USE OF INTELLIGENT COMPRESSION IN WAVEFORM LIKE DATA OF NEUTRINO OSCILLATION EXPERIMENTS

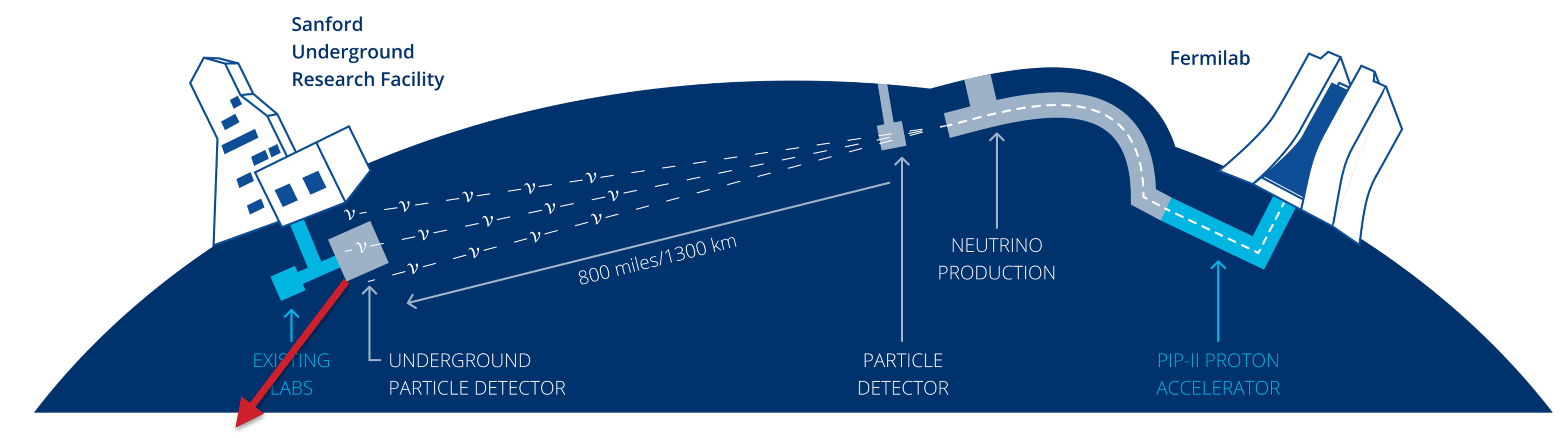
Emma Weiler (University of Manchester, United Kingdom)
Amit Bashyal (Argonne National Laboratory)

ABSTRACT

Many High Energy Physics experiments store their data in compressed format using lossless algorithms. The use of lossy compression algorithms is less common in the HEP. However, as the storage requirements of the HEP experiments will grow during the HL-LHC and DUNE era, intelligent lossy compression algorithms can enable significant storage optimization. The ability to highly compress data while preserving enough fidelity for subsequent physics analysis and interpretation could allow HEP experiments to limit storage needs so compressed objects can be part of down- stream/derived data products. This could allow experiments to use these objects when processing the derived data without reading the original raw data which is often much larger and harder to access, specially for upcoming experiments like DUNE [1].

MOTIVATION

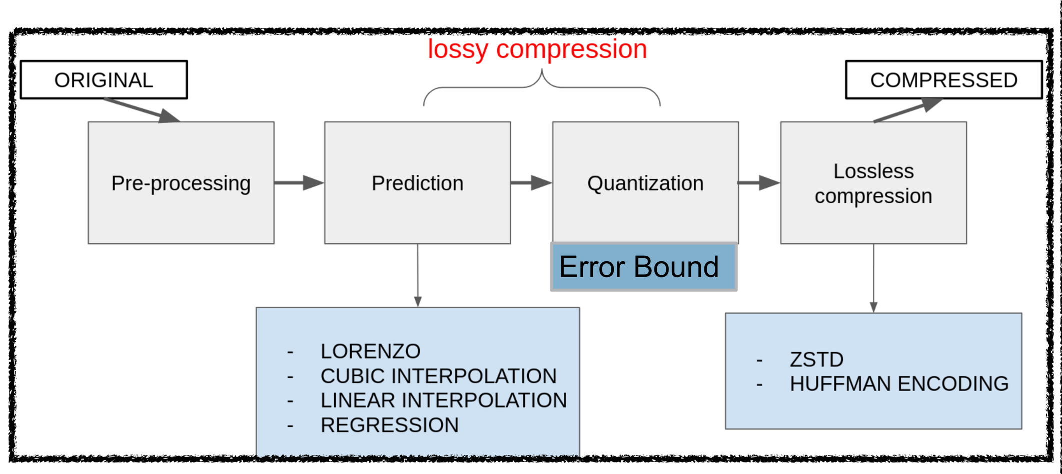
- Introduction of modern intelligent lossy compression algorithms in HEP field where raw data is typically lossless compressed.
- Oscillation experiment like DUNE will have large trigger data but simple data model due to large and homogenous far detector.
- Compressed data with enough fidelity can be used as resident data for inspection of reconstructed data.



Far detector of the DUNE experiment in SURF will consist of four 17 kiloton Liquid Argon Detectors in the underground cavern. Raw data from each trigger consists of readout from millions of channels from the detector electronics.

METHODOLOGY

- Test Framework that produces synthetic waveform like data.
- SZ3 libraries [2] to test compression and decompression of data with different parameters
- Use of statistical tests like Kolmogorov Smirnov (KS) Test to quantify data fidelity
- Faciliate the storage of compressed data with compression parameters in ROOT format (a widely used I/O subsystem in the HEP community) [3].
- Test with actual data of demonstrator experiment for the DUNE (called Proto-DUNE) [4]



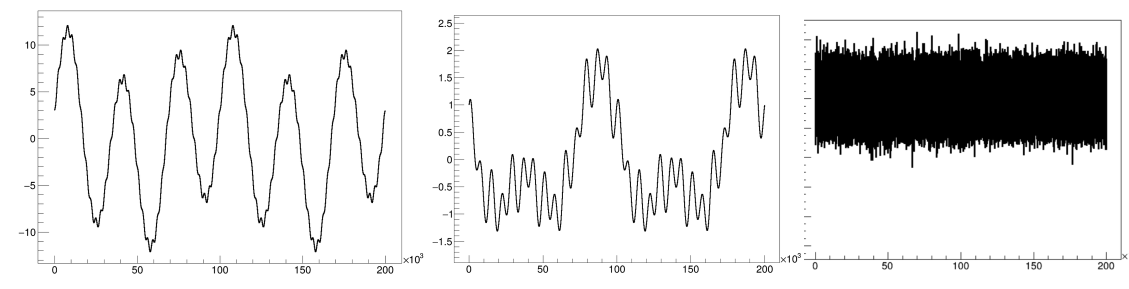
Different components of SZ3 compression algorithms. Parameters in blue boxes are control parameters that users can change. This study did not use lossless compression on SZ3 compressed data.

Important Metrics

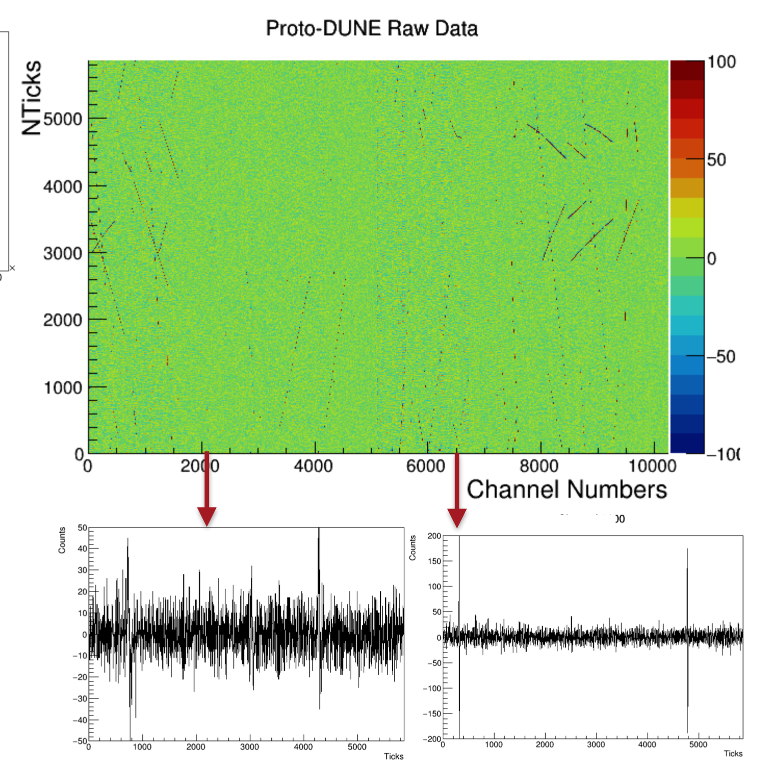
Prediction Metrics: Lorenzo, Cubic and Linear Interpolation and regression are SZ3 metrics that were explored for this work.

Error Bound (EB): Error bound quantifies the deviation of original data from decompressed one. Relation between data shape, predictors and EB values were explored.

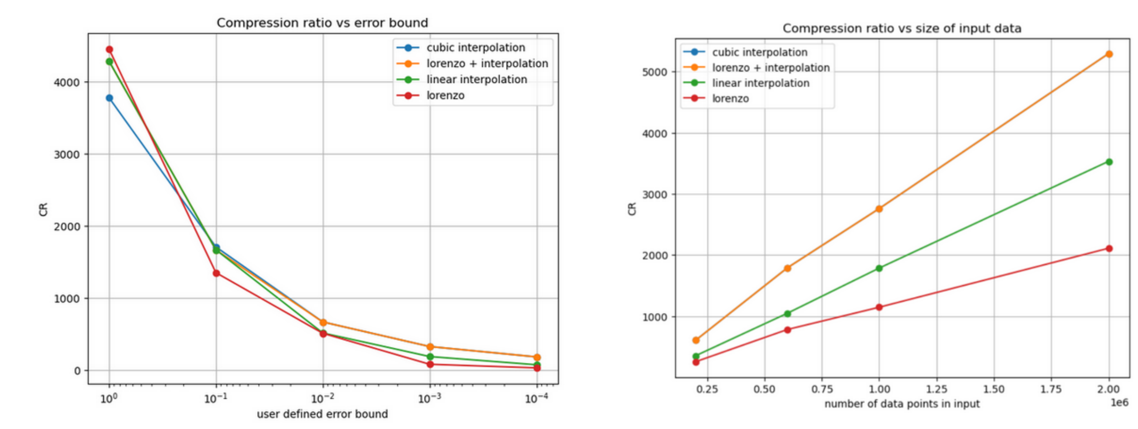
COMPRESSION OF DATA WITH DIFFERENT SHAPES



Some of the sample synthetic data that were used for the test.
Compression tests were done with synthetic data of different shapes and lengths.



Proto-DUNE raw data used in the compression test. X axis is the channel numbers and Y axis is the readout time. Color scale shows the amplitude of readout signal. Data (Amplitude) is *int* type. We show two 1-D projected channel readouts from the raw data with (channel 6500, right) and without (channel 2100, left) signal peak.

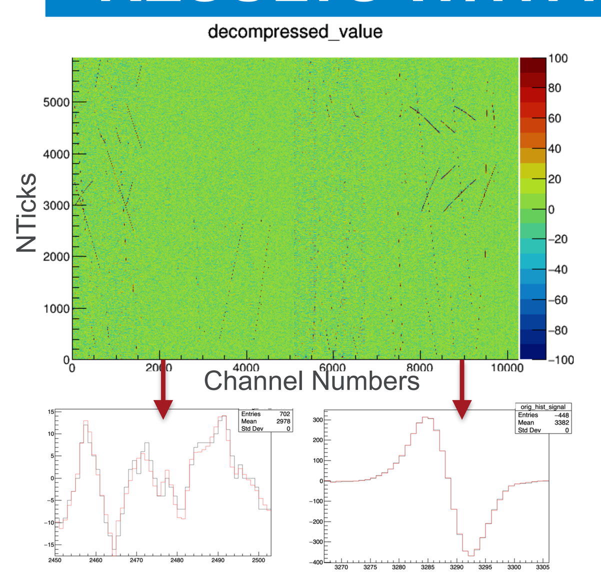


Left : Compression Ratio (CR) with different EB values using different prediction parameters. Higher data fidelity (lower EB values) results in smaller CR values for all predictor types.
Right : CR values of data with different length with EB values kept constant using different predictors. Compression of larger data size yields higher CR for waveform like data.

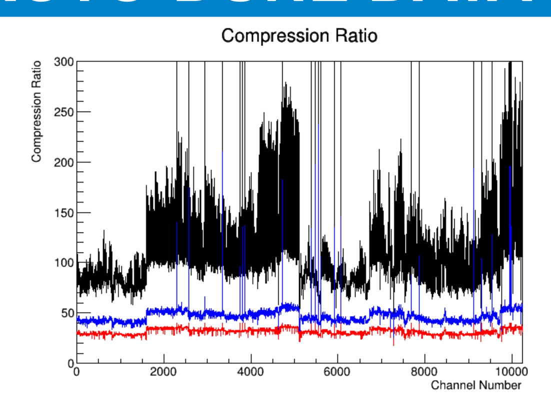
$$CR = \frac{sizeof(intput\ data\ type) \times input\ data\ size}{sizeof(char) \times compressed\ data\ size}$$

SZ3 compressed data is of *char* type

RESULTS WITH PROTO-DUNE DATA



Top : Output of SZ3 compressed data using Lorenzo predictor with EB = 4. **Compressed data is 32 times smaller than original data.**
Bottom : Original and decompressed readout for channel 2100 (left) and 9000 (right) respecitively)



Compression ratio as a function of channel numbers from the Proto-DUNE raw data using Lorenzo Predictor. **Red**, **Blue** and **Black** colors represent the compression ratio distribution for EB = **4**, **8** and **20** respectively.

Error Bounds	4	8	20
Lorenzo	32	46	118
Interpolation	43	87	64

Average Compression values for different EB values for Lorenzo and Linear Interpolation

RESULTS

- Waveform like data can be compressed with lossy compression algorithms like SZ3 with user defined fidelity and achieving significant storage savings.
- Raw data with larger length could yield larger compression ratio with proper choice of predictors improving the CR.

NEXT STEPS

- Apply DUNE reconstruction algorithms on raw and decompressed data will provide a more quantitative data fidelity for different EB values.
- Extend tests with other intelligent lossy compression algorithms like MGARD, IDEAL-EM.

REFERENCES

- B. Abi, R. Acciarri, and M.A. Acero. Volume i. introduction to dune. Journal of Instrumentation, 15(08):T08008, August 2020
- Xin Liang, Kai Zhao, Sheng Di, and et al. SZ3: A Modular Framework for Composing Prediction-Based Error-Bounded Lossy Compressors. arXiv e-prints, page arXiv:2111.02925, November 2021.
- ROOT: https://root.cern/
- A. Abed Abud et al. Design, construction and operation of the ProtoDUNE-SP Liquid Argon TPC. JINST, 17(01):P01005, 2022.
- MGARD. GitHub. https://github.com/CODARcode/MGARD
- IDEALEM. Lawrence Berkeley National Laboratory. https://sdm.lbl.gov/idealem/

This work was funded by Argonne National Laboratory LDRD seed project (Project ID = 2024 - 0387)