Published online 11 May 2023 *Nucleic Acids Research, 2023, Vol. 51, Web Server issue* ***W509–W519***
https://doi.org/10.1093/nar/gkad404

RNAincoder: a deep learning-based encoder for RNA and RNA-associated interaction

Yunxia Wang[1], Zhen Chen[1], Ziqi Pan[1], Shijie Huang[1], Jin Liu[1], Weiqi Xia[1], Hongning Zhang[1], Mingyue Zheng[2], Honglin Li[1,3], Tingjun Hou[1] and Feng Zhu[1,4,5,*]

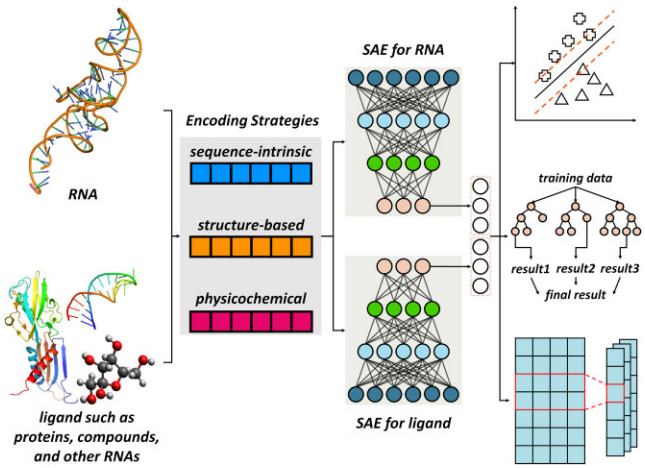
[1]College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China, [2]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China, [3]School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China, [4]Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Hangzhou 330110, China and [5]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, China

Received February 20, 2023; Revised April 12, 2023; Editorial Decision May 02, 2023; Accepted May 10, 2023

ABSTRACT

Ribonucleic acids (RNAs) involve in various physiological/pathological processes by interacting with proteins, compounds, and other RNAs. A variety of powerful computational methods have been developed to predict such valuable interactions. However, all these methods rely heavily on the 'digitalization' (also known as 'encoding') of RNA-associated interacting pairs into a computer-recognizable descriptor. In other words, it is urgently needed to have a powerful tool that can not only represent each interacting partner but also integrate both partners into a computer-recognizable interaction. Herein, *RNAin*coder (deep learning-based encoder for RNA-associated interactions) was therefore proposed to (a) provide a comprehensive collection of RNA encoding features, (b) realize the representation of any RNA-associated interaction based on a well-established deep learning-based embedding strategy and (c) enable large-scale scanning of all possible feature combinations to identify the one of optimal performance in RNA-associated interaction prediction. The effectiveness of *RNAin*coder was extensively validated by case studies on benchmark datasets. All in all, *RNAin*coder is distinguished for its capability in providing a more accurate representation of RNA-associated interactions, which makes it an indispensable complement to other available tools. *RNAin*coder can be accessed at https://idrblab.org/rnaincoder/

GRAPHICAL ABSTRACT



INTRODUCTION

Ribonucleic acids (RNAs) are mainly known to function as catalytic molecules in gene expression (1–3) and play fundamental roles in the regulation of diverse biological and pathological processes (4–6). Considerable research has proved that the interactions between RNA and other molecules including RNAs, proteins and compounds, are crucial to RNAs' functions (7–9). Related studies have gained huge momentum and spawned the development of a variety of powerful computational methods to predict such valuable interactions (8,10,11). All these methods rely heavily on the 'digitalization' (also known as 'encoding') of RNA-associated interacting pairs into a computer-recognizable descriptor (12), which asks for the development of functional tools that can digitalize RNAs, proteins and compounds (13–16).

*To whom correspondence should be addressed. Tel: +86 571 88208444; Fax: +86 571 88208444; Email: zhufeng@zju.edu.cn; prof.zhufeng@gmail.com

So far, various methods/tools aiming at accurately and efficiently digitalizing different types of molecules have been constructed (17–20). PROFEAT is a widely-used web server that can compute a total of 11 feature groups of popular descriptors for proteins and peptides (17). PseKRAAC has been developed to generate various kinds of pseudo amino acid compositions (18). PaDEL-Descriptor works as open-source software that calculates 797 molecular descriptors and 10 types of fingerprints with multiple frequently-used user interfaces (19). Mordred is a molecule descriptor calculator that generates >1800 descriptors (20). Besides these computational tools aiming primarily at encoding a certain type of molecule, there are other tools with hybrid functions (21,22). For example, PyDPI and Rcpi are standalone packages used for computing protein and small molecule features to study protein–protein interactions and compound-protein interactions (21,22).

Among these existing tools, some focus on encoding one type of molecule, such as protein, compound and RNA, and the others are merely used to encode interaction between protein and compound (17–19,21,22). However, there are currently no tools to encode RNA-associated interactions. Moreover, to the best of our knowledge, the encoding strategies in the servers encoding RNA are far from comprehensive (23,24). In other words, it is urgently needed to have a powerful tool for studying RNA-associated interactions, which can not only describe RNA and its interacting partners but also integrate both molecules into an interacting pair (25). However, no such tool has been available yet.

Herein, *RNAin*coder was therefore proposed to (a) provide a comprehensive collection of RNA encoding features (including sequence-intrinsic, physicochemical and structure-based ones), (b) realize the representation of any RNA-associated interaction based on a well-established deep learning-based embedding strategy and (c) enable large-scale scanning of all possible feature combinations to identify the one of optimal performance in RNA-associated interactions prediction. The usefulness of *RNAin*coder was extensively exhibited by three case studies in the last section of this work. All in all, when comparing with the strategies applied in the original publications, *RNAin*coder consistently achieved better predictive performances of RNA-associated interactions. *RNAin*coder was freely available at https://idrblab.org/rnaincoder/ and the local version was released at https://github.com/idrblab/rnaincoder/.

MATERIALS AND METHODS

Collection of the comprehensive strategies for encoding RNA

Currently, 380 RNA descriptors commonly applied in the RNA encoding process were collected and integrated into *RNAin*coder, which included 10 encoding feature groups, as shown in Table 1. These feature groups were grouped into three categories: 177 sequence-intrinsic features (subdivided into six feature groups), 195 physicochemical features (subdivided into three feature groups) and eight structure-based features (belonging to one feature group).

The sequence-intrinsic features enrolled in this study included six feature groups: codon related (CDR), open reading frame (ORF), guanine–cytosine related (GCR), *K*-mer (KME), global descriptor (GBD) and entropy density related (EDT). Specifically, when the length of the RNA sequence was over 200nt, the Fickett score (a subgroup of CDR), could achieve 94% sensitivity and 97% specificity for the identification of long non-coding RNA (lncRNA) (26). The ORF was a feasible and meaningful RNA feature group on account of a long and high-quality ORF for the protein-coding transcript (24). KME was a simple approach to encoding RNA sequences through the occurrence frequencies of *k* neighboring nucleic acids and has been successfully applied to the functional classification of lncRNAs (27). Besides, GCR, GBD and EDT have shown effective enhancement in RNA prediction (28), classification (7) and annotation (29).

Physicochemical features were descriptors related to RNA and its product. Physicochemical features applied in this study included three feature groups: Electron-ion interaction pseudopotential (EIIP) based spectrum features (EBS), nucleotide related (NTR) and pseudo protein related (PPR). To be specific, EIIP values were indications of the energy of delocalized electrons in nucleotides (28). NTR contained autocorrelation of dinucleotide features and pseudo dinucleotide composition (PseDNC). Autocorrelation of dinucleotide features was the correlation of identical physicochemical features between two nucleotide residues separated by a certain distance along the RNA sequence (30). PseDNC incorporated three angular parameters (twist, tilt and roll) and three translational parameters (shift, slide and rise) physicochemical features (31). The calculation process of PPR consisted of two steps: (a) All RNA sequences were transformed to corresponding amino acid sequences or pseudo-protein sequences according to the genetic code. (b) Calculate the physicochemical features of transformed protein sequences (32).

Structure-based features were several descriptors that depicted the established RNA secondary and tertiary structure, which were essential for many RNA functions (33). Particularly, the medium-scale feature and high-scale feature of RNA secondary structure could be well-displayed in dot-bracket notation (34). Therefore, structure-based features were critical to RNA representation.

The full names and descriptions of the 380 RNA encoding feature mentioned above were provided in Supplementary Table S1. The detailed descriptions and application of these encoding methods mentioned above were provided in Supplementary Methods, which included the corresponding parameters, as shown in Supplementary Tables S2–S4.

Collecting the strategies for encoding protein and compound

*RNAin*coder also provided the encoding features of proteins and compounds for the research of RNA-associated interactions, including RNA-protein and RNA-compound interactions. Both types of encoding features were based on previous publications which developed a tool for calculating structural and physicochemical features of proteins (17) and compounds (19). The protein encoding features were grouped in the same way as RNA encoding features because of the similar principle between RNA and protein (35,36).

Table 1. The comprehensive set of RNA encoding features with their brief descriptions

Feature group	Feature subgroup	No. of features	Brief description
Sequence-intrinsic features			
Codon related	Fickett score	1	It is a score to evaluate the variety of nucleotide positions and compositions between mRNAs and lncRNAs (26).
	Stop codon related features	4	Itis a set of features related to stop codon, including stop codon count, frequency, frame score and frequency frame score (73).
Open reading frame	Basic ORF features	4	This feature subgroup is calculated mainly based on the most basic information of open reading frames in RNA sequences, including length, coverage, *etc* (74).
	Entropy density profiles on ORF	20	It is a systematic linguistic description of RNA sequence based on short motif frequency and Shannon entropy theory of artificial language (75).
	Measurement of hexamer on ORF	7	It is a set of features to estimate the relative degree of hexamer usage bias and distinguish between mRNA and non-coding RNA (76).
Guanine-cytosine related	Guanine-cytosine related	7	This feature subgroup describes the efficiency of gene expression at a time of increased steady-state mRNA levels and efficient transcription (77).
K-mer	Transcript *k*-mer content	84	It is a commonly applied approach to code RNA sequences through the occurrence frequencies of k neighboring nucleic acids (78).
Global descriptor	Global transcript sequence descriptors	30	Itis a computing strategy for nucleotide composition, transition and distribution representation in an RNA sequence (79).
Entropy density related	Entropy density profiles on transcript	20	It is a model used to describe the properties of RNA transcript in the 20-dimensional phase space for calculating the coding potential based on amino acid usage (75).
Physicochemical features			
Pseudo protein Related	Pseudo protein related	5	It is a set of features to describe the physicochemical properties of pseudo protein translated from RNA by computational methods (80).
Nucleotide related	Autocorrelation of dinucleotide	136	It is an approach to measuring the autocorrelation between the same properties or cross-covariance between two different RNA properties (81).
	Pseudo dinucleotide composition	46	It is an approach to incorporating the contiguous local and global sequence-order information into the feature vector of the RNA (31).
EIIP-based spectrum	EIIP-based spectrum	8	It is a set of features that represent RNA sequence via electron-ion interaction pseudopotential values for each nucleotide (82).
Structure-based features			
Secondary Structure	Multi-scale secondary Structure information	8	Itis a feature subgroup that represents RNA from three levels: stability, secondary structure elements and multi-scale secondary structure-derived sequences (83).

Features for encoding protein. Specifically, 188 encoding features frequently adopted in protein function research were collected in *RNAin*coder, which included 20 sequence-intrinsic features, 147 physicochemical features, and 21 structure-based features, as shown in Supplementary Table S5.

Sequence-intrinsic features transformed protein sequences into computer-recognizable matrices, including amino acid composition and position specific scoring. Amino acid composition represented the content of each kind of amino acids and was used to predict protein family (37).

Physicochemical features covered the physicochemical characteristics of amino acids. The physicochemical features involved in this study were based on an electric charge, hydrophobicity, polarity, polarizability, solvent accessibility, surface tension and van der Waals volume. These descriptors were based on eight kinds of physicochemical features and had been applicated to analysis of protein arginine methylation (38,39).

Structure-based features described the structural characteristics of amino acids and peptides. These descriptors were mainly based on secondary structure and related solvent accessibility, which had been used for the prediction of protein–RNA interactions using machine learning models (40).

Features for encoding compound. Furthermore, the encoding features of compounds in *RNAin*coder were also grouped into three classes according to a previous publication (19). In particular, 2756 descriptors frequently adopted in small molecule research were collected, which included 1444 composition topology descriptors, 431 stereo-structural descriptors and 881 small molecules PubChem fingerprints, as shown in Supplementary Table S5.

The composition topology descriptors involved in this study included autocorrelation descriptors, Barysz matrix descriptors, constitutional descriptors, physicochemical descriptors and topology-related descriptors. Composition topological descriptors such as physicochemical descriptors had been used to predict drug aqueous solubility (41).

3D-shape functionality descriptors contained 3D functionality such as 3D autocorrelation, charged partial surface area, gravitational index, length over breadth, moment inertia, Petitjean shape index and radial distribution function. 3D autocorrelation descriptors such as spatial autocorrelation descriptors had been developed for molecular modeling (42).

Small molecule fingerprints used fixed-length arrays to digitize different compounds. PubChem fingerprint was mainly applied in this study. PubChem fingerprint characterized small molecules by the number of functional groups and had been used to present drug chemical structure in side effect prediction (43).

Deep learning-based embedded feature integration

The deep learning methods have made outstanding contributions in many RNA-related research fields (44,45) and keep an upward tendency in the application of RNA-associated interactions during the era of big data (29,46–48). The deep learning-based unsupervised learning algorithm can effectively reduce the dimensions of RNA encoding features and extract more discriminative features in the circumstance of insufficient prior knowledge (49). An autoencoder (AE) is applied to learn efficient data representations in an unsupervised manner, which included three layers: an input layer, a hidden layer and an output layer. AE-related variant stacked AE (SAE) (50) is widely used and has shown exceptional capacity in promoting the prediction of RNA-associated interactions. SAE was constructed and applied in *RNAin*coder, as shown in Figure 1.

Specifically, the SAE consisting of three autoencoders was utilized in *RNAin*coder to extract high-level embedded features from the encoding features of RNA and RNA-interacting molecules. The embedded features were obtained in the following steps: (i) The RNA encoding features were taken as input to train the AE1 via back-propagation algorithm, getting the hidden feature 1 and 1st hidden layer. (ii) The hidden feature 1 served as the input for AE2 subsequently to attain the hidden feature 2 and 2nd hidden layer. The AE3 training strategy followed the same way as AE2. (iii) 1st/2nd/3rd hidden layer from the AE1/2/3 and a classifier were incorporated as the SAE. The parameters in SAE got fine-tuned based on the label of the training dataset and then updated.

The SAEs applied to extract embedded features from encoding features of RNA and RNA-interacting molecules were trained respectively and each AE adopted the full-connection layer neural network to realize the compression and reduction processes (51). Ultimately, the embedded features for RNA and RNA-interacting molecules were concatenated and fed into the downstream classifier, such as machine learning algorithm (random forest (52), support vector machine (53), and extreme gradient boosting (54)) or deep learning models (recurrent neural networks (55) and convolutional neural networks (56)) to predict the RNA-associated interactions.

For proper evaluation of *RNAin*coder, several standard evaluation metrics have been used, including the area under the receiver operating characteristic curve (ROC-AUC), Matthews correlation coefficient (MCC), accuracy (ACC), precision (PRE), specificity (SP) and sensitivity (SN). Statistical significance assessment was calculated by one-way ANOVA with Dunnett's post hoc test. The statistical significance was denoted by $*P < 0.05$; $**P < 0.01$; $***P < 0.001$; $****P < 0.0001$.

Server implementation details and required format of input files

The *RNAin*coder server was hosted on a Linux server of an Intel(R) Xeon(R) Gold 6149 3.10 GHz CPUs with 8 cores and 64 GB of memory based on the Python web framework of Tornado (an asynchronous networking library). *RNAin*coder could be free and open to all users with no login requirement and could be accessed at https://idrblab.org/rnaincoder/ by diverse and popular web browsers including *Google Chrome*, *Mozilla Firefox*, *Safari* and *Internet Explorer 10* (or later).

For RNA or protein encoding, the input is a set of RNA or protein sequences in FASTA format, which can be uploaded as a single file. For small molecule compounds, the input is SMILE format, which can be uploaded as single files. For the label file of encoding RNA, the first row of the first 2 columns should be sequentially labeled as 'Seqname' and 'Label', which indicate the *sequence name* and *class of sample* respectively. The *sequence name* should be the RNA sequence name in the FASTA file; the *class of samples* refers to different RNA classes, which should be labeled with an ordinal number (e.g. 0, 1, 2, . . .). For encoding RNA-associated interaction, three files need to be uploaded. The first file is the RNA FASTA file and the last letter of the file name must be 'A'. The second file is an RNA or protein FASTA file and the last letter of the file name must be 'B'. For the label file of RNA-associated interaction, the first row of the first 3 columns should be sequentially labeled as 'A', 'B' and 'Label', which represent *A sequence name, B sequence name* and *the type of interaction*, respectively. The *A sequence name* and *B sequence name* should be RNA or protein sequence names in the FASTA file; the *type of interaction* refers to whether interactions between A and B exist (existing is 1 and non-existing is 0). Various exemplar files strictly following these requirements are fully provided and can be directly downloaded from the *RNAin*coder website. The local version of *RNAin*coder is provided on GitHub at https://github.com/idrblab/rnaincoder.

RESULTS AND DISCUSSION

Effective representation of comprehensive encoding strategies in *RNAin*coder

Due to the important biological function of RNAs (57,58), it remains crucial for wealthy assembled transcripts to annotate the different classes of RNAs and especially to distinguish protein-coding from non-coding RNAs after high-throughput RNA sequencing (59–61). An RNA

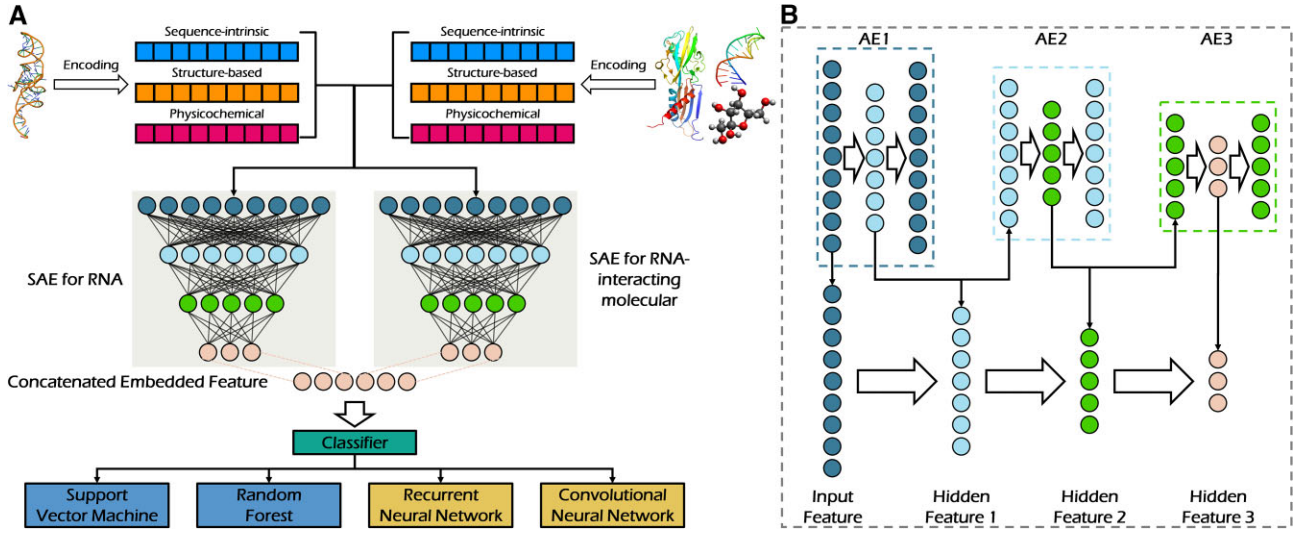


Figure 1. The workflow of (**A**) the deep learning-based embedding strategy for RNA-associated interactions and the framework of (**B**) the stacked autoencoder (SAE) in *RNAin*coder. The stacked autoencoder consisted of three autoencoders and each autoencoder included an encoder and a decoder based on a multilayer perceptron. Embedded features sequentially optimized by encoders in three pre-trained autoencoders would be paired and concatenated for the prediction of RNA-associated interactions.

classification dataset was collected from FEELnc (62) to evaluate the capability of *RNAin*coder for providing comprehensive RNA encoding features. This dataset consisted of 10 000 mRNAs (divided into two sets of 5000 mRNAs used for the training and testing model, respectively) and 10 000 lncRNAs (divided into two sets of 5000 lncRNAs used for the training and testing model, respectively). To illustrate the contribution of the comprehensive encoding features provided by *RNAin*coder in the prediction of RNA coding potential, the performance of *RNAin*coder was compared with state-of-the-art tools, FEELnc (62) and RNAsamba (63), based on the same training sets. The classifiers were random forest and neural network model from FEELnc and RNAsamba, respectively.

As shown in Figure 2, the classification performance of encoding features generated by *RNAin*coder (bar in yellow) achieved improvements at AUC, MCC, ACC, PRE, SP and SN compared with FEELnc (bar in purple). Specifically, *RNAin*coder obtained AUC of 0. 973, MCC of 0.852 and ACC of 0.926. Compared with the results reproduced via the encoding features in FEELnc (62), the AUC, MCC, and ACC achieved by encoding features in *RNAin*coder have been increased by 2.27%, 4.10% and 2.37%, respectively. Meanwhile, *RNAin*coder could also improve the performance of RNAsamba in the prediction of RNA coding potential, as shown in Supplementary Figure S1. For encoding features used in FEELnc, they are merely limited to characterizing the RNA sequence and lack the description of the physicochemical properties and structure of the RNA, which are crucial for distinguishing mRNA from lncRNAs (23). *RNAin*coder integrated a total of 380 encoding features and represented RNA from multiple perspectives (sequence-intrinsic, physicochemical and structure-based features). The encoding features used in RNAsamba have been fully covered by *RNAin*coder. Thus, *RNAin*coder got a better achievement in the identification of RNA coding potential by characterizing RNA more accurately than FEELnc and RNAsamba. It is demonstrated that *RNAin*coder is a powerful tool to provide comprehensive encoding strategies for the studied RNAs.

In addition to the above evaluation of *RNAin*coder on the classification of mRNA and lncRNA, the performance of *RNAin*coder was further verified on the classification of mRNA and ncRNA. First, the previously published tool, RNAming, was trained based on human mRNA and ncRNA dataset (46575 mRNA and 46269 ncRNA), and tested on rat mRNA and ncRNA dataset (9331 mRNA and 9331 ncRNA) for cross-species prediction (64). By directly adopting the classifier and the model construction strategy applied in RNAming, a new model was constructed in our study based on those encoding features of *RNAin*coder. As illustrated in Supplementary Figure S2, comparing with the original features used in RNAming, *RNAin*coder's features could extensively improve classification performance, which significantly elevated the values of MCC, ACC and PRE by 7.6%, 3.9% and 7.5%, respectively.

Superior performance achieved by the integration strategy in *RNAin*coder

RNAs play a crucial role in the physiological processes (65,66) and pathological processes (67) interacting with corresponding other molecules (RNA, protein and compound). Thus, it's necessary to further evaluate the performance of deep learning-based embedded feature integration (SAE), provided by *RNAin*coder in the prediction of RNA-associated interactions. Taking the prediction of RNA-protein interactions as an example, a lncRNA-protein interaction dataset containing 291 lncRNAs and 1460 proteins, named RPI1460, was collected from the latest published LPI-CSFFR (68). RPI1460 included 1460 positive pairs (lncRNA-protein interactive pairs) and 1460 negative pairs (lncRNA-protein noninteractive pairs). As a method of integrating two interacting molecules, *RNAin*coder

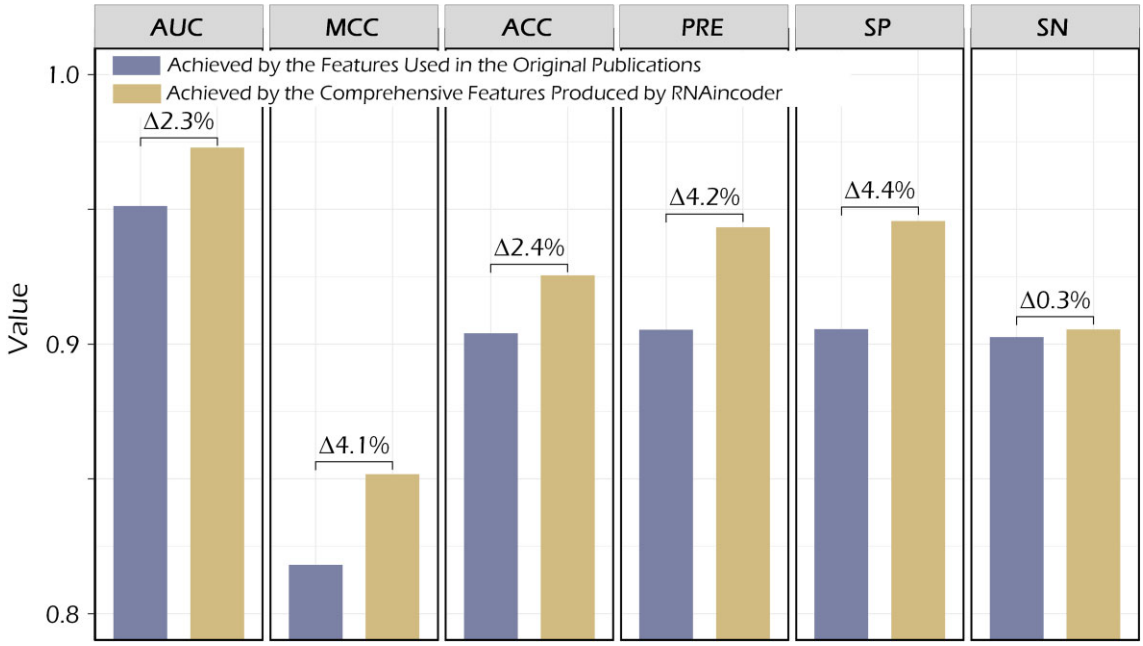


Figure 2. The comparison of performance between comprehensive encoding features provided by *RNAin*coder (bars in yellow) and the original encoding features from FEELnc (62) (bars in purple) in distinguishing protein-coding from non-coding RNAs. Their performance was compared using the metrics of receiver operating characteristic curve (ROC-AUC), Matthews correlation coefficient (MCC), accuracy (ACC), precision (PRE), specificity (SP) and sensitivity (SN) as the indicators and the classifiers from FEELnc (62). The training set and test set were all from FEELnc (62). Δ indicates the increase by *RNAin*coder over the original publication.

extracted and integrated them through SAE. LPI-CSFFR applied a sample direct concatenated method to generate the combined features. The predictive performances of *RNAin*coder and LPI-CSFFR were evaluated on benchmark datasets RPI1460 using five-fold cross-validation based on the convolutional neural networks (CNN) model from LPI-CSFFR (68).

As shown in Figure 3, SAE (boxplot in yellow) displayed a better predictive capacity than feature integration methods in LPI-CSFFR (boxplot in blue) based on the same encoding features and classification model CNN as the original publication (68). To be specific, it was worth indicating that the improvement of *RNAin*coder was obvious and the performance of SAE obtained a great increase of AUC by 5.17%, MCC by 10.6% and ACC by 6.72%. This improvement was quite considerable and was found to be statistically significant. Moreover, the comprehensive encoding features generated by *RNAin*coder (boxplot in orange) outperformed the encoding features in LPI-CSFFR (boxplot in yellow) based on the same deep learning-based integration and classifier. Meanwhile, it is clear to see in Figure 3 that both comprehensive encoding features and deep learning-based integration in *RNAin*coder (boxplot in orange) have achieved a great improvement of AUC by 7.47%, MCC by 15.5% and ACC by 8.58% compared with the encoding features and integration methods using in the original publication (boxplot in blue) (68) based on the same classifier. This improvement was also found to be statistically significant.

To further explore the representation ability of the embedded feature learned by the deep learning model in the prediction of RNA-protein interactions, a semi-supervised dimensionality reduction method (69) and a uniform manifold approximation and projection (UMAP) scatter diagram were used to represent the distribution of interaction and no interaction pairs from RPI1460, as shown in Figure 4 and Supplementary Figure S3, respectively. Specifically, the points in Figure 4A were the concatenation of the RNA encoding features provided by *RNAin*coder and the protein encoding features in LPI-CSFFR for all 1460 sample pairs. After feature extraction by SAE, the embedded features of RNA and protein were concatenated and presented in Figure 4B. It could be seen that the positives and negatives in the embedded feature space were more clearly distributed in two clusters than those in the original feature space. The same result can also be obtained from the visualization of the UMAP method. These results demonstrated that using deep learning-based embedded feature integration improved the feature representation ability of RNA-associated interactions. Using the same way to extract the RNA and protein encoding features in LPI-CSFFR, Figure 4c and d were produced by the above semi-supervised reduction method (69). Supplementary Figure S3c and S3d were also generated by the UMAP method. A similar result illustrated that the representation of positive and negative pairs using embedded features made the same type of sample cluster more closely than the other type of sample.

From the above visualization, overlapping area was observed and indicated that the interacting and non-interacting pairs were not completely separated. The reason might be that there were unannotated interacting pairs in non-interacting pairs of the training set. Particularly,

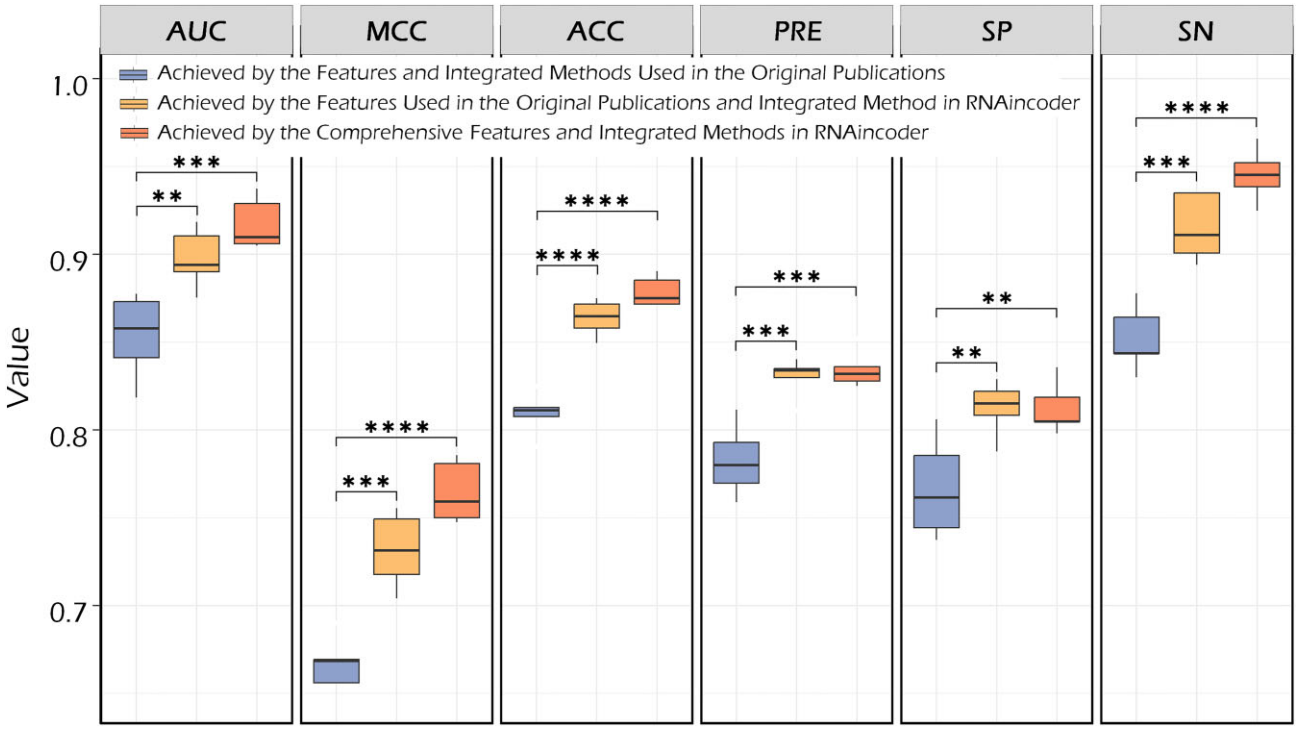


Figure 3. The comparison of performance between embedded features extracted by deep learning-based integration method from the original encoding features in LPI-CSFFR (68) (boxplot in blue), embedded features extracted by deep learning-based integration method from the original encoding features in LPI-CSFFR (68) (boxplot in yellow) and comprehensive encoding features in *RNAin*coder (boxplot in orange) in predicting RNA-protein interactions. Their performance was compared using the metrics of receiver operating characteristic curve (ROC-AUC), Matthews correlation coefficient (MCC), accuracy (ACC), precision (PRE), specificity (SP) and sensitivity (SN) as the indicators over 5-fold cross-validation and the classifiers from LPI-CSFFR (68). The statistical significance was denoted by $*P < 0.05$; $**P < 0.01$; $***P < 0.001$; $****P < 0.0001$.

the interacting pairs were established by calculating atom distances between RNA and protein, which came from RNA-protein complexes in the protein data bank database (70). Non-interacting pairs were generated by adopting the criteria from published literature (71) and were not experimentally validated. There might be interacting pairs among these non-interacting pairs.

Moreover, to provide a real-world test for further illustrating the benefit of *RNAin*coder for users, a dataset of 143 new interactions between 136 proteins and 3 RNAs which were detected by an CRISPR-assisted RNA-protein interaction detection method in the native cellular context was collected (72). Particularly, these 143 novel RPIs were adopted in our study to evaluate the performance of our *RNAin*coder and the LPI-CSFFR. As shown in Table 2, the numbers of 3 RNAs' real-world interaction with proteins (54,46), and (43) were given, and the prediction accuracies of *RNAin*coder and LPI-CSFFR equaled to 96.3–100% and 32.6–58.7%, respectively. It is clear that *RNAin*coder provides significantly better performance than the recent method in RPI prediction, and the improvements of *RNAin*coder from LPI-CSFFR were found to be 41.3–65.1%. The detailed prediction results of these 'real-world' examples were provided in Supplementary Table S6.

All in all, *RNAin*coder could effectively enhance the predictive performance in the identification of RNA-associated interactions using deep learning-based embedded feature integration, which learned the more discriminative features to represent RNA-associated interactions.

Good performance achieved by the large-scale scanning in *RNAin*coder

To demonstrate the variation among the best RNA encoding features for different datasets, the two data sets mentioned above were encoded by 10 individual feature groups in *RNAin*coder. As shown in Figure 5, the best-performing feature groups of two datasets were different. Particularly, open reading frame (shown in Figure 5A) and *K*-mer (shown in Figure 5B) were the optimal feature groups in the identification of RNA coding potential and the prediction of RNA-protein interactions, respectively.

This result inspired us to combine all encoding features (total 380 dimensions) and then conduct a large-scale scanning of all possible feature combinations to identify the best-performing feature combination. The process of large-scale scanning included: (a) ranking all combined 380 RNA encoding features according to the previously published feature ranking method (59), (b) generating 380 feature combinations by iteratively removing the last feature according to the feature rank from the previous step, (c) extracting the embedded feature through deep learning-based integration method (SAE) mentioned above, (d) obtaining the predictive result using the embedded feature as the input of the downstream classifier.

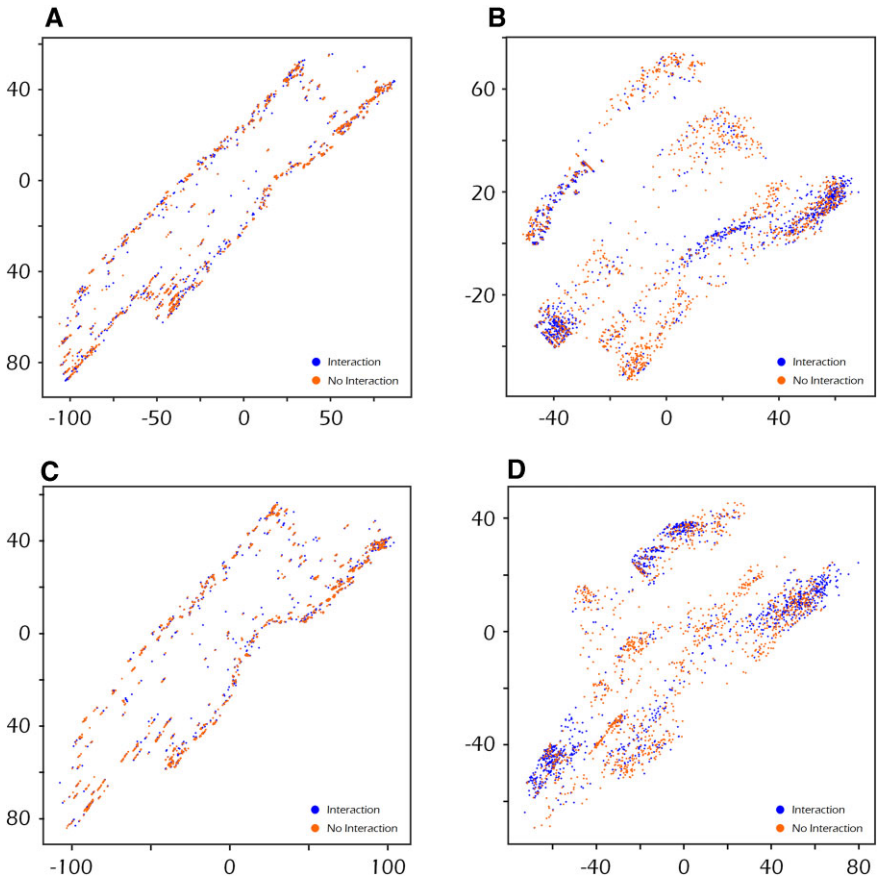


Figure 4. A semi-supervised dimensionality reduction (69) of the RNA-protein interactions dataset for (**A**) encoding features in *RNAin*coder, (**B**) embedded features extracted by deep learning-based integration method from encoding features in *RNAin*coder, (**C**) encoding features in LPI-CSFFR (68), (**D**) embedded features extracted by deep learning-based integration method from encoding features in LPI-CSFFR (68).

Table 2. The performances of *RNAin*coder and the LPI-CSFFR in predicting 143 real-world RPIs newly reported in (72)

RNA name	No. of real-world RPIs	LPI-CSFFR	*RNAin*coder	Improvement
XIST	54	25 (46.3%)	52 (96.3%)	50.0%
DANCR	46	27 (58.7%)	46 (100.0%)	41.3%
MALAT1	43	14 (32.6%)	42 (97.7%)	65.1%

As shown in Figure 5, the best-performing feature combinations (shown in Supplementary Table S7) were identified by large-scale scanning for the prediction of RNA coding potential and RNA-protein interactions, respectively. Particularly, for the identification of RNA coding potential, the performance of the optimal feature combination (bar in purple) achieved an improvement of AUC by 2.26%, MCC by 5.10% and ACC by 2.80% compared with the encoding features used in the original publication (62) (bar in green), as shown in Figure 5c. For the prediction of RNA-protein interactions, the performance of the optimal feature combination (boxplot in blue) obtained an increase of AUC by 6.54%, MCC by 15.5% and ACC by 9.04% compared with the encoding features used in the original publication (68) (boxplot in yellow), as shown in Figure 5d. This increase was also found to be statistically significant.

All in all, based on comprehensive RNA encoding features, *RNAin*coder effectively improved the predictive performance of RNA-associated interactions using a deep learning-based embedded feature integration and a large-scale scanning of all possible feature combinations.

CONCLUSIONS

The *RNAin*coder web server aims at providing an accurate representation of RNA-associated interactions based on collected comprehensive feature encoding methods and deep learning-based feature integration. First, it provides the user with comprehensive RNA encoding features (including sequence-intrinsic, physicochemical, and structure-based ones). Next, it helps the user to obtain a powerful representation of any RNA-associated interaction based on a well-established deep learning-based embedding strategy. Finally, it allows the user to identify the one of optimal feature sets by large-scale scanning of all possible feature combinations. The web server presented herein brings the first free and easy-to-use computational tool for encoding RNA-associated interactions. The *RNAin*coder web server

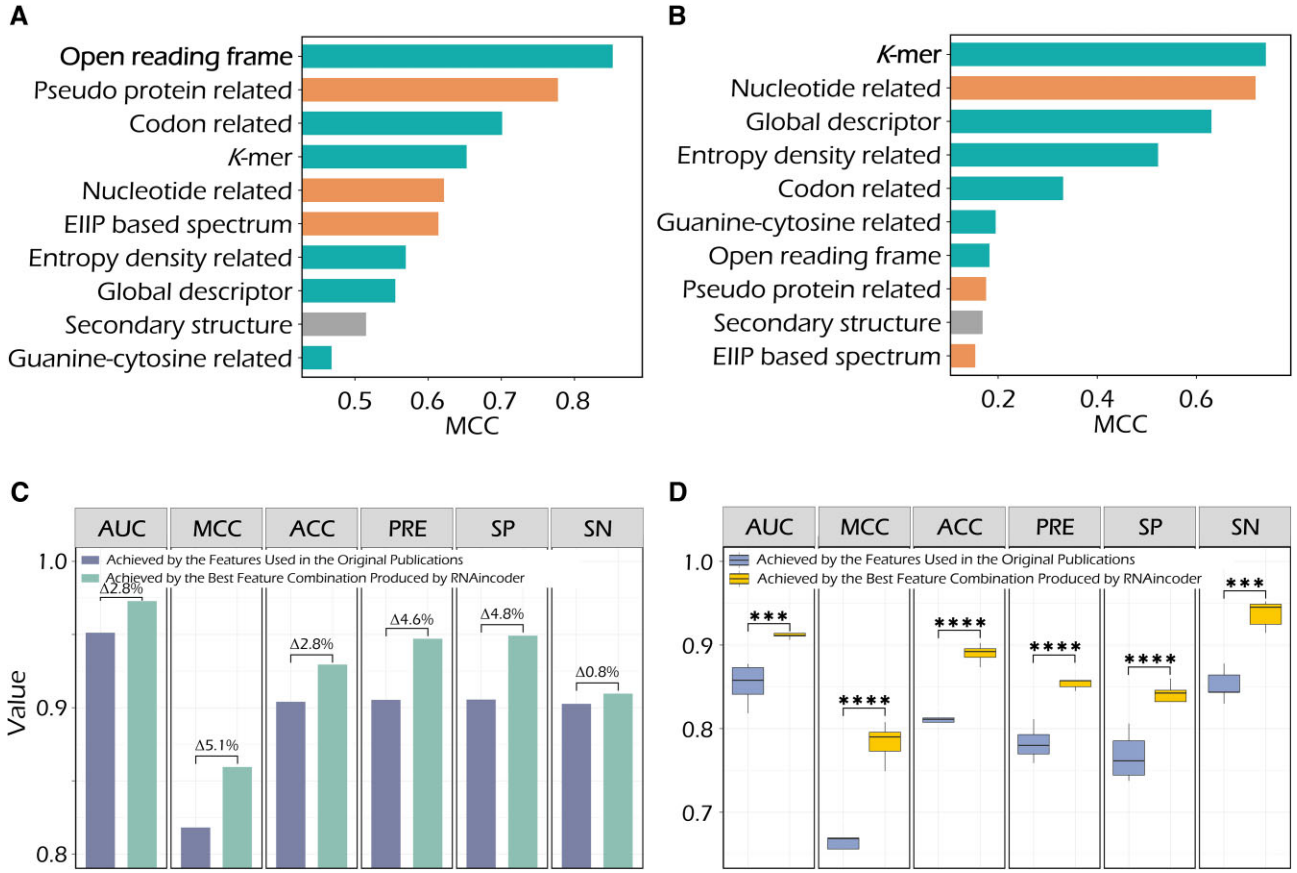


Figure 5. The performance ranking of 10 feature groups for identification of (**A**) RNA coding potential and (**B**) RNA–protein interactions. The comparison of performance between the best feature combination in *RNAin*coder and the original encoding features from previous publications (**C**) FEELnc (62) and (**D**) LPI-CSFFR (68) for identification of RNA coding potential and RNA-protein interactions, respectively. The assessed ten feature groups belong to three feature categories, and the feature groups colored in cyan, orange, and gray indicated sequence-intrinsic, physicochemical, and structure-based categories, respectively. Δ indicates the increase by *RNAin*coder over the original publication. The statistical significance was denoted by $*P < 0.05$; $**P < 0.01$; $***P < 0.001$; $****P < 0.0001$.

will assist in the advancement of RNA-related computational methods in various downstream tasks.

DATA AVAILABILITY

The authors declare that the data supporting the findings of this study are available within the article and its supplementary information files.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Natural Science Foundation of Zhejiang Province [LR21H300001]; National Natural Science Foundation of China [22220102001, U1909208, 81872798]; Leading Talent of the 'Ten Thousand Plan' – National High-Level Talents Special Support Plan of China; Fundamental Research Fund for Central Universities (2018QNA7023); 'Double Top-Class' University Project [181201*194232101]; Key R&D Program of Zhejiang Province [2020C03010]; Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine); Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare; Alibaba Cloud; Information Technology Center of Zhejiang University. Funding for open access charge: Natural Science Foundation of Zhejiang Province [LR21H300001].
Conflict of interest statement. None declared.

REFERENCES

1. Chen,L.L. (2020) The expanding regulatory mechanisms and cellular functions of circular RNAs. *Nat. Rev. Mol. Cell Biol.*, **21**, 475–490.
2. Goodall,G.J. and Wickramasinghe,V.O. (2021) RNA in cancer. *Nat. Rev. Cancer*, **21**, 22–36.
3. Keil,P., Wulf,A., Kachariya,N., Reuscher,S., Huhn,K., Silbern,I., Altmuller,J., Keller,M., Stehle,R., Zarnack,K. *et al.* (2023) Npl3 functions in mRNP assembly by recruitment of mRNP components to the transcription site and their transfer onto the mRNA. *Nucleic Acids Res.*, **51**, 831–851.
4. Willson,J. (2022) Getting organized with non-coding RNAs. *Nat. Rev. Genet.*, **23**, 1.
5. Palcau,A.C., Canu,V., Donzelli,S., Strano,S., Pulito,C. and Blandino,G. (2022) CircPVT1: a pivotal circular node intersecting long non-coding-PVT1 and c-MYC oncogenic signals. *Mol. Cancer*, **21**, 33.
6. Mou,X., Liew,S.W. and Kwok,C.K. (2022) Identification and targeting of G-quadruplex structures in MALAT1 long non-coding RNA. *Nucleic Acids Res.*, **50**, 397–410.

7. Cai,Z., Cao,C., Ji,L., Ye,R., Wang,D., Xia,C., Wang,S., Du,Z., Hu,N., Yu,X. *et al.* (2020) RIC-seq for global in situ profiling of RNA-RNA spatial interactions. *Nature*, **582**, 432–437.
8. Oliver,C., Mallet,V., Gendron,R.S., Reinharz,V., Hamilton,W.L., Moitessier,N. and Waldispuhl,J. (2020) Augmented base pairing networks encode RNA-small molecule binding preferences. *Nucleic Acids Res.*, **48**, 7690–7699.
9. Ramanathan,M., Porter,D.F. and Khavari,P.A. (2019) Methods to study RNA-protein interactions. *Nat. Methods*, **16**, 225–234.
10. Lai,D. and Meyer,I.M. (2016) A comprehensive comparison of general RNA-RNA interaction prediction methods. *Nucleic Acids Res.*, **44**, e61.
11. Armaos,A., Colantoni,A., Proietti,G., Rupert,J. and Tartaglia,G.G. (2021) catRAPID omics v2.0: going deeper and wider in the prediction of protein-RNA interactions. *Nucleic Acids Res.*, **49**, W72–W79.
12. Ryle,P.R. and Dumont,J.M. (1987) Malotilate: the new hope for a clinically effective agent for the treatment of liver disease. *Alcohol Alcohol.*, **22**, 121–141.
13. Yang,S., Wang,Y., Lin,Y., Shao,D., He,K. and Huang,L. (2020) LncMirNet: predicting lncRNA-miRNA interaction based on deep learning of ribonucleic acid sequences. *Molecules*, **25**, 4372.
14. Peng,C., Han,S., Zhang,H. and Li,Y. (2019) RPITER: a hierarchical deep learning framework for ncRNA-protein interaction prediction. *Int. J. Mol. Sci.*, **20**, 1070.
15. Philips,A., Milanowska,K., Lach,G. and Bujnicki,J.M. (2013) LigandRNA: computational predictor of RNA-ligand interactions. *RNA*, **19**, 1605–1616.
16. Mahmud,S.M.H., Chen,W., Liu,Y., Awal,M.A., Ahmed,K., Rahman,M.H. and Moni,M.A. (2021) PreDTIs: prediction of drug-target interactions based on multiple feature information using gradient boosting framework with data balancing and feature selection techniques. *Brief. Bioinform.*, **22**, bbab046.
17. Rao,H.B., Zhu,F., Yang,G.B., Li,Z.R. and Chen,Y.Z. (2011) Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.*, **39**, W385–W390.
18. Zuo,Y., Li,Y., Chen,Y., Li,G., Yan,Z. and Yang,L. (2017) PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics*, **33**, 122–124.
19. Yap,C.W. (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.*, **32**, 1466–1474.
20. Moriwaki,H., Tian,Y.S., Kawashita,N. and Takagi,T. (2018) Mordred: a molecular descriptor calculator. *J. Cheminform.*, **10**, 4.
21. Cao,D.S., Liang,Y.Z., Yan,J., Tan,G.S., Xu,Q.S. and Liu,S. (2013) PyDPI: freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies. *J. Chem. Inf. Model.*, **53**, 3086–3096.
22. Cao,D.S., Xiao,N., Xu,Q.S. and Chen,A.F. (2015) Rcpi: r/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics*, **31**, 279–281.
23. Hu,L., Xu,Z., Hu,B. and Lu,Z.J. (2017) COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. *Nucleic Acids Res.*, **45**, e2.
24. Kang,Y.J., Yang,D.C., Kong,L., Hou,M., Meng,Y.Q., Wei,L. and Gao,G. (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.*, **45**, W12–W16.
25. Weidmann,C.A., Mustoe,A.M., Jariwala,P.B., Calabrese,J.M. and Weeks,K.M. (2021) Analysis of RNA-protein networks with RNP-MaP defines functional hubs on RNA. *Nat. Biotechnol.*, **39**, 347–356.
26. Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.*, **10**, 5303–5318.
27. Kirk,J.M., Kim,S.O., Inoue,K., Smola,M.J., Lee,D.M., Schertzer,M.D., Wooten,J.S., Baker,A.R., Sprague,D., Collins,D.W. *et al.* (2018) Functional classification of long non-coding RNAs by k-mer content. *Nat. Genet.*, **50**, 1474–1482.
28. Han,T., Liu,C., Yang,W. and Jiang,D. (2019) Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions. *ISA Trans.*, **93**, 341–353.
29. Yang,C., Yang,L., Zhou,M., Xie,H., Zhang,C., Wang,M.D. and Zhu,H. (2018) LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics*, **34**, 3825–3834.
30. Zuo,Y., Zou,Q., Lin,J., Jiang,M. and Liu,X. (2020) 2lpiRNApred: a two-layered integrated algorithm for identifying piRNAs and their functions based on LFE-GM feature selection. *RNA Biol.*, **17**, 892–902.
31. Chen,W., Feng,P.M., Lin,H. and Chou,K.C. (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.*, **41**, e68.
32. Yang,S., Wang,Y., Zhang,S., Hu,X., Ma,Q. and Tian,Y. (2020) NCResNet: noncoding ribonucleic acid prediction based on a deep resident network of ribonucleic acid sequences. *Front. Genet.*, **11**, 90.
33. Koodli,R.V., Keep,B., Coppess,K.R., Portela,F., Eterna,p. and Das,R. (2019) EternaBrain: automated RNA design through move sets and strategies from an Internet-scale RNA videogame. *PLoS Comput. Biol.*, **15**, e1007059.
34. Avihoo,A., Churkin,A. and Barash,D. (2011) RNAexinv: an extended inverse RNA folding from shape and physical attributes to sequences. *BMC Bioinf.*, **12**, 319.
35. Zhang,P., Tao,L., Zeng,X., Qin,C., Chen,S., Zhu,F., Li,Z., Jiang,Y., Chen,W. and Chen,Y.Z. (2017) A protein network descriptor server and its use in studying protein, disease, metabolic and drug targeted networks. *Brief. Bioinform.*, **18**, 1057–1070.
36. Wen,J., Liu,Y., Shi,Y., Huang,H., Deng,B. and Xiao,X. (2019) A classification model for lncRNA and mRNA based on k-mers and a convolutional neural network. *BMC Bioinf.*, **20**, 469.
37. Zuo,Y., Chang,Y., Huang,S., Zheng,L., Yang,L. and Cao,G. (2019) iDEF-PseRAAC: identifying the defensin peptide by using reduced amino acid composition descriptor. *Evol. Bioinform. Online*, **15**, 1176934319867088.
38. Chen,Z., Zhao,P., Li,F., Marquez-Lago,T.T., Leier,A., Revote,J., Zhu,Y., Powell,D.R., Akutsu,T., Webb,G.I. *et al.* (2020) iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief. Bioinform.*, **21**, 1047–1057.
39. Chen,Z., Zhao,P., Li,C., Li,F., Xiang,D., Chen,Y.Z., Akutsu,T., Daly,R.J., Webb,G.I., Zhao,Q. *et al.* (2021) iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Res.*, **49**, e60.
40. Zhang,T., Zhang,H., Chen,K., Ruan,J., Shen,S. and Kurgan,L. (2010) Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr. Protein Pept. Sci.*, **11**, 609–628.
41. Tetko,I.V., Tanchuk,V.Y., Kasheva,T.N. and Villa,A.E. (2001) Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.*, **41**, 1488–1493.
42. Klein,C.T., Kaiser,D. and Ecker,G. (2004) Topological distance based 3D descriptors for use in QSAR and diversity analysis. *J. Chem. Inf. Comput. Sci.*, **44**, 200–209.
43. Liang,X., Zhang,P., Li,J., Fu,Y., Qu,L., Chen,Y. and Chen,Z. (2019) Learning important features from multi-view data to predict drug side effects. *J Cheminform*, **11**, 79.
44. Townshend,R.J.L., Eismann,S., Watkins,A.M., Rangan,R., Karelina,M., Das,R. and Dror,R.O. (2021) Geometric deep learning of RNA structure. *Science*, **373**, 1047–1051.
45. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
46. Zhao,T., Hu,Y., Peng,J. and Cheng,L. (2020) DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics*, **36**, 4466–4472.
47. Yi,H.C., You,Z.H., Huang,D.S., Li,X., Jiang,T.H. and Li,L.P. (2018) A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information. *Mol. Ther. Nucleic Acids*, **11**, 337–344.
48. Chuai,G., Ma,H., Yan,J., Chen,M., Hong,N., Xue,D., Zhou,C., Zhu,C., Chen,K., Duan,B. *et al.* (2018) DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol.*, **19**, 80.
49. Lee,J.A. and Verleysen,M. (2010) In: *IEEE World Congress on Computational Intelligence (WCCI 2010)*. Barcelona, SPAIN Vol. **1**, p.1.
50. Pan,X., Fan,Y.X., Yan,J. and Shen,H.B. (2016) IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked

autoencoder for accurate computational prediction. *Bmc Genomics (Electronic Resource)*, **17**, 582.
51. Xu,L., Xu,Y., Xue,T., Zhang,X. and Li,J. (2021) AdImpute: an imputation method for single-cell RNA-seq data based on semi-supervised autoencoders. *Front. Genet.*, **12**, 739677.
52. Liu,Z.P., Wu,L.Y., Wang,Y., Zhang,X.S. and Chen,L. (2010) Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics*, **26**, 1616–1622.
53. Cheng,C.W., Su,E.C., Hwang,J.K., Sung,T.Y. and Hsu,W.L. (2008) Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinf.*, **9**, S6.
54. Deng,L., Sui,Y. and Zhang,J. (2019) XGBPRH: prediction of binding hot spots at protein-RNA interfaces utilizing extreme gradient boosting. *Genes (Basel)*, **10**, 1.
55. Pan,X., Rijnbeek,P., Yan,J. and Shen,H.B. (2018) Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics (Electronic Resource)*, **19**, 511.
56. Wang,L., You,Z.H., Huang,D.S. and Zhou,F. (2020) Combining High Speed ELM Learning with a Deep Convolutional Neural Network Feature Encoding for Predicting Protein-RNA Interactions. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **17**, 972–980.
57. Amin,N., McGrath,A. and Chen,Y.-P.P. (2019) Evaluation of deep learning in non-coding RNA classification. *Nat. Mach. Intell.*, **1**, 246–256.
58. Wang,Q., Wei,L., Guan,X., Wu,Y., Zou,Q. and Ji,Z. (2014) Briefing in family characteristics of microRNAs and their applications in cancer research. *Biochim. Biophys. Acta*, **1844**, 191–197.
59. Tong,X. and Liu,S. (2019) CPPred: coding potential prediction based on the global description of RNA sequence. *Nucleic Acids Res.*, **47**, e43.
60. Hill,S.T., Kuintzle,R., Teegarden,A., Merrill,E. 3rd, Danaee,P. and Hendrix,D.A. (2018) A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. *Nucleic Acids Res.*, **46**, 8105–8113.
61. Zou,Q., Mao,Y., Hu,L., Wu,Y. and Ji,Z. (2014) miRClassify: an advanced web server for miRNA family classification and annotation. *Comput. Biol. Med.*, **45**, 157–160.
62. Wucher,V., Legeai,F., Hedan,B., Rizk,G., Lagoutte,L., Leeb,T., Jagannathan,V., Cadieu,E., David,A., Lohi,H. *et al.* (2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.*, **45**, e57.
63. Camargo,A.P., Sourkov,V., Pereira,G.A.G. and Carazzolle,M.F. (2020) RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences. *NAR Genom. Bioinform.*, **2**, lqz024.
64. Ramos,T.A.R., Galindo,N.R.O., Arias-Carrasco,R., da Silva,C.F., Maracaja-Coutinho,V. and do Rego,T.G. (2021) RNAmining: a machine learning stand-alone and web server tool for RNA coding potential prediction. *F1000Res*, **10**, 323.
65. Morlando,M., Ballarino,M., Fatica,A. and Bozzoni,I. (2014) The role of long noncoding RNAs in the epigenetic control of gene expression. *ChemMedChem*, **9**, 505–510.
66. Pan,X. and Shen,H.B. (2018) Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics*, **34**, 3427–3436.
67. Zhu,Y.P., Bian,X.J., Ye,D.W., Yao,X.D., Zhang,S.L., Dai,B., Zhang,H.L. and Shen,Y.J. (2014) Long noncoding RNA expression signatures of bladder cancer revealed by microarray. *Oncol. Lett.*, **7**, 1197–1202.
68. Huang,X., Shi,Y., Yan,J., Qu,W., Li,X. and Tan,J. (2022) LPI-CSFFR: combining serial fusion with feature reuse for predicting LncRNA-protein interactions. *Comput. Biol. Chem.*, **99**, 107718.
69. Tara,C., Joeyta,B. and Lior,P. (2021) The specious art of single-cell genomics. bioRxiv doi: https://doi.org/10.1101/2021.08.25.457696, 22 December 2022, preprint: not peer reviewed.
70. 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.
71. Cheng,Z., Huang,K., Wang,Y., Liu,H., Guan,J. and Zhou,S. (2017) Selecting high-quality negative samples for effectively predicting protein-RNA interactions. *BMC Syst. Biol.*, **11**, 9.
72. Yi,W., Li,J., Zhu,X., Wang,X., Fan,L., Sun,W., Liao,L., Zhang,J., Li,X., Ye,J. *et al.* (2020) CRISPR-assisted detection of RNA-protein interactions in living cells. *Nat. Methods*, **17**, 685–688.
73. Liu,S., Zhao,X., Zhang,G., Li,W., Liu,F., Liu,S. and Zhang,W. (2019) PredLnc-GFStack: a global sequence feature based on a stacked ensemble learning method for predicting lncRNAs from transcripts. *Genes (Basel)*, **10**, 1.
74. Clamp,M., Fry,B., Kamal,M., Xie,X., Cuff,J., Lin,M.F., Kellis,M., Lindblad-Toh,K. and Lander,E.S. (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 19428–19433.
75. Ouyang,Z., Zhu,H., Wang,J. and She,Z.S. (2004) Multivariate entropy distance method for prokaryotic gene identification. *J. Bioinform. Comput. Biol.*, **2**, 353–373.
76. Fickett,J.W. and Tung,C.S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **20**, 6441–6450.
77. Kudla,G., Lipinski,L., Caffin,F., Helwak,A. and Zylicz,M. (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.*, **4**, e180.
78. Myers,E.W., Sutton,G.G., Delcher,A.L., Dew,I.M., Fasulo,D.P., Flanigan,M.J., Kravitz,S.A., Mobarry,C.M., Reinert,K.H., Remington,K.A. *et al.* (2000) A whole-genome assembly of Drosophila. *Science*, **287**, 2196–2204.
79. Han,L.Y., Cai,C.Z., Ji,Z.L., Cao,Z.W., Cui,J. and Chen,Y.Z. (2004) Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res.*, **32**, 6437–6444.
80. Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **43**, 246–255.
81. Moran,P.A. (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.
82. Nair,A.S. and Sreenadhan,S.P. (2006) A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation*, **1**, 197–202.
83. Han,S., Liang,Y., Ma,Q., Xu,Y., Zhang,Y., Du,W., Wang,C. and Li,Y. (2019) LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. *Brief. Bioinform.*, **20**, 2009–2027.