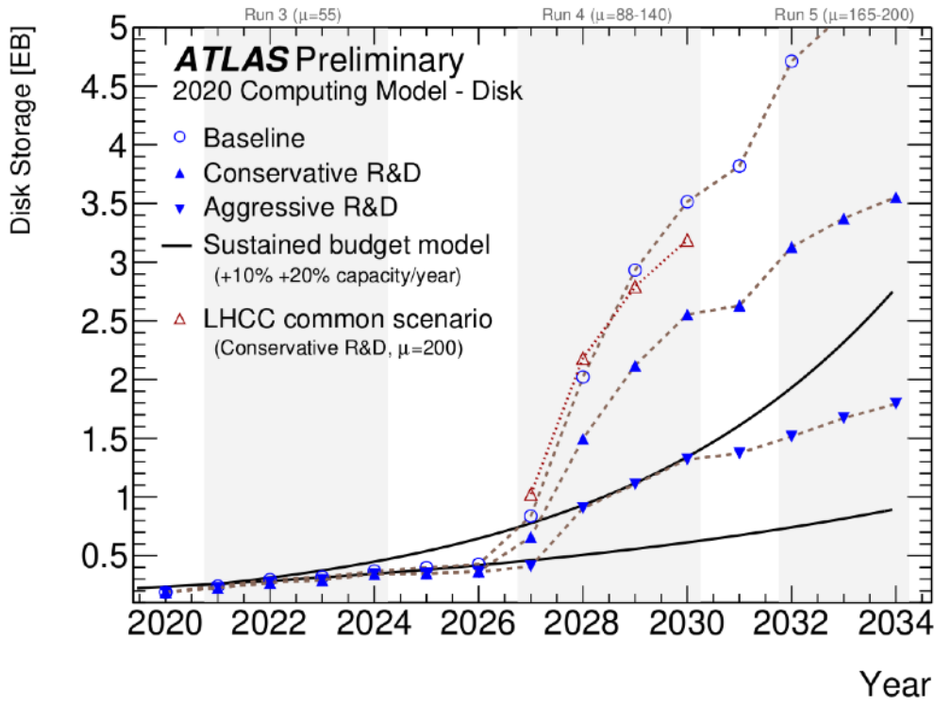
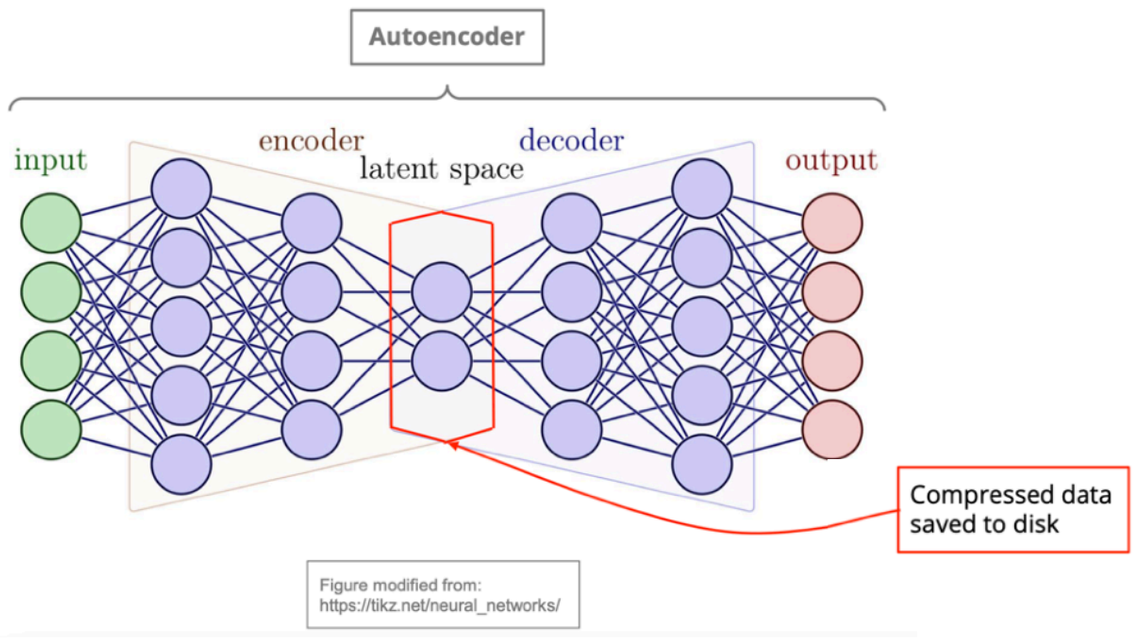
Master's thesis using Machine Learning

Compress particle physics data and more, and contribute to the *Baler* Open Source Software!

Motivation

At the ATLAS experiment, the projected data production in 2032 will be **6 times larger than the storage available for the given budget** [1]. This is issue is not unique to the ATLAS experiment. Storing and sharing increasingly large datasets is a challenge across disciplines in **scientific research and industry**. There is a high demand, inside and outside particle physics, to **effectively compress data more than conventional loss-less methods** do.





The idea and the tool

Baler *is* a software tool that started as a Lund Master's project that uses machine learning (autoencoder) to derive a **lossy compression method that is tailored to the user's input data**, achieving large data reduction and high fidelity where it matters. Baler has already been used to compress scientific research data [2] (note: N authors did a Master's project on Baler).

The Master's projects

Master's students can choose among different research directions to improve Baler:

1. To extend the compression to different data, e.g. photon science, neutrino physics.
2. To understand how the network generalises results when compressing results that are different with respect to the data it has been trained on.
3. To study how to minimise the environmental impact of Baler, by using network optimization techniques To understand how the network makes decisions on which variables to compress the most (interpretability), in collaboration with computer scientists.
4. To reproduce a full physics analysis with compressed data.

Who you will be working with

The Baler project is a **collaboration** between **physicists**, **computer scientists**, and **machine learning experts** at the universities of Lund, Manchester, Salford, Warwick and Uppsala, Tel Aviv, and Ohio State. This project will be based at the **Physics division of Lund University** and supervised by Oxana Smirnova, and co-supervised by Caterina Doglioni at the **University of Manchester**. Contact us at oxana.smirnova@fysik.lu.se and caterina.doglioni@manchester.ac.uk.

[1] Expected data storage needs as a function of expected data storage available for the ATLAS experiment. http://cds.cern.ch/record/2729668/files/, and CERN Courier article https://cerncourier.com/a/time-to-adapt-for-big-data/
[2] https://arxiv.org/abs/2305.02283 and description by one of the Research Software Engineers who worked with us: https://research-it.manchester.ac.uk/news/2023/05/15/baler-using-ai-to-help-efficiently-store-data-in-a-big-data-age/