Interpretation of RNA Universe and Coding Potential Using *IntRNA*

Yunxia Wang, Minjie Mou, Shijie Huang, Wei Zhang, Ziqi Pan, Jing Tang, Yihao Wang, Qingxia Yang, and Feng Zhu**

The interpretation of *RNA universe* and coding potential are long-standing issues in modern RNA studies, and three crucial questions remain unanswered: a) how to detect and interpret the coding potential of RNA, b) how to annotate the sophisticated taxonomy of the sncRNAs, and c) how to successfully distinguish between circular and linear lncRNAs. In this study, a multi-channel deep learning framework, *IntRNA*, is thus constructed to interpret *RNA universe* and coding potential. First, a large number of RNA encoding features are proposed, which dramatically enlarged the available feature space. Second, a method realizing image-like representation of RNA sequences is developed to describe the intrinsic correlation among the encoding features generated above. Third, a dual-path model is constructed, which consistently performed the best among existing methods in various benchmarks. *IntRNA*'s interpretability is also validated by analysis, and all source codes are accessible at: https://idrblab.org/intrna/ and https://github.com/idrblab/intrna.

1. Introduction

The annotation/classification of RNAs has emerged as a pivotal scientific challenge in advancing our understanding of sophisticated biological process and discovery of new therapeutic strategy.[1] A comprehensive classification system of RNA has been proposed,[2] which is frequently adopted by a variety of recent studies.[3–5] In this system, an *RNA universe* is explicitly described (as shown in **Figure 1**), which divides RNAs into messenger (mRNA) and non-coding (ncRNA) ones based on their coding potentials.[6] The ncRNAs are further categorized into small non-coding (sncRNA, which provides sophisticated classification taxonomy) and long non-coding (lncRNA, which can be further divided into linear and circular ones) ones based on their sequence length.[7] For instance, the lncRNA HOTAIR identified through precise annotation drives cancer metastasis by modulating chromatin states,[8] and deletion of the lncRNA Maenli locus on chromosome 2 causes severe congenital limb malformations.[9]

However, it remains extremely challenging to provide comprehensive and clear interpretation of the *RNA universe*.[10] Particularly, three key problems (shown in Figure 1) still need to be answered: a) how to detect and interpret the coding potential of RNAs,[11] b) how to annotate the sophisticated taxonomy of those sncRNAs,[12] and c) how to successfully distinguish between circular and linear lncRNAs.[13] Till now, a number of methods (derived from the classical machine learning strategy) have been constructed to achieve fast and accurate prediction of RNA coding potential based on sequence-intrinsic (such as CPPred,[14] CPAT,[15] COME,[16] and mRNN[17]), nucleotide counts-based (such as

Y. Wang, M. Mou, S. Huang, W. Zhang, Z. Pan, F. Zhu
College of Pharmaceutical Sciences
The Second Affiliated Hospital
Zhejiang University School of Medicine
State Key Laboratory of Advanced Drug Delivery and Release Systems
Zhejiang University
Hangzhou 310058, China
E-mail: zhufeng@zju.edu.cn

Y. Wang, F. Zhu
Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University
Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare
Hangzhou 330110, China

J. Tang
College of Basic Medicine
Chongqing Medical University
Chongqing 400016, China

Y. Wang
The High School Affiliated to Renmin University of China
No. 37 Zhongguancun Street. Haidian District, Beijing 100080, China

Q. Yang
Zhejiang Provincial Key Laboratory of Precision Diagnosis and Therapy for Major Gynecological Diseases
Women's Hospital
Zhejiang University School of Medicine
Hangzhou 310058, China
E-mail: yangqx@zju.edu.cn

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/advs.202509518

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/advs.202509518

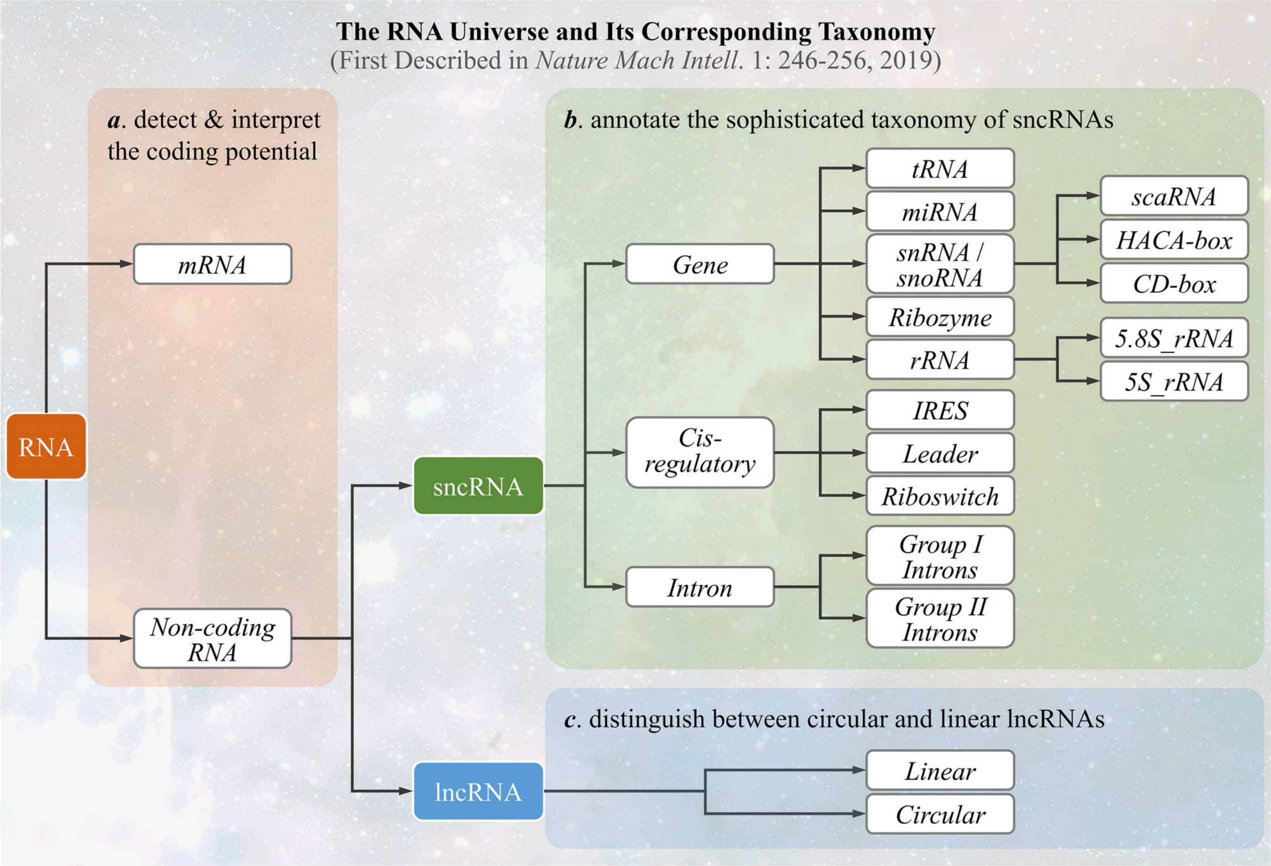

The RNA Universe and Its Corresponding Taxonomy
(First Described in Nature Mach Intell. 1: 246-256, 2019)
a. detect & interpret the coding potential
b. annotate the sophisticated taxonomy of sncRNAs
c. distinguish between circular and linear lncRNAs
RNA
mRNA
Non-coding RNA
sncRNA
lncRNA
Gene
Cis-regulatory
Intron
tRNA
miRNA
snRNA / snoRNA
Ribozyme
rRNA
IRES
Leader
Riboswitch
Group I Introns
Group II Introns
scaRNA
HACA-box
CD-box
5.8S_rRNA
5S_rRNA
Linear
Circular

Figure 1. The *RNA universe* together with its corresponding taxonomy, which were first proposed by previous study (*Nature Mach Intell*. 1: 246-256, 2019). To have a comprehensive understanding of the entire "*universe*", three key problems remained to be answered: a) how to detect and interpret the coding potential of RNAs; b) how to annotate the sophisticated taxonomy of small non-coding RNAs (sncRNAs); and c) how to successfully distinguish between circular and linear lncRNAs. Such problems (colored by light red, light green & light blue backgrounds) provided the key scientific questions that were planned to be answered in this research.

PLEK,[18] RNAmining,[19] and DeepCPP[20]) and physicochemical properties-directed (such as CPC2,[21] CPE-SLDI,[22] and NCResNet[23]) encoding descriptors of the ribonucleic acids.

Moreover, some tools are available for depicting the sncRNA taxonomy (such as nRC,[24] ncRNA-deep,[25] GraPPLE,[26] and EDeN[27]) based on various characteristics (including secondary structure, atoms-based graph, etc.), and some others are constructed to discriminate circular lncRNAs from linear ones (such as circDeep,[28] CirRNAPL,[29] PredcircRNA,[30] and StackCirRNAPred[31]) through fusing various types of sequence-intrinsic descriptor. Meanwhile, researchers have leveraged the language models to elevate the ability in RNA classification (such as RNA-MSM,[32] RNAErnie,[33] and RNABERT[34]), which offer additional choices on the interpretation of *RNA universe*.

As reported, the relation among the nucleotides of "long-distance interval" can largely determine RNAs' function annotation by shaping their structure,[35] but none of the methods discussed above has taken such critical features into consideration when encoding the *RNA universe*. Additionally, the performances of the latest Artificial Intelligence (AI) techniques are found susceptible to the order of studied features in their representation,[36] which asks for the construction of new strategy for capturing the innate relationship among RNA encoding features. Finally, there is a significant lack of interpretability in the available RNA annotation tools,[37,38] which hinders the provision of new biological insight for modern RNA research. In other words, there remains substantial rooms for improvement in the classification of *RNA universe* and the interpretation of RNA annotations (especially, in the explanation of the mechanism underlining RNA's coding potential).

In this study, a multi-channel deep learning framework, entitled *IntRNA*, was therefore developed to enable the annotation of *RNA universe* and interpretation of coding potentials with extensively elevated model performance. First, a large number of new encoding features (especially the ones describing the nucleotides of long-distance interval) were proposed, which dramatically enlarged (to over four times) the existing feature space that encodes RNAs. Second, a method that realized image-like representation of RNA sequences was constructed to describe the intrinsic correlation among the massive amount of those encoding features generated above. Third, to further enhance the interpretability of both *RNA universe* and coding potentials, a dual-path multi-channel model was developed, which was found consistently well-performing when comparing with the existing methods in a series of benchmark studies. Moreover, the interpretability of *IntRNA* was validated by a real-world study, which identified key

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202509518 by Zhejiang University, Wiley Online Library on [04/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202509518 by Zhejiang University, Wiley Online Library on [04/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

structural features determining RNA coding potential. All in all, our *IntRNA* was freely accessible at: https://idrblab.org/intrna/, and all the source codes were readily downloadable from GitHub at: https://github.com/idrblab/intrna.

2. Results and Discussion

2.1. The Framework Proposed in This Research and Its Characteristics

The *IntRNA* was constructed here to enable the annotation of *RNA universe* and interpretation of coding potential with enhanced performance. To achieve such goal, the most comprehensive list of encoding features (especially the ones describing the nucleotides of long-distance interval, as provided in **Figure 2**) among the existing methods were first produced to extensively expand (to over four times, as shown on the right side of **Figure 3**a) the feature space of RNA representation. Second, a novel method that realized the image-like representation of RNA sequences (illustrated in Figure 3b) was developed to provide the intrinsic correlations among encoding features. Third, a new model was further constructed based on a dual-path multi-channel deep learning algorithm (shown in Figure 3c). Furthermore, the results of ablation study on *IntRNA* model were given in Figure S1 (Supporting Information). Compared with the performances of the latest *IntRNA* model (which were measured by Matthews correlation coefficient), that of the ablated models (the one deprived of *G-features*, the one deprived of *D-features*, and the one deprived of *RNAImage*) decreased by 25.6%, 4.9%, 18.3%, respectively. All in all, such results highlighted the critical contribution of both the *G-features* and the *RNAImage* strategy proposed in this study.

2.2. Performance Comparison in Solving the Problems of *RNA Universe*

As provided in Figure 1, the classification system required the solving of three key problems: a) how to detect and interpret RNA coding potential, b) how to annotate the sophisticated taxonomy of sncRNAs, and c) how to successfully distinguish between circular and linear lncRNAs. Based on our literature review, *CPPred* was constructed for classifying mRNA and ncRNA,[14] *nRC* was made available to classify sncRNAs,[24] and *circDeep* was constructed for differentiating between circular and linear lncRNAs.[28] Therefore, the performances of those three popular methods were compared with that of the newly-constructed *IntRNA*. As provided in Table S1 (Supporting Information), the performances of the three existing tools (*CPPred*, *circDeep* and *nRC*) alongside *IntRNA* were compared based on five metrics, and the results of this comparison were described in **Figure 4**a. As illustrated, the bar plots of accuracy (ACC) and Matthews correlation coefficient (MCC) were utilized to give the performances of four methods. As a result, *IntRNA* consistently outperformed three existing tools in all key problems. Specifically, as observed in Figure 4a, *IntRNA* performed as well as *IntRNA* in the "binary classification between mRNAs and ncRNAs" (this is the problem what *CPPred* was constructed for). Similarly, *IntRNA* performed as well as the remaining two existing models on their intended task. While CPPred, circDeep, and nRC were originally developed for addressing specific RNA classification tasks, the comparative analysis is intended to demonstrate general applicability of the *IntRNA*.

2.3. Performance Comparison in Mining Cross-Species Coding Potential

The mining of RNA coding potential has been considered as a long-standing challenge in modern RNA research,[39–41] and several methods have therefore been constructed (the representative tools under this category were offered in Introduction section). However, the mining of cross-species coding potentials remained a huge challenge,[42] which asked for the construction of new powerful tool to address this critical issue. In this study, the performances of *IntRNA* and six existing tools (currently reproducible with source code provided) were thus compared based on the benchmark dataset provided in Table S2 (Supporting Information). As demonstrated on the left side of Figure 4b and Table S3 (Supporting Information), the evolutionary divergence between human (*homo sapiens*) and four typical species increased progressively, and the data used in this research to assess seven methods' performances in the detection and interpretation of RNAs' coding potential had a total of 180233 RNAs (including 112726 mRNAs & 67507 ncRNAs) from five different species: *Homo sapiens*, *Mus musculus*, *Danio rerio*, *Drosophila melanogaster* & *Saccharomyces cerevisiae*. Particularly, the data from *Homo sapiens* were split by the original publication[14] into the training (comprising 33360 mRNAs and 24163 ncRNAs) and the test (comprising 8557 mRNAs and 8241 ncRNAs) datasets. During the training process, the strategy of fivefold cross-validation was applied, and the test dataset was further enriched by integrating the data from the remaining four species.

As shown on the right side of Figure 4b, the performances of seven methods (shown in Table S4, Supporting Information) were systematically evaluated using the RNAs from four typical species, and *IntRNA* was identified consistently best-performing for all species. Particularly, with the increase of evolutionary divergence, the *IntRNA* provided much better performance enhancements comparing with other six methods. Particularly, a detailed comparison among the performances of seven methods was offered in Table S5 (Supporting Information). As described using two important metrics (AUC & MCC), the performance of *IntRNA* consistently outperformed the existing tools (highlighted using BOLD fonts), which underscored the good performance of *IntRNA* in mining the cross-species RNA coding potential. Besides the animal datasets, *IntRNA* also achieved good performance on a plant dataset, such as Arabidopsis thaliana (ACC = 0.9101, MCC = 0.8324, AUC = 0.9782), further demonstrating its broad cross-species applicability. This comparison highlights *IntRNA*'s generalization ability and is not intended to critique tools optimized for species-specific retraining.

2.4. Discovery of the RNA Coding Potential Based on Feature Ranking

The performance of *IntRNA* in mining coding potential could also be reflected by the left side of **Figure 5**a. As shown, the MCC

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

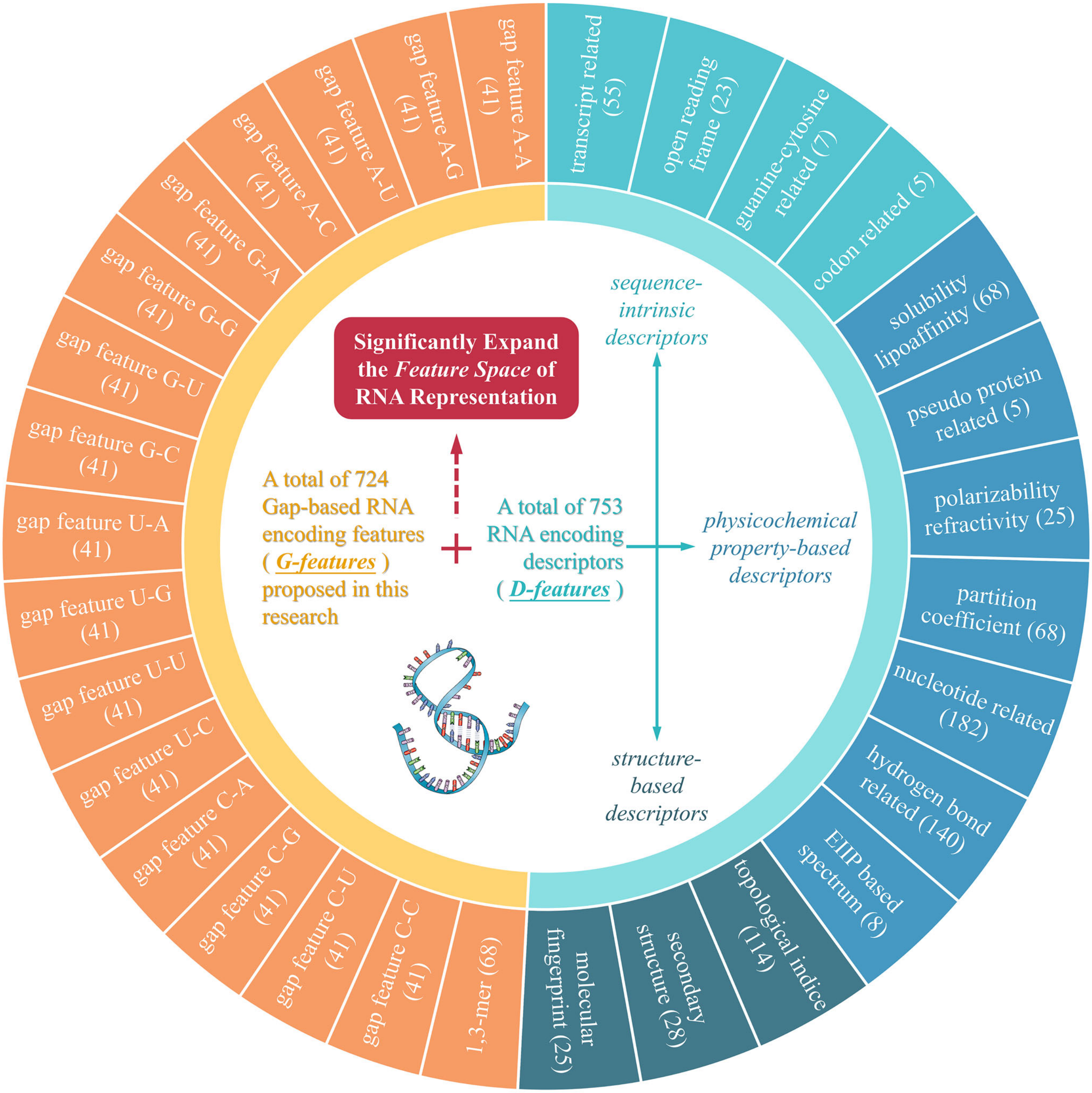

gap feature A-A (41)
gap feature A-G (41)
gap feature A-U (41)
gap feature A-C (41)
gap feature G-A (41)
gap feature G-G (41)
gap feature G-U (41)
gap feature G-C (41)
gap feature U-A (41)
gap feature U-G (41)
gap feature U-U (41)
gap feature U-C (41)
gap feature C-A (41)
gap feature C-G (41)
gap feature C-U (41)
gap feature C-C (41)
1,3-mer (69)
transcript related (55)
open reading frame (23)
guanine-cytosine related (7)
codon related (5)
solubility lipoaffinity (68)
pseudo protein related (5)
polarizability refractivity (25)
partition coefficient (68)
nucleotide related (182)
hydrogen bond related (140)
EIIP based spectrum (8)
topological indice (114)
secondary structure (28)
molecular fingerprint (25)
Significantly Expand the Feature Space of RNA Representation
A total of 724 Gap-based RNA encoding features (G-features) proposed in this research
A total of 753 RNA encoding descriptors (D-features)
sequence-intrinsic descriptors
physicochemical property-based descriptors
structure-based descriptors

Figure 2. The significantly expanded feature space for RNA representation. There were two types of encoding features: the gap-based RNA encoding features (*G-features*, a total of 724 features fully proposed in this study) and the RNA encoding descriptors (*D-features*, a total of 753 features, 439 (58%) out of which were the newly-proposed ones). For both types, their corresponding feature subgroups (17 groups and 14 groups for *G-features* and *D-features*, respectively) were provided in the outer-most leaves, and the 14 subgroups of *D-features* were further summarized to three classes (sequence-intrinsic, structure-based and physicochemical property-based). The number in each bracket indicated the total number of features in each feature subgroup.

value of *IntRNA* was significantly higher than that of other tools, which reached 94.2% and was higher than that (ranging from 68.0% to 93.2%) of available tools. To test the ability of *IntRNA* in capturing the critical features that endorsed RNA coding potential, a prediction model was first trained using the human coding potential dataset, and the importance of each feature was measured using permutation algorithm (as described in the Methods section), which resulted in the ranking of all features based on their importance. Second, the adjusted rand scores (ARSs) were calculated for measuring the ability of the top-*N* ranked features ($N = 1{\sim}1477$) to discover the coding potential.[43] The higher the ARS value was, the better ability of the top-*N* ranked features was

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202509518 by Zhejiang University, Wiley Online Library on [04/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

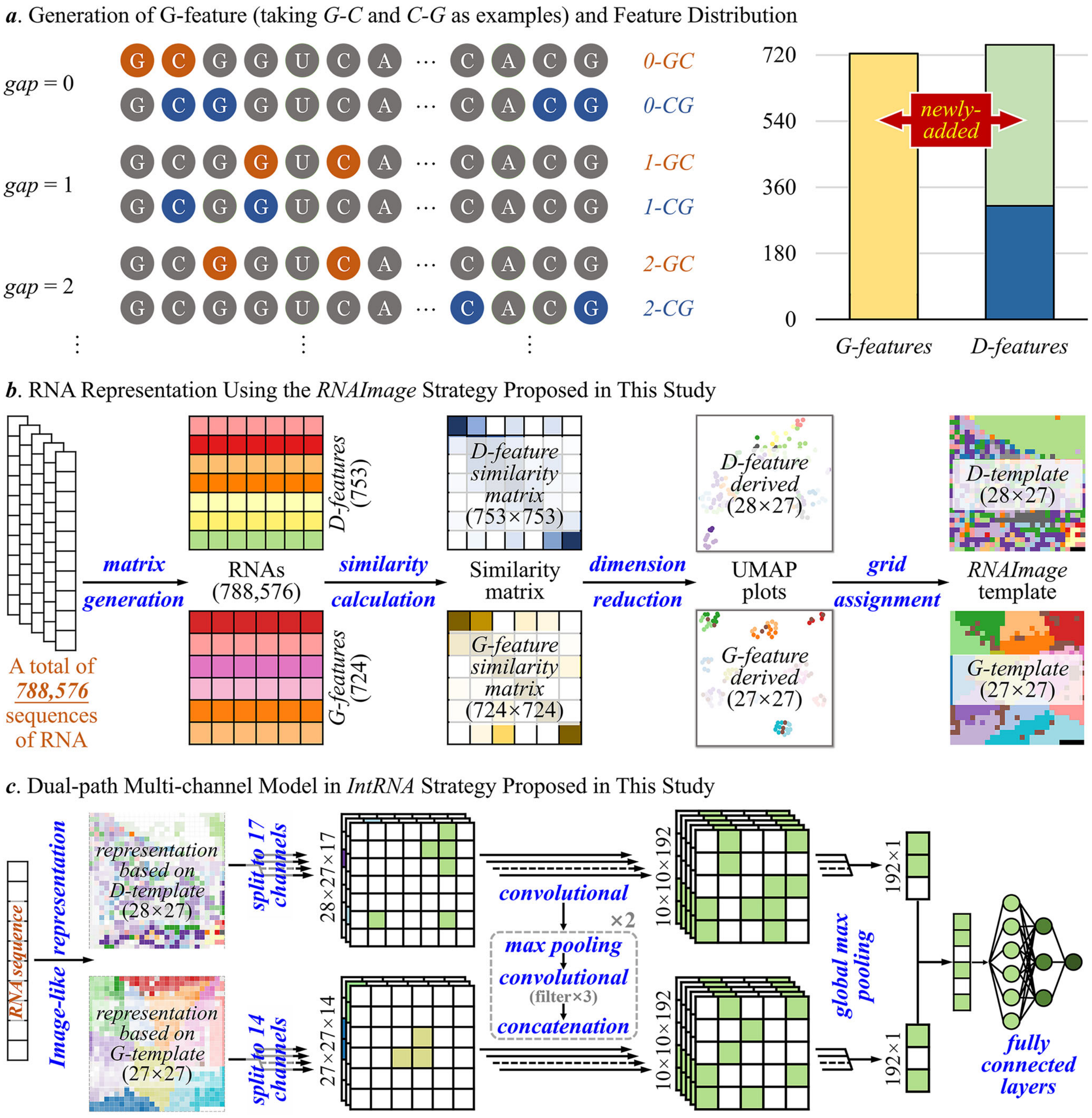

a. Generation of G-feature (taking G-C and C-G as examples) and Feature Distribution
gap = 0
G C G G U C A ··· C A C G
0-GC
0-CG
gap = 1
1-GC
1-CG
gap = 2
2-GC
2-CG
720
540
360
180
0
newly-added
G-features
D-features
b. RNA Representation Using the RNAImage Strategy Proposed in This Study
A total of 788,576 sequences of RNA
matrix generation
RNAs (788,576)
D-features (753)
G-features (724)
similarity calculation
Similarity matrix
D-feature similarity matrix (753×753)
G-feature similarity matrix (724×724)
dimension reduction
UMAP plots
D-feature derived (28×27)
G-feature derived (27×27)
grid assignment
RNAImage template
D-template (28×27)
G-template (27×27)
c. Dual-path Multi-channel Model in IntRNA Strategy Proposed in This Study
RNA sequence
Image-like representation
representation based on D-template (28×27)
representation based on G-template (27×27)
split to 17 channels
split to 14 channels
28×27×17
27×27×14
convolutional
×2
max pooling
convolutional (filter×3)
concatenation
10×10×192
10×10×192
global max pooling
192×1
192×1
fully connected layers

Figure 3. The framework of the *IntRNA* strategy proposed in this research. a) the generation of *G-feature* and feature distribution. New features for encoding the long-distance relations between nucleobases were proposed. Taking *G-C* and *C-G* as examples, the *G-features* were generated based on the gapped distances (from 0 to 40, described on the left side). A total of 724 *G-features* and 439 *D-features* proposed in this research, which were further enriched by 314 previously-reported *D-features*, were adopted to significantly expand the feature space of RNA representation (provided on the right side). b) RNA representation using *RNAImage* method proposed in this study. Two template images (*D-template* and *G-template*) were generated using a total of 788576 RNA sequences collected from two established databases (*RefSeq* and *ENCODE*) based on four sequential processes (from matrix generation, to similarity calculation, then to dimension reduction, and finally to grid assignment). c) *IntRNA* strategy proposed in this study for interpreting *RNA universe*. Each RNA sequence was represented as two images of different matrix dimensions (28 × 27 & 27 × 27) based on the newly-generated template images, and the image-like representations of all RNAs in a studied key problem were adopted to construct a dual-path multi-channel model, which was explicitly described above and in the section of Methods.

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202509518 by Zhejiang University, Wiley Online Library on [04/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

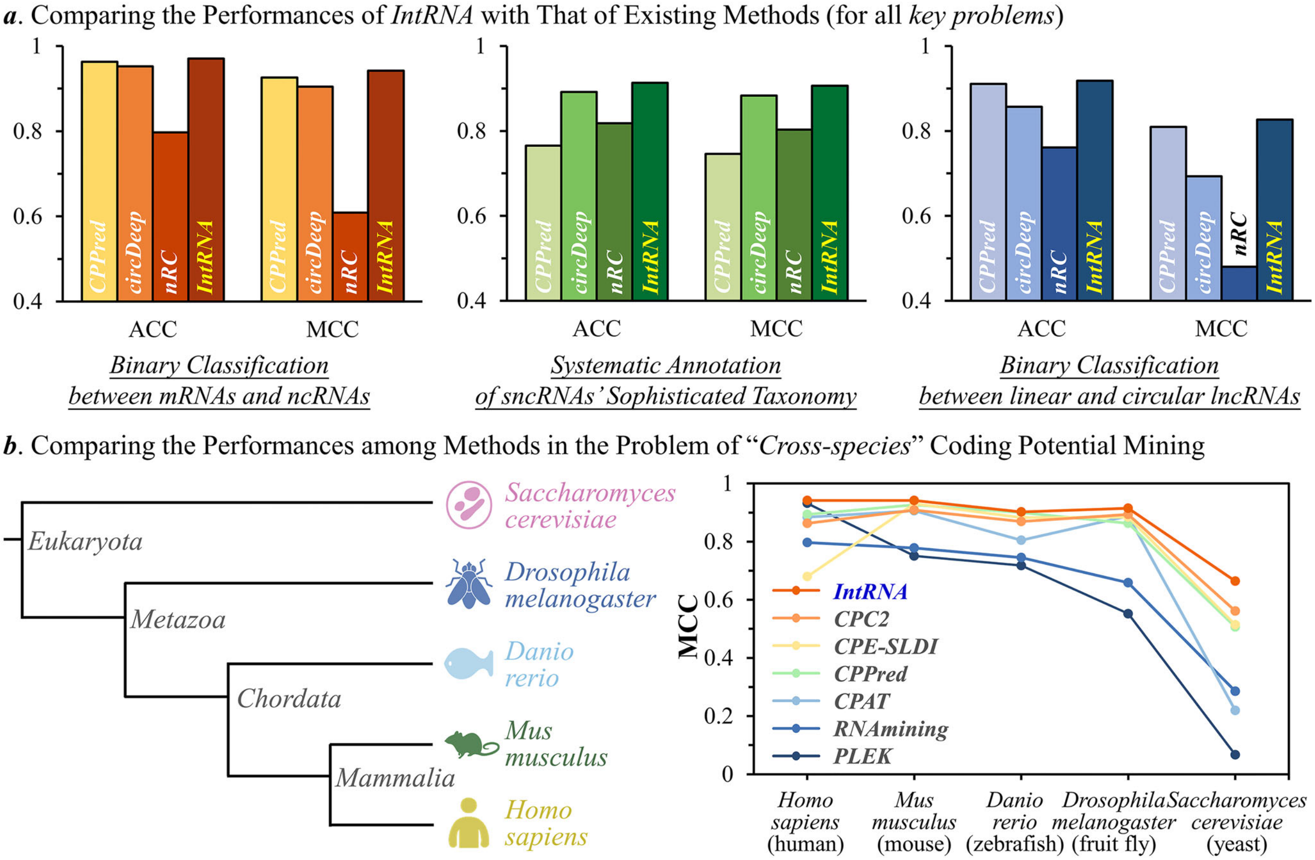


Figure 4. Performance comparison with existing methods and in the problem of cross-species coding potential minding. a) performance comparison among *IntRNA* and three existing methods (*CPPred*, *circDeep* and *nRC*). The performances of all four methods when solving three key problems (as provided in Figure 1) were assessed, and *IntRNA* was identified consistently best-performing among four methods. b) comparing the performances of *IntRNA* and six existing methods that were designed to mine the coding potential of RNAs. The evolutionary divergence between human (*homo sapiens*) and other four typical species (described on the left side) increased progressively. The performances of all seven methods were systematically assessed using the RNAs from four typical species (by training their models using the human RNAs, described on the right side), and *IntRNA* was identified consistently best-performing in all species. Particularly, with the increase of evolutionary divergence, the *IntRNA* demonstrated much better performance enhancement comparing with other six methods.

in identifying RNA coding potential. As provided on the right side of Figure 5a, the top-21 features (shown in Table S6, Supporting Information) were found to provide the highest ARS value, which highlighted the great contribution of these features. Third, the capacity of these features in discriminating mRNA from ncRNA was further demonstrated in Figure 5b. As described, the abilities of the top-*N* ($N = 5$, 21 and 50) ranked features in discovering coding potential were evaluated by hierarchical clustering, and mRNAs and ncRNAs were colored in GREEN and YELLOW, respectively. As a result, the classification accuracies of top-5, top-21, and top-50 features equaled to 67.7%, 90.0% and 66.9%, respectively, which further emphasized the importance of the identified top-21 features.

The role of identified features in coding potential prediction was also investigated. As shown in Figure S2a (Supporting Information), there were clear correlations between the features from the training dataset and that from the test one. Particularly, the Pearson's correlation coefficients of *D-feature* and *G-feature* reached 0.993 and 0.968, respectively. Moreover, an in-depth analysis on the best-ranking features (based on their importance score $IS > 0.8$, shown in Methods) was provided in Figure S2b (Supporting Information) for *D-feature* (on the left side) and *G-feature* (on the right side). If the *IS* was larger than 0.8, the corresponding feature was then highlighted. For *D-features*, "open reading frame", "EIIP based spectrum", "hydrogen bond" and some other groups were identified, and some of the identified groups had been reported to be closely related to RNA coding potential. For example, the open reading frame was frequently adopted to enable sequence-based discovery of mRNAs;[44] the EIIP based spectrum had been applied to differentiate mRNAs from ncRNAs;[45] the hydrogen bond was found helpful in facilitating coding potential estimation.[46] For *G-features* (right side of Su Figure S2b, Supporting Information), the numbers in squares indicated the amount of bases between the studied base pair, and the numbers of long-distance (number≥20) and short-distance (number<20) *G-features* were colored by YELLOW and BLACK, respectively. It was clear that the amount of long-distance features (with yellow-colored number) was significantly larger than that of the short-distance ones, which indicated the great contribution of "long-distance contacts" between base pairs. Similar conclusion

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202509518 by Zhejiang University, Wiley Online Library on [04/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

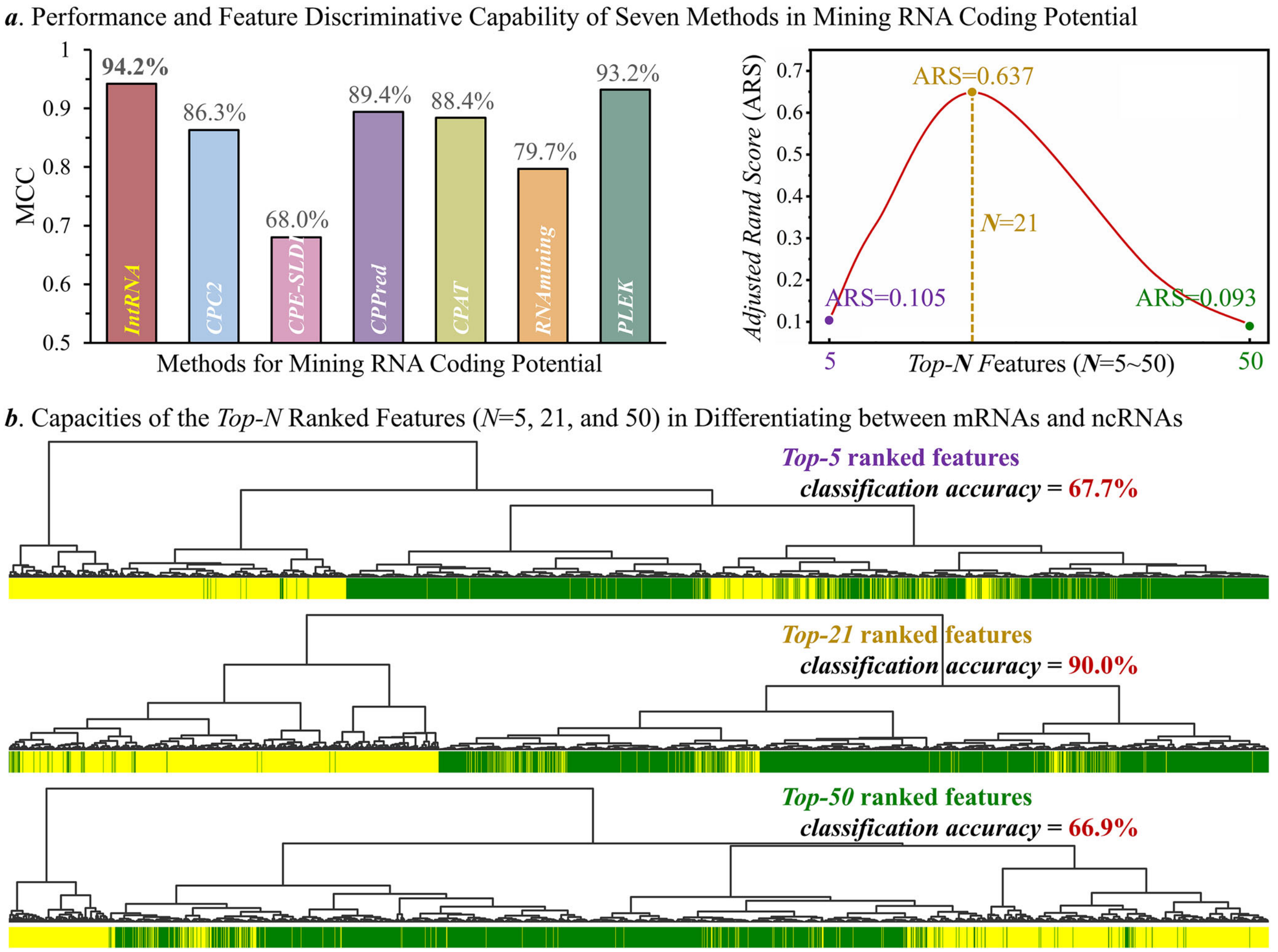


Figure 5. Performance comparison among seven methods designed to mine the coding potential of RNAs (by training their models using a benchmark dataset collected from *CPPred*, which contained 74321 human RNAs, and the way adopted here to split training and test datasets was exactly the same as that described in the original publication of *CPPred*). a) performance and feature discriminative capability of each method for mining RNA coding potential. Particularly, the values of matthews correlation coefficient (MCC) were used to measure the method performance (described on the left side), and the values of adjusted rand score (ARS) were utilized to measure the capacity of the *top-N* ranked features (N = 5~50) in detecting RNA coding potential (shown on the right side), which identified that the *top-21* ranked features performed the best. b) the capacities of the *top-N* ranked features (N = 5, 21, and 50) in differentiating between mRNAs and ncRNAs. The resulting classification accuracies for *top-5*, *top-21*, and *top-50* features were 67.7%, 90.0%, and 66.9%, respectively, which indicated the good differential capacity of those *top-21* ranked features that were identified in this analysis.

had been identified by previous work, which emphasized the importance of certain long-distance segments in the determination of RNA function.[47]

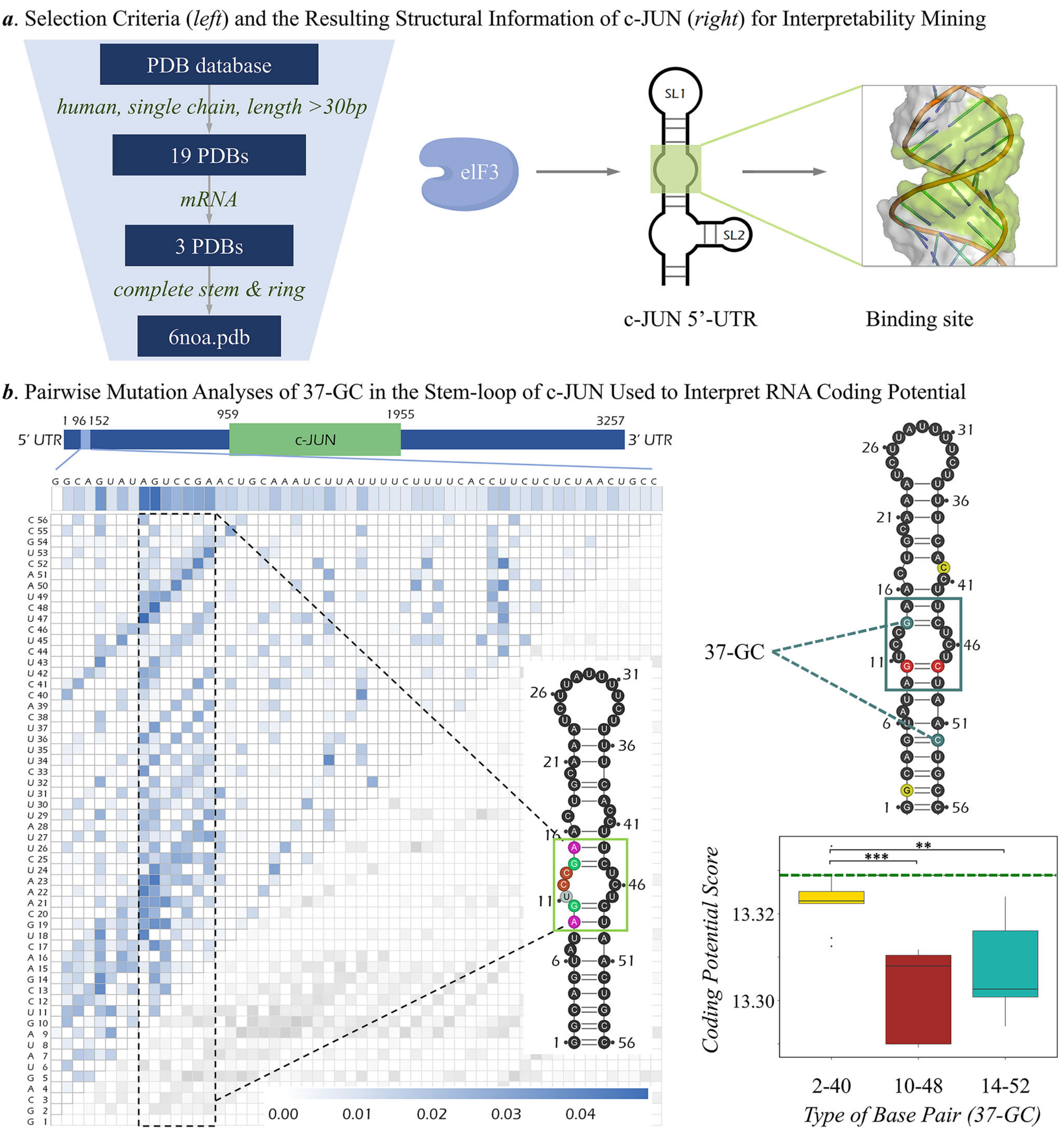
To demonstrate the generality of *IntRNA*'s interpretability, the feature importance analysis was extended to the other two key problems: how to annotate the sophisticated taxonomy of sncRNAs (shown in Figures S3 and S4, Supporting Information), and how to successfully distinguish between circular and linear lncRNAs (presented in Figure S5, Supporting Information) using the same permutation-based ranking approach. In the sncRNAs annotation, ribozyme RNAs showed enrichment in topological indices and partition coefficients, reflecting their catalytic folding and solvent-sensitive architecture.[48,49] Introns were marked by motif-level correlation features and entropy-based metrics, consistent with splicing signals and non-coding character.[50–52] To distinguish circular and linear lncRNA, top features included hydrogen bond donors/acceptors and solubility-related descriptors, which are consistent with the compact, base-paired, and stable nature of circular RNAs.[53–55] This comprehensive analysis highlights *IntRNA*'s ability to extract meaningful, task-specific features across RNA classes, reinforcing its utility not only as a high-performance classifier but also as a tool for uncovering mechanistic insights in RNA biology.

2.5. In-Depth Interpretation of Coding Potential by Molecular Example

To assess the interpretability of *IntRNA*, specific RNA structure was adopted to validate the RNA coding potential prediction. First, RNA structures were filtered from PDB database by following the criteria provided on the left side of **Figure 6a**, which

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202509518 by Zhejiang University, Wiley Online Library on [04/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



a. Selection Criteria (*left*) and the Resulting Structural Information of c-JUN (*right*) for Interpretability Mining

b. Pairwise Mutation Analyses of 37-GC in the Stem-loop of c-JUN Used to Interpret RNA Coding Potential

Figure 6. Interpretation of RNA coding potential by taking c-JUN as an example. a) selection criteria and the resulting structural information of c-JUN. The stem-loop highlighted in GREEN of c-JUN 5′-UTR mRNA was the key structure recognized by eIF3. b) pairwise mutation analyses of 37-GC in the stem-loop of c-JUN used to interpret RNA coding potential. The pairwise mutation analysis was conducted on c-JUN's 5′-UTR segment (containing 56 base pairs) to ensure the impact score of each base pair on RNA coding potential (described on the left side). 37-GC was found at three different positions (2-40, 10-48 and 14-52) in 5′-UTR region (as shown in the upper right side), and a boxplot of coding potential scores was applied for visualizing the importance of 37-GC found at different locations (2-40, 10-48 and 14-52) (provided in the lower right side). The green dashed line showed the coding potential scores of the original c-JUN mRNA sequence. The *p*-value <0.01 was denoted using two asterisks (**), while the *p*-value <0.001 were indicated using three asterisks (***).

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202509518 by Zhejiang University, Wiley Online Library on [04/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

identified one RNA sequence (PDB ID: 6NOA, c-JUN 5′-UTR mRNA). A stem-loop structure in this mRNA was identified (colored in GREEN on the right side of Figure 6a), which had been reported to be crucial for the recognition of this mRNA by protein eIF3 and for the specialized translation initiation of c-JUN.[56] eIF3 directly binds the stem-loop structure within the c-JUN 5′ untranslated region (5′ UTR), which is essential for specialized c-JUN translation and the dysregulation disrupts the precise control of c-JUN expression.[57]

On the left side of Figure 6b, a pairwise mutation analysis was performed on the c-JUN 5′-UTR segment (56 bp) using *IntRNA* to assess the contribution of any studied nucleobase pair to coding potential of RNA. The darker the color of a cell was, the greater contribution of the corresponding nucleobase pair to coding potential was. As described, the bases within the location from 9 to 15 offered the most significant contribution to the coding potential, which perfectly overlapped with the stem-loop structure highlighted in GREEN on the right side of Figure 6a and the left side of Figure 6b. Furthermore, the contribution to coding potential was different for pairwise mutations of the same *G-features* when a studied nucleobase pair was found in different sequence locations. Taking the gap-feature *G-C* as an example, its *G-feature* of 37 bases gap (37-GC) was discovered as important ($IS > 0.8$, right side of Figure S2b, Supporting Information) for RNA coding potentials. By mapping 37-GC onto the c-JUN 5′-UTR mRNA, three nucleobase pairs (colored in RED, BLUE, and YELLOW) were found. The red-colored and blue-colored pairs were located in the identified GREEN region, while the yellow-colored one was not. Therefore, it is interesting to compare the contributions of these three 37-GC pairs to the coding potential prediction. As shown on the right side of Figure 6b, the coding potential score (CPS) for all possible mutated sequences (a total of nine sequences for each pair, described in the step 5 of the Methods section) was calculated, and a green line was used to provide the CPS of the original sequence. The larger the CPS differences between the original sequence and those nine mutated ones were, the greater contributions of the corresponding 37-GC pairs made. As illustrated in the lower-right corner of Figure 6b, there were significant differences between the CPSs of yellow-colored 37-GC pair and that of the red-colored/blue-colored ones, which further led to distinct contribution made by the yellow-colored, red-colored, and blue-colored pairs. In other words, the 37-GC pairs located within the identified region (red and blue) contributed much more compared with that outside the region (yellow). All in all, *IntRNA* could effectively detect the impacts of sequence variations on coding potential and the roles of certain segments in RNA structure. Therefore, *IntRNA* could also be considered as a powerful method/tool for revealing the relationship between RNA structure and function.

2.6. In-Depth Interpretation of ncRNA Functional Potential by Molecular Example

To further evaluate *IntRNA*'s interpretability, SNORD44, a snoRNA known to guide site-specific 2′-O-methylation on 18S rRNA was analyzed.[58–60] As shown in Figure S6a (Supporting Information), structural and functional elements, including the conserved Box C and Box D motifs (highlighted in red), were focused. These motifs are essential for snoRNA stability, maturation, and snoRNP complex formation.[61] A pairwise mutation analysis of the 61-nt SNORD44 sequence revealed that nucleotides 9-12 and 24-29 contributed most significantly to binding potential, partially overlapping with the Box C and D' motifs (red color in Figure S6a, Supporting Information). As shown in Figure S6b (Supporting Information), these regions also had higher probability scores predicted by *IntRNA* compared to surrounding sequences. Furthermore, gap-based features, such as 14-AU and 18-AU (IS > 0.6; Figure S4c, Supporting Information) were analyzed. As shown in Figure S6c,d (Supporting Information), red-colored 14-AU and 18-AU pairs located within functional motifs showed greater changes in probability scores upon mutation than gray-colored pairs outside these regions.

In a similar way, circHIPK3, a circular RNA derived from exon 2 of the HIPK3 was adapted to further evaluate the interpretability and generalizability of *IntRNA*.[62] Previous studies have reported that miR-605[63] and miR-124[64] can bind within the first 80-nucleotides regions of circHIPK3, as shown in Figure S7a (Supporting Information). A pairwise mutation analysis of this 81-nt segment revealed that nucleotides 36-42 and 46-55 contributed most significantly to interaction potential, overlapping precisely with the predicted binding sites for miR-605 and miR-124. These regions also exhibited higher binding probability scores, as predicted by *IntRNA* (Figure S7b, Supporting Information), compared to adjacent non-functional sequences. Gap-based interaction features were further examined, including 5-CA and 13-AA, which demonstrated strong importance scores (IS > 0.6; Figure S5b, Supporting Information). As shown in Figure S7c,d (Supporting Information), red-colored 5-CA and 13-AA pairs located within miRNA binding motifs showed markedly greater differences in probability scores upon mutation than gray-colored counterparts outside these regions. Together, these results underscore *IntRNA*'s capability to accurately pinpoint structurally and functionally relevant motifs in non-coding RNAs, particularly those mediating miRNA-RNA interactions.

Building upon these capabilities, *IntRNA* provides a unified and interpretable framework with strong predictive performance across *RNA universe*. It is particularly well-suited for applications such as genome-wide RNA annotation, evolutionary transcriptomics, and the discovery of functionally relevant RNAs from bulk sequencing data.[65–68] Nevertheless, the method has certain limitations: it requires substantial computational resources for large-scale feature extraction and some features may exhibit limited utility depending on the specific classification task. Overall, *IntRNA* offers a robust and extensible foundation for RNA annotation, with clear opportunities for future improvement in scalability, adaptive feature selection, and integrative analysis.

3. Conclusion

The interpretation of *RNA universe* and coding potential are long-standing issues in modern RNA studies, and three crucial questions remain unanswered: a) how to detect and interpret the coding potential of RNA, b) how to annotate the sophisticated taxonomy of the sncRNAs and c) how to successfully distinguish between circular and linear lncRNAs. In this study, a multi-channel deep learning framework *IntRNA*, was thus constructed to

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202509518 by Zhejiang University, Wiley Online Library on [04/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

interpret *RNA universe* and coding potential. First, a large number of RNA encoding features were proposed, which dramatically enlarged the available feature space. Second, a method realizing image-like representation of RNA sequences was developed to describe the intrinsic correlation among the encoding features generated above. Third, a dual-path multi-channel model was developed, which consistently performed well when comparing with existing methods in a series of benchmark studies. The interpretability of *IntRNA* was tested by a real-world study, which identified the key feature determining coding potential.

4. Experimental Section

Two key modules were integrated into the latest *IntRNA*: 1) the image-like representation module "*RNAImage*" and 2) the dual-path multi-channel prediction module "CNN-based model". In order to construct the *IntRNA* model, five critical steps were employed here, which included: preparing datasets for model construction and performance assessments, generating new encoding features for enlarging the existing feature space, developing novel strategy enabling the representation of RNA as image, building the CNN-based prediction model for interpreting *RNA universe*, and the in-depth interpretation of the RNA coding potential based on feature ranking.

Step 1. Preparing Datasets for Model Construction and Performance Assessment: To construct the *IntRNA* model, a total of 788576 unique human RNAs and their sequences were first gathered from *ENCODE*[69] and *Refseq*.[70] Then, three benchmarks were collected by literature review to assess the performances of studied methods in solving the three key problems described in Figure 1. As detailed in Table S2a (Supporting Information), the dataset adopted to assess the methods' performances in the detection and interpretation of RNA coding potential[14] had a total of 180233 RNAs (including 112726 mRNAs & 67507 ncRNAs) from five different species: *Homo sapiens*, *Mus musculus*, *Danio rerio*, *Drosophila melanogaster* & *Saccharomyces cerevisiae*. Particularly, the data from *Homo sapiens* were split by the original publication[14] into the training (comprising 33360 mRNAs and 24163 ncRNAs) and the test (comprising 8557 mRNAs and 8241 ncRNAs) datasets. During the training process, the strategy of five-fold cross-validation was applied, and the test dataset was further enriched by integrating the data from the remaining four species.

Furthermore, the dataset adopted to evaluate the methods' performances in the annotation of the sophisticated taxonomy of sncRNAs[24] had a total of 8920 RNAs (as detailed in Table S2b, Supporting Information), and the partition of this dataset into training, validation and test followed the exactly same way as that in the original publication.[24] Additionally, the benchmark adopted to assess the methods' performance in distinguishing between circular and linear lncRNAs[28] contained a total of 51625 RNAs (comprising 31942 circular & 19683 linear lncRNAs), and the dataset partition into training, validation and test followed the same way as that in the original publication.[28]

Step 2. Generating New Encoding Features for Enlarging Existing Feature Space: In this study, a large number of new encoding features were generated to extensively expand the feature space of RNA representation. As shown in Figure 2, there were two types: the gap-based features (*G-features* containing a total of 724 newly-proposed features) and encoding descriptors (*D-features* comprising a total of 753 features and including 439 newly-proposed ones). For both types, their containing sub-groups (17 & 14 subgroups for *G-features* & *D-features*, respectively) were shown in the outermost layer, and the 14 sub-groups of *D-feature* were further summarized to three classes (sequence-intrinsic, structure-based & physicochemical property-based). Those numbers in the brackets indicated the total numbers of features in the feature subgroups.

The *G-feature* emphasized the distribution of nucleotides within RNA sequence, whereas the *D-feature* gave the sequence-intrinsic, structure-based & physicochemical property-based property of RNAs. As shown in Figure 3a, the way used in this study to generate *G-features* was provided by taking the pairs of *G-C* and *C-G* as example, which were calculated by counting the frequency of nucleotide pairs at various intervals (from 0 to 40). For instance, *2-GC* stood for the frequency of the nucleotide pair *G-C* separated by two nucleotides within an RNA sequence, which resulted in a value of 1 for *2-GC* in the example shown on the left side of Figure 3a. Similarly, the values resulted for *2-CG*, *0-CG*, *0-GC*, *1-GC*, and *1-CG* equaled to 1, 2, 1, 1, and 1, respectively.

Furthermore, a total of 753 *D-features* were generated in this study to encode the RNA sequences, 314 out of which were the classical encoding features (such as open reading frame, codon related, and secondary structures) proposed by previous reports,[71–77] described in the Method S1 (Supporting Information). For the remaining 439 features, they were newly-generated in this study based on specific set of physicochemical/structural properties (such as hydrogen bond basicity, topological polar surface area, and atom-specific path lengths). Particularly, 439 *D-features* were calculated using the strategy of Composition-Transition-Distribution (CTD). The detailed calculating method was systematically described in the Method S2 (Supporting Information). All in all, a total of 1477 encoding features were generated in this study to realize RNA representations, which included 724 *G-features* and 753 *D-features*. As illustrated on the right side of Figure 3a, a large number of encoding features/descriptors were generated in this study, which dramatically expanded (to over four times) the traditionally available feature space that encoded RNAs.

The feature sparsity was evaluated by calculating the fraction of zero values per feature across all samples in each task. As shown in Figure S8 (Supporting Information), most features exhibit low sparsity, indicating broad coverage and informative signal. To assess redundancy, pairwise cosine similarity analysis of *D-* and *G-features* was conducted. As shown in Figure S9 (Supporting Information), while most features are distinct, moderate similarity is observed among subsets derived from related physicochemical or structural properties.

Step 3. Developing Novel Strategy Enabling the Representation of RNA as Image: A strategy titled *RNAImage* was further proposed to realize the image-like representation of RNA for capturing the intrinsic correlations among the massive amount of RNA encoding features. As offered in Figure 3b, two template images (*D-template* & *G-template*) were first generated using a total of 788576 RNA sequences from two established knowledge bases (*ENCODE* and *RefSeq*) by following four sequential processes (from matrix generation[78] to similarity calculation,[79] then to dimension reduction,[80] and finally to grid assignment[81]), which resulted in the template images of 28 × 27 (*D-template*) and 27 × 27 (*G-template*). Second, each analyzed RNA sequence could be converted to two "images" by mapping their intensities of encoding features to the corresponding locations in those two newly-constructed template images based on our *RNAImage* strategy.

A detailed description on the application of *RNAImage* strategy was provided in Figure S10 (Supporting Information). First, each RNA sequence was converted into two feature-matrices of distinct matrix dimension (753 × 788576 and 724 × 788576 for *D-features* and *G-features*, respectively). Second, each feature-matrix was transformed to a similarity-matrix by calculating the Cosine Similarities among studied features, which led to two similarity-matrices of distinct dimension (753 × 753 and 724 × 724 for *D-features* and *G-features*, respectively). Third, UMAP approach was then adopted to reduce the dimension of each similarity-matrix, and *J-V algorithm* was further applied to each UMAP plot, which finally generated two template images of 28 × 27 (*D-template*) and 27 × 27 (*G-template*) dimension. Finally, each RNA sequence could be converted to two images by mapping their intensities of encoding features to the corresponding locations in two template images.

Step 4. Building the CNN-Based Prediction Model for Interpreting RNA Universe: A dual-path multi-channel model was constructed to facilitate the interpretation of *RNA universe*. As shown in Figure 3c, each RNA sequence was first converted to two images of different matrix dimension (28 × 27 & 27 × 27 for *D-features* & *G-features*, respectively) using *RNAImage* strategy described in step 3. Second, both images representing the studied RNA were then split to multiple channels according to the number of feature sub-groups provided in Figure 2, which resulted in 17 and 14

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202509518 by Zhejiang University, Wiley Online Library on [04/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

channels for *D-template* and *G-template*, respectively. Third, those channels were then fed into a convolutional layer containing a single filter in a dual path manner, and a convolutional block was separately adopted to extract information (including a max-pool layer, a convolutional layer with three filters and a concatenation layer), which was repeated twice to extract embedded features and resulted in a total of 192 matrices of 10 × 10 dimension in each path. Fourth, a global max pooling layer was separately applied to each path for converting the 192 matrices to a vector, and the vectors of two paths were then concatenated to generate new vector of 384 length. Finally, this newly-generated vector was forwarded to a fully connected layer for RNA classification.

The classifiers were implemented using *Python TensorFlow* (v2.3.0) and Keras libraries (v2.4.3). The model employed categorical crossentropy as the loss function to measure the discrepancy between predicted class probabilities and true labels. For each of the three tasks (binary classification of RNA coding potential, multi-class annotation of sncRNA taxonomy, and binary classification of circular vs linear lncRNAs), the final output layer of the model was adjusted accordingly. To assess the performances of *IntRNA* and available methods, six popular metrics were employed, which are specificity (SPE), sensitivity (SEN), precision (PRE), accuracy (ACC), area under the curve (AUC), and Matthews correlation coefficient (MCC). A grid search methodology was used to identify the optimal combination of hyperparameters (learning rate and batch size) in terms of the parameter tuning. Hyperparameter optimization for *IntRNA* was performed using MCC value. The batch size varied at 32, 64 and 96, and the learning rate was tested at 0.00001, 0.0001, 0.0002, and 0.0005. As provided in Figure S11 (Supporting Information), the *IntRNA* model exhibited insensitivity to both studied hyperparameters, as evidenced by the line charts providing MCC values. Training and validation accuracy and loss curves during training were provided in Figure S12 (Supporting Information) and parameter tuning of UMAP and dimensionality reduction method comparison were shown in Figure S13 (Supporting Information).

Step 5. In-Depth Interpretation of the RNA Coding Potential by Feature Ranking: To quantify the contribution of the studied RNA encoding features to the newly-developed model, the importance score of each feature was measured by permutation algorithm.[82] Particularly, the optimized cross-entropy loss (OCEL) of the constructed model was first calculated using the data (provided in Table S2a, Supporting Information) applied to train the model for detecting and interpreting RNA coding potential. Second, permutations were applied to each feature on RNAs' image-like representations, and a permuted cross-entropy loss (PCEL) was calculated for each feature. Third, the importance score (*IS*) of a studied feature was determined by the difference between its PCEL and OCEL. The higher the IS was, the greater the contribution of the feature to the *IntRNA* model was. This process above was iterated for all features until their ISs were fully obtained.

To measure the contribution of certain *G-feature* in detecting coding potential, pairwise mutation analysis was further applied to confirm the contribution of an identified nucleotide pair to *IntRNA* model. As provided in the Figure S14 (Supporting Information), the nucleotide *G* and *C* were fully mutated, which resulted in a total of nine mutated sequences. The probability scores generated by the fully connected layers of the *IntRNA* model provided in Figure 3c were used to quantitatively measure the coding potential of all those ten sequences (including nine mutated and one original), and the average probability score of those nine mutated sequences could thus be represented as:

$$P_{mutated} = \frac{1}{9} \times \sum_{i=1}^{3} \sum_{j=1}^{3} P_{i,j} \tag{1}$$

where $P_{i,j}$ indicated the probability score of the sequences (with mutations at the studied positions) predicted by *IntRNA* model. As a result, the contributions of the studied nucleotide pair to *IntRNA* model could therefore be described using the following Equation (2):

$$C_{nucleotidepair} = P_{original} - P_{mutated} \tag{2}$$

The higher the $C_{nucleotide\ pair}$ was, the greater contribution the studied pair made to coding potential. For visualizing RNA structure, the *RNArtist* (https://github.com/fjossinet/RNArtist) was adopted in this study to convert 3D RNA structure into a 2D plot.

Statistical Analysis: In this work, the statistical tests and the definition of *p*-values are provided in the figure legends. Statistical analysis was performed in R language.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The National Natural Science Foundation of China (82373790, 22220102001, 62201289); the Natural Science Foundation of Zhejiang (RG25H300001); the National Key R&D Programs of China (2024YFA1307503); the Information Technology Center of Zhejiang University.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Y.X.W., M.J.M., and S.J.H. contributed equally to this work as co-first authors. F.Z. conceived the idea and designed the research. Y.X.W., M.J.M., S.J.H., and Q.X.Y. developed the model and debug source codes. Y.X.W., M.J.M., S.J.H., and Q.X.Y. performed the benchmark data analyses. Y.X.W., M.J.M., S.J.H., W.Z., Z.Q.P., J.T., and Q.X.Y. contributed to statistics and data visualization. Y.X.W., M.J.M., Q.X.Y., and S.J.H. constructed the online server. F.Z., Q.X.Y., Y.X.W., and S.J.H. wrote the manuscript. All authors provided critical feedback to the research.

Data Availability Statement

The source codes and datasets are accessible at: https://github.com/idrblab/intrna.

Keywords

coding potential, deep learning, RNA annotation, RNA representation, RNA universe

Received: May 25, 2025
Revised: August 8, 2025
Published online:

[1] J. S. Mattick, P. P. Amaral, P. Carninci, S. Carpenter, H. Y. Chang, L. L. Chen, R. Chen, C. Dean, M. E. Dinger, K. A. Fitzgerald, T. R. Gingeras, M. Guttman, T. Hirose, M. Huarte, R. Johnson, C. Kanduri, P. Kapranov, J. B. Lawrence, J. T. Lee, J. T. Mendell, T. R. Mercer, K. J. Moore, S. Nakagawa, J. L. Rinn, D. L. Spector, I. Ulitsky, Y. Wan, J. E. Wilusz, M. Wu, *Nat. Rev. Mol. Cell Biol.* **2023**, *24*, 430.
[2] N. Amin, A. McGrath, Y. P. P. Chen, *Nat. Mach. Intell.* **2019**, *1*, 246.

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202509518 by Zhejiang University, Wiley Online Library on [04/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[3] D. Singh, J. Roy, *Nucleic Acids Res.* **2022**, *50*, 12094.

[4] C. Yang, M. Zhou, H. L. Xie, H. Q. Zhu, *Nat. Mach. Intell.* **2021**, *3*, 197.

[5] S. U. Umu, V. M. Paynter, H. Trondsen, T. Buschmann, T. B. Rounge, K. J. Peterson, B. Fromm, *Cell. Genom.* **2023**, *3*, 100348.

[6] H. Wang, J. S. Fleishman, S. Cheng, W. Wang, F. Wu, Y. Wang, Y. Wang, *Mol. Cancer* **2024**, *23*, 177.

[7] K. Nemeth, R. Bayraktar, M. Ferracin, G. A. Calin, *Nat. Rev. Genet.* **2024**, *25*, 211.

[8] R. A. Gupta, N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M. C. Tsai, T. Hung, P. Argani, J. L. Rinn, Y. Wang, P. Brzoska, B. Kong, R. Li, R. B. West, M. J. van de Vijver, S. Sukumar, H. Y. Chang, *Nature* **2010**, *464*, 1071.

[9] L. Allou, S. Balzano, A. Magg, M. Quinodoz, B. Royer-Bertrand, R. Schopflin, W. L. Chan, C. E. Speck-Martins, D. R. Carvalho, L. Farage, C. M. Lourenco, R. Albuquerque, S. Rajagopal, S. Nampoothiri, B. Campos-Xavier, C. Chiesa, F. Niel-Butschi, L. Wittler, B. Timmermann, M. Spielmann, M. I. Robson, A. Ringel, V. Heinrich, G. Cova, G. Andrey, C. A. Prada-Medina, R. Pescini-Gobert, S. Unger, L. Bonafe, P. Grote, et al., *Nature* **2021**, *592*, 93.

[10] L. Poliseno, M. Lanza, P. P. Pandolfi, *Cell Res.* **2024**, *34*, 609.

[11] W. Huang, T. Xiong, Y. Zhao, J. Heng, G. Han, P. Wang, Z. Zhao, M. Shi, J. Li, J. Wang, Y. Wu, F. Liu, J. J. Xi, Y. Wang, Q. C. Zhang, *Nat. Genet.* **2024**, *56*, 124.

[12] J. Shi, T. Zhou, Q. Chen, *Nat. Cell Biol.* **2022**, *24*, 415.

[13] C. X. Liu, L. L. Chen, *Cell* **2022**, *185*, 2016.

[14] X. Tong, S. Liu, *Nucleic Acids Res.* **2019**, *47*, 43.

[15] L. Wang, H. J. Park, S. Dasari, S. Wang, J. P. Kocher, W. Li, *Nucleic Acids Res.* **2013**, *41*, 74.

[16] L. Hu, Z. Xu, B. Hu, Z. J. Lu, *Nucleic Acids Res.* **2017**, *45*, e2.

[17] S. T. Hill, R. Kuintzle, A. Teegarden, E. Merrill, P. Danaee, D. A. Hendrix, *Nucleic Acids Res.* **2018**, *46*, 8105.

[18] A. Li, J. Zhang, Z. Zhou, *BMC Bioinformatics* **2014**, *15*, 311.

[19] T. A. R. Ramos, N. R. O. Galindo, R. Arias-Carrasco, C. F. da Silva, V. Maracaja-Coutinho, T. G. do Rego, *F1000Res.* **2021**, *10*, 323.

[20] Y. Zhang, C. Jia, M. J. Fullwood, C. K. Kwoh, *Brief Bioinform.* **2021**, *22*, 2073.

[21] Y. J. Kang, D. C. Yang, L. Kong, M. Hou, Y. Q. Meng, L. Wei, G. Gao, *Nucleic Acids Res.* **2017**, *45*, W12.

[22] X. G. Chen, S. Liu, W. Zhang, *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2022**, *19*, 1075.

[23] S. Yang, Y. Wang, S. Zhang, X. Hu, Q. Ma, Y. Tian, *Front. Genet.* **2020**, *11*, 90.

[24] A. Fiannaca, M. La Rosa, L. La Paglia, R. Rizzo, A. Urso, *BioData. Min.* **2017**, *10*, 27.

[25] T. M. R. Noviello, F. Ceccarelli, M. Ceccarelli, L. Cerulo, *PLoS Comput. Biol.* **2020**, *16*, 1008415.

[26] L. Childs, Z. Nikoloski, P. May, D. Walther, *Nucleic Acids Res.* **2009**, *37*, 66.

[27] N. Navarin, F. Costa, *Bioinformatics* **2017**, *33*, 2642.

[28] M. Chaabane, R. M. Williams, A. T. Stephens, J. W. Park, *Bioinformatics* **2020**, *36*, 73.

[29] M. Niu, J. Zhang, Y. Li, C. Wang, Z. Liu, H. Ding, Q. Zou, Q. Ma, *Comput. Struct. Biotechnol. J* **2020**, *18*, 834.

[30] X. Pan, K. Xiong, *Mol. BioSyst.* **2015**, *11*, 2219.

[31] X. Wang, Y. Liu, J. Li, G. Wang, *BMC Bioinformatics* **2022**, *23*, 563.

[32] Y. Zhang, M. Lang, J. Jiang, Z. Gao, F. Xu, T. Litfin, K. Chen, J. Singh, X. Huang, G. Song, Y. Tian, J. Zhan, J. Chen, Y. Zhou, *Nucleic Acids Res.* **2024**, *52*, 3.

[33] N. Wang, J. Bian, Y. C. Li, X. H. Li, S. Mumtaz, L. H. Kong, H. Y. Xiong, *Nat. Mach. Intell.* **2024**, *6*, 548.

[34] M. Akiyama, Y. Sakakibara, *NAR Genom. Bioinform.* **2022**, *4*, lqac012.

[35] R. J. L. Townshend, S. Eismann, A. M. Watkins, R. Rangan, M. Karelina, R. Das, R. O. Dror, *Science* **2021**, *373*, 1047.

[36] W. X. Shen, X. Zeng, F. Zhu, Y. L. Wang, C. Qin, Y. Tan, Y. Y. Jiang, Y. Z. Chen, *Nat. Mach. Intell.* **2021**, *3*, 334.

[37] K. Gupta, C. Yang, K. McCue, O. Bastani, P. A. Sharp, C. B. Burge, A. Solar-Lezama, *Genome Biol.* **2024**, *25*, 23.

[38] M. Budnik, J. Wawrzyniak, L. Grala, M. Kadzinski, N. Szóstak, *Artif. Intell. Rev.* **2024**, *57*, 254.

[39] J. J. Quinn, H. Y. Chang, *Nat. Rev. Genet.* **2016**, *17*, 47.

[40] H. Hwang, H. Jeon, N. Yeo, D. Baek, *Exp. Mol. Med.* **2024**, *56*, 1293.

[41] H. Gao, P. Gao, N. Ye, *Comput. Biol. Med.* **2024**, *168*, 107752.

[42] J. Wang, M. Horlacher, L. Cheng, O. Winther, *Bioinformatics* **2024**, *40*, btae065.

[43] L. Cao, C. Yang, L. Hu, W. Jiang, Y. Ren, T. Xia, M. Xu, Y. Ji, M. Li, X. Xu, Y. Li, Y. Zhang, S. Fang, *Gigascience* **2024**, *13*, giae003.

[44] A. Varabyou, B. Erdogdu, S. L. Salzberg, M. Pertea, *Nat. Comput. Sci.* **2023**, *3*, 700.

[45] A. S. Nair, S. P. Sreenadhan, *Bioinformation* **2006**, *1*, 197.

[46] D. W. Binzel, X. Li, N. Burns, E. Khan, W. J. Lee, L. C. Chen, S. Ellipilli, W. Miles, Y. S. Ho, P. Guo, *Chem. Rev.* **2021**, *121*, 7398.

[47] L. Adams, *Nat. Rev. Genet.* **2017**, *18*, 145.

[48] Z. Miao, R. W. Adamiak, M. Antczak, R. T. Batey, A. J. Becka, M. Biesiada, M. J. Boniecki, J. M. Bujnicki, S. J. Chen, C. Y. Cheng, F. C. Chou, A. R. Ferre-D'Amare, R. Das, W. K. Dawson, F. Ding, N. V. Dokholyan, S. Dunin-Horkawicz, C. Geniesse, K. Kappel, W. Kladwang, A. Krokhotin, G. E. Lach, F. Major, T. H. Mann, M. Magnus, K. Pachulska-Wieczorek, D. J. Patel, J. A. Piccirilli, M. Popenda, K. J. Purzycka, et al., *RNA* **2017**, *23*, 655.

[49] T. R. Sosnick, T. Pan, *Curr. Opin. Struct. Biol.* **2003**, *13*, 309.

[50] F. Hube, D. Ulveling, A. Sureau, S. Forveille, C. Francastel, *Nucleic Acids Res.* **2017**, *45*, 4768.

[51] D. Rearick, A. Prakash, A. McSweeny, S. S. Shepard, L. Fedorova, A. Fedorov, *Nucleic Acids Res.* **2011**, *39*, 2357.

[52] K. McNair, C. L. Ecale Zhou, B. Souza, S. Malfatti, R. A. Edwards, *Microorganisms* **2021**, *9*, 129.

[53] O. S. Lusky, M. Meir, A. Goldbourt, *Biophys Rep.* **2021**, *1*, 100027.

[54] L. L. Chen, *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 205.

[55] A. Huang, H. Zheng, Z. Wu, M. Chen, Y. Huang, *Theranostics* **2020**, *10*, 3503.

[56] A. S. Lee, P. J. Kranzusch, J. A. Doudna, J. H. Cate, *Nature* **2016**, *536*, 96.

[57] M. J. Walker, M. D. Shortridge, D. D. Albin, L. Y. Cominsky, G. Varani, *J. Mol. Biol.* **2020**, *432*, 1841.

[58] B. Bai, S. Yegnasubramanian, S. J. Wheelan, M. Laiho, *PLoS One* **2014**, *9*, 107519.

[59] L. Huang, S. Ashraf, J. Wang, D. M. Lilley, *EMBO Rep.* **2017**, *18*, 1631.

[60] S. Galardi, A. Fatica, A. Bachi, A. Scaloni, C. Presutti, I. Bozzoni, *Mol. Cell. Biol.* **2002**, *22*, 6663.

[61] Z. H. Huang, Y. P. Du, J. T. Wen, B. F. Lu, Y. Zhao, *Cell Death Discov.* **2022**, *8*, 259.

[62] J. Zhou, B. Wang, X. Bin, C. Xie, B. Li, O. Liu, Z. Tang, *Front. Med.* **2021**, *8*, 615417.

[63] Y. Li, F. Zheng, X. Xiao, F. Xie, D. Tao, C. Huang, D. Liu, M. Wang, L. Wang, F. Zeng, G. Jiang, *EMBO Rep.* **2022**, *23*, 56102.

[64] G. Chen, Y. Shi, M. Liu, J. Sun, *Cell Death Dis.* **2018**, *9*, 175.

[65] X. Zhou, H. Wu, *Brief Bioinform* **2024**, *26*, bbaf009.

[66] X. Yang, K. K. Mann, H. Wu, J. Ding, *Genome Biol.* **2024**, *25*, 198.

[67] Y. Wu, Z. Shi, X. Zhou, P. Zhang, X. Yang, J. Ding, H. Wu, *Commun. Biol.* **2024**, *7*, 923.

[68] H. Liu, D. Li, H. Wu, *IEEE J. Biomed. Health Inform.* **2023**, *28*, 538.

[69] P. C. Encode, *Nature* **2012**, *489*, 57.

[70] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin,

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202509518 by Zhejiang University, Wiley Online Library on [04/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, et al., *Nucleic Acids Res.* **2016**, *44*, D733.

[71] H. Zhu, Y. Yang, Y. Wang, F. Wang, Y. Huang, Y. Chang, K. C. Wong, X. Li, *Nat Commun.* **2023**, *14*, 6824.

[72] H. Zheng, A. Talukder, X. Li, H. Hu, *Brief. Bioinform.* **2021**, *22*, bbab285.

[73] M. Magnus, M. Antczak, T. Zok, J. Wiedemann, P. Lukasiak, Y. Cao, J. M. Bujnicki, E. Westhof, M. Szachniuk, Z. Miao, *Nucleic Acids Res.* **2020**, *48*, 576.

[74] S. Han, Y. Liang, Q. Ma, Y. Xu, Y. Zhang, W. Du, C. Wang, Y. Li, *Brief Bioinform.* **2019**, *20*, 2009.

[75] S. Zhang, J. Zhou, H. Hu, H. Gong, L. Chen, C. Cheng, J. Zeng, *Nucleic Acids Res.* **2016**, *44*, 32.

[76] Y. Wang, Z. Pan, M. Mou, W. Xia, H. Zhang, H. Zhang, J. Liu, L. Zheng, Y. Luo, H. Zheng, X. Yu, X. Lian, Z. Zeng, Z. Li, B. Zhang, M. Zheng, H. Li, T. Hou, F. Zhu, *Nucleic Acids Res.* **2023**, *51*, 110.

[77] Y. Wang, Z. Chen, Z. Pan, S. Huang, J. Liu, W. Xia, H. Zhang, M. Zheng, H. Li, T. Hou, F. Zhu, *Nucleic Acids Res.* **2023**, *51*, W509.

[78] T. Zhao, A. Zwaenepoel, J. Y. Xue, S. M. Kao, Z. Li, M. E. Schranz, Y. Van de Peer, *Nat Commun.* **2021**, *12*, 3498.

[79] V. J. Sadbhawna, S. Chaudhary, B. N. Subudhi, W. Lin, S. C. Guntuku, *IEEE Trans. Image Process* **2022**, *31*, 2027.

[80] L. Zheng, S. Shi, M. Lu, P. Fang, Z. Pan, H. Zhang, Z. Zhou, H. Zhang, M. Mou, S. Huang, L. Tao, W. Xia, H. Li, Z. Zeng, S. Zhang, Y. Chen, Z. Li, F. Zhu, *Genome Biol.* **2024**, *25*, 41.

[81] R. Jonker, A. Volgenant, *Computing* **1987**, *38*, 325.

[82] A. Altmann, L. Tolosi, O. Sander, T. Lengauer, *Bioinformatics* **2010**, *26*, 1340.

© 2025 The Author(s). Advanced Science published by Wiley-VCH GmbH

21983844, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202509518 by Zhejiang University, Wiley Online Library on [04/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License