From Sequence to Simulation: Trying RAG to Decode and Evolve RNA Molecules

Work in Progress, preliminary results

OBJECTIVE & OUTLINE

Construct a Retrieval-Augmented Generation (RAG) system to generate a descriptive summary of an RNA molecule based only on its sequence, using the following contextual sources:

1. Relevant text is retrieved from biology textbooks using question embeddings via transformers.
2. Image embeddings from RNA figures are also used to retrieve semantically related textual descriptions, enabling cross-modal retrieval.
3. 3D structure information generated from the RNA sequence using recent DRfold2 model, followed by structural feature extraction with DSSR (Dissecting the Spatial Structure of RNA).
4. Structural refinement through cyclic application of DRfold2 and RiboDiffusion, enabling mutual correction between RNA sequences and predicted 3D structures via bidirectional inference.
5. An attempt to model RNA evolution using a large language model (LLM)

Literature used:

***LLM**: Yanis Labrak et al., BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains, (**2024**)*

***DRfold2**: Li, Yang Li et al., Ab initio RNA structure prediction with composite language model and denoised end-to-end learning}, Cold Spring Harbor Laboratory, (**2025**)*

***Ribodiffusion**: Han Huang et al., RiboDiffusion: tertiary structure-based RNA inverse folding with generative diffusion models, Bioinformatics, (**2024**)*

***RNA-FM**: Jiayang Chen et al., Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions, bioRxiv, (**2022**)*

***DSSR**: Xiang-Jun Lu, DSSR-enabled innovative schematics of 3D nucleic acid structures with PyMOL, Nucleic Acids Research, (**2020**)*

Large Language Model for Biomedical Domain I

BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains

Yanis Labrak*[1,2] Adrien Bazoge*[3,4]
Emmanuel Morin[4] Pierre-Antoine Gourraud[3] Mickael Rouvier [1] Richard Dufour [1,4]
[1]LIA, Avignon Université [2]Zenidoc
[3]Nantes Université, CHU Nantes, Clinique des données, INSERM, CIC 1413, F-44000 Nantes, France
[4]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
{firstname.lastname}@univ-avignon.fr
{firstname.lastname}@univ-nantes.fr

Abstract

Large Language Models (LLMs) have demonstrated remarkable versatility in recent years, offering potential applications across specialized domains such as healthcare and medicine. Despite the availability of various open-source LLMs tailored for health contexts, adapting general-purpose LLMs to the medical domain presents significant challenges. In this paper, we introduce BioMistral, an open-source LLM tailored for the biomedical domain, utilizing Mistral as its foundation model and further pre-trained on PubMed Central. We conduct a comprehensive evaluation of BioMistral on a benchmark comprising 10 established medical question-answering (QA) tasks in English. We also explore lightweight models obtained through quantization and model merging approaches. Our results demonstrate BioMistral's superior performance compared to existing open-source medical models and its competitive edge against proprietary counterparts. Finally, to address the limited availability of data beyond English and to assess the multilingual generalization of medical LLMs, we automatically translated and evaluated this benchmark into 7 other languages. This marks the first large-scale multilingual evaluation of LLMs in the medical domain. Datasets, multilingual evaluation benchmarks, scripts, and all the models obtained during our experiments are freely released.

1 Introduction

In the rapidly evolving landscape of Natural Language Processing (NLP), generative Large Language Models (LLMs) like ChatGPT (OpenAI, 2023) and Vicuna (Zheng et al., 2023) have revolutionized human-computer interactions, demonstrating remarkable versatility and advanced capabilities across various tasks and domains. These

The emergence of open-source LLMs such as BLOOM (Workshop et al., 2023) and LLaMA (Touvron et al., 2023a) underscores the transformative potential of these models, facilitating their innovative use in specialized domains including medicine (Dave et al., 2023).

However, integrating LLMs into healthcare and medicine presents unique challenges and opportunities (He et al., 2023; Zhou et al., 2024). While preliminary adoption has opened new avenues for innovation, concerns about data privacy risks associated with proprietary models like MedPaLM-2 (Singhal et al., 2023b) and GPT-4 (Nori et al., 2023a) have arisen. The community's interest in specialized LLMs for healthcare has led to initiatives like PMC-LLaMA (Wu et al., 2023) and MedAlpaca (Han et al., 2023). However, the adoption of open-source medical models has been limited, primarily due to the lack of lightweight models allowing commercial use with performance comparable to larger or proprietary models. To address this gap, there is a need to develop specialized models based on open-source foundation ones like GPT-NeoX (Black et al., 2022), LLaMa 2 (Touvron et al., 2023b), or Mistral (Jiang et al., 2023), and to optimize them for use on consumer-grade devices while maintaining performance.

In this work, we present BioMistral 7B, a specialized LLM tailored for the biomedical domain, derived from Mistral 7B Instruct v0.1 (Jiang et al., 2023) and further pre-trained on PubMed Central. Our contributions include:

1. The construction of BioMistral 7B, the first open-source Mistral-based model for the biomedical domain, including the analysis of different evaluation strategies such as few-shot in-context learning and supervised fine-tuning.

BioMistral is built upon the Mistral foundation model and further pre-trained on **PubMed Central** to enhance its biomedical domain knowledge.
A comprehensive evaluation was conducted on a benchmark of **10 established medical question-answering (QA) tasks**.

Question and Answer Example

You are an expert in the origins of life and RNA structural evolution.

Task: Answer the question.

Question: **What is RNA hairpin?**
Please explain the roles of:
Stem length
Loop size

Answer:
An RNA hairpin is a secondary structure in RNA molecules, consisting of a stem of base pairs and a loop of unpaired nucleotides. The stem length and loop size are important factors that affect the stability and function of RNA hairpins. The stem length is the number of base pairs in the stem, and longer stems generally lead to more stable hairpins. The loop size is the number of nucleotides in the loop, and smaller loops generally lead to more stable hairpins. RNA hairpins play important roles in many biological processes, including the formation of RNA secondary structures, the regulation of gene expression, and the processing of RNA molecules.

```
model_name = "BioMistral/BioMistral-7B-TIES"
Quantization strategie for Biomistral:
 bnb_config = BitsAndBytesConfig(
        load_in_8bit=True,
        llm_int8_threshold=6.0,
        llm_int8_enable_fp32_cpu_offload=True
    )
```

Yanis Labrak et al., BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains, (2024)

Large Language Model for Biomedical Domain II

MEDALPACA - AN OPEN-SOURCE COLLECTION OF MEDICAL CONVERSATIONAL AI MODELS AND TRAINING DATA

A PREPRINT

arXiv:2304.08247v3 [cs.CL] 18 Mar 2025

Tianyu Han[1,+], Lisa C. Adams[2,+], Jens-Michalis Papaioannou[4], Paul Grundmann[4], Tom Oberhauser[4], Alexei Figueroa[4], Alexander Löser[4], Daniel Truhn[1,+], and Keno K. Bressem[2,5,+]

[1]Department of Radiology, University Hospital Aachen, Aachen, Germany
Email: {tianyu.han, dtruhn}@ukaachen.de

[2]Department of Diagnostic and Interventional Radiology, Technical University of Munich, School of Medicine and Health, Klinikum rechts der Isar, TUM University Hospital, Munich, Germany
Email: lisa.adams@tum.de

[4]Berliner Hochschule für Technik (BHT), Berlin, Germany
Email: {michalis.papaioannou, pgrundmann, tom.oberhauser, alexei.figueroa, aloeser}@bht-berlin.de

[5]Department of Cardiovascular Radiology and Nuclear Medicine, Technical University of Munich, School of Medicine and Health, German Heart Center, TUM University Hospital, Munich, Germany
Email: keno.bressem@tum.de

[+]Contributed equally

March 20, 2025

ABSTRACT

As large language models (LLMs) like OpenAI's GPT series continue to make strides, we witness the emergence of artificial intelligence applications in an ever-expanding range of fields. In medicine, these LLMs hold considerable promise for improving medical workflows, diagnostics, patient care, and education. Yet, there is an urgent need for open-source models that can be deployed on-premises to safeguard patient privacy. In our work, we present an innovative dataset consisting of over 160,000 entries, specifically crafted to fine-tune LLMs for effective medical applications. We investigate the impact of fine-tuning these datasets on publicly accessible pre-trained LLMs, and subsequently, we juxtapose the performance of pre-trained-only models against the fine-tuned models concerning the examinations that future medical doctors must pass to achieve certification.

Keywords Natural Language Processing · Artificial Intelligence · Medicine

1 Introduction

The advent of large language models (LLMs), trained using reinforcement learning through human feedback (RLHF) and exemplified by OpenAI's GPT series, has profoundly influenced the fields of natural language processing (NLP) and artificial intelligence (AI) research [1]. Their remarkable capacity to produce coherent, contextually apt, and intricate responses has increased their value across diverse domains. Notably, the medical field is poised to reap substantial benefits from the implementation of these models.

Context:
An RNA hairpin is a common secondary structure formed when a single strand of RNA folds back.

Question:
What are main functions of RNA hairpin?

Answer of BioMistral:
1. RNA hairpins are involved in many biological processes, including protein synthesis, gene regulation, and RNA splicing.
2. RNA hairpins can be used to design artificial RNA molecules with specific functions, such as RNA interference (RNAi) and aptamers.
3. RNA hairpins can be used as therapeutic targets for the development of drugs that target specific RNA molecules.
4. RNA hairpins can be used as biomarkers for the diagnosis of diseases.

Answer of medalpaca-7b:
1. RNA hairpin can act as a binding site for RNA-binding proteins.
2. RNA hairpin can also act as a site for RNA editing.

Medical Meadow, a collection of medical tasks was compiled for fine-tuning and evaluating the performance of large language models in the context of medicine.

medalpaca-7b is a large language model specifically fine-tuned for medical domain tasks. It is based on LLaMA (Large Language Model Meta AI) and contains 7 billion parameters. The primary goal of this model is to improve question-answering and medical dialogue tasks.

https://arxiv.org/abs/2304.08247

```
model_name = "medalpaca/medalpaca-7b"
Quantization strategie for medalpaca:
 bnb_config = BitsAndBytesConfig(
        load_in_8bit=True,
        llm_int8_threshold=6.0,
        llm_int8_enable_fp32_cpu_offload=True
    )
```

RNA Prompt analysis using Large Language Model (sequence comparison)

The answer is fully correct.

You are an expert in molecular biology. Your task is to analyze two RNA sequences and determine which is more likely to be found in a real biological organism. Justify your answer in one sentence.
Sequence 1: GGGGGGG
Sequence 2: AAGUCGCGCCGAAAAGGUGUCUCUU
Question: Which sequence is more biologically realistic?
Answer format: One sentence explaining which sequence is more likely and why.

1. **Sequence 2 is more biologically realistic because it has a more diverse nucleotide composition and is not a repetition of a single nucleotide.**

Including additional context would help make the answer clearer and more specific

```
pip install biopython
from Bio.Align import PairwiseAligner
```

```
<class 'Bio.Align.Alignment'>
target            0 GAG--GCG-------GGUG------ 10
                  0 .||--|||-------||||------ 25
query             0 AAGUCGCGCCGAAAAGGUGUCUCUU 25
```

The LLM has difficulty understanding this.


```
{
  "target_sequence": "GAGGCGGGUG",
  "query_sequence": "AAGUCGCGCCGAAAAGGUGUCUCUU",
  "aligned_target": "GAG--GCG-------GGUG------",
  "aligned_query": "AAGUCGCGCCGAAAAGGUGUCUCUU",
  "score": 0.2,
  "identical_nucleotide_counts": {
    "A-A": 1,
    "G-G": 6,
    "C-C": 1,
    "U-U": 1
  }
}
```


dictionaries **are better understood** by LLMs

USE for LLM prompt

RNA sequence to 3D structure

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.05.641632; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ab initio RNA structure prediction with composite language model and denoised end-to-end learning

Yang Li[1], Chenjie Feng[2], Xi Zhang[5], Yang Zhang[1,3,4,5*]

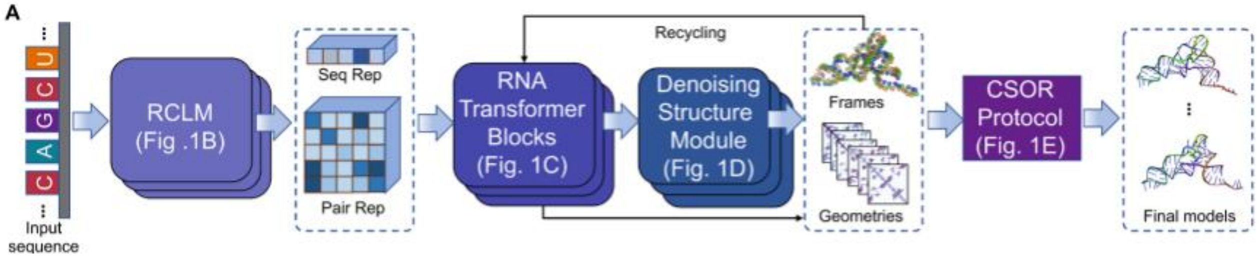
[1]Cancer Science Institute of Singapore, National University of Singapore, 117599, Singapore.
[2]School of Science, Ningxia Medical University, Yinchuan, 750004, China
[3]Department of Computer Science, School of Computing, National University of Singapore, 117417, Singapore.
[4]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 117596, Singapore.
[5]Center for AI and Computational Biology, Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Suzhou, 215123 China

*Correspondence should be addressed to Yang Zhang (email: zhang@zhanggroup.org)

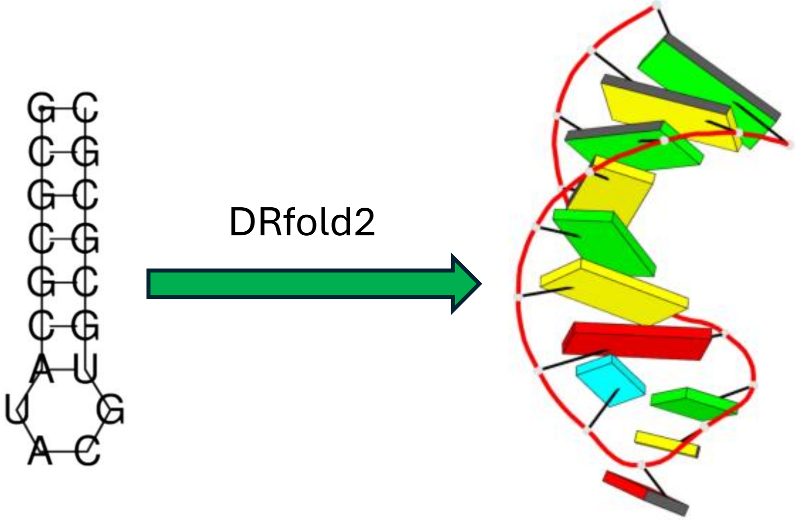
Abstract

RNA structures are essential for understanding their biological functions and developing RNA-targeted therapeutics. However, accurate RNA structure prediction from sequence remains a crucial challenge. We introduce DRfold2, a deep learning framework that integrates a novel pre-trained RNA Composite Language Model (RCLM) with a denoising structure module for end-to-end RNA structure prediction. DRfold2 achieves superior performance in both global topology and secondary structure predictions over other state-of-the-art approaches across multiple benchmark tests. Detailed analyses reveal that the improvements primarily stem from the RCLM's ability to capture co-evolutionary pattern and the effective denoising process, leading to a more than 100% increase in contact prediction precision compared to existing methods. Furthermore, DRfold2 demonstrates high complementarity with AlphaFold3, achieving statistically significant accuracy gains when integrated into our optimization framework. By uniquely combining composite language modeling, denoise-based end-to-end learning, and deep learning-guided post-optimization, DRfold2 establishes a distinct direction for advancing *ab initio* RNA structure prediction.

Keywords: RNA structure prediction, composite likelihood, RNA language model, end-to-end learning



Overview of DRfold2 pipeline for end-to-end RNA structure prediction.



Li, Yang Li et al., Ab initio RNA structure prediction with composite language model and denoised end-to-end learning}, Cold Spring Harbor Laboratory, (2025), doi = {10.1101/2025.03.05.641632}

DRfold2 enables the prediction of RNA 3D structures (in PDB format) from nucleotide sequences provided in FASTA format

RNA Prompt analysis using Large Language Model

The answer is only partially correct.

You are an expert in RNA structural biology.
Task: Determine which of the following RNA sequences is most likely to form a stable hairpin structure.
A stable hairpin forms when the RNA contains internal reverse-complementary regions (that can pair to form a stem) separated by a short unpaired loop.
RNA Sequences:

sequence1: GCGCGCAUACGUGCGCGC
sequence2: AAAAAAAAAAAAAAAAAA

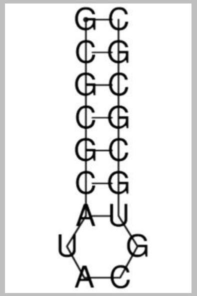
Question: Which sequence is most likely to form a stable hairpin structure, and why?
Answer format:
State the number of the chosen sequence, followed by a concise explanation (1–2 sentences) identifying the reverse-complementary regions that could form a stem and the loop region.

1. The first sequence is most likely to form a stable hairpin structure. The sequence contains two reverse-complementary regions (GCGCGC) separated by a short unpaired loop (CAU)

True correct answer: The sequence contains two reverse-complementary regions (GCGCGCA) separated by a short unpaired loop (GCAU).

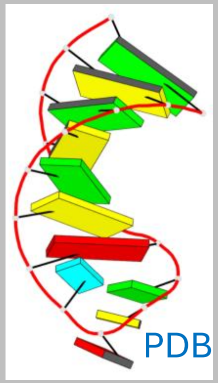
Conclusion: The RNA sequence alone is not sufficient; more contextual information about the RNA is needed in the prompt!

Including additional context would help make the answer clearer and more specific



DRfold2

Li, Yang Li et al., Ab initio RNA structure prediction with composite language model and denoised end-to-end learning}, Cold Spring Harbor Laboratory, (**2025**), doi = {10.1101/2025.03.05.641632}



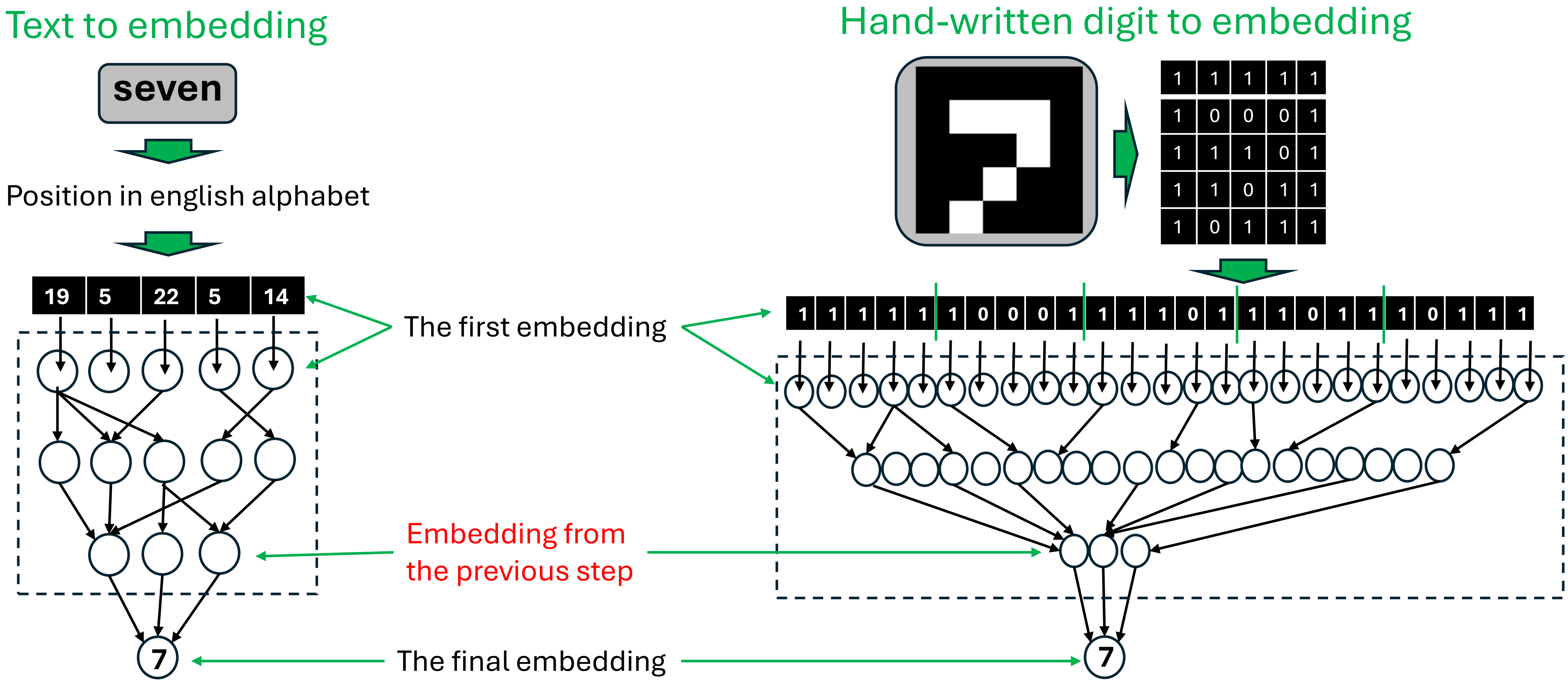
PDB file

DSSR
API: http://skmatic.x3dna.org/api


```
{
    "general_info": {
        "full_RNA_sequence": "GCGCGCAUACGUGCGCGC",
        "length": 18,
        "base_pairs": 7,
        "hydrogen_bonds": 25,
        "dot_bracket": [
            "(((((((....)))))))"
        ],
        "splayUnits": [
            "UA"
        ],
        "hairpins": [
            "AUACGU"
        ],
        "stacks": [
            "AU",
            "ACGU"
        ]
    },
    "helices_info": {
        "full_RNA_sequence": "GCGCGCAUACGUGCGCGC",
        "helix_0": {
            "base_pairs": 7,
            "strand_1": "GCGCGCA",
            "strand_2": "CGCGCGU",
            "helix_form": "AAAAAA"
        }
    }
}
```


Xiang-Jun Lu, DSSR-enabled innovative schematics of 3D nucleic acid structures with PyMOL, Nucleic Acids Research, Vol 48, number 13, p. e74-e74, (2020), doi = {10.1093/nar/gkaa426}

What is Embedding in neural networks



The final embedding is interesting as it captures the concept of "seven" from the text and/or the image.

The previous-step embedding is even more useful, as it enables direct comparison between text and image.

However, image embeddings with text embeddings are not aligned in general case (joint training needed in addition)

How to align text and image with CLIP model

Training Dataset : Images + Captions

C. Jia et al , Representation Learning With Noisy Text Supervision, 2021

Inference

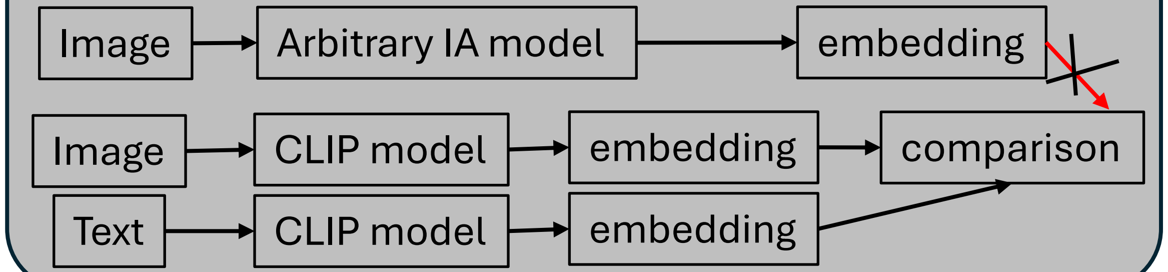


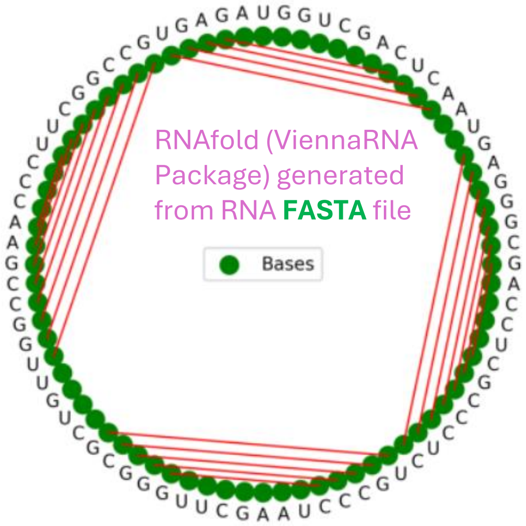
Image from:
https://commons.wikimedia.org/w/index.php?curid=23636924

text	similarity
"crocodile"	0.128
"dog"	0.203
"animal"	0.227
"cat",	0.274
"bengal cat"	0.333
"A striking bengal cat with a golden-orange coat and bold rosette markings, walking with one paw raised and looking upward, illuminated against a deep maroon background."	0.4

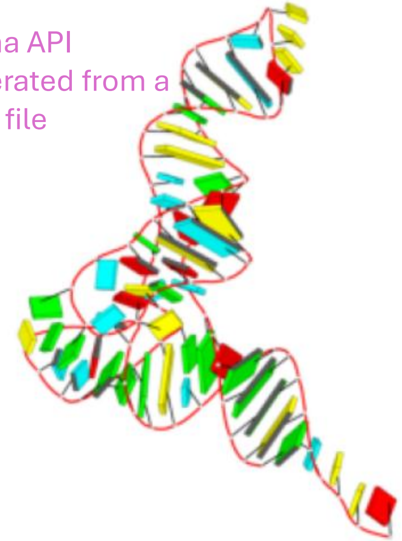
While OpenAI has never explicitly specified or shared the data used to train the original CLIP model, the CLIP paper mentions that the model was trained on **400 million image-text pairs** collected from the Internet.
Phrase from here: https://voxel51.com/blog/a-history-of-clip-model-training-data-advances/

RAG context may include paragraphs from RNA textbooks, retrieved using RNA molecule images

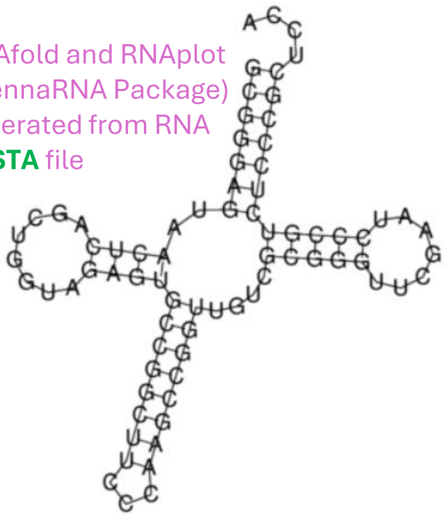
circular secondary structure



DSSR image

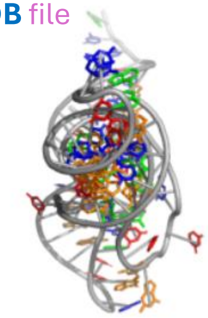


planar secondary structure

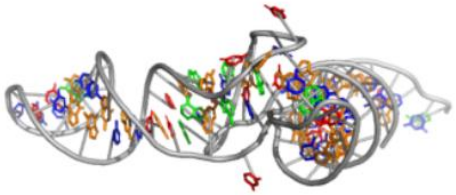


rna view front

rna view side



rna view top



Different images of the same $tRNA^{Gly}$, (RNAcentral ID: URS0002910775)

Approach I: Question Answer with **prompt engineering**

prompt

The instruction:
“You are an assistant trained to answer questions using the given context.”

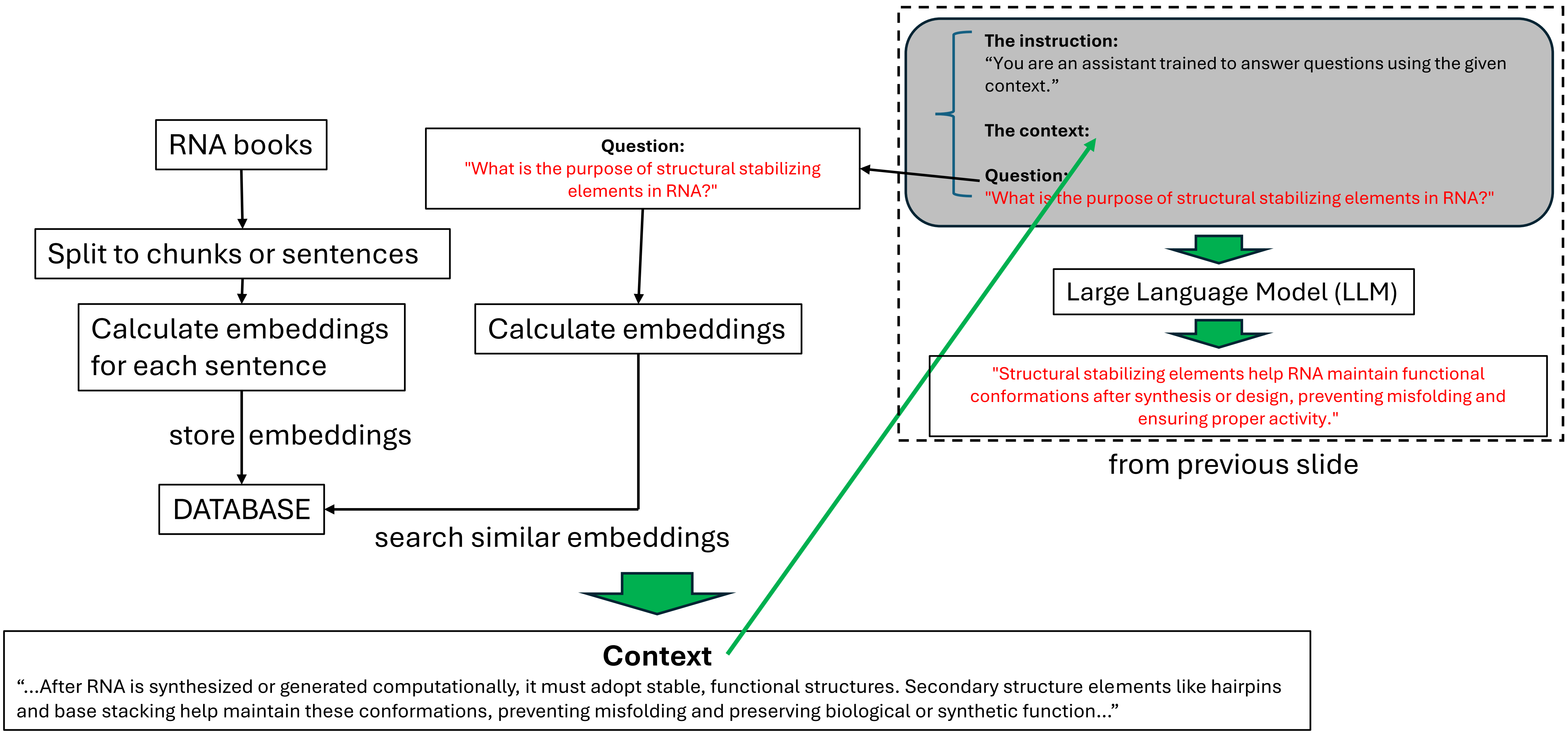
The context (small part of RNA book):
“…After RNA is synthesized or generated computationally, it must adopt stable, functional structures. Secondary structure elements like hairpins and base stacking help maintain these conformations, preventing misfolding and preserving biological or synthetic function…”

Question:
"What is the purpose of structural stabilizing elements in RNA?"

Large Language Model (LLM)

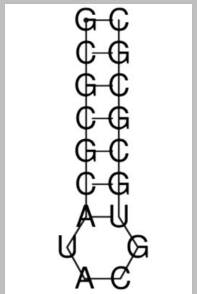
"Structural stabilizing elements help RNA maintain functional conformations after synthesis or design, preventing misfolding and ensuring proper activity."

Approach II: Question Answer with **R**etrieval **A**ugmented **G**eneration



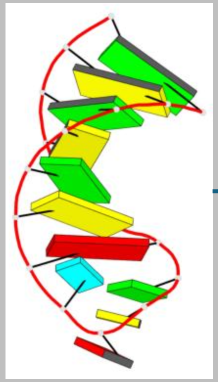
Generate Context from RNA sequence

Including additional context would help make the answer clearer and more specific



DRfold2

DRfold2: Li, Yang Li et al., Ab initio RNA structure prediction with composite language model and denoised end-to-end learning}, Cold Spring Harbor Laboratory, (2025), doi = {10.1101/2025.03.05.641632}



the extracted book paragraph is omitted to comply with copyright law

API: http://skmatic.x3dna.org/api


```
{
    "general_info": {
        "full_RNA_sequence": "GCGCGCAUACGUGCGCGC",
        "length": 18,
        "base_pairs": 7,
        "hydrogen_bonds": 25,
        "dot_bracket": [
            "(((((((....)))))))"
        ],
        "splayUnits": [
            "UA"
        ],
        "hairpins": [
            "AUACGU"
        ],
        "stacks": [
            "AU",
            "ACGU"
        ]
    },
    "helices_info": {
        "full_RNA_sequence": "GCGCGCAUACGUGCGCGC",
        "helix_0": {
            "base_pairs": 7,
            "strand_1": "GCGCGCA",
            "strand_2": "CGCGCGU",
            "helix_form": "AAAAAA"
        }
    }
}
```


the extracted book paragraph is omitted to comply with copyright law

Book example: Mattick J, Amaral P. RNA, the Epicenter of Genetic Information: A new understanding of molecular biology. Abingdon (UK): CRC Press; 2022 Sep 20. PMID: 37847807.

LLM-Based RNA Function Interpretation: Prompt and Answer

You are an expert in the origins of life and RNA structural evolution.

Context:
Context for question: the extracted book paragraph is omitted to comply with copyright law
Context for images: the extracted book paragraph is omitted to comply with copyright law

Task: Answer the question based on the given non-coding RNA sequence.

[1] Non-coding RNA sequence:
GCGGGAGUAACUCAGCUGGUAGAGUGCCGGCUUCCCAAGCCGGUUGUCGCGGGUUCGAAUCCCGUCUCCCGCUCCA

[2] Structural features of the non-coding RNA:
- 3 hairpins
- **helix_0** consists of **12 base pairs**
- **helix_1** consists of **14 base pairs**
- 13 base-stacking interactions

Question:
What are the two most likely functions of this RNA sequence? Provide both a primary and a secondary hypothesis, with brief reasoning for each.

LLM input

tRNA^{Gly}
Rnacentra_id:
URS0002910775

DRfold2

DSSR

RNA textbooks

LLM output

Answer:
The two most likely functions of this RNA sequence are: 1) a structural RNA (e.g., ribosomal RNA, transfer RNA, or ribozyme) and 2) a regulatory RNA (e.g., microRNA, small interfering RNA, or long non-coding RNA). The primary hypothesis is that this RNA sequence is a structural RNA because it contains a large number of base-stacking interactions (13), which are characteristic of structural RNAs. The secondary hypothesis is that this RNA sequence is a regulatory RNA because it contains three hairpins, which are characteristic of regulatory RNAs. However, it is important to note that this RNA sequence may have both structural and regulatory functions.

LLM for RNA design

PROMPT for LLM:

You are an expert in the origins of life and RNA structural evolution.

Task: Propose a plausible ancestral RNA sequence, relying on both the current RNA sequence and on the reconstructed one.

Current RNA sequence:
GGGGCCAUAGCUCAAUUGGCAGAGCGCCGCCUUUGCAAGGCGGAGGCUAGGGGUUCGAUUCCCCUUGGCUCCA

Hairpins of the current RNA sequence:
CAAUUGGCAG, UUGCA, GUUCGAUUC

Reconstructed sequence from 3D structure of the current RNA sequence:
GGGGGCGUAGCUCAGUUGGUAGAGCACUGCCUUUCCAAGGCAGAGGUCAGGGGUUCGAAUCCCCUCGCCUCCA
(Sequence identity: 82.2%)

Requirements for a plausible ancestral RNA sequence:
- Nucleotide changes should be present, affecting no more than approximately 1% of the total nucleotides in the ancestral RNA sequence.
- Length must be between 68 and 73 nucleotides.

The length Between L-5 and L

Output format:
Return the ancestral RNA sequence in uppercase letters, no explanation.

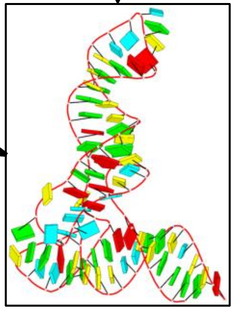
LLM input

Rnacentral_id:
URS00028C2F46
OR any random sequence

DRfold2

DSSR

RiboDiffusion



LLM

https://github.com/PavelPll/RNA_RAG

LLM output

Example 1:
Input:
GGGGCCAUAGCUCAAUUGGCAGAGCGCCGCCUUUGCAAGGCGGAGGCUAGGGGUUCGAUUCCCCUUGGCUCCA
Output:
GGGGCCUAGCUCAAUUGGCAGAGCGCCUUUGCAAGGCGGAGGCUAGGGGUUCGAUUCCCCUUGGCUCCA

Input/output alignement

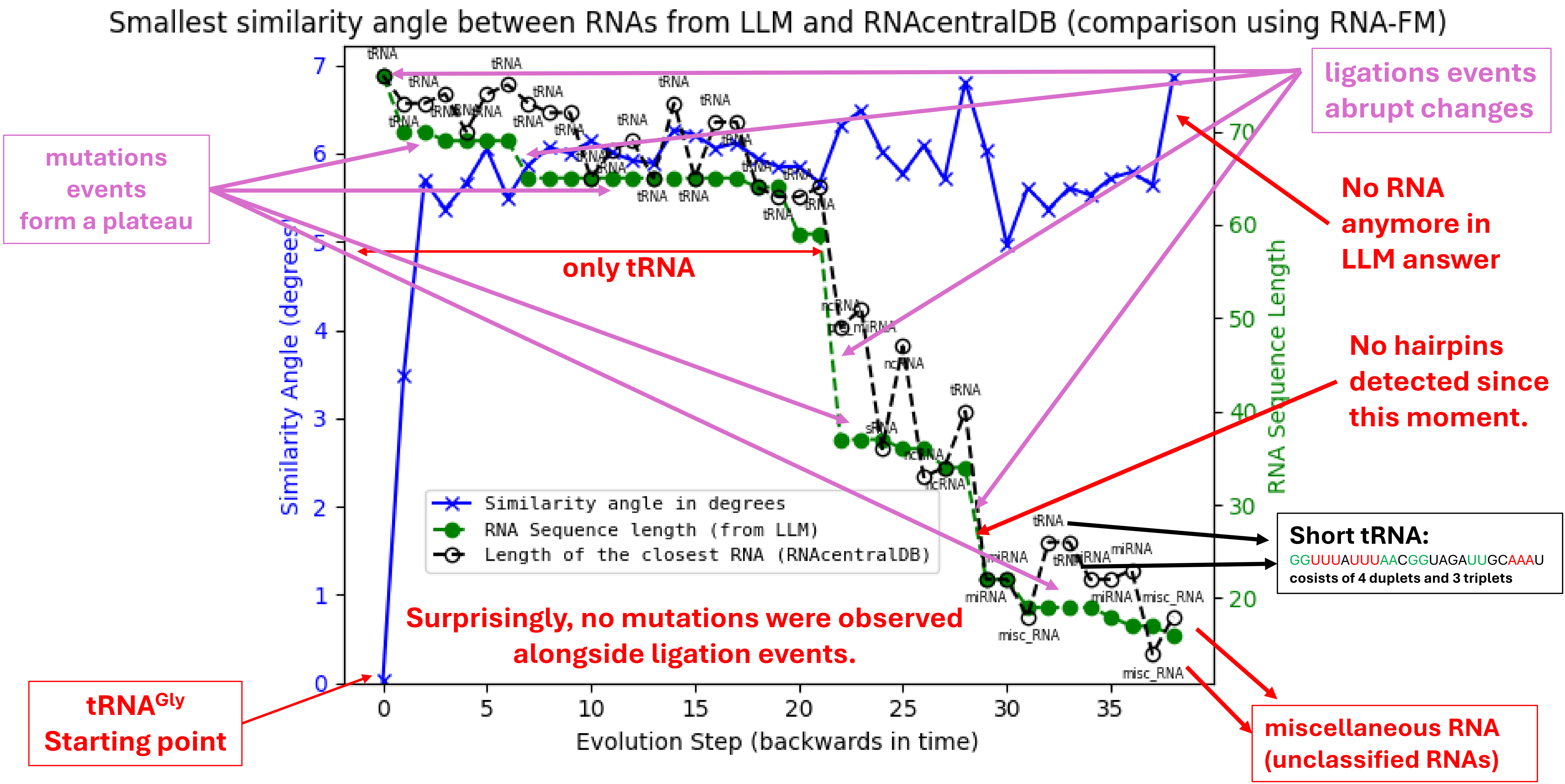
```
<class 'Bio.Align.Alignment'>
target            0 GGGGCCAUAGCUCAAUUGGCAGAGCGCCGCCUUUGCAAGGCGGAGGCUAGGGGUUCGAUU
                  0 ||||||-|||||||||||||||||---|||||||||||||||||||||||||||||||||
query             0 GGGGCC-UAGCUCAAUUGGCAGAG---CGCCUUUGCAAGGCGGAGGCUAGGGGUUCGAUU

target           60 CCCCUUGGCUCCA 73
                 60 ||||||||||||| 73
query            56 CCCCUUGGCUCCA 69

Score: 133.0
```

This is the first attempt to use a large language model (LLM) for RNA design. Further work is needed, particularly to account for evolutionary scenarios where RNA length changes abruptly

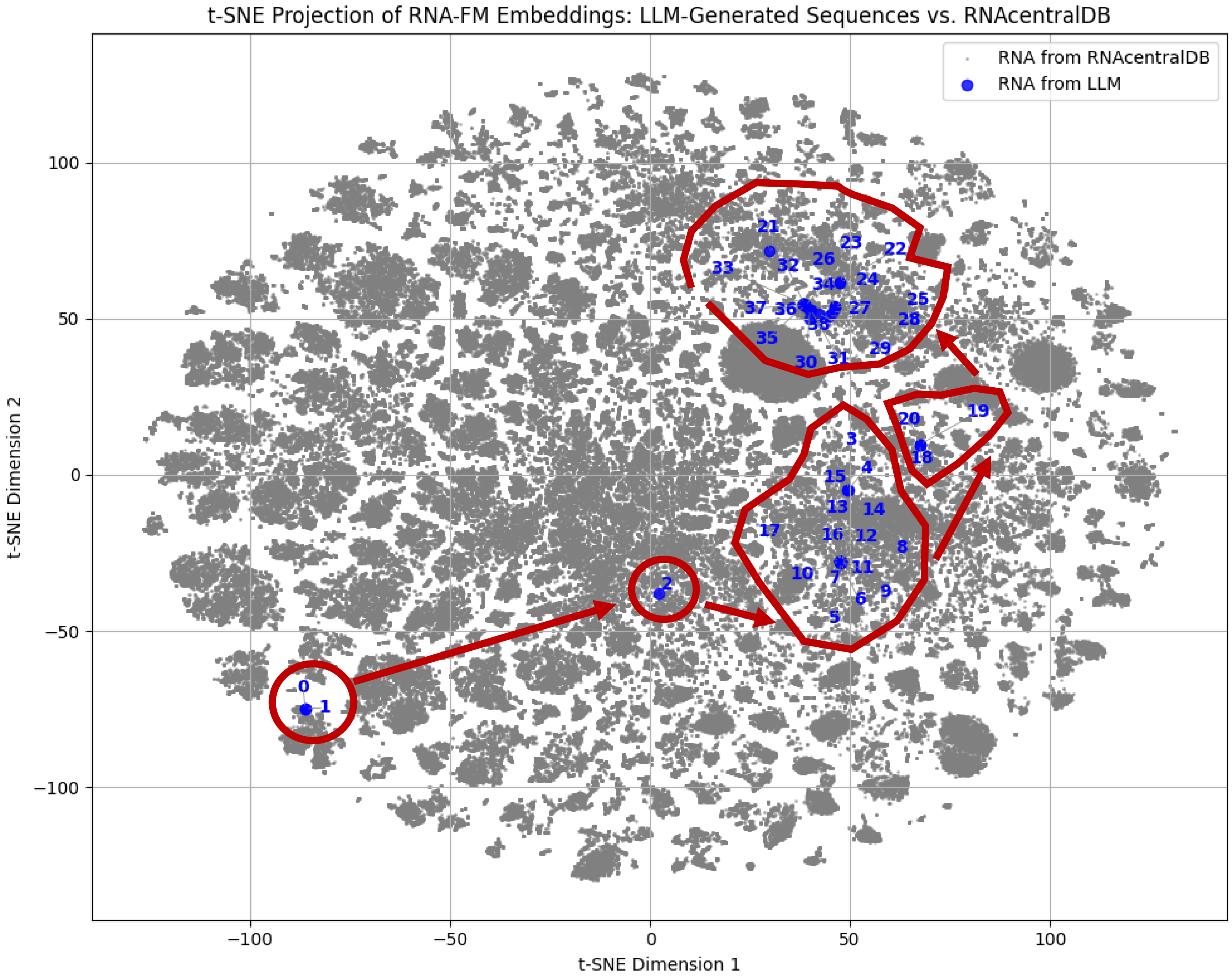
Using Large Language Models to Simulate the Reverse Evolution of tRNA^{Gly}



https://github.com/PavelPll/RNA_RAG/tree/main

From tRNA^{Gly} to short tRNA and miscellaneous RNA

How Accurately do LLM-Designed RNAs Represent Real RNA?



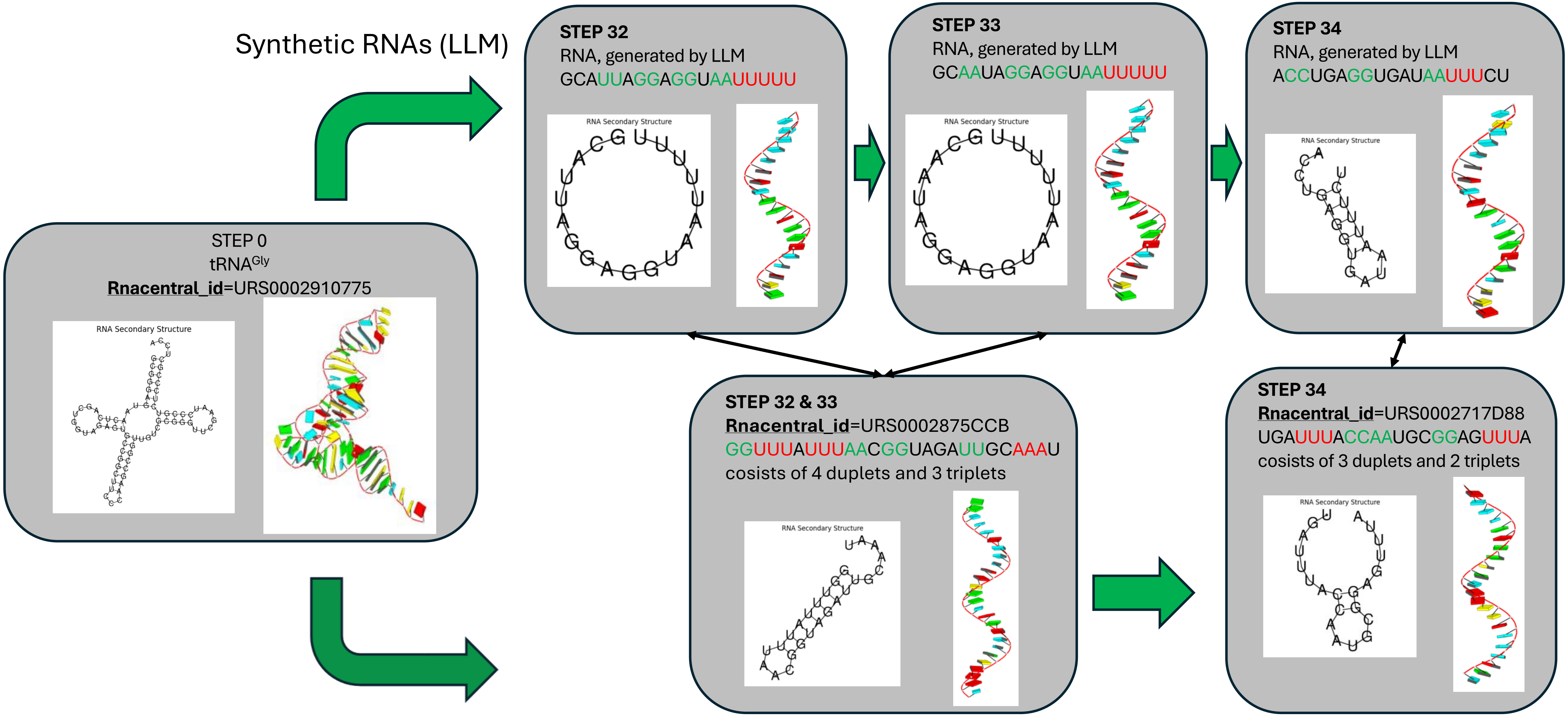
0, 1, … are evoution step (backwards in time)

400,000 tRNA sequences from the RNAcentral database with lengths less than 80 nucleotides

Evoution step (backwards in time)

Step	RNA sequence
0 $tRNA^{Gly}$	GCGGGAGUAACUCAGCUGGUAGAGUGCCGGCUUCCCAAGCCGGUUGUCGCGGGUUCGAAUCCCGUCUCCCGCUCCA
1	GCGGGAGUAACUCAGCUGGUAGAGUGCCGGCUUCCCAAGCCGGUUGUCGCGGGUUCGAAUCCCGUCUCCC
2	GCGGGAGUAAUUCAGCUGGUAGAAUGCUGGCUUUGCAAGCCAGUGGUCGCCGGUUCGAUUCCGGUCAUCC
3	GCGGAGUAAUUCAGCUGGUAGAAUGCUGGCUUUGCAAGCCAGUGGUCGCCGGUUCGAUUCCGGUCAUCC
4	GGGGGAUAGUGCAGCUGGUAGCAUGCUGGCCUUACAAGCCAGUGGUCCUCGGUUCGAUUCCGAGUUCCC
5	GGGGGAUAGUGCAGCUGGUAGCAUGUUGGCCUUACAAGCCAAUGGUCCUGGGUUCAAAUCCCAGUCUCC
6	AGGAGAUCGUGCAGUUGGUAGCAUGUUGGCCUAACAAGCCAAUGGUCCCGGGUUCAAAUCCCAGUCUCC
7	AGGAGAUCGUGCAGUUGGUAGCAUGUUGGCCUAACAAGCCAAUGGUCCCGGGUUCAAAUCCCAGU
8	GGGCGUUGGUGCAGUUGGUAGCAUUCUGGCCUAACACGCCAGGGGUCCCCGGUUCAAAUCCGGGA
	…
28	GGUGUUCCCGGCAGGCUGGUGCUGCCAGGACACC
	…
32	GCAUUAGGAGGUAAUUUUU
33	GCAAUAGGAGGUAAUUUUU
	…
37	ACUUAGCAGAUAAGUUC
38	ACCUCGAGGCUGGGCU

RNA Sequences, 3D Structures, and Closest RNAcentral Matches Over Time



Short homopolymeric stretches — clusters of identical repeated nucleotides — were observed in RNA sequences at early stages of evolution.

Origin and Significance of Homopolymeric Stretches (GGG, AA) in the Proto-RNA World

CGUAUGCAGUC

<u>High Temperature (Hot)</u>

During the early stage of evolution, when nucleotides are homogeneously mixed, the RNA chain exhibits maximal entropy

Thus: $\Delta S > 0$ results in $\boldsymbol{\Delta G = -T\Delta S < 0}$

AAAAAGGGGG

<u>Low Temperature (Cold)</u>

Homopolymeric stretches (k-mers) reduce sequence entropy, $\Delta S < 0$

Stacking interactions among identical nucleotides (π–π overlap) lower the enthalpy, $\Delta H < 0$,

Thus: $\Delta S < 0$ & $\Delta H < 0$ can results in $\boldsymbol{\Delta G = \Delta H - T\Delta S < 0}$

This is biologically meaningful — in early evolution, such sequences may have emerged due to simpler synthesis, replication, or selection constraints.

The negative ΔH (enthalpy gain) can offset the loss in entropy (negative ΔS), potentially making the formation of homopolymeric stretches thermodynamically favorable ($\Delta G < 0$) despite reduced sequence complexity

CONCLUSIONS (part 1)

- **RAG (LLM) can be used not only to query existing RNA sequences but also to generate new ones, potentially providing an advantage when guiding RNA generation with conditional diffusion models.**

- **An attempt to model RNA evolution using RAG reveals several distinct patterns:**
  - No hairpins are detected beyond a certain reverse evolution timestep
  - The LLM fails to generate RNA sequences beyond approximately 15 nucleotides in length
  - After ligation, the LLM retains part of the 5' end
  - Interestingly, significant ligation events occurred without any accompanying mutations.
  - Homopolymeric stretches (*k*-mers, e.g., GGG, AAAA) tend to appear during the early stages of evolution

- **The formation of homopolymeric stretches (*k*-mers) in the proto-RNA world is also thermodynamically favorable ($\Delta G < 0$), particularly at low temperatures. Next three slides show an analysis of homopolymeric stretches (*k*-mers) in the peptidyl transferase center of the ribosome.**

Number of Homopolymeric Stretches of Length *k* (*k*-mers) in a random RNA Sequence of Length *L*

k-mers = A,…,A or G,…,G or C,…,C or U,…,U
repeated *k* times

$$A(k) = \sum_{base} p_{base}^{k} = 4\frac{1}{4^{k}}$$

Probability of *k* consecutive identical nucleotides (we don't yet care about neighbors)

$$B(k) = \sum_{base} p_{base}^{k}\,(1 - p_{base})$$

Probability of *k* consecutive identical nucleotides with a different nucleotide adjacent on one side

$$C(k) = \sum_{base} p_{base}^{k}(1 - p_{base})^{2}$$

Probability of observing k consecutive identical nucleotides flanked on both sides by different nucleotides

$$N(k, L) = (L - k + 1)$$

Number of sliding windows of length k

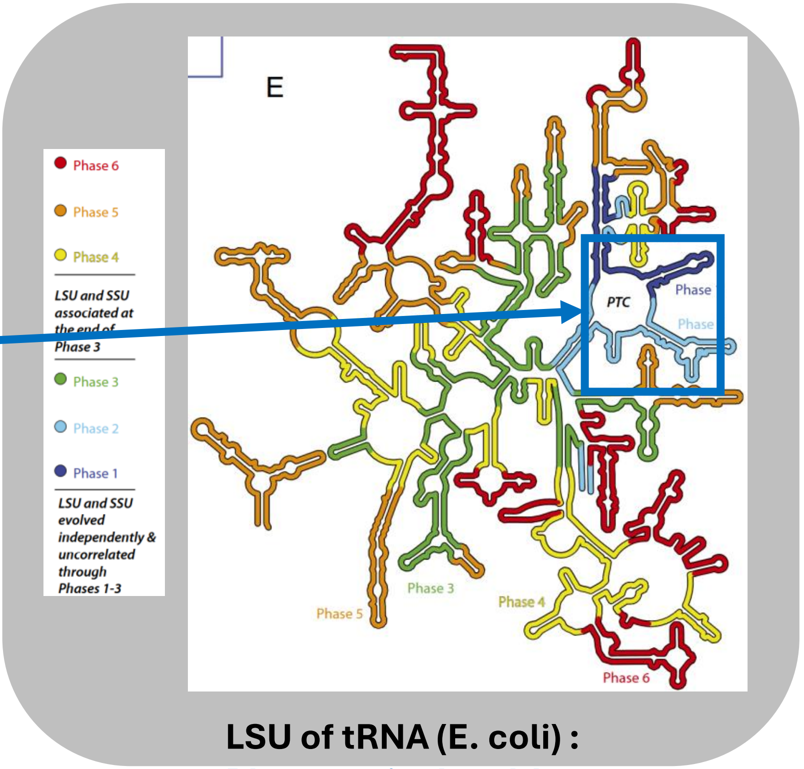
$$\boxed{N_{k-mers}(k, L) = B(k) + (N(k, L) - 2)C(k) + B(k)}$$

<u>A random RNA sequence of length 89 nucleotides contains, on average:</u>

- **0.76 of 4-mers (AAAA or GGGG or CCCC or UUUU)**
- **3.1 of 3-mers (AAA or GGG or CCC or UUU)**
- **12.5 of 2-mers (AA or GG or CC or UU)**
- **50 of 1-mers (A or G or C or U)**

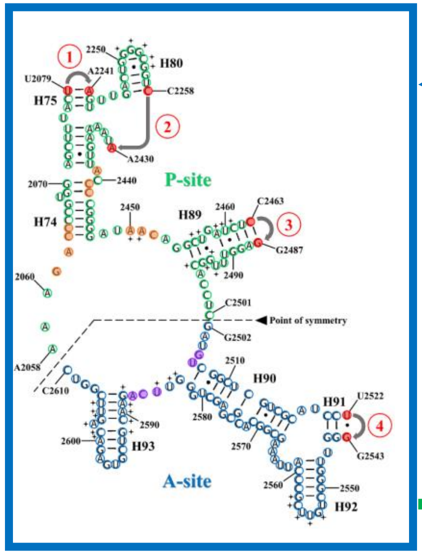
What is the oldest know RNA ?

rRNA evolution: Age Variability Across Different Regions



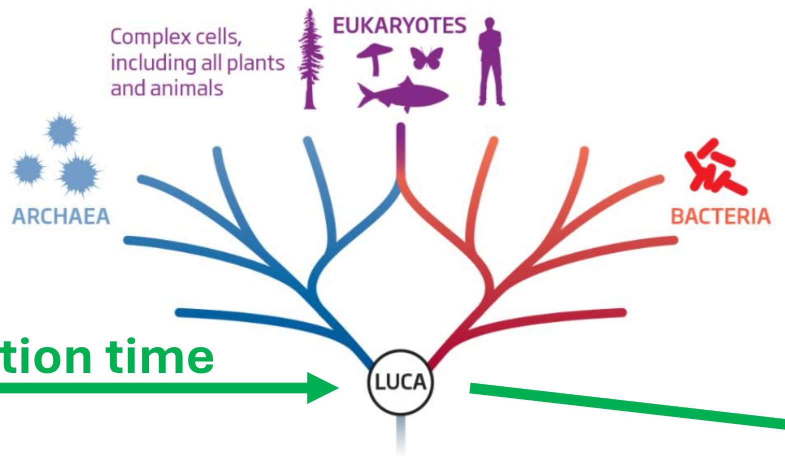
LSU of tRNA (E. coli) :
Blue part is the oldest one

the OLDEST functional part: always the same!



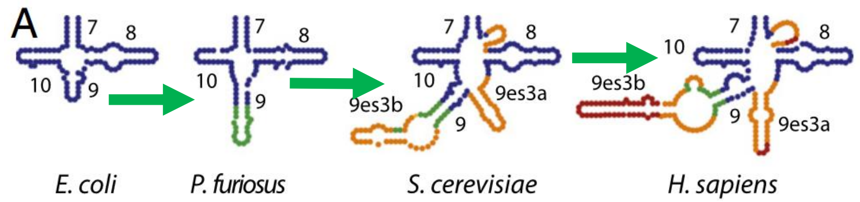
Secondary structure of the pseudosymmetrical region (**SymR**; *Agmon et al., 2005*), derived from the LSU secondary structure of Thermus thermophilus (*Petrov et al., 2013*). (*Madhan R. Tirumalai et al., 2021*)

evolution time



Last Universal Common Ancestor (LUCA)

Image from:
https://www.pulseheadlines.com/
earths-universal-common-ancestor-volcanic-origins/43890/

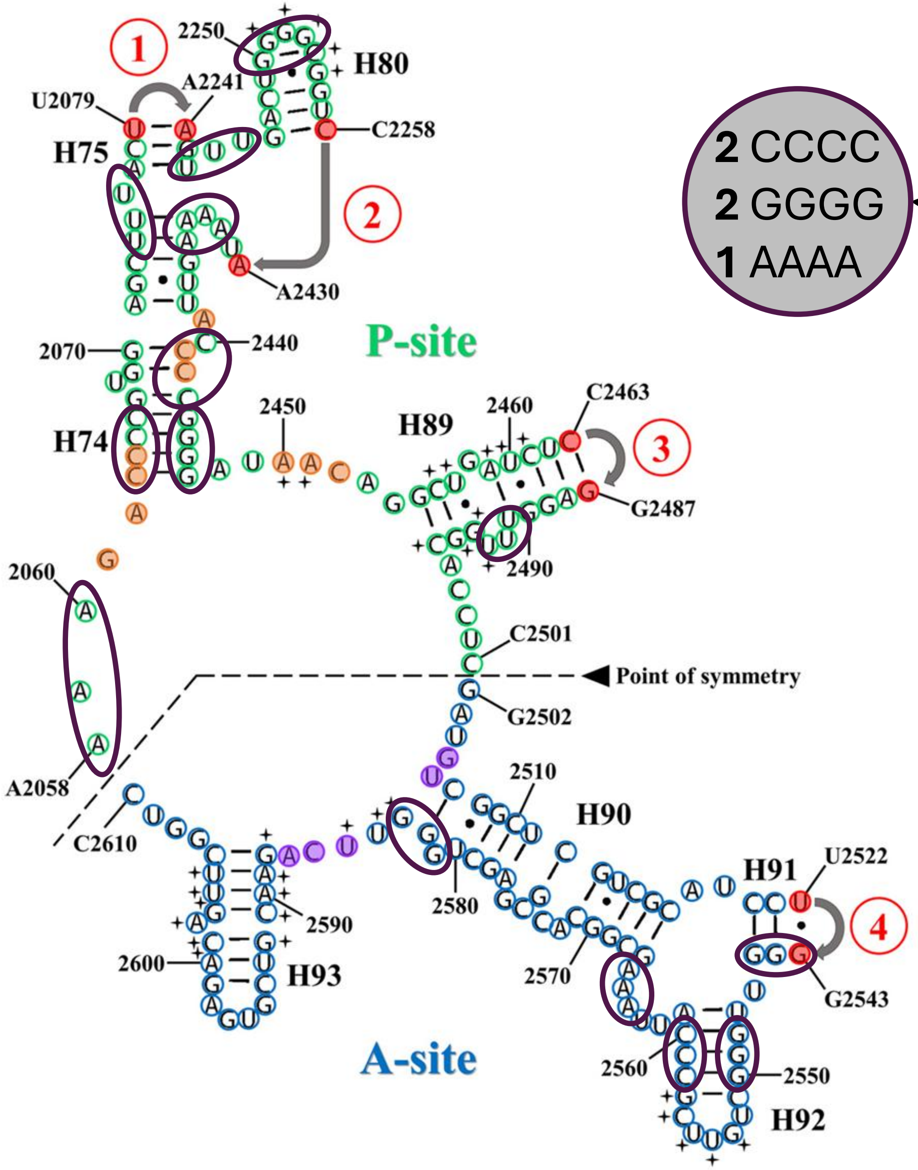


Molecular level chronology of the evolution of the large ribosomal subunit (LSU) rRNA. Each accretion step adds to previous rRNA but leaves the underlying **core unperturbed** (Anton S. Petrov et al., PNAS, 2015)

Peptidyl Transferase Center (PTC) is the oldest part of ribosomes and contains no proteins

This symmetry (**SymR**) suggests that the ancient ribosome may have been **a dimer of identical or nearly identical RNA molecules**, later evolving into the asymmetrical modern ribosome with **PTC**.

The null hypothesis H_0 is: the RNA sequence is random (i.e. uniform bases)



2 CCCC
2 GGGG
1 AAAA

P-site sequence deviates significantly from randomness

k-mers	N of k-mers (random sequence)	N of k-mers (P-site)	p-value	H_0 hypothesis
4-mers	0.76	5	0.0000005	rejected
3-mers	3.1	4	0.564	fail to reject
2-mers	12.5	8	0.096	rejected
1-mers	50	41	0.044	rejected

whereas the A-site sequence does not

k-mers	N of k-mers (random sequence)	N of k-mers (P-site)	p-value	H_0 hypothesis
4-mers	0.76	0	0.366	fail to reject
3-mers	3.1	5	0.228	fail to reject
2-mers	12.5	10	0.357	fail to reject
1-mers	50	54	0.447	fail to reject

Ancient evolutionary core of the ribosome
SymR includes the Peptidyl Transferase Center
See previous slide for details

The presence of homopolymeric stretches (k-mers) in P-site supports the well-known view that the P-site is evolutionarily older than the A-site — that is, early ribosomes may have functioned with just one substrate, similar to the modern P-site.

CONCLUSIONS (part 2)

RAG modeling suggests that homopolymeric k-mers (e.g., GGG, AAAA) may have emerged during early RNA evolution. This finding is thermodynamically supported, as such sequences are favorable ($\Delta G < 0$), especially at low temperatures.

Altogether, this supports the idea that the P-site of the Peptidyl Transferase Center is evolutionarily older than the A-site, with early ribosomes possibly operating using a single substrate at the P-site.

What's next ?

1. Further improve the context for RAG using the rnaGPT model

Only RNA Sequence (e.g., "AUGGCUUAGCU...")
The Question (e.g., "What is the function of this RNA?")

→ RNA-GPT → answer

at this time, **RNA-GPT** is not freely available for use

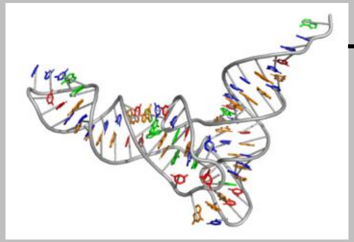
Yijia Xiao et al. , RNA-GPT: Multimodal Generative System for RNA Sequence Understanding, **2025**

2. Further improve the context for RAG using the HealthGPT model

HealthGPT is a medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation
HealthGPT **understands medical images** (e.g. answer questions about them)

Q: What type of RNA molecule is shown in the image?

+



→ HealthGPT →

ANSWER: The image shows a ribosomal RNA (rRNA) molecule. Ribosomal RNA is a type of RNA that is a key component of the ribosome, the cellular machinery responsible for protein synthesis. The image depicts the secondary structure of the rRNA molecule, which is formed by the folding and pairing of the RNA strands.

This is tRNA not rRNA...
HealthGPT-M3 is available
HealthGPT-L14 weights are still not released yet

Tianwei Lin et al., HealthGPT: A Medical Large Vision-Language Model for Unifying Comprehension and Generation via Heterogeneous Knowledge Adaptation, **2025**

3. Construct an agent (on top of my RAG stack) in order to intelligently decide or coordinate:

- **Agent decides:** "Do I need structural, visual, or/and textual info for this question?“
- **Agent breaks query into subtasks**, calling tools/models iteratively