# Transformer Language Models for Genomic Sequences

Šimeček Petr1, Martínek Vlastimil1, Čechák David1

1 Central European Institute of Technology (CEITEC) MU, Brno

Since long short-term memory (LSTM) architecture, neural networks have been proven helpful for natural language processing (NLP) tasks. But what about the language of genomic sequences written in a four-letter alphabet? In the last few years, researchers have demonstrated that neural networks can be used to learn to identify functional elements in genomic sequences, and that the resulting models can predict the function of previously uncharacterized genomic regions.

The "revolution" had come in 2018 with the transformer architercure having the ability to detect complex dependencies between elements of a series thanks to a mechanism of attention or self-attention. This talk will review known pre-trained models, explain how they can be fine-tuned to a specific task, and show the usage on a couple of genomic benchmarks. We will demonstrate that the embeddings (=numeric representations) of genomic sequences contain additional information that the model has not been implicitly taught. We conclude by discussing practical aspects like the hyperparameter search and deployment to Hugging Face Models / Spaces.

*Maximum abstract length is ½ of A4 page.*