Enhancing RNA Motif Representation with Contrastive Learning and Language Models for Sequence-Structure Analysis

Vinay Chaudhari[1], Md. Sharear Saon[2], Grace Fu[3], Brent M. Znosko[2], Jie Hou[1*]

[1]Department of Computer Science, Saint Louis University, St. Louis, MO, 63112, [2]Department of Chemistry, Saint Louis University, St. Louis, MO, 63112, and [3]Parkway South High School, Manchester, MO, 63021, USA

Introduction

RNA secondary structural motifs, such as stems, loops, and bulges, are fundamental units of RNA folding and structure. These motifs influence critical biological processes, including gene regulation and molecular interactions. Despite extensive RNA sequence data, accurately predicting 3D structures remains challenging due to the scarcity of experimentally resolved RNA structures and the limitations of traditional tools to identify homologous RNA motifs.

This work focuses on enhancing RNA motif representation by integrating contrastive learning and RNA language models. By combining sequence and structure embeddings, we aim to improve motif clustering, classification, and sequence-structure alignment.

Methods

A. Sequence-Based Learning

RNA motif sequences are processed using the RNA language model[1] for embedding extraction:

Method 1: Default embeddings from RNA-FM.

Method 2: Fine-tuned embeddings from RNA-FM, optimized for motif-specific sequence classification.

B. Structure-Based Learning

RNA motif 3D structures are encoded using a 3D ResNet model:

Method 1: Train 3D ResNet for multi-class classification (25 motif types)

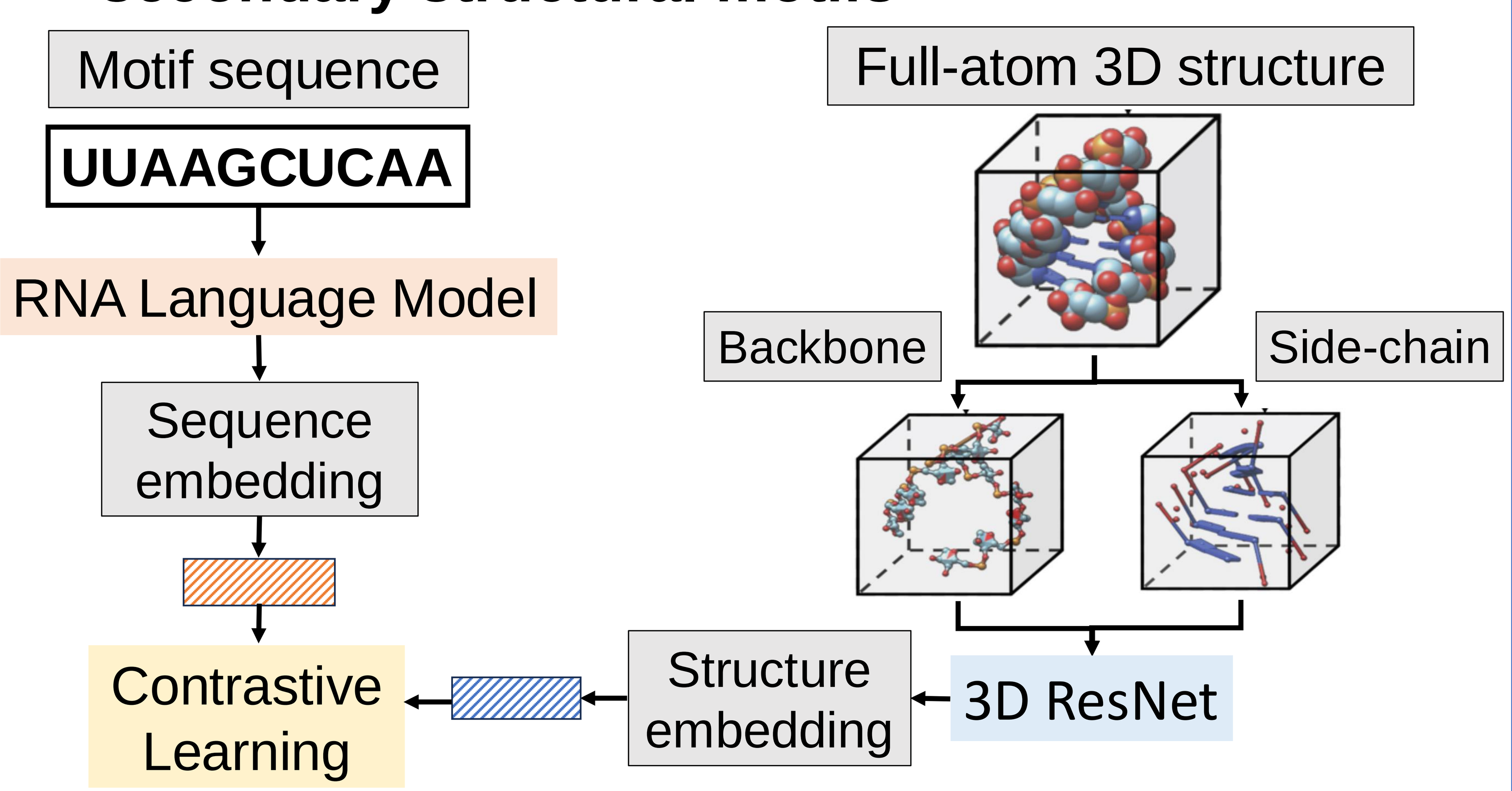
Method 2: Apply contrastive learning directly to structural data for motif representation.

C. Contrastive Learning

To integrate sequence and structure information, we

1) align sequence and structure embeddings for RNA motifs.
2) capture relationships between motifs' sequence features and their structural characteristics.

D. Framework of Contrastive Learning for RNA secondary structural motifs



References

1. Chen, J., Hu, Z., Sun, S., Tan, Q., Wang, Y., Yu, Q., ... & Li, Y. (2022). Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint arXiv:2204.00300*.

2. Richardson, K. E., Kirkpatrick, C. C., & Znosko, B. M. (2020). RNA CoSSMos 2.0: an improved searchable database of secondary structure motifs in RNA three-dimensional structures. *Database*, *2020*, baz153.

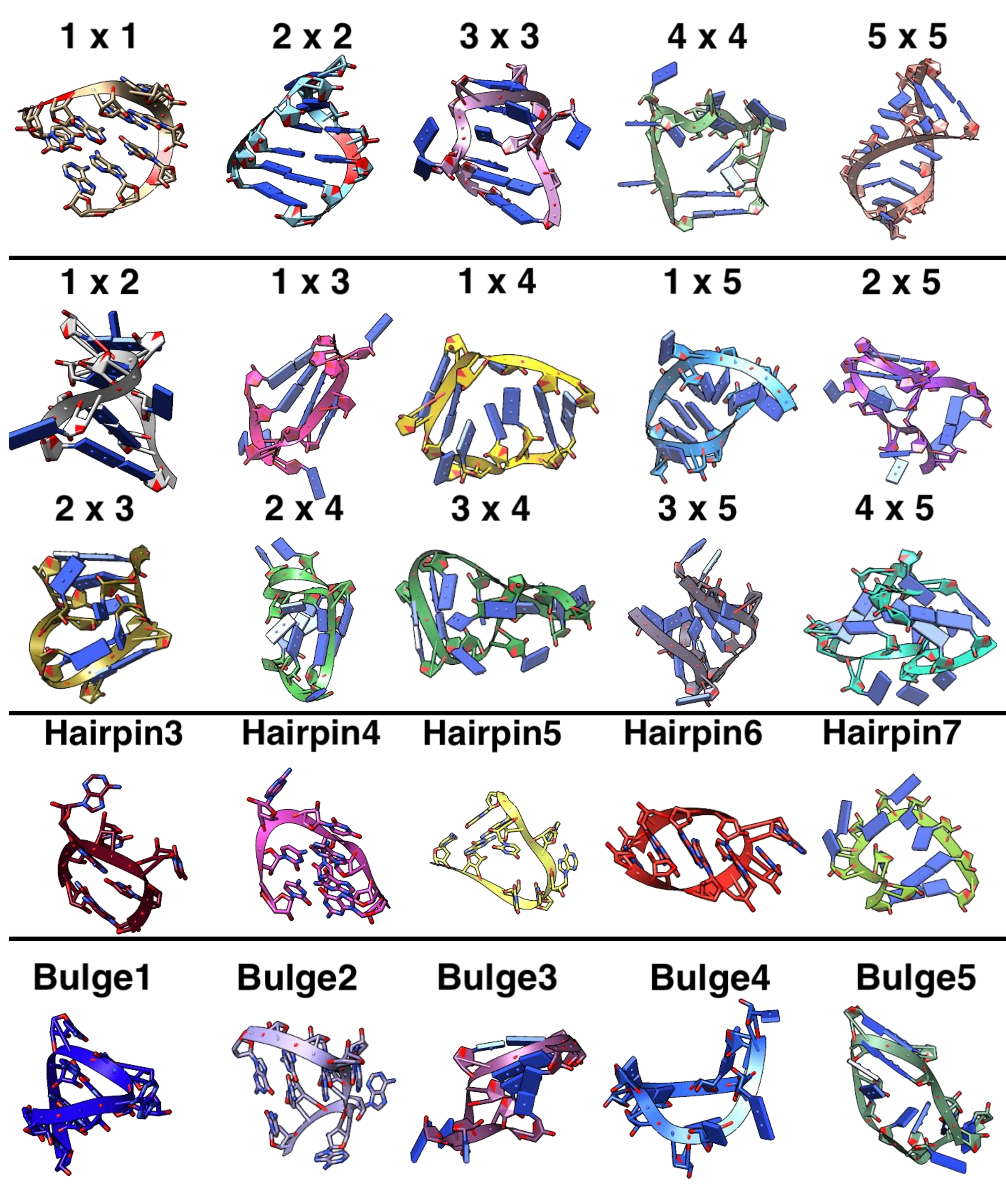
Acknowledgements

NIH NIGMS 1R15GM155891-01(JH) 2R15GM085699-04(BZ)

Dataset

Statistics and Visualization of Motifs in CoSSMos[2] Database

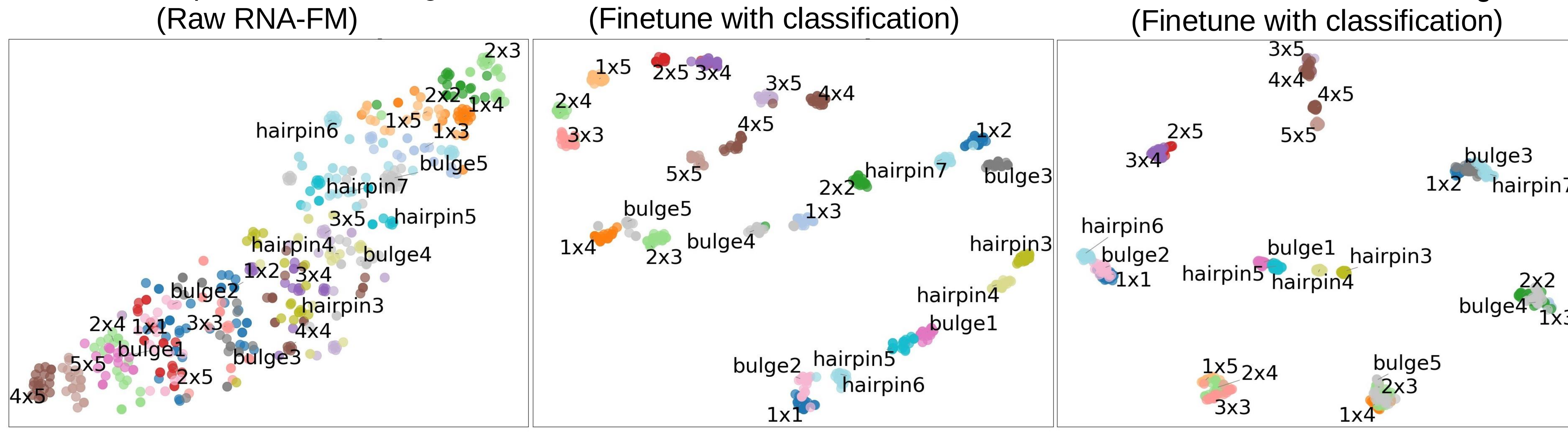
Type	Motif	All PDB	X-ray	NMR	cryo-EM
Symmetric Internal Loops	1x1	5605	1390	465	3750
	2x2	3291	1374	130	1787
	3x3	3078	1169	186	1723
	4x4	114	21	30	63
	5x5	341	169	0	172
Asymmetric Internal Loops	1x2	4298	1519	275	2504
	1x3	3770	1806	121	1843
	1x4	2339	1044	85	1210
	1x5	458	158	0	300
	2x3	3409	1459	89	1861
	2x4	1413	554	39	820
	2x5	585	143	30	412
	3x4	1645	497	19	1129
	3x5	115	74	0	41
	4x5	325	106	9	210
Hairpin Loops	hairpin3	3274	1440	243	1591
	hairpin4	42973	18228	2048	22697
	hairpin5	17836	7033	801	10002
	hairpin6	13890	5512	688	7690
	hairpin7	11561	5039	318	6204
Bulge Loops	bulge1	37048	14568	1318	21162
	bulge2	13258	5565	310	7383
	bulge3	3474	1876	224	1354
	bulge4	851	205	79	567
	bulge5	318	151	84	83
	Total	175269	71100	7591	96558



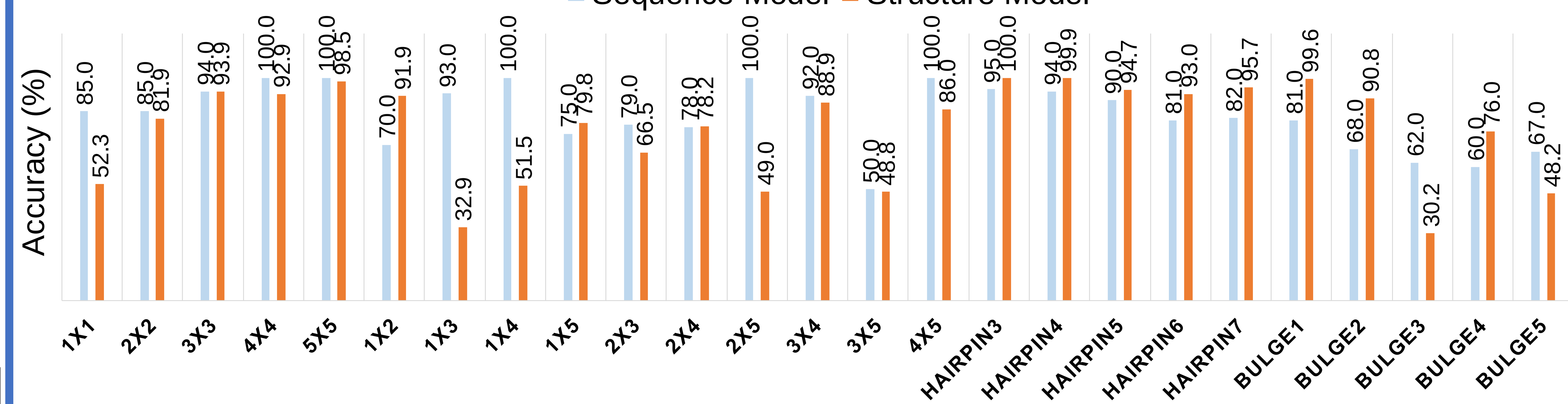
Results

I. Visualization of RNA motif embeddings for three different deep learning models

Motif Sequence Embedding (Raw RNA-FM) — Motif Sequence Embedding (Finetune with classification) — Motif Structure Embedding (Finetune with classification)



II. Evaluation of Motif classification using sequence model and structure model



III. Motif structure and sequence embedding rearrangement through contrastive learning

Before Contrastive Learning — After Contrastive Learning