Machine Learning Guided RNA Structure Prediction

Utkarsh Upadhyay - Jülich Supercomputing Centre, Germany
Oskar Taubert - Karlsruher Institut für Technologie, Germany
Alexander Schug - Jülich Supercomputing Centre, Germany

JÜLICH Forschungszentrum

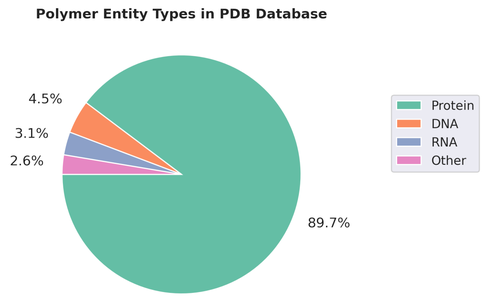
MOTIVATION

•RNA structure prediction helps researchers understand the function and behavior of RNA molecules to aid the development of RNA-based therapeutics and synthetic biology applications.

•Machine learning methods developed for proteins are not directly transferable to RNAs because of a large data gap.

•We develop methods to generate contact maps because they represent spatial pairwise inter-residue interactions.

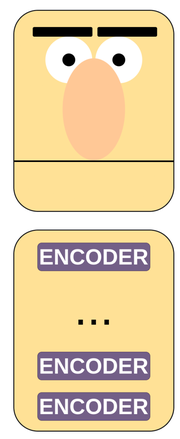
•We have worked on methods that took accuracy from 47%(DCA) to 77%(CoCoNet) and now to 87%(Barnacle) i.e., doubling accuracy while reducing false positives by five-fold.



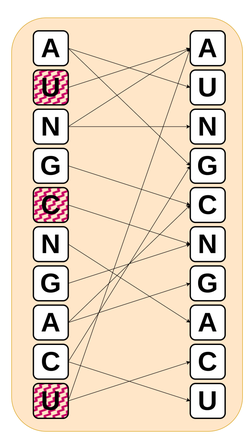
INTRODUCTION

We construct a modified version of BERT, a commonly used machine learning architecture. Its input is an RNA sequence with some residues or groups of residues masked and the model has to learn to predict the masked parts in an unsupervised manner.

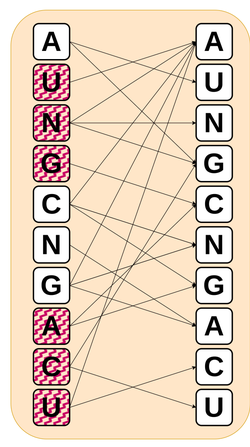
By this training the model captures interrelation in the residues of an RNA sequence. This information can then be finetuned for various other tasks to generate contact maps, distance maps, secondary structures and 3D structures as well.



BERT



Attention Mechanism (Single Residue)

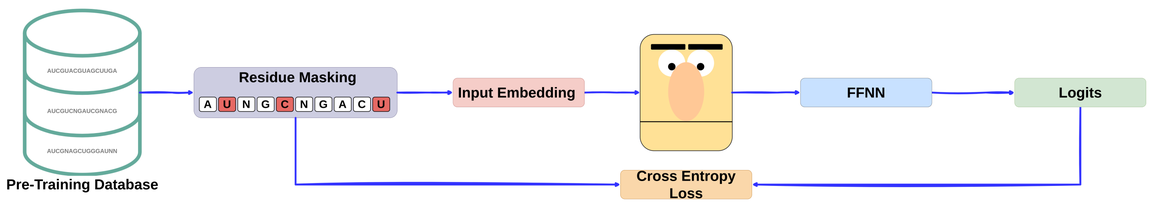


Attention Mechanism (Multi-Residue)

TRAINING APPROACH

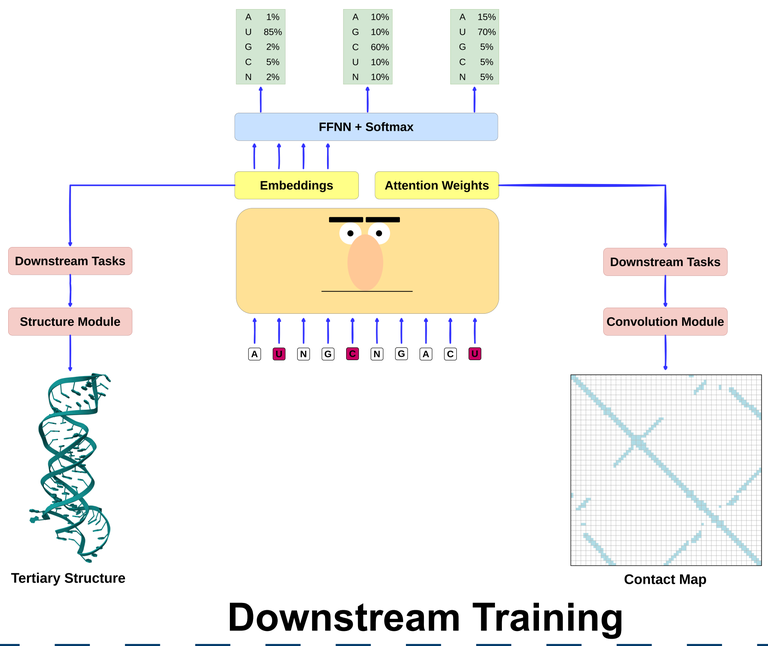
We are using standard masking technique, where -

- 15% of the overall tokens are randomly selected for prediction
- 80% out of the 15% are replaced with '[MASK]'
- 10% out of the 15% are replaced with a random token
- 10% out of the 15% are unchanged.



Pre-Training

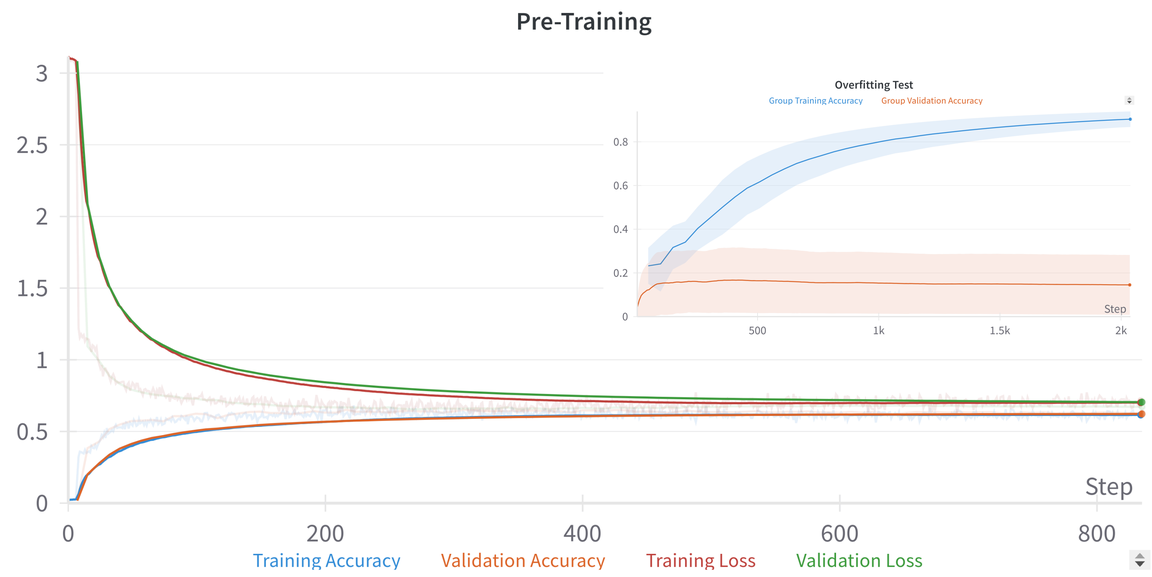
We finetune this pre-trained model for contact map prediction but is not straightforward, we can't directly use computer vision techniques. We experiment with different downstream modules.



Downstream Training

RESULTS

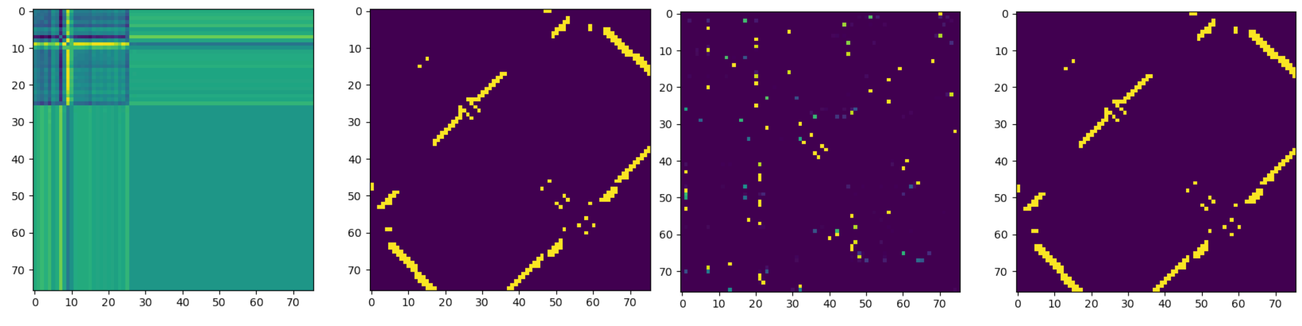
We pre-trained our model on 250K RNA sequences.



For downstream training, we prepare targets using annotated structures in the PDB database

Using this pre-trained model for downstream tasks allows us to predict contact maps directly from the sequence of RNA.

Here we can see how downstream training helps the model learn better



REFERENCES

Taubert, O. (2023). RNA contact prediction by data efficient deep learning. Nature. https://doi.org/10.1038/s42003-023-05244-9

Devlin, J. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. https://arxiv.org/abs/1810.04805

A. Elnaggar et al., "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 7112-7127, 1 Oct. 2022, doi: 10.1109/TPAMI.2021.3095381.