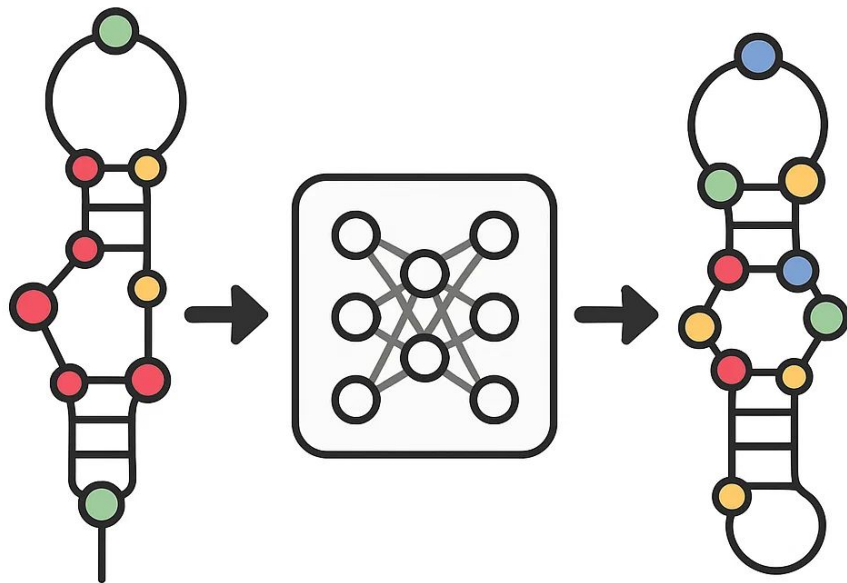


RNA Structure Prediction with GNN

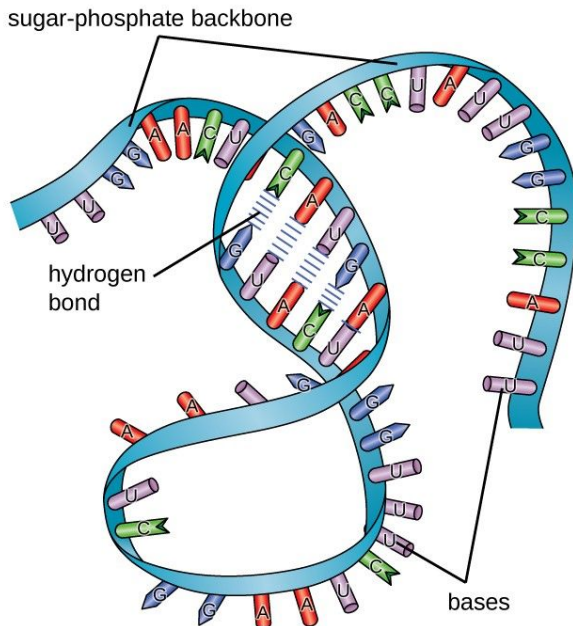
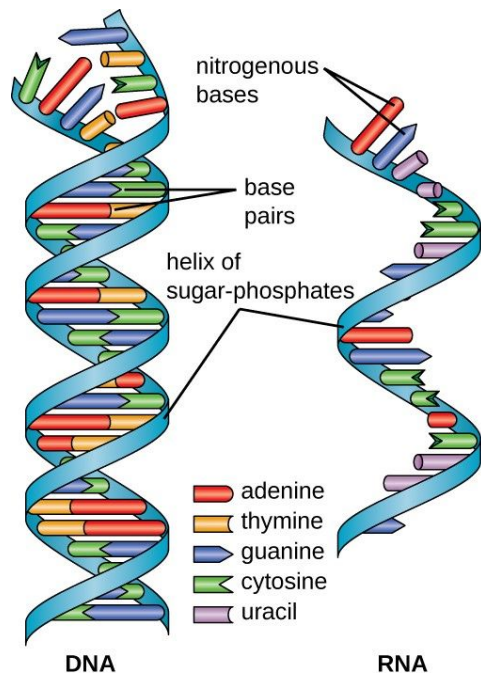


Springboard - Capstone 3 - Shen Dong
June 2025

Diversity in RNA structure and function

RNA Type	Primary Function	Structure Type	Typical Size
mRNA	Carries genetic code from DNA to the ribosome for protein synthesis	Linear	~1,000–10,000 nt
tRNA	Brings amino acids to the ribosome during translation	Cloverleaf	~70–90 nt
rRNA	Forms the core structural and functional components of the ribosome	Complex secondary structure	120–5,000 nt
snRNA	Involved in RNA splicing (removing introns from pre-mRNA)	Hairpin	~150 nt
snoRNA	Guides chemical modifications of rRNA and other RNAs	Hairpin	~60–300 nt
miRNA	Regulates gene expression by silencing mRNA	Hairpin	~21–25 nt
siRNA	Involved in RNA interference; silences specific mRNA	Double-stranded	~21–23 nt
piRNA	Protects germline cells from transposable elements	Single-stranded	~24–31 nt
lncRNA	Regulates gene expression at transcriptional and post-transcriptional levels	Linear	>200 nt
ncRNA	General term for RNA that does not encode proteins	Various	Variable
gRNA	Guides Cas proteins to specific DNA sequences for cutting/editing	Hairpin	~100 nt
crRNA	Part of bacterial immune system; guides Cas to target sequences	Hairpin	~39–42 nt
tmRNA	Rescues stalled ribosomes in bacteria	Structured	~350–400 nt
rasiRNA	Silences repetitive DNA in heterochromatin regions	Double-stranded	~24–30 nt
exRNA	RNA found outside cells, involved in cell-cell communication	Various	Variable

Complexity of RNA structure determination



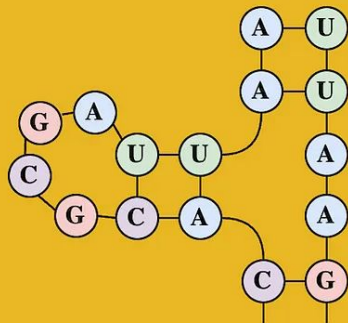
- RNA molecules fold into complex 3D shapes essential for function.
- Experimental 3D methods (X-ray, cryo-EM) are time-consuming and low-throughput.
- Deep learning, especially Graph Neural Networks (GNNs), can learn structural patterns directly from data.
- This project integrates various features such as sequence, thermodynamics, secondary structure, and evolutionary signals into a GNN to predict 3D coordinates.

Features



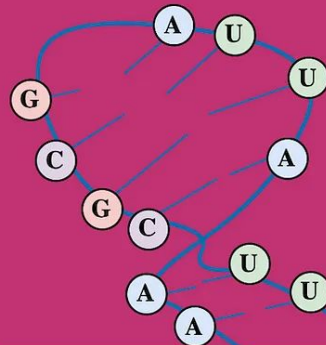
Raw sequence

One-hot encoded
Primary sequence



Secondary structure

Dot-bracket structure
MFE
BPPM



Tertiary structure

3D coordinates for
each residu.

x, y, z

Target feature

A	C	G	U	C	C	U	G
A	G	G	G	C	A	U	C
A	A	G	U	A	C	G	G
A	C	U	U	C	C	U	A

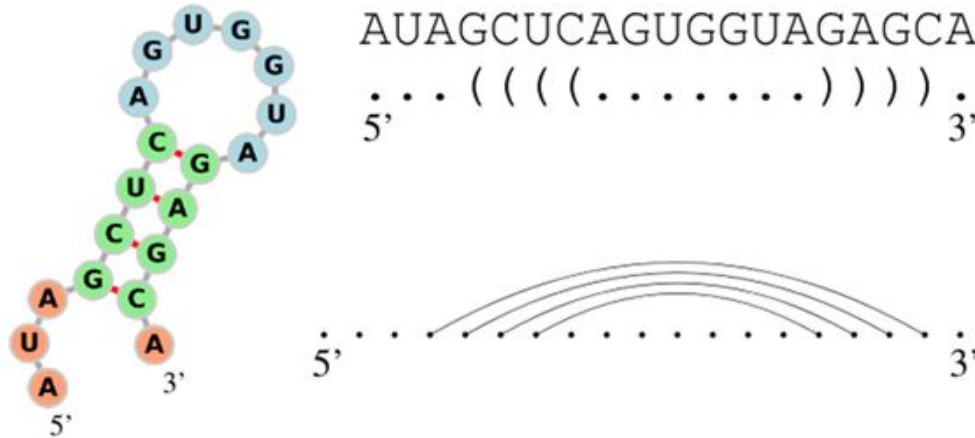
MSA

Entropy
PSSM
MI-Matrix

Secondary Structures features - RNAfold

- thermodynamically stable secondary structure with the Minimum Free Energy (MFE) using dynamic programming (Zuker algorithm).
- Outputs:
 - the dot-bracket annotation,
 - MFE value
 - the Base Pairing Probability Matrix (BPPM) of the best structure

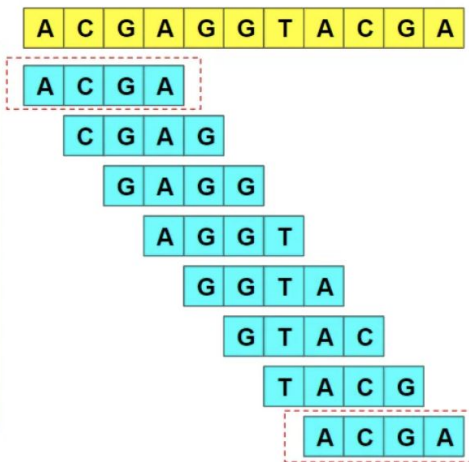
Example of dot-bracket annotation



- . represents an unpaired nucleotide,
- (and) represent paired nucleotides (forming base pairs),
- Nested and stacked brackets show more complex structural relationship

Secondary Structures features - Kmers

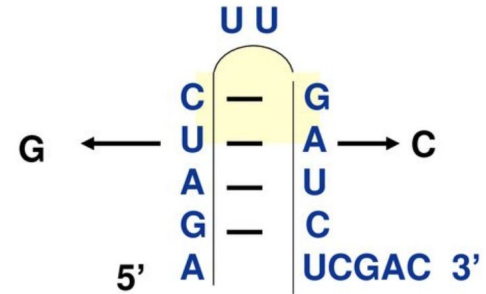
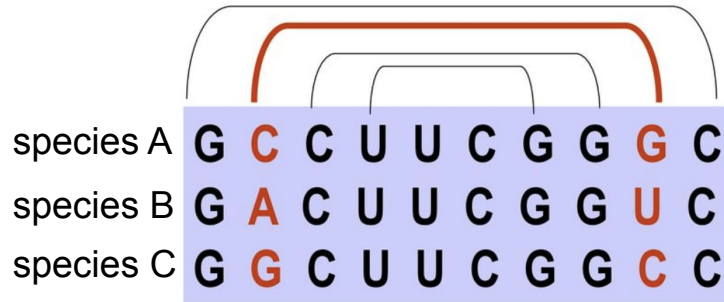
	Total	Distinct	Unique
ACGA	2	1	0
CGAG	1	1	1
GAGG	1	1	1
AGGT	1	1	1
GGTA	1	1	1
GTAC	1	1	1
TACG	1	1	1



- **Local Pattern Recognition:** Capture sequence motifs (e.g., loops, stems) by analyzing triplets of nucleotides.
- **Structural Context Clues:** Tri-mers reflect base-pairing tendencies and help identify folding patterns.
- **Improved Model Inputs:** Add biologically rich features beyond single nucleotides for more accurate predictions.

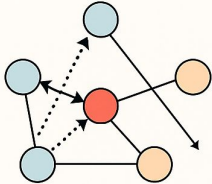
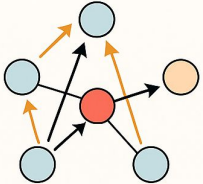
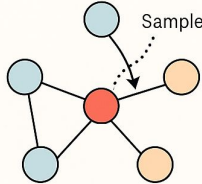
MSA-based features

Multiple sequence alignment



	Entropy	PSSM	MI Matrix
Definition	Measures randomness at a position	Frequencies of nucleotides at a position	Pairwise co-variation between positions
Input	One column from MSA	One column from MSA	Two columns from MSA
Interprets...	Diversity of nucleotides	Composition of nucleotides	Dependency or correlation between positions
Used for...	Detecting conserved regions	Encoding positional sequence data	Detecting structural dependencies

Graph Neural Network

Graph Convolutional Network (GCN)	Graph Attention Network (GAT)	GraphSAGE (Sample and Aggregate)
 <ul style="list-style-type: none">Aggregates features from immediate neighbors	 <ul style="list-style-type: none">Uses attention scores to weigh neighbors	 <ul style="list-style-type: none">Samples and aggregates information from neighbors

Graph Convolutional Network (GCN)

- Implements the **GCNConv** layer from PyTorch Geometric
- Each node aggregates features from its **immediate neighbors**
- Weights are shared across the graph — similar to CNNs for images

Key Properties:

- Local neighborhood averaging:** Combines information from nearby nodes
- Efficient and scalable** for sparse biological graphs like RNA
- Suitable for **static, undirected graphs** with rich node features

Stacked Architecture:

- Two GCN layers followed by a **fully connected (linear)** output layer
- Learns hierarchical, structural features layer by layer

Nodes and Edges

NODE = each residu	EDGE
One-hot nucleotide (4-dim): A/U/C/G.	Sequence adjacency: connect ($i \leftrightarrow i+1$)
k-mer counts (D^k -dim).	Base-pair edges: from BPPM, if $bppm[i,j] > 0.5$, connect ($i \leftrightarrow j$)
Structure vector (3-dim): dot-bracket one-hot.	MSA MI edges: if $mi_matrix[i,j] > 0.1$, connect ($i \leftrightarrow j$)
MFE scalar (1-dim).	
Entropy (MSA) (1-dim).	
PSSM (MSA) (5-dim).	

GNN model architecture

- **Model:** Two-layer Graph Convolutional Network (GCN).
 - Layer 1: GCNConv(in_channels=D_total, out_channels=128) + ReLU.
 - Layer 2: GCNConv(in_channels=128, out_channels=128) + ReLU.
 - Readout: Linear(128 \rightarrow 3) to predict (x,y,z).
- **Input:**
 - **x**: node feature tensor ($L \times D_{\text{total}}$).
 - **edge_index**: graph connectivity ($2 \times E$).
 - **edge_attr** (optional) as weights ($E \times 1$).
- **Output:** Predicted coordinates **pred** of shape ($L \times 3$).

Distance based loss and TM-Score

1. Loss function:

- coord_loss = MSE between the true coordinates and the predicted coordinates
- dist_loss = MSE between true inter-residues distance and predicted inter-residues
 - Invariant to global rotations/translations, only relative distances matter.
- Total loss = $\text{coord_loss} + \alpha * \text{dist_loss}$.
 - Balances accuracy of structure and preservation of spatial relationships using a weight α .

→ Total loss used to compute the gradients of the loss w.r.t. all model parameters using backpropagation.

2. TM-Score:

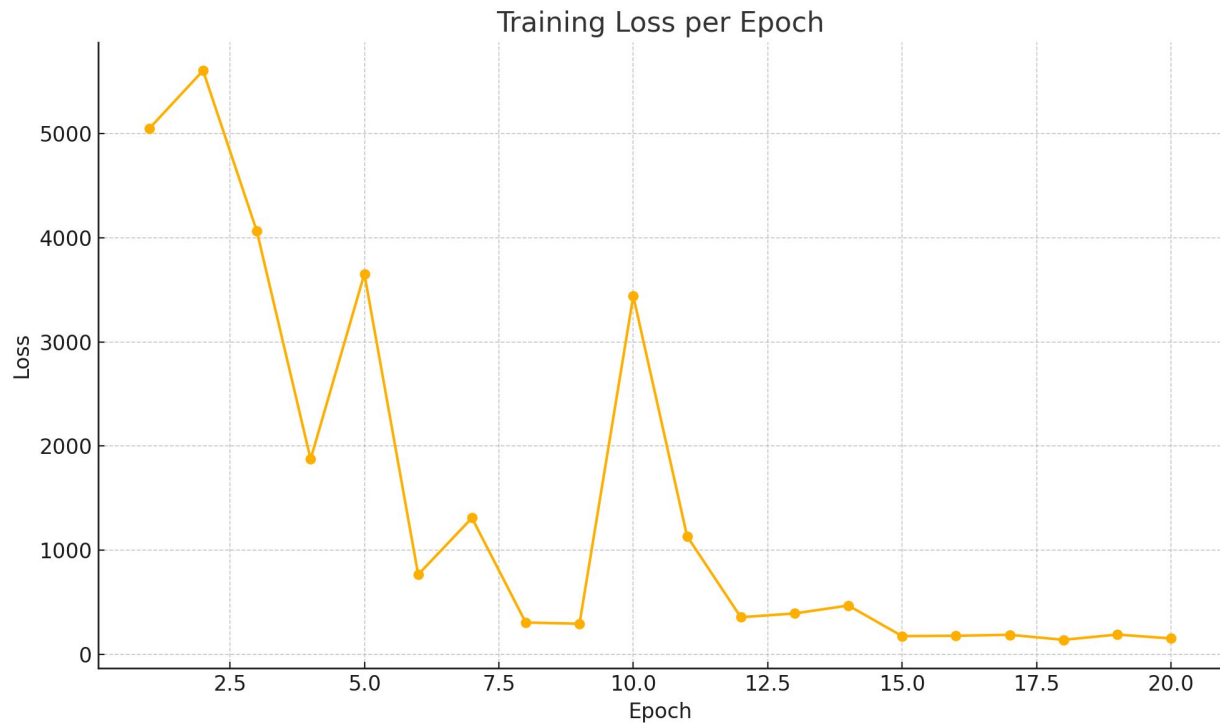
- After predicting pred and converting to NumPy,

- $\text{TM-score} = (1/L) \sum_i [1 / (1 + (d_i / d_0)^2)]$,

where $d_i = \|\text{pred}[i] - \text{true}[i]\|$, and d_0 is length-dependent.

→ TM-score is between 0 to 1, closer to 1 means a better 3D alignment between the predicted and true structure

Convergence of the training loss



Validation set:

● Loss = 1.8

● TM-score = 0.81

Discussion

- **Strengths:**
 - **Rich feature integration**—combining sequence, secondary structure, MFE, k-mers, and evolutionary signals.
 - **Distance-based loss** yields robust training, invariant to rigid motions.
 - Adding **MSA (entropy, PSSM, MI)** significantly improves learning of tertiary contacts.
- **Limitations:**
 - Dataset size (~338 RNAs) is modest, more data would boost performance.
 - Empiric thresholds ($\text{BPPM} > 0.5$, $\text{MI} > 0.1$) might not generalize to all RNAs.
 - Two-layer GCN may be insufficient for very long RNAs or complex tertiary motifs.

Future Directions

- **Scale to Larger Datasets**

Incorporate more RNA sequences with diverse structures to improve generalizability and robustness.

- **Upgrade the GNN Architecture**

Explore advanced models (e.g., Graph Attention Networks, equivariant GNNs, or geometric transformers) for capturing spatial and sequence relationships more effectively.

- **Use Pretrained RNA Embeddings**

Incorporate contextual embeddings from existing models trained on large RNA/protein datasets. Concatenate learned embeddings with current node features

- **Pretraining on RNA Graphs**

Apply self-supervised learning (e.g., masked node/edge prediction) on large unlabeled RNA datasets before fine-tuning on 3D tasks.