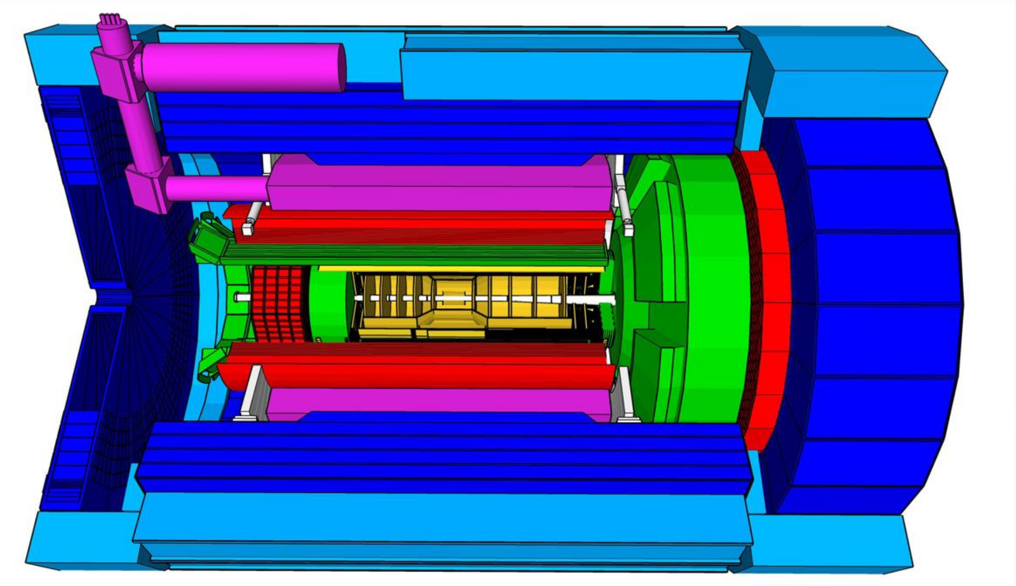
JULY 16, 2026

July 2026 Joint ePIC/EICUG Meeting

AI-READY DATA FOR BIC



Zhiwan Xu, Sylvester Joosten, Minho Kim, Chun Yuen Tsang and Maria Żurek
zhiwan.xu@anl.gov
Argonne National Lab

NP AMSC DATA PROVIDERS PROGRAM (DAPP)

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

AI FOR NUCLEAR PHYSICS

The transformative power of AI for science

Traditional Scientific AI

- Classify objects/pid/events
- Reconstruct particle tracks
- Detect anomalies
- Optimize simulations

→ One model = One task

Foundation Models for Physics

- Learn diverse exp./sim. data
- Support multiple scientific tasks
- Reason across modalities
- Agentic scientific workflows
- Improve as new data arrive

→ One evolving model = Multi tasks

- **The limiting factor is not just the model**
- **It is the quality, scale, and accessibility of the training data**

AI-READY DATA DESIGN

Data design principle

- Physics detector measurement (x, y, z, A, t) is naturally a point-cloud data structure.
- We aim to release low-level pre-reconstruction detector data, rather than reconstructed objects such as clusters, to provide the closest equivalent to the raw detector measurements for AI training.

➢ Why not ROOT/existing storage format? → iterative lookup; no separation for features or labels

➢ Why not EMP4hep/existing reconstruction formats? → tightly coupled to detector-specific software, hard to generalize across experiments. Designed for physicist-driven analysis, not AI training.

- 4 Design principles: unified representation, explicit uncertainty, metadata-defined semantics, minimal scope. Document ready: 10.5281/zenodo.19710980

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.

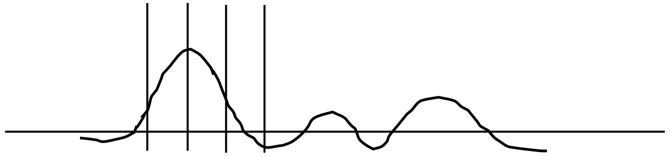


TABLE II. Structure of the labels array, with contributor-specific fields repeated for each retained contributor $n \in \{0, \ldots, N-1\}$.

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 AI-READY DATASET

Barrel Imaging Electromagnetic Calorimter

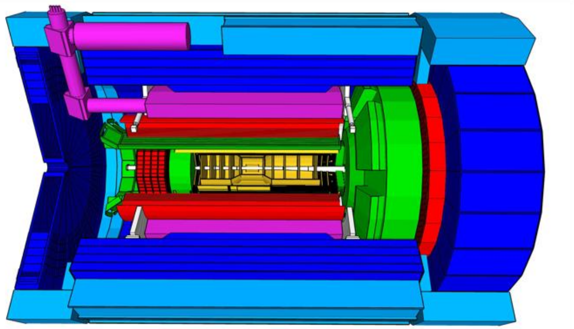
- Advantage of BIC: Hybrid dual readout: AstroPix silicon pixels (500 μm, single-sample) for spatial imaging + Pb/ScFi waveform per SiPM (25 ns multi-sample) for energy sampling.
- 3 arrays: measurements (readout), labels (per-hit truth + top-3 contributors), event_label (MC particles + decay chain). arxiv:2606.07667
- 6M labeled DIS data (10x100 GeV, Q2min = 1, 10, 100, 1000), and 6.5 M single particle data (16 pids).

Event Generator PYTHIA8; SINGLE → dd4hep environment GEANT4 → edm4hep → eicrecon → AI-ready data

edm4hep → SimCalorimeterHit

eicrecon → ImagingRecHits, ImagingProcessed Hits Contributions, CALOROC Hits (low/highgain ADC), (Combined)Pulse, Attenuated Hits & Contribution

AI-ready data → measurements, labels, Event labels



BIC AI-READY DATASET

Metadata for the ‘hit sample’

- Based on low-level detector simulation of BIC and are pre-reconstruction.

TABLE III. Structure of the `measurements` array for the BIC.

Field	Type	Description
`event`	`uint64`	Event index defined per simulation collision.
`subsystem`	`uint16`	0 = AstroPix, 1 = Pb/ScFi.
`hit`	`uint16`	A single channel response from AstroPix or Pb/S
`sample`	`uint16`	Usually 0 for AstroPix. 0–N for Pb/ScFi wavefor
`x, y, z`	`float32`×3	Position of the AstroPix hit or Pb/ScFi SiPM rea
`deposit_energy`	`float32`	Deposited energy (GeV) converted from ToT or w
`time`	`float32`	Readout time at the measurement (ns).
`x_unc, y_unc, z_unc`	`float32`×3	Diagonal position uncertainties (mm).
`deposit_energy_unc`	`float32`	Deposited energy uncertainty from calibrated reso
`time_unc`	`float32`	Uncertainty on the measurement time from DAQ(

- Uncertainties defined accordingly:

$$\sigma_x = \sqrt{\sigma_r^2 \cos^2\varphi_{\rm sec} + \sigma_\varphi^2 \sin^2\varphi_{\rm sec}}, \qquad \sigma_y = \sqrt{\sigma_r^2 \sin^2\varphi_{\rm sec} + \sigma_\varphi^2 \cos^2\varphi_{\rm sec}}.$$

$$\sigma_E = \begin{cases} 0.0457\, E_{\rm dep}, & E_{\rm dep} > 1.2210^{-4}\ \text{GeV} \\ 5.8910^{-8}\, E_{\rm dep}^{-0.499}, & E_{\rm dep} \le 1.2210^{-4}\ \text{GeV} \end{cases}$$

Metadata

- **x, y, z:** `float32` x 3, spatial scalars, unit: mm. Cartesian coordinates of the readout unit in the laboratory frame, with the z-axis along the proton beam direction. For AstroPix, this is the center of the fired pixel within the silicon bulk. For Pb/ScFi, this is the position of the SiPM readout unit at the longitudinal fiber end, not the physical location of the particle interaction along the fiber. The SiPM readout x,y location is defined at the center of the grid cell, and its z position is at one of the two ScFi ends: z = - 2587.5 mm or 1762.5 mm. The true interaction z(hit) must be reconstructed from the time difference between the two fiber ends (t(+z) and t(-z)), as z(hit) = (t(+z) - t(-z))·veff /2 , based on the effective velocity of veff=160 mm/ns. Coordinate convention: the nominal interaction point is at the origin (0, 0, 0).
- **deposit_energy:** `float32` scalar, unit: GeV. The measured signal amplitude converted to deposited energy. For AstroPix, it is derived from the charge collected by the pixel and corresponds to calibrated ToT (time over threshold). For Pb/ScFi, it is energy attenuated along the fiber to the ends of the detector, which is converted independently for each 25 ns waveform sample from either the 14-bit ADC value (dynamic range 0.1 MeV to 10 GeV in true energy) or the electric photon count (1700 e.p. is 0.1 GeV in deposited energy not corrected for sampling fraction). The energy spans approximately five orders of magnitude (˜1e-5 to ˜10 GeV); logarithmic scaling is recommended before training. The truth counterpart is `deposit_energy` in `labels`.

BIC AI-READY DATASET

Metadata for the ‘label’

- Ready to use label, processed matched with measurement hit.

TABLE IV. Structure of the labels array for the BIC (contributor fields repeated for n

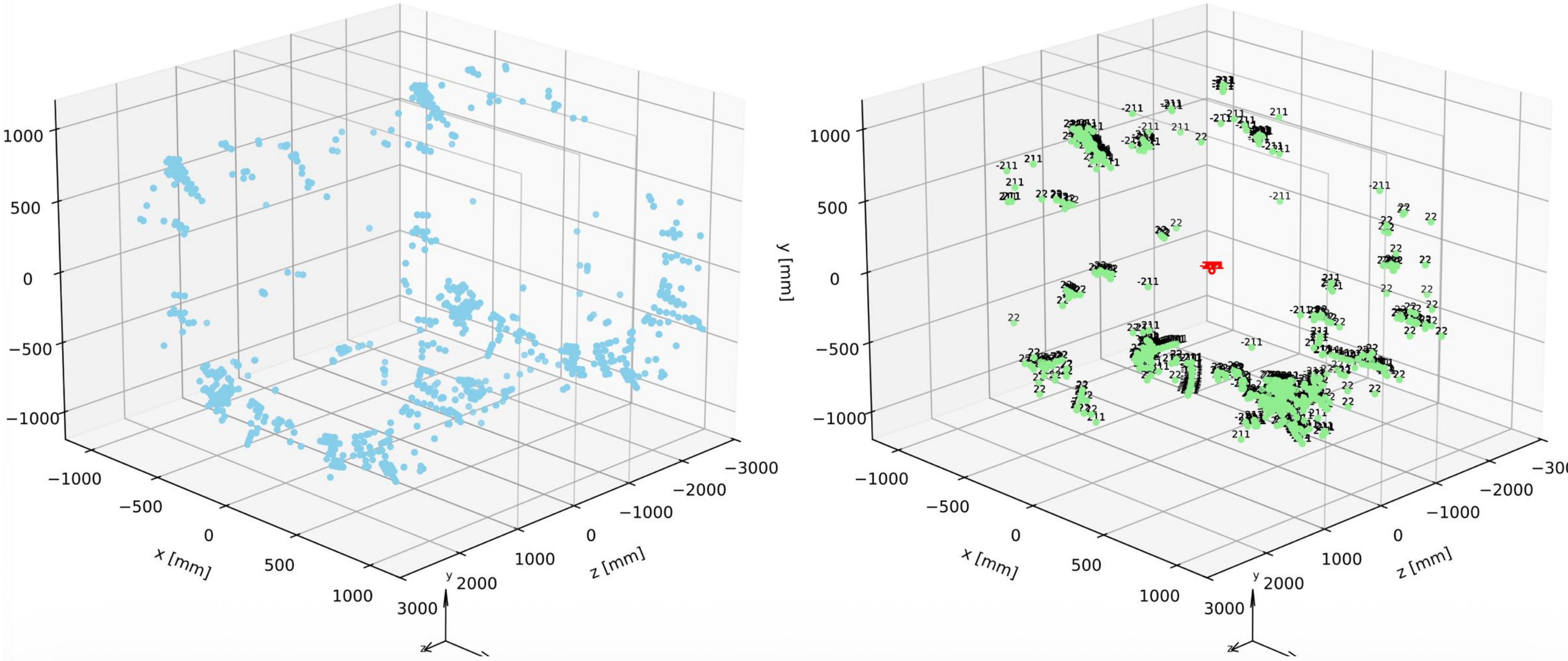
Field	Type	Description
event	uint64	Event index.
subsystem	uint16	Subsystem identifier.
hit	uint16	Hit index.
sample	uint16	Sample index.
x, y, z	float32×3	Energy-weighted truth position (mm).
deposit_energy	float32	Total truth energy deposit (GeV).
time	float32	Earliest truth interaction time (ns).
n_contributions	uint16	Total contributing particles.
con{n}_particle	uint16	Index into event-level truth record.
con{n}_x, y, z	float32×3	Truth interaction position (mm).
con{n}_deposit_energy	float32	Deposited energy from contributor n (G
con{n}_time	float32	Truth interaction time (ns).
con{n}_px, py, pz	float32×3	Momentum at production (GeV/c).
con{n}_energy	float32	Energy at production (GeV).
con{n}_pid	int32	PDG ID.
con{n}_parent_pid	int32	PDG ID of parent.
con{n}_parent_energy	float32	Parent energy (GeV).

Metadata

- **deposit_energy:** `float32` scalar, unit: GeV. Total truth energy deposited in this measurement, summed over all contributing particles. This is the truth counterpart to the `deposit_energy` column in `measurements`, and the sum of all `con{n}_deposit_energy` values equals this field up to rounding from contributions beyond MAX_CONTRIB.
- **con{n}_deposit_energy:** `float32` scalar, unit: GeV. Energy deposited by the n-th contributing particle in a single pixel or a single fiber. For ScFi only: summed over all shower particles by energy-weighted within the fiber. The sum over all valid contributors of this value would approximate the aggregate `deposit_energy` field in this array. Contributors are sorted by this field in descending order. Sentinel: 0 for empty slots.
- **con{n}_energy:** `float32` scalar, unit: GeV. Total energy of the n-th contributing particle at its production vertex. Together with `con{n}_px, py, pz`, this completes the particle four-vector. As with momentum, the value corresponds to the traced incident particle after primary particle fetching. Sentinel: 0 for empty slots.

AI-READY DATA VISULIZATION

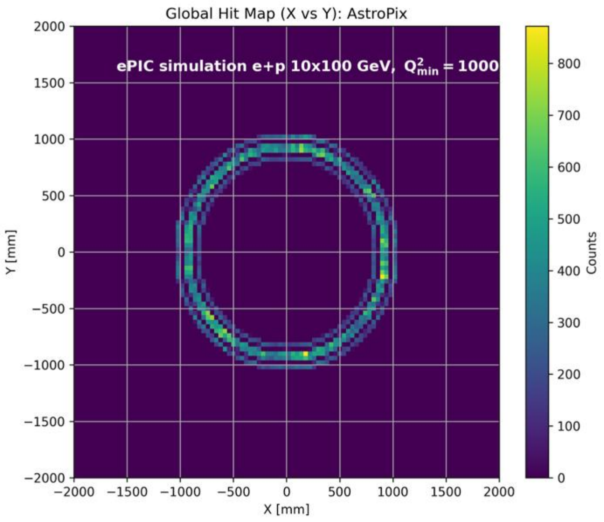
DIS data 10x100 GeV, min Q2=1000, npz format, 1k events ~ 5M

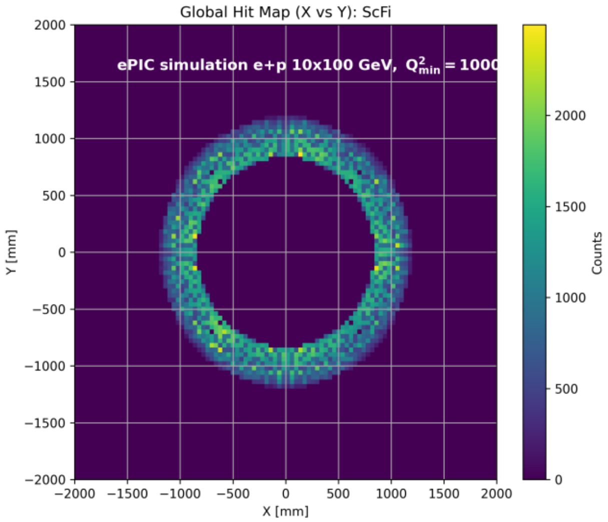


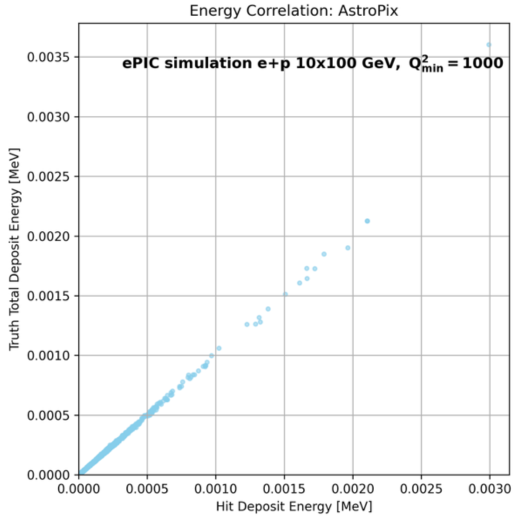
Data – Label visualization. Well demonstrated!

AI-READY DATA VISULIZATION

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

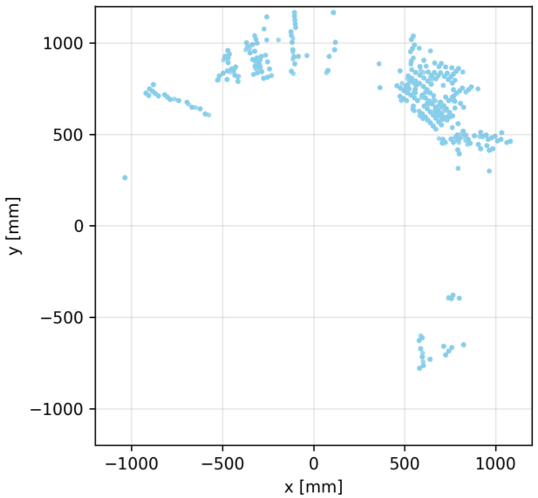


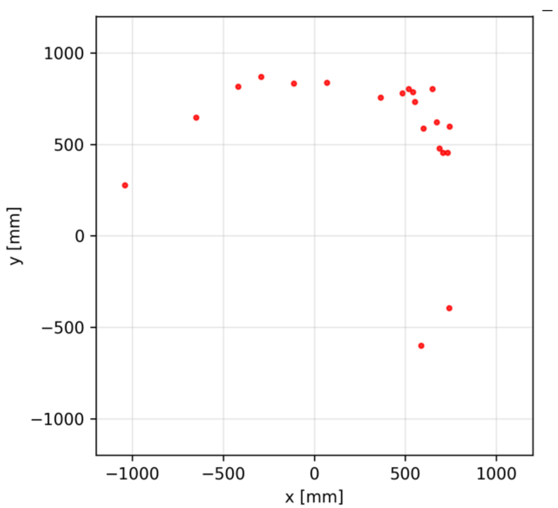




Clear distinctions between AstroPix (sub sys 0) and Pb/ScFi (sub sys 1)

Data – Label visualization. Well demonstrated!



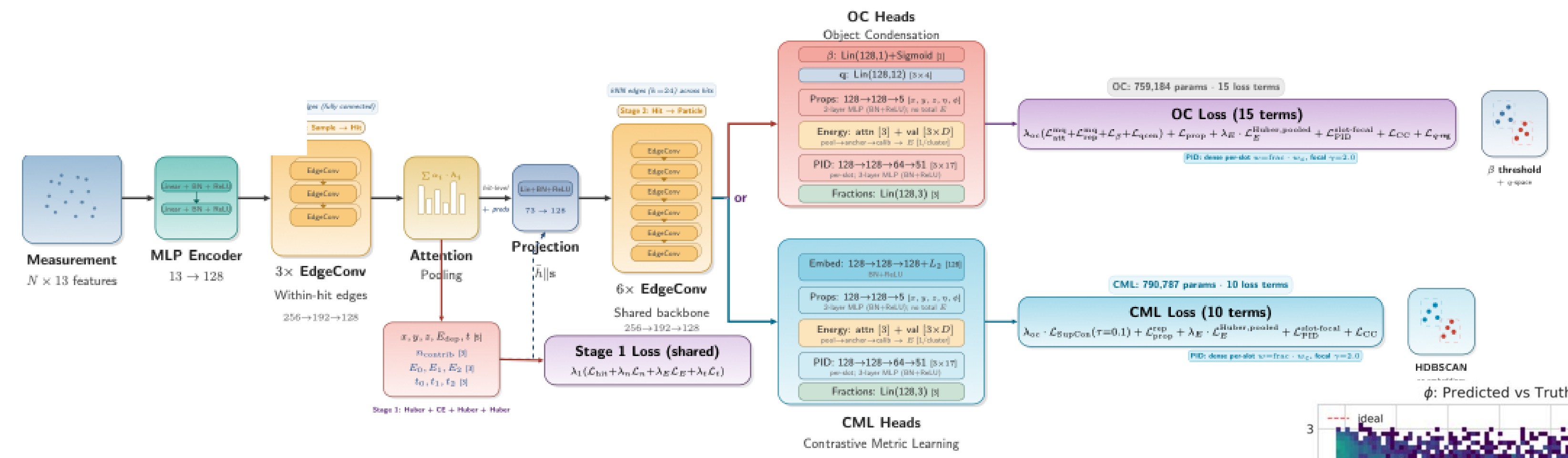


GNN AI-BASED RECONSTRUCTION

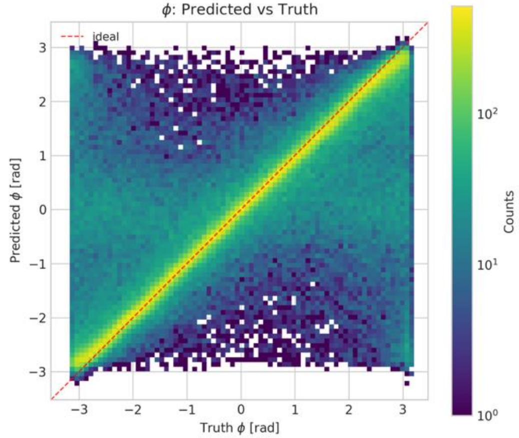
Validation of the AI-ready BIC data

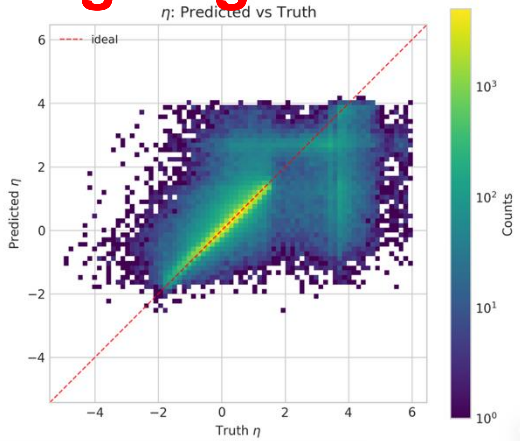
- GNN with Object Condensation vs Contrastive Metric Learning for shower clustering and reconstruction in the BIC.

TwoStageGNN v7.2 Architecture



Still Ongoing!





SUMMARY

- We have established an AI-ready low-level detector dataset for the BIC, preserving raw detector information through realistic digitization and waveform readout simulation, across both physics and single particle events.
- Building on this foundation, we are developing end-to-end GNN-based AI models for clustering and reconstruction.

Thank you.


Argonne
NATIONAL LABORATORY
U.S. DEPARTMENT of ENERGY