

## RESEARCH ARTICLE SUMMARY

## MOLECULAR BIOLOGY

## The biochemical basis of microRNA targeting efficacy

Sean E. McGeary\*, Kathy S. Lin\*, Charlie Y. Shi, Thy M. Pham, Namita Bisaria, Gina M. Kelley, David P. Bartel†

**INTRODUCTION:** MicroRNAs (miRNAs) are short RNAs that guide repression of mRNA targets. Each miRNA associates with an Argonaute (AGO) protein to form a complex in which the miRNA recognizes mRNA targets, primarily through pairing to sites that match its extended seed region (miRNA nucleotides 1 to 8) while the AGO protein recruits factors that promote destabilization and translational repression of bound targets. The miRNA targetome is vast, involving most mammalian mRNAs, and miRNA regulatory effects are consequential, with severe developmental or physiological defects often observed after deleting a broadly conserved miRNA (or set of paralogous miRNAs). Deeper understanding of these regulatory roles would be facilitated by a better understanding of miRNA targeting efficacy.

**RATIONALE:** In principle, targeting efficacy should be a function of the affinity between AGO-miRNA complexes and their target sites, in that greater affinity for a target site would cause increased occupancy at that site and thus increased repression of the target mRNA. However, the set of measured miRNA-target

binding affinities has been sparse, and standard thermodynamic models of RNA-RNA pairing poorly predict affinities that have been measured. These limitations have prevented construction of an informative biochemical model of targeting efficacy, such that the best predictive performances have instead relied on indirect, correlative approaches. Here, we adapted RNA bind-n-seq (RBNS) and a convolutional neural network (CNN) to study miRNA-target interactions, thereby obtaining the quantity and diversity of affinity values needed to better understand and predict miRNA targeting efficacy.

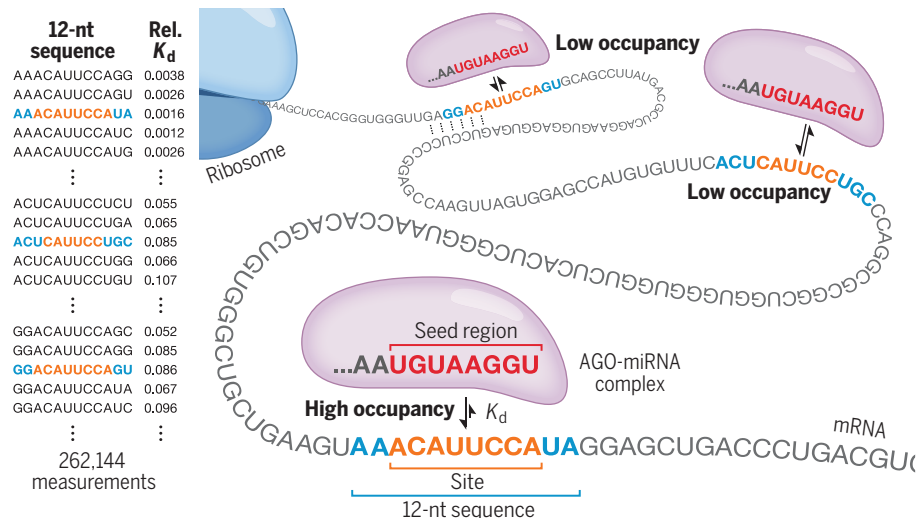
**RESULTS:** Analysis of motifs enriched in RNA sequences bound to the AGO2-miR-1 complex provided unbiased identification of all miR-1 binding sites  $\leq 12$  nucleotides (nt) in length, and a newly developed computational procedure simultaneously inferred the relative dissociation constants ( $K_d$  values) of all of these sites. Repeating this procedure with AGO2 loaded with five other miRNAs (let-7a, miR-7, miR-124, miR-155, and lsy-6) revealed pronounced miRNA-specific differences in the relative affinities of canonical site types (defined as sites

with  $\geq 6$ -nt contiguous matches to the seed region). The analyses also revealed that each miRNA has a distinct repertoire of noncanonical site types and that dinucleotides flanking both sides of each site influence affinity by as much as 100-fold, primarily because of their impact on site accessibility. Most of the noncanonical sites paired to the seed region but did so with imperfections that reduced affinity to levels below those of the top four canonical sites. Nonetheless, for miR-124 and miR-155, noncanonical sites were identified with affinities approaching that of the top canonical site. These high-affinity noncanonical sites were larger and correspondingly rarer in mRNA sequences, which showed that canonical seed pairing is the most efficient way to achieve high-affinity binding.

The miRNA-specific differences in site repertoire and relative binding affinities corresponded to differential repression in cells, thereby enabling

construction of a biochemical model of miRNA-mediated repression. This biochemical model predicts the occupancy at each site as a function of the  $K_d$  measured for the 12-nt sequence encompassing the site. The model outperformed the best correlative model, explaining  $\sim 60\%$  of the relevant variation observed after transfecting a miRNA into cells. Although partly attributable to inclusion of noncanonical sites, the improved performance was primarily due to more accurate representation of the effects of canonical sites. Improved performance was extended to miRNAs without RBNS data by building a CNN that was trained with both RBNS-derived  $K_d$  values and mRNA-transfection fold-change measurements to predict binding affinity between any miRNA and any 12-nt sequence.

**CONCLUSION:** We replaced correlative models of targeting efficacy with a principled, biochemical model that explains and predicts about half of the variability attributable to the direct effects of miRNAs on their targets. The success of the model shows that site binding affinity is the major determinant of miRNA-mediated repression. It also shows that although active AGO-miRNA complexes are occupied primarily by canonical sites, noncanonical sites measurably contribute to repression in the cell. Repression efficacy predicted by this model will be available on the TargetScan website to provide improved guidance for placing miRNAs into gene-regulatory networks. ■



**Biochemical modeling of targeting efficacy.** RBNS generates relative  $K_d$  values for an AGO-miRNA and 262,144 different 12-nt sequences with at least a weak match to the miRNA (left). Values for sites found within an mRNA (colored 12-nt sequences) are used to estimate site occupancy, thereby enabling prediction of mRNA repression. Either a shorter match to the seed region (upper right) or suboptimal flanking nucleotides that promote occlusive mRNA structure (upper middle) can reduce occupancy. Rel.  $K_d$ , relative  $K_d$ .

The list of author affiliations is available in the full article online.

\*These authors contributed equally to this work.

†Corresponding author. Email: dbartel@wi.mit.edu

Cite this article as S. E. McGeary et al., *Science* 366, eaav1741 (2019). DOI: 10.1126/science.aav1741

## RESEARCH ARTICLE

## MOLECULAR BIOLOGY

## The biochemical basis of microRNA targeting efficacy

Sean E. McGeary<sup>1,2,3\*</sup>, Kathy S. Lin<sup>1,2,3,4\*</sup>, Charlie Y. Shi<sup>1,2,3</sup>, Thy M. Pham<sup>1,2,3</sup>, Namita Bisaria<sup>1,2,3</sup>, Gina M. Kelley<sup>1,2,3</sup>, David P. Bartel<sup>1,2,3,4†</sup>

MicroRNAs (miRNAs) act within Argonaute proteins to guide repression of messenger RNA targets. Although various approaches have provided insight into target recognition, the sparsity of miRNA-target affinity measurements has limited understanding and prediction of targeting efficacy. Here, we adapted RNA bind-n-seq to enable measurement of relative binding affinities between Argonaute-miRNA complexes and all sequences  $\leq 12$  nucleotides in length. This approach revealed noncanonical target sites specific to each miRNA, miRNA-specific differences in canonical target-site affinities, and a 100-fold impact of dinucleotides flanking each site. These data enabled construction of a biochemical model of miRNA-mediated repression, which was extended to all miRNA sequences using a convolutional neural network. This model substantially improved prediction of cellular repression, thereby providing a biochemical basis for quantitatively integrating miRNAs into gene-regulatory networks.

MicroRNAs (miRNAs) are ~22-nucleotide (nt) regulatory RNAs that derive from hairpin regions of precursor transcripts (1). Each miRNA associates with an Argonaute (AGO) protein to form a silencing complex, in which the miRNA pairs to sites within target transcripts and the AGO protein promotes destabilization and/or translational repression of bound transcripts (2). miRNAs are grouped into families on the basis of the sequence of their extended seed (nucleotides 2 to 8 of the miRNA), which is the region of the miRNA most important for target recognition (3). The 90 most broadly conserved miRNA families of mammals each have an average of >400 preferentially conserved targets, such that mRNAs from most human genes are conserved targets of at least one miRNA (4). Most of these 90 broadly conserved families are required for normal development or physiology, as shown by knockout studies in mice (1).

Deeper understanding of these numerous biological functions would be facilitated by a better understanding of miRNA targeting efficacy, with the ultimate goal of correctly predicting the effects of each miRNA on the output of each expressed gene. In principle, targeting efficacy should be a function of the affinity between AGO-miRNA complexes and their target sites, in that greater affinity to a target site would cause increased occupancy at that site and thus increased repression of the tar-

get mRNA. Until very recently, binding affinities have been known for only a few target sequences of only three miRNAs (5–11). In a recent study, high-throughput imaging and cleavage analyses provide extensive binding and slicing data for two of these three miRNAs: let-7a and miR-21 (12). Although these measurements provide insight and enable a quantitative model that predicts the efficiency of miR-21-directed slicing in cells (12), the sparsity of binding-affinity data still limits insight into how targeting might differ between different miRNAs and prevents construction of an informative biochemical model of targeting efficacy relevant to the vastly more prevalent, nonslicing mode of miRNA-mediated repression.

With insufficient affinity measurements, the most informative models of targeting efficacy rely instead on indirect, correlative approaches. These models focus on mRNAs with canonical 6- to 8-nt sites matching the miRNA seed region (Fig. 1A) and train on features known to correlate with targeting efficacy (including the type of site as well as various features of site context, mRNAs, and miRNAs), by using datasets that monitor mRNA changes that occur after introducing a miRNA (13–16). Although the correlative model implemented in TargetScan7 performs as well as the best in vivo cross-linking approaches at predicting mRNAs most responsive to miRNA perturbation, it nonetheless explains only a small fraction of the mRNA changes observed upon introducing a miRNA [coefficient of determination ( $r^2$ ) = 0.14] (14). This low value indicates that prediction of targeting efficacy has room for improvement, even when accounting for the fact that experimental noise and secondary effects of inhibiting direct targets place a ceiling on the variability attributable to direct targeting. Therefore, we adapted RNA bind-

n-seq (RBNS) (17) and a convolutional neural network (CNN) to the study of miRNA-target interactions, with the goal of obtaining the quantity and diversity of affinity measurements needed to better understand and predict miRNA targeting efficacy.

## The site-affinity profile of miR-1

As previously implemented, RBNS provides qualitative relative binding measurements for an RNA-binding protein to a virtually exhaustive list of binding sites (17, 18). A purified RNA-binding protein is incubated with a large library of RNA molecules that each contain a central random-sequence region flanked by constant primer-binding regions. After reaching binding equilibrium, the protein is pulled down and any copurifying RNA molecules are reverse transcribed, amplified, and sequenced. To extend RBNS to AGO-miRNA complexes (Fig. 1B), we purified human AGO2 loaded with miR-1 (19) (fig. S1A) and set up five binding reactions, each with a different concentration of AGO2-miR-1 (range of 7.3 to 730 pM, logarithmically spaced) and a constant concentration of an RNA library with a 37-nt random-sequence region (100 nM). We also modified the protein-isolation step of the RBNS protocol, replacing protein pull down with nitrocellulose filter binding, reasoning that the rapid wash step of filter binding would improve retention of low-affinity molecules that would otherwise be lost during the wash steps of a pull down. This modified method was highly reproducible, with high correspondence observed between enrichments for the same 9-nt  $k$ -mers (where  $k$ -mer is any sequence of length  $k$ ) in two independent experiments using different preparations of both AGO2-miR-1 and the RNA library (fig. S1B;  $r^2 = 0.86$ ).

When analyzing our AGO-RBNS results, we first examined enrichment of the canonical miR-1 sites, comparing the frequency of these sites in RNA bound in the 7.3 pM AGO2-miR-1 sample with that of the input library. As expected from the site hierarchy observed in meta-analyses of site conservation and endogenous site efficacy (3), the 8mer site (perfect match to miR-1 nucleotides 2 to 8 followed by an A) was most enriched (38-fold), followed by the 7mer-m8 site, then the 7mer-A1 site, and the 6mer site (Fig. 1, A and C). Little if any enrichment was observed for either the 6mer-A1 site or the 6mer-m8 site at this lowest concentration of 7.3 pM AGO2-miR-1 (Fig. 1, A and C), consistent with their weak signal in previous analyses of conservation and efficacy (4, 14, 20). Enrichment of sites was quite uniform across the random-sequence region, which indicated minimal influence from either the primer-binding sequences or supplementary pairing to the 3' region of the miRNA (fig. S1D). Although sites with supplementary pairing can have enhanced efficacy and affinity (3, 5, 21),

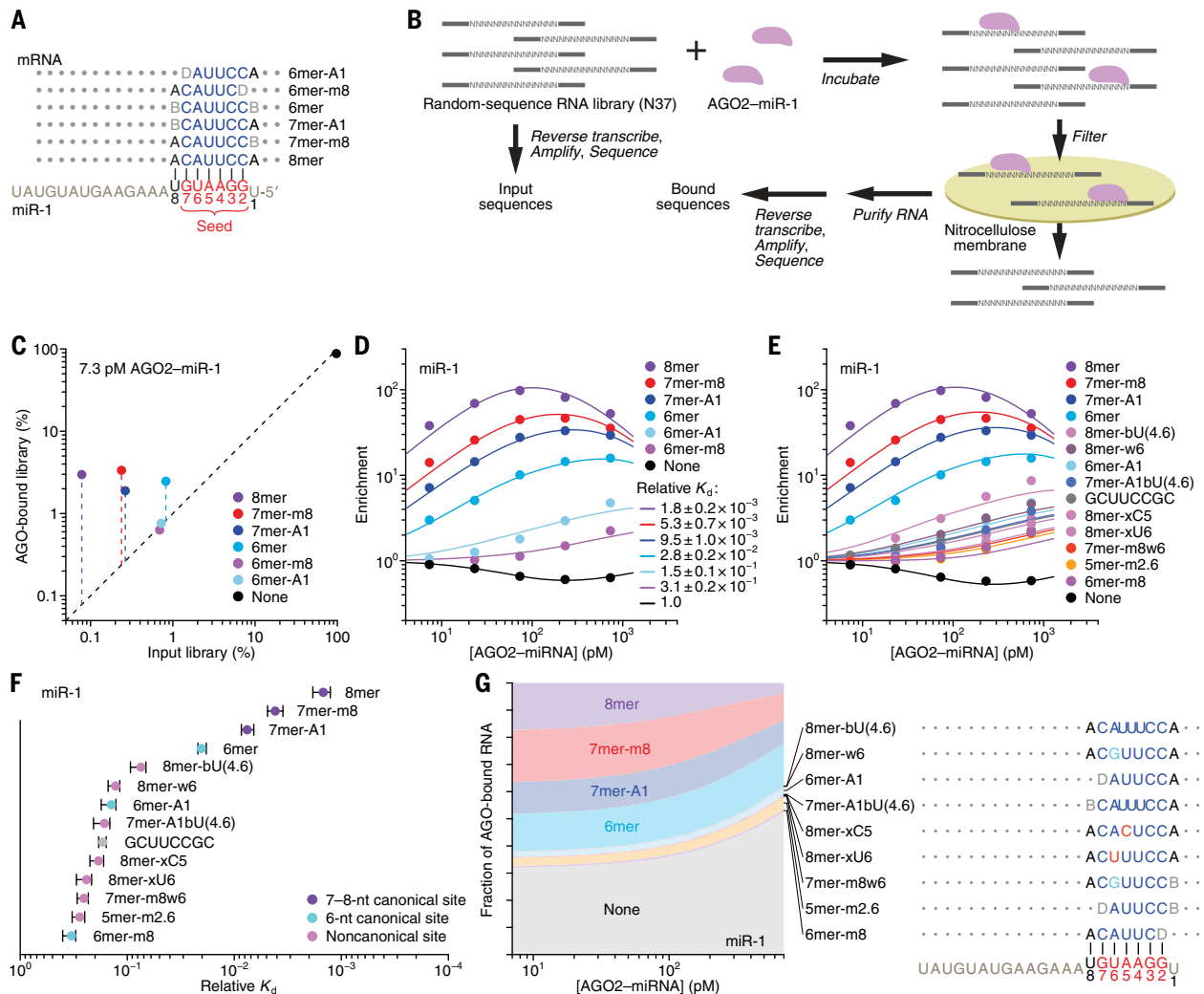
<sup>1</sup>Howard Hughes Medical Institute, Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA.

<sup>2</sup>Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA. <sup>3</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

<sup>4</sup>Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

\*These authors contributed equally to this work.

†Corresponding author. Email: dbartel@wi.mit.edu



**Fig. 1. AGO-RBNS reveals binding affinities of canonical and previously uncharacterized miR-1 target sites.**

(A) Canonical sites of miR-1. These sites have contiguous pairing (blue) to the miRNA seed (red), and some include an additional match to miRNA nucleotide 8 or an A opposite miRNA nucleotide 1 (B represents C, G, or U; D represents A, G, or U). (B) AGO-RBNS. Purified AGO2-miR-1 is incubated with excess RNA library molecules that each have a central block of 37 random-sequence positions (N37). After reaching binding equilibrium, the reaction is applied to a nitrocellulose membrane and washed under vacuum to separate library molecules bound to AGO2-miR-1 from those that are unbound. Molecules retained on the filter are purified, reverse transcribed, amplified, and sequenced. These sequences are compared with those generated directly from the input RNA library. (C) Enrichment of reads containing canonical miR-1 sites in the 7.3 pM AGO2-miR-1 library. Shown is the abundance of reads containing the indicated site (key) in the bound library plotted as a function of the respective abundance in the input library. Dashed vertical lines depict the enrichment in the bound library; dashed diagonal line shows  $y = x$ . Reads containing multiple sites were assigned to the site with greatest enrichment. (D) AGO-RBNS profile of the canonical miR-1 sites. Plotted is the enrichment of reads with the indicated canonical site (key) observed at each of the five AGO2-miR-1 concentrations of the AGO-RBNS experiment, determined as in (C). Points show the observed values, and lines show the enrichment predicted from the mathematical model fit simultaneously to all of the data. Also shown for each site are  $K_d$  values obtained from fitting the model, listing the geometric mean  $\pm$  the 95% confidence interval determined by resampling the read data, removing data for

one AGO-miR-1 concentration and fitting the model to the remaining data, and repeating this procedure 200 times (40 times for each concentration omitted). (E) AGO-RBNS profile of the canonical and the newly identified noncanonical miR-1 sites (key). Sites are listed in the order of their  $K_d$  values and named and colored based on the most similar canonical site, indicating differences from this site with b (bulge), w (G-U wobble), or x (mismatch) followed by the nucleotide and its position. For example, the 8mer-bU(4.6) resembles a canonical 8mer site but has a bulged U at positions that would normally pair to miRNA nucleotides 4, 5, or 6. Everything else is the same as in (D). (F) Relative  $K_d$  values for the canonical and the newly identified noncanonical miR-1 sites determined in (E). Sites are classified as either 7- to 8-nt canonical sites (purple), 6-nt canonical sites (cyan), noncanonical sites (pink), or a sequence motif with no clear complementarity to miR-1 (gray). The solid vertical line marks the reference  $K_d$  value of 1.0 assigned to reads lacking an annotated site. Error bars indicate 95% confidence interval on the geometric mean, as in (D). (G) The proportion of AGO2-miR-1 bound to each site type. Shown are proportions inferred by the mathematical model over a range of AGO2-miR-1 concentrations spanning the five experimental samples, plotted in the order of site affinity (top to bottom), using the same colors as in (E). On the right is the pairing of each noncanonical site, diagrammed as in (A), indicating Watson-Crick pairing (blue), wobble pairing (cyan), mismatched pairing (red), bulged nucleotides (compressed rendering), and terminal noncomplementarity (gray; B represents C, G, or U; D represents A, G, or U; H represents A, C, or U; V represents A, C, or G). The GCUUCCGC motif is omitted because it did not match miR-1 and did not mediate repression by miR-1 (fig. S5B).



the minimal influence of supplementary pairing reflected the rarity of such sites in our library.

Analysis of enrichment of the six canonical sites across all five AGO2–miR-1 concentrations illustrated two hallmarks of this experimental platform (17). First, as the concentration increased from 7.3 to 73 pM, enrichment for each of the six site types increased (Fig. 1D), which was attributable to an increase in signal over a constant low background of library molecules isolated even in the absence of AGO2–miR-1. Second, as the AGO2–miR-1 concentration increased beyond 73 pM, 8mer enrichment decreased, and at the highest AGO2–miR-1 concentration, enrichment of the 7mer-m8 and 7mer-A1 sites decreased (Fig. 1D). These waning enrichments indicated the onset of saturation for these high-affinity sites (17). These two features, driven by AGO–miRNA-independent background and partial saturation of the higher-affinity sites, respectively, caused differences in enrichment values for different site types to be highly dependent on the AGO2–miR-1 concentration; the lower AGO2–miR-1 concentrations provided greater discrimination between the higher-affinity site types, the higher AGO2–miR-1 concentrations provided greater discrimination between the lower-affinity site types, and no single concentration provided results that quantitatively reflected differences in relative binding affinities.

To account for background binding and ligand saturation, we developed a computational strategy that simultaneously incorporated information from all concentrations of an RBNS experiment to calculate relative  $K_d$  values. Underlying this strategy was an equilibrium-binding model that predicts the observed enrichment of each site type across the concentration series as a function of the  $K_d$  values for each miRNA site type (including the “no-site” type), as well as the stock concentration of purified AGO2–miR-1 and a constant amount of library recovered as background in all samples. Using this model, we performed maximum likelihood estimation (MLE) to fit the relative  $K_d$  values, which explained the observed data well (Fig. 1D). Moreover, these relative  $K_d$  values were robustly estimated, as indicated by comparing values obtained using results from only four of the five AGO2–miR-1 concentrations ( $r^2 \geq 0.994$  for each of the 10 pairwise comparisons; fig. S1, F and G). These quantitative binding affinities followed the same hierarchy as observed for site enrichment, but the differences in affinities were of greater magnitude (Fig. 1D and fig. S1C).

Up to this point, our analysis was informed by the wealth of previous computational and experimental data showing the importance of a perfect 6- to 8-nt match to the seed region (3). However, the ability to calculate the rel-

ative  $K_d$  of any  $k$ -mer of length  $\leq 12$  nt (the 12-nt limit imposed by the sparsity of reads with longer  $k$ -mers) provided the opportunity for a de novo search for sites, without bias from any previous knowledge. In this search, we (i) calculated the enrichment of all 10-nt  $k$ -mers in the bound RNA in the 730 pM AGO2–miR-1 sample, which was the sample with the most sensitivity for detecting low-affinity sites; (ii) determined the extent of complementarity between the 10 most enriched  $k$ -mers and the miR-1 sequence; (iii) assigned a site most consistent with the observed  $k$ -mers; and (iv) removed all reads containing this newly identified site from both the bound and input libraries. These four steps were iterated until no 10-nt  $k$ -mer remained that was enriched  $\geq 10$ -fold, thereby generating 14 sites for AGO2–miR-1. We then applied our MLE procedure to calculate relative  $K_d$  values for this expanded list of sites (Fig. 1, E and F).

This unbiased approach demonstrated that the 8mer, 7mer-m8, 7mer-A1, and 6mer sites to miR-1 were the highest-affinity site types of lengths  $\leq 10$  nt. It also identified eight previously uncharacterized sites with binding affinities resembling those of the 6mer-m8 and the 6mer-A1 (Fig. 1F). Comparison of these sites to the sequence of miR-1 revealed that miR-1 can tolerate either a wobble G at position 6 or a bulged U somewhere between positions 4 and 6 and achieve affinity at least 7- to 11-fold above that of the remaining no-site reads and that it can tolerate either a mismatched C at position 5 or a mismatched U at position 6 and achieve affinity four- to fivefold above that of the no-site reads. The GCUUCCGC motif also passed our cutoffs, which was more difficult to explain, because it had contiguous complementarity to positions 2 to 5 of miR-1 flanked by noncomplementary GC dinucleotides on both sides. Nonetheless, among the 1,398,100 possible motifs  $\leq 10$  nt, this was the only one that satisfied our criteria yet was difficult to attribute to miRNA pairing.

Our analytical approach and its underlying biochemical model also allowed us to infer the proportion of AGO2–miR-1 bound to each site (Fig. 1G). The 8mer site occupied 3.8 to 17% of the silencing complex over the concentration course, whereas the 7mer-m8, by virtue of its greater abundance, occupied a somewhat greater fraction of the complex. In aggregate, the marginal sites—including the 6mer-A1, 6mer-m8, and seven noncanonical sites—occupied 6.1 to 9.8% of the AGO2–miR-1 complex. Moreover, because of their very high abundance, library molecules with no identified site occupied 32 to 53% of the complex (Fig. 1G). These results support the inference that the summed contributions of background binding and low-affinity sites to intracellular AGO occupancy are of the same order of magnitude as those of canonical sites, suggesting

that an individual AGO–miRNA complex spends about half its time associated with a vast repertoire of background and low-affinity sites (22, 23). This phenomenon would help explain why sequences without recognizable sites often cross-link to AGO in cells.

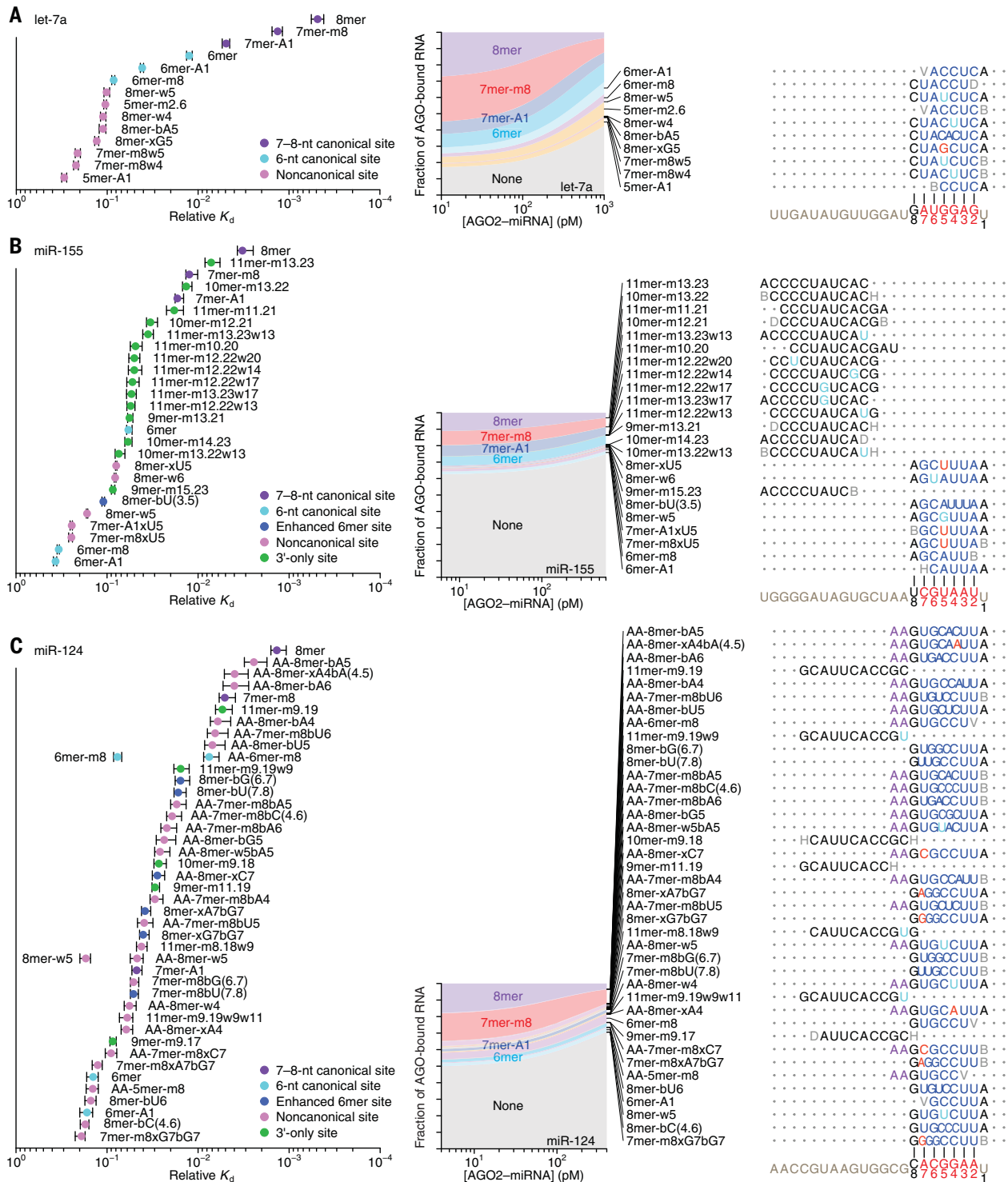
Our results confirmed that AGO2–miR-1 binds the 8mer, 7mer-m8, 7mer-A1, and 6mer sites most effectively and revealed the relative binding affinities and occupancies of these sites. In addition, our results uncovered weak yet specific affinity to the 6mer-A1 and 6mer-m8 sites plus seven noncanonical sites, all with affinities outside the dynamic range of recent high-throughput imaging experiments (12). Although alternative binding sites for miRNAs have been proposed on the basis of high-throughput in vivo cross-linking studies (24–28), our approach provided quantification of the relative strength of these sites without the confounding effects of differential cross-linking efficiencies, potentially enabling their incorporation into a quantitative framework of miRNA targeting.

#### Distinct canonical and noncanonical binding of different miRNAs

We extended our analysis to five additional miRNAs, including let-7a, miR-7, miR-124, and miR-155 of mammals, chosen for their sequence conservation as well as the availability of data examining their regulatory activities, intracellular binding sites, or in vitro binding affinities (1, 5, 6, 24, 25), and lsy-6 of nematodes, which is thought to bind unusually weakly to its canonical sites (29) (Fig. 2 and fig. S2, B and C). In the case of let-7a, previous biochemical analyses have determined the  $K_d$  values of some canonical sites (5, 6, 12), and our values agreed well, which further validated our high-throughput approach (fig. S1H).

The site-affinity profile of let-7a resembled that of miR-1, except the 6mer-m8 and 6mer-A1 sites for let-7a had greater binding affinity than essentially all of the noncanonical sites (Fig. 2A). As with miR-1, the noncanonical sites each paired to the seed region but did so imperfectly, typically with a single wobble, single mismatch, or single-nucleotide bulge, but these imperfections differed from those observed for miR-1 (Figs. 1F and 2A).

The site-affinity profiles of miR-124, miR-155, lsy-6, and miR-7 resembled those of miR-1 and let-7a. All but one included the six canonical sites (with miR-7 missing the 6mer-m8 site), and all contained noncanonical sites with extensive yet imperfect pairing to the miRNA seeds, the imperfections tending to occur at different positions and with different mismatched or bulged-nucleotide identities for different miRNAs (Fig. 2, B and C, and fig. S2, B and C). In contrast to the noncanonical sites of miR-1 and let-7a, more of the noncanonical sites of the other four miRNAs had affinities interspersed



**Fig. 2. Distinct canonical and noncanonical binding of different miRNAs.** (A to C) Relative  $K_d$  values and proportional occupancy of established and newly identified sites of let-7a (A), miR-155 (B), and miR-124 (C). The two miR-124 sites that were present as a 5'-AA-extended form in addition to an unextended form are shown on the same line (C). Relative  $K_d$  values are plotted as in Fig. 1F but in some cases with additional categories, either for 3'-only sites (green) [(B) and (C)] or for 6-nt canonical sites enhanced by either additional

wobble-pairing or additional Watson-Crick complementarity separated by a bulged nucleotide (blue) [(B) and (C)]. The proportion of AGO2-miRNA bound to each site type is estimated and shown as in Fig. 1G. These analyses also detected a GCACUUUA motif for let-7a and AACGAGGA motif for miR-155, which were assigned relative  $K_d$  values of  $7.1 \pm 0.8 \times 10^{-2}$  and  $6 \pm 1 \times 10^{-2}$ , respectively. These motifs are excluded because each did not match its respective miRNA and did not mediate repression by its respective miRNA (fig. S5B).

with those of the top four canonical sites. Moreover, the profiles for miR-155, miR-124, and *lsy-6* also included sites with extended (9- to 11-nt) complementarity to the miRNA 3' region. These sites had estimated  $K_d$  values that were derived from reads with little more than chance complementarity to the miRNA seed, and they had uniform enrichment across the length of the random-sequence region (fig. S1E), which indicated that these sites represented an alternative binding mode dominated by extensive pairing to the 3' region without involvement of the seed region (Fig. 2, B and C, and fig. S2B). We named them "3'-only sites."

In some respects, the 3'-only sites resembled noncanonical sites known as centered sites, which are reported to function in mammalian cells (30). Like 3'-only sites, centered sites have extensive perfect pairing to the miRNA, but for centered sites, this pairing begins at miRNA positions 3 or 4 and extends 11 to 12 nt through the center of the miRNA (30). Our unbiased search for sites did not identify centered sites for any of the six miRNAs. We therefore directly queried the region of each miRNA to which extensive noncanonical pairing was favored, determining the affinity of sequences with 11-nt segments of perfect complementarity to the miRNA sequence, scanning from miRNA position 3 to the 3' end of the miRNA (Fig. 3A). For miR-155, miR-124, and *lsy-6*, sequences with 11-nt sites that paired to the miRNA 3' region bound with greater affinity than did those with a canonical 6mer site, whereas for *let-7a*, miR-1, and miR-7, none of the 11-nt sites conferred stronger binding than did the 6mer. Moreover, for all six miRNAs, the 11-nt sites that satisfied the criteria for annotation as centered sites conferred binding  $\leq 2$ -fold stronger than that of the 6mer-m8 site, which also starts at position 3 but extends only 6 nt. These results called into question the function of centered sites, although we cannot rule out the possibility that centered sites are recognized by some miRNAs and not others. Indeed, the newly identified 3'-only sites functioned for only miR-155, miR-124, and *lsy-6*, and even among these, the optimal region of pairing differed, occurring at positions 13 to 23, 9 to 19, and 8 to 18, respectively (Fig. 3A).

When evaluating other types of noncanonical sites proposed to confer widespread repression in mammalian cells (20, 24), we found that all but two bound with affinities difficult to distinguish from background. One of these two was the 5-nt site matching miRNA positions 2 to 6 (5mer-m2.6) (20), which was bound by miR-1, *let-7a*, and miR-7 but not by the other three miRNAs (fig. S3). The other was the pivot site (24), which was bound by miR-124 [e.g., 8mer-bG(6.7); Fig. 2C] and *lsy-6* [e.g., 8mer-bA(6.7); fig. S2B] but not by the other four miRNAs (fig. S4). Thus, these two previously identified

noncanonical site types resembled the newly identified noncanonical sites with extensive yet imperfect pairing to the seed region, in that they function for only a limited number of miRNAs.

In addition to the differences in noncanonical site types observed for each miRNA, we also observed pronounced miRNA-specific differences in the relative affinities of the canonical site types. For example, for miR-155, the affinity of the 7mer-A1 nearly matched that of the 7mer-m8, whereas for miR-124, the affinity of the 7mer-A1 was  $>9$ -fold lower than that of the 7mer-m8. These results implied that the relative contributions of the A at target position 1 and the match at target position 8 can substantially differ for different miRNAs. Although prior studies show that AGO proteins remodel the thermodynamic properties of their loaded RNA guides (5, 6), our results show that the sequence of the guide strongly influences the nature of this remodeling, leading to differences in relative affinities across canonical site types and a distinct repertoire of noncanonical site types for each miRNA.

### The energetics of canonical binding

With the relative  $K_d$  values for the canonical binding sites of six miRNAs in hand, we examined the energetic relationship between the A at target position 1 (A1) and the match at miRNA position 8 (m8), within a framework analogous to a double-mutant cycle (Fig. 3B, left). The apparent binding-energy contributions of the m8 and A1 ( $\Delta\Delta G_{m8}$  and  $\Delta\Delta G_{A1}$ , respectively) were largely independent, as inferred from the relative  $K_d$  values of the four site types. That is, for each miRNA, the  $\Delta\Delta G_{m8}$  inferred in the presence of the A1 (using the ratio of the 8mer and 7mer-A1  $K_d$  values) resembled that inferred in the absence of the A1 (using the ratio of the 7mer-m8 and 6mer  $K_d$  values), and vice versa (Fig. 3B).

The relative  $K_d$  values for canonical sites of six miRNAs provided the opportunity to examine the relationship between the predicted free energy of site pairing and measured site affinities. We focused on the 6mer and 7mer-m8 sites because they lack the A1, which does not pair to the miRNA (Fig. 1A) (8, 31). Consistent with the importance of base pairing for site recognition and the known relationship between predicted seed-pairing stability and repression efficacy (29), affinity increased with increased predicted pairing stability, although this increase was statistically significant for only the 7mer-m8 site type (Fig. 3C;  $p = 0.09$  and  $0.005$  for the 6mer and 7mer-m8 sites, respectively). However, for both site types, the slope of the relationship was significantly less than expected from  $K_d = e^{-\Delta G/RT}$ , where  $\Delta G$  is the change in free energy,  $R$  is the universal gas constant, and  $T$  is temperature ( $p = 0.008$  and  $8 \times 10^{-5}$ , respectively). When considered

together with the previous analysis of a miRNA with enhanced seed-pairing stability, these results indicated that in remodeling the thermodynamic properties of the loaded miRNAs, AGO not only enhances the affinity of seed-matched interactions but also dampens the intrinsic differences in seed-pairing stabilities that would otherwise impose much greater inequities between the targeting efficacies of different miRNAs (6). Thus, although *lsy-6*, which has unusually poor predicted seed-pairing stability (29), did indeed have the weakest site-binding affinity of the six miRNAs, the difference between its binding affinity and that of the other miRNAs was less than might have been expected.

### Correspondence with repression observed in the cell

To evaluate the relevance of our in vitro binding results to intracellular miRNA-mediated repression, we examined the relationship between the relative  $K_d$  measurements and the repression of endogenous mRNAs after miRNA transfection into HeLa cells. When examining intracellular repression attributable to 3'UTR (3' untranslated region) sites of the transfected miRNA, we observed a pronounced relationship between AGO-RBNS-determined  $K_d$  values and mRNA fold changes (Fig. 3, D to I;  $r^2 = 0.80$  to  $0.97$ ). For instance, the different relative affinities of the 7mer-A1 and 7mer-m8 sites, most extremely observed for sites of miR-155 and miR-124, were nearly perfectly mirrored by the relative efficacy of these sites in mediating repression in the cell (Fig. 3, F and G). A similar correspondence between relative  $K_d$  values and repression was observed for the noncanonical sites that had both sufficient affinity and sufficient representation in the HeLa transcriptome to be evaluated using this analysis (Fig. 3, D to I). These included the pivot sites for miR-124 and *lsy-6* and the bulge-G7-containing sites for miR-7 (Fig. 3, G to I).

Analysis of mRNA changes observed after miRNA transfection was not suitable for measuring efficacy of the highest-affinity noncanonical sites because these sites lacked sufficient representation in endogenous 3'UTRs. Therefore, we implemented a massively parallel reporter assay designed to examine the efficacy of every site type identified by AGO-RBNS, each in 184 different 3'UTR sequence contexts (fig. S5A). This assay showed that 3'-only sites and other high-affinity-but-rare noncanonical site types do mediate repression in cells and that their efficacies tend to track with their affinities (fig. S5B). In sum, we found a strong correspondence between intracellular repression and in vitro binding affinity, regardless of miRNA identity and regardless of whether the target site is canonical or noncanonical or within an endogenous or a reporter mRNA. This result supported a model in which repression is a function of miRNA occupancy, as dictated by

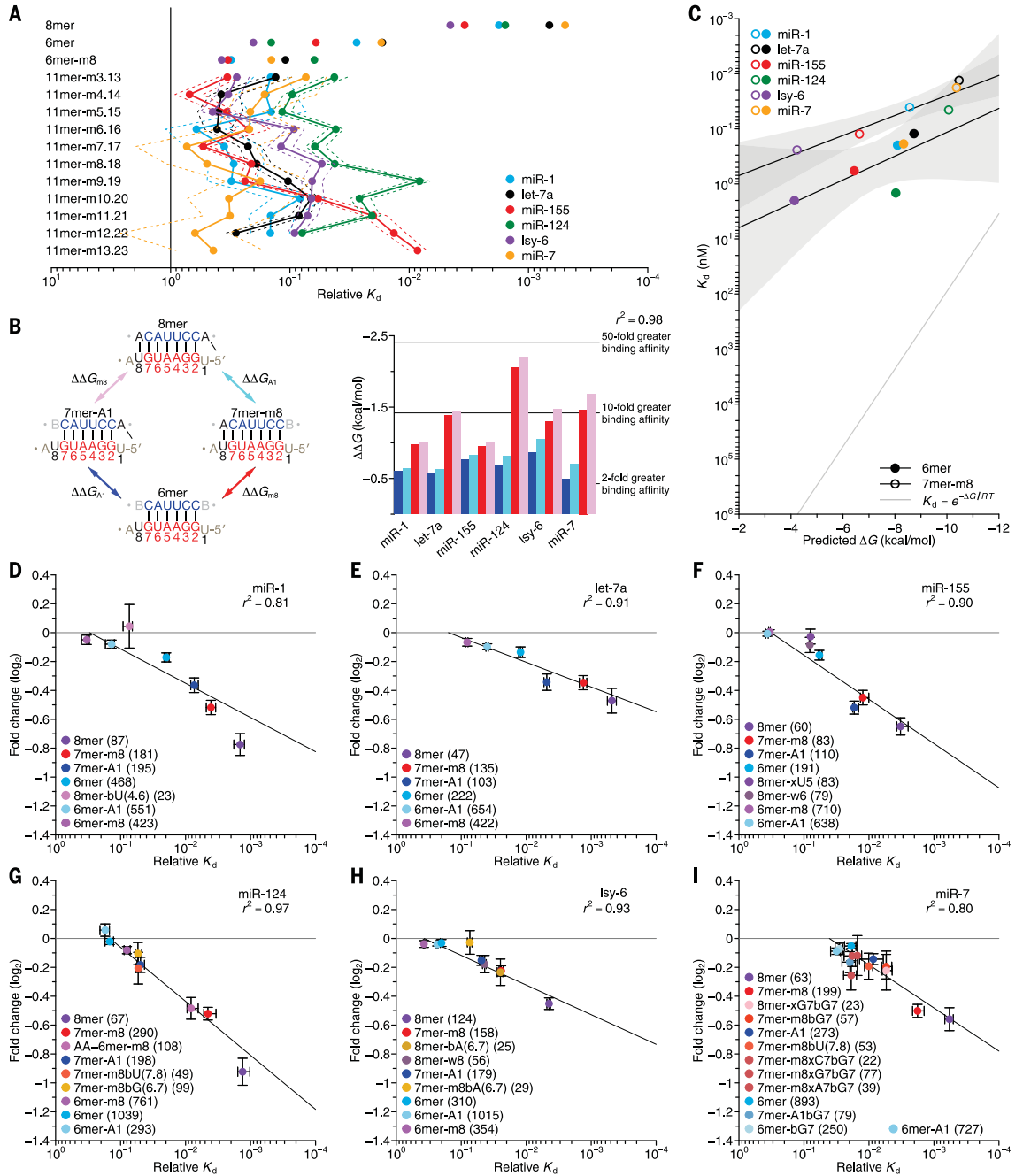


**Fig. 3. Additional analyses of binding affinities and the correspondence between binding affinity and repression efficacy. (A)** Diverse functionality and position dependence of 11-nt 3'-only sites. Relative  $K_d$  values for each potential 11-nt 3'-only site are plotted for the indicated miRNAs (key).

For reference, values for the 8mer, 6mer, and 6mer-m8 sites are also plotted. The solid vertical line marks the reference  $K_d$  value of 1.0, as in Fig. 1F. The solid and dashed lines indicate geometric mean and 95% confidence interval, respectively, determined as in Fig. 1D.

**(B)** The independent contributions of the A1 and m8 features. On the left, a double-mutant cycle depicts the affinity differences observed among the four top canonical sites for miR-1, as imparted by the independent contributions of the A1 and m8 features and their potential interaction. On the right, the apparent binding contributions of the A1 ( $\Delta\Delta G_{A1}$ , blue and cyan) or m8 ( $\Delta\Delta G_{m8}$ , red and pink) features are plotted, determined from the ratio of relative  $K_d$  values of either the 7mer-A1

and the 6mer (blue), the 8mer and the 7mer-m8 (cyan), the 7mer-m8 and the 6mer (red), or the 8mer and the 7mer-A1 (pink) for the indicated AGO2-miRNA complexes. The  $r^2$  reports on the degree of  $\Delta\Delta G$  similarity for both the m8 and A1 features using either of the relevant site-type pairs across all six complexes. **(C)** The relationship between the observed relative  $K_d$  values and predicted pairing stability of the 6mer (filled circles) and 7mer-m8 (open circles) sites of the indicated AGO-miRNA complex (key), under the assumption that the  $K_d$  value for library molecules without a site was 10 nM for all AGO-miRNA complexes. The two black lines are the best fit of the relationship observed for each of the site types (gray regions, 95% confidence interval). The gray line shows the expected relationship with the



predicted stabilities given by  $K_d = e^{-\Delta G/RT}$ . **(D to I)** The relationship between repression efficacy and relative  $K_d$  values for the indicated sites of miR-1 (D), let-7a (E), miR-155 (F), miR-124 (G), Isy-6 (H), and miR-7 (I). The number of sites of each type in the 3'UTRs is indicated (parentheses). To include information from mRNAs with multiple sites, multiple linear regression was applied to determine the log fold-change attributable to each site type (error bars, 95% confidence interval). The relative  $K_d$  values are those of Figs. 1 and 2 and fig. S2 (error bars, 95% confidence interval). Lines show the best fit to the data, determined by least-squares regression, weighting residuals using the 95% confidence intervals of the log fold-change estimates. The  $r^2$  values were calculated using similarly weighted Pearson correlations.

site affinity, and thus miRNA- and site-specific differences in binding affinities explain substantial differences in repression.

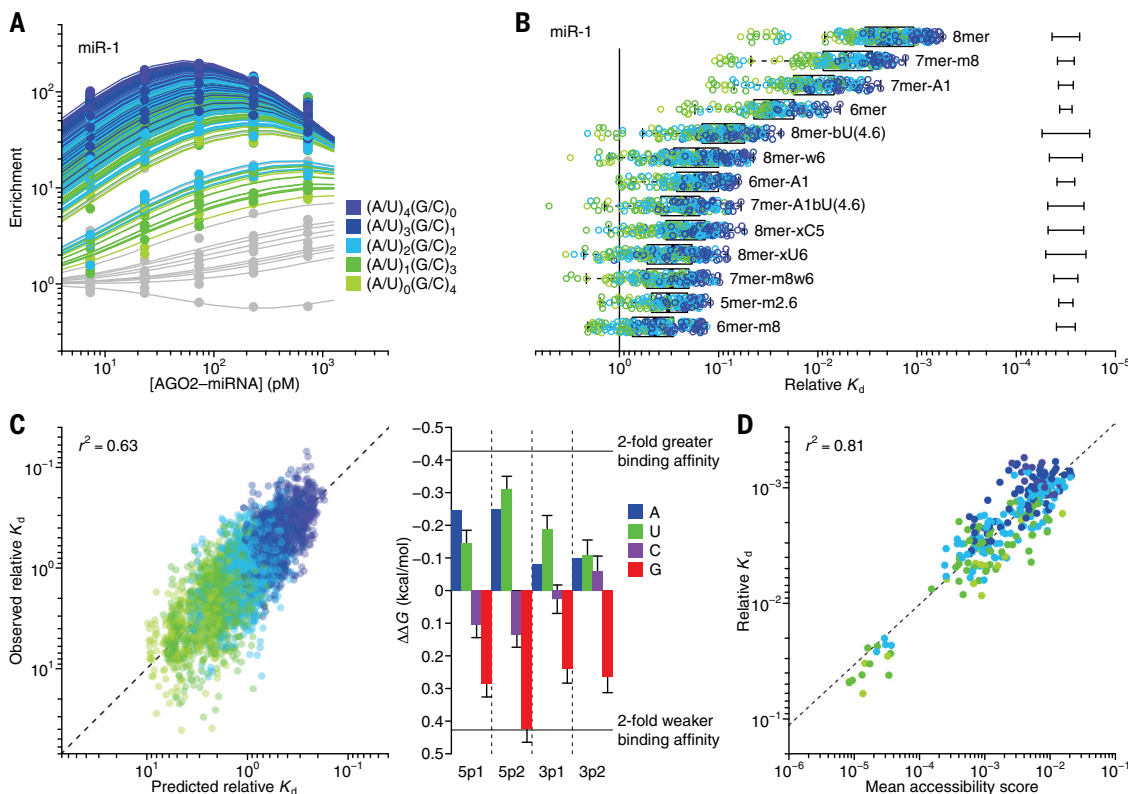
### The strong influence of flanking dinucleotide sequences

AU-rich nucleotide composition immediately flanking miRNA sites has long been associated with increased site conservation and efficacy in cells (13, 31, 32), but the mechanistic basis of this phenomenon has not been investigated, presumably because of the sparsity of affinity measurements. The AGO-RBNS data provided the means to overcome this limitation. We first separated the miR-1 8mer site into 256 different 12-nt sites, on the basis of the dinucleotide sequences immediately flanking each side of the 8mer, and determined relative  $K_d$

values for each (Fig. 4A). This analysis revealed a ~100-fold range in values, depending on the identities of the flanking dinucleotides, with binding affinity strongly tracking the AU content of the flanking dinucleotides. Extending this analysis across all miR-1 site types (Fig. 4B), as well as to sites to the other five miRNAs (fig. S6, A to E), yielded similar results. The effect of the flanking-dinucleotide context was of such magnitude that it often exceeded the affinity differences observed between miRNA-site types. Indeed, for each miRNA, at least one 6-nt canonical site in its most favorable context had greater affinity than that of the 8mer site in its least favorable context (Fig. 4B and fig. S6, A to E).

To identify general features of the flanking-dinucleotide effect across miRNA sequences

and site types, we trained a multiple linear regression model on the complete set of flanking-dinucleotide  $K_d$  values corresponding to all six canonical site types of each miRNA, fitting the effects at each of the four positions within the two flanking dinucleotides. The output of the model agreed well with the observed  $K_d$  values (Fig. 4C, left;  $r^2 = 0.63$ ), which indicated that the effects of the flanking dinucleotides were largely consistent between miRNAs and between site types of each miRNA. The output of the model also corresponded with the efficacy of intracellular repression, which indicated that these effects on  $K_d$  values were consequential in cells (fig. S6F). A and U nucleotides each enhanced affinity, whereas G nucleotides reduced affinity and C nucleotides were intermediate or neutral



**Fig. 4. The influence of flanking dinucleotide sequence context.** (A) AGO-RBNS profile of miR-1 sites, showing results for the 8mer separated into 256 different 12-nt sites on the basis of the identities of the two dinucleotides immediately flanking the 8mer. For each 12-nt site, the points and line are colored on the basis of the AU content of the flanking dinucleotides (key). For context, results of Fig. 1E are replotted in gray. Everything else is the same as in Fig. 1E. (B) Relative  $K_d$  values for each miR-1 site identified in Fig. 1F separated into 144 to 256 sites as in (A) on the basis of the identities of the flanking dinucleotides. The points are colored as in (A). Error bars indicate median 95% confidence interval across all  $K_d$  values. Everything else is the same as in Fig. 1F. (C) Consistency of flanking-dinucleotide effect across miRNA and site type. At the left is a comparison of observed relative  $K_d$  values and results of a mathematical model that used multiple linear regression to predict the influence of flanking dinucleotides. Plotted are results for all flanking dinucleotide contexts of all six canonical site types, for all six

miRNAs, normalized to the average affinity of each canonical site. Predictions of the model are those observed in a sixfold cross-validation, training on the results for five miRNAs and reporting the predictions for the held-out miRNA. The points for five outliers are not shown. The  $r^2$  quantifies the agreement between the predicted and actual values, considering all points. On the right, the model coefficients (multiplied by  $-RT$ , where  $T = 310.15$  K) corresponding to each of the four nucleotides of the 5' (5p) and 3' (3p) dinucleotides in the 5'-to-3' direction are plotted (error bars, 95% confidence interval). (D) Relationship between the mean structural-accessibility score and the relative  $K_d$  for the 256 12-nt sites containing the miR-1 8mer flanked by each of the dinucleotide combinations. Points are colored as in (A). Linear regression (dashed line) and calculation of  $r^2$  were performed using log-transformed values. For an analysis of the relationship between 8mer flanking-dinucleotide  $K_d$  and structural accessibility over a range of window lengths and positions relative to the 8mer site, see fig. S6G.



(Fig. 4C, right). Moreover, the identity of the 5' flanking dinucleotide, which must come into close proximity with the central RNA-binding channel of AGO (7), contributed more to binding affinity than did the 3' flanking sequence (Fig. 4C, right).

One explanation for this hierarchy of flanking nucleotide contributions, with  $A \approx U > C > G$ , is that it inversely reflected the propensity of these nucleotides to stabilize RNA secondary structure that could occlude binding of the silencing complex. To investigate this potential role for structural accessibility in influencing binding, we compared the predicted structural accessibility of 8mer sites in the input and bound libraries of the AGO2-miR-1 experiment, using a score for predicted structural accessibility previously optimized on data examining miRNA-mediated repression (14, 33). This score is based on the predicted probability that the 14-nt segment at target positions 1 to 14 is unpaired. We found that predicted accessibilities of sites in the bound libraries were substantially greater than those for sites in the input library and that the difference was greatest for the samples with the lower AGO2-miR-1 concentrations (fig. S6G), as expected if the accessibility score was predictive of site accessibility and if the most accessible sites were the most preferentially bound.

To build on these results, we examined the relationship between predicted structural accessibility and binding affinity for each of the 256 flanking dinucleotide possibilities. For each input read with a miR-1 8mer site, the accessibility score of that site was calculated. The sites were then differentiated on the basis of their flanking dinucleotides into 256 12-nt sites, and the geometric mean of the structural-accessibility scores of each of these extended sites was compared with the AGO-RBNS-derived relative  $K_d$  value (Fig. 4D and fig. S6H). A notable correlation was observed ( $r^2 = 0.82$ ,  $p < 10^{-15}$ ), with all 16 sites containing a 5'-flanking GG dinucleotide having both unusually poor affinities and unusually low accessibility scores. Moreover, sampling reads from the input library to match the predicted accessibility of sites in the bound library recapitulated the flanking dinucleotide preferences observed in the bound library (fig. S6I,  $r^2 = 0.79$ ). Taken together, our results demonstrate that local sequence context has a large influence on miRNA-target binding affinity and indicate that this influence results predominantly from the differential propensities of flanking sequences to favor structures that occlude site accessibility.

### A biochemical model predictive of miRNA-mediated repression

Inspired by the finding that measured affinities strongly corresponded to the repression

observed in cells (Fig. 3, D to I), we set out to build a biochemical framework that predicts the degree to which a miRNA represses each mRNA. Biochemical principles have been used to model miR-21-directed mRNA slicing (12). However, previous efforts that used biochemical principles to model aspects of the predominant mode of miRNA-mediated repression, including competition between endogenous target sites (23, 34, 35) and the influence of miRNAs on reporter gene-expression noise (36), were severely limited by the sparsity of the data. Our ability to measure the relative binding affinity of a miRNA to any 12-nt sequence enabled modeling of the quantitative effects of the six miRNAs on each cellular mRNA.

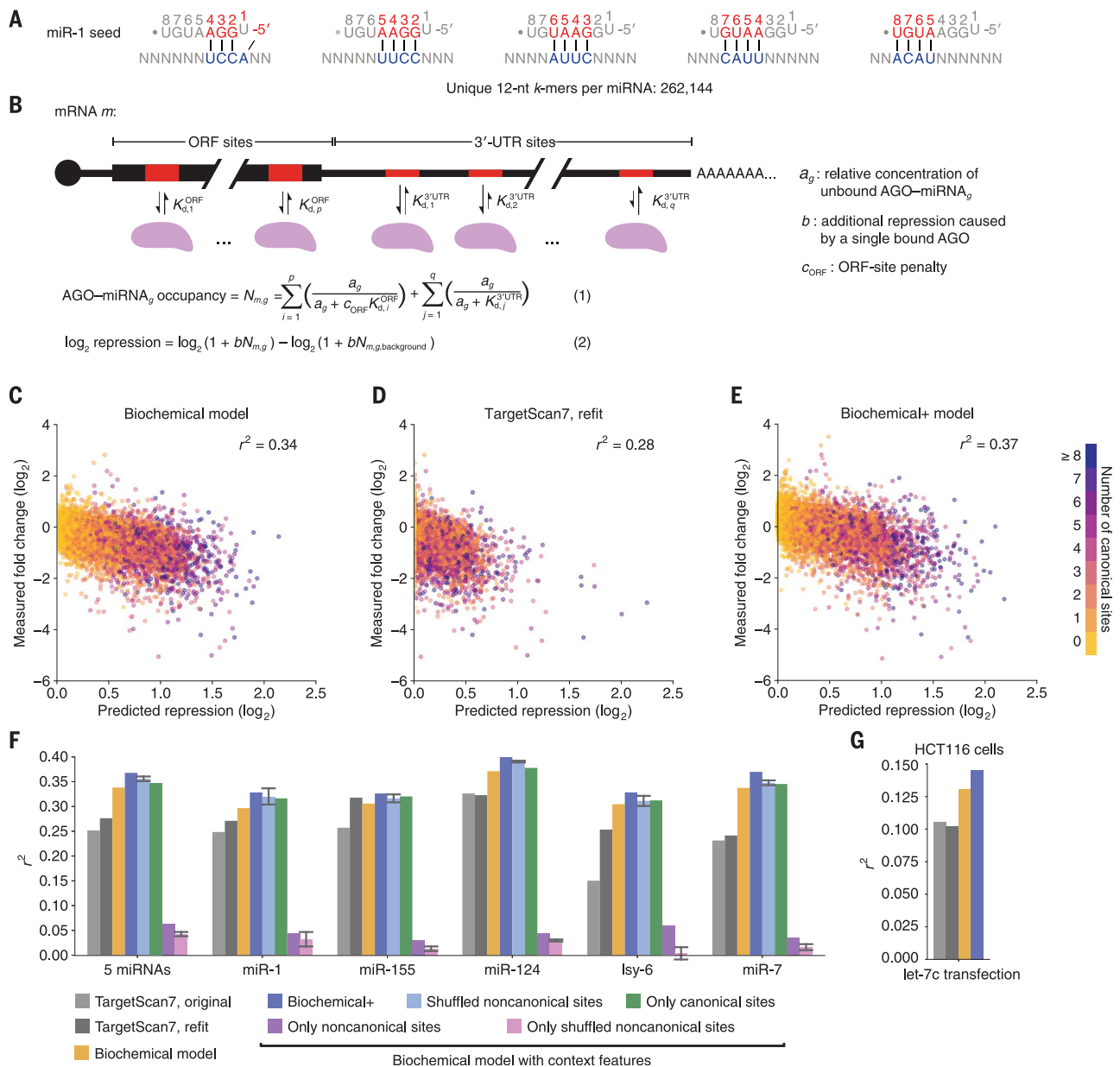
We first reanalyzed all six AGO-RBNS experiments to calculate, for each miRNA, the relative  $K_d$  values for all 262,144 12-nt  $k$ -mers that contained at least four contiguous nucleotides of the canonical 8mer site (Fig. 5A). These potential binding sites included the canonical sites and most of the noncanonical sites that we had identified, each within a diversity of flanking sequence contexts (Figs. 1F and 2). For each mRNA  $m$  and transfected miRNA  $g$ , the steady-state occupancy  $N_{m,g}$  (i.e., average number of AGO-miRNA complexes loaded with miRNA  $g$  bound to mRNA  $m$ ) was predicted as a function of the  $K_d$  values of the potential binding sites contained within the mRNA open reading frame (ORF) and 3'UTR, as well as the concentration of the unbound AGO-miRNA $_g$  complex  $a_g$ , which was fit as a single value for each transfected miRNA (Fig. 5B, equation 1). This occupancy value enabled prediction of a biochemically informed expectation of repression, assuming that the added effect of the miRNA on the basal decay rate scaled with the basal rate and  $N_{m,g}$  (Fig. 5B, equation 2). To isolate the effects of a transfected miRNA over background, we further offset our prediction of repression by a background-binding term (Fig. 5B,  $N_{m,g,background}$ ).

The calculation of predicted repression required an estimate of how much a single bound AGO affected the mRNA decay rate (Fig. 5B,  $b$ ), which was fit as a global value. Additionally, to account for the observation that sites in ORFs are less effective than those in 3'UTRs (3), our model included a penalty term for sites in ORFs, which was also fit as a global value (Fig. 5B). Because no appreciable repression was observed from sites in 5'UTRs, our model did not consider these sites.

Our biochemical model was fit against repression observed in HeLa cells transfected with one of five miRNAs with RBNS-derived measurements (let-7a was excluded because of its high endogenous expression in HeLa cells). A strong correspondence was observed when comparing mRNA changes measured upon miRNA transfection with those predicted by the model (fig. S7A,  $r^2 = 0.30$  to 0.37).

The overall performance of our biochemical model (Fig. 5C,  $r^2 = 0.34$ ) exceeded those of the 30 target-prediction algorithms ( $r^2 \leq 0.14$ ) that were also tested on changes in mRNA levels observed in response to miRNA transfection (14). We reasoned that in addition to our biochemical framework and the use of experimentally measured affinity values, other aspects of our analysis might have contributed to this improvement. For example, the miRNAs chosen for RBNS have high efficacy in transfection experiments, and our RNA-sequencing (RNA-seq) datasets generally had stronger signal over background compared to microarray datasets used to train and test previous target-prediction algorithms. Indeed, when evaluated on the same five datasets, the performance of the latest TargetScan model (TargetScan7) improved from an  $r^2$  of 0.14 to an  $r^2$  of 0.25 (fig. S7B). To explore the possibility that TargetScan7 might also benefit from training on this type of improved data, we generated transfection datasets for 11 additional miRNAs and retrained TargetScan7 on the collection of 16 miRNA-transfection datasets (again omitting the let-7a dataset), putting aside one dataset each time in a 16-fold cross-validation. Training and testing TargetScan on improved datasets further increased the  $r^2$  to 0.28 for the five miRNAs with AGO-RBNS data (Fig. 5D). Nonetheless, the biochemical model still outperformed the retrained TargetScan by >20%, which showed that the use of measured affinity values in a biochemical framework substantially increased prediction performance.

Many features known to correlate with targeting efficacy were captured by our biochemical model. Indeed, the contribution of certain features, such as site type (3), predicted seed-pairing stability (29), and nucleotide identities at specific miRNA or site positions (14), are expected to be represented more accurately in the miRNA-specific  $K_d$  values of the 12-nt  $k$ -mers than when generalized across miRNAs. However, these  $K_d$  values did not fully capture other factors that influence the affinity between miRNAs and their target sites in cells, including the structural accessibility of sites within their larger mRNA contexts and the contribution of supplementary pairing to the miRNA 3' region, which influences about 5% of sites (3). Without sufficient biochemical data quantifying these effects, we approximated their influence using scoring metrics known to correlate with miRNA targeting efficacy (13, 14) and allowed them to modify the  $K_d$  values additively in log space (i.e., linearly in free-energy space). Incorporating each of these metrics slightly improved the performance of the biochemical model, as did incorporating a score for the evolutionary conservation of the site (4), which helped account for additional unknown or imperfectly captured factors that influence targeting efficacy (fig. S7C).



**Fig. 5. AGO-RBNS  $K_d$  values enable a predictive model of miRNA-mediated repression in cells.**

(A) The 262,144 12-nt *k*-mers with at least four contiguous matches to the extended seed region of miR-1, for which relative  $K_d$  values were determined. Relative  $K_d$  values were similarly determined for the analogous *k*-mers of the other five miRNAs. (B) Biochemical model for estimating miRNA-mediated repression of an mRNA using the relative  $K_d$  values of the 12-nt *k*-mers in the mRNA. (C) Performance of the biochemical model as evaluated using the combined results of five miRNAs. Plotted is the relationship between mRNA changes observed after transfecting a miRNA and those predicted by the model. Each point represents the mRNA from one gene after transfection of a miRNA and is colored according to the number of canonical sites in the mRNA 3'UTR (key). For easier visual comparison between mRNAs, y-axis points for the

same mRNA are adjusted by the extrapolated expression level of the mRNA with no transfected miRNA. The Pearson's  $r^2$  between measured and predicted values is for unadjusted values and is reported in the upper right. (D) Performance of the retrained TargetScan7 model. Everything else is the same as in (C). (E) Performance of the biochemical+ model. Everything else is the same as in (C). (F) Model performances and the contribution of cognate noncanonical sites to performance of the biochemical+ model. Results for each model (key) are plotted for individual miRNAs and for all five miRNAs combined (error bars, standard deviation). (G) Performances of models tested on mRNA changes observed after transfecting let-7c into HCT116 cells engineered to have reduced endogenous miRNA expression (37). This analysis used the average  $a_g$  fit for the five miRNAs in (F). Everything else is the same as in (F).

Simultaneously incorporating all three metrics to generate what we call the “biochemical+ model” improved the  $r^2$  by 9% to 0.37 (Fig. 5E).

To examine how well our models generalized to another cell type and to a miRNA

family not used for fitting (let-7), we evaluated them on repression data collected after transfecting let-7c into HCT116 (human colon cancer) cells that had been engineered to not express endogenous miRNAs (37). Although

these data had a considerably lower signal-to-noise ratio, which lowered all  $r^2$  values, our biochemical models substantially outperformed TargetScan7 (Fig. 5G). This improvement extended to predicting repression after

transfecting miR-124 and miR-7 into human embryonic kidney (HEK) 293 cells (38) (fig. S8A). Additional analyses showed that the biochemical+ model performed at least as well as in vivo cross-linking immunoprecipitation sequencing (CLIP-seq) approaches in identifying the mRNAs most repressed upon miRNA transfection or most derepressed upon miRNA knockout (25, 38, 39) (fig. S8, B to D). Furthermore, for individual CLIP clusters enriched in wild type relative to miR-155 knockout, we observed a correlation between the occupancy predicted by our  $K_d$  values and the observed enrichment of the cluster [Spearman's rank-order correlation ( $r_s$ ) = 0.46,  $p < 10^{-7}$ ; fig. S8E], supporting the conclusion that  $K_d$  values measured in vitro reflect intracellular AGO binding.

When provided with  $K_d$  values for only the 12-nt  $k$ -mers that contained one of the six canonical sites, the biochemical+ model captured somewhat less variance (Fig. 5F, green bars;  $r^2 = 0.35$ ), and conversely when provided with  $K_d$  values for only the 12-nt  $k$ -mers lacking a canonical site, the model still retained some predictive power (Fig. 5F, purple bars;  $r^2 = 0.06$ ,  $p < 10^{-15}$ , likelihood-ratio test). As a control, we repeated the analysis after replacing the non-canonical sites (and their  $K_d$  values) of each miRNA with those of another miRNA, performing this shuffling and reanalysis for all 309 possible shuffle permutations. When using each of these shuffled controls, performance decreased, both when considering all sites (Fig. 5F, light blue bars) and when considering only the noncanonical sites (Fig. 5F, pink bars), as expected if the modest improvement conferred by including noncanonical sites were due, at least in part, to miRNA pairing to those sites. This advantage of cognate over shuffled noncanonical sites was largely maintained when evaluating the results for individual miRNAs (Fig. 5F). Together, our results showed that noncanonical sites can mediate intracellular repression but that their impact is dwarfed by that of canonical sites because high-affinity noncanonical sites are not highly abundant within transcript sequences. Thus, the improved performance over TargetScan achieved by the biochemical model was primarily from more accurate modeling of the effects of canonical sites.

### CNN for predicting site $K_d$ values from sequence

Our findings that binding preferences differ substantially between miRNAs and that these differences are not well predicted by existing models of RNA duplex stability in solution posed a major challenge for applying our biochemical framework to other miRNAs. Because performing AGO-RBNS for each of the known miRNAs would be impractical, we attempted to predict miRNA-target affinity from sequence using the six sets of relative

$K_d$  values and 16 miRNA-transfection datasets already in hand. Bolstered by recent successful applications of deep learning to predict complex aspects of nucleic acid biology from sequence (40–43), we chose a CNN for this task.

The overall model had two components. The first was a CNN that predicted relative  $K_d$  values for the binding of miRNAs to 12-nt  $k$ -mers (fig. S9A), and the second was the previously described biochemical model that links intracellular repression with relative  $K_d$  values (Fig. 6A). The training process simultaneously tuned both the neural network weights and the parameters of the biochemical model to fit both the relative  $K_d$  values and the mRNA repression data, with the goal of building a CNN that accurately predicts the relative  $K_d$  values for all 12-nt  $k$ -mers of a miRNA of any sequence.

For the CNN, we chose to include only the first 10 nucleotides of the miRNA sequence, which includes the position 1 nucleotide, the seed region, and the two downstream nucleotides that could pair to a 12-nt  $k$ -mer. Because the  $k$ -mers were not long enough to include sites with 3' supplementary pairing, we excluded the 3' region of the miRNA. Pairs of 10-nt truncated miRNA sequences and 12-nt  $k$ -mers were each parameterized as a 10-by-12-by-16 matrix, with the third dimension representing the 16 possible pairs of nucleotides that could be present at each pair of positions in the miRNA and target. The first layer of the CNN was designed to learn important single-nucleotide interactions, the second layer was designed to learn dinucleotide interactions, and the third layer was designed to learn position-specific information.

The training data for the CNN consisted of more than 1.5 million relative  $K_d$  values from six AGO-RBNS experiments and 68,112 mRNA expression estimates derived from 4257 transcripts in 16 miRNA transfection experiments. Five miRNAs had data in both sets. Because some repression was attributable to the passenger strands of the transfected duplexes (fig. S9B), the model considered both strands of each transfected duplex, which allowed the neural network to learn from another 16 AGO-loaded guide sequences.

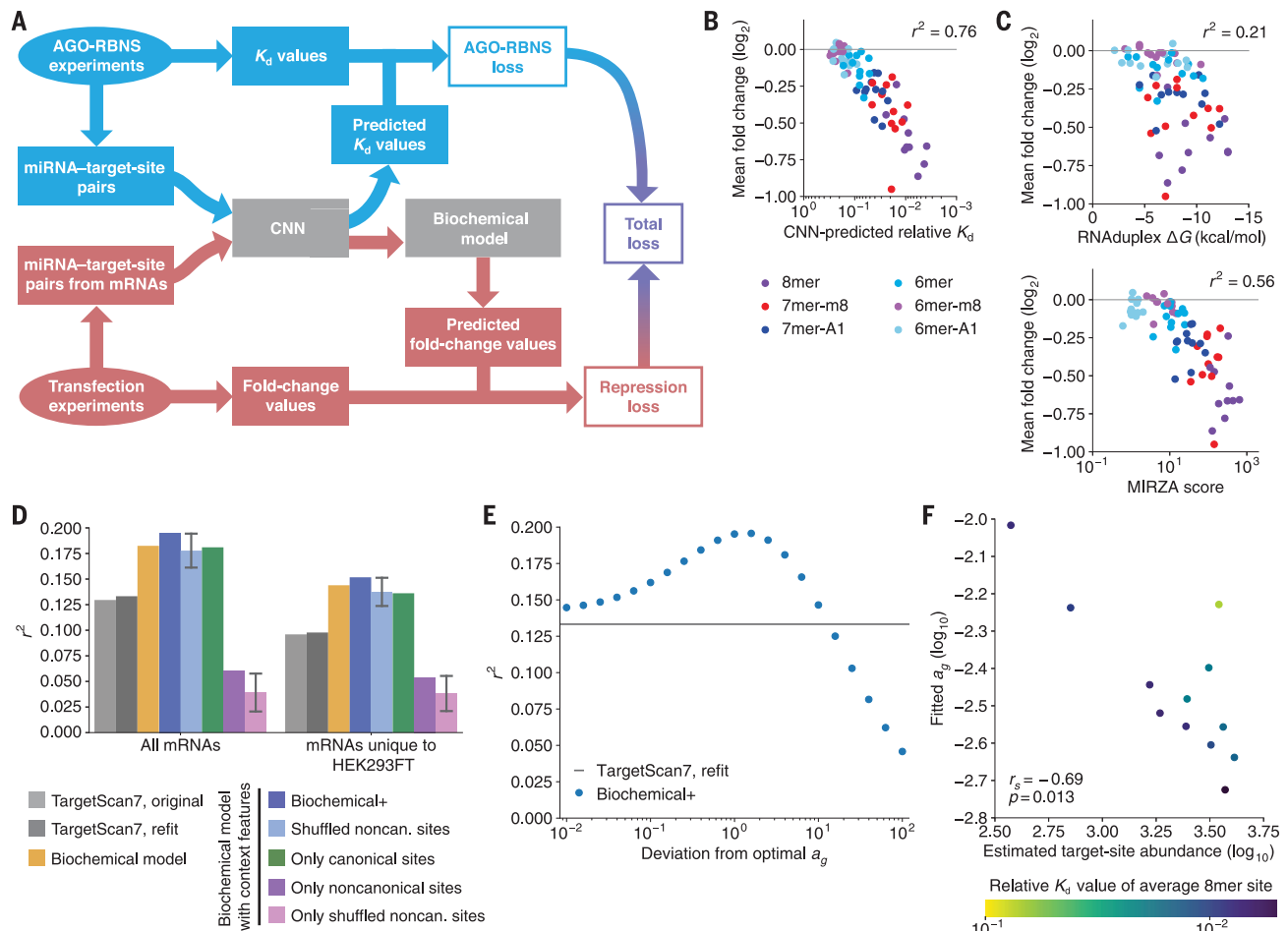
To test how well the CNN-predicted relative  $K_d$  values enabled our approach to be generalized to other miRNAs and another cell type, we generated 12 miRNA-transfection datasets in HEK293FT cells, choosing miRNAs that were not appreciably expressed in HEK293 cells (44) and that had not been used in any training (fig. S10). For each miRNA duplex in the test set, the CNN was used to predict relative  $K_d$  values for 12-nt  $k$ -mers to both the miRNA and passenger strands. As observed with the experimentally derived relative  $K_d$  values (Fig. 3, D to I), substantial correspondence was observed between CNN-predicted

relative  $K_d$  values for the six canonical site types of the transfected miRNAs and the mean repression that these site types conferred in cells (Fig. 6B and fig. S11). This correspondence ( $r^2 = 0.76$ ) substantially exceeded that observed for predictions of RNA-duplex stability in solution (45) and predictions derived from cross-linking results (27) (Fig. 6C;  $r^2 = 0.21$  and  $0.56$ , respectively). Aside from accurately predicting the relative efficacy of sites to the same miRNA, the CNN was better able to stratify sites of the same type to different miRNAs (e.g., Fig. 6B, purple dots;  $r^2 = 0.52$ ,  $p = 0.02$ ). Analysis of other site types suggested that the CNN had some ability to identify effective noncanonical sites for new miRNAs (fig. S11).

When the CNN-predicted  $K_d$  values and HeLa-derived global parameters were used as input for the biochemical and biochemical+ models to predict repression of individual mRNAs in HEK293FT cells, the results mirrored those observed when using relative  $K_d$  values derived from AGO-RBNS. Median ( $r^2 = 0.21$ ) and overall performance ( $r^2 = 0.18$ ) for the test set both exceeded those of TargetScan ( $r^2 = 0.12$  and  $0.13$ , respectively); overall performance improved ( $r^2 = 0.20$ ) when using the biochemical+ model, implying a 50% improvement over TargetScan, and performance dropped slightly when either shuffling or omitting noncanonical sites (Fig. 6D and fig. S12A; the main exception being the results for miR-190a, for which the performance of the biochemical+ model resembled that of TargetScan when only considering the canonical sites but substantially dropped when also considering noncanonical sites). The overall improvement over TargetScan was maintained when focusing on mRNAs that were expressed in HEK293FT cells but not HeLa cells (Fig. 6D). The CNN-predicted relative  $K_d$  values also enabled the biochemical+ model to outperform TargetScan and cross-linking approaches in predicting the effects of deleting or adding a miRNA in other cellular contexts (46–48) (fig. S12, B to D).

Although our models were improved over previous models, the highest  $r^2$  value achieved by our models for any of our datasets was 0.37 (Fig. 5F and fig. S12A), implying that they explained only a minority of the variability in mRNA fold changes occurring upon introducing a miRNA. However, even perfect prediction of the direct effects of miRNAs was not expected to explain all of the variability; some variability was due to the secondary effects of repressing the primary targets, and some was due to experimental noise. To estimate the maximal  $r^2$  that could be achieved by predicting the primary effects of miRNA targeting, we attempted to quantify and subtract the fraction of the fold-change variability attributable to the other two causes. For each dataset, the





**Fig. 6. A CNN for predicting binding affinity from sequence.** (A) Schematic of overall model architecture for training on RBNS data and transfection data simultaneously. "Loss" refers to squared loss. (B) The relationship between repression efficacy and CNN-predicted relative  $K_d$  values for the 12 test miRNAs. Everything else is the same as in Fig. 3, D to I. (C) The relationship between repression efficacy and RNA duplex-predicted free-energy values (45) (top) or MIRZA scores (27) (bottom) for the canonical sites of the 12 test miRNAs. Everything else is the same as in (B). (D) Performance of the biochemical and biochemical+ models when provided the CNN-predicted relative  $K_d$  values and tested on the 12 datasets examining the effects of

transfecting miRNAs into HEK293FT cells. On the left are results obtained when considering all mRNAs, and on the right are results obtained when considering mRNAs expressed in HEK293FT cells but not in HeLa cells. Everything else is the same as in Fig. 5F, except shuffling results were for 250 random permutations rather than all possible permutations. (E) Performance of the biochemical+ model on the HEK293FT test set while allowing the  $a_g$  values to deviate from the optimal fitted values. (F) Relationship between fitted  $a_g$  and estimated target-site abundance (29) for the guide strands of the 12 duplexes transfected into HEK293FT cells. Points are colored by the average relative  $K_d$  value of the 8mer site to each miRNA. The Spearman  $r_s$  and  $p$  value for the relationship are shown.

fraction attributable to experimental noise was estimated by examining the reproducibility between replicates in our transfection experiments, and the fraction attributable to secondary effects was inferred by assuming that primary miRNA effects only repress mRNAs, whereas secondary effects affect mRNAs in either direction (with effects distributed log normally). After accounting for these other sources of variability, the biochemical+ model provided with experimentally determined affinity values explained ~60% of the variability attributable to direct targeting (fig. S12E, median of five datasets), and when provided with CNN-predicted values, it explained ~50% of the variability attributable to direct targeting (fig. S12F, median of 12 datasets).

### Insights into miRNA targeting

The observation that canonical sites are not necessarily those with the highest affinity raises the question of how canonical sites are distinguished from noncanonical ones and whether making such a distinction is useful. Our results show that two criteria readily distinguish canonical sites from noncanonical ones. First, with only one exception, all six canonical site types were identified for each of the six miRNAs (the exception being the 6mer-m8 site for miR-7), whereas the noncanonical site types were typically identified for only one miRNA and never for more than three. Second, the four highest-affinity canonical sites occupied most of the specifically bound AGO2, even for miR-124, which had the largest and

highest-affinity repertoire of noncanonical sites (Figs. 1F and 2 and fig. S2, B and C). This greater role for canonical sites was presumably because perfect pairing to the seed region is the most efficient way to bind the silencing complex; to achieve equivalent affinity, the noncanonical sites must be longer and therefore less abundant. The ubiquitous function and more efficient binding of canonical sites explains why these site types have the greatest signal in meta-analyses of site conservation, thereby explaining why they were the first site types to be identified (31) and justifying the continued distinction between canonical and non-canonical site types.

The potential role of pairing to miRNA nucleotides 9 and 10 has been controversial.

Although some target-prediction algorithms (such as TargetScan) do not reward pairing to these nucleotides, most algorithms assume that such pairing enhances site affinity. Likewise, although one biochemical study reports that pairing to position 9 reduces site affinity (6), another reports that it increases affinity (12). We found that extending pairing to nucleotide 9 or 10 neither enhanced nor diminished affinity in the context of seed-matched sites (Fig. 4), whereas extending pairing to nucleotide 9 or 10 enhanced affinity in the context of 3'-only sites (Fig. 2, C and D). These results support the idea that extensive pairing to the miRNA 3' region unlocks productive pairing to nucleotides 9 to 12, which is otherwise inaccessible (1).

The biochemical parameters fit by our model provided additional insights into miRNA targeting. In the framework of our model, the fitted value of 1.8 observed for the parameter  $b$  suggested that a typical mRNA bound to an average of one silencing complex will experience a near tripling of its decay rate, which would lead to a ~60% reduction in its abundance. In the concentration regimes of our transfection experiments, this occupancy can be achieved with two to three median 7mer-m8 sites. In addition, our fitted value for the ORF-site penalty suggested that the translation machinery reduces site affinity by 5.5-fold.

Another parameter was  $a_g$ , that is, the intracellular concentration of AGO loaded with the transfected miRNA and not bound to a target site. Whereas values of the other parameters could be fit globally in HeLa cells and then used for testing,  $a_g$  was fit separately for each miRNA and passenger strand of each transfection experiment. Nonetheless, when  $a_g$  values were allowed to deviate from the fitted values, the biochemical+ model still outperformed TargetScan in predicting test-set repression over a 100-fold range of values (Fig. 6E), which indicated that even with rough estimates of miRNA abundances, our modeling framework had an advantage over other predictive methods in new contexts. Information that might be used to more accurately estimate  $a_g$  values should come with the determination of these values for more miRNAs in more cellular contexts, together with the observation that, as expected (29, 49), fitted  $a_g$  values are higher for miRNAs with lower predicted target abundance and lower general affinity for their targets (Fig. 6F).

Our work replaced the correlative models of targeting efficacy with a principled biochemical model that explains and predicts about half of the variability attributable to the direct effects of miRNAs on their targets, raising the question of how the understanding and prediction of miRNA-mediated repression might be further improved. Acquiring site-affinity profiles for additional miRNAs with diverse

sequences will improve the CNN-predicted miRNA-mRNA affinity landscape and further flesh out the two major sources of targeting variability revealed by our study, that is, the widespread differences in site preferences observed for different miRNAs and the substantial influence of local (12-nt) site context. We suspect additional improvement will come with increased ability to predict the other major cause of targeting variability, which is the variability imparted by mRNA features more distant from the site. This variability is captured only partially by the three features added to the biochemical model to generate the biochemical+ model. Perhaps the most promising strategy for accounting for these more distal features will be an unbiased machine-learning approach that uses entire mRNA sequences to predict repression, leveraging substantially expanded repression datasets as well as site-affinity values. In this way, the complete regulatory landscape, as specified by AGO within this essential biological pathway, might ultimately be computationally reconstructed.

#### Methods summary

AGO2-miRNA complexes were generated by adding synthetic miRNA duplexes to lysate from cells that overexpressed recombinant AGO2, and then these complexes were purified on the basis of affinity to the miRNA seed. RNA libraries were generated by in vitro transcription of synthetic DNA templates. For AGO-RBNS, purified AGO2-miRNA complex was incubated with a large excess of library molecules, and after reaching binding equilibrium, library molecules bound to AGO2-miRNA complex were isolated and prepared for high-throughput sequencing. Examination of  $k$ -mers enriched within the bound library sequences identified miRNA target sites, and relative  $K_d$  values for each of these sites were simultaneously determined by maximum likelihood estimation, fitting to AGO-RBNS results obtained over a 100-fold range in AGO2-miRNA concentration.

Intracellular miRNA-mediated repression was measured by performing RNA-seq on HeLa cells that had been transfected with a synthetic miRNA duplex. For sites that were sufficiently abundant in endogenous 3'UTRs, efficacy was measured on the basis of their influence on levels of endogenous mRNAs of HeLa cells. Site efficacy was also evaluated using massively parallel reporter assays, which provided information for the rare sites as well as the more abundant ones. The biochemical and biochemical+ models of miRNA-mediated repression were constructed and fit using the measured  $K_d$  values, and the repression of endogenous mRNAs was observed after transfecting miRNAs into HeLa cells. The CNN was built using TensorFlow, trained using the mea-

sured  $K_d$  values and the repression observed in the HeLa transfection experiments, and tested on the repression of endogenous mRNAs observed after transfecting miRNAs into HEK293T cells. Results were also tested on external datasets examining either intracellular binding of miRNAs by CLIP-seq or repression of endogenous mRNAs after miRNAs had been transfected, knocked down, or knocked out. The details of each of these methods are described in the supplementary materials.

#### REFERENCES AND NOTES

- D. P. Bartel, Metazoan microRNAs. *Cell* **173**, 20–51 (2018). doi: [10.1016/j.cell.2018.03.006](https://doi.org/10.1016/j.cell.2018.03.006); pmid: 29570994
- S. Jonas, E. Izaurralde, Towards a molecular understanding of microRNA-mediated gene silencing. *Nat. Rev. Genet.* **16**, 421–433 (2015). doi: [10.1038/nrg3965](https://doi.org/10.1038/nrg3965); pmid: 26077373
- D. P. Bartel, MicroRNAs: Target recognition and regulatory functions. *Cell* **136**, 215–233 (2009). doi: [10.1016/j.cell.2009.01.002](https://doi.org/10.1016/j.cell.2009.01.002); pmid: 19167326
- R. C. Friedl, K. K. H. Farh, C. B. Burge, D. P. Bartel, Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19**, 92–105 (2009). doi: [10.1101/gr.082701.108](https://doi.org/10.1101/gr.082701.108); pmid: 18955434
- L. M. Wee, C. F. Flores-Jasso, W. E. Salomon, P. D. Zamore, Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. *Cell* **151**, 1055–1067 (2012). doi: [10.1016/j.cell.2012.10.036](https://doi.org/10.1016/j.cell.2012.10.036); pmid: 23178124
- W. E. Salomon, S. M. Jolly, M. J. Moore, P. D. Zamore, V. Serebrov, Single-molecule imaging reveals that Argonaute reshapes the binding properties of its nucleic acid guides. *Cell* **162**, 84–95 (2015). doi: [10.1016/j.cell.2015.06.029](https://doi.org/10.1016/j.cell.2015.06.029); pmid: 26140592
- N. T. Schirle, J. Sheu-Gruttadauria, I. J. MacRae, Structural basis for microRNA targeting. *Science* **346**, 608–613 (2014). doi: [10.1126/science.1258040](https://doi.org/10.1126/science.1258040); pmid: 25359968
- N. T. Schirle, J. Sheu-Gruttadauria, S. D. Chandross, C. Joo, I. J. MacRae, Water-mediated recognition of t1-adenosine anchors Argonaute2 to microRNA targets. *eLife* **4**, e07646 (2015). doi: [10.7554/eLife.07646](https://doi.org/10.7554/eLife.07646); pmid: 26359634
- M. H. Jo et al., Human Argonaute 2 has diverse reaction pathways on target RNAs. *Mol. Cell* **59**, 117–124 (2015). doi: [10.1016/j.molcel.2015.04.027](https://doi.org/10.1016/j.molcel.2015.04.027); pmid: 26140367
- S. M. Klum, S. D. Chandross, N. T. Schirle, C. Joo, I. J. MacRae, Helix-7 in Argonaute2 shapes the microRNA seed region for rapid target recognition. *EMBO J.* **37**, 75–88 (2018). doi: [10.15252/emj.201796474](https://doi.org/10.15252/emj.201796474); pmid: 28939659
- S. D. Chandross, N. T. Schirle, M. Szczepaniak, I. J. MacRae, C. Joo, A dynamic search process underlies microRNA targeting. *Cell* **162**, 96–107 (2015). doi: [10.1016/j.cell.2015.06.032](https://doi.org/10.1016/j.cell.2015.06.032); pmid: 26140593
- W. R. Becker et al., High-throughput analysis reveals rules for target RNA binding and cleavage by AGO2. *Mol. Cell* **75**, 741–755.e11 (2019). doi: [10.1016/j.molcel.2019.06.012](https://doi.org/10.1016/j.molcel.2019.06.012); pmid: 31324449
- A. Grimson et al., MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. *Mol. Cell* **27**, 91–105 (2007). doi: [10.1016/j.molcel.2007.06.017](https://doi.org/10.1016/j.molcel.2007.06.017); pmid: 17612493
- V. Agarwal, G. W. Bell, J.-W. Nam, D. P. Bartel, Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015). doi: [10.7554/eLife.05005](https://doi.org/10.7554/eLife.05005); pmid: 26267216
- R. Guimnyy, M. Zavolan, Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. *Nucleic Acids Res.* **43**, 1380–1391 (2015). doi: [10.1093/nar/gkv050](https://doi.org/10.1093/nar/gkv050); pmid: 25628353
- M. D. Paraskevopoulou et al., DIANA-microT web server v5.0: Service integration into miRNA functional analysis workflows. *Nucleic Acids Res.* **41**, W169–W173 (2013). doi: [10.1093/nar/gkt393](https://doi.org/10.1093/nar/gkt393); pmid: 23680784
- N. Lambert et al., RNA Bind-n-Seq: Quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol. Cell* **54**, 887–900 (2014). doi: [10.1016/j.molcel.2014.04.016](https://doi.org/10.1016/j.molcel.2014.04.016); pmid: 24837674
- D. Dominguez et al., Sequence, structure, and context preferences of human RNA binding proteins. *Mol. Cell* **70**, 854–867.e9 (2018). doi: [10.1016/j.molcel.2018.05.001](https://doi.org/10.1016/j.molcel.2018.05.001); pmid: 29883606
- C. F. Flores-Jasso, W. E. Salomon, P. D. Zamore, Rapid and specific purification of Argonaute-small RNA complexes from

- crude cell lysates. *RNA* **19**, 271–279 (2013). doi: [10.1261/rna.036921.112](https://doi.org/10.1261/rna.036921.112); pmid: [23249751](https://pubmed.ncbi.nlm.nih.gov/23249751/)
20. D. Kim *et al.*, General rules for functional microRNA targeting. *Nat. Genet.* **48**, 1517–1526 (2016). doi: [10.1038/ng.3694](https://doi.org/10.1038/ng.3694); pmid: [27776116](https://pubmed.ncbi.nlm.nih.gov/27776116/)
  21. J. Brennecke, A. Stark, R. B. Russell, S. M. Cohen, Principles of microRNA-target recognition. *PLoS Biol.* **3**, e85 (2005). doi: [10.1371/journal.pbio.0030085](https://doi.org/10.1371/journal.pbio.0030085); pmid: [15723116](https://pubmed.ncbi.nlm.nih.gov/15723116/)
  22. R. Denzler, V. Agarwal, J. Stefano, D. P. Bartel, M. Stoffel, Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Mol. Cell* **54**, 766–776 (2014). doi: [10.1016/j.molcel.2014.03.045](https://doi.org/10.1016/j.molcel.2014.03.045); pmid: [24793693](https://pubmed.ncbi.nlm.nih.gov/24793693/)
  23. R. Denzler *et al.*, Impact of microRNA levels, target-site complementarity, and cooperativity on competing endogenous RNA-regulated gene expression. *Mol. Cell* **64**, 565–579 (2016). doi: [10.1016/j.molcel.2016.09.027](https://doi.org/10.1016/j.molcel.2016.09.027); pmid: [27871486](https://pubmed.ncbi.nlm.nih.gov/27871486/)
  24. S. W. Chi, G. J. Hannon, R. B. Darnell, An alternative mode of microRNA target recognition. *Nat. Struct. Mol. Biol.* **19**, 321–327 (2012). doi: [10.1038/nsmb.2230](https://doi.org/10.1038/nsmb.2230); pmid: [22343717](https://pubmed.ncbi.nlm.nih.gov/22343717/)
  25. G. B. Loeb *et al.*, Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. *Mol. Cell* **48**, 760–770 (2012). doi: [10.1016/j.molcel.2012.10.002](https://doi.org/10.1016/j.molcel.2012.10.002); pmid: [23142080](https://pubmed.ncbi.nlm.nih.gov/23142080/)
  26. A. Helwak, G. Kudla, T. Dudnakova, D. Tollervey, Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153**, 654–665 (2013). doi: [10.1016/j.cell.2013.03.043](https://doi.org/10.1016/j.cell.2013.03.043); pmid: [23622248](https://pubmed.ncbi.nlm.nih.gov/23622248/)
  27. M. Khorshid, J. Hausser, M. Zavolan, E. van Nimwegen, A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nat. Methods* **10**, 253–255 (2013). doi: [10.1038/nmeth.2341](https://doi.org/10.1038/nmeth.2341); pmid: [23334102](https://pubmed.ncbi.nlm.nih.gov/23334102/)
  28. S. Grosswendt *et al.*, Unambiguous identification of miRNA: target site interactions by different types of ligation reactions. *Mol. Cell* **54**, 1042–1054 (2014). doi: [10.1016/j.molcel.2014.03.049](https://doi.org/10.1016/j.molcel.2014.03.049); pmid: [24857550](https://pubmed.ncbi.nlm.nih.gov/24857550/)
  29. D. M. Garcia *et al.*, Weak seed-pairing stability and high target-site abundance decrease the proficiency of *Isy-6* and other microRNAs. *Nat. Struct. Mol. Biol.* **18**, 1139–1146 (2011). doi: [10.1038/nsmb.2115](https://doi.org/10.1038/nsmb.2115); pmid: [21909094](https://pubmed.ncbi.nlm.nih.gov/21909094/)
  30. C. Shin *et al.*, Expanding the microRNA targeting code: Functional sites with centered pairing. *Mol. Cell* **38**, 789–802 (2010). doi: [10.1016/j.molcel.2010.06.005](https://doi.org/10.1016/j.molcel.2010.06.005); pmid: [20620952](https://pubmed.ncbi.nlm.nih.gov/20620952/)
  31. B. P. Lewis, C. B. Burge, D. P. Bartel, Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005). doi: [10.1016/j.cell.2004.12.035](https://doi.org/10.1016/j.cell.2004.12.035); pmid: [15652477](https://pubmed.ncbi.nlm.nih.gov/15652477/)
  32. C. B. Nielsen *et al.*, Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* **13**, 1894–1910 (2007). doi: [10.1261/ma.768207](https://doi.org/10.1261/ma.768207); pmid: [17872505](https://pubmed.ncbi.nlm.nih.gov/17872505/)
  33. H. Tafer *et al.*, The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.* **26**, 578–583 (2008). doi: [10.1038/nbt1404](https://doi.org/10.1038/nbt1404); pmid: [18438400](https://pubmed.ncbi.nlm.nih.gov/18438400/)
  34. A. D. Bosson, J. R. Zamudio, P. A. Sharp, Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. *Mol. Cell* **56**, 347–359 (2014). doi: [10.1016/j.molcel.2014.09.018](https://doi.org/10.1016/j.molcel.2014.09.018); pmid: [25449132](https://pubmed.ncbi.nlm.nih.gov/25449132/)
  35. M. Jens, N. Rajewsky, Competition between target sites of regulators shapes post-transcriptional gene regulation. *Nat. Rev. Genet.* **16**, 113–126 (2015). doi: [10.1038/nrg3853](https://doi.org/10.1038/nrg3853); pmid: [25488579](https://pubmed.ncbi.nlm.nih.gov/25488579/)
  36. J. M. Schmiedel *et al.*, MicroRNA control of protein expression noise. *Science* **348**, 128–132 (2015). doi: [10.1126/science.aaa1738](https://doi.org/10.1126/science.aaa1738); pmid: [25838385](https://pubmed.ncbi.nlm.nih.gov/25838385/)
  37. P. S. Linsley *et al.*, Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Mol. Cell Biol.* **27**, 2240–2252 (2007). doi: [10.1128/MCB.02005-06](https://doi.org/10.1128/MCB.02005-06); pmid: [17242205](https://pubmed.ncbi.nlm.nih.gov/17242205/)
  38. J. Hausser, M. Landthaler, L. Jaskiewicz, D. Gaidatzis, M. Zavolan, Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome Res.* **19**, 2009–2020 (2009). doi: [10.1101/gr.091181.109](https://doi.org/10.1101/gr.091181.109); pmid: [19767416](https://pubmed.ncbi.nlm.nih.gov/19767416/)
  39. M. Hafner *et al.*, Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129–141 (2010). doi: [10.1016/j.cell.2010.03.009](https://doi.org/10.1016/j.cell.2010.03.009); pmid: [20371350](https://pubmed.ncbi.nlm.nih.gov/20371350/)
  40. B. Alipanahi, A. Delong, M. T. Weirauch, B. J. Frey, Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015). doi: [10.1038/nbt.3300](https://doi.org/10.1038/nbt.3300); pmid: [26213851](https://pubmed.ncbi.nlm.nih.gov/26213851/)
  41. R. Tunney *et al.*, Accurate design of translational output by a neural network model of ribosome distribution. *Nat. Struct. Mol. Biol.* **25**, 577–582 (2018). doi: [10.1038/s41594-018-0080-2](https://doi.org/10.1038/s41594-018-0080-2); pmid: [29967537](https://pubmed.ncbi.nlm.nih.gov/29967537/)
  42. J. T. Cuperus *et al.*, Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome Res.* **27**, 2015–2024 (2017). doi: [10.1101/gr.224964.117](https://doi.org/10.1101/gr.224964.117); pmid: [29097404](https://pubmed.ncbi.nlm.nih.gov/29097404/)
  43. K. Jaganathan *et al.*, Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019). doi: [10.1016/j.cell.2018.12.015](https://doi.org/10.1016/j.cell.2018.12.015); pmid: [30661751](https://pubmed.ncbi.nlm.nih.gov/30661751/)
  44. P. Landgraf *et al.*, A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**, 1401–1414 (2007). doi: [10.1016/j.cell.2007.04.040](https://doi.org/10.1016/j.cell.2007.04.040); pmid: [17604727](https://pubmed.ncbi.nlm.nih.gov/17604727/)
  45. R. Lorenz *et al.*, ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26–14 (2011). doi: [10.1186/1748-7188-6-26](https://doi.org/10.1186/1748-7188-6-26); pmid: [22115189](https://pubmed.ncbi.nlm.nih.gov/22115189/)
  46. K. Zhang *et al.*, A novel class of microRNA-recognition elements that function only within open reading frames. *Nat. Struct. Mol. Biol.* **25**, 1019–1027 (2018). doi: [10.1038/s41594-018-0136-3](https://doi.org/10.1038/s41594-018-0136-3); pmid: [30297778](https://pubmed.ncbi.nlm.nih.gov/30297778/)
  47. I. Lipchina *et al.*, Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. *Genes Dev.* **25**, 2173–2186 (2011). doi: [10.1101/gad.17221311](https://doi.org/10.1101/gad.17221311); pmid: [22012620](https://pubmed.ncbi.nlm.nih.gov/22012620/)
  48. S. W. Eichhorn *et al.*, mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Mol. Cell* **56**, 104–115 (2014). doi: [10.1016/j.molcel.2014.08.028](https://doi.org/10.1016/j.molcel.2014.08.028); pmid: [25263593](https://pubmed.ncbi.nlm.nih.gov/25263593/)
  49. A. Arvey, E. Larsson, C. Sander, C. S. Leslie, D. S. Marks, Target mRNA abundance dilutes microRNA and siRNA activity. *Mol. Syst. Biol.* **6**, 363 (2010). doi: [10.1038/msb.2010.24](https://doi.org/10.1038/msb.2010.24); pmid: [20404830](https://pubmed.ncbi.nlm.nih.gov/20404830/)

## ACKNOWLEDGMENTS

We thank K. Heindl, T. Eisen, and T. Bepler for helpful discussions; Y. Zhou for providing processed CLIP data from miR-20a overexpression; and members of the Bartel lab for comments on this manuscript. **Funding:** This work was supported by NIH grants GM118135 (D.P.B.) and GM123719 (N.B.). D.P.B. is an investigator of the Howard Hughes Medical Institute. **Author contributions:** S.E.M. developed AGO-RBNS and associated analyses, which he implemented with help from T.M.P. and N.B. K.S.L. devised and implemented the biochemical model and CNN. C.Y.S., G.M.K., and T.M.P. performed transfection and sequencing experiments. C.Y.S. and S.E.M. designed and performed the massively parallel reporter assay. S.E.M., K.S.L., and D.P.B. designed the study and wrote the manuscript with input from other authors. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Sequencing data are available in the Gene Expression Omnibus (accession number GSE140220), and computational tools are deposited in GitHub (<https://github.com/smcgeary/agorbn> and [https://github.com/kslin/miRNA\\_models](https://github.com/kslin/miRNA_models)).

## SUPPLEMENTARY MATERIALS

[science.sciencemag.org/content/366/6472/eaav1741/suppl/DC1](https://science.sciencemag.org/content/366/6472/eaav1741/suppl/DC1)  
Materials and Methods

Figs. S1 to S12  
Tables S1 and S2  
References (50–55)  
Data S1 to S3

[View/request a protocol for this paper from Bio-protocol.](#)

21 August 2018; resubmitted 24 September 2019

Accepted 16 November 2019

Published online 5 December 2019

10.1126/science.aav1741



## The biochemical basis of microRNA targeting efficacy

Sean E. McGeary, Kathy S. Lin, Charlie Y. Shi, Thy M. Pham, Namita Bisaria, Gina M. Kelley and David P. Bartel

*Science* **366** (6472), eaav1741.

DOI: 10.1126/science.aav1741 originally published online December 5, 2019

### Biochemical prediction of miRNA targeting

MicroRNAs (miRNAs) regulate most human messenger RNAs and play essential roles in diverse developmental and physiological processes. Correctly predicting the function of each miRNA requires a better understanding of miRNA targeting efficacy. McGeary *et al.* measured binding affinities between six miRNAs and synthetic targets, built a biochemical model of miRNA-mediated repression, and expanded it to all miRNAs using a convolutional neural network. This approach offers insights into miRNA targeting and enables more accurate prediction of intracellular miRNA repression efficacy than previous algorithms.

*Science*, this issue p. eaav1741

#### ARTICLE TOOLS

<http://science.sciencemag.org/content/366/6472/eaav1741>

#### SUPPLEMENTARY MATERIALS

<http://science.sciencemag.org/content/suppl/2019/12/04/science.aav1741.DC1>

#### RELATED CONTENT

<http://stm.sciencemag.org/content/scitransmed/10/465/eaau6414.full>  
<http://stm.sciencemag.org/content/scitransmed/9/401/eaam9327.full>  
<http://stm.sciencemag.org/content/scitransmed/8/347/347ra94.full>  
<http://stm.sciencemag.org/content/scitransmed/5/212/212ra162.full>

#### REFERENCES

This article cites 54 articles, 12 of which you can access for free  
<http://science.sciencemag.org/content/366/6472/eaav1741#BIBL>

#### PERMISSIONS

<http://www.sciencemag.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of Service](#)

---

*Science* (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

Copyright © 2019 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works