Towards Foundation Models for Next-Generation Experiments at the EIC: A Perspective

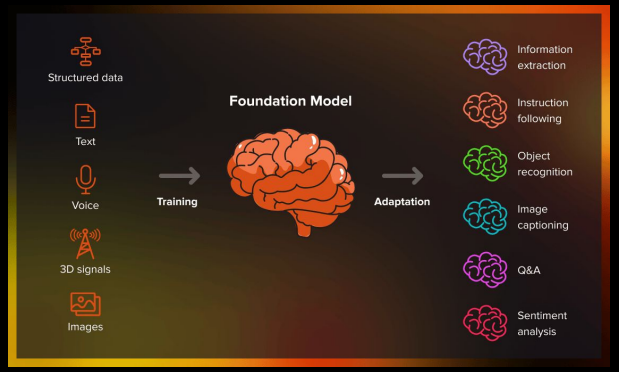
Cristiano Fanelli

2026 RHIC/AGS Annual Users' Meeting & RHIC Science Symposium

William & Mary
Chartered 1693

What are Foundation Models?

- What is a Foundation Model?
  - Large-scale, pre-trained AI models
  - Learn general representations from large-scale datasets
  - Adaptable to various downstream tasks
- Common characteristics:
  - Generalization across related tasks
  - Transfer learning and fine-tuning
  - Scalability to large datasets/workflows
  - Potential multimodal capabilities — Can process heterogeneous data modalities
- Why do they matter for nuclear and particle physics?
  - Address growing computational demands from increasing data volumes
  - Enhance data analysis and reconstruction in NP/HEP experiments
  - Support fast simulation and theoretical modeling
  - Allow stronger theory–experiment integration



FM in HEP/NP: Emerging Paradigms

- Foundation Models are gaining increasing attention in the Physics community, though applications in HEP/NP are still at an early stage:
  - Relatively large scale pre-trained models
  - Capable of supporting multiple downstream tasks, e.g., fast simulation, reconstruction, denoising, and analysis.
- Multiple promising FM paradigms are currently being explored:
  - Transformer-based / GPT style (see [1])
  - Diffusion-based approaches (see [2])
  - State-space model (see [3])

[1] J. Birk, A. Hallin, G. Kasieczka. "OmniJet-α: the first cross-task foundation model for particle physics." *Machine Learning: Science and Technology* 5.3 (2024): 035031

[2] V. Mikuni, B. Nachman. "OmniLearn: A method to simultaneously facilitate all jet physics tasks." *arXiv:2404.16091* (2024).

[3] D. Park, et al. "FM4NPP: A scaling foundation model for nuclear and particle physics." *arXiv:2508.14087 (2025).*

- What follows is primarily based on works [4–6], which leverage Mixture-of-Experts (MoE) transformer-based FM:

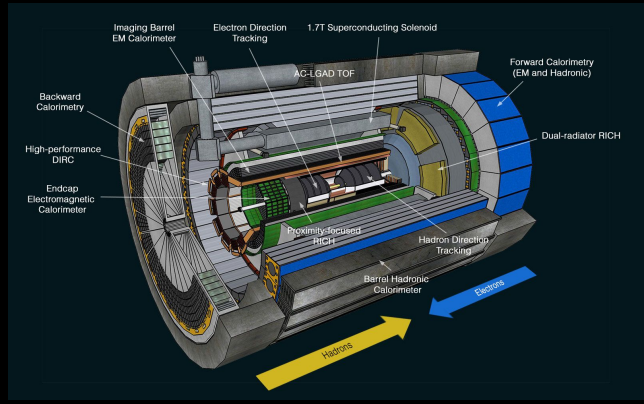
[4] J. Giroux, CF. "Towards foundation models for experimental readout systems combining discrete and continuous data." *Machine Learning: Science and Technology 7.1 (2026): 015031.*

[5] CF, J. Giroux, C. Granger, J. Stevens. "Application of a Mixture of Experts-based Foundation Model to the GlueX DIRC Detector." *arXiv:2604.24775 (2026).*

[6] C. Cardona, CF et al. "Generalizable Foundation Models for Calorimetry via Mixtures-of-Experts and Parameter Efficient Fine Tuning." *arXiv:2603.28804 (2026)*

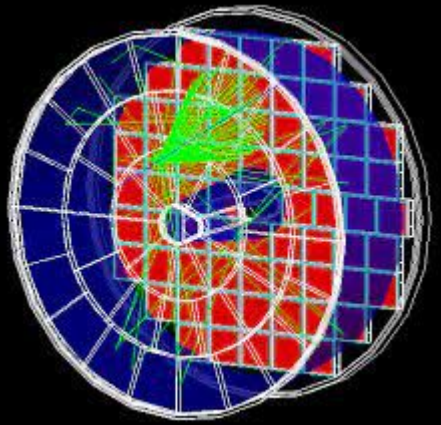
FM for the future EIC: Perspectives

- The EIC naturally aligns with Foundation Models through its large-scale multimodal data, holistic event reconstruction challenges, and AI-driven computing ecosystem.
- In this talk, I will use Cherenkov detectors as a case study.
- I will show how FM can support multiple downstream tasks:
  - Fast simulation
  - Reconstruction / PID
  - Noise filtering

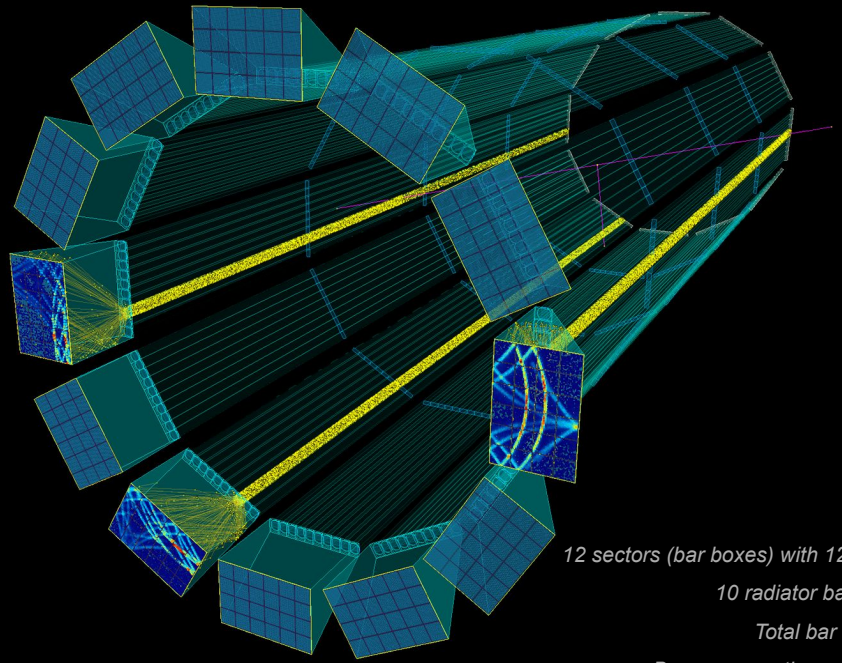


- The presented FM framework can be viewed as a generalization of our previous ML efforts for DIRC detectors:
  - [7] CF, J. Giroux, J. Stevens. "Deep (er) reconstruction of imaging Cherenkov detectors with swin transformers and normalizing flow models." *Machine Learning: Science and Technology 6.1 (2025): 015028*.
  - [8] J. Giroux, M. Martinez, and CF. "Generative models for fast simulation of Cherenkov detectors at the electron–ion collider." *Machine Learning: Science and Technology 6.4 (2025): 040501*.
- I will then discuss how these approaches have the potential to generalize to:
  - Other detector subsystems
  - Multimodal detector integration for holistic reconstruction, modeling and interpretation of physics events

Cherenkov Detectors in epIC/EIC



pfRICH (electron endcap)



hpDIRC (barrel)

12 sectors (bar boxes) with 12 optical boxes

10 radiator bars per bar box

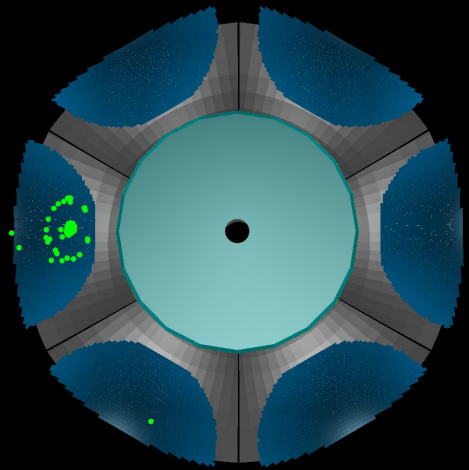
Total bar length ~5.48m

*Bar cross-section ~ 3.5 cm*1.7 cm*

(Baseline design) 6x4 MCP-PMTs units/box

Each PMT is 16x16 pixels

100ps precision on photon arrival time



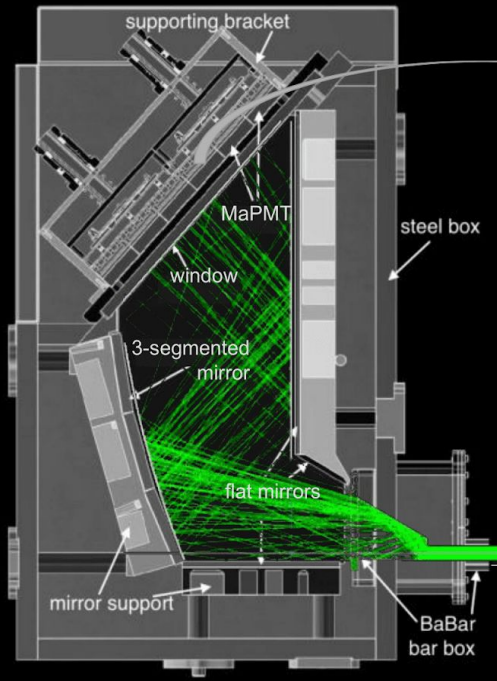
dRICH (hadron endcap)

ePIC

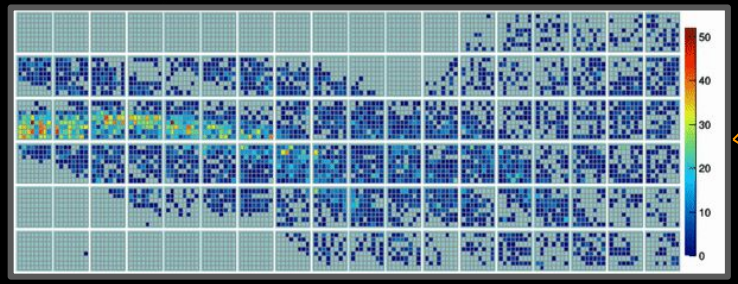
DIRC Detectors

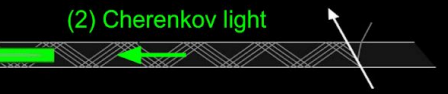
- In this talk I will focus on DIRC detectors. Goal is to do PID from their hit patterns.
- DIRC detectors have complex and sparse space–time hit patterns in the (x, y, t) readout.

48 fused silica bars segmented into 4 bar boxes
Two optical boxes (distilled water and reflective mirrors)
6 x 18 PMT (8 x 8 pixels) array for photon detection.
Provides location and timing information for photons



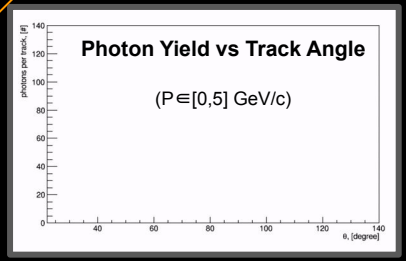
Hit patterns depend on kinematics
Shown below: collection of multiple tracks at given < $|\boldsymbol{p}|$, θ , ϕ >



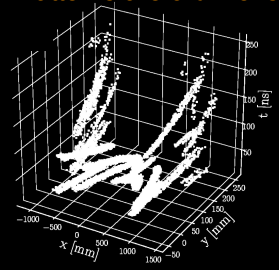


pions/kaons/protons

Patterns are sparse
Photon yield per particle



Patterns are 3-dimensional



Modern Architectures and Advances

1. **High-Fidelity Fast Simulation:**
   Developed generative models capable of producing photon hit distributions with fidelity comparable to Geant4, but at a fraction of the computational cost—critical given the expense of tracking optical photons through complex geometries. (hpDIRC standalone sim)

   J. Giroux, M. Martinez, C. Fanelli "*Generative Models for Fast Simulation of Cherenkov Detectors at the Electron-Ion Collider.*" Machine Learning: Science and Technology 6 (2025): 040501 [link]

2. **Enhanced Particle Identification:**
   Achieved improved PID performance across the full detector phase space, with reduced computational cost compared to traditional reconstruction methods. (GlueX DIRC sim)

   C. Fanelli, J. Giroux, and J. Stevens. "*Deep(er) reconstruction of imaging Cherenkov detectors with swin transformers and normalizing flow models.*" Machine Learning: Science and Technology 6.1 (2025): 015028. [link]

3. **Towards Foundation Models for DIRC:**
   Recently introduced a unified model architecture capable of performing both reconstruction and fast simulation, enabling simultaneous achievement of (1) and (2) within a single framework.

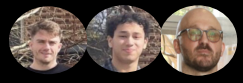
   J. Giroux, C. Fanelli, "*Towards Foundation Models for Experimental Readout Systems Combining Discrete and Continuous Data*." Machine Learning: Science and Technology, 7.1 (2026): 015031 [link]

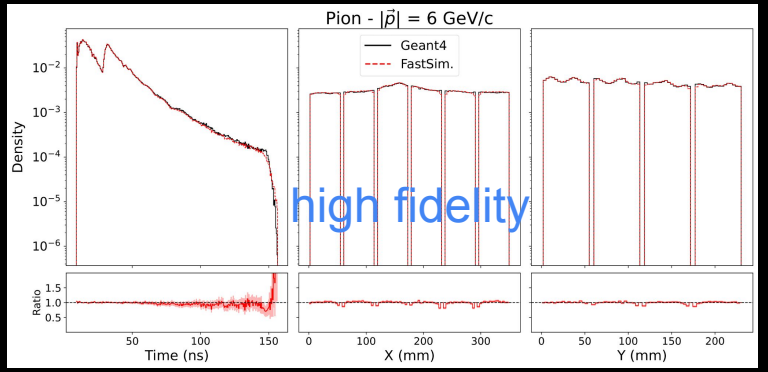
Fast Simulation - hpDIRC @ePIC (EIC)

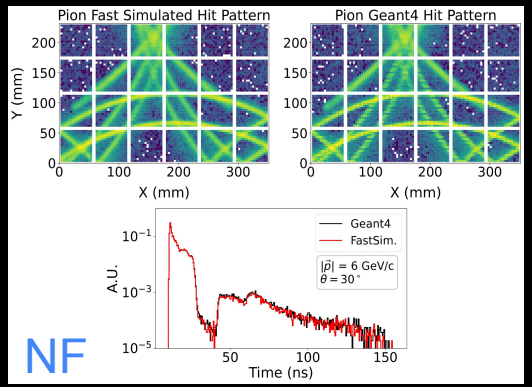
Architectures: Normalizing Flows (NF), Continuous Normalizing Flows (CNF), Conditional Flow Matching (CFM), Denoising Diffusion Probabilistic Models (DDPM), Score Based Generative Models (SB)

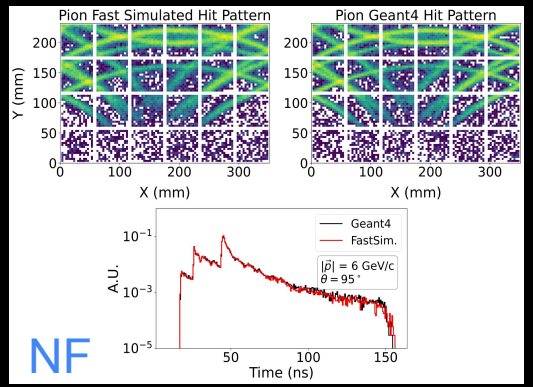
- **Suite of SOTA Generative Models** – Compare modern SOTA generative algorithms in the space of DIRC simulation
- **Hit-Level Learning** – Model conditioned on kinematic parameters (|p|,θ)
- **Agnostic to Photon Yield** – Ensure model independence from photon yield
- **Abstract away Fixed Input Size** – Address limitations with discrete distributions; data preprocessing transform DIRC readout (row, col) to (x,y) in mm and uniformly smear over PMT pixels

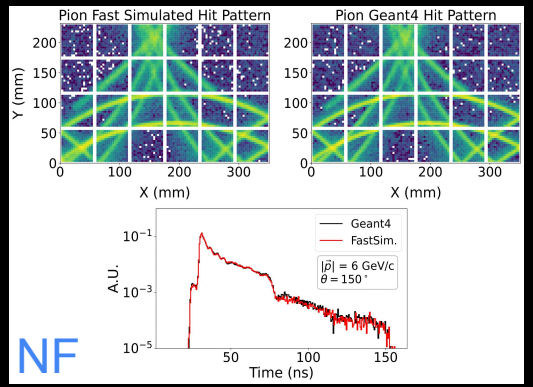
J. Giroux, James, M. Martinez, and CF. "*Generative Models for Fast Simulation of Cherenkov Detectors at the Electron-Ion Collider:*" *2025 Mach. Learn.: Sci. Technol. 6 040501*









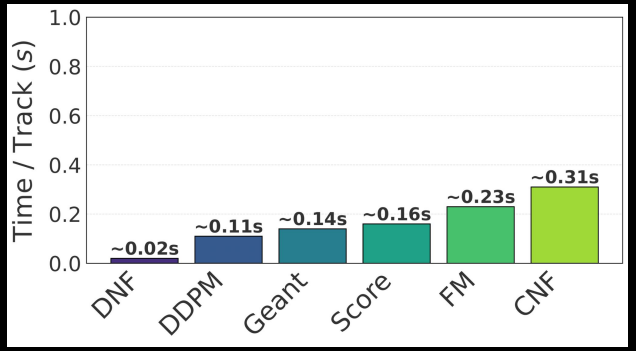


Simulation is fast - O(0.5)μs per hit (effective)

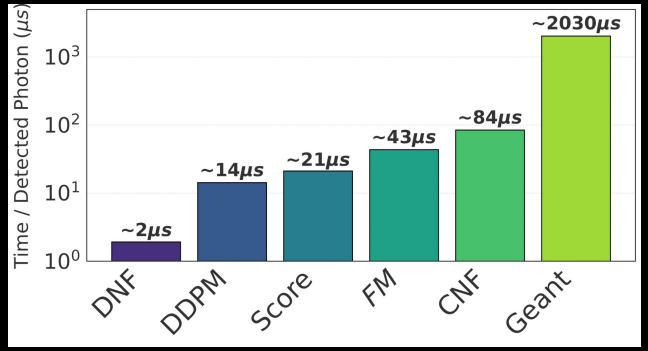
(hpDIRC standalone sim)

Fast Simulation - hpDIRC at EIC

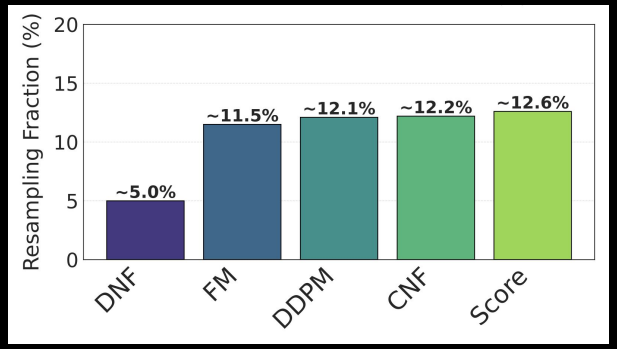
- Ring and time structures follow correct kinematic dependencies for both particle types (π/K)
  - See paper for more in depth evaluation
- We have created an open source suite of SOTA algorithms for the hpDIRC (easily adapted to other detectors)
- Our fast simulation is self-contained, fast and capable of being run on CPU or GPU



Track Generation (CPU)



Photon Generation - Large PDFs (models leverage GPU)

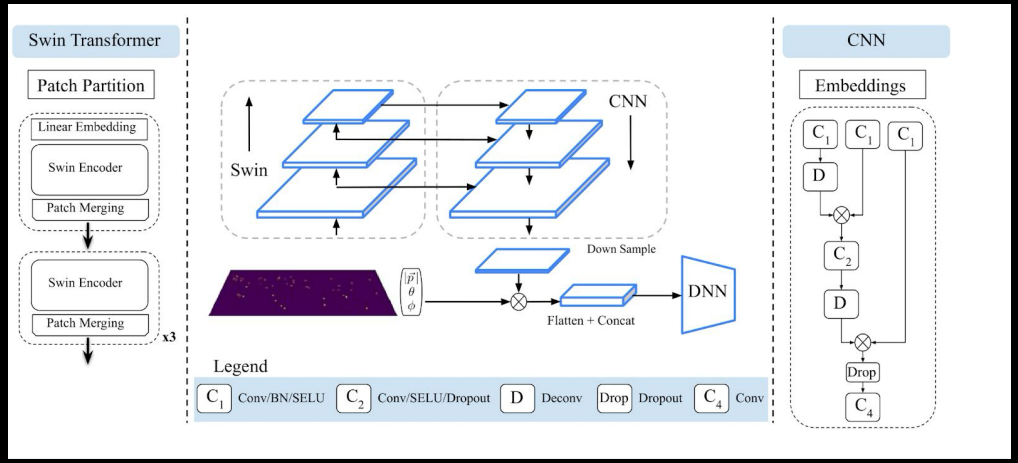


Resampling Fraction

Deep(er)RICH: PID

- Individual tracks do form “images” in optical boxes
  - Sparse point representations
- Possibility of overlapping hits
  - Same *x*,*y* - different times
  - Construct these as images as FIFO
  - Tends to be low percentage of overlap

CF, J. Giroux, J. Stevens. “*Deep(er)RICH: Deep(er) reconstruction of imaging Cherenkov detectors with swin transformers and normalizing flow models*” Machine Learning: Science and Technology 6.1 (2025): 015028.



- Hierarchical Vision Transformer (Swin) - encoder style feature extraction
  - Windowed attention - higher throughput
- Combine information through CNN - utilize skip connections for different resolutions
- Inject kinematics as concatenated information to DNN

NF-based PID (2nd method)

CF, J. Giroux, J. Stevens. "*Deep(er)RICH: Deep(er) reconstruction of imaging Cherenkov detectors with swin transformers and normalizing flow models*" Machine Learning: Science and Technology 6.1 (2025): 015028.

- Recall our bijection

$$x_k = f_\theta(z,k) = f_{\theta_N} \circ f_{\theta_{N-1}} \circ ... f_{\theta_1}(z_0,k)$$

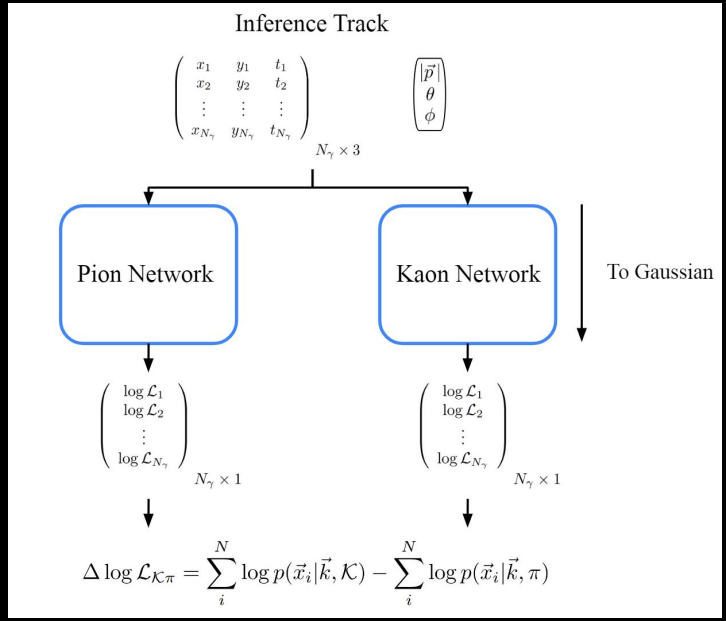
- Recall our analytical computation of the likelihood under a change of variables

$$\log p(x|k) = \log q(f_\theta^{-1}(x)|k) + \sum_{i=1}^{N} \log \left| det \left(\frac{\partial f_{\theta_i}^{-1}(x)}{\partial x} \right) \right|$$

- We can compute the DLL under the base distribution - summed contribution over hits

$$\Delta \log \mathcal{L}_{K\pi} = \sum_i^N \log p(\vec{x}_i|\vec{k},K) - \sum_i^N \log p(\vec{x}_i|\vec{k},\pi)$$

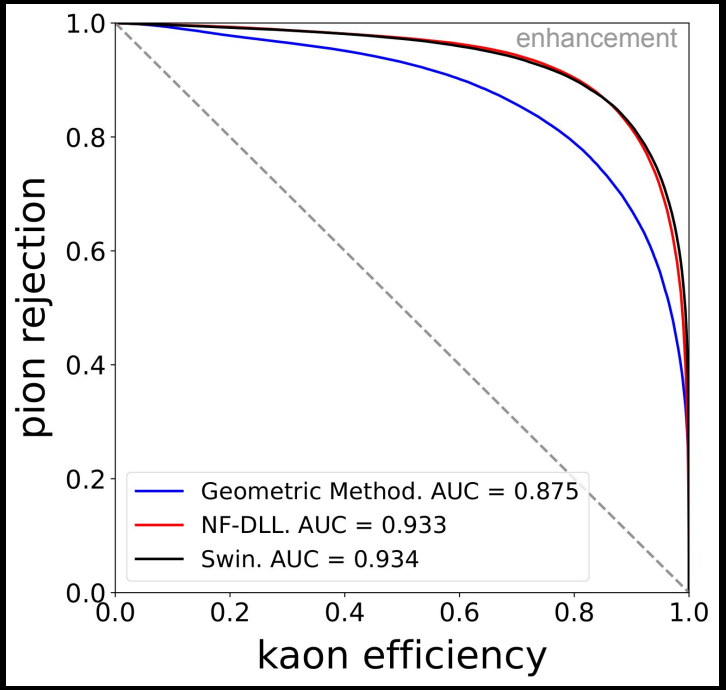
—the hypothesis of π/K represented by individual networks—

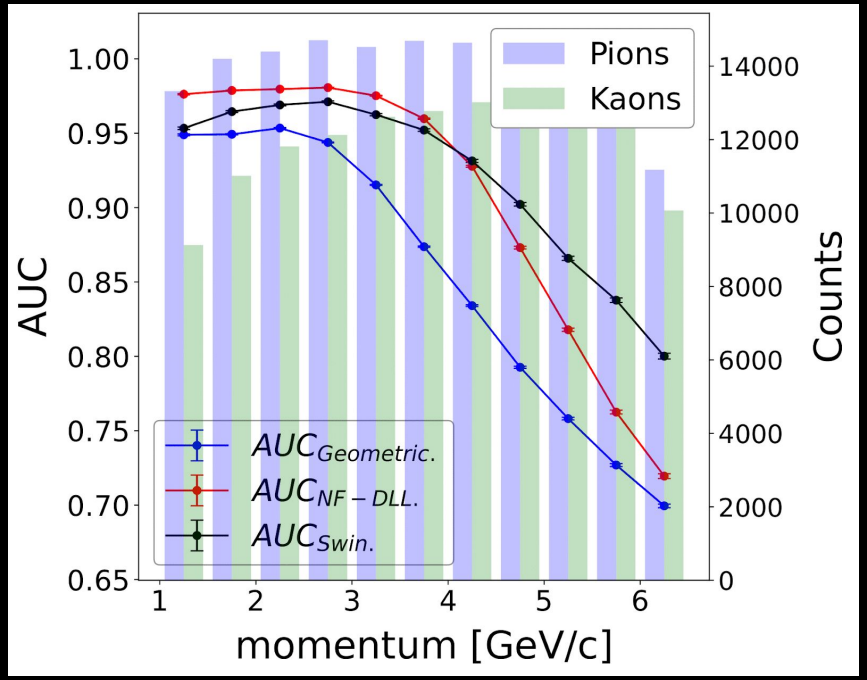


- Normalizing Flow - Likelihood based PID
- PID hypotheses represented through independent models
- Analytic likelihood computation from NF in base distribution
- Compute Delta-Log Likelihood

PID Performance - GlueX

CF, J. Giroux, J. Stevens. “Machine Learning: Science and Technology 6.1 (2025): 015028.





Github

PID is fast - O(10)*μs* per track with transformer (effective)

Bonus: NF for PID. This method is slightly slower.

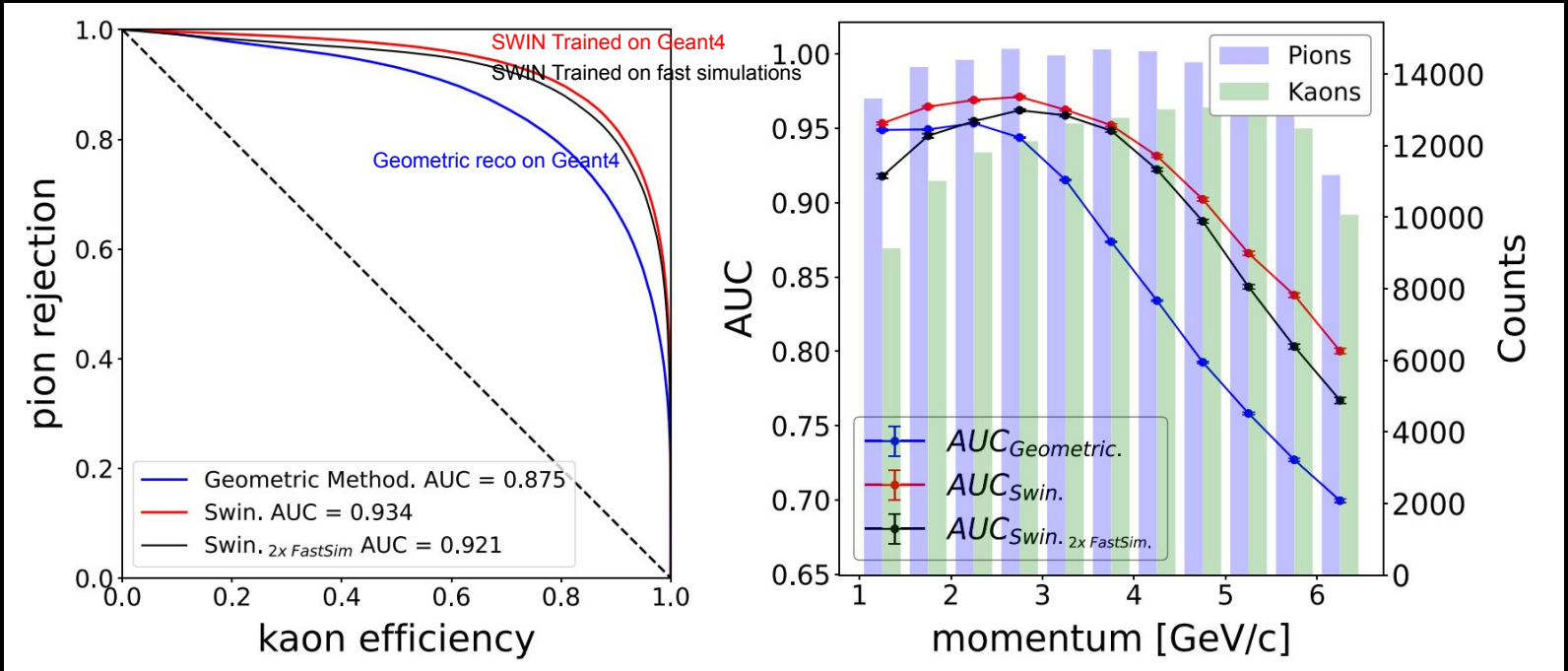
All code is open source and pre-trained models are provided.

(GlueX DIRC sim)

How to measure fidelity of synthetic data?

- Fidelity of synthetic data is typically assessed through dedicated metrics. These are primarily “relative”: they enable comparisons between datasets but do not define an absolute notion of fidelity.
- Information-theoretic fidelity metric: Datasets consistent with the same underlying physics admit similar Shannon-optimal compression under physics-aware arithmetic coding; discrepancies manifest as excess code length (in bits). arxiv:2602.19476
- In these slides:
  - we use 1D ratio plots to compare synthetic and real distributions.
  - We also evaluate fidelity through downstream physics performance, applying a PID classifier and comparing the resulting performance with that obtained using detailed simulations such as Geant4.

Fidelity of Fast Sim - DIRC in GlueX

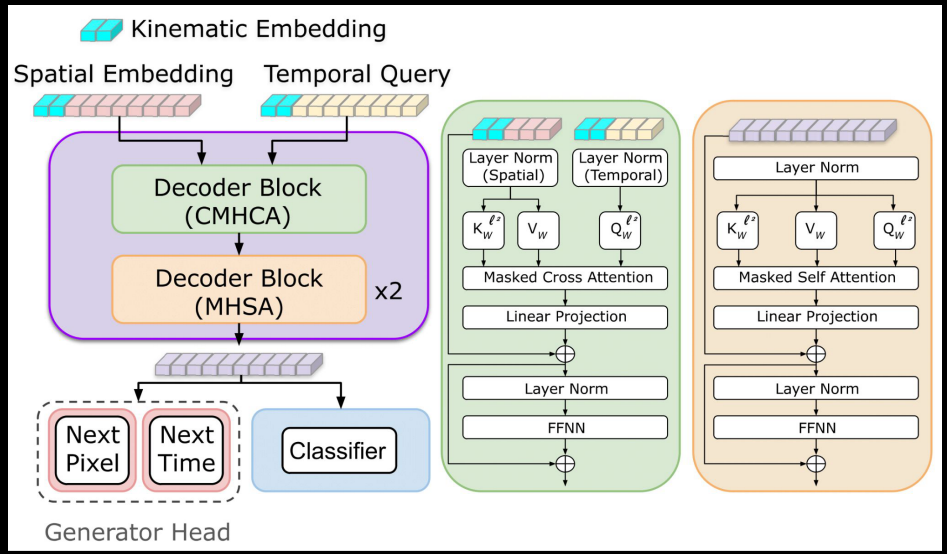


- An independent classifier (SWIN transformer) is trained on Geant4 data and compared to training on fast simulations.
- For comparison, performance obtained from a standard method (geometric) are also shown.

Foundation Model - hpDIRC

- Foundation Models capable of generalizing to multiple tasks
  - Pre-trained backbone structure (transformer based)
- *Fine-tune* to different tasks
  - Generation
  - Classification
  - Noise Filtering
- Represent hits in *tokenized* space

J. Giroux and C Fanelli "*Towards Foundation Models for Experimental Readout Systems Combining Discrete and Continuous Data.*" *Machine Learning: Science and Technology 7.1 (2026): 015031*



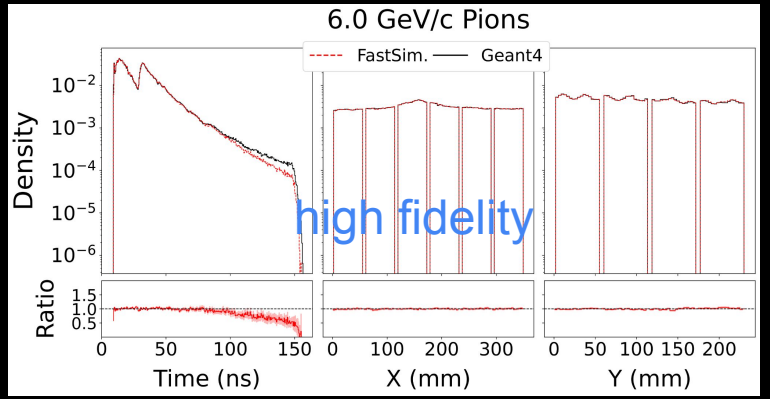
$$\text{spatial} \rightarrow \{|\vec{p}|, \theta, \text{SOS}_p, p_1, \ldots, p_n, \text{EOS}_p\}$$
$$\text{time} \rightarrow \{|\vec{p}|, \theta, \text{SOS}_t, t_1, \ldots, t_n, \text{EOS}_t\}$$

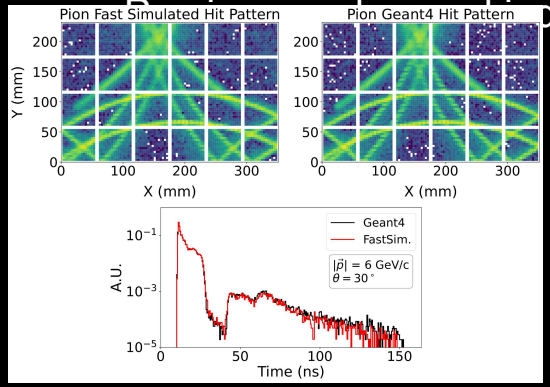
Github

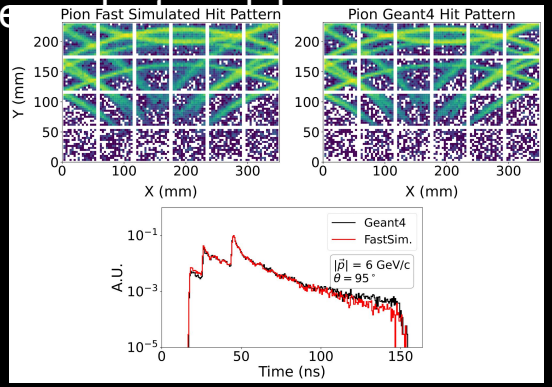
All code is open source and pre-trained models are provided.

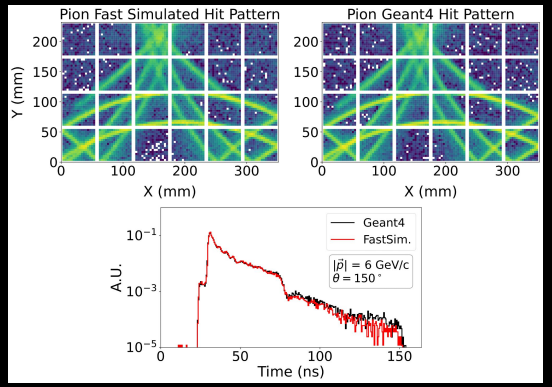
Foundation Model - Fast Sim

- Fast simulation through *next token* prediction
- Directly learns variability in photon yield
  - Model conditioned on kinematic parameters (|p|, θ)
  - No external modeling of photon yield required
- Class conditional (particle type) generation through a fixed routing *Mixture of Experts* (MoE)







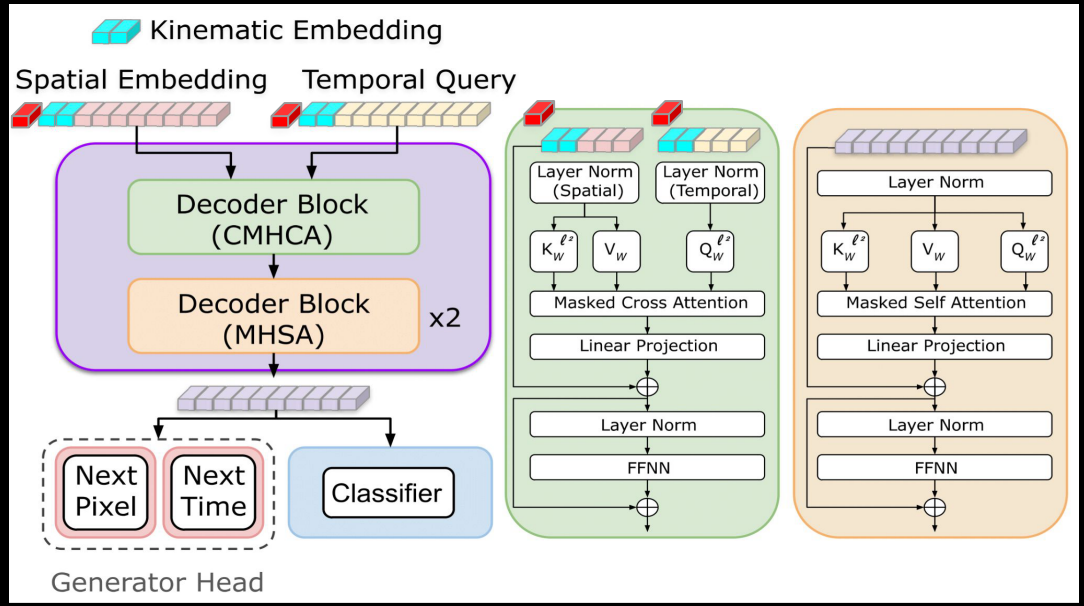


Simulation is fast - O(0.02)*s* per track (effective)

(hpDIRC standalone sim)

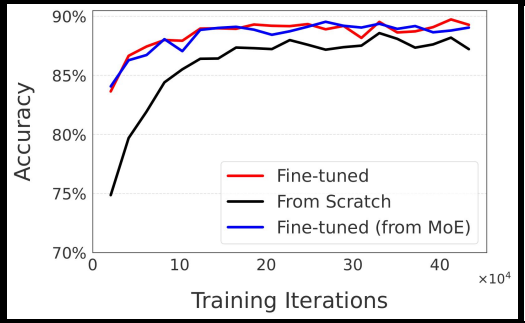
Foundation Model - From Sim to PID

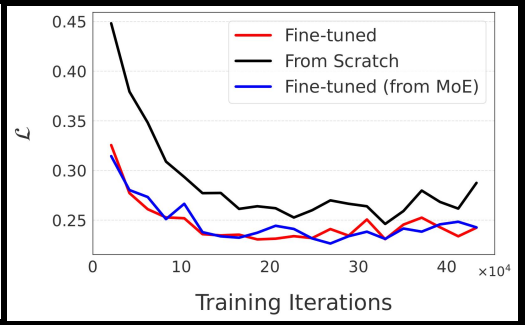
- Our model also supports classification (π / K)
  - Additional token - ***CLS Token***
  - Remove causal masking
  - Can be **fine-tuned**

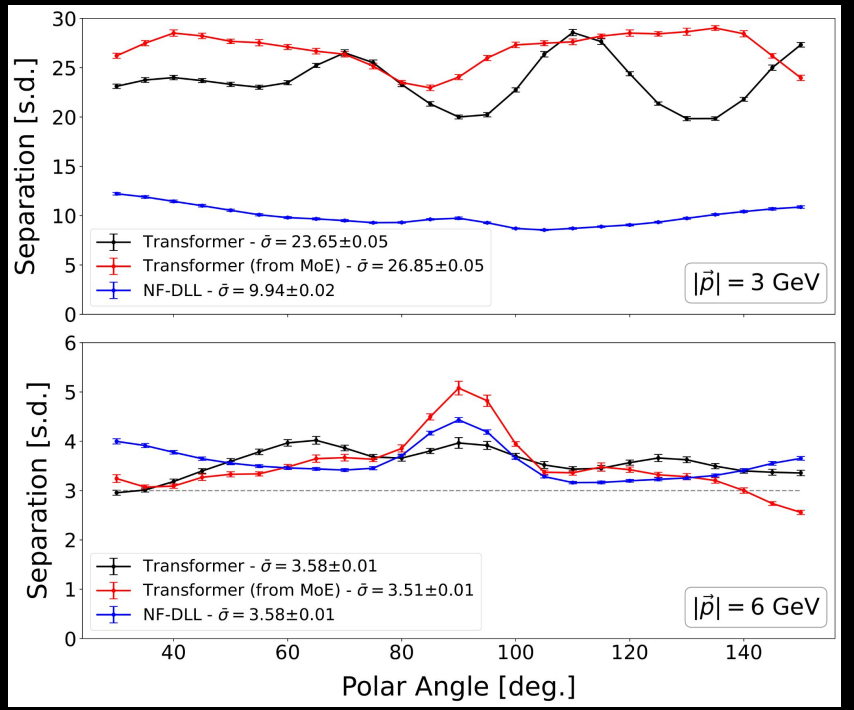


Foundation Model - PID

- Classification (π/K) through fine-tuning fast simulation model (sequence level)
  - Decrease in required training time
  - Increased performance
- Reaching separation requirement of 3σ at 6 GeV/c







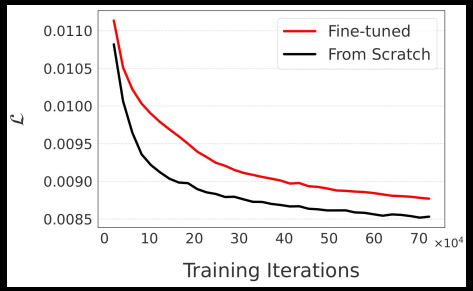
Fine-tuned models are initialized with weights from a separately trained generative model (e.g., trained on π or K);
"from MoE" models are initialized using weights from a multi-expert architecture with 4 experts

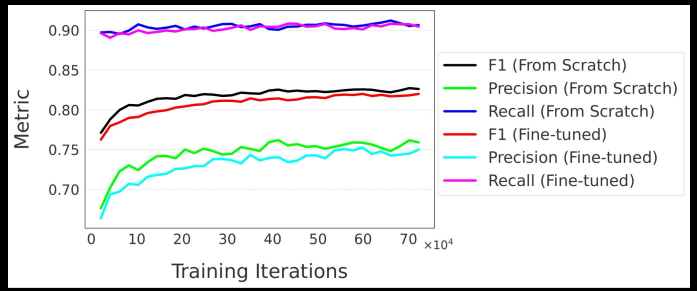
(hpDIRC standalone sim)

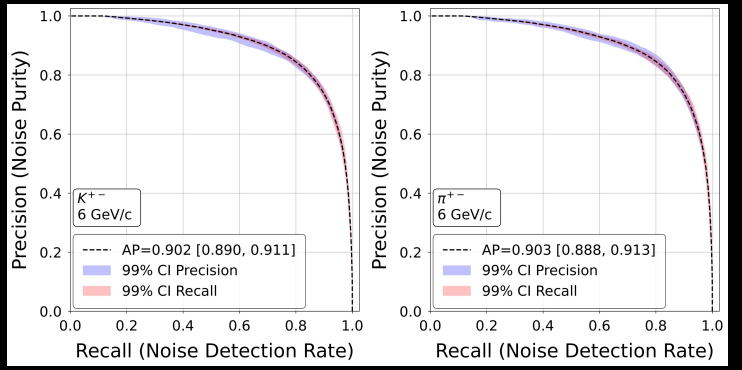
Foundation Model - Noise Filter

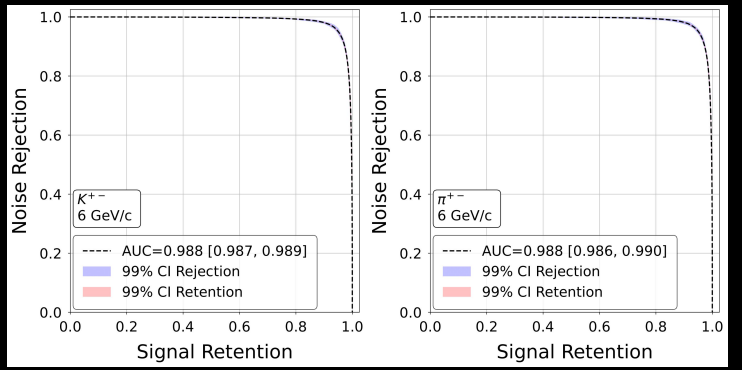
Noise filtering (proof of principle)

- Simulated dark rate of ~ 100 khz/cm^2
- Classification of noise hits (token level)
- Fine-tuning not valuable here
  - Prior attention heads have learned information under a more global context
  - Need to unlearn and realign attention







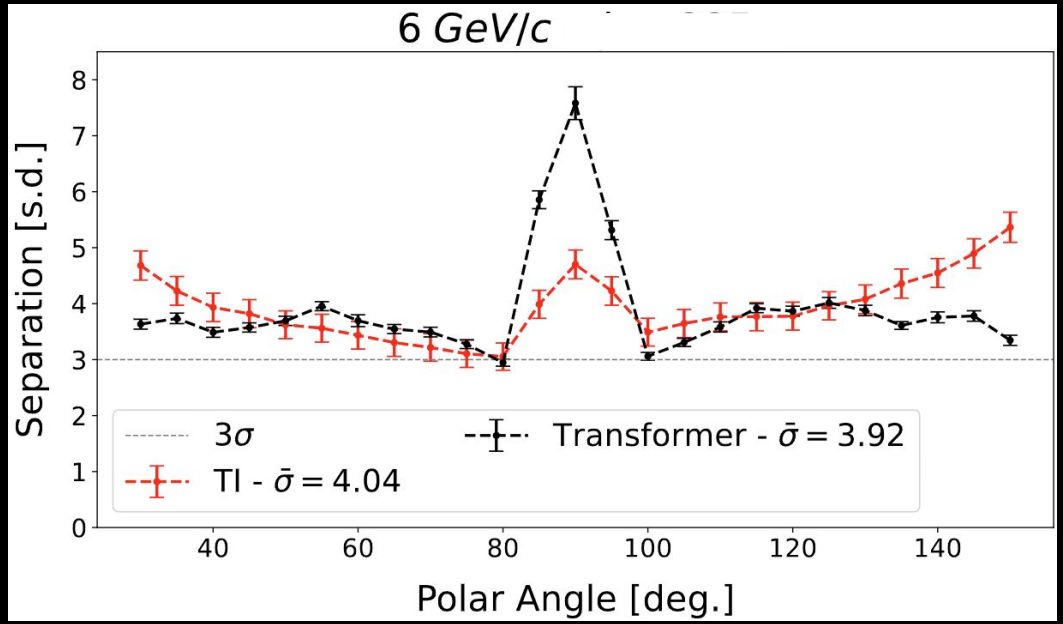


Preliminary Studies

(hpDIRC standalone sim)

Ongoing Studies

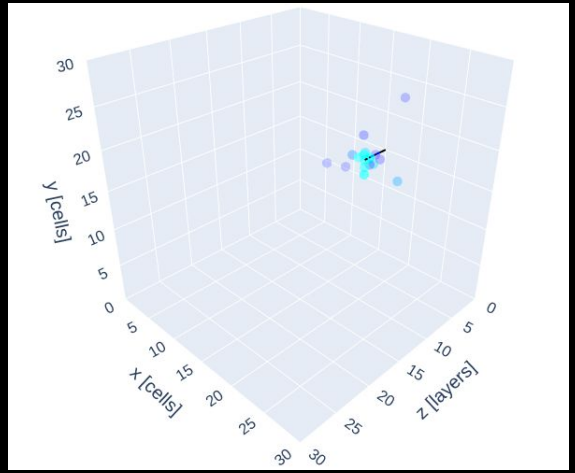
- π / K separation using Foundation Models
  - Full ϕ range considered - model is trained over all bars continuously
  - Magnetic field on

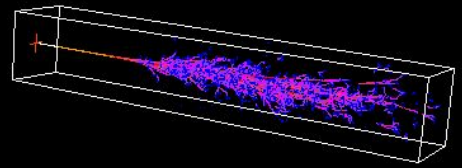


Extensible to full particle kinematics

Towards FM for Calorimetry

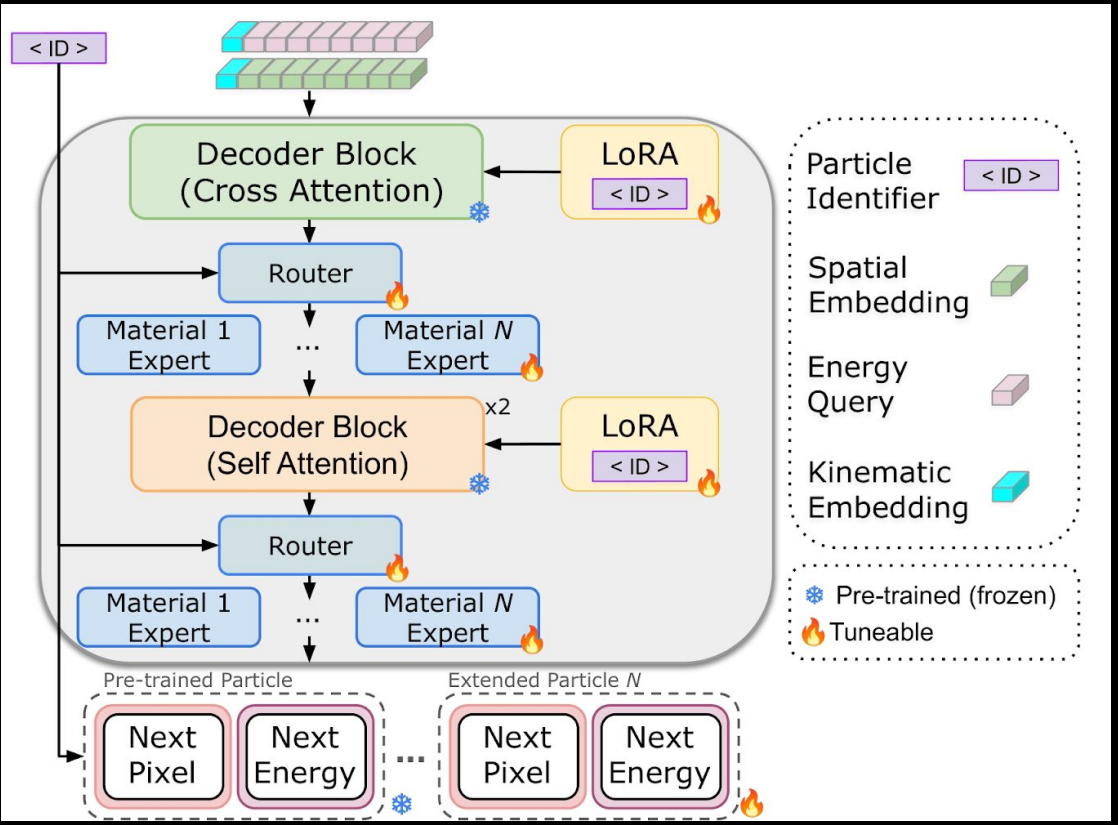
- Calorimeters naturally align with FM paradigms:
  - Discrete / pixelated detector structure
  - Strong correlations between particle kinematics and energy deposition patterns
  - Large simulation demands for detector design and reconstruction
- Generalizable FM for calorimetry using:
  - Transformer-based autoregressive architectures
  - Mixture-of-Experts
  - Parameter-Efficient Fine-Tuning (LoRA)
- Target Capabilities
  - Fast Sim / Reco & Analysis
  - Adaptation to new particle species, new absorber materials, new geometries
- Design Considerations
  - Efficient fine-tuning with limited new samples
  - Incorporation of new detector knowledge without retraining the full model
  - Reduced catastrophic forgetting and model misalignment





Source:
International Large Detector

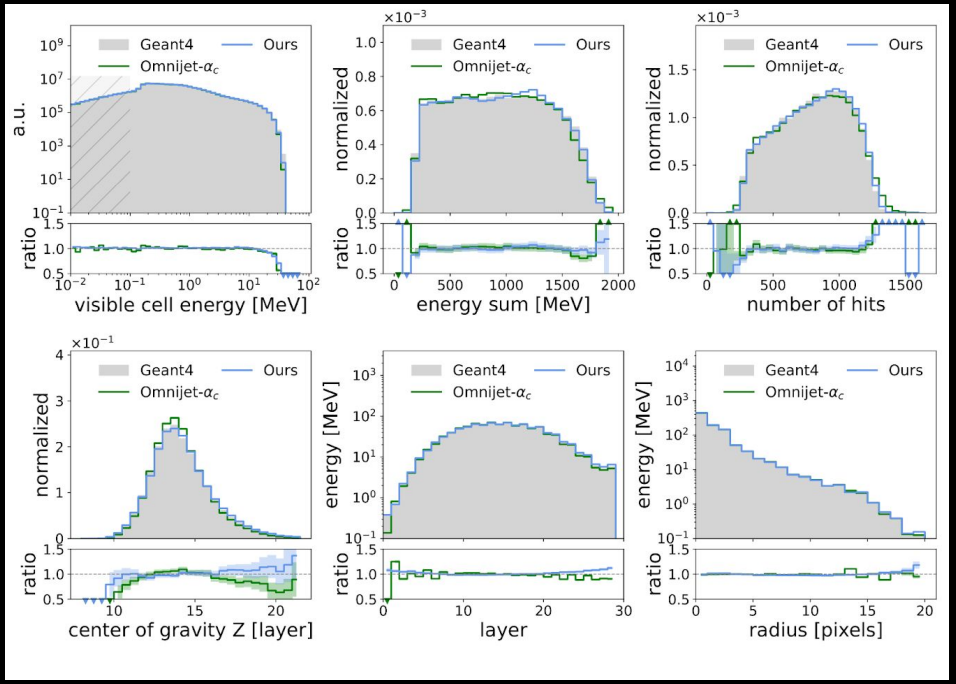
Generalizable FM for Calorimetry



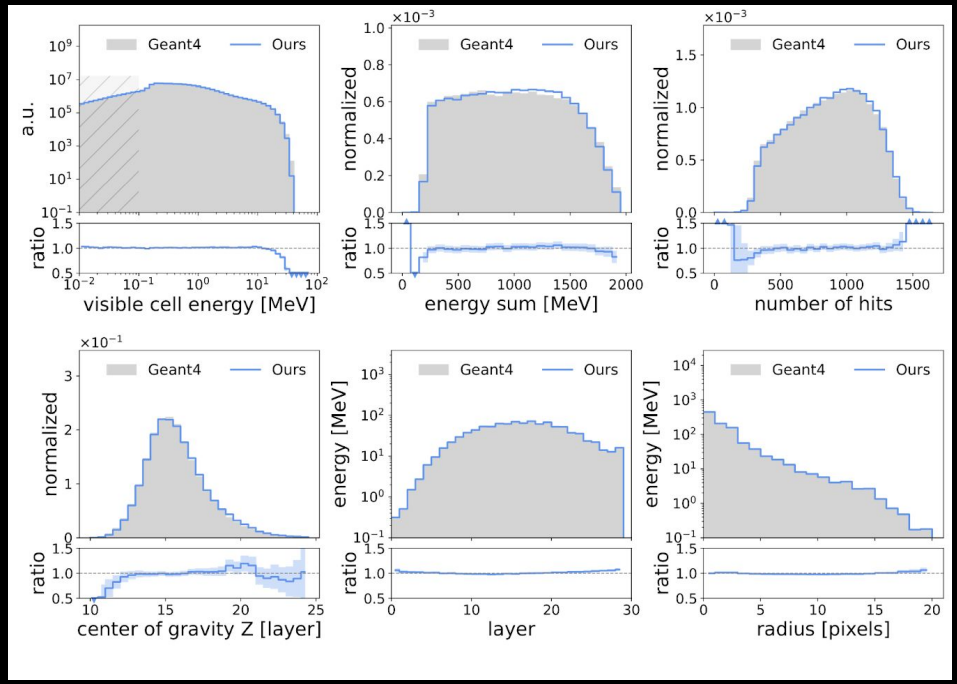
- Pre-training
  - Model absorber materials through experts
- Material Fine Tuning
  - Frozen backbone + new expert(s)
- Particle/Material Fine Tuning
  - Frozen backbone
  - LoRA handles shifts in particle dynamics (Attention)
  - MoE handless material marginals
  - Particle specific vocab projections

C. Cardona, CF et al. "Generalizable Foundation Models for Calorimetry via Mixtures-of-Experts and Parameter Efficient Fine Tuning."
arXiv:2603.28804 (2026)

Pre-Training - Mixture of Materials

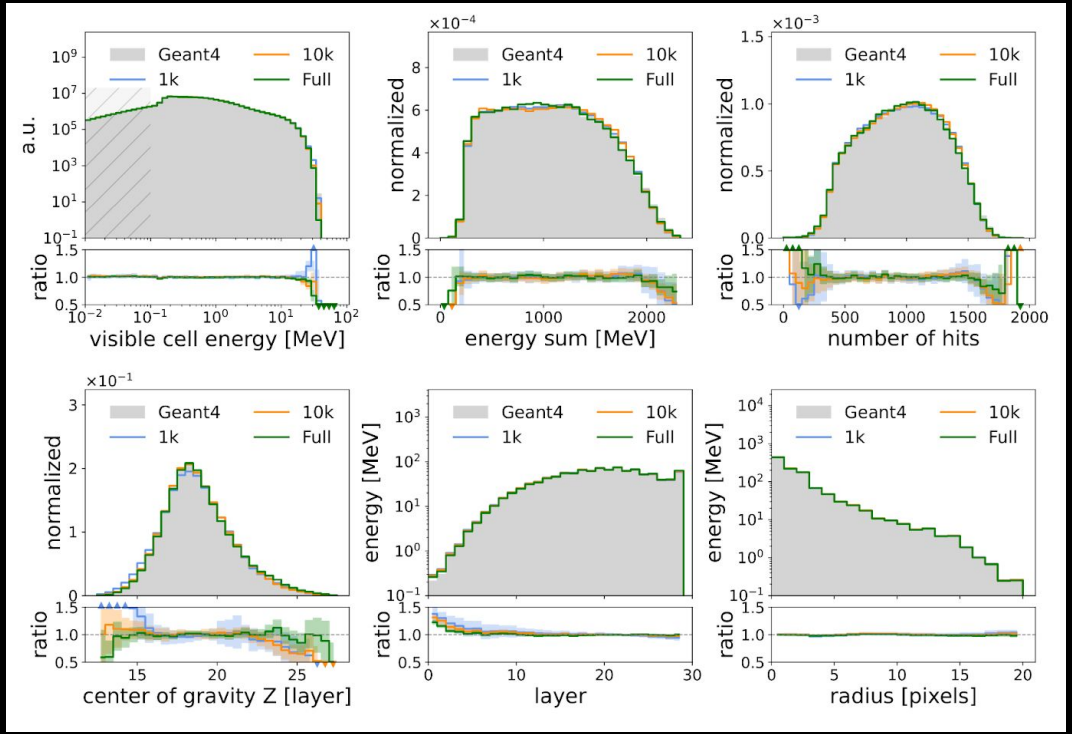


(a) Photons in Tungsten



(b) Photons in Tantalum

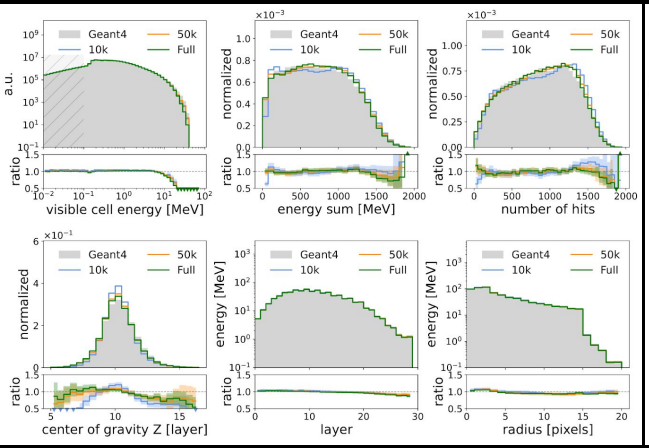
Fine-tuning - Addition of an Expert



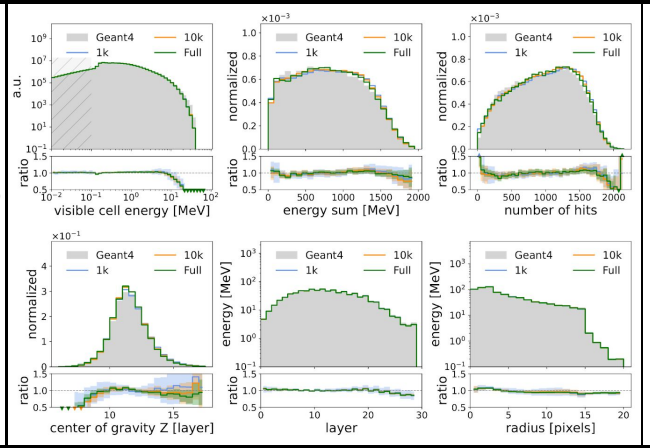
(c) Photons in Lead

- High sample efficiency
- Converges to performance of the full dataset in **~10k samples**
- Pre-training is limited
  - Expect improvements with scale

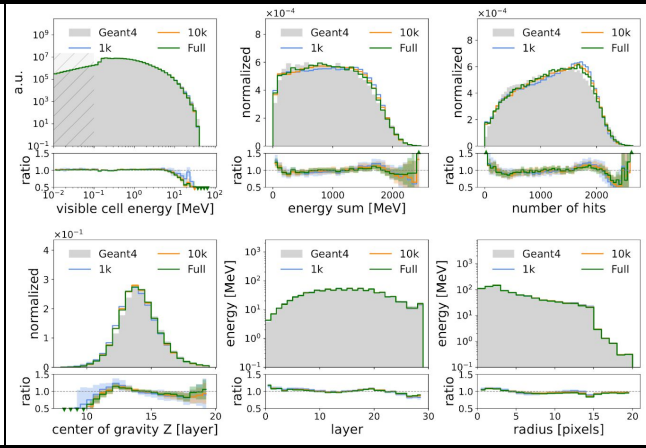
Extension to Different Particles



(a) Electrons in Tungsten



(b) Electrons in Tantalum

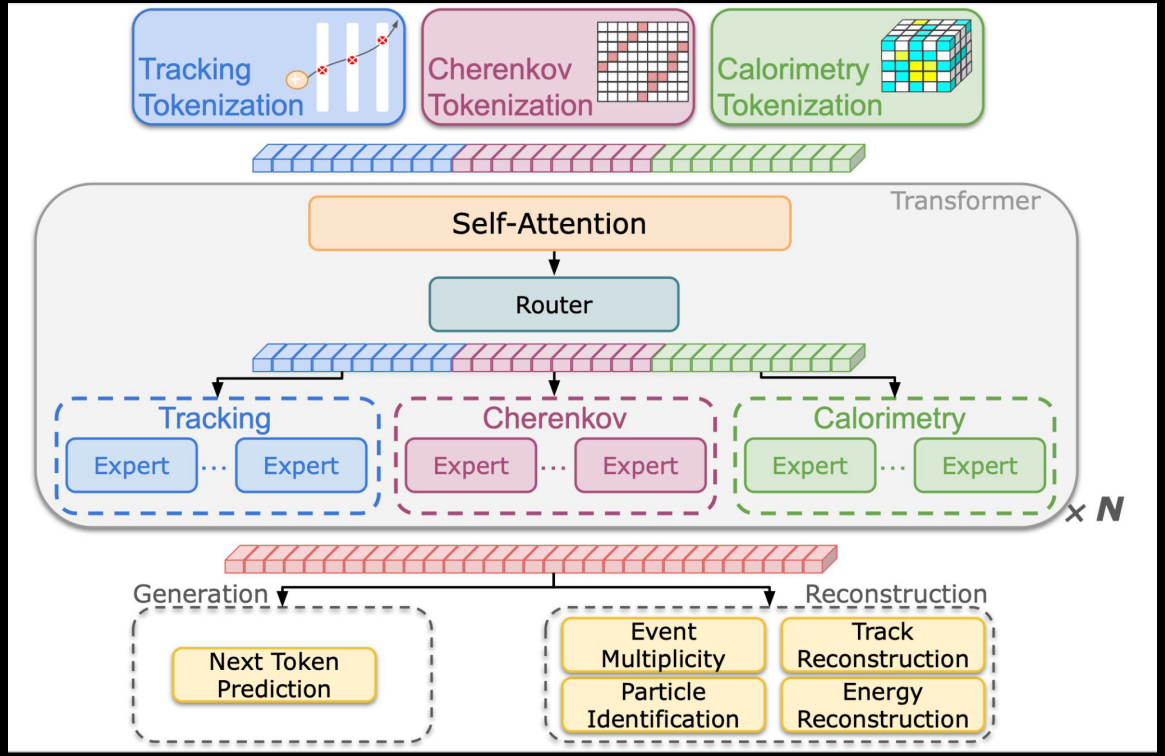


(c) Electrons in Lead

- Particle/Material Fine Tuning
  - Frozen backbone
  - LoRA handles shifts in particle dynamics (Attention)
  - MoE handless material marginals
  - Particle specific vocab projections
- Relatively high sample efficiency
  - Converges in ~50k samples
- Adaptation modules can now be frozen
  - New materials through experts

Extension to Multiple Detectors

And, more ambitiously…



Leveraging a unified latent representation across detector subsystems

Conclusions

- Deep learning approaches can outperform classical methods for Cherenkov detector reconstruction and analysis.
- Foundation Models provide a promising framework for learning detector and physics representations in HEP/NP.
  - Shared backbone architectures can support multiple downstream tasks:
    - High-fidelity + fast simulation
    - (Near real-time) PID / reconstruction
    - Denoising and other analysis tasks
  - Unified architectures may reduce fragmentation across ML workflows and detector subsystems.
- Early studies on DIRC detectors demonstrate the feasibility of FM-inspired approaches in our field.
- These approaches are beginning to generalize to additional detector systems, including calorimetry (ILD example).
- Long-term perspective:
  - Toward multimodal, event-level FM integrating multiple detector subsystems within unified AI frameworks.
  - Potential applications for future experiments such as ePIC at EIC.

Backup

https://www.bnl.gov/rhicagsaum/index.php