

RNA Sequence Analysis Using Transformer Models

Work in progress!

Pavel Polyakov

OUTLINE

❖ Objectives

❖ The idea of transformers

❖ Dataset

- What is RNA molecule
- RNACentral database

❖ RNA autoencoder

- Vanilla transformer*
- BERT-like transformer*
- BERT-like transformer (RNA-FM model)

❖ Results

- Word embeddings arithmetic
- Peptidyl Transferase Center (PTC): structure and evolution

❖ Conclusions

❖ What's Next?

[*https://github.com/PavelPll/RNA_transformer](https://github.com/PavelPll/RNA_transformer)

OBJECTIVES

1. Generate hidden representations of RNA molecules using modern Natural Language Processing (NLP) approaches
2. Try to use these representations to describe some biological processes

The idea of transformers

Embedding Layer

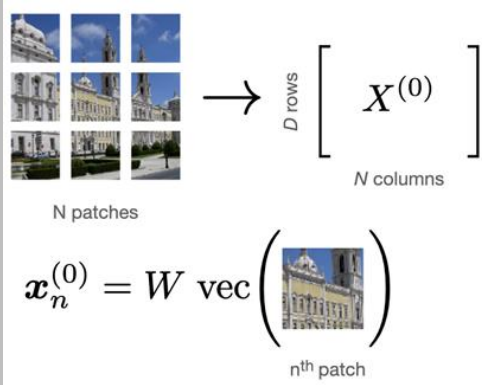
N tokens

$$X^{(0)} = \begin{pmatrix} \mathbf{x}_0^{(0)} & \dots & \mathbf{x}_n^{(0)} & \dots & \mathbf{x}_N^{(0)} \end{pmatrix} \quad \begin{matrix} \updownarrow \\ D \end{matrix}$$



Encoding an image

$\mathbf{x}_n^{(0)}$ is a token

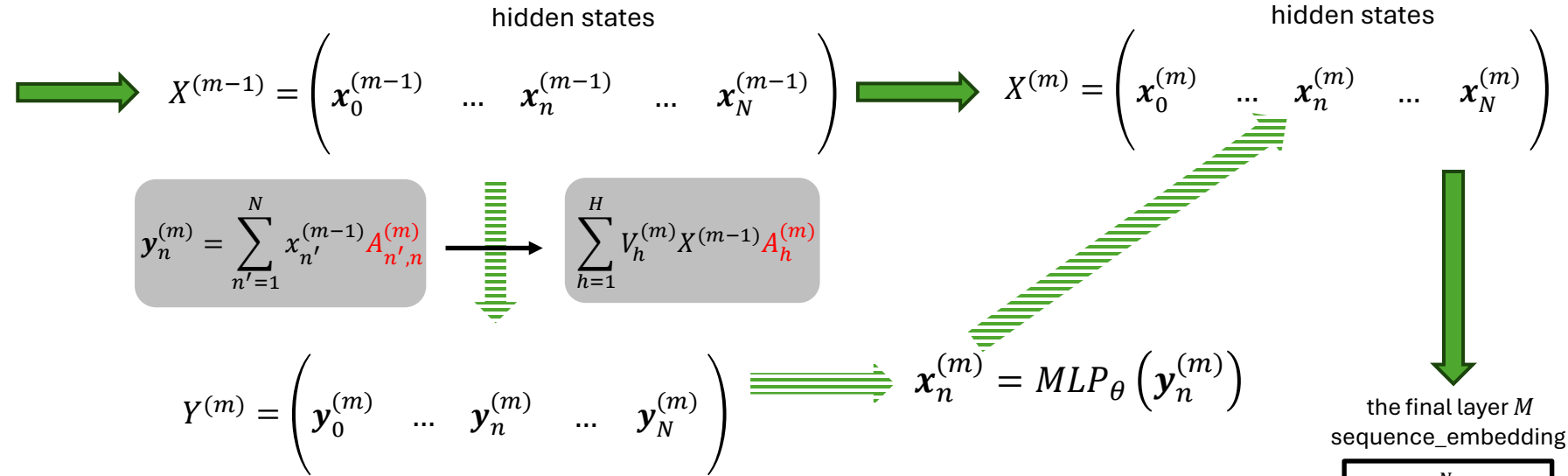


vec operator: Each patch is reshaped into a vector by the vec operator.

matrix W : maps a vector (the patch) to a D dimensional vector $\mathbf{x}_n^{(0)}$.

[Dosovitskiy et al., 2021]

Iteratively applying a transformer block



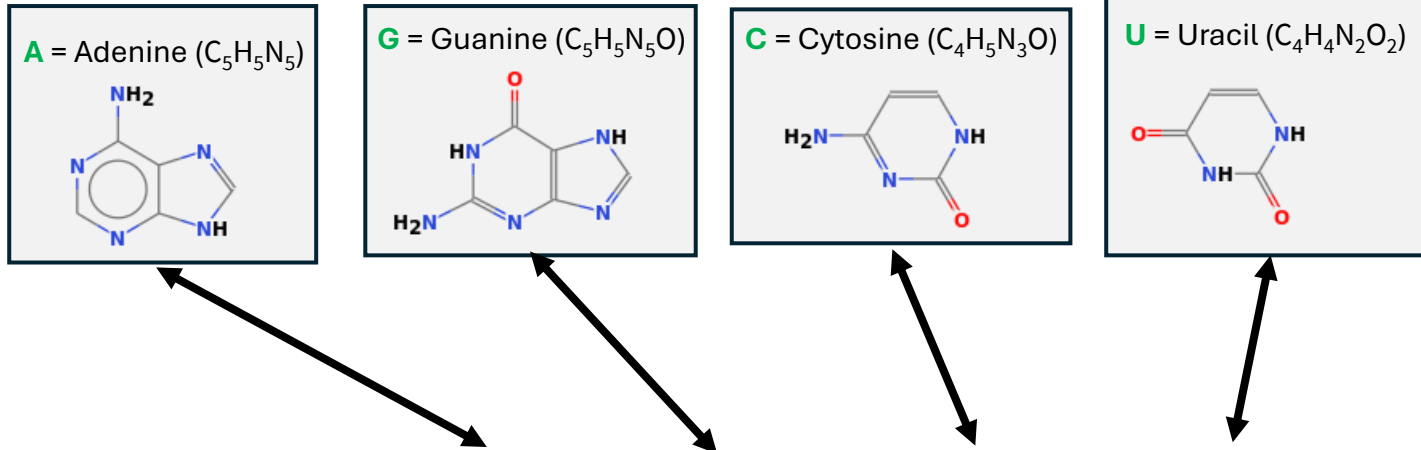
The attention matrix

$$A_{n,n'} = \frac{\mathbf{x}_n^T \mathbf{x}_{n'}}{\sum_{n''=1}^N \mathbf{x}_n^T \mathbf{x}_{n''}} \rightarrow \frac{\exp(\mathbf{x}_n^T \mathbf{x}_{n'})}{\sum_{n''=1}^N \exp(\mathbf{x}_n^T \mathbf{x}_{n''})} \rightarrow \frac{\exp(\mathbf{x}_n^T U^T U \mathbf{x}_{n'})}{\sum_{n''=1}^N \exp(\mathbf{x}_n^T U^T U \mathbf{x}_{n''})} \rightarrow \frac{\exp(\mathbf{x}_n^T U_k^T U_q \mathbf{x}_{n'})}{\sum_{n''=1}^N \exp(\mathbf{x}_n^T U_k^T U_q \mathbf{x}_{n''})}$$

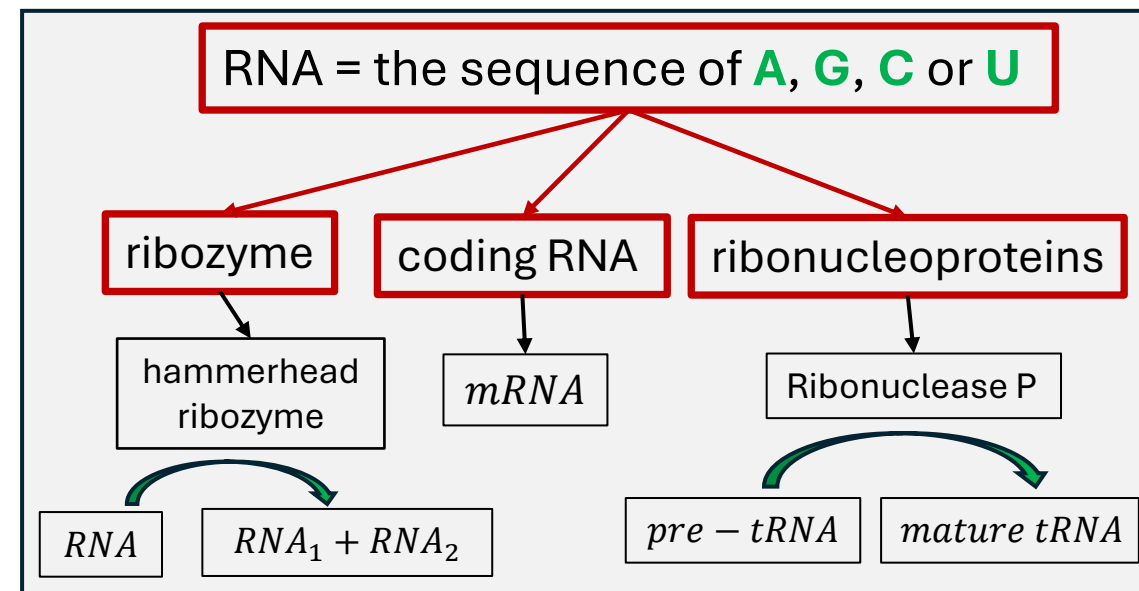
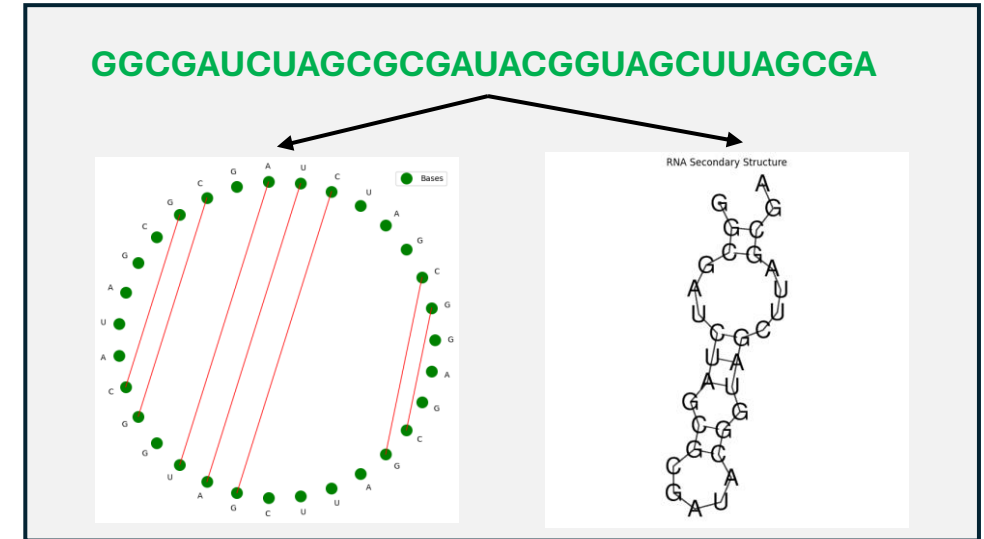
$$A_{n,n'} = \frac{\exp(k_{n,n'}^T q_{n,n'})}{\sum_{n''=1}^N \exp(k_{n,n''}^T q_{n,n''})}, \quad \text{where} \quad \begin{cases} q_{h,n}^{(m)} = U_q^{(m)} \mathbf{x}_n^{(m-1)}, & \text{queries} \\ k_{h,n}^{(m)} = U_k^{(m)} \mathbf{x}_n^{(m-1)}, & \text{keys} \end{cases}$$

[Turner R.E., <https://arxiv.org/abs/2304.10557>, 2023]

What is RNA molecule ?



	Adenine	Guanine	Cytosine	Uracil
O	0	1	1	2
N	5	5	3	2
C	5	5	4	4
H	5	5	5	4
$\Delta_f H_{solid}^0, kJ/mol$	96.9	-183.9	-221	-424.4
$\Delta_c H_{solid}^0, kJ/mol$	-2779.0	-2498.2	-2067	-1721.3
$M_w, g/mol$	135	151	111	112
Hydrogen bonds	2	3	3	2

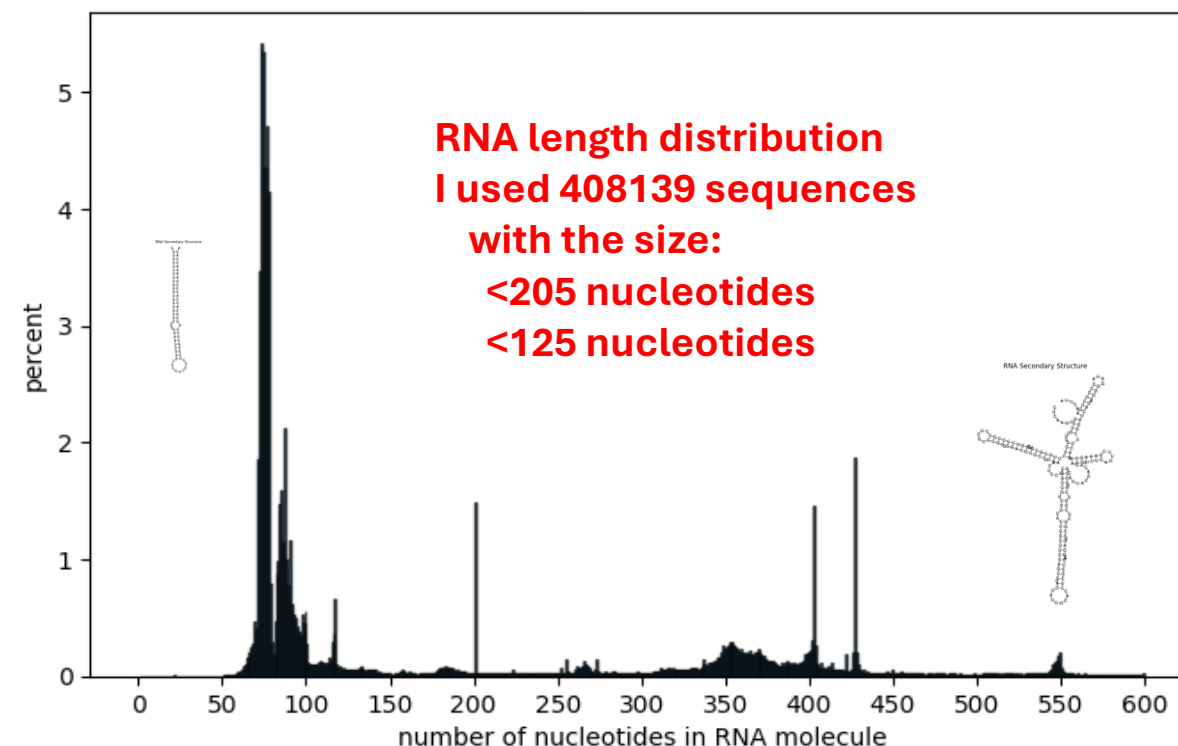
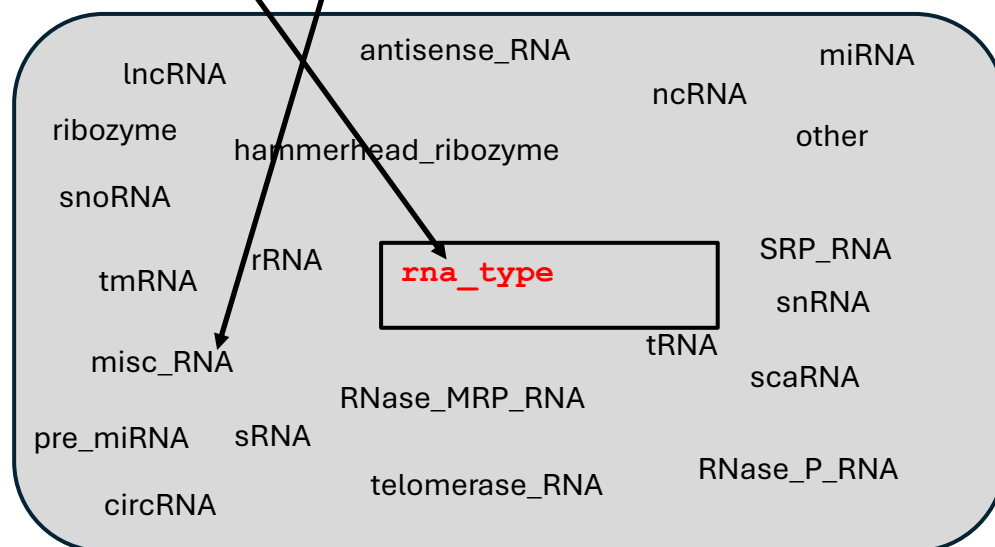
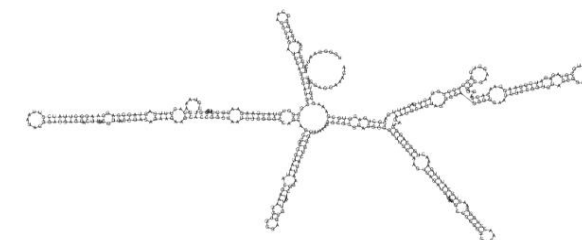


The dataset: RNACentral database

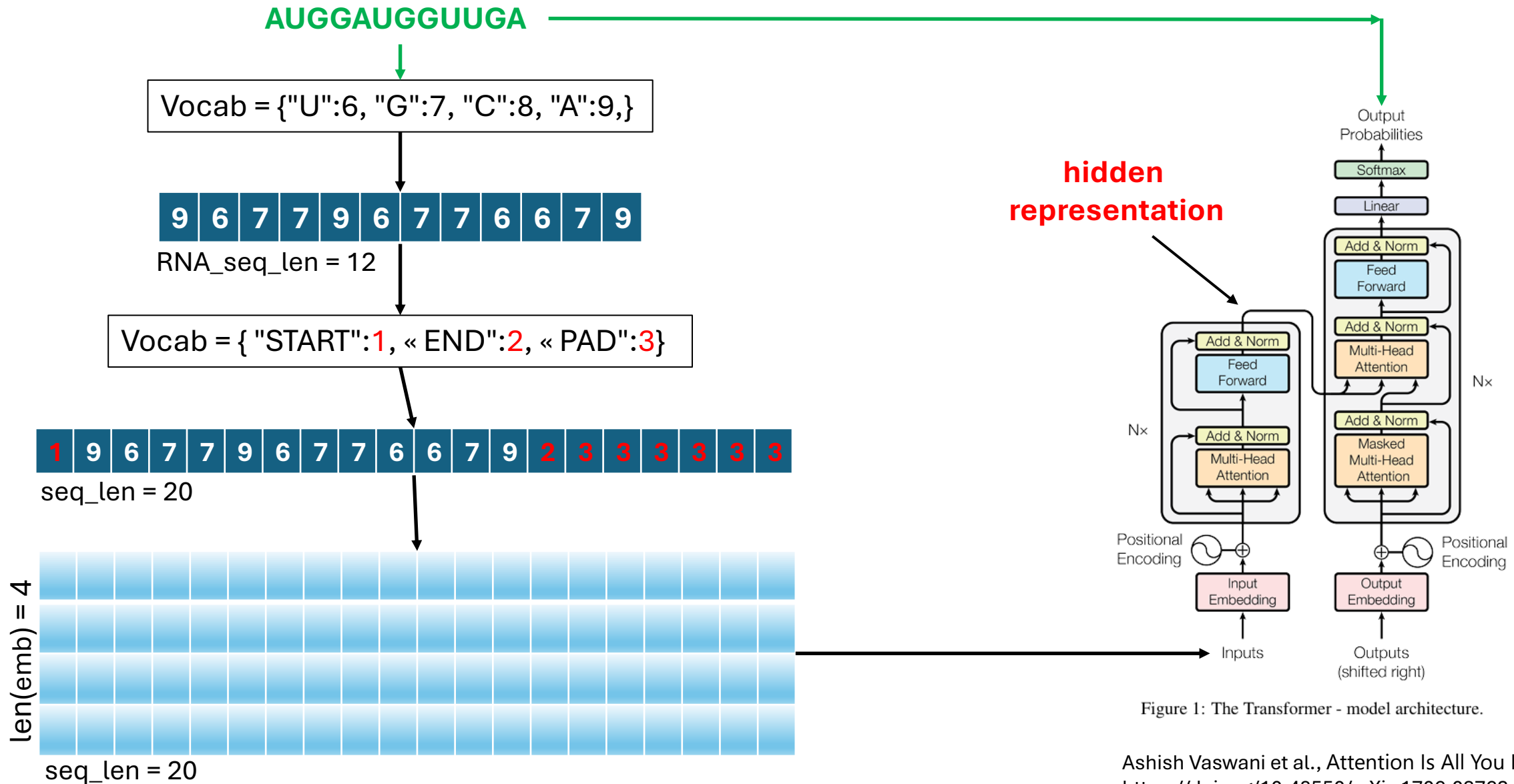
RNACentral Browsable API

https://rnacentral.org/api/v1/rna/?page=3&page_size=100 gives:

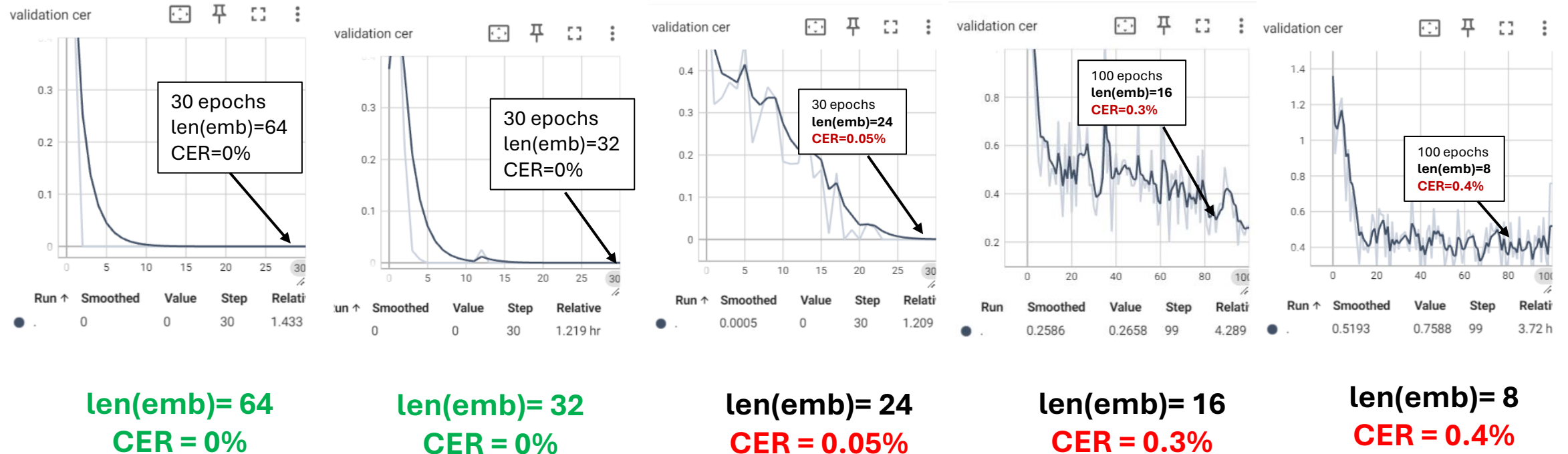
```
{ "url": "http://rnacentral.org/api/v1/rna/URS0002915621",
  "rnacentral_id": "URS0002915621",
  "md5": "fee3fe68dbd91ee898bffd9d4b89b2e9",
  "sequence": "AUGGAUGGUUGAUCAGAGAACGUACAUUUUAUAAAUGGUGUAUGUCAAUUGAUCCACAGUCCCU",
  "length": 64,
  "xrefs": "http://rnacentral.org/api/v1/rna/URS0002915621/xrefs",
  "publications": "http://rnacentral.org/api/v1/rna/URS0002915621/publications",
  "is_active": true,
  "description": "pre_miRNA from 0 species",
  "rna_type": "pre_miRNA",
  "count_distinct_organisms": 4,
  "distinct_databases": [ "Rfam" ] }, ...
```



RNA autoencoder (vanilla transformer)



RESULTS for RNA autoencoder validation



CER (Character Error Rate) calculates the proportion of incorrect nucleotides (insertions, deletions, and substitutions) relative to the total number of nucleotides.

CER changes from 0% to 0.4% with decreasing the length of embeddings.

RNA classification (BERT-like transformer)

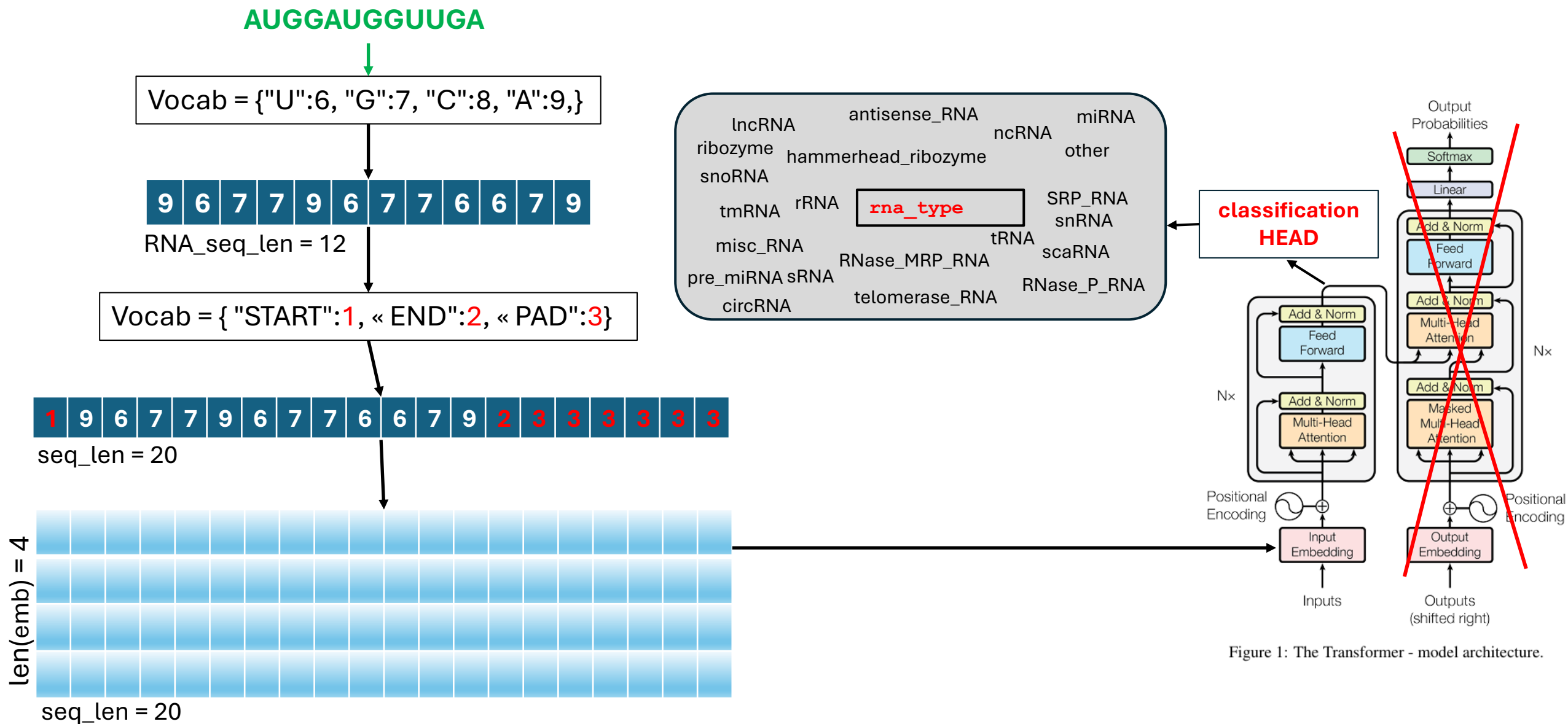


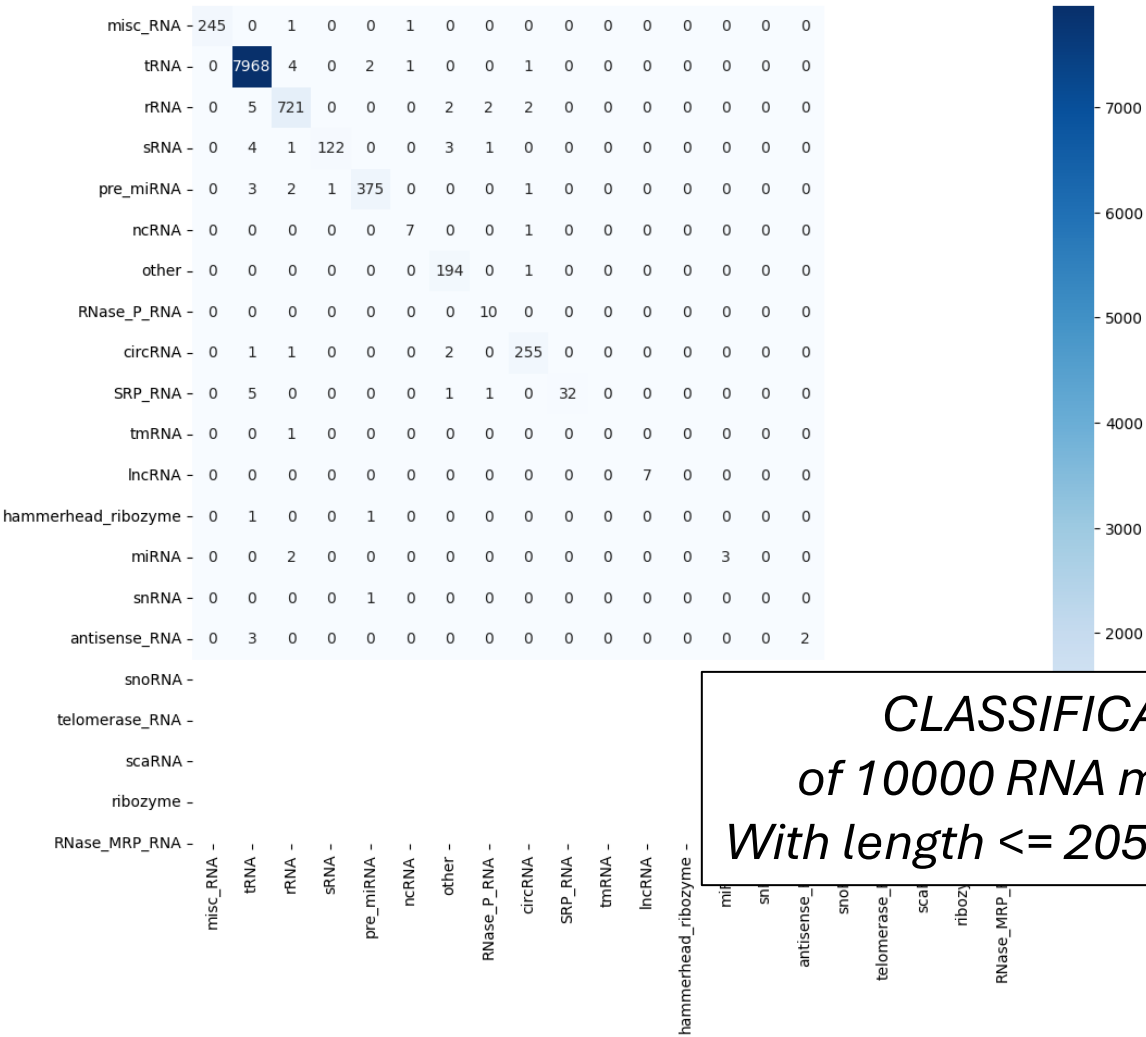
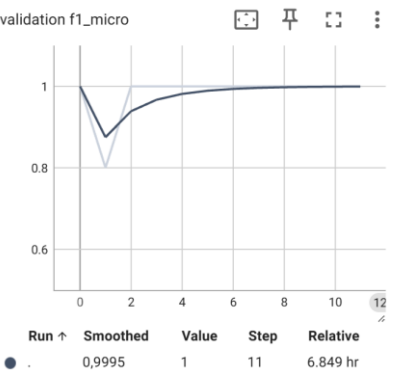
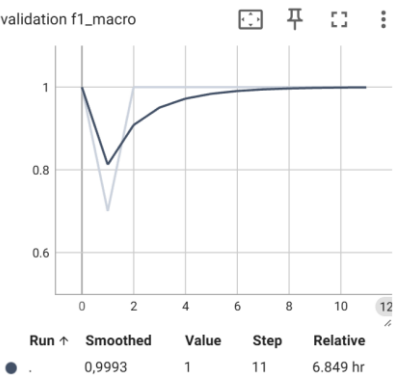
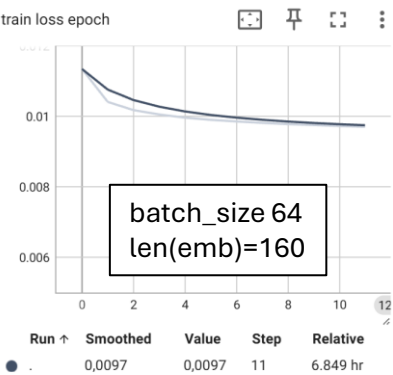
Figure 1: The Transformer - model architecture.

RESULTS for RNAClassificator validation

VALIDATION

f1_micro 0.9941

f1_macro 0.7351456585589439



CLASSIFICATION
of 10000 RNA molecules
With length <= 205 nucleotides

RNA autoencoder (BERT-like transformer)

AUGGAUGGUUGA

Vocab = {"U":6, "G":7, "C":8, "A":9,}

9 6 7 7 9 6 7 7 6 6 7 9

RNA_seq_len = 12

Vocab = { "START":1, « END":2, « PAD":3}

1 9 6 7 7 9 6 7 7 6 6 7 9 2 3 3 3 3 3 3

seq_len = 20



seq_len = 20

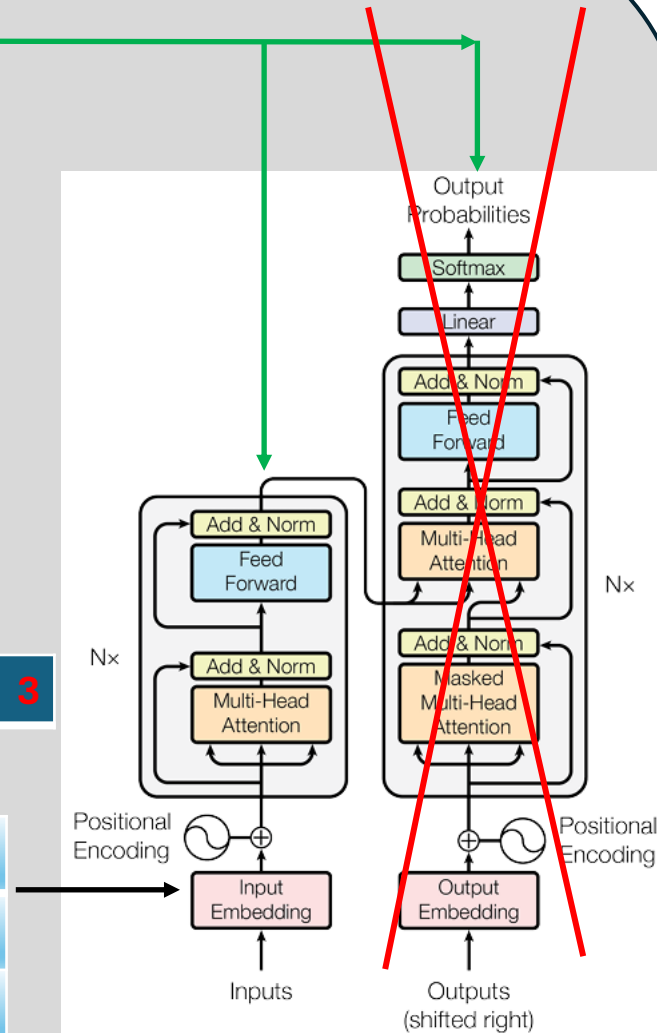
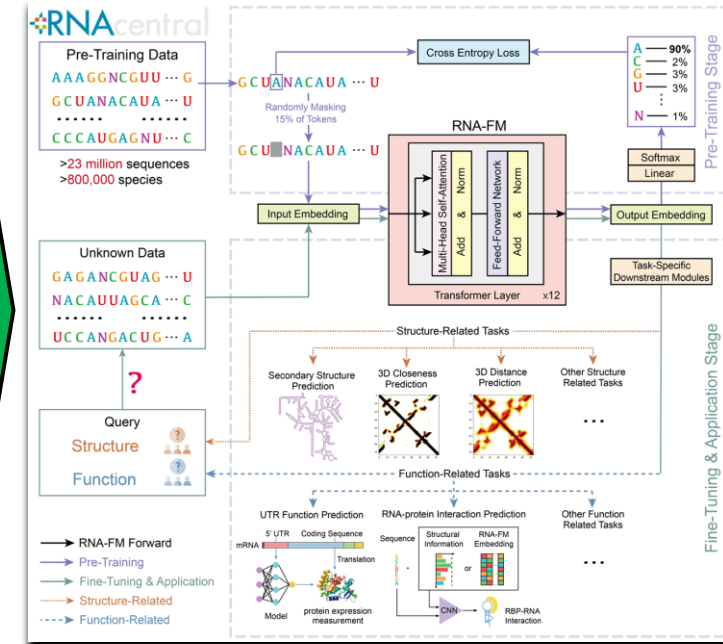


Figure 1: The Transformer - model architecture.

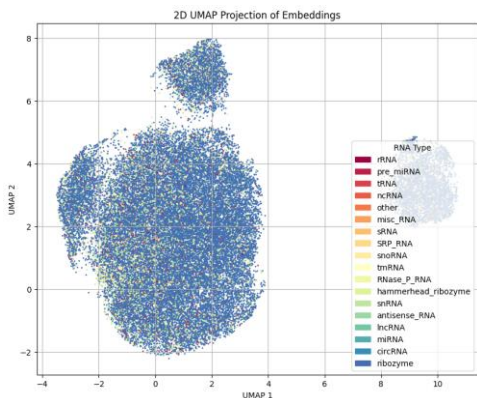
**STATE of the ART
RNAFM model**



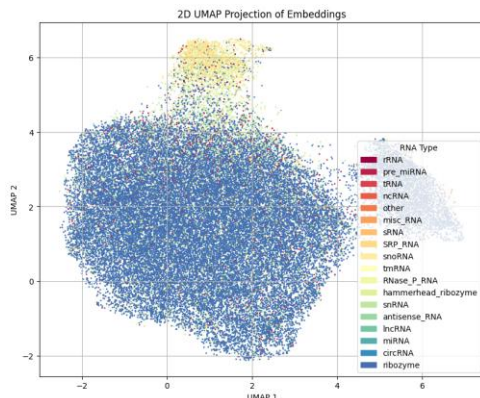
J.Chen et al.,
<https://www.biorxiv.org/content/10.1101/2022.08.06.503062v2>

RESULTS for RNA autoencoder validation (for 40 000 RNAs)

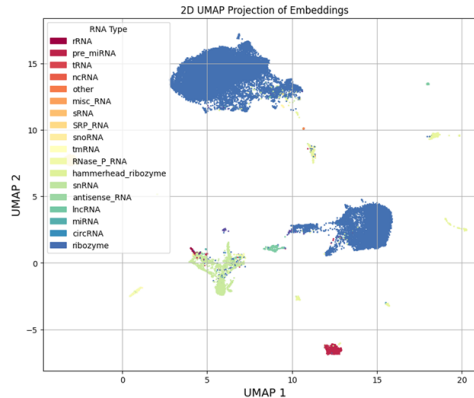
reconstruction loss (1)
random embedding
batch_size = 64,
len(emb) = 64



reconstruction loss (1)
custom embedding
batch_size = 64,
len(emb) = 64

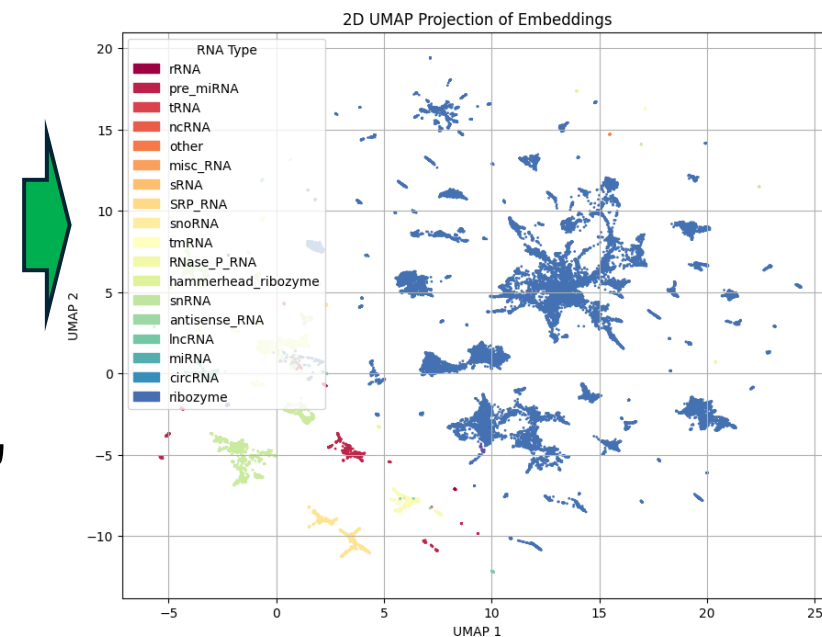


reconstruction loss (0.9)
classification loss (0.1)
custom embedding
batch_size = 64
len(emb) = 160



- secondary/3D structure prediction
- SARS-CoV-2 genome structure and evolution prediction
- protein-RNA binding preference modeling
- gene expression regulation modeling

STATE of the ART
RNAFM model



4 vanilla transformer-based encoder blocks
400 000 sequences
 $L \times 64$ embedding matrix for each RNA (length L)
RTX 4060, 8GB for 12 hours

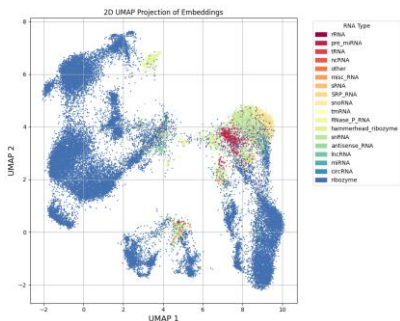
12 transformer-based bidirectional encoder blocks
23 million sequences
 $L \times 640$ embedding matrix for each RNA (length L)
eight A100 GPUs of 80 GB memories for one month

IMPROVEMENT direction

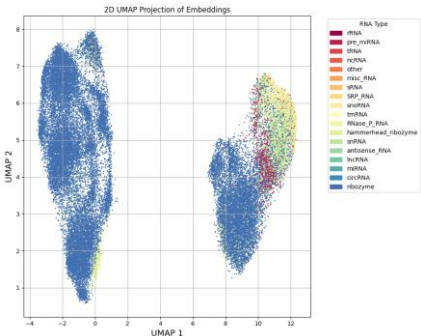
Further Improvement (for 40 000 RNAs)

vanilla transformer-based

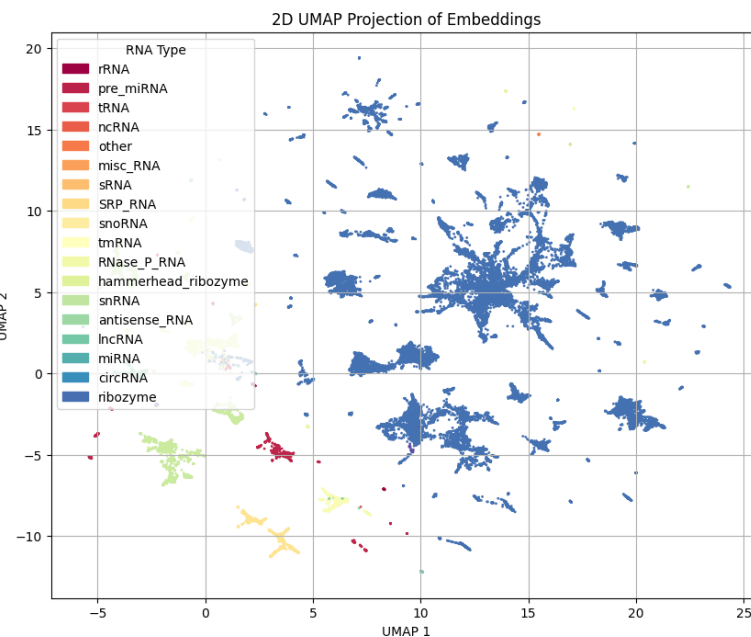
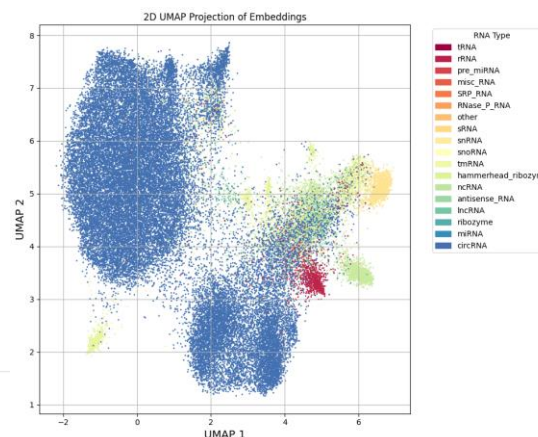
reconstruction loss (1)
random embedding
len(emb) = 160



reconstruction loss (1)
custom embedding
len(emb) = 160



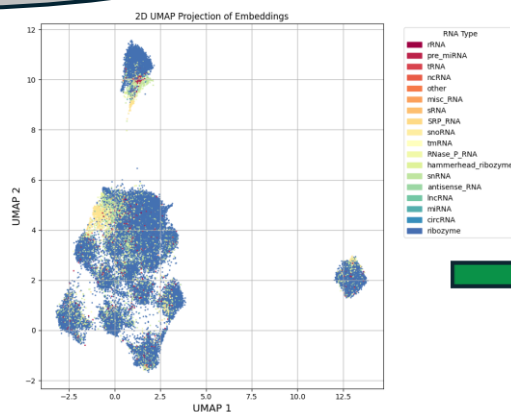
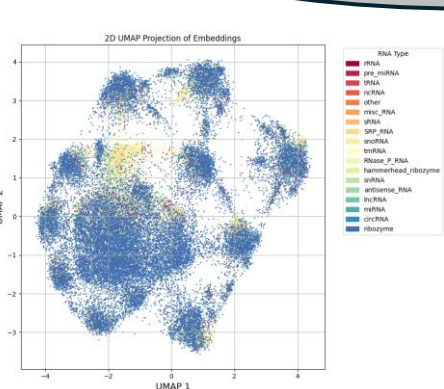
Composite autoencoder



**STATE of the ART
RNAFM model**

- *secondary/3D structure prediction*
- *SARS-CoV-2 genome structure and evolution prediction*
- *protein-RNA binding preference modeling*
- *gene expression regulation modeling*

BERT-like transformer



IMPROVEMENT direction

**23 million sequences
12 blocks, 16 heads
 $L \times 640$ embedding matrix for each RNA (length L)
eight A100 GPUs of 80 GB memories for one month**

**0.4 million sequences
4 blocks, 8 heads
 $L \times 64$, 160 embedding matrix for each RNA
length $L_{max} = 205$ nucleotides
RTX 4060, 8GB for 12 hours**

Word embeddings arithmetic

ENGLISH language

Laptop \approx Computer + Portable
Smartphone \approx Phone + Smart

Fast + More \approx Faster
Happy + Not \approx Sad

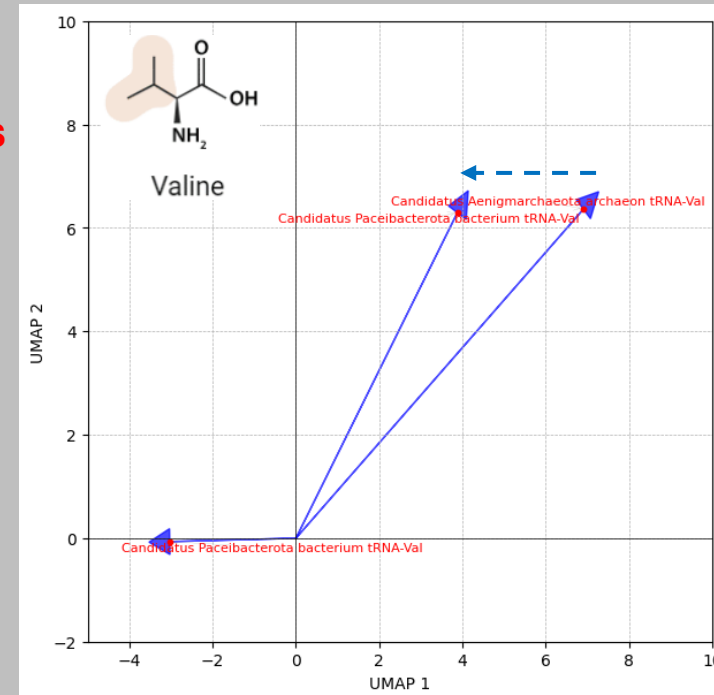
King – Man + Woman \approx Queen
Paris – France + Italy \approx Rome

Amino acids from:
<https://www.rapidnovor.com/structure-of-an-amino-acid/>

**archaea and
bacteria often
coexist in various
environments**

Candidatus
Paceibacterota
bacterium
tRNA-Val

RNA language



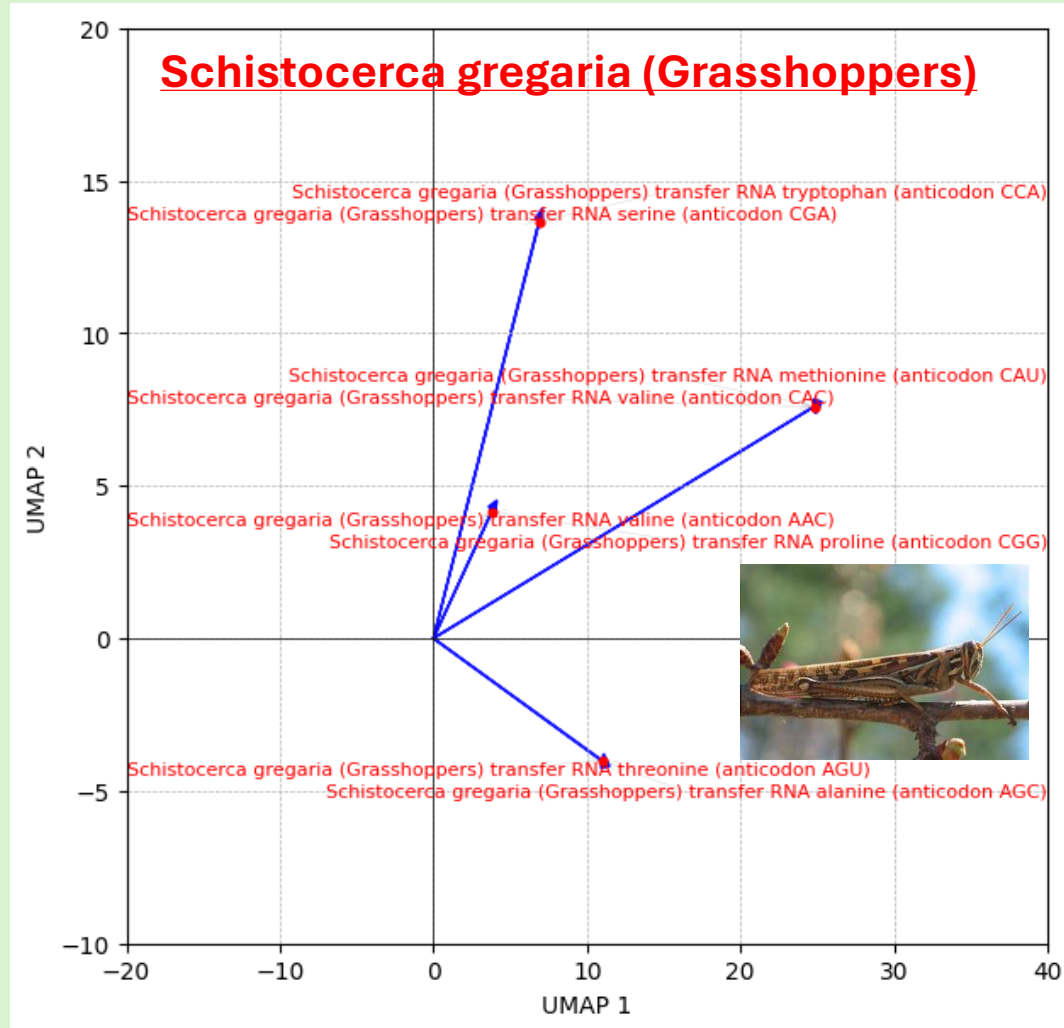
Candidatus
Paceibacterot
a bacterium
tRNA-Val

$$\text{tRNA}(73\text{nt}) = \text{tRNA}(73\text{nt}) + \text{tRNA}(74\text{nt})$$

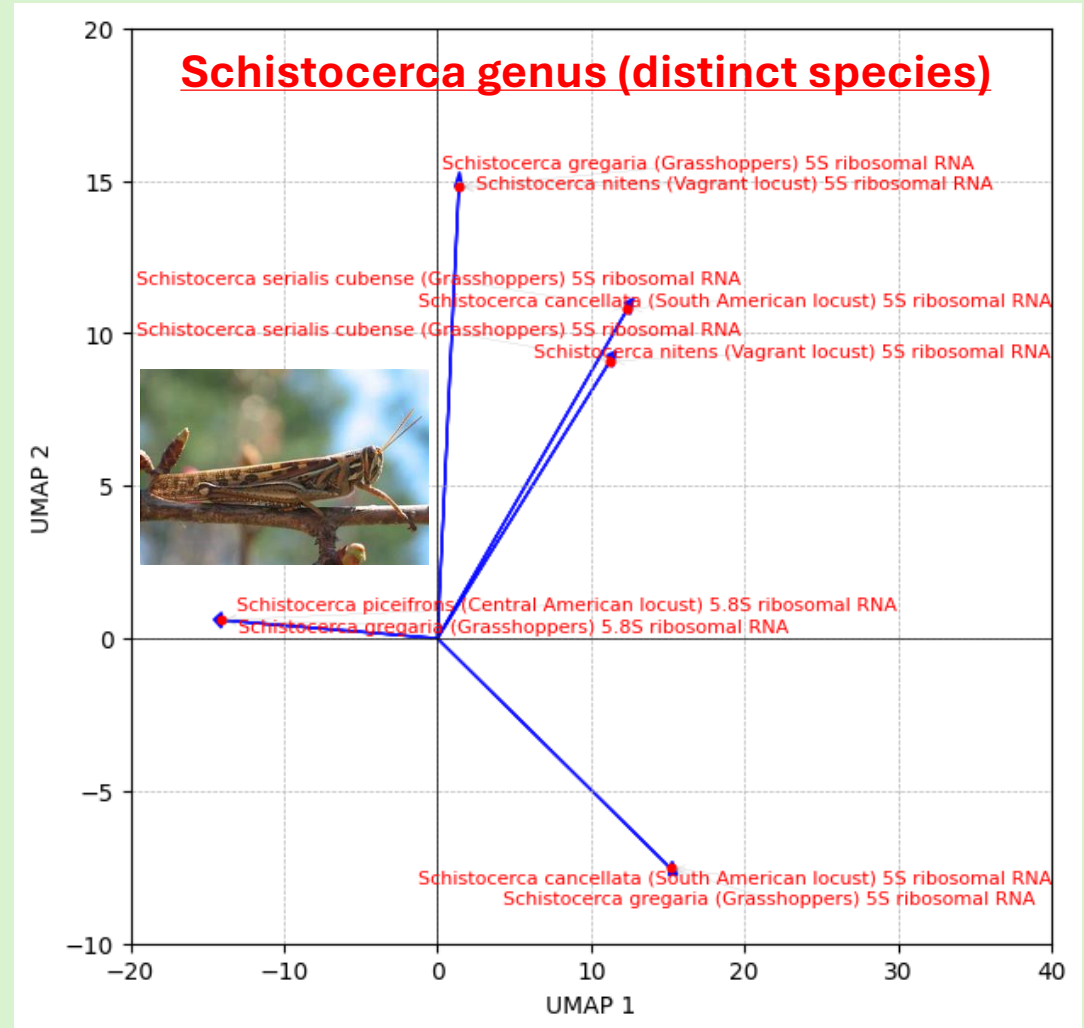
Candidatus Aenigmarchaeota
archaeon tRNA-Val

Word arithmetic: tRNA and rRNA of Schistocerca

Transfer RNA (tRNA)



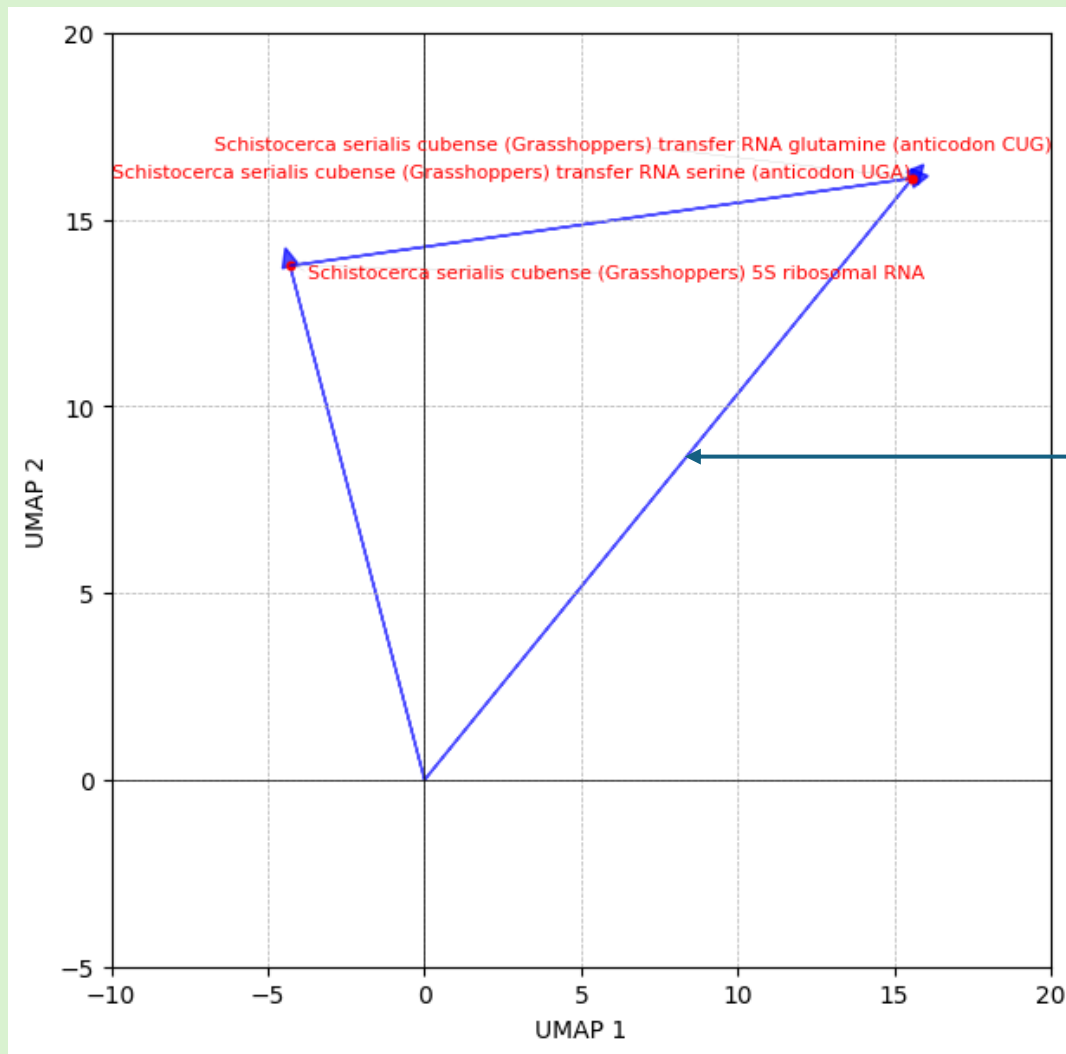
5S Ribosomal RNA (rRNA)



Each arrow consists of two sub-arrows to identify two similar RNAs that share the same hidden representation.

The image of Schistocerca from here: https://th.bing.com/th/id/OIP.Nzhi77J6_VosvgaRUmvWzAAAAA?rs=1&pid=ImgDetMain

Schistocerca serialis cubense (word embedding arithmetic)



Schistocerca serialis cubense
5S ribosomal RNA



$$tRNA(72nt) = rRNA(121nt) + tRNA(75nt)$$

transfer RNA glutamine
(anticodon CUG)

transfer RNA serine
(anticodon UGA)

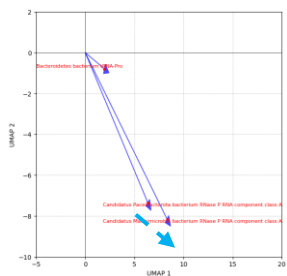
Schistocerca image from: https://th.bing.com/th/id/OIP.Nzhi77J6_VosvgaRUmvWzAAAAA?rs=1&pid=ImgDetMain

Dependence of tRNA on other ncRNAs

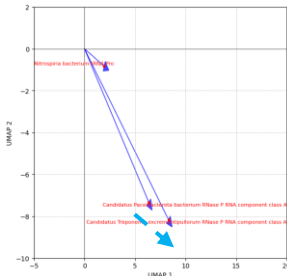
ncRNA Type	Regulates tRNA-modifying enzyme genes at the genomic level.	tRNA-Modifying Enzymes Physically Interacts with tRNA-Modifying Enzymes
lncRNA	Most versatile —acts before or after mRNA is made, affects transcription, translation, or protein function.	Yes
miRNA	Regulate gene expression post-transcriptionally (after mRNA is made)	No direct interaction
circRNA	binds to miRNAs , preventing them from regulating tRNA-modifying enzymes .	No direct interaction
sRNA	Bacterial equivalent of miRNA, regulates post-transcriptionally but without a RISC complex.	No direct interaction
snoRNA	Does not regulate tRNA gene expression	snoRNA binds to a tRNA-modifying enzyme (e.g., methyltransferase or pseudouridine synthase).
RNase P RNA	Does not regulate tRNA gene expression	Cleaves the 5' leader sequence , generating the mature 5' end of tRNA, allowing CCA addition at the 3' end and amino acid charging.



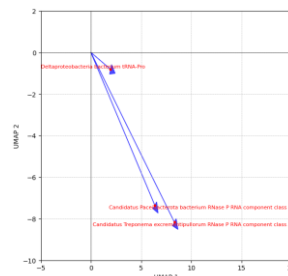
-RNase_P_RNA(387nc)+
RNase_P_RNA(**388nc**)+
tRNA(74nc) = 0



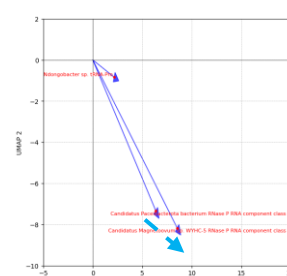
-tRNA(77nc)+
RNase_P_RNA(**413nc**)-
RNase_P_RNA(**362nc**) = 0



-tRNA(74nc)-
RNase_P_RNA(**388nc**)+
RNase_P_RNA(**413nc**) = 0



-RNase_P_RNA(406nc)+
tRNA(77nc)+
RNase_P_RNA(**362nc**) = 0



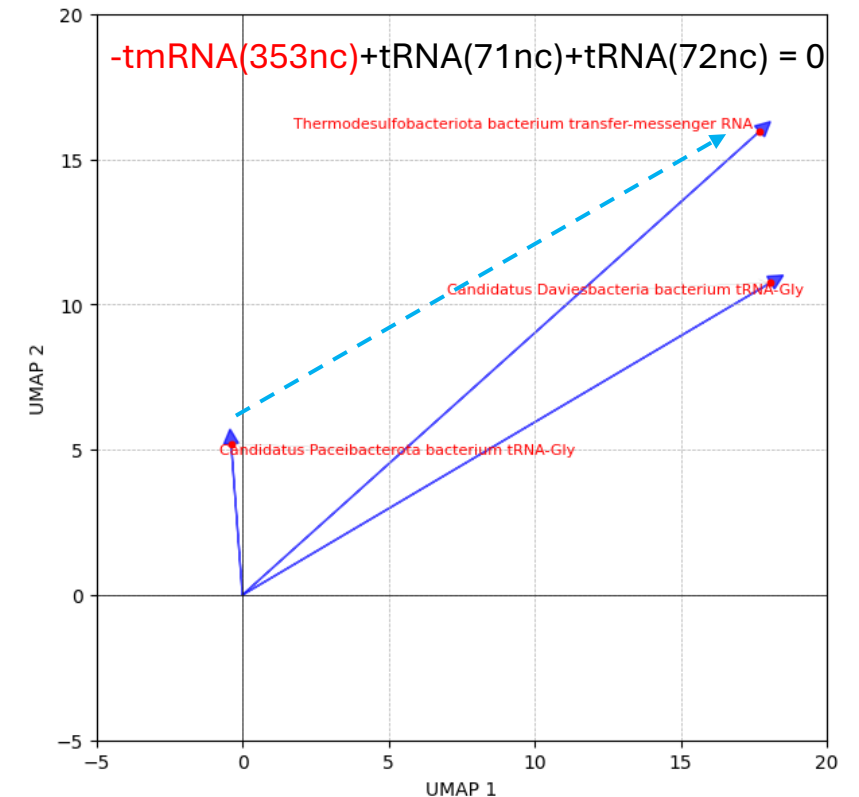
Two RNase P RNA with tRNA of proline

tmRNA (Transfer-Messenger RNA)

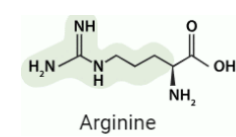
Rescues stalled ribosomes when an mRNA lacks a stop codon.

Acts as both **tRNA** and **mRNA**.

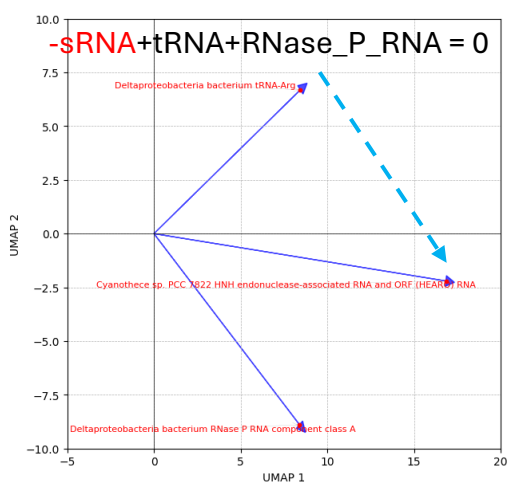
It is charged with **alanine (Ala)**.



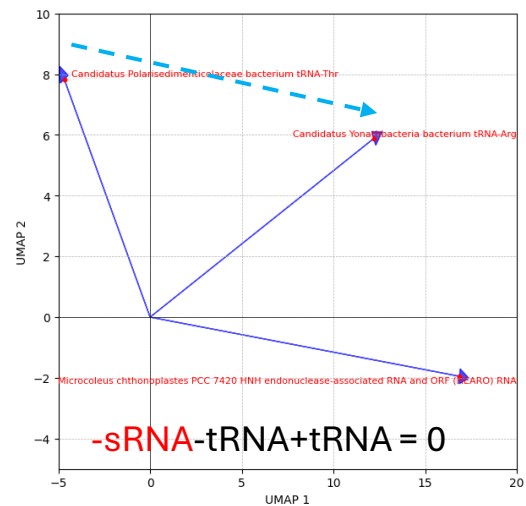
Cyanobacteria and tRNA of Arginine (3N atoms)



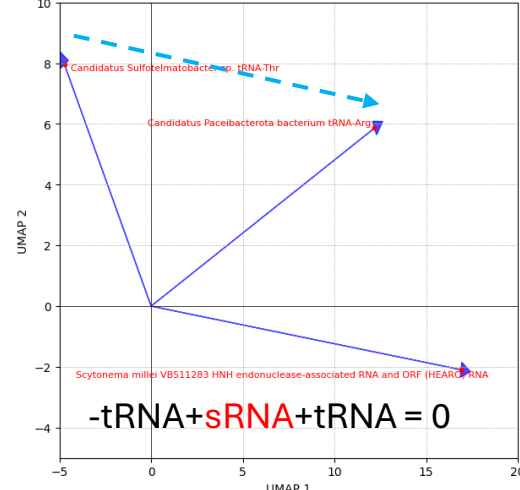
Cyanothece sp. PCC 7822



Microcoleus chthonoplastes PCC 7420

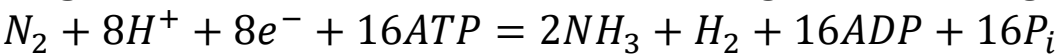


Scytonema millei VB511283

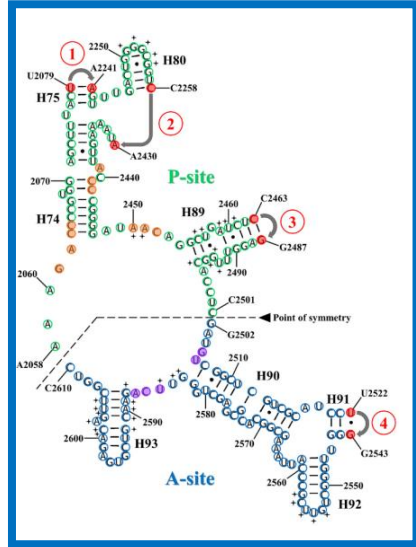


Organism Type	Estimated Origin (B years)	Related Events
LUCA (Last Universal Common Ancestor) (<i>Early prokaryotic ancestor of all life</i>)	~3.5–3.8	First known life on Earth, anaerobic metabolism, no oxygen in atmosphere
First Cyanobacteria (Ancestors of Cyanothece sp. PCC 7822)	~2.7	Oxygenic photosynthesis begins, starts producing atmospheric oxygen
Great Oxygenation Event (GOE)	~2.5	Oxygen starts accumulating in atmosphere, first mass extinction of anaerobes
Filamentous Cyanobacteria (Ancestors of Microcoleus chthonoplastes PCC 7420)	~2.0	Stromatolites become widespread, more advanced cyanobacteria evolve
Heterocyst-forming Cyanobacteria (Ancestors of Scytonema millei VB511283)	~1.5	More efficient nitrogen fixation, evolution of heterocysts

Cyanobacteria are among the oldest known life forms on Earth. **Cyanobacterium converts N₂ (dinitrogen) into NH₃.** This process, known as **biological nitrogen fixation**, allows to assimilate nitrogen for building essential biomolecules



rRNA evolution: Age Variability Across Different Regions

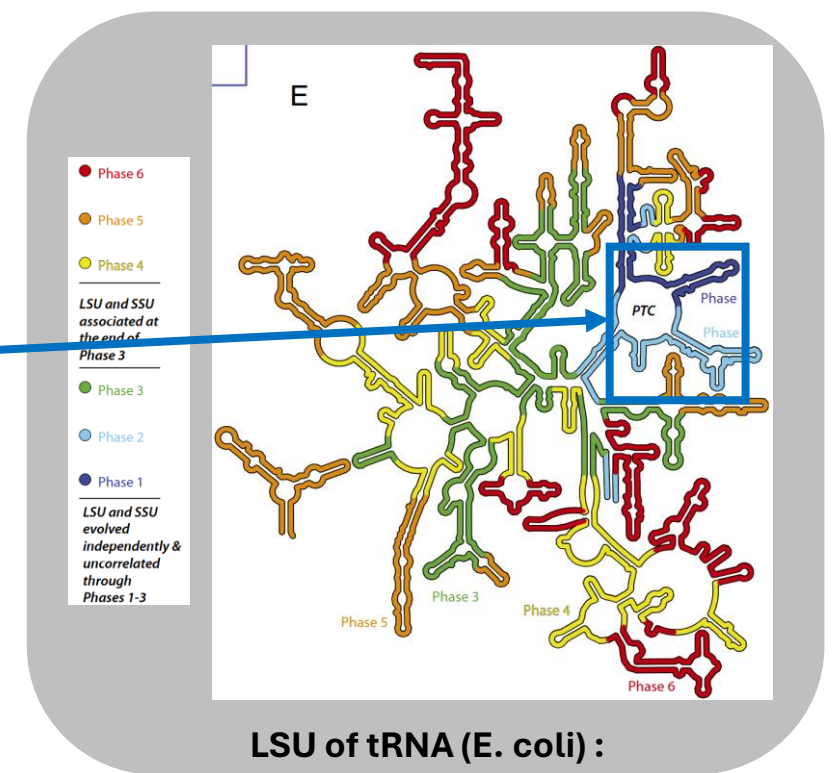


the OLDEST functional part: always the same!

evolution time

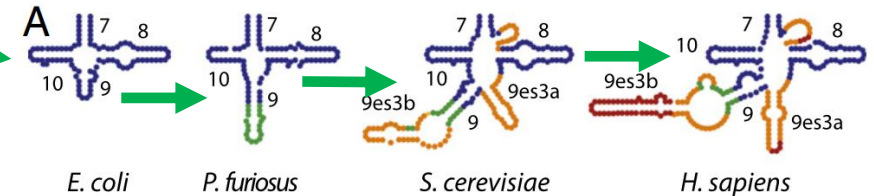
Last Universal Common Ancestor (LUCA)

<https://www.pulseheadlines.com/earths-universal-common-ancestor-volcanic-origins/43890/>



LSU of tRNA (E. coli) :

Blue part is the oldest one



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)

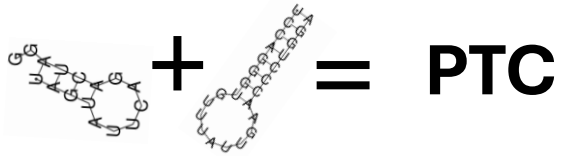
Peptidyl Transferase Center (PTC) is the oldest part of ribosomes.

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**.

Peptidyl Transferase Center (PTC) Sequences

the idea

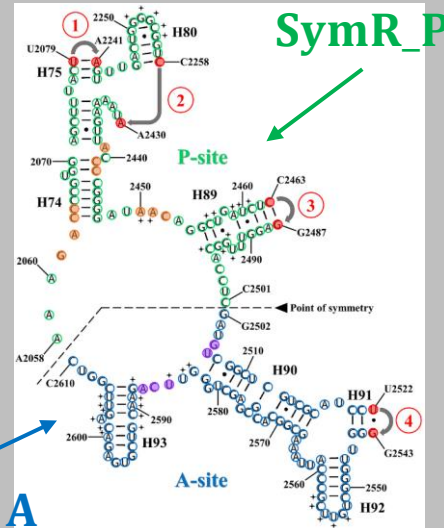
the dimerization of two similar RNA structures



“The peptidyl transferase center (PTC) evolved from a primitive system in the RNA world comprising tRNA-like molecules formed by **duplication of minihelix-like small RNA**”

Tamura, J. Biosci, 2011

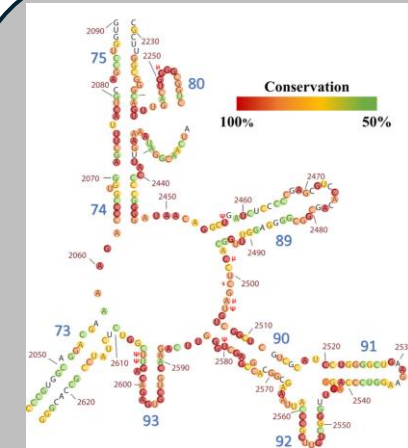
pseudosymmetrical region



$$\text{SymR_PA} = \text{SymR_P} + \text{SymR_A}$$

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)

PTC



PTC2 = red
PTC3 = PTC2 + orange
PTC4 = PTC3 + yellow
PTC5 = PTC4 + green

Nucleotide CONSERVATION level:

Red circles: 100% conservation (78 nt).
Orange circles: 90 to 99.9% conservation (68 nt)
Yellow circles: 70 to 89.9% (52nt)
Green circles: 50 to 69.9% conservation (49nt)
Black letters: less than 50% conservation (35nt)

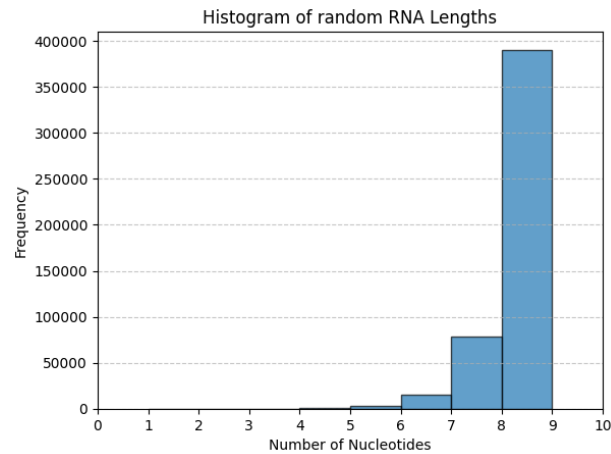
(Bernier et al.; Faraday Discuss, 2014)
(Madhan R. Tirumalai et al., 2021)

SymR_P is older than SymR_PA
PTC2 is older than PTC3, PTC4, PTC5

Transformer Hidden representation of the oldest part (PTC) in two different basis sets

BASIS I

488280 random RNAs
With length $\leq 8\text{ncl}$

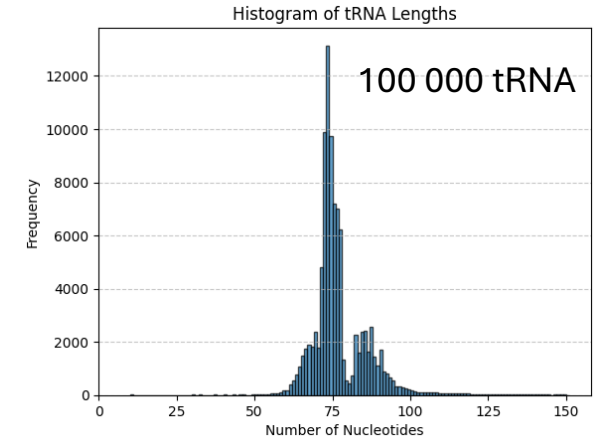


Peptidyl Transferase Center (**PTC**) is the oldest part of ribosomes.
 PTC2: $549\text{nc} = 78\text{nc}(\text{defined}) + 471\text{nc}(\text{undefined})$
 PTC3: $564\text{nc} = 146\text{nc}(\text{defined}) + 418\text{nc}(\text{undefined})$
 PTC4: $583\text{nc} = 198\text{nc}(\text{defined}) + 385\text{nc}(\text{undefined})$
 PTC5: $583\text{nc} = 247\text{nc}(\text{defined}) + 336\text{nc}(\text{undefined})$

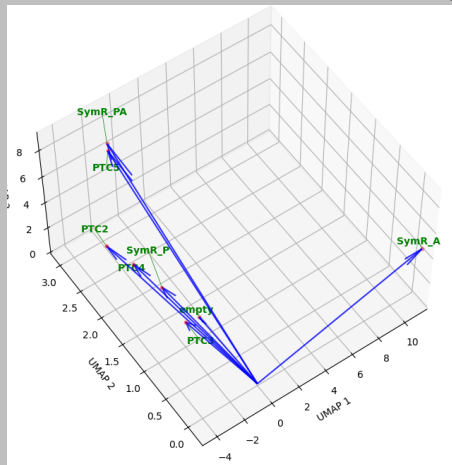
Pseudosymmetrical region (**SymR**; Agmon et al., 2005), derived from the LSU secondary structure of *Thermus thermophilus*
SymR_A: $109\text{nc} = 89\text{nc}(\text{defined}) + 20\text{nc}(\text{undefined})$
SymR_P: $444\text{nc} = 89\text{nc}(\text{defined}) + 375\text{nc}(\text{undefined})$
 (Petrov et al., 2013).
 (Madhan R. Tirumalai et al., 2021)

BASIS II

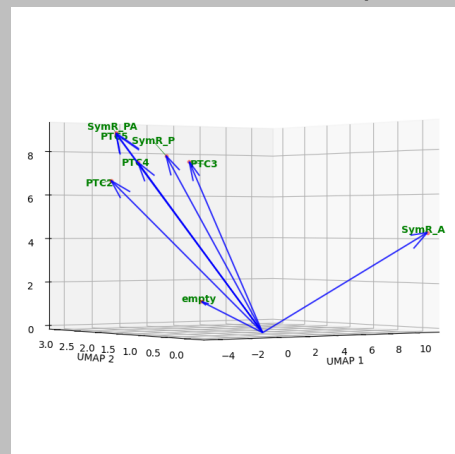
100 000 tRNAs
 $50\text{ncl} \leq \text{length} \leq 125\text{ncl}$



Visualization of the same 3D plot from two distinct viewpoints

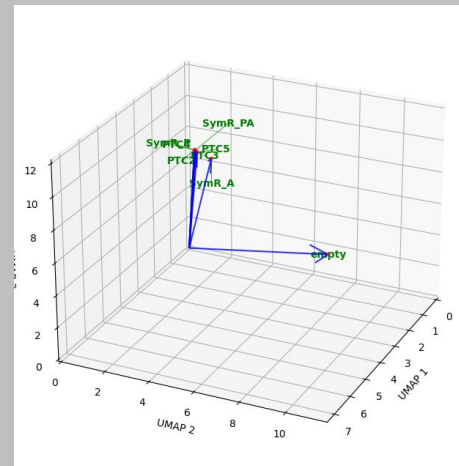


SymR_A \perp all

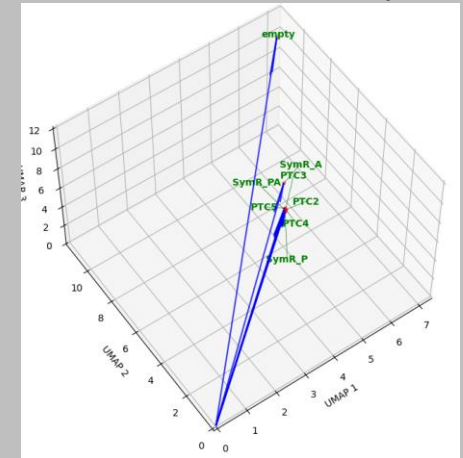


PTC5 = SymR_PA

Visualization of the same 3D plot from two distinct viewpoints



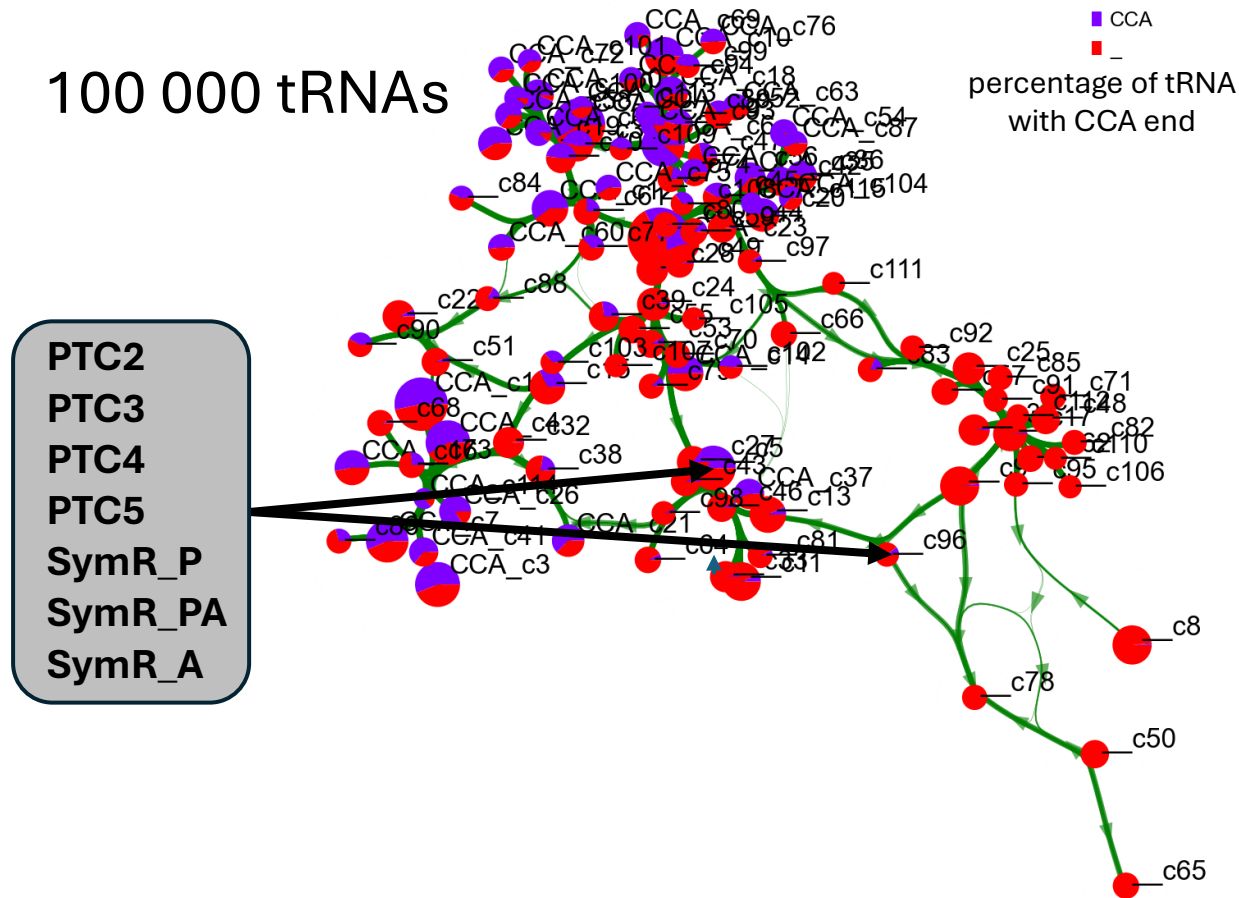
SymR_A \neq (PTC2 = PTC3 = PTC4 = PTC5 = SymR_PA = SymR_P)



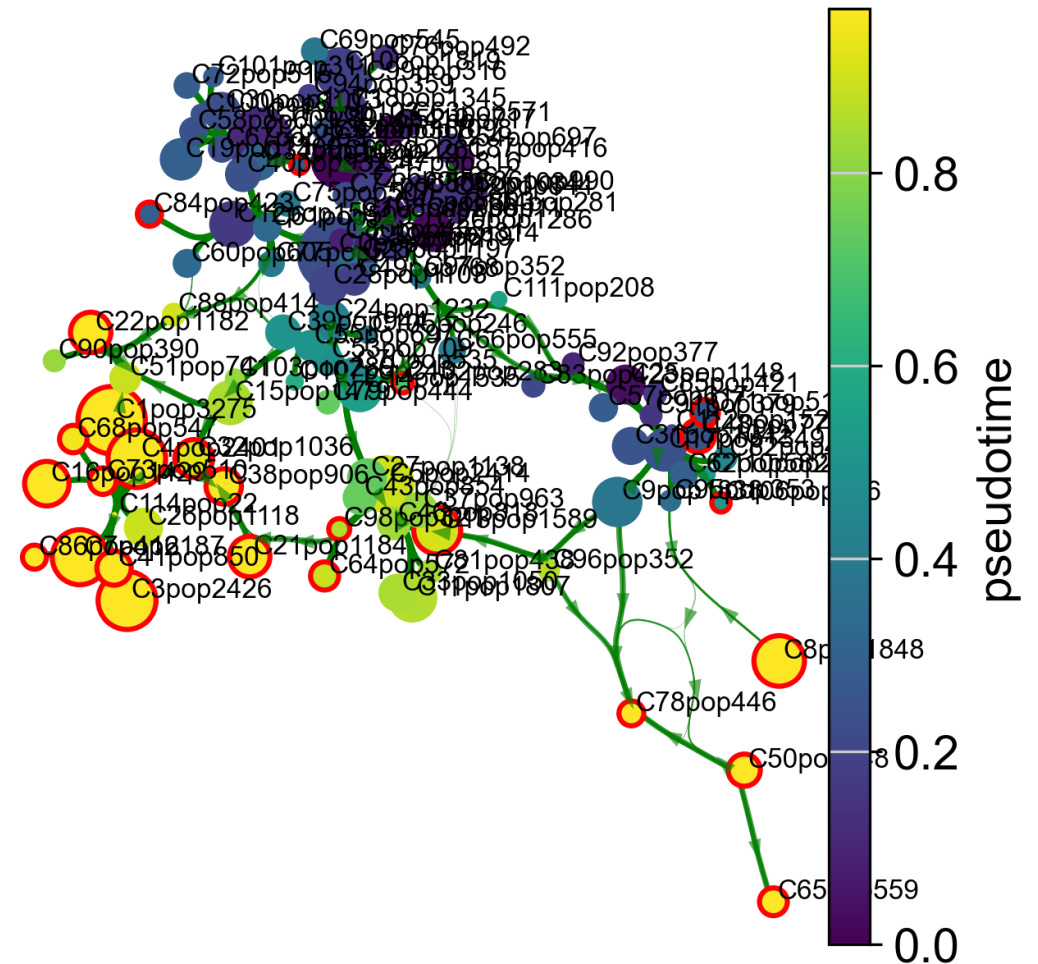
Single-Cell Inspired Analysis of tRNA and oldest rRNA

Cluster Composition. K=200. ncomp = 640

100 000 tRNAs



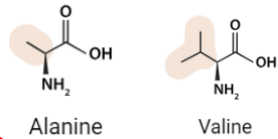
Pseudotime



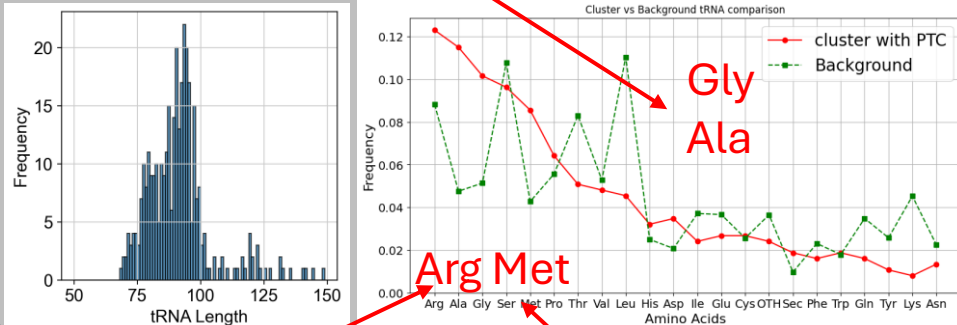
Peptidyl transferase center was detected in the 5th and 96th clusters.

Single-Cell Inspired Analysis of tRNA and oldest rRNA multiple runs

two of the simplest and oldest amino acids.



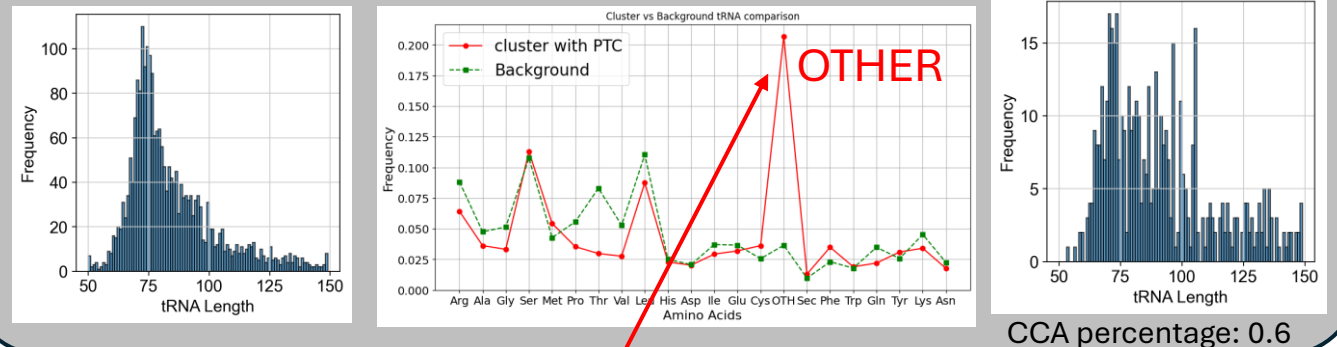
Cluster96 with **PTC** (CCA=0.11, L=93.52nc):



essential for stabilizing the negatively charged RNA phosphate backbone encoded by the universal start codon

Four standard amino acids tRNA were detected above the background

Cluster5 with **PTC** (CCA=0.5, L= 97.15nc)



other tRNA and pre-tRNA

No standard amino acids tRNA were detected above the background

Peptidyl transferase center was detected in the 5th and/or 96th clusters.

Amino acid images from: <https://www.rapidnovor.com/structure-of-an-amino-acid/>

Method used from : Generalized and scalable trajectory inference in single-cell omics data with VIA, Stassen et al., Nature Communications, 2021

CONCLUSIONS

- ❖ The vanilla transformer with custom embeddings and a masked training paradigm allows for deeper hidden representations than a BERT-like transformer
- ❖ Different ncRNAs are connected among each other by simple additive equations, some preliminary interpretations are proposed:
 - *Archaea and bacteria often coexist in various environments*
 - *Biological nitrogen fixation by Cyanobacteria*
- ❖ It is possible to compress the hidden representations of various PTCs into a single vector in 3D space by switching from a random basis to a tRNA basis
- ❖ The PTC is either associated with tRNA for Gly, Ala, Arg and Met, or with other pre-tRNAs not directly related to the standard 21 amino acids

What's Next?

- 1. Refinement of RNA similarity measure:** Instead of using a single scalar (such as L2 distance), one can extract the individual attention head outputs before concatenation into the final embedding. This approach enables a more detailed characterization of RNA sequences, allowing for distinctions where **two RNAs may be similar in one aspect but different in another.**
- 2. Refinement of autoencoder embeddings:** Use Graph Attention Network (GAT) to incorporate the graph structure and adjust the node (RNAs) representation based on the importance of its neighbors.