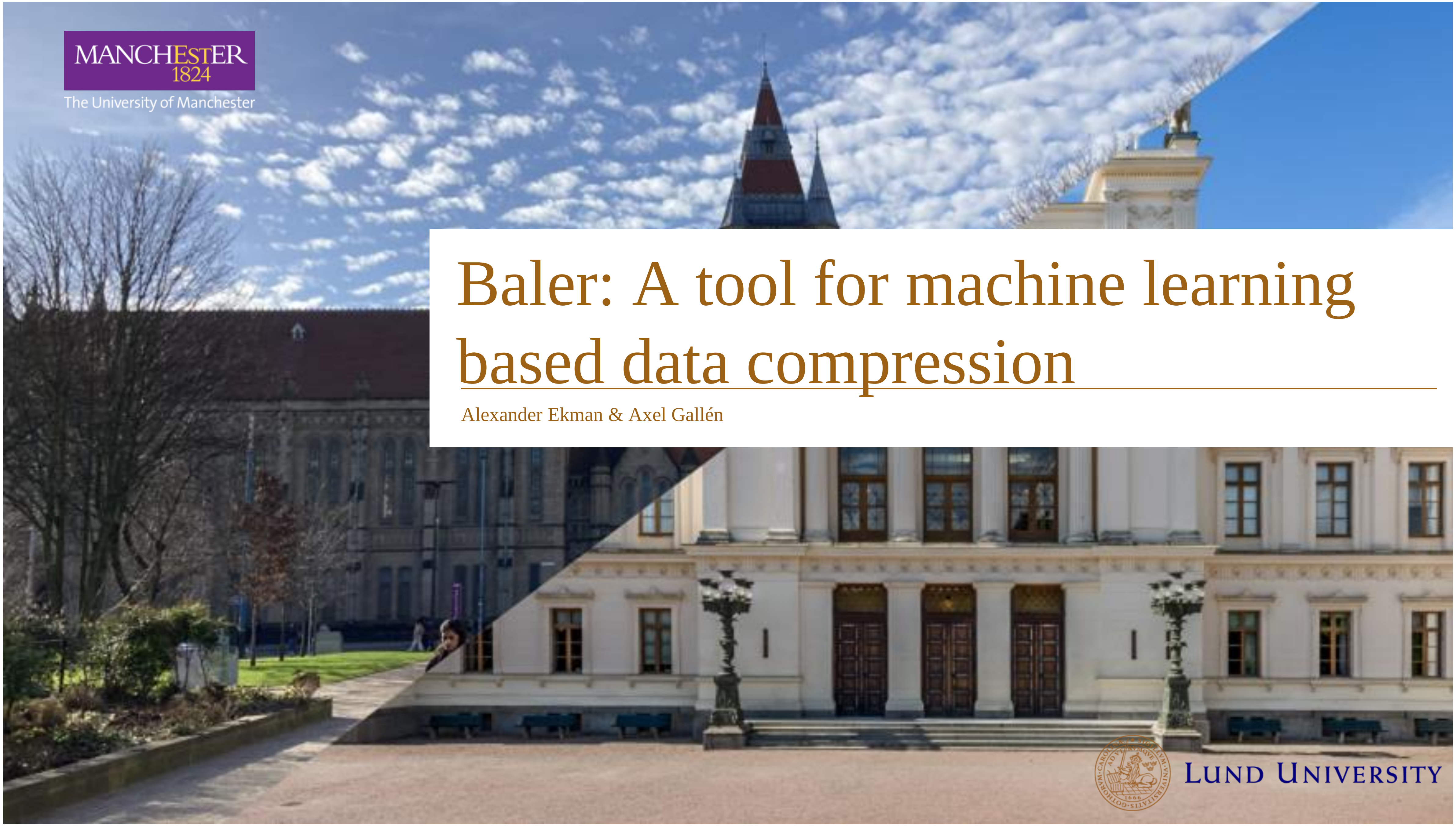
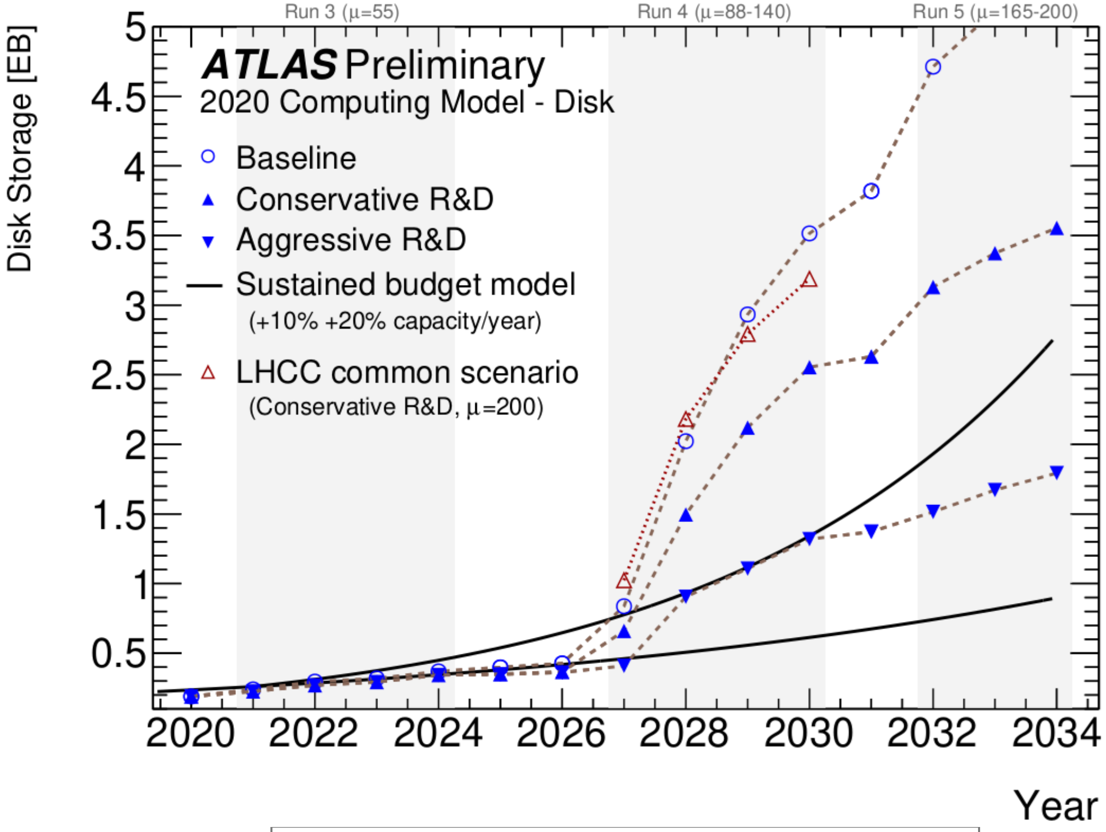


Baler: A tool for machine learning based data compression

Alexander Ekman & Axel Gallén

The problem

- Problem: Too much data, too little storage
- 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/

A Solution

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

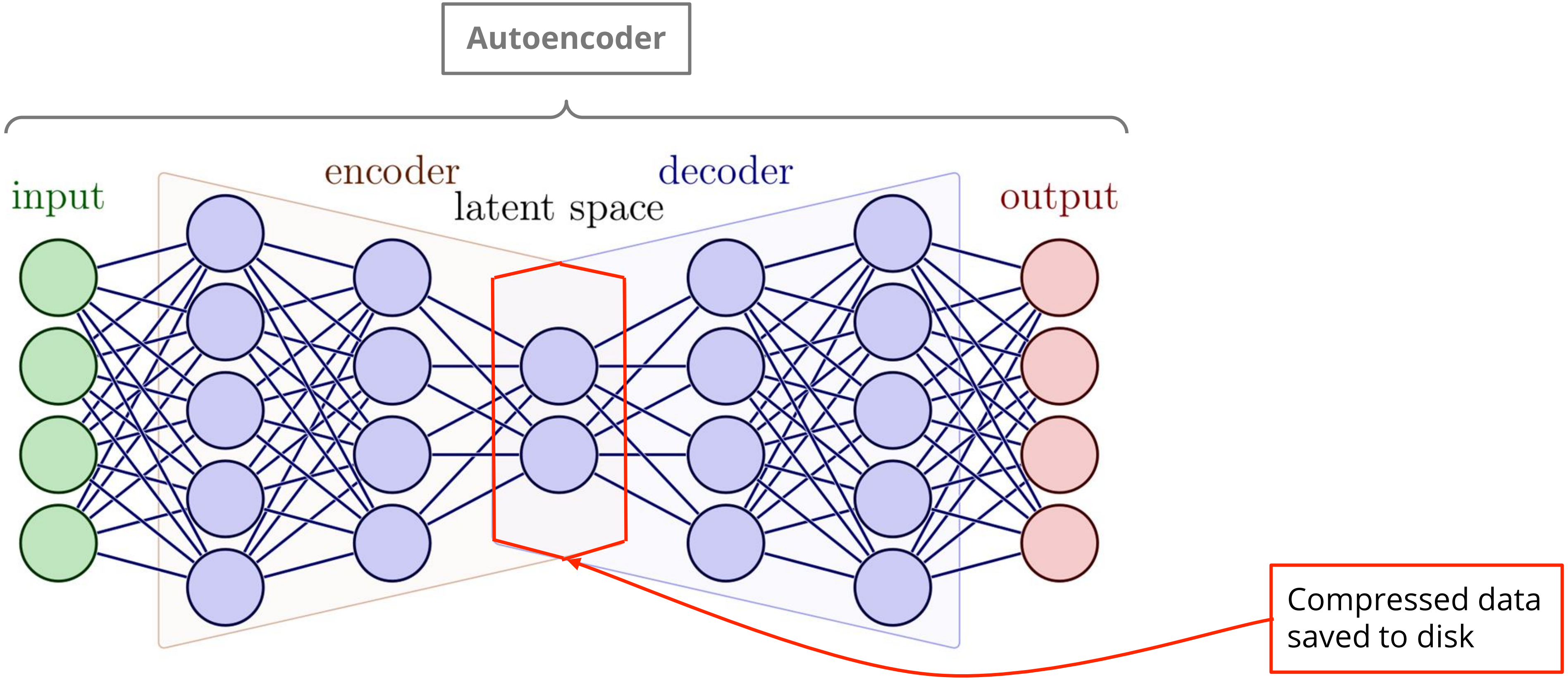


Figure modified from: https://tikz.net/neural_networks/

Lossy compression

- Works well in cases where more data is better
  - Particle physics: where more events compensate for the loss in precision

- Works well where the only option is to delete the data
  - Computational Fluid dynamics: No infrastructure to store generated data for long times after publication

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
  - https://github.com/baler-collaboration/baler
- Simple to run with python through Poetry

```
poetry run python baler --project=CMS --mode=train
```

- Docker implementation also available
  - Docker-Sponsored Open Source program

arXiv:submit/4872497 [hep-ex] 3 May 2023

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

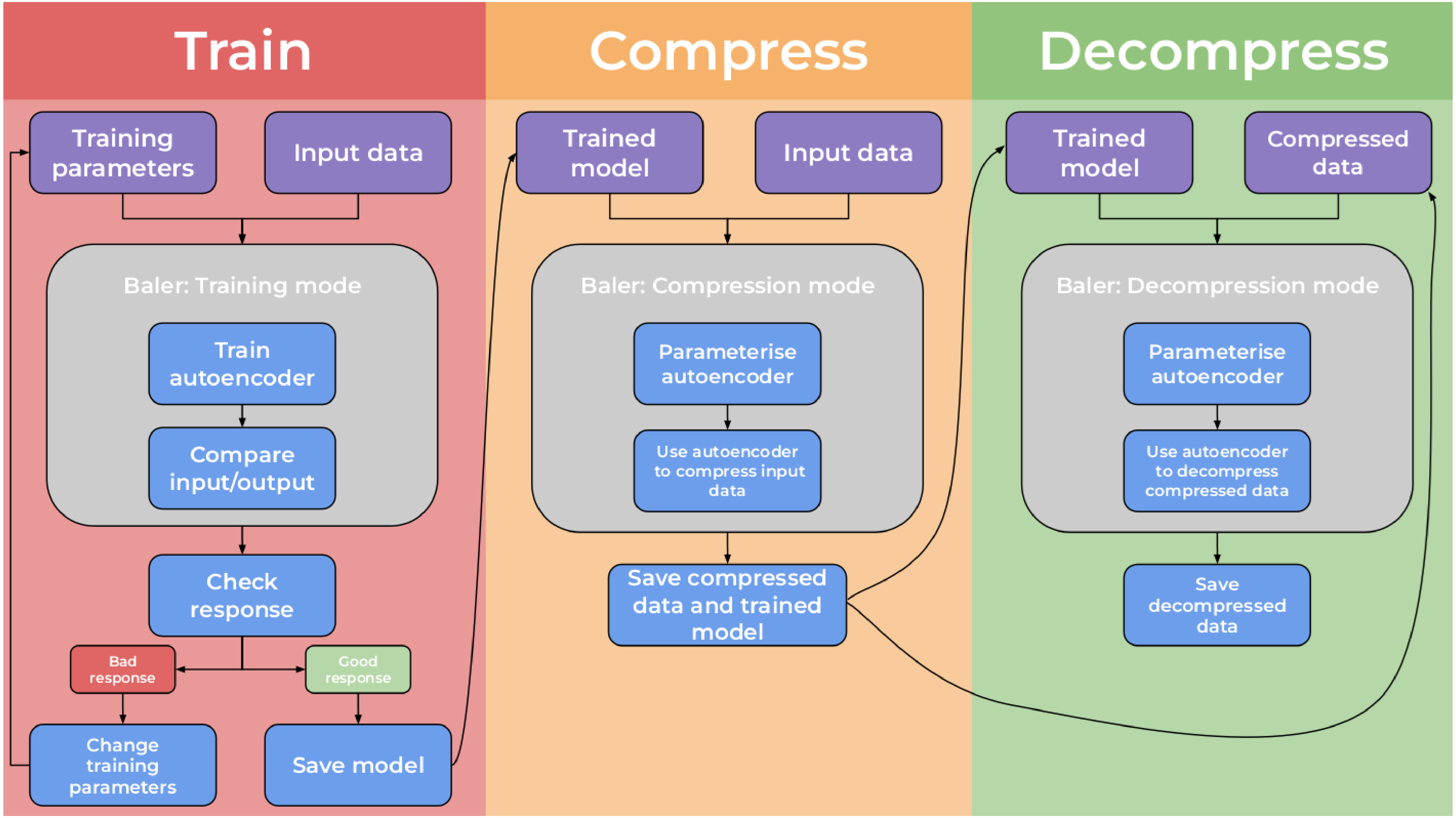
Many different fields of science share a common issue; storing ever-growing datasets. By the end of the next decade, the Large Hadron Collider (LHC) experiments will have over an order of magnitude more data to analyze than currently [1–3]; the Square Kilometre Array (SKA) experiment is expected to record 8.5EB of data over its 15-year lifespan [4] and fields such as Computational Fluid Dynamics (CFD) rely on TB-sized simulation samples that need to be stored and shared. Without significant R&D, the datasets expected to be collected by big-data science experiments are projected to exceed the available storage resources (see e.g. Fig. 2 of Ref. [1] for the case of the ATLAS experiment at the LHC). This cross-disciplinary issue is not limited to scientific research and extends to industrial operations [5].

1.1 Lossy data compression in high energy physics

A common mitigation strategy to this problem involves compressing data using lossless algorithms, see e.g. Refs. [6–8]. Once the storage limit is reached, one is forced to discard parts of the dataset, or only save certain features of the data. Generally, this can be done without impacting the overall scientific program of the experiments, for example by using a data selection system called *trigger* that only stores data satisfying certain pre-determined characteristics that ensure the dataset will be aligned with the experiment's main scientific goals. However, saving only a subset of data is not ideal for processes where additional

https://arxiv.org/abs/2305.02283

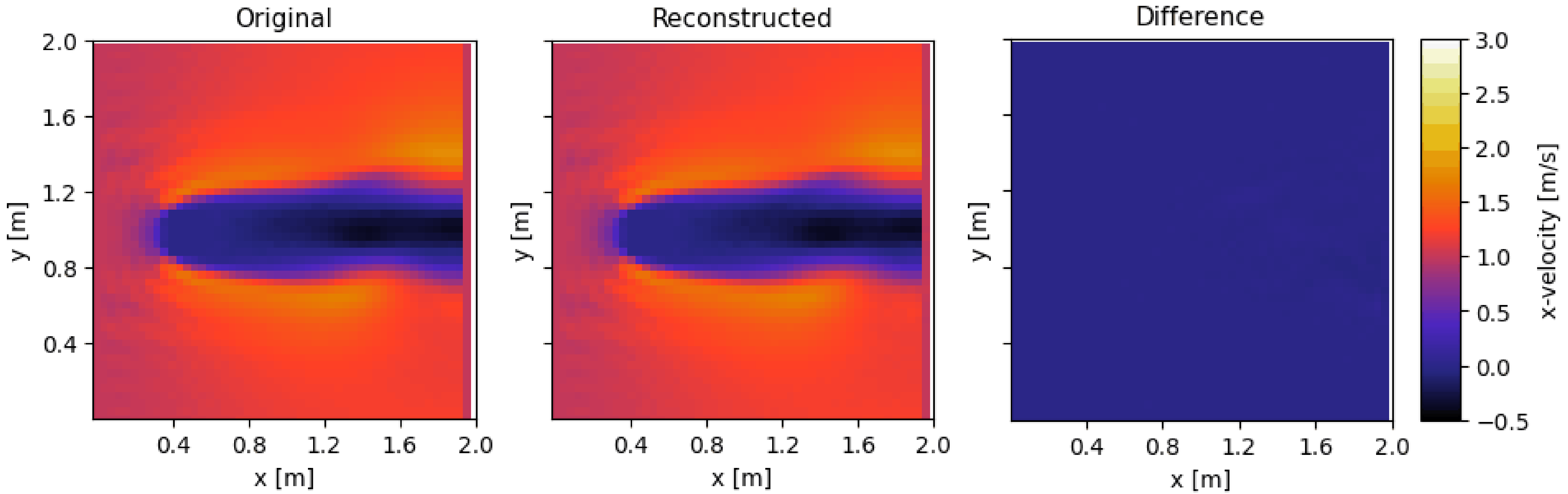
Baler Workflow



Computational Fluid Dynamics

- Data consists of 2D slice of the x-velocity component for a liquid flowing over a cube
- The compressed file is 0.5% the size of the input
- We present:
  - Data **before** and **after** compression+decompression
  - **Difference** between before and after

Original.npy	1,2 MB
Compressed.npy	6,1 kB



Methodology

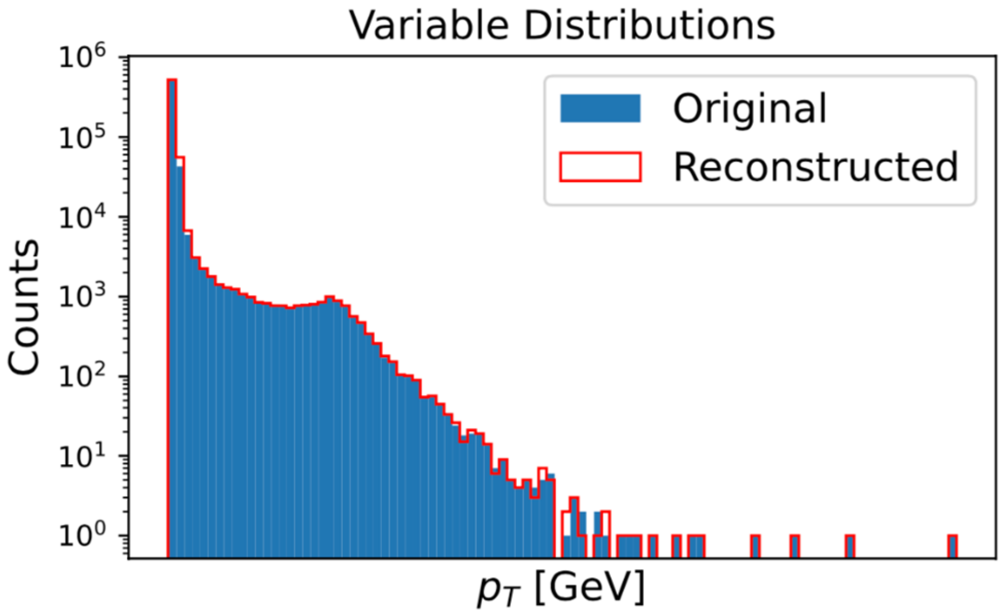
- HEP Data
  - Open CMS Data (DOI:10.7483/OPENDATA.CMS.KL8H.HFVH)
  - ~ 600 000 jets
  - 24 variables per jet compressed to 14 variables -> 58% original size

- Evaluation Metrics:

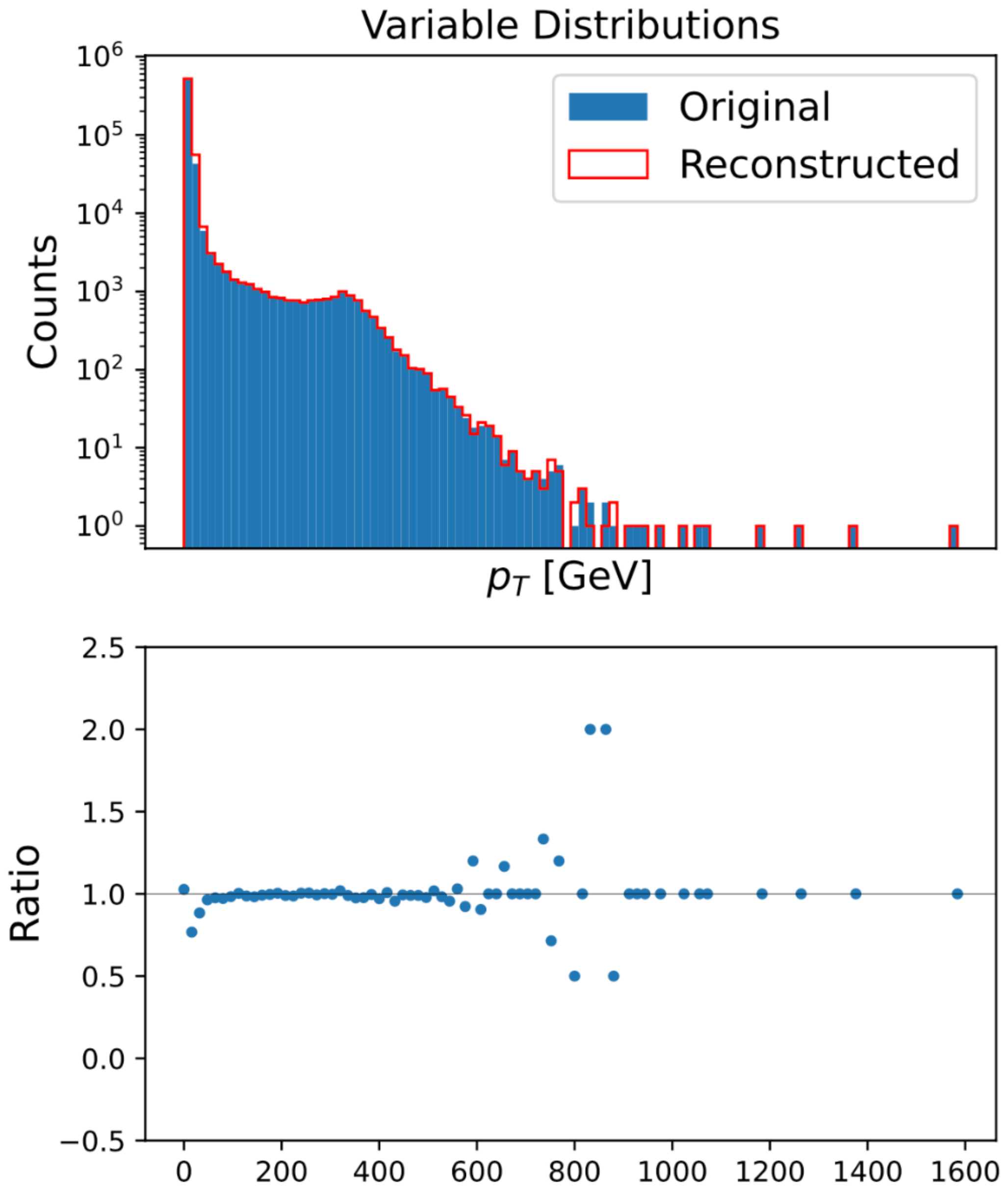
$$\text{Relative Difference} = \frac{\text{reconstructed} - \text{original}}{\text{original}}$$

$$\text{Difference} = \text{reconstructed} - \text{original}$$

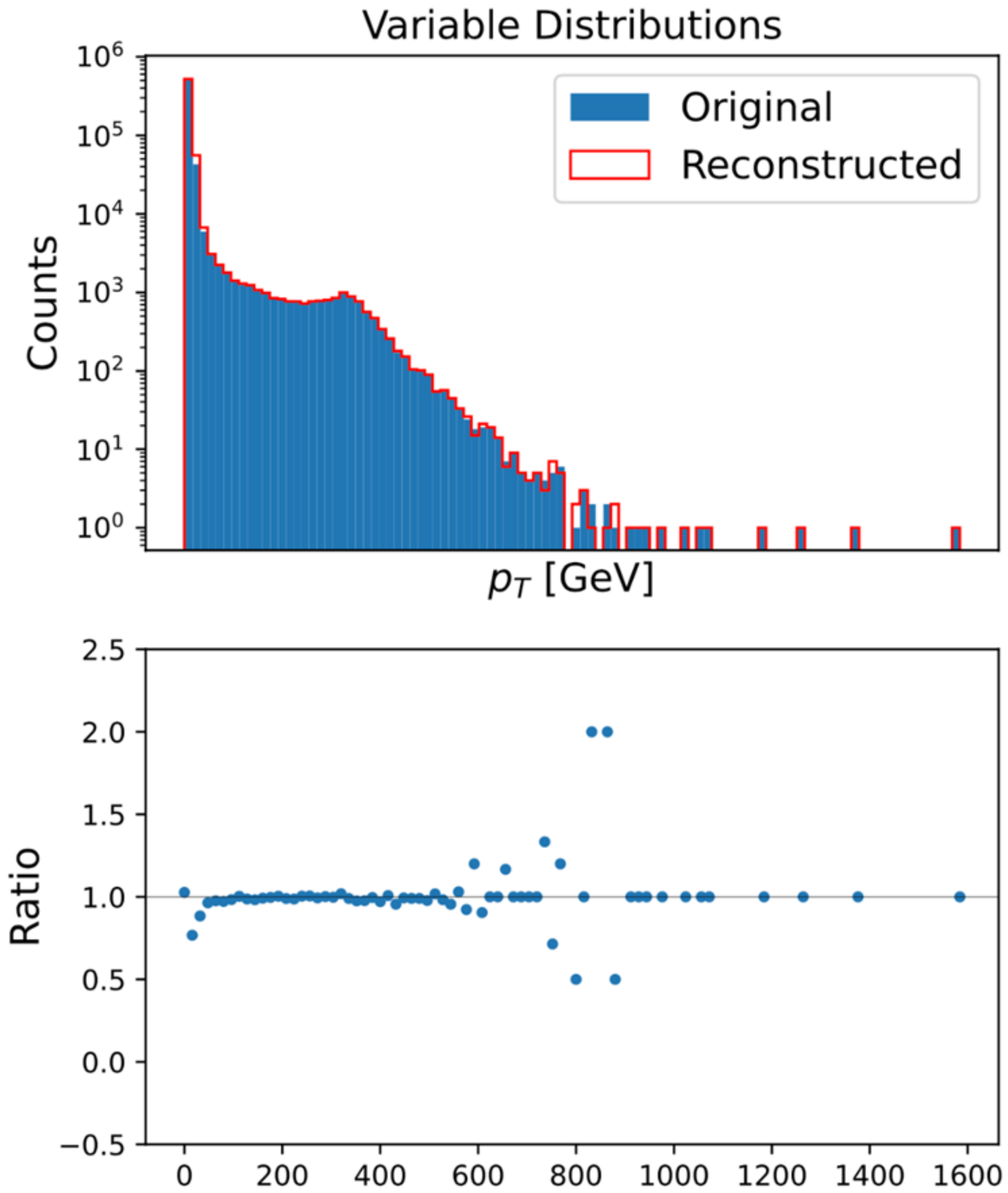
Results in HEP: Transverse Momentum

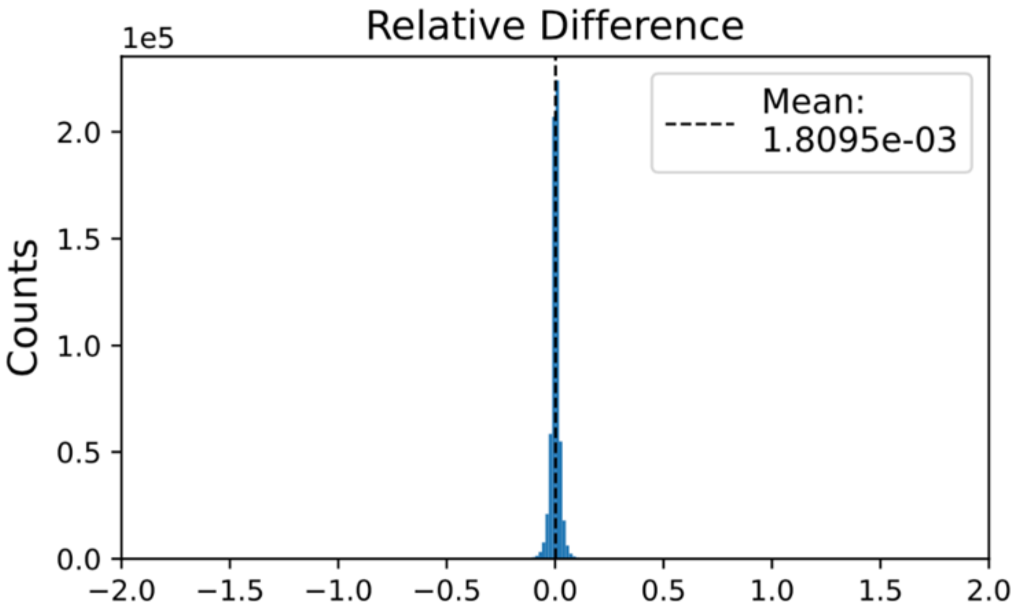


Results in HEP: Transverse Momentum

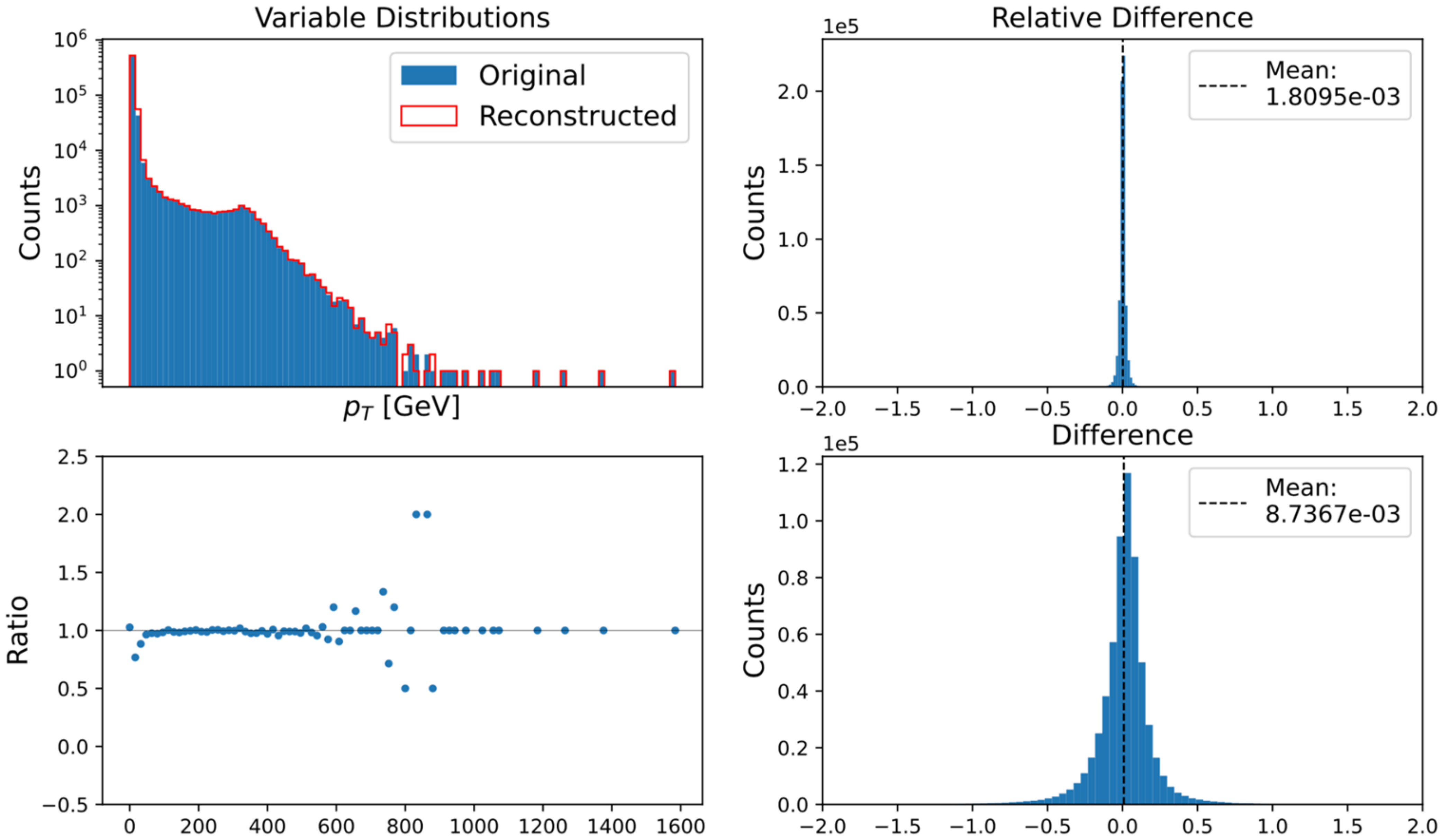


Results in HEP: Transverse Momentum

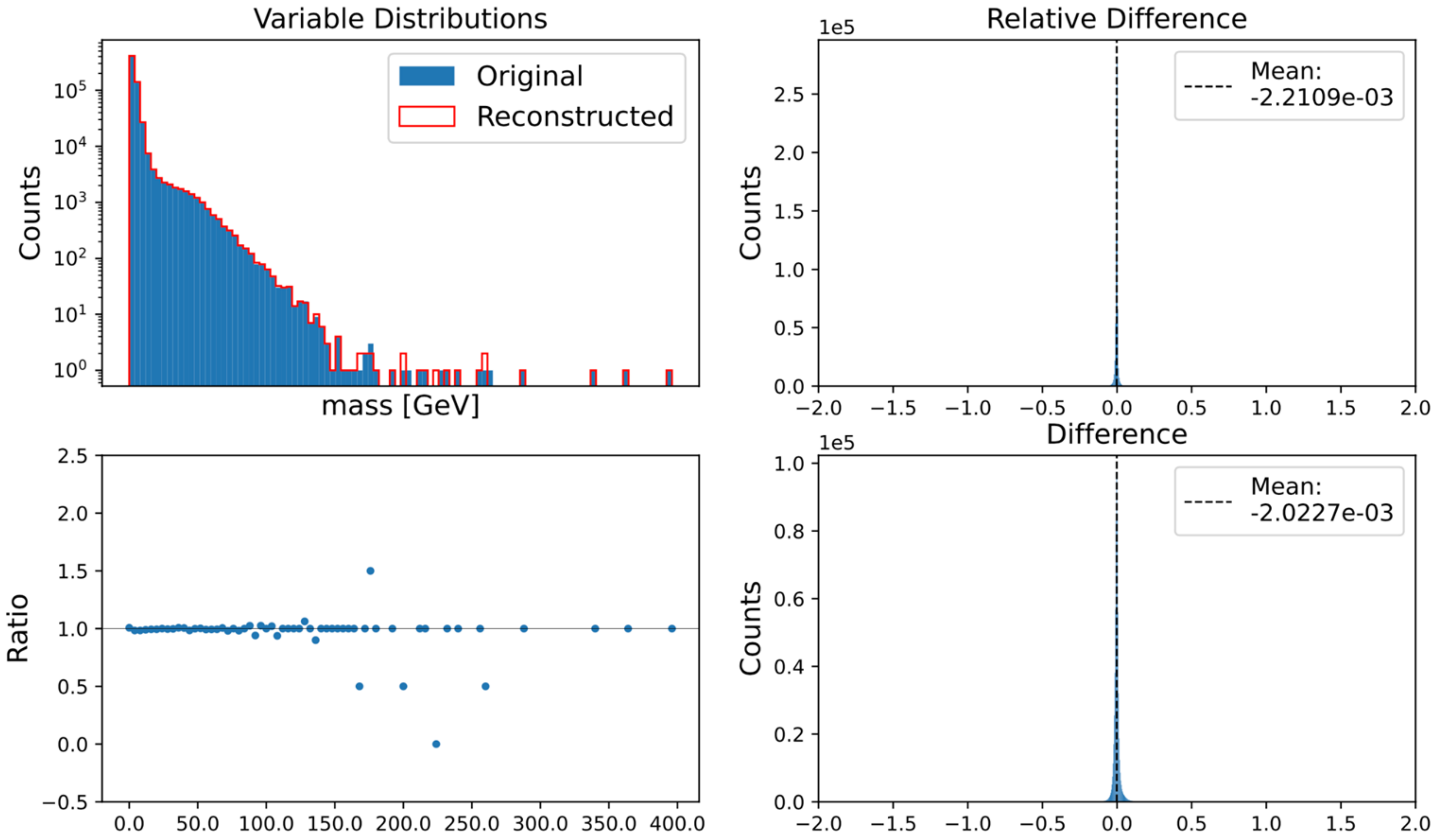




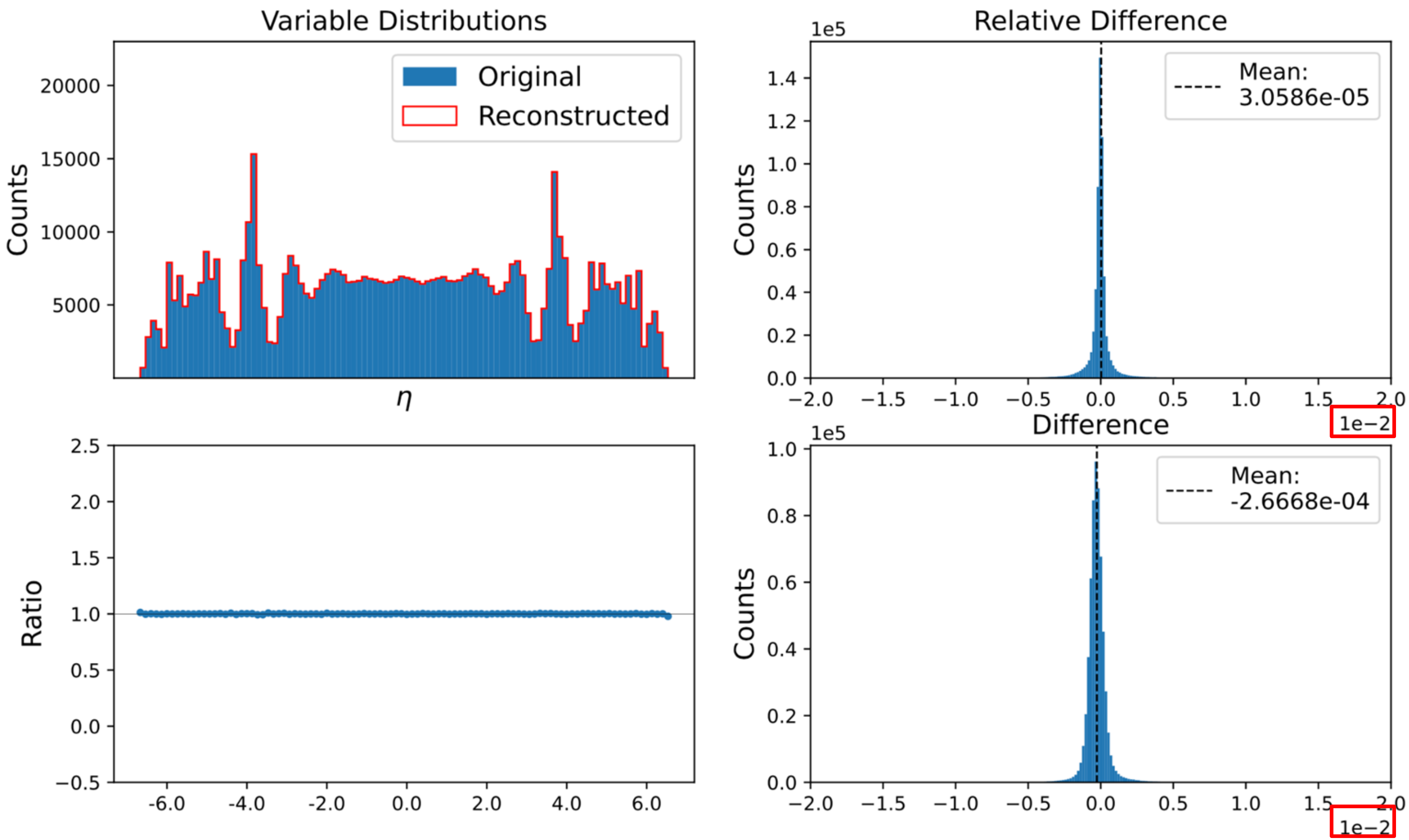
Results in HEP: Transverse Momentum



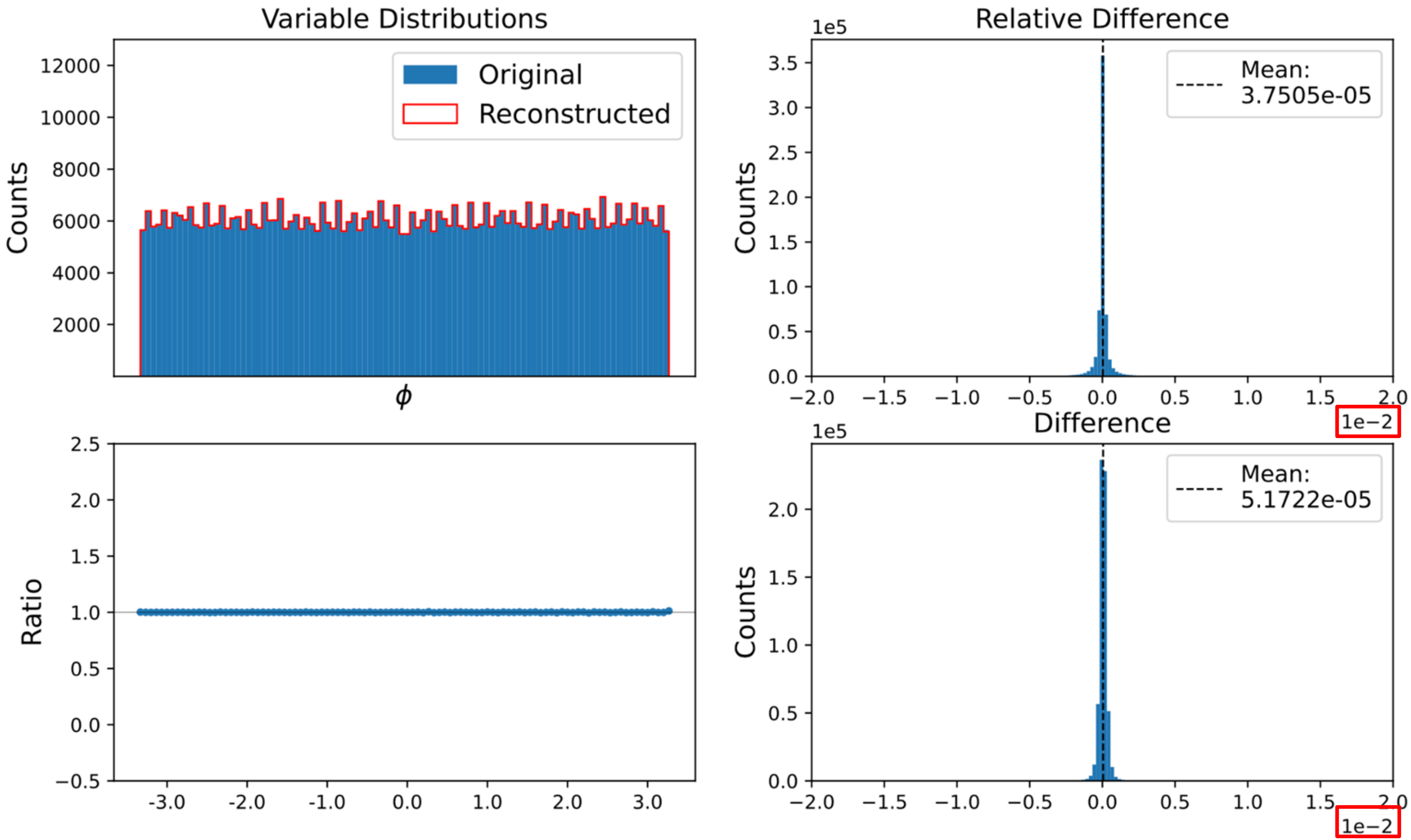
Results in HEP: Mass



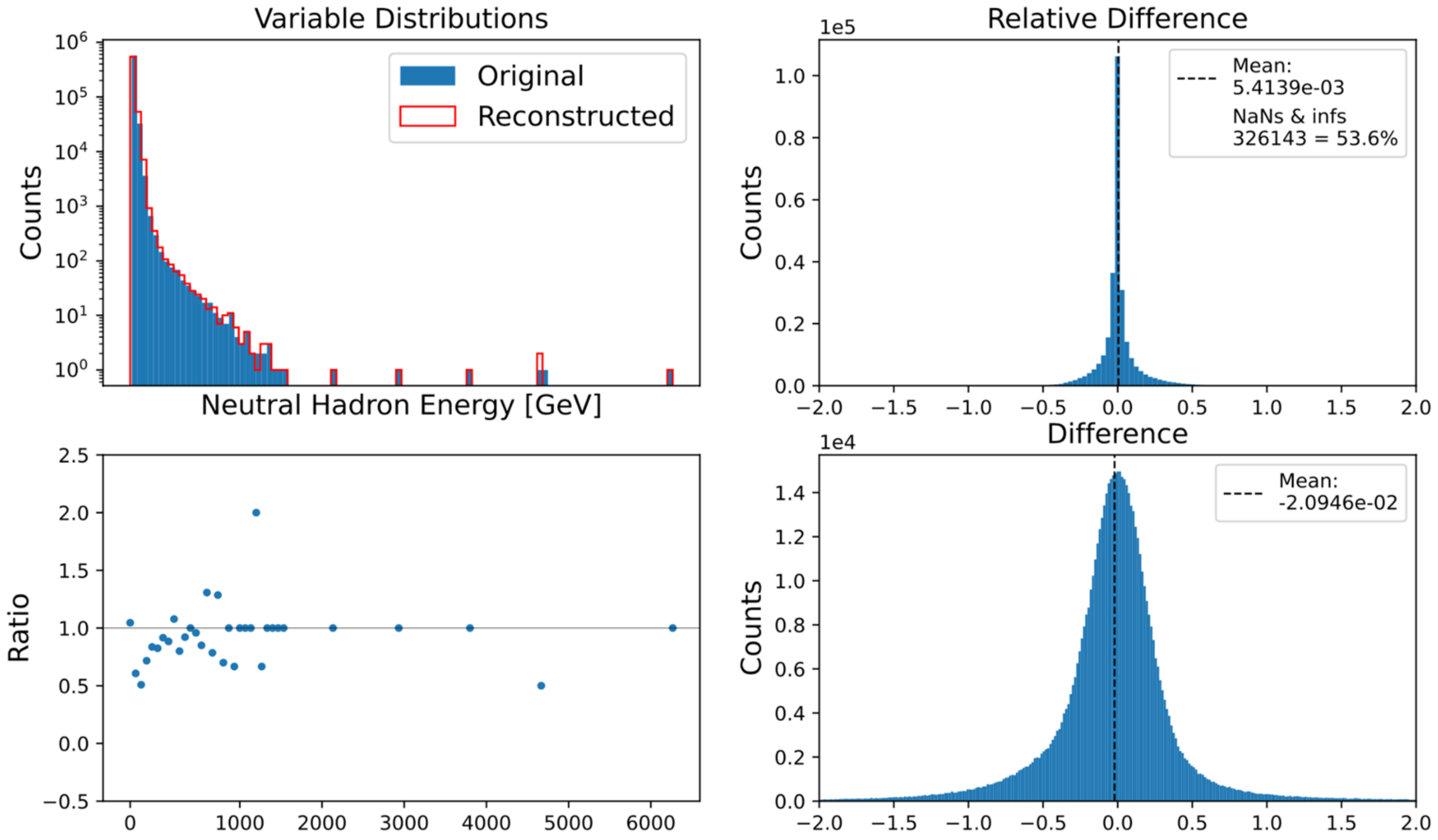
Results in HEP: Pseudorapidity, η



Results in HEP: Polar Angle, ϕ



Results in HEP: Neutral Hadron Energy



HEP gzip dilemma

- HEP

– Baler ->	OK reconstruction	58% original file size
– gzip ->	Perfect reconstruction	25% original file size

- Reason for the big difference:
  - A lot of repeating values in HEP data is beneficial for methods like gzip

- Future work:
  - Run on other datasets
  - Evaluate impact on full physics analysis

CFD Auxiliary file dilemma

- CFD
  - Baler -> Good reconstruction 0.5% original file size
  - gzip -> Lossless reconstruction 50% original file size

- Reason for the big difference:
  - Few repeating values in CFD data

- One problem… Auxiliary files
  - Input CFD data size: ~1.2 MB
  - Decoder: ~600 MB

- Future work:
  - Run on large 3D time series datasets

Summary

- Open-source tool for machine learning based compression
- HEP results:
  - Compression to 58% of input size
  - On average jet pT and mass differ on order of 0.2%, eta and phi 0.003%
  - Other 20 variables have varying performance
- CFD results:
  - Huge compression to 0.5% of input size, small error, but large auxiliary files
- Future improvements:
  - Try more “suitable” input files
  - Apply methods employed by the other 4 “major” AE-compression research groups

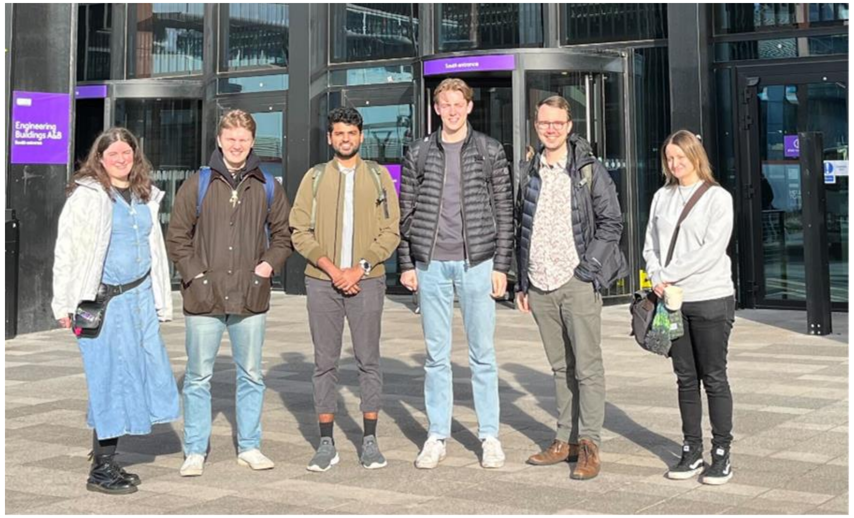
Further reading

- Read our Arxiv paper: https://doi.org/10.48550/arXiv.2305.02283

- Leading AE compressor: https://doi.org/10.48550/arXiv.2105.11730
- Others:
  - https://doi.org/10.1109/TBDATA.2021.3066151
  - https://doi.org/10.1103/PhysRevFluids.5.114602
  - https://doi.org/10.48550/arXiv.2210.09262

The Baler Team

- Big thank you from the Baler team!
- For more details see: https://arxiv.org/abs/2305.02283
- Try our working examples at our GitHub repository
  - https://github.com/baler-collaboration/baler

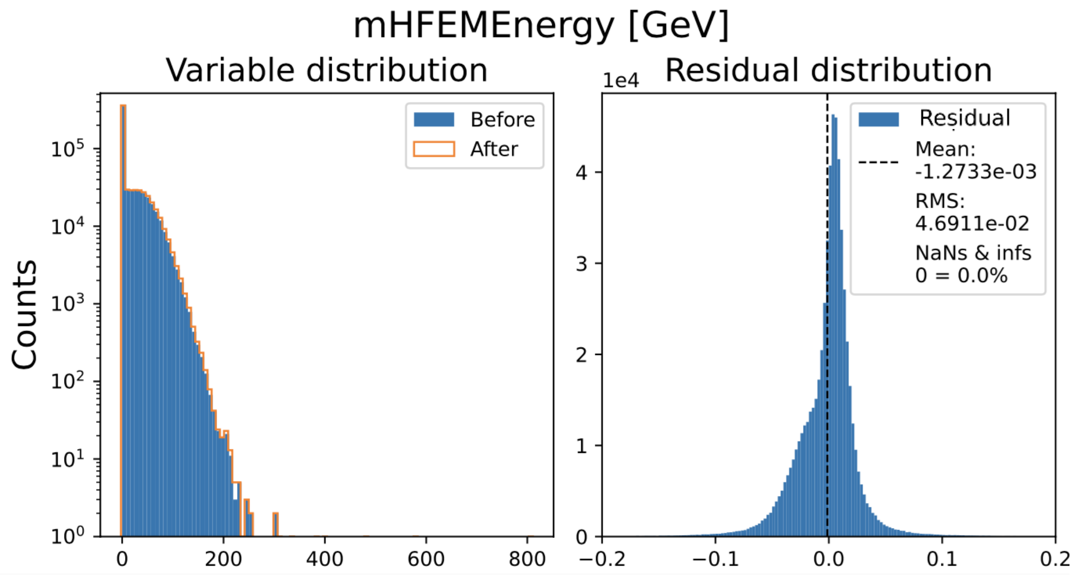


Marta Camps Santasmasas	(UoM, CFD)
Nicole Skidmore	(UoM, HEP)
Caterina Doglioni	(UoM, HEP)
Pratik Jawahar	(UoM, HEP)
Oliver Woolland	(UoM, RSE)
Alma Orucevic-Alagic	(Lund, CS)
Fritjof Bengtsson	(Lund, CS)
Axel Gallén	(Lund, HEP)
Alexander Ekman	(Lund, HEP)

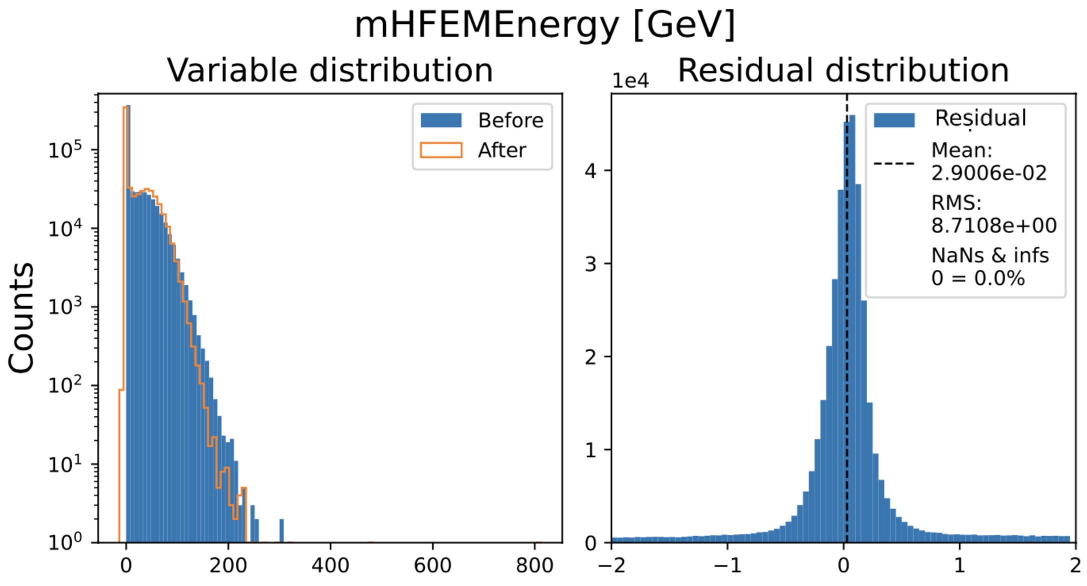
Backup slides

1.7x vs 6x compression

1.7x compression



6x compression



Full variable list (see https://arxiv.org/abs/2305.02283)

Table 2: Residual and Response distribution means and RMS values for all variables in the dataset. These values are presented at $R = 1.7$, and all values have been averaged over 5 runs, with an added statistical error of two standard deviations.

Variable ($R = 1.7$)	Response		Residual	
	Mean	RMS	Mean	RMS
p_T	$-1.07 \times 10^{-3} \pm 1.34 \times 10^{-2}$	$2.09 \times 10^{-2} \pm 3.56 \times 10^{-3}$	$-1.44 \times 10^{-2} \pm 1.04 \times 10^{-1}$	$2.12 \times 10^{-1} \pm 5.29 \times 10^{-2}$
η	$3.75 \times 10^{-4} \pm 6.11 \times 10^{-4}$	$8.12 \times 10^{-1} \pm 1.17$	$-1.12 \times 10^{-3} \pm 2.67 \times 10^{-3}$	$2.09 \times 10^{-3} \pm 1.45 \times 10^{-3}$
ϕ	$3.44 \times 10^{-4} \pm 8.64 \times 10^{-4}$	$1.93 \times 10^{-1} \pm 4.32 \times 10^{-1}$	$2.45 \times 10^{-4} \pm 1.80 \times 10^{-3}$	$9.91 \times 10^{-4} \pm 1.12 \times 10^{-3}$
mass	$2.39 \times 10^{-1} \pm 7.87$	$4.38 \times 10^{3} \pm 4.47 \times 10^{3}$	$-8.05 \times 10^{-3} \pm 2.51 \times 10^{-2}$	$3.98 \times 10^{-2} \pm 1.42 \times 10^{-2}$
mJetArea	$6.12 \times 10^{-5} \pm 1.81 \times 10^{-4}$	$3.13 \times 10^{-4} \pm 1.48 \times 10^{-4}$	$3.21 \times 10^{-5} \pm 8.90 \times 10^{-5}$	$1.10 \times 10^{-4} \pm 5.77 \times 10^{-5}$
mChargedHadronEnergy	$1.58 \times 10^{-3} \pm 1.70 \times 10^{-2}$	$2.85 \times 10^{-2} \pm 1.30 \times 10^{-2}$	$1.68 \times 10^{-2} \pm 1.43 \times 10^{-1}$	$1.71 \times 10^{-1} \pm 7.33 \times 10^{-2}$
mNeutralHadronEnergy	$7.05 \times 10^{-2} \pm 9.88 \times 10^{-2}$	$2.22 \times 10^{-1} \pm 6.59 \times 10^{-2}$	$2.77 \times 10^{-1} \pm 5.23 \times 10^{-1}$	$6.94 \times 10^{-1} \pm 2.26 \times 10^{-1}$
mPhotonEnergy	$-2.75 \times 10^{-2} \pm 7.48 \times 10^{-2}$	$6.84 \times 10^{-2} \pm 1.09 \times 10^{-1}$	$-8.00 \times 10^{-2} \pm 1.87 \times 10^{-1}$	$1.52 \times 10^{-1} \pm 1.77 \times 10^{-1}$
mElectronEnergy	$-7.71 \times 10^{-2} \pm 1.05 \times 10^{-1}$	$1.44 \times 10^{-1} \pm 7.47 \times 10^{-2}$	$1.71 \times 10^{-2} \pm 5.32 \times 10^{-2}$	$8.40 \times 10^{-2} \pm 4.15 \times 10^{-2}$
mMuonEnergy	$1.29 \times 10^{-2} \pm 1.97 \times 10^{-2}$	$8.04 \times 10^{-2} \pm 9.77 \times 10^{-2}$	$1.18 \times 10^{-2} \pm 1.46 \times 10^{-2}$	$3.15 \times 10^{-2} \pm 7.05 \times 10^{-3}$
mHFHadronEnergy	$-1.10 \times 10^{-2} \pm 4.66 \times 10^{-2}$	$1.77 \times 10^{-1} \pm 2.48 \times 10^{-2}$	$-3.15 \times 10^{-1} \pm 1.07$	$1.85 \pm 7.31 \times 10^{-1}$
mHFEMEnergy	$1.78 \times 10^{-3} \pm 7.40 \times 10^{-3}$	$1.41 \times 10^{-2} \pm 3.63 \times 10^{-3}$	$1.22 \times 10^{-2} \pm 8.26 \times 10^{-2}$	$6.93 \times 10^{-2} \pm 5.54 \times 10^{-2}$
mChargedHadronMultiplicity	$-1.00 \times 10^{-3} \pm 5.04 \times 10^{-3}$	$4.48 \times 10^{-3} \pm 4.90 \times 10^{-3}$	$-3.13 \times 10^{-3} \pm 1.82 \times 10^{-2}$	$9.68 \times 10^{-3} \pm 1.50 \times 10^{-2}$
mNeutralHadronMultiplicity	$-1.22 \times 10^{-4} \pm 1.29 \times 10^{-3}$	$8.76 \times 10^{-4} \pm 9.42 \times 10^{-4}$	$-1.19 \times 10^{-4} \pm 1.51 \times 10^{-3}$	$9.89 \times 10^{-4} \pm 1.20 \times 10^{-3}$
mPhotonMultiplicity	$-1.14 \times 10^{-3} \pm 3.62 \times 10^{-3}$	$2.72 \times 10^{-3} \pm 4.14 \times 10^{-3}$	$-2.69 \times 10^{-3} \pm 7.44 \times 10^{-3}$	$4.92 \times 10^{-3} \pm 7.12 \times 10^{-3}$
mElectronMultiplicity	$1.07 \times 10^{-3} \pm 3.87 \times 10^{-3}$	$2.37 \times 10^{-3} \pm 2.37 \times 10^{-3}$	$-1.54 \times 10^{-5} \pm 9.96 \times 10^{-5}$	$2.11 \times 10^{-4} \pm 1.75 \times 10^{-4}$
mMuonMultiplicity	$1.12 \times 10^{-3} \pm 1.22 \times 10^{-3}$	$2.51 \times 10^{-3} \pm 6.69 \times 10^{-4}$	$5.67 \times 10^{-5} \pm 1.16 \times 10^{-4}$	$2.41 \times 10^{-4} \pm 6.35 \times 10^{-5}$
mHFHadronMultiplicity	$-1.34 \times 10^{-3} \pm 1.84 \times 10^{-3}$	$2.53 \times 10^{-3} \pm 1.94 \times 10^{-3}$	$-2.67 \times 10^{-3} \pm 3.33 \times 10^{-3}$	$4.44 \times 10^{-3} \pm 4.05 \times 10^{-3}$
mHFEMMultiplicity	$2.41 \times 10^{-4} \pm 2.51 \times 10^{-3}$	$1.98 \times 10^{-3} \pm 1.33 \times 10^{-3}$	$5.98 \times 10^{-4} \pm 4.16 \times 10^{-3}$	$3.08 \times 10^{-3} \pm 2.95 \times 10^{-3}$
mChargedEmEnergy	$-7.72 \times 10^{-2} \pm 1.05 \times 10^{-1}$	$1.44 \times 10^{-1} \pm 7.48 \times 10^{-2}$	$1.72 \times 10^{-2} \pm 5.30 \times 10^{-2}$	$8.40 \times 10^{-2} \pm 4.15 \times 10^{-2}$
mChargedMuEnergy	$1.29 \times 10^{-2} \pm 1.97 \times 10^{-2}$	$8.05 \times 10^{-2} \pm 9.78 \times 10^{-2}$	$1.18 \times 10^{-2} \pm 1.46 \times 10^{-2}$	$3.15 \times 10^{-2} \pm 7.07 \times 10^{-3}$
mNeutralEmEnergy	$-1.73 \times 10^{-2} \pm 5.42 \times 10^{-2}$	$5.89 \times 10^{-2} \pm 8.87 \times 10^{-2}$	$-6.70 \times 10^{-2} \pm 2.57 \times 10^{-1}$	$1.75 \times 10^{-1} \pm 1.81 \times 10^{-1}$
mChargedMultiplicity	$-9.83 \times 10^{-4} \pm 5.04 \times 10^{-3}$	$4.46 \times 10^{-3} \pm 4.88 \times 10^{-3}$	$-3.07 \times 10^{-3} \pm 1.83 \times 10^{-2}$	$9.74 \times 10^{-3} \pm 1.51 \times 10^{-2}$
mNeutralMultiplicity	$-8.97 \times 10^{-4} \pm 1.42 \times 10^{-3}$	$1.56 \times 10^{-3} \pm 1.93 \times 10^{-3}$	$-5.36 \times 10^{-3} \pm 7.37 \times 10^{-3}$	$7.34 \times 10^{-3} \pm 6.60 \times 10^{-3}$