Sustainable AI Data Compression Case Study: Profiling BOA on BEDE

Supervisors: Prof. Caterina Doglioni (caterina.doglioni@manchester.ac.uk), Senior RSE and SSI Fellow Michael Sparks (michael.sparks@manchester.ac.uk), PhD students Akshat Gupta and Sanjiban Sengupta, University of Manchester

Project Overview: This 8-week project will develop a reproducible case study to profile the environmental costs of a novel deep learning workflow for data compression, embedded in the UKRI-funded **Sustainable Training Initiative for NetDRIVE (SusTraIN)** project. The SusTraIN project aims to transition Digital Research Infrastructures (DRIs) toward net-zero by creating training materials for researchers and Digital Research Technical Professionals (dRTPs). SusTraIN is **led by three PIs at the N8-CIR Universities of Manchester, Liverpool and Sheffield.**

Overarching goal: The intern will create a **training resource** in the form of an **executable workflow** that **quantifies the energy consumption** of an advanced compression technique for high energy physics data that requires large-scale AI training to be run on a High-Performance Computing (HPC) system such as Bede. *BOA Constrictor* (https://arxiv.org/abs/2511.11337, submitted to Machine Learning in Science and Technology) is a Mamba-network based lossless data compressor, and an open source project led by two Early Stage Researchers (PhDs at the University of Manchester) and co-supervised by the academic supervisor of this project.

Project context and outcomes: Modern AI/ML models are among the primary drivers of the rising environmental impact in digital research. In addition to obtaining a **robust estimate of the energy consumption** for the training and inference of data compression with the BOA algorithm, the outcomes of this project address a **training gap** identified in the SusTraIN project: researchers and dRTPs often lack the practical tools and knowledge to measure and mitigate this energy consumption. This project complements the SusTraIN case studies with an AI-native, data-intensive reproducible workflow that can be executed on a UK supercomputer.

Learning outcomes and theme alignment This project provides the student with training in both the **Machine Learning** theme (implementing and running BOA) and **Research Software Engineering** best practices (developing reproducible, version-controlled code and FAIR metadata).

Work environment, supervision and interaction with other students/members of the group: the student will be supervised by the PI and by a senior Research Software Engineer (M. Sparks) and will work within a **cohort of four other interns** (supported by EU, UKRI and Google Summer of Code funding) f**ocused on Research Software Engineering and Green Computing**, supervised by PDRAs and dRTPs within the group of the PI and collaborators. The successful experience of previous years' cohorts demonstrates that this is an environment where students can thrive and receive 1:1 supervision on a daily basis, and exchange knowledge in a meaningful way for their project. This project will receive **direct input** from one of the other students who will be developing **simulations of the environmental impact of embedded carbon** in hard disks needed to store extra data if compression is not implemented.

Dissemination: This internship will produce quantifiable results that will be shown at the EuCAIF Conference in late August 2026 (https://indico.physi.uni-heidelberg.de/event/1277/), and become part of **SusTraIN's Work Package 2 (WP2)** that will include hands-on tutorials and case studies ready for inclusion in the EVERSE-TeSS training catalogue and in the SPARK-HUB UKRI platform, ensuring visibility across the UK and European research communities.