TREASURE

Future Collaborations

Data Compression

Marco Montella, with input from:

Akshat Gupta, Sanjiban Sengupta, Caterina Doglioni,
Antonio Boveia, Thomas Elliott

Treasure & Friends Meeting 2026.06.17

Our Initiative: ML lossless compression

- Explore ML-driven data compression methods to:
  - Expand Real Time Analysis on collider data in a fixed-budget model
  - Reduce data-recording & storage costs across experiments

Real Time Analysis dataset "forecast"

ATLAS (TLA)	CMS (Scouting)	LHCb (Real-time)	DUNE (streaming)
Record EF trigger-level objects at 400 kHz (Goal: ~50% of events seen by EF)	L1 scouting + parking streams (100's of kHz of lightweight events)	30 MHz real-time reconstruction via Allen/GPU	Streaming readout, 100+ interactions per spill, full ND-LAr waveforms (~1 PB/yr)

- Both Lossy and Lossless compression models under study for ATLAS, CMS, LHCb, and DUNE by a cross-experiment team involving information theory experts
  - As summarised in Antonio Boveia's talk at April TREASURE Workshop

- **TODAY's GOALS: → Outline recent progress**
  → Highlight potential areas for collaborations

Lossless Compression: Recent Progress

BOA constrictor: a Mamba-based lossless compressor for scientific data

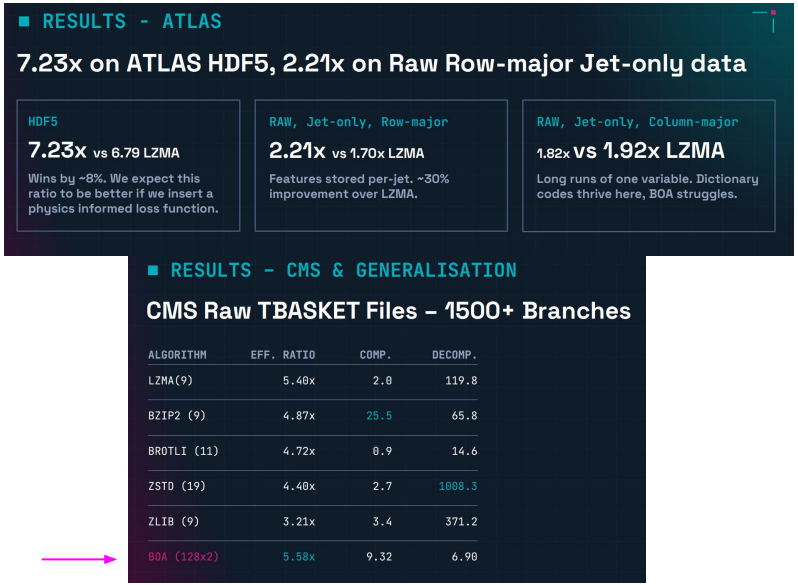
Akshat Gupta *et al* 2026 *Mach. Learn.: Sci. Technol.* 7 035014

FRESHLY PUBLISHED!

In-depth presentation of work by A. Gupta at CHEP2026

HIGHLIGHTS

Very competitive compression ratios across data formats / experiments



More work necessary to be competitive with standard libraries

Short-term directions of work

- **FOUR Summer Students** to work on compression in collaboration with University of Manchester / OSU.

Stefan Ionescu (OSU)

Research: Lossless compression on High-Luminosity ATLAS Trigger-Level data

Shanu Sahlot (GSoC / UoM)

Research: Physics informed bytestream compression implementation

Zhengkai Sun (UoM)

Research: Compression resources vs Disk resources tradeoff

- **N8CIR**

Research: Sustainable data compression

Potential Collaboration / Alignment

- **Several areas of potential alignment:**
  - Understanding orthogonality between tokenization and ranged compression
  - Assess the feasibility of a “general map” for BOA with acceptable cross-experiment performance
  - Compression as a potential vehicle for open data distribution

  - Submitted cross-experiment Phase-I *Genesis* proposal on AI Compression for HEP
    for broader benchmarking across many datasets, optimization, and entropy floor studies
    Regardless of review outcome, the core team will pursue this with the aim of joining a Phase-II proposal

We would be happy to kickstart a more in-depth discussion to potentially steer our upcoming work along directions of common interest!