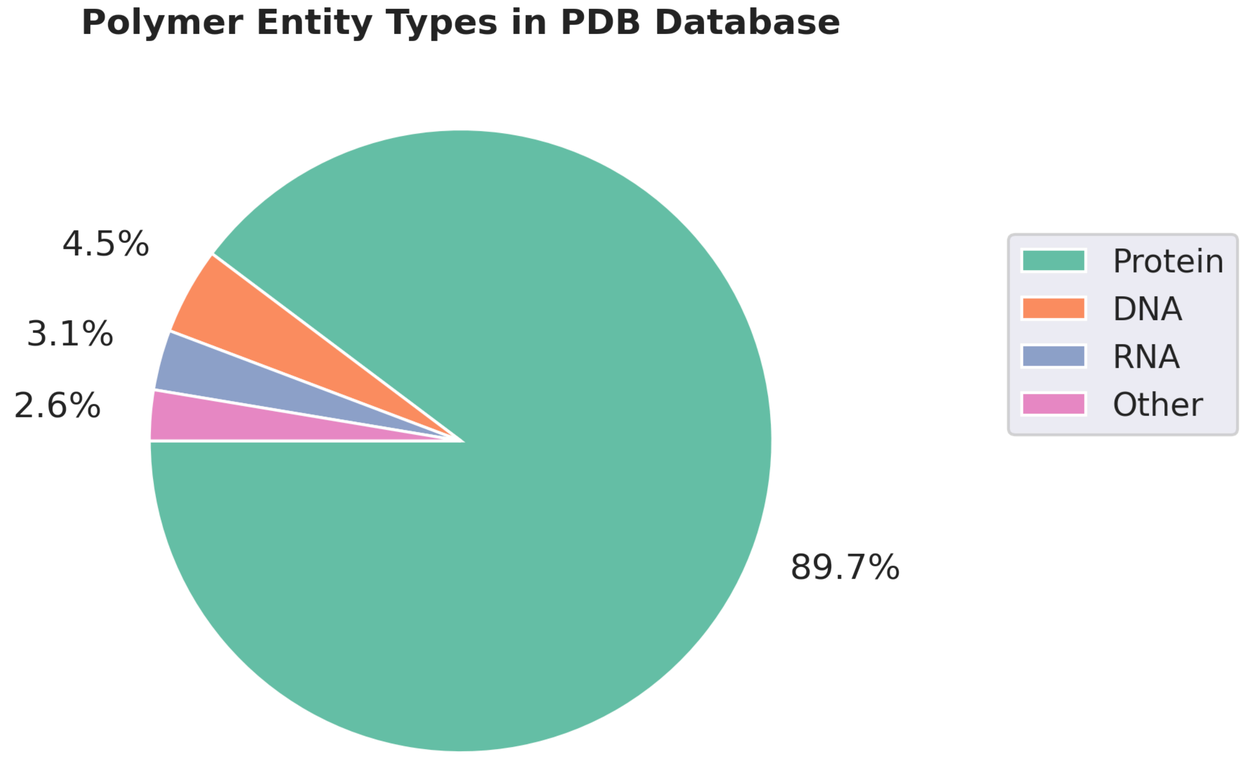
NucleicBERT: Deciphering The Language of Nucleic Acids

Utkarsh Upadhyay - Jülich Supercomputing Centre, Germany
Alexander Schug - Jülich Supercomputing Centre, Germany

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 a large language model trained on abundant RNA sequence datasets and use its representation for supervised tasks.

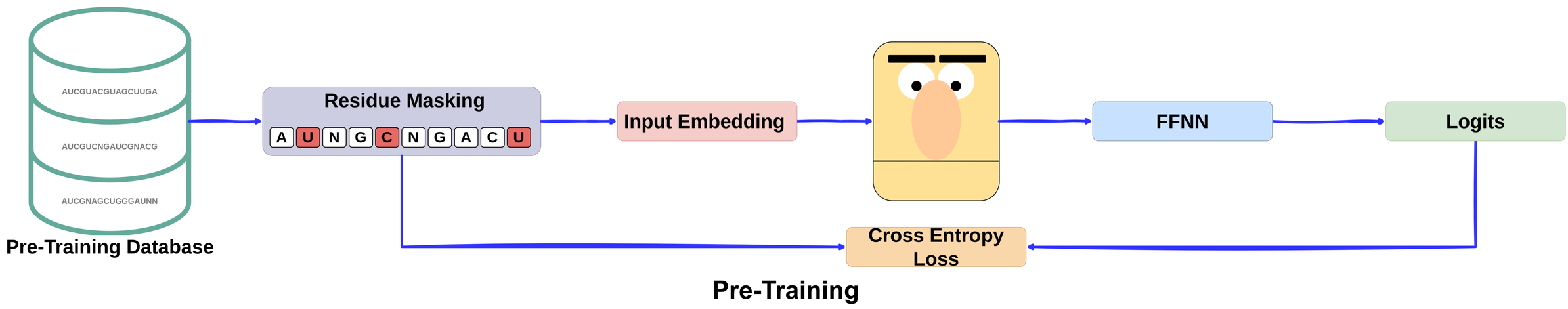


TRAINING

Each nucleotide is treated as a token (word) and each RNA sequence is like a sentence. BERT is based on the masked language model (MLM), which is trained to reconstruct the random masked parts of a sequence based on the surrounding context. This bidirectional context understanding helps BERT capture rich contextual information. We believe that by using this information in the form of embeddings as input for supervised tasks, we can perform per-nucleotide prediction tasks for which experimental data is available.

Non-Coding RNA	30944624
DNA	1910781
mRNA	52988782
Unspecified	1116154779

Pre-Training Dataset



We finetune this pre-trained model for three different downstream tasks: contact map prediction, distance map prediction and secondary structure. Each task requires a different neural network architecture and different objective functions.

REFERENCES

Upadhyay, U., Pucci, F., Herold, J., & Schug, A. (2024). NucleoSeeker: Precision Filtering of RNA Databases to Curate High-Quality Datasets. bioRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2024.12.06.626307

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

Fu, L., Cao, Y., Wu, J., Peng, Q., Nie, Q., & Xie, X. (2021). UFold: fast and accurate RNA secondary structure prediction with deep learning. Nucleic Acids Research, 50(3), e14. https://doi.org/10.1093/nar/gkab1074

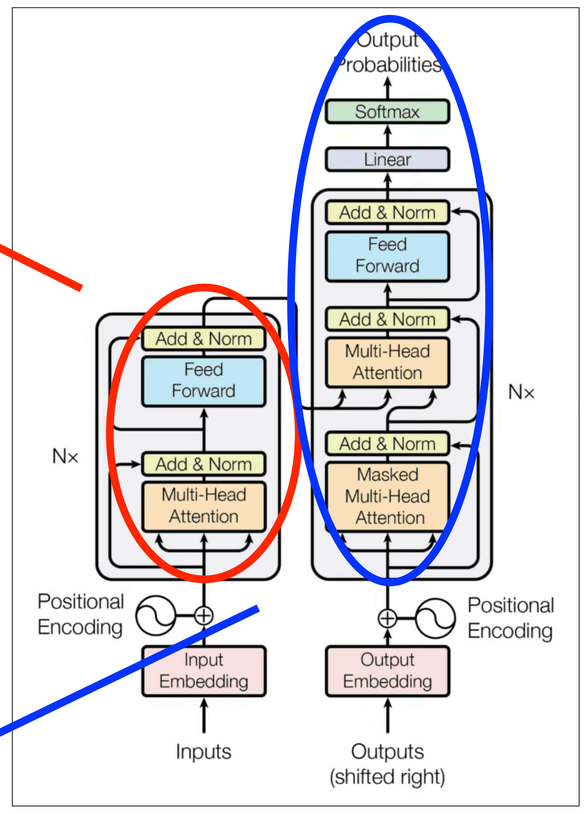
INTRODUCTION

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.

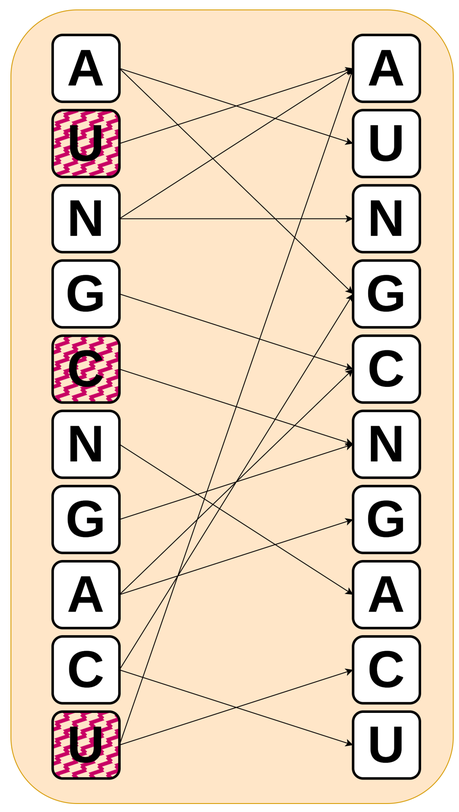
By this training the model captures interrelation in the residues of an RNA sequence. This information can then be finetuned for various other tasks to generate contact maps, distance maps, secondary structures and 3D structures as well.

BERT (Bidirectional Encoder Representations from Transformers) captures context bidirectionally, providing a deeper understanding of sequence interactions.

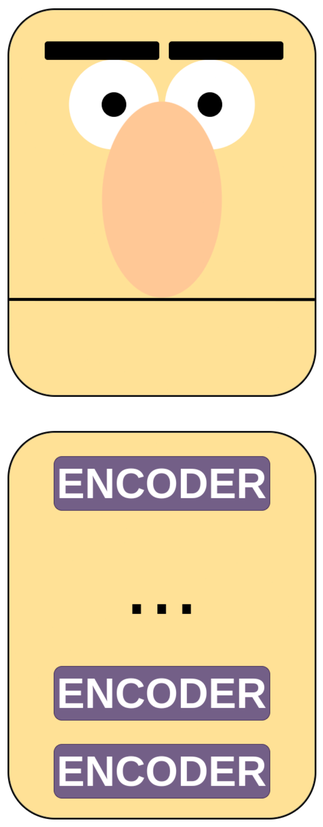
GPT (Generative Pre-trained Transformer) processes sequences unidirectionally, generating predictions from left to right.



Attention Mechanism



Attention Mechanism



BERT

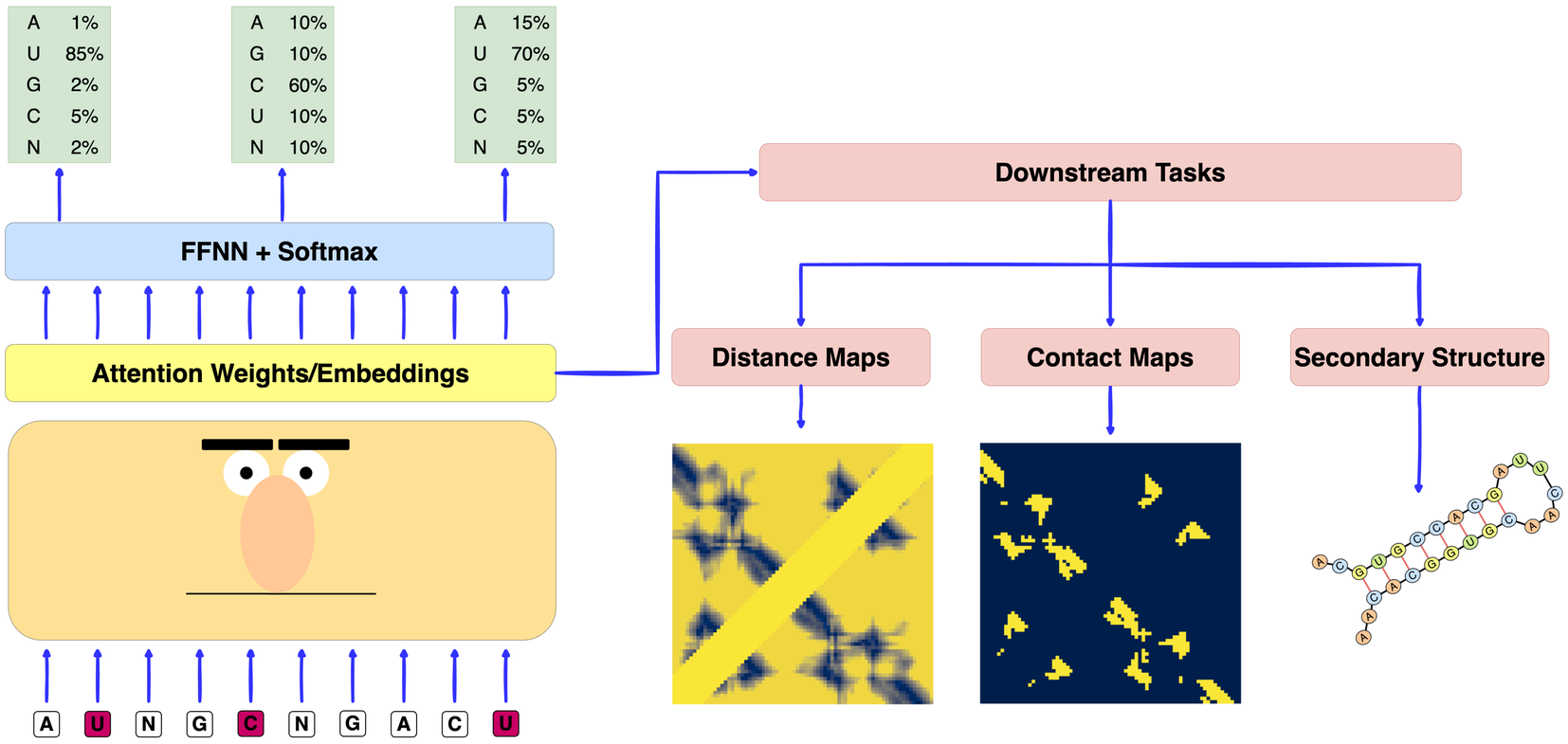
RESULTS

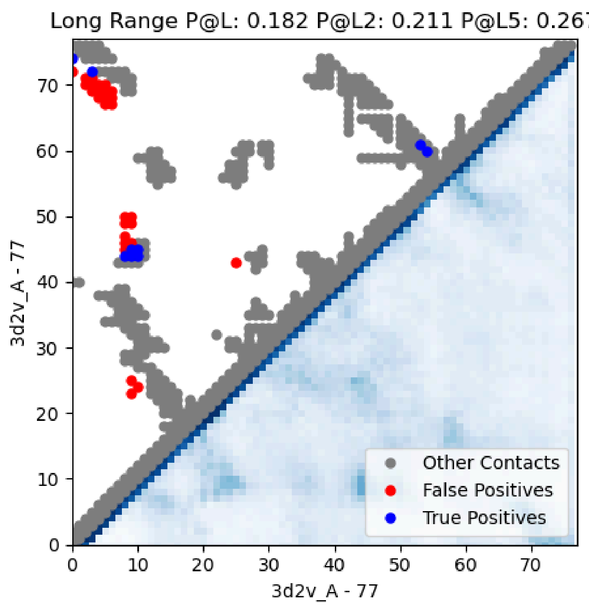
We created our dataset using NucleoSeeker for contact map prediction and the dataset from UFold for secondary structure prediction.

With the help of this model, we avoid the step of MSA creation and directly predict structural properties from the sequence.

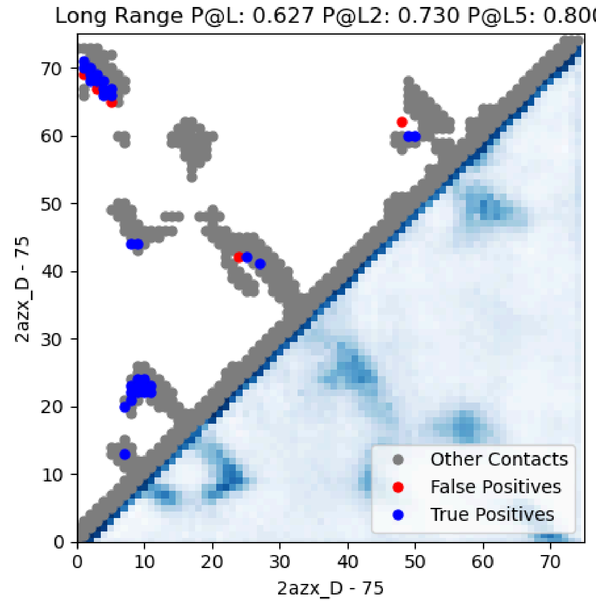
We start with distance map prediction, where distances are binned into 20 classes. Later, these distance classes are used for contact map prediction as well.

Secondary structure module is trained on WUSS notation of each sequence in the training dataset





Performance of NucleicBERT on Contact Prediction



CONTACT

Utkarsh Upadhyay
E-Mail: u.upadhyay@fz-juelich.de

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputers JUWELS and JURECA at Jülich Supercomputing Centre (JSC).