BALER: Machine-Learning-Based Compression of Scientific Data in Real Time

James Smith, University of Manchester

EuCAIFCon 2025, Cagliari

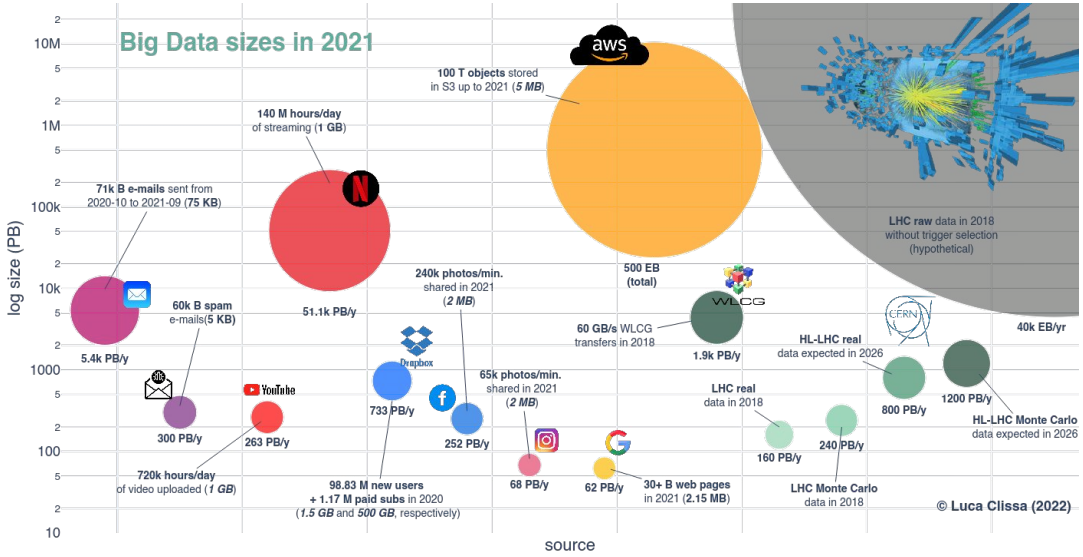
Currently active team:
Caterina Doglioni, Elena Gramellini (Academics),
Pratik Jawahar (PhD students),
Khwaish Anjum, Keerath Dhariwal, Oscar Fuentes, Sarah Hayat, Natalia Stepniak (Undergraduate/Master's students),
Leonid Didukh (Industry)

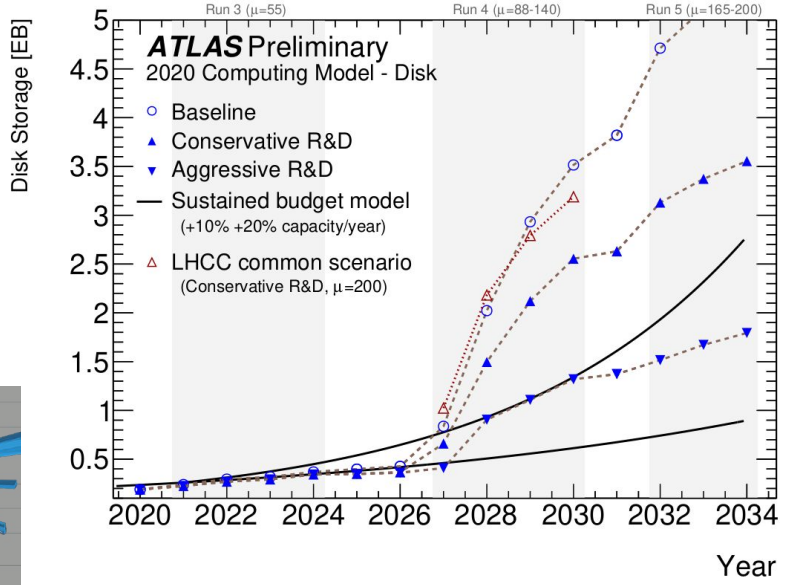
MANCHESTER 1824
The University of Manchester

European Research Council
Established by the European Commission

The Problem

- **Too much data, too little storage, bandwidth**
- Not unique to LHC Experiments
- High demand for compression





ATLAS HL-LHC Computing Conceptual Design Report
Calafiura, P ; Catmore, J ; Costanzo, D ; Di Girolamo, A
http://cds.cern.ch/record/2729668/

https://cloud.datapane.com/reports/dkjK28A/big-data-2021/ - Image by Luca Clissa

A Solution

- One approach: Lossy compression
- One problem: Lossy compression needs to be tailored
- Solution: **Lossy Machine Learning based compression**

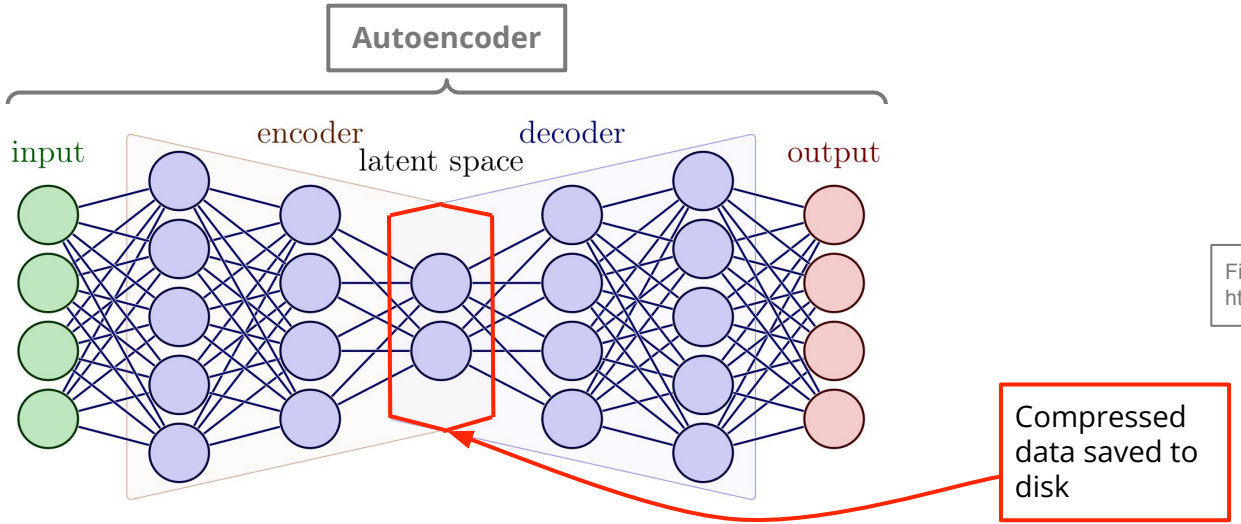


Figure modified from: https://tikz.net/neural_networks/

Our Tool: “Baler”

- We have created a tool called “**Baler**” to help investigate the viability of this compression
- **Multidisciplinary** tool
- Distributed and developed as an **open source project** [GitHub: baler-collaboration/baler]
- Simple to install as a command line tool
  - `Poetry run python baler --project=CMS --mode=train`
  - **Docker** and **pip** versions also available

https://arxiv.org/abs/2305.02283

Baler - Machine Learning Based Compression of Scientific Data

F. Bengtsson[1] C. Doglioni[2] P.A. Ekman[1] A. Gallén[1] P. Jawahar[2] A. Orucevic-Alagic[1] M. Camps Santasmasas[2] N. Skidmore[2] O. Woolland[2]

[1] Lund University
[2] University of Manchester

Abstract: Storing and sharing increasingly large datasets is a challenge across scientific research and industry. In this paper, we document the development and applications of Baler - a Machine Learning based data compression tool for use across scientific disciplines and industry. Here, we present Baler's performance for the compression of High Energy Physics (HEP) data, as well as its application to Computational Fluid Dynamics (CFD) toy data as a proof-of-principle. We also present suggestions for cross-disciplinary guidelines to enable feasibility studies for machine learning based compression for scientific data.

1 Introduction

Many different fields of science share a common issue; storing ever-growing datasets. By the end of the next decade, the Large Hadron Collider (LHC) experiments will have over an

[hep-ex] 3 May 2023

EPJ Web of Conferences All issues Series Forthcoming About

All issues › Volume 295 (2024) › EPJ Web of Conf., 295 (2024) 09023 › Abstract

Open Access

Issue	EPJ Web of Conf. Volume 295, 2024 26th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2023)
Article Number	09023
Number of page(s)	8
Section	Artificial Intelligence and Machine Learning
DOI	https://doi.org/10.1051/epjconf/202429509023
Published online	06 May 2024

EPJ Web of Conferences **295**, 09023 (2024)
https://doi.org/10.1051/epjconf/202429509023

Baler - Machine Learning Based Compression of Scientific Data

Fritjof Bengtsson Folkesson[1*], Caterina Doglioni[2**], Per Alexander Ekman[1***], Axel Gallén[1****], Pratik Jawahar[2†], Marta Camps Santasmasas[2‡] and Nicola Skidmore[2§]

CHEP 2023 proceedings

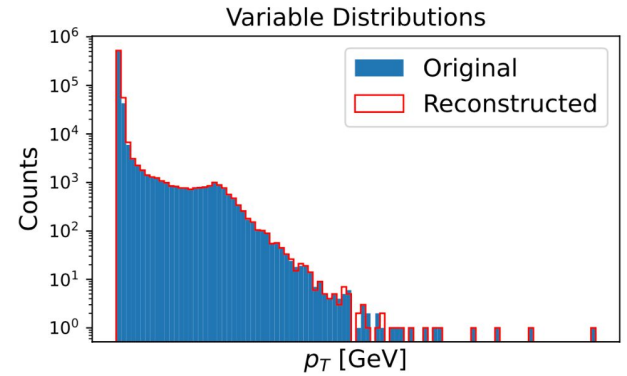
Model Details

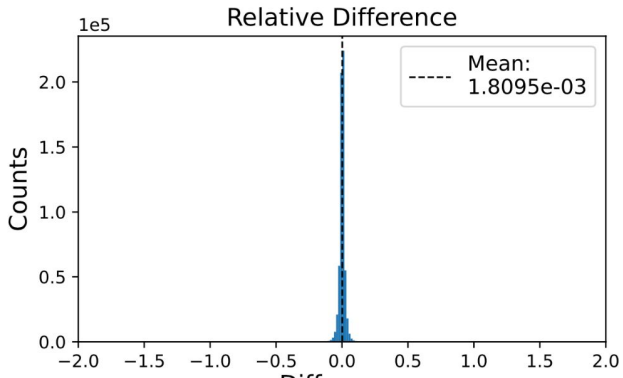
- Initially used simple AEs, but a range of models are supported
  - Any implementation supported by pyTorch is possible! (we also investigated TF's own compression libraries)
- CNN elements used, particularly useful for CFD/image data
- New Transformer-based model evaluated
  - 3 transformer encoder layers (X->256->128 dimensions)
    - Canned layers from pyTorch
  - 3 autoencoder layers (X->256->128->Latent space)
  - Mirrored decoder layers (3 autoencoder, 3 transformer
- Ideas for future models
  - GNN or HyperEdge networks for whole-event, multi-object compression?

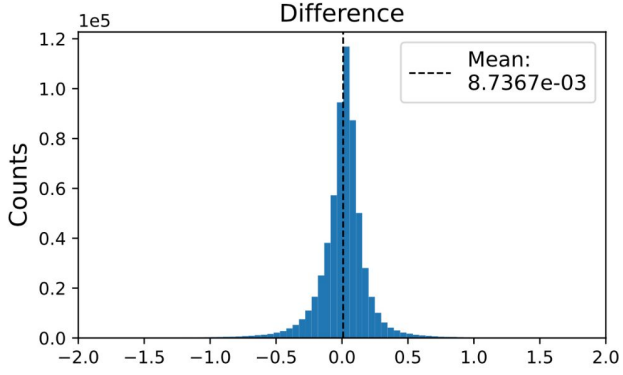
Results: Jet Transverse Momentum

- Open CMS Data
  - ~ 600 000 jets
- 24 variables per jet compressed to 14 variables
  - Transverse momentum one of these variables
- **58% original size**

DOI:10.7483/OPENDATA.CMS.KL8H.HFVH

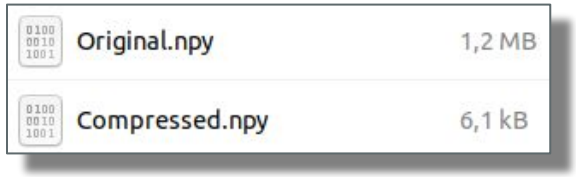


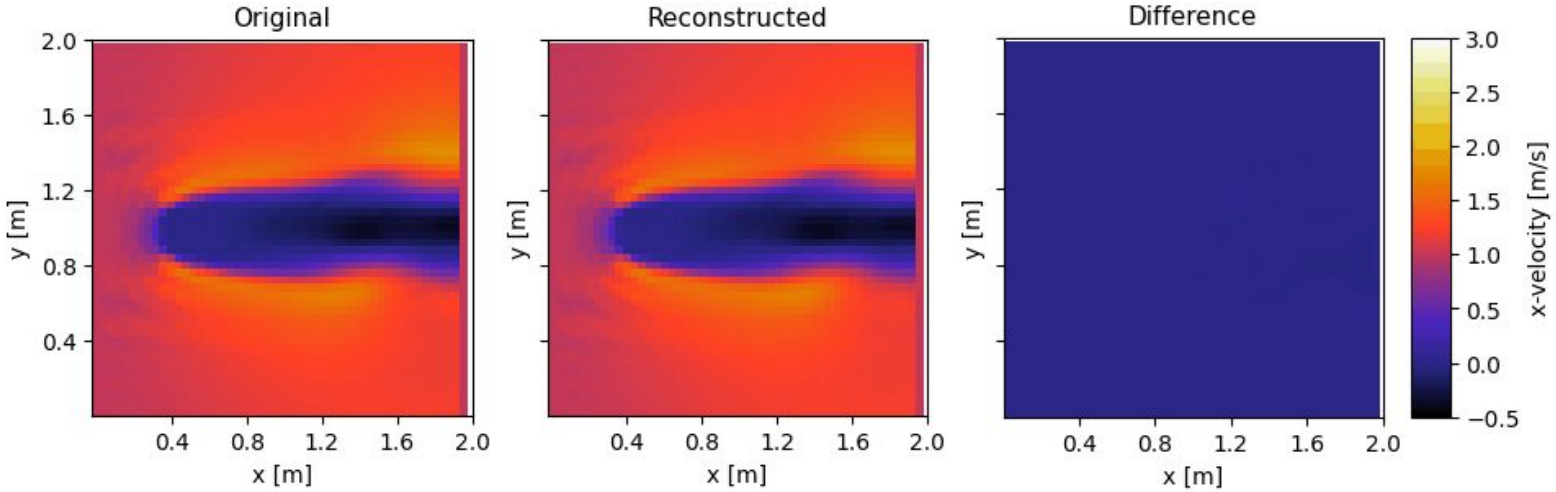




Results: CFD

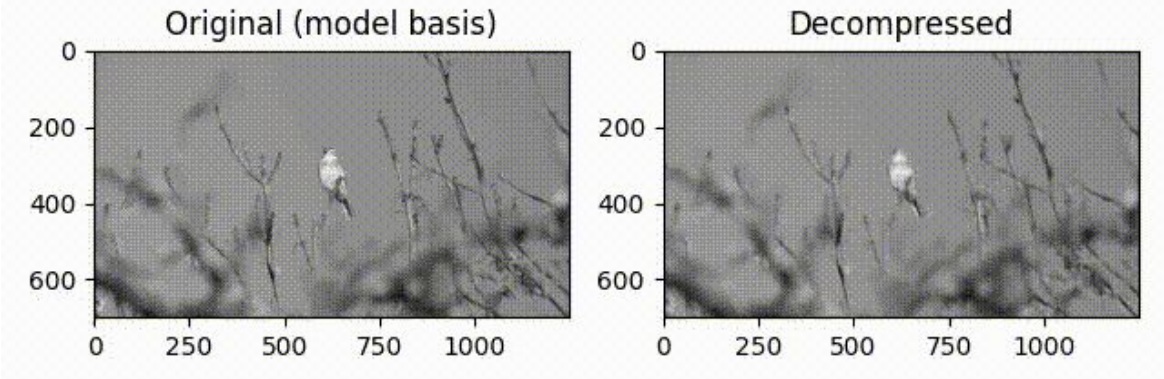
- Data consists of 2D slice of a liquid flowing over a cube
- The compressed file is **0.5%** the size of the input
- Model larger than input (4.2 vs 1.2 MB)





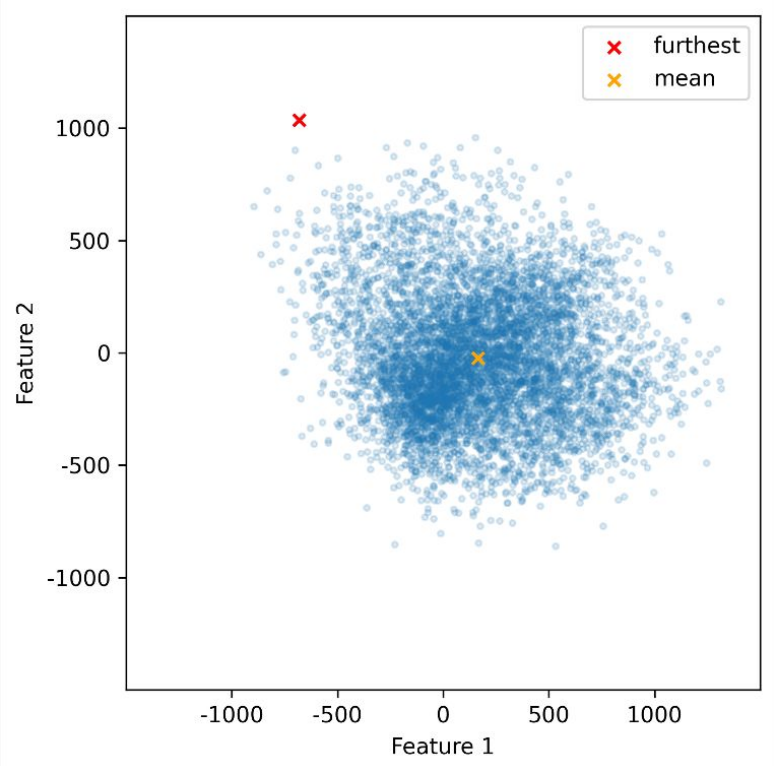
Online vs offline (Video)

- Previously applied model trained on one dataset to the same dataset (***offline***)
- Can also apply to similar but unseen datasets (***online***)
  - Eliminate the cost of the model size!
- Useful for compressing **live data** (triggers, networks, etc)



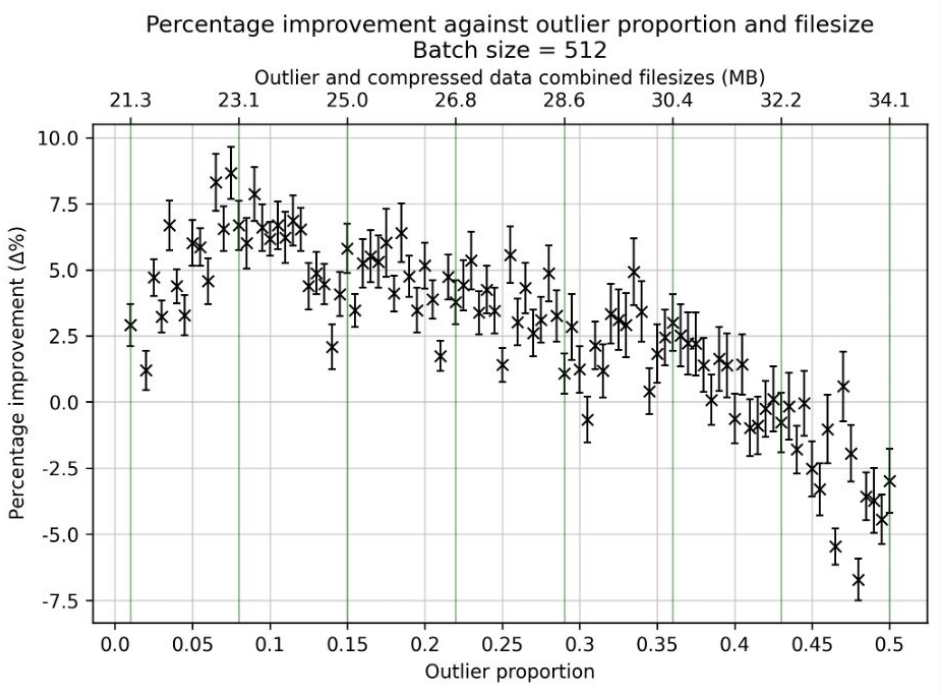
Anomaly Detection for Outlier Removal

- Online performance degraded by **outliers**
- Exploring use of **anomaly detection** to separate outliers
  - Outliers could be stored in full for further analysis
- Simplest form: remove X% of points furthest from mean in latent space
- Performance **evaluation ongoing**
- Also exploring clustering and categorisation



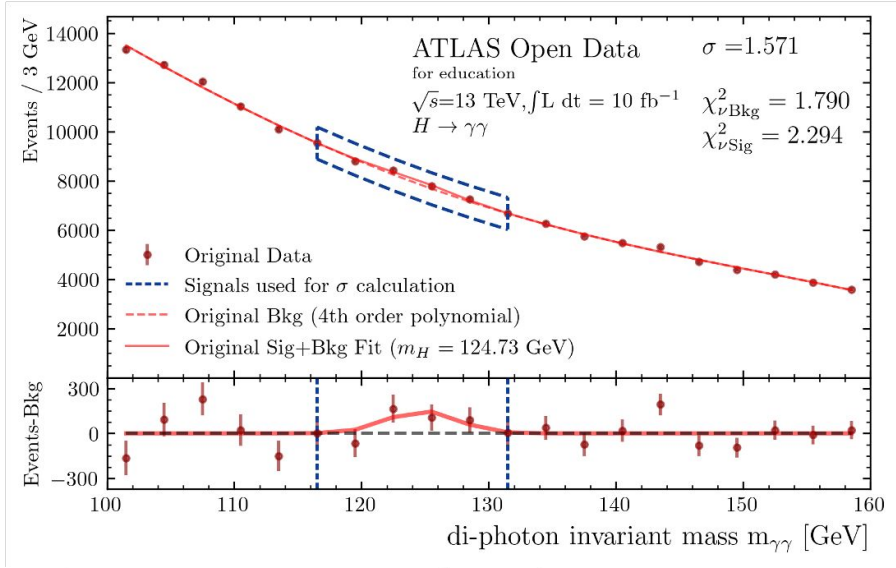
Anomaly Detection for Outlier Removal

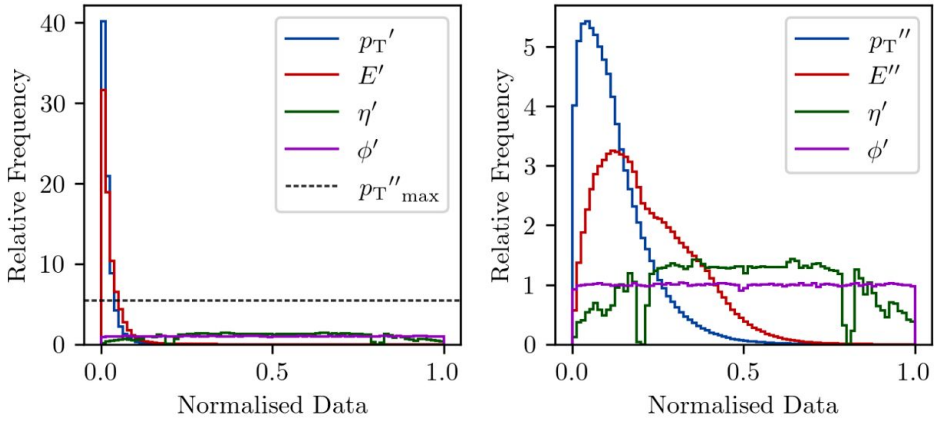
- Online performance degraded by **outliers**
- Exploring use of **anomaly detection** to separate outliers
  - Outliers could be stored in full for further analysis
- Simplest form: remove X% of points furthest from mean in latent space
- Performance **evaluation ongoing**
- Also exploring clustering and categorisation



End-to-End analysis

- Evaluate effect of compression on discovery sensitivity rather than just object reconstruction performance
- Use ATLAS Open Datasets & Education notebooks - can we rediscover the Higgs?
  - ~10 fb^{-1} of heavily-preprocessed Run-2 data, notebooks for H→γγ and H→ZZ→4l
- Normalisation required to bring variables to similar scales
  - min-max used for η and phi
  - log-scale min-max for p_T and E
- Evaluate both simple and transformer AEs



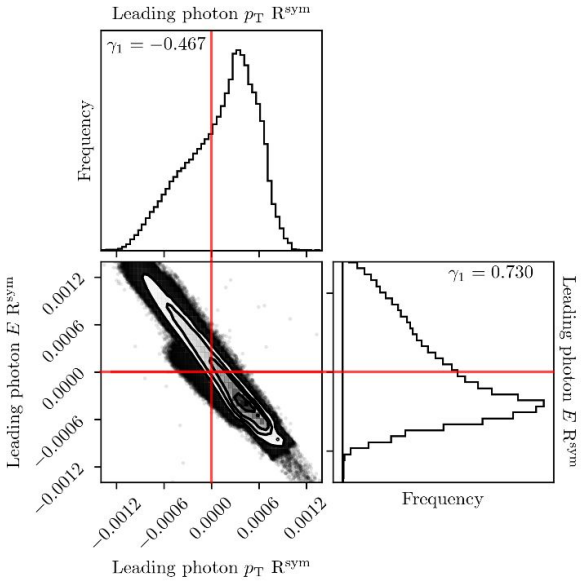


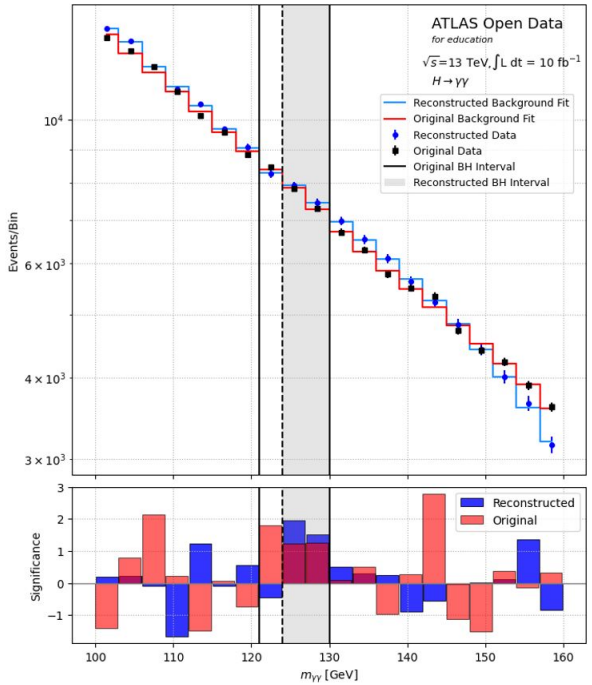
Results

Reconstruction errors in single values can be correlated, leading to greater errors in composite variables such as $m_{\gamma\gamma}$!

With tuning, decompressed dataset can yield similar significances and 25% smaller file size!

H→γγ, transformer





Compression used?	Global p-value	Z [σ]	File Size [MB]
No	0.1510	1.03 ± 0.03	474
Yes	0.1820	0.91 ± 0.05	356

Next steps

- Model, training and normalisation improvements underway
- New larger datasets
  - Compression performance was limited by small numbers of variables in open dataset
  - Training performance limited by small number of events
  - Move to PHYSLITE datsets (research open data) - 36 fb^{-1} of (mostly) full-fat ATLAS data!
- New models
  - Improve physics inference using new loss functions and model architectures
- ROOT integration?
- FPGA applications?

Software Sustainability and Community

- Funded by local software sustainability grants
- How can we improve climate impact?
- Make software more efficient
- Improve documentation and findabilty - prevent re-making software!
- Share cross-discipline expertise

Project driven by Early-Career Researchers

- Main contributors for this project: undergraduate/Master's students and summer students/interns
  - Huge amount of high-quality work, but seasonal and limited prior experience
  - Strong tutorials and documentation essential for rapid onboarding
- Managed by PhDs/Post-Docs, limited academic involvement

Interested? Feedback? Contact us!

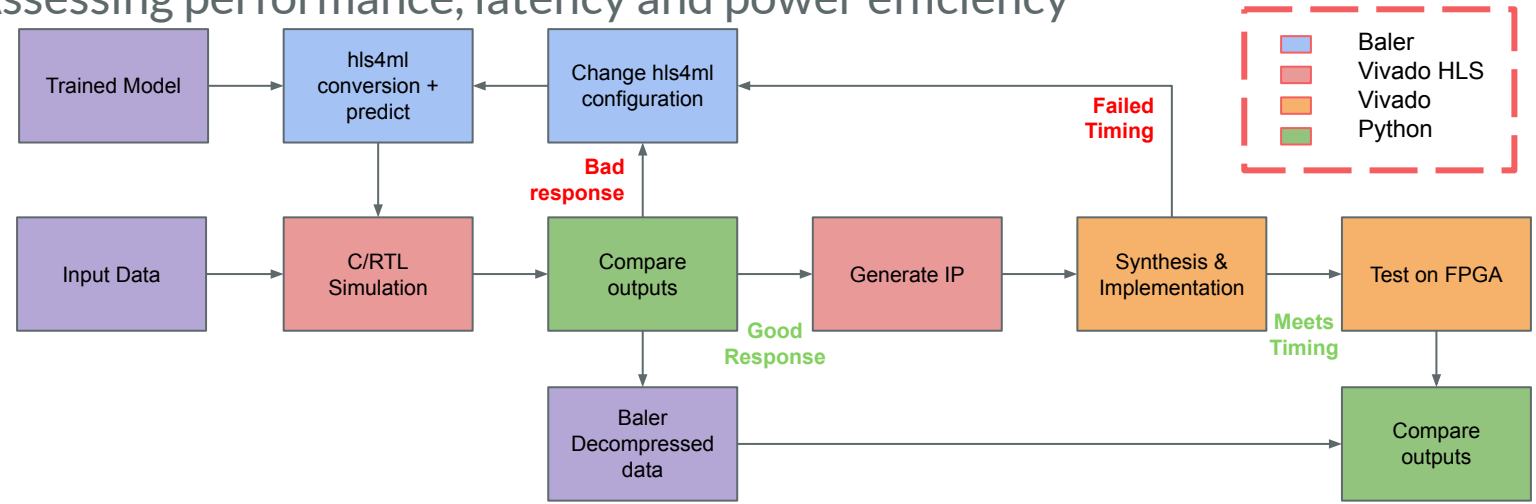
- We are a friendly, cross-discipline team with significant involvement from **ECRs** and **industry**
- Summer, Bachelor's/Master's and PhD projects very welcome and **can be supported**

- https://github.com/baler-collaboration/baler
- james.smith-7@manchester.ac.uk
- caterina.doglioni@manchester.ac.uk

Backup

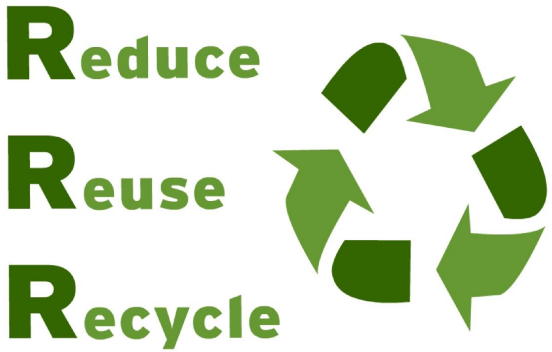
Baler on FPGA: Workflow

- Prototype version for developing and running BALER on an **FPGA**
  - Using vivado HLS code
- Useful in **bandwidth-restricted cases**
  - Network cards, detector readout, triggers, transmitters
- Assessing performance, latency and power efficiency



Software Sustainability (energy & more)

- Funded by local **software sustainability** grants
- How can we improve climate impact?
  - **Reduce** software resource usage
    - Efficient software
      - Trade-off between performance and consumption
    - Share cross-discipline expertise
  - **Reuse** software
    - Open-source
    - Well-written so it can be extended
    - Generic as possible
  - **Recycle** old software
    - Good documentation!
    - Good publicity
    - Preserve code and datasets (github, zenodo)



Community Development

- Project driven by Early-Career Researchers
  - Main contributors for this project: undergraduate/Master's students and summer students/interns
    - Huge amount of high-quality work, but seasonal and limited prior experience
    - Strong tutorials and documentation essential for rapid onboarding
  - Managed by PhDs/Post-Docs, limited academic involvement
  - Well defined, well planned short projects useful for students and academics alike
  - Important to reward junior members and share knowledge across academia (ESCAPE, EVERSE)
- Range of funding sources are important, large and small
  - Small 'pump-priming' grants useful for buying prototype equipment, hiring RSEs
  - Large national and international grants important for academic stability (Horizon, ESCAPE)
- Industry connections fruitful for datasets and best practices, but difficult to find
  - Difficult to convince we don't want money or a job!