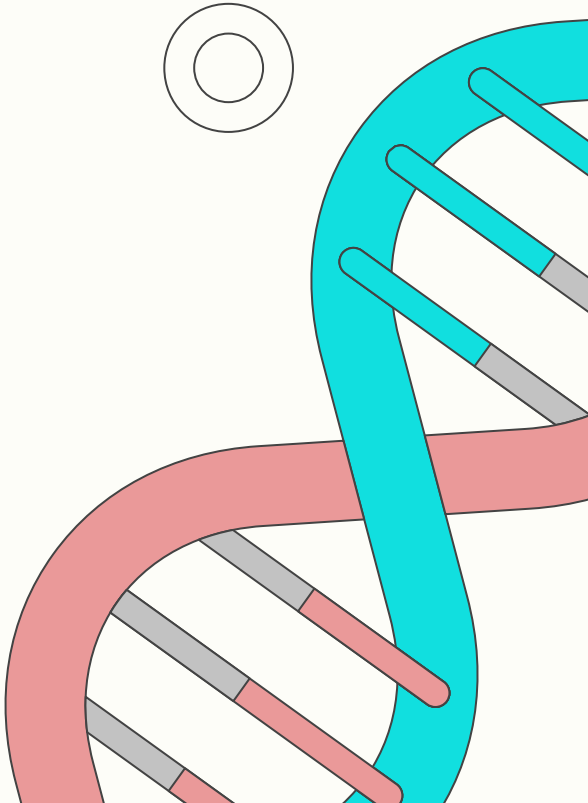
Predicting RNA 3D Structures – Stanford Kaggle Challenge

Violeta Kastreva, Kristiyan Garchev
Link



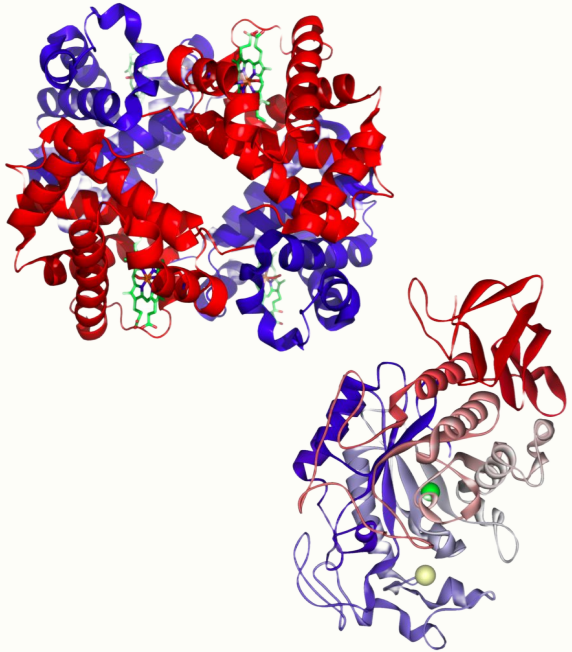
Project Motivation

Predicting RNA 3D structure is a key challenge in:

- Bioinformatics
- Molecular biology
- Drug discovery
- Synthetic biology

Stanford Kaggle competition aims to advance this with ML

Groundtruth is derived from experimental molecular structure data



Dataset Overview

Source: Stanford RNA Folding Dataset
~5000+ molecules, many with multiple samples

Contains:

- RNA sequences
- MSA (Multiple Sequence Alignments) for over 50% of samples
- Chain sequences (proteins, ligands) of all products from the actual experimental 3D structure discovery of the main chain

They come in formats .csv, .fasta and .cif

From the dataset we could also precompute:

- Pairwise atom distance matrices
- Backbone and sidechain angles (ϕ, θ)

Real vs Synthetic Data

In addition to real data, we use a large synthetic dataset (450k+ samples) generated by RNA folding models (by RFdiffusion) [Link]

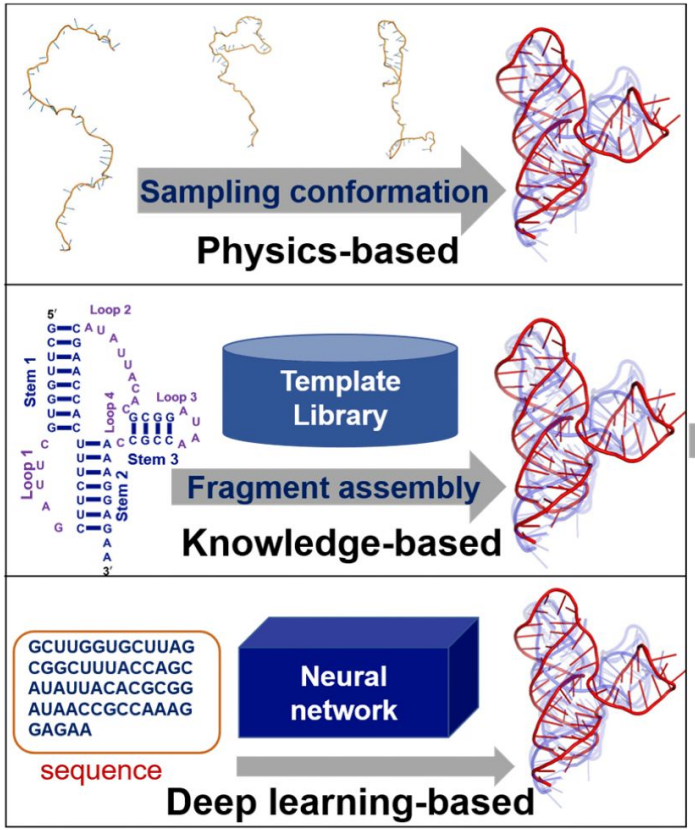
Synthetic data is used:

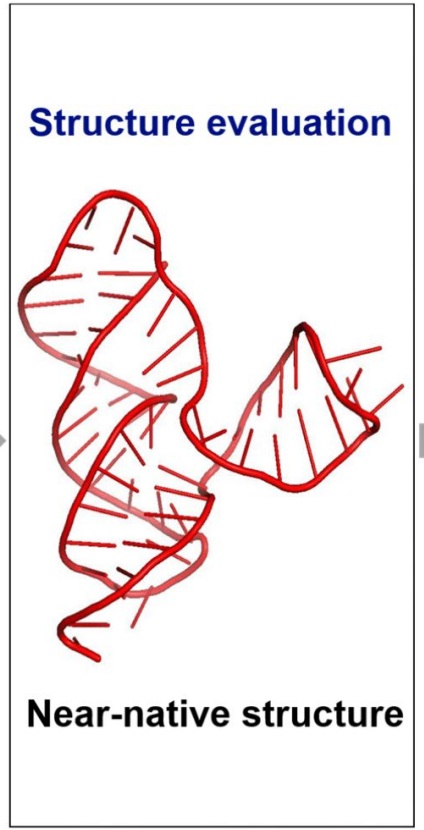
- In early training stages
- To teach stereochemical rules of RNA folding

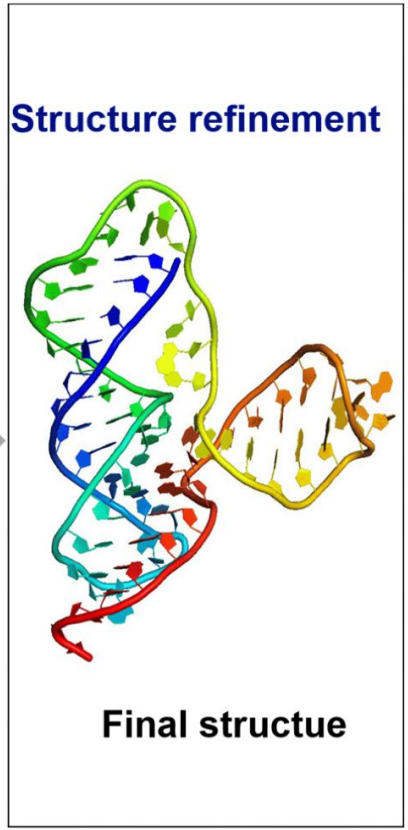
Unfortunately, synthetic data cannot capture true biological complexity, so we use it for initial model calibration to the rules of chemistry.

Both datasets have distinct sequence length distributions → must be handled carefully.

Synthetic Data Generation Pipeline







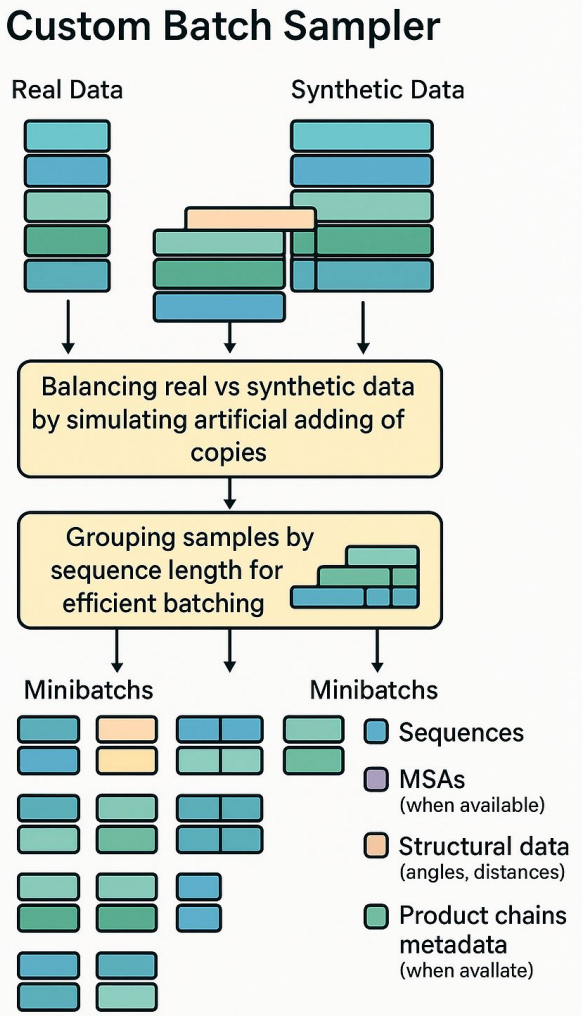
Data Usage & Sampling Strategy

Custom batch sampler for:

- Balancing real vs synthetic data by simulating artificial adding of copies of the samples of the smaller dataset to each batch.
- Grouping samples by sequence length for efficient batching

Attempting to use all modalities:

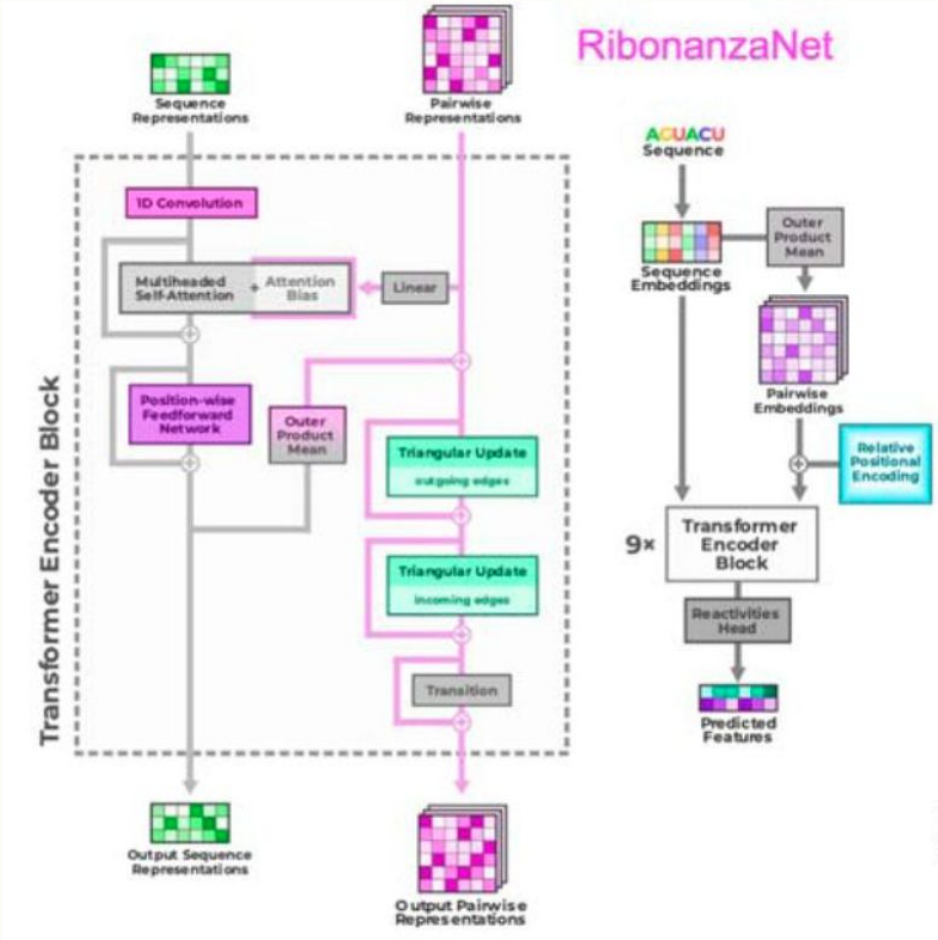
- Sequences
- MSAs (when available)
- Structural data (angles, distances)
- Product Chains metadata (when available)



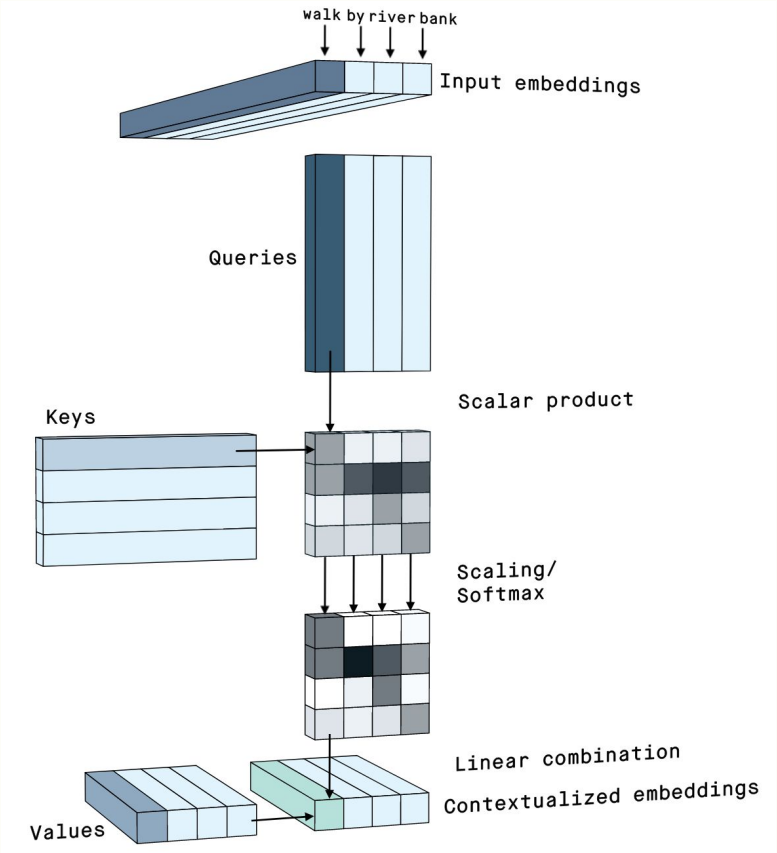
Model Architectures

We plan to use as base model designs & adapt:

- RibonanzaNet
  - Based on EvoFormer block (from AlphaFold & previous Kaggle winners)
  - Originally accepts only sequences + optional MSA
- RibonanzaNet-DDPM
  - A diffusion-based version
  - Strong performance on generative structure prediction



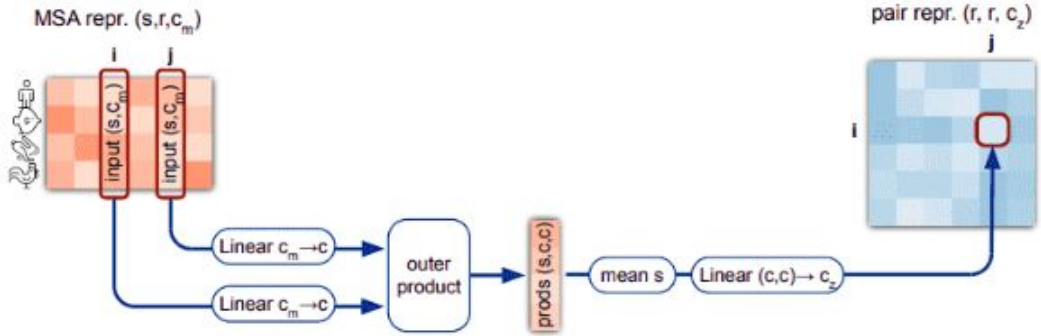
RibonanzaNet



RibonanzaNet

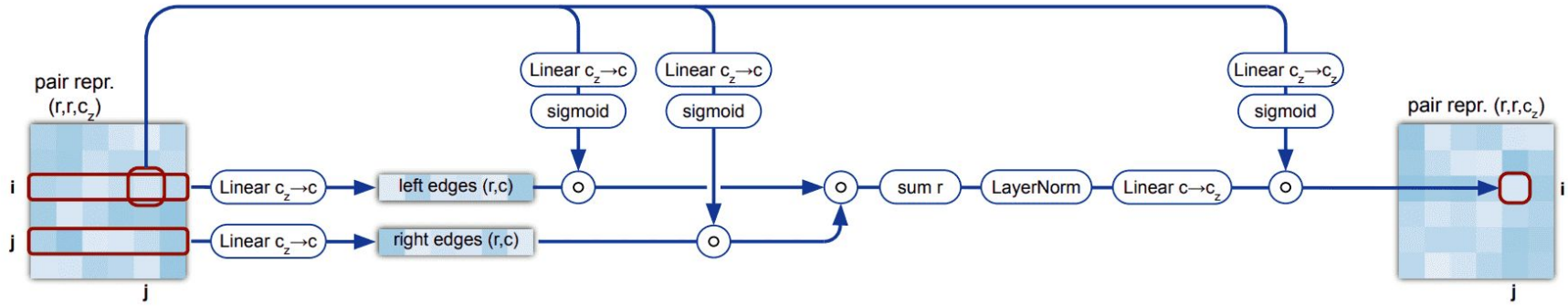
1.6.4 Outer product mean

The "Outer product mean" block transforms the MSA representation into an update for the pair representation (Suppl. Fig. 5 and Algorithm 10). All MSA entries are linearly projected to a smaller dimension $c = 32$ with two independent Linear transforms. The outer products of these vectors from two columns i and j are averaged over the sequences and projected to dimension c_z to obtain an update for entry ij in the pair representation. This is a memory intensive operation, as it requires constructing high-dimensional intermediate tensors. See section 1.11.8 for implementation details.



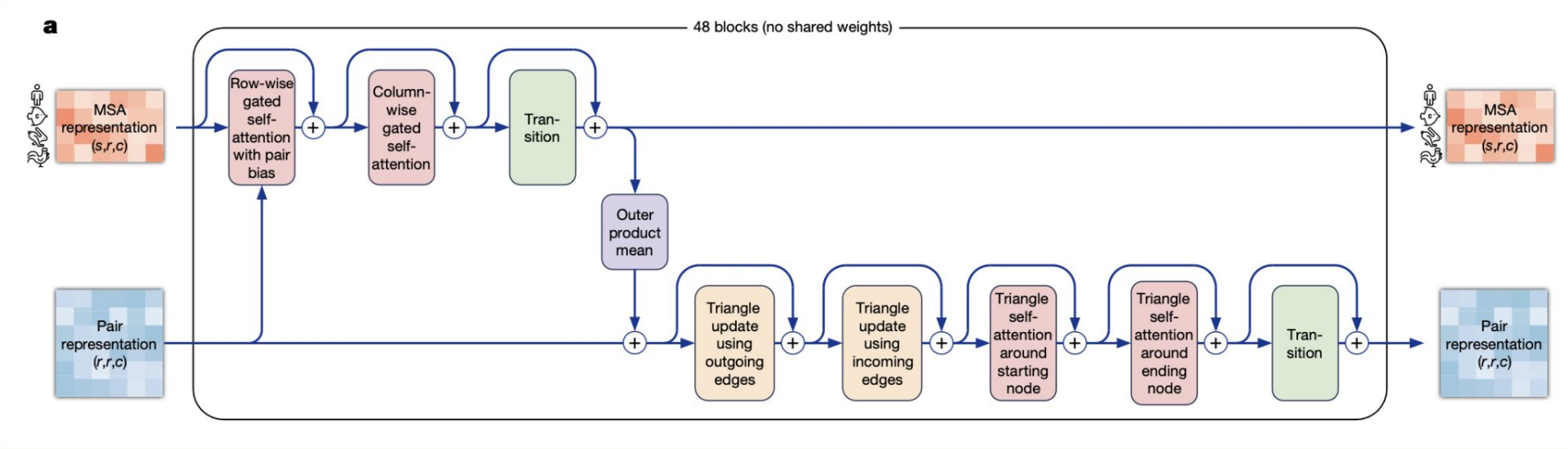
Supplementary Figure 5 | Outer product mean. Dimensions: s: sequences, r: residues, c: channels.

RibonanzaNet

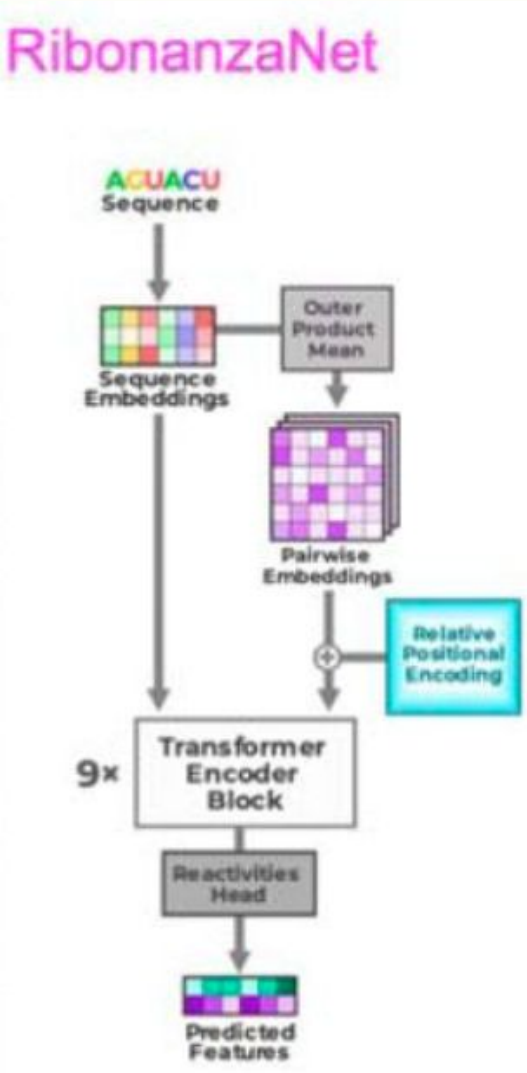


Supplementary Figure 6 | Triangular multiplicative update using "outgoing" edges. Dimensions: r: residue
c: channels.

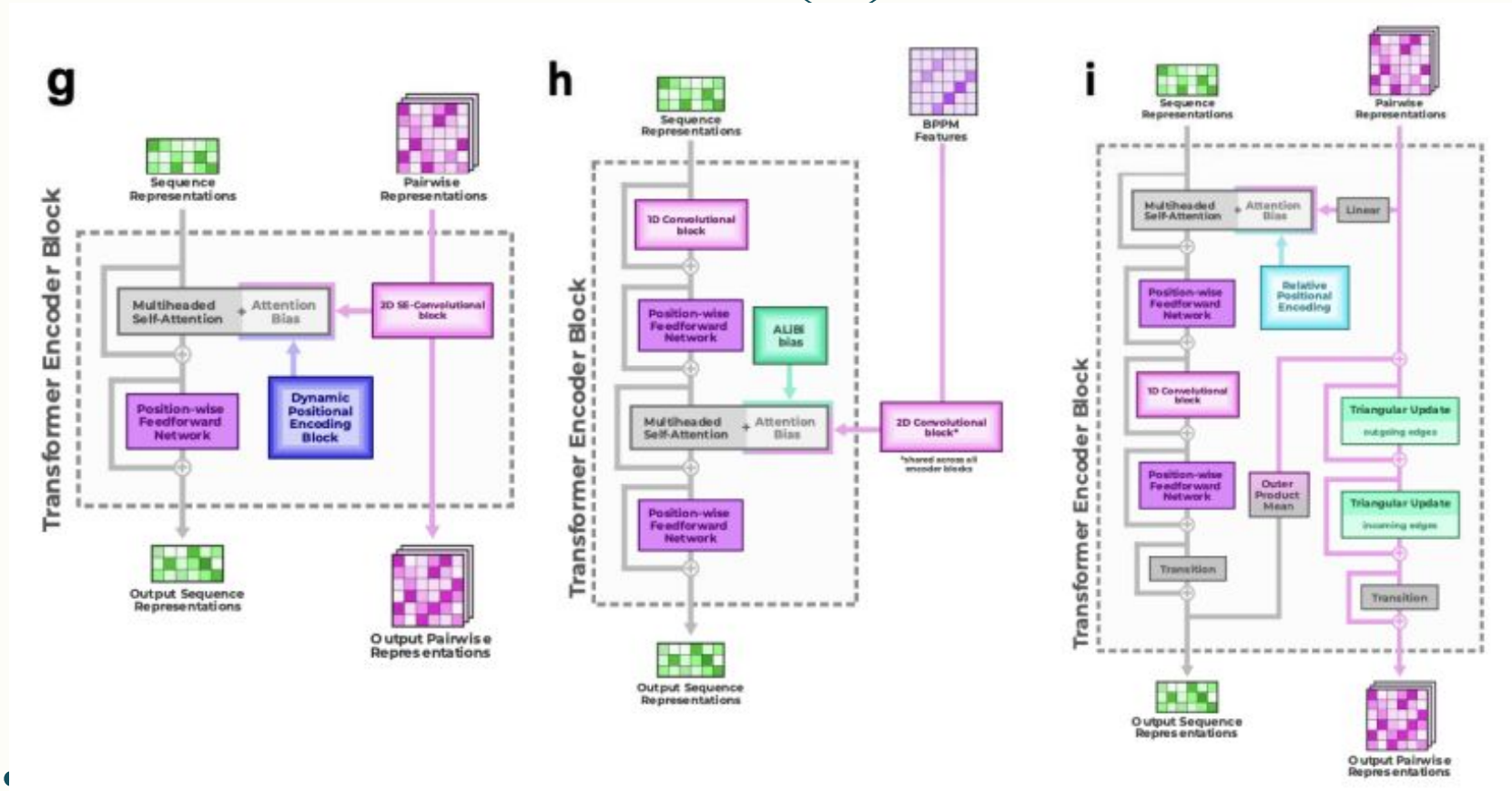
RibonanzaNet



RibonanzaNet



RibonanzaNet-DDPM



Model Modifications

However, these baseline models don't neither accept all the input we'd like to utilize, nor do they output the exact type of data we need.

We will adapt the architecture to match the required groundtruth: angles, distances, and coordinates.

Current Optimization Focus

Improve training with:

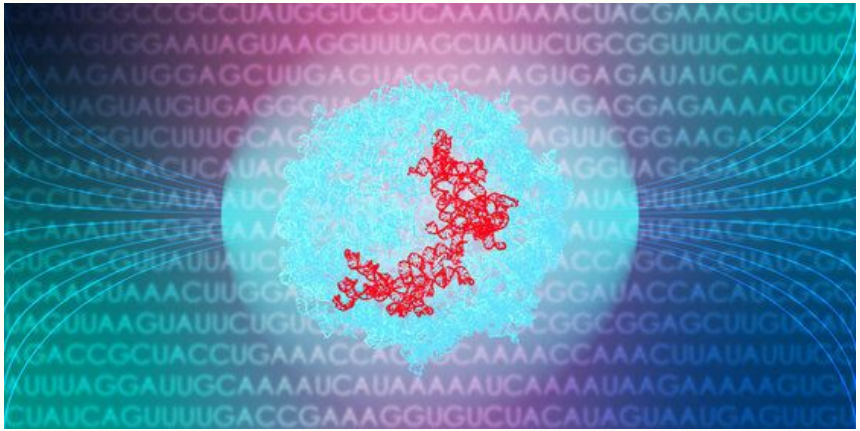
- Custom loss functions for angular continuity
- Efficient batching for long sequences

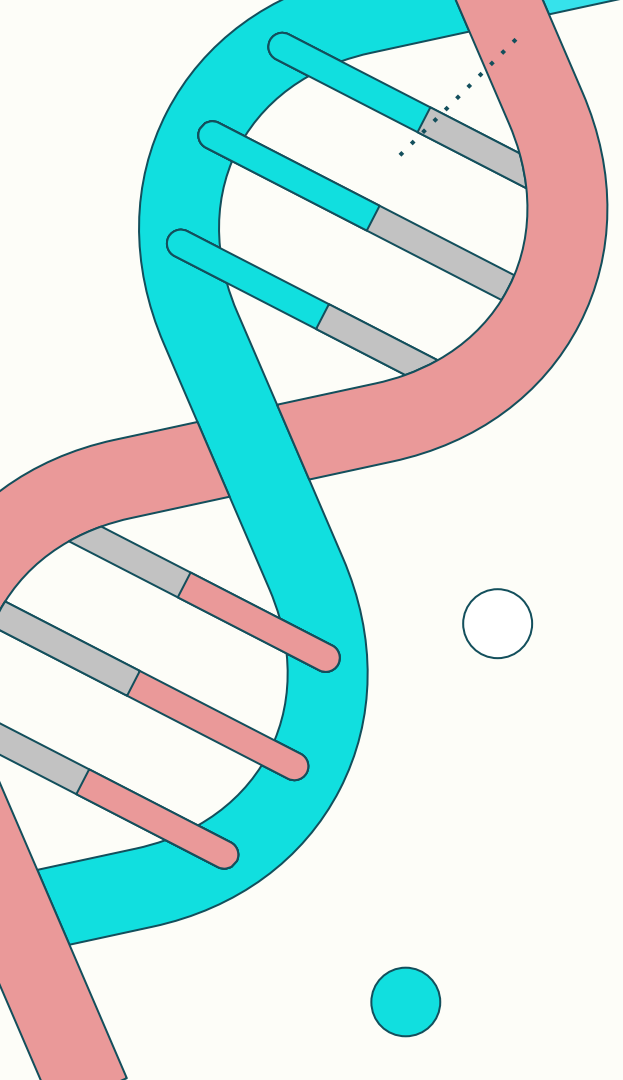
Investigating integration of:

- All-chain sequence information
- Product-chain-to-main-chain interactions

Future Work

- Experimenting with hyperparameters
- Compare results
- Write a scientific paper





Thanks!

Do you have any questions?