BMC Bioinformatics

RESEARCH

Open Access

Predicting RNA sequence-structure likelihood via structure-aware deep learning

You Zhou[1,2], Giulia Pedrielli[1,2*], Fei Zhang[3] and Teresa Wu[1,2]

*Correspondence:
Giulia.Pedrielli@asu.edu

[1] School of Computing and Augmented Intelligence, Arizona State University, 699 S Mill Ave, Tempe, AZ 85281, USA
[2] ASU-Mayo Center for Innovative Imaging, Arizona State University, 699 S Mill Ave, Tempe, AZ 85281, USA
[3] Department of Chemistry, Rutgers University, 73 Warren St, Newark, NJ 07102, USA

Abstract

Background: The active functionalities of RNA are recognized to be heavily dependent on the structure and sequence. Therefore, a model that can accurately evaluate a design by giving RNA sequence-structure pairs would be a valuable tool for many researchers. Machine learning methods have been explored to develop such tools, showing promising results. However, two key issues remain. Firstly, the performance of machine learning models is affected by the features used to characterize RNA. Currently, there is no consensus on which features are the most effective for characterizing RNA sequence-structure pairs. Secondly, most existing machine learning methods extract features describing entire RNA molecule. We argue that it is essential to define additional features that characterize nucleotides and specific sections of RNA structure to enhance the overall efficacy of the RNA design process.

Results: We develop two deep learning models for evaluating RNA sequence-secondary structure pairs. The first model, NU-ResNet, uses a convolutional neural network architecture that solves the aforementioned problems by explicitly encoding RNA sequence-structure information into a 3D matrix. Building upon NU-ResNet, our second model, NUMO-ResNet, incorporates additional information derived from the characterizations of RNA, specifically the 2D folding motifs. In this work, we introduce an automated method to extract these motifs based on fundamental secondary structure descriptions. We evaluate the performance of both models on an independent testing dataset. Our proposed models outperform the models from literatures in this independent testing dataset. To assess the robustness of our models, we conduct 10-fold cross validation. To evaluate the generalization ability of NU-ResNet and NUMO-ResNet across different RNA families, we train and test our proposed models in different RNA families. Our proposed models show superior performance compared to the models from literatures when being tested across different independent RNA families.

Conclusions: In this study, we propose two deep learning models, NU-ResNet and NUMO-ResNet, to evaluate RNA sequence-secondary structure pairs. These two models expand the field of data-driven approaches for learning RNA. Furthermore, these two models provide the new method to encode RNA sequence-secondary structure pairs.

Keywords: Deep learning, RNA, Secondary structure prediction

Background

RNA molecules play an important role in protein synthesis, gene regulation, and catalysis [1, 2]. It is composed of four types of nucleotides, adenine (A), uracil (U), cytosine (C), and guanine (G) [3]. RNA studies have witnessed an important growth in the areas of materials in nanotechnology applications [3]. For example, RNA aptamers can be utilized as biosensors [4], riboswitches that exist in non-coding parts of messenger RNA (mRNA) can control gene expression [5], and RNA strands can be used to construct nanoscaffolds for therapeutic applications in nanomedicine [6]. In many of these applications, the functionality of the designed RNA molecule is highly influenced by its geometrical structure [1, 7]. In general, the structure can be represented in its primary (sequence of nucleotides), secondary (sequence and pairings between nucleotides), and tertiary (sequence, pairings, and 3D displacement of nucleotides) form, with increasing associated computational and experimental complexity. Although RNA tertiary structure can provide insights into the actual 3D geometry, the scope of this study is specifically directed towards the investigation of RNA secondary structure. In fact, RNA secondary structure exhibits greater stability compared to the tertiary structure folding [8]. Moreover, learning RNA secondary structure can help understand and predict the tertiary structure folding [8, 9]. Therefore, developing a comprehensive understanding of RNA secondary structure, along with its associated patterns, can enhance our understanding of RNA and its functional mechanisms.

In this study, we present a series of deep learning models designed to assess RNA sequence-secondary structure pairs, focusing exclusively on pseudoknot-free structures. To describe the distinctive shapes within the secondary structure, we employ the term *motif*, which will be defined in section "Methods". The underlying concept involves using specific sub-structures to capture the localized patterns formed by subsets of nucleotides.

Motivation

Several scientific contributions have focused on the analysis of RNA secondary structures with three core research areas: (i) *secondary structure prediction*: given a sequence, we can predict the secondary structure that the RNA will adopt [10]; (ii) *inverse folding prediction*: given a secondary structure, we can predict the most possible sequence to achieve the target structure [11]; (iii) problems (i)-(ii) have in common the need to evaluate the *quality* of a *sequence-structure* pair [12] because an RNA sequence or an RNA structure with its corresponding predicted RNA structure or predicted RNA sequence need to be evaluated with respect to the likelihood of co-existence of this RNA sequence-structure pair [12]. Recently, it has been shown how methods from (iii) can be used in (i) and, potentially, (ii) in the form of *experts* that evaluate intermediate structures [13]. In the following, we briefly review the state-of-the-art approaches in (i), (ii), and (iii).

Within the research field of *RNA secondary structure prediction*, several studies have focused on the use of minimum free energy (MFE) as a key metric for RNA folding. The underlying assumption is that the structure with the lowest free energy is also the most likely structure the RNA will adopt. It is important to highlight that, generally, free

energy cannot be calculated in closed form due to (a) the incomplete understanding of the RNA molecular interactions, and (b) the impractical computational cost of detailed kinetic simulation tools. As a result, several approximate models have been proposed in the literature [14–17] to estimate the free energy associated with a given secondary structure. An example of methods in this class is RNAfold [10], which uses the approach in [18] to approximate MFE. MXfold2 [19], SPOT-RNA [20], and SPOT-RNA2 [9] utilize deep learning (DL) to predict the RNA secondary structure. Specifically, MXfold2 predicts the RNA secondary structure by maximizing a score which is the sum of a DL model generated folding score and the contribution from thermodynamic parameters. SPOT-RNA employs Transfer Learning [21] where the input is the outer concatenation of the one hot encoding of the RNA sequence, and the output is an upper triangular matrix which represents the predicted base-pairing information. The SPOT-RNA2 adds three features to the RNA sequence as the input: the predicted probability of base pairing obtained from a Linear Partition algorithm [22], the Position Specific Score Matrix (PSSM), and the information of Direct Coupling Analysis (DCA) [23, 24]. The output of SPOT-RNA2 is still an upper triangular matrix representing the predicted base-pairing information. ExpertRNA is a Reinforcement Learning algorithm that uses the rollout method [25–28] to predict RNA secondary structure [13]. ExpertRNA predicts the dot bracket notation of RNA secondary structure by position from 5′ end to 3′ end. In each position, ExpertRNA uses RNAfold to generate multiple intermediate candidate structures [10] to be assessed by RNA sequence-secondary structure pair evaluation model, ENTRNA, presented in [12].

In *RNA inverse folding* area, NUPACK [29] is among the most commonly used methods. NUPACK formulates the inverse folding as an optimization problem whose objective is to minimize the ensemble defect defined as the average of wrongly paired nucleotides' counts on the ensemble of unpseudoknotted structures. When designing the RNA sequence, NUPACK decomposes the target structure into sub-structures that are then optimized. RNAiFold employs the MFE as the objective to predict the RNA sequence for a given RNA secondary structure [11, 30]. The RNAiFold includes two approaches, a Constraint Programming (CP) [31] based algorithm and Large Neighborhood Search (LNS) based algorithm. The constraints of the CP ensure the solutions can follow the RNA folding rules, possess desired features of RNA design, and fold into the corresponding target RNA secondary structures. The only difference between CP and LNS is regarding the search method. The CP searches the entire space while the LNS fixes parts of the solution space and explores the unfixed parts.

From the review, we note that *evaluating the RNA sequence-secondary structure pair* is a fundamental aspect of the methodologies, irrespective of their specific applications. Example evaluating metric is the probability from the Boltzmann distribution which is used in the partition-based methods such as the one proposed in [32] and LinearPartition [22]. The partition-based methods rely on the Boltzmann distribution where the molecule with lower energy has higher probability to exist. However, the consensus regarding which metric should be utilized to evaluate RNA secondary structure has not been made [12]. There has been a notable rise in the utilization of machine learning-based approaches that incorporate features beyond MFE-based methods. This is justified by the recognition that RNA molecules often exhibit stability levels higher than

what is predicted by MFE calculations [12]. Additionally, considering the folding kinetics becomes important when dealing with large RNAs [1]. In this direction, ENTRNA is a classifier [12] that utilizes domain knowledge to determine features for encoding the information of RNA sequence-structure pairs and develops the machine learning model (i.e. logistic regression) to evaluate the pair of RNA sequence and secondary structure. ENTRNA's output score is defined by the authors as *foldability*, and it can be interpreted as the likelihood that the RNA sequence and RNA secondary structure coexist. Considering *foldability* in place of free energy has proven to improve classification and design tasks [12, 13]. In fact, foldability can play a role across several tasks in RNA related studies.

- Folding prediction is usually performed by means of sequential algorithms (e.g., Reinforcement Learning (RL), Dynamic Programming). In this application, foldability can be used as reward function underlying the evaluation of the possible intermediate foldings. An example is given in [13].
- Emerging Generative Artificial Intelligence (GenAI) techniques (e.g., Transformers, Large Language Models) are increasingly being used for both prediction of structures and sequences. These techniques can generate a large number of candidate solutions, thus generating a challenge in experimental validation, which is extremely expensive. A metric such as foldability can be used to downselect solutions thus possibly reducing the experimental effort.

Contribution and paper structure

In order to tackle the aforementioned challenges, we propose a deep learning based approach including two deep learning models, NU-ResNet and NUMO-ResNet, whose scores function as foldability. Our contributions are:

(1) For the challenge that there is no consensus on which features are most effective for characterizing RNA sequence-secondary structure pairs, the NU-ResNet explicitly encodes the RNA sequence-secondary structure pairs by using an innovative image representation, a 3D matrix. Thus, a convolutional neural networks (CNN), ResNet-18 [33], can be employed to automatically extract features from our proposed 3D matrix.
(2) Given the challenge posed by existing machine learning method, which extract features characterizing the entire RNA molecule (e.g. GC percentage), our proposed 3D matrix is designed to incorporate nucleotide-level information including the types of nucleotides and their base pairing. The NUMO-ResNet extends the NU-ResNet by incorporating sub-structure (i.e. motif) information including motifs' types and their free energy, which is encoded into a 2D matrix. We revise the architecture based on ResNet-18 to take both of our proposed 3D matrix and 2D matrix as inputs to develop NUMO-ResNet.

To our knowledge, this is the first paper which utilizes neural networks to evaluate the pair of RNA sequence and RNA secondary structure. The performance of NU-ResNet

and NUMO-ResNet is superior to state-of-the-art approach, ENTRNA. In addition, the NU-ResNet and NUMO-ResNet exhibit the unique advantages compared to the equilibrium probability when being tested on an independent dataset and across RNA families.

The remainder of the paper is structured as follows: in Methods section, we elaborate the detailed algorithms for encoding the pair of RNA sequence and RNA secondary structure as well as the two neural network architectures utilized in our proposed framework. The Results section introduces the data sets utilized in this research, the models' performance comparison with data-driven approach as well as model-driven approach, the analysis of NU-ResNet and NUMO-ResNet, and the testing of the NU-ResNet, NUMO-ResNet, ENTRNA, and Equilibrium Probability across different RNA families. In Discussion section, we discuss the characteristics and potential use of our two proposed models as well as the future work with respect to this research. In Conclusion section, we summarize the work and introduce the potential directions for future research.

Methods

Figure 1 shows an example of RNA secondary structure represented as a graph using the software VARNA [34]. The RNA graph $G_{\text{RNA}} = (V, E, F)$ has vertices V, edges E, and faces F. Each element in the set of vertices $v \in V$ has an associated label $\ell(v) \in \{C, G, A, U\}$, based on its composition being, cytosine, guanine, adenine, and uracil, respectively. Edges $e \in E$ of G_{RNA} are of two types: *hydrogen bonds* connecting two paired nucleotides (we refer to this subset of edges as E_H), also commonly referred to as *interior edges*; and *phosphodiester bonds* connecting any two adjacent nucleotides (we refer to this subset of edges as E_P), commonly referred to as *exterior*

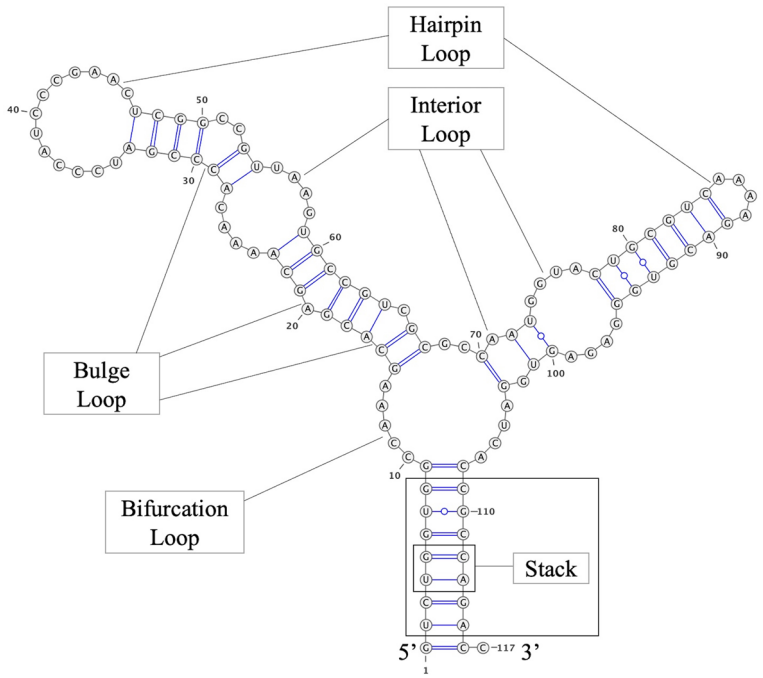


Fig. 1 An example of RNA represented by RNA secondary structure. This RNA is 5 s_ Paracoccus-denitrificans-1, extracted from the Mathews laboratory data set [36]. The 5 motifs considered in our approach are indicated: hairpin, interior, bulge, bifurcation loops, and stack

edges. As a result, $E = E_H \cup E_P$. All edges in E are labelled using the dot-bracket notation presented in [35]. An interior edge is represented by a pair of open and closed brackets ("(", ")"), and an exterior edge is represented by two consecutive dots (".").

Specifically, the dot-bracket notation of pseudoknot-free RNA is comprised of three elements, namely:

- *dot* (".") is used when representing an unpaired nucleotide;
- *open bracket* ("(") is used when a nucleotide is the origin of a base pair. This nucleotide will be paired with one and only one nucleotide;
- *closed bracket* (")") is used when a nucleotide closes a base pair. For symmetry, this nucleotide will be paired with one and only one "open bracket"nucleotide.

As shown below, the information given in graph form by Fig. 1 in the paper can be encoded, without any information loss, using two strings: one containing the characters of the bases in the RNA molecule, and the second reporting the dot-bracket notation to express the secondary structure information.

Sequence:
GUCUGGUGGCCAAAGCACGAGCAAAACACCCGAUCCCAUCCCGAACUC
GGCCGUUAAGUGCCGUCGCGCCAAUGGUACUGCGUCAAAAGACGUGGG
AGAGUGGAUCACCGCCAGACC
Secondary structure:
((((((((((.....(((((.(((....((((((.............))))..))....)))))))..))..(.((....(((((((....)))))))....)).).....)))))))))).

Given each element $f \in F$, a face is the 2-d region defined by the tuple $\langle e_p^i, \ldots, e_h^j, \ldots \rangle$, i.e., a closed region bounded by consecutive hydrogen and phosphodiester bonds. The number and arrangement of edges determines the size and the shape of each face. Given the constraints that nucleotides have to satisfy when bonding, 5 shapes exist for faces in 2-d RNA structures (see Fig. 1 for a depiction). In the rest of the paper, we refer to the shape of the faces as *motifs*.

The definition of *motifs* are as follows.

- Stack: A face having two interior edges which are separated by one exterior edge on each side.
- Hairpin loop: A face only having one interior edge.
- Interior loop: A face having two interior edges separated by more than one exterior edge on each side.
- Bulge loop: A face having two interior edges separated by more than one exterior edge on one side and exactly one exterior edge on the other side.
- Bifurcation loop: A face having more than two interior edges.

Section "Nucleotide-level features-informed residual neural network model (NU-ResNet)" introduces our first DL model in which the nucleotide level information of an RNA sequence-structure pair is encoded into a 3D matrix. We name this model NU-ResNet. Section "Nucleotide-level features and Motifs-informed residual neural network model (NUMO-ResNet)" introduces our second DL model, which adds motifs to the features encoded by NU-ResNet, we refer to this as NUMO-ResNet.

Nucleotide-level features-informed residual neural network model (NU-ResNet)

In this section, we introduce our first deep learning model, which uses nucleotide level properties to evaluate sequence-structure pairs. We refer to this model as NU-ResNet since it uses a Residual Neural Network as basis architecture and it uses nucleotide level features to encode the input.

Input layer encoding A key contribution of this work is the design and implementation of the input layer encoding, which we detail herein. The input layer for the first model describes the structure of the molecule. We encode the secondary structure as a 3D matrix of size $[L \times L \times B]$, where L is the length of RNA sequence (number of nucleotides), $B = 4$ is the number of bits we need for the one-hot encoding of the nucleotide type (C,G,A,U) and the base pairing information, which we will refer to as the channels for the input of our NU-ResNet. As a result, a cell in the 3D matrix with index $(i,j), i = 1,\ldots,L; j = 1,\ldots,L$ is a 4-elements vector with each element in $\{0,1\}$. A key motivation behind the choice of transforming the one-hot encoding into a 3D matrix is driven by the consideration that deep learning approaches are particularly effective with imaging data which are essentially 2D or 3D matrixes, in that they are designed to extract features from this type of input format. In the following, we detail the construction of the 3D matrix starting from input sequence-structure information. The 3D matrix $\mathbf{G}$ is initialized with all zeros.

Diagonal elements encoding We set

$$\mathbf{G}_{ii\cdot} = \begin{cases} [0,0,1,0] & \text{if type is cytosine } (C) \\ [0,0,0,1] & \text{if type is guanine } (G) \\ [1,0,0,0] & \text{if type is adenine } (A) \\ [0,1,0,0] & \text{if type is uracil } (U) \end{cases}, i = 1,\ldots,L$$

Out-of-diagonal elements encoding Next, we encode all the edges $e \in E_H$ as vectors of the 3D matrix, denoted as $\mathbf{G}_{ij\cdot}$, which satisfy $(i,j) \in E_H$, i.e., all the nucleotides that form a base pair.

$$\mathbf{G}_{ij\cdot} = \begin{cases} [1,1,0,0] & \text{if base pair is } (AU) \\ [0,0,1,1] & \text{if base pair is } (CG) \\ [0,1,0,1] & \text{if base pair is } (UG) \end{cases}, (i,j) \in E_H$$

Specifically, the values are obtained as the sum of the one-hot encoding vectors of the two nucleotides involved in the pairing. For symmetry, $(i,j) \in E_H$ implies $(j,i) \in E_H$, hence the vector $\mathbf{G}_{ij\cdot}$ is equivalent to the vector $\mathbf{G}_{ji\cdot}, \forall i,j = 1,\ldots,L$, where the equality between vectors is interpreted as the equality of all elements.

Example In order to clarify the approach, we show an example that leads to the generation of the 3D RNA matrix. We start considering a fictitious RNA sequence "CAG GAGCUCUUC" with corresponding secondary structure ".(((...))).", i.e., it contains 3 base pairs. The visualization of the 3D RNA matrix from this example is shown in Fig. 2.

Deep learning architecture design and architecture training Convolutional Neural Networks (CNN) have shown very good performance within the image recognition literature [33, 37, 38]. Within this broad family of learning models, ResNets are designed to resolve the degradation of training accuracy when the depth of the neural network increases [33]. ResNets have shown robust performance on image classification tasks. Rather than learning the sophisticated functions to depict the relationship

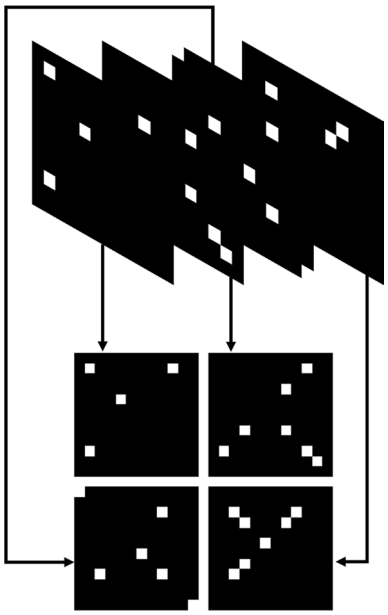


Fig. 2 Example of 3D RNA matrix. The combination of four stacked subfigures on the top is the input of NU-ResNet. The flat form of four stacked subfigures are shown at the bottom. The 4 subfigures at the bottom represent the encoded G_{ij1}, G_{ij2}, G_{ij3}, and G_{ij4}, respectively. Fundamentally, each subfigure is a 2D matrix, which refers to a channel in the 3D RNA matrix. In this 3D RNA matrix example, the white box and black box correspond to 1 and 0 values of the matrix, respectively

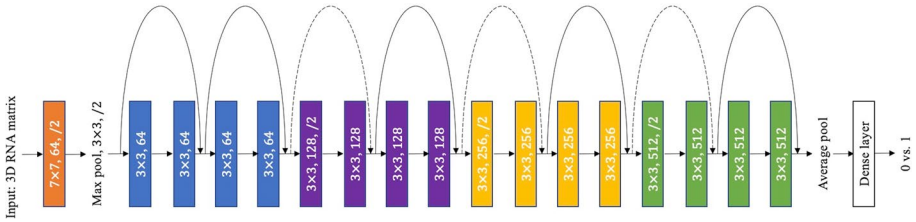


Fig. 3 NU-ResNet architecture. The black straight and solid arrows and the curved dashed arrows represent identity shortcuts, additive operators that combine the input and the output of a layer returning the residuals to learn. The black solid curved arrows are used in cases where the input and output have the same dimensionality, and the black dashed curve is used when output dimension is increased compared to input dimension [33]. Each identity shortcut corresponds to a *building block* which is a network sub-structure including 2 convolutional layers, 2 Batch Normalization, 2 rectified linear activation units (ReLU) [39] (enabling non-linear models), and 1 identity shortcut. The Batch Normalization [40] is used between each convolutional layer and the ReLU activation to normalize the intermediate representation [33, 40]. The kernel size of the first convolutional layer is 7 and the kernel size of any convolutional layer within the building block is 3. The number of output channels from the convolutional layers in the 4 types of building block is 64, 128, 256, and 512, respectively. The parameters in the convolutional layer at the beginning of architecture, in the eight building blocks, and in the fully connected layer (i.e. Dense layer) need to be learned during training

between input and output directly, ResNets learn how to approximate residual functions by using stacked nonlinear layers [33]. For these reasons, we choose the ResNet-18 as the primary architecture in this research.

Figure 3 shows the NU-ResNet architecture with detailing the building block including all the components from the beginning of an identity shortcut to its end, where an identity shortcut is represented by a black solid or dashed curve. The

difference between identity shortcuts with black solid and dashed curve is also illustrated in Fig. 3.

The model is trained using RNA molecules of different lengths each corresponding to 3D RNA matrixes of different sizes (L). To allow training, we transform the input by "padding" with zeros along each of the $B = 4$ channels. After padding, the size of each 3D RNA matrix equals to $\mathcal{L} \times \mathcal{L} \times 4$, where $\mathcal{L}$ is any value greater than the size of the longest RNA we are building the model to evaluate. This value can be given as input by the user or be set to the size of longest sequence across the training, testing, and validation datasets.

When training NU-ResNet, we employ Adam [41] as the optimizer of the models parameters. The learning rate and weight decay of optimizer for NU-ResNet are 0.0001 and 0.15, respectively. The learning rate decreases exponentially with gamma value equaling to 0.95, the batch size during the training is 20, and the model is trained for 100 epochs. The models associated with the best validation accuracy and the best validation loss are saved as models to be deployed.

Nucleotide-level features and motifs-informed residual neural network model (NUMO-ResNet)

As mentioned in section "Motivation", several approaches account for RNA features that knowingly impact the folding. NUMO-ResNet extends the fully data-driven (black-box) model in section "Nucleotide-level features-informed residual neuralnetwork model (NU-ResNet)" to include features *localized* to sub-structures of the molecule that can potentially impact the RNA stability. In particular, we include the motifs present in the structure, and characterize them with the associated free energy, and the motif types (see section "Motivation" for the definition of free energy and Sect. 2, Fig. 1 for the definition of the motifs).

Input layer encoding In the following, we detail the approach to automatically derive the motifs contained in the structure, and how to calculate the free energy associated to the different sub-structures given the type of motif and the number of nucleotides involved. The nucleotide type is encoded in the same way as in NU-ResNet.

Automatic identification and encoding of the motifs NUMO-ResNet adds to the base types used in NU-ResNet, the motifs associated to each nucleotide (e.g., stack, hairpin loop, interior loop, bulge loop, bifurcation loop, and no motif). Since a motif involves a face (sub-structure) of a molecule, based on its location within the sub-structure, the same nucleotide may be involved in two motifs. For this reason, for each nucleotide $i = 1, \ldots, L$, we can assign two distinct vectors to encode all the possible motifs the base is involved in. This results into two matrixes $\mathbf{M}^1_{i\cdot}, \mathbf{M}^2_{i\cdot}$ encoded as follows:

$$\mathbf{M}^k_{i\cdot} = \begin{cases} [1,0,0,0,0,0] & \text{stack} \\ [0,1,0,0,0,0] & \text{hairpin loop} \\ [0,0,1,0,0,0] & \text{interior loop} \\ [0,0,0,1,0,0] & \text{bulge loop} \\ [0,0,0,0,1,0] & \text{bifurcation loop} \\ [0,0,0,0,0,1] & \text{no motif} \end{cases}, i = 1, \ldots, L; k = 1, 2.$$

While the encoding of the nucloetide type is a direct translation from the input sequence information, an algorithm is necessary to recover the type and characteristics of the

motifs given the sequence and structure information. We develop a procedure to automatically extract motif information for each nucleotide in RNA sequence.

In the following, we show the key steps of the procedure to automatically extract motif information for each nucleotide in RNA sequence.

- *Initialization*: As previously mentioned, the RNA structure is given as input encoded as a string of dots and brackets, which we refer to as Σ of size L (number of nucleotides). We create a set for each motif type and initialize it to the empty set, namely the set of stacks $S^{\text{stack}} = \emptyset$, the set of hairpin loops $S^{\text{hairpin}} = \emptyset$, the set of bulge loops $S^{\text{bulge}} = \emptyset$, the set of interior loops $S^{\text{interior}} = \emptyset$, and the set of bifurcation loops $S^{\text{bifurcation}} = \emptyset$.
- *Step 1.a Base Pair Identification*: All base pairs within Σ are translated into a tuple $(j,k), j,k \in \{1,\ldots,L\}$ representing the index of the origin and destination nucleotide, respectively. The collection of all the tuples forms the base pairs index set $\mathbf{B}$ whose elements $\mathbf{b}_i, i = 1,\ldots,|\mathbf{B}|$ are defined as $\mathbf{b}_i = (j,k)$. We adopt the `rna-tools` [42] Python package for the automatic identification of all base pairs within this RNA secondary structure. The implemented algorithm runs with complexity $O(L)$.
- *Step 1.b Motif Elicitation* As previously explained, each motif is bounded by at least one base pair. Hence, given the collection of base pairs from Step 1.a, $\mathbf{B}$, we verify which motifs the base pairs are defined in. More specifically:

  – Stack: If $(\mathbf{b}_i(1)+1, \mathbf{b}_i(2)-1) \in \mathbf{B}$, then

  $$\begin{aligned} S^{\text{stack}} \leftarrow S^{\text{stack}} \cup \{&\mathbf{b}_i, (\mathbf{b}_i(1)+1, \mathbf{b}_i(2)-1), \\ &(\mathbf{b}_i(1), \mathbf{b}_i(1)+1), (\mathbf{b}_i(2)-1, \mathbf{b}_i(2))\}, \end{aligned}$$

  where $\{\mathbf{b}_i, (\mathbf{b}_i(1)+1, \mathbf{b}_i(2)-1)\}$ are interior edges (hydrogen bonds), while $\{(\mathbf{b}_i(1), \mathbf{b}_i(1)+1), (\mathbf{b}_i(2)-1, \mathbf{b}_i(2))\}$ are exterior edges (phosphodiester bonds);

  – Hairpin loop: If the structure elements $\Sigma_j = \text{“.”}, \forall j = \mathbf{b}_i(1)+1,\ldots,\mathbf{b}_i(2)-1$, then

  $$\begin{aligned} S^{\text{hairpin}} \leftarrow S^{\text{hairpin}} \cup \{&\mathbf{b}_i, (\mathbf{b}_i(1)+1, \mathbf{b}_i(1)+2), \\ &\ldots, (\mathbf{b}_i(2)-2, \mathbf{b}_i(2)-1)\}, \end{aligned}$$

  where all edges are phosphodiester bonds, except the hydrogen bond $\mathbf{b}_i$;

  – Bulge loop: Similar to the stack, we have two consecutive base pairs whose bases are either at the origin or destination. In particular, if the bulge is *on the side of the destination* bases, we have that $\mathbf{b}_i(1)+1 = \mathbf{b}_{i+1}(1)$,with a bulge of size $\mathbf{b}_i(2) - \mathbf{b}_{i+1}(2) > 1$.Then,

  $$\begin{aligned} S^{\text{bulge}} \leftarrow S^{\text{bulge}} \cup \{&\mathbf{b}_i, \mathbf{b}_{i+1}, (\mathbf{b}_i(1), \mathbf{b}_{i+1}(1)), \\ &(\mathbf{b}_{i+1}(2), \mathbf{b}_{i+1}(2)+1), \ldots, (\mathbf{b}_i(2)-1, \mathbf{b}_i(2))\}, \end{aligned}$$

  where all edges are phosphodiester bonds, except the hydrogen bonds $\mathbf{b}_i, \mathbf{b}_{i+1}$.In case the bulge is *on the side of the origin* bases, we have that $\mathbf{b}_i(2) - 1 = \mathbf{b}_{i+1}(2)$, with a bulge of size $\mathbf{b}_{i+1}(1) - \mathbf{b}_i(1) > 1$. Then,

$$S^{\text{bulge}} \leftarrow S^{\text{bulge}} \cup \{\mathbf{b}_i, \mathbf{b}_{i+1}, (\mathbf{b}_i(1), \mathbf{b}_i(1)+1), \ldots, (\mathbf{b}_{i+1}(1)-1, \mathbf{b}_{i+1}(1)), (\mathbf{b}_{i+1}(2), \mathbf{b}_i(2))\},$$

where all edges are phosphodiester bonds, except the hydrogen bonds $\mathbf{b}_i, \mathbf{b}_{i+1}$;

– Interior loop: Similar to the bulge loop, we have two consecutive base pairs whose bases are either at the origin or destination. *On the side of the origin and destination* bases, we have that $\mathbf{b}_{i+1}(1) - \mathbf{b}_i(1) > 1$ and $\mathbf{b}_i(2) - \mathbf{b}_{i+1}(2) > 1$. Then,

$$S^{\text{interior}} \leftarrow S^{\text{interior}} \cup \{\mathbf{b}_i, \mathbf{b}_{i+1}, (\mathbf{b}_i(1), \mathbf{b}_i(1)+1), \ldots, (\mathbf{b}_{i+1}(1)-1, \mathbf{b}_{i+1}(1)), (\mathbf{b}_{i+1}(2), \mathbf{b}_{i+1}(2)+1), \ldots, (\mathbf{b}_{i+1}(1)-1, \mathbf{b}_{i+1}(1))\}$$

where all edges are phosphodiester bonds, except the hydrogen bonds $\mathbf{b}_i, \mathbf{b}_{i+1}$.

– Bifurcation loop: Let $\mathbf{z}$ represent the number of base pairs in the loop. If there are more than two base pairs (i.e. $\mathbf{z} > 2$) in this loop, then

$$\begin{aligned} S^{\text{bifurcation}} \leftarrow S^{\text{bifurcation}} \cup \{&\mathbf{b}_i, \mathbf{b}_{i+1}, \ldots, \mathbf{b}_{i+\mathbf{z}-1}, \\ &(\mathbf{b}_i(2), \mathbf{b}_i(2)+1), \ldots, (\mathbf{b}_{i+1}(1)-1, \mathbf{b}_{i+1}(1)), \\ &(\mathbf{b}_{i+1}(2), \mathbf{b}_{i+1}(2)+1), \ldots, (\mathbf{b}_{i+2}(1)-1, \mathbf{b}_{i+2}(1)), \ldots, \\ &(\mathbf{b}_{i+\mathbf{z}-2}(2), \mathbf{b}_{i+\mathbf{z}-2}(2)+1), \ldots, (\mathbf{b}_{i+\mathbf{z}-1}(1)-1, \mathbf{b}_{i+\mathbf{z}-1}(1)) \\ &(\mathbf{b}_{i+\mathbf{z}-1}(2), \mathbf{b}_{i+\mathbf{z}-1}(2)+1), \ldots, (\mathbf{b}_i(1)-1, \mathbf{b}_i(1))\} \end{aligned}$$

where all edges are phosphodiester bonds, except the hydrogen bonds $\mathbf{b}_i, \mathbf{b}_{i+1}, \ldots, \mathbf{b}_{i+\mathbf{z}-1}$.

• *Step 2 Obtain the sequence of each motif* The sequence of the motif includes all the paired nucleotides and unpaired nucleotides in the motif that are aligned in order of index.

Calculating and encoding free energy The free energy of a molecule is the sum of the free energies calculated at the sub-structure (motif) level [18]. In particular, the free energy of a motif in RNA is a function of the involved bases. In our model, the type and the free energy of a motif are its two core features. Because the motif is composed of nucleotides, we assign the type and the free energy of a motif as the features for each nucleotide that forms it. Since a nucleotide could belong to two adjacent motifs, both of these two adjacent motifs' types and free energy values should be assigned to this nucleotide as features. To summarize all these features, we propose a 2D matrix, which we refer to as the *nucleotide localized information matrix*, to encode the motifs' free energy, motifs' types, and nucleotides' types for each nucleotide in the RNA molecule. The size of nucleotide localized information matrix is $L \times 5$, where L is the length of RNA sequence and 5 is the number of motif-driven features.

After representing the categorical variables by one-hot encoding, the nucleotide localized information matrix becomes a $L \times 18$ matrix where 4 columns are for 4-elements nucleotide one-hot encoding, 12 columns are for two 6-elements motif one-hot encoding, and 2 columns are for free energy of two motifs. We pad 0 in the bottom of each nucleotide localized information matrix to uniform the size of all nucleotide localized information matrixes. As a result, each nucleotide localized information matrix has a size of $\mathcal{L} \times 18$ ($\mathcal{L}$ has the same meaning as defined in section "Nucleotide-level

features-informed residual neuralnetwork model (NU-ResNet)"). Different from the binary valued features from the categorical variables, the free energy of motifs are numerical in nature. To prevent learning difficulties, we rescale these values to the [0, 1] interval by utilizing the cumulative distribution function (CDF) of normal distribution with μ and σ, where μ is the mean of normal distribution and σ is the standard deviation of normal distribution. Here, the reason why we utilize CDF of normal distribution to rescale the free energy is that the rate of CDF of normal distribution converging to 0 and 1 can be controlled by the parameter σ. We wish the scaling function converging to 0 and 1 neither too fast nor too slow. In other words, we expect to customize a scaling function so that it can be more sensitive to the different free energy values. The nature of CDF of normal distribution exactly satisfies this requirement.

Example An example of nucleotide localized information matrix is shown in Table 1. The fictitious RNA used in this example is same with the one used in Fig. 2.

Architecture design To account for the novel features, we revise the architecture proposed for NU-ResNet. Specifically, we have two inputs, 3D RNA matrix and nucleotide localized information matrix, for each RNA. We develop a neural network model which uses two parallel ResNet-18 with removing the last fully connected layer to generate features from each input and concatenates these generated features. The concatenated features are followed by stacked fully connected layers to perform classification. We name this revised ResNet-18 as NUMO-ResNet. The detailed architecture of NUMO-ResNet is shown in Fig. 4.

When training NUMO-ResNet, all the hyperparameters are set as the same with NU-ResNet except the weight decay. The weight decay set for NUMO-ResNet is 0.1.

Results

In this section, we analyze the performance of NU-ResNet and NUMO-ResNet, and analyze them against the state-of-the-art approaches, ENTRNA presented in [12] and equilibrium probability from the ensemble of the RNA structures proposed in [32].

Section "Data sets" characterizes the data sets utilized in this research. Section "Models Comparison" presents the comparison of performance of NU-ResNet

Table 1 An example of nucleotide localized information matrix

nt	Motif1	Motif1 FE	Motif 2	Motif2 FE
C	None	0	None	0
A	Stack	– 210	None	0
G	Stack	– 150	Stack	– 210
G	Hairpin loop	590	Stack	– 150
A	Hairpin loop	590	None	0
G	Hairpin loop	590	None	0
C	Hairpin loop	590	None	0
U	Hairpin loop	590	Stack	– 150
C	Stack	– 150	Stack	– 210
U	Stack	– 210	None	0
U	None	0	None	0
C	None	0	None	0

In this table, nt and FE are the abbreviations of nucleotide and free energy, respectively

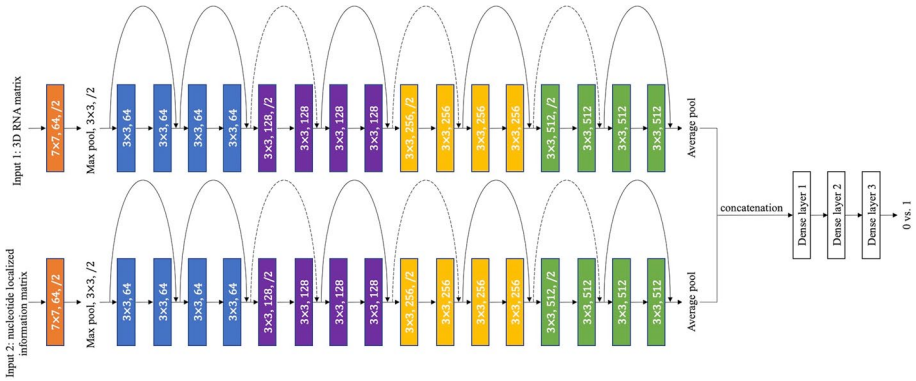


Fig. 4 The details of architecture of NUMO-ResNet

and NUMO-ResNet with data-driven approach, ENTRNA, and model-driven approach, equilibrium probability from the ensemble of the RNA secondary structures. Section "Analysis of NU-ResNet and NUMO-ResNet" introduces the analysis of NU-ResNet and NUMO-ResNet, including the comparison between NU-ResNet and NUMO-ResNet, convergence behavior of proposed models, and model robustness analysis of proposed models. Section "Performance of NU-ResNet and NUMO-ResNet across independent RNA families" shows the performance of testing NU-ResNet, NUMO-ResNet, ENTRNA, and Equilibrium Probability across independent RNA families.

Data sets

The samples utilized in this research are extracted from Protein Data Bank (PDB), in particular from the RNA STRAND database [43]. This is a commonly used data set in the literature [9, 12, 20]. When generating the dataset for our analysis, we only consider RNAs validated by X-Ray or NMR, thus ensuring the availability of the ground truth for each sequence. Synthetic RNAs and RNAs with pseudoknots are not considered.

Both our deep learning models require positive and negative samples for training. While the positive samples are the RNA sequence-structure pairs in the data set, we use the Positive-Unlabeled (PU) Learning method [12, 44] to generate multiple negative samples for the same RNA structure. Specifically, we use RNAinverse [45] and incaRNAtion [46] to generate 101 negative sequence candidates for each RNA structure. Not all the generated sequences are accepted as negative samples. Similar to the approach in [12, 47], we accept negative samples that satisfy three requirements:

- *repetition constraint*: as first requirement, we ask that any sub-sequence of an RNA sequence can have at most r consecutive identical nucleotides. In this analysis, we set $r = 6$;
- The second constraint is that the only allowed base pairs in RNA sequence is AU, CG, and GU;
- The third constraint is that the most or least occurring nucleotide within the sequence is either G or C.

These three constraints can help us to select the negative samples that are "reliable". Upon screening the candidates against the constraints, we calculate the five features, Normalized Sequence Segment Entropy with Segment Size 3, GC Percentage, Ensemble Diversity, Expected Accuracy, Pseudoknot-free RNA normalized free energy, proposed in [12] and compute the Euclidean distance between each negative sample candidate and the corresponding positive sample. The negative sample candidate with the largest distance from the positive sample is finally selected and included in the data set. As a result, each RNA structure has associated a *positive* and a *negative* sequence.

The positive samples that are the ground truth within the data set have the base pairs other than AU, CG, and GU pairs. In order to keep consistency with negative samples, we only consider the AU, CG, and GU pairs in the positive samples. Otherwise, whether having base pairs except AU, CG, and GU pairs will become a main feature to classify the positive samples and negative samples, which is not what we expect.

The longest RNA sequence within the data set utilized in this research has 408 nucleotides, and the shortest RNA sequence has 12 nucleotides. To unify the size of inputs of NU-ResNet and NUMO-ResNet, we pad both the 3D RNA matrix and the nucleotide localized information matrix with 0. We choose to use 410 as the maximum length $\mathcal{L}$, resulting in 3D RNA matrixes with size $[410 \times 410 \times 4]$ and nucleotide localized information matrixes with size $[410 \times 18]$. When utilizing the CDF of normal distribution to rescale free energy to the $[0, 1]$ interval in the nucleotide localized information matrix, we set $\mu = 0$ and $\sigma = 5$.

In this work, the RNA sequence-structure pairs are randomly selected to generate the training (TrDS, with 81% of the inputs), validation (VDS, with 9% of the inputs), and testing (TeDS, with 10% of the inputs) datasets. This data split setting is the same as proposed in PreRBP-TL [48]. The TrDS, VDS, and TeDS have 259 RNAs, 29 RNAs, and 32 RNAs, respectively. Considering the negative samples results in 518, 58, and 64 samples, respectively.

In the following, we analyze the performance of both our models with the associated largest validation accuracy and lowest validation loss. For these models, we also show the 10-fold CV performance under the combined TrDS, VDS, and TeDS datasets.

Models comparison

In order to evaluate the performance of NU-ResNet and NUMO-ResNet, we compare them with data-driven approach, ENTRNA [12], and model-driven approach, equilibrium probability proposed in [32]. The data-driven approach uses the Machine Learning to develop the model where the RNA sequence-secondary structure pairs are encoded by using feature engineering and the parameters of the model are learnt from the training of the model. The model-driven approach develops the model based on the Physics knowledge. Specifically, the ENTRNA evaluates an RNA sequence-secondary structure pair based on its features, while the equilibrium probability approach evaluates an RNA sequence-secondary structure pair based on its free energy. In section "Models comparison with data-driven approach" introduces the comparison of NU-ResNet and NUMO-ResNet with data-driven approach, ENTRNA. Section "Models comparison with model-driven approach" introduces the comparison of NU-ResNet and NUMO-ResNet with model-driven approach, equilibrium probability.

Models comparison with data-driven approach

Performance Metrics

Since NU-ResNet, NUMO-ResNet, and ENTRNA are trained as binary classification models, we propose several metrics to comprehensively analyze the trained architectures. The metrics we utilize include the area under curve receiver operator characteristic (AUCROC), the Matthews correlation coefficient (MCC), accuracy, precision, recall, and specificity. The AUCROC has a threshold invariant characteristic which can comprehensively evaluate the classification model. In addition, AUCROC has been proven to be equivalent to the probability that a randomly chosen positive sample can be ranked higher than a randomly chosen negative sample by the classification model [49]. Here, $MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, $\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP}$, $\text{Precision} = \frac{TP}{TP+FP}$, $\text{Recall} = \frac{TP}{TP+FN}$, and $\text{Specificity} = \frac{TN}{TN+FP}$. Within these formulas, *TP*, *TN*, *FP*, and *FN* refer to the number of true positive, true negative, false positive, and false negative respectively. AUCROC, accuracy, precision, recall, and specificity are all defined in the range $[0, 1]$. The MCC metric is defined in the range $[-1, 1]$. For all of the metrics, higher value indicates better performance.

NU-ResNet and NUMO-ResNet compared to ENTRNA

As previously mentioned, we record the models with best validation accuracy and best validation loss resulting from training. The set of parameters of the model with best validation accuracy and best validation loss is referred to as $\boldsymbol{\vartheta}_a^*$ and $\boldsymbol{\vartheta}_\ell^*$ respectively. The performance of the resulting NU-ResNet and NUMO-ResNet on the TeDS is shown in Table 2. We retrain and test the state-of-the-art RNA sequence-secondary structure pair evaluation model, ENTRNA, on TrDS and TeDS, respectively. The performance of ENTRNA on the TeDS is also shown in Table 2. It can be observed how all of four models outperform ENTRNA on the TeDS.

In Table 2, we observe that NU-ResNet with $\boldsymbol{\vartheta}_a^*$ outperforms ENTRNA and on all metrics except for the recall where they achieve the same performance. NU-ResNet with $\boldsymbol{\vartheta}_\ell^*$ outperforms ENTRNA on 5 out of 6 metrics (i.e. accuracy, AUCROC, MCC, precision and specificity). In addition, NUMO-ResNet models with $\boldsymbol{\vartheta}_\ell^*$ outperform ENTRNA on all metrics. The NUMO-ResNet with $\boldsymbol{\vartheta}_a^*$ outperforms ENTRNA on 5 out of 6 metrics (i.e. accuracy, AUCROC, MCC, precision, and specificity). The performance of NU-ResNet and NUMO-ResNet is superior to the performance of ENTRNA

Table 2 Models performance on TeDS

Metric	NU-ResNet		NUMO-ResNet		ENTRNA
	$\boldsymbol{\vartheta}_a^*$	$\boldsymbol{\vartheta}_\ell^*$	$\boldsymbol{\vartheta}_a^*$	$\boldsymbol{\vartheta}_\ell^*$	
Accuracy	93.75%	92.19%	90.63%	96.88%	73.44%
AUCROC	0.9736	0.9824	0.9824	0.9912	0.7275
MCC	0.875	0.8442	0.8141	0.9375	0.5130
Precision	93.75%	93.55%	93.33%	96.88%	66.67%
Recall	93.75%	90.63%	87.5%	96.88%	93.75%
Specificity	93.75%	93.75%	93.75%	96.88%	53.13%

Models with parameters that optimize the validation accuracy are referred to as $\boldsymbol{\vartheta}_a^*$, while models with parameters that optimize the validation loss are referred to as $\boldsymbol{\vartheta}_\ell^*$

except the recall. Here, the low precision, low specificity, and high recall of ENTRNA indicates that the model is more inclined to classify the sample as positive.

From the model comparisons among NU-ResNet, NUMO-ResNet, and ENTRNA. We have following two conclusions.

- The overall classification performance of NU-ResNet or NUMO-ResNet is superior to ENTRNA on TeDS.
- The experiments indicate the effectiveness of the methods utilized by NU-ResNet and NUMO-ResNet to encode the RNA sequence-structure pair.

Models comparison with model-driven approach

Equilibrium probability

According to the method proposed in [32], the equilibrium probability is defined with respect to the set of all pseudoknot-free RNA structures for a given RNA sequence [32]. Specifically, the authors define the equilibrium probability as

$$p(str_i) = \frac{\exp(-\frac{E(str_i)}{RT})}{\sum_{i=1}^{N} \exp(-\frac{E(str_i)}{RT})},$$

where str_i is i-th RNA structure, $E(str_i)$ is the associated free energy of str_i, N is the number of all RNA structures in ensemble. Finally, R is the gas constant and T is the thermodynamic temperature [22]. From this formula, we observe that the RNA structure with lower energy has higher equilibrium probability for a given sequence.

In this analysis, we adopt the ViennaRNA package [10] to calculate the equilibrium probability of the RNA sequence-structure pairs. We firstly use **mfe()** to obtain the MFE structure for a given RNA sequence and the corresponding free energy of this MFE structure. Then, we use **exp_params_rescale()** with setting the parameter equal to MFE value to rescale Boltzmann factor for computing partition function. Finally, we use **pf()** and **pr_structure()** to cauculate the partition function and the associated equilibrium probability for the given RNA structure, respectively.

NU-ResNet and NUMO-ResNet compared to Equilibrium Probability

We calculate the equilibrium probability based on the ensemble for all data on TeDS. In TeDS, there are 32 RNAs in total. On 13 out of these 32 RNAs, the equilibrium probability of their corresponding positive samples are 0. This is because that their free energies are much greater than the free energies of associated RNAfold [10] predicted structures which are obtained by approximating MFE. For 32 positive samples in TeDS, there are 22 samples whose free energies that are greater than the free energies of the associated RNAfold predicted structures. And among these 32 positive samples, only 7 positive samples' structures are same with the corresponding RNAfold predicted RNA structures, which is 21.88%. In terms of the free energy comparison between the positive sample and negative sample of each RNA in TeDS, 28 out of 32 RNAs whose corresponding positive sample's free energy is less than

the corresponding negative sample's free energy. In other words, using free energy value to directly classify the positive sample and negative sample has 87.5% accuracy, which is still lower than the accuracy of NU-ResNet or NUMO-ResNet. By removing all the zero and extremely small values of equilibrium probability, there are 7 RNAs within the TeDS having associated positive sample's equilibrium probability less than the associated negative sample's equilibrium probability, which means 21.88% samples are evaluated wrongly. For the 7 RNAs whose positive samples have free energies equaling to the free energy of RNAfold predicted structures, the equilibrium probability classifies all the 7 associated positive samples correctly and 3 out of 7 associated negative samples correctly when threshold equals to 0.5. When setting the threshold as 0.6, the equilibrium probability classifies all the associated 7 positive samples and 7 negative samples correctly. Hence, the equilibrium probability has admirable performance when dealing with the RNAs whose ground truth structures are same with RNAfold predicted structures in nature. However, when dealing with the RNAs whose ground truth structures are not same with RNAfold predicted structures, the equilibrium probability has the limitation. The data-driven approach is a direction to overcome this limitation. Therefore, our data-driven approaches, NU-ResNet and NUMO-ResNet, are good complement to RNA sequence-secondary structure evaluation research field.

The significant difference in the performance between four proposed models and equilibrium probability is mainly from the different mechanisms between these two types of the approaches. The NU-ResNet and NUMO-ResNet are data-driven approaches. However, the equilibrium probability is Physics-based approach. In these 32 RNAs, there are two positive samples whose equilibrium probabilities are greater than 1. This is because that these two positive samples have AC pair in their structures which leads to their free energies are less than the corresponding ensemble free energies. The NU-ResNet and NUMO-ResNet neglect the base pairs other than AU, CG, as well as GU pairs and limit the output score ranging from 0 to 1. The equilibrium probability only considers the AU, CG, and GU pairs when they build the ensemble. However, when dealing with some ground truth RNAs which have base pairs other AU, CG, and GU pairs, these ground truth RNAs could have free energies less than that of ensemble, which causes that the corresponding equilibrium probability is greater than 1. The advantage of data-driven approaches is that they learn the knowledge from the data, which can benefit the domain by using the knowledge learnt from big data.

From the model comparisons among NU-ResNet, NUMO-ResNet, and equilibrium probability. We have following two conclusions.

- The data-driven approaches, NU-ResNet and NUMO-ResNet, can learn the knowledge from the data source directly. The model-learnt knowledge is able to benefit the RNA evaluation domain.
- The data-driven approaches, NU-ResNet and NUMO-ResNet, are good complement to RNA sequence-secondary structure pair evaluation field because purely using free energy to evaluate RNA sequence-secondary structure pair has limitation in classification performance. The good classification performance can benefit the RNA secondary structure prediction and RNA inverse folding field.

Analysis of NU-ResNet and NUMO-ResNet

Comparison between proposed models

Table 2 shows that NUMO-ResNet with $\boldsymbol{\vartheta}_{\ell}^{*}$ outperforms NU-ResNet with $\boldsymbol{\vartheta}_{a}^{*}$ and $\boldsymbol{\vartheta}_{\ell}^{*}$ in all 6 metrics. In AUCROC, the NUMO-ResNet with $\boldsymbol{\vartheta}_{a}^{*}$ is superior to NU-ResNet with $\boldsymbol{\vartheta}_{a}^{*}$ and is equal to NU-ResNet with $\boldsymbol{\vartheta}_{\ell}^{*}$.

The results of the experiments follow our expectation because NUMO-ResNet incorporates more features compared to NU-ResNet. Intuitively, NUMO-ResNet should at least have the same performance with NU-ResNet because NUMO-ResNet has the whole input that NU-ResNet has. The results of experiments also show the advance of motif-based features extracted by NUMO-ResNet compared to the input only incorporating sequence and structure information employed by NU-ResNet.

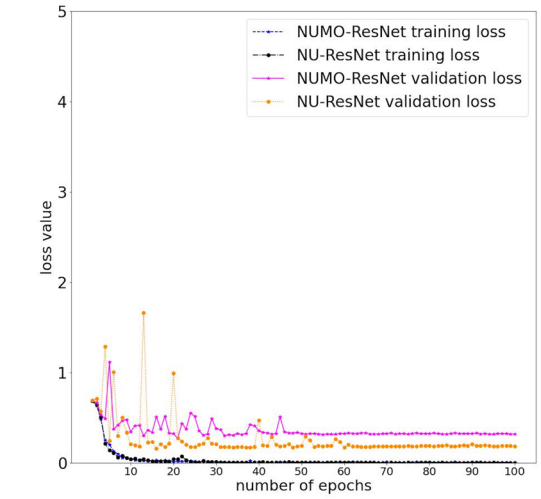
Convergence behavior of NU-ResNet training compared to NUMO-ResNet training Here, we provide insights into the training process of the proposed models. In particular, we analyze the validation loss and accuracy metrics as a function of the training effort (i.e., number of epochs). Figure 5a shows that NU-ResNet validation loss presents larger fluctuation than NUMO-ResNet. Figure 5b confirms this observation when accuracy is considered: NU-ResNet has larger fluctuation in validation accuracy compared to NUMO-ResNet. This finding suggests that the motif-based features extracted by NUMO-ResNet do have positive effects on model when it learns the RNA data because its validation loss and validation accuracy are more stable compared to NU-ResNet.

Models robustness analysis Since we utilize a weighted sampler to sample the data during the training which has randomness, the performance of trained models on testing data may be affected by this randomness. To verify the robustness of the trained models, we perform a 10-fold CV on both NU-ResNet and NUMO-ResNet.

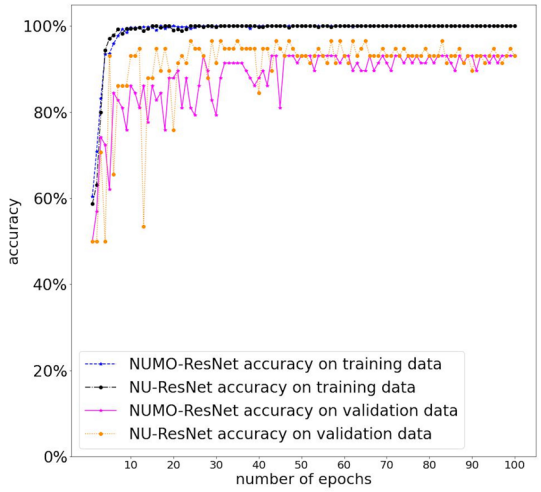
Similar to the previous analysis, for each iteration of the validation routine, we consider two models, one with the best validation accuracy, and one with the best validation loss. Table 3 shows the 10-fold CV results from our models as the average of the performance resulting from 10 iterations of the approach. The 10-fold CV results presented in Table 3 confirm that NU-ResNet and NUMO-ResNet are capable of tackling different groups of RNAs across the data set used in this research.

Performance of NU-ResNet and NUMO-ResNet across independent RNA families

Inspired by the findings introduced in [50], we conduct the experiments to analyze the performance of NU-ResNet and NUMO-ResNet across independent RNA families. Specifically, we train and validate the NU-ResNet and NUMO-ResNet only on Transfer RNA and Ribosomal RNA because they have first two largest data sizes compared to other RNA families in the data set utilized in this research. The training and validation data have 118 RNAs and 14 RNAs respectively. By considering the negative samples, there are 236 and 28 samples in training data and validation data respectively. The ratio between training data and validation data is consistent with the ratio between TrDS and VDS in section "Data sets". Then we test the trained models on all remaining RNA families individually. In addition to RNA families within the PDB data set we utilize in this research, we also include the Riboswitch data from [51] as an independent testing data set. The statistics of each RNA family are summarized in Table 4. We also retrain



(a)



(b)

Fig. 5 **a**: The training and validation loss. **b**: The training and validation accuracy

Table 3 10-fold cross validation results from NU-ResNet and NUMO-ResNet

Metric	NU-ResNet		NUMO-ResNet	
	ϑ_a^*	ϑ_ℓ^*	ϑ_a^*	ϑ_ℓ^*
Accuracy	98.13 %	97.19%	95.47%	93.44%
AUCROC	0.9939	0.9948	0.9749	0.9768
MCC	0.9629	0.9444	0.9111	0.8702
Precision	98.16%	97.54%	96.91%	93.48%
Recall	98.13%	96.88%	94.06%	93.75%
Specificity	98.13%	97.5%	96.88%	93.13%

Models with parameters that optimize the validation accuracy are referred to as ϑ_a^*, while models with parameters that optimize validation loss are referred to as ϑ_ℓ^*

Table 4 Nanoparticles length range grouped by family

Family (abbreviation)	Number	Length range
Transfer RNA (t)	60	13–152
5S Ribosomal RNA (5Sr)	12	24–122
16S Ribosomal RNA (16Sr)	2	14–14
23S Ribosomal RNA (23Sr)	7	19–27
other Ribosomal RNA (or)	51	12–408
Group I intron (GI)	3	18–22
Group II intron (GII)	4	27–70
Signal Recognition Particle RNA (SRP)	10	28–192
Viral and Phage (V &P)	18	12–38
Small Nuclear RNA (sn)	6	20–66
Ribonuclease P RNA (RNP)	15	21–270
Internal Ribosome Entry Site (IR)	4	14–30
Hairpin Ribozyme (HpR)	11	38–226
Hammerhead Ribozyme (HhR)	9	40–82
Riboswitch (Rbo)	4	71–154
other Ribozyme (oR)	17	17–159
other RNA (O)	91	16–304
Total	324	12–408

Table 5 Models performance across RNA families on Accuracy, AUCROC, and MCC

Ctgy	NU-ResNet			NUMO-ResNet			ENTRNA		
	Acc	AUC	MCC	Acc	AUC	MCC	Acc	AUC	MCC
GI	50%	0.44	0	50%	0.33	0	50%	0.33	0
GII	75%	1	0.5774	62.5%	0.8125	0.2582	62.5%	0.8125	0.3780
SRP	100%	1	1	90%	0.96	0.8	66.67%	0.6790	0.3333
V &P	63.89%	0.8364	0.2817	52.78%	0.7037	0.0642	66.67%	0.6790	0.3333
sn	66.67%	1	0.4472	50%	0.4722	0	66.67%	0.6111	0.3536
RNP	100%	1	1	83.33%	0.88	0.7071	63.33%	0.76	0.3922
IR	100%	1	1	75%	1	0.5774	75%	0.75	0.5774
HpR	90.91%	0.9835	0.8321	77.27%	0.9504	0.5669	50%	0.9917	0
HhR	100%	1	1	55.56%	0.7778	0.1140	50%	0.6914	0
oR	97.06%	0.9792	0.9428	64.71%	0.6747	0.3333	67.65%	0.7474	0.4629
Rbo	50%	–	0	25%	–	0	75%	–	0
O	94.51%	0.9546	0.8921	80.22%	0.8721	0.6140	65.93%	0.7313	0.3349
C	86.36%	0.9458	0.7390	66.67%	0.7862	0.3658	62.12%	0.6889	0.2920
C+	90.26%	0.9520	0.8114	73.16%	0.8272	0.4846	63.95%	0.7085	0.3102

Here, the NU-ResNet and NUMO-ResNet are $\hat{\theta}_L^*$. C refers to the combination of all RNA categories in this table except the other RNA (O). C+ refers to the combination of C and other RNA (O). Ctgy, Acc, and AUC are the abbreviations of category, Accuracy, and AUCROC, respectively

and test the ENTRNA in same data sets for data-driven approaches' comparison. The results are shown in Tables 5 and 6. In addition, we test the equilibrium probability for comparing the NU-ResNet and NUMO-ResNet with model-driven approach. In addition to RNA families within the PDB data set we utilize in this research, we also include the Riboswitch data from [51] as an independent testing data set.

Data-driven models performance across RNA families

Table 6 Models performance across RNA families on Precision, Recall, and Specificity

Ctgy	NU-ResNet			NUMO-ResNet			ENTRNA		
	Pre	Rec	Spe	Pre	Rec	Spe	Pre	Rec	Spe
GI	0%	0%	100%	0%	0%	100%	50%	100%	0
GII	100%	50%	100%	66.67%	50%	75%	57.14%	100%	25%
SRP	100%	100%	100%	90%	90%	90%	66.67%	66.67%	66.67%
V &P	66.67%	55.56%	72.22%	55.56%	27.78%	77.78%	66.67%	66.67%	66.67%
sn	100%	33.33%	100%	50%	33.33%	66.67%	62.5%	83.33%	50%
RNP	100%	100%	100%	100%	66.67%	100%	57.69%	100%	26.67%
IR	100%	100%	100%	100%	50%	100%	66.67%	100%	50%
HpR	100%	81.82%	100%	87.5%	63.64%	90.91%	50%	100%	0%
HhR	100%	100%	100%	57.14%	44.44%	66.67%	50%	100%	0%
oR	100%	94.12%	100%	77.78%	41.18%	88.24%	60.71%	100%	35.29%
Rbo	100%	50%	–	100%	25%	–	100%	75%	–
O	97.65%	91.21%	97.80%	86.67%	71.43%	89.01%	62.18%	81.32%	50.55%
C	94.05%	78.22%	94.85%	77.78%	48.51%	85.57%	58.13%	92.08%	30.93%
C+	95.86%	84.38%	96.28%	82.61%	59.38%	87.23%	59.86%	86.98%	40.43%

Here, the NU-ResNet and NUMO-ResNet are ϑ_ℓ^*. C refers to the combination of all RNA categories in this table except the other RNA (O). C+ refers to the combination of C and other RNA (O). Ctgy, Pre, Rec, and Spe are the abbreviations of category, Precision, Recall, and Specificity, respectively

We select from the PDB data set the Transfer RNA, 5S Ribosomal RNA, 16S Ribosomal RNA, 23S Ribosomal RNA, and other Ribosomal RNA to from the training data set. We use the same hyperparameters setting introduced in section "Methods" to retrain the NU-ResNet and NUMO-ResNet. Then we test the resulting NU-ResNet and NUMO-ResNet models "out of sample" on Group I intron, Group II intron, SRP RNA, Viral and Phage, Small Nuclear RNA, Ribonuclease P RNA, Internal Ribosome Entry Site, Hairpin Ribozyme, Hammerhead Ribozyme, Riboswitch, other Ribozyme, and other RNA individually. We also retrain and test ENTRNA using the same data sets. Because the model complexity of NUMO-ResNet is higher than that of NU-ResNet, more training data are expected for NUMO-ResNet compared to NU-ResNet. However, in order to avoid the overlap RNA families between training data and testing data, we need to exclude the other RNA from the training data set, which causes that the size of the training data decreased compared to the TrDS.

In Table 5, there are 11 testing RNA families in total. The NU-ResNet has better or equal performance compared to ENTRNA in 9, 9, and 10 testing RNA families based on accuracy, AUCROC, and MCC, respectively. The NUMO-ResNet has better or equal performance compared to ENTRNA in 7, 7, and 7 testing RNA families based on accuracy, AUCROC, and MCC, respectively.

Table 6 shows that both of NU-ResNet and NUMO-ResNet have balanced performance in Precision, Recall, and Specificity across all testing RNA families except the Group I intron, which implies that NU-ResNet and NUMO-ResNet are not biased when they are tested on most of these new RNA families. However, ENTRNA has the biased performance on Group I intron, Hairpin Ribozyme, and Hammerhead Ribozyme. In terms of the model performance on handling both of positive samples and negative samples across different RNA families, the NU-ResNet and NUMO-ResNet show more balanced capability than ENTRNA.

Tables 5and 6 show that both the NU-ResNet and NUMO-ResNet models outperform the competitors over the SRP RNA, Ribonuclease P RNA, Internal Ribosome Entry Site, and Hammerhead Ribozyme families.

The results for the aggregate data sets "C" and "C+" in Table 5 shows that NU-ResNet and NUMO-ResNet outperform ENTRNA across all metrics. Because "C" data set and the training data have no overlap RNA families, the NU-ResNet still has the AUCROC as 0.9458, which is consistent with the AUCROC in Table 2. This shows the generalizability of NU-ResNet. On the other hand, NUMO-ResNet shows a decrease in the performance. We believe such difference is not due to lesser generalizability of the model, rather to the reduced size of the training data set compared to the TrDS and the larger number of parameters required by NUMO-ResNet as compared to the NU-ResNet model.

By testing the NU-ResNet, NUMO-ResNet, and ENTRNA across different RNA families, we can obtain the following conclusions.

- The overall testing performance of NU-ResNet and NUMO-ResNet across different RNA families is superior to ENTRNA.
- The experiments show that the NU-ResNet has the admirable performance when handling the data from the new RNA families.

Equilibrium probability performance across RNA families

In order to compare NU-ResNet and NUMO-ResNet with equilibrium probability in handling different RNA families, we test the equilibrium probability in each data set listed in Table 5. The performance of equilibrium probability in different RNA families is shown in Table 7. Specifically, in Table 7, there are in total 80.73% RNAs whose positive

Table 7 The performance of equilibrium probability across RNA families

Families	$FE_{pos} > FE_{RNAfoldstr}$	$FE_{pos} = FE_{RNAfoldstr}$	$FE_{pos} < FE_{neg}$
GI	100%	0%	33.33%
GII	100%	0%	100%
SRP	90%	0%	90%
V &P	77.78%	22.22%	83.33%
sn	83.33%	0%	83.33%
RNP	86.67%	13.33%	100%
IR	75%	0%	50%
HpR	100%	0%	36.36%
HhR	100%	0%	77.78%
oR	70.59%	17.65%	64.71%
Rbo	75%	25%	–
O	75.82%	7.69%	83.52%
C	85.15%	9.90%	–
C+	80.73%	8.85%	–

Here the column $FE_{pos} > FE_{RNAfoldstr}$ refers to the percentage of RNAs whose positive sample has greater free energy than the free energy of the corresponding RNAfold predicted structure. The column $FE_{pos} = FE_{RNAfoldstr}$ refers to the percentage of RNAs whose positive sample has the same free energy as the corresponding RNAfold predicted structure. The column $FE_{pos} < FE_{neg}$ refers to the percentage of RNAs whose positive sample has lower free energy than the negative sample. C refers to the combination of all RNA categories in this table except the other RNA (O). C+ refers to the combination of C and other RNA (O). Ctgy is the abbreviation of category

sample has the free energy greater than the free energy of the corresponding RNAfold predicted structures.

Table 7 shows in column $FE_{pos} < FE_{neg}$ the percentage of RNAs that exhibit positive samples with lower free energy compared with the negative samples within each RNA family. As an example all the data within the Group II intron and Ribonuclease P RNA families exhibit this property. As a result, a classifier that uses free energy to distinguish positive from negative samples (like the equilibrium probability does) will achieve a good performance. On the other hand, families such as Group I intron and Hairpin Ribozyme exhibit this property for only 33.33% and 36.36% of the cases. Considering only RNAs which have negative samples in the "C" and "C+", we observe that 75.26% and 79.26% exhibit positive samples with lower free energy compared to the negative sample, thus resulting in an accuracy of 75.26% and 79.26% for the equilibrium probability method. This performance is superior to both ENTRNA and NUMO-ResNet, while NU-ResNet still shows the best results.

From the testing performance of NU-ResNet, NUMO-ResNet and Equilibrium Probability across different RNA families. We have the following conclusions.

- Using free energy to evaluate the RNA sequence-secondary structure pairs has the different performance in different RNA families. The NU-ResNet and NUMO-ResNet have more consistent performance in different RNA families.
- Leveraging the knowledge learnt from data-driven approaches can benefit the classification performance of the models across independent RNA families.

Discussion

In this work, we propose two deep learning models, NU-ResNet and NUMO-ResNet, to evaluate the pair of RNA sequence and RNA secondary structure. And we propose two matrixes, 3D RNA matrix and nucleotide localized information matrix, to encode the RNA sequence-secondary structure pairs. The 3D RNA matrix can be used to explicitly encode the information of the RNA sequence and RNA structure. And the nucleotide localized information matrix incorporates the motif and free energy information of RNA. The NU-ResNet and NUMO-ResNet exhibit the distinguished performance in the experiments. And they outperform the state-of-the-art data-driven RNA sequence-secondary structure pair evaluation model, ENTRNA, and physics-based model, equilibrium probability from the ensemble, in an independent testing data set. The 10-fold CV results show the robustness of NU-ResNet and NUMO-ResNet. The high level embeddings from NU-ResNet or NUMO-ResNet are model-automatically-extracted-features of RNA, which can be used in other downstream work in the future research. NU-ResNet and NUMO-ResNet show a consistent performance when they are being tested across independent RNA families. The NU-ResNet model works especially well across different RNA families, which shows that the model has the robust generalization ability even when handling the data from new RNA families.

There are several avenues for further to explore. Firstly, we can enhance the model development by incorporating RNA with pseudoknots. Secondly, we can devise novel methods for generating higher-quality negative samples. Thirdly, we can introduce confidence interval for the prediction results obtained from NU-ResNet and NUMO-ResNet.

Conclusion

In this research, we present a deep learning framework for the evaluation of RNA sequence-structure pairs. Within this framework, we introduce two models, NU-ResNet and NUMO-ResNet. Both models exhibit superior performance compared to state-of-the-art RNA sequence-structure pair evaluation models across multiple metrics. The two models rely on different inputs. Particularly, NUMO-ResNet incorporates motif-based features, which enhance the training process stability to a considerable degree. The NU-ResNet model shows a robust generalization ability when handling the data from new RNA families. It is important to highlight that this study exclusively focuses on pseudoknot-free structures. Our future efforts will involve addressing pseudoknotted RNA structures as well as evaluating RNA sequence-tertiary structure pairs, and the outcomes will be reported separately. Furthermore, we plan to explore incorporating uncertainty quantification techniques into our models in future, further enhancing their reliability and robustness.

Acknowledgements
The authors thank Arizona State University Research Computing for providing us with the computational nodes.

Author Contributions
YZ contributed to the code, algorithms, formulation of the research problem, experiments, experimental analysis, and main writing of the manuscript. GP led the project, contributed to the formulation of the research problem as well as the main writing of the manuscript, and supervised the model development and experimental analysis. TW contributed to the formulation of the research problem as well as the review and edit of the manuscript, and supervised the model development and experimental analysis. FZ contributed to the domain knowledge consulting of Biology as well as the review and edit of the manuscript.

Funding
This research is supported by U.S. National Science Foundation grant 2007861.

Data availibility
The corresponding data for this research is available at https://github.com/yzhou617/NU_ResNet_NUMO_ResNet

Declarations

Competing interests
No Conflict of interest is declared.

Code availability
The corresponding source code for this research is available at https://github.com/yzhou617/NU_ResNet_NUMO_ResNet

Received: 7 March 2024 Accepted: 27 August 2024
Published online: 30 September 2024

References

1. Low JT, Weeks KM. Shape-directed RNA secondary structure prediction. Methods. 2010;52(2):150–8.
2. Brenner S. The ancient molecule. Nature. 1994;367:228–9.
3. Guo P. The emerging field of RNA nanotechnology. Nat Nanotechnol. 2010;5(12):833–42.
4. Oguro A, Ohtsu T, Nakamura Y. An aptamer-based biosensor for mammalian initiation factor eukaryotic initiation factor 4a. SAN DIEGO Elsevier Inc. 2009;388(1):102–107
5. Winkler WC, Breaker RR. Regulation of bacterial gene expression by riboswitches. PALO ALTO Annual Rev. 2005;59(1):487–517.
6. Jaeger L, Voss N, Bindewald E, Yaghoubian AJ, Shapiro BA, Afonin KA, Jacovetty E. In vitro assembly of cubic RNA-based scaffolds designed in silico. Nat Nanotechnol. 2010;5(9):676–82.
7. Pyle AM. Metal ions in the structure and function of RNA. J Biol Inorg Chem. 2002;7:679–90.
8. Tinoco I, Bustamante C. How RNA folds. J Mol Biol. 1999;293:271–81.
9. Singh J, Paliwal K, Zhang T, Singh J, Litfin T, Zhou Y, Gorodkin J. Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. Bioinformatics. 2021;37:2589–600.

10. Lorenz R, Bernhart SH, Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL Viennarna package 2.0. Algorithms for molecular biology 2011;6:26–26
11. Garcia-Martin JA, Clote P, Dotu I. Rnaifold: a web server for RNA inverse folding and molecular design. Nucleic Acids Res. 2013;41(W1):465–70.
12. Su C, Weir JD, Zhang F, Yan H, Wu T. Entrna: a framework to predict RNA foldability. BMC Bioinf. 2019;20:373–373.
13. Liu M, Poppleton E, Pedrielli G, Sulc P, Bertsekas DP. Expertrna: a new framework for RNA secondary structure prediction. INFORMS J Comput. 2022;34(5):2464–84.
14. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol. 1999;288(5):911–40.
15. Xia T, SantaLucia J, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watsonâcrick base pairs. Biochemistry. 1998;37(42):14719–35.
16. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP. Efficient parameter estimation for RNA secondary structure prediction. Bioinformatics. 2007;23(13):19–28.
17. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP. Computational approaches for RNA energy parameter estimation. RNA. 2010;16(12):2304–18.
18. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981;9(1):133–48.
19. Sato K, Akiyama M, Sakakibara Y. RNA secondary structure prediction using deep learning with thermodynamic integration. Nat Commun. 2021;12:941–941.
20. Singh J, Hanson J, Paliwal K, Zhou Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat Commun. 2019;10:5407–13.
21. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–59.
22. Zhang H, Zhang L, Mathews DH, Huang L. Linearpartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. Bioinformatics. 2020;36(1):258–67.
23. Zhang T, Singh J, Litfin T, Zhan J, Paliwal K, Zhou Y. Rnacmap: a fully automatic pipeline for predicting contact maps of RNAs by evolutionary coupling analysis. Bioinformatics. 2021;37(20):3494–500.
24. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue residue contact predictions in a sequence- and structure-rich era. In Proceedings of the National Academy of Sciences—PNAS. 2013;110(39):15674–9.
25. Bertsekas DP, Tsitsiklis JN. Neuro-dynamic programming 1996
26. Bertsekas DP, Tsitsiklis JN, Wu C. Rollout algorithms for combinatorial optimization. J Heuristics. 1997;3(3):245–62.
27. Bertsekas DP. Reinforcement learning and optimal control (2019).
28. Bertsekas, DP. Rollout, policy iteration, and distributed reinforcement learning (2020).
29. Zadeh JN, Wolfe BR, Pierce NA. Nucleic acid sequence design via efficient ensemble defect optimization. J Comput Chem. 2011;32(3):439–52.
30. Garcia-Martin JA, Clote P, Dotu I. Rnaifold: a constraint programming algorithm for RNA inverse folding and molecular design. J Bioinform Comput Biol. 2013;11(02):1350001.
31. Van Hentenryck P, Michel L. Constraint-based local search (2005).
32. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers: original research on biomolecules 1990;29(6–7), 1105–1119
33. He K, Zhang X, Ren S. Sun, J. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016;2016:770–8.
34. Darty K, Denise A, Ponty Y. Varna: interactive drawing and editing of the RNA secondary structure. Bioinformatics. 2009;25(15):1974–5.
35. Antczak M, Popenda M, Zok T, Zurkowski M, Adamiak RW, Szachniuk M. New algorithms to represent complex pseudoknotted RNA structures in dot-bracket notation. Bioinformatics. 2018;34(8):1304–12.
36. Reuter JS, Mathews DH. Rnastructure: software for RNA secondary structure prediction and analysis. BMC Bioinf. 2010;11(1):129–129.
37. Alex K, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90.
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR (2015)
39. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. ICML, 807–814 (2010)
40. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv (2015.
41. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv (2014).
42. Magnus M, Antczak M, Zok T, Wiedemann J, Lukasiak P, Cao Y, Bujnicki JM, Westhof E, Szachniuk M, Miao Z. RNA-puzzles toolkit: a computational resource of RNA 3d structure benchmark datasets, structure manipulation, and evaluation tools. Nucleic Acids Res. 2019;48(2):576–88.
43. Andronescu M, Bereg V, Hoos HH, Condon A. RNA strand: The RNA secondary structure and statistical analysis database. BMC Bioinf. 2008;9(1):340–340.
44. Liu B, Dai Y, Li X, Lee WS, Yu PS. Building text classifiers using positive and unlabeled examples. In IEEE International Conference on Data Mining, ICDM, 179–186 (2003)
45. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatsh Chem. 1994;125(2):167–88.
46. Reinharz, V., Ponty, Y., WaldispÃ¼hl, J.: A weighted sampling algorithm for the design of rna sequences with targeted secondary structure and nucleotide distribution. Bioinformatics 29(13), 308–315 (2013)
47. Williams, S., Lund, K., Lin, C., Wonka, P., Lindsay, S., Yan, H.: Tiamat: a three-dimensional editing tool for complex dna structures. In DNA Computing: 14th International Meeting on DNA Computing, 90–101 (2008)
48. Zhang J, Yan K, Chen Q, Liu B. Prerbp-tl: prediction of species-specific RNA-binding proteins based on transfer learning. Bioinformatics. 2022;38(8):2135–43.

49. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27(8):861–74.
50. Szikszai M, Wise M, Datta A, Ward M, Mathews DH. Deep learning models for RNA secondary structure prediction (probably) do not generalize across families. Bioinformatics. 2022;38(16):3892–9.
51. Wayment-Steele HK, Kladwang W, Strom AI, Lee J, Treuille A, Becka A, Participants E, Das R. Rna secondary structure packages evaluated and improved by high-throughput experiments. Nat Methods. 2022;19(10):1234–42.

Publisher's Note