# Genomic Benchmarks QC: Automated Quality Control for Genomic Machine Learning Datasets

Maršálková Eva1,2, Grešová Katarína3,4, Šimeček Petr1, Alexiou Panagiotis3,4

*1 Central European Institute of Technology, Masaryk University, Brno, Czech Republic*

*2 National Centre for Biomolecular Research, Faculty of Science, Masaryk University, 61137 Brno, Czech Republic*

*3 Centre for Molecular Medicine and Biobanking, University of Malta, Msida, Malta*

*4 Department of Applied Biomedical Science, Faculty of Health Sciences, University of Malta, Msida, Malta*

As machine learning becomes increasingly prominent in genomics, the quality and composition of training data become crucial to developing robust and generalizable models. While AlphaFold2’s success in structural bioinformatics was made possible by a carefully curated benchmark of experimentally determined protein structures, genomics still lacks comparable, high-quality datasets for many of its key challenges. When training machine learning models, they can become only as good as the datasets they are built upon—if the training dataset contains biases, the models will learn those biases instead of capturing the biological features. This results in models that perform well on internal validation but fail to generalize across datasets. For instance, through simulated data, we can show that when GC-content bias exceeds a certain threshold, convolutional neural networks begin to learn this bias instead of recognizing actual sequence motifs.

To address this issue, we present a tool for automated quality control of genomic datasets. It evaluates properties such as nucleotide composition, GC-content, sequence length distribution, and duplication rates, producing a detailed report that helps researchers check, detect and mitigate misleading data patterns. The tool is easily accessible through GitHub [1] and as a Python package, designed to reduce overhead and promote the development of biologically meaningful, high-quality genomic datasets. We have applied this tool to the datasets from our previous work in Genomic Benchmarks v1 [2] to evaluate their quality and identify potential candidates for inclusion in a release of an improved and unbiased version 2 of this benchmarking dataset collection.

[1] <https://github.com/katarinagresova/GenBenchQC>

[2] Grešová, K., Martinek, V., Čechák, D. et al. Genomic benchmarks: a collection of datasets for genomic sequence classification. BMC Genom Data 24, 25 (2023). https://doi.org/10.1186/s12863-023-01123-8