

# Baler: Machine Learning-Based Data Compression



The University of Manchester

**Axel Gallén**  
**Uppsala University**  
**ICHEP 2024**

Currently active team:

James Smith (postdoc), Pratik Jawahar (PhD student),  
Aleko Lilius, Khwaish Anjum, Malena Duroux, Kaarel Kvisalu,  
Chakravarty Varadarajan, Samuel Hill (UG/Master's students),  
Jacob Forsell, Fritjof Bengtsson, Yuyang Jin, Leonid Didukh (industry)



UPPSALA  
UNIVERSITET



European Research Council  
Established by the European Commission



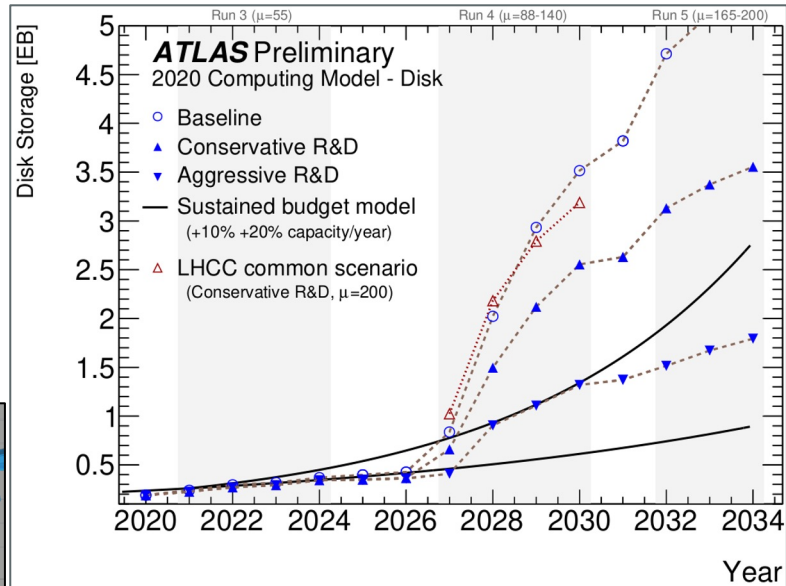
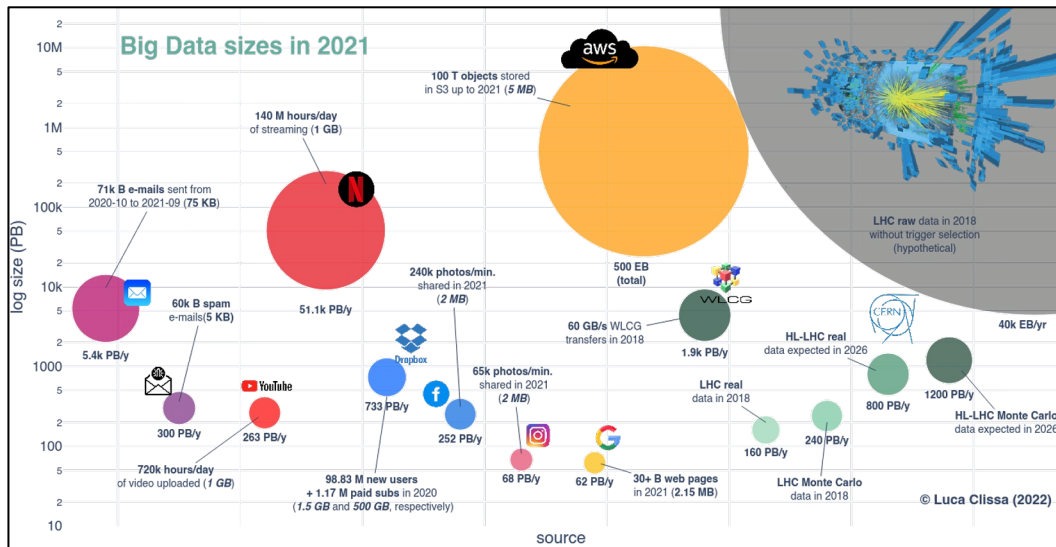
**WARWICK**  
THE UNIVERSITY OF WARWICK



**LUND**  
UNIVERSITY

# The 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/>

<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**

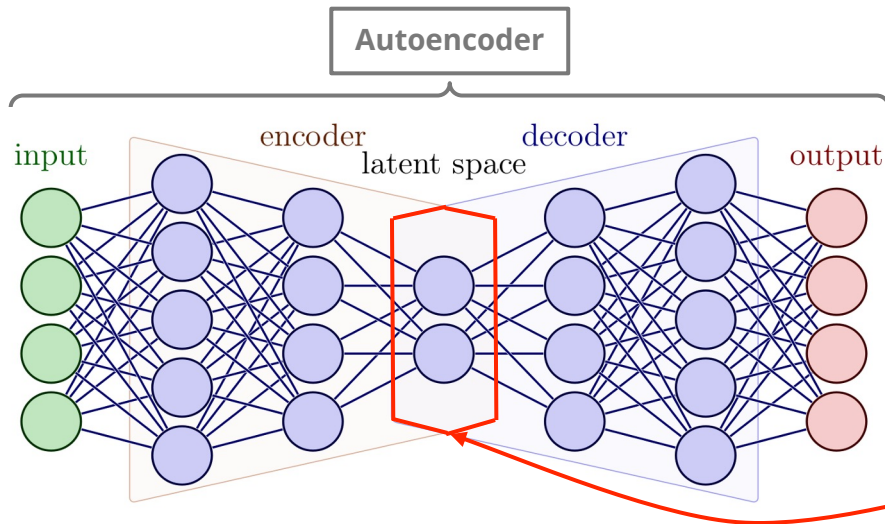


Figure modified from:  
[https://tikz.net/neural\\_networks/](https://tikz.net/neural_networks/)

Compressed  
data saved to  
disk

# Our Tool: Baler



- We have created a tool called **Baler** to help investigate the viability of this compression method
- **Multidisciplinary** tool
- Distributed and developed as an **open source** project [[GitHub: baler-collaboration/baler](https://github.com/baler-collaboration/baler)]
- Simple to install as a **pip** package or as command line tool
  - `pip install baler-compressor`
  - `Poetry run python baler --project=CMS --mode=train`
  - **Docker** version also available

[hep-ex] 3 May 2023

## Baler - Machine Learning Based Compression of Scientific Data

F. Bengtsson<sup>1</sup> C. Doglioni<sup>2</sup> P.A. Ekman<sup>1</sup> A. Gallén<sup>1</sup> P. Jawahar<sup>2</sup> A. Orucevic-Atagic<sup>1</sup> M. Camps Santasmasas<sup>2</sup> N. Skidmore<sup>2</sup> O. Woodland<sup>2</sup>

<sup>1</sup>Leeds University  
<sup>2</sup>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

**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
Article Number	09023
Number of page(s)	8
Section	Artificial Intelligence and Machine Learning
DOI	<a href="https://doi.org/10.1051/epjconf/202429509023">https://doi.org/10.1051/epjconf/202429509023</a>
Published online	06 May 2024

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

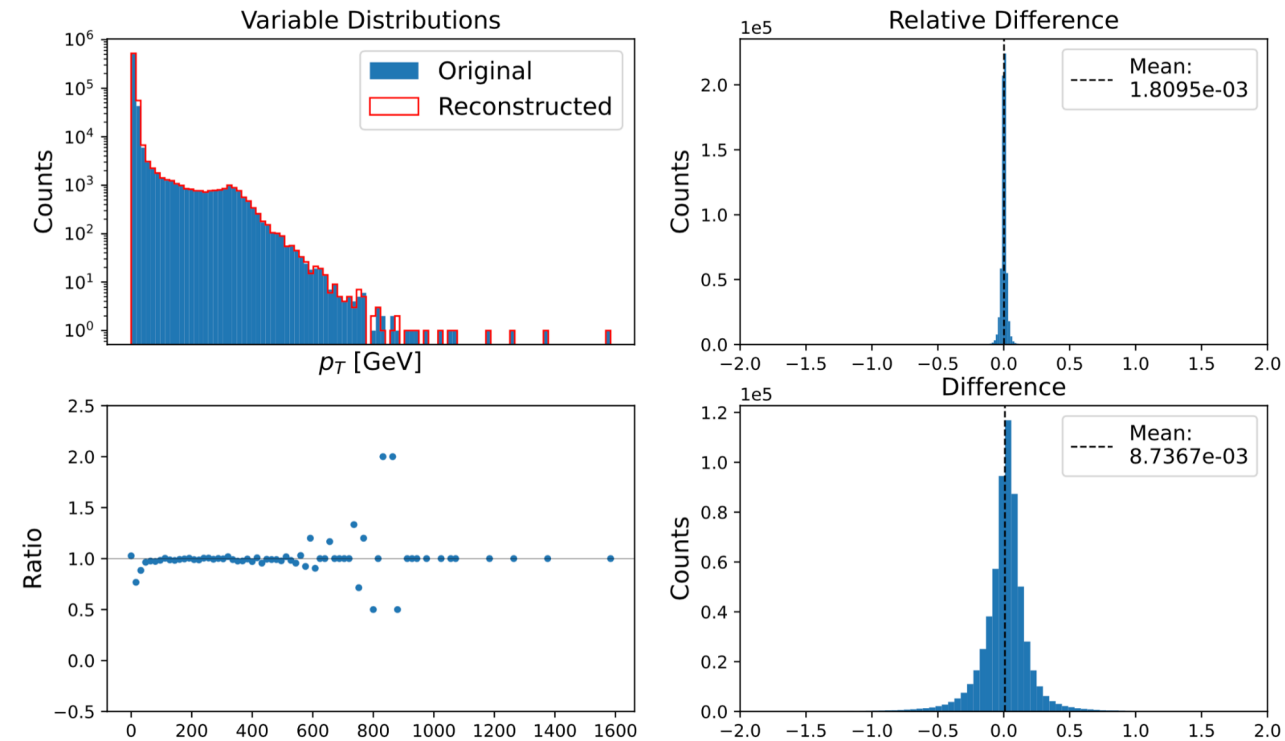
## Baler - Machine Learning Based Compression of Scientific Data

Fritjof Bengtsson Folkesson<sup>1\*</sup>, Caterina Doglioni<sup>2\*\*</sup>, Per Alexander Ekman<sup>1\*\*\*</sup>, Axel Gallén<sup>1\*\*\*\*</sup>, Pratik Jawahar<sup>2†</sup>, Marta Camps Santasmasas<sup>2‡</sup> and Nicola Skidmore<sup>2‡</sup>

# 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](https://doi.org/10.7483/OPENDATA.CMS.KL8H.HFVH)

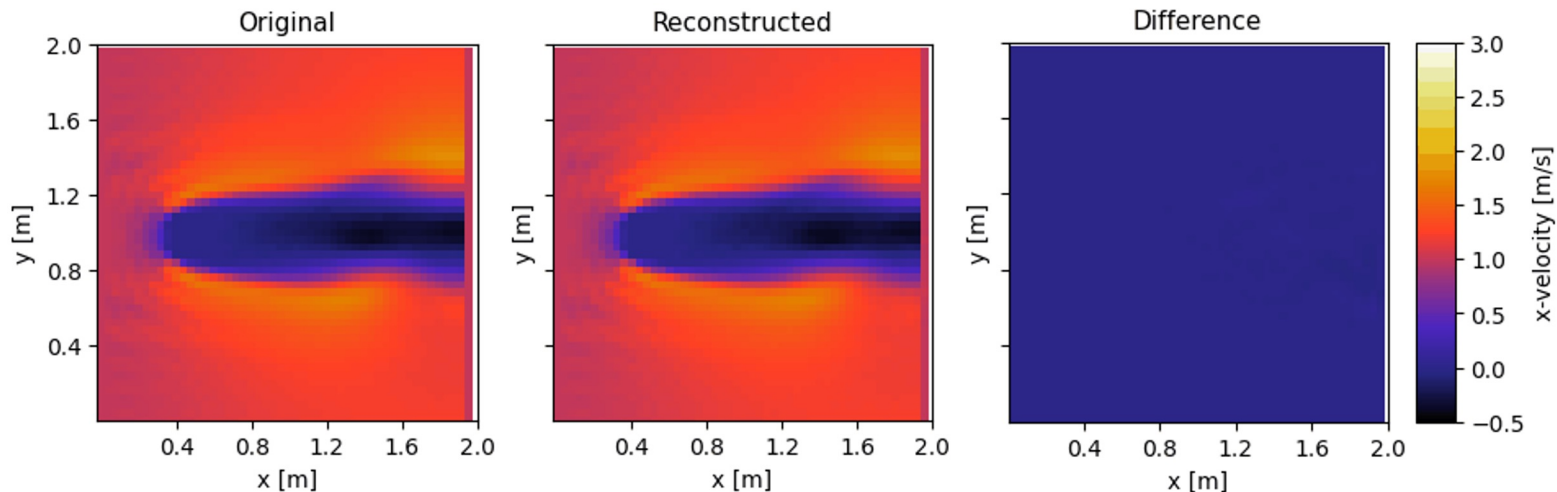


Full results: <https://arxiv.org/abs/2305.02283>

# Results: CFD

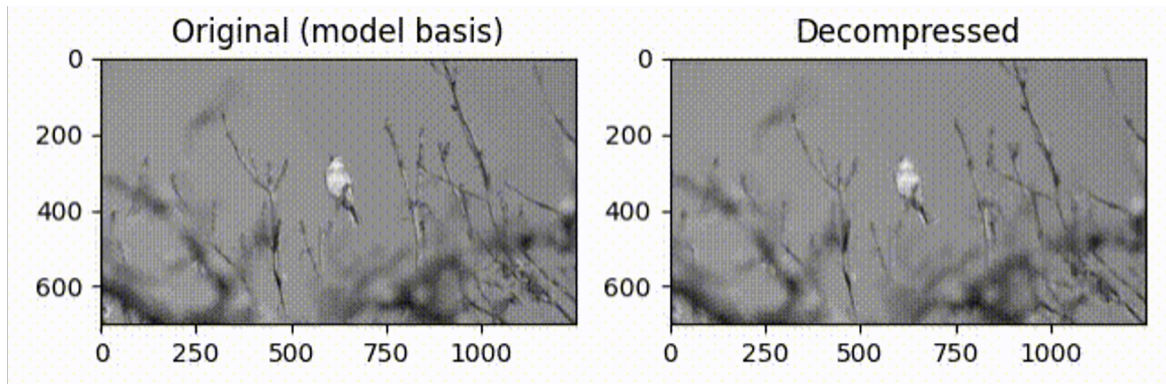
- Data consists of 2D slice of a liquid flowing over a cube
- The compressed file is **0.5%** the size of the input
- **Issue:** Model larger than input (4.2 MB vs 1.2 MB)

 0100 0010 1001	Original.npy	1,2 MB
 0100 0010 1001	Compressed.npy	6,1 kB



# Online vs offline

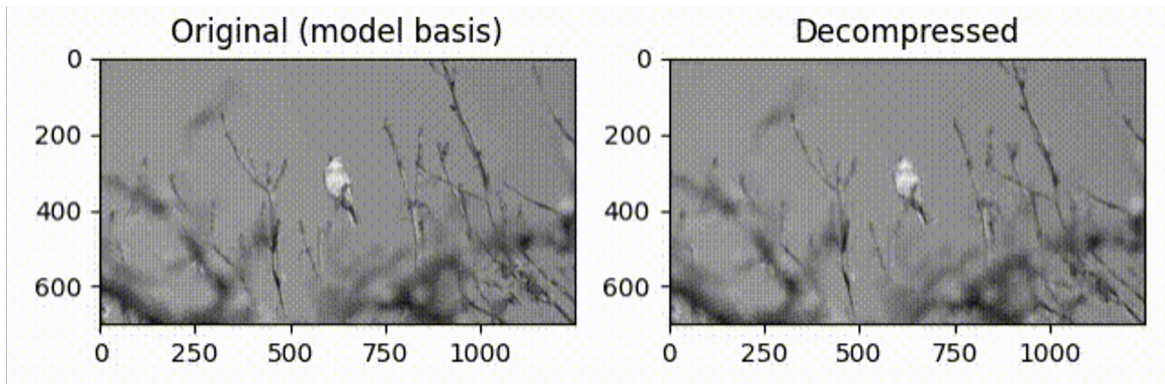
- Previously applied model trained on one dataset to the same dataset (*offline*)





# Online vs offline

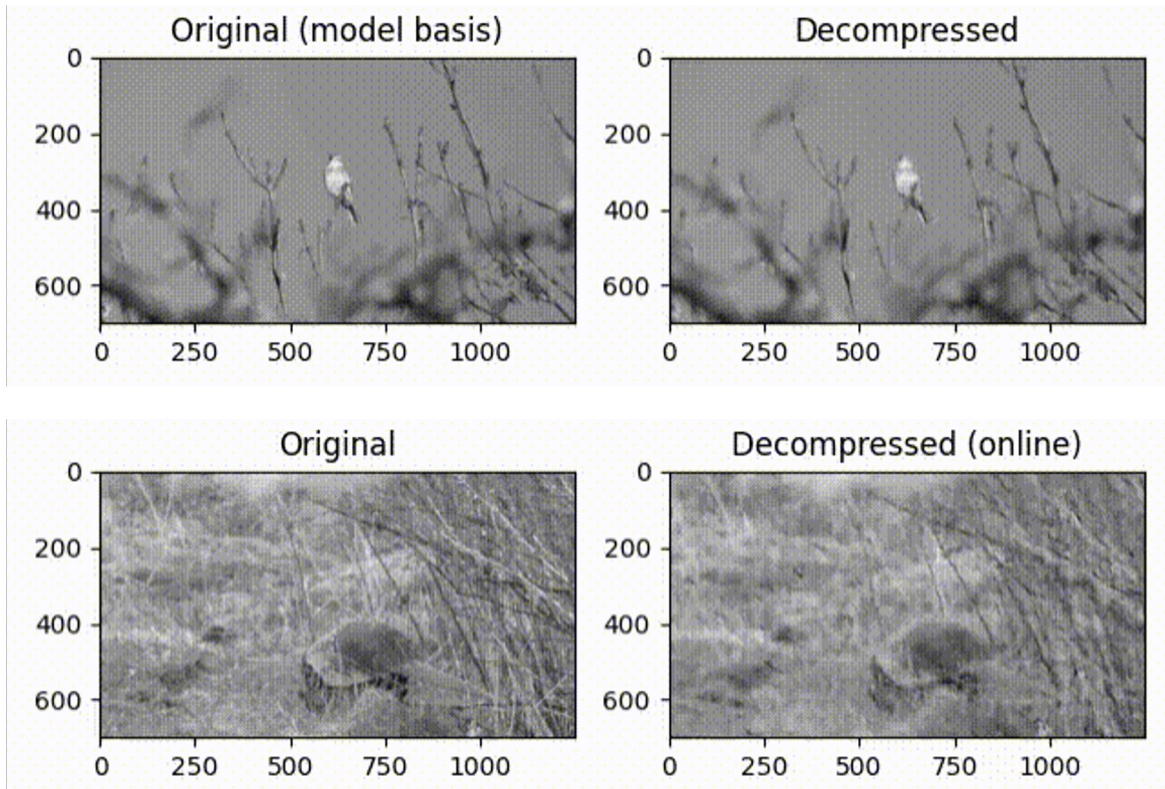
- Previously applied model trained on one dataset to the same dataset (**offline**)
- Can also apply to similar but unseen datasets (**online**)
  - Eliminate the cost of the model size!
- Useful for compressing **live data** (triggers, networks, etc)





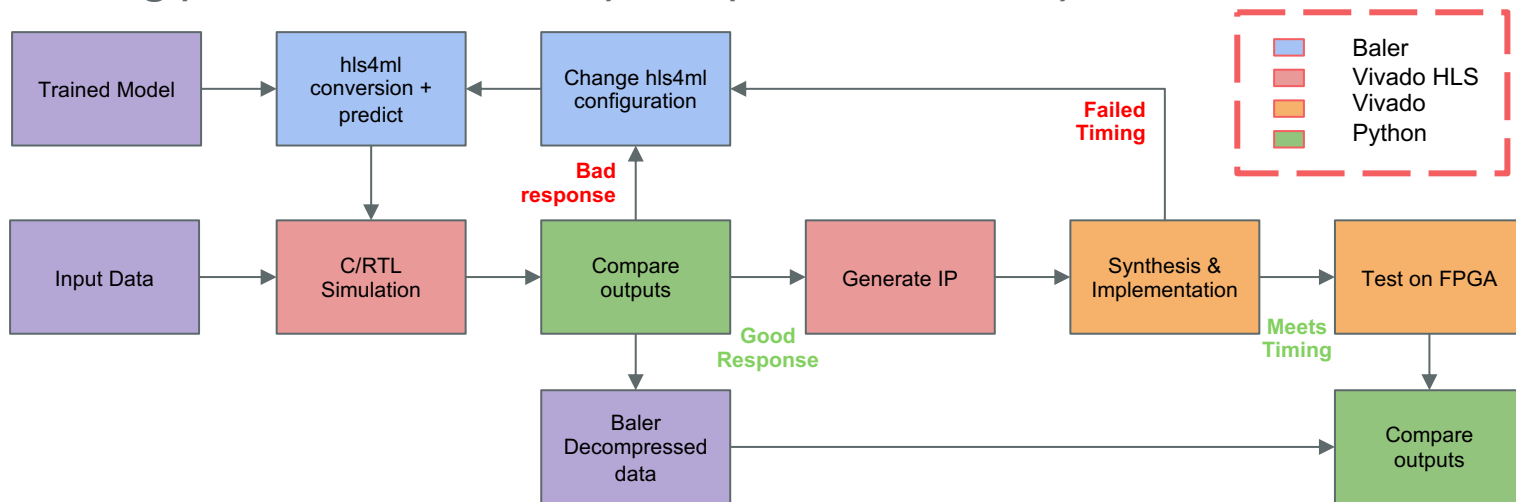
# Online vs offline

- 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)



# Baler on FPGA: Workflow

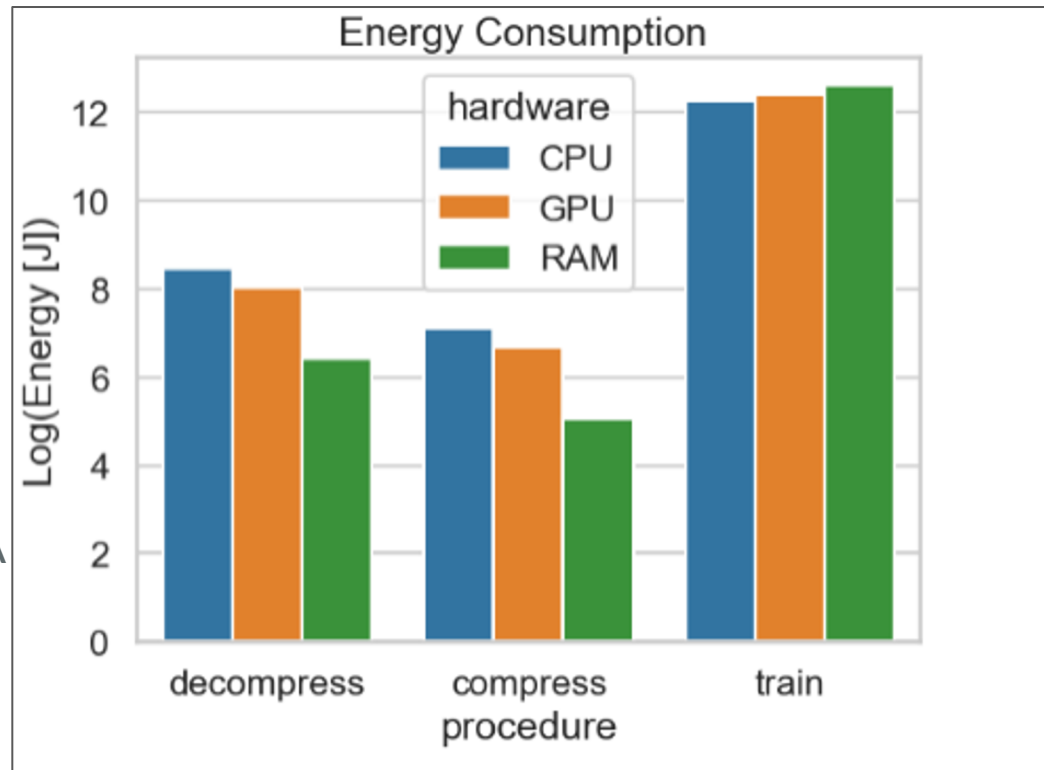
- Prototype version for developing and running Baler on an **FPGA**
  - Using vivado HLS code
- Useful in **bandwidth-restricted cases**
  - Network cards, detector readout, triggers, transmitters
- Assessing performance, latency and power efficiency



# Recent Results

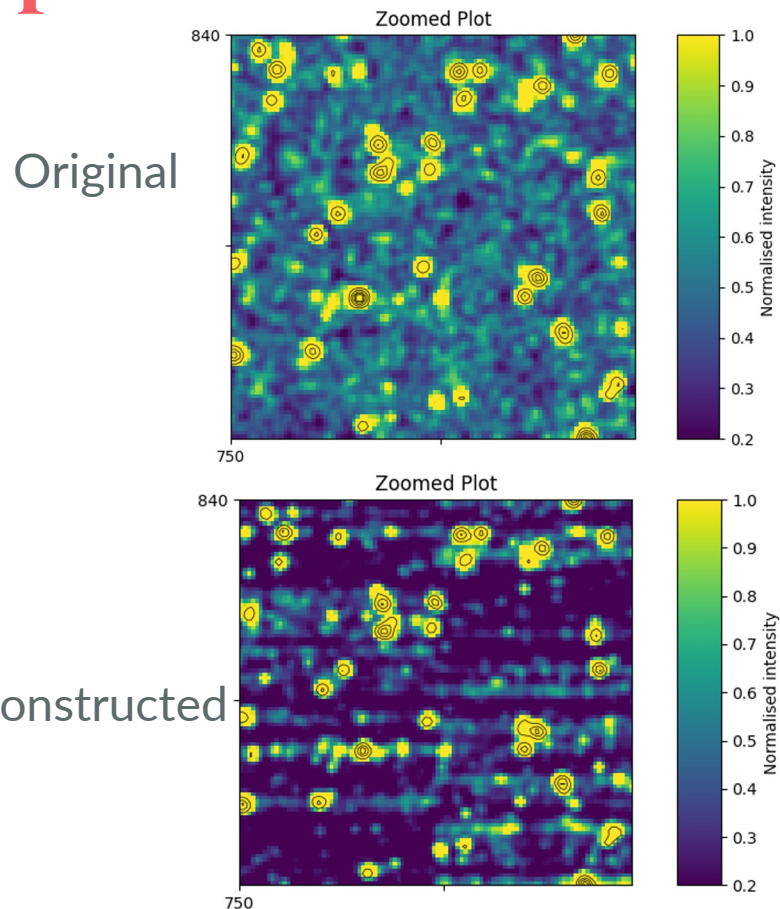
# Environmental impact of Baler

- Study by Leonid Didukh
- Need to compare energy consumption of CPU vs GPU
  - Faster not always better!
  - Energy is a main cost of big data
  - Substantial carbon footprint
  
- In future plan to compare FPGA power consumption as well



# Mössbauer imaging - potential solution to storage issues

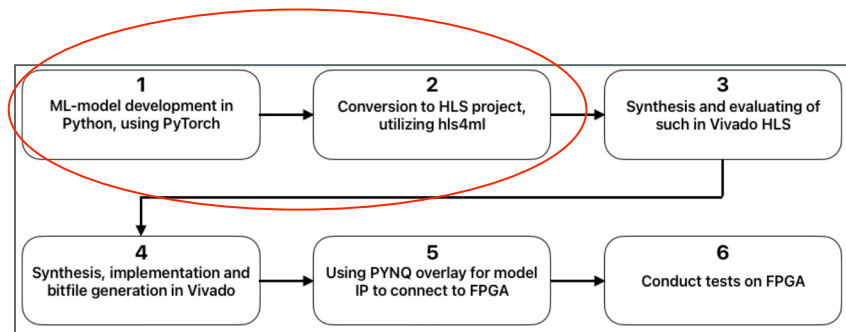
- Khwaisih Anjum, DESY
- Taking ~ 2 TB of images every ~ 5 days
  - Data acquisition uptime: > 75%
- Currently data is discarded when storage limit is reached
  - Saving compressed with some loss in quality is ok!



# Performance of ML-Based Bandwidth Compression on FPGAs

- Aleko Lilius, Lund University
  - Collaboration with the MAX IV laboratory
- Real-time compression of images on FPGAs
- Throughput increase of  $\sim 16x$  was achieved (compared to desktop CPUs)
  - Depending on model size
- Several key factors regarding ML performance on FPGAs was found

Done with Baler



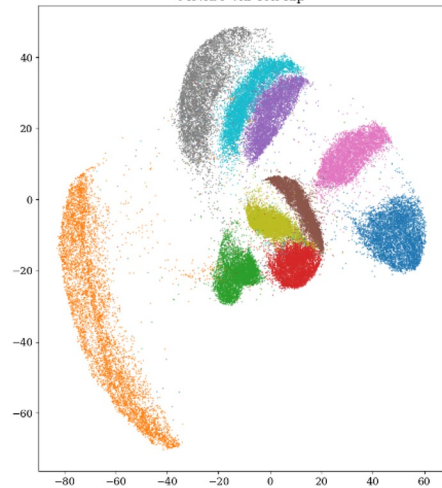
Model (Encoder)	Processing Unit	Time (s)	Throughput (inferences/s)
DNN Large	CPU	0.95	189473
DNN Large	FPGA	1.26	142377
DNN Reduced	CPU	0.94	191489
DNN Reduced	FPGA	0.23	768481
DNN Tiny	CPU	0.86	209302
DNN Tiny	FPGA	0.05	3472422



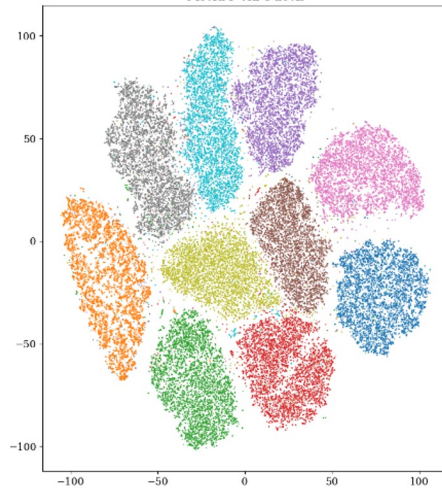
# Exploring Baler's Dimensionality Reduction Capacities via Latent Space Visualisation

- Malena Duroux, Manchester
- Visualization and comparison of latent space representations across Autoencoders on the MNIST dataset
- Interesting adjacent study on the Autoencoder subject

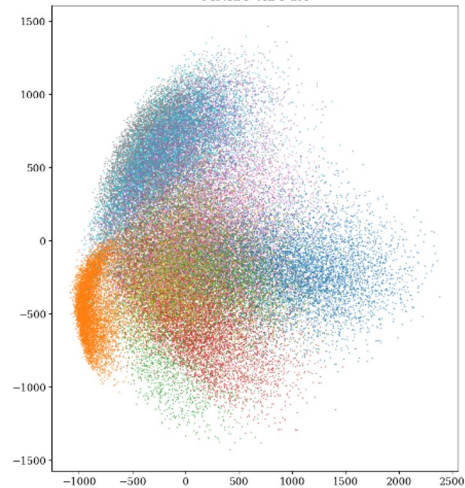
MNIST via TriMap



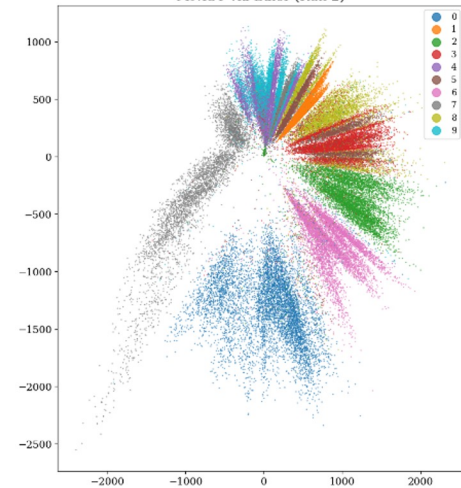
MNIST via t-SNE



MNIST via PCA



MNIST via Baler (Run 2)





# Summary

- **Baler** is a new toolkit for **compressing data** using **autoencoders**
- Capable of **impressive** compression results, but requires saving a **large model**
  
- Next steps: **FPGAs** for network or trigger applications & **online lossy compression**
  
- Careful management of a project can provide short tasks suitable both for junior members and academics with limited time

# Interested? Feedback? Contact us!

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

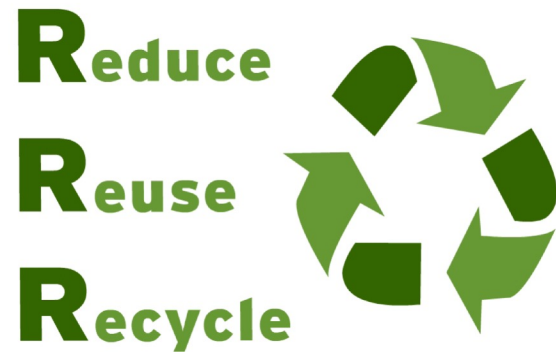
- <https://github.com/baler-collaboration/baler>
- [baler-compression-members@cern.ch](mailto:baler-compression-members@cern.ch)
- [axel.lars.gallen@cern.ch](mailto:axel.lars.gallen@cern.ch)
- [james.smith-7@manchester.ac.uk](mailto:james.smith-7@manchester.ac.uk)
- [caterina.doglioni@manchester.ac.uk](mailto:caterina.doglioni@manchester.ac.uk)



# Backup

# 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)

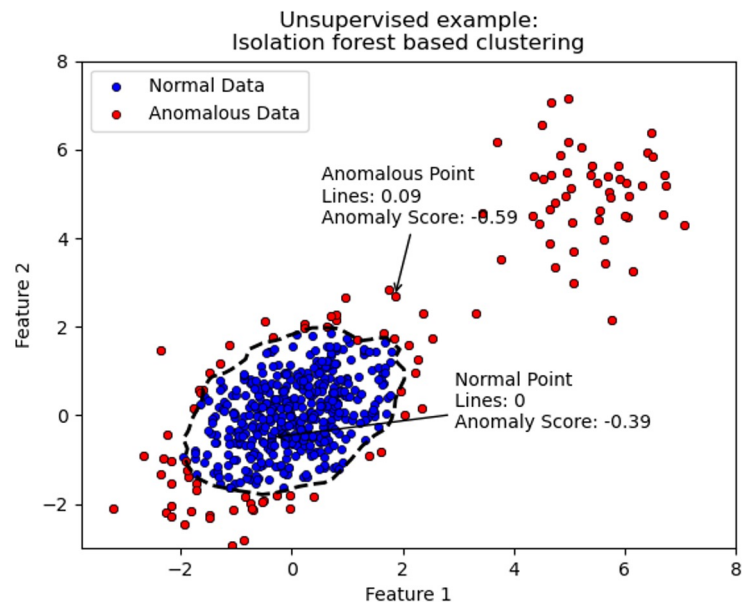


# Community Development

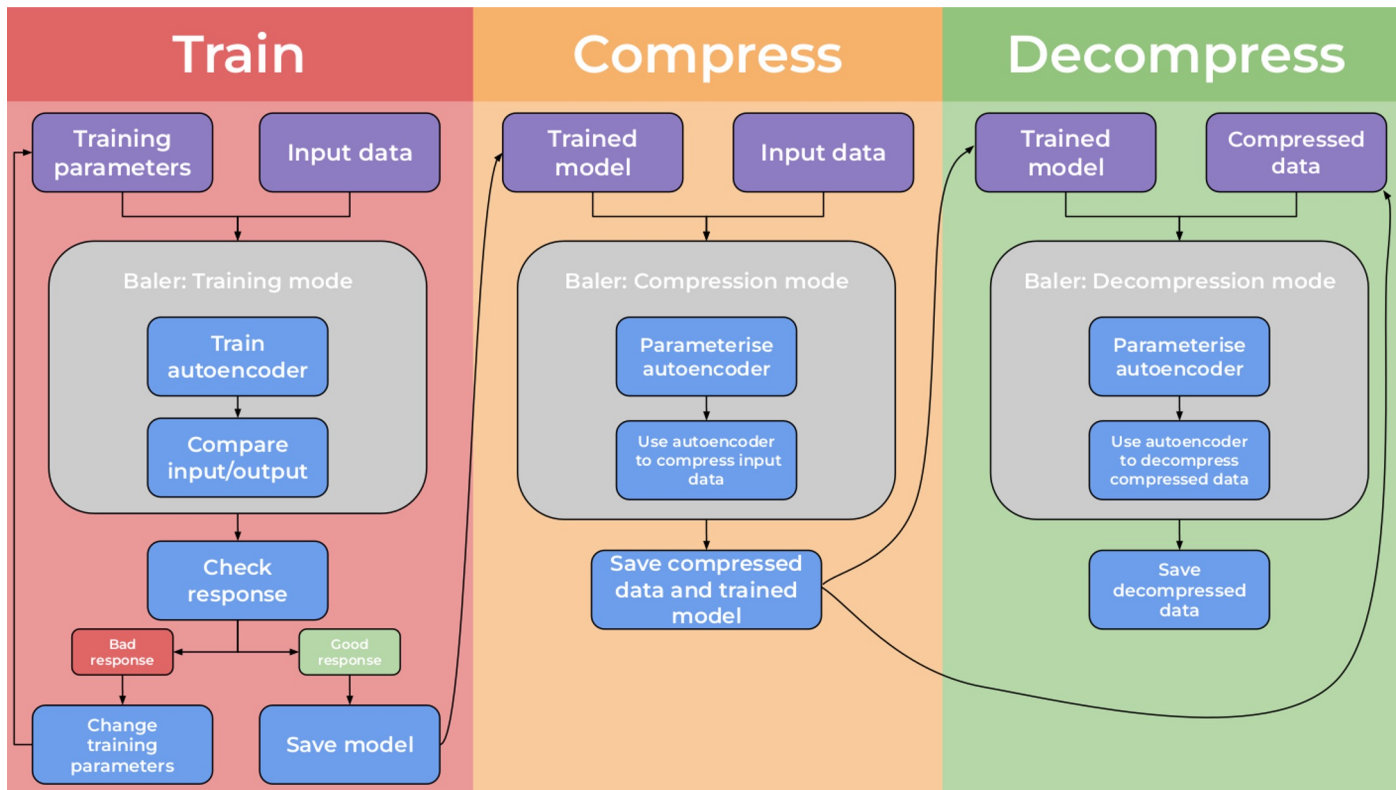
- Project driven by Early-Career Researchers
  - Main contributors for this project: undergraduate/Master's students and summer students/interns
    - Huge amount of high-quality work, but seasonal and limited prior experience
    - Strong tutorials and documentation essential for rapid onboarding
  - 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!

# 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
- Use a simplified version of Baler to build a **probability distribution** of points in latent space
- **Remove points** that significantly disagree, **iterate recursively**
- Performance **evaluation ongoing**
- Also exploring clustering and categorisation



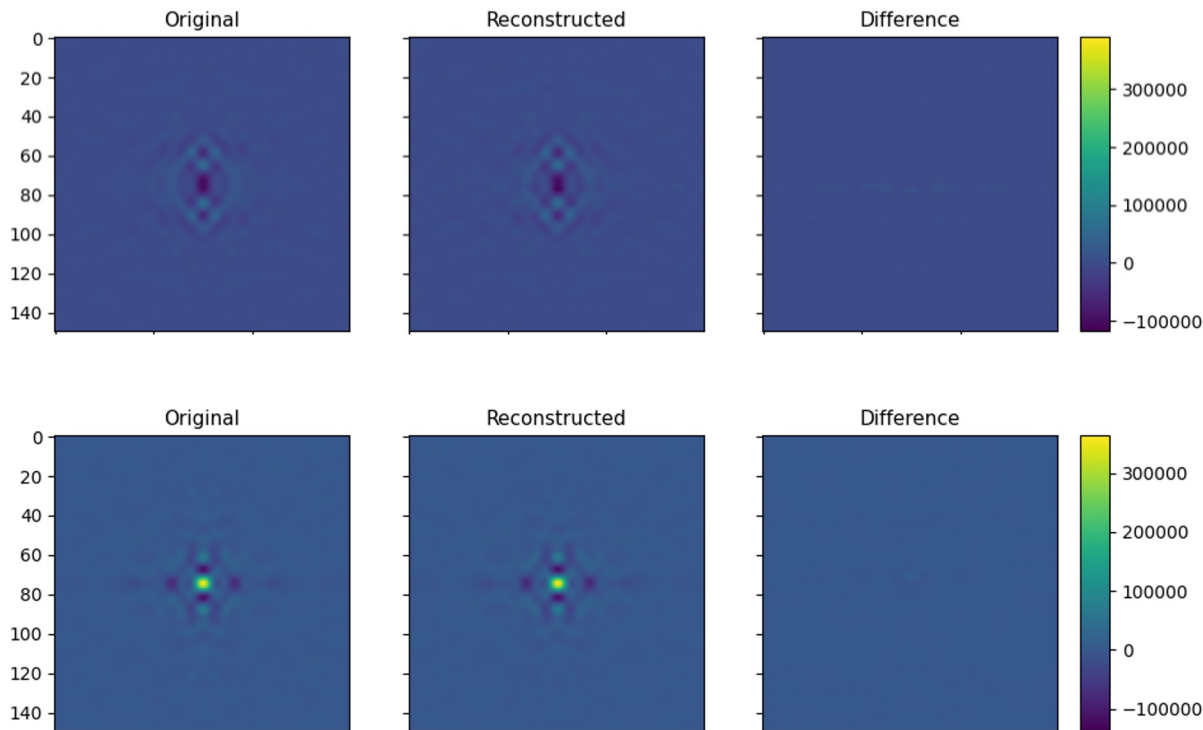
# Workflow





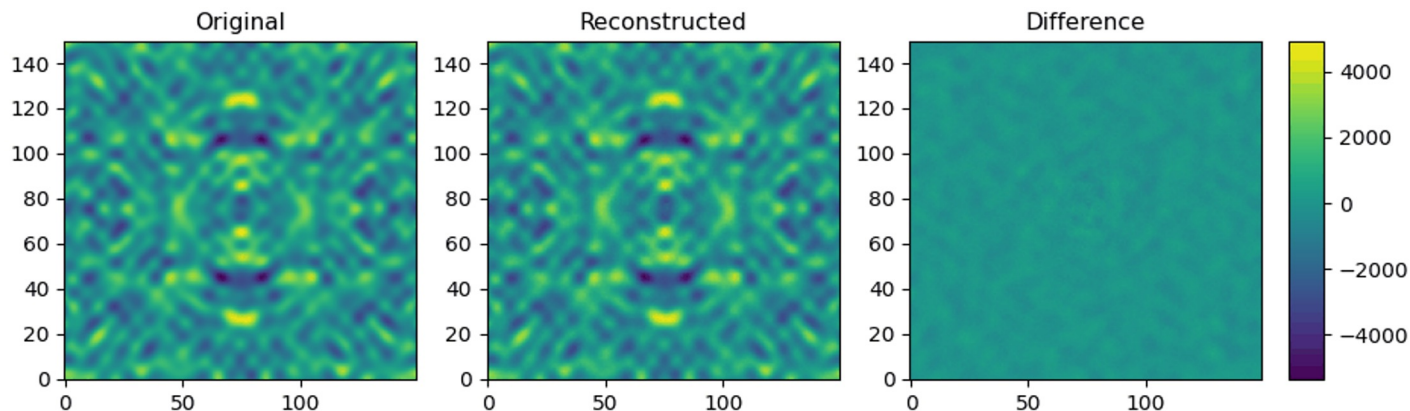
# Online vs offline (X-Ray Diffraction)

- 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)



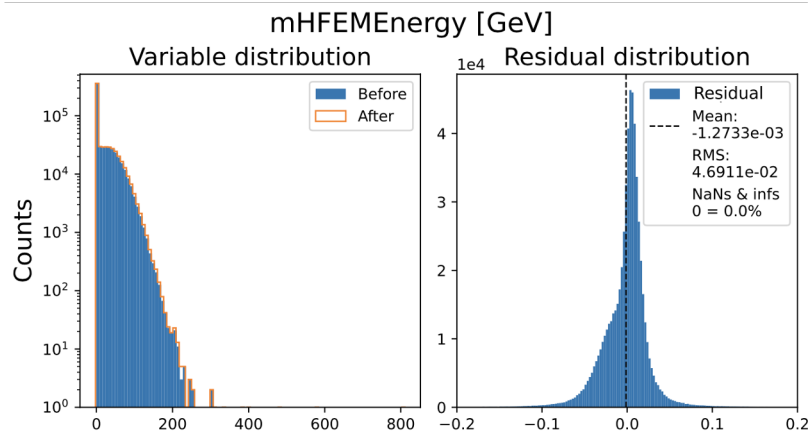
# X-Ray Diffraction

- “4M simulated diffraction images of chaperone 3iyf”
  - In actuality 151x151x151 array, which I split into two 75x150x15 arrays
- Train on one half to compress down to 0.001% the original size
- Used for compression of the other half
  - Actually great performance

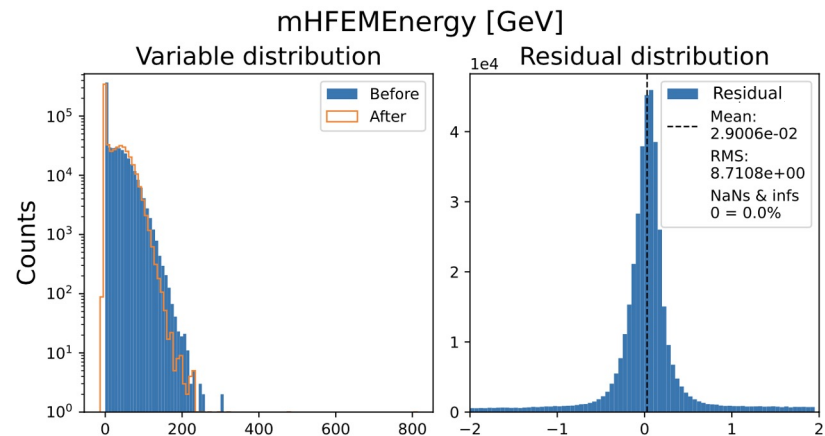


# 1.7x vs 6x compression

1.7x compression



6x compression



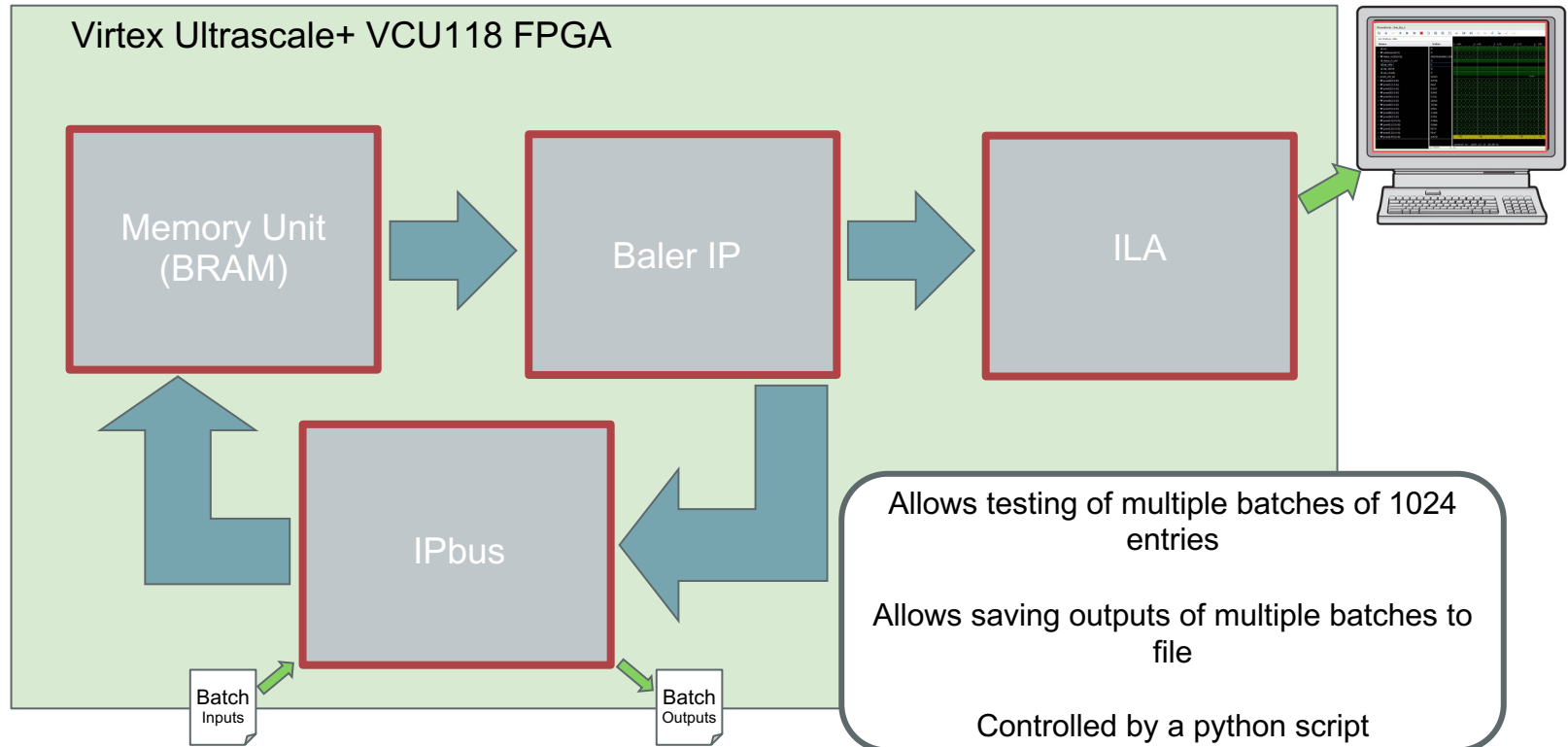
# Full variable list (see

<https://arxiv.org/abs/2305.02283>)

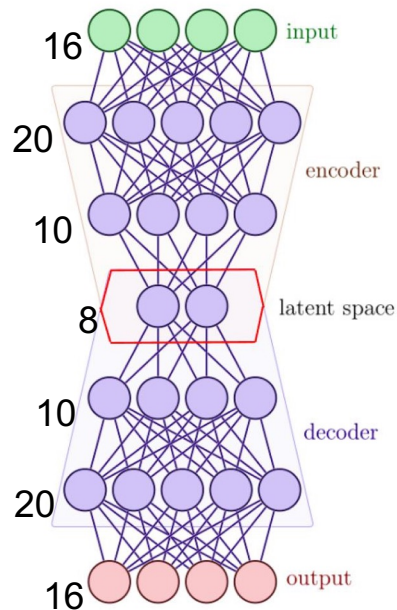
**Table 2:** Residual and Response distribution means and RMS values for all variables in the dataset. These values are presented at  $R = 1.7$ , and all values have been averaged over 5 runs, with an added statistical error of two standard deviations.

Variable ( $R = 1.7$ )	Response		Residual	
	Mean	RMS	Mean	RMS
$p_T$	$-1.07 \times 10^{-3} \pm 1.34 \times 10^{-2}$	$2.09 \times 10^{-2} \pm 3.56 \times 10^{-3}$	$-1.44 \times 10^{-2} \pm 1.04 \times 10^{-1}$	$2.12 \times 10^{-1} \pm 5.29 \times 10^{-2}$
$\eta$	$3.75 \times 10^{-4} \pm 6.11 \times 10^{-4}$	$8.12 \times 10^{-1} \pm 1.17$	$-1.12 \times 10^{-3} \pm 2.67 \times 10^{-3}$	$2.09 \times 10^{-3} \pm 1.45 \times 10^{-3}$
$\phi$	$3.44 \times 10^{-4} \pm 8.64 \times 10^{-4}$	$1.93 \times 10^{-1} \pm 4.32 \times 10^{-1}$	$2.45 \times 10^{-4} \pm 1.80 \times 10^{-3}$	$9.91 \times 10^{-4} \pm 1.12 \times 10^{-3}$
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	$-8.97 \times 10^{-4} \pm 1.42 \times 10^{-3}$	$1.56 \times 10^{-3} \pm 1.93 \times 10^{-3}$	$-5.36 \times 10^{-3} \pm 7.37 \times 10^{-3}$	$7.34 \times 10^{-3} \pm 6.60 \times 10^{-3}$

# Vivado Project - (in progress)



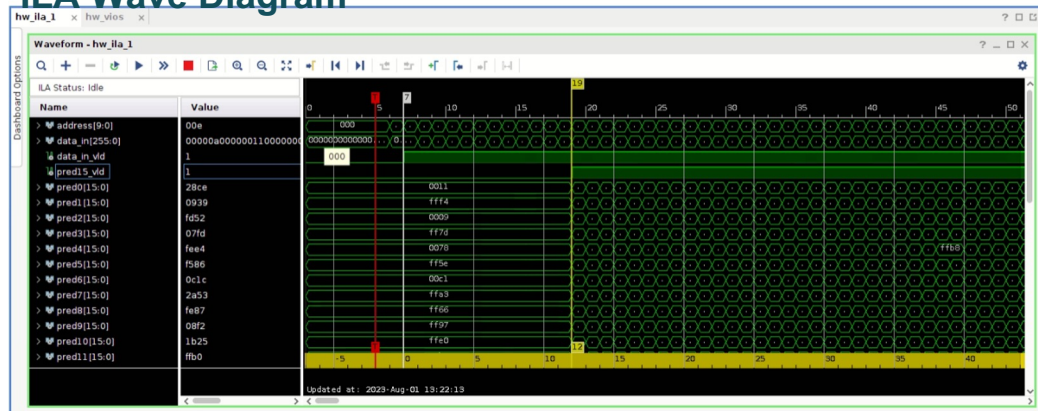
# Prototype Specifications



## Resource Utilization

Name	CLB LUTs (1182240)	CLB Registers (2364480)	CARRY8 (147780)	F7 Muxes (591120)	F8 Muxes (295560)	CLB (147780)	LUT as Logic (1182240)	LUT as Memory (591840)	Block RAM Tile (2160)	DSPs (6840)	Bonded IOB (832)	HPIOB M (384)	HPIOB S (384)	HPIOB DIFFN BUF (720)	GLOBAL CLOCK BUFFERS (1800)	MMCM (30)	BSCAN2E (12)
<b>baler_top</b>	24545	10229	2535	125	28	5129	23862	683	38	653	2	1	1	1	2	1	1
<b>baler (tiny_model_0)</b>	21889	4948	2462	0	0	4284	21889	0	0	653	0	0	0	0	0	0	0
clk_inst (clk_wiz_0)	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0
dbg_hub (dbg_hub)	461	753	7	0	0	159	429	32	0	0	0	0	0	0	1	0	1
ila_inst (ila_0)	2080	4272	66	125	28	794	1429	651	30.5	0	0	0	0	0	0	0	0
mem_inst (blk_mem_0)	0	0	0	0	0	0	0	0	7.5	0	0	0	0	0	0	0	0
vio_inst (vio_0)	99	231	0	0	0	52	99	0	0	0	0	0	0	0	0	0	0

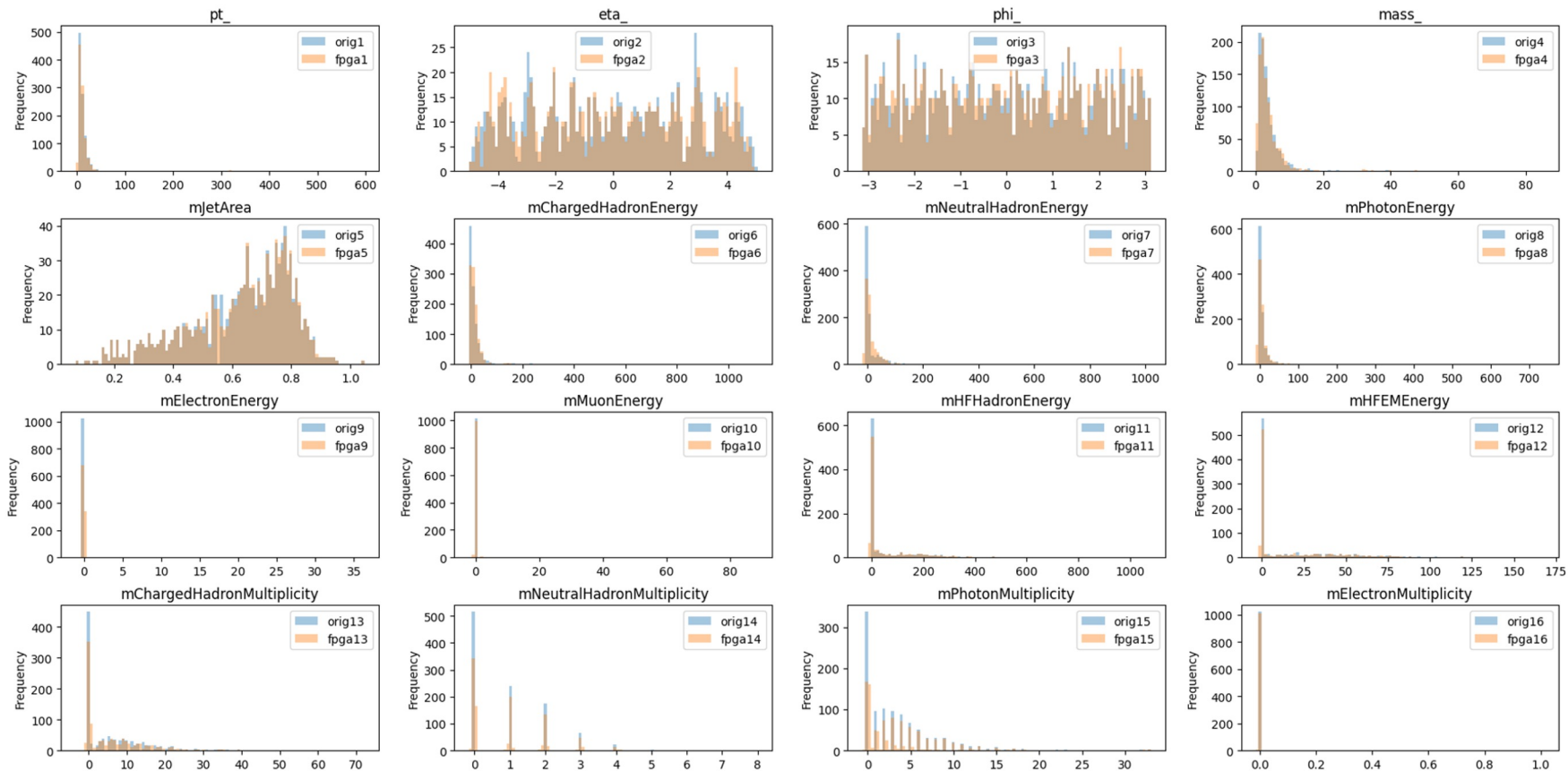
## ILA Wave Diagram



## Synthesis Timing Estimation

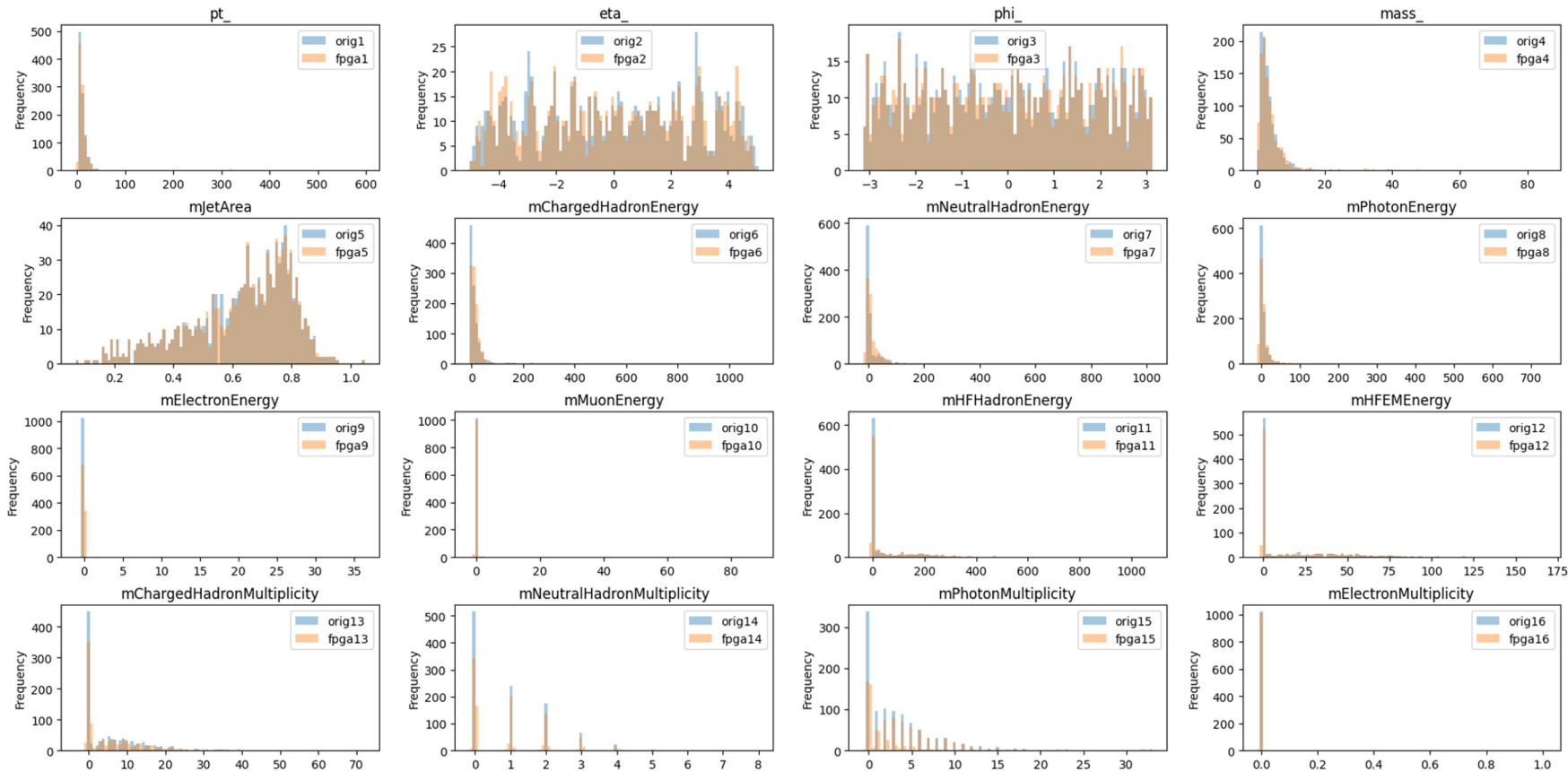
Latency (cycles)		Latency (absolute)		Interval (cycles)		Type
min	max	min	max	min	max	
12	12	60.000 ns	60.000 ns	1	1	function

# Preliminary Results: Data vs FPGA





# Preliminary Results: GPU vs FPGA



# Community Development: ECRs

- Main contributors for this project: undergraduate/Master's students and summer students/interns  
(IRIS-HEP Ukraine Fellows + GSoC through HSF + Trilateral Data Science exchange programs)
  - **Huge amount of high quality** work completed by these junior members
  - Students need **training** - often limited or no prior software or ML experience
- Leadership and development decisions made by PhDs and PostDocs
  - **James Smith (Mancs)**
  - Support from (busy) academics (C. Doglioni (Mancs), N. Skidmore (Warwick))
- Short-term members can be **variable and seasonal**
  - Good **OS code** and **documentation** essential for **fast onboarding**
  - **Timeline planning** is important both for student, and for summer breaks when there are only academics with too many other projects
  - **Careful selection** key - a poorly designed/matched project can cost more time than you gain!
  - **Well defined, short projects** good both for students and academics with limited time!

# Community Development: Funding & Resources

- Use all the funding sources you can find, however small!
- BALER Members are part of EVERSE, SMARTHEP, and other large grants (Horizon 2020 ERC Consolidator grant, national grants)
- Received smaller grants to fund specific projects for short timelines - useful for hiring RSEs from local pool, materials (FPGA boards), for semi-annual meetings, etc
- Received funding from departmental/institute level as well
  - Received two “pump-prime” grants, but also funding for strategic international collaborations
- Also working with local (Swedish) industry students, “in-kind” resources

# Community Development: Industry

- It's easy to get academics interested in your project...
  - But limited time/funding
  - 5-10 people with 5-10% of time each leads to slow progress
  - Work with RSEs / engineers! Speak to your research support team
- Industry much harder to attract
  - Email & cold-call a lot
  - Easier to contact companies in same city as your institute that have prior experience with your institute!
  - Must stress you're not after a job / their money!
  - However can provide useful experience in best practices and making your code easy to use by other people