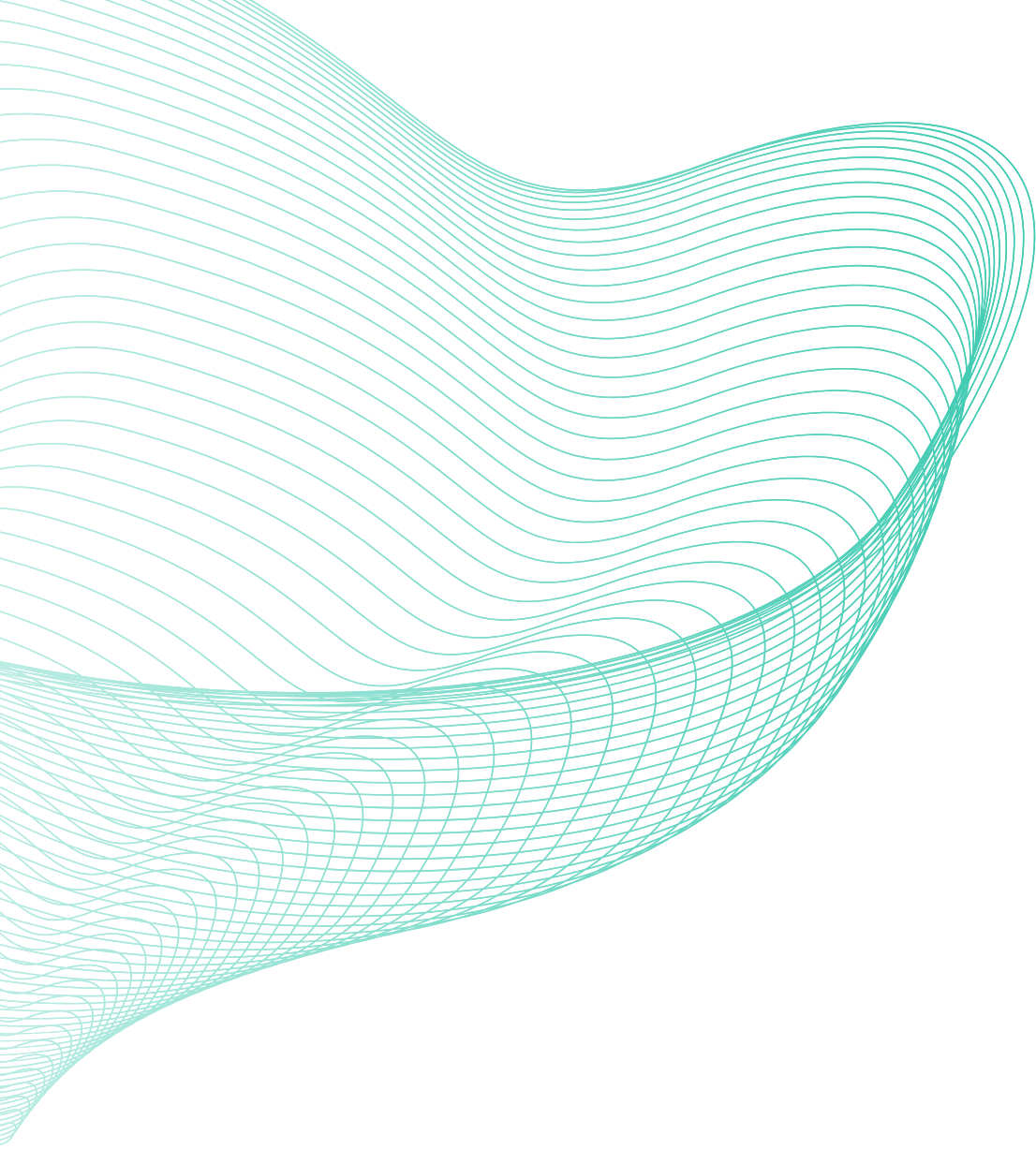# RNAgrail: Graph neural network and diffusion model for RNA 3D structure prediction

Marek Justyna, Craig Zirbel, Maciej Antczak, Marta Szachniuk

Agenda

RNA vs protein

Domain background.

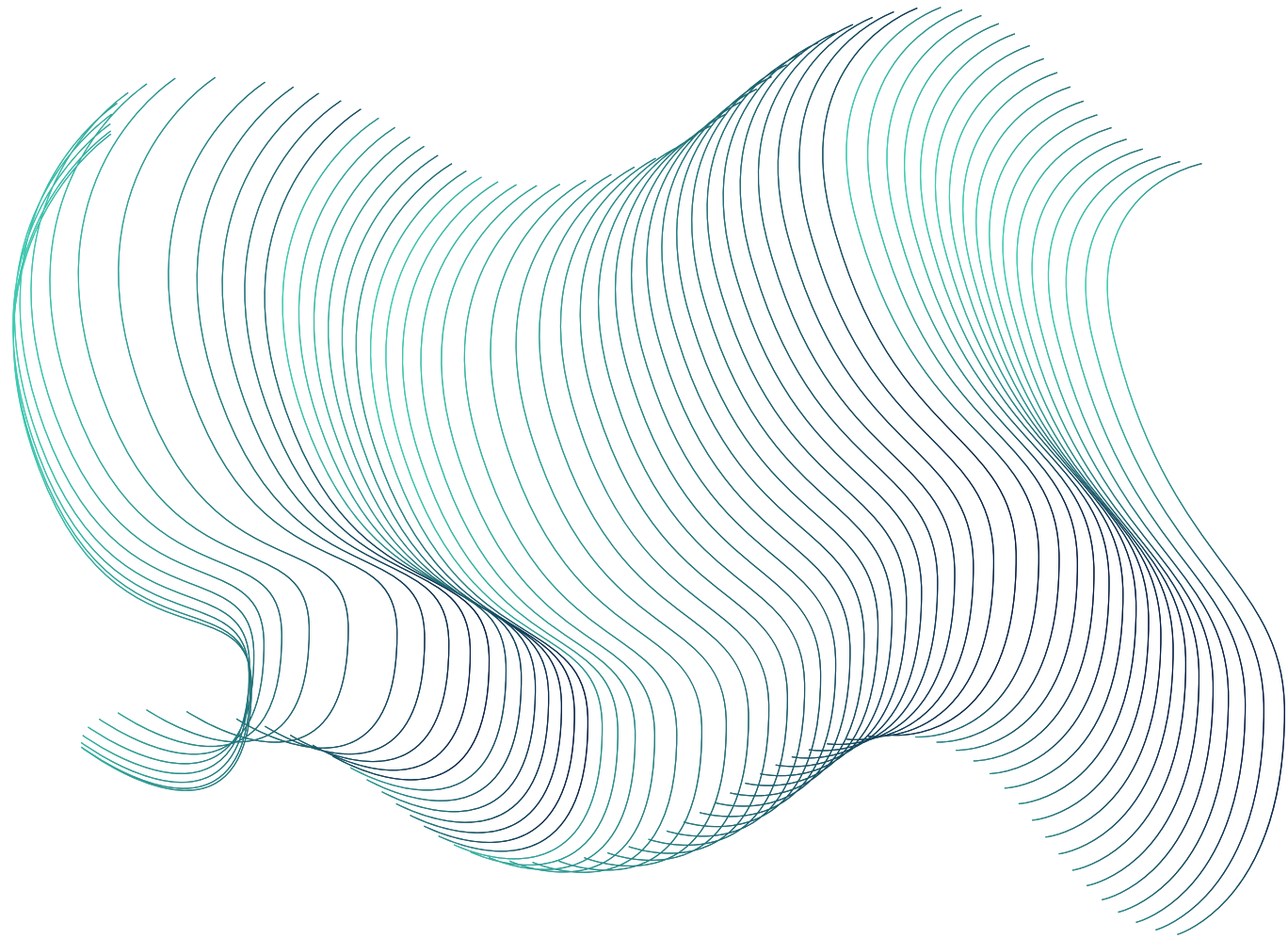**Challenges in RNA structure prediction**

If you can predict protein structure, then why is it so difficult to predict RNA?

Our approach

Generative Graph Neural Network for RNA structure prediction.

Results & Summary

Comparison of our method to AlphaFold 3.
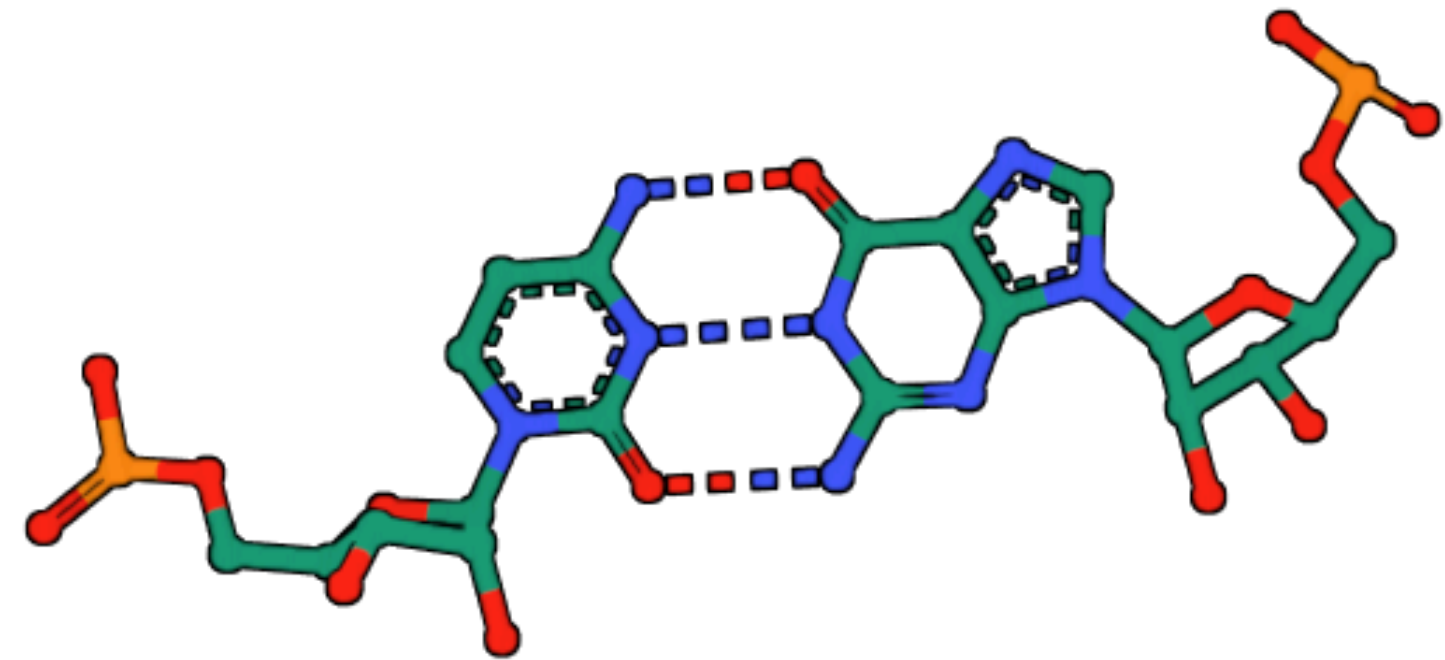
RNA vs protein.

RNA vs protein

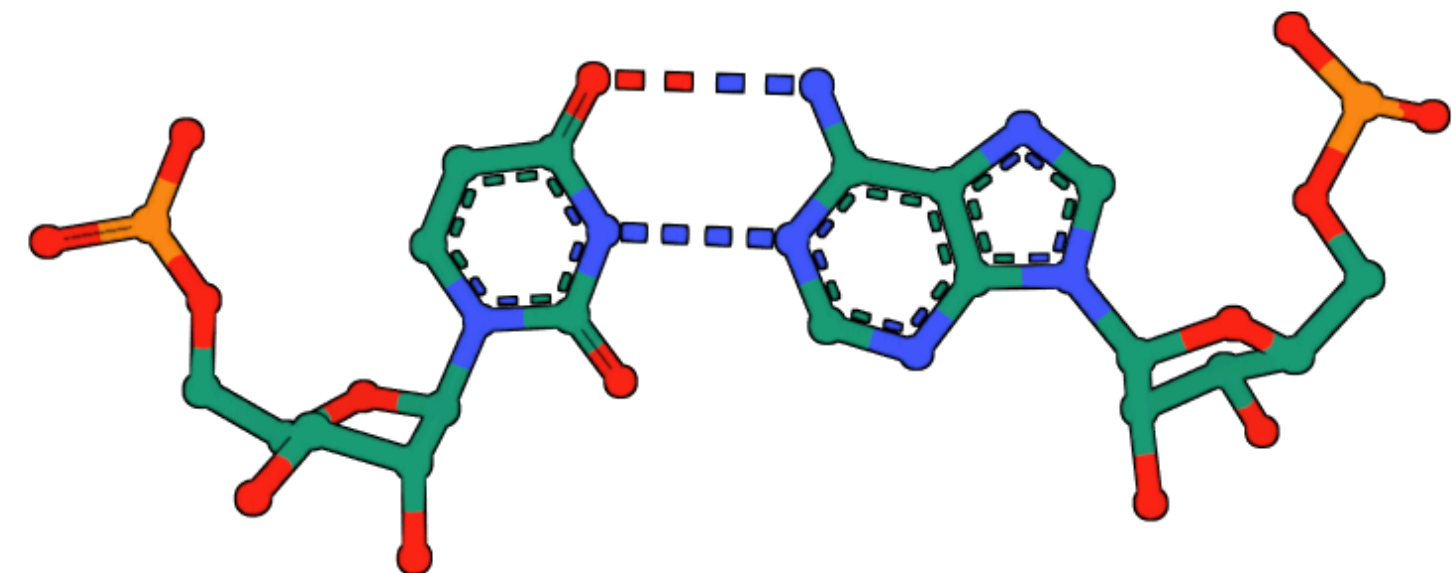4 nucleotides in RNA

VS

20 aminoacids in protein

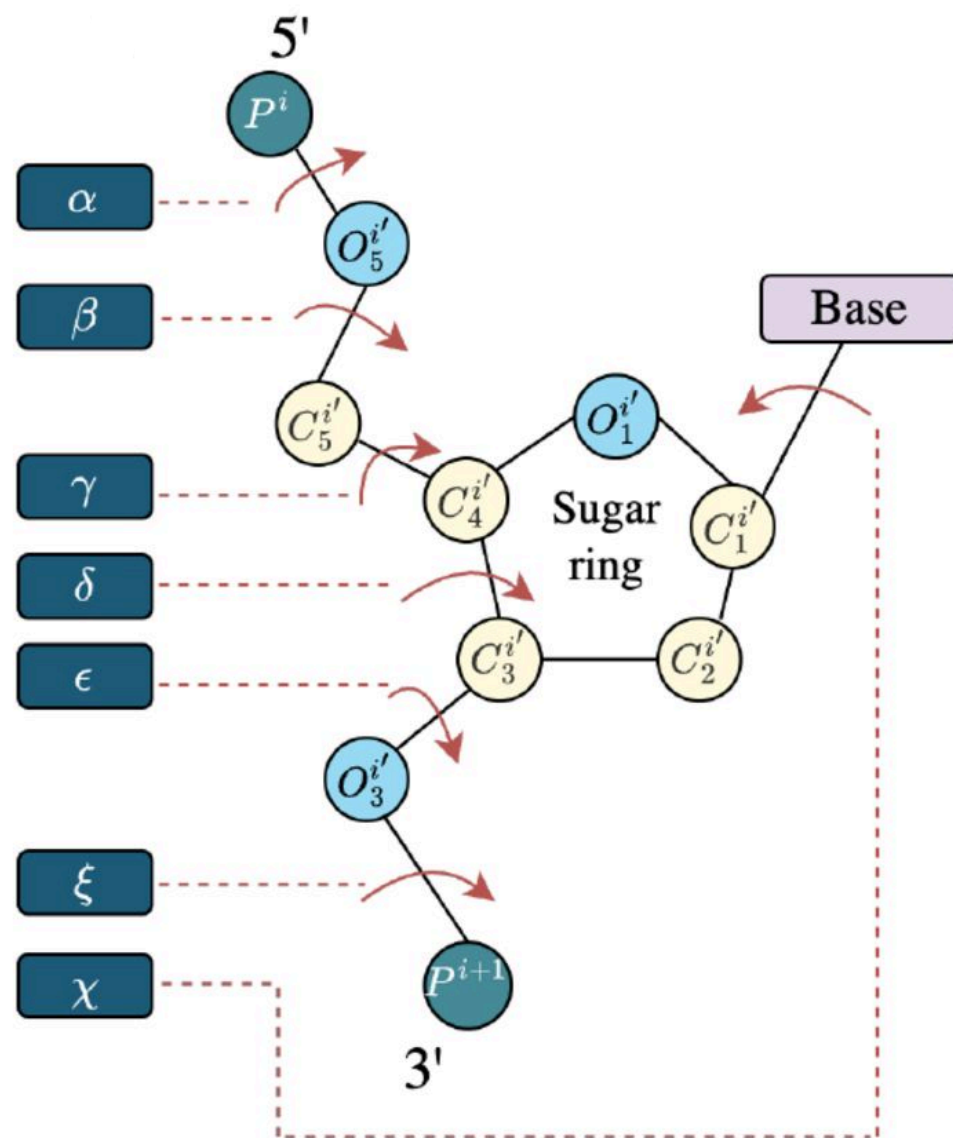Cytosine   Guanine

Uracil   Adenine
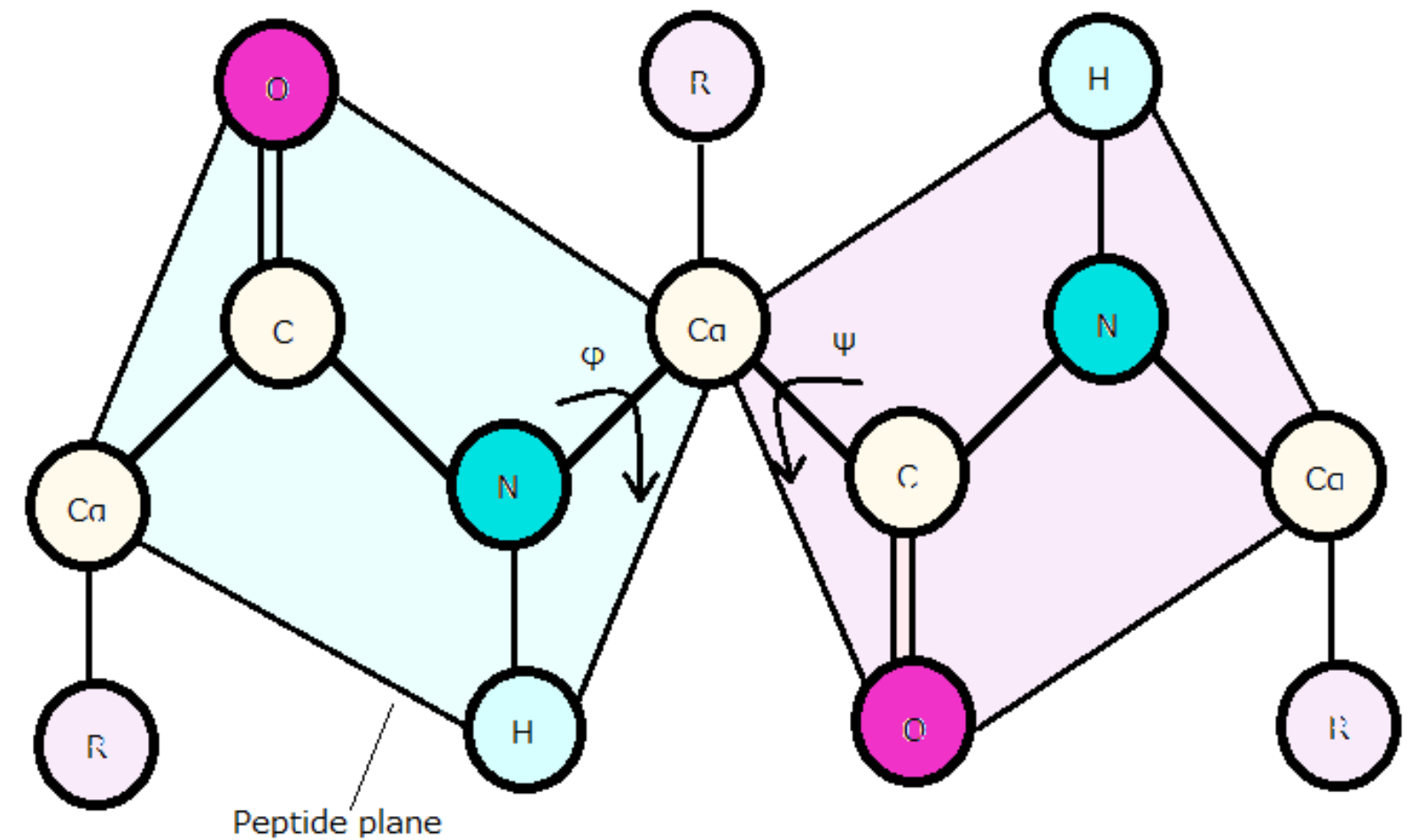
7 degrees of freedom

7 torsion angles in RNA

2 torsion angles in protein

Bernard, C., Postic, G., Ghannay, S., & Tahi, F. (2024). RNA-TorsionBERT: leveraging language models for RNA 3D torsion angles prediction.

http://cib.cf.ocha.ac.jp/bitool/DIHED2/

AlphaFold2 architecture

Jumper *et al*. (2021) Highly accurate protein structure prediction with AlphaFold.
Nature 596, 583–589 (doi: 10.1038/s41586-021-03819-2)
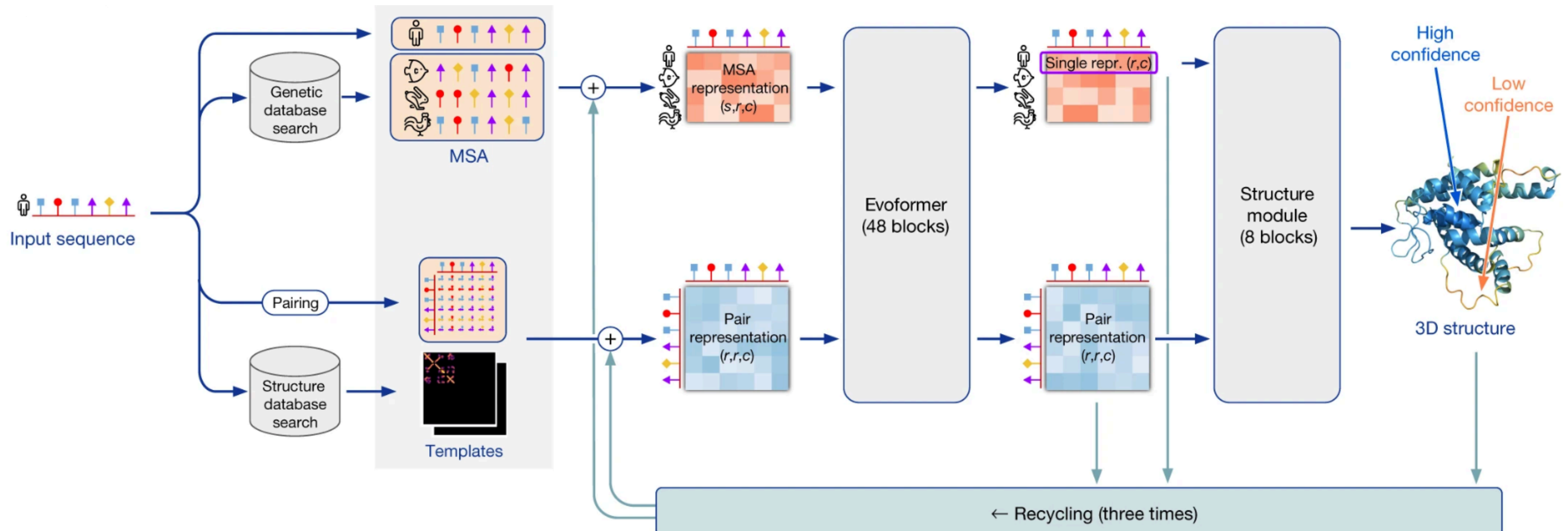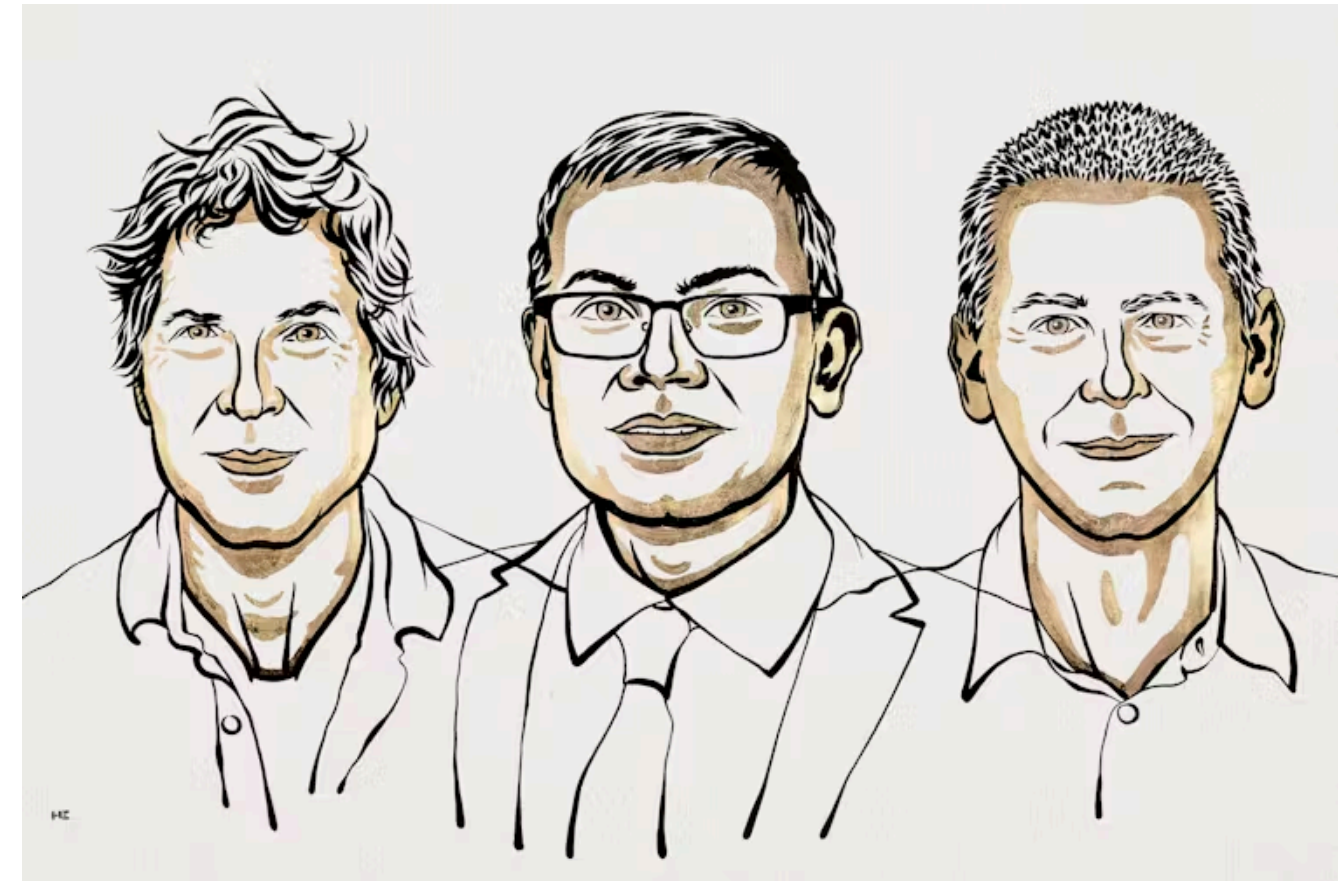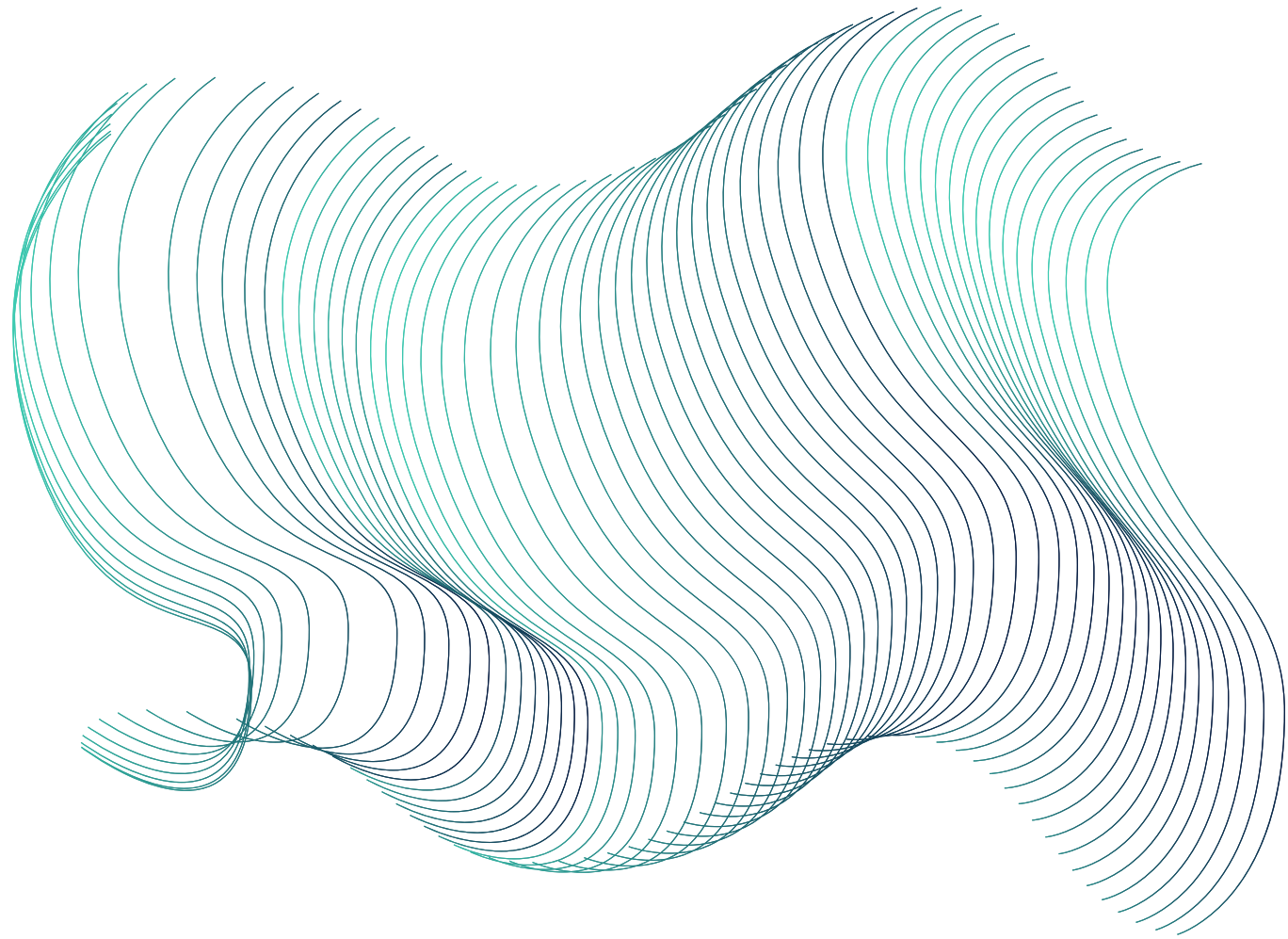
Nobel prizes in 2024

Victor Ambros and Gary Ruvkun for their discovery of microRNA (miRNA) and its function in the post-transcriptional regulation of genes.
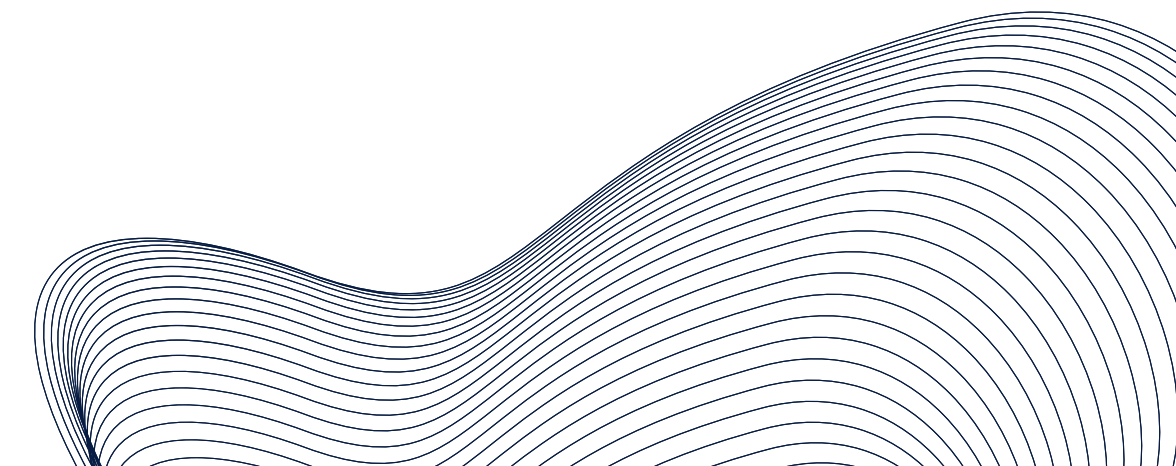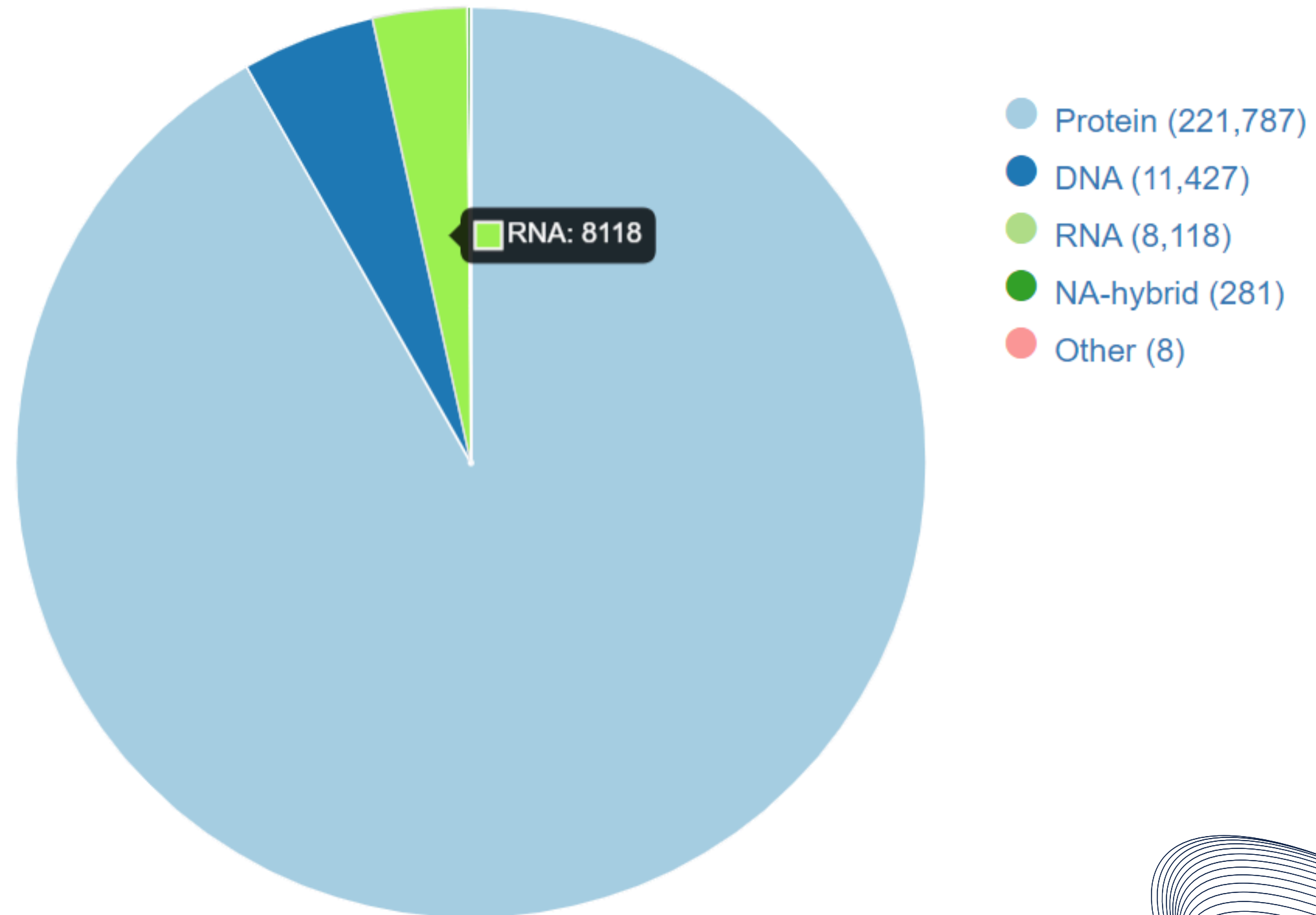

David Baker for for computational protein design, Demis Hassabis and John M. Jumper of Google DeepMind for protein structure prediction.

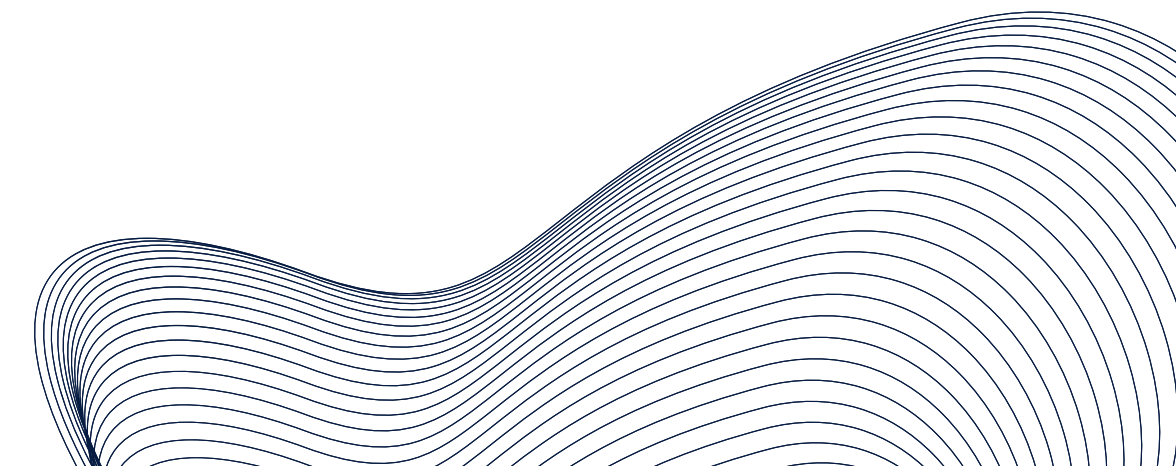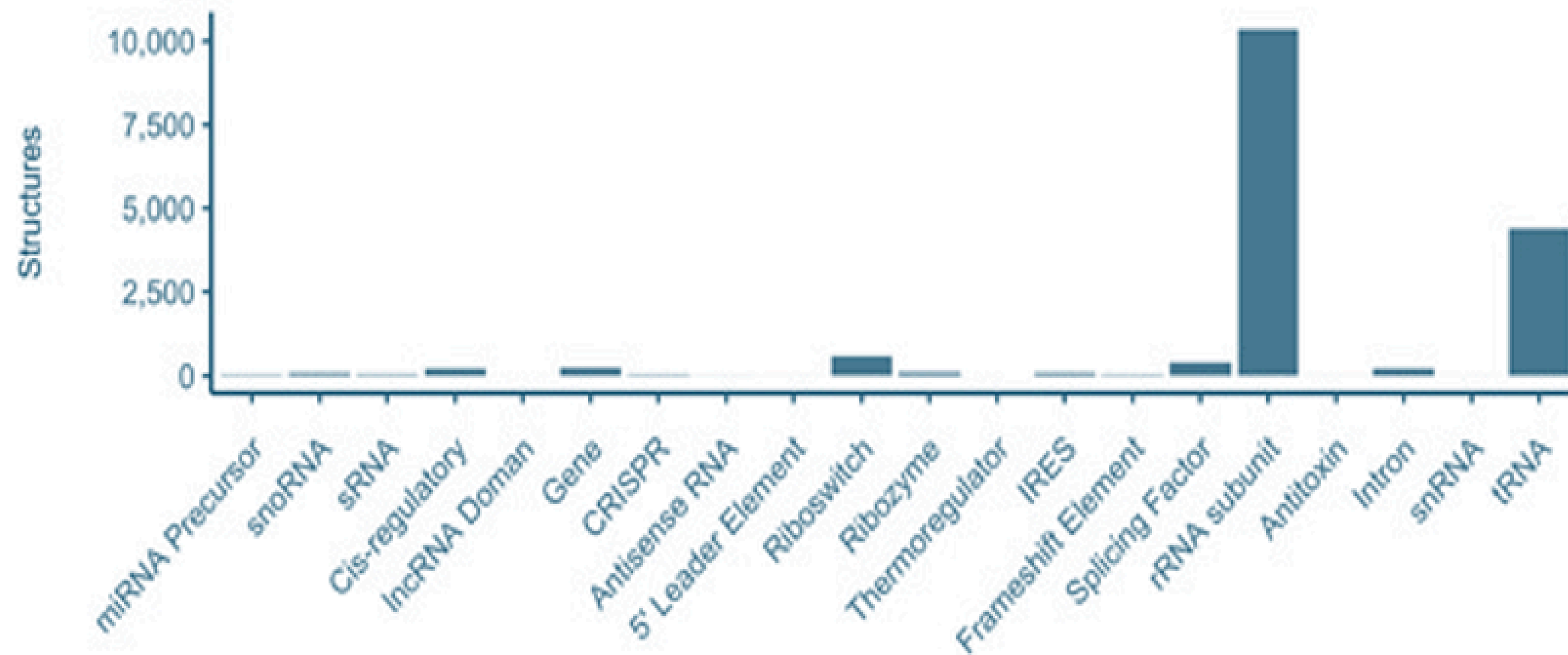# Challenges in RNA 3D structure prediction

Why is it so difficult to predict RNA 3D structure?

Legend:
- Protein (221,787)
- DNA (11,427)
- RNA (8,118)
- NA-hybrid (281)
- Other (8)

RNA: 8118

Why is it so difficult to predict RNA 3D structure?

Challenges

The key challenges which hinder the researchers from developing the AlphaFold for RNA.

Schneider *et al*. (2023) **When will RNA get its AlphaFold moment?**, *Nucleic Acids Research* 51, 9522-9532 (doi: 10.1093/nar/gkad726)

RNA content in the Protein Data Bank

The number of high-resolution experimentally determined protein structures is about 100 more abundant than for RNAs.

Data imbalance

Almost 80% of structures belong to rRNA and tRNA. This disproportion introduces huge bias in the dataset.

Quality of experimental RNA data

Many of the deposited experimental structures of RNA have base pairs incorrectly classified, inconsistences in valence, and backbone geometry.
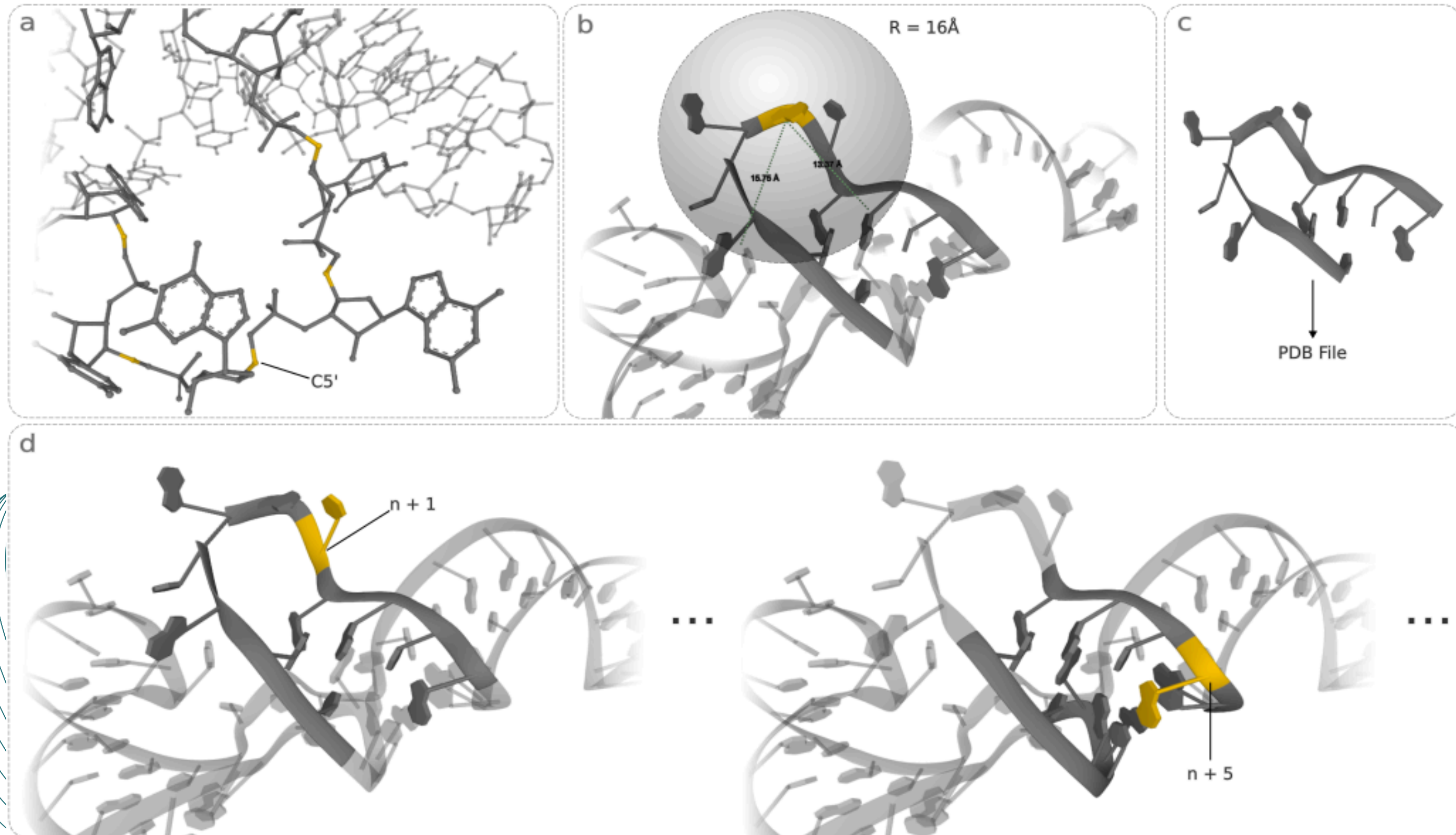
Sequences and sequence alignments

MSA is a powerful strategy for structure prediction (both RNA and protein). Creating high-quality RNA alignments is difficult and often requires the manual work of an expert.
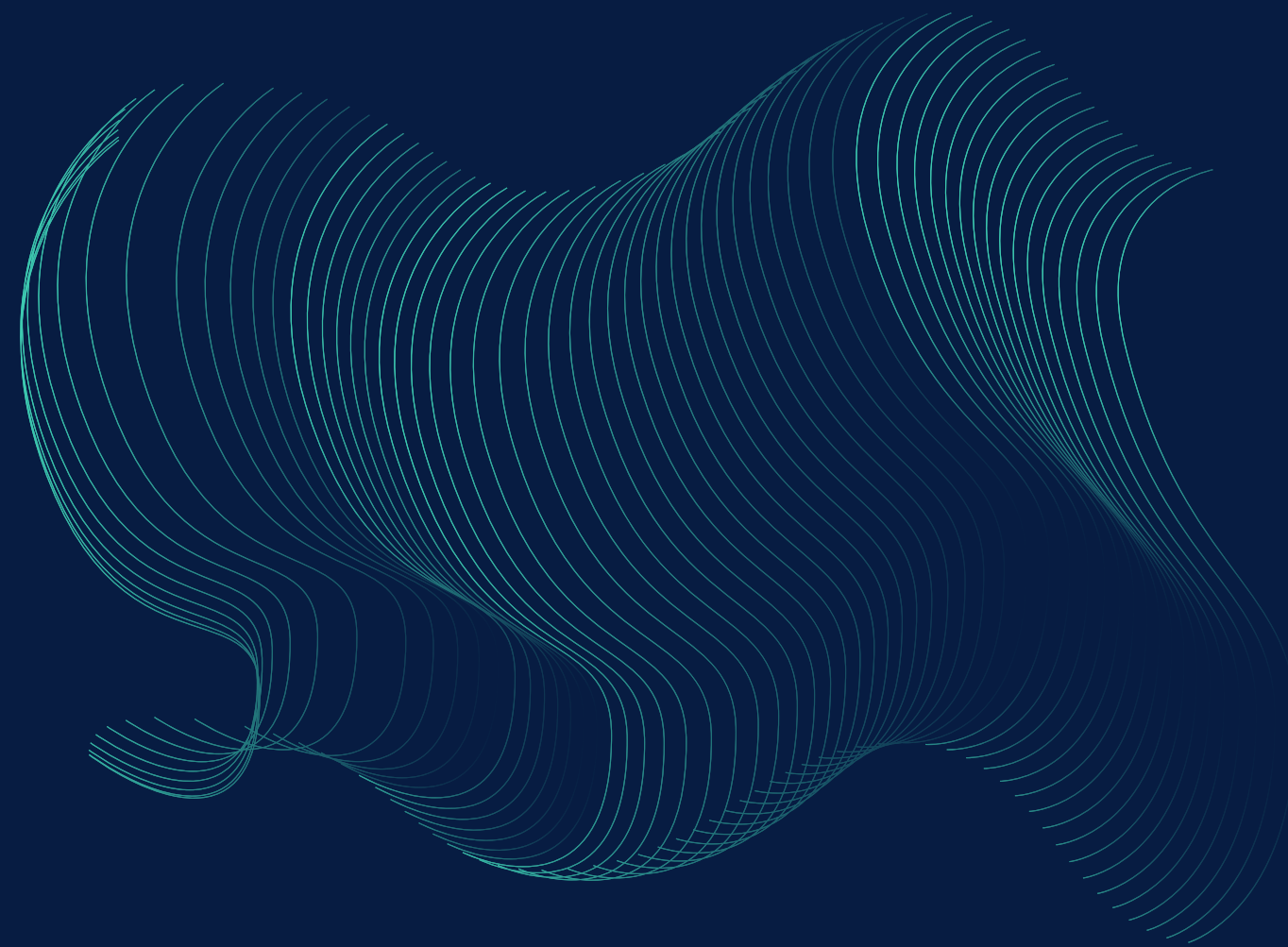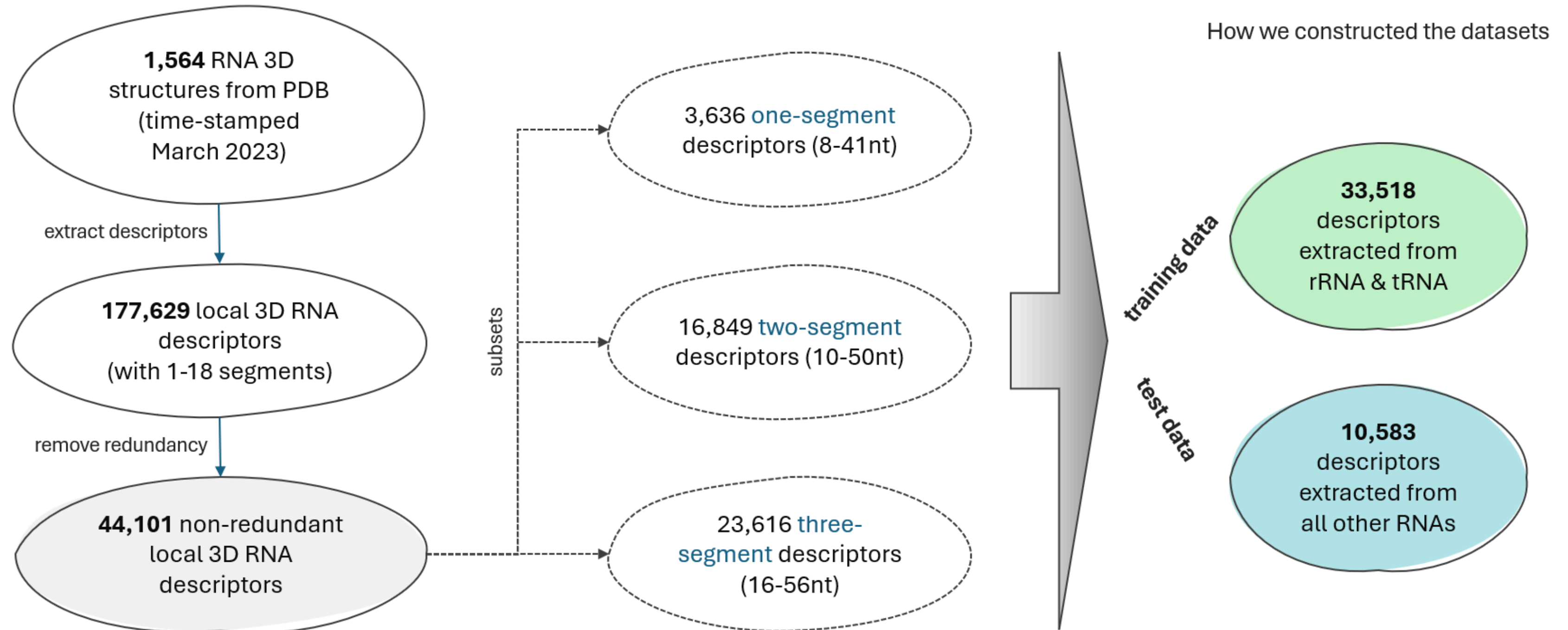
Our approach

Generative Model for RNA 3D structure prediction

Local 3D RNA descriptors

Local 3D RNA descriptors (2)

How we constructed the datasets

1,564 RNA 3D structures from PDB (time-stamped March 2023)

extract descriptors

177,629 local 3D RNA descriptors (with 1-18 segments)

remove redundancy

44,101 non-redundant local 3D RNA descriptors

subsets

3,636 one-segment descriptors (8-41nt)

16,849 two-segment descriptors (10-50nt)

23,616 three-segment descriptors (16-56nt)

training data

33,518 descriptors extracted from rRNA & tRNA

test data

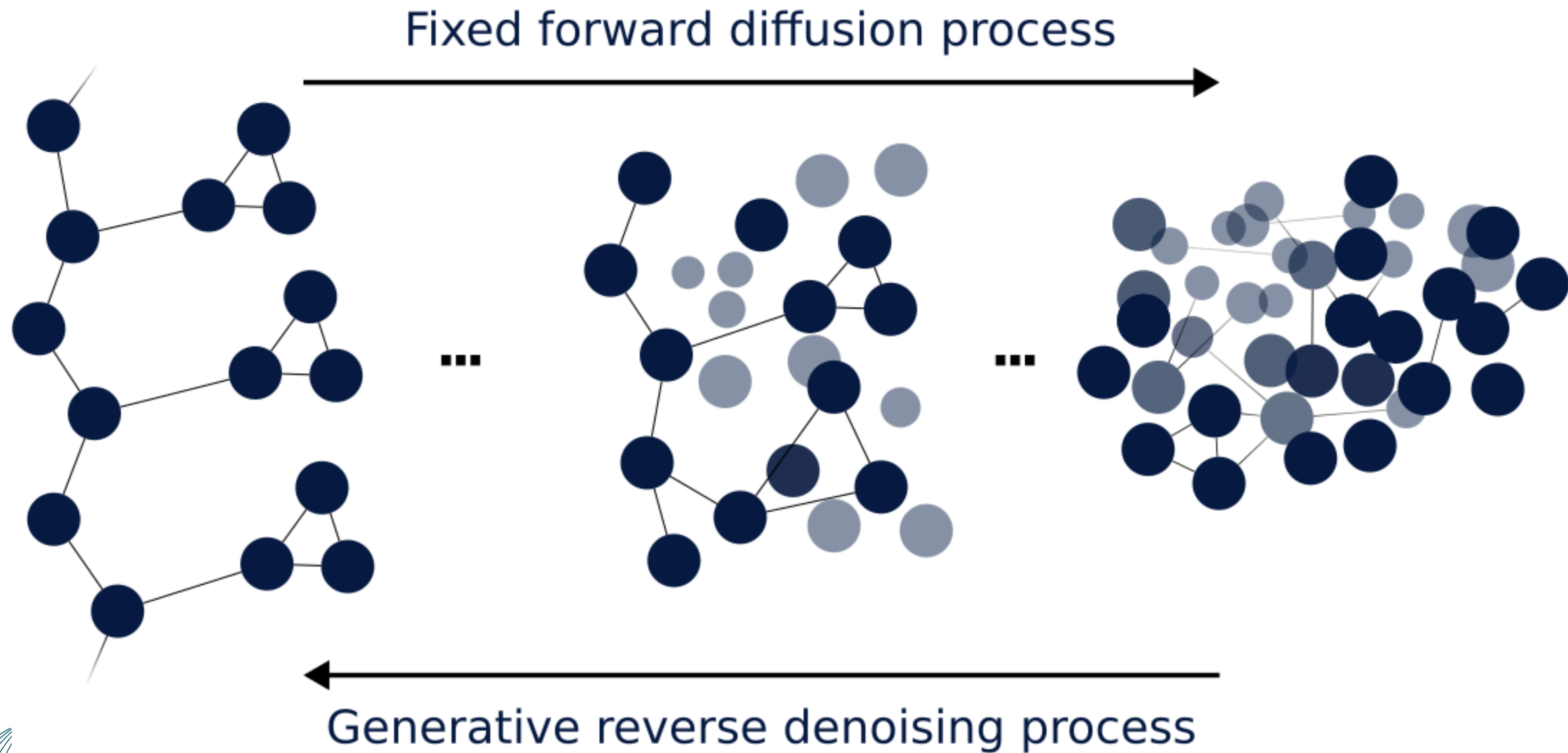**10,583** descriptors extracted from all other RNAs

Local 3D RNA descriptors recur in non-homologous structures, indicating that similar local environments are found across diverse RNAs.
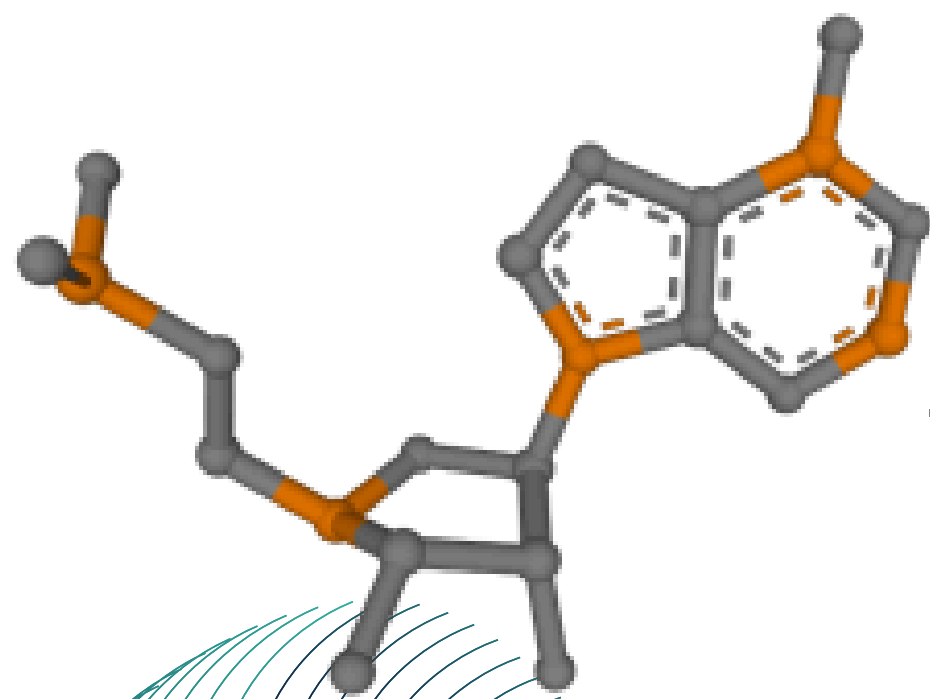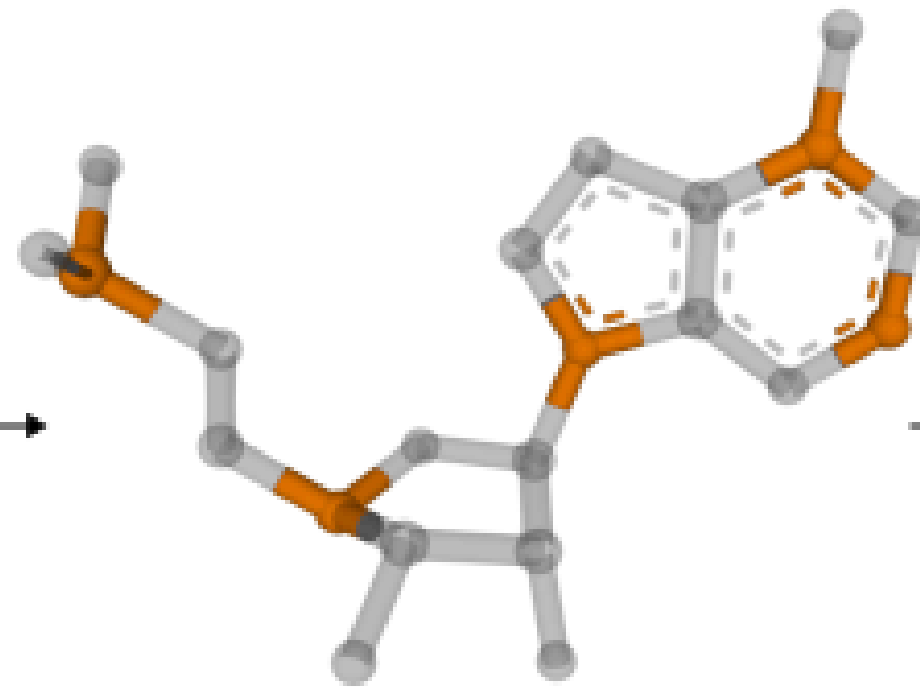
Our model

Why generative model?

Fixed forward diffusion process

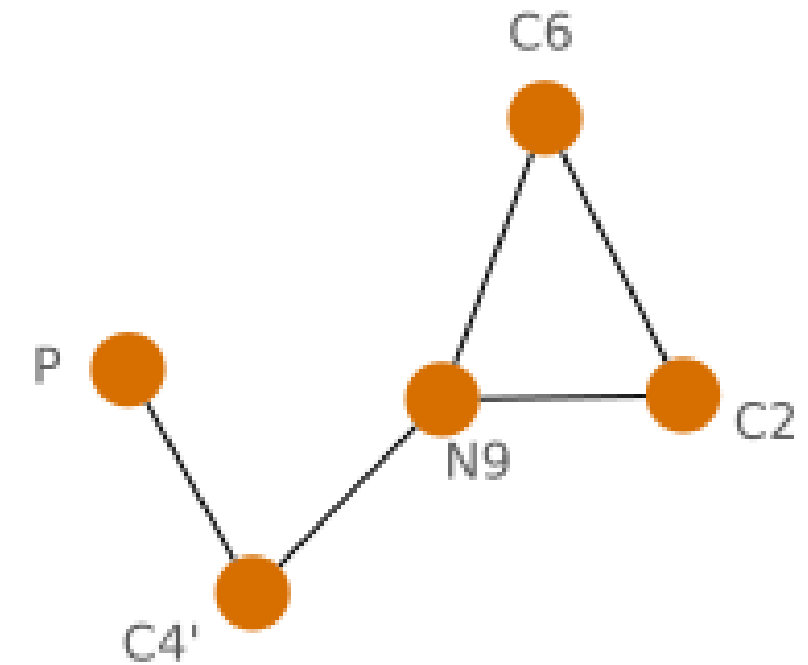Generative reverse denoising process
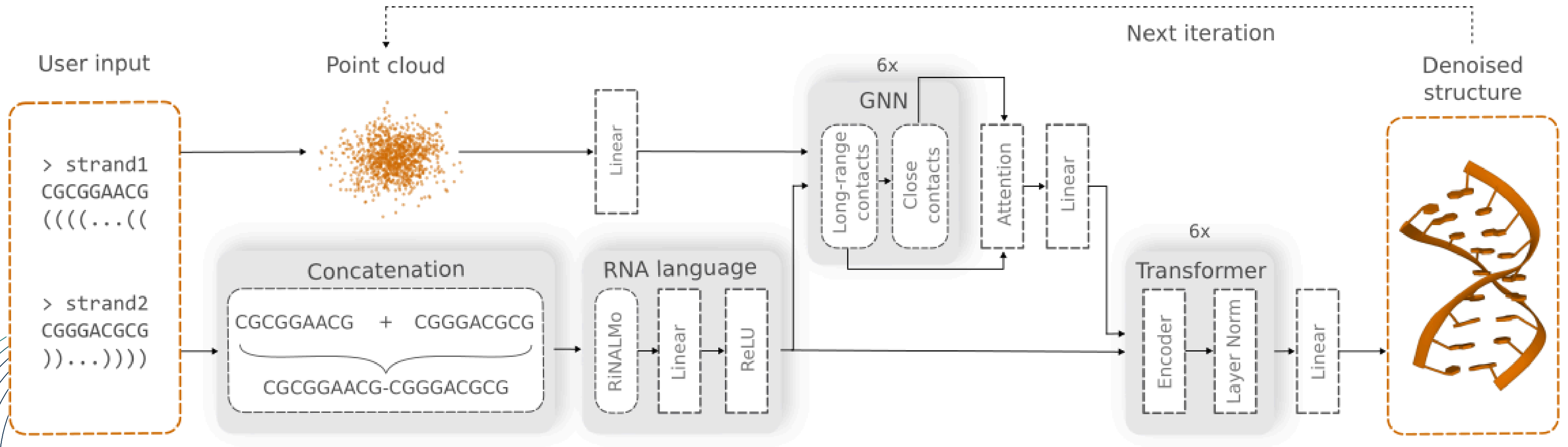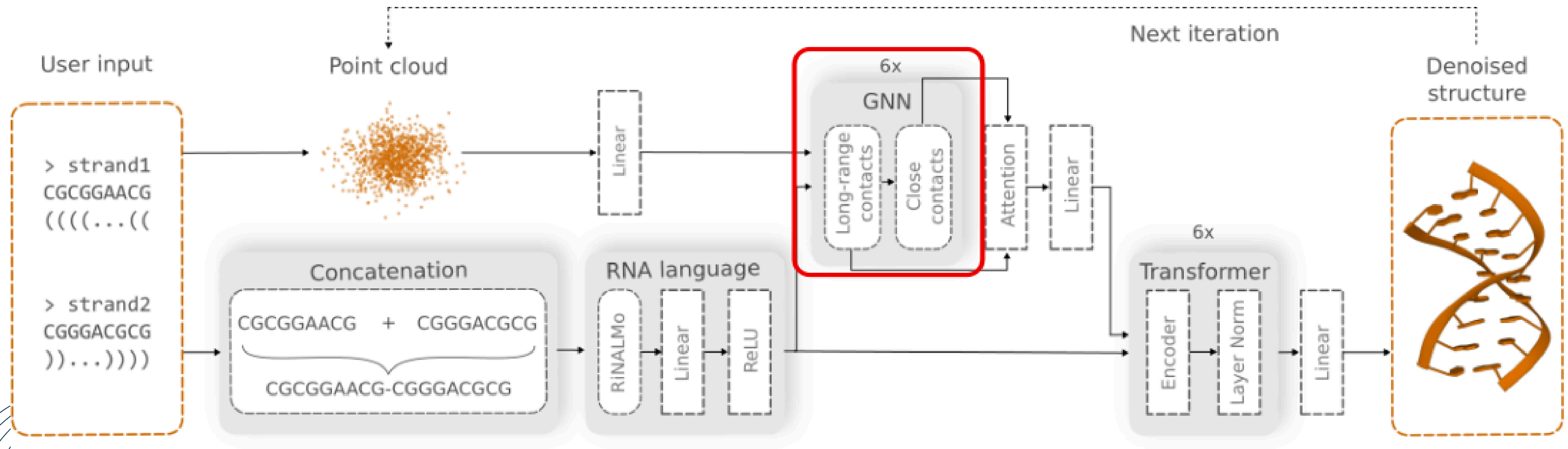
RNA 3D structure representation

All-atom structure

Coarse-grained model
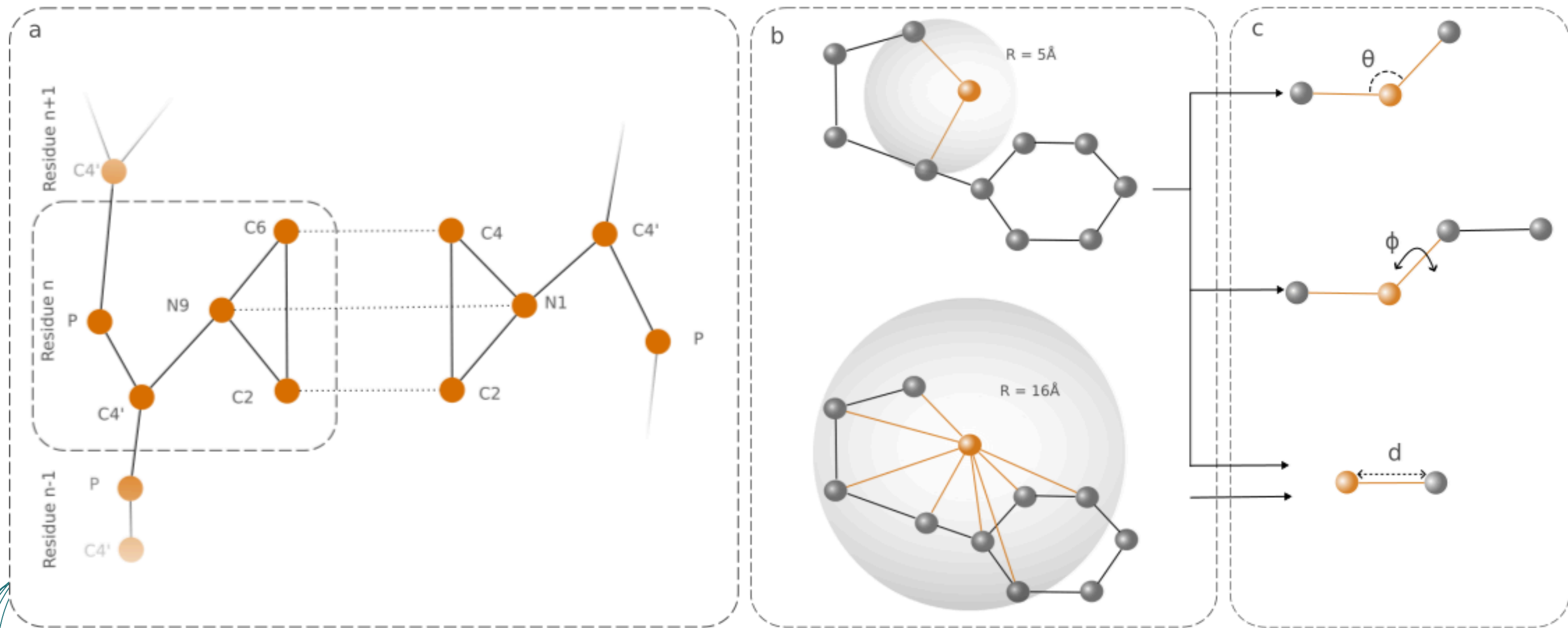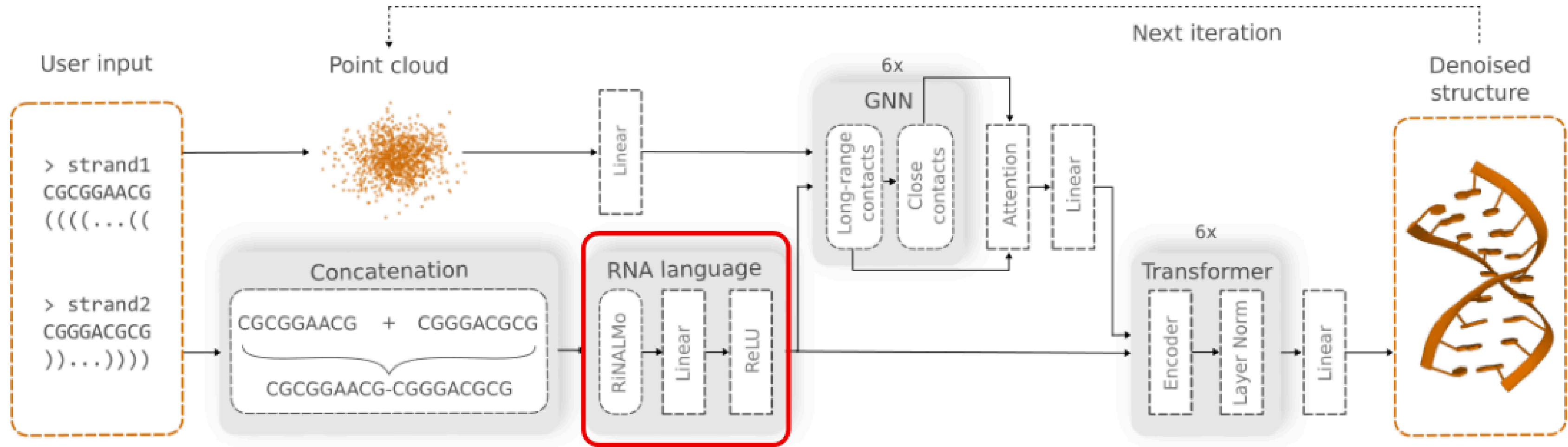
Graph representation

Graph Neural Network

RNA 3D structure representation

RNA language

User input

> strand1
CGCGGAACG
((((...((

> strand2
CGGGACGCG
))...))))

Point cloud

Next iteration

Denoised structure

Concatenation

CGCGGAACG + CGGGACGCG

CGCGGAACG-CGGGACGCG

RNA language

RiNALMo → Linear → ReLU

Linear

GNN 6x

Long-range contacts → Close contacts → Attention → Linear
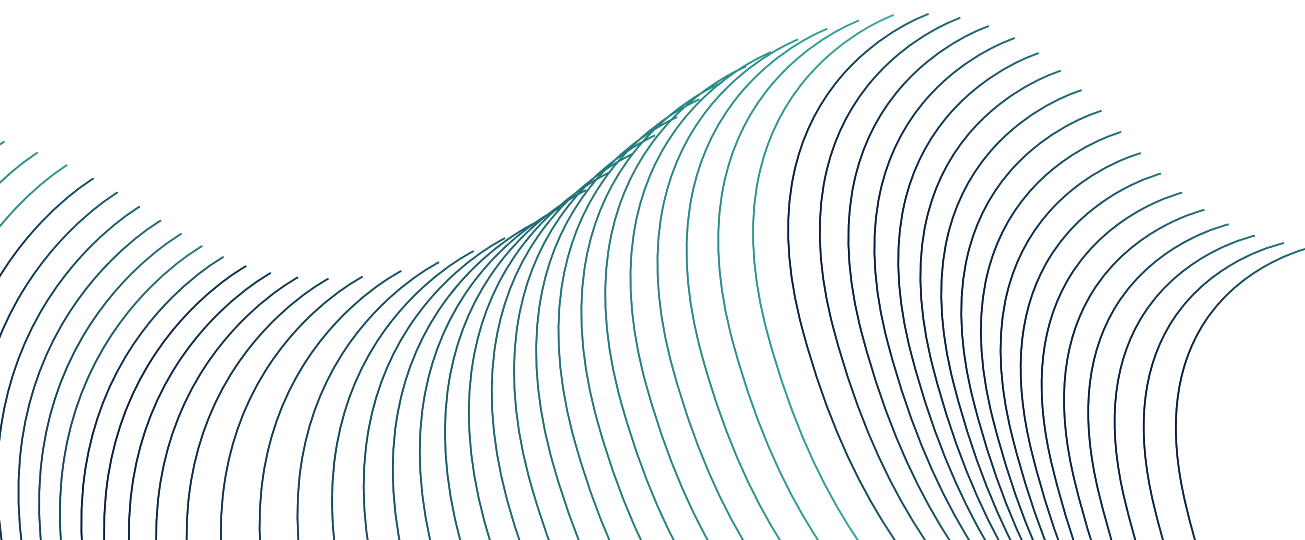
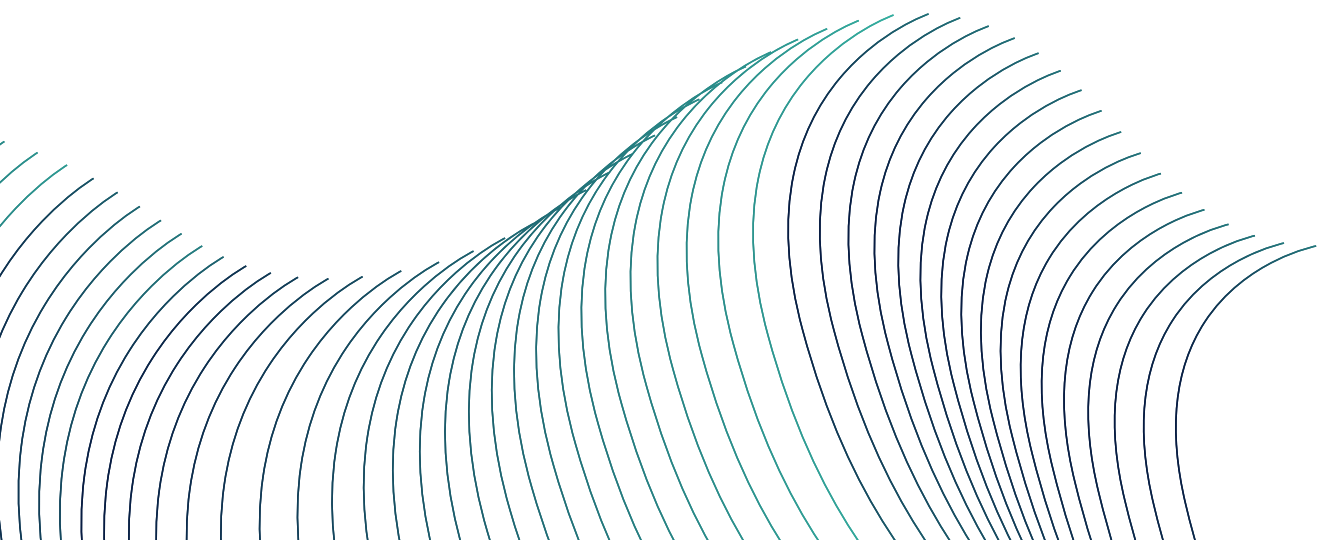Transformer 6x

Encoder → Layer Norm → Linear

G: 0.21, 0.37...
C: 0.47, 0.16...
G: 0.53, 0.42...
C: 0.78, 0.36...

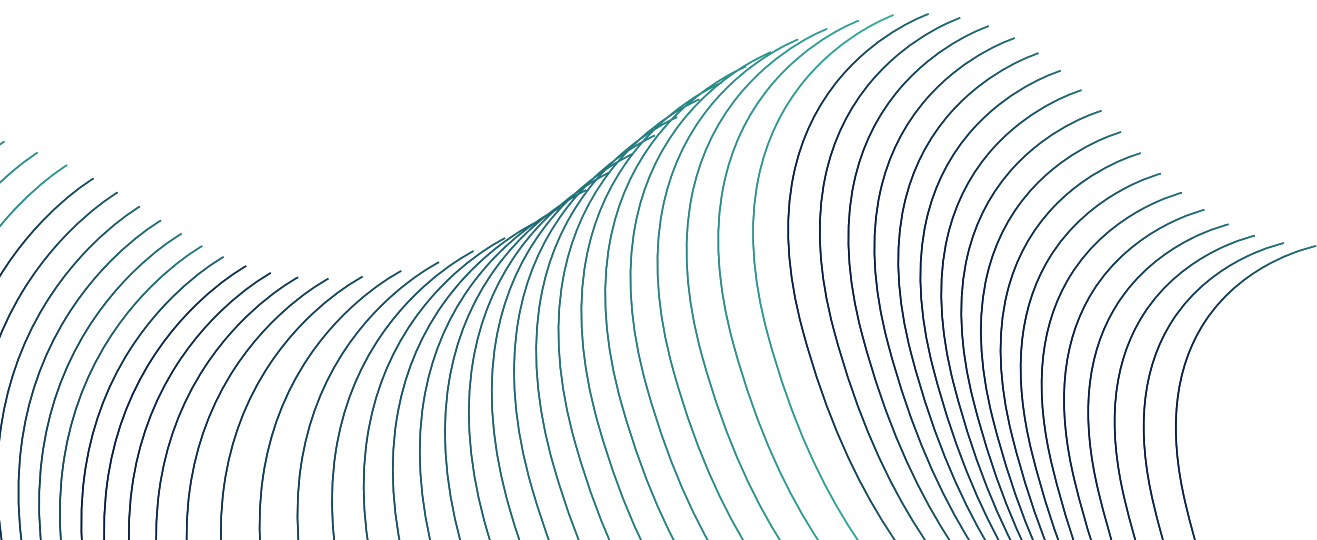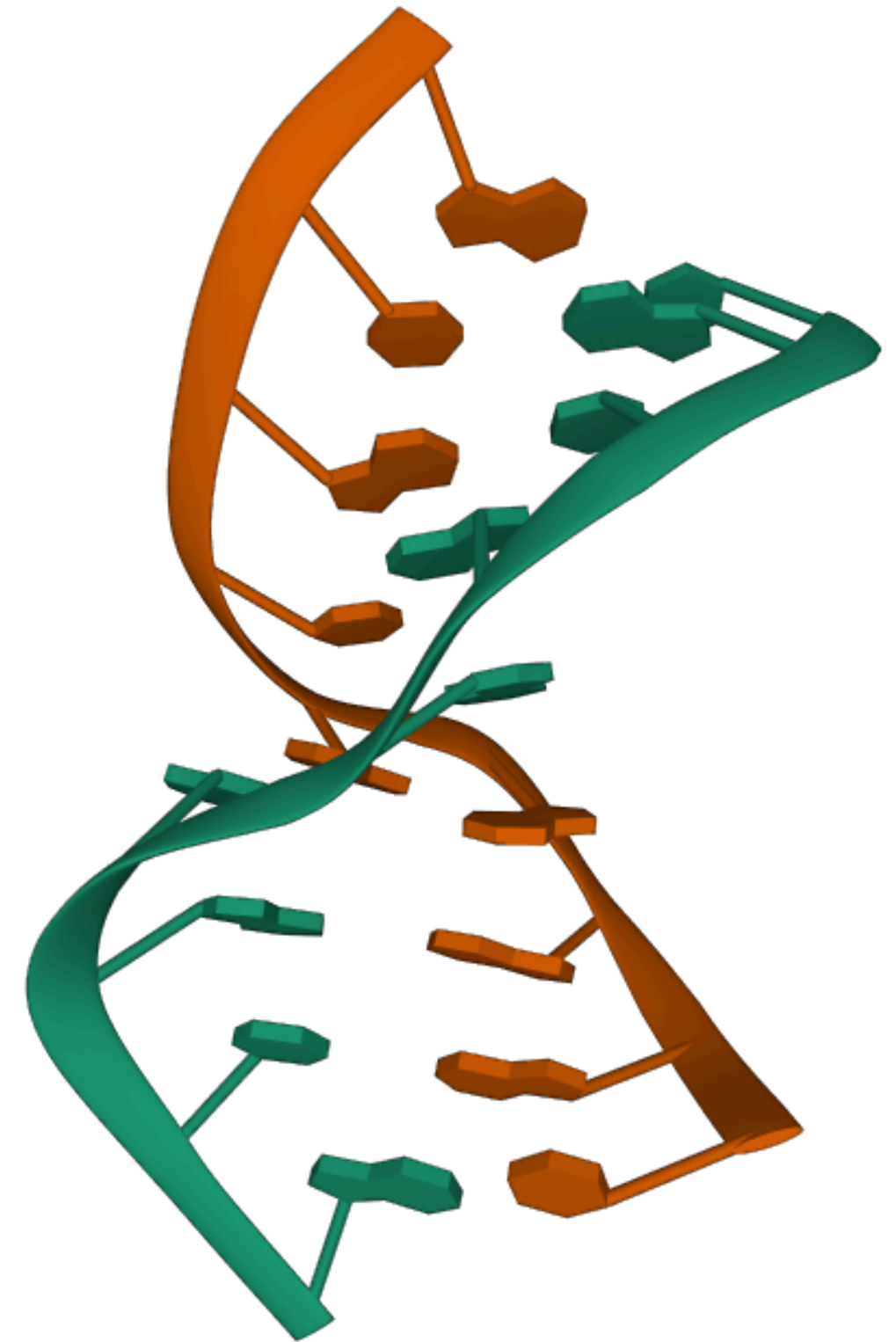Nucleotides represented as vector embeddings

Sequence context matters

No language model considered.
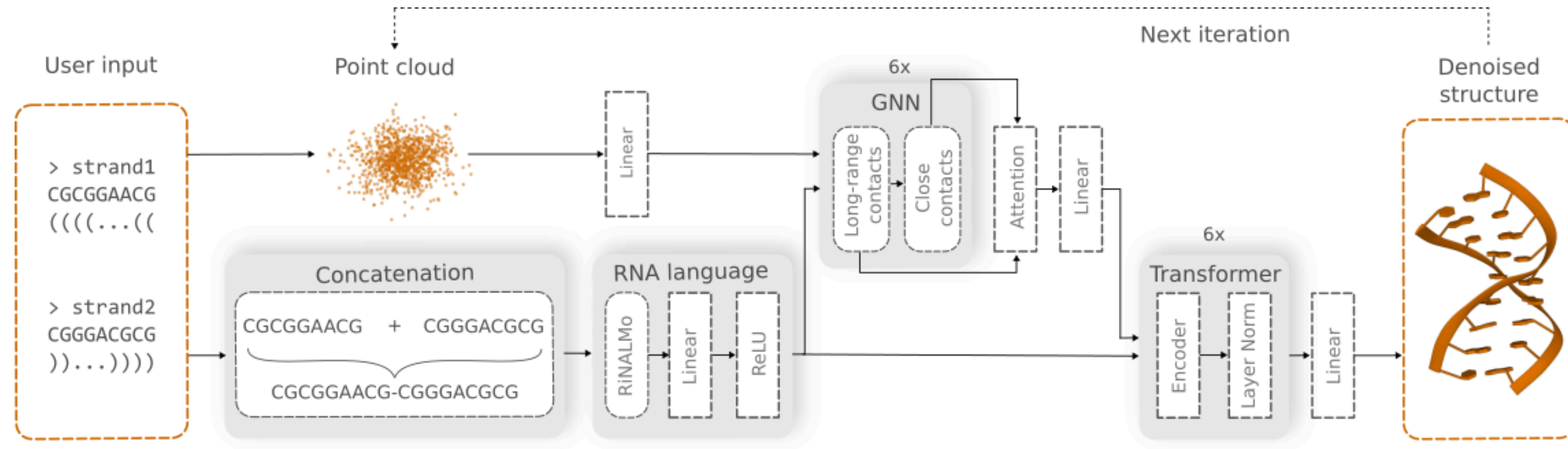
Sequence context matters

RNA language model considered.

Our method vs. AlphaFold3

Results

8D29 6YW5 6JXM

RNAgrail

3.99 Å 5.43 Å 5.13 Å

AlphaFold 3

4.95 Å 8.89 Å 14.46 Å
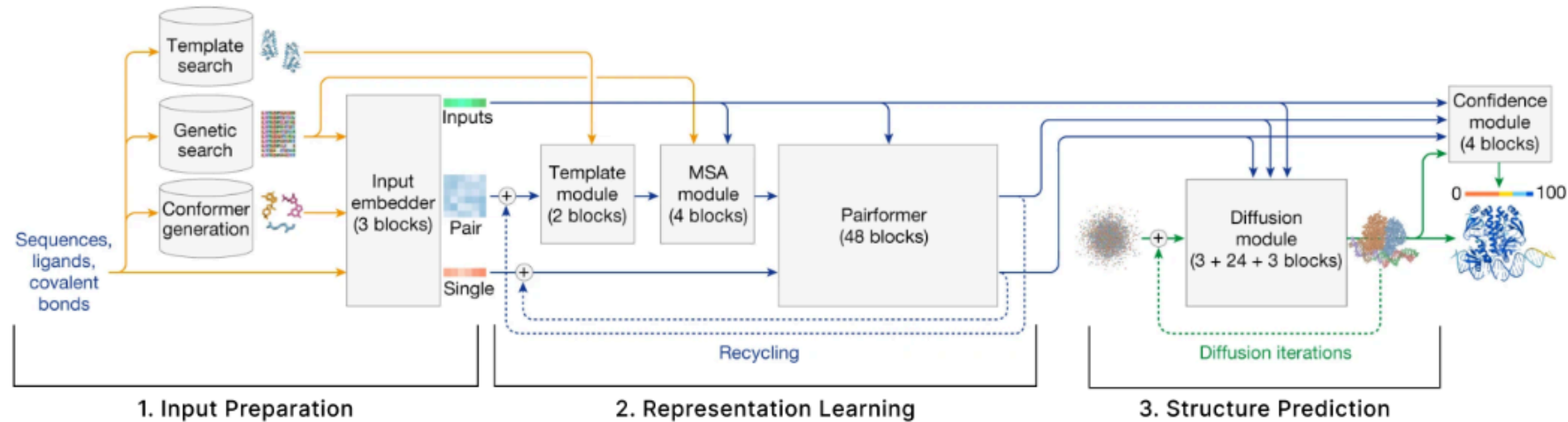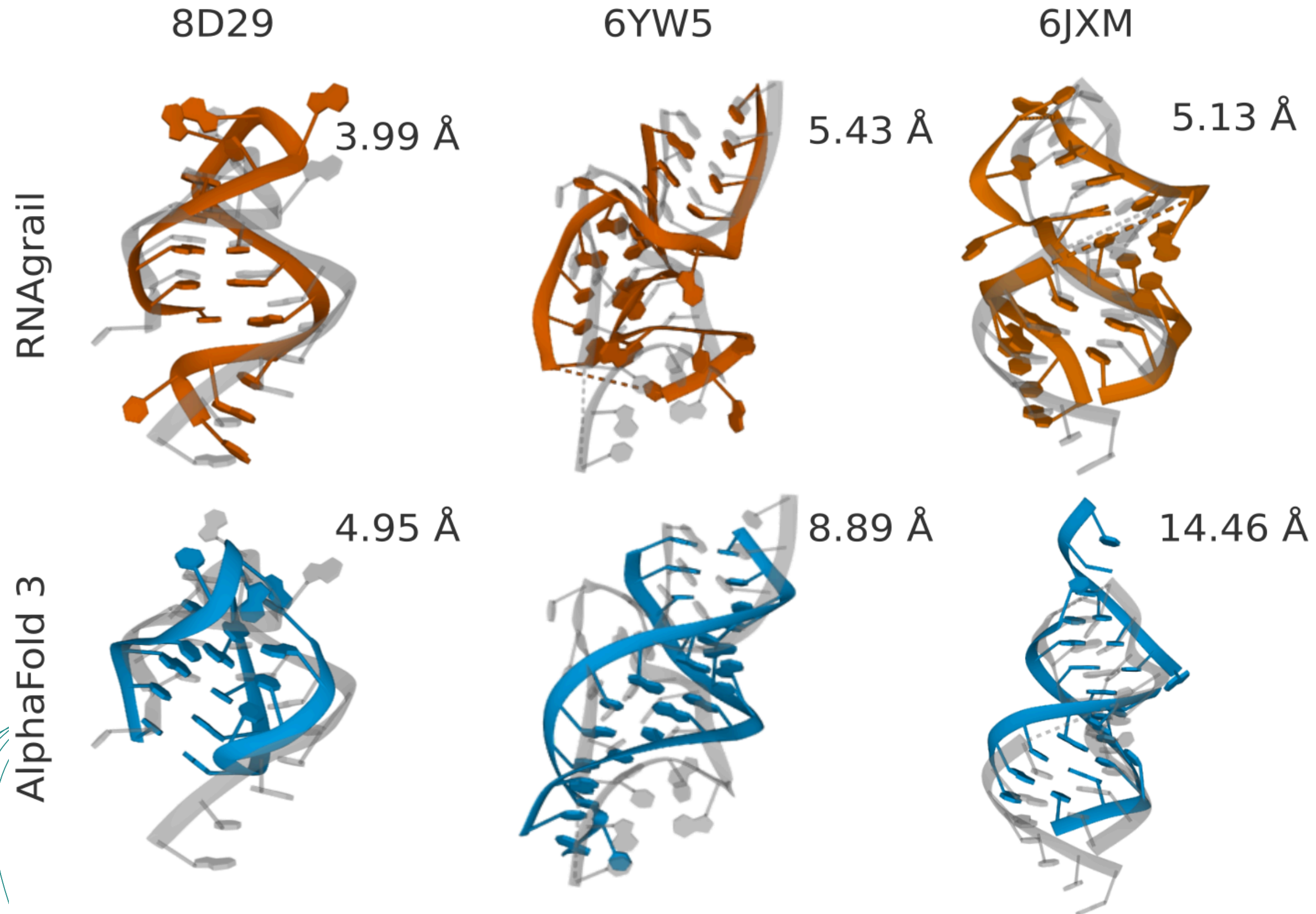
Results (2)
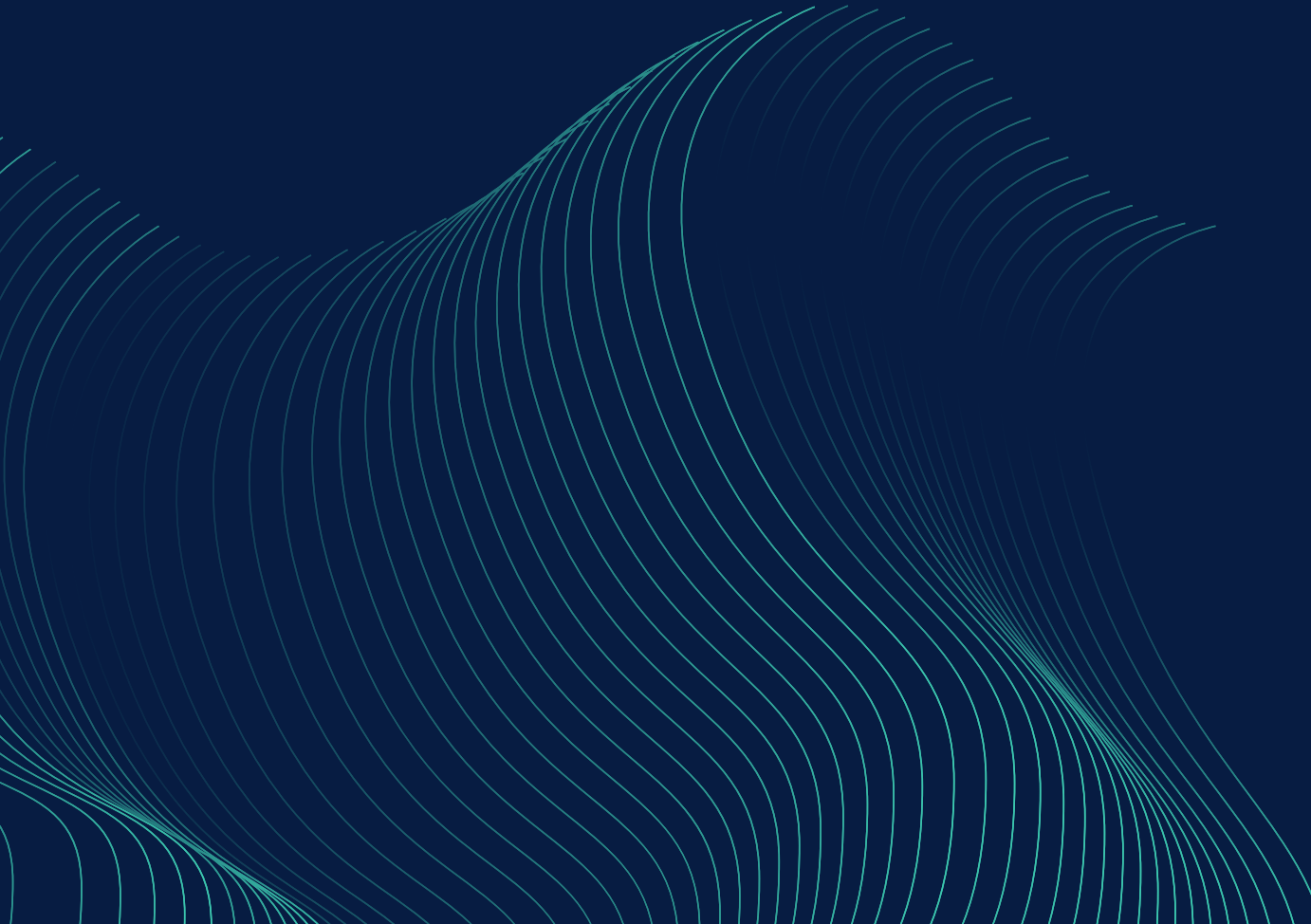
Ground-truth

RNAGrail

AlphaFold3
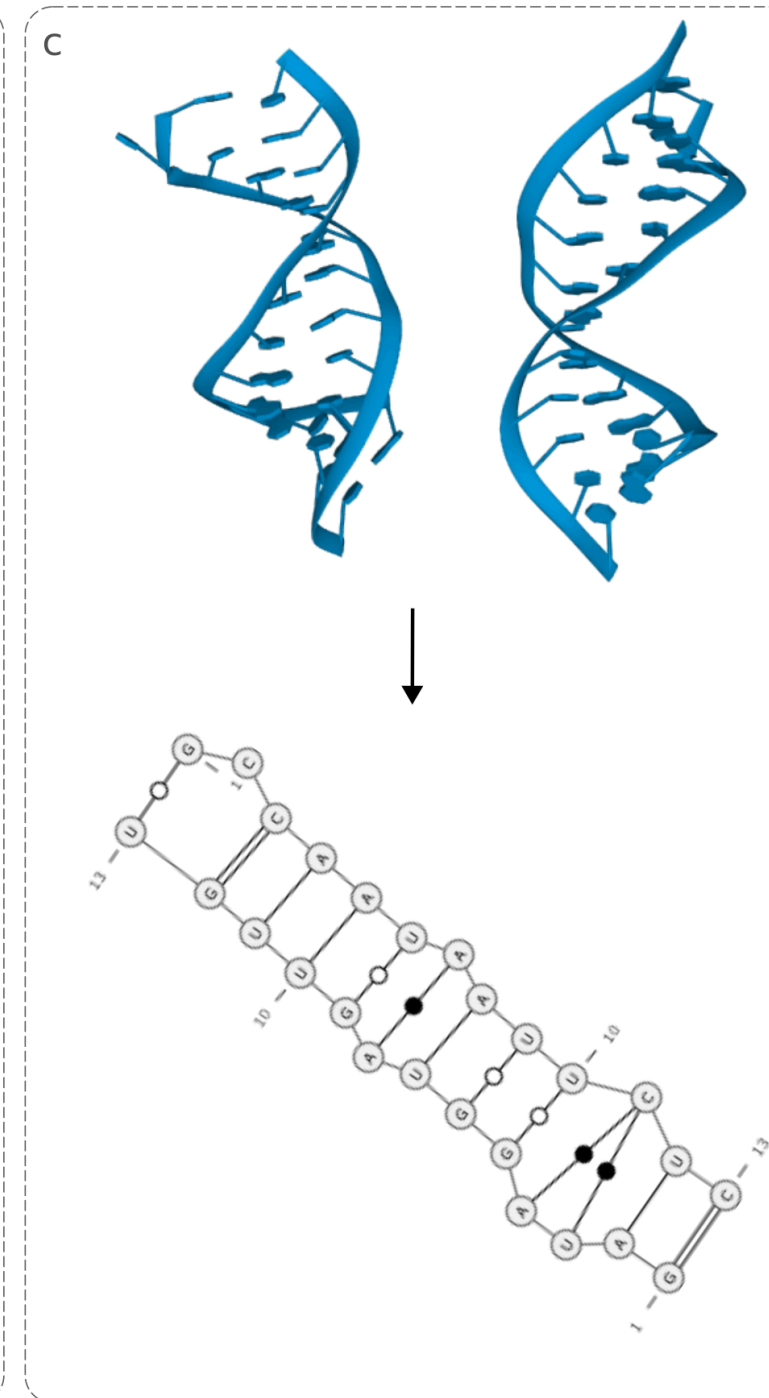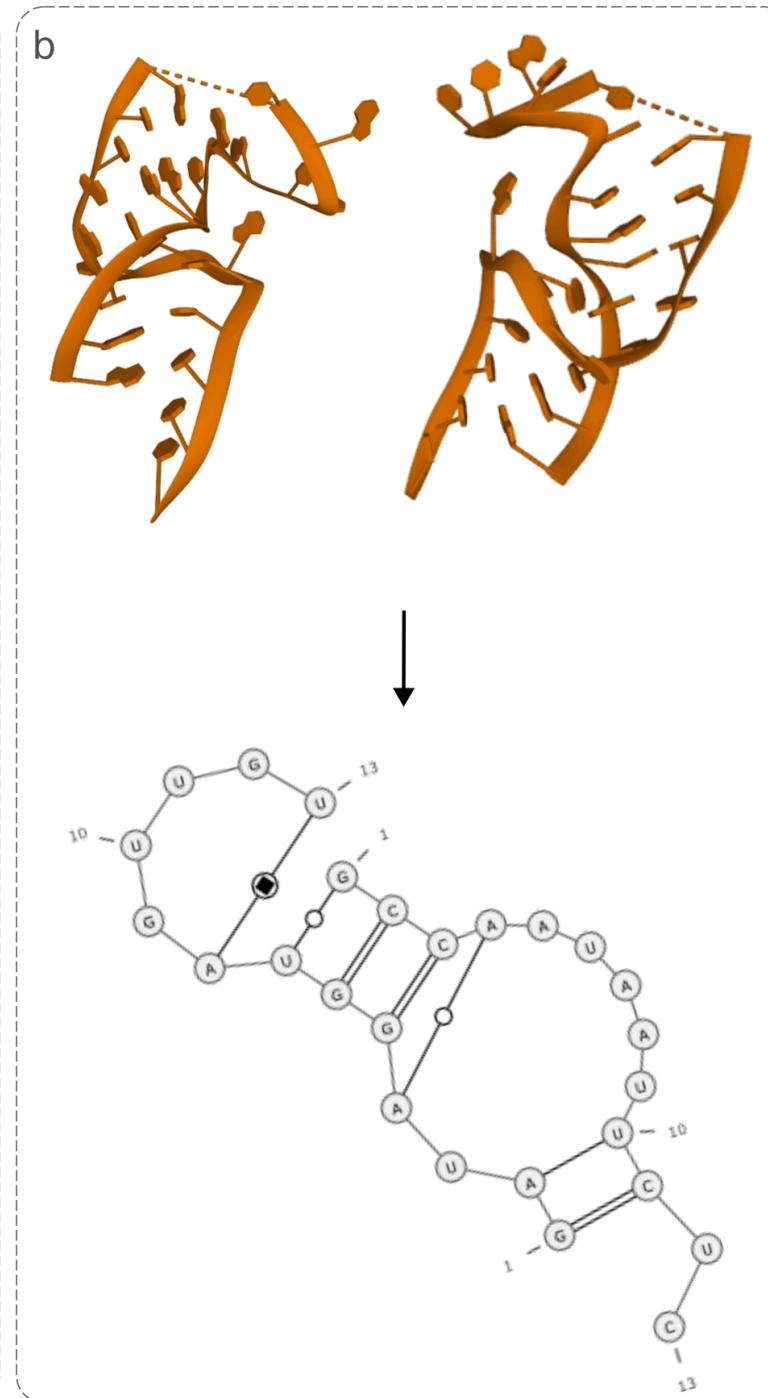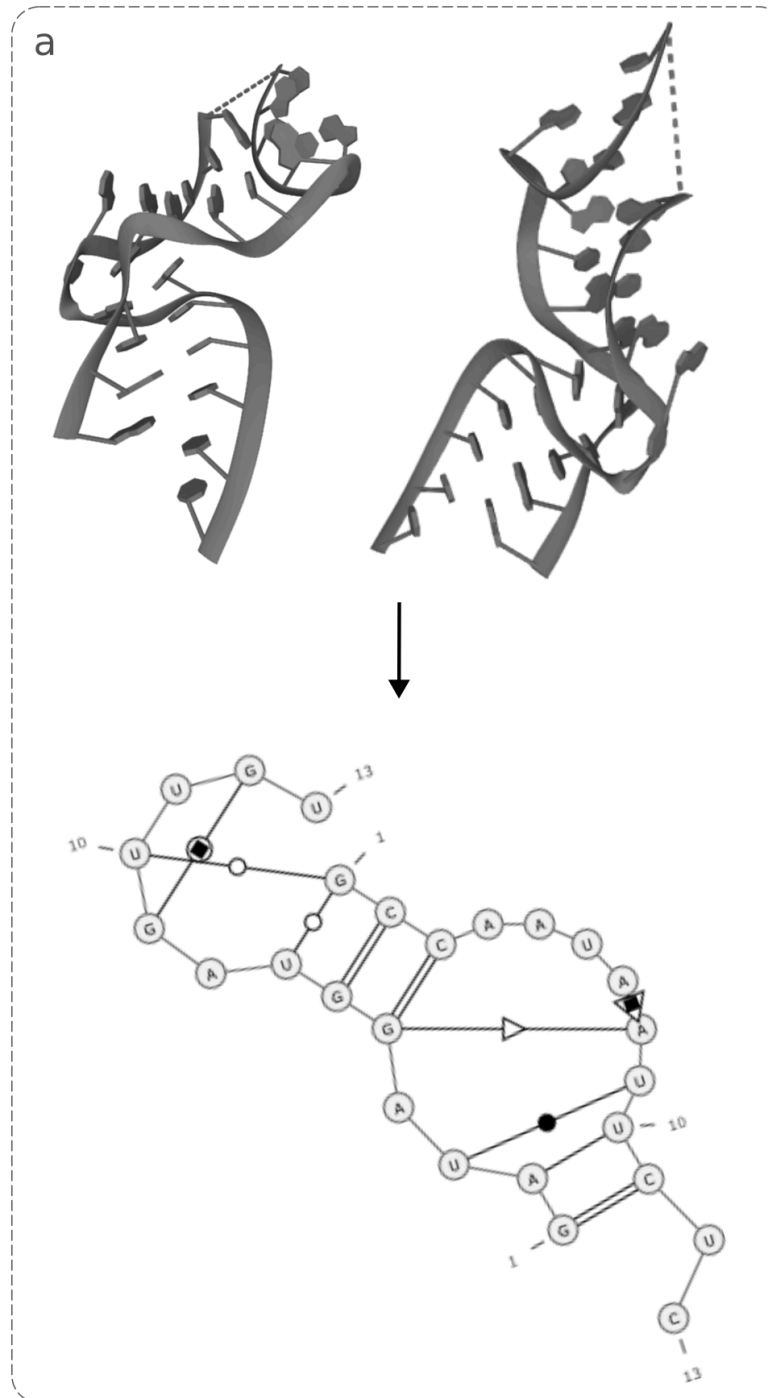
Summary

- Using local 3D RNA descriptors is a novel approach in the RNA 3D structure prediction domain.

- Our generative method predicts reliable RNA 3D structures without relying on structural templates.

- Embeddings derived from the RNA language model significantly increase the model performance.

- This model can be further extended to bigger structures

- User defined 2D structure is a useful feature for domain experts.

Endnotes

OpenSource

The training and test datasets and a pre-trained model weights are available at **Zenodo**. The source code is available at **GitHub**.

NeurIPS24 workshops

RNAgrail was accepted for **Machine Learning in Structural Biology** workshop at **NeurIPS24**.

RNAgrail: graph neural network and diffusion model for RNA 3D structure prediction

Marek Justyna[1*]
mjustyna@cs.put.poznan.pl

Craig L. Zirbel[2]
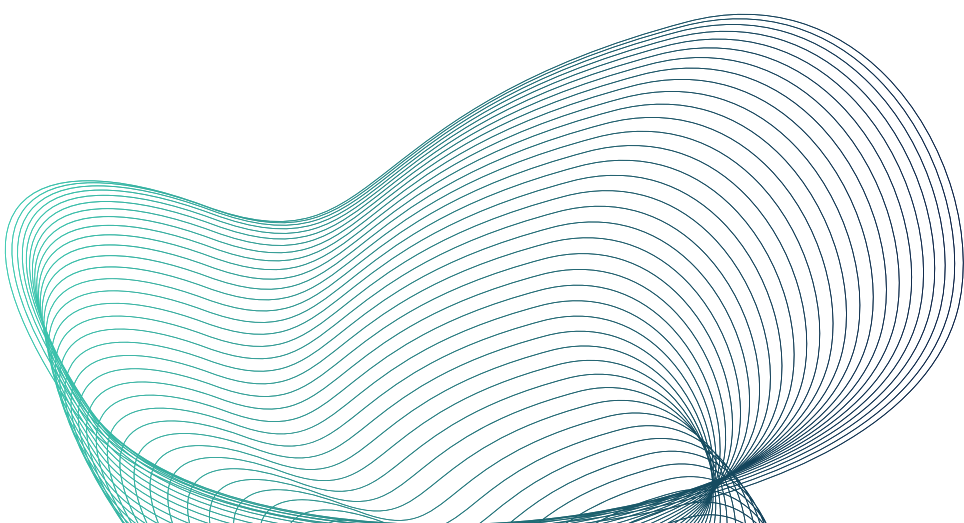zirbel@bgsu.edu

Maciej Antczak[1,3]
mantczak@cs.put.poznan.pl

Marta Szachniuk[1,3]
mszachniuk@cs.put.poznan.pl

[1] Institute of Computing Science, Poznan University of Technology, Poland
[2] Department of Mathematics and Statistics, Bowling Green State University, USA
[3] Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poland

Abstract

The function of RNA is intrinsically tied to its 3D structure traditionally explored by X-ray crystallography, NMR, and Cryo-EM. However, these experiments often lack atomic-level resolution, creating the need for accurate *in silico* RNA structure prediction tools. This need has driven advances in artificial intelligence (AI), which has already revolutionized protein structure prediction. Unfortunately, similar breakthroughs in the RNA field remain limited due to sparse and unbalanced structural data. Here, we introduce RNAgrail, a novel RNA 3D structure prediction method that focuses on RNA substructures using a denoising diffusion probabilistic model (DDPM). Unlike AlphaFold 3 (AF3), considered by many to be an oracle, RNAgrail allows expert users to define base pair constraints, offering superior flexibility and precision. Our method outperformed AF3 by 12% in terms of mean RMSD and by 24% in terms of mean eRMSD. Additionally, it perfectly reproduced the canonical secondary structure outperforming Af3 by 40% in terms of interaction network fidelity (INF). RNAgrail demonstrated robustness across diverse RNA motifs and families. Despite being trained exclusively on rRNA and tRNA, it effectively generalizes to new RNA families, thus, addressing one of the major challenges in RNA 3D structure prediction. These results underscore the potential of focusing on small RNA components and integrating user-defined constraints to significantly enhance RNA 3D structure prediction, setting a new standard in RNA modeling.

Introduction

Ribonucleic acid (RNA) is fundamental to a wide range of biological processes across all living organisms. It plays a critical role in gene transcription regulation, protein synthesis, and many other cellular functions. RNA also constitutes the genetic material of some pandemic-causing viruses, including HIV and SARS-CoV-2. In medicine, this molecule serves as a valuable biomarker for cancer detection and a target in cancer therapeutics [36, 38]. Understanding the full spectrum of RNA functions is based heavily on structural studies, with a particular focus on deciphering the three-dimensional shape of this molecule.

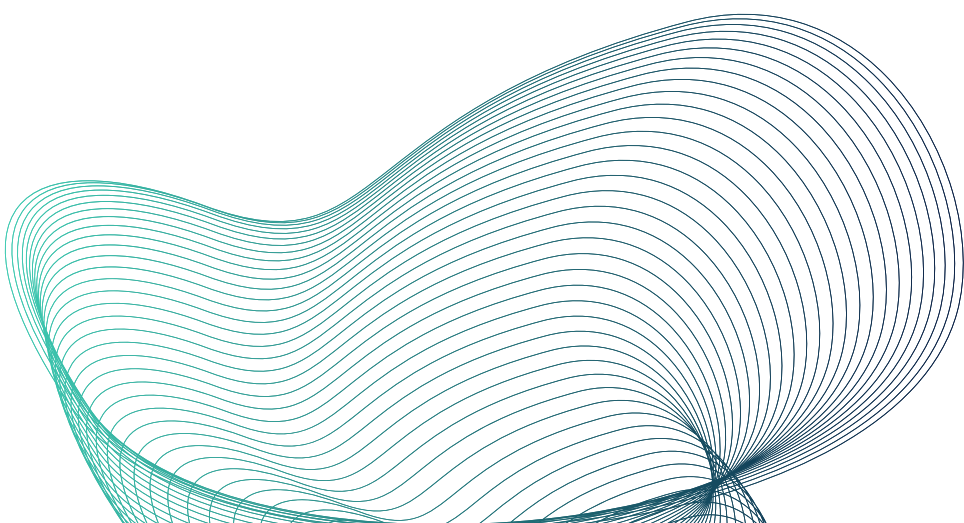*Corresponding author

Acknowledgements

Craig Zirbel

Maciej Antczak

Marta Szachniuk