GNN + Attention-Based Cluster Finding in Live Experiment in HallC, JLab

Joint ePIC/EICUG Meeting (AI Town Hall), July 16, 2026

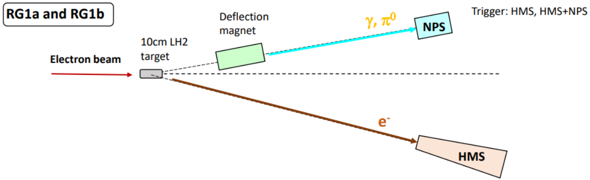
THE CATHOLIC UNIVERSITY OF AMERICA

Jefferson Lab

Chi-Kin Tam

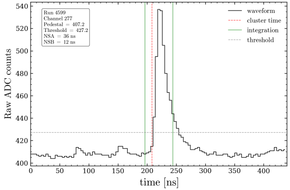
tamc@cua.edu / ckin@jlab.org

- high precision measurements of cross sections of γ, π^0
- `HMS - NPS` coincidence, $\theta \in [5.5°, 60°]$
- 30 columns and 36 rows of $PbWO_4$ crystals, each is $2 \times 2 \times 20$ cm^3



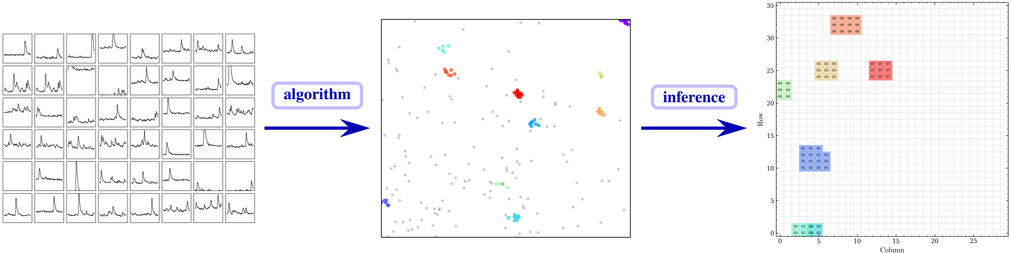
<u>Loose trigger mode for pseudo-streaming</u> :

- each readout with PMT, connected to Jlab **FADC250** module
- generate pulse integral and rise time
- VXS Trigger Processor (VTP) module triggers waveform readout



AI/ML Algorithm - Task Definition

- **Input** : Point Cloud of signals with the Waveform as features.
  - ϟ pedestal provided to the model, either from VME config or calibration.
- **Truth** : Cluster IDs, either from `Geant4` simulation, `VTP` FPGA logic or `HCana` reconstruction.
- **Goal** : Assign cluster memberships to each block.

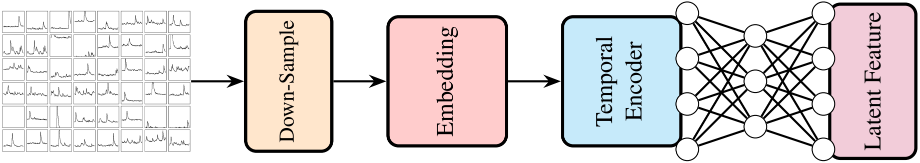


Input data waveform + geometry

Latent space

Prediction

Architecture : Encoding Waveform



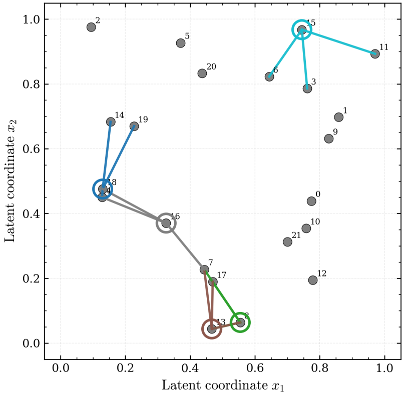
- Unlike in the literature where high-level features (e.g., peak time, pulse integral) are used, we directly process raw waveforms.
- Input **pedestal-subtracted waveform**, dimension [1080 channels][110 samples]
- Down-Sampling the background waveform to reduce memory overhead.
  - For each WF associated with a cluster, keep 3 – 5 WFs associated with noise.
- Embedding waveform sample into higher dimension.
- Attention-based or Long-Short Term Memory (LSTM) to compress the temporal information of the waveform.

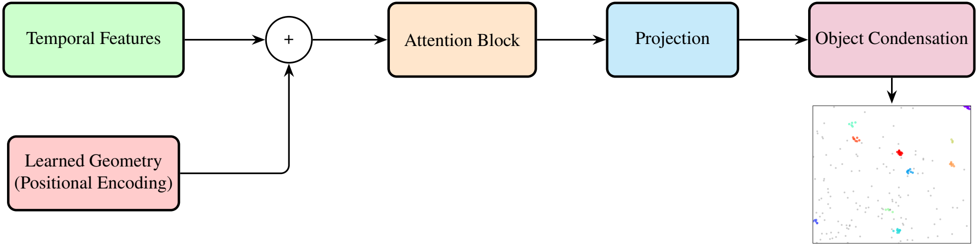
Architecture : GravNet

- takes the NPS geometry, dimension [1080 channels] [2 dim]
- project to
  - coordinate in latent space
  - feature space
- build kNN graph → aggregate features from neighbors scaled by **Grav**itational potential !

$$\overbrace{\mathbf{x}_i^{(k)}}^{\text{node } i} = \gamma^{(k)}\left(\mathbf{x}_i^{(k-1)}, \bigoplus_{j} \phi^{(k)}\left(\mathbf{x}_i^{(k-1)}, \mathbf{x}_j^{(k-1)}\right)\right)$$

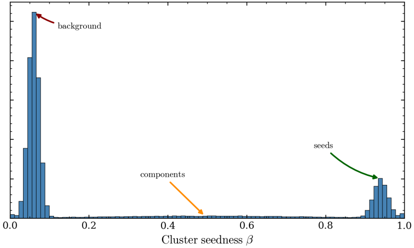
- used as **Positional encoding** in downstream attention





Model Outputs : Object Condensation Head

- Input : learned geometry from GravNet + latent WF features
- Attention mechanism to correlate different blocks
- Compress into object representation
  - latent space coordinate $x \in \mathbb{R}^2$
  - Cluster seedness score $\beta \in (0.0, 1.0)$



Loss function

Attractive loss

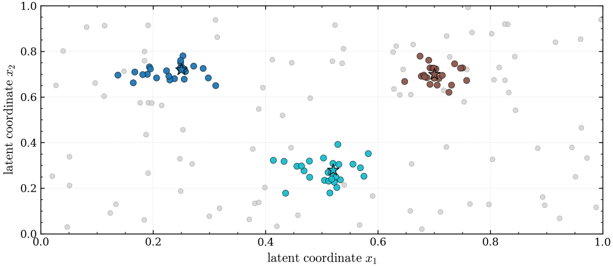
$$V_{\text{attr}}(x) = ||x - x_{\text{seed}}||^2 q_{\text{seed},k}$$

- Signals associated to the same cluster are pulled towards their cluster seed in latent space.
- Minimize latent x distance for points (k) in each cluster.

Repulsive loss

$$V_{\text{repul}}(x) = \max(0, \lambda - ||x - x_{\beta}||) q_{\beta k}$$

- Signals from different objects are pushed away.
- Minimize overlap so different clusters are well-separated.



Coward Loss

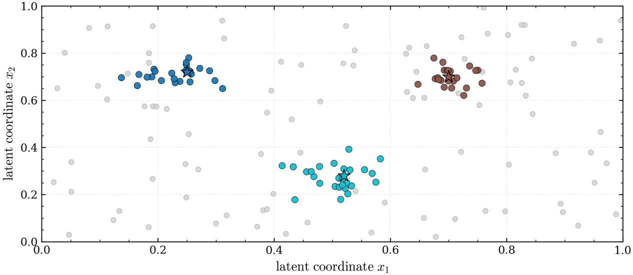
$$V_{\text{coward}} = \frac{1}{K}\sum_i (1-\beta_i)$$

- apply to true cluster seeds only
- encourage cluster seeds to have high $\beta \approx 1$

Noise loss

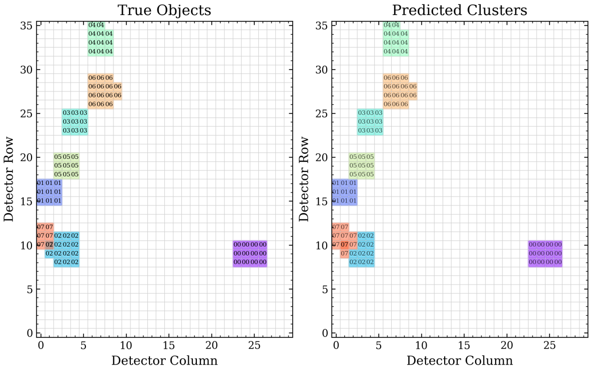
$$V_{\text{noise}} = \underbrace{s_B}_{\text{s/n ratio}} \sum_i n_i \beta_i$$

- apply to background / noise nodes only
- regularize backgrounds to stay $\beta \approx 0$



- Total loss is just $\mathcal{L} = aV_{\text{attr}} + bV_{\text{repul}} + cV_{\text{coward}} + dV_{\text{noise}} \leftarrow$ need to fine tune the weights

Inference Results - Simulation

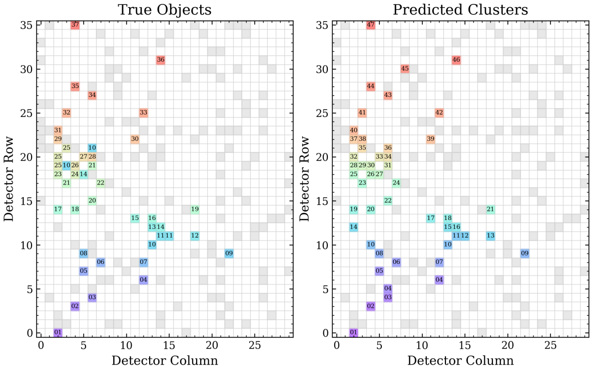


Geant4 simulation:

- primary vertex from `SIMC`
- simulate two γ hits
- overlapping events to increase cluster density
- high-level features as input
  - energy deposited in each channel
  - signal arrival time

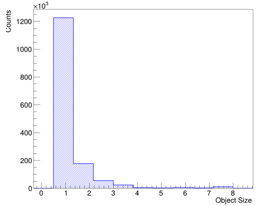
- ✔ Model can cleanly assign cluster ownership to points, effectively separating objects from each other.
- ✖ No background generated simulation.

Inference Results - Hcana (Preliminary)



<u>HCana reconstruction</u>:

- reconstructed calorimeter clusters
- real signal mixed with noise and singleton clusters
- preliminary preprocessing pipeline



- ✔ The model separates real signal from noise and identifies cluster seeds.
- ✖ It still struggles to pull points from the same cluster together consistently.
- Next steps: improve preprocessing and reduce decoder memory overhead.

- Developed models for cluster reconstruction using high-level features and waveform with Attention + GNN architecture.
- Integrated ONNX Runtime inference into JANA2 via ONNX JService.

Using High-level features with background events in EIC simulation

Multi-Scale Background Rejection and Physics Process Discrimination in Electron-Ion Collider Experiments Using Machine Learning

Collaborator Contributions

17 Jul 2026, 15:30

Sandeep Shiraskar (CUA)

More on beam test preparation streaming readout

NPS Prototype beam test with streaming readout and AI/ML-based clustering algorithms in HallC, JLab

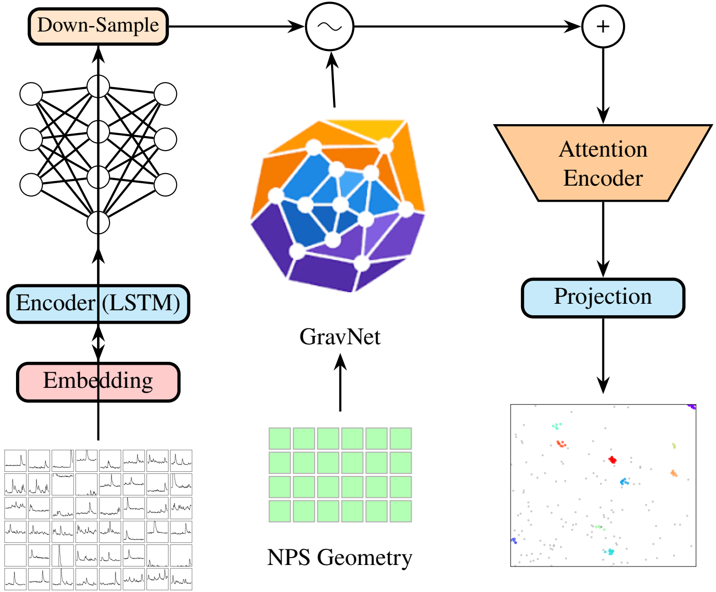
Collaborator Contributions

17 Jul 2026, 15:30

Chi Kin Tam (Catholic University of America)

Collaborations : Tanja Horn, Dmitry Romanov, Sergey Furletov, Benjamin Raydo, Brad Sawatzky, Joshua Crafts, Avnish Singh, Chi-Kin Tam, and other NPS collaborators.

- A declarative implementation of NPS reconstruction in JANA2 : `https://github.com/tck199732/test_jana2_nps.git`
- **Pytorch (geometric) for cluster reconstruction**: `https://github.com/JeffersonLab/nps-sro-ml`
- straight-forward **ONNX Runtime** implementation: `https://github.com/tck199732/ONNX-Runtime-Inference/tree/graph-data`
- Geant4 simulation of NPS detector response with primary vertex from SIMC (by Avnish) : `https://github.com/avnishphy/HallC_SIMC_Geant`


Down-Sample
~
+
Attention
Encoder
Encoder (LSTM)
GravNet
Projection
Embedding
NPS Geometry

Clustering from RAW waveform : VTP

- FPGA-based data processing for triggering
- support 16 FADC250 modules, fiber optics serial links to other crates
- 1Gbps Ethernet connection for configuration, 10/40Gbps Ethernet for CODA ROC readout
- equiped for streaming readout

What VTP module sees :

- Energies and rise times of each pulses, could be multiple pulses per block
- readout window : 1 μs

For each 3x3 grid :

- Filter out pulses within 20 ns window. Triggered if
  - At least 1 hit
  - $E_{\mathrm{seed}} > 150$ MeV
  - E_{seed} is local maximum