NC-Bench and NCfold: A Benchmark and Closed-Loop Framework for RNA Non-Canonical Base-Pair Prediction

Heqin Zhu[1,*], Ruifeng Li[2,*], Ao Chang[1], Mingqian Li[2], Hongyang Chen[2,✉], Peng Xiong[1,✉], Shaohua Kevin Zhou[1,✉]

[1]University of Science and Technology of China, [2]Zhejiang University

ICLR

data and codes →

Motivation

- Lack of a standardized benchmark for non-canonical (NC) base pair prediction
- Scarcity and imbalanced distribution of NC data
- Difficulty of existing methods in capturing complex structures

Contribution

- Construction of NC-Bench, the first benchmark for NC base pair prediction
- Proposal of the NCfold framework for effective utilization of RNA foundation model (RFM) base-to-base priors
- Provision of a novel evaluation standard for RNA structure prediction

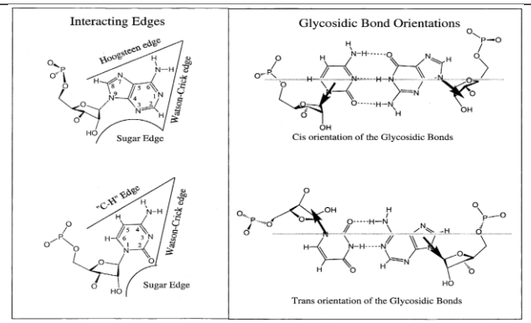


Fig from Leontis-Westhof 2001

Non-canonical (NC) base pairs are essential for RNA structural stability, catalytic activity, and biological regulation, playing a fundamental role in determining RNA structure and function.

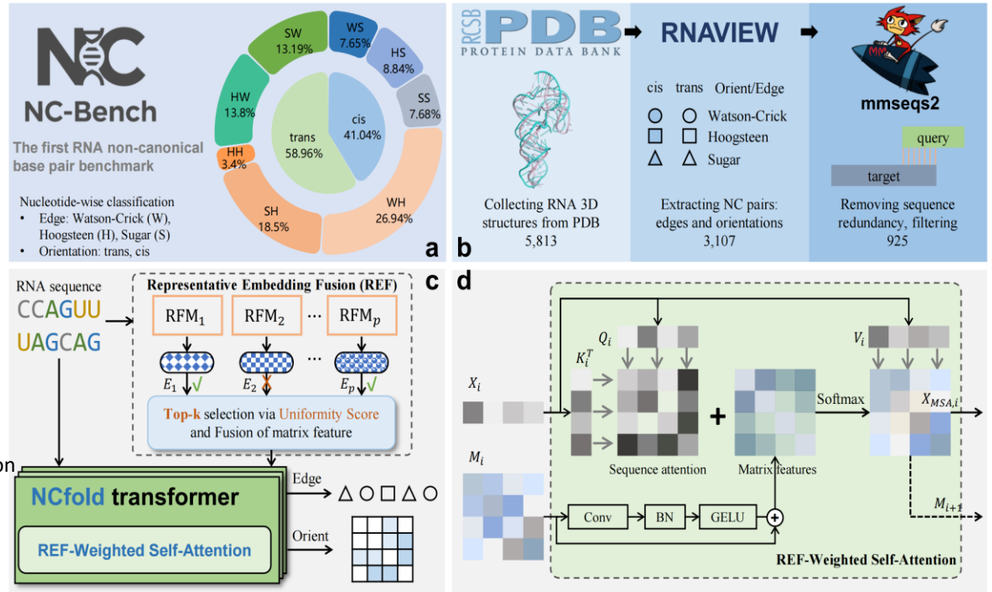
NC-Bench Pipeline & NCfold Architecture

NC-Bench (a, b)

- The first NC-pair benchmark
- Edgex (W, H, S)
- Orientations (trans, cis)
- PDB->RNAVIEW->mmseqs2

NCfold (c, d)

- NC base-to-base knowledge priors from RNA foundation model
- Tok-k selection according to uniformity (IsoScore)
- REF: Representative Embedding Fusion
- REF-weighted self-attention: add the base-to-base matrix feature to the attention map of sequence modeling



Benchmark Results

- NCfold performs better than existing 7 RFMs and 7 traditional machine learning methods.

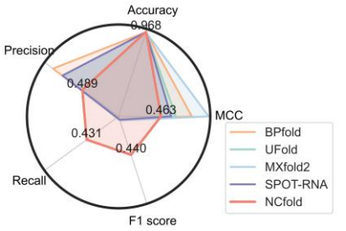
Model	Pair-Edge					Orientation				
	MCC	ACC	P	R	F1	MCC	ACC	P	R	F1
Random Forest	-0.020	0.285	0.232	0.269	0.168	0.002	0.976	0.369	0.377	0.372
Gradient Boosting	0.042	0.554	0.274	0.293	0.258	0.002	0.924	0.371	0.379	0.365
XGBoost	-0.008	0.439	0.246	0.265	0.221	0.000	0.979	0.376	0.385	0.380
SGD	-0.028	0.178	0.106	0.211	0.101	0.111	0.792	0.396	0.539	0.369
Logistic Regression	0.007	0.106	0.096	0.173	0.093	0.026	0.050	0.327	0.434	0.037
KNN	0.014	0.345	0.255	0.279	0.188	0.000	0.979	0.376	0.385	0.380
MLP	0.000	0.638	0.185	0.281	0.218	0.163	0.939	0.395	0.472	0.402
RNA-FM	0.000	0.638	0.185	0.281	0.218	0.000	0.984	0.352	0.359	0.355
structRFM	0.000	0.638	0.185	0.281	0.218	0.005	0.972	0.358	0.361	0.357
RNAErnie	0.000	0.638	0.185	0.281	0.218	0.024	0.948	0.357	0.372	0.354
SpliceBERT	0.000	0.638	0.185	0.281	0.218	0.003	0.972	0.350	0.358	0.352
UTR-LM	0.000	0.638	0.185	0.281	0.218	-0.001	0.983	0.352	0.358	0.355
AIDO.RNA-650M	0.000	0.638	0.185	0.281	0.218	0.003	0.963	0.358	0.359	0.355
RiNALMo	0.000	0.638	0.185	0.281	0.218	0.007	0.955	0.348	0.357	0.345
NCfold (top-1)	0.211	0.596	0.375	0.386	0.341	0.285	0.966	0.474	0.524	0.482
NCfold (top-2)	0.245	0.628	0.400	0.409	0.365	0.312	0.951	0.487	0.544	0.486
NCfold (top-3)	0.219	0.603	0.370	0.376	0.336	0.265	0.948	0.463	0.520	0.466
NCfold (top-4)	0.194	0.605	0.347	0.366	0.321	0.256	0.935	0.458	0.516	0.456

Zero-shot Results

Existing methods:

- only can predict paring status (paired, unpaired)
- Trained on canonical pairs
- High precision but extremely low recall

Model	Pairing Status				
	MCC	ACC	P	R	F1
BPfold	0.799	0.965	0.885	0.021	0.033
UFold	0.625	0.965	0.752	0.027	0.047
MXfold2	**0.984**	0.965	**0.984**	0.000	0.000
SPOT-RNA	0.573	0.965	0.757	0.022	0.039
NCfold	0.463	**0.968**	0.489	**0.431**	**0.440**



Visualization and Top-k Ablation

- NCfold predicts accurate NC pairs
- RFM count of different top-k (Best: top-2)