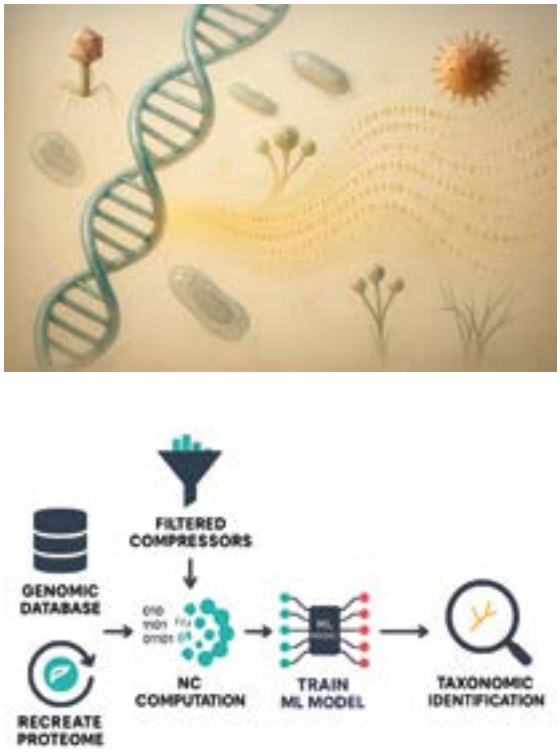
Enhancing metagenomic classification with compression-based features

Jorge Miguel Silva[1], João Rafael Almeida[1]

Metagenomics relies heavily on next-generation sequencing but efficiently classifying novel or complex microbial samples can be challenging, especially when relying on cumbersome, reference-based methods. In our study, we propose a streamlined, alignment-free strategy that captures how compressible each sequence is (both at the DNA and amino-acid level) as a predictive signature for taxonomic identification. By analyzing the "compression fingerprint" produced by 16 well-known data compressors, we effectively identify organisms across viruses, bacteria, archaea, fungi, and protozoa, achieving up to 95% classification accuracy. This unified approach saves significant computational effort by sidestepping extensive reference databases, and it also excels in low-coverage of highly diverse genomes (such as protozoa and unknown viruses) due to the use of data compression. The resulting compression-driven pipeline, validated on thousands of public-domain sequences, shows how the interplay of data compression, machine learning, and computational genomics can offer a robust new lens for metagenomic research.



1 – IEETA & Department of Electronics, Telecommunications and Informatics, University of Aveiro.

FIGURE 1
Compression-Based Taxonomic Identification Pipeline.