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

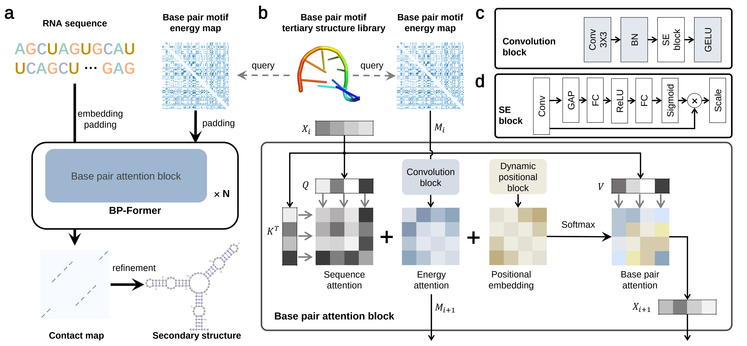
Heqin Zhu[1,2,3], Fenghe Tang[1,2,3], Quan Quan[4], Ke Chen[1,2], Peng Xiong[1,2]✉, S. Kevin Zhou[1,2,3,5,6]✉

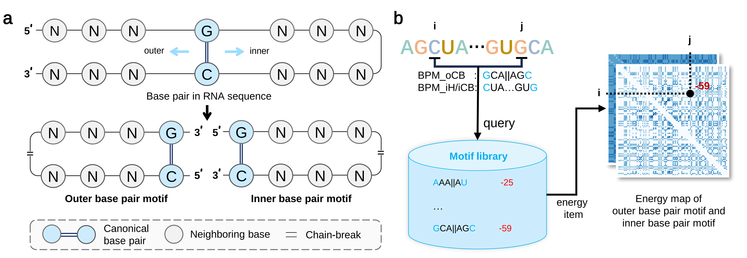
Corresponding author(s). E-mail(s): xiongxp@ustc.edu.cn; skevinzhou@ustc.edu.cn;
1. School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China (USTC), Hefei, Anhui, 230026, China.
2. Suzhou Institute for Advanced Research, USTC, Suzhou, Jiangsu, 215123, China.
3. Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE), Suzhou Institute for Advance Research, USTC, Suzhou, Jiangsu, 215123, China.
4. Key Laboratory of Intelligent Information Processing of Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China.
5. Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology, Suzhou, Jiangsu, 215123, China.
6. State Key Laboratory of Precision and Intelligent Chemistry, USTC, Hefei, Anhui, 230026, China.

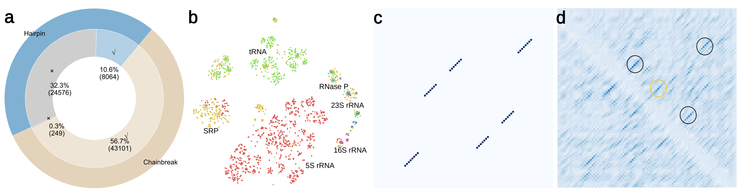
Abstract

Deep learning methods have demonstrated great performance for RNA secondary structure prediction. However, generalizability is a common unsolved issue on unseen out-of-distribution RNA families, which hinders further improvement of the accuracy and robustness of deep learning methods. Here we construct a base pair motif library that enumerates the complete space of the 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 learns relationship between RNA sequence and the energy map of base pair motif. Experiments on sequence-wise and family-wise datasets have demonstrated the great superiority of BPfold compared to other state-of-the-art approaches in accuracy and generalizability. We hope this work contributes to integrating physical priors and deep learning methods for the further discovery of RNA structures and functionalities.

Overview of BPfold and BP motifs



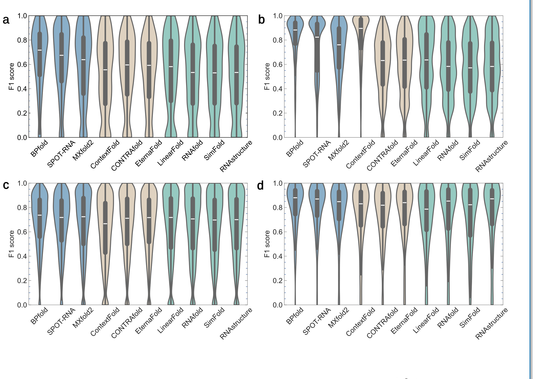


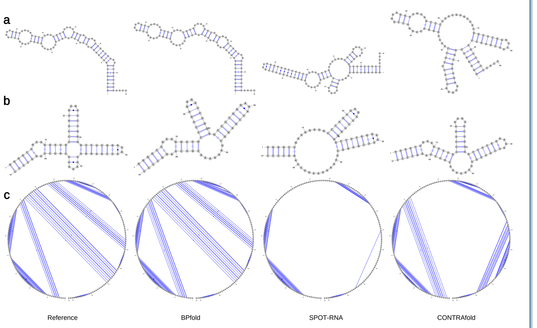


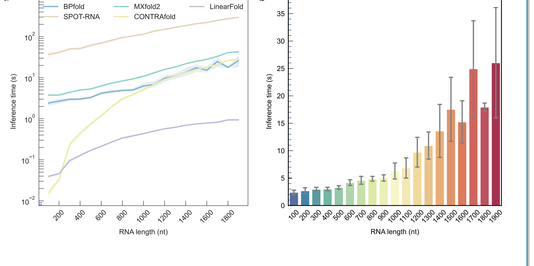
Description

- **BPfold architecture**. **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 refinement procedures, we obtain the final predicted secondary structure. **b** The detailed structure of the proposed base pair attention block which jointly fuses the sequence features X_i and energy matrix features M_i for enhanced learning of base pair interactions. When computing self-attention, Q, K, and V represent query, key, and value matrixs, respectively. **c** The detailed structure of convolution block in base pair attention block. **d** The detailed structure of squeeze & excitation (SE) block in convolution block.
- **Base pair motif library**. **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. **b** For any canonical base pair (i, j) from an RNA sequence of L nucleotides, we firstly 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 $L \times L$.
- **Analysis of BP motifs**. **a** Pie visualization of the data coverage of hairpin and chainbreak base pair motifs in RNAStrAlign dataset. **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 (n=3,966 RNAs). **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 black circles.

Comparison with SOTA

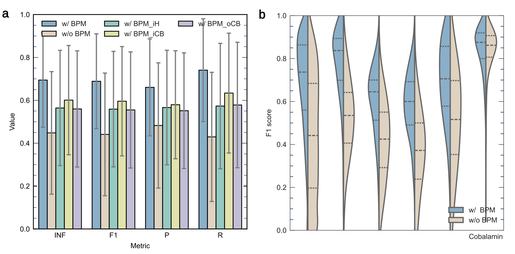


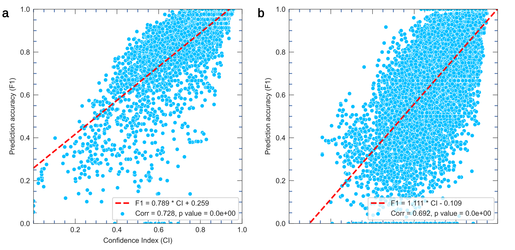




- Performance comparison on bpRNA-TS0, ArchiveII, Rfam12.3-14.10, and PDB datasets.
- Visualization of predicted RNA secondary structures of three examples.
- **a** Comparison of inference time. **b** Inference time of BPfold on RNA targets within different lengths.

Ablation study and confidence index





- **a** Ablation study of BPfold under five configurations of BP motifs (BPM) on family-wise dataset Rfam12.3-14.10. **b** Violinplot.
- Correlation of F1 score and confidence index on ArchiveII and Rfam12.3-14.10 datasets.

doi: 10.1101/2024.10.22.619430

codes: github.com/heqin-zhu/BPfold

mysite: heqin-zhu.github.io

MIRACLE: miracle.ustc.edu.cn