RNA Sequence Analysis using Transformer Models

Work in Progress, preliminary results

OBJECTIVES

1. Generate hidden representations of RNA molecules using modern Natural Language Processing (NLP) approaches

2. Compare these representations with those produced by existing state-of-the-art methods

3. Try to use these representations to describe some RNA-related processes.

OUTLINE

- **Objectives**

- **The idea of transformers**
- **Dataset**
  - What is RNA molecule
  - RNAcental 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**

The idea of transformers

Embedding Layer

N tokens

$$X^{(0)} = \begin{pmatrix} \boldsymbol{x}_0^{(0)} & \dots & \boldsymbol{x}_n^{(0)} & \dots & \boldsymbol{x}_N^{(0)} \end{pmatrix} \; D$$

Iteratively applying a transformer block

hidden states

$$X^{(m-1)} = \begin{pmatrix} \boldsymbol{x}_0^{(m-1)} & \dots & \boldsymbol{x}_n^{(m-1)} & \dots & \boldsymbol{x}_N^{(m-1)} \end{pmatrix}$$

hidden states

$$X^{(m)} = \begin{pmatrix} \boldsymbol{x}_0^{(m)} & \dots & \boldsymbol{x}_n^{(m)} & \dots & \boldsymbol{x}_N^{(m)} \end{pmatrix}$$

$$\boldsymbol{y}_n^{(m)} = \sum_{n'=1}^{N} x_{n'}^{(m-1)} A_{n',n}^{(m)} \rightarrow \sum_{h=1}^{H} V_h^{(m)} X^{(m-1)} A_h^{(m)}$$

$$Y^{(m)} = \begin{pmatrix} \boldsymbol{y}_0^{(m)} & \dots & \boldsymbol{y}_n^{(m)} & \dots & \boldsymbol{y}_N^{(m)} \end{pmatrix}$$

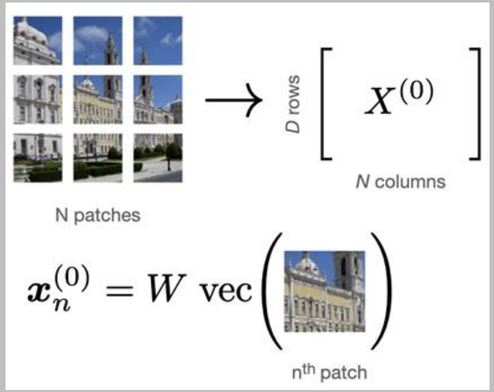
$$\boldsymbol{x}_n^{(m)} = MLP_\theta\left(\boldsymbol{y}_n^{(m)}\right)$$

the final layer M
sequence_embedding

$$\boxed{h = \sum_{n=1}^{N} \boldsymbol{x}_n^{(M)}}$$

Encoding an image

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



vec operator: Each patch is reshaped into a vector by the vec operator.
matrix $\boldsymbol{W}$: maps a vector (the patch) to a D dimensional vector $\boldsymbol{x}_n^{(0)}$.
[Dosovitskiy et al., 2021]

The attention matrix

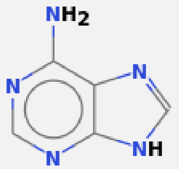
$$A_{n,n'} = \frac{x_n^T x_{n'}}{\sum_{n''=1}^{N} x_{n''}^T x_{n'}} \rightarrow \frac{\exp\left(x_n^T x_{n'}\right)}{\sum_{n''=1}^{N} \exp\left(x_{n''}^T x_{n'}\right)} \rightarrow \frac{\exp\left(x_n^T \boldsymbol{U}^T \boldsymbol{U} x_{n'}\right)}{\sum_{n''=1}^{N} \exp\left(x_{n''}^T \boldsymbol{U}^T \boldsymbol{U} x_{n'}\right)} \rightarrow \frac{\exp\left(x_n^T \boldsymbol{U}_k^T \boldsymbol{U}_q x_{n'}\right)}{\sum_{n''=1}^{N} \exp\left(x_{n''}^T \boldsymbol{U}_k^T \boldsymbol{U}_q x_{n'}\right)}$$

$$\boxed{A_{n,n'} = \frac{\exp\left(k_n^T q_{n'}\right)}{\sum_{n''=1}^{N} \exp\left(k_{n''}^T q_{n'}\right)}, \quad \text{where} \quad \begin{cases} q_{h,n}^{(m)} = U_q^{(m)} x_n^{(m-1)}, & \text{queries} \\ \boldsymbol{k}_{h,n}^{(m)} = \boldsymbol{U}_k^{(m)} \boldsymbol{x}_n^{(m-1)}, & \textbf{\textit{keys}} \end{cases}}$$

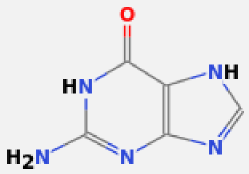
[Turner R.E., https://arxiv.org/abs/2304.10557, 2023]

What is RNA molecule ?

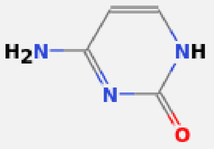
A = Adenine ($C_5H_5N_5$)



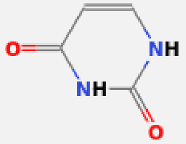
G = Guanine ($C_5H_5N_5O$)



C = Cytosine ($C_4H_5N_3O$)

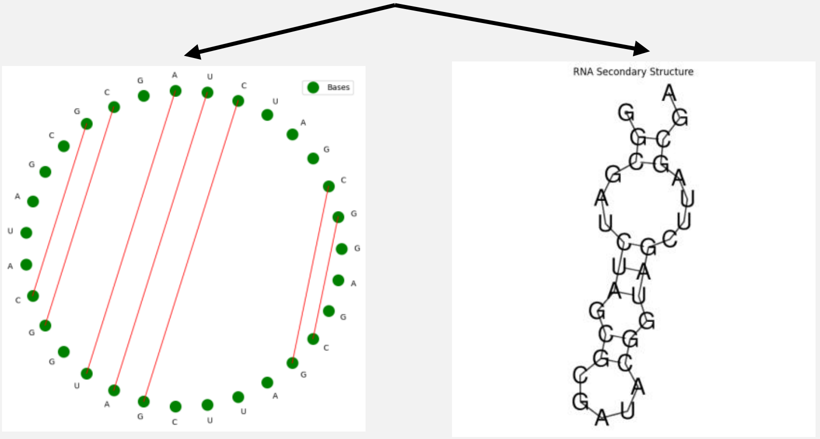


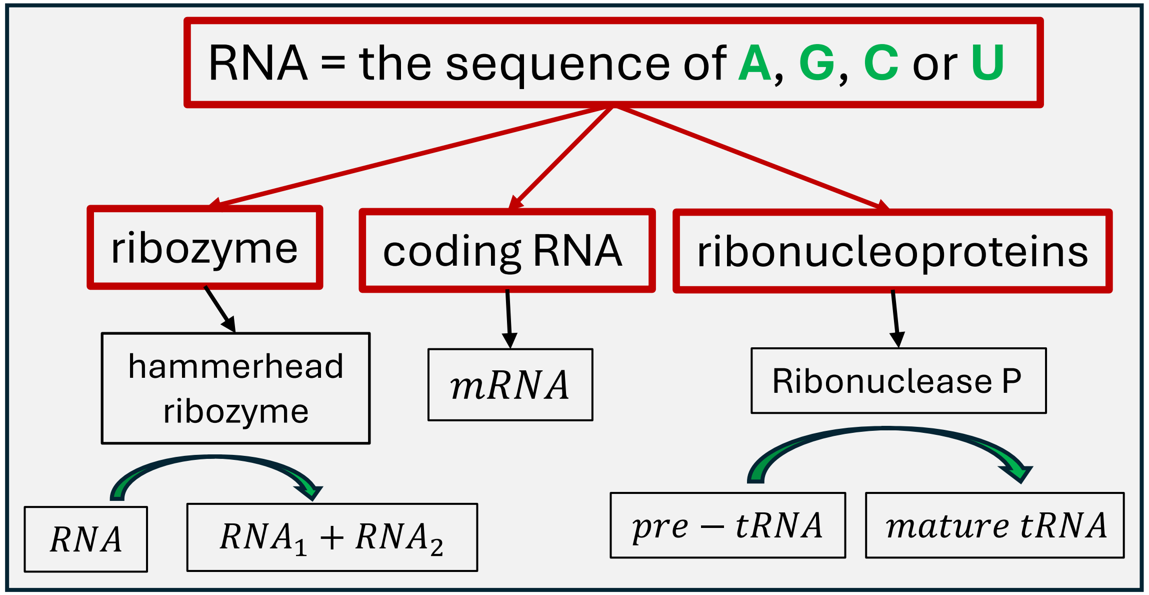
U = Uracil ($C_4H_4N_2O_2$)



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

GGCGAUCUAGCGCGAUACGGUAGCUUAGCGA





RNA = the sequence of A, G, C or U

ribozyme

coding RNA

ribonucleoproteins

hammerhead ribozyme

$mRNA$

Ribonuclease P

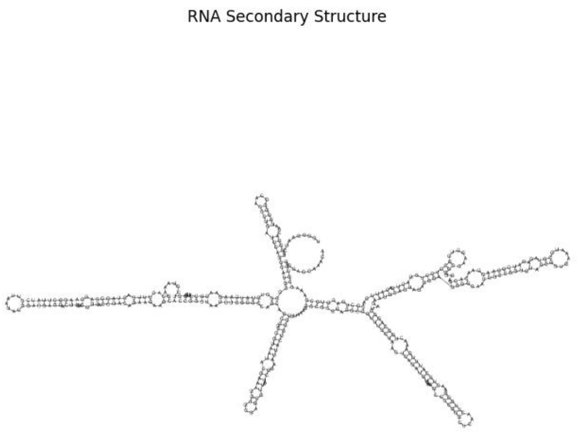
RNA

$RNA_1 + RNA_2$

$pre-tRNA$

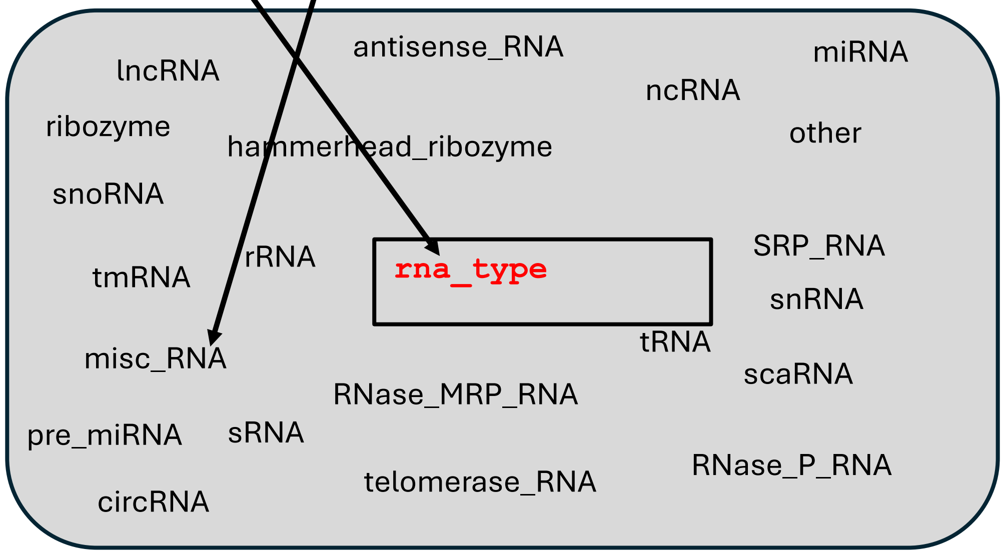
$mature\ tRNA$

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" ] }, …
```



lncRNA

antisense_RNA

miRNA

ncRNA

ribozyme

hammerhead_ribozyme

other

snoRNA

rRNA

rna_type

SRP_RNA

tmRNA

snRNA

tRNA

misc_RNA

scaRNA

RNase_MRP_RNA

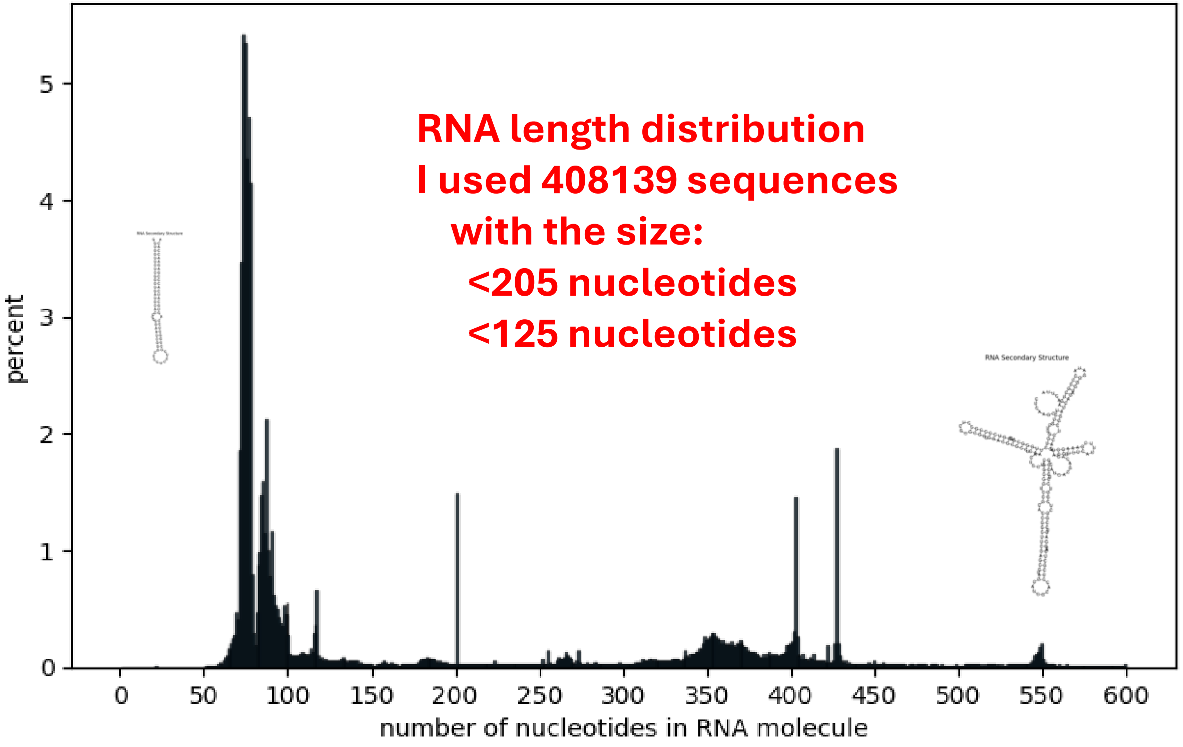
pre_miRNA

sRNA

telomerase_RNA

RNase_P_RNA

circRNA



RNA length distribution
I used 408139 sequences
with the size:
<205 nucleotides
<125 nucleotides

RNA autoencoder (vanilla transformer)

AUGGAUGGUUGA

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

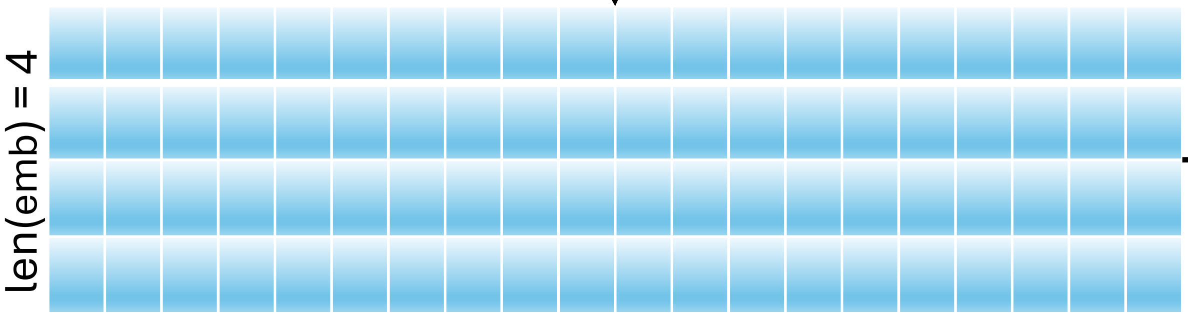
9	6	7	7	9	6	7	7	6	6	7	9

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

hidden representation

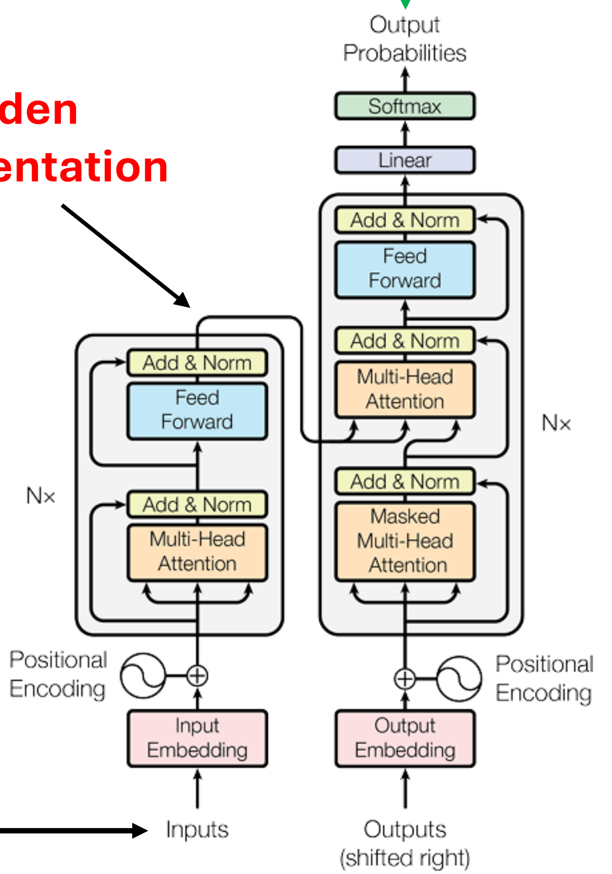
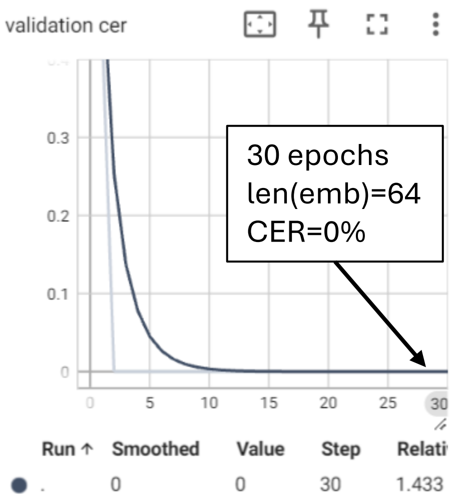


Figure 1: The Transformer - model architecture.

Ashish Vaswani et al., Attention Is All You Need
https://doi.org/10.48550/arXiv.1706.03762

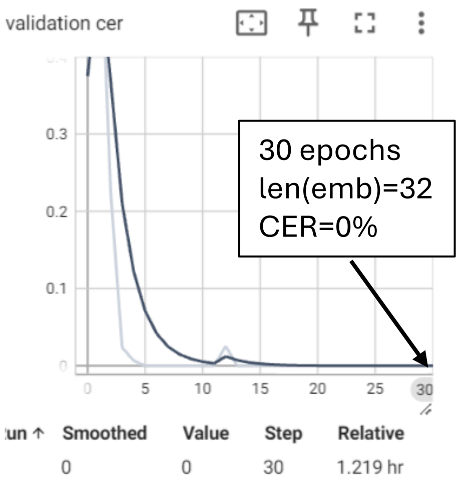
RESULTS for RNA autoencoder validation



30 epochs
len(emb)=64
CER=0%

Run ↑	Smoothed	Value	Step	Relati
● .	0	0	30	1.433

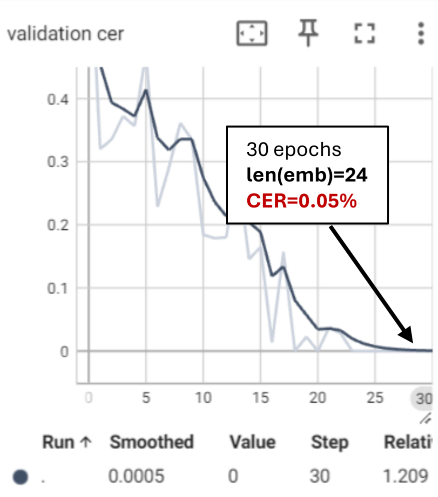
len(emb)= 64
CER = 0%



30 epochs
len(emb)=32
CER=0%

un ↑	Smoothed	Value	Step	Relative
	0	0	30	1.219 hr

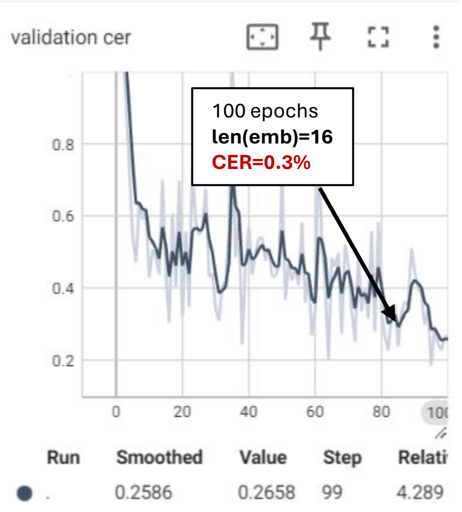
len(emb)= 32
CER = 0%



30 epochs
len(emb)=24
CER=0.05%

Run ↑	Smoothed	Value	Step	Relati
● .	0.0005	0	30	1.209

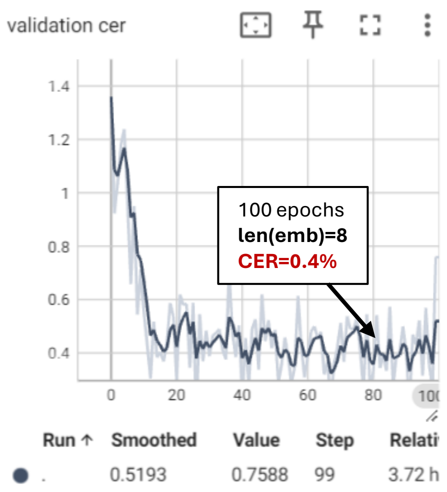
len(emb)= 24
CER = 0.05%



100 epochs
len(emb)=16
CER=0.3%

Run	Smoothed	Value	Step	Relati
● .	0.2586	0.2658	99	4.289

len(emb)= 16
CER = 0.3%



100 epochs
len(emb)=8
CER=0.4%

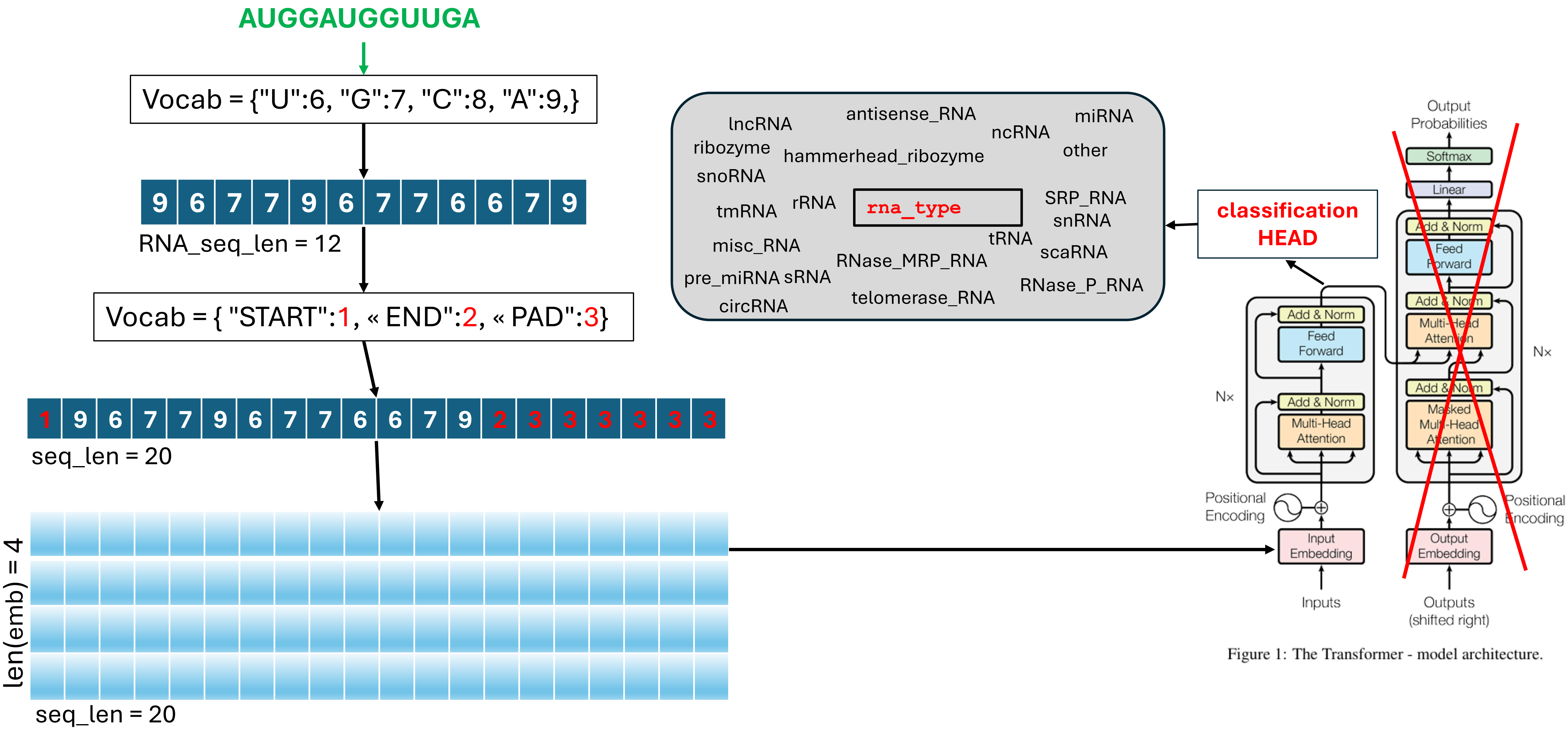
Run ↑	Smoothed	Value	Step	Relati
● .	0.5193	0.7588	99	3.72 h

len(emb)= 8
CER = 0.4%

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)



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

len(emb) = 4

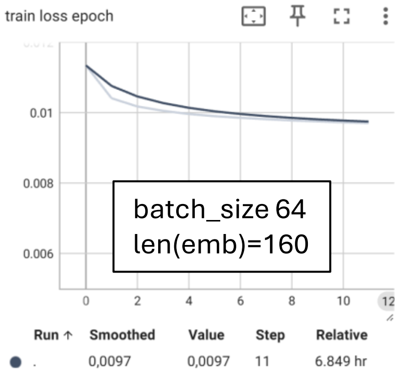
seq_len = 20

lncRNA antisense_RNA miRNA ncRNA ribozyme hammerhead_ribozyme other snoRNA tmRNA rRNA **rna_type** SRP_RNA snRNA misc_RNA tRNA scaRNA RNase_MRP_RNA pre_miRNA sRNA RNase_P_RNA circRNA telomerase_RNA

classification HEAD

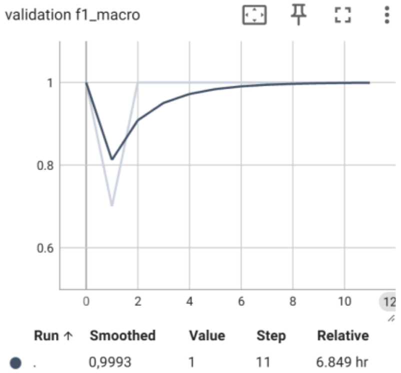
Figure 1: The Transformer - model architecture.

RESULTS for RNAclassificator validation

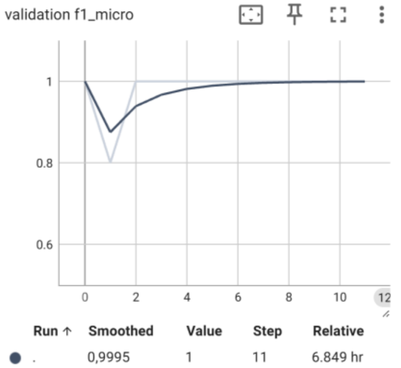


batch_size 64
len(emb)=160

Run ↑	Smoothed	Value	Step	Relative
● .	0,0097	0,0097	11	6.849 hr

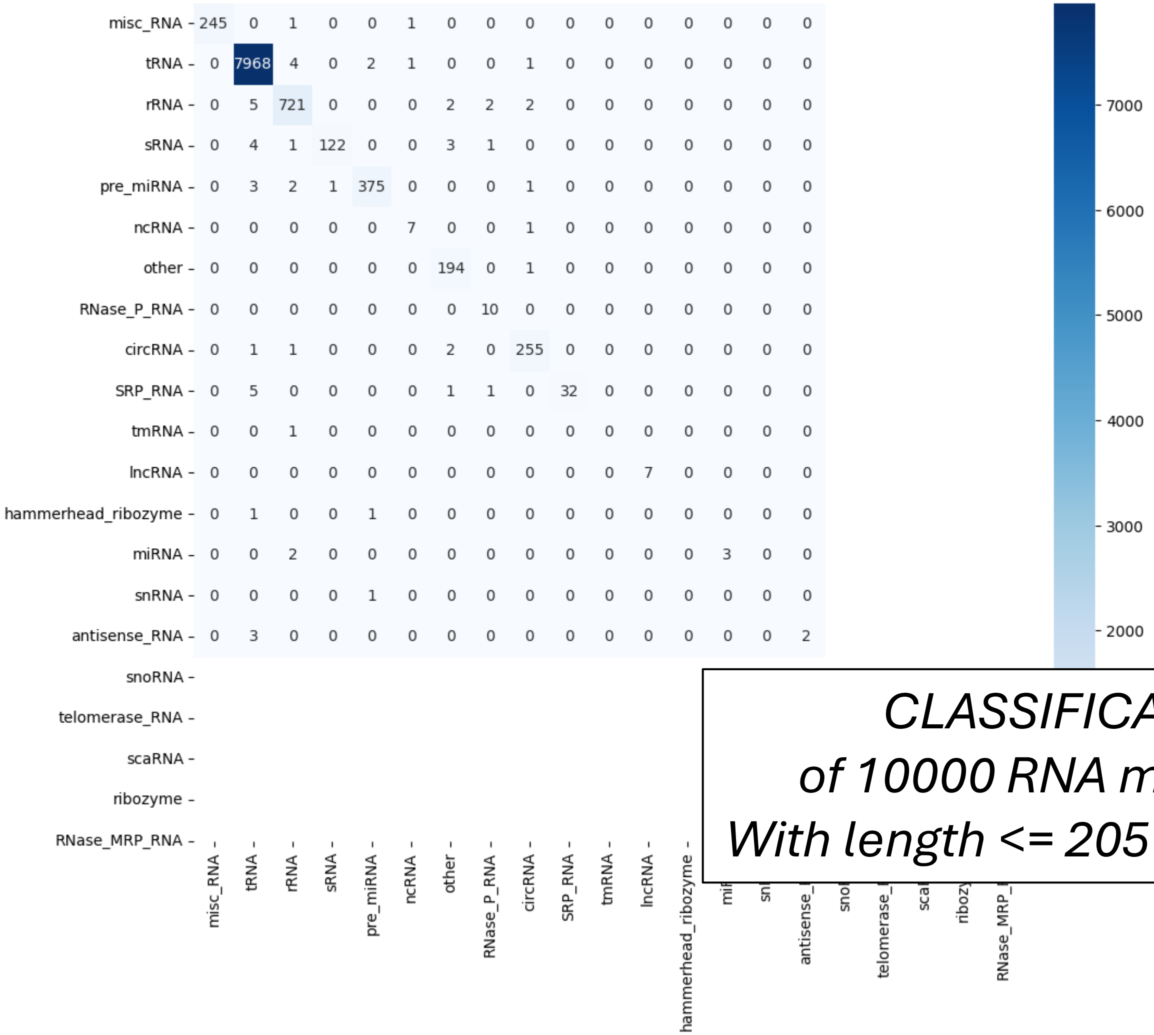


Run ↑	Smoothed	Value	Step	Relative
● .	0,9993	1	11	6.849 hr



Run ↑	Smoothed	Value	Step	Relative
● .	0,9995	1	11	6.849 hr

VALIDATION
f1_micro 0.9941
f1_macro 0.7351456585589439



	misc_RNA	tRNA	rRNA	sRNA	pre_miRNA	ncRNA	other	RNase_P_RNA	circRNA	SRP_RNA	tmRNA	lncRNA	hammerhead_ribozyme	mi	sn	antisense_	sno	telomerase_	sca	riboz	RNase_MRP_
misc_RNA	245	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0					
tRNA	0	7968	4	0	2	1	0	0	1	0	0	0	0	0	0	0					
rRNA	0	5	721	0	0	0	2	2	2	0	0	0	0	0	0	0					
sRNA	0	4	1	122	0	0	3	1	0	0	0	0	0	0	0	0					
pre_miRNA	0	3	2	1	375	0	0	0	1	0	0	0	0	0	0	0					
ncRNA	0	0	0	0	0	7	0	0	1	0	0	0	0	0	0	0					
other	0	0	0	0	0	0	194	0	1	0	0	0	0	0	0	0					
RNase_P_RNA	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0					
circRNA	0	1	1	0	0	0	2	0	255	0	0	0	0	0	0	0					
SRP_RNA	0	5	0	0	0	0	1	1	0	32	0	0	0	0	0	0					
tmRNA	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0					
lncRNA	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0					
hammerhead_ribozyme	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0					
miRNA	0	0	2	0	0	0	0	0	0	0	0	0	0	3	0	0					
snRNA	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0					
antisense_RNA	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	2					
snoRNA																					
telomerase_RNA																					
scaRNA																					
ribozyme																					
RNase_MRP_RNA																					

CLASSIFICATION
of 10000 RNA molecules
With length <= 205 nucleotides

RNA autoencoder (BERT-like transformer)

AUGGAUGGUUGA

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

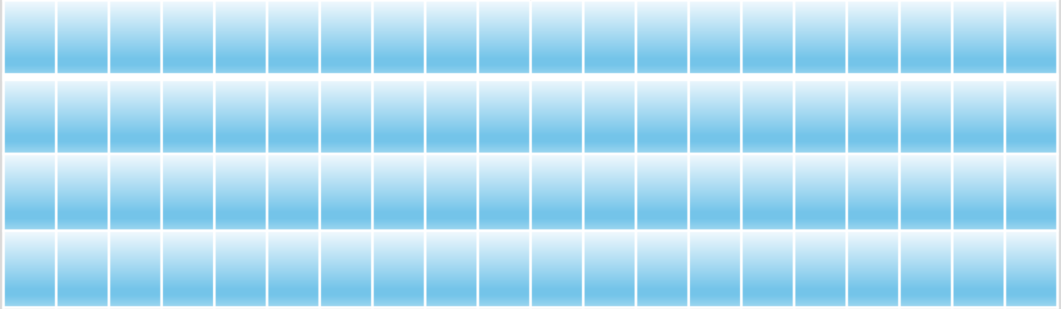
9	6	7	7	9	6	7	7	6	6	7	9

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

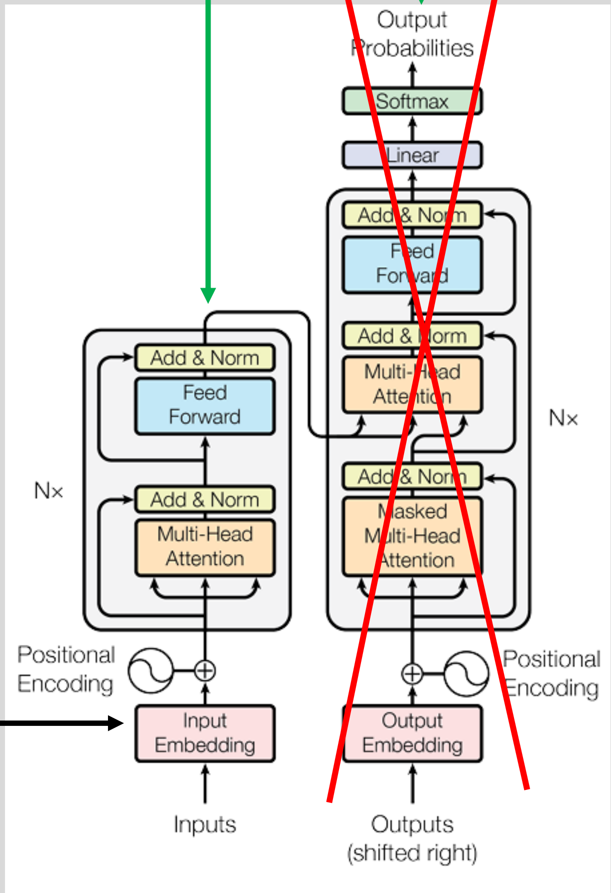
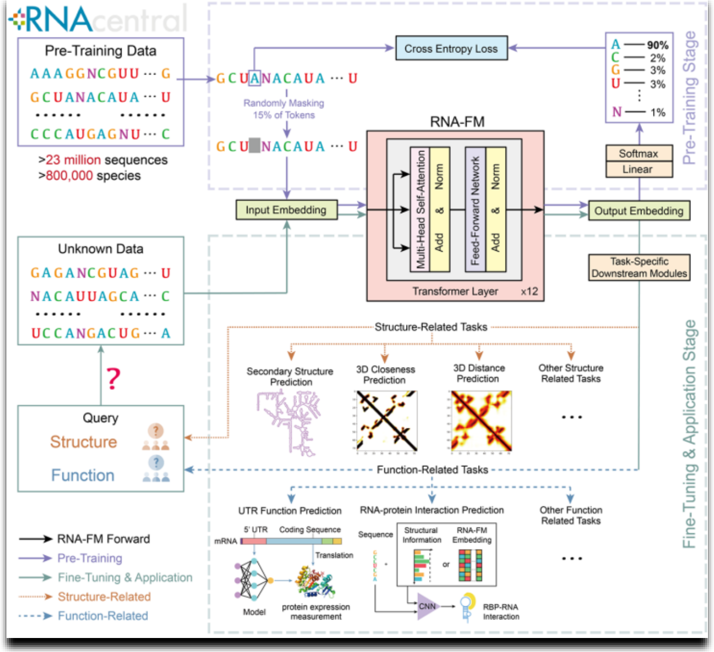


Figure 1: The Transformer - model architecture.

STATE of the ART RNAFM model



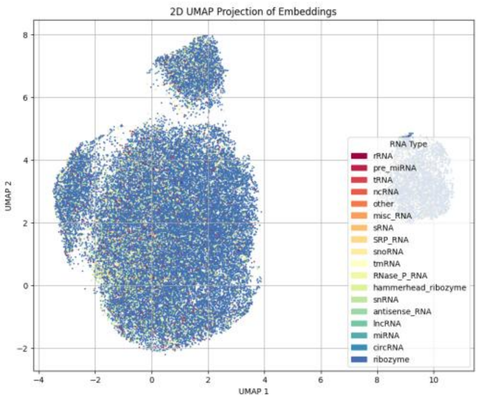
J.Chen et al.,
https://www.biorxiv.org/content/10.1101/2022.08.06.503062v2

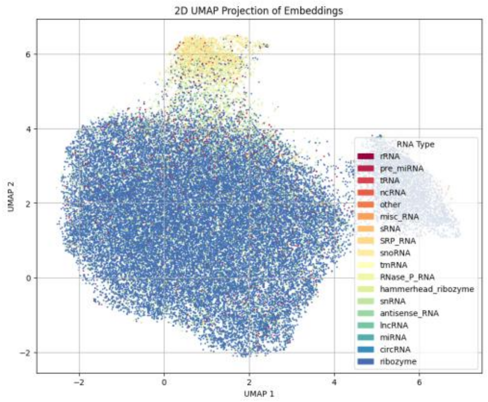
RESULTS for RNA autoencoder validation (for 40 000 RNAs)

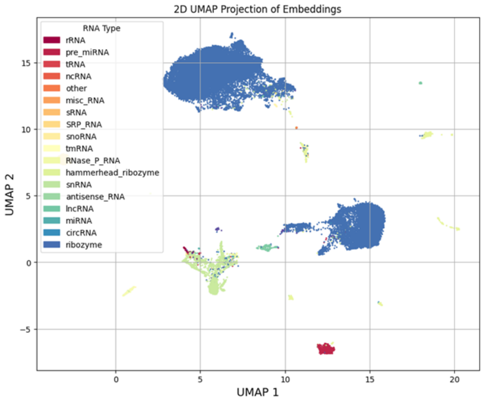
reconstruction loss (1)
random embedding
batch_size = 64,
len(emb) = 64

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





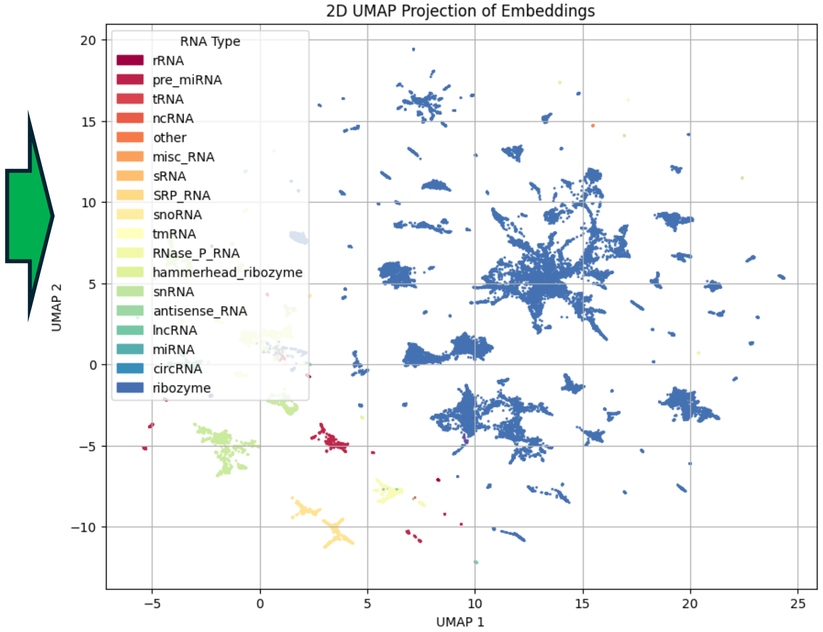


4 vanilla transformer-based encoder blocks
400 000 sequences
L × 64 embedding matrix for each RNA (length L)
RTX 4060, 8GB for 12 hours

IMPROVEMENT direction

- *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



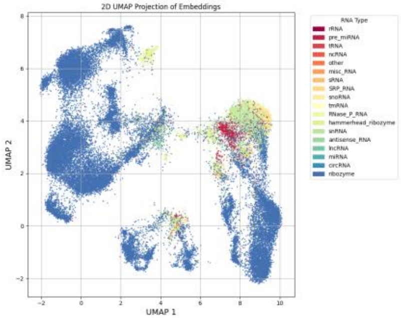
12 transformer-based bidirectional encoder blocks
23 million sequences
L × 640 embedding matrix for each RNA (length L)
eight A100 GPUs of 80 GB memories for one month

Further Improvement (for 40 000 RNAs)

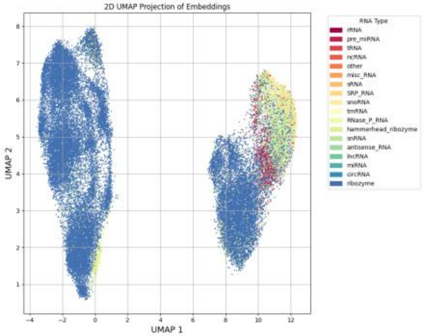
- *secondary/3D structure prediction*
- *SARS-CoV-2 genome structure and evolution prediction*
- *protein-RNA binding preference modeling*
- *gene expression regulation modeling*

vanilla transformer-based

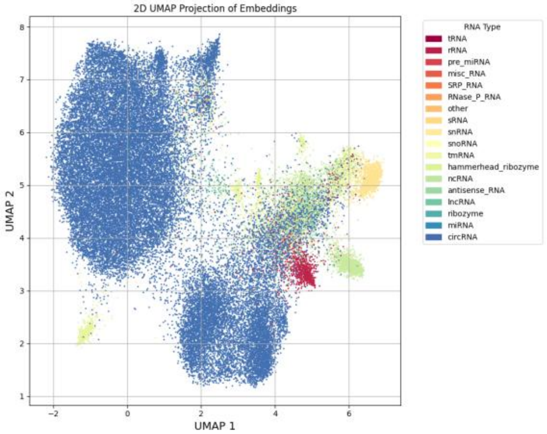
reconstruction loss (1)
random embedding
len(emb) = 160



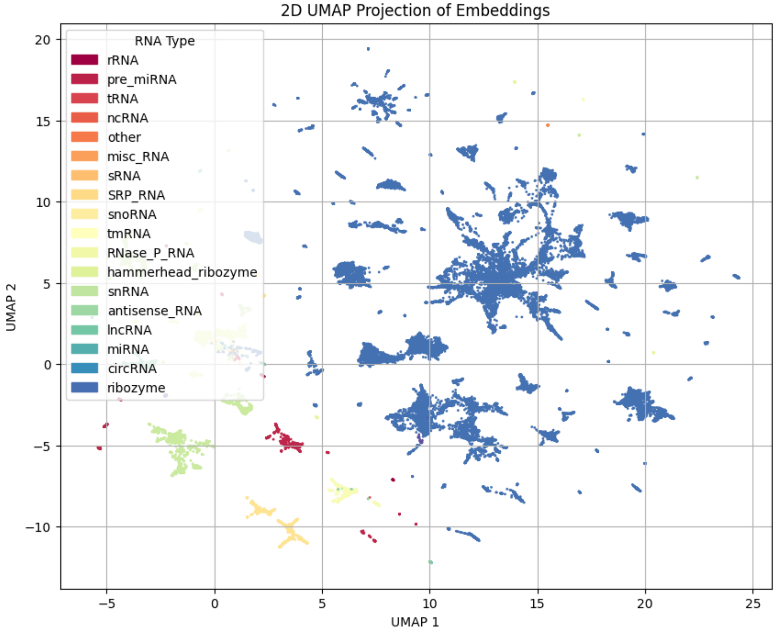
reconstruction loss (1)
custom embedding
len(emb) = 160



Composite autoencoder

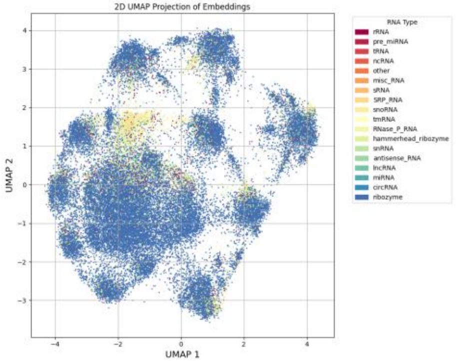


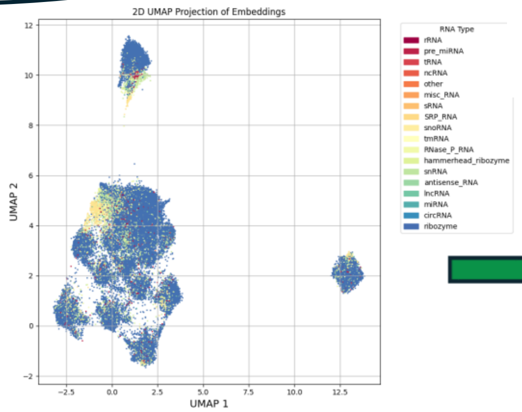
STATE of the ART
RNAFM model



0.4 million sequences
4 blocks, 8 heads
L × 64, 160 embedding matrix for each RNA
length Lmax = 205 nucleotides
RTX 4060, 8GB for 12 hours

BERT-like transformer





IMPROVEMENT direction

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

Word embeddings arithmetic with UMAP

ENGLISH language

Laptop ≈ Computer + Portable
Smartphone ≈ Phone + Smart

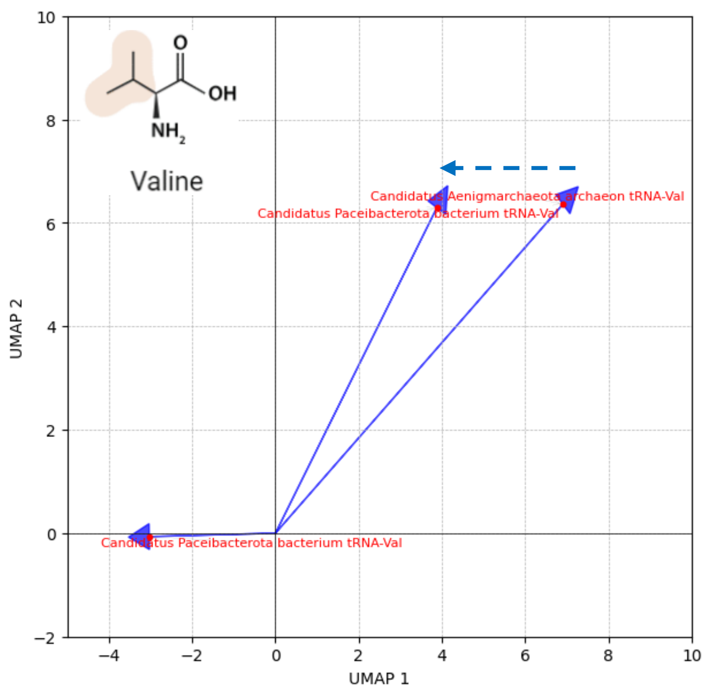
Fast + More ≈ Faster
Happy + Not ≈ Sad

King – Man + Woman ≈ Queen
Paris – France + Italy ≈ Rome

RNA language

archaea and bacteria often coexist in various environments

Candidatus Paceibacterota bacterium **tRNA-Val**



Candidatus Paceibacterota bacterium **tRNA-Val**

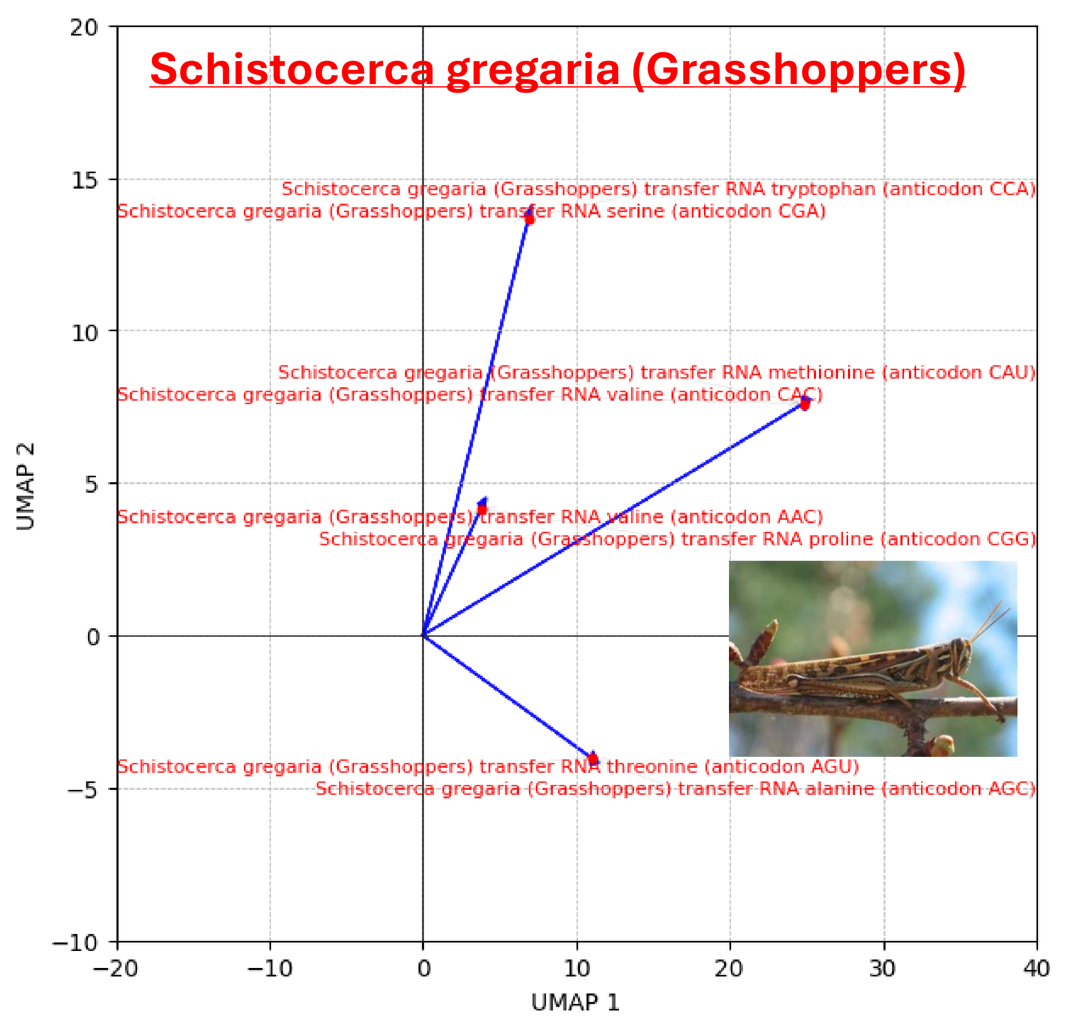
tRNA(73nt) = tRNA(73nt) + tRNA(74nt)

Candidatus Aenigmarchaeota archaeon **tRNA-Val**

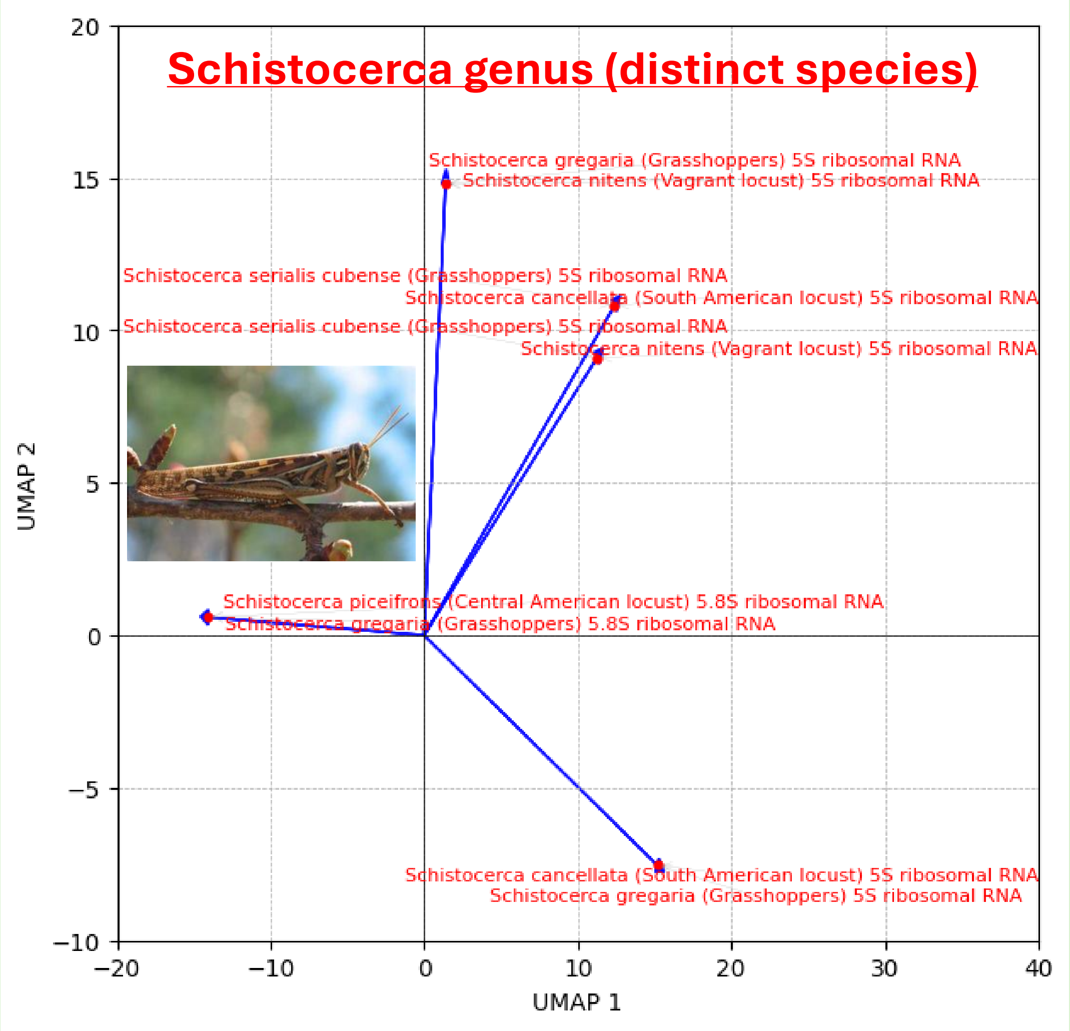
Amino acids from:
https://www.rapidnovor.com/structure-of-an-amino-acid/

Word arithmetic (UMAP): tRNA and rRNA of Schistocerca

Transfer RNA (tRNA)



5S Ribosomial RNA (rRNA)

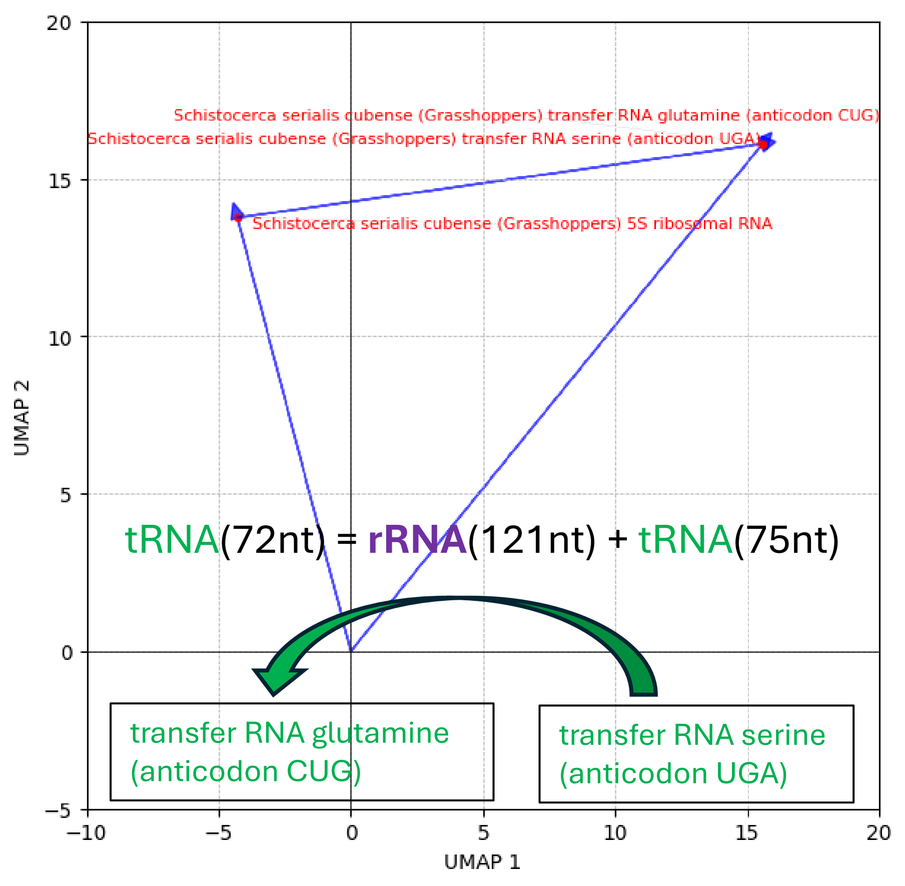


Each arrow consists of two sub-arrows to identify two similar RNAs that share the same hidden representation.

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

Word embedding arithmetic: more examples using UMAP

Schistocerca serialis cubense **5S ribosomal RNA**

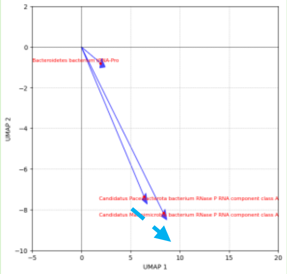


tRNA(72nt) = **rRNA**(121nt) + tRNA(75nt)

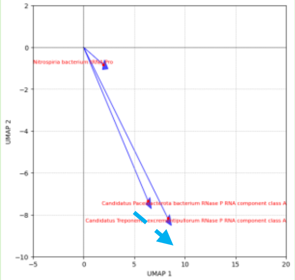
transfer RNA glutamine (anticodon CUG)

transfer RNA serine (anticodon UGA)

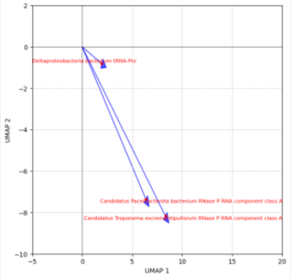
-RNase_P_RNA(387nc)+
RNase_P_RNA(**388nc**)+
tRNA(74nc) = 0



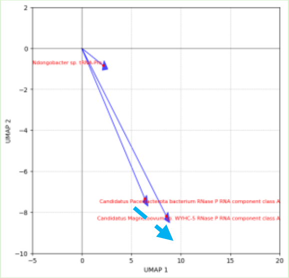
-tRNA(77nc)+
RNase_P_RNA(**413nc**)-
RNase_P_RNA(**362nc**) = 0



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

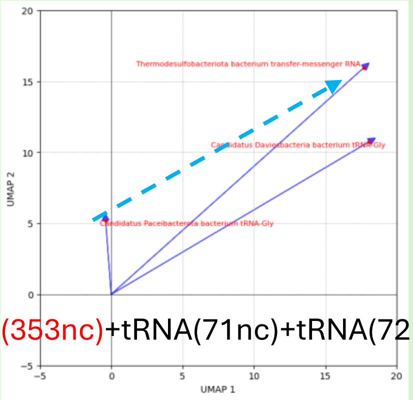


-RNase_P_RNA(406nc)+
tRNA(77nc)+
RNase_P_RNA(**362nc**) = 0



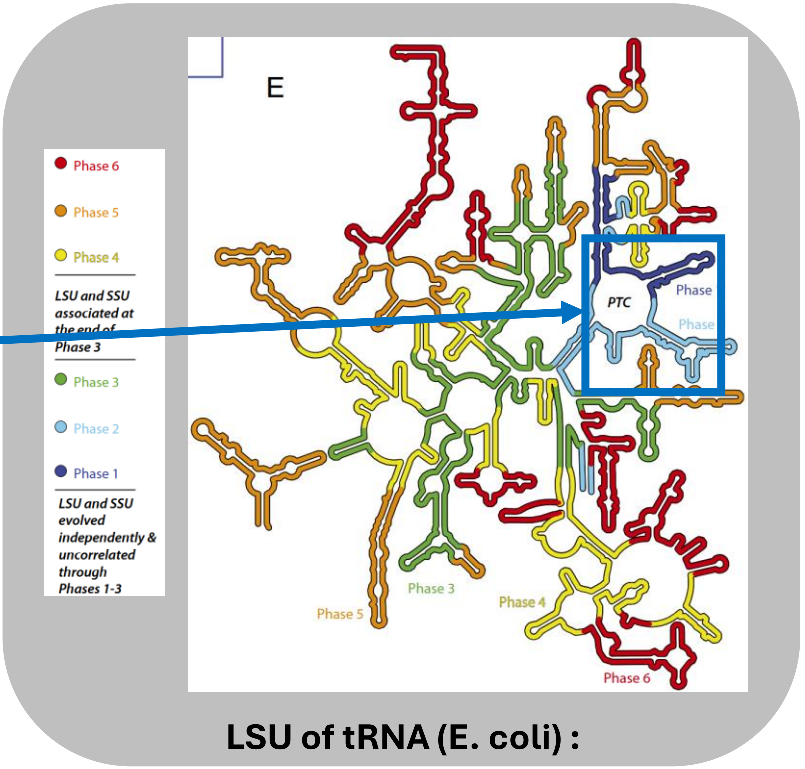
Two **RNase P RNA** with tRNA of proline

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)**.



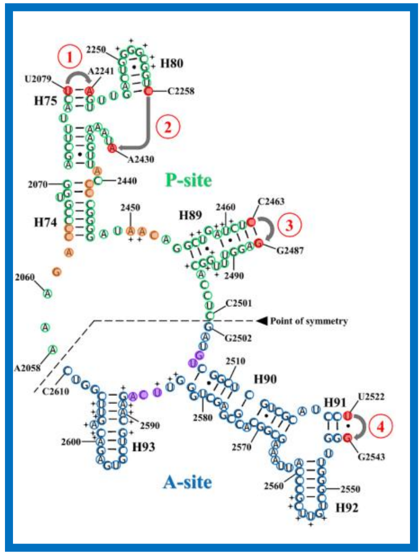
-tmRNA(353nc)+tRNA(71nc)+tRNA(72nc) = 0

rRNA evolution: Age Variability Across Different Regions


E
Phase 6
Phase 5
Phase 4
LSU and SSU associated at the end of Phase 3
Phase 3
Phase 2
Phase 1
LSU and SSU evolved independently & uncorrelated through Phases 1-3
PTC
Phase
Phase
Phase 5
Phase 3
Phase 4
Phase 6

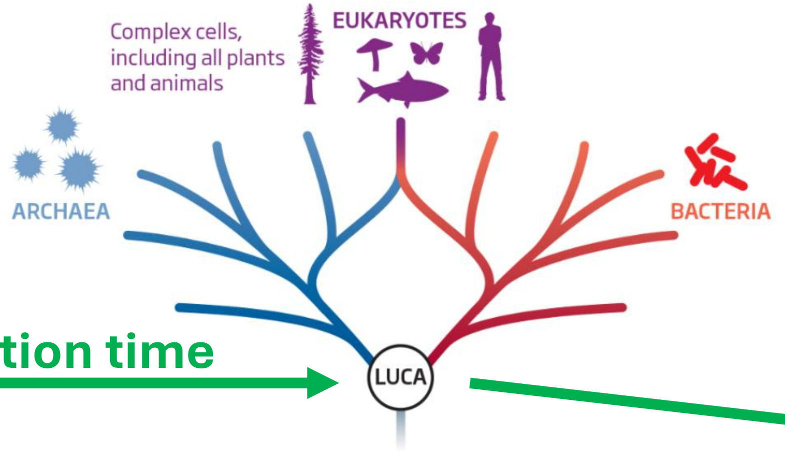
LSU of tRNA (E. coli) :
Blue part is the oldest one

the OLDEST functional part: always the same!



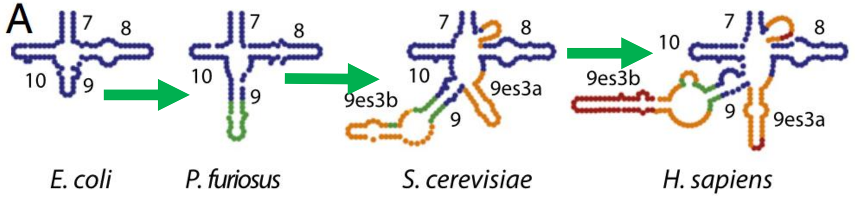
Secondary structure of the pseudosymmetrical region (**SymR**; *Agmon et al., 2005*), derived from the LSU secondary structure of Thermus thermophilus (*Petrov et al., 2013*). (*Madhan R. Tirumalai et al., 2021*)

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)

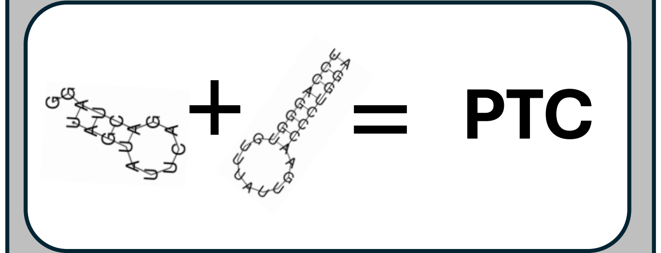
Peptidyl Transferase Center (PTC) is the oldest part of ribosomes.

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

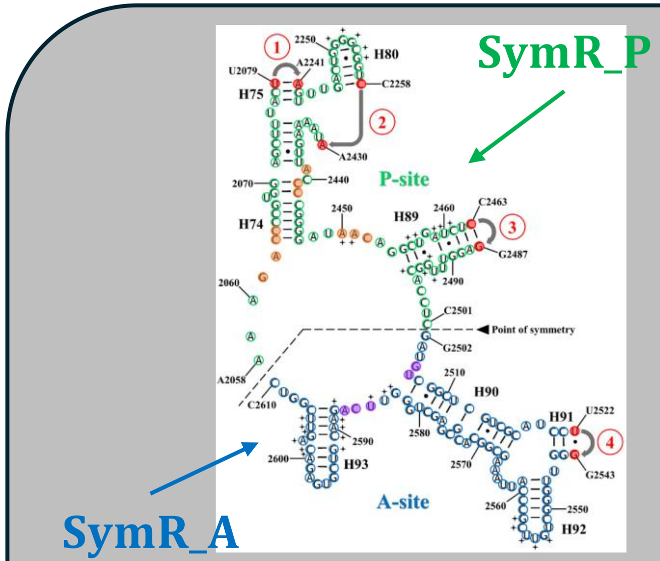
the dimerization of two similar RNA structures



“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

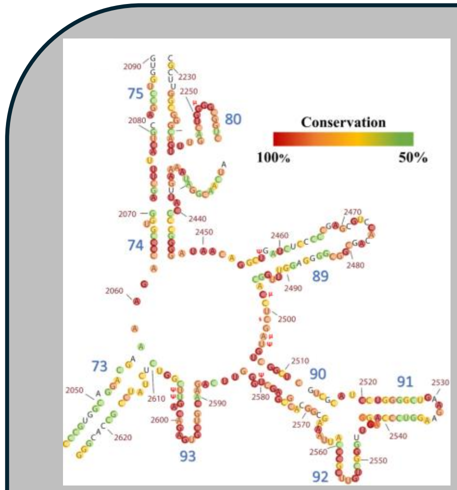
pseudosymmetrical region



SymR_PA = SymR_P + 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)

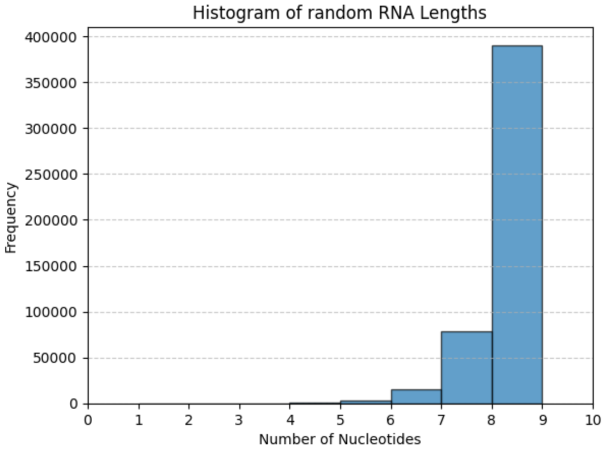
SymR_P is older than **SymR_PA**
PTC2 is older than PTC3, PTC4, PTC5

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

BASIS I

488280 random RNAs

With length <=8ncl



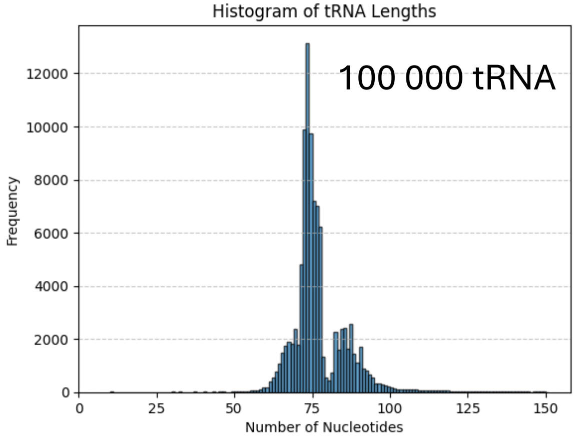
Peptidyl Transferase Center (**PTC**) is the oldest part of ribosomes.
PTC2: 549nc = 78nc(defined) + 471nc(undefined)
PTC3: 564nc = 146nc(defined) + 418nc(undefined)
PTC4: 583nc = 198nc(defined) + 385nc(undefined)
PTC5: 583nc = 247nc(defined) + 336nc(undefined)

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

BASIS II

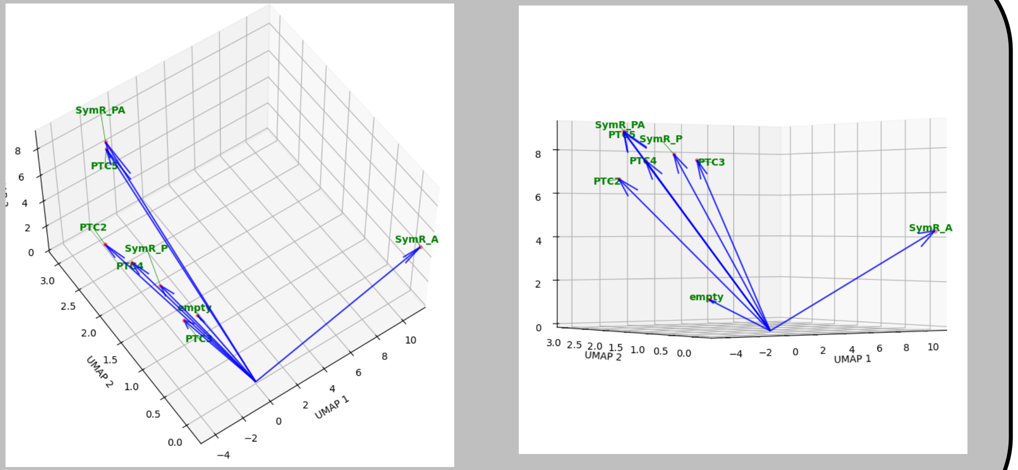
100 000 tRNAs

50ncl <= length <= 125ncl



UMAP dimensionality reduction was applied to obtain this result

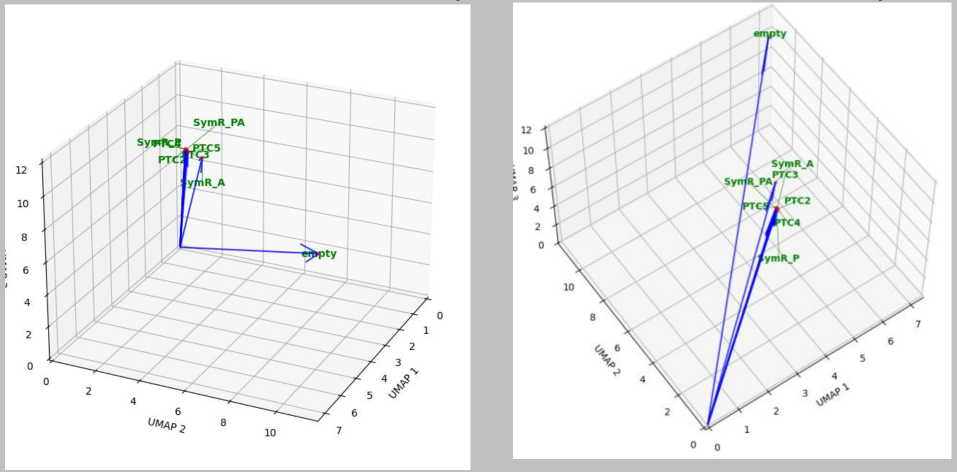
Visualization of the same 3D plot from two distinct viewpoints



SymR_A $\perp$ all

PTC5 = SymR_PA

Visualization of the same 3D plot from two distinct viewpoints

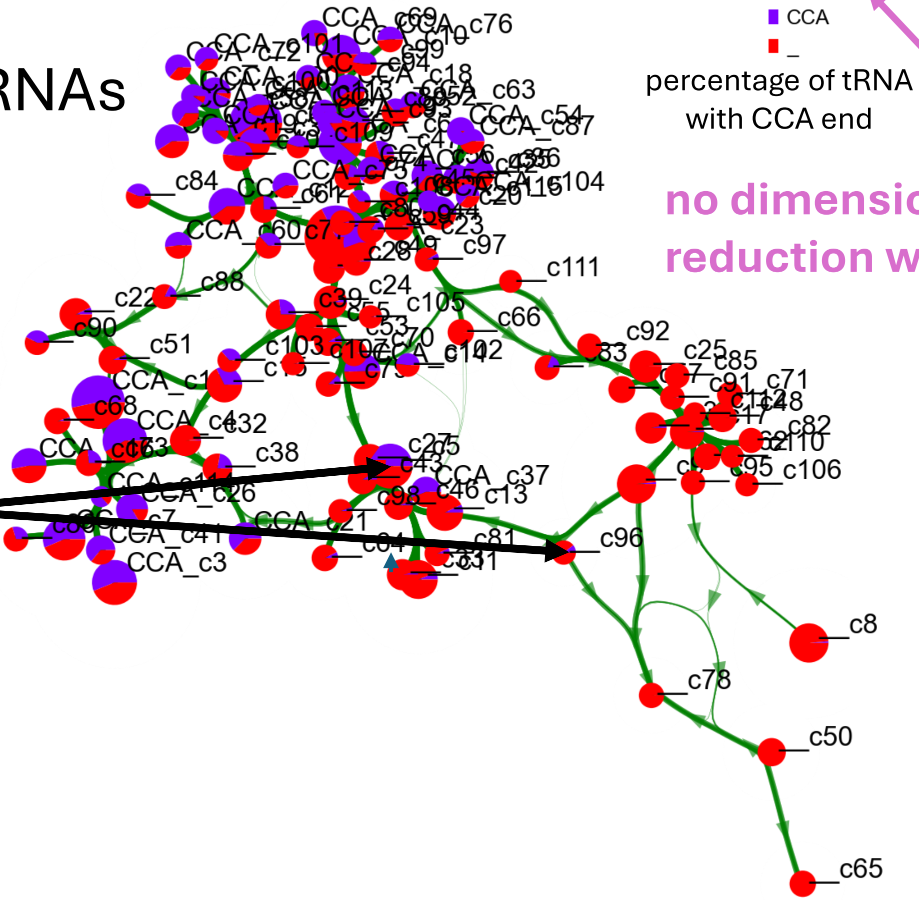


SymR_A $\neq$ (PTC2 = PTC3 = PTC4 = PTC5 = SymR_PA = SymR_P)

Single-Cell Inspired Analysis of tRNA and oldest rRNA

100 000 tRNAs

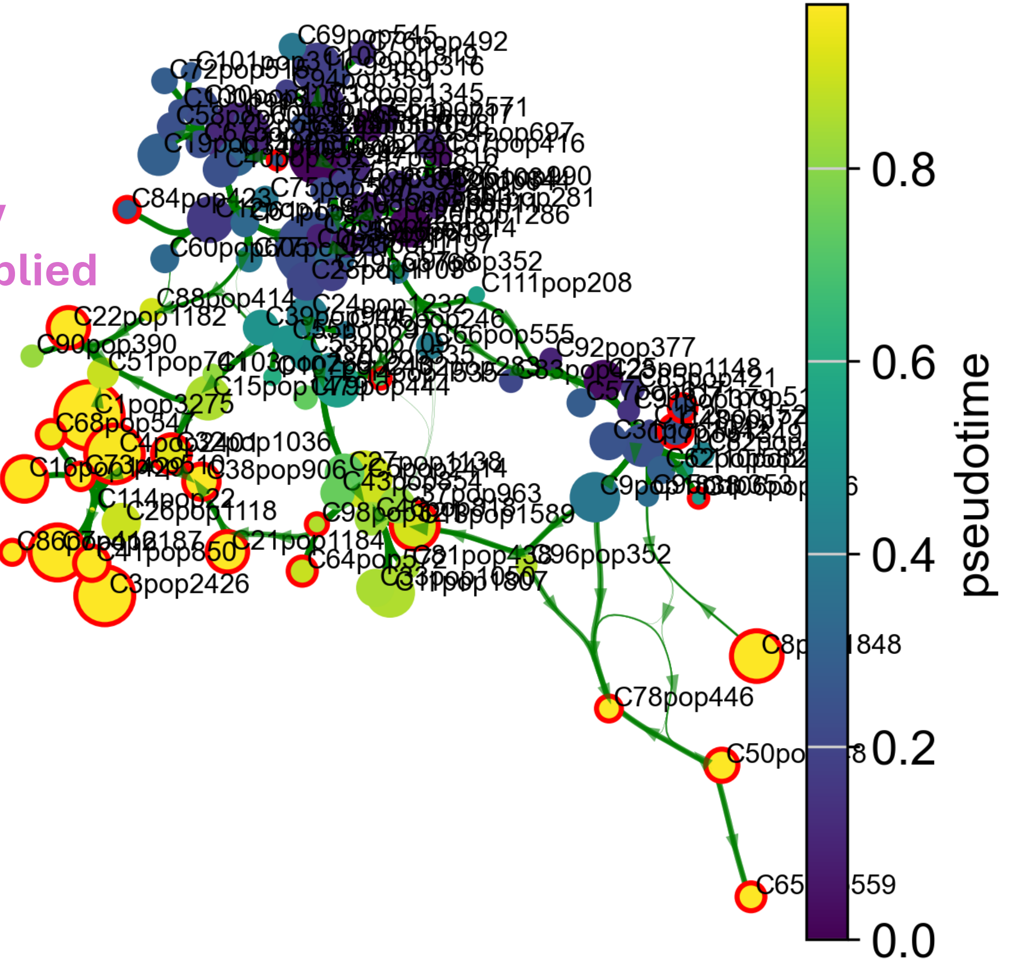
Cluster Composition. K=200. ncomp = 640



PTC2
PTC3
PTC4
PTC5
SymR_P
SymR_PA
SymR_A

no dimensionality reduction was applied

Pseudotime

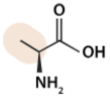


Peptidyl transferase center was detected in the 5th and 96th clusters.

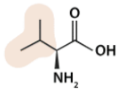
Method used from : Generalized and scalable trajectory inference in single-cell omics data with VIA, Stassen et al., Nature Communications, 2021

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

two of the simplest and oldest amino acids.

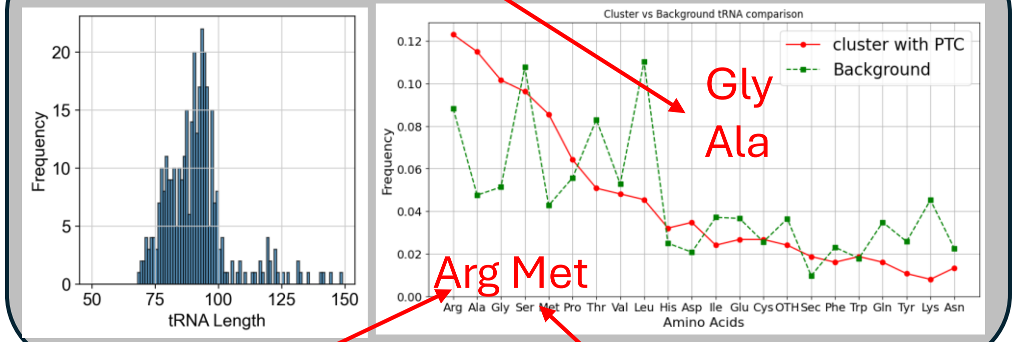


Alanine



Valine

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

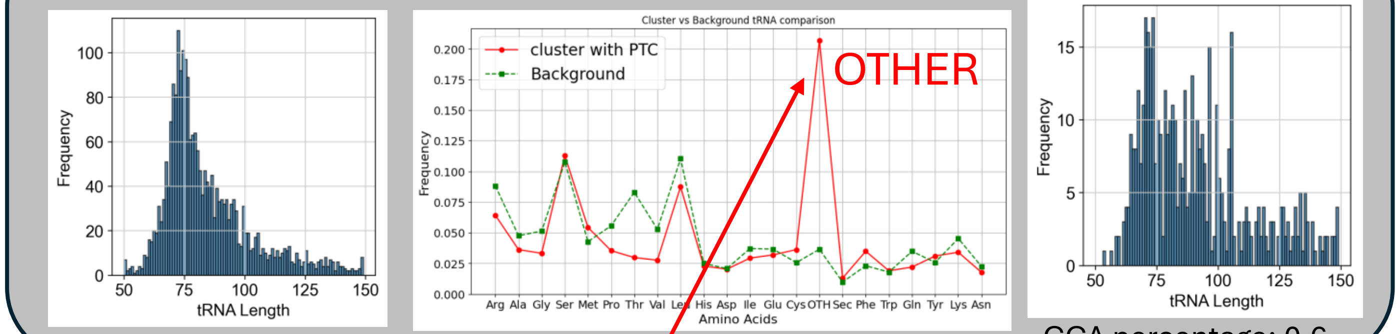


essential for stabilizing the negatively charged RNA phosphate backbone

encoded by the universal start codon

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

OTHER (OTH)
aver length: 90.84nc



CCA percentage: 0.6

other tRNA and pre-tRNA

no dimensionality reduction was applied to obtain this result

Four standard amino acids tRNA were detected above the background

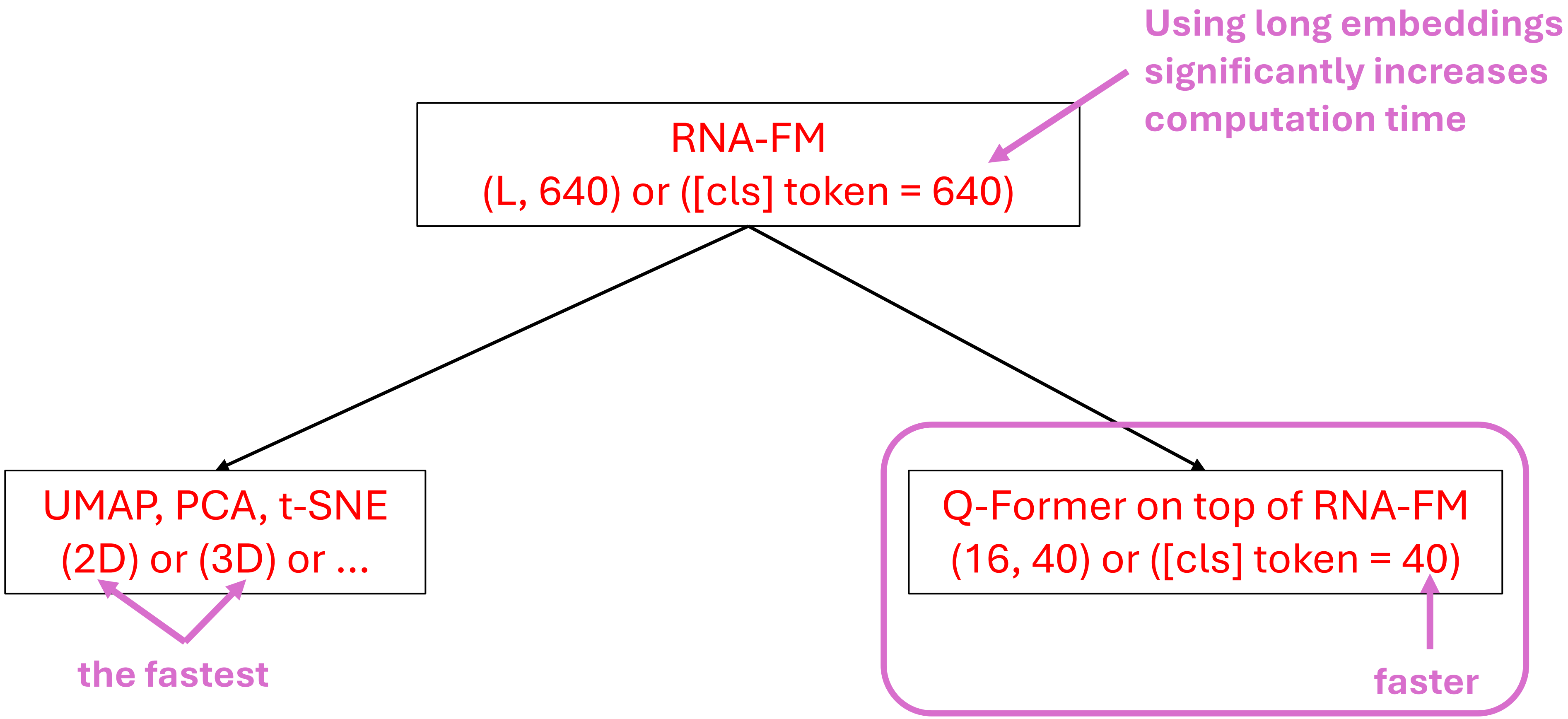
No standard amino acids tRNA were detected above the background

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

What is the correct way to use dimensionality reduction for comparing high-dimensional embeddings?



Using long embeddings significantly increases computation time

RNA-FM
(L, 640) or ([cls] token = 640)

UMAP, PCA, t-SNE
(2D) or (3D) or …

the fastest

Q-Former on top of RNA-FM
(16, 40) or ([cls] token = 40)

faster

Latent Diffusion Model for Controlable RNA Sequence generation

Latent Diffusion Models for Controllable RNA Sequence Generation

Kaixuan Huang[1]* Yukang Yang[1]* Kaidi Fu[2]♮ Yanyi Chu[3] Le Cong[3] Mengdi Wang[1]†
[1]Princeton University [2]Tsinghua University [3]Stanford University

Abstract

This work presents RNAdiffusion , a latent diffusion model for generating and optimizing discrete RNA sequences of variable lengths. RNA is a key intermediary between DNA and protein, exhibiting high sequence diversity and complex three-dimensional structures to support a wide range of functions. We utilize pretrained BERT-type models to encode raw RNA sequences into token-level, biologically meaningful representations. A Query Transformer is employed to compress such representations into a set of fixed-length latent vectors, with an autoregressive decoder trained to reconstruct RNA sequences from these latent variables. We then develop a continuous diffusion model within this latent space. To enable optimization, we integrate the gradients of reward models—surrogates for RNA functional properties—into the backward diffusion process, thereby generating RNAs with high reward scores. Empirical results confirm that RNAdiffusion generates non-coding RNAs that align with natural distributions across various biological metrics. Further, we fine-tune the diffusion model on mRNA 5' untranslated regions (5'-UTRs) and optimize sequences for high translation efficiencies. Our guided diffusion model effectively generates diverse 5'-UTRs with high Mean Ribosome Loading (MRL) and Translation Efficiency (TE), outperforming baselines in balancing rewards and structural stability *trade-off*. Our findings hold potential for advancing RNA sequence-function research and therapeutic RNA design.

1 Introduction

Diffusion models demonstrate exceptional performances in modelling continuous data, with applications in images synthesis [98, 100, 16], point clouds generation [92], video synthesis [58], reinforcement learning [3, 62, 79], time series [112] and molecule structure generation [122]. An important advantage of diffusion models is that their generation process can be "controlled" to achieve specific objectives via incorporating additional *guidance* signal. The guidance can steer the backward process toward generating samples with desired properties, without additional training [34, 17, 31].

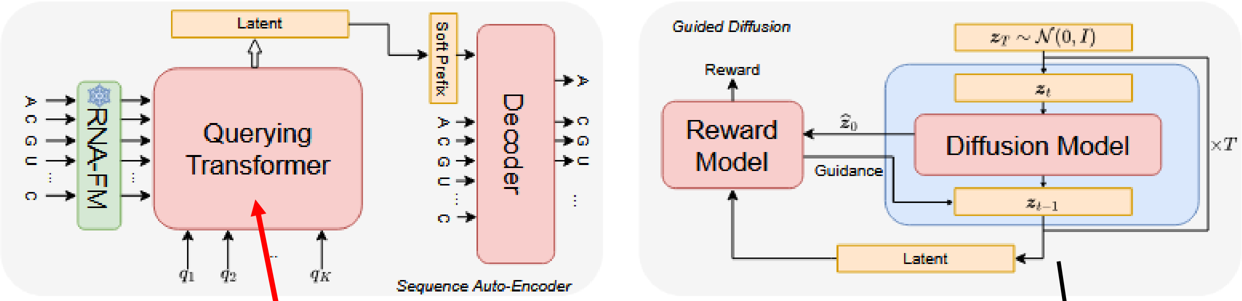


Figure 1: RNAdiffusion : Latent diffusion model for RNA sequences. Three parts of RNAdiffusion : (1) RNA **sequence auto-encoder**, consisting of a pretrained RNA-FM model, a Querying Transformer, and a decoder, for translating between the sequence space and the latent space; (2) **Guided diffusion** model with a pre-trained score network, for generating latent RNA embeddings under external guidance; (3) **Latent reward model**, trained on the latent space to predict functional properties of RNA, for computing guidance of diffusion.

Q-Former reduces the embedding size from (L, 640) in RNA-FM to a fixed (16, 40), eliminating dependence on sequence length L

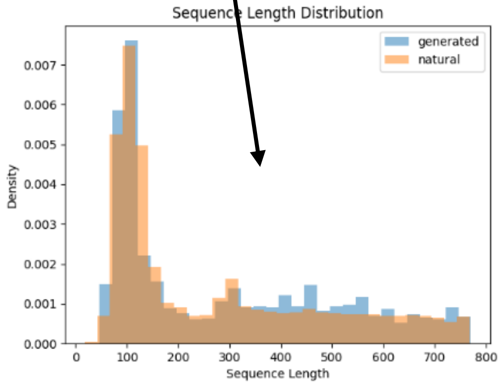


Figure 2: **Sequence length comparison** between the natural **ncRNA** test set and generated sequences (sample size: 20000).

Word embeddings arithmetic (using Q-Former)

Q-Former combined with parallel vector search was employed as an alternative to exact equation-based methods with standard dimensionality reduction.

ENGLISH language

Laptop ≈ Computer + Portable
Smartphone ≈ Phone + Smart

Fast + More ≈ Faster
Happy + Not ≈ Sad

King – Man + Woman ≈ Queen
Paris – France + Italy ≈ Rome

RNA language

Myotis brandtii mir-9229 microRNA precursor family

pre_miRNA(76nt) || pre_miRNA(76nt) + pre_miRNA(76nt)

--> Myotis brandtii mir-9229 microRNA precursor family (URS00027E3601)
embedding: [-4.48729768e-02 -6.57163262e-02 4.18678485e-03 -5.97885512e-02
 -7.67776966e-02 ...
sequence: GAGACACCCUUAUGGGGCAAGACUUGCUUCAGUGGGGCUUUGGUGCUCAUUGAGUCUUCCCCCUGAGUGUGUCCCU
paper: https://rnacentral.org/api/v1/rna/URS00027E3601/publications

--> Myotis brandtii mir-9229 microRNA precursor family (URS00027E23DB)
embedding: [-0.04290817 -0.06729076 0.0020279 -0.05250832 -0.07260264 ...
sequence: AAGACACCCUUAUGGGGCAAGACUUGCUUCAGUGGGGCUUUGGUGCUCACUGAGUCUUCCCCCUGAGUGUGUCUUU
paper: https://rnacentral.org/api/v1/rna/URS00027E23DB/publications

--> Myotis brandtii mir-9229 microRNA precursor family (URS00027E4E40)
embedding: [-0.04220863 -0.06273255 0.0013684 -0.06472154 -0.08135836 ...
sequence: GAGACACCCUUAUGGGACAGGAUGUGCUUCAGUGGGGCUUUGGUGCUCACUGAGUCUGCCCCCUGAGUGUGUCACU
paper: https://rnacentral.org/api/v1/rna/URS00027E4E40/publications

Image from:
https://fr.wikipedia.org/wiki/Myotis_brandtii

Word embeddings arithmetic for tRNA (using Q-Former)

Cordylochernes scorpioides tRNA-Gly (the same species)

```
tRNA (71nc), tRNA (71nc), tRNA (71nc)
--> Cordylochernes scorpioides tRNA-Gly (URS00026DEE46)
embedding: [-0.03151628 -0.11068475  0.00586255 -0.08975317 -0.01870812 ...
sequence: AGCAUCGGCCGGGAAUCGAACCCGGGCCGCCCGCGUGGCAGGCGAGCAUUCUACCACUGAACCACCGAUGA
paper: https://rnacentral.org/api/v1/rna/URS00026DEE46/publications
--> Cordylochernes scorpioides tRNA-Gly (URS00027A2828)
embedding: [-3.9467324e-02 -1.1325409e-01  3.7442837e-03 -8.5267112e-02
 -1.6314739e-02 ...
sequence: UGCAUCGGCCGGGAAUCGAACCGGGGCCGCCCGCGUGGCAGGCGAGAAUUCUACCAGUGAGCCACCGAUGC
paper: https://rnacentral.org/api/v1/rna/URS00027DF2F6/publications
--> Cordylochernes scorpioides tRNA-Gly (URS0002728A5D)
embedding: [-0.02521602 -0.11237185  0.00560259 -0.09093822 -0.02415926 ...
sequence: UGCAUCCGCCGGGAAUCGAACCCGGGCCGCCCACGUGGCAGGCGAGCAUUCUACCACUGAACCACCGAUGG
paper: https://rnacentral.org/api/v1/rna/URS00027DF2F6/publications
```

https://inaturalist-open-data.s3.amazonaws.com/photos/93668183/original.jpeg

Schistocerca tRNA-Thr (**different** species)

```
tRNA (75nc), tRNA (75nc), tRNA (75nc)
--> Schistocerca serialis cubense (Grasshoppers) transfer RNA threonine (anticodon UGU) (URS000282ACFD)
embedding: [-0.0268122  -0.04015758  0.00049825 -0.05642947 ...
sequence:
GCCCUCGGUGGCUCAGAUGGAUAGAGCGUCUGCCGUGUAAGCAGGACAUCCCGGGUUCGAGUCCCGGUCGGGGCA
paper: https://rnacentral.org/api/v1/rna/URS000282ACFD/publications
--> Schistocerca cancellata (South American locust) transfer RNA threonine (anticodon UGU) (URS000283E7A4)
embedding: [-0.03692201 -0.04039897 -0.00416766 -0.05146854 -0.07276659  ...
sequence:
GCCCGCGGUGGCUUAGAUGGACAGAGCGUCUGCCAUGUAAGCAGGAGAUCCCGGGUUCGAGUCCCGGUCGGGGCA
paper: https://rnacentral.org/api/v1/rna/URS00027DF2F6/publications
--> Schistocerca gregaria (Grasshoppers) transfer RNA threonine (anticodon UGU) (URS000282D22D)
embedding: [-0.02303507 -0.03340865 -0.00222736 -0.05441515 -0.07606529  ...
sequence:
GCCCUCGAUGGCUCAGUUGGAUAGAGCGCCUGCCAUGUAAGCAGGAGGUGCCGGGUUCGAGUCCCGGUCGGGGCA
paper: https://rnacentral.org/api/v1/rna/URS00027DF2F6/publications
```

https://tse1.mm.bing.net/th/id/OIP.NkSvBG2WVFgIUTVTX2U4twHaEo?r=0&rs=1&pid=ImgDetMain&o=7&rm=3

Word embeddings arithmetic for 5S rRNA (using Q-Former)

Helianthus annuus 5S ribosomal RNA

```
rRNA (119nc) || rRNA (119nc) + rRNA (119nc)
--> Helianthus annuus 5S ribosomal RNA (URS000266B49E)
embedding: [-0.0088359  -0.00282581 -0.00113988 -0.04848261 -0.10705304 ...
sequence: GGUUGCGAUCAUACCAGCACUAAUGCACCGGAUCCGAUCAGAACUCCGCAGUUAAGCGUGCUUGGGUGAGAGUAGUACUAGGAUGGGUGACCCCCUGGGAAGUCCUCGUGUUGCAACCC
paper: https://rnacentral.org/api/v1/rna/URS000266B49E/publications

--> Helianthus annuus 5S ribosomal RNA (URS0002658146)
embedding: [-0.0091403  -0.00924383 -0.00049806 -0.05139868 -0.10899518 ...
sequence: GGUUGCGAUCAUACCAGCACUAAUGCACCGGAUCCCAUCAGAACUCUACAGUUAAGCGUGUUUGGGCGAGAGUAGUACUAGGAUGGGUGACCCCCUGGGAAGUCCUCGUGUUGCAACCC
paper: https://rnacentral.org/api/v1/rna/URS0002658146/publications

--> Helianthus annuus 5S ribosomal RNA (URS000266C6B1)
embedding: [-0.00760111 -0.00236688 -0.00298203 -0.05139999 -0.10575107  ...
sequence: GGUUGCGAUCAUACAAGCACUAAUGCACCGGAUCCCAUCAGAACUCCGCAGUUAAGCGUGCUUGUGCGAGAGUAGUACUAGGAUGGGUGACCCCCUGGGAAGUCCUCGUGUUGCAACCC
paper: https://rnacentral.org/api/v1/rna/URS000266C6B1/publications
```

Word embeddings arithmetic for bacteria (using Q-Former)

```
tRNA (87nc) || tRNA (85nc) + tRNA (73nc)
--> Candidatus Eiseniibacteriota bacterium tRNA-Leu (URS00028CF2C0)
embedding: [ 2.6688760e-02  7.8196831e-02 -3.6661938e-02  ...
sequence: GCCCGAGUGGCGGAACUGGCAGACGCGCUAGAUUCAGGUUCUAGUGUUCGCAAGGACGUGGAGGUUCGAGUCCUCUCUCGGGCACCA
paper: https://rnacentral.org/api/v1/rna/URS00028CF2C0/publications
--> Deltaproteobacteria bacterium tRNA-Leu (URS00026BC59E)
embedding: [ 3.88834961e-02  7.53495395e-02 -5.43716773e-02  ...
sequence: GCCCAAGUGGCGGAACUGGCAGACGCGCUAGAUUCAGGUUCUAGUGGGCUAAUCCCCCGUGGAAGUUCGAGUCUUCUCUUGGGCA
paper: https://rnacentral.org/api/v1/rna/URS00026BC59E/publications
--> Gaiellaceae bacterium tRNA-Thr (URS00028A2E7F)
embedding: [ 0.02313009  0.07265869 -0.0230953  ...
sequence: GCCGGAGUAGCUCAGCUGGUAGAGCAGCUGAUUUGUAAUCAGCAGGUCGUGGGUUCGAGUCCCUCCUCCGGCU
paper: https://rnacentral.org/api/v1/rna/URS00028A2E7F/publications
2.16
```

```
tRNA (75nc) || tRNA (77nc) + tRNA (87nc)
--> Candidatus Cloacimonadota bacterium tRNA-Gln (URS00028E2231)
embedding: [ 0.00905947  0.08511419 -0.03471698  ...
sequence: UGGGCAGUCGCCAAGUGGUAAGGCAGCAGGUUUUGGUCCUGCCAUCCGGGGGUUCAAAUCCUCCCUGCCCAGCCA
paper: https://rnacentral.org/api/v1/rna/URS00028E2231/publications
--> Alphaproteobacteria bacterium tRNA-Pro (URS00028CFD83)
embedding: [ 0.00279988  0.06899755 -0.02794215  ...
sequence: CGGGCGGUGGCGCAGCCUGGUAGCGCACCAGACUGGGGGUCUGGGGGUCGCAGGUUCAAAUCCUGUCCGCCCGACCA
paper: https://rnacentral.org/api/v1/rna/URS00028CFD83/publications
--> Erysipelotrichales bacterium tRNA-Leu (URS00026BF45E)
embedding: [ 0.02560369  0.08812317 -0.04586785  ...
sequence: GCCCGGAUGACGAAAUUGGUAGACGUAGCAGACUCAAAAUCUGCCGGUGUCAAAGCCGUGCCGGUUCGAGUCCGGCUCCGGGCACCA
paper: https://rnacentral.org/api/v1/rna/URS00026BF45E/publications
2.27
```

Do these three bacteria coexist in the same environment?

Word embeddings arithmetic for bacteria and archaea

```
tRNA (75nc) || tRNA (77nc) + tRNA (92nc)
--> Candidatus Cloacimonadota bacterium tRNA-Gln (URS00028E2231)
embedding: [ 0.00905947  0.08511419 -0.03471698  ...
sequence: UGGGCAGUCGCCAAGUGGUAAGGCAGCAGGUUUUGGUCCUGCCAUCCGGGGGUUCAAAUCCUCCCUGCCCAGCCA
paper: https://rnacentral.org/api/v1/rna/URS00028E2231/publications
--> Arcobacteraceae bacterium tRNA-Met (URS0002879DE9)
embedding: [ 0.01143912  0.08863723 -0.02804913  ...
sequence: GUCAAGGUAGCUCAGCUGGUUAGAGCGCUGGUCUCAUAAGCCGGAGGUCGAGGGUUCGAGUCCCUCCCUUGACACCA
paper: https://rnacentral.org/api/v1/rna/URS0002879DE9/publications
--> Nitrososphaeraceae archaeon tRNA-OTHER (URS00028C35F0)
embedding: [ 0.01913258  0.09050473 -0.04348399  ...
sequence: AGCCCGGUAGAGAAUUAAAAACCGCUGAAUGUAGUGGCCAAGCAUAGAGGCCUUUGGAGCCUUUGACCCCAGUUCGAAUCUGGGCCGGGCUA
paper: https://rnacentral.org/api/v1/rna/URS00028C35F0/publications
 2.31
```

Word embeddings arithmetic for bacteria and fungi

```
tRNA (75nc) || tRNA (72nc) + tRNA (129nc)
--> Candidatus Cloacimonadota bacterium tRNA-Gln (URS00028E2231)
embedding: [ 0.00905947  0.08511419 -0.03471698  ...
sequence: UGGGCAGUCGCCAAGUGGUAAGGCAGCAGGUUUUGGUCCUGCCAUCCGGGGGUUCAAAUCCUCCCUGCCCAGCCA
paper: https://rnacentral.org/api/v1/rna/URS00028E2231/publications
--> Owenweeksia sp. TMED14 tRNA-Gln(ttg) (URS00026EF88A)
embedding: [ 0.00647415  0.08554987 -0.03140881  ...
sequence: UGCCCCAUCGUCUAAAGGCAGGACAGCGGUUUUUGGUACCGUCAGUCUAGGUUCGAGUCCUAGUGGGGCAAC
paper: https://rnacentral.org/api/v1/rna/URS00026EF88A/publications
--> Elasticomyces elasticus tRNA-Leu (URS00028A207F)
embedding: [ 0.01019155  0.08791753 -0.04669886  ...
sequence: GCCGGUUAUGGUGUAGUGGUAAGCAUACCCGCUUCAGCUUGUUGGUGAUUUCCAUCGAAGGAUUGAGUAAUCGAACCUUCGUGGAAUUGACUUCCGCGGGUGACCUAAGUUCGAUCCUUAGUGGCGGCG
paper: https://rnacentral.org/api/v1/rna/URS00028A207F/publications
2.51
```

Bacterial–fungal interactions: ecology, mechanisms and challenges, Deveau et al. 2018
https://academic.oup.com/femsre/article/42/3/335/4875924

To what extent can these organisms occupy overlapping ecological niches and coexist within the same habitat?

```
pre_miRNA (76nc) || pre_miRNA (88nc) + pre_miRNA (76nc)

--> Myotis lucifugus (little brown bat) mir-9229 microRNA precursor family (URS00027DFAA8)
embedding: [ 0.03975813  0.08132407 -0.08302216  ...
sequence: AGACACCCUUGUGGGGCAAGACUUGCUUCAGUGGGGGCAUUGGUGCUCAAUGAGUCUGCCCCCUGAGUGUGUCCCU
paper: https://rnacentral.org/api/v1/rna/URS00027DFAA8/publications

--> Musca domestica (house fly) microRNA mir-67 (URS00027DF700)
embedding: [ 0.0321277   0.07590463 -0.0834218 ...
sequence: UCUUGCUUUGACUCACUCAACCUGGGUGUGAUGUGUGUAUUUCGUUUUGGCUAUCCAUCACAACCUCCUUGAGUGAGCGAUAGCAGGA
paper: https://rnacentral.org/api/v1/rna/URS00027DF700/publications

--> Myotis brandtii mir-9229 microRNA precursor family (URS00027DCC00)
embedding: [ 0.04045758  0.07466664 -0.08986323  ...
sequence: GAGACACCCCUCUGGGGCUAGACUUGCUUCGAUGGGGCUUUGGUGCUCACUGAGUCUGCCCCCUGAGUGUGUCCUU
paper: https://rnacentral.org/api/v1/rna/URS00027DCC00/publications
1.96 degrees
```

CONCLUSIONS

- The vanilla transformer with custom embeddings and a masked training paradigm allows for deeper hidden representations than a BERT-like transformer

- A preliminary interpretation of simple additive relationships among different ncRNAs suggests they may reflect coordinated roles in the coexistence of Archaea and Bacteria across diverse environments.

- 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, with UMAP employed for dimensionality reduction. 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, without dimensionality reduction.

- Using Q-Former may be more appropriate than UMAP, PCA, or t-SNE for dimensionality reduction when precisely comparing high-dimensional embeddings

PTC - Peptidyl Transferase Center of Ribosome

What’s Next?

1. **Refinement of RNA similarity measure:** Instead of relying on a single scalar value (such as L2 distance or cosine similarity), one can analyze the outputs of individual attention heads prior to their aggregation into the final embedding. This 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.