Deep generalizable prediction of RNA secondary structure via base pair motif energy

Heqin Zhu[1,2], Fenghe Tang[1,2], Quan Quan[3], Ke Chen[1,2], Peng Xiong[1,2*], S. Kevin Zhou[1,2,3,4*]

[1]School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, 230026, P.R. China.
[2]Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu, 215123, P.R. China.
[3]Key Laboratory of Intelligent Information Processing of Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, P.R. China.
[4]Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui, 230026, P.R. China.

*Corresponding author(s). E-mail(s): xiongxp@ustc.edu.cn; s.kevin.zhou@gmail.com;

Abstract

RNA secondary structure plays essential roles in modeling RNA tertiary structure and further exploring the function of non-coding RNAs. Computational methods, especially deep learning methods, have demonstrated great potential and performance for RNA secondary structure prediction. However, the generalizability of deep learning models is a common unsolved issue in the situation of unseen out-of-distribution cases, which hinders the further improvement of accuracy and robustness of deep learning methods. Here we construct a base pair motif library which enumerates the complete space of locally adjacent three-neighbor base pair and records the thermodynamic energy of corresponding base pair motifs through de novo modeling of tertiary structures, and we further develop a deep learning approach for RNA secondary structure prediction, named BPfold, which employs hybrid transformer and convolutional neural network architecture and an elaborately designed base pair attention block to jointly learn representative features and relationship between RNA sequence and the energy map of base pair

motif generated from the above motif library. Quantitative and qualitative experiments on sequence-wise datasets and family-wise datasets have demonstrated the great superiority of BPfold compared to other state-of-the-art approaches in both accuracy and generalizability. The significant performance of BPfold will greatly boost the development of deep learning methods for predicting RNA secondary structure and the further discovery of RNA structures and functionalities.

Keywords: RNA secondary structure prediction, Base pair motif, Base pair attention, Deep learning

1 Introduction

RNA secondary structure plays vital roles in modeling the RNA tertiary structure through long-range base pairing interactions [1], demonstrating the remarkable versatility and functional mechanisms of RNA in biological systems and cellular processes, such as catalytic functionality [2, 3], regulatory functions [4], and intro splicing events [5]. Generally, RNA secondary structure forms a sequential of stem and loop regions, with stem regions composed of canonical paired nucleotide bases [6, 7] and loop regions composed of unpaired bases. Furthermore, loop regions exhibit various patterns [8], such as hairpin loops, bulge loops, internal loops, junction loops, and pseudoknot [9].

Discovering the secondary structure of RNA is important and necessary for modeling the tertiary structure and exploring potentialities. Experimental methods such as X-ray crystallography [10], nuclear magnetic resonance spectroscopy [11], and cryogenic electron microscopy [12] achieve great success in detecting three-dimensional RNA structures and further figuring out the secondary structure of RNA [13, 14], but this procedure poses enormous difficulties due to the high experimental costs and resolution limits. As such, computational prediction of RNA secondary structures is an efficient and low-cost alternative.

During the past three decades, various computational methods have been developed to predict RNA secondary structures, such as comparative sequence analysis and thermodynamic models. Comparative sequence analysis [15–18] predicts structures by searching for homologous sequences, which is extremely effective and accurate when the target sequence is hit in the homologous sequences database. However, the number of known RNA families is a few thousand in Rfam [19, 20], resulting in poor generalizability and accuracy of the comparative methods for unknown sequences. Thermodynamic models [21–33] aim at finding the best structure from thermodynamically stable candidates that are selected through free energy minimization. These methods assign each structure with a score, with parameters obtained from experiments, such as Vienna RNAfold [26, 34], RNAstructure [27, 35], and Eternafold [28]). Thermodynamic energy-based models can be effectively solved via dynamic programming when the target secondary structure is only composed of nested base pairing. However, dynamic programming methods have problems dealing with non-nested pairs

(pseudo-knots), non-canonical base pairing, or other complex structures, reaching a performance ceiling at about 80% [36, 37].

Recently, deep learning methods [38–41, 41–43] raise significant attention, greatly boost the speed of prediction, and achieve high accuracy. As data-driven approaches, deep learning methods utilize the benefit of big data and learn the implicit features and intrinsic of data distribution in hidden space via deep neural networks. Once the neural network has learned to build the mapping and the relationship between input data (RNA sequence) and output data (RNA secondary structure), it can predict the secondary structure of an unknown arbitrary input RNA sequence. For instance, Singh et al. [40] propose SPOT-RNA, an ensemble of two-dimensional deep neural networks equipped with transfer learning on a high-quality dataset. Fu et al. [39] develop a U-shaped fully convolutional image-to-image network, named UFold, which converts an RNA sequence into an image-like representation to predict RNA secondary structure. Wang et al. [44] develop a multi-purpose RNA language model, RNAErnie [44], with motif-aware pre-training and type-guided fine-tuning, which demonstrates satisfying performance when applied to RNA secondary structure prediction task. Sato et al. [43] propose MXfold2, a novel method that learns RNA folding scores by integrating Turner's nearest-neighbor free energy parameters into deep neural networks.

Although existing deep learning (DL) methods behave well in currently known test datasets, their performances degrade rapidly as sequence similarity decreases in situation of unseen data compared to traditional methods [45, 46], which indicates poor generalizability and the possibility of overfitting on training datasets. To mitigate this, researchers resort to integrating auxiliary information into deep learning models. For example, UFold [39] uses an alternative representation of the RNA sequence derived from CDPfold [47] to enhance the relationship between the input RNA sequence and thermodynamic prior of base pairs, SPOT-RNA [40] takes full advantage of evolutionary information, and MXfold2 [43] regularizes the learning of the model with a penalty loss on the folding score for deviating too far from thermodynamic estimation. With the help of auxiliary information, these methods have made some progress in improving prediction accuracy. However, a general **data insufficiency** problem plaques RNA structure prediction, including secondary structure prediction. Unlike protein structure prediction, which possesses a sufficiently large number of data to represent the underlying distribution, which guarantees the effectiveness of DL methods such as AlphaFold [48, 49], *the number of available RNA structure data is very low.* Therefore, for RNA secondary structure prediction, how to train a reliable DL model under such data insufficiency and further to deal with out-of-distribution samples is an unsolved problem, hindering further improvements in accuracy and generalizability of DL learning models.

It is known that enriching data at secondary structure level is extremely hard for both experimental and computational methods. Instead, it is more computationally efficient to predict the tertiary structure of short-sequence RNA motifs. Luckily, RNA secondary structure is mainly dependent on the structure motifs [50]. Motivated by these and different from previous attempts that integrate knowledge prior into data-driven models incompletely, we propose to leverage the local short-distance

interactions of base pairs and enumerate the whole space of adjacent neighboring patterns of all canonical base pairs, named as base pair motif, aiming at enriching data at base-pair level completely.

In this paper, we propose BPfold, a deep learning model integrated with thermodynamic energy from the complete space of the upstream and downstream of three-neighbor base pair motifs for predicting RNA secondary structures. BPfold comprises two key components:

(1) *Base pair motif energy.* A base pair motif is a canonical base pair (i.e., A-U, U-A, G-C, C-G, G-U, and U-G) together with its local spatially adjacent bases, which dominates the local structure of the base pair. We explore the entire space of the base pair motifs of r neighbors and compute their energy by de novo modeling the tertiary structure of the base pair motif. After storing the computed tertiary structures and corresponding energy items in the motif library, we can quickly obtain the base pair motif energy for any base pair of any arbitrary input RNA sequence. Thereby, this auxiliary input energy fully covers the data distribution at base-pair level, mitigating current insufficient database, eliminating the major hurdle posed by generalizability of de novo DL models.

(2) *Base pair attention.* In BPfold neural network, we elaborately design a base pair attention block, which combines transformer [51] and convolution [52] layers and enables information integration between RNA sequence and base pair motif energy. The base pair attention block aggregates the attention map of the RNA sequence and the base pair motif energy to effectively learn the base pair knowledge from the RNA sequence. BPfold takes advantage of the deep learning approach, predicts the accurate RNA secondary structure in seconds, and has great generalizability that accounts for the learned knowledge of thermodynamic energy.

We conduct sequence-wise and family-wise cross-validation experiments on multiple benchmark datasets to evaluate the accuracy and generalizability of BPfold. ArchiveII [53] (3,966 RNAs) and bpRNA-TS0 [54] (1,305 RNAs) are sequence-wise datasets while Rfam14.5-14.10 [19, 20] (2,048 RNAs) contains cross-family RNA sequences and PDB [55] (116 RNAs) consists of high-quality experimentally validated RNA structures. Quantitative and qualitative results demonstrate the superiority of proposed BPfold in accuracy and generalizability against other deep learning methods and traditional methods. We expect that BPfold will take a meaningful step toward fast and robust prediction of RNA secondary structures.

2 Results

2.1 Overview of the BPfold approach.

In this work, we present BPfold (Fig. 1a), a novel deep learning approach integrated with thermodynamic energy for RNA secondary structure prediction. Aiming at improving the generalizability and accuracy of deep learning-based model, we compute the base pair motif energy and design a base pair attention neural network block. As demonstrated in Fig. 2, a base pair motif is a canonical base pair (i.e., A-U, U-A, G-C, C-G, G-U, and U-G) together with their neighboring bases. We compute the

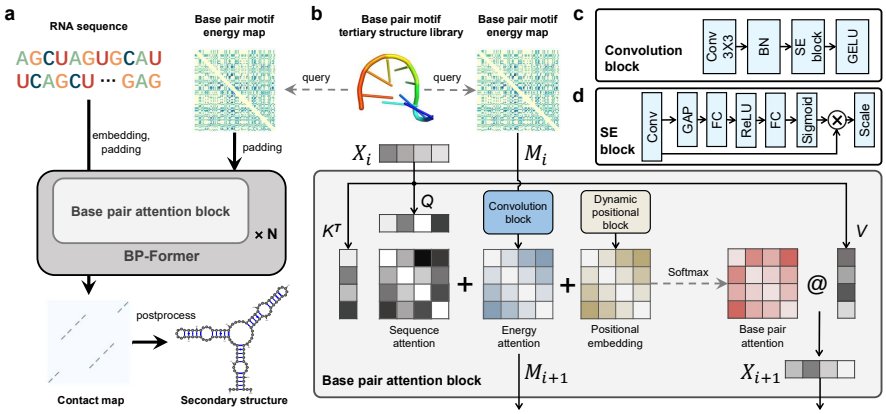

Fig. 1: **Overview of our BPfold approach for modeling RNA secondary structures**. **a** BPfold takes RNA sequence and corresponding base pair motif energy map generated from base pair motif library as inputs, consisting of transformer blocks with designed base pair attention, and outputs contact map. After applying physical constraints to the contact map in post-processing procedures, we obtain the final predicted secondary structure. **b** The detailed structure of proposed base pair attention block which jointly fuses the sequence features and energy matrix features for enhanced learning of base pair interactions. **c** The detailed structure of convolution block in base pair attention block. **d** The detailed structure of squeeze & excitation (SE) block in convolution block.

de novo RNA tertiary structure and obtain its thermodynamic energy of the complete space of r-neighbor base pair motif to mitigate the insufficient coverage of base pair in existing datasets. Furthermore, as illustrated in Fig. 1b, we equip BPfold with custom-designed base pair attention block (described in Section 4.2), which applies an attention mechanism to the base pair motif energy (described in Section 4.1) and RNA sequence feature to perfectly learn representative knowledge from RNA sequence and thermodynamic energy.

2.2 Establishing the base pair motif library.

We build the base pair motif library by modeling the tertiary structures of all three-neighbor base pair motifs and storing corresponding energy items in the motif library. Specifically, each tertiary structure of base pair motif is computed by our previous de novo RNA structure modeling method BRIQ [56], which employs Monte Carlo (MC) [57] algorithm to sample candidate RNA tertiary structures and evaluates the BRIQ energy score of each sampled tertiary structure. BRIQ energy score is a combined energy score of physical energy using density functional theory (DFT) and statistical energy calibrated by quantum mechanism, supplying a trade-off measurement of thermodynamic energy between computational speed and accuracy. Furthermore, each energy score of base pair motif is normalized according to its sequence length and motif category. After obtaining all energy items, given an RNA

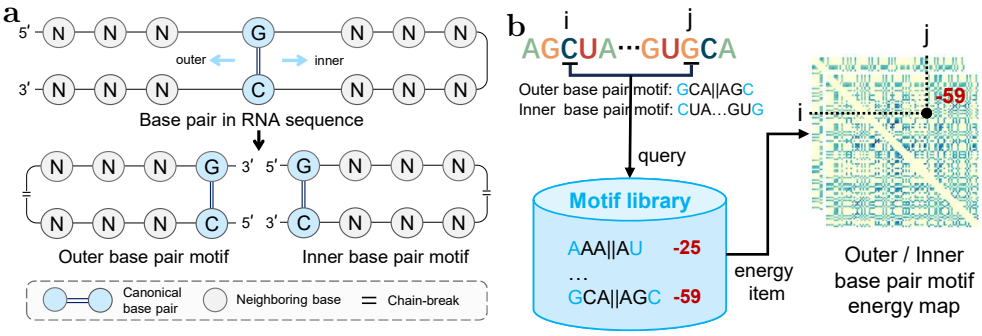

Fig. 2: **Diagram of generating outer/inner base pair motif from a canonical base pair and constructing energy maps of base pair motifs for an RNA sequence. a** For any canonical base pair (i.e., A-U, U-A, G-C, C-G, G-U, and U-G) in an RNA sequence, the upstream and downstream three neighboring bases (denoted as N, an arbitrary base of A, U, G, or C) of the base pair form two base pair motifs, an inner base pair motif with neighboring bases extending to the middle of the RNA sequence, and an outer base pair motif with neighboring bases extending to both ends of the RNA sequence. Note that the inner base pair motif can be categorized into inner hairpin base pair motif and inner chain-break base pair motif (demonstrated in this figure) in accordance with the distance of the paired bases. **b** For any canonical base pair (i, j) from an RNA sequence of L nucleotides, we first find the corresponding outer/inner base pair motifs of this base pair, and then query the energy items in the base pair motif library, which forms the (i, j) element of the outer/inner energy maps in a shape of $L \times L$.

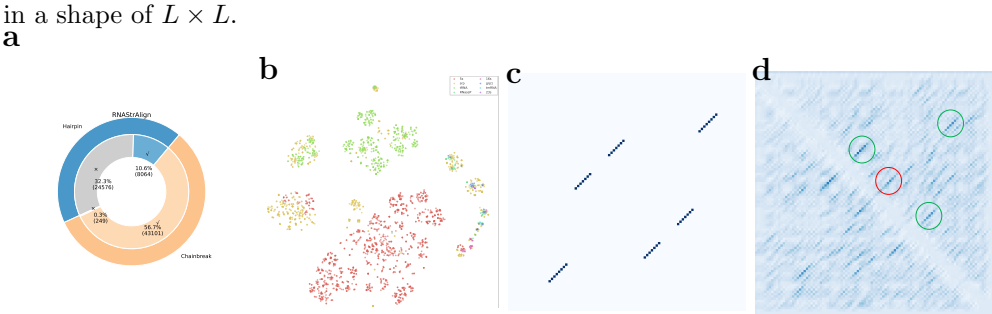

Fig. 3: **Analysis of base pair motif library**. **a** Pie visualization of the data coverage of hairpin and chainbreak base pair motifs in current datasets. The outer ring of each pie chart represents the hairpin and chainbreak distribution in the whole base pair motif library (75990 motifs, blue for hairpin motifs, orange for chainbreak motifs). The inner ring represents the hairpin and chainbreak distribution in each dataset (blue for hairpin motifs, orange for chainbreak motifs, and grey for missing motifs). **b** t-SNE visualization of the latent feature map of base pair motif energy map at the third convolutional layer from various RNA families in ArchiveII dataset [53]. **c** Ground truth heatmap visualization of the secondary structure of an example RNA sequence. **d** Heatmap visualization of the extracted latent feature map of the same RNA sequence from subfig **c**. The corrected responses of base pair interactions are annotated in green circles.

sequence of length L, for any base i and base j, we built two energy maps in a shape of $L \times L$. One for outer base pair motif and the other for inner base pair motif, denoted as M^{μ} and M^{ν}, respectively, which are used as input thermodynamic information by the BPfold neural network.

After constructing the base pair motif library, we make detailed analysis of this library, which is displayed in Fig. 3. Firstly, we demonstrate the coverage of base pair motif in current datasets. As Fig. 3a shows, RNAStrAlign dataset contains adequete chainbreak base pair motifs with only 249 chainbreak motifs missing. However, as for the hairpin motifs, RNAStrAlign misses 24,576 motifs, weighting 32.3% of total 75990 motifs, covering a small data distribution, which hinders the pattern recognition of hairpin base pair motifs for deep learning models (See the data coverage of the other four datasets and the intersection of base pair motifs from these datasets in Supplementary Fig. 1 and Supplementary Fig. 2, respectively). Furthermore, we store the intermediate results of base pair motif energy map when BPfold predicts secondary structures from the third convolutional layer, and then apply t-SNE [58] decompression to these feature maps to visualize the latent embeddings of base pair energy maps from various RAN families. As Fig. 3b demonstrates, BPfold learns discriminative embeddings of family-wise RNAs effectively, projecting the features of energy maps to a wide-range scattered latent vectors. In addition, we visualize the heatmap of the stored intermediate feature map of the base pair energy map from the third convolutional layer in Fig. 3d. Compared with the ground truth heatmap displayed in Fig. 3c, the feature map correctly captures the interactions of base pairs, annotated in green circles, which indicates that base pair motif energy assists the neural network with the interpretation of base pair interactions. Although there are false positive interactions annotated in red circles, these weak responses will be eliminated by subsequent transformer layers and post-processing procedures. In view of above analysis of the base pair motif library, we expect that our proposed base pair motif library that takes into account tertiary structures and energy items can positively contribute to any computational method.

2.3 Assessing the effectiveness of base pair motif energy.

To investigate the effectiveness of the proposed base pair motif energy, which is the key contribution of BPfold, we conduct an ablation study to demonstrate the performances of BPfold with and without base pair motif energy. In this experiment, we train BPfold with and without base pair motif energy on the training datasets of RNAStrAlign [59] and bpRNA-1m [54], respectively, and evaluate them on family-wise datasets Rfam14.5-14.10 [19, 20] and PDB [55] and sequence-wise datasets ArchiveII [53] and bpRNA-TS0 [54] (See dataset details in Section 4.4 and Supplementary Table 1).

The total results of each dataset and specific RNA families (Dicty_Class_I_RNA, bZIP, Hairpin-meta1 in Rfam14.5-14.10 and RNaseP, tRNA, tmRNA in ArchiveII) in Fig. 4 and Supplementary Table 2, 3 show that BPfold with base pair motif energy achieves much better performances than BPfold without base pair motif energy, especially on unseen RNA families, such as Dicty, bZIP, and Hairpin-meta1, indicating that the thermodynamic energy greatly improves the generalizability and accuracy of deep

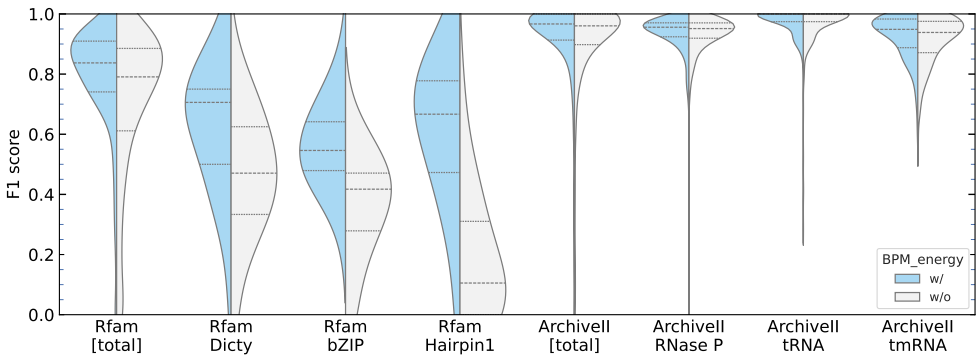

Fig. 4: **Illustrations of the performances of BPfold with and without base pair motif energy on family-wise dataset Rfam14.5-14.10** [19, 20] **and sequence-wise dataset ArchiveII** [53]. The first four pairs of violins show the F1 scores of the total Rfam14.5-14.10 dataset together with its three unseen RNA families (*i.e.*, Dicty_Class_I_RNA, bZIP, and Hairpin-meta1), while the last four pairs of violins show the F1 scores of the total ArchiveII dataset together with its three trained RNA families (*i.e.*, RNaseP, tRNA and tmRNA). Quartile scores are shown as dashed lines. With the integration of base pair motif energy, BPfold improves the prediction accuracy by a large gap, especially on unseen RNA families.

learning models. Moreover, since the training datasets have no similar RNA families of Rfam14.5-14.10, which means that Dicty, bZIP, and Hairpin-meta1 in test datasets are unseen data, BPfold without base pair motif energy obtains an extremely low accuracy, achieving F1 scores of 0.460, 0.356, and 0.186 for Dicty, bZIP, and Hairpin-meta1, respectively, compared with 0.585, 0.551, and 0.609 of BPfold with base pair motif energy. As Supplementary Tables 2 and 3 demonstrate, on sequence-wise datasets, BPfold with base pair motif energy achieves F1 scores of 0.701 and 0.918 for bpRNA-TS0 and ArchiveII datasets, respectively, surpassing BPfold without energy's F1 scores of 0.666 and 0.903. In contrast, on family-wise datasets, BPfold with energy achieves F1 scores of 0.791 and 0.781 for Rfam14.5-14.10 and PDB datasets, respectively, much better than BPfold without energy's F1 scores of 0.701 and 0.678, suggesting that base pair motif energy mitigates the lack of data distribution at base-pair level of thermodynamic energy aspect and comprehensively improves the performances on unseen data, indicating better generalizability and robustness. We also plot the results on ArchiveII datasets together with its nine RNA families in Supplementary Fig. 3.

Additionally, we make head-to-head comparisons of BPfold with and without base pair motif energy on bpRNA-Ts0, ArchiveII, Rfam14.5-14.10, and PDB datasets with the same configuration. As shown in Supplementary Fig. 4, head-to-head analysis demonstrates that the accuracy of BPfold with base pair motif energy is above that of BPfold without base pair motif energy in the majority of sample points, indicating the advantage of base pair motif energy is much more clear on improving accuracy and generalizability of learning-based data-driven neural networks.

2.4 Assessing the effectiveness and robustness of BPfold.

In this section, we compare proposed BPfold with other nine state-of-the-art methods, including deep learning methods and traditional thermodynamic models. As for data-driven models, the training datasets are mainly the same, namely RNAStrAlign [59] and bpRNA [54], except SPOT-RNA that employs PDB dataset [55] for transfer learning. All methods are evaluated on sequence-wise test datasets that contain distinguished sequences from training datasets and family-wise test datasets that consist of unseen RNA families as out-of-distribution validation. Furthermore, we conduct experiments on RNA sequences with various lengths to evaluate the prediction time of BPfold at inference stage. We also visualize some RNA secondary structures predicted by BPfold to demonstrate the performance of BPfold.

2.4.1 Evaluating BPfold on sequence-wise datasets.

To evaluate the performance of our BPfold model on sequence-wise datasets, we report the results of BPfold on ArchiveII [53] test set and bpNRA-TS0 [54] test set, compared with traditional methods, including Contextfold [22], Contrafold [21], Eternafold [28], Linearfold [25], Vienna RNAfold [26, 34], RNAsoft [24], RNAStructure (Fold) [27, 35], together with deep learning methods SPOT-RNA [40] and MXfold2 [43]. Same as previous deep learning methods [39, 41, 43], BPfold is trained on RNA sequences with lengths no more than 600 nucleotides from RNAStrAlign [59] and bpRNA [54] datasets.

Table 1: Sequence-wise evaluation of RNA secondary structure prediction approaches on bpRNA-TS0 (1,305 RNAs) and ArchiveII (3,966 RNAs) datasets. The macro-average F1 score, precision and recall of canonical pairs are reported. The best results are in **bold** while the second best results are <u>underlined</u>.

Method	bpRNA-TS0			ArchiveII		
	F1	Precision	Recall	F1	Precision	Recall
BPfold	**0.701**	**0.670**	**0.757**	**0.918**[a]	**0.922**	**0.921**
SPOT-RNA	<u>0.625</u>	<u>0.598</u>	<u>0.694</u>	0.730	0.763	0.723
MXfold2	0.570	0.515	0.685	0.746	0.734	0.767
ContextFold	0.516	0.477	0.599	<u>0.818</u>	<u>0.824</u>	<u>0.820</u>
ContraFold	0.541	0.482	0.656	0.617	0.593	0.651
EternaFold	0.539	0.475	0.663	0.599	0.573	0.636
LinearFold	0.530	0.539	0.582	0.606	0.629	0.605
ViennaRNA	0.508	0.446	0.631	0.577	0.551	0.613
RNAsoft	0.507	0.448	0.626	0.568	0.550	0.594
RNAStructure	0.507	0.446	0.628	0.573	0.548	0.607

[a]The macro-average F1 score is possibly smaller than the macro-average precision and macro-average recall meanwhile.

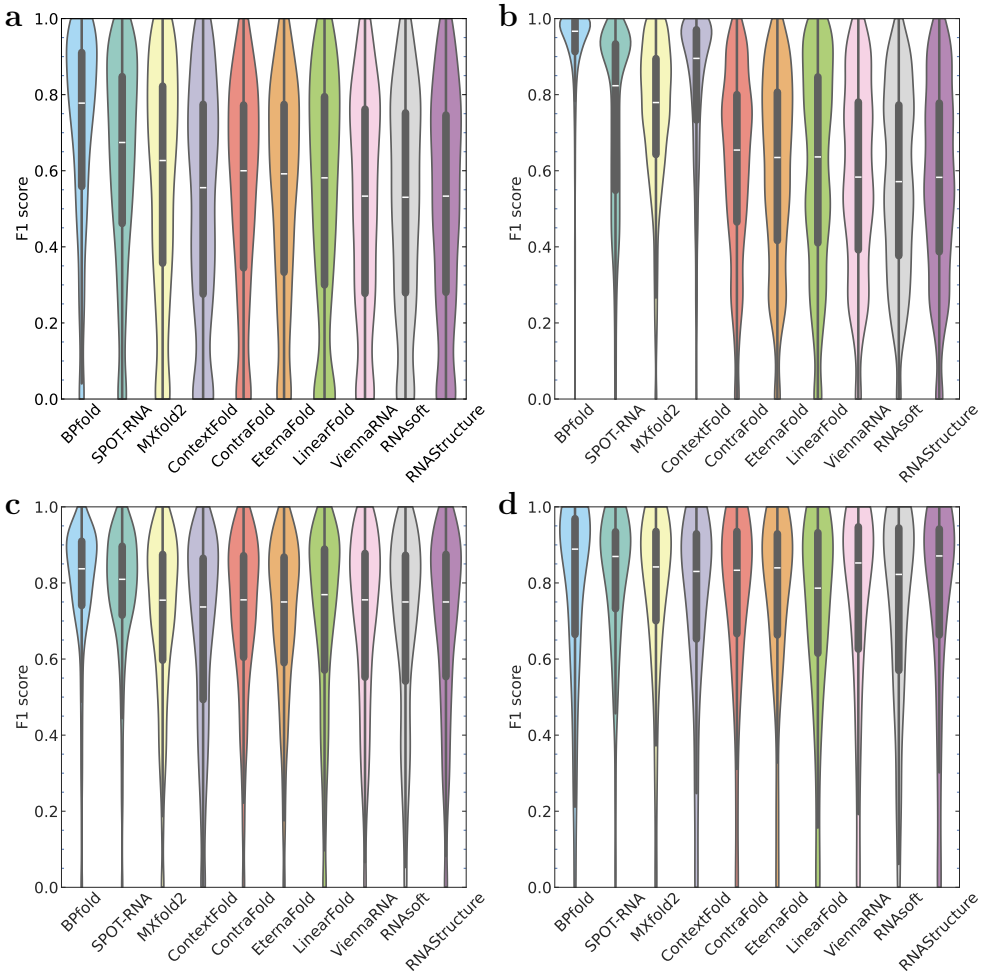

Fig. 5: **Performance comparison of BPfold and nine existing RNA secondary structure prediction methods on sequence-wise datasets and family-wise datasets**. **a** The macro-average F1 scores of predicted structures on sequence-wise dataset bpRNA-TS0. **b** The macro-average F1 scores of predicted structures on sequence-wise dataset ArchiveII. **c** The macro-average F1 scores of predicted structures on family-wise dataset Rfam14.5-14.10. **d** The macro-average F1 scores of predicted structures on high-quality dataset PDB.

As Fig. 5a and Table 1 demonstrate, on bpNRA-TS0 dataset, traditional methods achieve an average F1 score in the range of [0.507, 0.541], dropping behind deep learning methods such as 0.570 of Mxfold2 and 0.625 of SPOT-RNA. With the base pair motif energy, BPfold further improves the performance and obtains an average F1 score of 0.701, leading ahead of other methods by a remarkable gap. Compared with previous state-of-the-art method SPOT-RNA, BPfold achieves an 12.2% increase in

F1 score. As for ArchiveII dataset, Fig. 5b and Table 1 demonstrate similar rankings except Contextfold. BPfold also reaches the highest prediction accuracy, achieving an average F1 score of 0.918, outperforming the second-best method Contextfold with an F1 score of 0.818, leading ahead by a huge gap and a 12.2% increase in F1 score.

2.4.2 Evaluating BPfold on family-wise datasets.

With the same training and data processing procedures, we further evaluate our BPfold model on family-wise datasets of unseen data and out-of-distribution situation to verify its model generalizability. We collect Rfam14.5-14.10 dataset from Rfam [19, 20] database by retaining the newly added unseen families of version 14.10 from version 14.5, which contains 2048 RNA sequences after removing similar sequences by CD-HIT-EST [60] with a threshold of 80%.

Table 2: Family-wise evaluation of RNA secondary structure prediction approaches on Rfam14.5-14.10 (2,048 RNAs) and PDB (116 RNAs) datasets. The macro-average F1 score, precision, and recall of canonical pairs are reported. The best results are in **bold** while the second best results are underlined.

Method	Rfam14.5-14.10			PDB		
	F1	Precision	Recall	F1	Precision	Recall
BPfold	**0.791**	**0.777**	0.824	0.781	0.821	0.759
SPOT-RNA	0.777	0.745	**0.832**	**0.808**	**0.867**	**0.772**
MXfold2	0.707	0.655	0.789	0.778	0.837	0.744
ContextFold	0.663	0.631	0.712	0.737	0.795	0.702
ContraFold	0.710	0.647	0.805	0.756	0.789	0.737
EternaFold	0.704	0.638	0.803	0.758	0.785	0.741
LinearFold	0.700	0.687	0.746	0.718	0.813	0.672
ViennaRNA	0.695	0.633	0.789	0.747	0.776	0.728
RNAsoft	0.683	0.626	0.769	0.736	0.770	0.714
RNAStructure	0.691	0.629	0.784	0.752	0.775	0.736

As Fig. 5c and Table 2 demonstrate, deep learning methods obtain satisfying performances even though the family-wise test dataset is out of distribution, such as Mxfold2 obtains F1 score of 0.707 on Rfam14.5-14.10 and SPOT-RNA achieves an F1 score of 0.777. As mentioned above, Mxfold2 take advantage of integrating free energy minimization into deep learning model and SPOT-RNA utilizes evolutionary information. However, these kinds of auxiliary information are not complete and bias-free, which hinders the overall performance of unseen data. In contrast, BPfold outperforms any other deep learning methods and traditional methods with the help of thermodynamic energy, exploring the complete space of 3-neighbor base pair motif and mitigating the out-of-distribution data at base-pair level, achieving the best F1 score of 0.791.

Furthermore, we evaluate BPfold on PDB, a widely used benchmark dataset that contains high-resolution RNA X-ray tertiary structures. We divide this set into three test sets as SPOT-RNA does, namely TS1, TS2, and TS3, which contains 60, 38, and 18 sequences, respectively. As Fig. 5d, Table 2 and Supplementary Table 4 demonstrate the F1 score, precision, and recall metrics of canonical pairs (i.e., A-U, U-A, G-C, C-G, G-U and U-G), BPfold achieves an F1 score of 0.781, falling behind the best score 0.808 achieved by SPOT-RNA for the reason that SPOT-RNA uses PDB dataset in training process, comparable with the F1 score of 0.778 achieved by MXfold2. However, BPfold outperforms any other traditional methods, demonstrating high prediction accuracy and strong generalizability in detecting dense canonical pairs of RNA secondary structures on this high-quality experimentally validated dataset.

2.4.3 Visualizing the performance of BPfold

In theory, our proposed BPfold is efficient at inference stage for predicting RNA secondary structures, with a time complexity of $O(L^2)$ for an L nucleotides RNA sequence, comparable with other deep learning methods, and much faster than traditional methods with a time complexity of $O(L^4) \sim O(L^6)$. Experimentally, we evaluate BPfold on a selected dataset with various RNAs in length. The results demonstrated in Supplementary Fig. 5 show that BPfold predicts RNA secondary structures within 10 seconds for RNAs no longer than 1,000 nucleotides, and within 40 seconds for RNAs no longer than 2,000 nucleotides.

Apart from the above quantitative experimental analysis, we also provide qualitative visualization of the RNA secondary structures predicted by BPfold to verify the detailed interactions of each nucleotide. To achieve this, we utilize the RNA visualization tool VARNA [63] to draw the figure, which takes various formats of RNA secondary structures, such as dot-bracket notation, bpseq, and ct format. For comparison, we also show the structures of native annotations (ground truth structures) and other two excellent methods, the deep-learning method SPOT-RNA and the traditional method ContraFold. As Fig. 6 visualizes, each column displays the structures of one method and each row displays one example. In these samples, structures predicted by BPfold are similar to native structures, more accurate and robust than other methods. Specifically, Fig. 6a shows the superiority of BPfold in precision on an example of brucella abortus S19 signal recognition particle RNA, with 97% precision and 94% sensitivity, respectively. Meanwhile, Fig. 6b displays an example of delta-J-delta-K domain of EMCV IRES [61] from PDB [55] dataset (PDB-ID=2NC1) in radiate style, where lines with blue solid circle denote non-canonical pairs. As it shows, BPfold has the ability of predicting non-canonical pairs, outperforming other methods, with 100% precision and 93% sensitivity, respectively. Fig. 6c displays the structures of the lariat capping ribozyme [62] from PDB [55] dataset (PDB-ID=4P95) in circular style. This RNA sequence is a long sequence with 192 nucleotides and dense interactions. BPfold successfully models the long-range connections and predicts the most accurate structure than other methods, with 93% precision and 96% sensitivity, respectively. Furthermore, we display the effect of applying removing lone pairs in post-processing procedures. As Supplementary Fig. 6a demonstrates, compared with Supplementary Fig. 6b, there are many long-distance lone pairs marked in red, which are irrational

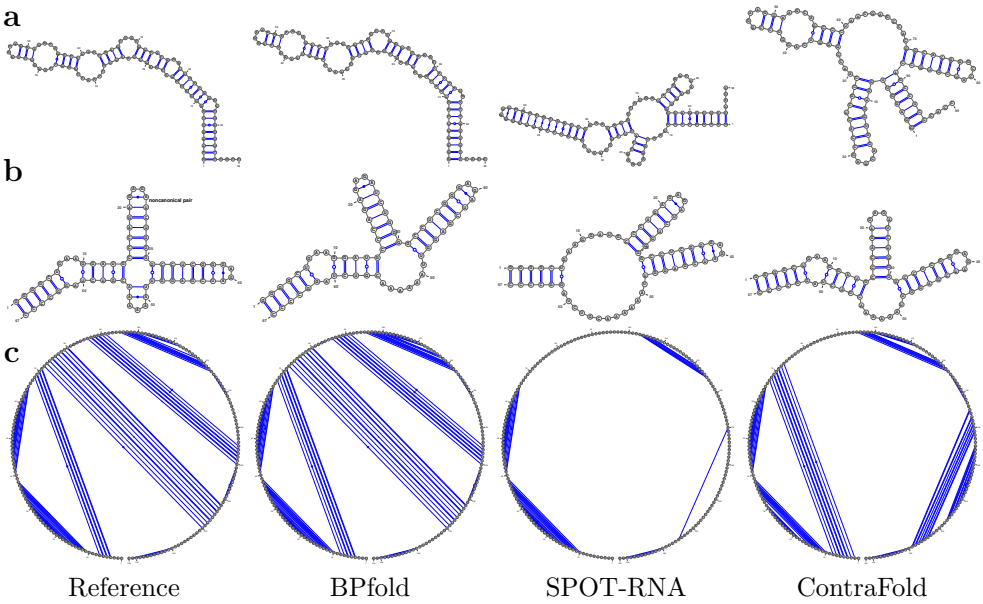

Fig. 6: **Visualization of RNA secondary structure examples predicted by BPfold along with deep learning method SPOT-RNA [40] and traditional method ContraFold [21]. a** An example of brucella abortus S19 signal recognition particle from ArchiveII [53] dataset (srp_Bruc.suis._AE014291), displayed in radiate style. From left to right, these structures are from native, BPfold, SPOT-RNA, and ContraFold, respectively. BPfold predicts the most accurate interactions compared with native reference structure, with 97% precision and 94% sensitivity, respectively. **b** The solution structure of the delta-J-delta-K domain of EMCV IRES [61] from PDB [55] dataset (PDB-ID=2NC1), displayed in radiate style, where lines with blue solid circle denote non-canonical pairs. From left to right, these structures are from native, BPfold, SPOT-RNA, and ContraFold, respectively. BPfold can correctly predict canonical pairs and non-canonical pairs, with 100% precision and 93% sensitivity, respectively. **c** The crystal structures of the lariat capping ribozyme [62] from PDB [55] dataset (PDB-ID=4P95), displayed in circular style. In such a dense prediction situation, BPfold predicts the most accurate base interactions than other methods, with 93% precision and 96% sensitivity, respectively.

and unstable, distorting the local structure of loop region. After post-processing, these lone pairs are removed.

2.5 Constructing confidence index for reliable prediction.

To measure the reliability of the predicted secondary structures, we further construct a confidence index, which supplies a quality assessment of predicted secondary structures in case the native structures are not available. In BPfold, the neural network directly generates the contact map $\tilde{Y} \in R^{L \times L}$, then we apply structural constraints on $\tilde{Y}$ in

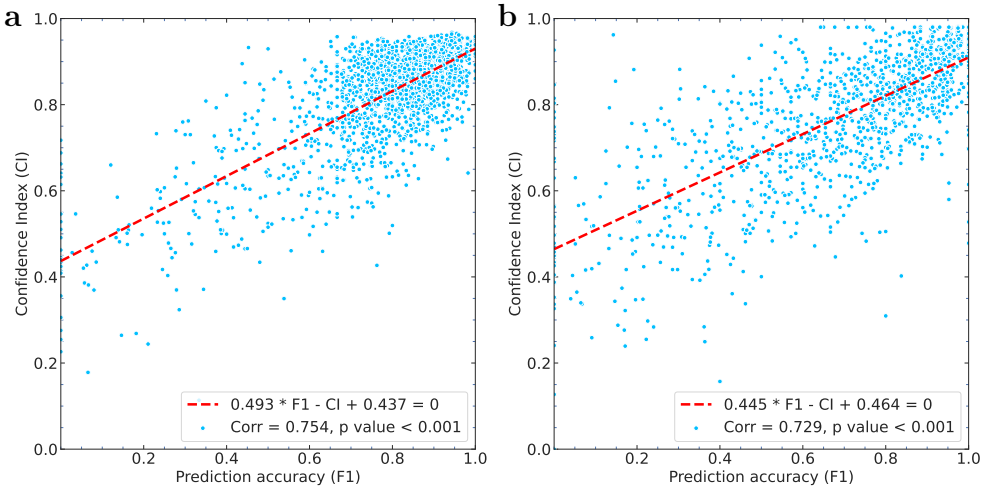

Fig. 7: **The Pearson's correlation between F1 score and the estimated confidence index on Rfam14.5-14.10 and bpRNA-TS0 datasets.** We compute the cosine similarity between the contact map generated from the neural network and the contact map after post-processing. The Pearson correlation coefficient between the prediction accuracy (F1 score) and confidence index reaches 0.754 and 0.729, respectively. The approximate relation between F1 and CI, together with the correlation coefficients are displayed in the bottom right corner of each figure. **a** Rfam14.5-14.10 [19, 20]. **b** bpRNA-TS0 [54].

post-processing procedures, obtaining the final contact map $\tilde{Y}_{\text{post}} \in R^{L\times L}$. Therefore, the difference between $\tilde{Y}$ and $\tilde{Y}_{\text{post}}$ reflects the reliability and quality of predicted structures to some extent, which means that the less difference between $\tilde{Y}$ and $\tilde{Y}_{\text{post}}$ indicates the more reliability and accuracy of the output contact map that meets the structural constraints. As a result, to form the confidence index, we compute the cosine similarity between $\tilde{Y}$ and $\tilde{Y}_{\text{post}}$ and scale it to the range of $[0, 1]$, which can be formulated as:

$$\begin{aligned}\text{confidence index} &= k\cos(\tilde{Y}, \tilde{Y}_{\text{post}}) + b \\ &= k\frac{\tilde{Y}\cdot\tilde{Y}_{\text{post}}}{\|\tilde{Y}\|\|\tilde{Y}_{\text{post}}\|} + b \\ &= k\frac{\sum_{i=1}^{L}\sum_{j=1}^{L}\tilde{y}_{ij}\tilde{y}_{\text{post}_{ij}}}{\sqrt{\sum_{i,j=1}^{L}\tilde{y}_{ij}^2}\sqrt{\sum_{i,j=1}^{L}\tilde{y}_{\text{post}_{ij}}^2}} + b\end{aligned}$$

where $k = 2.218$ and $b = -0.559$ are empirically selected coefficients for adjusting the range of confidence index to $[0, 1]$.

To demonstrate the validity of the proposed confidence index, we compute the Pearson's correlation coefficient between confidence indexes and F1 scores on the four test datasets (*i.e.*, bpRNA-TS0, ArchiveII, Rfam14.5-14.10, PDB) together with their

mixed total datasets. As Supplementary Table 5 shows, we achieve high correlation coefficients of 0.729, 0.771, 0.754, 0.632, and 0.745 with low p-values, respectively, which indicates the correlation between confidence index and F1 score and further suggests that our proposed confidence index provides us a referable and reliable insight of the predicted RNA secondary structures. Fig. 7 and Supplementary Fig. 7 display direct views of the strong correlation between the designed confidence index and the F1 score metric on different datasets.

3 Discussions

Our method, BPfold, aiming at improving the generalizability and accuracy of deep learning approaches, creatively proposes the complete space of 3-neighbor base pair motif and its thermodynamic energy to tackle the problem of data insufficiency and out-of-distribution situation at base-pair level, and further bridges the knowledge prior with input RNA sequence by the elaborately designed base pair attention neural network block. The designed transformer network architecture and energy map representation facilitate the identification of long-distance interactions and the extension to unseen long RNA sequences. Experimental results on sequence-wise datasets and family-wise datasets confirm the superiority of BPfold over other methods in accuracy, generalizability, robustness, and inference speed. Additionally, we construct a confidence index for BPfold to provide a reference for the reliability of predicted RNA secondary structures. BPfold is publicly available and serves as an effective tool for RNA structure modeling.

More importantly, the proposed base pair motif and the idea of combining thermodynamic prior with input RNA sequences can be integrated into any other data-driven models. For instance, UFold [39] processes an image-like representation of RNA sequences and an alternative matrix representation [47] in consideration of possible hydrogen bonds of canonical base pairs, in which this matrix representation can be readily replaced by the energy map of base pair motif presented in this study. By doing so, the neural network achieves a complete perception of the interactions of 3-neighbor base pair, much more accurate than the roughly counted number of hydrogen bonds. Besides, the computed energy scores of base pair motifs by the BRIQ [56] force field supplies a reliable estimation of base pair interactions, which works effectively in other thermodynamic energy-based methods. In fact, EternaFold [28] obtains thermodynamic measures via high-throughput experiments to update the parameters in Contrafold [21]. Similarly, it is possible to apply the framework of Contrafold with the statistical energy scores of base pair motifs. Note that this set of base pair motif energy scores can be further updated with the development of BRIQ or other tertiary structure de novo modeling methods.

Despite the innovations of BPfold in RNA sequence analysis, BPfold confronts several challenges. Firstly, on the one hand, the base pair motif can be further extended and more reliable by computing more neighboring bases, such as four or more upstream and downstream bases. In this study, we only model the tertiary structures of 3-neighbor base pair motifs using BRIQ [56] due to the large computation cost. On the other hand, we could introduce more knowledge prior in extra forms rather than

base pair motifs, such as RNA family information, to make a better understanding of the input RNA sequences and further enhance the generalization and accuracy of the models. Secondly, existing RNA secondary structure datasets consist of RNA sequences whose lengths are mainly less than 600 nucleotides or shorter, and existing deep learning methods, including BPfold, are also trained on these sequences. Therefore, to relieve the performance degradation brought by long sequences, BPfold applies a dynamic positional embedding for scalable feature embedding in base pair attention block. However, the long sequence problem still remains unsolved; we believe that the key is to enrich the distribution of long sequences in the current training dataset. Thirdly, prediction of non-canonical pairs and pseudo knots is difficult for data-driven methods since there are few annotations of these interactions for deep learning approaches to learn from. Regardless of that, BPfold has the ability of predicting non-canonical pairs and pseudo knots (Exemplified in Fig. 6), the accuracy of BPfold for these interactions would be greatly improved when corresponding data annotations for training are adequate in the future.

The generalizability of deep learning models on newly discovered unseen RNA families is an inevitable issue in current research. Technically speaking, there are various common techniques to relieve the overfitting of models. As for training, we can apply early-stopping to stop the learning of models before the model tends to overfitting on training data. Besides, as for model, we also apply a multiple-fold cross-validation to decrease the systematic error and select the best hyper-parameters of model architecture according to the amount of data. In this study, we innovatively propose the base pair motif and energy map from data aspect to fully cover the data distribution at base-pair level. BPfold receives the information not only the RNA sequence but also the energy map of all paired canonical pairs in this RNA sequence, which allows the specific prediction of each input RNA sequence and hinders the overfitting of similar RNA sequences. As a result, evidenced in experiments on family-wise dataset Rfam14.5-14.10, BPfold achieves the best performances against other methods, revealing the great generalizability and accuracy in modeling RNA secondary structures from unseen RNA families. Apart from base pair motif, high-throughput chemical probing data [64, 65] such as SHAPE [66] can be applied to provide the information of chemical activity of nucleotide bases. Note that because the auxiliary information involved in existing deep learning methods such as BPfold, MXfold2, UFold, and SPOT-RNA is implicitly learned by neural networks, further improvements in modeling RNA secondary structures can be achieved when learning approaches can explicitly integrate physical laws into neural networks.

In summary, we show a great potential of BPfold in improving generalizability and accuracy with base pair motif for RNA secondary structure prediction, hoping that BPfold brings inspiration in further development of deep learning methods and understanding of RNA structures and biological functions.

4 Methods

4.1 Base pair motif energy as thermodynamic prior.

The performance and generalizability of deep learning models are highly dependent on training data, which currently may be hampered by the lack of structural data and the limitation of data diversity. To tackle this, we aim at improving the coverage of data at base-pair level and bringing thermodynamic energy to RNA sequence modeling. In view of the locality of the secondary structures of RNA, we define the *base pair motif*, a canonical base pair (i.e., A-U, U-A, G-C, C-G, G-U, and U-G) together with the r upstream and downstream neighboring bases of each base.

Specifically, base pair motifs can be divided into three categories, for any two bases indexed as $i, j, (i < j)$ of an RNA sequence: (1) **Inner hairpin base pair motif**: While neighboring bases extend to inner sequence and $j - i \leq 2r$, the downstream of base i and the upstream of base j are continuous in sequence and form a hairpin loop, with the base pair motif denoted as $[i, i{+}1, \ldots, j]$. There are $\sum_{i=3}^{6} 4^i = 5440$ sequences for each canonical base pair of this category. (2) **Inner chain-break base pair motif**: While neighboring bases extend to inner sequence and $j{-}i{>}2r$, the downstream of base i is not continuous with the upstream of base j, forming a chain-break. Therefore, the base pair motif consists of two chains, denoted as $[i, i{+}1, \ldots, i{+}r : j{-}r, j{-}r{+}1, \ldots, j]$, where : represents chain-break of RNA sequence. There are $4^6 = 4096$ sequences for each canonical base pair of this category. (3) **Outer base pair motif**: While neighboring bases extend to outer both ends of the sequence, the base pair motif is denoted as $[j, j{+}1, \ldots, j{+}r : i{-}r, i{-}r{+}1, \ldots, i]$. Besides that, we also deal with special corner cases while base i or base j has no sufficient r neighboring bases upstream or downstream. There are $\sum_{i,j=1,(i,j)\neq(3,3)}^{3} 4^i \times 4^j = 3229$ sequences for each canonical base pair of this category. In total, there are $75990 = 6 \times (5440 + 4096 + 3129)$ base pair motifs for all six kinds of canonical base pairs (i.e., AU, UA, G-C, C-G, GU, and U-G). In the formation of these, the paired bases are always located at the beginning and the end of each category of base pair motif.

After enumerating all possible sequences of r-neighbor base pair motifs, we use our previous de novo RNA tertiary structures modeling method BRIQ [56] to model the tertiary structure of each base pair motif and extract energy score $E_{\text{bpm}}(L_{\text{bpm}}, I_{\text{break}})$ estimated by BRIQ force field of whole motif sequence, where L_{bpm} is the length of base pair motif and I_{break} represents whether there is a chain-break in this motif. Then we compute the energy score E_{bpm} of each motif by eliminating the influence of single strand and normalizing the value with the maximum energy of the motif with the same length in the same category, which can be formulated as:

$$E_{\text{bpm}} = \frac{E_{\text{bpm}}(L_{\text{bpm}}, I_{\text{break}}) - E_{\text{single-strand}}(L_{\text{bpm}})}{\max\{E_{\text{bpm}}(L_{\text{bpm}}, I_{\text{break}})\}}.$$

Finally, we establish a base pair motif library with these energy items and tertiary structures. As Fig. 2 demonstrates, with this energy table, for any two bases of RNA sequence with index being $i, j, (i{<}j)$, we query the outer/inner base pair motif energy items of corresponding outer/inner base pair motif for any canonical base pair (i.e.,

A-U, U-A, G-C, C-G, G-U and U-G) and we set the energy of any other non-canonical pair to zero. Therefore, for an RNA sequence with L nucleotides, we can obtain the outer/inner energy maps M^{μ} and M^{ν} in a shape of $L \times L$.

4.2 Deep neural network with base pair attention.

As shown in Figure 1, BPfold is a deep neural network consisting of consecutive N modified transformer blocks. In each transformer block, there is an elaborately designed base pair attention block (Fig. 1b), composed of hybrid convolutional block (Fig. 1c) which applies the squeeze-and-excitation (SE) block [67](Fig. 1d) to adaptively re-calibrate the channel-wise feature response and explicitly build the relationship of energy maps, along with an enhanced self-attention mechanism that aggregates attention map from RNA sequence and base pair motif energy, learning thermodynamic knowledge in the complete space of r-neighbor base pair motif to improve the generalizability of model in situation of unseen RNA sequence and families.

The original transformer block [51] is composed of a multi-head self-attention module (MSA), followed by a feed-forward network (FFN) which consists of a two-layer multi-layer perceptron (MLP) with a GELU activation. A LayerNorm (LN) layer is adopted before each MSA and each FFN, and a residual shortcut is adopted after each module.

To integrate the thermodynamic priors in the form of energy maps into this attention mechanism, we design the base pair attention block. As illustrated in Figure 1b, when processing self-attention, a base pair attention block applies several 3×3 convolutional layers (denoted as $CONV$) to the energy map to establish the relationship among base pairs. Furthermore, This attention block adds the thermodynamic feature map to the attention map of sequence features, imposing the thermodynamic relationship of base pairs on RNA sequences. More specifically, the input of the neural network is an RNA sequence of length L, containing four bases, *i.e.*, A, C, G, U (other unknown bases will be converted to the above four bases). The input RNA sequences are firstly padded with "START", "END" and "EMPTY" tokens to a uniform length $L_{\max}$ to deal with variable lengths, which are then encoded into a D-dimensional embedding using trainable parameters, forming the input feature X in a shape of $L_{\max} \times D$. Meanwhile, the energy maps M^{μ} and M^{ν} of outer base pair motif and inner base pair motif are prepared according to the input RNA sequence and the stored energy table. Similar to the input feature of RNA sequences, both M^{μ} and M^{ν} energy maps are padded with zero to the shape of $L_{\max} \times L_{\max}$. The base pair attention block processes can be formulated as follows: (1) Obtaining base pair attention maps:

$$\begin{aligned} M_1 &= \underset{\theta_i}{CONV}([M^{\mu} : M^{\nu}]); \\ M_{i+1} &= \underset{\theta_i}{CONV}(M_i), i = 1, 2; \\ M_i &= M_3, i = 4, 5, 6 \ldots, N; \end{aligned} \tag{1}$$

where M_i is the i-th base pair attention block, $[M^{\mu} : M^{\nu}]$ denotes the concatenation of base pair energy maps M^{μ} and M^{ν}, and θ represents learnable parameters. (2)

Integrating base pair attention maps with sequence feature maps into transformer block:

$$
\begin{aligned}
\tilde{X}_i &= \underset{\theta_i}{LN}(X_i); \\
Q_i = \underset{\theta_i}{Q}(\tilde{X}_i); \quad K_i &= \underset{\theta_i}{K}(\tilde{X}_i); \quad V_i = \underset{\theta_i}{V}(\tilde{X}_i); \\
X_{MSA_i} &= softmax(\frac{Q_i K_i^T + M_i}{\sqrt{D_k}} V_i); \\
\tilde{X}_{MSA_i} &= \underset{\theta_i}{LN}(X_{MSA_i} + \tilde{X}_i); \\
X_{i+1} &= \tilde{X}_{MSA_i} + \underset{\theta_i}{FFN}(\tilde{X}_{MSA_i})).
\end{aligned} \tag{2}
$$

where $i = 1, 2, \ldots, N$ with N being the number of transformer blocks. After N base pair attention blocks, we obtain the output orthogonal matrix $\tilde{Y}$ in shape of $L_{\max} \times L_{\max}$ by applying matrix multiplication between X_N and its transpose X_N^T, namely $\tilde{Y} = X_N \times X_N^T$, which represents the possibility score of each nucleotide being paired with other nucleotides.

4.3 Training strategy and post-processing.

To relieve the impact of efficiency brought by padding, we apply a length-matching strategy for sampling a mini-batch at the training stage. Specifically, we set a series of buckets $\{B_0, B_1, B_2, \ldots\}$ and assign each input RNA sequence to a bucket $B_i, i = \lfloor \frac{L}{L_p} \rfloor$ where L is the length of RNA sequence and L_p is the predefined interval of buckets. Mini-batches are sampled from the same bucket, which leads to a controllable padding size L_p. This length-matching strategy is especially effective in dealing with the varying lengths of input RNA sequences from tens for short sequences to thousands for large sequences.

BPfold is implemented in PyTorch framework [68] and trained by minimizing the binary cross-entropy between the predicted contact matrix $\tilde{Y}$ and the true contact matrix Y using the ADAM optimization algorithm. The number of parameters is listed in Supplementary Table 6. To leverage the imbalanced distribution of paired bases and unpaired bases, we adopt a positive weight $\omega = 300$ to derive the loss function as below:

$$
loss(\tilde{Y}, Y; \theta) = -\sum_{ij} [\omega Y_{ij} \log(\tilde{Y}_{ij}) + (1 - Y_{ij}) \log(1 - \tilde{Y}_{ij})]. \tag{3}
$$

where θ denotes all learnable parameters of the neural network, $i \in \{1, 2, \ldots, L\}$ and $j \in \{1, 2, \ldots, L\}$ denote the row and column, respectively, index of matrices.

When predicting an RNA sequence, following previous work [39, 42], we apply post-processing procedures to the output contact map $\tilde{Y}$ generated by BPfold neural network to impose physical constraints on RNA secondary structures and rule out invalid base pairs. Specifically, the following rules of constraints are considered for post-processing: (1) Only Waston-Crick pairs (*i.e.*, A-U, U-A, G-C, C-G) and Wobble pair (G-U, U-G) are allowed for canonical pairs while others are allowed for non-canonical base pairs; (2) A loop region has at least two bases for ruling out sharp loops; (3)

Overlapping pairs are discarded. We encode these constraints as matrix transformations and apply them to the output contact matrix. Since the output contact matrix is already a symmetric matrix ($\tilde{Y} = X_N \times X_N^T$), we do not explicitly declare symmetric post-processing. (4) Lone pairs are removed. Lone pair has no consecutive neighboring helix and is not stable enough to form a base pair in most situations. We verify all paired bases and remove lone pairs to rule out long-distance unstable interactions.

4.4 Datasets and evaluation.

We utilize several widely used open-source benchmark datasets for evaluating the performances of our proposed BPfold and comparing it with state-of-the-art RNA secondary structure prediction methods. Specifically, these benchmark datasets are as follows:

- **RNAStralign** [59]: This dataset contains 37,149 RNA sequences from eight RNA families, with sequence lengths ranging from approximately 30 to 3,000 nucleotides (nt). Similar to previous work [39, 42, 43], we remove redundant sequences and invalid secondary structures, and obtain a total of 29,647 unique sequences. Furthermore, we filter sequences with lengths no more than 600 nt, forming a training dataset that consists of 19,313 sequences.
- **bpRNA-1m** [54]: This dataset contains 102,318 RNA sequences from 2,588 RNA families. Following MXfold2 [43], we use CD-HIT program [60] to remove similar sequences with a cut-off of 80% and split the processed dataset into two sub-dataset for training and testing, named TR0 and TS0, which contain 12,114 and 1,305 sequences respectively, with sequence lengths ranging from 22 to 499 nt.
- **ArchiveII** [53]: This dataset is the most widely used benchmark dataset for evaluation of RNA secondary structures, consisting of 3,966 RNA sequences from ten RNA families, *i.e.*, 5s rRNA 16s rRNA, 23s rRNA, tRNA, tmRNA, telomerase RNA RNase P, SRP, group I Intron, and group II Intron. Among them, 3,911 sequences have a length below 600 nt while the other 55 RNA sequences are all from group II Intron which have a max length of 1800 nt.
- **Rfam14.5-14.10**: This dataset initially collects 10,035 RNA sequences that are newly added to the latest version of Rfam [19, 20], namely from Rfam14.5 to Rfam14.10. After using CD-HIT-EST [60] to remove redundant sequences with sequence similarity of more than 80%, this dataset contains 2,048 unique sequences. We employ this dataset for family-wise evaluation.
- **PDB** [55]: This dataset is a benchmark dataset, consisting of 116 RNA sequences with high-resolution ($<3.5\AA$) RNA X-ray structures, with sequence length ranging from 32 to 355 nt. According to previous study [39–41], PDB dataset is divided into three sub datasets, i.e., TS1, TS2, and TS3, with 60, 38, and 18 sequences, respectively.

For performance evaluation of predicted RNA secondary structures, we use the positive predictive value (PPV), the sensitivity (SEN), and their harmonic mean F1 score to assess the quality of base pair prediction, which is a binary classification

problem. These metrics are defined as below:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \mathrm{F1} = \frac{2 \times \mathrm{PPV} \times \mathrm{SEN}}{\mathrm{PPV}+\mathrm{SEN}}. \tag{4}$$

where TP, FP, and FN denote true positive (the number of correctly predicted base pairs), false positive (the number of incorrectly predicted base pairs), and false negative (the number of base pairs whose reference structures are not predicted), respectively.

Supplementary information. Supplementary information is available for this paper.

Declarations

Funding

Supported by Natural Science Foundation of China under Grant 62271465, Suzhou Basic Research Program under Grant SYG202338, and Open Fund Project of Guangdong Academy of Medical Sciences, China (No. YKY-KF202206).

Conflict of interest

The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The datasets used in this study are available at TODO.

Materials availability

Not applicable.

Code availability

The source code and software of BPfold is publicly available at https://github.com/heqin-zhu/BPfold.

Author contribution

H.Z. constructed the motif library, designed the network architectures, conducted the experiments, analyzed the results, and wrote the paper. P.X. and S.K.Z. conceived

the idea, supervised the study, and designed the experiments. P.X. also participated in the design of the core components. F.T. carried out part of the experiments and participated in the design of networks. Q.Q. and K.C. helped in the analysis of data results. All authors read, contributed to the discussion, and approved the final paper.

References

[1] Butcher, S.E., Pyle, A.M.: The molecular interactions that stabilize rna tertiary structure: Rna motifs, patterns, and networks. Accounts of chemical research **44**(12), 1302–1311 (2011)

[2] Doudna, J.A., Cech, T.R.: The chemical repertoire of natural ribozymes. Nature **418**(6894), 222–228 (2002)

[3] Strulson, C.A., Molden, R.C., Keating, C.D., Bevilacqua, P.C.: Rna catalysis through compartmentalization. Nature chemistry **4**(11), 941–946 (2012)

[4] Morris, K.V., Mattick, J.S.: The rise of regulatory rna. Nature Reviews Genetics **15**(6), 423–437 (2014)

[5] Black, D.L.: Mechanisms of alternative pre-messenger rna splicing. Annual review of biochemistry **72**(1), 291–336 (2003)

[6] Lemieux, S., Major, F.: Rna canonical and non-canonical base pairing types: a recognition method and complete repertoire. Nucleic acids research **30**(19), 4250–4263 (2002)

[7] Lemieux, S., Major, F.: Rna canonical and non-canonical base pairing types: a recognition method and complete repertoire. Nucleic acids research **30**(19), 4250–4263 (2002)

[8] Walter, N.G., Burke, J.M., Millar, D.P.: Stability of hairpin ribozyme tertiary structure is governed by the interdomain junction. nature structural biology **6**(6), 544–549 (1999)

[9] Brierley, I., Pennell, S., Gilbert, R.J.: Viral rna pseudoknots: versatile motifs in gene expression and replication. Nature Reviews Microbiology **5**(8), 598–610 (2007)

[10] Jackson, R.W., Smathers, C.M., Robart, A.R.: General strategies for rna x-ray crystallography. Molecules **28**(5), 2111 (2023)

[11] Zhang, H., Keane, S.C.: Advances that facilitate the study of large rna structure and dynamics by nuclear magnetic resonance spectroscopy. Wiley Interdisciplinary Reviews: RNA **10**(5), 1541 (2019)

[12] Ognjenović, J., Grisshammer, R., Subramaniam, S.: Frontiers in cryo electron

microscopy of complex macromolecular assemblies. Annual review of biomedical engineering **21**(1), 395–415 (2019)

[13] Zhang, J., Fei, Y., Sun, L., Zhang, Q.C.: Advances and opportunities in rna structure experimental determination and computational modeling. Nature methods **19**(10), 1193–1207 (2022)

[14] Hüttenhofer, A., Vogel, J.: Experimental approaches to identify non-coding rnas. Nucleic acids research **34**(2), 635–646 (2006)

[15] Parsch, J., Braverman, J.M., Stephan, W.: Comparative sequence analysis and patterns of covariation in rna secondary structures. Genetics **154**(2), 909–921 (2000)

[16] Fox, G.E., Woese, C.R.: 5s rna secondary structure. Nature **256**(5517), 505–507 (1975)

[17] Gardner, P.P., Giegerich, R.: A comprehensive comparison of comparative rna structure prediction approaches. BMC bioinformatics **5**, 1–18 (2004)

[18] Havgaard, J.H., Gorodkin, J.: Rna structural alignments, part i: Sankoff-based approaches for structural alignments. RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods, 275–290 (2014)

[19] Sato, K., Akiyama, M., Sakakibara, Y.: Rna secondary structure prediction using deep learning with thermodynamic integration. Nature communications **12**(1), 941 (2021)

[20] Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., *et al.*: Rfam 14: expanded coverage of metagenomic, viral and microrna families. Nucleic Acids Research **49**(D1), 192–200 (2021)

[21] Do, C.B., Woods, D.A., Batzoglou, S.: Contrafold: Rna secondary structure prediction without physics-based models. Bioinformatics **22**(14), 90–98 (2006)

[22] Zakov, S., Goldberg, Y., Elhadad, M., Ziv-Ukelson, M.: Rich parameterization improves rna structure prediction. Journal of Computational Biology **18**(11), 1525–1542 (2011)

[23] Hamada, M., Kiryu, H., Sato, K., Mituyama, T., Asai, K.: Prediction of rna secondary structure using generalized centroid estimators. Bioinformatics **25**(4), 465–473 (2009)

[24] Andronescu, M., Aguirre-Hernandez, R., Condon, A., Hoos, H.H.: Rnasoft: a suite of rna secondary structure prediction and design software tools. Nucleic acids research **31**(13), 3416–3422 (2003)

[25] Huang, L., Zhang, H., Deng, D., Zhao, K., Liu, K., Hendrix, D.A., Mathews, D.H.: Linearfold: linear-time approximate rna folding by 5'-to-3'dynamic programming and beam search. Bioinformatics **35**(14), 295–304 (2019)

[26] Lorenz, R., Bernhart, S.H., Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L.: Viennarna package 2.0. Algorithms for molecular biology **6**, 1–14 (2011)

[27] Mathews, D.H., Turner, D.H.: Prediction of rna secondary structure by free energy minimization. Current opinion in structural biology **16**(3), 270–278 (2006)

[28] Wayment-Steele, H.K., Kladwang, W., Strom, A.I., Lee, J., Treuille, A., Becka, A., Eterna Participants, Das, R.: RNA secondary structure packages evaluated and improved by high-throughput experiments. Nature Methods **19**(10), 1234–1242 (2022) https://doi.org/10.1038/s41592-022-01605-0

[29] Andronescu, M., Condon, A., Hoos, H.H., Mathews, D.H., Murphy, K.P.: Efficient parameter estimation for rna secondary structure prediction. Bioinformatics **23**(13), 19–28 (2007)

[30] Andronescu, M., Condon, A., Hoos, H.H., Mathews, D.H., Murphy, K.P.: Computational approaches for rna energy parameter estimation. RNA **16**(12), 2304–2318 (2010)

[31] Delli Ponti, R., Marti, S., Armaos, A., Tartaglia, G.G.: A high-throughput approach to profile rna structure. Nucleic acids research **45**(5), 35–35 (2017)

[32] Seemann, S.E., Gorodkin, J., Backofen, R.: Unifying evolutionary and thermodynamic information for rna folding of multiple alignments. Nucleic Acids Research **36**(20), 6355–6362 (2008)

[33] Zuker, M.: Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research **31**(13), 3406–3415 (2003)

[34] Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., Schuster, P., *et al.*: Fast folding and comparison of rna secondary structures. Monatshefte fur chemie **125**, 167–167 (1994)

[35] Reuter, J.S., Mathews, D.H.: Rnastructure: software for rna secondary structure prediction and analysis. BMC bioinformatics **11**, 1–9 (2010)

[36] Seetin, M.G., Mathews, D.H.: Rna structure prediction: an overview of methods. Bacterial regulatory RNA: methods and protocols, 99–122 (2012)

[37] Xu, X., Chen, S.-J.: Physics-based rna structure prediction. Biophysics reports **1**, 2–13 (2015)

[38] Gong, T., Ju, F., Bu, D.: Accurate prediction of rna secondary structure including pseudoknots through solving minimum-cost flow with learned potentials. Communications Biology **7**(1), 297 (2024)

[39] Fu, L., Cao, Y., Wu, J., Peng, Q., Nie, Q., Xie, X.: Ufold: fast and accurate rna secondary structure prediction with deep learning. Nucleic acids research **50**(3), 14–14 (2022)

[40] Singh, J., Hanson, J., Paliwal, K., Zhou, Y.: Rna secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nature communications **10**(1), 5407 (2019)

[41] Singh, J., Paliwal, K., Zhang, T., Singh, J., Litfin, T., Zhou, Y.: Improved rna secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. Bioinformatics **37**(17), 2589–2600 (2021)

[42] Chen, X., Li, Y., Umarov, R., Gao, X., Song, L.: RNA secondary structure prediction by learning unrolled algorithms. In: International Conference on Learning Representations (2020)

[43] Sato, K., Akiyama, M., Sakakibara, Y.: Rna secondary structure prediction using deep learning with thermodynamic integration. Nature communications **12**(1), 941 (2021)

[44] Wang, N., Bian, J., Li, Y., Li, X., Mumtaz, S., Kong, L., Xiong, H.: Multi-purpose rna language modelling with motif-aware pretraining and type-guided fine-tuning. Nature Machine Intelligence (2024) https://doi.org/10.1038/s42256-024-00836-4

[45] Qiu, X.: Sequence similarity governs generalizability of de novo deep learning models for rna secondary structure prediction. PLOS Computational Biology **19**(4), 1011047 (2023)

[46] Zhang, J., Fei, Y., Sun, L., Zhang, Q.C.: Advances and opportunities in rna structure experimental determination and computational modeling. Nature methods **19**(10), 1193–1207 (2022)

[47] Zhang, H., Zhang, C., Li, Z., Li, C., Wei, X., Zhang, B., Liu, Y.: A new method of rna secondary structure prediction based on convolutional neural network and dynamic programming. Frontiers in genetics **10**, 467 (2019)

[48] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., *et al.*: Highly accurate protein structure prediction with alphafold. nature **596**(7873), 583–589 (2021)

[49] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al.: Accurate structure

prediction of biomolecular interactions with alphafold 3. Nature, 1–3 (2024)

[50] Watkins, A.M., Rangan, R., Das, R.: Farfar2: improved de novo rosetta prediction of complex global rna folds. Structure **28**(8), 963–976 (2020)

[51] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[52] LeCun, Y., Bengio, Y., *et al.*: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks **3361**(10), 1995 (1995)

[53] Sloma, M.F., Mathews, D.H.: Exact calculation of loop formation probability identifies folding motifs in rna secondary structures. RNA **22**(12), 1808–1818 (2016)

[54] Danaee, P., Rouches, M., Wiley, M., Deng, D., Huang, L., Hendrix, D.: bprna: large-scale automated annotation and analysis of rna secondary structure. Nucleic acids research **46**(11), 5381–5394 (2018)

[55] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. Nucleic acids research **28**(1), 235–242 (2000)

[56] Xiong, P., Wu, R., Zhan, J., Zhou, Y.: Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving rna model refinement. Nature communications **12**(1), 2777 (2021)

[57] Liu, F., Ou-Yang, Z.-c.: Monte carlo simulation for single rna unfolding by force. Biophysical journal **88**(1), 76–84 (2005)

[58] Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)

[59] Tan, Z., Fu, Y., Sharma, G., Mathews, D.H.: Turbofold ii: Rna structural alignment and secondary structure prediction informed by multiple homologs. Nucleic acids research **45**(20), 11570–11581 (2017)

[60] Li, W., Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**(13), 1658–1659 (2006)

[61] Imai, S., Kumar, P., Hellen, C.U., D'Souza, V.M., Wagner, G.: An accurately preorganized ires rna structure enables eif4g capture for initiation of viral translation. Nature structural & molecular biology **23**(9), 859–864 (2016)

[62] Meyer, M., Nielsen, H., Oliéric, V., Roblin, P., Johansen, S.D., Westhof, E., Masquida, B.: Speciation of a group i intron into a lariat capping ribozyme.

Proceedings of the National Academy of Sciences **111**(21), 7659–7664 (2014)

[63] Darty, K., Denise, A., Ponty, Y.: Varna: Interactive drawing and editing of the rna secondary structure. Bioinformatics **25**(15), 1974 (2009)

[64] Zhang, J., Fei, Y., Sun, L., Zhang, Q.C.: Advances and opportunities in rna structure experimental determination and computational modeling. Nature methods **19**(10), 1193–1207 (2022)

[65] Wayment-Steele, H.K., Kladwang, W., Strom, A.I., Lee, J., Treuille, A., Becka, A., Participants, E., Das, R.: Rna secondary structure packages evaluated and improved by high-throughput experiments. Nature methods **19**(10), 1234–1242 (2022)

[66] Morandi, E., Hemert, M.J., Incarnato, D.: Shape-guided rna structure homology search and motif discovery. Nature Communications **13**(1), 1722 (2022)

[67] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

[68] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., *et al.*: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035 (2019)