



OPEN

## A deep learning method for miRNA/isomiR target detection

Amlan Talukder<sup>1</sup>, Wencai Zhang<sup>2</sup>, Xiaoman Li<sup>2</sup>✉ & Haiyan Hu<sup>1,3</sup>✉

Accurate identification of microRNA (miRNA) targets at base-pair resolution has been an open problem for over a decade. The recent discovery of miRNA isoforms (isomiRs) adds more complexity to this problem. Despite the existence of many methods, none considers isomiRs, and their performance is still suboptimal. We hypothesize that by taking the isomiR–mRNA interactions into account and applying a deep learning model to study miRNA–mRNA interaction features, we may improve the accuracy of miRNA target predictions. We developed a deep learning tool called DMISO to capture the intricate features of miRNA/isomiR–mRNA interactions. Based on tenfold cross-validation, DMISO showed high precision (95%) and recall (90%). Evaluated on three independent datasets, DMISO had superior performance to five tools, including three popular conventional tools and two recently developed deep learning-based tools. By applying two popular feature interpretation strategies, we demonstrated the importance of the miRNA regions other than their seeds and the potential contribution of the RNA-binding motifs within miRNAs/isomiRs and mRNAs to the miRNA/isomiR–mRNA interactions.

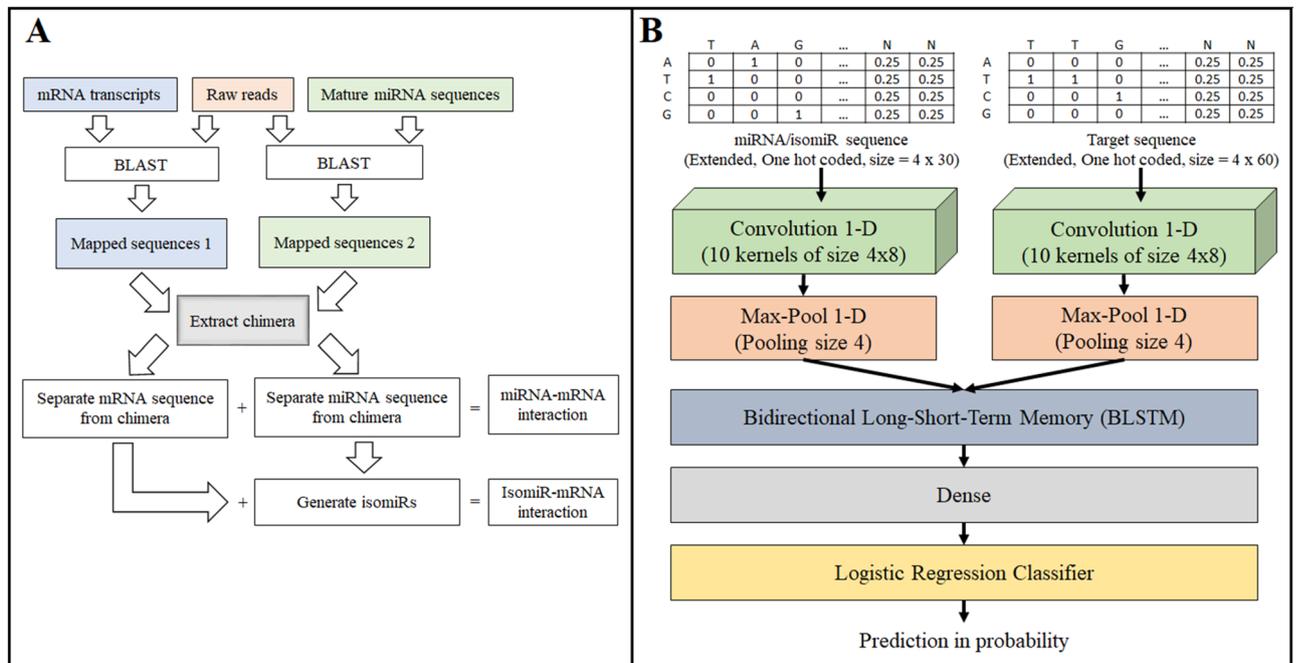
MicroRNAs (miRNAs) are ~22 nucleotides (nt) long single-stranded non-coding RNAs that play important roles in gene regulation and disease progression<sup>1–5</sup>. During metazoan miRNA biogenesis, miRNA genes are transcribed into pri-miRNAs, which are cut by the enzymes Drosha and DGCR8 to create the hairpin-structured pre-miRNAs. The pre-miRNAs are then exported to the cytoplasm and processed by the enzyme Dicer to produce the duplex miRNAs. Finally, the miRNAs are matured from either or both strands of the duplex miRNAs. These mature miRNAs directly bind and interact with their target mRNAs in different cell types through the context-specific choices of target sites, which results in the degradation and/or translation repression of the target mRNAs<sup>1–3</sup>. It is thus important to investigate how miRNAs select their target sites and identify their target mRNAs.

The discovery of different types of miRNA isoforms (isomiRs) makes the overall investigation of miRNA target sites and target genes more challenging while fascinating<sup>6</sup>. During metazoan miRNA biogenesis, isomiRs are created by imprecise cleavage of pri-miRNAs and/or pre-miRNAs<sup>7</sup>, the addition/deletion of nt to/from the ends of mature miRNAs<sup>8,9</sup>, and the modification of one or more nt in the middle of mature miRNAs. Correspondingly, the resulted isomiRs are classified into addition, deletion, and polymorphic isomiRs, respectively. Based on the modified end, both addition and deletion isomiRs can be further grouped as 3' isomiRs and 5' isomiRs. The 3' isomiRs are more abundant and share the same seed region (positions 2–7) as their original miRNAs while the 5' isomiRs have different seed regions and hence different target mRNAs. The resulting isomiRs can also be hybrids of the aforementioned types.

The widespread isomiRs in different cell types call for a revisit to the miRNA target site selection and target mRNA identification problem<sup>10–13</sup>. Under any given experimental condition, there are likely diverse changes in sequence and expression of individual miRNAs. The diversity implies the existence of different isomiRs of varied abundance and the conventional miRNAs actively interacting with their respective target genes under an experimental condition. Previous studies have shown that such a mixture of miRNAs and their isomiRs under specific experimental conditions are common instead of sequencing artifacts<sup>14,15</sup>. The available methods and tools for miRNA target prediction take only miRNA–mRNA interactions into account, which may unknowingly consider isomiR–mRNA interactions for miRNA–mRNA interactions in training and thus result in high false-positive rates and sub-optimal performance<sup>16,17</sup>. It is thus critical to study isomiRs and miRNAs together for their target site and target mRNA identification.

The CLASH (crosslinking, ligation and sequencing of hybrids) and CLEAR-CLIP (covalent ligation of endogenous Argonaute-bound RNAs with cross-linking immunoprecipitation) provide an unprecedented opportunity to study miRNA targets under the context of isomiRs<sup>18,19</sup>. Both experiments provide the chimeric reads composed

<sup>1</sup>Department of Computer Science, University of Central Florida, Orlando, FL 32816, USA. <sup>2</sup>Burnett School of Biomedical Science, University of Central Florida, Orlando, FL 32816, USA. <sup>3</sup>Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, FL 32816, USA. ✉email: xiaoman@mail.ucf.edu; haihu@cs.ucf.edu



**Figure 1.** (A) The pipeline to obtain miRNA/isomiR–mRNA interactions. (B) The DMISO model structure.

of pairs of miRNA variants and their interacting target sites in mRNAs. These data, especially the CLASH data, have been widely used to study non-canonical target sites, which show the importance of regions other than the seed regions in miRNAs<sup>20–23</sup>. However, these data have not been explored to study how miRNAs and their isomiRs interact with their targets.

In this study, we attempted to predict target sites and target mRNAs of miRNAs and isomiRs by considering miRNA–mRNA interactions together with isomiR–mRNA interactions for the first time. Our purpose was to present a tool that is exclusively trained to predict isomiR–mRNA interactions and also capitalize the sequence information learned from isomiR–mRNA interactions to improve the prediction performance on the canonical miRNA–mRNA interactions. As the features of miRNA target sites and miRNA–mRNA interactions are not fully understood, and deep learning-based approaches have shown better performance in analyzing genomic and epigenomic data<sup>24</sup>, we designed a deep learning method and tool for miRNA and isomiR target prediction (DMISO). Tested by cross-validation and on independent datasets, we showed that, on average, DMISO had a precision of 95% and a recall of 90%. Compared with three popular tools and two recently developed deep learning-based tools, DMISO showed superior performance in almost every metric to the five tools. The DMISO tool and its codes are freely available at <http://hulab.ucf.edu/research/projects/DMISO>.

## Materials and methods

**Identification of miRNA–mRNA and isomiR–mRNA interactions in CLASH.** To train and test DMISO, we obtained miRNA–mRNA and isomiR–mRNA interactions from the CLASH experiments<sup>18</sup>. We downloaded the CLASH data in the HEK293 cell line (GSE50452), which contained raw reads from six human samples. Each sample consisted of both single and chimeric reads. Only the chimeric reads comprised the miRNA/isomiR sequences and their interacting target site sequences in mRNAs. Here and henceforth, we refer miRNAs as the traditional ones in miRBase<sup>25</sup> and use “miRNAs/isomiRs” to represent the miRNAs and their isomiRs.

We identified miRNA–mRNA interactions from the chimeric reads similar to the original study<sup>18</sup> (Fig. 1A and Supplementary Table S1). In brief, we downloaded raw reads, removed adapters from the reads, discarded duplicate reads, and finally mapped the remaining reads against two databases separately with BLAST version 2.10.1+<sup>26</sup>. One database was the protein-coding transcript sequences from GENCODE version 38<sup>27</sup>. The other was the human mature miRNA sequences from miRBase version 22.1<sup>25</sup>. We required a BLAST hit with an e-value  $\leq 0.1$  to claim the mapping of a read. The mapped reads along the antisense strand of a transcript or having alignment loops were discarded. To control the mapping quality of the miRNA portion of a chimeric read, which was much shorter than the mRNA portion and thus can be harder to differentiate from the sequencing errors, we required that the same miRNA portion occurred in at least 10 chimeric reads. We chose 10 here because the chance to observe the same isomiR at least 10 times was about  $8.87E-08$ , given the average miRNA length 22nt, the Illumina sequencer error rate 0.001, and the number of reads mapped to a miRNA was fewer than 1000 for most (>95%) miRNAs. Moreover, this choice enabled enough data for training the deep learning models. We allowed a maximum gap or distance of 4 nt between the mapped miRNA portion and the mapped mRNA portion in a chimeric read as previously<sup>18</sup>. The miRNA and mRNA portion of a chimeric read may be mapped to multiple miRNA and mRNA transcripts, respectively. If a read was mappable to multiple miRNA or mRNA transcripts,

to retain the most significant miRNA–mRNA pair, we used the following criteria in order: (1) the pair with the smaller BLAST e-values; and (2) the pair with the larger BLAST bit scores if the e-values were the same.

With the identified miRNA–mRNA candidate pairs, we compared the aligned miRNA portion of the chimeric reads with the corresponding miRNAs to define miRNA–mRNA and isomiR–mRNA interactions. If a read perfectly matched a miRNA, we claimed this candidate pair as a miRNA–mRNA pair. Otherwise, if the nt sequencing quality scores at the variation positions (compared with the miRNA sequence) are larger than 30, this candidate pair is an isomiR–mRNA pair. To select isomiRs with confidence, we also required that the isomiR sequences were seen in at least 10 chimeric reads. We further classified these isomiRs in isomiR–mRNA pairs into the following eight types: 5' isomiR (addition, deletion and replacement), 3' isomiR (addition, deletion and replacement), single nucleotide polymorphic (SNP) isomiR, multiple nucleotide polymorphic (MNP) isomiR. An isomiR may belong to multiple types.

**Training data and cross-validation.** For the obtained miRNA/isomiR–mRNA pairs, we extended the 3' end of the mRNA section of the chimeric sequences by 25 nt to have more complete target sites. The extended mRNA target sites shorter than 30 nt were filtered out as previously<sup>18</sup>. The extended pairs were considered as positive interaction pairs. For every positive pair, a negative pair was generated with the same miRNA or isomiR and a negative site in the 3' untranslated region of the corresponding positive mRNA transcript as previously<sup>28,29</sup>. The negative site was required to be at least 10 nt away from the positive sites and had free folding energy < 10 kcal/mol measured by the RNACoFold tool<sup>30</sup>. We created our training dataset by randomly choosing 80% of the positive and negative interactions. We tested DMISO on the training data using tenfold cross-validation. We also tested DMISO on the remaining 20% interactions that were not used for training.

**Independent data.** In addition to the remaining 20% CLASH test data, we extracted miRNA/isomiR–mRNA pairs from CLEAR-CLIP data as independent test data. Similar to the above analysis of the CLASH data, we analyzed the CLEAR-CLIP chimeric reads in 12 human samples from the hepatocyte-derived carcinoma cell line HuH-7.5 (GSE73059)<sup>19</sup>. We defined 14,684 positive miRNA/isomiR–mRNA pairs, all of them involved isomiRs instead of the conventional miRNAs.

We also obtained another independent dataset from the recently updated miRTarBase database release 8.0<sup>31</sup>. This database contains experimentally validated functional and non-functional miRNA target sites, which are considered as positives in this dataset. We extended the 3' end of the mRNAs in the interactions and discarded the interactions that did not have mapping mRNA and miRNA ids in the respective databases used here and those interactions where the mRNA sequences were shorter than 30 nt. After this filtering, we obtained 14,144 miRNA–mRNA sequence pairs, 13,926 of which were functional and 226 of which were non-functional interactions based on the original study<sup>31</sup>. This dataset did not have any negatives.

**Deep learning model.** We designed a deep learning method called DMISO for miRNA/isomiR target sites and target mRNA identification. DMISO takes the miRNAs/isomiRs and their corresponding mRNA target site sequences as input and outputs a binary number to indicate whether a miRNA/isomiR interacts with its corresponding mRNA site. The architecture of DMISO is composed of two separate branches containing convolutional neural network (CNN) layers, a long short-term memory (LSTM) layer, and a fully connected neural network layer (Fig. 1B). The two convolutional layers are for the miRNA/isomiR and target site sequences, respectively. The LSTM layer combines the features detected by the two convolutional layers. The output of the LSTM layer is fed into a fully connected neural network to predict the label of the interaction.

The convolutional layer in each branch is 1-dimensional and consists of an array of 10 kernels, each with a size of  $4 \times 8$ . The kernels act as sliding windows to capture spatial features in input sequences by scanning the sequences. The convolutional layer does not have any padding around the input (padding = “valid”). The kernels are convolved across the input by 1 step (stride = 1). After the 10 kernels, the outputs of the two convolutional layers become the matrices of size  $10 \times 23$  and  $10 \times 53$ , respectively. The next layer in each branch is a 1-dimensional max pooling layer with a pooling size 4, which captures the maximum values within each  $10 \times 4$  window, sliding by 1 step (stride = 1), across the output of the respective convolutional layers. The output of the max-pooling layers in the miRNA/isomiR and target site branches is  $10 \times 20$  and  $10 \times 50$  matrices, respectively. Rectified Linear Unit activates the neurons in the convolutional layers of the two branches and the neurons in the dense layer. After the max-pooling step, the outputs of the two branches are merged to create a  $10 \times 70$  matrix and fed into a bidirectional LSTM (BLSTM) layer. The BLSTM layer processes the spatially connected features from both left to right and from right to left, generating a  $20 \times 70$  matrix output, which is then flattened to a vector of length 1400 and fed into a dense layer. The dense layer is a fully connected neural network with 100 neurons, which outputs a vector of size 100. This vector is used as an input to a logistic regression unit to generate the final prediction, where the sigmoid function is used.

Before training DMISO, the miRNA/isomiR and target site sequences are converted into  $4 \times 30$  and  $4 \times 60$  matrices, respectively, by applying one-hot encoding on every nucleotide in the sequences. That is, ‘A’, ‘T’, ‘C’, ‘G’ and ‘N’ are encoded into  $[1, 0, 0, 0]^T$ ,  $[0, 1, 0, 0]^T$ ,  $[0, 0, 1, 0]^T$ ,  $[0, 0, 0, 1]^T$  and  $[0.25, 0.25, 0.25, 0.25]^T$ , respectively. The fixed lengths 30 and 60 are the average length of the processed miRNAs/isomiRs and target sites in chimeric reads, respectively. To keep the fixed lengths, we removed extra nt from the ends of longer sequences and added additional ‘N’s to the ends of shorter sequences.

Batch normalization was used to train DMISO with mini-batches of 100 samples at a time. We calculated the loss of each prediction using the binary cross-entropy loss function, which is minimized by the Adam optimizer with a learning rate of 0.001<sup>32</sup>. To avoid overfitting, we had a dropout layer with 25% dropout rate after merging the two branches and two dropout layers with 50% dropout rate after the BLSTM layer and the dense layer. L1

regularization with the parameter value 0.01 was applied on the two convolutional layers and the dense layers to reduce overfitting. For the implementation of the deep learning model, Keras 2.3.1 version was used (<https://github.com/keras-team/keras/releases/tag/2.3.1>). DMISO model is executed with two inputs: miRNA/isomiR sequence and mRNA sequence. The model provides an output probability score from 0 to 1 and a binary prediction value of 0 and 1.

**Feature identification.** Many machine learning methods have been developed to select features<sup>33–43</sup>. Deep Learning models are infamous for being a black box when it comes to understanding the underlying features. But recent studies have focused on various strategies that can reveal the features or patterns learned by different types of machine learning models<sup>39–43</sup>. Here, two of the most popular feature identification methods, convolutional kernel analysis and input perturbation, were applied to discover important features for miRNA/isomiR–mRNA interactions<sup>24</sup>.

The convolutional kernel analysis method is suitable for a deep learning model that contains a convolutional layer<sup>24,39–41</sup>. This method is used to interpret the kernel weights of the convolutional layer after training the model. In this study, the miRNA/isomiR and mRNA sequences were scanned separately by the  $k$  length kernels of the two convolutional layers in DMISO, which captured the composition of  $k$ -mers in sequences that were important for the interaction between the miRNA/isomiR and mRNA sequences. Since the convolutional layers are the first layers in DMISO, the  $k$ -mer patterns captured should represent important features specific to the miRNA/isomiR and target sequences.

The input modification technique is another popular feature interpretation method<sup>24,39,41</sup>, where a part of the input is perturbed with random noise and the changes in the model prediction is recorded. The change in the model prediction after the modification to a part of the input represents the sensitivity of the model to that part of the input. Therefore, this method can help reveal the model's sensitivity patterns to different regions in input sequences. Here, we masked every contiguous region of length 4 in input sequences with “N” and recorded the respective changes in the prediction probabilities of the output layer. The changes should show important regions in terms of target binding.

**Comparison with existing tools.** DMISO was compared with three popular tools, TargetScan version 7.2<sup>44</sup>, miRanda 3.3a<sup>45</sup> and RNA22 version 2<sup>46</sup>, and two recently published deep learning-based tools, miRAW<sup>47</sup> and miTAR<sup>48</sup>, on the 20% CLASH test data, the CLEAR-CLIP data, and the miRTarBase data. TargetScan and miRanda take two separate files for the miRNA and mRNA sequences as inputs, while miRAW and miTAR take the interactions (miRNA–mRNA sequence pairs) as inputs. In the case of isomiR–mRNA pairs in the testing data, we used the isomiR sequences in place of miRNA sequences in the input. To run RNA22, the input miRNA/isomiR and mRNA sequences must be uploaded the RNA22 server with day-wise traffic restrictions. Because of this, obtaining results from the RNA22 server on a large dataset like ours is a time-consuming process. Therefore, while the other four tools were executed on the test datasets, RNA22 was evaluated by overlapping the test data sets with the pre-computed predictions of RNA22 on human (<https://cm.jefferson.edu/rna22-full-sets-of-predictions/>). A test interaction was considered a predicted positive by a tool, if the miRNA id and mRNA gene id of the test interaction matched with any of the predicted interactions as well as the mRNA target sequence locations overlapped with the corresponding predicted target sites.

## Results

**The characteristics of the identified interactions in CLASH.** We identified 12,170 miRNA–mRNA and 58,043 isomiR–mRNA interactions from the six CLASH samples (Supplementary Table S2). We observed each of the eight types of isomiRs, with a 3' addition isomiR hsa-miR-4268 occurring the most frequently in 3,565 interactions, while 96 isomiRs occurring only 10 times (Supplementary Fig. S1). Moreover, 98 isomiRs were found to participate in at least 100 interactions. Consistent with the previous studies<sup>6</sup>, there were more 3' isomiRs than other types in the CLASH data set (Supplementary Table S2). The number of isomiRs with nucleotide addition were more than other types. The number of SNP and MNP isomiRs were similar in the dataset. Despite the varied frequency of different types, there were at least 9 isomiRs from each of the eight types (Supplementary Table S3). Note that, these statistics were based on all the documented miRNAs in miRbase database. We also cross-referenced the miRNAs with curated miRNA database miRGeneDB<sup>49</sup> and found that 66 of the documented 268 miRNAs in miRGeneDB were in the CLASH dataset. Of them the most frequent isomiR was a 3' addition isomiR of hsa-miR-615-3p that occurred in 824 interactions. The number of 3' isomiRs was still higher (218) than 5' isomiRs (57) and polymorphic isomiRs (10).

Since we considered isomiRs covered by at least 10 reads, all de novo identified isomiRs were supported by their recurrent occurrence. The mean and median number of reads supporting these isomiRs were 53 and 19, respectively. There were 200 miRNAs with at least one identified isomiR. The number of identified isomiRs for a miRNA varied from 1 to 98. On average, a miRNA had around 6 isomiRs of different types. Despite the existence of different types of isomiRs, it was evident that one miRNA preferred specific types of isomiRs. In other words, for a given miRNA, a specific type of isomiRs occurred much more frequently. In fact, for all 67 miRNAs with at least 100 isomiR–mRNA interactions, at least one type of isomiRs occurred in significantly higher than expected frequencies (corrected Binomial test  $p$  value < 0.01).

The identified isomiRs were likely to be condition-specific. We compared the identified isomiRs in the CLASH interactions with the isomiRs in the CLEAR-CLIP interactions. The CLASH samples were from a healthy kidney cell line, while the CLEAR-CLIP samples were from a carcinomic liver cell line. Among the 1226 isomiRs and exact miRNAs identified in the CLASH data, 1203 (98.12%) were not identified in the CLEAR-CLIP data. If we considered the 1095 isomiRs and exact miRNAs supported by at least 50 reads in the CLASH data, 1078

Data set	Tools	Pos	Neg	AUROC	AUPR	F1	Precision	Recall	Specificity
CLASH test	DMISO	14,053	8319	0.9945	0.9969	0.9581	0.9286	0.9895	0.8714
	miRanda	14,053	8319	0.5873	0.6921	0.3032	0.9851	0.1792	0.9954
	RNA22	14,053	8319	0.5002	0.6283	0.0023	0.7273	0.0011	0.9993
	TargetScan	14,053	8319	0.5568	0.6686	0.2181	0.9563	0.1231	0.9905
	miRAW	14,053	8319	0.6119	0.6889	0.6552	0.7305	0.5940	0.6298
	miTAR	14,053	8319	0.5970	0.6832	0.5300	0.7639	0.4057	0.7882
CLEAR-CLIP	DMISO	14,684	1323	0.9427	0.9945	0.9398	0.9844	0.8990	0.8420
	miRanda	14,684	1323	0.5230	0.9211	0.1135	0.9790	0.0603	0.9856
	RNA22	14,684	1323	0.4993	0.9173	0.0004	0.6000	0.0002	0.9985
	TargetScan	14,684	1323	0.6001	0.9337	0.3670	0.9901	0.2252	0.9751
	miRAW	14,684	1323	0.5704	0.9283	0.7030	0.9367	0.5626	0.5782
	miTAR	14,684	1323	0.6904	0.9482	0.6596	0.9793	0.4972	0.8836

**Table 1.** Performance comparison on CLASH 20% test data and CLEAR-CLIP data.

Tools	Recall (overall)	Recall (functional)	Recall (non-functional)
DMISO	0.9164	0.9168	0.8938
miRanda	0.7045	0.7054	0.3540
RNA22	0.0199	0.0195	0.0221
TargetScan	0.7632	0.7643	0.3053
miRAW	0.6734	0.6696	0.5531
miTAR	0.0090	0.0090	0.0044

**Table 2.** Performance comparison on miRTarBase data.

(98.45%) were still not identified in the CLEAR-CLIP data. It was thus not the isomiR quality that made the difference of isomiRs in different experiments. In other words, isomiRs and their interacting sites are likely to be condition-specific.

We investigated the difference between the unique target sites of the 5' and 3' isomiRs and those of the exact miRNAs in the CLASH dataset. Among the 5742 CLASH target sites of exact miRNAs, 355 were common with the 5' isomiRs' while 2353 were common with the 3' isomiRs'. The lower number of common targets of exact miRNAs with the 5' than the 3' isomiRs corroborates that the 5' isomiRs have altered target specificity. Interestingly, the 3' isomiRs targeted much more unique mRNAs (5021) than other types, indicating that the 3' miRNA regions may also be important for miRNA/isomiR targeting.

**DMISO had good performance on miRNA and isomiR target prediction.** We evaluated DMISO by tenfold cross-validation on the training data (Supplementary Table S4). It showed over 99% area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPR), and more than 93% F1 scores, precision, and recall.

We also evaluated DMISO on three independent datasets: the left-out 20% CLASH test data, the CLEAR-CLIP data, and the miRTarBase data (Table 1). The performance of DMISO on the left-out CLASH test data was similar to that of the above cross-validation on the CLASH training data. That is, the AUROC and AUPR were over 99%, and the F1 score, precision and recall were more than 93%. When DMISO was tested on the CLEAR-CLIP data, the performance was slightly lower (94% AUROC, 99% AUPR, 94% F1, 98% precision and 90% recall). Because we did not have negative pairs in the miRTarBase dataset, we could only evaluate the recall of DMISO on this independent dataset (Table 2). DMISO had a recall of 92%, almost as good as the recall on the CLEAR-CLIP and 20% left-out CLASH test data.

The above analysis was on all interactions in the three test datasets. We further studied how well DMISO predicted the interactions involving different isomiR types instead of exact or wild-type miRNAs (Supplementary Table S5). DMISO showed consistently good performances on different types of isomiR–mRNA interactions. For instance, DMISO had an AUROC of 94%, an AUPR of 99%, a F1 score of 93%, a precision of 98%, and a recall of 89% on the CLEAR-CLIP data for the 5' isomiR–mRNA interactions. Overall, although the performance was close for different types of isomiR–mRNA interactions, it was slightly better on 3' isomiR–mRNA interactions on both CLASH and CLEAR-CLIP datasets.

We also compared DMISO with five other tools: miRanda, RNA22, TargetScan, miRAW, and miTAR. The input to every tool was the pairs of positive or negative miRNA/isomiR–mRNA target sites. DMISO showed a superior performance to all five tools in terms of every metric considered (Tables 1 and 2). For instance, DMISO had an AUROC of 94% and an AUPR of 99% on the CLEAR-CLIP data, compared with the highest AUROC of 69% and the highest AUPR of 95% by the other five tools. The five existing tools had a smaller precision and a much smaller recall than DMISO, which may be because DMISO was the first tool that took the isomiR–mRNA

interactions into account to train the models. It also highlighted the importance of considering such interactions for future miRNA target predictions.

Among the other five tools, the deep learning-based miRAW and miTAR had slightly higher AUROC and AUPR than those of the three classical tools (TargetScan, miRanda and RNA22) and much higher recall scores. This indicates that miRAW was able to capture the information in the non-seed regions, unlike the three tools that focused on the seed regions. This implied the importance of the non-seed regions for miRNA target site identification<sup>28,50,51</sup>. On the miRTarBase dataset, DMISO had a recall that was at least 10% larger than the recall scores of other tools. In contrast, miRAW and miTar had a worse recall than miRanda and TargetScan, suggesting that the deep learning models used in the two tools might not consider certain well-studied features of miRNA target interaction captured by the DMISO model (Table 2).

Among the above three independent test datasets, only the CLASH test data contained target sites for both miRNAs and isomiRs; the CLEAR-CLIP data had only isomiR–mRNA interactions; and the miRTarBase dataset had only miRNA–mRNA interactions. Therefore, to investigate how well each tool predicted miRNA–mRNA interactions only and isomiR–mRNA interactions only, we compared DMISO with the five tools on the separated miRNA–mRNA interactions and isomiR–mRNA interactions. DMISO again showed a much higher AUROC, AUPR and F1 than other tools on predicting the target sites of both miRNAs and isomiRs (Supplementary Table S6).

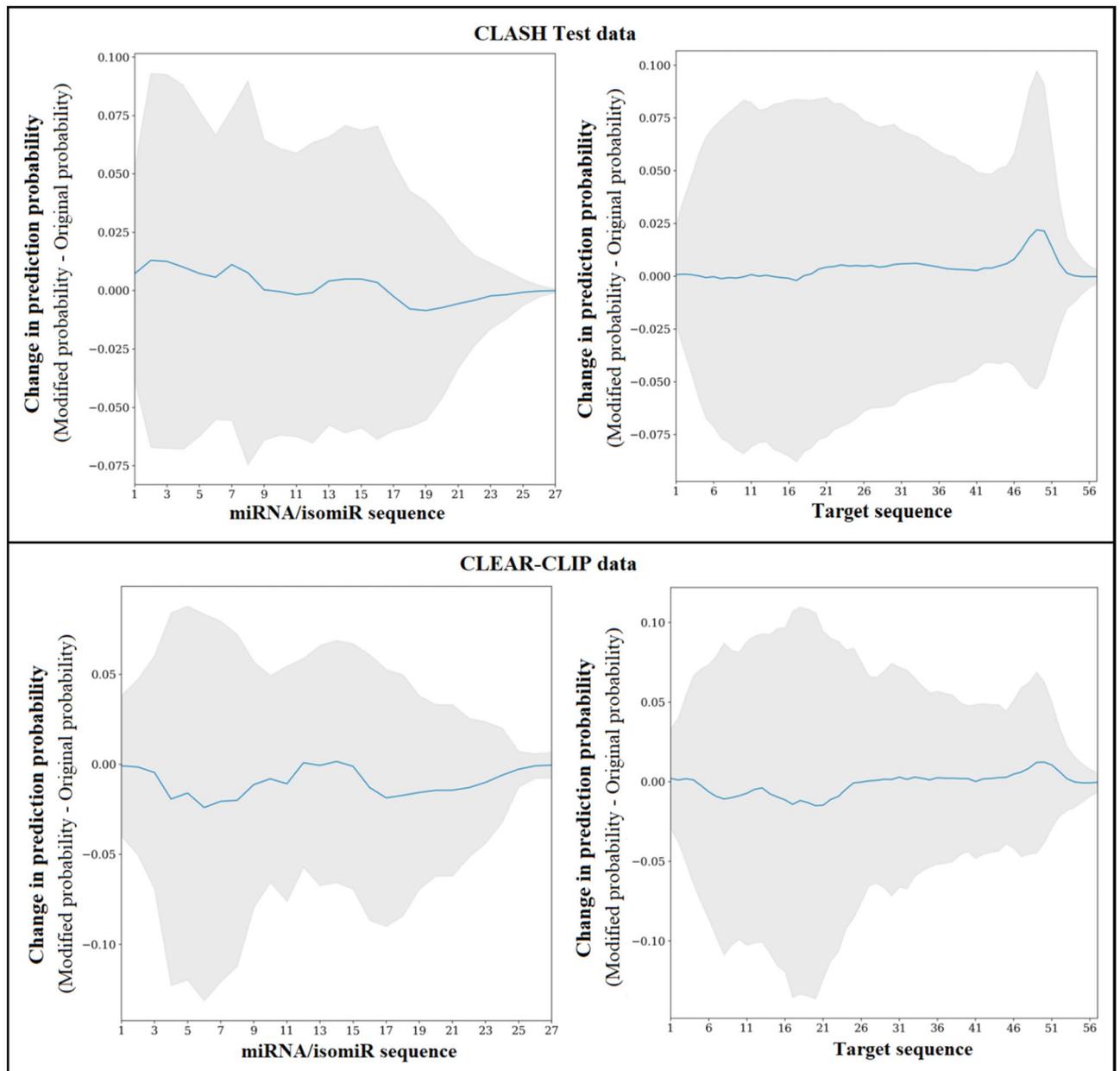
**Features analyzed.** Despite the high accuracy of deep learning models to solve a problem, these models are infamous for being unable to interpret what they learn from the data. To address this problem, we used two feature interpretation methods, convolutional kernel analysis and input perturbation<sup>24</sup>, as an attempt to unwind the learning process of the DMISO model.

Since DMISO has convolutional layers, we applied the convolutional kernel analysis method to extract the higher-level features learned by the model. In this process, the 10 kernel matrices, each of size  $4 \times 8$ , of the two convolutional layers for the miRNA (isomiR) and target branches were analyzed to find the weights of the A, T, C and G learned by the layers. In the beginning of the training process, the weight values of the kernels were randomly initialized in the forward pass and updated through backpropagation of the loss. After 500 epochs of training, 3 of the 10 kernel matrices in the miRNA (isomiR) branch and 7 of the 10 kernel matrices in the mRNA branch picked up certain weight values that might be commensurate with the protein binding motifs in the miRNA and mRNA sequences. For the remaining kernels in the respective branches, the weight values were too similar (25%) to be assigned to one of the 4 nucleotides. To find out the binding protein motifs, we compared the  $4 \times 8$  kernel matrices of both the miRNA and target branches with the JASPAR motif database on vertebrates<sup>52</sup>. The top motifs that each kernel matched with high significance were the protein binding motifs GATA1::TAL1, ZFP42, RARA::RXRG, RARA::RXRA, ESR2, ZFP42, ZBTB26, etc., all of which are CCHH and CCCH type Zinc finger proteins. Zinc finger proteins are well known as RNA binding proteins, which are essential to bind with ribonucleoproteins in a RNA-induced silencing complex<sup>53</sup>. This analysis shows that DMISO was able to recognize the binding profile of RNA binding motifs in miRNA/isomiR and mRNA sequences through the kernels of the convolutional layers, which has not been considered in the existing miRNA target prediction tools.

The other feature interpretation method that we used was the input perturbation technique. In this case, for every miRNA and target sequence in a dataset, a 4-nucleotide long mask consisting of “N” was applied on the sequence starting at every position of the miRNA and the target sequences. The changed value of DMISO’s prediction probability after applying the mask was recorded for the region in the corresponding sequences. The mask was then slid across the miRNA and the target sequences by one nucleotide every time. The average changes in the prediction probability after scanning all the miRNA and target sequences within a dataset gave us the regions that were most significant to DMISO within all the miRNA and target sequences.

The input perturbation method was applied to the 20% CLASH test data and CLEAR-CLIP datasets (Fig. 2). For each position, the mean (middle blue) and variation (gray area) of the sensitivity of DMISO to the changes at that position for all miRNA/isomiR and mRNA sequences in a dataset was recorded. In both datasets, positions 1–9 of the miRNAs/isomiRs had the most variation, which confirmed the importance of the seed region. However, not all positions in the seed were of the same importance. For instance, the first position was associated with the lowest variation while positions 3–5 were the highest in the seed for both datasets. The sensitivity variation of DMISO was also similar for both datasets from position 1 until position 18, and then dramatically decreased to the end. This might be because of the difference in miRNA/isomiR lengths (the shortest miRNA/isomiR was 17 nt long). It also suggested the importance of almost all miRNA positions instead of only the seed for target binding<sup>50</sup>. When the target mRNA sequences were changed, DMISO reacted more on the modifications in the 5′ regions of the target sequences, which further implied the importance of the 3′ regions of miRNAs/isomiRs that bound the 5′ region of the target mRNA sequences. In addition to the high sensitivity variation in the 5′ region, the positions around 50 (3′ region) of the target sequences showed a spike in sensitivity variation in both datasets. Since the 3′ region of the target corresponded to the seed region of the miRNA/isomiR, this confirmed the importance of the high-quality match in this seed region as well.

We also clustered miRNAs based on the binding sensitivity scores at every position of miRNAs. Our intuition was, if the model showed similar reaction patterns to the changes to two miRNAs, the two miRNAs might have similar binding patterns or features (Supplementary Fig. S2). We found that most of these miRNA clusters had many common targets, and a cluster of miRNAs often work in the same pathways (Supplementary Table S7).



**Figure 2.** The changes in DMISO's prediction probabilities with modification to different regions of the input miRNA (isomiR) and target sequences.

## Discussion

We developed a new method DMISO to efficiently predict miRNA/isomiR targets from mere sequence inputs (<http://hulab.ucf.edu/research/projects/DMISO>). The abundance of isomiRs in different cell lines and cell types makes it impossible to ignore for miRNA target prediction. The consideration of isomiRs in DMISO enables us to consider the intricate sequence changes that contribute to the miRNA/isomiR–mRNA interactions and thus make a more informed decision in identifying miRNA and isomiR target sites and targets. DMISO showed high performance on the cross-validation and three independent datasets. It outperformed the existing tools on these datasets, including three popular tools and two recently published deep learning-based tools. The better performance may be partially explained by the more generalized training data that contains both canonical and non-canonical isomiR/miRNA target sites, the consideration of the target difference of the isomiRs and miRNAs, the deep learning framework used, etc. Note that the CLASH and CLEAR-CLIP datasets we used here are the only available datasets that provide both miRNAs/isomiRs and their interacting mRNA target sites. The performance of DMISO may be further evaluated when more data is available in the future.

We applied two feature interpretation methods to understand miRNA/isomiR–mRNA target site features. The convolutional kernel analysis suggested the role of the RNA-binding protein-specific regions in miRNAs/isomiRs and the target mRNAs to form an interaction. The input perturbation technique confirmed that the 5' region of the miRNA/isomiR and the 3' region of the target sequence were highly important for their interactions.

Additionally, it showed that the middle part of the miRNA (isomiRs) and the target sequence might also have significant contributions.

We used 80% CLASH interactions to train DMISO and the remaining 20% interactions to test the tool in the above analysis. Although the miRNA target sites in the training and test datasets do not overlap, the training and testing datasets contained target sites of common miRNAs. This may raise a question if DMISO could do well to predict target sites of unknown miRNAs. To address this question, we trained a new DMISO model with the interactions for half of the total 217 CLASH miRNAs (109 miRNAs) and tested it on the interactions for the other half (108 miRNAs). We also compared the performance of DMISO on this testing dataset with the five tools. DMISO still outperforms other tools (Supplementary Table S8), which shows that the performance of DMISO is miRNA-independent.

We also analyzed the performance of the model after incorporating the abundance of miRNAs/isomiRs and mRNA transcripts. The read coverage information of miRNAs/isomiRs and mRNAs were added to the inputs of the final logistic regression layer. After training the model with 500 epochs, the model showed 99% and 94% AUROC, 100% and 99% AUPR scores on the CLASH test and CLEAR-CLIP datasets, respectively, which slightly increased compared with the original DMISO model on the respective datasets. This analysis indicates that the abundance of information can improve the model. Since it is not straightforward to obtain the abundance of information in practice, we preferred the original trained DMISO model.

We did not use available tools to detect isomiR in chimeric reads because such tools took away the flexibility to define a specific length of sequence change in an isomiR. In addition to considering only the top one pair of miRNA/isomiR–mRNA target sites from a chimeric read above, we also tried the top five potential pairs from each chimeric read to generate training and testing data. The conclusions made above still held, especially the better performance of DMISO (Supplementary Table S9). With a better understanding of isomiR–mRNA interactions, we may convert chimeric reads to miRNA/isomiR–mRNA target sites better in the future.

### Data availability

The CLASH and CLEAR-CLIP data were downloaded respectively from accession number GSE50452<sup>18</sup> and GSE73059<sup>19</sup> of the NCBI Genome Expression Omnibus. The miRTarBase release 8.0 data<sup>31</sup> was downloaded from the miRNA target sites provided by original literatures listed at [https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase\\_2022/php/download.php](https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2022/php/download.php). The TargetScan tool was downloaded from TargetScan Human version 7.2 code at [https://www.targetscan.org/cgi-bin/targetscan/data\\_download.vert72.cgi](https://www.targetscan.org/cgi-bin/targetscan/data_download.vert72.cgi). The miRanda tool version 3.3a was downloaded from <https://anaconda.org/bioconda/miranda/files>. The RNA22 version 2 full sets of prediction was downloaded for “Homo Sapiens, mRNA, ENSEMBL65, miRBase18” from <https://cm.jefferson.edu/rna22-full-sets-of-predictions/>. The miRAW tool was downloaded from [https://bitbucket.org/bipous/miraw\\_dl4mirna\\_binaries/src/master/](https://bitbucket.org/bipous/miraw_dl4mirna_binaries/src/master/). The miTAR tool was downloaded from <https://github.com/tjgu/miTAR>.

### Code availability

The source code and tool are available at <http://hulab.ucf.edu/research/projects/DMISO>.

Received: 29 March 2022; Accepted: 14 June 2022

Published online: 23 June 2022

### References

- Bartel, D. P. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
- Macfarlane, L. A. & Murphy, P. R. MicroRNA: Biogenesis, function and role in cancer. *Curr. Genom.* **11**, 537–561. <https://doi.org/10.2174/138920210793175895> (2010).
- Pritchard, C. C., Cheng, H. H. & Tewari, M. MicroRNA profiling: Approaches and considerations. *Nat. Rev. Genet.* **13**, 358–369. <https://doi.org/10.1038/nrg3198> (2012).
- Ruppert, S. M. *et al.* JunD/AP-1-mediated gene expression promotes lymphocyte growth dependent on interleukin-7 signal transduction. *PLoS ONE* **7**, e32262. <https://doi.org/10.1371/journal.pone.0032262> (2012).
- Wang, Y., Goodison, S., Li, X. & Hu, H. Prognostic cancer gene signatures share common regulatory motifs. *Sci. Rep.* **7**, 4750. <https://doi.org/10.1038/s41598-017-05035-3> (2017).
- Neilsen, C. T., Goodall, G. J. & Bracken, C. P. IsomiRs—The overlooked repertoire in the dynamic microRNAome. *Trends Genet.* **28**, 544–549. <https://doi.org/10.1016/j.tig.2012.07.005> (2012).
- Gebert, L. F. R. & Macrae, I. J. Regulation of microRNA function in animals. *Nat. Rev. Mol. Cell Biol.* <https://doi.org/10.1038/s41580-018-0045-7> (2019).
- Burroughs, A. M. *et al.* A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness. *Genome Res.* **20**, 1398–1410. <https://doi.org/10.1101/gr.106054.110> (2010).
- Wyman, S. K. *et al.* Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res.* **21**, 1450–1461. <https://doi.org/10.1101/gr.118059.110> (2011).
- Lee, L. W. *et al.* Complexity of the microRNA repertoire revealed by next-generation sequencing. *RNA* **16**, 2170–2180. <https://doi.org/10.1261/rna.2225110> (2010).
- Cloonan, N. *et al.* MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol.* **12**, R126. <https://doi.org/10.1186/gb-2011-12-12-r126> (2011).
- Llorens, F. *et al.* A highly expressed miR-101 isomiR is a functional silencing small RNA. *BMC Genom.* **14**, 104. <https://doi.org/10.1186/1471-2164-14-104> (2013).
- Loher, P., Londin, E. R. & Rigoutsos, I. IsomiR expression profiles in human lymphoblastoid cell lines exhibit population and gender dependencies. *Oncotarget* **5**, 8790–8802. <https://doi.org/10.18632/oncotarget.2405> (2014).
- Martí, E. *et al.* A myriad of miRNA variants in control and Huntington’s disease brain regions detected by massively parallel sequencing. *Nucleic Acids Res.* **38**, 7219–7235. <https://doi.org/10.1093/nar/gkq575> (2010).
- Telonis, A. G., Loher, P., Jing, Y., Londin, E. & Rigoutsos, I. Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity. *Nucleic Acids Res.* **43**, 9158–9175. <https://doi.org/10.1093/nar/gkv922> (2015).

16. Pinzón, N. *et al.* microRNA target prediction programs predict many false positives. *Genome Res.* **27**, 234–245. <https://doi.org/10.1101/gr.205146.116> (2017).
17. Ding, J., Li, X. & Hu, H. CCmiR: A computational approach for competitive and cooperative microRNA binding prediction. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btx606> (2017).
18. Helwak, A., Kudla, G., Dudnakova, T. & Tollervey, D. Mapping the human miRNA interactome by CLASH reveals frequent non-canonical binding. *Cell* **153**, 654–665. <https://doi.org/10.1016/j.cell.2013.03.043> (2013).
19. Moore, M. J. *et al.* MiRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. *Nat. Commun.* **6**, 1–17. <https://doi.org/10.1038/ncomms9864> (2015).
20. Bhattacharya, A. & Cui, Y. SomamiR 2.0: A database of cancer somatic mutations altering microRNA-ceRNA interactions. *Nucleic Acids Res.* **44**, D1005–1010. <https://doi.org/10.1093/nar/gkv1220> (2016).
21. Ding, J., Li, X. & Hu, H. MicroRNA modules prefer to bind weak and unconventional target sites. *Bioinformatics* **31**, 1366–1374 (2014).
22. Liu, W. & Wang, X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biol.* **20**, 18. <https://doi.org/10.1186/s13059-019-1629-z> (2019).
23. Wang, X. Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics* **32**, 1316–1322. <https://doi.org/10.1093/bioinformatics/btw002> (2016).
24. Talukder, A., Barham, C., Li, X. & Hu, H. Interpretation of deep learning in genomics and epigenomics. *Brief. Bioinform.* **22**, bbaa177. <https://doi.org/10.1093/bib/bbaa177> (2021).
25. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162. <https://doi.org/10.1093/nar/gky1141> (2019).
26. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinform.* <https://doi.org/10.1186/1471-2105-10-421> (2009).
27. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773. <https://doi.org/10.1093/nar/gky955> (2019).
28. Ding, J., Li, X. & Hu, H. TarPmiR: A new approach for microRNA target site prediction. *Bioinformatics* **32**, btw318 (2016).
29. Li, J. *et al.* Identifying mRNA sequence elements for target recognition by human Argonaute proteins. *Genome Res.* **24**, 775–785. <https://doi.org/10.1101/gr.162230.113> (2014).
30. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* <https://doi.org/10.1186/1748-7188-6-26> (2011).
31. Huang, H. Y. *et al.* MiRTarBase 2020: Updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.* **48**, D148–D154. <https://doi.org/10.1093/nar/gkz896> (2020).
32. Kingma, D. P. & Ba, J. in *3rd International Conference on Learning Representations*. (ed. Yoshua Bengio and Yann LeCun).
33. Kim, S. & Xing, E. P. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet.* **5**, e1000587. <https://doi.org/10.1371/journal.pgen.1000587> (2009).
34. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B Methodol.* **58**, 267–288 (1996).
35. Zhou, J., Liu, J., Narayan, V. A. & Ye, J. Modeling disease progression via fused sparse group lasso. *KDD* **1095–1103**, 2012. <https://doi.org/10.1145/2339530.2339702> (2012).
36. Ding, J., Cai, X., Wang, Y., Hu, H. & Li, X. ChIPModule: Systematic discovery of transcription factors and their cofactors from ChIP-seq data. In *Pacific Symposium on Biocomputing*. *Pacific Symposium on Biocomputing*, 320–331 (2013).
37. Talukder, A., Saadat, S., Li, X. & Hu, H. EPIP: A novel approach for condition-specific enhancer–promoter interaction prediction. *Bioinformatics* **35**, 3877–3883. <https://doi.org/10.1093/bioinformatics/btz641> (2019).
38. Zhao, C., Li, X. & Hu, H. PETModule: A motif module based approach for enhancer target gene prediction. *Sci. Rep.* **6**, 30043. <https://doi.org/10.1038/srep30043> (2016).
39. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838. <https://doi.org/10.1038/nbt.3300> (2015).
40. Quang, D. & Xie, X. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkw226> (2016).
41. Singh, S., Yang, Y., Póczos, B. & Ma, J. Predicting enhancer–promoter interaction from genomic sequence with deep neural networks. *Quant. Biol.* **7**, 122–137. <https://doi.org/10.1007/s40484-019-0154-0> (2019).
42. Cha, M. *et al.* A two-stream convolutional neural network for microRNA transcription start site feature integration and identification. *Sci. Rep.* **11**, 5625. <https://doi.org/10.1038/s41598-021-85173-x> (2021).
43. Barham, C., Cha, M., Li, X. & Hu, H. In *2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB)*. 22–28.
44. Agarwal, V., Bell, G., Nam, J. & Bartel, D. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* <https://doi.org/10.7554/eLife.05005> (2015).
45. Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.* **11**, R90. <https://doi.org/10.1186/gb-2010-11-8-r90> (2010).
46. Miranda, K. *et al.* A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. *Cell* **126**, 1203–1217. <https://doi.org/10.1016/j.cell.2006.07.031> (2006).
47. Pla, A., Zhong, X. & Rayner, S. miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts. *PLoS Comput. Biol.* <https://doi.org/10.1371/journal.pcbi.1006185> (2018).
48. Gu, T., Zhao, X., Barbazuk, W. B. & Lee, J.-H. miTAR: A hybrid deep learning-based approach for predicting miRNA targets. *BMC Bioinform.* **22**, 1–16. <https://doi.org/10.1186/s12859-021-04026-6> (2021).
49. Fromm, B. *et al.* MirGeneDB 2.1: Toward a complete sampling of all major animal phyla. *Nucleic Acids Res.* **50**, D204–D210. <https://doi.org/10.1093/nar/gkab1101> (2021).
50. Talukder, A., Li, X. & Hu, H. Position-wise binding preference is important for miRNA target site prediction. *Bioinformatics* **36**, 3680–3686. <https://doi.org/10.1093/bioinformatics/btaa195> (2020).
51. Wang, X. W. Composition of seed sequence is a major determinant of microRNA targeting patterns. *Bioinformatics* **30**, 1377–1383. <https://doi.org/10.1093/bioinformatics/btu045> (2014).
52. Fornes, O. *et al.* JASPAR 2020: Update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92. <https://doi.org/10.1093/nar/gkz1001> (2019).
53. Stefl, R., Skrisovska, L. & Allain, F. H. T. in *EMBO Reports*, vol. 6, 33–38 (European Molecular Biology Organization, 2005).

## Author contributions

H.H. and X.L. conceived and designed the study. A.T. performed the experiments. A.T., X.L., and H.H. analyzed the data. A.T., X.L. and H.H. wrote the manuscript. All authors proofread the manuscript.

## Funding

This work has been supported by the National Science Foundation [Grants 2120907, 1661414 and 2015838].

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-14890-8>.

**Correspondence** and requests for materials should be addressed to X.L. or H.H.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022