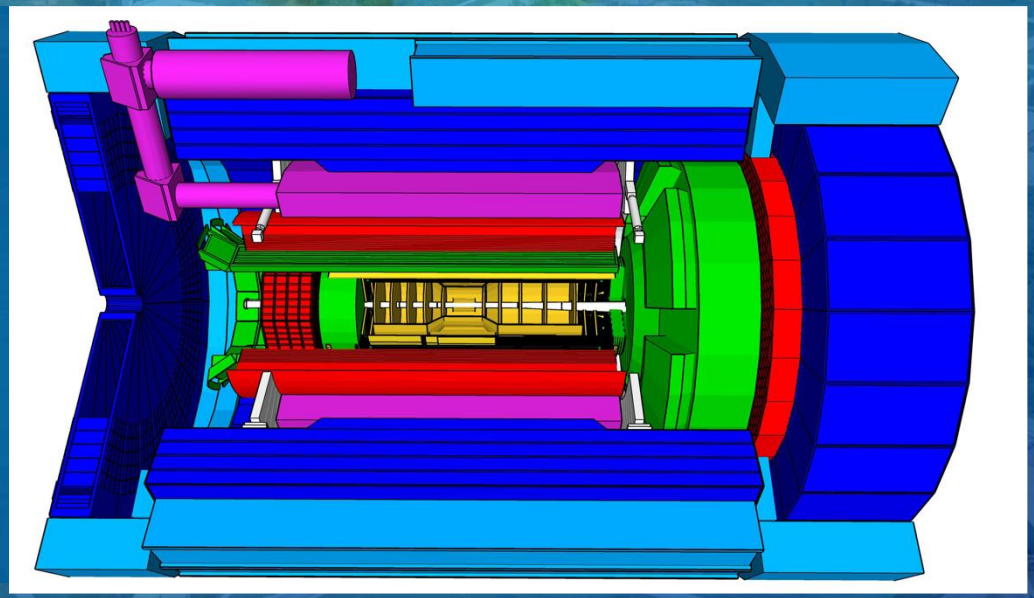
JUN 16, 2026

TOWARDS AI-BASED RECONSTRUCTION -- AI-READY DATA



Zhiwan Xu
Argonne National Lab

NP AMSC DATA PROVIDERS PROGRAM (DAPP)

Design principle: 10.5281/zenodo.19710980
BIC application + metadata: arxiv:2606.07667

AI-READY DATA DESIGN

Design principle

- Physical measurement (x, y, z, A, t) is naturally a "point-cloud" data structure: sparse, heterogeneous, uniquely distinctive from industrial LLM training data.
- 4 Design principles: unified representation, explicit uncertainty, metadata-defined semantics, minimal scope
- Available in Zenodo: 10.5281/zenodo.19710980.

➢ Why not ROOT/HDF5? → Nested tree structures require iterative lookup; no clear separation between trainable features and truth labels

➢ Why not existing reconstruction formats? → Designed for physicist-driven analysis, not AI training; tightly coupled to detector-specific software, hard to generalize across experiments.

AI-READY DATA DESIGN

Measurement and Label array for supervised learning

TABLE I. Structure of the `measurements` array.

Field	Type	Description
`event`	`uint64`	Event identifier.
`detector`	`uint16`	Subdetector identifier.
`hit`	`uint16`	Hit index within a subdetector.
`sample`	`uint16`	Sample index within a hit.
`x, y, z`	`float32`×3	Spatial coordinates associated with the sample.
`pos_cov_xx, xy, xz, yy, yz, zz`	`float32`×6	Upper triangle of the position covariance matrix.
`amplitude`	`float32`	Measured signal amplitude.
`amplitude_uncertainty`	`float32`	Uncertainty on the amplitude.
`time`	`float32`	Time associated with the measurement.
`time_uncertainty`	`float32`	Uncertainty on the time measurement.

TABLE II. Structure of the `labels` array, with contributor-specific fields repeated for each retained contributor $n \in \{0, \ldots, N-1\}$.

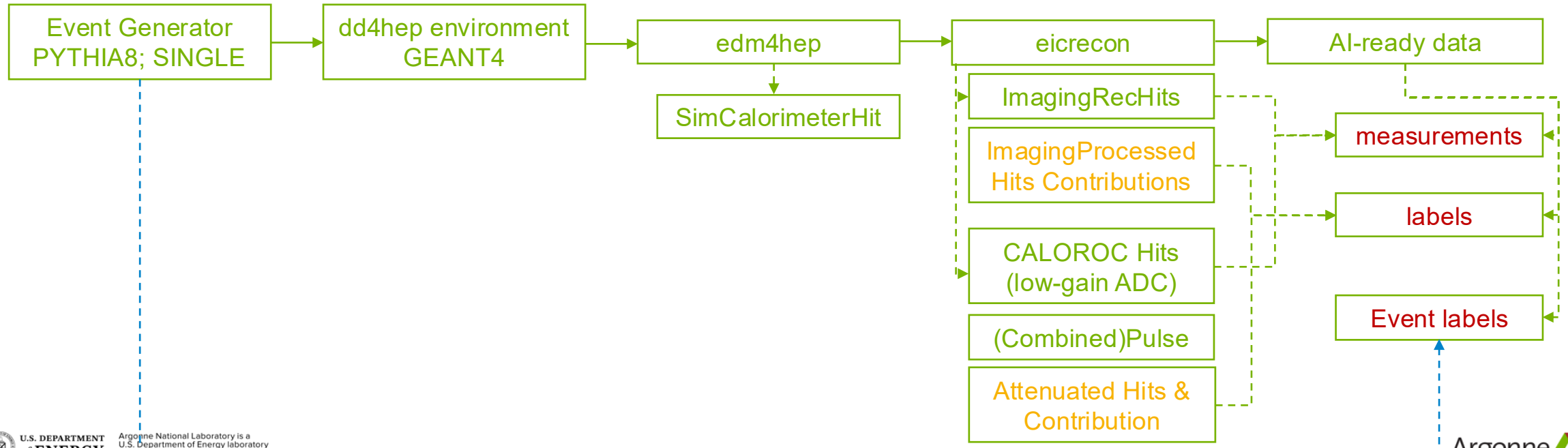
Field	Type	Description
`event`	`uint64`	Event identifier.
`detector`	`uint16`	Subdetector identifier.
`hit`	`uint16`	Hit index within a subdetector.
`sample`	`uint16`	Sample index within a hit.
`x, y, z`	`float32`×3	Truth spatial coordinates.
`deposit_energy`	`float32`	Truth total deposited energy.
`time`	`float32`	Truth time associated with the measurement.
`n_contributions`	`uint16`	Total number of contributing particles.
`con{n}_particle`	`uint32`	Index of contributor n in the event-level record.
`con{n}_x, y, z`	`float32`×3	Interaction spatial coordinate of contributor n.
`con{n}_deposit_energy`	`float32`	Deposited energy from contributor n.
`con{n}_time`	`float32`	Time of interaction of contributor n.
`con{n}_px, py, pz`	`float32`×3	Momentum vector of contributor n at production.
`con{n}_energy`	`float32`	Energy of contributor n at production.
`con{n}_pid`	`int32`	PDG identifier of contributor n.
`con{n}_parent_pid`	`int32`	PDG identifier of the immediate parent of contributor n.
`con{n}_parent_energy`	`float32`	Energy of the immediate parent of contributor n.

- Suitable for a variety of detector readouts, including tracking, calorimeter, spectator detectors. The uncertainty covariance matrix is key information for AI to learn detector resolution.
- Label array design principles: granularity, contributors, spatiotemporal, ancestry, minimality.

BIC CASE APPLICATION

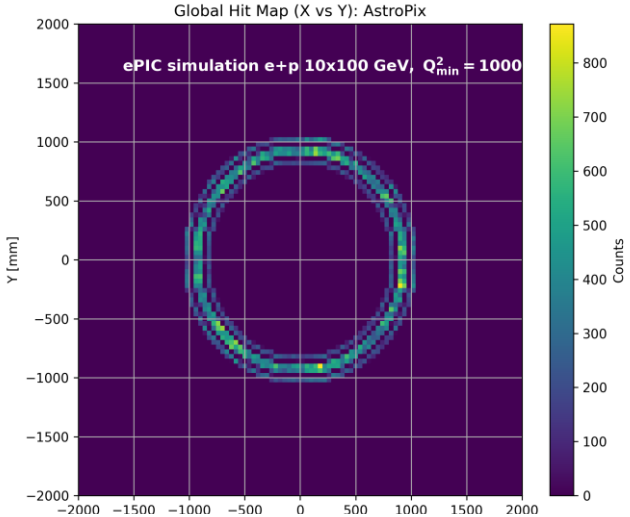
DAPP Argonne led detector data thrust

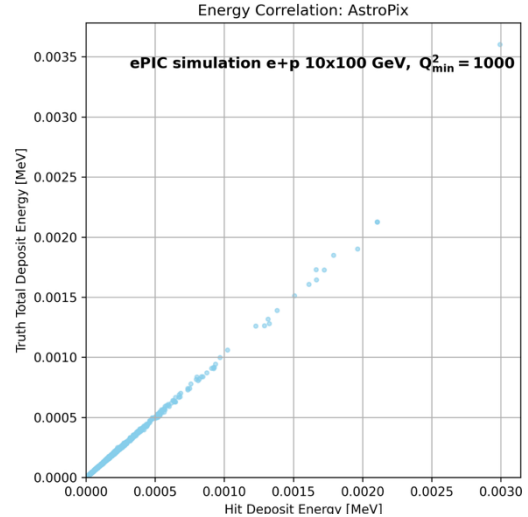
- Advantage: Hybrid dual readout: AstroPix silicon pixels (500 μm, single-sample) for spatial imaging + Pb/ScFi waveform per SiPM (25 ns multi-sample) for energy sampling.
- 3 arrays: measurements (readout), labels (per-hit truth + top-3 contributors), event_label (MC particles + decay chain). arxiv:2606.07667

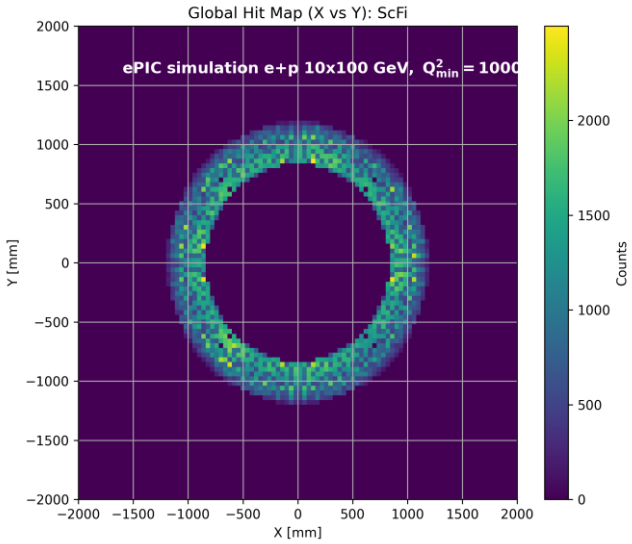


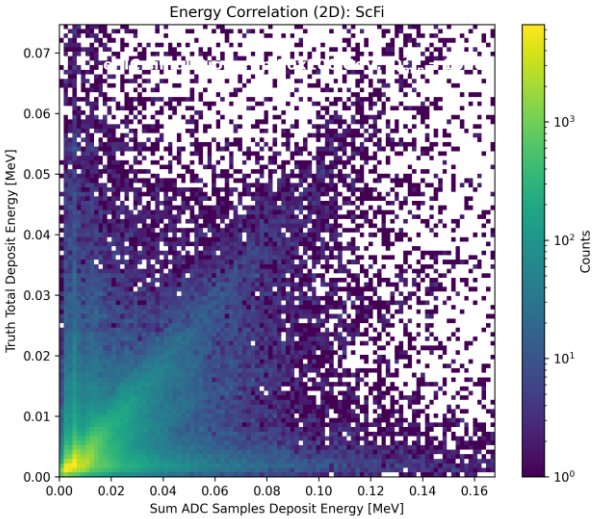
AI-READY DATA VISULIZATION

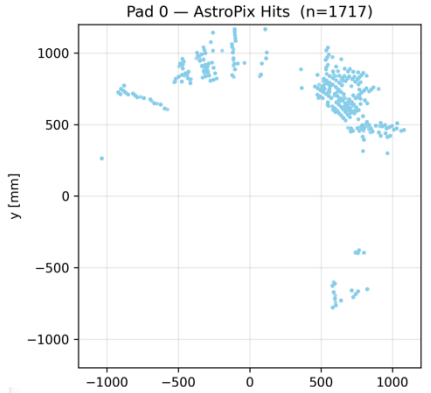
DIS data 10x100 GeV, min Q2=1000, npz format, 1k events ~ 5M

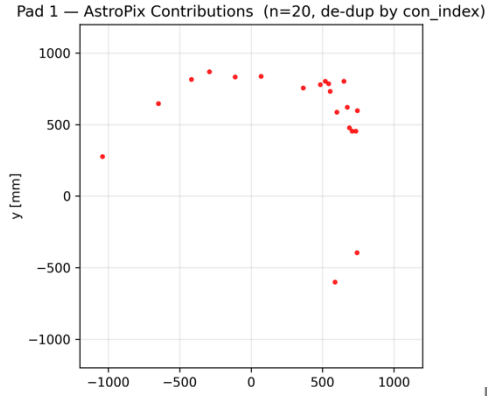


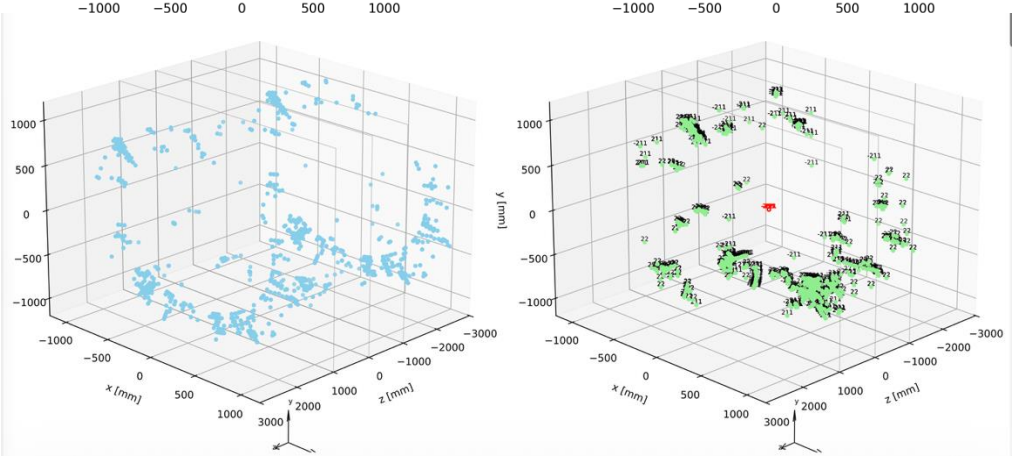










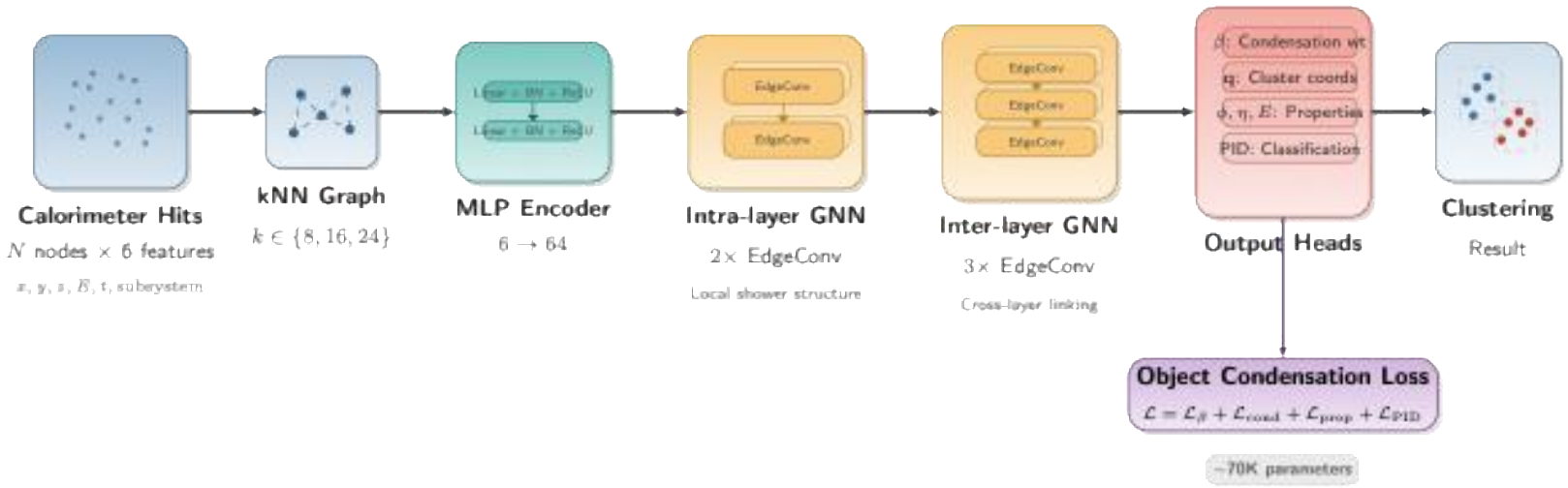


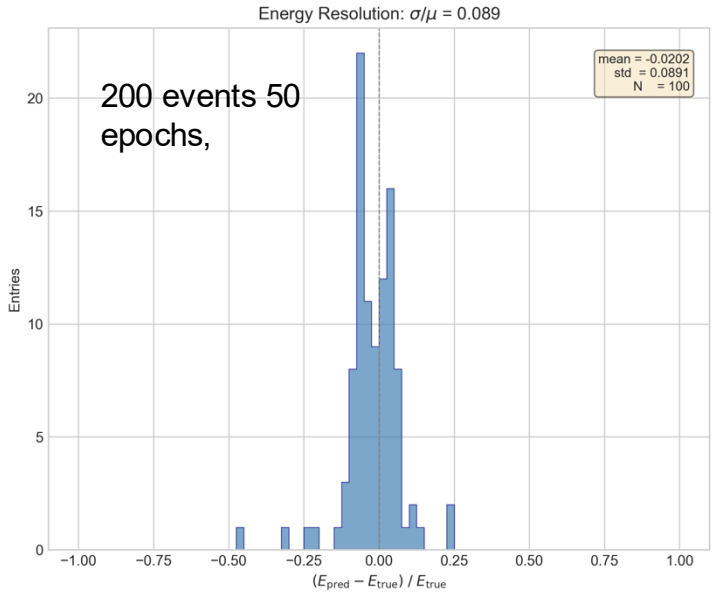
Data – Label visualization. Well demonstrated!

FIRST GLIMPSE OF AI-BASED RECONSTRUCTION

Build AI models based on the AI-ready BIC data

- AI applications: reconstruction (clustering, particle ID, energy regression), anomaly detection, fast simulation, jet tagging, and physics inference from raw detector data.
- (Ongoing) AI demonstration with hierarchical GNN with Object Condensation for shower cluster reconstruction in the BIC. Toy training on 10 GeV γ simulation: E_res~9%





Thank you!


Argonne
NATIONAL LABORATORY
U.S. DEPARTMENT
of ENERGY