Evaluating the SZ3 compressor on data from the ATLAS experiment

Author: Amy Byrnes (UIC, ALCF)

Mentors: Serhan Mete (ANL, ATLAS),

Peter Van Gemmeran (ANL, ATLAS), Jahred Adelman (NIU, ATLAS)

Advisor: Michael E. Papka (ALCF, UIC)

Motivation

- Data collected from HL-LHC will strain storage capacity
- There is a hard limit to how much we can shrink data via lossless compression

What is data compression?

- **Compression** is a technique that allows us to store the same amount of information using fewer bits

- In the context of floating-point data:

  - In the standard encoding for floats, each float is 32 bits, or 4 bytes

  - In a sequence of *compressed* floats, the average bits/value is <32

Lossless vs Lossy compression

▶ **Lossless compression**

   ▶ Original data can be recovered exactly from compressed data

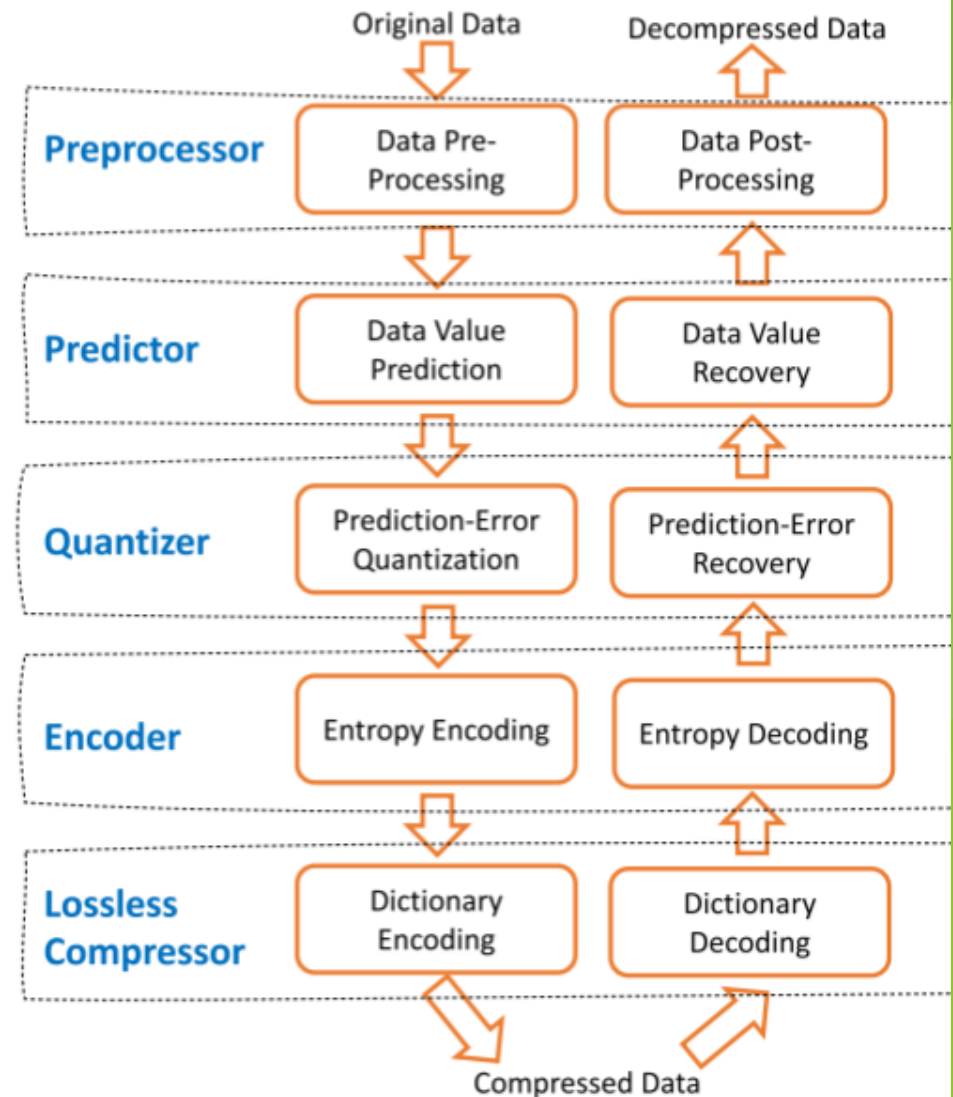   ▶ Major trade-off is **compression ratio** vs **computation**
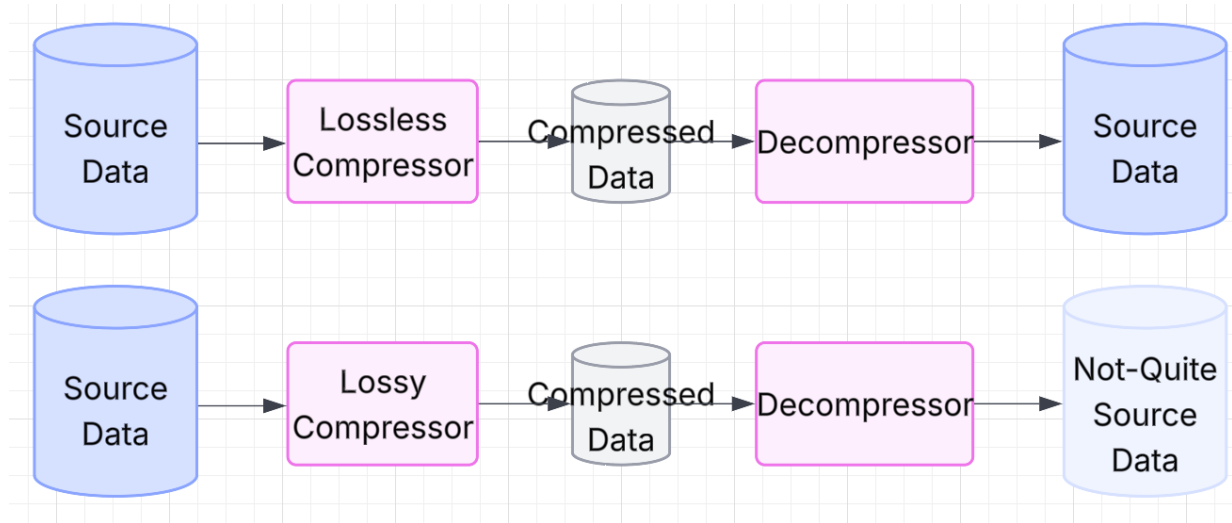
▶ **Lossy compression** refers to

   ▶ Data is permanently altered by the compression process

   ▶ Additional trade-off: **compression ratio** vs **distortion rate**

SZ3

- "A Modular Error-bounded Lossy Compression Framework for Scientific Datasets"
- Developed from 2016 at Argonne National Laboratory
- Designed primarily for HPC applications
- Major components:
  - Predictors
  - Error-bounded quantizer
 - Ability to compose a custom compression pipeline

Pipeline/Module Abstraction

	Original Data	Decompressed Data
Preprocessor	Data Pre-Processing	Data Post-Processing
Predictor	Data Value Prediction	Data Value Recovery
Quantizer	Prediction-Error Quantization	Prediction-Error Recovery
Encoder	Entropy Encoding	Entropy Decoding
Lossless Compressor	Dictionary Encoding	Dictionary Decoding

Compressed Data

Methodology

▶ Dataset:

- ▶ DAOD_PHYSLITE.37019878
- ▶ 21 files
- ▶ 13.4 GB
- ▶ Data size for each branch ranges from 0.27 – 10.14 MB

▶ Branches:

- ▶ AnalysisJetsAuxDyn.pt or "Jet pt"
- ▶ AnalysisJetsAuxDyn.phi or "Jet phi"
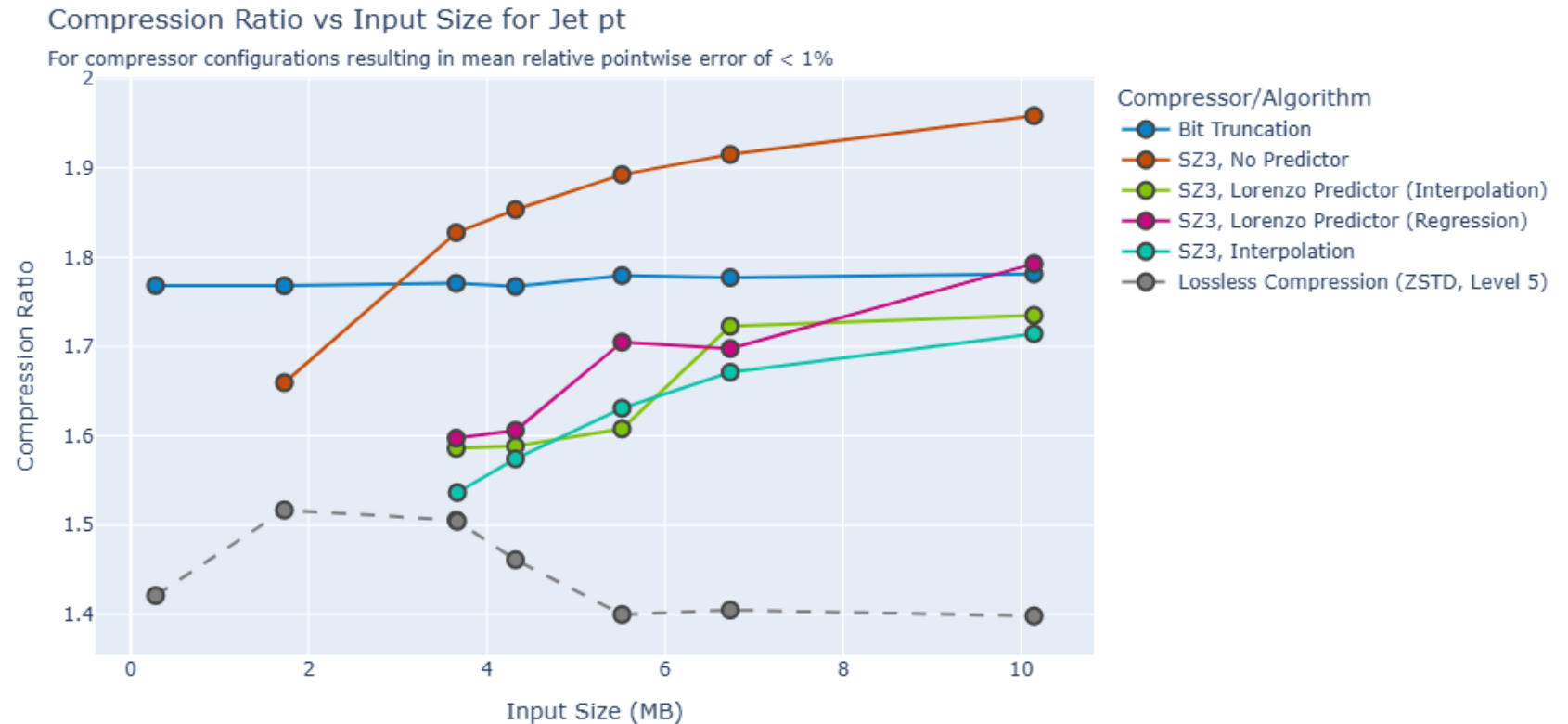- ▶ AnalysisJetsAuxDyn.eta or "Jet eta"
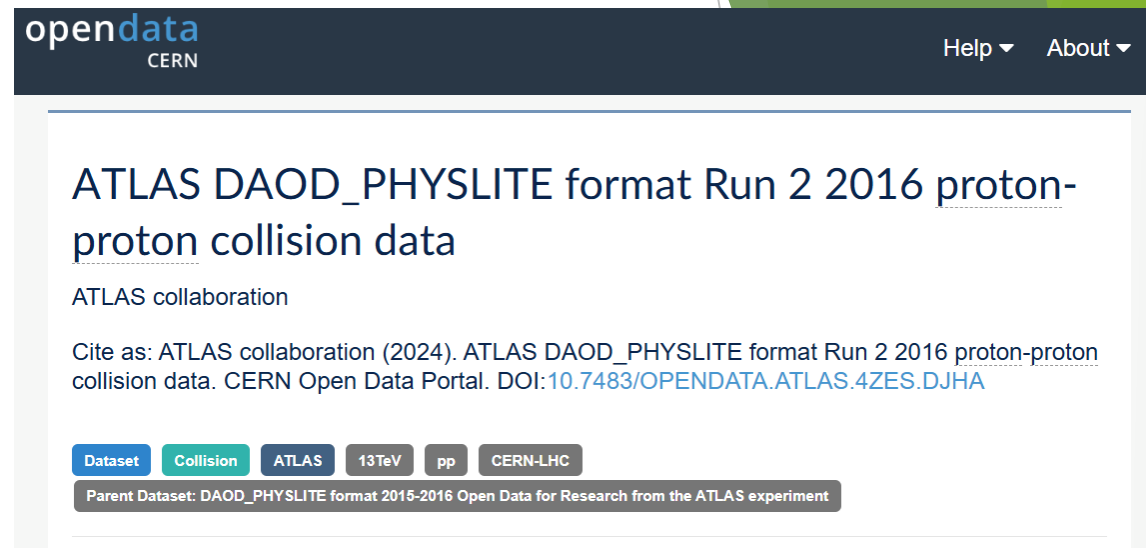
Methodology

- Bit truncation
  - Truncate floats to have < 23 mantissa bits
    - 16, 15, 14, 13, 12, 11, 10, 9, 8
  - Compress truncated floats with zlib at level 5
- SZ3
  - Lossy algorithms
    - Lorenzo predictor, regression
    - Lorenzo predictor, interpolation
    - Interpolation predictor
    - No predictor
  - Absolute error bounds:
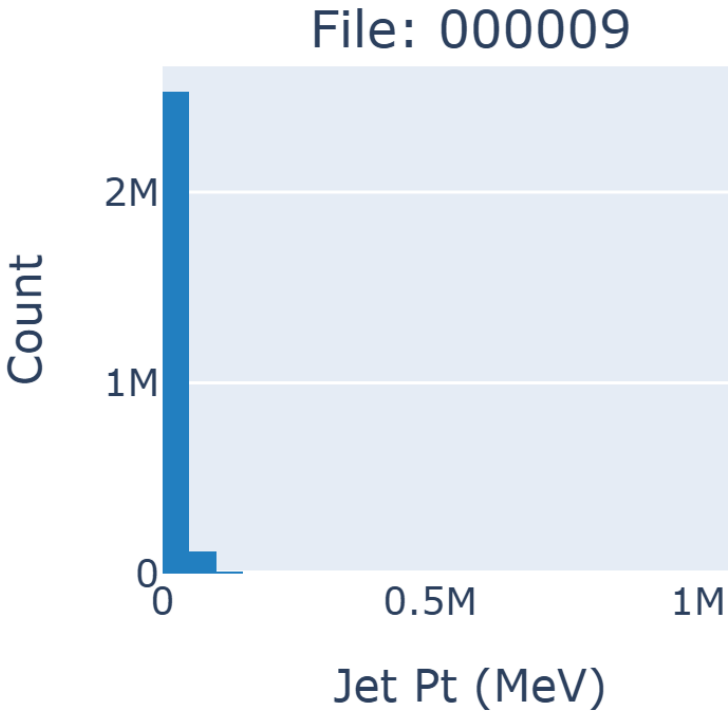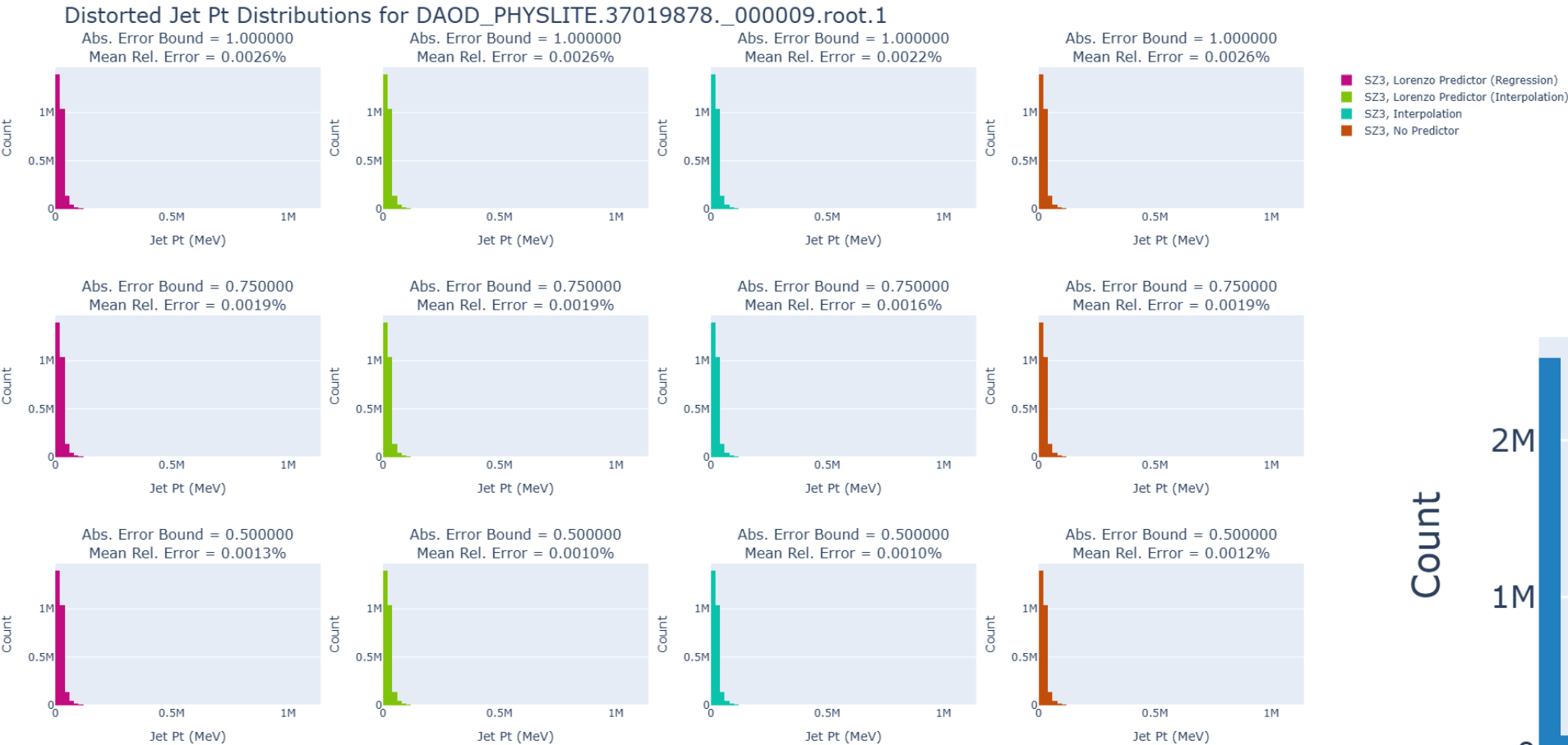   - 1, 0.75, 0.5, 0.25, 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001

SZ3 Performance Overview

- Observation 1: SZ3's performance depends on input size

- Observation 2: SZ3 "fails" when input size or error bound are too low

  - "Failure" in this case means quantizer cannot guarantee data will remain within the requested error bound
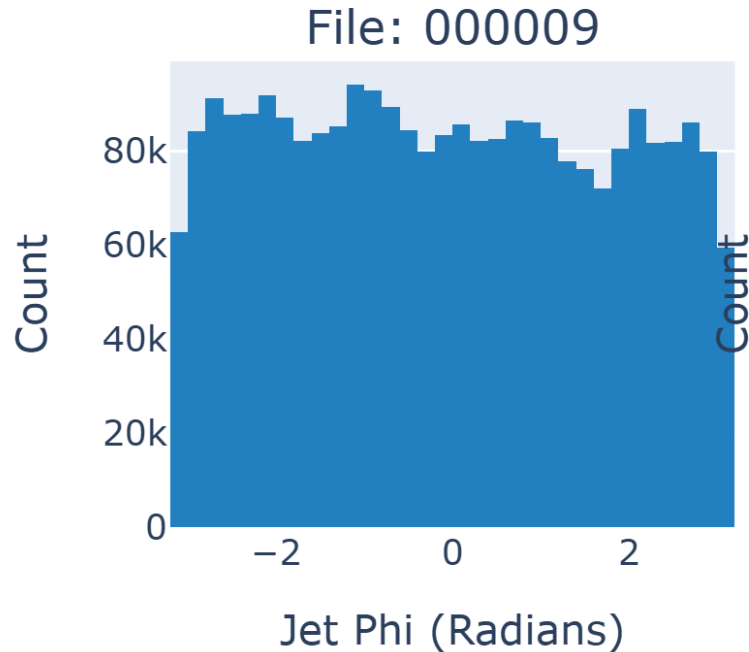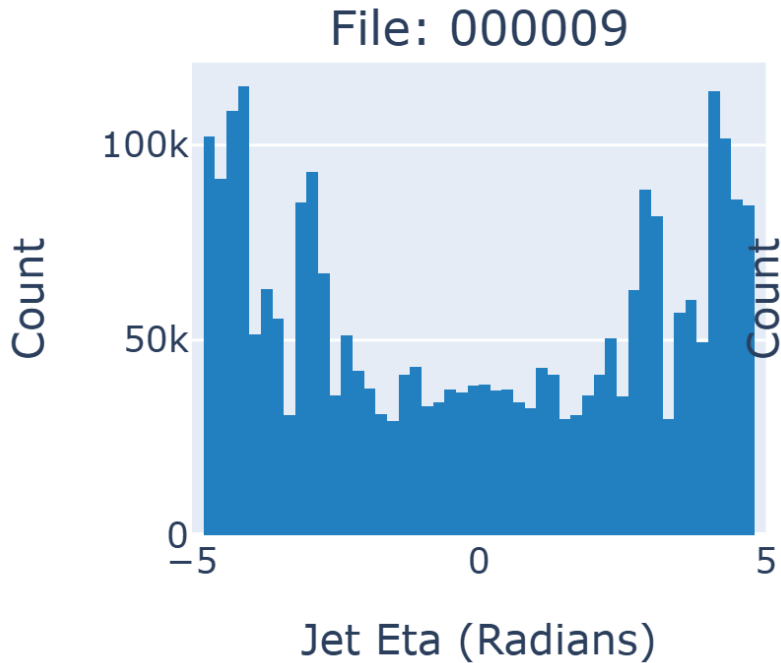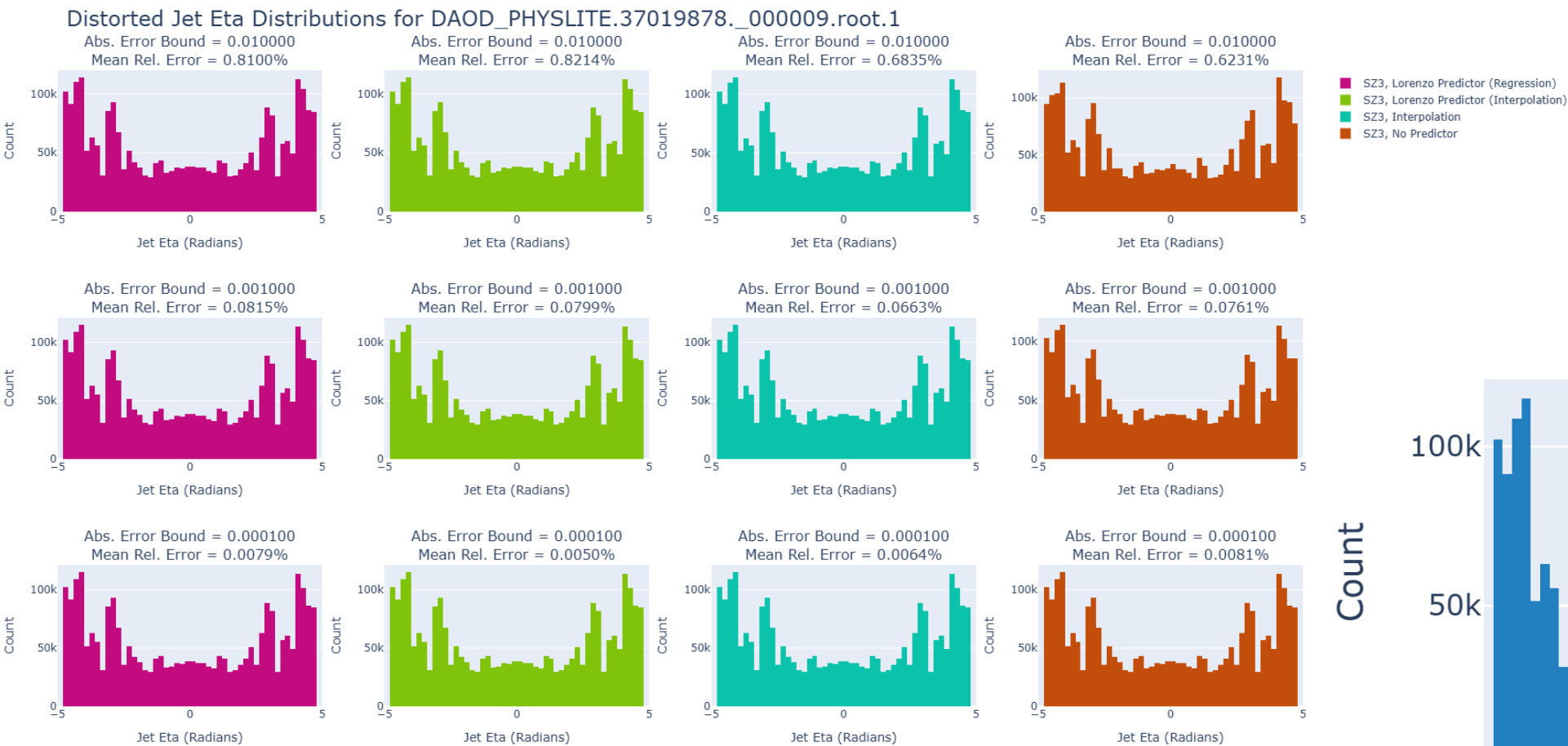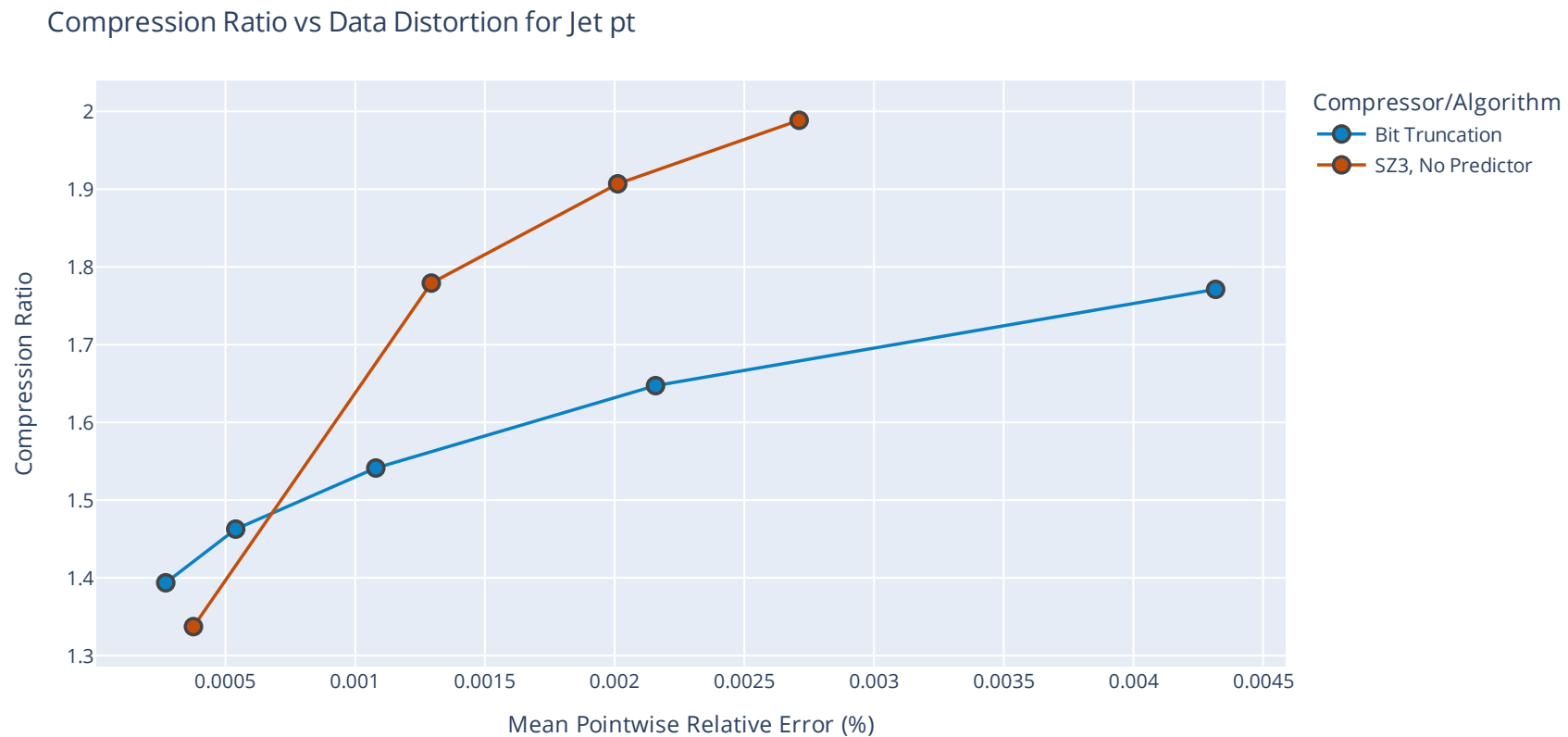

Compression Ratio vs Input Size for Jet pt

For compressor configurations resulting in mean relative pointwise error of < 1%

Compressor/Algorithm
- Bit Truncation
- SZ3, No Predictor
- SZ3, Lorenzo Predictor (Interpolation)
- SZ3, Lorenzo Predictor (Regression)
- SZ3, Interpolation
- Lossless Compression (ZSTD, Level 5)

Distorted data: Jet pt

Distorted Jet Pt Distributions for DAOD_PHYSLITE.37019878._000009.root.1

File: 000009

Distorted data: Jet phi

Distorted Jet Phi Distributions for DAOD_PHYSLITE.37019878._000009.root.1

File: 000009

Distorted data: Jet phi

Distorted Jet Eta Distributions for DAOD_PHYSLITE.37019878._000009.root.1

File: 000009
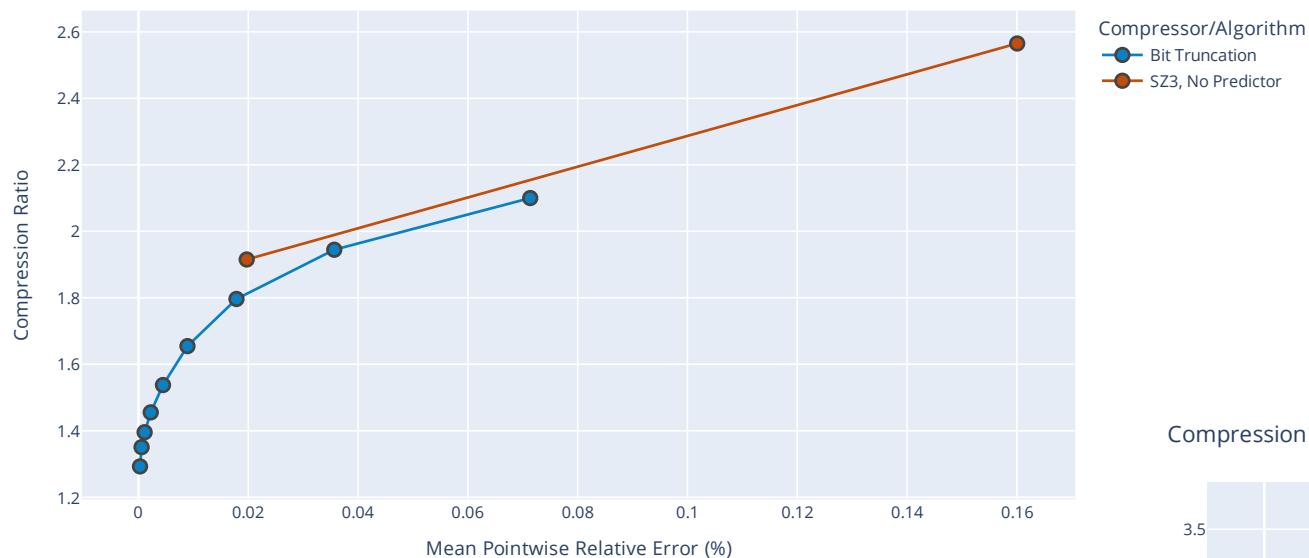
SZ3 vs Bit Truncation: Distortion

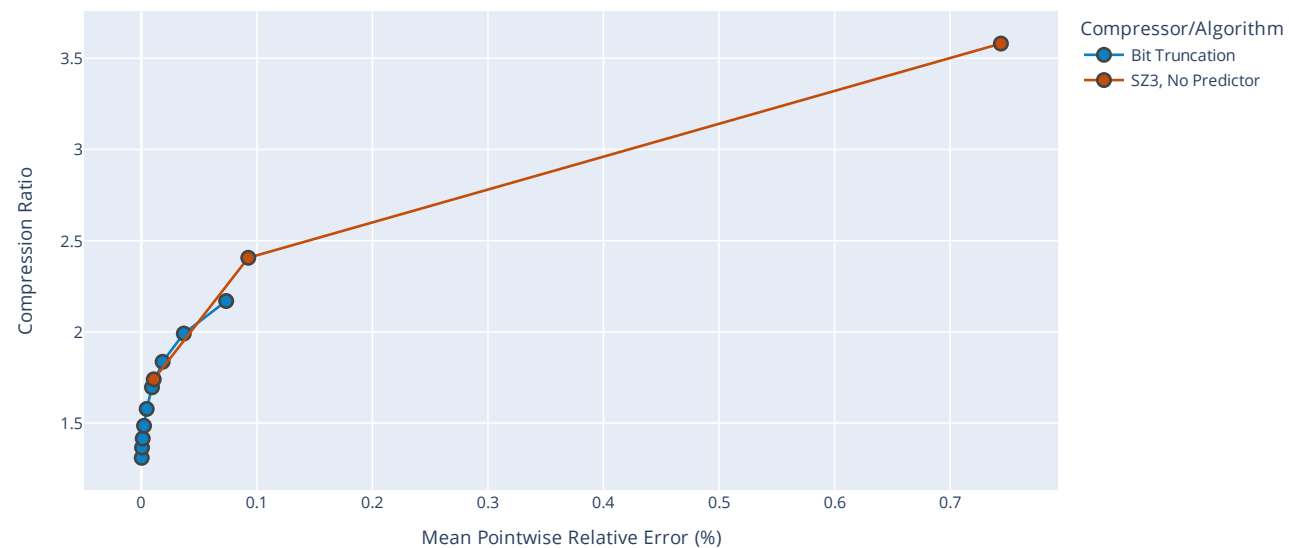Compression Ratio vs Data Distortion for Jet pt

SZ3 vs Bit Truncation: Distortion

Compression Ratio vs Data Distortion for Jet phi

Compression Ratio vs Data Distortion for Jet eta

References & Acknowledgements

▶ SZ3 Framework: Xin Liang, Kai Zhao, Sheng Di, Sihuan Li, Robert Underwood, Ali M Gok, Jiannan Tian, Junjing Deng, Jon C Calhoun, Dingwen Tao, Zizhong Chen, and Franck Cappello. "SZ3: A modular framework for composing prediction-based error-bounded lossy compressors", IEEE Transactions on Big Data (TBD 22).

▶ "ATLAS DAOD_PHYSLITE format Run 2 2016 proton-proton collision data" https://opendata.cern.ch/record/80001