Molecular Therapy
Nucleic Acids
Original Article

SMTRI: A deep learning-based web service for predicting small molecules that target miRNA-mRNA interactions

Huan Xiao,[1,5] Yihao Zhang,[1,5] Xin Yang,[2,3,4,5] Sifan Yu,[1] Ziqi Chen,[2,3,4] Aiping Lu,[2,3,4] Zongkang Zhang,[1] Ge Zhang,[2,3,4] and Bao-Ting Zhang[1]

[1]School of Chinese Medicine, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR 999077, China; [2]Law Sau Fai Institute for Advancing Translational Medicine in Bone & Joint Diseases, Hong Kong Baptist University, Hong Kong SAR 999077, China; [3]Institute of Integrated Bioinformedicine and Translational Science, Hong Kong Baptist University, Hong Kong SAR 999077, China; [4]Institute of Precision Medicine and Innovative Drug Discovery, Hong Kong Baptist University, Hong Kong SAR 999077, China

Mature microRNAs (miRNAs) are short, single-stranded RNAs that bind to target mRNAs and induce translational repression and gene silencing. Many miRNAs discovered in animals have been implicated in diseases and have recently been pursued as therapeutic targets. However, conventional pharmacological screening for candidate small-molecule drugs can be time-consuming and labor-intensive. Therefore, developing a computational program to assist mature miRNA-targeted drug discovery *in silico* is desirable. Our previous work (https://doi.org/10.1002/advs.201903451) revealed that the unique functional loops formed during Argonaute-mediated miRNA-mRNA interactions have stable structural characteristics and may serve as potential targets for small-molecule drug discovery. Developing drugs specifically targeting disease-related mature miRNAs and their target mRNAs would avoid affecting unrelated ones. Here, we present SMTRI, a convolutional neural network-based approach for efficiently predicting small molecules that target RNA secondary structural motifs formed by interactions between miRNAs and their target mRNAs. Measured on three additional testing sets, SMTRI outperformed state-of-the-art algorithms by 12.9%–30.3% in AUC and 2.0%–18.4% in accuracy. Moreover, four case studies on the published experimentally validated RNA-targeted small molecules also revealed the reliability of SMTRI.

INTRODUCTION

Recent studies about the structural and functional information of RNAs[1–3] have put them in the spotlight to replace the long-dominating proteins as promising therapeutic targets.[4] MicroRNA (miRNA) is now a well-validated target of all RNA classes.[5–7] Mature miRNA is a type of short (∼22 nt) single-stranded non-coding RNA molecule that integrates with Argonaute protein to form miRNA-induced silencing complex (miRISC), which then negatively regulates post-transcriptional gene expression by either degrading mRNA or inhibiting mRNA translation.[8] Diseases caused by miRNA-induced mRNA silencing can be rescued by screening drug-like molecules specifically targeting those miRNAs to alter their abundance and adjust downstream translation efficiency accordingly.[9]

Various approaches have been applied to screen small molecules (SMs) targeting miRNAs. Wet lab-based methods include SM microarray screening, high-throughput screening, and fragment-based screening.[10,11] Computational methods, like artificial intelligence (AI)-assisted tools, have also recently been developed to promote miRNA drug discovery by predicting SM-miRNA associations *in silico*. For example, Jamal et al.[12] were the pioneers in employing machine learning (ML) models (naive Bayes and random forest) to mine miR-21 inhibitors from large SM datasets. Both classifiers were trained with SMs represented in two-dimensional (2D) molecular descriptors and produced a prediction accuracy of nearly 0.80. Wang et al.[13] also used the random forest algorithm to predict SM-miRNA association. Their model RFSMMA utilized the similarities of SMs and miRNAs as feature vectors to represent SM-miRNA pairs, which were then fed into a random forest classifier to train a predictive model. Zhao et al.[14] presented a symmetric non-negative matrix factorization model for SM-miRNA association prediction (SNMFSMMA). First, they applied symmetric non-negative matrix factorization to perform matrix decomposition on the integrated similarity matrixes of SMs and miRNAs, respectively. Second, they calculated the Kronecker product of the newly integrated similarity matrixes from the previous step and obtained the SM-miRNA pair

Received 12 September 2023; accepted 12 August 2024;
https://doi.org/10.1016/j.omtn.2024.102303.

[5]These authors contributed equally

Correspondence: Zongkang Zhang, School of Chinese Medicine, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR 999077, China.
E-mail: maxzhangzk@cuhk.edu.hk
Correspondence: Ge Zhang, Law Sau Fai Institute for Advancing Translational Medicine in Bone & Joint Diseases, Hong Kong Baptist University, Hong Kong SAR 999077, China.
E-mail: zhangge@hkbu.edu.hk
Correspondence: Bao-Ting Zhang, School of Chinese Medicine, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR 999077, China.
E-mail: zhangbaoting@cuhk.edu.hk

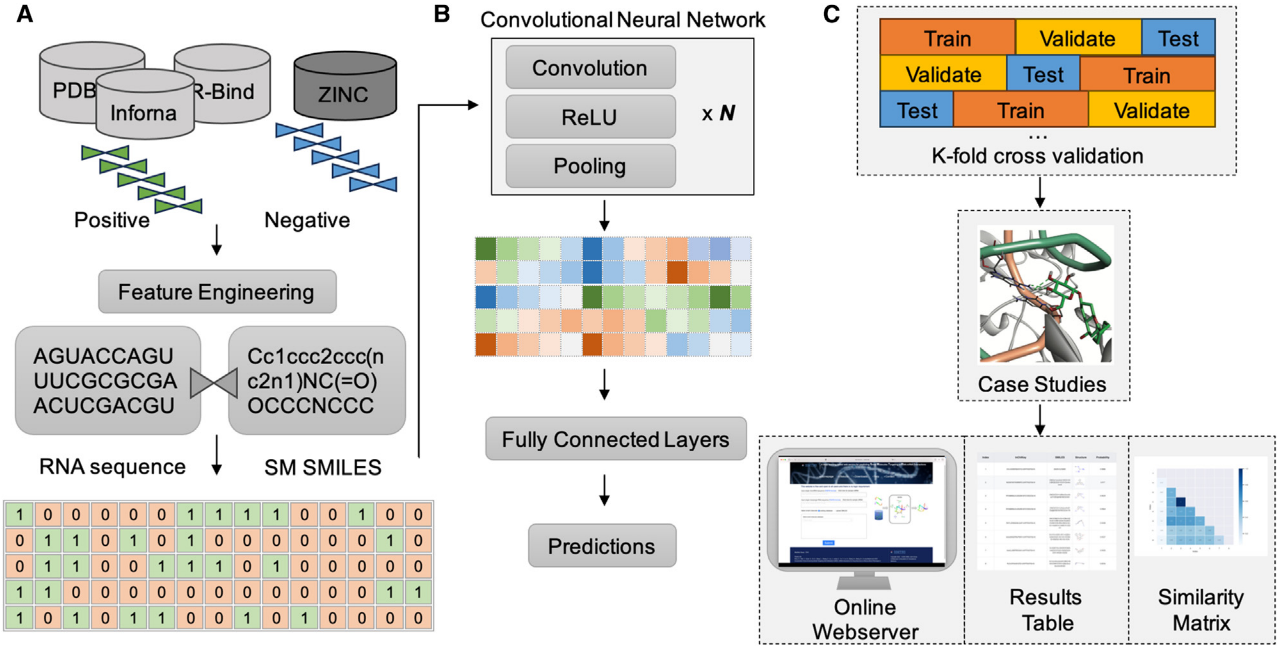


Figure 1. Overview of SMTRI workflow

(A) Data collection from public databases and feature engineering of RNA motif-SM associations. (B) Deep learning model construction. (C) Training procedures, case studies, and launch of online web server.

similarity. Finally, they implemented regularized least squares to map the SM-miRNA pairs to the association probabilities. More recently, some other studies attempted to use deep learning models for this task. Shen et al.[15] proposed a joint learning computational framework (SMAJL) built on the restricted Boltzmann machine to predict SM-miRNA associations. SMAJL used enhancing matrix completion to obtain vector features from the network representations of SM clinical and miRNA functional similarities. Next, it extracted SM chemical structural features and miRNA secondary structural features from RDKit and Pse-in-One. All acquired features were fed into the joint learning model to make predictions. Oliver et al.[16] developed a tool, RNAmigos, which took graph representations of RNA structures to predict binding SMs. A relational graph convolutional network was used as the core model to operate on the RNA representation encoded by an augmented base pairing network and compute node embeddings. The embeddings underwent a pooling process and a multilayer perceptron to generate a molecular fingerprint for a potential SM. The fingerprint was then used to search a library of SMs for active binders.

Disney et al. have proposed a miRNA biogenesis-dependent method to develop SM drugs that directly degrade the miRNAs.[17] However, few studies have proposed strategies targeting disease-specific miRNA-mRNA interactions. We tried to develop a miRNA biogenesis-independent method that targets the downstream after miRNA biogenesis (i.e., miRNA-mRNA interactions). Previously, our work revealed that mature miRNAs and their target mRNAs exhibit structural uniqueness and high-stability loops (e.g., bulge loop, internal loop), which could be postulated as therapeutic targets for drug discovery in translation repression-related diseases.[9] In this paper, building upon the previous foundation, we developed biogenesis-independent algorithms incorporating a convolutional neural network (CNN) based model for predicting SMs that directly target the RNA structural motifs formed by miRNA-mRNA interactions. We called the program SMTRI (Small Molecules Targeting miRNA-mRNA Interactions) (Figure 1). Unlike most other AI methods that nonselectively encoded the whole RNA sequence to predict binding SMs,[12–15] SMTRI selectively used the pocket region of an RNA sequence—RNA motifs—eliminating much useless information. The program extracted RNA motifs in letter-bracket notations, which retain the nucleic acid compositions and order information in the targets. The dataset of experimentally verified RNA motif-SM associations used in this model was downloaded from RNALigands,[18] PDB,[19] PubChem,[20] and RPocket[21] databases. In addition, the program implemented both deep learning and statistical methods in feature engineering to fully represent the RNA motifs and SMs in numerical forms.

To evaluate the performance of SMTRI, we compared our method with state-of-the-art (SOTA) algorithms as well as tested its ability in case studies. The results showed that SMTRI surpassed the other methods on independent testing sets for predicting true targeting SMs. To promote and facilitate the usage of SMTRI, a user-friendly web server was launched at http://www.smtri.net/. Our program narrows down the range of SMs that require experimental verification, which plays a guiding role in future research and experiments.

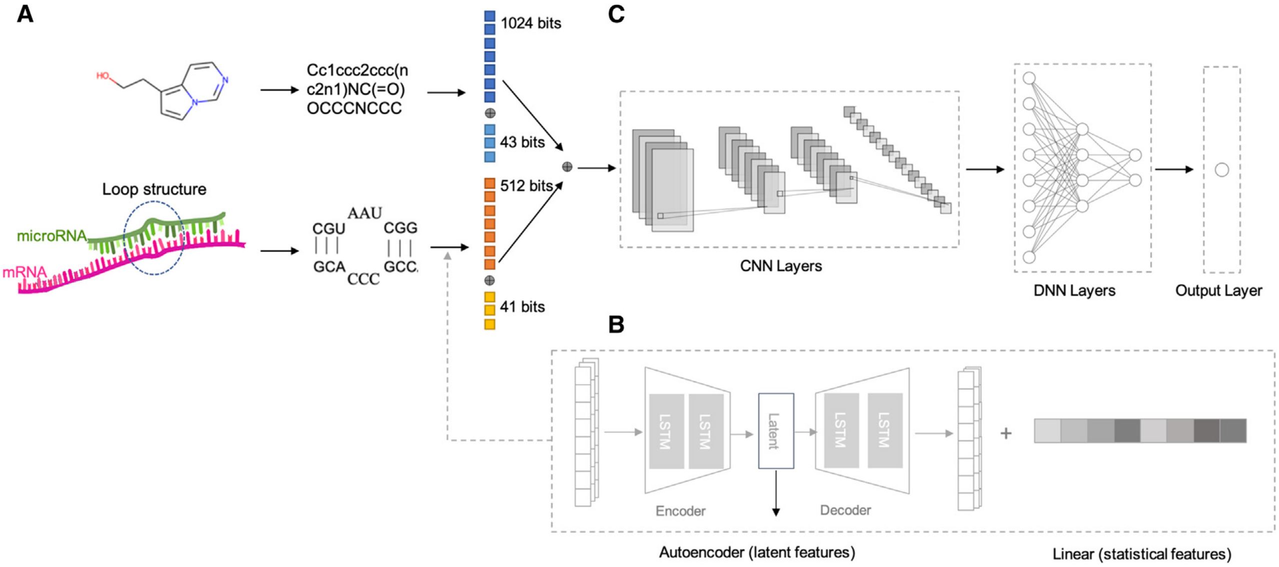


Figure 2. The detailed deep learning framework implemented in SMTRI

(A) The schematic view of RNA motif-SM inputs. (B) The schematic view of feature encoding methods for RNA motifs. (C) The schematic view of the CNN-based model.

RESULTS

In this study, we obtained RNA motif-SM association data from public databases and curated valid ones to construct positive and negative samples (described in materials and methods). Then, we transformed the data into numerical forms by extracting sequential and structural features from RNA motifs and SMs, respectively, and concatenating them linearly. In this process, a long short-term memory-based stacked autoencoder (LSTM-SAE) was employed to learn latent high-level features from the sequences of RNA motifs, while a statistical approach was established to calculate features from their nucleic acid compositions. The molecular fingerprints and descriptors were used to represent SMs in feature vectors. Next, we trained a CNN-based model to predict the binding probability scores of RNA motif-SM combinations (Figure 2). The reason for choosing CNNs as the core model rather than recurrent neural networks, LSTMs, gated recurrent units, and other complex models is that the latter performed much worse on our feature data or were too complicated for our data volume. We evaluated the model's predictive ability on three additional testing datasets and compared its performance with three SOTA algorithms to demonstrate the effectiveness of our approach. Furthermore, we conducted four case studies with specific examples that proved the robustness of SMTRI. Finally, we launched our program on the web server incorporating the above CNN-based model and two SM databases.

RNA motif-SM associations

When disease-causing miRNA interacts with its target mRNA, the conserved seed region (mostly situated at positions 2–7 from the miRNA 5′ end) of miRNA has completely complementary base pairing to the 3′ UTR of target mRNA. The rest of the miRNA sequence generally has partially complementary base pairing to mRNA, leaving unpaired bases to form cleft-like motifs that SMs can bind with (Figure S1A). Identifying those ligand-binding pockets in RNA structures is critical for RNA-targeted SM drug discovery. We utilized RNA22[22] to generate possible secondary structural interactions between miRNA and its target mRNA in dot-bracket notations. We then extracted the motifs in letter-bracket notations from the interactions (Figure S2). Platforms like RNALigands[18] provided available RNA motif-SM associations or structures of RNA-SM associations. We compiled RNA secondary structural motifs and their associated targeting SMs from these databases for deep learning. Figure S1B illustrates the proportion of different loops in all curated RNA motifs. Internal loops (59.09%) and bulge loops (32.14%) accounted for a significant proportion because they are most likely to be formed from the interactions between short mature miRNAs (∼22 nt) and their target mRNAs. Other loops (8.78%) included hairpin loops, exterior loops, and multi-branch loops (Figure S1C). The associated SMs have relatively concentrated properties (Figures S1D–S1G—for example, the molecular mass is between 450 and 650, and the number of atoms is between 60 and 90. These molecular properties assisted in predicting potential targeting SMs through deep learning.

Performance evaluation

To evaluate the prediction performance of SMTRI, an independent testing set was held out for assessing the final model. We further compared SMTRI with three SOTA algorithms—XGBoost, NB,[12] and RFSMMA[13]—on three additional testing sets collected from PDB,[19] PubChem (AID: 2899),[20] and RPocket,[21] respectively. The evaluation metrics involved area under the receiver operating characteristic curve (AUC), accuracy, precision, recall, F1 score, Matthew's correlation coefficient (MCC), and kappa (Equations S1–S6).

Table 1. The performance of SMTRI on an independent testing set

	AUC	Accuracy	Precision	Recall	F1 score	MCC	Kappa
SMTRI	0.978	0.963	0.894	0.927	0.910	0.887	0.887

A 5-fold cross-validation was implemented in training procedures to fine-tune the model parameters. We randomly divided the known RNA motif-SM associations into five equal parts; one part was selected as the validating set in turn, and the remaining four parts were regarded as the training samples. The corresponding receiver operating characteristic curves in validation results were depicted five times, and the average AUC values (0.963 ± 0.009) were taken to evaluate the model (Figure S3). After that, we re-trained the model with optimized parameters using all available data and tested it on an independent testing set (82 positive samples, 324 negative samples). The SMTRI model achieved good results (Table 1; Figure S4) in AUC (0.978), accuracy (0.963), precision (0.894), recall (0.927), F1 score (0.910), MCC (0.887), and kappa (0.887).

We next compared the performance of SMTRI with XGBoost, NB, and RFSMMA in the task of predicting RNA motif-SM associations. These benchmark models were tuned using the training set by searching the hyper-parameter spaces. With the optimal parameters, the benchmarks and SMTRI were tested on three additional testing sets from PDB, PubChem (AID: 2899), and RPocket, with 5-fold cross-validation. Evaluation metrics with means and SDs were obtained to assess their average performances on each testing set. The methodologies, including detailed parameter configurations for these methods, were thoroughly documented in the supplemental information. The comparative results were encapsulated in Tables 2, 3, and 4.

PDB testing set

The PDB set is a medium-sized testing set with 125 positive and 345 negative samples. SMTRI performed best in all metrics on this dataset, leading to an increase in AUC by 18.8%–23.0%, an increase in accuracy by 12.3%–13.4%, an increase in precision by 26.2%–27.0%, an increase in recall by 11.4%–22.2%, an increase in F1 score by 18.3%–23.7%, an increase in MCC by 27.7%–33.5%, and an increase in kappa by 27.3%–33.0% (Table 2).

PubChem testing set

The PubChem set (AID: 2899) is a large testing set with 882 positive and 2,600 negative samples. SMTRI ranked at the top in all metrics, leading to an increase in AUC by 24.9%–30.3%, an increase in accuracy by 7.5%–18.4%, an increase in precision by 12.5%–39.5%, an increase in recall by 11.2%–39.4%, an increase in F1 score by 25.7%–33.0%, an increase in MCC by 30.3%–46.1%, and an increase in kappa by 31.4%–44.5% (Table 3).

RPocket testing set

The RPocket set is the smallest testing set, with only 28 positive and 72 negative samples. SMTRI dominated all metrics on this dataset, leading to an increase in AUC by 12.9%–24.2%, an increase in accuracy by 2.0%–11.0%, an increase in precision by 6.8%–17.4%, an increase in recall by 6.0%–28.0%, an increase in F1 score by 5.2%–23.4%, an increase in MCC by 8.9%–29.6%, and an increase in kappa by 7.0%–28.6% (Table 4).

In general, SMTRI excels in predicting true positives across all three testing sets. The true positive rate (recall) of SMTRI is significantly higher than other methods, which is an important indicator for downstream drug discovery. The results demonstrated that SMTRI has visible advantages compared with other models and is superior in predicting RNA motif-SM associations.

Case studies

To further evaluate the discrimination ability of SMTRI, we carried out four case studies based on the data of RNA-SM interactions experimentally verified in the publications (Table S1). Since existing studies of SMs that target miRNA-mRNA interactions were limited and most of them have already been used for our model training, two of the four cases were not conducted directly on mature miRNAs. However, the secondary structural motifs in the RNA sequences reported in these studies can still be used for testing purposes. The first case was our previous research[9] of the verified 7-hydroxyflavone-β-D-glucoside (OC-3), which targets the interaction between mature hsa-miR-214-3p and TRAF3 (Figure 3A). The second case was a synthesized molecule 6′-fluorosisomicin targeting protozoal cytoplasmic rRNA A-site (PDB: 5Z1I).[23] The palindromic RNA duplex comprising two identical protozoal cytoplasmic rRNA sequences provides binding pockets for 6′-fluorosisomicin (Figure 3B).

Table 2. The performance of SMTRI compared with other models on a PDB testing set

PDB	AUC	Accuracy	Precision	Recall	F1 score	MCC	Kappa
SMTRI	0.947 ± 0.087*	0.923 ± 0.072*	0.875 ± 0.173*	0.787 ± 0.226*	0.823 ± 0.196*	0.781 ± 0.236*	0.775 ± 0.238*
XGBoost	0.759 ± 0.041	0.800 ± 0.032	0.613 ± 0.055	0.673 ± 0.071	0.640 ± 0.056	0.504 ± 0.076	0.502 ± 0.076
NB	0.742 ± 0.052	0.794 ± 0.037	0.605 ± 0.062	0.635 ± 0.107	0.617 ± 0.076	0.478 ± 0.092	0.475 ± 0.093
RFSMMA	0.717 ± 0.027	0.789 ± 0.020	0.610 ± 0.054	0.565 ± 0.050	0.586 ± 0.048	0.446 ± 0.055	0.445 ± 0.054

The highest score in each column is indicated with an asterisk.

Table 3. The performance of SMTRI compared with other models on a PubChem testing set

PubChem	AUC	Accuracy	Precision	Recall	F1 score	MCC	Kappa
SMTRI	0.935 ± 0.079*	0.878 ± 0.113*	0.828 ± 0.207*	0.782 ± 0.142*	0.783 ± 0.153*	0.719 ± 0.212*	0.703 ± 0.227*
XGBoost	0.632 ± 0.014	0.712 ± 0.008	0.436 ± 0.038	0.472 ± 0.026	0.453 ± 0.028	0.258 ± 0.030	0.258 ± 0.030
NB	0.686 ± 0.008	0.694 ± 0.011	0.433 ± 0.026	0.670 ± 0.020	0.526 ± 0.020	0.332 ± 0.018	0.315 ± 0.020
RFSMMA	0.666 ± 0.012	0.803 ± 0.010	0.703 ± 0.026	0.388 ± 0.029	0.498 ± 0.022	0.416 ± 0.019	0.389 ± 0.023

The highest score in each column is indicated with an asterisk.

The third case was a designed molecule, Targapremir-210, which targets the secondary structural motifs in the single-stranded has-mir-210 precursor (Figure 3C).[24] The fourth case was a study of SM inhibitor Isis-11 ((7R)-7-[(dimethylamino)methyl]-1-[3-(dimethylamino) propyl]-7,8-dihydro-1H-furo[3,2-e]benzimidazol-2-amine) that targets the RNA motif in the hepatitis C virus (HCV) internal ribosomal entry site (IRES) RNA (PDB: 2KTZ) (Figure 3D).[25]

We randomly collected 46 irrelevant SMs from the Zinc[26] and Natural Product databases,[27] together with the above four candidate SMs, to form a list of 50 for testing. The RNA motifs were first reproduced from the involved RNA sequences. SMTRI then made predictions on each RNA motif-SM pair. We set the threshold of binding probability at 0.85, which meant the SM was deemed to target the motif if the predicted score was above 0.85. Figure 3 (lower row) also depicts the score distributions of 50 candidate SMs in 4 case studies. We can infer from the pictures that SMTRI can separate the true targeting ones from the vast majority of SMs. The detailed results are reported below.

Case 1: OC-3, targeting the interaction between hsa-miR-214-3p and TRAF3

Table S2 lists three candidate SMs predicted to have a >85% chance of targeting hsa-miR-214-3p-TRAF3. The other 47 SMs, not listed in the table, had a <85% chance of targeting this interaction. The red-marked SM 11 (OC-3) was predicted to have a 99.64% chance of targeting hsa-miR-214-3p-TRAF3, the highest among the other possible candidate SMs (Figure 3E).

Case 2: 6′-fluorosisomicin, targeting the protozoal cytoplasmic rRNA A-site

Table S3 lists seven candidate SMs predicted to have a >85% chance of targeting protozoal cytoplasmic rRNA A-site. SM 5 (6′-fluorosisomicin), colored red, exhibited an 85.70% chance of targeting the rRNA (Figure 3F).

Case 3: Targapremir-210, targeting the has-mir-210 precursor

Table S4 lists three candidate SMs predicted to have a >85% chance of targeting the has-mir-210 precursor. All the other candidates were very unlikely (<10% chance) to hit the target. Targapremir-210 was the red-marked SM 8, with a dominant targeting probability of 99.73%. Thus, as expected, the real targeting SM stood out from the prediction result (Figure 3G).

Case 4: Isis-11, targeting the HCV IRES RNA

Table S5 lists ten candidate SMs predicted to have a >85% chance of targeting the HCV IRES RNA. SM 10 (Isis-11) emerged as the best candidate molecule for targeting the RNA motif, with a probability score of 99.95% (Figure 3H).

These case studies demonstrated that SMTRI could identify high-affinity SMs that target RNA motifs formed from miRNA-mRNA interactions with relatively high accuracy.

Online web server

SMTRI has been implemented as a user-friendly and freely available web server. It provides an easy-to-use interface to predict potential high-affinity SMs that target miRNA-mRNA interactions. The data flow of SMTRI is shown in Figure S5. On the homepage, users are required to provide a pair of miRNA and mRNA in FASTA format and specify the candidate SMs by either selecting a database provided by us (Drug Bank Database or Natural Product Database)[27,28] or uploading a list of SM simplified molecular-input line-entry system (SMILES) (Figure 4A). By default, SMTRI will run RNA22[22] on miRNA and mRNA sequences to identify their binding modes, from which SMTRI extracts the RNA motifs in string format.

Table 4. The performance of SMTRI compared with other models on a RPocket testing set

RPocket	AUC	Accuracy	Precision	Recall	F1 score	MCC	Kappa
SMTRI	0.950 ± 0.100*	0.870 ± 0.112*	0.749 ± 0.326*	0.840 ± 0.206*	0.763 ± 0.263*	0.704 ± 0.300*	0.677 ± 0.314*
XGBoost	0.820 ± 0.109	0.840 ± 0.080	0.665 ± 0.304	0.780 ± 0.177	0.685 ± 0.221	0.609 ± 0.246	0.583 ± 0.249
NB	0.821 ± 0.135	0.850 ± 0.114	0.681 ± 0.171	0.767 ± 0.200	0.711 ± 0.167	0.615 ± 0.243	0.607 ± 0.243
RFSMMA	0.708 ± 0.133	0.760 ± 0.116	0.575 ± 0.238	0.560 ± 0.233	0.529 ± 0.221	0.408 ± 0.257	0.391 ± 0.267

The highest score in each column is indicated with an asterisk.

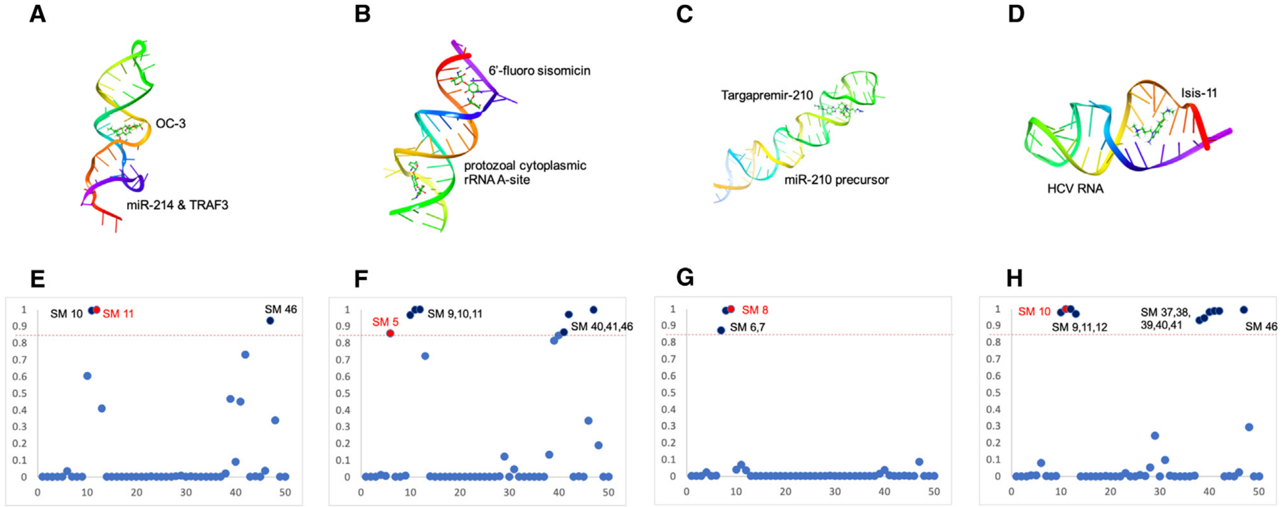


Figure 3. The 3D structures of RNA motif-SM interactions in 4 case studies, and the results of screening targeting SMs from 50 candidates

The x axes in the scatterplots (E–H) indicate the serial numbers of the 50 candidate SMs, and the y axes indicate the predicted binding scores (range from 0 to 1) of the SMs. The red dashed lines at the scores of 0.85 in (E)–(H) represent the partition thresholds. Therefore, an SM is considered to have a high affinity for an RNA motif if its predicted probability is >0.85. (A) Case 1: OC-3 (SM 11 in E, scores 0.9964), targeting the interaction between miR-214 and TRAF3. (B) Case 2: 6′-fluorosisomicin (SM 5 in F, scores 0.8570), targeting the protozoal cytoplasmic rRNA A-site. (C) Case 3: Targapremir-210 (SM 8 in G, scores 0.9973), targeting the miRNA-210 precursor. (D) Case 4: Isis-11 (SM 10 in H, scores 0.9995), targeting the HCV RNA.

Sequence-based RNA motifs and SMILES-based SMs are further transformed into their numerical features, which, after combination, will be input into the trained classifier to produce a probability score. Finally, SMTRI will summarize the top 20 predicted SMs with a >85% likelihood to bind to the miRNA-mRNA interactions. These potential targeting SMs will be displayed in the results table (Figure 4B), with affinity levels (probabilities) presented in descending order, in which their InChIKey, canonical SMILES, 2D structure images, and predicted probability scores are listed for reference. In addition, a similarity matrix for pairwise comparison of the SM structures will also be given based on their fingerprints (Figure 4C). As structurally similar molecules tend to have similar biological activities or functions, the similarity matrix provides another perspective to cluster suitable SMs for experimental validation in addition to the binding probabilities.

DISCUSSION

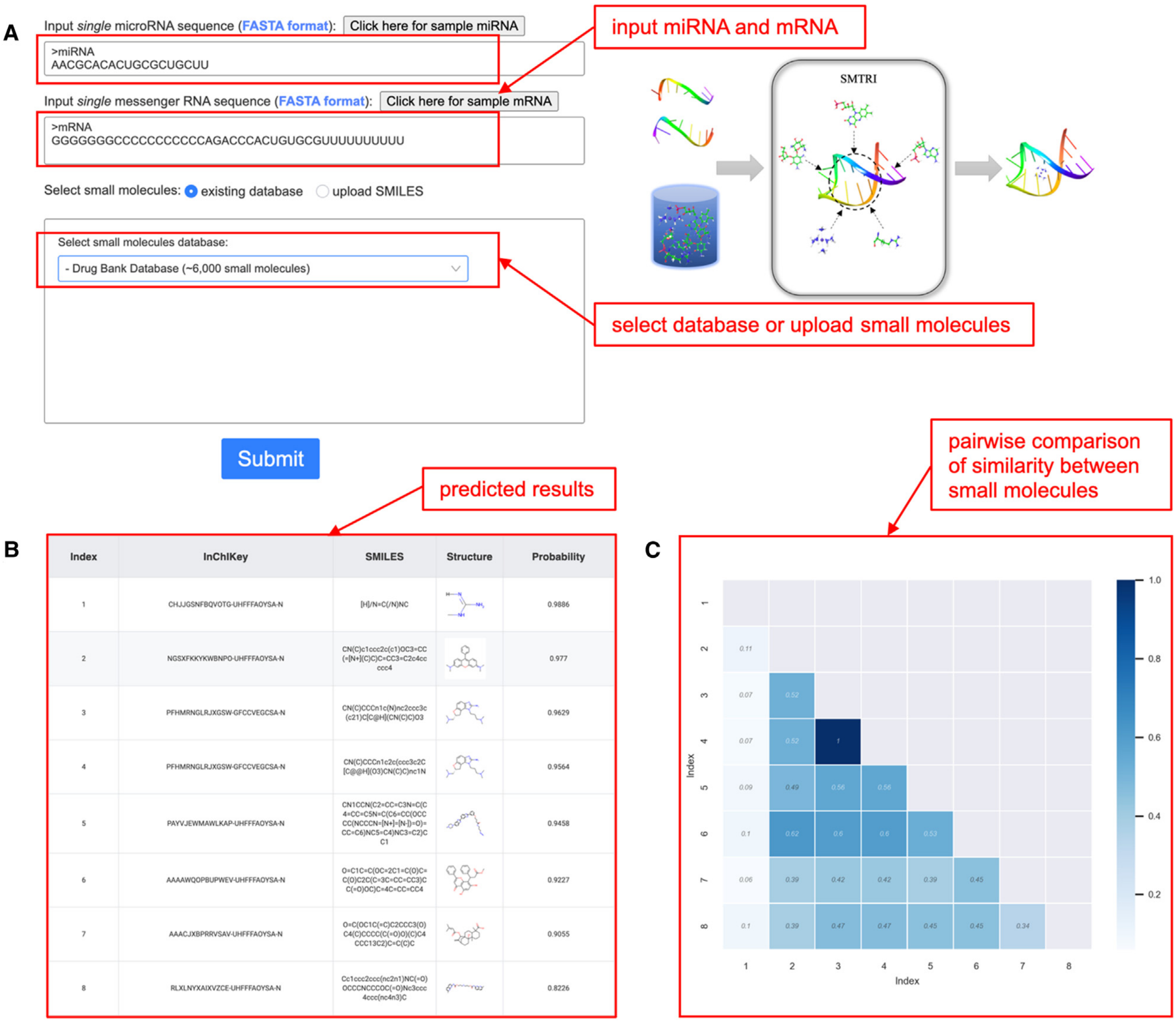
RNA-targeted SM drug discovery has emerged as an alternative solution to disease treatment in addition to conventional protein-targeted SM drug strategies and RNA-targeted oligonucleotide therapeutics. Among the various RNA classes, miRNA is the most promising target for developing SM drugs.

So far, therapeutic antagomiRs have not been clinically approved for marketing by the US Food and Drug Administration in the past ten years, indicating the concern in druggability.[29–31] There is a long way to clinical translation for therapeutic antagomiRs. Regarding the research and development approach, SMs rather than antagomiRs are mature drug forms with no critical concerns in druggability.[32–34] This is why we focus on the approach of SMs.

The target complex miRISC, formed by mature miRNA and Argonaute protein, is considered transient in some studies.[35–38] In our previously published work,[9] we identified two SMs that could effectively target the miRISC complex. Our *in vitro* and *in vivo* data have demonstrated the efficacy of these two molecules. One SM was found to target miR-214-ATF4 mRNA to promote ATF4 protein expression and enhance osteogenic potential. Another SM was found to target miR-214-TRAF3 mRNA to promote TRAF3 protein expression and inhibit osteoclast activity. Our findings indicate that targeting the interactions within the miRISC complex is feasible, even though they could be transient.

The current miRNA-targeted SM drug discovery is limited, however, by the high time and labor costs of the traditional experimental method. To address this problem, we presented SMTRI, a computational tool for predicting lead compounds targeting miRNA-mRNA interactions. SMTRI applied deep learning techniques in both feature representation and result prediction. The comparison between SOTA algorithms and the validation in case studies showed that SMTRI could effectively predict high-affinity SMs for the RNA motifs. Finally, SMTRI was integrated into the web service to offer convenience for other researchers.

The most distinct factor that contributed to the achievement of SMTRI was the use of RNA motif segments instead of whole RNA sequences as the targets. Most similar studies would encode the full-length RNAs without preferences, which not only wasted computing resources but also retained a large amount of redundant and useless information for ML. A few studies, like RNAmigos,[16] utilize the three-dimensional (3D) structural data of RNAs to achieve better prediction but require complex structural inputs and longer running



Input *single* microRNA sequence (FASTA format): Click here for sample miRNA

>miRNA
AACGCACACUGCGCUGCUU

Input *single* messenger RNA sequence (FASTA format): Click here for sample mRNA

>mRNA
GGGGGGGCCCCCCCCCCCAGACCCACUGUGCGUUUUUUUUUU

Select small molecules: ● existing database ○ upload SMILES

Select small molecules database:
- Drug Bank Database (~6,000 small molecules)

Index	InChIKey	SMILES	Structure	Probability
1	CHJJGSNFBQVOTG-UHFFFAOYSA-N	[H]/N=C(/N)NC		0.9886
2	NGSXFKKYKWBNPO-UHFFFAOYSA-N	CN(C)c1ccc2c(c1)OC3=CC(=[N+](C)C)C=CC3=C2c4ccccc4		0.977
3	PFHMRNGLRJXGSW-GFCCVEGCSA-N	CN(C)CCCn1c(N)nc2ccc3c(c21)C[C@H](CN(C)C)O3		0.9629
4	PFHMRNGLRJXGSW-GFCCVEGCSA-N	CN(C)CCCn1c2c(ccc3c2C[C@@H](O3)CN(C)C)nc1N		0.9564
5	PAYVJEWMAWLKAP-UHFFFAOYSA-N	CN1CCN(C2=CC=C3N=C(C4=CC=C5N=C(C6=CC(OCCCC(NCCCN=[N+]=[N-])=O)=CC=C6)NC5=C4)NC3=C2)CC1		0.9458
6	AAAAWQOPBUPWEV-UHFFFAOYSA-N	O=C1C=C(OC=2C1=C(O)C=C(O)C2C(C=3C=CC=CC3)CC(=O)OC)C=4C=CC=CC4		0.9227
7	AAACJXBPRRVSAV-UHFFFAOYSA-N	O=C(OC1C(=C)C2CCC3(O)C4(C)CCCC(C(=O)O)(C)C4CCC13C2)C=C(C)C		0.9055
8	RLXLNYXAIXVZCE-UHFFFAOYSA-N	Cc1ccc2ccc(nc2n1)NC(=O)OCCCNCCCOC(=O)Nc3ccc4ccc(nc4n3)C		0.8226

Figure 4. The interface of the SMTRI website

(A) Users can input a pair of miRNA and mRNA sequences in FASTA format and input SMs by choosing a database or uploading user-preferred ones in SMILES format. (B) The results table shows the InChIKeys, SMILES strings, structure images, and calculated probability scores of the predicted SMs. The probability score indicates how confidently the SM can bind to the miRNA-mRNA pair. (C) The similarity matrix for pairwise comparisons between predicted SMs is also given as part of the results.

times. In our study, we greatly reduced the storage space for RNA representations by focusing only on the binding pockets, whose average length was 6.4 nt. Another essential factor that influenced the effectiveness of SMTRI was that we employed multiple ways to fully extract the features, especially for RNA motifs. Not limited to nucleotide-level statistical features, we utilized an LSTM-SAE to map sequence-level RNA motifs into latent space. Consequently, the prediction accuracy of SMTRI was more convincing than other methods.

However, there are also weaknesses in our proposed method. The experimentally confirmed RNA motif-SM associations used in this study are not enough. More RNA motif-SM associations need to be verified by experiments, which would promote the robustness of our model. Moreover, the diversity of website functionality and display needs to be improved in the next version. An interface for uploading experimentally verified data will be developed to accept data from other research groups. Despite the imperfections, we believe that our work has brought new solutions to RNA-targeted SM drug discovery and provided better-guiding information for wet experiments.

MATERIALS AND METHODS

Data acquisition and processing

RNA secondary structural motifs and bound SMs were obtained from RNALigands,[18] PDB,[19] PubChem,[20] and RPocket.[21] RNALigands curated RNA-SM data from three public databases—Inforna,[39] R-Bind,[40] and PDB[19]—which provided ready-made RNA motif-SM pairs. For data cleaning and filtering, pairs that contained either invalid RNA motifs or nonexistent SM names were removed. The leftover data have five RNA motifs: bulge loop, internal loop, hairpin loop, exterior loop, and multi-branch loop. All the data with bulge loops and internal loops were reserved, while only a few with hairpin loops, exterior loops, and multi-branch loops were reserved. The reason is that mature miRNAs are short (~22 nt) and likely to form bulge loops and internal loops with target mRNAs. In this case, forming hairpin loops or multi-branch loops is almost impossible. Finally, all corresponding SMs in the remaining data were converted into SMILES format. Hence, we screened 809 valid RNA motif-SM pairs as positive samples.

We further collected RNA-SM complexes from PDB and RPocket that are not covered by RNALigands. For this part of the data, we first constructed RNA secondary structures from their primary sequences via RNAfold.[41] We then manually extracted the RNA motifs from the dot-bracket notations of their secondary structures to form the RNA motif-SM pairs. The PDB set contains 125 positive pairs, and the RPocket set contains 28 positive pairs.

We also downloaded the publicly available dataset (AID: 2289) of high-throughput screens on SM modulators of hsa-miR-21-5p from PubChem. We filtered out 882 true positives according to the data processing strategy in Jamal et al.[12]: by using an Excel-based approach, compounds with a PubChem Activity Score between 40 and 100 were considered active (3,282 counts); FLuc inhibitors were then eliminated from these active compounds utilizing the counter-screen of mir-21 project (AID: 588342). We retained 2,600 of the total negatives in AID: 2289 to form the complete PubChem set.

To generate negative samples for all datasets except the PubChem set, SMs from the Zinc database[26] that did not pair with existing RNA motifs were chosen to make up most of the negative ones. SMs only with a cosine similarity of <0.95 to the original paired SMs were selected in this procedure to enhance the credibility of made negative samples. A few SMs from positive samples were also used to make negatives by mismatching them with RNA motifs. Likewise, SMs with a cosine similarity of <0.85 to the original paired ones were considered. The ratio between positive and negative samples was set at ~1:4 for the RNALigands set (4,051 samples in total) and ~1:3 for the PDB (470 samples in total) and RPocket (100 samples in total) sets.

We used 90% of the data from the RNALigands set for training purposes and reserved 10% of the data as an independent testing set. We set up three additional testing sets from the PDB, PubChem, and RPocket sets, respectively.

Feature engineering

After data processing, the RNA motifs were stored in letter-bracket notation, and the SMs were stored in SMILES format. We then transformed these strings into numerical features.

For RNA motifs, we extracted statistical features from nucleotides and latent features from primary sequences. First, an RNA motif in miRNA-mRNA interaction was read from the 5′ end of miRNA to the 3′ end of mRNA that forms a closed loop structure to produce a string sequence made up of “A,” “C,” “G,” and “U” letters (Figure S2). Second, statistical features (41 bits) were extracted from the nucleotide compositions and combinations in the string sequence, such as the sequence length, GC content, ratio of each nucleotide, ratio of each combination of two neighboring nucleotides, and length ratio of miRNA (Table S6). Third, we used an LSTM-SAE to encode the string sequence into 512-bit latent features. Finally, statistical and latent features were linearly concatenated into a unique 553-bit numeric vector.

For SMs, we extracted molecular fingerprints and descriptors from SMILES strings. First, a SMILES string was converted into 1,024-bit extended 3D fingerprints,[42] which outperformed the other 15 tested molecular fingerprinting approaches in our classification task (Table S7). Second, 43-bit molecular descriptors were generated for the SMILES string through the RDKit (http://www.rdkit.org) toolkit. Finally, molecular fingerprints and descriptors were linearly concatenated into a unique 1,067-bit numerical vector. We then connected the RNA motif feature vector and SM feature vector together to form a 1,620-bit feature vector for deep learning (Figure 2A).

LSTM-based stacked autoencoder

We constructed an LSTM-SAE to map the input into latent space and extracted high-dimensional features for the input. The original input was the string sequence of RNA motif in arbitrary length, consisting of “A,” “C,” “G,” and “U” letters. We mapped the sequence into a fixed-length one-hot vector (5 × 16 matrix) through one-hot encoding techniques.

The LSTM-SAE architecture consisted of four LSTM layers (Figure 2B): two LSTM encoders and two LSTM decoders. One LSTM encoder (512 bits) was stacked on top of another (64 bits) to expand the dimensionality of the input vector (5 bits). One LSTM decoder (64 bits) was stacked on top of another (512 bits) to reduce the dimensionality of the latent vector (512 bits) to the original dimension of the input. Thus, in this process, the encoders took the input vector and encoded it into a latent representation vector (512 bits). Then, the decodes took that vector as input and reconstructed the original vector. The cost function of this LSTM-SAE was the categorical cross-entropy of the difference between the input vector and the reconstructed vector.

We trained the LSTM-SAE with RNA motif sequences in the training set and saved the encoder parts as a model for feature engineering.

Deep convolutional neural networks

SMTRI CNN architecture consisted of 17 layers: an input elayer, three one-dimensional (1D) convolutional layers, three 1D max-pooling layers, three fully connected layers, six batch normalization layers, and one output layer. A 1D convolutional layer $(3 \times 3, 2 \times 2, 2 \times 2)$

and a 1D max-pooling layer were stacked 3 times, followed by 3 fully connected layers (512-, 64-, and 8-hidden units, respectively) and 1 output layer (Figure 2C). Batch normalization was applied after each convolutional and fully connected layer to re-center and re-scale the input for the successive layer. The activation function used in all hidden layers was a rectified linear unit. The activation function used in the output layer was sigmoid, which squashed vectors from the last hidden layer to a value between 0 and 1. A kernel regularizer (L2: 0.01) was applied in each convolutional and fully connected layer by adding a penalty term to the network weights to prevent overfitting. The total number of parameters in this architecture was ∼0.86 million.

The input layer had 1,620 neurons corresponding to a 1,620-bit input feature vector of an RNA motif-SM pair. The output layer had only one neuron to output a probability score, which indicated the likelihood of an SM targeting a given RNA motif. A higher probability score means a greater chance that the SM and RNA will interact.

For training processes, the model adopted binary cross-entropy as the loss function and the Adamax optimization algorithm (learning rate: 0.001) as the optimizer and was trained with a batch size of 32 in 78 epochs to minimize the loss. In 5-fold cross-validation, the training set was divided into five parts, one of which took turns making the validation set. The learning curves were depicted according to training and validation loss at each epoch to diagnose underfitting and overfitting issues. The TensorFlow and Keras version 2.11.0 were adopted for model construction.

Website construction

The server's back end was developed using Java via Springboot3 Framework, while the front end was developed using Java, JavaScript, and HTML5. The data were stored in the Apache HBase 2.9 (https://hbase.apache.org/) database, with Redis (https://redis.io/) as a cache database for high-speed data transmission. The web server is hosted on a Linux machine with a CentOS 7.9 64-bit operating system.

SMTRI has been tested on web browsers, including Microsoft Edge, Safari, Chrome, and Firefox, on different operating systems (Linux, MacOS, and Windows). In addition to interface interaction, SMTRI can be programmatically accessed through Curl commands.

DATA AND CODE AVAILABILITY

SMTRI is freely available at http://smtri.net. Documentation on how to use the web server is available at http://smtri.net/#/help. The code is available at https://github.com/huan-xiao/SMTRI. The notebooks reproducing the results of case studies and the comparisons of SOTA algorithms were also uploaded to GitHub. A docker image (https://hub.docker.com/u/hilaryhsiao) of SMTRI was generated to be run locally with easy-to-follow instructions at https://github.com/huan-xiao/SMTRI/blob/main/README.md. The RNA motif-SM data, all the saved best models, and the prediction results are available at https://zenodo.org/records/11201636.

ACKNOWLEDGMENTS

This work was supported by the Hong Kong General Research Fund (CUHK 14108322, CUHK 14109721, CUHK 14103121, CUHK 14103420, HKBU 12114416, HKBU 12101117, HKBU 12100918, HKBU 12101018, HKBU 12103519, and HKBU 14100218), the CUHK Direct Grant (4054714), the Interdisciplinary Research Clusters Matching Scheme of Hong Kong Baptist University (RC-IRCs/17-18/02), and the Theme-based Research Scheme (T12-201-20R).

AUTHOR CONTRIBUTIONS

Conceptualization: B.-T.Z. Methodology: H.X. Formal analysis: H.X. Data curation: H.X. and Y.Z. Validation: Y.Z. Resources: S.Y., Z.C., and G.Z. Software: H.X. Writing – original draft: H.X. Writing – review & editing: Y.Z. and X.Y. Visualization: X.Y. Supervision: A.L., Z.Z., G.Z., and B.-T.Z. Project administration: B.-T.Z. Funding acquisition: Z.Z., G.Z., and B.-T.Z.

DECLARATION OF INTERESTS

A patent on SMTRI has been filed by The Chinese University of Hong Kong, with B.-T.Z., H.X., and Z.Z. as inventors.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2024.102303.

REFERENCES

1. Belter, A., Gudanis, D., Rolle, K., Piwecka, M., Gdaniec, Z., Naskręt-Barciszewska, M.Z., and Barciszewski, J. (2014). Mature miRNAs form secondary structure, which suggests their function beyond RISC. PLoS One *9*, e113848.
2. Liu, B., Childs-Disney, J.L., Znosko, B.M., Wang, D., Fallahi, M., Gallo, S.M., and Disney, M.D. (2016). Analysis of secondary structural elements in human microRNA hairpin precursors. BMC Bioinf. *17*, 112.
3. Mauger, D.M., Cabral, B.J., Presnyak, V., Su, S.V., Reid, D.W., Goodman, B., Link, K., Khatwani, N., Reynders, J., Moore, M.J., and McFadyen, I.J. (2019). mRNA structure regulates protein expression through changes in functional half-life. Proc. Natl. Acad. Sci. USA *116*, 24075–24083.
4. Aguilar, R., Spencer, K.B., Kesner, B., Rizvi, N.F., Badmalia, M.D., Mrozowich, T., Mortison, J.D., Rivera, C., Smith, G.F., Burchard, J., et al. (2022). Targeting Xist with compounds that disrupt RNA structure and X inactivation. Nature *604*, 160–166.
5. Matsui, M., and Corey, D.R. (2017). Non-coding RNAs as drug targets. Nat. Rev. Drug Discov. *16*, 167–179.
6. Adams, B.D., Parsons, C., Walker, L., Zhang, W.C., and Slack, F.J. (2017). Targeting noncoding RNAs in disease. J. Clin. Invest. *127*, 761–771.
7. Warner, K.D., Hajdin, C.E., and Weeks, K.M. (2018). Principles for targeting RNA with drug-like small molecules. Nat. Rev. Drug Discov. *17*, 547–558.
8. Jonas, S., and Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. Nat. Rev. Genet. *16*, 421–433.
9. Zhuo, Z., Wan, Y., Guan, D., Ni, S., Wang, L., Zhang, Z., Liu, J., Liang, C., Yu, Y., Lu, A., et al. (2020). A Loop-Based and AGO-Incorporated Virtual Screening Model Targeting AGO-Mediated miRNA-mRNA Interactions for Drug Discovery to Rescue Bone Phenotype in Genetically Modified Mice. Adv. Sci. *7*, 1903451.
10. Zhao, R., Fu, J., Zhu, L., Chen, Y., and Liu, B. (2022). Designing strategies of small-molecule compounds for modulating non-coding RNAs in cancer therapy. J. Hematol. Oncol. *15*, 14.
11. Childs-Disney, J.L., Yang, X., Gibaut, Q.M.R., Tong, Y., Batey, R.T., and Disney, M.D. (2022). Targeting RNA structures with small molecules. Nat. Rev. Drug Discov. *21*, 736–762.
12. Jamal, S., Periwal, V.; Open Source Drug Discovery Consortium, and Scaria, V. (2012). Computational analysis and predictive modeling of small molecule modulators of microRNA. J. Cheminf. *4*, 16.
13. Wang, C.C., Chen, X., Qu, J., Sun, Y.Z., and Li, J.Q. (2019). RFSMMA: A New Computational Model to Identify and Prioritize Potential Small Molecule-MiRNA Associations. J. Chem. Inf. Model. *59*, 1668–1679.
14. Zhao, Y., Chen, X., Yin, J., and Qu, J. (2020). SNMFSMMA: using symmetric nonnegative matrix factorization and Kronecker regularized least squares to predict potential small molecule-microRNA association. RNA Biol. *17*, 281–291.

15. Shen, C., Luo, J., Lai, Z., and Ding, P. (2020). Multiview Joint Learning-Based Method for Identifying Small-Molecule-Associated MiRNAs by Integrating Pharmacological, Genomics, and Network Knowledge. J. Chem. Inf. Model. *60*, 4085–4097.

16. Oliver, C., Mallet, V., Gendron, R.S., Reinharz, V., Hamilton, W.L., Moitessier, N., and Waldispühl, J. (2020). Augmented base pairing networks encode RNA-small molecule binding preferences. Nucleic Acids Res. *48*, 7690–7699.

17. Tong, Y., Lee, Y., Liu, X., Childs-Disney, J.L., Suresh, B.M., Benhamou, R.I., Yang, C., Li, W., Costales, M.G., Haniff, H.S., et al. (2023). Programming inactive RNA-binding small molecules into bioactive degraders. Nature *618*, 169–179.

18. Sun, S., Yang, J., and Zhang, Z. (2022). RNALigands: a database and web server for RNA-ligand interactions. RNA *28*, 115–122.

19. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

20. Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., and Bryant, S.H. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. *37*, W623–W633.

21. Zhou, T., Wang, H., Zeng, C., and Zhao, Y. (2021). RPocket: an intuitive database of RNA pocket topology information with RNA-ligand data resources. BMC Bioinf. *22*, 428.

22. Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell *126*, 1203–1217.

23. Kanazawa, H., Saavedra, O.M., Maianti, J.P., Young, S.A., Izquierdo, L., Smith, T.K., Hanessian, S., and Kondo, J. (2018). Structure-Based Design of a Eukaryote-Selective Antiprotozoal Fluorinated Aminoglycoside. ChemMedChem *13*, 1541–1548.

24. Costales, M.G., Haga, C.L., Velagapudi, S.P., Childs-Disney, J.L., Phinney, D.G., and Disney, M.D. (2017). Small Molecule Inhibition of microRNA-210 Reprograms an Oncogenic Hypoxic Circuit. J. Am. Chem. Soc. *139*, 3446–3455.

25. Paulsen, R.B., Seth, P.P., Swayze, E.E., Griffey, R.H., Skalicky, J.J., Cheatham, T.E., 3rd, and Davis, D.R. (2010). Inhibitor-induced structural change in the HCV IRES domain IIa RNA. Proc. Natl. Acad. Sci. USA *107*, 7263–7268.

26. Sterling, T., and Irwin, J.J. (2015). ZINC 15–Ligand Discovery for Everyone. J. Chem. Inf. Model. *55*, 2324–2337.

27. Sorokina, M., Merseburger, P., Rajan, K., Yirik, M.A., and Steinbeck, C. (2021). COCONUT online: Collection of Open Natural Products database. J. Cheminf. *13*, 2.

28. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082.

29. Iacomino, G. (2023). miRNAs: The Road from Bench to Bedside. Genes *14*, ââ314.

30. Zhu, Y., Zhu, L., Wang, X., and Jin, H. (2022). RNA-based therapeutics: an overview and prospectus. Cell Death Dis. *13*, 644.

31. Saenz-Pipaon, G., and Dichek, D.A. (2023). Targeting and delivery of microRNA-targeting antisense oligonucleotides in cardiovascular diseases. Atherosclerosis *374*, 44–54.

32. Monroig, P.D.C., Chen, L., Zhang, S., and Calin, G.A. (2015). Small molecule compounds targeting miRNAs for cancer therapy. Adv. Drug Deliv. Rev. *81*, 104–116.

33. Disney, M.D., and Angelbello, A.J. (2016). Rational Design of Small Molecules Targeting Oncogenic Noncoding RNAs from Sequence. Acc. Chem. Res. *49*, 2698–2704.

34. Angelbello, A.J., Chen, J.L., Childs-Disney, J.L., Zhang, P., Wang, Z.F., and Disney, M.D. (2018). Using Genome Sequence to Enable the Design of Medicines and Chemical Probes. Chem. Rev. *118*, 1599–1663.

35. O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. (2018). Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front. Endocrinol. *9*, 402.

36. Mazumder, A., Bose, M., Chakraborty, A., Chakrabarti, S., and Bhattacharyya, S.N. (2013). A transient reversal of miRNA-mediated repression controls macrophage activation. EMBO Rep. *14*, 1008–1016.

37. Kudlow, B.A., Zhang, L., and Han, M. (2012). Systematic analysis of tissue-restricted miRISCs reveals a broad role for microRNAs in suppressing basal activity of the C. elegans pathogen response. Mol. Cell *46*, 530–541.

38. Wilczynska, A., and Bushell, M. (2015). The complexity of miRNA-mediated repression. Cell Death Differ. *22*, 22–33.

39. Disney, M.D., Winkelsas, A.M., Velagapudi, S.P., Southern, M., Fallahi, M., and Childs-Disney, J.L. (2016). Inforna 2.0: A Platform for the Sequence-Based Design of Small Molecules Targeting Structured RNAs. ACS Chem. Biol. *11*, 1720–1728.

40. Donlic, A., Swanson, E.G., Chiu, L.Y., Wicks, S.L., Juru, A.U., Cai, Z., Kassam, K., Laudeman, C., Sanaba, B.G., Sugarman, A., et al. (2022). R-BIND 2.0: An Updated Database of Bioactive RNA-Targeting Small Molecules and Associated RNA Secondary Structures. ACS Chem. Biol. *17*, 1556–1566.

41. Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R., and Hofacker, I.L. (2008). The Vienna RNA websuite. Nucleic Acids Res. *36*, W70–W74.

42. Axen, S.D., Huang, X.P., Cáceres, E.L., Gendelev, L., Roth, B.L., and Keiser, M.J. (2017). A Simple Representation of Three-Dimensional Molecular Structure. J. Med. Chem. *60*, 7393–7409.