Two Tales of RNA Folding

A peek to the RNA Secondary Structure Datasets and RNA Folding Playground

蔡禕濤 Cai Yitao

40th TBI Winterseminar

universität wien

Introduction

- RNA Secondary Structure Dataset Analysis
- RNA Folding Playground

RNA Secondary Structure Representations

RNA Sequence Primary

AGGCUCGAGGAGGUUACCCCACGGGGUCA
GUCAUAGUGUUUGAUAGCAUAUUUCUCGCG
AUGCGCGCUCACAGAACGACGGGAUAUUCA
CCCGACGCACGCUCGGAAGGCCUACGGAU

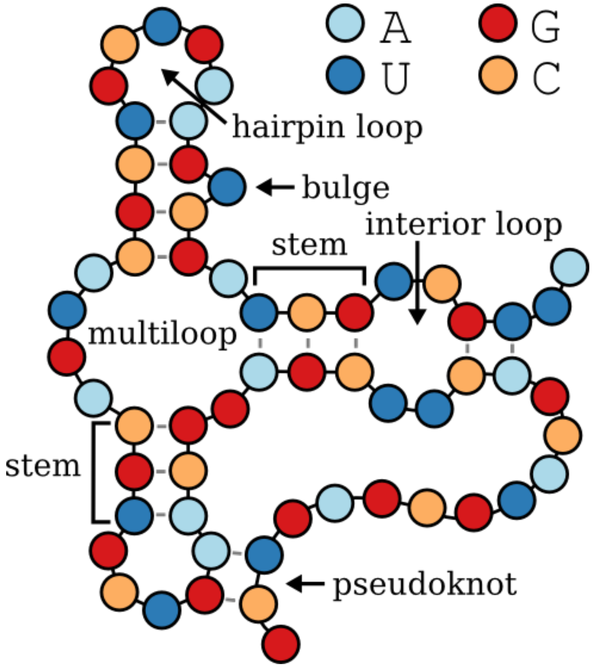
DotBracket

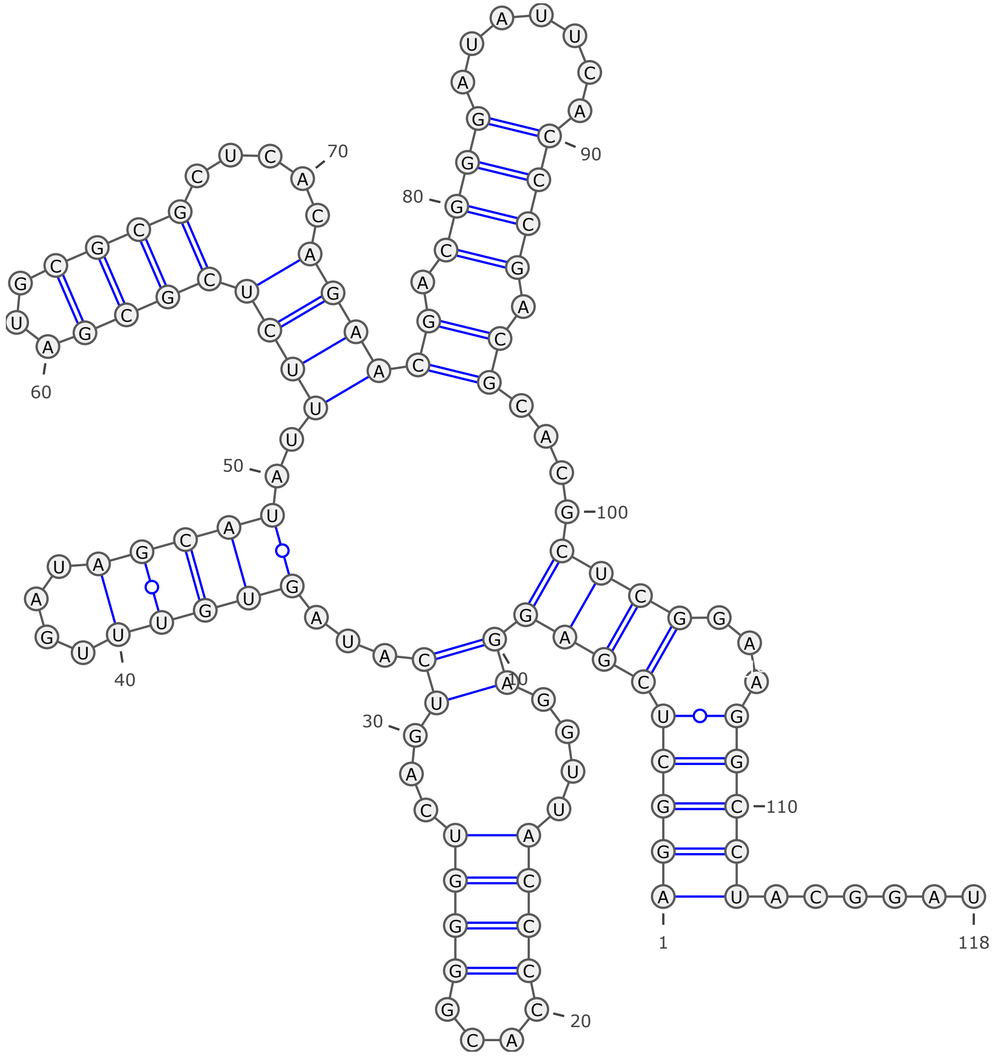
((((((((((….((((….))))…))…(((((….)))))..((((((((…))))).....))))

((.((((…….)))).))….))))…)))))……

RNA Structure Motifs

- Hairpin Loop
- Bulge
- Interior Loop
- Multiloop
- Stem
- Pseudoknot

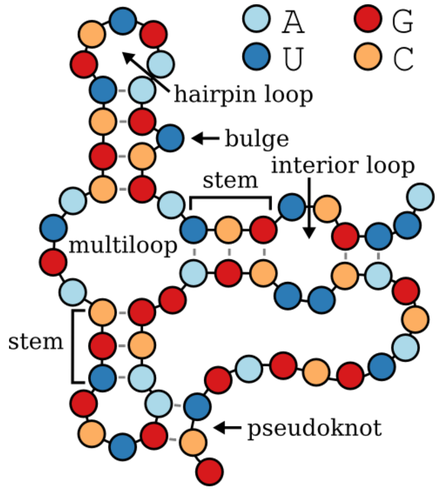




Graph representation of a folded RNA Secondary Structure

What is RNA Secondary Structure Prediction ?

- RNA Secondary Structure Prediction can be decomposed as Structure Motifs Prediction.



$$P(S \mid w) \equiv P(m_0 \mid w) \prod_{i=1}^{M} P(m_i \mid m_{<i}, w) .$$

Where $w \in \{A, C, G, U\}^N$, and $m_i \in \{Stem, Hairpin, Multiloop, Bulge, InteriorLoop\}$ M is the number of motifs

$$P(S \mid w) \equiv P(b_0^{i,j} \mid w) \prod_{k=1}^{K} P(b_k^{i,j} \mid b_{<k}^{i,j}, w) .$$

Where $w \in \{A, C, G, U\}^N$. $i, j \in 1 : N$ and $i < j$, $|i - j| \geq 4$, $b^{i,j} \in \{\{A, U\}\{G, C\}\{G, U\}\}$

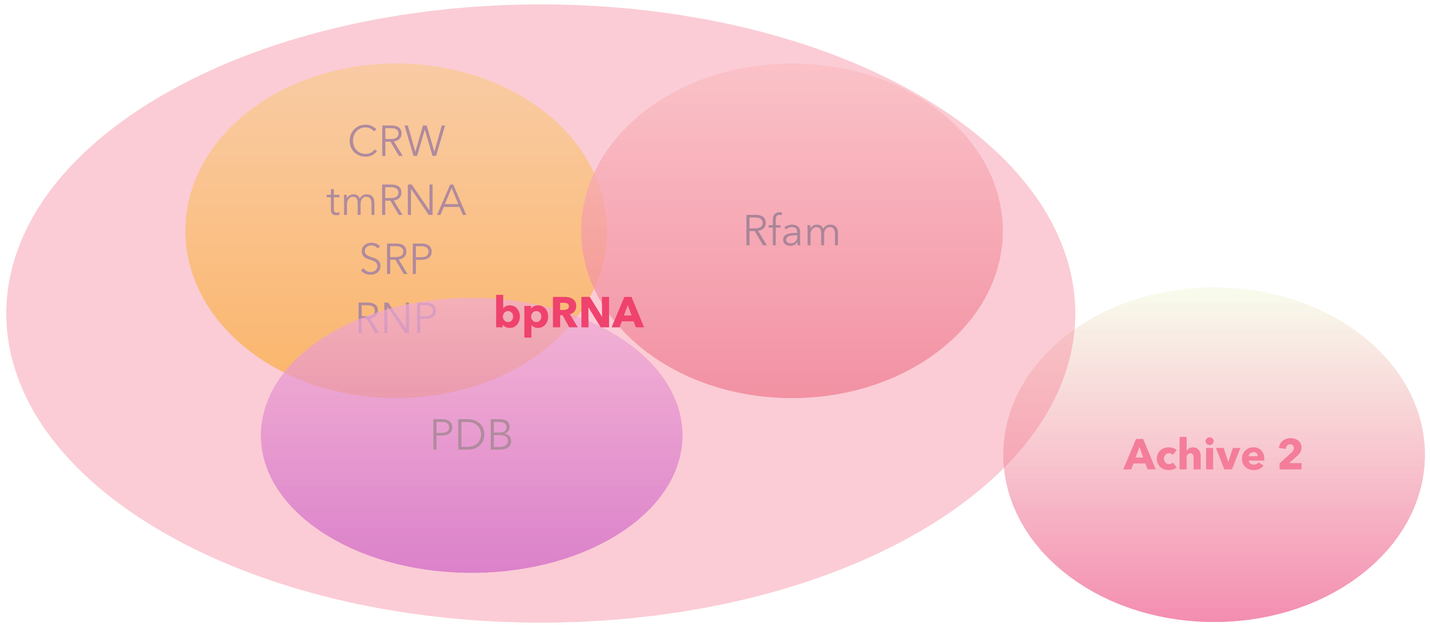
- The Objective is to Maximize the Likelihood of observing a specific RNA Secondary Structure(motifs) given an RNA strand.

Real World Dataset and Synthetic Dataset

- Training and Test set could include the same family of RNA which have very similar structure so the datasets are easily biased.

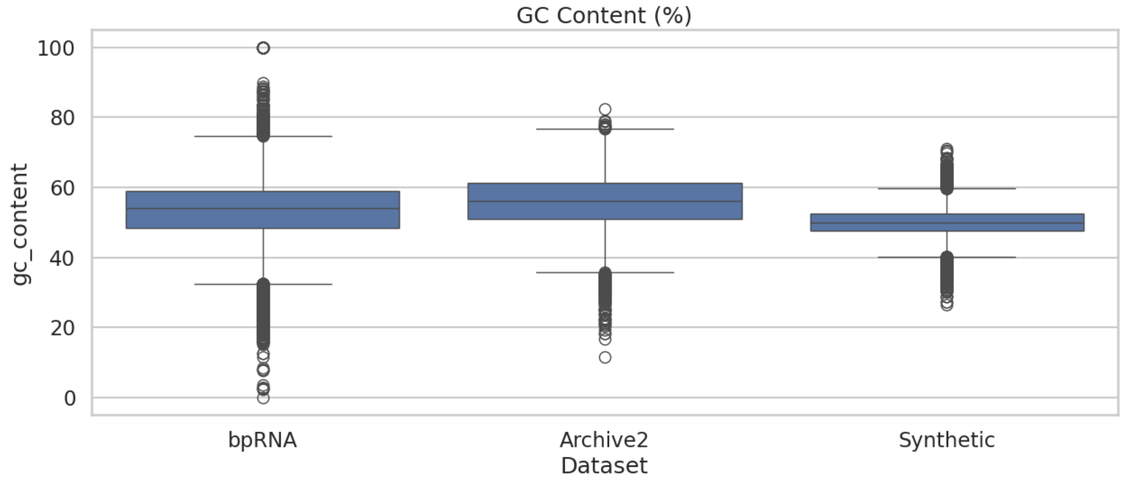
- We can create Synthetic Dataset to benchmark deep learning models with the assumption that if the model performs well in the synthetic dataset, they should also perform well in other real world datasets.

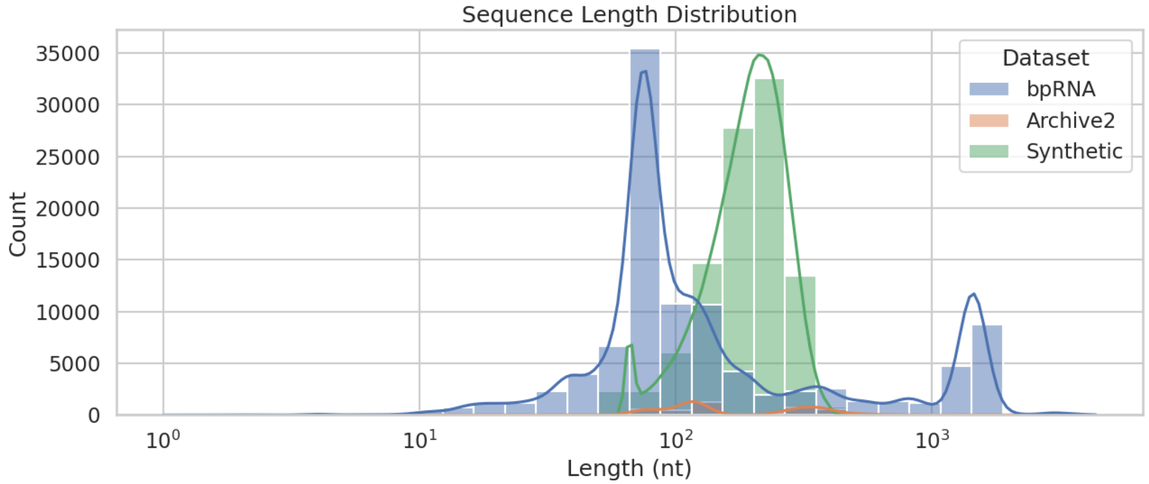
CRW	The Comparative RNA Web (CRW) Site	55,600
tmRNA	tmRNA Database	728
SRP	Signal Recognition Particle Database	959
SPR	Sprinzl tRNA Database (tRNAdb)	623
RNP	The RNase P Database	466
RFAM	The RNA Family Database	43,273
PDB	RCSB Protein Data Bank	669

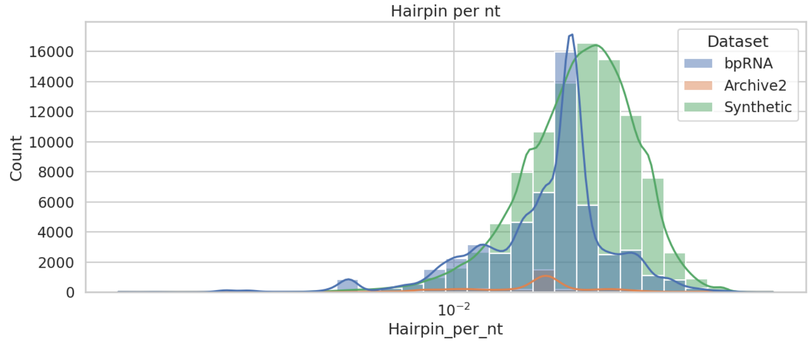


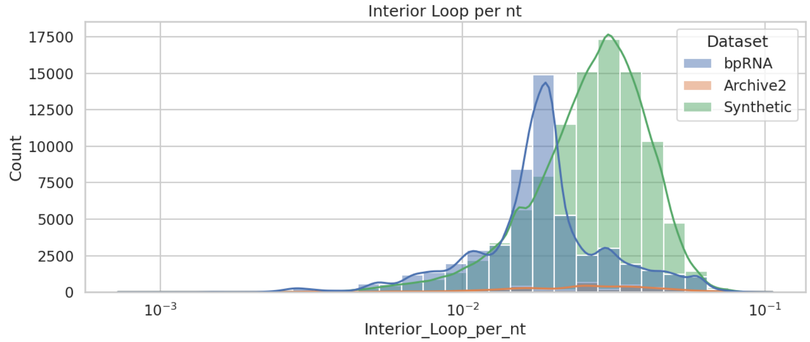
Ref:
Christoph Flamm et.al. Caveats to Deep Learning Approaches to RNA Secondary Structure Prediction
bpRNA dataset: https://bprna.cgrb.oregonstate.edu/index.html

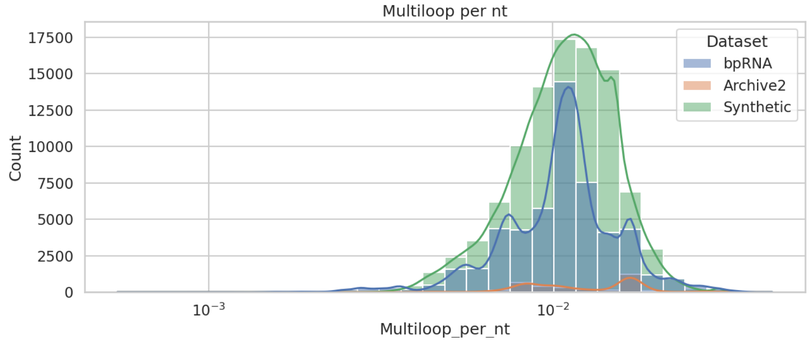
Comparison between Real World Datasets and Synthetic Dataset

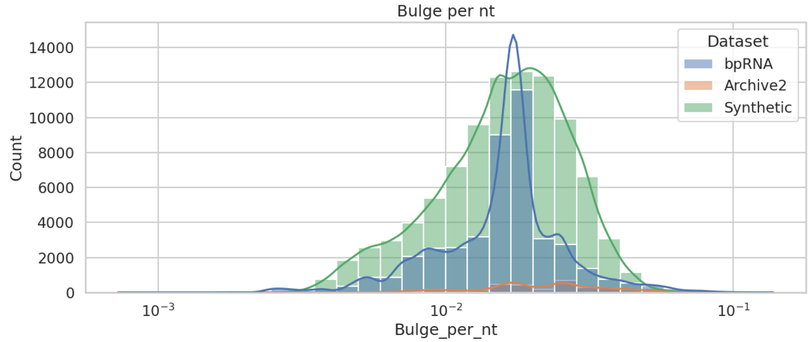


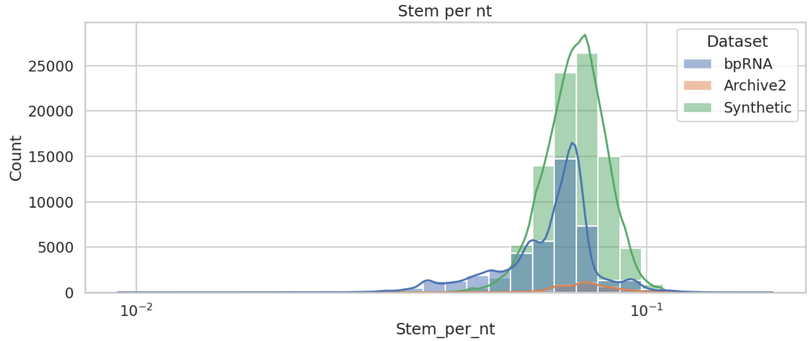


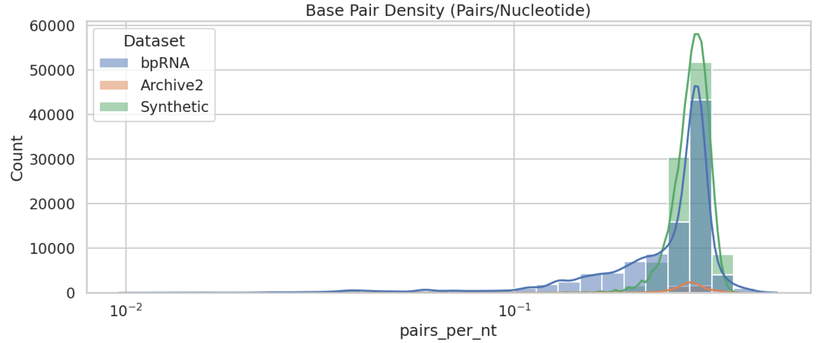












Building Synthetic Benchmark Datasets to Challenge the Generalization of AI models

- **Controlling Complexity**: RNAfold can generate random sequences with varying degrees of structural complexity. d= ensemble diversity

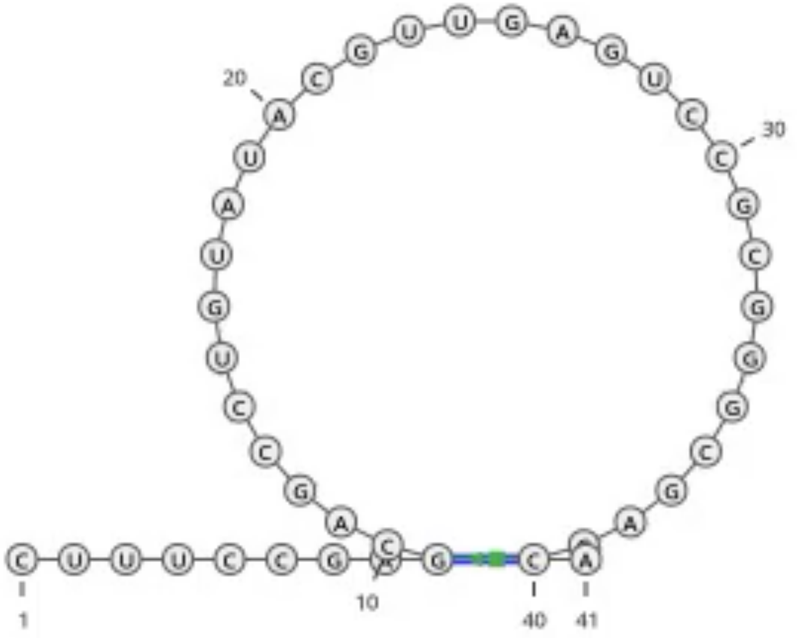
  Easy dataset should satisfy : $d < \theta_1$

  Medium dataset should satisfy : $\theta_1 \leq d \leq \theta_2$

  Hard dataset should satisfy : $d > \theta_2$

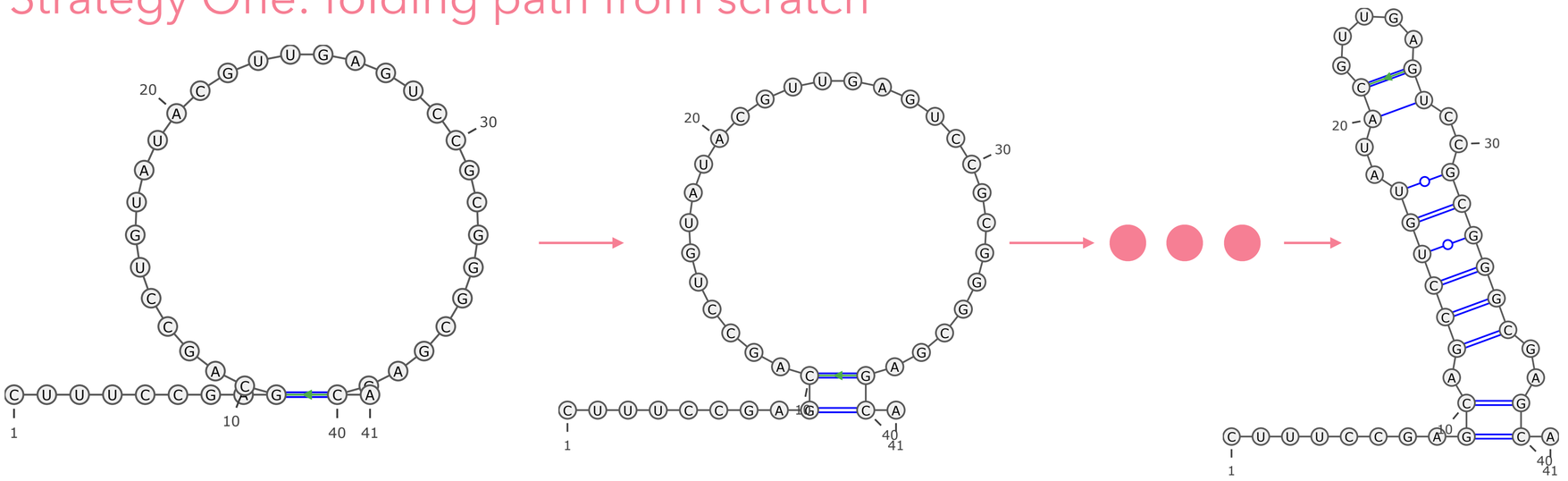
- **Perturbation of Energy Model to Test the Generalization of AI models**
  - AI models trained on datasets derived from standard energy models (e.g., RNAfold) may overfit to specific thermodynamic rules.
  - Perturbed energy models with noise simulate alternative folding dynamics, helping to test whether the model can generalize to unseen conditions or unexpected structural features.

RNA Folding Playground



RNA Folding Playground

Strategy One: folding path from scratch



Actions

Action Space:

If we assume that
1. In each position the agent have two options of actions, i.e. add base pair or remove base pair,

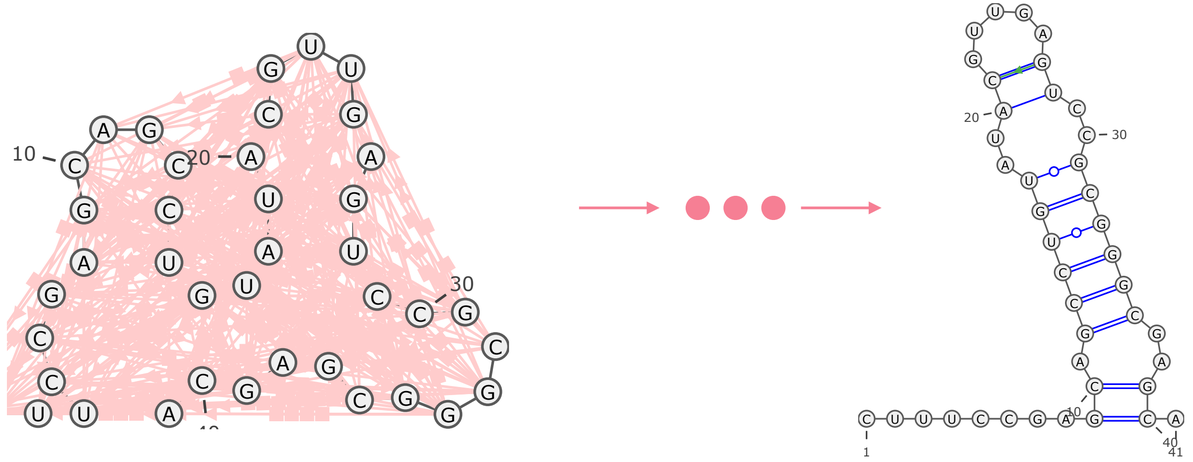
2. The base pair (i, j) = (j, i) and |i-j| >= 4

Then all the possible actions would be

$$\frac{n^2 - 4n + 6}{2}$$

RNA Folding Playground

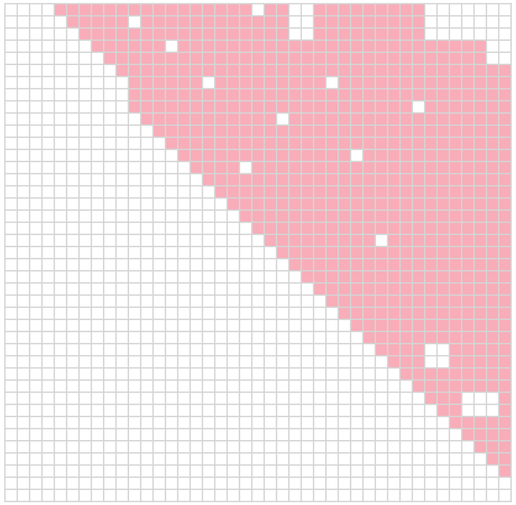
Strategy Two: folding path from fully connected canonical RNA graph to target RNA graph.



Action Space:

If we assume that

1. In each position the agent have two options of actions, i.e. add base pair or remove base pair,
2. The base pair (i, j) = (j, i) and |i-j| >= 4



Then the allowable N actions would be

$\propto$ Number of all possible canonical pairs M

Summary

- RNA Secondary Structure Prediction can be decomposed to the prediction of base pairs.
- RNA Folding process learning can be fun like playing a game.

Acknowledgement:(Surname Alphabetical Order)

Christopher Flamm

Ivo Hofacker

Hua-Ting Yao

TBI Team

StruDL program

Thank You!