# Genomic Benchmarks: A Collection of Datasets for Genomic Sequence Classification

Grešová Katarína1, Šimeček Petr1, Martinek Vlastimil1, Čechák David1, Alexiou Panagiotis1

*1Panagiotis Alexiou Research Group, Centre for Molecular Medicine, Central European Institute of Technology, Masaryk University, Brno, Czechia.*

Advances in Next Generation Sequencing have allowed quicker, cheaper and more precise sequencing of many species' genomes. However, to extract biologically significant information from obtained genomes, we need to identify various types of functional genomic elements within them. Experimental genomic annotation is expensive, time consuming and research is focused on only a few model organisms, resulting in a low number of well annotated species and a high number of sequenced, but unannotated species. Machine learning models are able to learn and generalize given information. They may be trained to learn structure of functional elements in well annotated species and then use this information to annotate new genomes from given species or even new species (i.e. cross-species annotation).

However, machine learning models are highly dependent on large amounts of high quality data to train. It is also challenging to compare the quality of different models since authors often use different datasets for evaluation and quality metrics may also be heavily influenced by data preprocessing techniques and other technical differences.

Here we propose Genomic Benchmarks: A collection of curated and easily accessible sequence classification datasets in the field of genomics. The proposed collection is based on a combination of existing datasets obtained from published papers and novel datasets constructed from mining publicly available databases. The main aim of this effort is to create a repository for shared datasets that will make machine learning for genomics more comparable and reproducible, while reducing the overhead of researchers that want to enter the field. Our repository will be especially useful for researchers with backgrounds in machine learning and statistics that aim to implement state of the art machine learning algorithms in the genomic fields but are limited by lack of domain knowledge.