OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae616
https://doi.org/10.1093/bib/bbae616
Problem Solving Protocol

miCGR: interpretable deep neural network for predicting both site-level and gene-level functional targets of microRNA

Xiaolong Wu[1,2,‡], Lehan Zhang[2,3,‡], Xiaochu Tong[2,3], Yitian Wang[2,3], Zimei Zhang[2], Xiangtai Kong[2,3], Shengkun Ni[2,3], Xiaomin Luo[2,3], Mingyue Zheng[2,3], Yun Tang[1,*], Xutong Li[2,3]

[1]School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China
[2]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China
[3]School of Pharmacy, University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, China

[*]Corresponding authors. Xutong Li, Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China; University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, China. Tel.: +86-13821547729. E-mail: lixutong@simm.ac.cn; Yun Tang, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China. Tel.: +86-021-64251052. E-mail: ytang234@ecust.edu.cn
[‡]Xiaolong Wu and Lehan Zhang have contributed equally to this work.

Abstract

MicroRNAs (miRNAs) are critical regulators in various biological processes to cleave or repress translation of messenger RNAs (mRNAs). Accurately predicting miRNA targets is essential for developing miRNA-based therapies for diseases such as cancer and cardiovascular disease. Traditional miRNA target prediction methods often struggle due to incomplete knowledge of miRNA-target interactions and lack interpretability. To address these limitations, we propose miCGR, an end-to-end deep learning framework for predicting functional miRNA targets. MiCGR employs 2D convolutional neural networks alongside an enhanced Chaos Game Representation (CGR) of both miRNA sequences and their candidate target site (CTS) on mRNA. This advanced CGR transforms genetic sequences into informative 2D graphical representations based on sequence composition and subsequence frequencies, and explicitly incorporates important prior knowledge of seed regions and subsequence positions. Unlike one-dimensional methods based solely on sequence characters, this approach identifies functional motifs within sequences, even if they are distant in the original sequences. Our model outperforms existing methods in predicting functional targets at both the site and gene levels. To enhance interpretability, we incorporate Shapley value analysis for each subsequence within both miRNA sequences and their target sites, allowing miCGR to achieve improved accuracy, particularly with more lenient CTS selection criteria. Finally, two case studies demonstrate the practical applicability of miCGR, highlighting its potential to provide insights for optimizing artificial miRNA analogs that surpass endogenous counterparts.

Keywords: MicroRNA; target prediction; chaos game representation; deep learning; convolutional neural network

Introduction

MicroRNAs (miRNAs) are endogenous short (~22 nucleotides) non-coding RNAs that play critical regulatory roles in many biological processes by targeting messenger RNAs (mRNAs) for cleavage or translational repression [1]. Primarily through binding to the 3′ untranslated regions (UTRs) of target mRNAs, miRNAs modulate the expression levels of their targets [2]. Unlike plants, which require near-perfect base pairing, miRNA function in animals is primarily driven by the "seed region" (nucleotides 2–8) at the 5' end of the miRNA [3, 4]. This enables a single miRNA to regulate multiple genes within a signaling pathway, exerting a broad influence on cellular processes like cell growth, differentiation, proliferation, and apoptosis [5–8]. Notably, dysregulation of miRNA levels is implicated in the development and progression of various diseases, including cardiovascular disease and cancer. This suggests the exploration of miRNA-based therapeutic strategies, such as miRNA mimics or antagonists, to restore normal miRNA levels and potentially treat these diseases.

Identifying miRNA targets is a critical step in developing them as therapeutic agents. Two key questions need to be addressed: (i) which mRNA transcripts are functionally targeted by a specific miRNA (gene-level targets), and (ii) which specific regions (sites) on these mRNAs serve as binding sites for the miRNA (site-level targets) [9, 10]. Computational tools typically initiate the process by predicting candidate target sites (CTS) within the 3' UTRs of target mRNAs based on predefined criteria encompassing both canonical and non-canonical functional characteristics [11]. Canonical sites exhibit strong Watson-Crick (WC) base pairing between the miRNA's seed region and the mRNA typically involving at least five consecutive complementary base

Received: July 8, 2024. **Revised:** October 29, 2024. **Accepted:** November 12, 2024

pairs [11]. Conversely, non-classical sites may display interspersed non-canonical base pairing alongside some WC complementarity. Following CTS identification, functional validation of the miRNA-CTS interaction is essential to differentiate true functional targets from non-functional predictions. Generally, a miRNA-mRNA pair is predicted to be functional if at least one miRNA-CTS pair is predicted as functional, ultimately leading to the identification of reliable gene-level miRNA targets.

Two main types of features are employed for prediction: manually curated features and raw sequences. Manually curated features, such as interaction patterns, thermodynamic characteristics, sequence conservation, target site accessibility, and the number of binding sites, are readily interpretable but can lead to false negatives due to the complexity of miRNA-mRNA interactions [12]. Tools like TargetScanHuman [13] and miRDB [14] utilize these features for prediction. Recently, deep learning (DL) methods have emerged, directly utilizing raw base sequences as input to bypass the need for pre-defined features. This approach offers the potential for improved accuracy. Examples include miRAW (deep feed-forward neural network) [9], deepTarget [10], cnnMirTarget [15], and TargetNet (convolutional neural networks [CNNs]) [16], and miTAR (hybrid CNN-RNN [recurrent neural networks]) [17], TEC-miTarget (Transformer-CNN) [18]. However, these DL-based methods often encode raw miRNA and mRNA sequences as one-hot vectors or utilize embedding layers to generate dense vectors. These strategies, which represent a one-dimensional characterization of the sequences, may limit the model's ability to capture long-range dependencies within the sequences.

Here, we introduce miCGR, an end-to-end deep learning framework miCGR designed for predicting the functional targets of miRNAs. The proposed method leverages chaos game representation (CGR) [19–23] to transform miRNA and mRNA CTS sequences into 2D images. This approach captures information on similar motifs, even when they are distantly located in the original sequences, unlike traditional one-dimensional characterization methods. Furthermore, miCGR emphasizes the importance of prior knowledge for accurate prediction by explicitly integrating seed regions and subsequence positions by representing them in additional image channels. Two separate CNNs extract latent features from the miRNA and CTS image representations. These features are subsequently concatenated and fed into a feed-forward neural network to predict miRNA-CTS binding. A transcript is considered a target if at least one CTS site is predicted positive. To assess miCGR's performance, we compared it with existing models on independent external test sets, evaluating its accuracy in predicting both site-level and gene-level functional targets. We further employed Shapley explanation methods [24, 25] to quantify the contribution of sub-fragments within the miRNA and CTS sequences to the predictions. These findings not only enhance our understanding of miRNA-mRNA interactions but also hold promises for the development of more effective miRNA-based therapeutics.

Method

Processing microRNA target datasets

Public miRNA-mRNA interaction data repositories contain two types of pairing data: gene-level and site-level interaction data. Gene-level interaction datasets indicate whether miRNAs target specific mRNAs, while site-level datasets provide labels for each CTS within the UTRs of mRNAs. To evaluate the proposed approach, we collected a site-level interaction dataset from miRAW [9] for model training, validation and internal testing (miRAW-site-level dataset). Besides, we collected two gene-level interaction datasets from miRAW (balanced-gene-level dataset) and cnnMirTarget (imbalanced-gene-level dataset) [15] for external test. Corresponding miRNA sequence information was obtained from miRbase [24], and UTR sequence for mRNAs were sourced from the UCSC Table Browser [25], respectively, utilizing the longest UTR sequence for each mRNA. Supplementary Table S1 available online at http://bib.oxfordjournals.org/ provides an overview of the dataset sources, the number of positive and negative sample pairs, and the counts of mRNA and miRNA sequences.

MiRAW-site-level dataset: In the miRAW study, Albert Pla et al. [9] collected a large amount of experimentally validated miRNA-mRNA pairs from two resources: DIANA-TarBase [26] and miRTarBase [27]. They constructed a site-level dataset by cross-referencing the miRNA-mRNA pairs with PAR-Clip [28], CLASH [29] and TargetScanHuman 7.2 [13]. This resulted in a total of 33,142 positive and 32,284 negative site-level pairs, encompassing both canonical and non-canonical interactions. After filtering the obsolete miRNAs without sequence information from current version of miRbase [30], the final site-level dataset retained 8642 genes, with 32,466 positive and 32,105 negative pairs. For each pair, the miRNA sequence information and the CTS sequence (30 nucleotides) with five additional upstream and downstream nucleotides were recorded, resulting in a final CTS size of 40 nucleotides ($n = 40$).

Balanced-gene-level dataset: In the miRAW study, Albert Pla et al. [9] constructed 10 balanced gene-level datasets for model external testing, each containing 548 positive and 548 negative gene-level pairs, totaling 4023 genes. The intersection between the miRAW-site-level dataset and the balanced-gene-level dataset included 1896 genes, representing 47.1% of the latter. The miRNA sequence information and the UTR sequence information were recorded in each pair.

Imbalanced-gene-level dataset: In the cnnMirTarget study, Zheng et al. generated an unbalanced gene-level dataset from DIANA-TarBase [31] and miRTarBase [32]. This dataset initially comprised 7815 positive and 281 negative gene-level pairs. After excluding pairs overlapping with the miRAW site-level dataset and obsolete miRNAs without sequence information according to the current version of miRBase, the dataset was refined to include 7615 positive and 274 negative gene-level interaction pairs across 2689 genes. The intersection between the miRAW-site-level dataset and the imbalanced-gene-level dataset consisted of 927 genes, accounting for 34.5% of the latter.

The chaos game representation algorithm

Chaos Game Representation (CGR) is a mathematical visualization technique used to represent genetic sequences in a 2D space. This method provides a unique way to capture the complexity and patterns within the genetic sequence [21, 33, 34]. The principle of the CGR algorithm involves using a loop iteration function to determine the position of each subsequence in a square plot and filling the regions in the plot with the frequency of occurrence of each subsequence [22]. Figure 1 summarizes the basic concept of the CGR algorithm when the subsequence length is 4 (k-mer = 4). Initially, a square plot is set up, with each corner representing one of the four nucleotides (A, U, C, G). The plot is then partitioned into $N \times N$ regions, where $N = 2^{k\text{-mer}}$, and each grid in the plot represents the frequency of subsequences of length k within the genetic sequence (Fig. 1A and 1B). As more subsequences are plotted, patterns begin to emerge that reflect the underlying

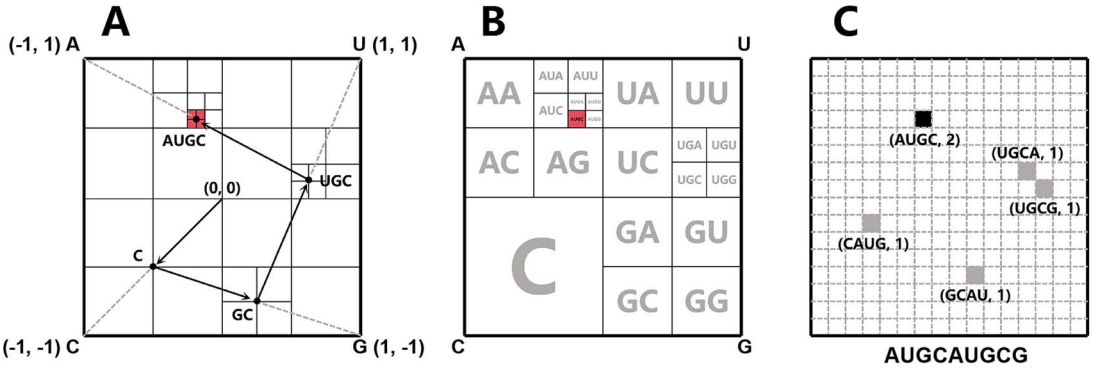

Figure 1. Chaos game representation (k-mer = 4) for genetic sequence. (**A**) the iterative process of plotting a subsequence "AUGC" using CGR algorithm. Starting from the center of the square, follow the inverted order of the sequence (i.e., "CGUA"). The first nucleotide "C" positions the point halfway between the center (0,0) and the "C" corner; for the next nucleotide "G", the point is plotted halfway between the current point and the "G" corner. This process repeats for each nucleotide in this subsequence. (**B**) the mapping for subsequence "AUGC" in CGR representation space. A grid-based counting approach is used for determining the position of the subsequence. Dividing the CGR into four quadrants, the upper left quadrant contains points representing subsequences starting with the nucleotide "a". further subdivision of this quadrant yields subsequences starting with "AA", "AC", "AG", and "AU". Continuing this process for 4-mers, the CGR image is partitioned into N × N regions, where $N = 2^{k\text{-mer}}$, or 16, and the sequence "AUGC" is eventually mapped. (**C**) the CGR visualization of a DNA sequence "AUGCAUGCG". With k-mer = 4, this sequence is divided into "AUGC", "UGCA", "GCAU", "CAUG", "AUGC", and "UGCG". For each subsequence, its region in (B) is identified, with the color depth representing the subsequence's frequency. Regions with frequent subsequences will appear denser. The numbers in parentheses represent frequencies.

structure of the genetic sequences, with higher frequency numbers exhibited by deeper colors (Fig. 1C). The resulting plot can be analyzed to identify motifs, repeats, and other features in the DNA sequence. Points or regions with similar prefixes in their sequences will cluster together (e.g., "UCGA" is close to "UGCG" in Fig. 1C), revealing underlying biological patterns.

The neural model architecture

CNN is a deep learning method which is good at learning information from data in grid form, such as pictures [17, 35]. Encoded by the method mentioned above, both miRNA and CTS sequence were transformed into 2D arrays. In our implementation, three different sizes of convolution kernels were employed to extract more diverse local features, which contributed to the improvement of model performance [36]. Additionally, non-linearization by rectified linear unit and BatchNorm2D operations are applied after each convolutional layer, which preventing overfitting and mitigating gradient vanishing. The size of images extracted using the CGR algorithm varies with different k-mer sizes, which leads to different specific parameters in the CNN.

The evaluation metrics

To evaluate the performance of the model, six commonly used binary classification metrics were employed, including accuracy, sensitivity, specificity, F1 score, positive predictive value (PPV), and negative predictive value (NPV). The formulas for each metric are shown in Table 1, where TP represents true positive, FP represents false positive, TN represents true negative, and FN represents false negative. F1 score is an indicator commonly used in statistics to measure the accuracy of binary classification (or multi-task binary classification) models. It is an evaluation of the harmonic average of precision and recall of a classification model. As a result, F1 score can better reflect the true performance of the model.

Benchmark models comparison

For model-based methods, including miRAW, TargetNet, deepTarget, cnnMirTarget, TEC-miTarget, and miTAR, we implemented benchmark models using the default hyperparameter settings specified in their original publications. In our re-implementation, all methods were trained on the miRAW site-level dataset, with the balanced gene-level dataset from miRAW serving as the external test set. Since the original publications also trained these models on the miRAW site-level dataset, we selected the superior results from either the reported outcomes or our re-implementation (except for cnnMirTarget, which used a different training dataset), as summarized in Table 7. For web-based tools (TargetScanHuman 7.2 and miRBD 6.0), we directly retrieved the prediction results from their respective websites. MiCGR was run 30 times, while the re-implementation experiments were conducted once without repetition.

Table 1. The evaluation metrics and their calculations.

Metric	Calculation
Accuracy	(TP + TN) / (TP + FP + TN + FN)
Sensitivity (recall)	TP / (TP + FN)
Specificity	TN / (TN + FP)
F1 score	2 ∗ (precision ∗ recall) / (precision + recall)
PPV (precision)	TP / (TP + FP)
NPV	TN / (TN + FN)

Functional enrichment analysis

The top 100 predicted target genes for hsa-miR-552-3p were submitted to the web tool Metascape [37] for functional enrichment analysis. The enrichment results were downloaded directly from the website.

Multiple sequence alignment

The top 10 predicted target miRNAs for PCSK9 were subjected to multiple sequence alignment using the MUSCLE algorithm [38].

Results

The architecture of miCGR

As illustrated in Fig. 2, miCGR utilizes enhanced CGR to transform miRNA and CTS sequences of mRNA into 2D images, specifically CGR representations for both miRNAs and mRNAs. These

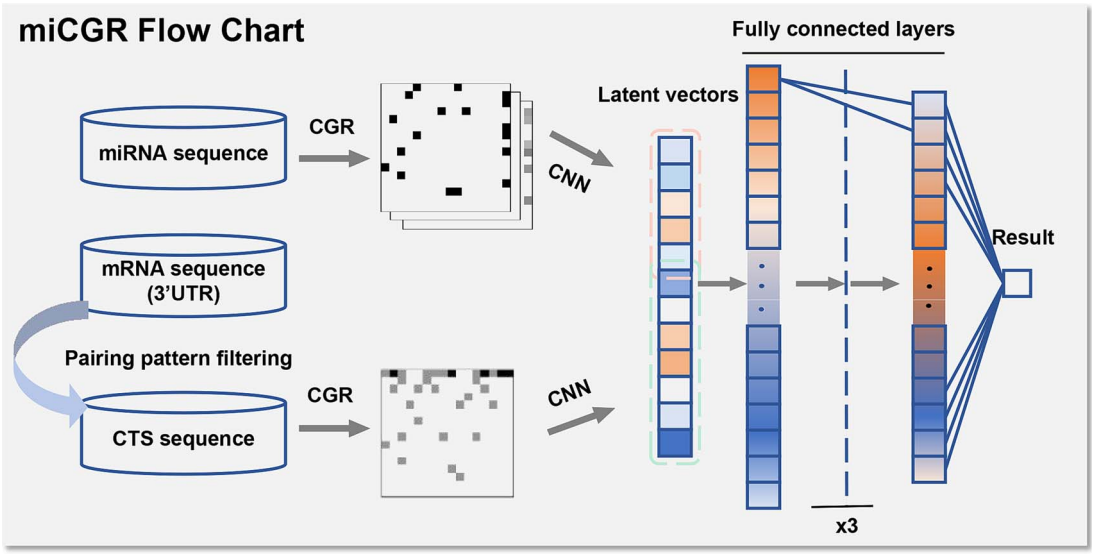

Figure 2. Schematic illustration of the flow of miCGR.

Table 2. Pairing patterns which determine candidate target sites.

Pairing pattern		Description
Stringent pairing patterns [11] (consecutive)	8-mer	WC pairings among the miRNA nucleotides 2–8 with an A on the nucleotide 1
	7-mer-m8	WC pairings among the miRNA nucleotides 2–8 with a B on the nucleotide 1
	7-mer-A1	WC pairings among the miRNA nucleotides 2–7 with an A on the nucleotide 1 and a Ø on nucleotide 8
	6-mer	WC pairings among the miRNA nucleotides 2–7 with an B on the nucleotide 1 and a Ø on nucleotide 8
	6-mer-A1	WC pairings among the miRNA nucleotides 2–6 with an A on the nucleotide 1 and a Ø on nucleotide 7
	offset-7-mer	WC pairings among the miRNA nucleotides 3–9 and a Ø on nucleotide 2
	offset-6-mer	WC pairings among the miRNA nucleotides 3–8 with an A on the nucleotide 1 and Ø on nucleotide 2 and nucleotide 9
Lenient pairing patterns	10-mer-m6	At least 6 WC pairings among the miRNA nucleotides 1–10
	10-mer-m7	At least 7 WC pairings among the miRNA nucleotides 1–10
	offset-9-mer-m7	At least 7 WC pairings among the miRNA nucleotides 2–10

[a]Stringent pairing patterns consist of contiguous complementary base pairing, meaning there are no gaps or mismatches, while lenient pairing patterns allow for gaps or mismatches within the pairing. In this context, any nucleotide except adenine (A) is represented as B, and any non-Watson-Crick pairing is denoted as ∅, following the conventions established in [11]. Here, WC is short for Watson-Crick.

images are then input into separate CNNs. These CNNs extract latent features from the image representations, which are then concatenated and fed into a feed-forward neural network to predict functional/non-functional miRNA-CTS binding. MiCGR was trained exclusively on a site-level dataset. The gene-level miRNA-mRNA prediction builds upon this foundation, whereby a transcript is considered a target if at least one CTS site is predicted to be positive. A detailed description of the datasets collected and utilized in miCGR can be found in the **Methods** section and Supplementary Table S1 available online at http://bib.oxfordjournals.org/.

The enhanced chaos game representation module

Given the established importance of the seed region for miRNA-mRNA binding [39–42], we modified the CGR algorithm to explicitly leverage this prior knowledge. The seed region is a conserved heptamer situated at positions 2–8 on the 5′ end of the miRNA. Stringent pairing patterns follow contiguous WC base pairing—requiring perfect complementary pairing without gaps or mismatches—between the miRNA and CTS within the seed region. Notably, perfect WC base pairing between the miRNA and mRNA is not always necessary; pairing can still occur if the seed sequence demonstrates certain complementarity with the CTS. To robustly identify CTS sequences and encompass the majority of known classical and non-classical functional sites, miCGR employs lenient pairing patterns derived from miRAW [9]. These robust lenient pairing patterns, alongside stringent classical contiguous pattern for performance comparison, are detailed in Table 2, with a diagram illustrating the lenient pairing pattern "offset-9-mer-m7" shown in Fig. 3A.

To capture the specificity of seed region, our approach encodes both the entire miRNA sequence (Fig. 3B, left) and a dedicated channel representing solely the seed region sequence (Fig. 3B, middle). Furthermore, we explicitly incorporate the relative positions of all subsequences within the miRNA into the model. This integration is achieved by replacing the original subsequence frequencies in the CGR representation with their respective relative positions (Fig. 3B, right). The relative position is defined as $(L - P)/L$, where L is the length of the sequence and P is position of the first nucleotide of the k-mer fragment in the original sequence (see Code for more detail). This additional channel enables miCGR

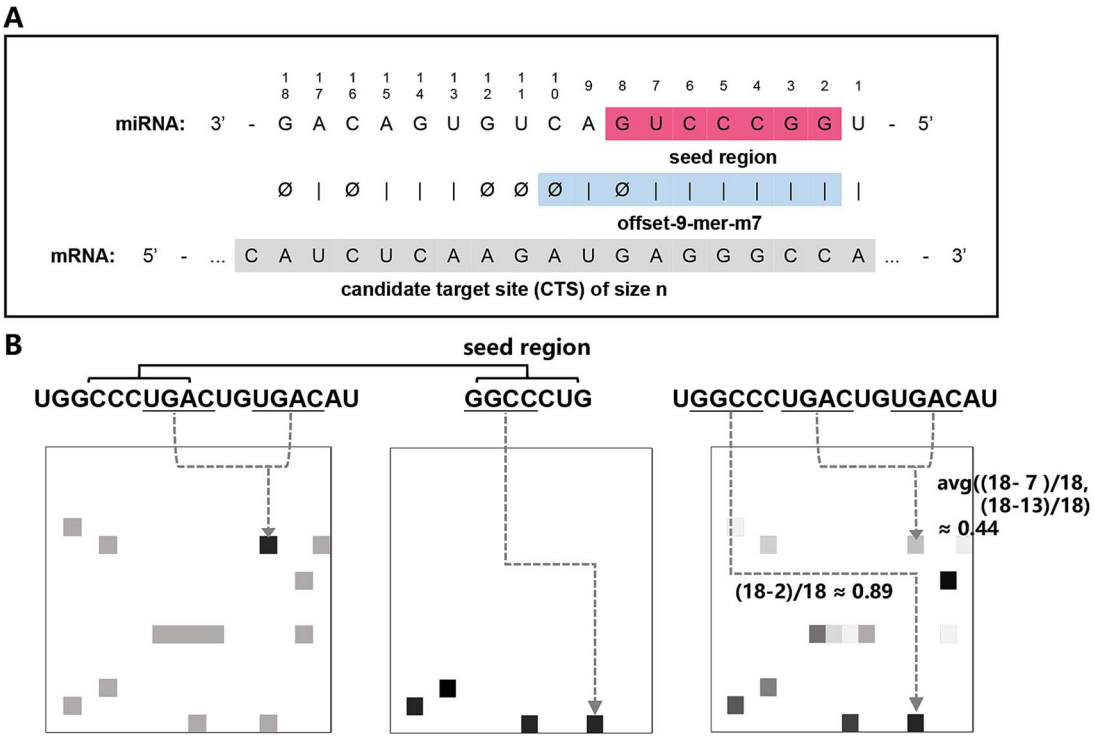

Figure 3. (**A**) Schematic diagram of the lenient pairing pattern "offset-9-mer-m7" between miRNA and mRNA. "GGCCCUG" (5' to 3') denotes the seed region of the miRNA, "Ø|Ø||||||" represents the pairing region of offset-9-mer-m7 between miRNA and mRNA, and "CAUCUCAAGAUGAGGGCCA" indicates the CTS of size n. since the CTS sequence (30 nucleotides) includes five additional upstream and downstream nucleotides, n = 40. **(B)** Modified chaos game representations (k-mer = 4) for a miRNA sequence. This includes the original CGR representation for the entire miRNA (left), the representation for the seed region of the miRNA (middle), and the representation for the relative positions of subsequences in the whole miRNA (right). For example, with the sequence "UGGCCCUGACUGUGACAU", the region for "UGAC" in the original CGR representation (left) records its frequency as 2. The seed region "GGCCCUG" is represented in the middle panel. In the representation for relative positions, the value for the "GGCC" region is (18–2)/18 ≈ 0.89. For the "UGAC" sequence, which appears twice, the region's value is the average of the two relative positions, calculated as avg((18–7)/18, (18–13)/18) ≈ 0.44.

to discern sequential features specific to nucleotides across distinct miRNA regions.

The combined representation of these three 2D arrays yields a three-channel image for each miRNA. For mRNA, CTS sequences are encoded into single-channel images using the original CGR algorithm.

Model implementation and optimization

The miCGR model was trained on the MiRAW-site-level dataset, which was split into training, validation, and internal testing sets in a 16:4:5 ratio. Gene-level miRNA-mRNA predictions were inferred from site-level predictions, with a transcript classified as a target if at least one CTS was predicted to be positive. The model was trained to minimize cross-entropy between the true labels and the predicted results as follows:

$$loss = -\frac{1}{n}\sum [y \log p + (1-y)\log(1-p)]$$

where p refers to the prediction result and y refers to the true label that refers to whether a CTS can serve as a binding site for miRNA.

The Adam optimizer [43] was used for training, and early stopping was employed to prevent overfitting by terminating the process when no further improvement was seen on the validation set. A grid search strategy was implemented to fine-tune key hyperparameters, including k-mer size, batch size, learning rate, and dropout rate, with their respective search ranges detailed in Table 3. Optimal performance was achieved with a k-mer size of 6, batch size of 128, learning rate of 0.0005, and dropout rate of 0.5. The loss curves for the training, validation, and test datasets are illustrated in Supplementary Fig. S1. The impact of different k-mer sizes on performance is presented in Table 4. Training the model required ~ 0.5 hours, spanning 165 epochs at 13 seconds per epoch, using a NVIDIA TESLA V100 GPU on an Intel platform.

Table 3. Optimal hyperparameters and performance of miCGR on miRAW-site-level dataset for different k-mer.

Hyperparameters	Searching space
k-mer	2; 3; 4; 5; **6**; 7
batch size	32; 64; **128**; 256
learning rate	0.01; 0.005; 0.001; **0.0005**; 0.0001
Dropout	0.0; 0.2; **0.5**

[a]Results in bold text are the best value for hyperparameters.

Model performance on miRAW-site-level dataset

Given that gene-level miRNA-mRNA prediction builds upon site-level predictions, we first evaluated our model's performance at the site level using the miRAW-site-level dataset. The Enhanced Chaos Game Representation incorporates three channels for each miRNA, which represent the miRNA sequence, the seed region sequence, and the relative positions of all subsequences. To evaluate the influence of integrating prior knowledge about the seed region and sequential features of subsequences within the miRNA on miCGR's predictive efficacy, we conducted ablation experiments with three variants: miCGR_NOS, which excludes the seed region channel; miCGR_NOP, which excludes the

Table 4. Performance comparison of miCGR on the miRAW-site-level dataset with varying hyperparameters and different k-mers.

K-mer	Batch Size	Learning rate	Dropout	F1 score	Accuracy	Sensitivity	Specificity	PPV	NPV
2	256	0.001	0.5	0.9415	0.9406	0.9433	0.9379	0.9397	0.9415
3	32	0.0001	0.5	0.9476	0.9467	0.9502	0.9432	0.9449	0.9486
4	64	0.0001	0.5	0.9600	0.9595	0.9597	0.9593	0.9603	0.9587
5	128	0.0001	0.5	0.9623	0.9618	**0.9627**	0.9608	0.9619	**0.9617**
6	128	0.0005	0.5	**0.9661**	**0.9658**	0.9623	**0.9712**	**0.9710**	0.9616
7	32	0.0001	0.5	0.9604	0.9600	0.9588	0.9611	0.962	0.9579

[a]Results presented in bold text indicate the best performance.

Table 5. Performance comparison of methods on the miRAW-site-level dataset.

Methods	F1 score	Accuracy	Sensitivity	Specificity	PPV	NPV
miCGR	**0.9661**	**0.9658**	**0.9623**	**0.9712**	**0.9710**	**0.9616**
miCGR_NOS	0.9545	0.9549	0.9542	0.9550	0.9551	0.9539
miCGR_NOP	0.9535	0.9517	0.9553	0.9539	0.9561	0.9508
miCGR_NOSP	0.9412	0.9430	0.9444	0.9487	0.9483	0.9401
miCGR_NOCGR	0.9480	0.9477	0.9493	0.9460	0.9469	0.9487
TEC-miTarget	0.9645	0.9647	0.9585	0.9710	0.9706	0.9590
miTAR	0.9652	0.9654	0.9609	0.9697	0.9695	0.9613
miRAW	0.935	0.935	0.935	0.938	0.935	0.9320
deepTarget	-	0.922	-	-	-	-

[a]miCGR_NOS: Excludes the seed region sequences while retaining the relative positional information of miRNA subsequences. miCGR_NOP: Excludes the relative positional information while retaining the seed region sequences. miCGR_NOSP: Excludes both the seed region and relative positional information. miCGR_NOCGR: Excludes CGR modules while retaining both the seed region sequences and relative positional information, processed with one-dimensional CNNs. The models TEC-miTarget, miTAR, miRAW, and deepTarget were re-implemented using the default hyperparameter settings specified in their original publications. We report the highest performance results in this table from either our re-implementation or the outcomes published by Yang et al. [18] and Gu et al. [44], as miCGR utilizes the same test set as these studies. Results presented in bold text indicate the best performance.

relative positional information channel; and miCGR_NOSP, which excludes both. As expected, models lacking prior knowledge exhibited decreased performance, with both seed region and relative positional information contributing equally to miCGR's effectiveness (Table 5).

To further assess the significance of the CGR modules, we conducted an additional ablation study with a variant, miCGR_NOCGR, where CGR modules were removed while retaining the original input information. In this variant, the input sequences (the entire sequence and seed region sequence for miRNA, and the entire CTS sequence for mRNA) were each encoded as one-hot vectors. Additionally, the relative positions of all miRNA subsequences were represented as a one-dimensional vector. These vectors were processed independently using standard one-dimensional CNNs and subsequently stacked, following the method applied in deepTarget [10]. The performance of miCGR_NOCGR surpassed that of deepTarget, which employs a similar one-dimensional CNN approach but lacks seed region sequences and relative positional information, thereby highlighting the importance of this prior knowledge. Furthermore, performance comparisons between miCGR and miCGR_NOCGR indicate that the CGR modules significantly enhance miCGR's predictive accuracy compared to the one-dimensional CNN models (Table 5).

Furthermore, we compared miCGR with recent models including miTAR, miRAW, and deepTarget on the miRAW-site-level dataset. The results indicate that the miCGR model shows improvements over these models across all six evaluation metrics (Table 5). Notably, while miTAR performs slightly worse than miCGR, it outperforms miCGR_NOSP. MiTAR incorporates an RNN module to capture continuous positional information of miRNA nucleotides. These results underscore the importance of such knowledge for site-level miRNA-mRNA prediction.

Model performance on gene-level dataset

Since the miCGR model was able to achieve acceptable prediction results on site-level dataset, we were eager to know if it would perform well in the gene-level prediction problem. The algorithm is supposed to predict the binding status between miRNA and all CTS sequences across the 3'UTR of the target gene. If at least one CTS site is predicted as positive, the gene is considered a positive prediction, while the gene is considered a negative prediction if all CTS sites are predicted as negative. For each gene, the longest mRNA sequence was used.

Initially, we evaluated the effect of various pairing pattern conditions on the miCGR model. The performance on the balanced-gene-level dataset is depicted in Table 6. Under the stringent pairing pattern condition, miCGR exhibited high specificity but low sensitivity, indicating that the model trained under this condition struggled to detect functional CTSs. Conversely, under the lenient pairing pattern condition, while the accuracy and F1 score of miCGR under the 10-mer-m7 and 10-mer-m6 pairing patterns were slightly superior to those under the most relaxed offset-9-mer-m7 pairing pattern, the specificity was notably diminished as the CTS count significantly increased. These results indicate that there is a trade-off between sensitivity and specificity. While a stricter pairing pattern reduces miCGR's sensitivity, a looser one compromises its specificity, potentially averting a high false positive rate in practical applications.

Therefore, the subsequent assessment of candidate models' predictive capabilities at the gene level will be based on the offset-9-mer-m7 pairing pattern, as miCGR demonstrates a more balanced performance across all six evaluation metrics, indicating more robust prediction results with a relatively smaller CTS count.

Secondly, we compared functional miRNA target classification performance of seven different prediction algorithms: miRAW [9],

Table 6. Performance comparison of miCGR on the balanced gene-level dataset with different pairing patterns.

Methods		F1 score	Accuracy	Sensitivity	Specificity	PPV	NPV	Average CTS count
miCGR under stringent pairing patterns		0.6854	0.7216	0.6103	**0.8318**	**0.7822**	0.6833	~2
miCGR under	offset-9-mer-m7	0.8009	0.7902	0.8479	0.7331	0.7590	0.8296	~8
lenient pairing	10-mer-m7	0.8197	**0.7986**	0.9200	0.6782	0.7392	0.8954	~24
patterns	10-mer-m6	**0.8223**	0.7927	**0.9635**	0.6234	0.7172	**0.9452**	~81

[a]The definitions for "stringent pairing pattern" and "lenient pairing patterns" can be found in Table 2. The term "Average CTS count" denotes the average count of candidate sites per gene within the 3'UTR for the corresponding pairing pattern. Results presented in bold text indicate the best performance.

Table 7. Performance comparison of methods on the gene-level datasets.

dataset	Methods	F1 score	Accuracy	Sensitivity	Specificity	PPV	NPV
balanced-gene-level[a]	miCGR	**0.8009**	**0.7902**	0.8479	0.7331	0.7590	0.8296
	miRAW-7-2:10-NF	0.7440	0.6880	0.9053	0.4710	0.6311	0.8327
	TargetNet-all[b]	0.7739	0.7251	0.9411	0.5091	0.6572	0.8966
	deepTarget2020	0.7683	0.7673	0.7689	0.7628	0.7649	0.7700
	TEC-miTarget	0.7817	0.7796	0.7892	0.7701	**0.7744**	0.7852
	miTAR	0.7995	0.7684	**0.9624**	0.5675	0.6925	**0.9428**
	miRBD6.0	0.2866	0.5511	0.1807	**0.9215**	0.6958	0.5294
	TargetScanHuman7.2	0.2050	0.5191	0.1241	0.9142	0.5906	0.5107
imbalanced-gene-level	miCGR	**0.9269**	**0.8661**	**0.8748**	0.5776	0.9856	**0.1222**
	cnnMirTarget	0.7899	0.6597	0.6627	0.5765	-	-
	deepTarget2020	0.8755	0.7822	0.7890	0.560	0.9833	0.0735
	miTAR	0.8997	0.8212	0.8262	0.6552	0.9876	0.1018
	TEC-miTarget	0.8115	0.6893	0.6886	0.7111	0.9876	0.0638
	miRBD6.0	0.6106	0.4571	0.4409	**0.9075**	**0.9925**	0.0551
	TargetScanHuman7.2	0.4733	0.3314	0.3112	0.8932	0.9878	0.0446

[a]miCGR is under the pairing pattern of offset-9-mer-m7. The results tested on balanced-gene-level dataset were averaged across the ten datasets. [b]The results for TargetNet reflect the best performance reported by Min et al. [16]. Target utilizes the exact same gene-level test set as the others. [c]Results presented in bold text indicate the best performance.

miTAR [44], TargetNet [16], deepTarget [10], TEC-miTarget [18], cnnMirTarget [15], miRBD [14], and TargetScanHuman [13].

The results presented in Table 7 highlight that the advantage of miCGR primarily manifests in more challenging imbalanced datasets. Specifically, on balanced test datasets, miCGR, under the offset-9-mer-m7 pairing pattern criteria, outperformed other state-of-the-art algorithms in terms of general classification performance measures, namely, F1 score and accuracy, although it has moderate levels of sensitivity and specificity. However, on imbalanced test datasets, miCGR not only maintained the highest F1 score and accuracy but also achieved significantly better sensitivity compared to the originally top-performing model miTAR, which exhibited a substantial decline.

On the one hand, the advantage of miCGR can be attributed to its lenient pairing pattern. Algorithms like TargetNet-all and miTAR simply utilize sliding windows when addressing gene-level prediction problems, potentially leading to excessive false positives. In contrast, methods that impose restrictive matching rules, such as miRBD6.0 and TargetScanHuman7.2, exhibited high specificity on both datasets but displayed low sensitivity, as they filtered out most non-canonical sites.

On the other hand, miCGR facilitates the identification of functional motifs within sequences, even when they are distant in the original sequence. This is evident when compared to miTAR. Although miTAR performs comparably with miCGR in the site-level (Table 5) and the balanced gene-level dataset, it experiences a significant decline in the imbalanced gene-level dataset. This could be attributed to the fact that although miTAR can capture continuous positional information of input sequences, this strategy fails when key nucleotides are far apart. Moreover, other deep learning methods such as cnnMirTarget and deepTarget2020, which lack positional information, are unable to discern nucleotides across different regions and consequently fail in this scenario.

Additionally, miCGR aligns more closely with biological principles, particularly the importance of pairings dominated by the seed region. This is evident when compared to TEC-miTarget. Although TEC-miTarget performed well on the miRAW-site-level dataset, it performed poorly on both the balanced and imbalanced gene-level datasets. This might be because TEC-miTarget adopted a CTS selection rule of 'at least nine WC pairings among the miRNA nucleotides 1–13', which diminishes the significance of pairings within the seed region of the miRNA.

To evaluate the predictive performance of miCGR on entirely novel miRNA-mRNA pairs—where neither the miRNAs nor the mRNAs were present in the training set—we excluded all mRNAs and miRNAs used in the training data when constructing both the balanced and imbalanced gene-level datasets. As shown in Supplementary Table S2 available online at http://bib.oxfordjournals.org/, miCGR maintained comparable performance on samples with unseen miRNAs and mRNAs, achieving accuracies of 0.7608 and 0.8839 for the balanced and imbalanced gene-level datasets, respectively. These results highlight miCGR's capability to accurately predict novel miRNA-mRNA interactions.

Learn from Interpretability on Site-level to Improve the Predictive Power on Gene-level

While traditional deep learning models often lack transparency in explaining how different CTS selection strategies impact predictive power [24, 25], understanding how miCGR distinguishes positive and negative binding pairs is crucial since miRNA-mRNA

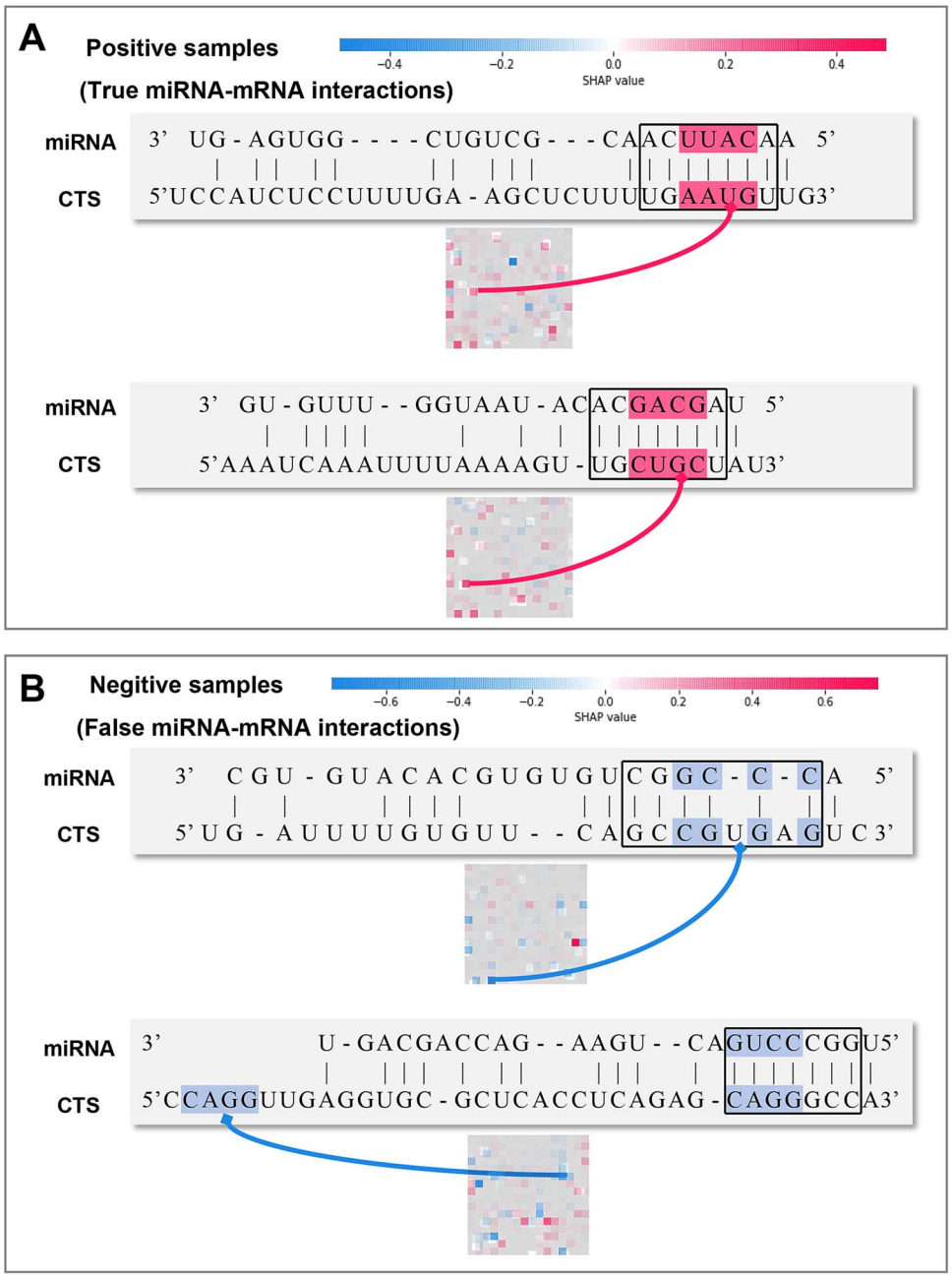

Figure 4. The SHAP analysis results for **(A)** two true miRNA-mRNA interaction pairs (positive samples) and **(B)** two false miRNA-mRNA interaction pairs (negative samples). Fragments colored pink indicate a positive contribution (Shapley value), while fragments colored blue indicate a negative contribution. The black boxes highlight the seed regions within the miRNA sequences.

interactions can extend beyond WC base pairing. To address this, we employed the SHAP method [45], a game-theoretic approach for interpreting machine learning predictions, to analyze feature contributions (Shapley values) to the final outcome. We applied this method to a well-performing 4-mer miCGR model for illustration, but it is applicable to models using different k-mer lengths.

To understand the model's classification of miRNA sequences as positive (binding) or negative (non-binding), we calculated Shapley values for the chaos game representations of two positive and two negative samples. Figure 4A displays two positive samples (true miRNA-mRNA interactions). The SHAP analysis reveals that continuous matching segments within the miRNA seed region significantly contribute to positive predictions. This finding aligns with previous reports indicating that animal miRNA and CTS recognition primarily occurs through complementary pairing in the seed region.

For the negative samples (false miRNA-mRNA interactions) in Fig. 4B, the SHAP analysis shows that discontinuous matching segments within the seed region have a lower positive prediction contribution (upper panel). Additionally, the presence of multiple possible matches with the miRNA's 3′ end sequence emerges as a negative indicator (lower panels).

The SHAP analysis suggests that CTS sequences containing multiple matching sites (m.s.) for the miRNA seed region might act as negative regulators due to competition between binding sites. To investigate this hypothesis, we filtered out CTSs with more than one matching site during evaluation on a balanced

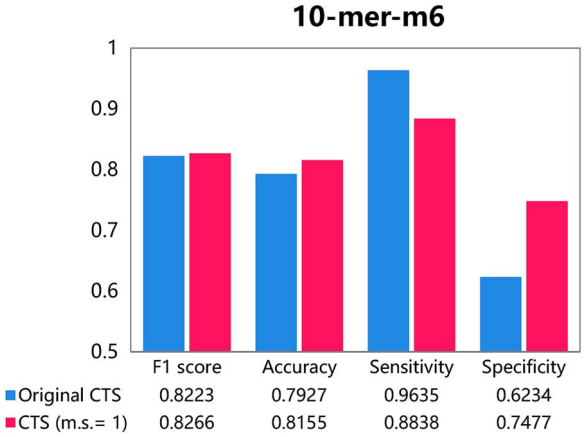

Figure 5. Comparison of miCGR performance between CTS (original) and CTS (m.s. = 1) on the balanced gene-level dataset. m.s. stands for matching sites.

gene-level dataset. As shown in Fig. 5, miCGR exhibited a slight improvement in accuracy, particularly under less stringent CTS selection criteria, such as 10-mer-m6. This improvement can be attributed to the reduced average CTS count of 10-mer-m6 from ∼ 81 to 15 (Table 6), which might help reduce the false positive rate. Consequently, the decrease in false positives indicates enhanced specificity for miCGR while causing relatively minor damage to sensitivity. Results for all CTS selection criteria are provided in Supplementary Table S3 available online at http://bib.oxfordjournals.org/.

Case studies in practical application

To evaluate the practical application of our model, we conducted two case studies: (i) predicting target genes for hsa-miR-552-3p, and (ii) predicting miRNAs that target PCSK9. The experimentally supported functional interactions between miRNAs and genes were retrieved from the latest version of miRTarBase [46] and TarBase [47] database.

Hsa-miR-552-3p is a small non-coding RNA located on chromosome 1p34.3, and its expression level is significantly dysregulated in tissues or cells of various tumors [48, 49]. For hsa-miR-552-3p, functionally targeted genes were predicted across 18,716 human genes, and all the prediction results can be found in our GitHub repository. As exhibited in Table 8, five experimentally functional target genes were ranked among the top 1% prediction results for hsa-miR-552-3p: GPR137C (rank 52nd), AR (rank 70th), DDX47 (rank 127th), MTDH (rank 129th), C12orf49 (rank 131st). For comparative analysis, we selected two well-established methods: TargetScanHuman7.2 [13] and miRDB 6.0 [50]. The top five true positive mRNA predictions for hsa-miR-552-3p from each model are presented in Table 8. MiCGR shows a higher enrichment of positive predictions, demonstrating that miCGR outperformed these commonly used methods. Notably, targets identified by these methods are also recognized by miCGR, but with better rankings; for instance, DDX47 is ranked in the top 0.68% by miCGR, compared to the top 15.09% by TargetScanHuman7.2. Additionally, the functional enrichment analysis depicted in Fig. 6A indicates that the target genes predicted by miCGR are predominantly involved in cancer-related pathways, corroborating findings from previous studies [49].

PCSK9, mainly produced by the liver, has a crucial role in the lifecycle of the LDL receptor [51]. The expression level of PCSK9 was associated with multiple cardiovascular diseases [52–54]. For PCSK9, functional miRNAs were predicted across 2656 human miRNAs, and all the prediction results can be found in our GitHub repository. As shown in Table 8, miCGR successfully predicted hsa-miR-877-5p among the top 10 prediction results. Additionally, hsa-miR-25-3p, which has been reported to have no interaction with PCSK9 [55], was ranked at the bottom (2043rd) in the prediction results. These findings demonstrate the strong practicality of our miCGR model. Similar to the prediction for hsa-miR-552-3p, miCGR demonstrates a greater enrichment of positive predictions for PCSK9 compared to TargetScanHuman7.2 and miRDB 6.0 (Table 8). Additionally, the multiple sequence alignment analysis presented in Fig. 6B indicates that the top 10 predicted miRNAs for PCSK9 share sequence similarities; however, these similarities extend beyond the seed region, as we employed lenient pairing pattern rules to select CTS.

Conclusion

Dysregulation of genes controlled by miRNAs and changes in miRNA expression are associated with various diseases, including cancer, cardiovascular, metabolic, and neurodegenerative

Table 8. Performance comparison of methods for case studies.

Methods	Top 5 true positive mRNA predictions for hsa-miR-552-3p (rank/total, %)	Top 5 true positive miRNA predictions for PCSK9 (rank/total, %)
miCGR	GPR137C (52/18716, 0.28%)	hsa-miR-877-5p (8/2656, 0.30%)
	AR (70/18716, 0.37%)	hsa-miR-320a-3p (11/2656, 0.41%)
	DDX47 (127/18716, 0.68%)	**hsa-miR-7845-5p (18/2656, 0.68%)**
	MTDH (129/18716, 0.69%)	hsa-let-7e-5p (31/2656, 1.17%)
	C12orf49 (131/18716, 0.70%)	hsa-miR-26b-5p (46/2656, 1.73%)
TargetScanHuman7.2	BTF3L4 (164/4791, 3.42%)	**hsa-miR-7845-5p (21/1892, 1.11%)**
	YBX1 (236/4791, 4.93%)	hsa-miR-128-3p (137/1892, 7.24%)
	RBN27 (398/4791, 8.31%)	hsa-miR-30c-1-3p (161/1892, 8.51%)
	DDX47 (723/4791, 15.09%)	hsa-miR-335-5p (182/1892, 9.62%)
	CBFA2T2 (764/4791, 15.95%)	hsa-miR-148b-3p (331/1892, 17.49%)
miRDB6.0[a]	BTF3L4 (94/513, 18.32%)	hsa-miR-191-5p (10/49, 20.41%)
	RBM27 (102/513, 19.88%)	hsa-miR-7845-5p (25/49,51.02%)
	MIER3 (184/513, 35.87%)	

[a]Fewer than five true positives were identified for miRDB6.0 [50], as it filtered the results to retain only positive predictions. [b]Results presented in bold text indicate the overlapping genes predicted by miCGR and the other methods.

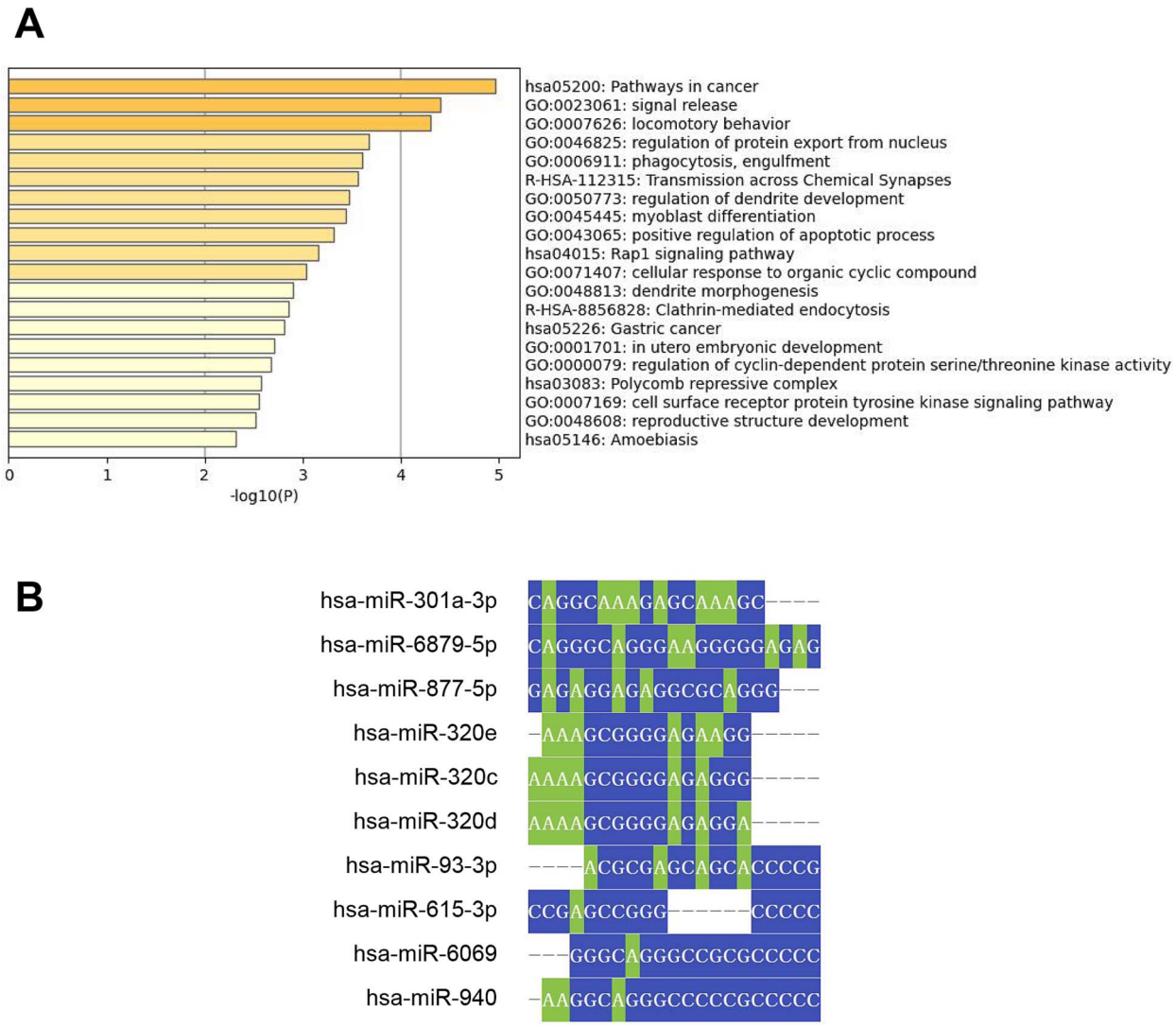

Figure 6. Prediction results for hsa-miR-552-3p and PCSK9. (A) Bar graph illustrating enriched terms from both gene ontology (GO) and the Kyoto Encyclopedia of genes and genomes (KEGG) for the top 100 predicted genes associated with hsa-miR-552-3p, color-coded by p-values. (B) Multiple sequence alignments of the top 10 predicted miRNAs for PCSK9, generated using the MUSCLE algorithm.

diseases. Therefore, reliably predicting potential miRNA targets related to these diseases is of particular importance. However, the interaction between miRNA and its targets is complex, often involving many putative miRNA recognition sites in mRNA.

In this study, we propose a deep learning model called miCGR, which is based on the enhanced Chaos Game Representation (CGR) algorithm. Unlike one-dimensional-based sequence characterization methods, the enhanced CGR helps to reveal some functional motifs of the sequences even when they are far away by converting genetic sequences into 2D graphical representations. By incorporating prior biological knowledge and position information through the enhanced CGR, miCGR can differently treat nucleotide fragments within and outside the seed region. Through training and hyperparameter tuning, our miCGR outperforms currently reported state-of-art algorithms in terms of accuracy and F1 scores on multiple validation datasets, including the site-level validation dataset, the balanced gene-level validation dataset, and the imbalanced gene-level validation dataset. Additionally, leveraging the SHAP interpretability analysis tool, we discover that the model can capture certain binding patterns between miRNA and CTS. Finally, two case studies demonstrate the practical applicability of miCGR, highlighting its potential to provide insights for optimizing artificial miRNA analogs that surpass endogenous counterparts.

Nevertheless, predicting gene-level targets for miRNA still faces many challenges. On one hand, the biological principles controlling the binding between miRNAs and target sites remain unclear, and the regulatory mechanisms involved are not well understood. On the other hand, mRNA transcripts, which serve as targets for miRNA, are often very long, with many genes that have numerous transcripts due to alternative splicing. Consequently, there are numerous potential binding sites, making accurate filtering difficult due to the lack of clear binding patterns. Additionally, the current lack of negative data restricts the model's ability to deeply learn the composition and binding patterns of CTS. With further research on miRNA and the accumulation of additional biological experimental data, it is anticipated that our miCGR model will show improved predictive and practical application capabilities in the future.

Key Points

- MiCGR uses an enhanced Chaos Game Representation (CGR) to convert genetic sequences into 2D graphical representations to predict functional miRNA targets. Unlike one-dimensional sequence characterization methods, miCGR's enhanced CGR retrieves information on similar motifs even when they are far apart in the original sequences, extracting information from nucleotide sequences more effectively.
- MiCGR further enhances the CGR algorithm by integrating seed region and subsequence position information into an additional image channel, emphasizing the seed region's importance for accurate prediction. This allows miCGR to treat nucleotide fragments within and outside the seed region differently, leading to better performance.
- Leveraging the SHAP interpretability analysis tool, miCGR can capture certain binding pattern rules between miRNA and CTS, improving its predictive power. The model outperforms existing methods in predicting functional targets at both site and gene levels and demonstrates practical applicability through case studies, offering insights for optimizing artificial miRNA analogs.

Acknowledgements

The authors would like to acknowledge their colleague, mentor and friend, Hualiang Jiang (1965–2022), who took part in the work and in the preparation of the original manuscript.

Author contributions

X.T.L., T.Y. conceived the project and were responsible for the decision to submit the manuscript. X.L.W. and L.H.Z. implemented the miCGR model and conducted computational analysis. X.C.T., Y.T.W, Z.M.Z., X.T.K., and S.K.N. accessed, verified, and analyzed the data. X.L.W., L.H.Z, X.T.L, Y.T., M.Y.Z. and X.M.L wrote the paper. All authors discussed the results and commented on the manuscript.

Supplementary Data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: The authors declare that they have no competing interests.

Funding

We gratefully acknowledge financial support from National Key Research and Development Program of China (2023YFC2305904 and 2022YFC3400504 to M.Y.Z.), the Strategic Priority Research Program of the Chinese Academy of sciences (XDB0830203 and XDB0830200), National Natural Science Foundation of China (82204278 to X.T.L, T2225002 to M.Y.Z., and 82273855 to M.Y.Z.), Shanghai Institute of Materia Medica Chinese Academy of Sciences (SIMM0220232001), SIMM-SHUTCM Traditional Chinese Medicine Innovation Joint Research Program (E2G805H), and Shanghai Municipal Science and Technology Major Project.

Data availability

The source codes and related data of miCGR are available at: https://github.com/myzhengSIMM/miCGR. The raw data are accessible at: https://doi.org/10.1371/journal.pcbi.1006185 and https://doi.org/10.1371/journal.pone.0232578.

References

1. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;**116**:281–97. https://doi.org/10.1016/S0092-8674(04)00045-5.
2. O'Brien J, Hayder H, Zayed Y. *et al.* Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Front Endocrinol* 2018;**9**:402. https://doi.org/10.3389/fendo.2018.00402.
3. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009;**136**:215–33. https://doi.org/10.1016/j.cell.2009.01.002.
4. Seok H, Ham J, Jang ES. *et al.* MicroRNA target recognition: insights from transcriptome-wide non-canonical interactions. *Mol Cells* 2016;**39**:375–81. https://doi.org/10.14348/molcells.2016.0013.
5. Kwak PB, Iwasaki S, Tomari Y. The microRNA pathway and cancer. *Cancer Sci* 2010;**101**:2309–15. https://doi.org/10.1111/j.1349-7006.2010.01683.x.
6. Tang C, Zhang Y. Detailed role of let-7e in human diseases. *Pathology-Research and Practice* 2024;**260**:155436. https://doi.org/10.1016/j.prp.2024.155436.
7. Eldakhakhny B, Sutaih AM, Siddiqui MA. *et al.* Exploring the role of noncoding RNAs in cancer diagnosis, prognosis, and precision medicine. *Non-coding RNA Research* 2024;**9**:1315–23. https://doi.org/10.1016/j.ncrna.2024.06.015.
8. Lu TX, Rothenberg ME. MicroRNA. *J Allergy Clin Immunol* 2018;**141**:1202–7. https://doi.org/10.1016/j.jaci.2017.08.034.
9. Albert P, Zhong X, Simon R. *et al.* miRAW: a deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts. *PLoS Comput Biol* 2018;**14**:e1006185. https://doi.org/10.1371/journal.pcbi.1006185.
10. Lee B. Deep learning-based microRNA target prediction using experimental negative data. *IEEE Access* 2020;**8**:197908–16. https://doi.org/10.1109/ACCESS.2020.3034681.
11. Kim D, Sung YM, Park J. *et al.* General rules for functional microRNA targeting. *Nat Genet* 2016;**48**:1517–26. https://doi.org/10.1038/ng.3694.
12. Peterson SM, Thompson JA, Ufkin ML. *et al.* Common features of microRNA target prediction tools. *Front Genet* 2014;**5**:23. https://doi.org/10.3389/fgene.2014.00023.
13. Agarwal V, Bell GW, Nam JW. *et al.* Predicting effective microRNA target sites in mammalian mRNAs. *eLife Sciences* 2015;**4**:e05005. https://doi.org/10.7554/eLife.05005.
14. Wang X. Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics* 2016;**32**:1316–22. https://doi.org/10.1093/bioinformatics/btw002.
15. Zheng X, Chen L, Li X. *et al.* Prediction of miRNA targets by learning from interaction sequences. *PloS One* 2020;**15**:e0232578. https://doi.org/10.1371/journal.pone.0232578.
16. Min S, Lee B, Yoon S. TargetNet: functional microRNA target prediction with deep neural networks. *Bioinformatics* 2022;**38**:671–7. https://doi.org/10.1093/bioinformatics/btab733.

17. Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In: *2017 international conference on engineering and technology (ICET)*, pp. 1–6. Antalya, Turkey: IEEE, 2017. https://doi.org/10.1109/ICEngTechnol.2017.8308186.
18. Yang T, Wang Y, He Y. TEC-miTarget: enhancing microRNA target prediction based on deep learning of ribonucleic acid sequences. *BMC bioinformatics* 2024;**25**:159. https://doi.org/10.1186/s12859-024-05780-z.
19. Hoang T, Yin C, Yau ST. Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. *Genomics* 2016;**108**:134–42. https://doi.org/10.1016/j.ygeno.2016.08.002.
20. Joseph J, Sasikumar R. Chaos game representation for comparison of whole genomes. *BMC bioinformatics* 2006;**7**:243. https://doi.org/10.1186/1471-2105-7-243.
21. Lochel HF, Heider D. Chaos game representation and its applications in bioinformatics. *Comput Struct Biotechnol J* 2021;**19**:6263–71. https://doi.org/10.1016/j.csbj.2021.11.008.
22. Löchel HF, Eger D, Sperlea T. *et al.* Deep learning on chaos game representation for proteins. *Bioinformatics* 2020;**36**:272–9. https://doi.org/10.1093/bioinformatics/btz493.
23. Thind AS, Sinha S. Using chaos-game-representation for analysing the SARS-CoV-2 lineages. *Newly Emerging Strains and Recombinants, Current genomics* 2023;**24**:187–95. https://doi.org/10.2174/0113892029264990231013112156.
24. Gunning D, Stefik M, Choi J. *et al.* XAI—Explainable artificial intelligence, science. *Robotics* 2019;**4**:eaay7120. https://doi.org/10.1126/scirobotics.aay7120.
25. Arrieta AB, Díaz-Rodríguez N, Ser JD. *et al.* Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 2020;**58**:82–115.
26. Vlachos IS, Paraskevopoulou MD, Karagkouni D. *et al.* DIANA-TarBase v7. 0: indexing more than half a million experimentally supported miRNA: mRNA interactions. *Nucleic Acids Res* 2015;**43**:D153–9. https://doi.org/10.1093/nar/gku1215.
27. Chou C-H, Chang N-W, Shrestha S. *et al.* miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res* 2016;**44**:D239–47. https://doi.org/10.1093/nar/gkv1258.
28. Grosswendt S, Filipchyk A, Manzano M. *et al.* Unambiguous identification of miRNA: target site interactions by different types of ligation reactions. *Mol Cell* 2014;**54**:1042–54. https://doi.org/10.1016/j.molcel.2014.03.049.
29. Helwak A, Kudla G, Dudnakova T. *et al.* Mapping the human miRNA Interactome by CLASH reveals frequent noncanonical binding. *Cell* 2013;**153**:654–65. https://doi.org/10.1016/j.cell.2013.03.043.
30. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 2019;**47**:D155–62. https://doi.org/10.1093/nar/gky1141.
31. Karagkouni D, Paraskevopoulou MD, Chatzopoulos S. *et al.* DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. *Nucleic Acids Res* 2018;**46**:D239–45. https://doi.org/10.1093/nar/gkx1141.
32. Hsu S-D, Lin F-M, Wu W-Y. *et al.* miRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucleic Acids Res* 2011;**39**:D163–9. https://doi.org/10.1093/nar/gkq1107.
33. Barnsley M, Hurd AJ. Fractals everywhere. *American Journal of Physics* 1989;**57**:1053–3. https://doi.org/10.1119/1.15823.
34. Altschul SF, Gish W, Miller W. *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
35. Li Z, Liu F, Yang W. *et al.* A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems* 2021;**33**:6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827.
36. Al-Qizwini M, Barjasteh I, Al-Qassab H. *et al.* Deep learning algorithm for autonomous driving using googlenet. *2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 89–96. Los Angeles, CA, USA, 2017. https://doi.org/10.1109/IVS.2017.7995703.
37. Zhou Y, Zhou B, Pache L. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets, nature. *Communications* 2019;**10**:10. https://doi.org/10.1038/s41467-019-09234-6.
38. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2010;**32**:1792–7. https://doi.org/10.1093/nar/gkh340.
39. Sussman JL, Lin D, Jiang J. *et al.* Protein data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* 1998;**54**:1078–84. https://doi.org/10.1107/S0907444998009378.
40. Lyon JA, Tennant MR, Danielson B. Introducing protein data bank, molecular modeling database, and Cn3D. *Journal of electronic resources in medical libraries* 2006;**3**:1–20. https://doi.org/10.1300/J383v03n03_01.
41. Loring JF, Wen X, Lee J. *et al.* A gene expression profile of Alzheimer's disease. *DNA Cell Biol* 2001;**20**:683–95. https://doi.org/10.1089/10445490152717541.
42. Domon B, Aebersold R, Aebersold R. Mass spectrometry and protein analysis. *Science* 2006;**312**:212–7. https://doi.org/10.1126/science.1124619.
43. Anighoro A, Bajorath J, Rastelli G. Polypharmacology: challenges and opportunities in drug discovery: Miniperspective. *J Med Chem* 2014;**57**:7874–87. https://doi.org/10.1021/jm5006463.
44. Gu T, Zhao X, Barbazuk WB. *et al.* miTAR: a hybrid deep learning-based approach for predicting miRNA targets. *BMC Bioinformatics* 2021;**22**:96. https://doi.org/10.1186/s12859-021-04026-6.
45. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 2017;**30**:4768–77.
46. Huang H-Y, Lin Y-C-D, Cui S. *et al.* miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions. *Nucleic Acids Res* 2022;**50**:D222–30. https://doi.org/10.1093/nar/gkab1079.
47. Skoufos G, Kakoulidis P, Tastsoglou S. *et al.* TarBase-v9. 0 extends experimentally supported miRNA–gene interactions to cell-types and virally encoded miRNAs. *Nucleic Acids Res* 2024;**52**:D304–10. https://doi.org/10.1093/nar/gkad1071.
48. Zhu L, Zhang S, Chen S. *et al.* Exosomal miR-552-5p promotes tumorigenesis and disease progression via the PTEN/-TOB1 axis in gastric cancer. *J Cancer* 2022;**13**:890–905. https://doi.org/10.7150/jca.66903.
49. Zou Y, Zhao X, Li Y. *et al.* miR-552: an important post-transcriptional regulator that affects human cancer. *J Cancer* 2020;**11**:6226–33. https://doi.org/10.7150/jca.46613.
50. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res* 2020;**48**:D127–31. https://doi.org/10.1093/nar/gkz757.
51. Barale C, Melchionda E, Morotti A. *et al.* PCSK9 biology and its role in atherothrombosis. *Int J Mol Sci* 2021;**22**:5880. https://doi.org/10.3390/ijms22115880.
52. Seidah NG, Prat A. The multifaceted biology of PCSK9. *Endocr Rev* 2022;**43**:558–82. https://doi.org/10.1210/endrev/bnab035.

53. Han L, Wu L, Yin Q. *et al.* A promising therapy for fatty liver disease: PCSK9 inhibitors. *Phytomedicine* 2024;**128**:155505. https://doi.org/10.1016/j.phymed.2024.155505.
54. Senftleber NK, Andersen MK, Jørsboe E. *et al.* GWAS of lipids in Greenlanders finds association signals shared with Europeans and reveals an independent PCSK9 association signal. *Eur J Hum Genet* 2024;**32**:215–23. https://doi.org/10.1038/s41431-023-01485-8.
55. Decourt C, Janin A, Moindrot M. *et al.* PCSK9 post-transcriptional regulation: role of a 3′UTR microRNA-binding site variant in linkage disequilibrium with c.1420G. *Atherosclerosis* 2020;**314**:63–70. https://doi.org/10.1016/j.atherosclerosis.2020.10.010.

Briefings in Bioinformatics, 2025, **26(1)**, bbae616
https://doi.org/10.1093/bib/bbae616
Problem Solving Protocol