Autoencoder based compression for high energy physics

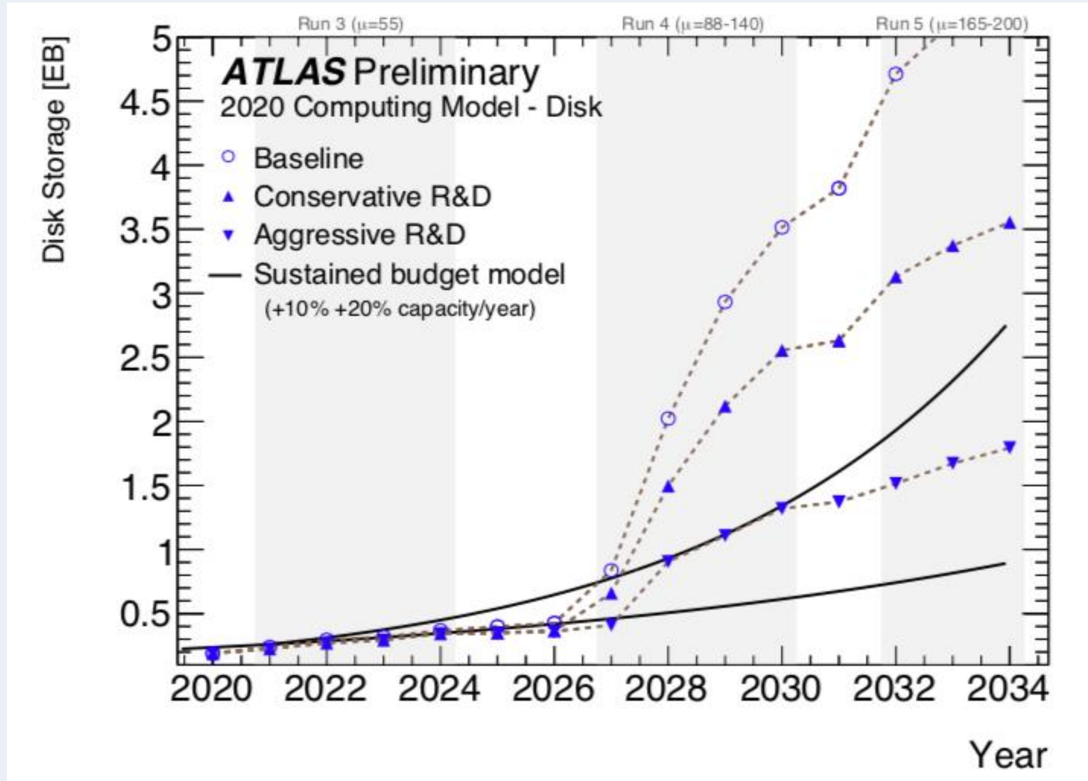
Antonio Boveia[1], Caterina Doglioni[2], Honey Gupta[3], Lukas Heinrich[4], Love Kildetoft[2], Baptiste Ravina[5], Erik Wallin[2], Eric Wulff[2]
[1] Ohio State University, [2] Division of Particle Physics, [3] Google Summer of Code (HEP Software Foundation) 2020, [4] CERN, [5] University of Glasgow,

Abstract

Data is what drives all research forward, regardless of scientific discipline. At the Large Hadron Collider (LHC) in Switzerland, the data stream of registered events can reach about 60 million megabytes per second, making it physically impossible to save all the produced data with current storage technology. This means that data selection has to be performed at an early stage in the experimental process, using so-called *trigger* systems. While these systems are sophisticated, it is possible that rare processes with high-rate backgrounds are discarded due to lack of storage space. This could potentially mean missing out on new discoveries, and thus a risk of not reaching the goal of the LHC in itself. Data compression techniques can reduce the size of data while giving a sufficiently faithful representation of the uncompressed data. For high-energy physics, using new data-compression techniques as a part of the data selection process would allow for further storage savings without waiting for major technological advancements in storage media, and thus increasing the amount of data that can be recorded. One of the compression techniques that have been under recent investigation uses autoencoder networks. This is a machine-learning based approach, which utilizes a certain type of neural network known as an autoencoder. Autoencoders are, in their most basic form, a neural network with multiple layers where the number of inputs is equal to the number of outputs, and the input and target datasets are the same. If the dimension of the hidden layer is (much) smaller than the dimension of the input and output layers, an autoencoder will be tasked with finding an effective representation of the input data, which can then be reconstructed in the output layer. As such, autoencoders area good candidate for data compression. In this contribution, autoencoder-based compression will be evaluated, by using the results presented in the Master's thesis *Deep Autoencoders for Compression in High Energy Physics* by Eric Wulff as a foundation [1], as well as work from Erik Wallin (Bachelor's student, 2020) [2], Honey Gupta (Google Summer of Code student in 2020 with the High Energy Physics Software Foundation) [3] and Love Kildetoft (Bachelor's student, 2021) [4].

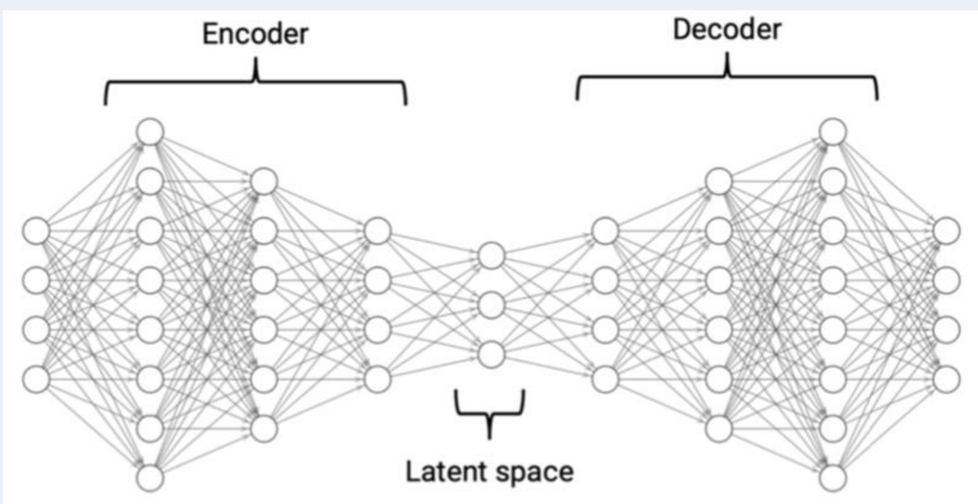
Why data compression?

- The LHC provides billions of proton-proton collisions per second, corresponding to a extremely large amount of data.
- It is not possible to save all this data and the corresponding simulation due to storage constrains within the current budget, especially in Run-4 and Run-5 (see Fig. 1)
- *Data compression techniques* are already in place for LHC experiments, either *lossless* (where the original precision is retained) or *lossy* (where some of the original precision is lost, but the compressed data is still sufficient for analysis)
  - Can we find **new compression techniques** with a better performance than current ones?
  - Can we implement those techniques **directly in the experiment's data selection (trigger) system** to record more LHC data?



Projected disk storage requirements of ATLAS between 2020 and 2034
Source: ATLAS Public Computing & Software Results.

Data compression using machine learning: *autoencoders*



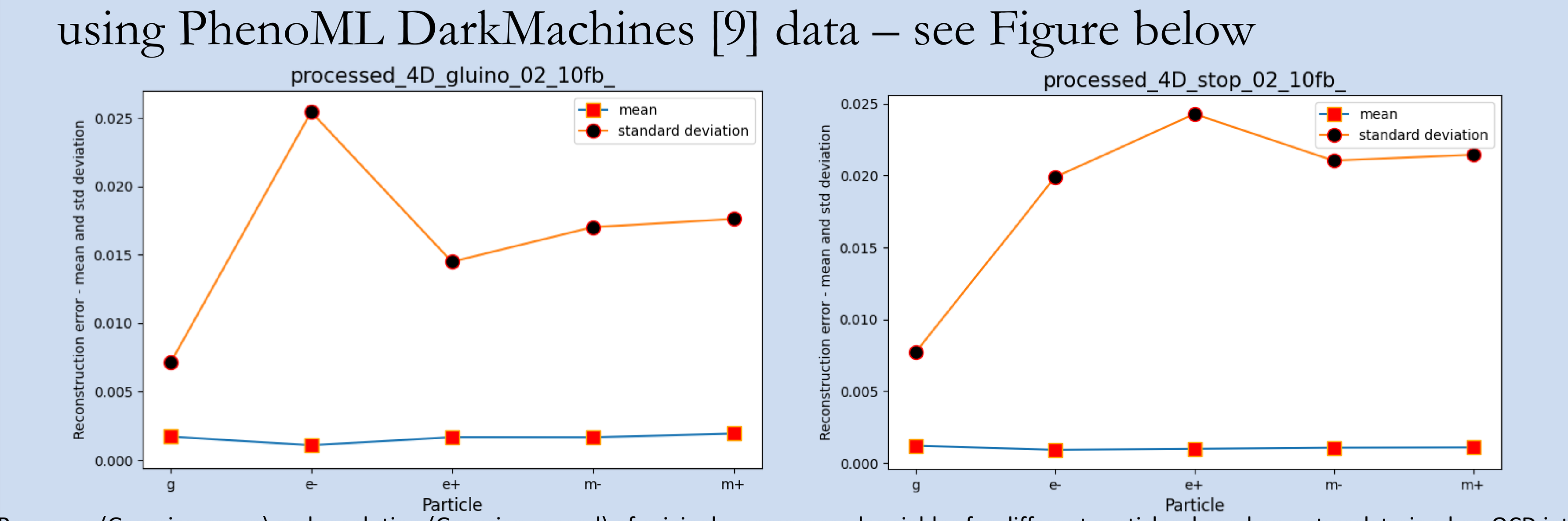
Sketch of an autoencoder architecture
Source: Erik Wulff's thesis [1]

- Autoencoders are a class of neural networks [4] which can be used for data compression, as their purpose is to provide an effective representation of the input data
- The basic working principles of an autoencoder are:
  - It has an equal number of inputs and outputs
  - It has a "latent space" in the middle, where the dimension is smaller than the input/output dimension.

  → As such, an autoencoder performs a dimensionality reduction of the data, and then reconstructs it back to the original number of dimensions, effectively finding an effective yet compressed representation
- The other advantages of an autoencoder are that:
  - They can be used with minimal or no supervision in training → good for data taking
  - They can have rather small / simple network architectures → good for running on constrained resources (see also Ref. [6] and the upcoming ML4Jets conference session [7])

Highlights of selected studies

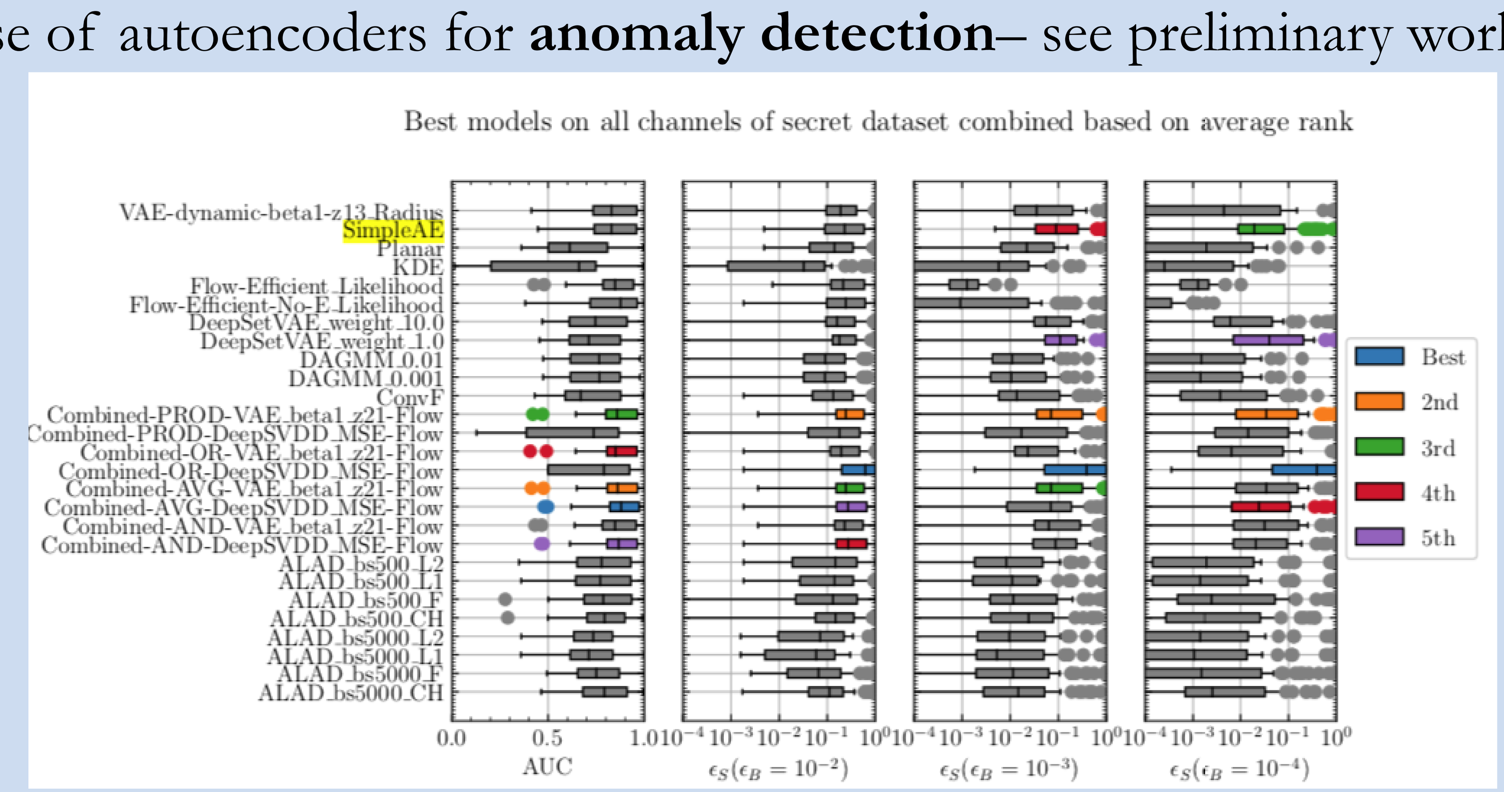
- **Prototype autoencoder** constructed for compression of jets → good performance of compression from 27 up to to 10 dimensions, with precision loss comparable to detector resolution [1]
- **Computational and memory cost** of AE compression in terms of time execution on a single machine, as well as chaining the compression and comparison with another technique known as float truncation [2] [3] → promising on a single CPU, need more tests on production systems \
- Use of network on **jet images**, mimicking calorimeter cells [8]
- How well the network performs on **event-level data**, not limited to jets, using PhenoML DarkMachines [9] data – see Figure below



Response (Gaussian mean) and resolution (Gaussian spread) of original – compressed variables for different particles, based on network trained on QCD jets and applied to new physics models. Performance better than detector resolution (~5%). Source:Honey Gupta's GSoC project [4]

Outlook

- More thorough testing to put current **network in production** to be considered for HL-LHC trigger data compression
- Study **event-level networks** (ongoing, Master's project by S. Astrand)
- Study different autoencoder **architectures** (ongoing, Google Summer of Code / HSF 2021, G. Dialektakis)
- Further tests of **raw data** compression (e.g. calorimeter cells)
- Use of autoencoders for **anomaly detection**– see preliminary work in [9]



Performance of simple autoencoder (SimpleAE) used for compression with hyperparameters tailored for data compression with respect to other methods in the DarkMachines Anomaly Score Challenge. Source:DarkMachines Anomaly Score Challenge [9]

For further reading

[1] Deep Autoencoders for Compression in High Energy Physics, Eric Wulff, Link *[2] Tests of Autoencoder Compression of Trigger Jets in the ATLAS Experiment,* Erik Wallin, Link *[3] Deep-compression for High Energy Physics data,* Honey Gupta Link *[4] Evaluation of float-truncation based compression techniques for the ATLAS jet trigger,* Love Kildetoft Link *[5] The rise of deep learning,* CERN Courier, July 2018, https://cerncourier.com/the-rise-of-deep-learning, *CERN/ATLAS resources on machine learning:* https://iml.web.cern.ch, https://cerncourier.com/learning-machine-learning/, https://atlas.cern/tags/machine-learning *[6] A reconfigurable neural network ASIC for detector front-end data compression at the HL-LHC,* G. Di Guglielmo et al. *https://arxiv.org/abs/2105.01683* *[7] ML4Jets Conference 2021 https://indico.cern.ch/event/980214/* *[8] Investigation of Autoencoders for Jet Images in Particle Physics,* J. Lastow *Link [9] PhenoML / Darkmachines dataset,* Darkmachines collaboration, *https://arxiv.org/abs/2105.14027*