miRNAs were taken as a training set, and the other 20% as the test set for a total of 100 Bootstrap resamplings. A submodel trained on the training set was used to predict the test sets' repression, and then compared to the measured repression using Pearson r^2 correlation. This process was repeated using Bergman et al. model on the same test set. The median correlation r^2 is shown as vertical line. (B) Model's mean improvement compared to Bergman et al. model using a right-tailed Wilcoxon signed-rank test on Pearson r^2 on test set. The difference (in percentage) of each model is shown in each bar. Models with a significant result ($p < 0.05$) are marked with * . (C-D) Model performance evaluation on different cell types. (C) Model's mean improvement compared to Bergman et al. model using a right-tailed Wilcoxon signed-rank test on Pearson r^2 on foreign datasets. The difference (in percentage) of each model is shown in each bar. Models with a significant result ($p < 0.05$) are marked with * . (D) Scatter plot of measured repression versus predicted repression of the 2 different datasets (McGeary HEK293FT cells and Agarwal HCT116 cells), when using the final Agarwal HeLa Integrative model (trained on 100% of the Agarwal HeLa set).

between a little under 22k canonical sites for the McGeary HEK293FT cells, and over 209k canonical sites for the Agarwal HeLa cells. Since training size difference impact model training, we found its best to compare an adjusted training size for each cell type, using 50% of the mRNAs and miRNAs as the training set for the McGeary HeLa cells, vs 70% for the McGeary HEK293FT cells, resulting in an estimated 5.7k sites.

In our comparison, we obsereved that both McGeary HeLa cells and HEK293FT cells showed better results when evaluated on the same cell type (Fig. 5). These results were significant for the both the HeLa cells and HEK293FT cells ($p < 10^{-15}$, $p < 0.05$ respectively) leading to a notable ~15–70% improvement in Pearson r^2. Given the value in creating cell specific models and understanding the differences between them, further investigation may contribute to a deeper understanding of miRNA functionality.

2.4. *miRNA site distribution*

Until now, interactions between proximal binding sites have been deemed insignificant, leading to the assumption that they act

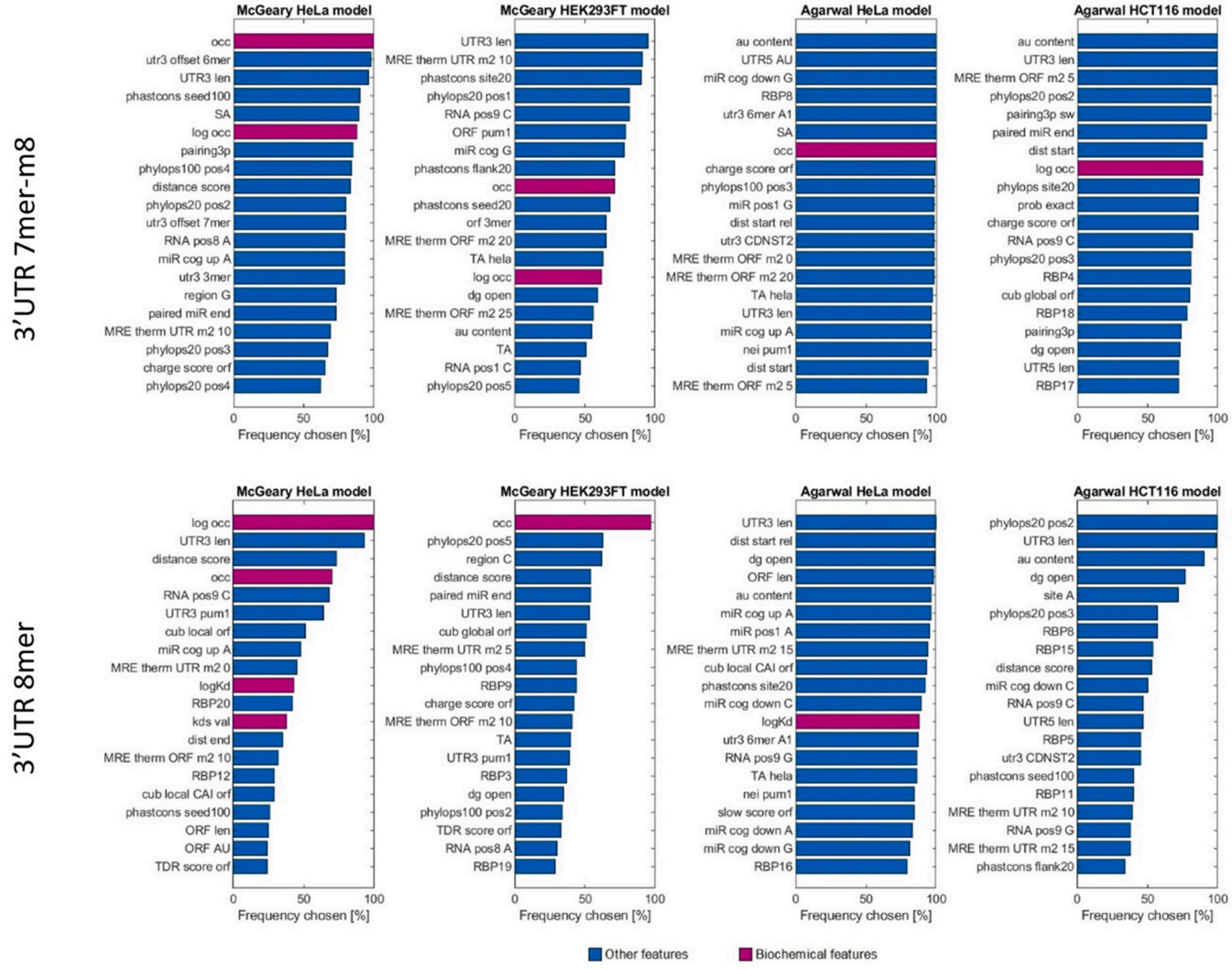

Fig. 4. Importance of biochemical features. Feature importance of the novel biochemical features (Kd, log Kd, occ, log occ) compared to the ranking of the rest of the 152 features. Features were ranked according to their feature selection frequency. Presented here the top 20 features for 3'UTR sites for a seed type of 7mer-m8 and 8mer. An extended list of features presented in Table S1.

independently, with their cumulative effect being equal to the sum of each partial repression. This hypothesis simplifies the development of computational tools, as considering binding site interaction adds another layer of complexity to an already complex problem. However, taking into account binding site interaction may result in a model that more accurately represents the silencing mechanism in the cell, particularly considering that most mRNAs are regulated by multiple miRNAs.

When considering binding site interactions, we expect to observe one of the following phenomena: (a) competition, (b) cooperation, or (c) indifference. Due to limited resources in the cell, sites may compete for AGO-loaded miRNA, resulting in an impact on the mRNA that is less than the sum of each potential site's contribution. This phenomenon has been previously reported on a genome-wide scale, under the hypothesis that certain mRNAs regulate miRNAs by acting as sponges, redirecting miRNA activity [48,49]. Similarly, competition with endogenous miRNAs has been proposed to affect miRNA transfection experiments, leading to an indirect change in mRNA levels in the cell that is not solely due to the transfected miRNA repression activity [42,45]. Just as competition is prevalent among endogenous and transfected miRNAs, and among different mRNAs for the same miRNA, competition may also occur among different binding sites for the same (miRNA, mRNA) pair. This competition may be exacerbated in cases where weaker sites attract the miRNA, diverting them from stronger sites.

Surprisingly, it has been demonstrated that a short distance between binding sites (up to 35 nt spacing) may enable cooperative binding interactions, leading to increased repression compared to the sum of each potential site [48,49]. Theorizing that this distance influences the recruitment of relevant proteins to the local area, enhancing the effectiveness of the degradation process by stabilizing the miRNA-mRNA complex [33,34,50,51]. This distance may facilitate better allocation of resources toward the functional sites area, creating a synergistic effect between these sites [29,33].

As both competition and cooperation may occur in a distance-dependent manner, we opted to investigate the distribution of binding site distances in our largest dataset, the Agarwal HeLa cells, and compare this distribution to a randomized HeLa genome distribution (see random genome generation in methods section). Here, we observe that sites tend to have a significantly shorter distance between them than expected by chance (Fig. 6; $p < 10^{-325}$), suggesting a preference for the cooperative mode. However, only ~8% of the sites are in a cooperative allowing distance from one another (<35nt). As suggested before, this is a conserved spacing between sites, allowing for optimal cooperation [33,34,52]. Therefore, we anticipate that our data will be predominantly influenced by competition, with minimal cooperation among sites.

Table 2
Feature selection frequency of biochemical features across the different models.

Training Dataset	Region	Feature	Frequency Chosen			
			6mer	7mer-A1	7mer-m8	8mer
Agarwal HeLa	ORF	Kd features	33%	2%	6%	2%
		Occupancy features	77%	42%	80%	13%
	3'UTR	Kd features	27%	37%	12%	95%
		Occupancy features	99%	100%	100%	92%
Agarwal HCT	ORF	Kd features	32%	2%	1%	3%
		Occupancy features	56%	18%	90%	14%
	3'UTR	Kd features	11%	82%	9%	6%
		Occupancy features	100%	17%	100%	28%
McGeary HeLa	ORF	Kd features	88%	73%	42%	50%
		Occupancy features	100%	91%	58%	29%
	3'UTR	Kd features	53%	98%	62%	53%
		Occupancy features	100%	99%	100%	100%
McGeary HEK	ORF	Kd features	21%	9%	9%	5%
		Occupancy features	38%	16%	14%	29%
	3'UTR	Kd features	67%	28%	22%	3%
		Occupancy features	98%	67%	95%	99%

Highlighted are sites where at least one of the features was ranked among the top 20 most chosen features.

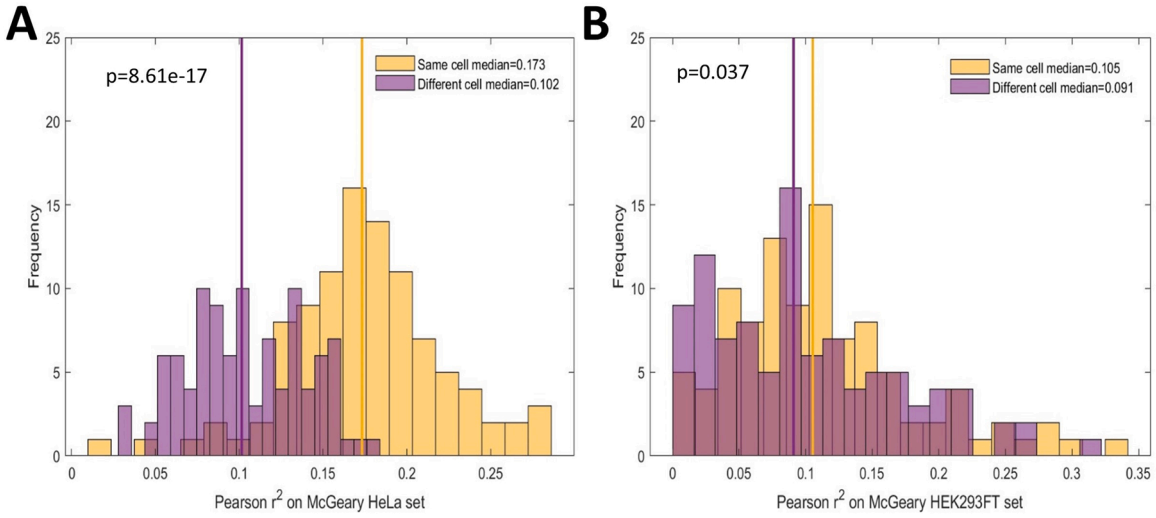

Fig. 5. Cell Specificity Performance. A performance assessment of cell-specific models. Each submodel trained was used to predict the validation sets' repression and compared to the measured repression. Then, we created two groups: the first holds models performances that were trained on the same type of cell as the test set; the second holds models performances that were trained on a different type of cell. We compared the two groups' performances using a right-tailed Wilcoxon signed-rank test on Pearson r^2 to determine which groups overall performance was better. Here, we compared between models trained on the McGeary HeLa cells, and models trained on the McGeary HEK293FT cells. for this analysis, 100 models were trained on 100 subgroups of mRNA and miRNA per dataset. Each model was then used to predict repression on its corresponding test set, as well as the 100 test sets of the other cell type. (A) Model performance on the McGeary HeLa cell dataset, the models trained on this type of cell outperformed models trained on the HEK293FT cells significantly ($p < 10^{-17}$). (B) Model performance on the McGeary HEK293FT cell dataset, the models trained on this type of cell outperformed models trained on the HeLa cells significantly ($p < 0.05$).

2.5. miRNA binding site interaction model improves performance

As it has been suggested in the past, miRNAs binding sites interact with each other in various ways. On one hand, there's evidence supporting competition between binding sites on different mRNAs, as well as competition between different miRNAs for factors involved in the silencing process. Since cellular resources are finite, sites may compete for the AGO loaded miRNA. On the other hand, miRNA binding sites may cooperate to achieve more efficient silencing, however it requires adjacent target sites. Following these guidelines, we presume distance is

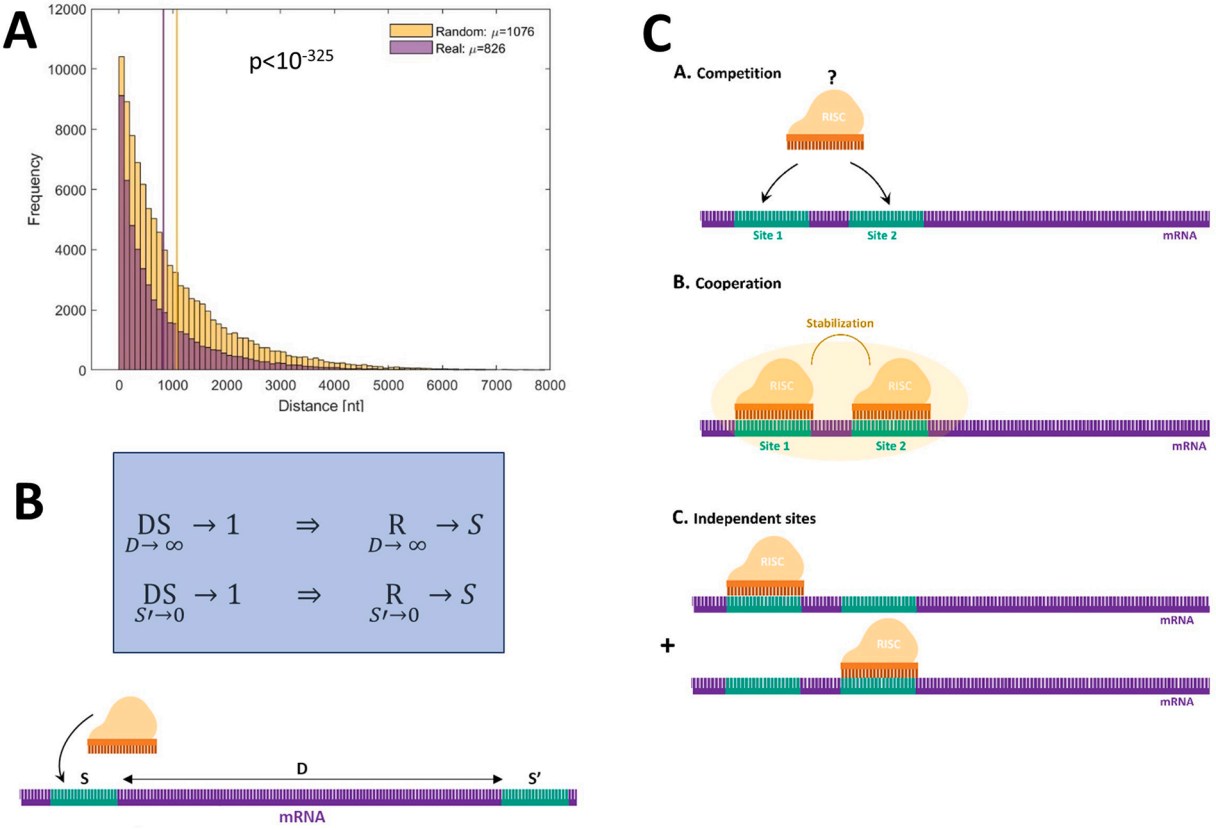

Fig. 6. miRNA Target Site Distance Distribution and miBSIM Schematic. (A) Nearest canonical target site distance distribution of the Agarwal HeLa training dataset compared to a randomized HeLa genome. For each target site, the nearest site's distance was calculated, resulting in this distribution. Only miRNA-mRNA pairs with multiple sites were considered for this analysis. Here we compared between the distance distribution of sites upon the Agarwal HeLa genome and the subsequent 74 miRNAs, to the distance distribution of a randomized HeLa genome and same miRNA set via Wilcoxon rank sum test. The median distance of the real genome was significantly shorter compared to the randomized genome ($p < 10^{-325}$). (B) Parameters used in the correction factor DS equation (Eqs. 3–10). We based our model following the idea adjacent binding sites will interact, and this interaction may be contingent on the distance between the sites D, as well as the current sites nominal repression S and the neighbor sites' nominal repression S'. We modeled DS assuming that nonadjacent sites will have negligible interaction between them, as well as sites that have little to no repression and affective interaction with the miRNA. (C) Guiding principles behind the miBSIM, illustration of the interaction between adjacent sites. We expect to see one of the three following interactions between adjacent sites
: a. competition, binding sites will compete for the limited resources and therefore the total effect will be less than the sum of each predicted site. b. cooperation, binding sites will cooperate guiding resources to the most optimal site and increasing local stability allowing for greater repression than the sum of each predicted site. c. independence, sites will act independently from one another, multiple sites increase the probability of repression but will not influence each other, hence the repression will be the sum of each sites predicted contribution.

a strong factor in determining the nature of the interaction, as well as the affinity of each site to the miRNA.

To incorporate this interplay, we have added a computational step that follows the integrative model's prediction, which includes neighboring sites' information (see miBSIM training in the methods section). For each potential site, we calculate a correction factor, DS, and use that to scale the prediction of our integrative model (Eqs. 2–10). This correction factor is dependent on the neighboring sites' predicted repression and the distance between the sites. It follows the logic that nonadjacent sites will have little to no interaction, as well as sites with very low affinity to the miRNA. DS was then optimized using the Hill Climbing algorithm on the Agarwal HeLa dataset to adjust the equation's constants for each of our training sets, starting from 100 different starting points (see Hill Climbing algorithm in the methods section). We repeated this process for 8 different formulas of DS with slight variations among them, all of which follow the concepts mentioned above (Eqs. 3–10). To allow both competition and cooperation in our model, we varied our starting points to include both positive and negative constants, which change the correction factor to either enhance the current prediction or reduce it. For each training set, the best-performing constants were chosen to be evaluated as the final model.

The correction factors DS_i were all designed to follow the same logic under similar constraints, allowing for either competition or cooperation determined by the optimized constants. Starting with all possible correction factors DS_i (Eqs. 3–10) for 25 out of the 100 training sets described in the training process of the Integrative model and evaluating their performance, we observe that DS_2 and DS_3 achieved the best improvement on the test set compared to the Integrative model, while the other models resulted in lower and sometimes negative improvements perhaps due to overfitting (Fig. 7A). Both DS_2 and DS_3 consider information from all neighboring sites, which might point to a more complex network of interactions between sites, favoring or possibly avoiding certain binding site grouping patterns. Most interestingly, DS_8, which divides the interactions into 2 subgroups and therefore might capture a cooperative interaction among adjacent sites and competitive interaction among distant sites, did not perform well. It is possible that in this case the model was overfitted, reasoning that in this case 5 parameters were needed to be optimized and perhaps the training set size was insufficient. In addition, only ~8% of sites among our data account for proximate sites; as such it is possible there was not enough representation among the chosen training set to accurately optimize proximate parameters.

Taking into account the suggested correction factor, we focused on developing DS_3 as its initial results were the best when considering

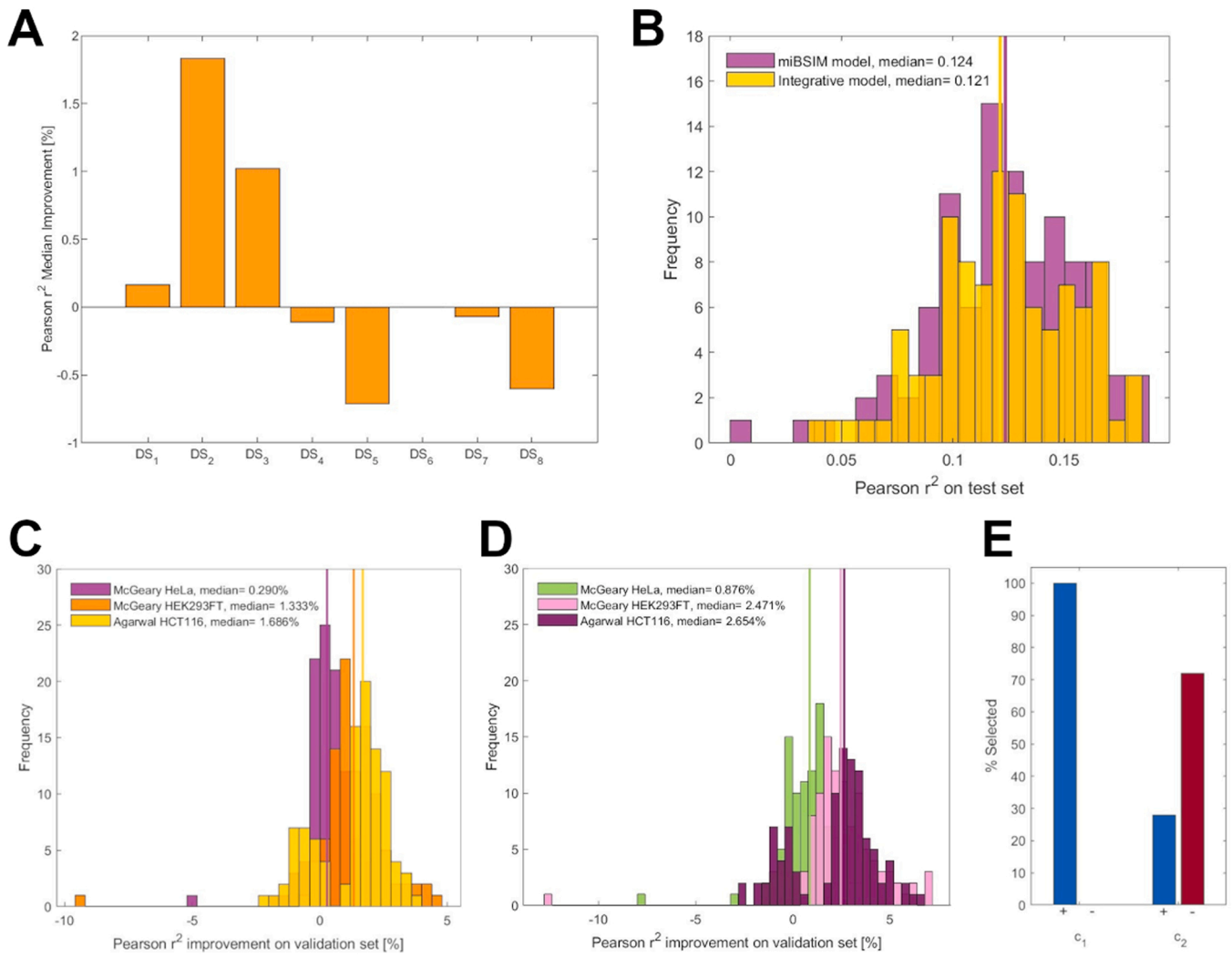

Fig. 7. miBSIM's training Performance. miBSIM (DS_3, Eq. 5) was trained on the Agarwal HeLa set, using the same training and test partitions as used previously to train the Integrative model (totaling in 100 different training and test sets). For each training set we optimized our constants c_1 and c_2 using Hill climbing (see Hill climbing algorithm in methods section), starting from 100 different uniformly distributed starting points: c1 [− 0.75,0.75], c2 [− 15000, 15000]. The final model for each training set was chosen as the best performing constants. (A) Improvement upon Integrative model for alternative correction factors (Eqs. 3–10) on the test set. Here we assessed 25 different training and test sets. This step was used to choose the final miBSIM model (DS_3) and to understand the nature of our data. (B) Comparison between the Integrative model's performance to miBSIM's performance on the test set. The miBSIM performance was significantly higher using a right-tailed Wilcoxon signed-rank test on Pearson r^2 ($p = 4.21e\text{-}6$), with a median improvement of 1.02%. (C) Improvement upon Interactive model on foreign datasets. Comparing the Integrative model's performance to the miBSIM's performance on each set using a right-tailed Wilcoxon signed-rank test the miBSIM significantly improved the Integrative model's performance ($p = 1.31e\text{-}5$, $p = 9.39e\text{-}17$, $p = 1.06e\text{-}12$ for the McGeary HeLa cells, the McGeary HEK293FT cells and the Agarwal HCT116 cells respectively). (D) Improvement upon Interactive model on foreign datasets considering only (miRNA,mRNA) pairs that had multiple sites. Comparing the Integrative model's performance to the miBSIM's performance on each set using a right-tailed Wilcoxon signed-rank test the miBSIM significantly improved the Integrative model's performance ($p = 3.41e\text{-}9$, $p = 3.84e\text{-}17$, $p = 2.20e\text{-}13$ for the McGeary HeLa cells, the McGeary HEK293FT cells and the Agarwal HCT116 cells respectively). (E) Trends in model's constants optimization. Frequency of sign(c_i) for all optimized DS_3 models. Blue/Red bars indicate positive/negative constants chosen for the final model.

outliers. For this correction factor, we optimized constants for all 100 training sets. Assessing the miBSIM on the test sets, we observed a median improvement of 1.02% compared to the integrative model (Fig. 7B; $p = 4.21e\text{-}6$ using a right-tailed Wilcoxon signed-rank test on Pearson r^2). When assessing all 100 optimized models on foreign datasets, we achieved a median improvement of 0.290%, 1.33%, 1.69% on the McGeary HeLa, McGeary HEK293FT and Agarwal HCT116 sets respectively (Fig. 7C; $p = 1.31e\text{-}5$, $p = 9.39e\text{-}17$, $p = 1.06e\text{-}12$ using a right-tailed Wilcoxon signed-rank test on Pearson r^2). Albeit significant, these improvements are relatively small. This might arise from the fact that multiple sites do not represent the majority of (miRNA,mRNA) pairs in our sets, as 57.1% of the Agarwal HeLa dataset is comprised of mRNAs that have a single miRNA binding site and these will not be affected by the correction factor. Repeating this analysis including only (miRNA, mRNA) pairs with multiple sites has increased the improvement of miBSIM compared to the integrative model on foreign data by 0.876%, 2.47%, 2.65% on the McGeary HeLa, McGeary HEK293FT and Agarwal HCT116 sets respectively (Fig. 7D; $p = 3.41e\text{-}9$, $p = 3.84e\text{-}17$, $p = 2.20e\text{-}13$ using a right-tailed Wilcoxon signed-rank test on Pearson r^2). Moreover, binding site interactions might explain very little of the repression variance. As such, we will not see a large change in performance due to the contribution of said interactions.

When observing the constants selected by the optimization process, we notice that the final constants varied. While the constant scaling repression strength,c_1, was consistently selected to be positive across all sets, the constant scaling distance,c_2,was mostly negative (Eq. 5,

Fig. 7E). Examining the optimized formula, we see that these trends indicate a negative effect between binding sites, which becomes stronger as the distance between sites is larger. Although competition between sites has been less investigated, our optimization process might suggest this interplay among binding sites. However, this does not necessarily mean that proximate sites do not cooperate. Site cooperativity has been shown to occur mainly within 13–35 nt spacing between sites [33,34,52]. Since only ~8% of our dataset's sites have a neighboring site with a distance under 35 nt, it is possible that the training set had a bias toward competitive interactions that occur at larger distances (Fig. 6B,C).

Finally, the entire Agarwal HeLa set was used to optimize the DS_3 correction factor. When using the final miBSIM model to predict miRNA-mediated repression on the McGeary HEK293FT and the Agarwal HCT116 sets, we obtained correlations of $r = 0.439$ and $r = 0.389$ (Fig. 8B). Comparing these results to previous model performances, including TargetScan 8.0 [30], Bergman et al. [31] and the Integrative model, miBSIM outperformed these models, showing a 0.958% and 0.290% improvement over the Integrative model, a 6.43% and 0.531% improvement compared to the Bergman et al. model, and a 35.9% and 51.6% improvement on TargetScan respectively (Fig. 8A). Following these results, we set to investigate how well our model represents functional miRNA-mRNA interactions. For this analysis, we used miRTarBase 9.0 [53], a database of experimentally validated miRNA-mRNA interactions. Maintaining only miRNAs and mRNAs present in both the test sets and the miRTarBase database, we performed a chi-squared enrichment analysis (Fig. 8C). A predicted interaction was defined as a miRNA-mRNA repression prediction of under − 0.1, which represents the average predicted repression for both test sets. In both cases enrichment was significant ($p = 2.24e\text{-}9$, $p = 2.70e\text{-}65$ for the HEK293FT and HCT116 cells respectively), and was robust to different selections of repression thresholds ranging between including all non-zero repression values up until to the 75th percentile of the

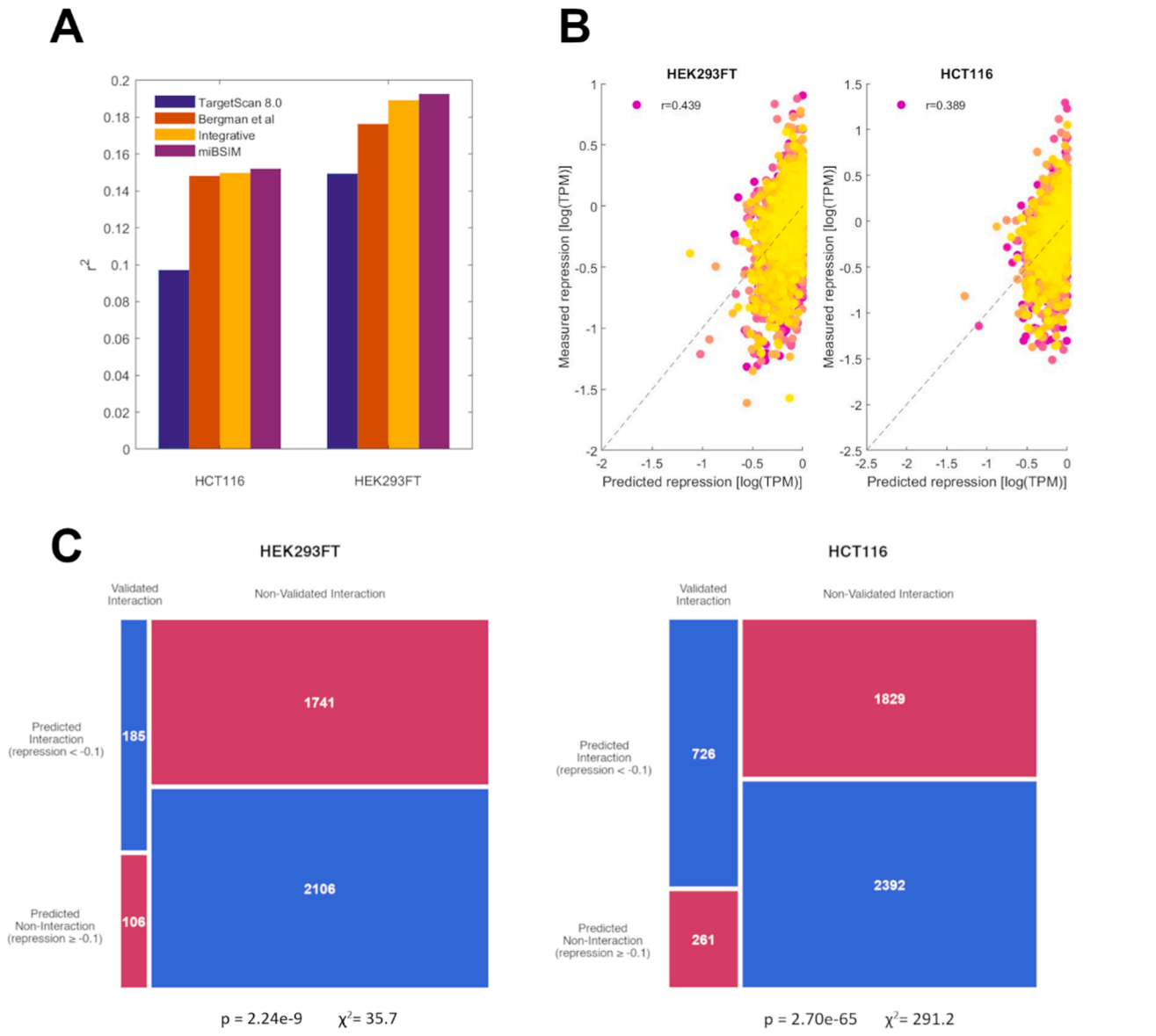

Fig. 8. miBSIM's final model performance. Final miBSIM model, trained on the complete Agarwal HeLa dataset. (A) Performance comparison of the final miBSIM model to previous models: TargetScan 8.0 [30], Bergman et al. [31] and the Integrative model. Results for each model performance tested on McGeary HEK293FT and Agarwal HCT116 cells. (B) Scatter plot of measured repression versus predicted repression of the 2 different datasets (McGeary HEK293FT cells and Agarwal HCT116 cells), when using the final miBSIM model. (C) Enrichment of validated mRNA-miRNA interactions by miRTArBase [53] in the interactions predicted by miBSIM by chi-squared test. Contengency plot of mRNA-miRNA pairs with at least one 7–8nt 3'UTR binding site. True positives and negatives in blue, false positives and negatives in red. Only mRNA and miRNAs appearing in both datasets were considered.

predicted repression (including only higher predicted repression), mirroring a larger amount of false negatives.

3. Discussion

miRNAs play key roles in post-transcriptional gene regulation, including mRNA degradation and translation inhibition. Despite the establishment of a basic set of canonical rules, attempts at unveiling and modeling the intricate details of miRNA-mRNA interactions remains a daunting task. Given their important regulatory functions, their association with numerous diseases, and their potential in therapeutic engineering, understanding all aspects of these interactions is of interest.

Several models have attempted to predict mRNA repression due to miRNA-mRNA interactions. This research area poses specific challenges due to the multitude of biophysical factors influencing mRNA downregulation by miRNAs, including competition or synergism among miRNAs and mRNAs. Moreover, experimental biases such as batch effects and low SNR limit model training and performance, as models cannot surpass the measured variance. As these models provide a strong starting point we created an integrative model, comprising traditional and novel features from established models. By integrating these models, we aimed to highlight different aspects of the miRNA-mediated repression and leverage them for repression prediction. For instance, incorporating ORF sites and features used by Bergman et al. [31] demonstrated the contribution of binding sites located at the end of the coding sequence to repression. Furthermore, the addition of the biochemical feature Kd, predicted by the McGeary CNN tool [46], provided another layer of information by estimating the affinity between the AGO-loaded miRNA and the binding site's sequence. Combining both the traditional works and these novel models has yielded a more robust model capable of predicting miRNA-mediated repression more accurately.

In our attempt of creating cell-specific models, we have demonstrated that models trained on the same cell type as the test data exhibit significantly better predictive performance. This finding aligns with the idea that different cell types may exhibit preferences for distinct features, allowing for the optimization of cell-specific miRNA profiles. One possible explanation for this phenomenon is the influence of mRNA levels within the cell on miRNA activity, thereby affecting the performance of trained features. A hypothesis proposed by Riolo et al. [48] suggests that most mRNA targets act as competitive inhibitors of miRNA, thereby reducing overall miRNA activity [49]. Additionally, it has been suggested that the regulation of the RISC complex association with miRNAs may vary across different cell types, potentially influencing the inhibitory potential of specific miRNAs [54]. Similarly, differences in miRNA levels may indicate varying levels of competition among present miRNAs, including endogenous miRNAs. Moreover, variations in mRNA secondary structure, influenced by numerous factors specific to each cell type, can directly impact accessibility and thermodynamic features. Developing cell-specific models has the potential to enhance predictive power and uncover unique interaction characteristics specific to each cell type. Consequently, investing in cell-specific models could support applications such as cancer subtyping [55] and the development of more effective treatments.

Developing our model further, we have added a computational step that includes interactions between miRNAs, considering for the first time both the distribution and strength of miRNA binding sites along the mRNA. Our model's improvement upon adding this step suggests that binding sites likely operate in a competitive manner, with this signal intensifying as sites become further apart. However, capturing binding site interaction proved to be challenging, primarily due to the prevalence of single sites governing the majority of our data, and among the multiple sites most tend to have large distances between them. Given these hurdles, the observed improvement underscores the value of this additional step. Previous research suggests that a cooperative effect is more pronounced for targets regulated by multiple distinct miRNAs compared to those targeted by identical miRNAs [33]. Further investigation into synergistic interplay, potentially incorporating non-proficient sites [56], could shed additional light on this phenomenon. Given the initially weak signal and the challenges in accurately modelling miRNA activity in vitro, this improvement is particularly noteworthy. It's essential to recognize that the scarcity of high-quality, large-scale experiments and the limited scope of studied tissues present significant challenges in this field. These constraints impede our ability to capture and analyze complex interactions comprehensively and to develop cell-specific models. Further advancements in the field that may allow for the development of tools capable of predicting the function of multiple miRNAs simultaneously, potentially incorporating endogenous miRNA site interaction. binding interaction might be more significant and crucial to understand miRNA-mediated repression.

The models developed in this study are expected to help various biomedical researchers. Clinicians and researchers studying human diseases can utilize this tool to detect meaningful mutations and SNPs that may have phenotypic effects related to disease. For example, a cancerous mutation in a transcript that affects the interaction with miRNAs can significantly increase or decrease the expression of the transcript, which can ultimately influence various characteristics of the cancer cell (e.g. growth rate, immune system evasion, apoptosis), thus directly affecting the survival of the cancer cell. Our models can also be utilized by synthetic biologists to introduce novel miRNA-mRNA interactions into RNA-based therapies. For instance, through the introduction of miRNA binding sites related to non-target tissues (e.g. healthy tissues but not cancer tissue), we can increase the specificity of an mRNA-based cancer therapy. Finally, our models can be used by biophysicists and molecular biologists to study genome evolution and various intracellular biophysical phenomena. For example, by comparing the miRNA site distribution in the genomes of various mammals, one can study the evolution of the dynamics of mRNA-miRNA interactions.

4. Methods

4.1. Repression datasets

miRNA-mediated repression was derived from normalized expression data, which measures the fold change in mRNA levels following the transfection of a miRNA into the cell.

Agarwal datasets: normalized transfection data published by Agarwal et al. [30]. An aggregation of different groups' microarray data processed and normalized by the TargetScan group. Two cell-lines were selected from this database: HeLa cells, encompassing 3912 mRNAs and 74 miRNAs, and HCT116 cells, consisting of 4477 mRNAs and 7 miRNAs. These sets were used as the training and tests sets for the Bergman et al. model [31] respectively.

McGeary datasets: normalized transfection data published by McGeary et al. [46] attained by transfection of synthetic miRNAs followed by RNA-seq. miRNAs were chosen for their sequence conservation, the availability of data examining their regulatory activities, intracellular binding sites, or in vitro binding affinities. Similar to the Agarwal datasets, two cell lines were chosen: HeLa cells, comprising 3958 mRNAs and 17 miRNAs, and HEK239FT cells, comprising 4113 mRNAs and 12 miRNAs. The HeLa dataset was used to fit the Biochemical+ model and train the repression CNN, while the HEK239FT cells were used as the test set for the CNN model. Repression of mRNA m by miRNA t for this set was calculated in the following manner:

$$r_{m,t} = \beta_{m,t} - \overline{\beta_{m,t*}} \tag{1}$$

Where $\beta_{m,t}$ $[\log TPM]$ is its batch-normalized expression, and $\overline{\beta_{m,t*}}$ is its average expression in all other transfection experiments, this is assumed to encompass the mRNA expression baseline [46].

mRNA sequences were extracted following the published transcript

annotations per experiment from Ensembl database (Agarwal HeLa and HCT116 cells, HEK239FT cells), and hg19 NCBI RefSeq database (McGeary HeLa cells).

4.2. Random genome generation

To assess the distances between proximate sites, we compared distance distribution of the Agarwal HeLa dataset to those of a randomized HeLa genome. The process of randomizing the genome involved permuting synonymous codons coding for the same amino acids, thereby preserving codon frequency while retaining the amino acid sequence. Additionally, the UTR regions were randomized by permuting their nucleotides, thus preserving nucleotide frequencies within each region.

4.3. Site detection

Site detection is based on a set of canonical pairing rules comprising four different pairings with the 5' end of the miRNA in the 3'UTR region of the mRNA. These rules have been established through previous miRNA-mRNA interaction studies and supported experimentally [10,18,47,57–60]. The process involves searching the mRNA in the last third of the ORF and 3'UTR for one of the following: (i) a complementary sequence to positions 2–6 nt of the miRNA (6mer), (ii) a complementary sequence to positions 2–6 nt of the miRNA followed by an 'A' (7mer-A1), (iii) a complementary sequence to positions 2–7 nt of the miRNA (7mer-m8), or (iv) a complementary sequence to positions 2–7 nt of the miRNA followed by an 'A' (8mer).

Non-canonical sites, such as interactions with other regions of either the miRNA or mRNA, bulges, and mismatches, are abundant in the human transcriptome [32]. Given the countless combinations of possible interactions, it raises a question of what constitutes a statistically and biologically relevant interaction. Many studies have focused on examining alternative binding sites, broadening the scope of binding possibilities [32,61–63]. However, these non-canonical interactions have been shown to be insufficient for repression and comprise the minority of conserved miRNA targets, perhaps caused by incompatible structural alignment with the RISC complex or due to the site's placement in a less favorable context [10,11,46,56,58,60,64]. Considering the low signal-to-noise ratio (SNR) of the current data, we expect these interactions to fall below detection levels. Therefore, we follow earlier models, regarding these potential interactions as insignificant and focusing solely on canonical interactions [10,30,31,46]. On the other hand, ORF sites have the potential to encompass an interesting interplay between miRNA activity and ribosomal traversal [11]. Bergman et al. showed that models including the last third of the ORF outperform models including sites solely in the 3'UTR or considering both sites in the entire ORF and 3'UTR [31]. Thus, we included sites in the last third of the ORF along with features related to the translation process to better represent this region.

4.4. Model features

The models incorporateda comprehensive list of thermodynamic, conservation, sequence, and translation related features, integrated from a number of established models. The base feature set consists of the features described by Bergman et al. [31] and were based on previous works by TargetScan [30], miRmap [27], TarPmir [24], DIANA—microT—CDS [65] and MIRZA-G [23]. Notably, we introduced two novel features: the binding efficacy Kd and the site occupancy occ predicted by the Biochemical+ and CNN model [46]. The complete feature list appears in Table 1 and extended in the Supplementary (Table S1).

Thermodynamic features were estimated using the ViennaRNA package [66]. Thermodynamic features include binding energy between miRNA and mRNA target site, minimum free energy of the miRNA-mRNA duplex, the energy needed to keep the RNA strand open relative to the RISC complex physical size, seed pairing stability, structural accessibility of binding site, as well as the number of potential non-canonical sites based on predicted binding energy to the miRNA. We expect features representing weak mRNA folding, high site accessibility and features representing strong miRNA-mRNA binding will exhibit positive correlations with binding efficacy [56,67–70].

Biochemical features, particularly site affinity, Kd, have demonstrated a strong correlation with miRNA-mediated repression. Kd values were estimated by a CNN tool published by McGeary et al. [46], which looks at the 12nt sequence comprising the target sites and flanking nucleotides. Kd values represent the dissociation constant fitted to a set of experiments analyzing the interactions between AGO loaded miRNAs and a library of RNA sequences. Trained on the McGeary HeLa cells, the CNN tool predicts the Kd value for any given miRNA and 12nt sequence. Additionally, the Biochemical+ tool [46], also trained on the McGeary HeLa cells, estimates the steady-state occupancy, which represents the average number of miRNA molecules bound to mRNA. Given that these values aim to predict site efficacy, a positive correlation with miRNA-mediated repression is expected.

Sequence features include the miRNA, mRNA and the target sites sequence patterns. These features include AU content, length of each region, frequency of each base, pairing of the 3' end of the miRNA relative to the target site, distance between the site and the start of the region (ORF/3'UTR), number of non-canonical sites, target site abundance in a reference set of mRNAs, and number of sites of RNA-binding proteins (RBPs). This family of features are the most diverse in terms of biological interpretation. For instance, features such as high AU content, short region length, and short distance to end of region, often indicate higher site accessibility resulting in a positive correlation with binding efficiency [11,30,43,44,60,65,67,71,72]. Conversely, other features such as non-canonical sites or RBPs, as well as target site abundance, may contribute to the destabilization of the mRNA, accelerating the degradation process, or act in competition to the examined canonical site [30,31,56,73,74].

Evolution features hold an important place, as miRNA target sites are often highly conserved [10,75]. Our model incorporates two conservation scores: (a) phastCons and (b) phyloP, obtained from [76]. These scores, assigned per nucleotide, compare the Homo sapiens genome to either 99 or 19 vertebrates using Multiple Sequence Alignment (MSA) and a phylogenetic Hidden Markov Model (HMM). Additionally, we utilize the binomial probability of the site, based on the binomial distribution, as well as the exact probability using the Spatt program [77]. We anticipate that highly conserved sites will facilitate more effective interactions with miRNAs, indicative of a selection for a more efficient interaction context.

Translation features were incorporated both for a window around the site (applicable for ORF target sites) and the entire ORF. Included are the codon adaptation index (CAI), amino acid charge, typical decoding rate (TDR), tRNA adaptation index (tAI) as well as the presence of codons the code for Proline and Aspartic Acid as they been shown to halt the ribosome. These features may suggest ribosomal traversal affects the miRNA-mediated repressions efficiency [11,31,78]. Since ribosomal speed has been shown to correlate positively with binding efficiency [31], we expect that features contributing to more efficient ribosomal translation (such as CAI, TDR, tAI) will correlate positively with binding efficiency, while features slowing down the ribosome (such as amino acid charge, Proline+, Aspartic Acid+) will correlate negatively with binding efficiency.

4.5. Integrative model training

We used a linear regression model trained with Elastic-net regularization, which offers built-in feature selection and reduces training bias. For each type of miRNA binding site, we trained a separate computational model categorized by region (ORF, 3'UTR) and site type (6mer, 7mer-A1, 7mer-m8, 8mer), resulting in eight unique models per dataset.

Because Elastic-net regularization is not deterministic, we conducted multiple training iterations and selected the model that performed best on the test set.

4.6. miBSIM training

To incorporate binding site interaction into our model, we designed an additional scaling step that incorporates information about neighboring sites. We utilized the same training and test sets as the base model, but this time the training set was not filtered, containing multiple sites per (miRNA,mRNA) pair. We start by using our basic model to predict nominal repression of each site S. Next, we calculate a correction factor per site, and multiply it by the nominal site repression S essentially scaling it (Eq. 2):

$$R = DS_i * S \tag{2}$$

The correction factor incorporates the nearest binding site of the same miRNA as described in Eq. 3:

$$DS_1 = 1 - \frac{|S'|}{c_1} e^{-\frac{D}{c_2}} \tag{3}$$

Where S' is the nearest site nominal repression prediction, D [nt] is the distance to the nearest site, c_1 and c_2 are optimized constants. This equation operates under the assumption that distant sites will not interact with each other, meaning that the predicted repression is solely affected by the features utilized in our basic model. Additionally, weak sites will have a low impact on nearby sites, resulting in minimal change to the nominal repression. To optimize our correction step, we used Hill Climbing to adjust the constants in Eq. 3. Since Hill Climbing is susceptible to converging to a local maximum, we used 100 different uniformly distributed starting values for the constants. Starting points were chosen proportionally to the values they scale, $c_1 \epsilon [-1,1]$ $c_2 \epsilon [-5000,5000]$, without constraining them through the algorithm's iterations. Starting steps were 0.5500 for c_1 and c_2 respectively.

In the effort of finding an equation best fitting target interaction, we repeated this process with several additional equations, all following the same concept of cooperativity relying on distance and the repression strength of other sites as described in Eqs. 4–10. First, we used the same concept as described in Eq. 3, but we considered all target sites upon the mRNA (Eqs. 4 and 5). Ideally, the constants here would vary for each site, perhaps affected by distance, or by repression strength. However, such an approach would significantly complicate the model and will require extensive computational resources. Therefore, we simplified this equation by using a single constant for each parameter, leaving two constant c_1 and c_2 to optimize.

$$DS_2 = 1 - \sum_{j}^{N} \frac{\left|S_j'\right|}{c_1} e^{-\frac{D_j}{c_2}} \tag{4}$$

$$DS_3 = 1 - \frac{1}{N} \sum_{j}^{N} \frac{\left|S_j'\right|}{c_1} e^{-\frac{D_j}{c_2}} \tag{5}$$

Where N is the total number of target site among this specific (miRNA, mRNA) pair. Then, we discarded the neighboring site's repression, as perhaps it's not the efficacy of the neighboring site that dominates interaction between sites but distance:

$$DS_4 = 1 - c_1 e^{-\frac{D}{c_2}} \tag{6}$$

Next, we suggest using the relative neighboring sites repression strength using the difference between them:

$$DS_5 = 1 - \frac{|S' - S|}{c_1} e^{-\frac{D}{c_2}} \tag{7}$$

Additionally, we suggested using a power formula replacing the exponential we used previously:

$$DS_6 = 1 - \frac{|S'|}{c_1} D^{-c_2} \tag{8}$$

Furthermore, we looked at the strongest potential sites and the mean route to that site:

$$DS_7 = 1 - \frac{\max_j |S'|}{c_1} e^{-\frac{\overline{D}_{\max}}{c_2}} \tag{9}$$

Where, $\max_j|S'|$ is the highest repression among all neighboring sites (excluding current target site), $\overline{D}_{\max}$ is the average distance between all sites between the current and highest repressed site.

Finally, we attempted using a more complex algorithm, dividing the formula into two correction factors determined by the distance between the sites:

$$DS_8 = \begin{cases} 1 - \frac{|S'|}{c_1} e^{-\frac{D}{c_2}} & D < TH \\ 1 - \frac{|S'|}{c_3} e^{-\frac{D}{c_4}} & TH \le D \end{cases} \tag{10}$$

Where TH is a threshold distance. Here, 5 parameters were optimized, hence requiring more starting points in the attempt to account for the possibility that distance may change the interaction from a competition to synergy and vice versa. Although all formulas share similarities, each embodies a unique concept. They all aim to leverage the spatial distribution of miRNA binding sites to enhance the repression process, whether by boosting local affinity to the AGO-loaded miRNA or potentially facilitating the transportation of miRNAs from a distant location to a site with greater repression potential. Consequently, we trained multiple models, each utilizing a different correction factor, to determine the best-performing one.

4.7. Hill climbing algorithm

Starting at a given point $\left(c_1^{(0)}, c_2^{(0)}\right)$ and a given performance $r^{(0)}$ (the Pearson r correlation of the current predicted repression and the measured repression of the training set), we calculate the performance of $c_1^{(0)} \pm$ step. If the performance improved- continue to the next point, if not- reduce step size by half and repeat. This is done in iterations until no improvements have been made for more than 10 iterations. Since here we need to optimize in 2 dimensions, each time we have seven points to investigate to find the optimal constants.

4.8. Performance assessment

The performances of miRNA-mediated repression models were assessed using Pearson correlation, comparing the predicted miRNA-mRNA pair log(fold change) to the transfected miRNAs measured log (fold change).

Since the integrative model incorporates feature selection and is highly dependent on the training set, we performed a 100-fold cross-validation. Cross-validation is asymptotically identical to Akaike information criterion (AIC) and can demonstrate that there is no overfitting. All datasets were randomly partitioned into two disjoint sets: a training set, which contained 80% of the mRNAs and miRNAs, and a test set, which comprised 20%. This partitioning process was repeated 100 times to perform cross-validation. Consequently, each dataset consisted of 100 different subsets of train-test pairs, allowing for a robust evaluation of model performance and feature selection assessment that is not due to overfitting.

It is important to note that although other measurements such as root mean square error are commonly used to evaluate linear regression models, assessing the biological relevance of a predictive model,

specifically for miRNA models, traditionally involves using the Pearson coefficient and cross-validation on a set containing no overlap with the training data. Additionally, we expect other measurements to yield similar results, since the sample data is large enough to properly use cross-validation. More importantly, the main models mentioned here-—Bergman et al., TargetScan, Biochemical+ , miRmap, and PUMA—all evaluate their performances using Pearson correlation [27,30–31,37, 46]. Therefore, following their lead, we chose to use these measures, which allows for a direct comparison to previous publications.

Author statement

SB and TT designed the models, analyzed the data, and wrote the paper.

CRediT authorship contribution statement

Sharon Bader: Conceptualization, Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Tamir Tuller:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors, SB and TT, do not have any Conflict of Interest.

Acknowledgments

We thank Shaked Bergman for providing data and his previous model, as well as for his comments and helpful discussions. The study was also supported by the Koret-UC Berkeley-Tel Aviv University Initiative in Computational Biology and Bioinformatics.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.04.015.

References

[1] Bartel DP. Metazoan MicroRNAs. Cell, Vol. 173. Cell Press; 2018. p. 20–51.
[2] Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell 2009;Vol. 136:215–33.
[3] Drury RE, O'Connor D, Pollard AJ. The clinical application of MicroRNAs in infectious disease. Frontiers in Immunology, Vol. 8. S.A.: Frontiers Media,; 2017.
[4] O'Connell RM, Rao DS, Chaudhuri AA, Baltimore D. Physiological and pathological roles for microRNAs in the immune system. Nat Rev Immunol 2010;Vol. 10: 111–22.
[5] Cui Q, Yu Z, Purisima EO, Wang E. Principles of microRNA regulation of a human cellular signaling network. Mol Syst Biol 2006;2.
[6] Ranganathan K, Sivasankar V. MicroRNAs - Biology and clinical applications. Journal of Oral and Maxillofacial Pathology, Vol. 18. Wolters Kluwer Medknow Publications,; 2014. p. 229–34.
[7] Zolboot N, Du JX, Zampa F, Lippi G. MicroRNAs instruct and maintain cell type diversity in the nervous system. Frontiers in Molecular Neuroscience, Vol. 14. S.A.: Frontiers Media; 2021.
[8] Macfarlane LA, Murphy PR. MicroRNA: biogenesis, function and role in cancer. Curr Genom 2010;11:537–61.
[9] Fabian MR, Sonenberg N, Filipowicz W. Regulation of mRNA translation and stability by microRNAs. Annu Rev Biochem 2010;Vol. 79:351–79.
[10] Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell, Vol. 120. Cell Press,; 2005. p. 15–20.
[11] Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 2007;27(1):91–105.
[12] Krützfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, Manoharan M, et al. Silencing of microRNAs in vivo with "antagomirs. Nature 2005;438(7068):685–9.
[13] Bueno MJ, De Castro IP, Malumbres M. Control of cell proliferation pathways by microRNAs. Cell Cycle, Vol. 7. Taylor and Francis Inc.,; 2008. p. 3143–8.
[14] Ambros V. miRNAs found by genomics and reverse genetics. Nature 2004;431: 350–5.
[15] Du T, Zamore PD. microPrimer: the biogenesis and function of microRNA. Development 2005;Vol. 132:4645–52.
[16] Wijnhoven BPL, Michael MZ, Watson DI. MicroRNAs and cancer. Br J Surg 2007; Vol. 94:23–30.
[17] Chen CZ. MicroRNAs as oncogenes and tumor suppressors. N Engl J Med 2005;353 (17):1768–71.
[18] Bartel DP. Review MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004;116:281–97.
[19] Schuck J, Gursinsky T, Pantaleo V, Burgyán J, Behrens SE. AGO/RISC-mediated antiviral RNA silencing in a plant in vitro system. Nucleic Acids Res 2013;41(9): 5090–103.
[20] Hydbring P, Badalian-Very G. Clinical applications of microRNAs. F1000Res 2013; 2:136.
[21] Hanna J, Hossain GS, Kocerha J. The potential for microRNA therapeutics and clinical research. Frontiers in Genetics, Vol. 10. S.A.: Frontiers Media,; 2019.
[22] Sladitschek HL, Neveu PA. Bidirectional promoter engineering for single cell microRNA sensors in embryonic stem cells. PLoS One 2016;11(5).
[23] Gumienny R, Zavolan M. Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. Nucleic Acids Res 2015;43(3):1380–91.
[24] Ding J, Li X, Hu H. TarPmiR: a new approach for microRNA target site prediction. Bioinformatics 2016;32(18):2768–75.
[25] Maragkakis M, Alexiou P, Papadopoulos GL, Reczko M, Dalamagas T, Giannopoulos G, et al. Accurate microRNA target prediction correlates with protein repression levels. BMC Bioinforma 2009;10:295.
[26] Preusse M, Theis FJ, Mueller NS. miTALOS v2: analyzing tissue specific microRNA function. PLoS One 2016;11(3).
[27] Vejnar CE, Zdobnov EM. MiRmap: comprehensive prediction of microRNA target repression strength. Nucleic Acids Res 2012;40(22):11673–83.
[28] Pinzón N, Li B, Martinez L, Sergeeva A, Presumey J, Apparailly F, et al. MicroRNA target prediction programs predict many false positives. Genome Res 2017 Feb 1; 27(2):234–45.
[29] Hon LS, Zhang Z. The roles of binding site arrangement and combinatorial targeting in microRNA repression of gene expression. Genome Biol 2007;8(8).
[30] Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife 2015;4(e05005).
[31] Bergman S, DIament A, Tuller T. New computational model for miRNA-mediated repression reveals novel regulatory roles of miRNA bindings inside the coding region. Bioinformatics 2020;36(22–23):5398–404.
[32] Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell 2013;153(3): 654–65.
[33] Rinck A, Preusse M, Laggerbauer B, Lickert H, Engelhardt S, Theis FJ. The human transcriptome is enriched for miRNA-binding sites located in cooperativity-permitting distance. RNA Biol 2013;10(7):1125–35.
[34] Broderick JA, Salomon WE, Ryder SP, Aronin N, Zamore PD. Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. RNA 2011;17(10):1858–69.
[35] Nam JW, Rissland OS, Koppstein D, Abreu-Goodger C, Jan CH, Agarwal V, et al. Global analyses of the effect of different cellular contexts on microRNA targeting. Mol Cell 2014;53(6):1031–43.
[36] Sood P, Krek A, Zavolan M, Macino G, Rajewsky N. Cell-type-specific signatures of microRNAs on target mRNA expression. PNAS 2006;103(8):2746–51.
[37] Kuijjer ML, Fagny M, Marin A, Quackenbush J, Glass K. PUMA: PANDA using MicroRNA associations. Bioinformatics 2020;36(18):4765–73.
[38] Briskin D, Wang PY, Bartel DP. The biochemical basis for the cooperative action of microRNAs. Proc Natl Acad Sci 2020;117(30):17764–74.
[39] Lai X, Gupta SK, Schmitz U, Marquardt S, Knoll S, Spitschak A, et al. MiR-205-5p and miR-342-3p cooperate in the repression of the E2F1 transcription factor in the context of anticancer chemotherapy resistance. Theranostics 2018;8(4):1106–20.
[40] Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 2010;Vol. 11:733–9.
[41] Elkon R, Agami R. Removal of AU bias from microarray mRNA expression data enhances computational identification of active microRNAs. PLoS Comput Biol 2008;4(10).
[42] Saito T, Sætrom P. Target gene expression levels and competition between transfected and endogenous microRNAs are strong confounding factors in microRNA high-throughput experiments. Silence 2012;3(1).
[43] Hausser J, Landthaler M, Jaskiewicz L, Gaidatzis D, Zavolan M. Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. Genome Res 2009;19(11):2009–20.
[44] Wen J, Parker BJ, Jacobsen A, Krogh A. MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action. RNA 2011;17(5): 820–34.
[45] Khan AA, Betel D, Miller ML, Sander C, Leslie CS, Marks DS. Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. Nat Biotechnol 2009;27(6):549–55.
[46] McGeary SE, Lin KS, Shi CY, Pham TM, Bisaria N, Kelley GM, et al. The biochemical basis of microRNA targeting efficacy. Science 2019;366(6472).
[47] Doench JG, Sharp PA. Specificity of microRNA target selection in translational repression. Genes Dev 2004;18(5):504–11.
[48] Riolo G, Cantara S, Marzocchi C, Ricci C. miRNA targets: from prediction tools to experimental validation. Methods and Protocols, Vol. 4. MDPI AG,; 2021. p. 1–20.

[49] Marques TM, Gama-Carvalho M. Network Approaches to Study Endogenous RNA Competition and Its Impact on Tissue-Specific microRNA Functions. Biomolecules, Vol. 12. MDPI,; 2022.
[50] Lai X, Schmitz U, Gupta SK, Bhattacharya A, Kunz M, Wolkenhauer O, et al. Computational analysis of target hub gene repression regulated by multiple and cooperative miRNAs. Nucleic Acids Res 2012;40(18):8818–34.
[51] Gam JJ, Babb J, Weiss R. A mixed antagonistic/synergistic miRNA repression model enables accurate predictions of multi-input miRNA sensor activity. Nat Commun 2018;9(1).
[52] Sætrom P, Heale BSE, Snøve O, Aagaard L, Alluin J, Rossi JJ. Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucleic Acids Res 2007;35(7):2333–42.
[53] Huang HY, Lin YCD, Cui S, Huang Y, Tang Y, Xu J, et al. MiRTarBase update 2022: An informative resource for experimentally validated miRNA-target interactions. Nucleic Acids Res 2022;50(D1):D222–30.
[54] Flores O, Kennedy EM, Skalsky RL, Cullen BR. Differential RISC association of endogenous human microRNAs predicts their inhibitory potential. Nucleic Acids Res 2014;42(7):4629–39.
[55] Xu A, Chen J, Peng H, Han GQ, Cai H. Simultaneous interrogation of cancer omics to identify subtypes with significant clinical differences. Front Genet 2019;10 (MAR).
[56] Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol 2010 Oct;18(10):1139–46.
[57] Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP. Prediction of Mammalian MicroRNA Targets that they could have many more regulatory functions than those uncovered to date (Lagos-Quintana et al. Cell 2003;115:787–98.
[58] Brennecke J, Stark A, Russell RB, Cohen SM. Principles of microRNA-target recognition. : PLoS Biol 2005:0404–18.
[59] Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM. Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. Cell 2005 Dec 16;123(6):1133–46.
[60] Nielsen CB, Shomron N, Sandberg R, Hornstein E, Kitzman J, Burge CB. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. RNA 2007 Nov;13(11):1894–910.
[61] Gumienny R, Zavolan M. Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. Nucleic Acids Res 2015 Feb 18;43(3):1380–91.
[62] van Dongen S, Abreu-Goodger C, Enright AJ. Detecting microRNA binding and siRNA off-target effects from expression data. Nat Methods 2008;5(12):1023–5.
[63] Yilmazel B, Hu Y, Sigoillot F, Smith JA, Shamu CE, Perrimon N, et al. Online GESS: prediction of miRNA-like off-target effects in large-scale RNAi screen data by seed region analysis. BMC Bioinforma 2014;15(1).
[64] Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 2005;433(7027):769–73.
[65] Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG. Functional microRNA targets in protein coding sequences. Bioinformatics 2012;28(6):771–6.
[66] Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA package 2.0. Algorithms Mol Biol 2011;6(26):1–14.
[67] Robins H, Press WH. Human microRNAs target a functionally distinct population of genes with AT-rich 3 UTRs. PNAS 2005;102(43):15557–62.
[68] Ameres SL, Martinez J, Schroeder R. Molecular basis for target RNA recognition and cleavage by human RISC. Cell 2007;130(1):101–12.
[69] Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet 2007;39(10):1278–84.
[70] Long D, Lee R, Williams P, Chan CY, Ambros V, Ding Y. Potent effect of target structure on microRNA function. Nat Struct Mol Biol 2007;14(4):287–94.
[71] Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinforma 2007;8.
[72] Majoros WH, Lekprasert P, Mukherjee N, Skalsky RL, Corcoran DL, Cullen BR, et al. MicroRNA target site identification by integrating sequence and binding information. Nat Methods 2013;10(7):630–3.
[73] Denzler R, Agarwal V, Stefano J, Bartel DP, Stoffel M. Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. Mol Cell 2014;54(5):766–76.
[74] Arvey A, Larsson E, Sander C, Leslie CS, Marks DS. Target mRNA abundance dilutes microRNA and siRNA activity. Mol Syst Biol 2010;6.
[75] Friedman RC, Farh KKH, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 2009;19(1):92–105.
[76] Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, et al. The UCSC genome browser database. Nucleic Acids Res 2003;Vol. 31:51–4.
[77] Nuel G, Regad L, Martin J, Camproux AC. Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data. Algorithms Mol Biol 2010;5(15):1–18.
[78] Gu S, Jin L, Zhang F, Sarnow P, Kay MA. Biological basis for restriction of microRNA targets to the 3′ untranslated region in mammalian mRNAs. Nat Struct Mol Biol 2009;16(2):144–50.