Toward Unified Deep Learning Models to Simulation and PID with Cherenkov Detectors: the hpDIRC case

Cristiano Fanelli

IVth AI4EIC workshop, MIT/AIFI, Oct 27-29, 2025

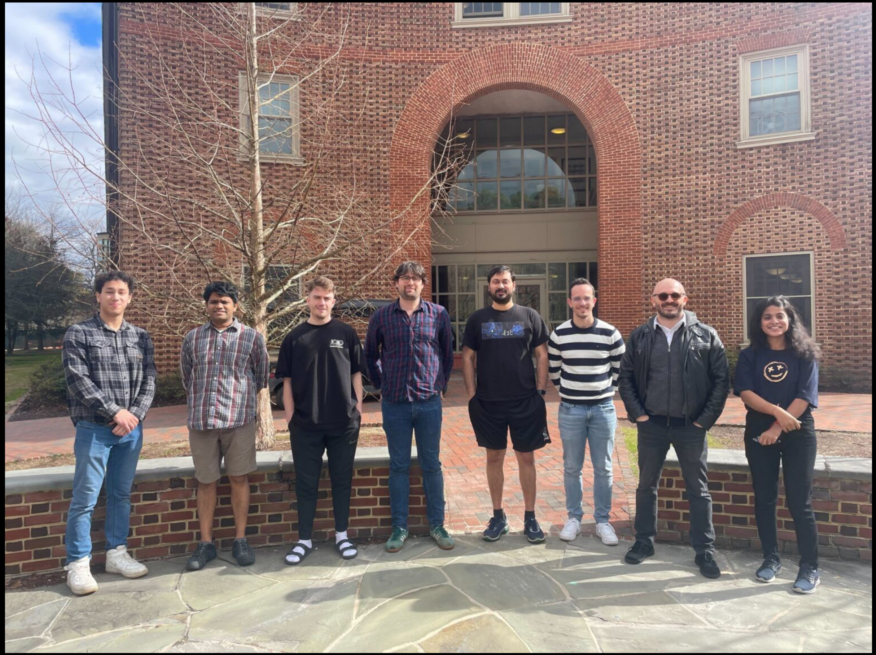
Outline & Acknowledgements

- Physics Motivation: short introduction to Imaging Cherenkov Detectors
- Role of AI/ML in Imaging Cherenkov Detectors
- Novel approaches to Reconstruction and Simulations
- Combining all tasks together: Foundation Models

CAREER

Some portions of this work are now supported by the NSF CAREER Award

People & Acknowledgements



James Giroux
PhD thesis (2023-)

Mike Martinez
BSc Honors Thesis (2025)

- DL for Reco/PID
- Gen AI / Fast Sim
- Foundation Models

AI for the Physical Sciences
Dept. Data Science, W&M

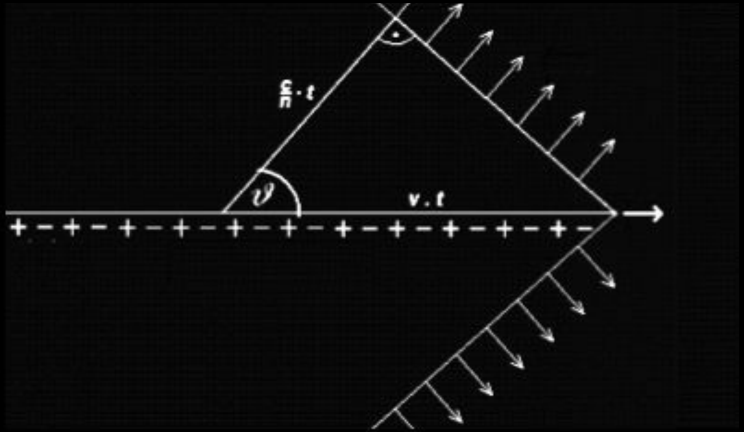
Imaging Cherenkov Detectors

Imaging Cherenkov Detectors

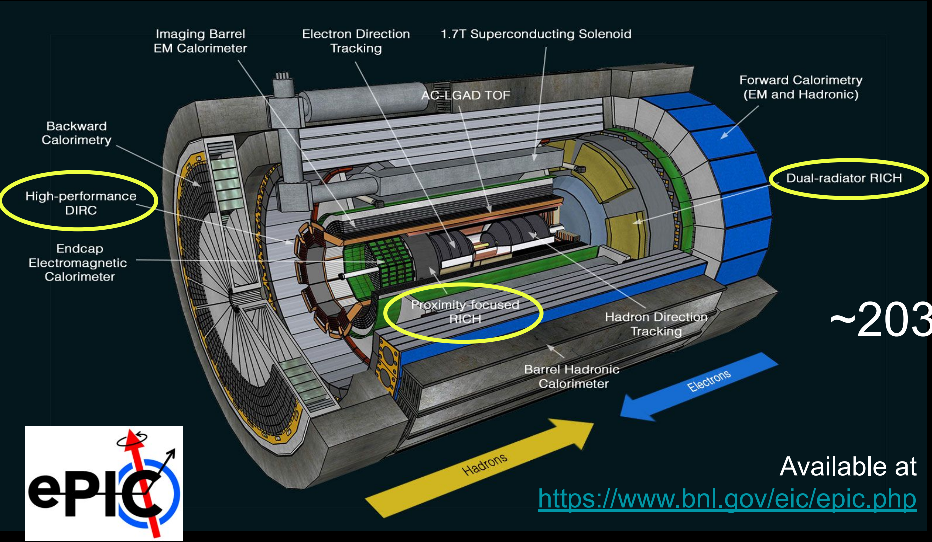
- Imaging Cherenkov detectors are largely used in many medium- and high-energy nuclear and particle physics experiments as particle identification (PID) systems

- They leverage the Cherenkov effect—light emitted when charged particles traverse a medium faster than the speed of light in that medium—to measure particle velocity with high precision. Combined with momentum information from tracking detectors, they enable the determination of particle mass, allowing for the separation of hadron species over wide momentum ranges.

- Their versatility and accuracy make them indispensable for studies of hadron structure, quark–gluon plasma properties, and searches for new physics.



AI/ML & Cherenkov Detectors



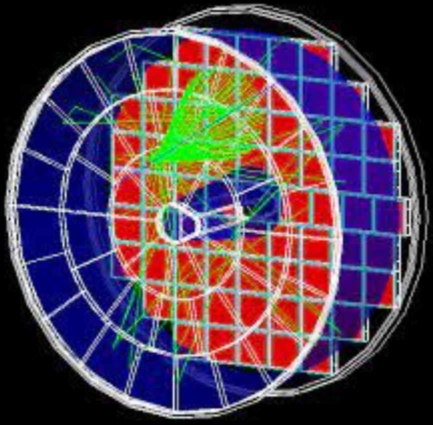
~2030

- Cherenkov detectors constitute the backbone of PID (DIRC, dRICH, pfRICH)
- They represent a major simulation bottleneck in that optical photons involve multiple photons that need to be tracked through complex surfaces → need for fast high fidelity simulations
- All Cherenkov detectors rely on pattern recognition of ring images in the reconstruction, which may become particularly complex like in the case of the DIRC → need to enhance reconstruction

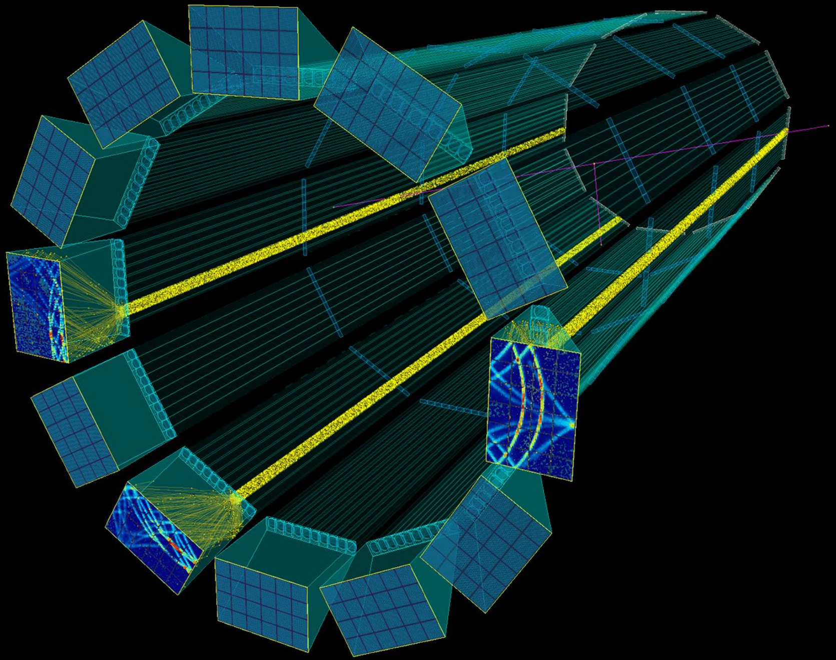
Additional Desiderata:

- Reconstruction at the "event-level" rather than "track-level" (*e.g.*, two tracks with overlapping patterns in the same optical box) — N.b. over 10% of SIDIS events involve at least two charged tracks with momenta above 1 GeV/c detected simultaneously in one sector of the hpDIRC
- Possibility of learning directly from real data the detector response.
- Faster algorithms to cope with near real-time analysis

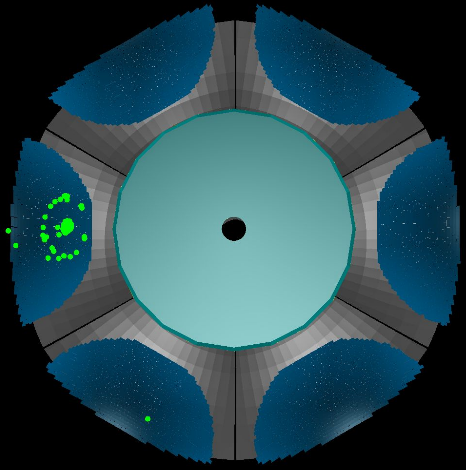
Cherenkov Detectors in epIC/EIC



pfRICH (electron endcap)

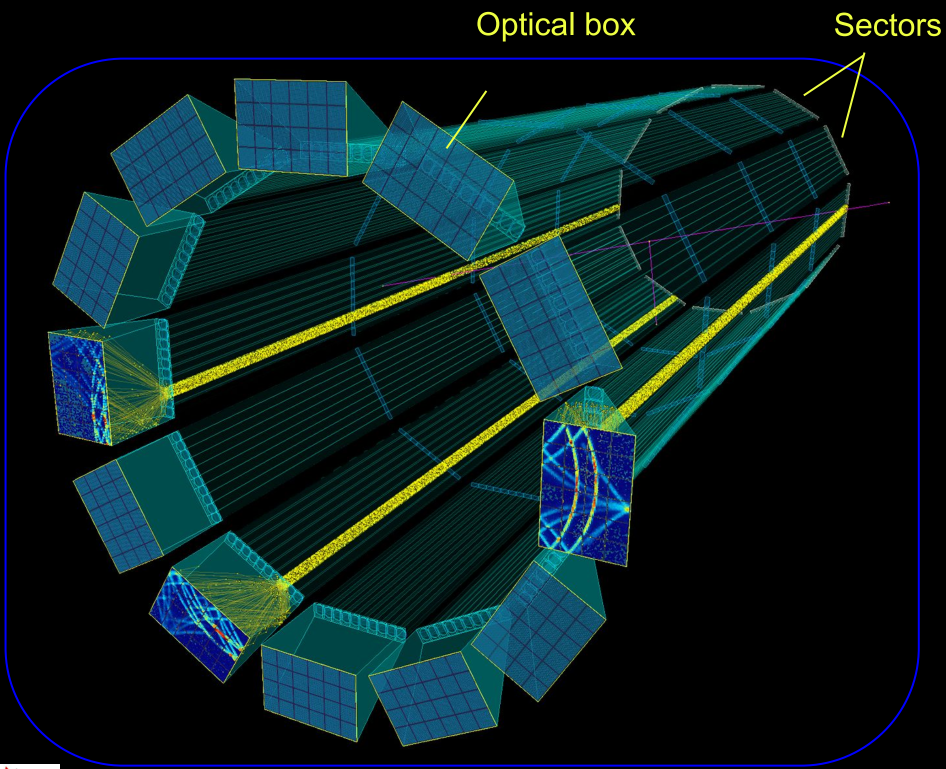


hpDIRC (barrel)



dRICH (hadron endcap)

DIRC Detectors (hpDIRC in ePIC)



12 sectors (bar boxes) with 12 optical boxes

10 radiator bars / sector

Total bar length ~5.48m

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

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

16x16 pixels/PMT

100ps precision on photon arrival time

DIRC Detectors (GlueX DIRC)

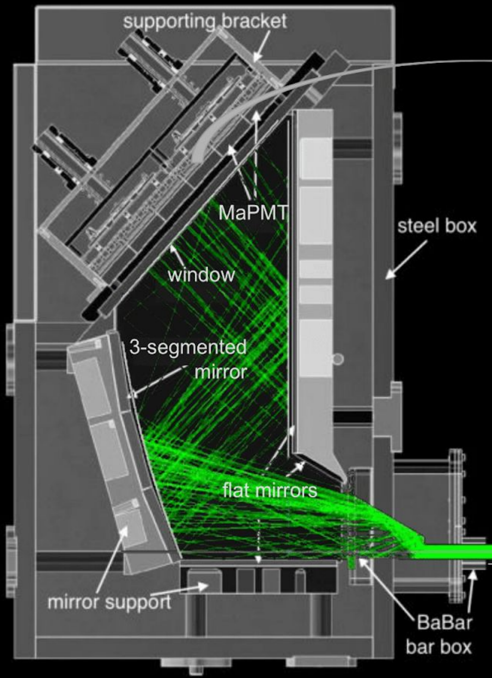
- In this talk I will focus on DIRC detectors. Goal is to do PID from their hit patterns.
- DIRC detectors have complex and sparse space–time hit patterns in the (x, y, t) readout.

48 fused silica bars segmented into 4 bar boxes

Two optical boxes (distilled water and reflective mirrors)

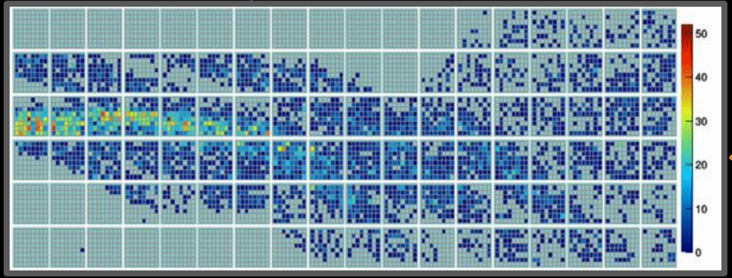
6 x 18 PMT (8 x 8 pixels) array for photon detection.

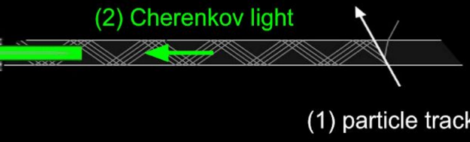
Provides location and timing information for photons



Hit patterns depend on kinematics

Shown below: collection of multiple tracks at given < |***p***| , θ , ϕ >

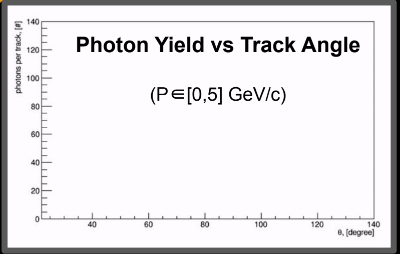




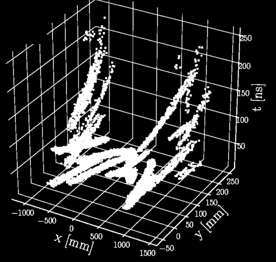
pions/kaons/protons

Patterns are sparse

Photon yield per particle

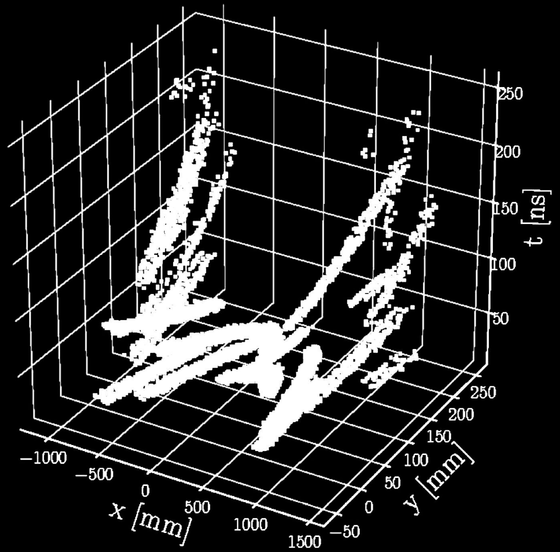


Patterns are 3-dimensional

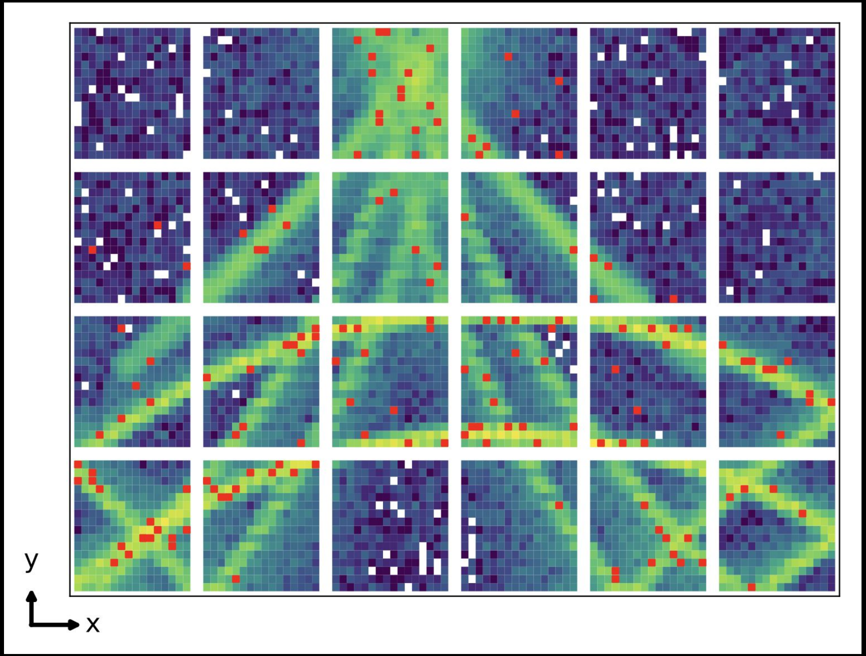


Cherenkov: A Beautiful Playground for ML

Complex Topologies



Sparse Data



(shown space only for illustrative purposes)

DeepRICH: Deep Reco of Imaging CHerenkov (2020)

Deep Learning and DIRC Detector

- Machine learning for imaging Cherenkov detectors has grown significantly in recent years—particularly in the context of the EIC.
- Our 2020 study (link here) was the first to explore deep learning approaches for DIRC-like detectors:

PAPER • OPEN ACCESS

DeepRICH: learning deeply Cherenkov detectors

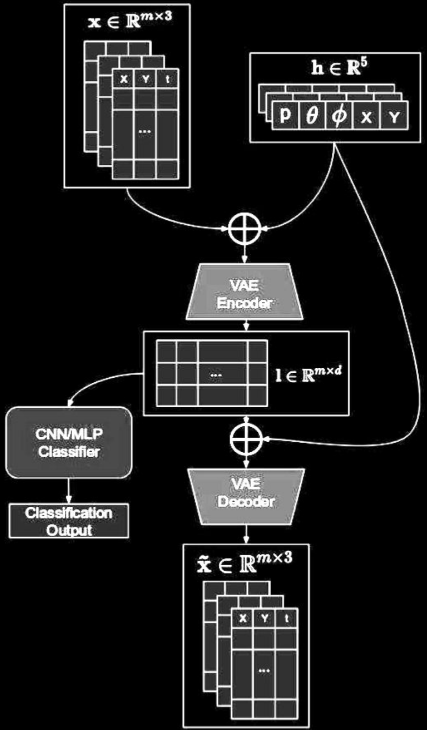
Cristiano Fanelli and Jary Pomponi

Published 27 April 2020 • © 2020 The Author(s). Published by IOP Publishing Ltd

Machine Learning: Science and Technology, Volume 1, Number 1

Citation Cristiano Fanelli and Jary Pomponi 2020 *Mach. Learn.: Sci. Technol.* **1** 015010

DOI 10.1088/2632-2153/ab845a



- This work helped demonstrate the potential of neural networks to:
  - Capture complex optical features directly from photon hit patterns
  - Offer alternatives to traditional reconstruction pipelines
  - Enable faster, data-driven inference for PID + Fast Simulation

Deep(er)RICH: Deeper Reco of Imaging CHerenkov

Modern Architectures and Advances

Since our initial work, we have significantly advanced ML for Cherenkov by leveraging and integrating modern architectures into novel solutions tailored to the DIRC reconstruction challenges at EIC.

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.

   J. Giroux, M. Martinez, C. Fanelli "Generative Models for Fast Simulation of Cherenkov Detectors at the Electron-Ion Collider." — 2025 Mach. Learn.: Sci. Technol. 6 040501 [link]

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

   C. Fanelli, J. Giroux, and J. Stevens. "Deep (er) reconstruction of imaging Cherenkov detectors with swin transformers and normalizing flow models." — 2025 Mach. Learn.: Sci. Technol. 6 015028 [link]

3. **Towards (Proto-) 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." arXiv:2505.08736 (2025). — submitted to Mach. Learn.: Sci. Technol.[link]

1) Fast Simulations

Github

Fast Sim with NF - DIRC @GlueX (JLab)

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

Architecture: Normalizing Flow (NF)

- **Density Transformation** – Define a bijective function and apply a change of variables, conditioning on kinematics parameters to maximize likelihood of expected hit pattern under a base distribution

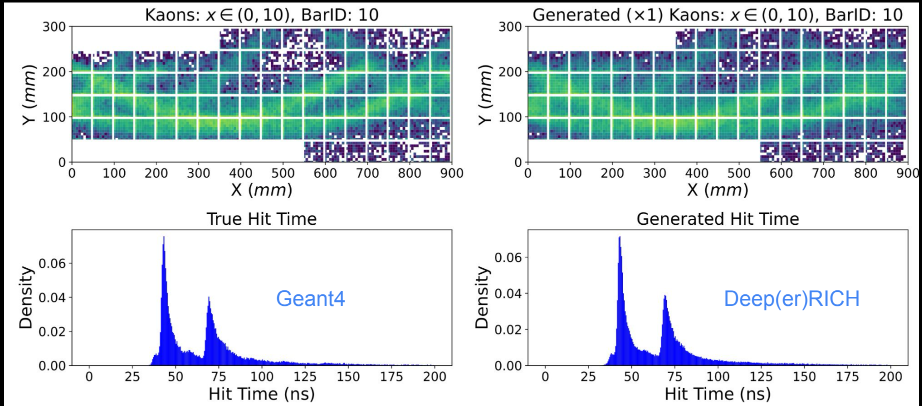
Use chain of bijections

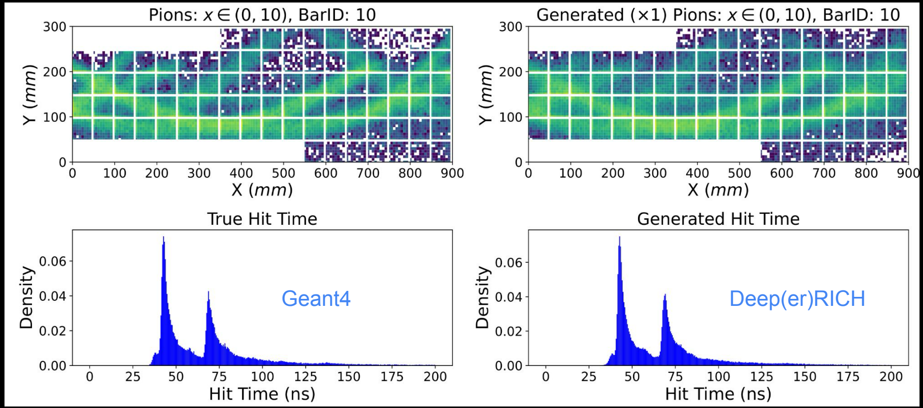
$$x_k = f_\theta(z, k) = f_{\theta_N} \circ f_{\theta_{N-1}} \circ \dots f_{\theta_1}(z_0, k)$$

Trained through MLE - exact likelihood computation at inference

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

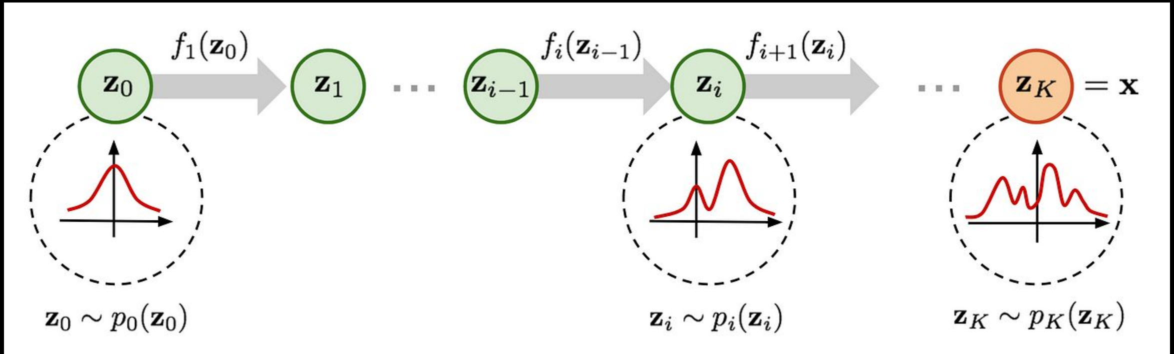
- **Hit-Level Learning** – Model conditioned on kinematic parameters (|p|,θ,ϕ)
- **Agnostic to Photon Yield** – Ensure model independence from photon yield, which is captured via a lookup table as a function of the kinematics.
- **Abstract away Fixed Input Size** – Address NF limitations with discrete distributions; data preprocessing transform DIRC readout (row, col) to (x,y) in mm and uniformly smear over PMT pixels



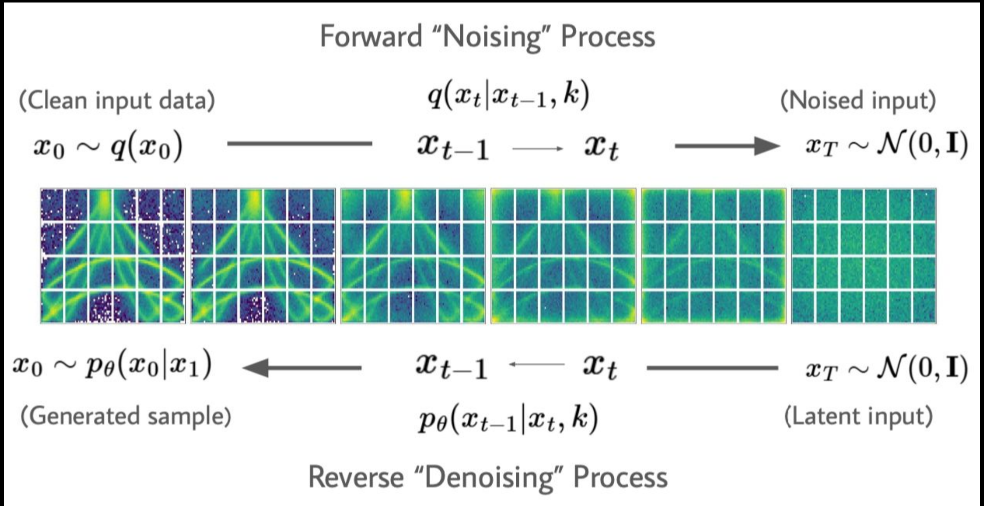


Simulation is fast - O(0.5)μs per hit (effective)

Comparing Different Generative AI



Normalizing Flows



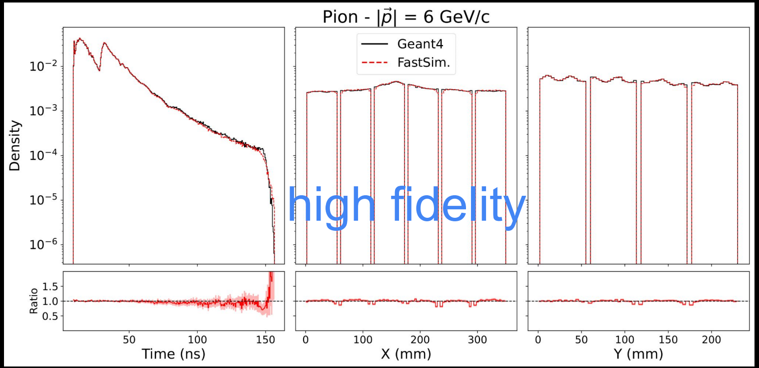
Diffusion Models

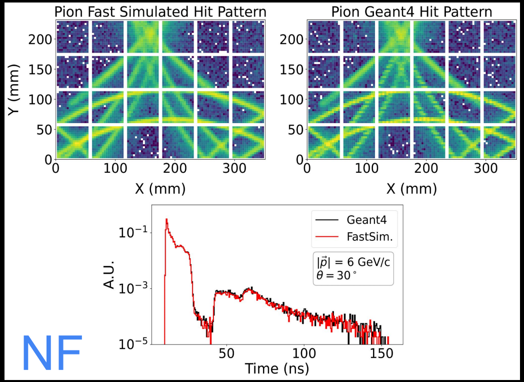
Fast Simulation - hpDIRC in ePIC

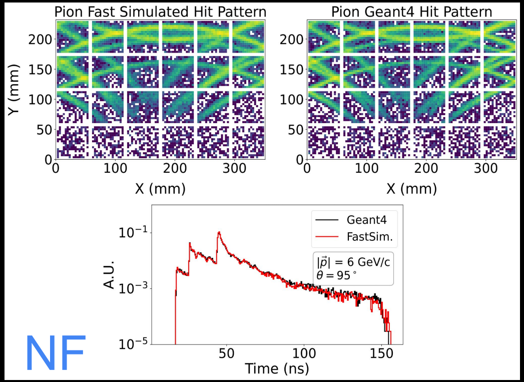
J. Giroux, M. Martinez, and CF. "Generative Models for Fast Simulation of Cherenkov Detectors at the EIC." *2025 Mach. Learn.: Sci. Technol. 6 040501*.

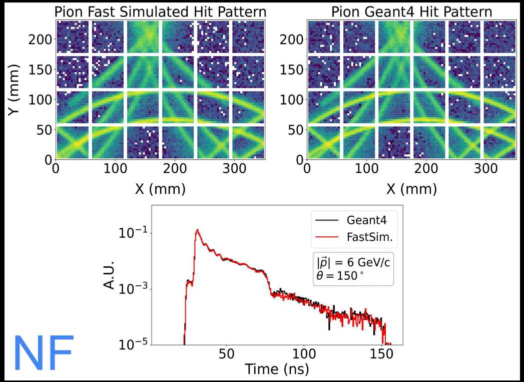
Architectures: Normalizing Flows (NF), Continuous Normalizing Flows (CNF), Conditional Flow Matching (CFM), Denoising Diffusion Probabilistic Models (DDPM), Score Based Generative Models (SB)

- **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







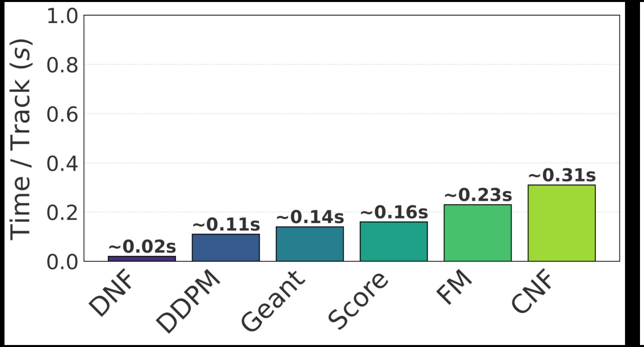


(hpDIRC standalone sim)

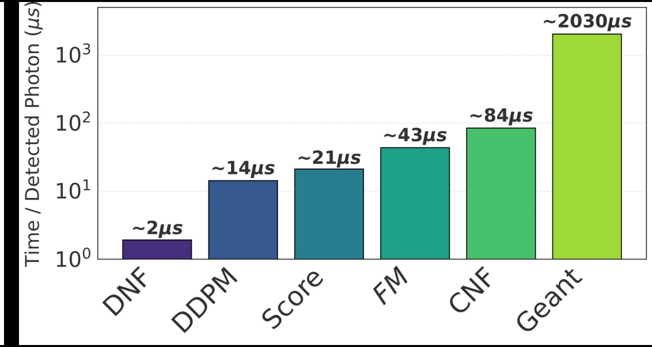
Simulation is fast - O(0.5)μs per hit (effective)

Fast Simulation - hpDIRC in ePIC

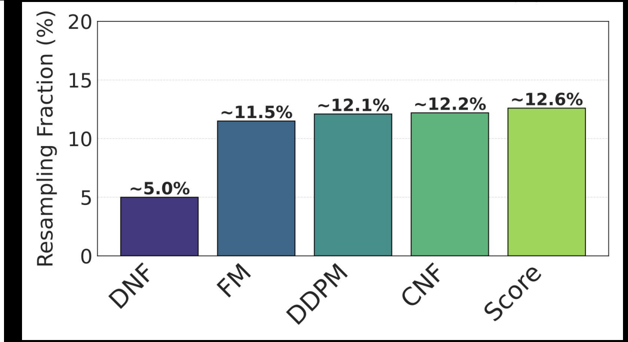
- Ring and time structures follow correct kinematic dependencies for both particle types (π/K)
  - See paper for more in depth evaluation
- We have created an open source suite of SOTA algorithms for the hpDIRC (easily adapted to other detectors)
- Our fast simulation is self-contained, fast and capable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



Track Generation (CPU)



Photon Generation - Large PDFs (GPU)

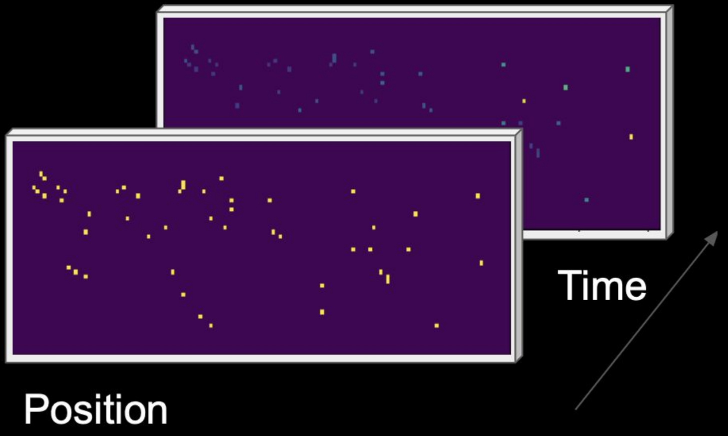


Resampling Fraction

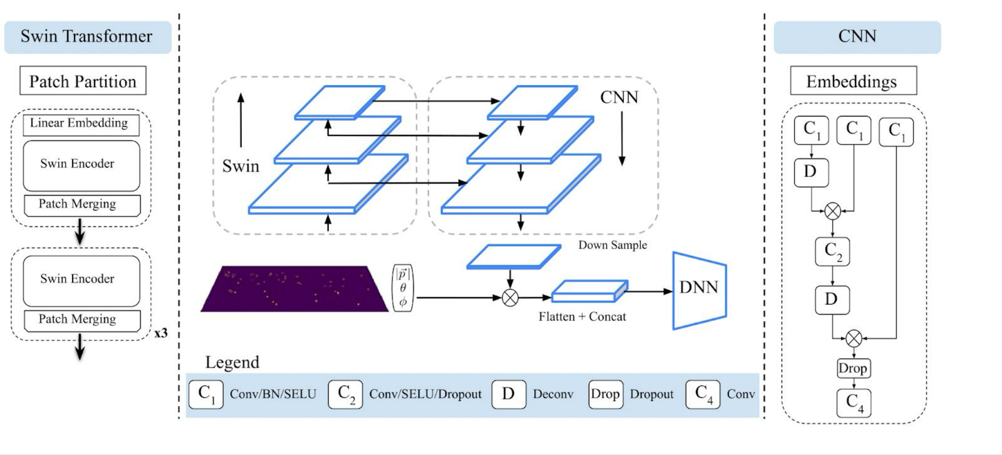
2) Particle Identification

Github

Transformer-based PID (1st method)



CF, J. Giroux, J. Stevens. "Deep(er)RICH"
Machine Learning: Science and Technology 6.1 (2025): 015028.



- 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

- 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"
Machine Learning: Science and Technology 6.1 (2025): 015028.

- Recall our bijection

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

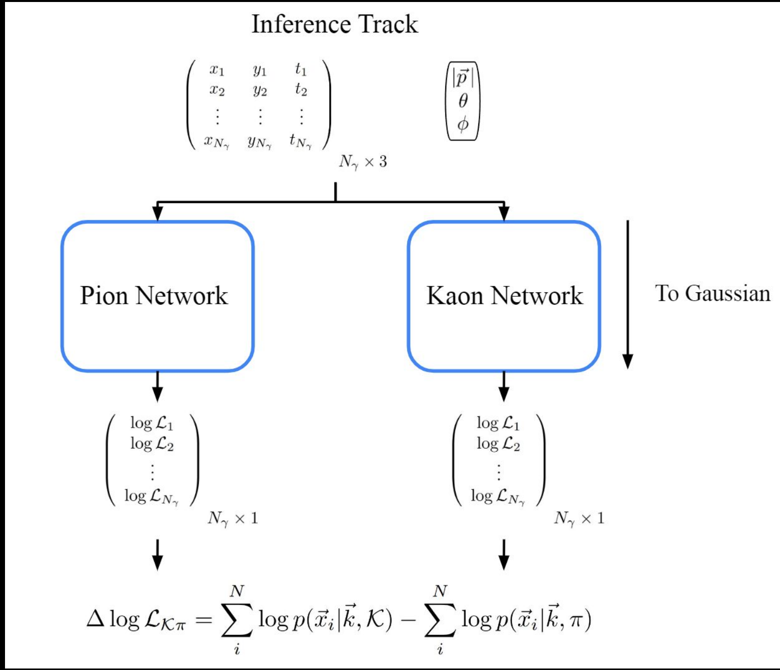
- Recall our analytical computation of the likelihood under a change of variables

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

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

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

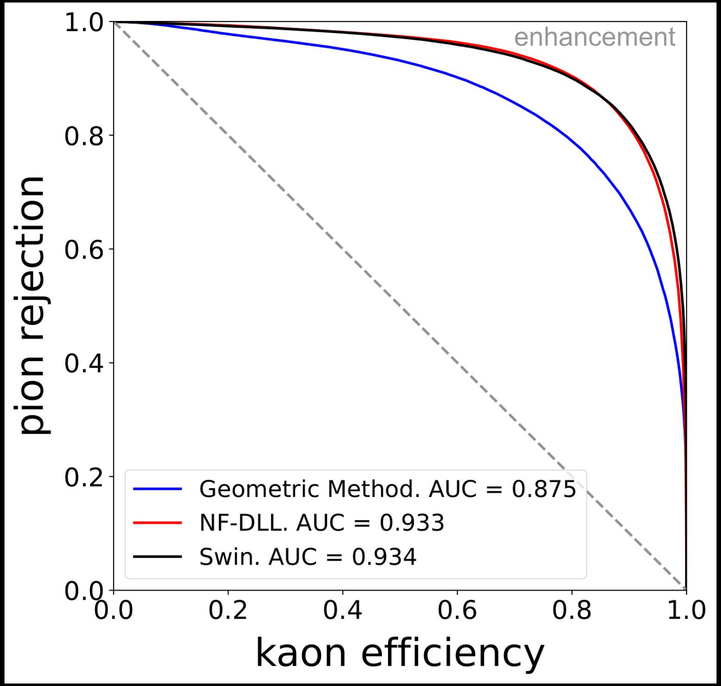
—the hypothesis of π/K represented by individual networks—

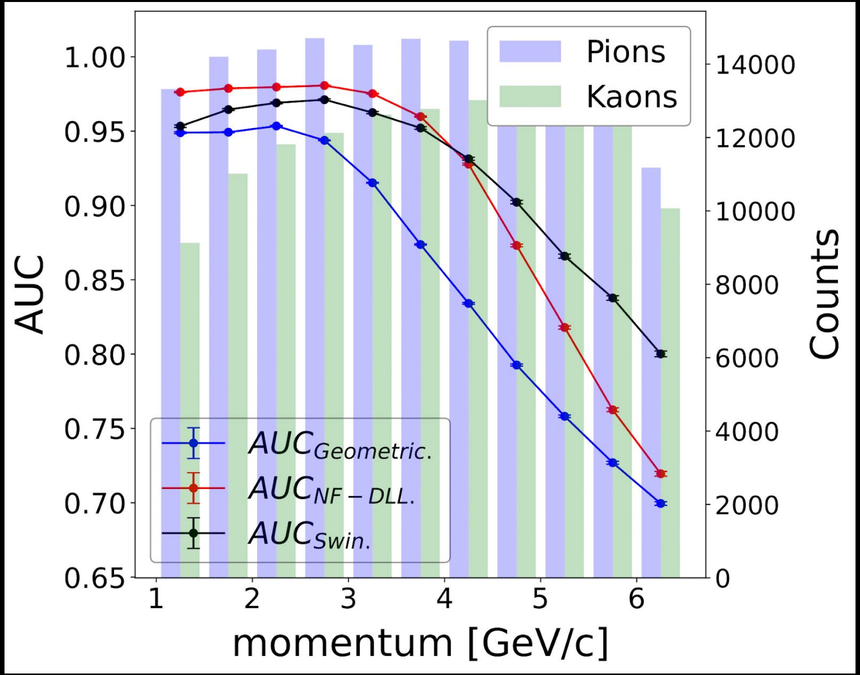


- Normalizing Flow - Likelihood based PID
- PID hypotheses represented through independent models
- Analytic likelihood computation from NF in base distribution
- Compute Delta-Log Likelihood

PID Performance - GlueX

CF, J. Giroux, J. Stevens. "Deep(er)RICH"
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.

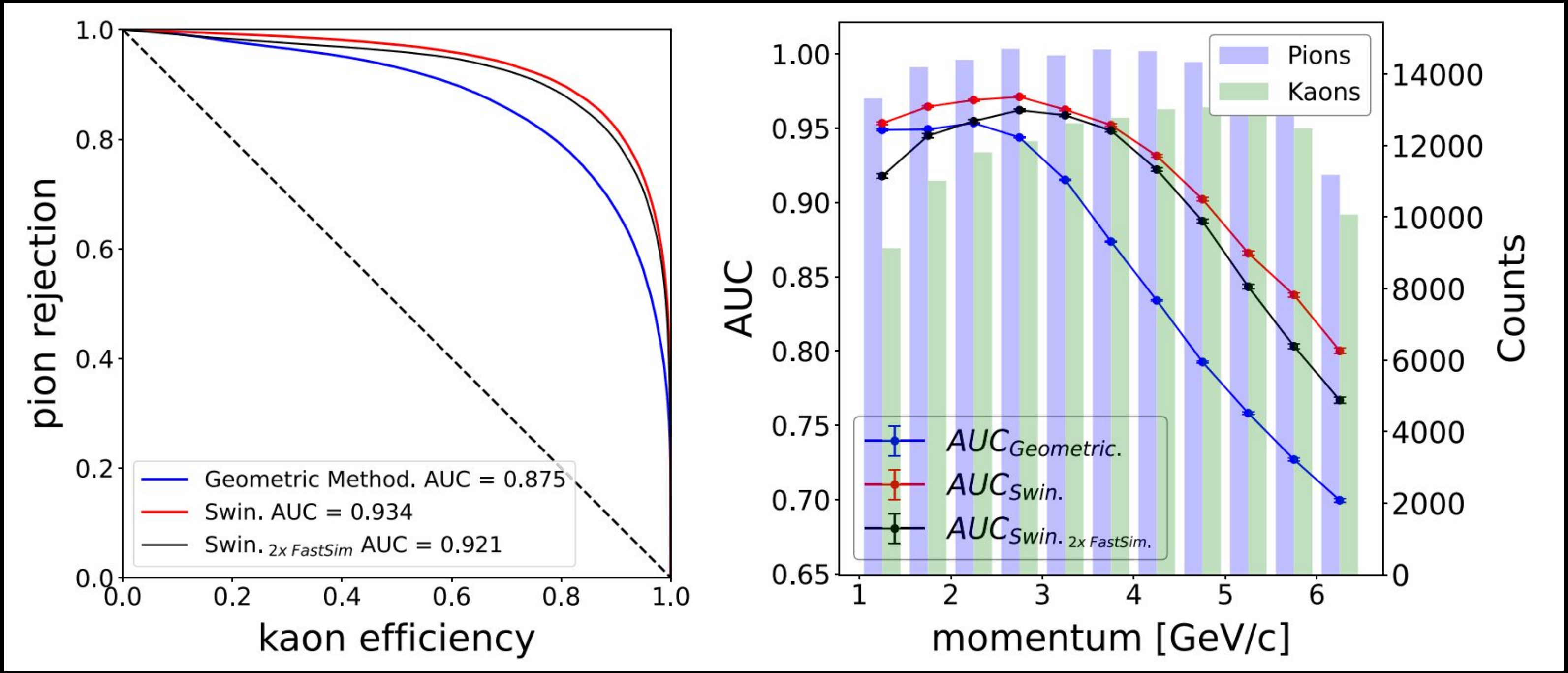
(GlueX DIRC sim)

Fidelity of Fast Sim

How to measure fidelity of synthetic data?

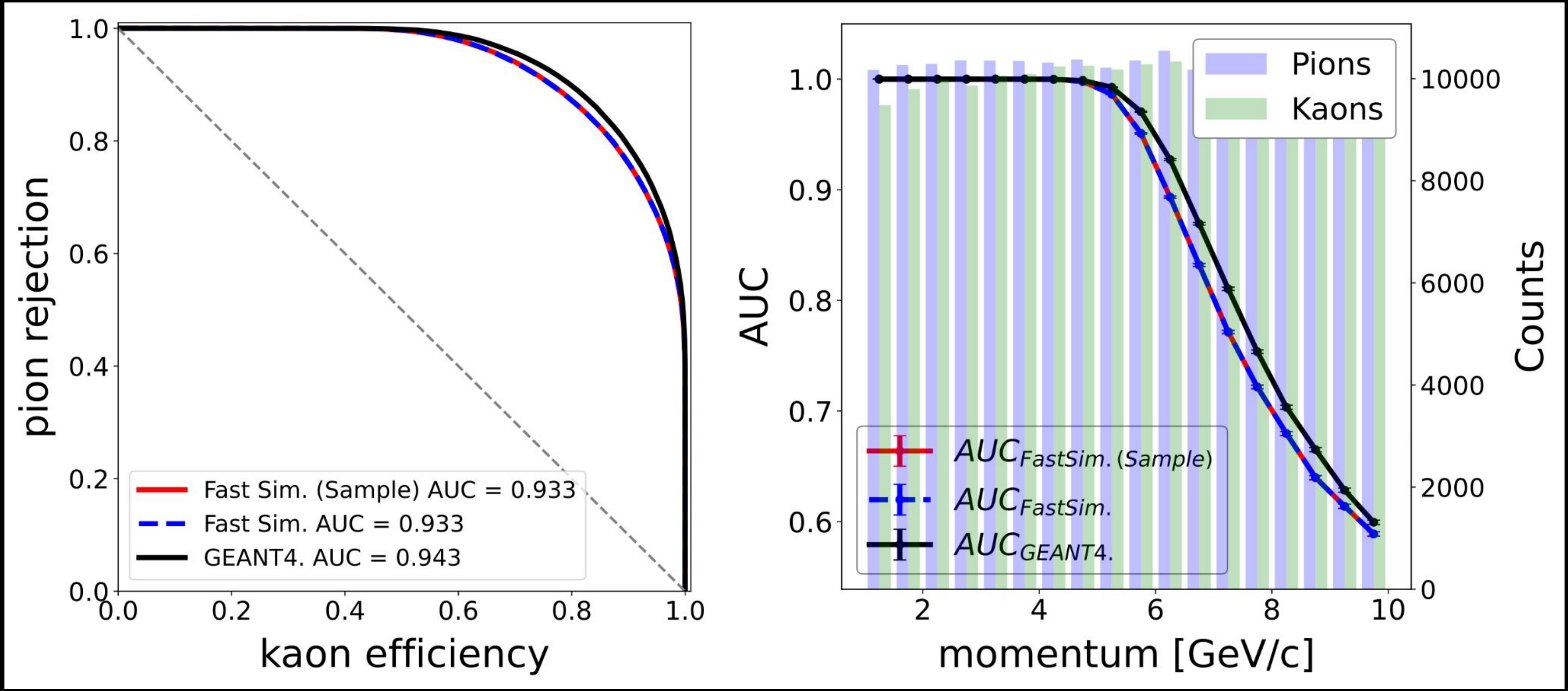
- Fidelity of synthetic data is typically evaluated using specific metrics. These metrics are meaningful primarily in a relative sense—they help compare the quality of different datasets but do not provide an absolute measure of fidelity.
- In these slides, we made use of 1D ratio plots comparing synthetic distributions to the real (true) ones.
- For a synthetic dataset related to a PID detector, we also apply a PID classifier to assess if the PID performance leveraging synthetic data are the "same" as if using more advanced 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.

Fidelity of Fast Sim - hpDIRC in ePIC



- An independent classifier (CNF) is trained on (i) fast simulated samples with photon yields matching those from Geant4; (ii) fast simulated samples with photon yields derived from our LUT approximation; (iii) independent Geant4 sample

3) (Proto-)Foundation Model

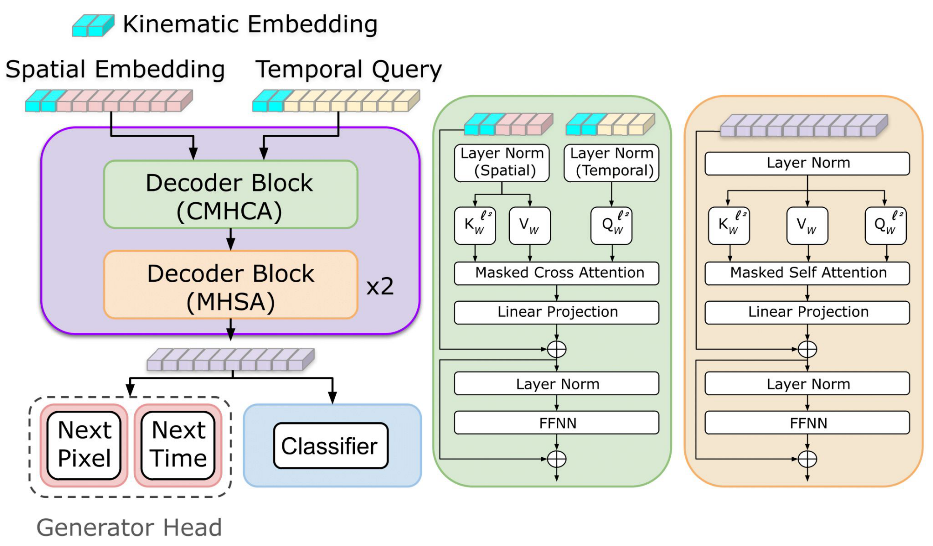
Github

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 CF "Towards Foundation Models for Experimental Readout Systems Combining Discrete and Continuous Data." *arXiv:2505.08736* (2025).

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

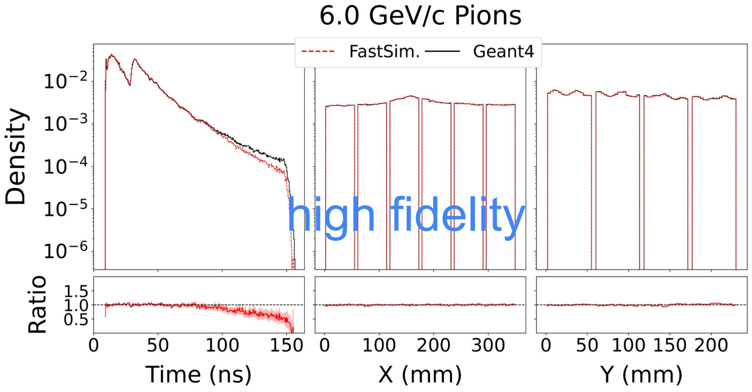


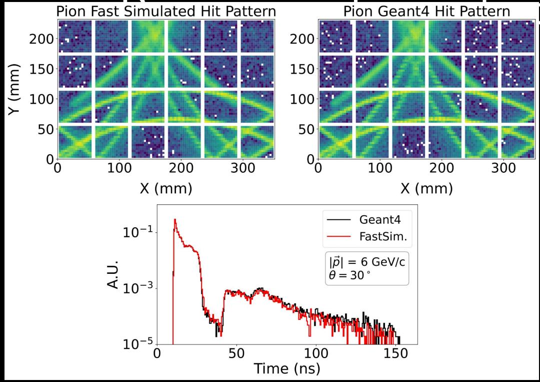
Github

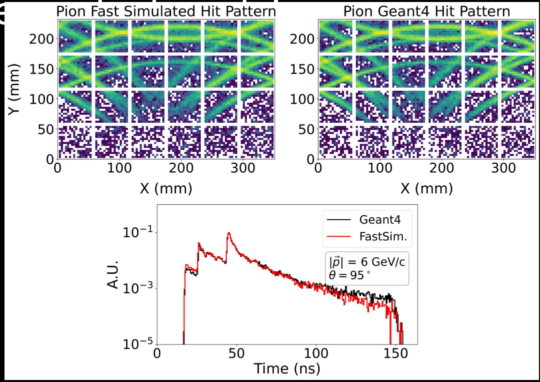
All code is open source and pre-trained models are provided.

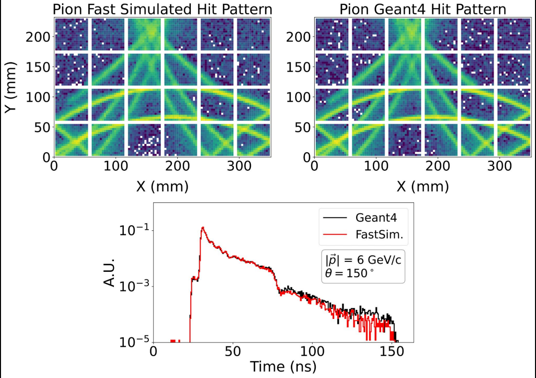
Foundation Model - Fast Sim

- 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)









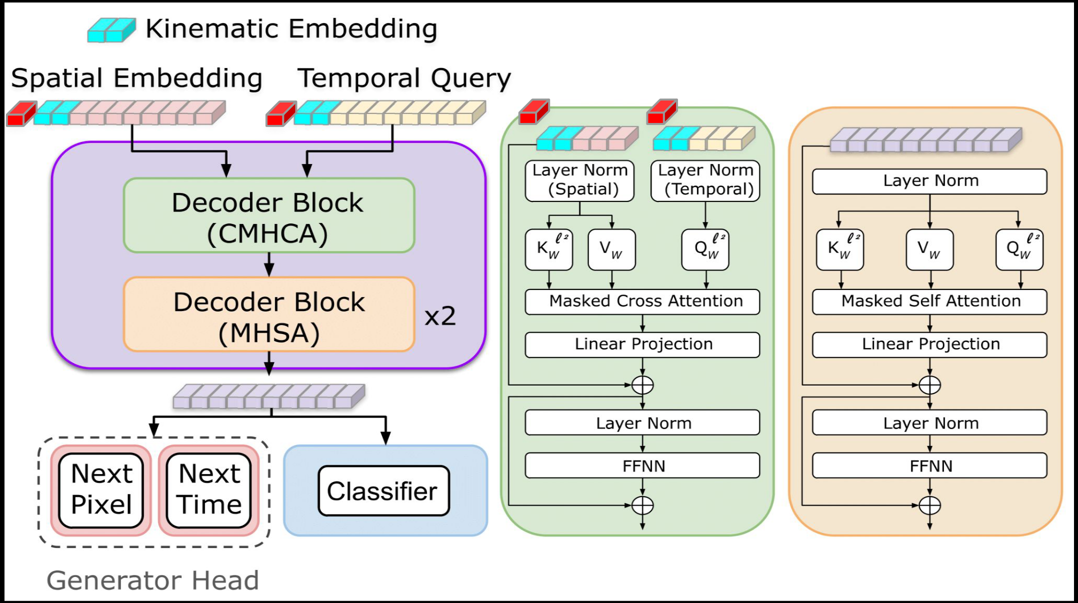
Simulation is fast - O(0.02)s per track (effective)

(hpDIRC standalone sim)

Foundation Model - From Sim to PID

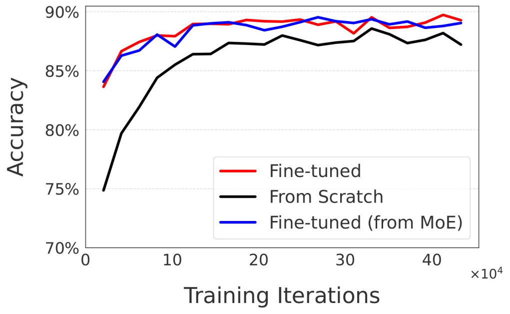
- Our model also supports classification (π / K)
  - Additional token - ***CLS Token***
  - Remove causal masking
  - Can be **fine-tuned**

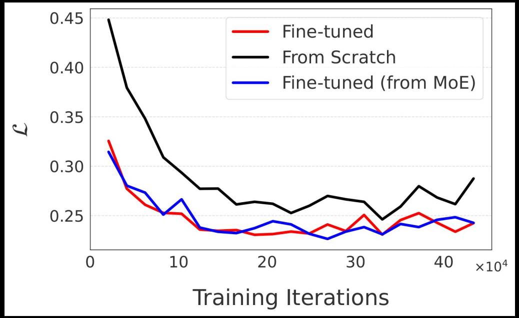
CLS Token

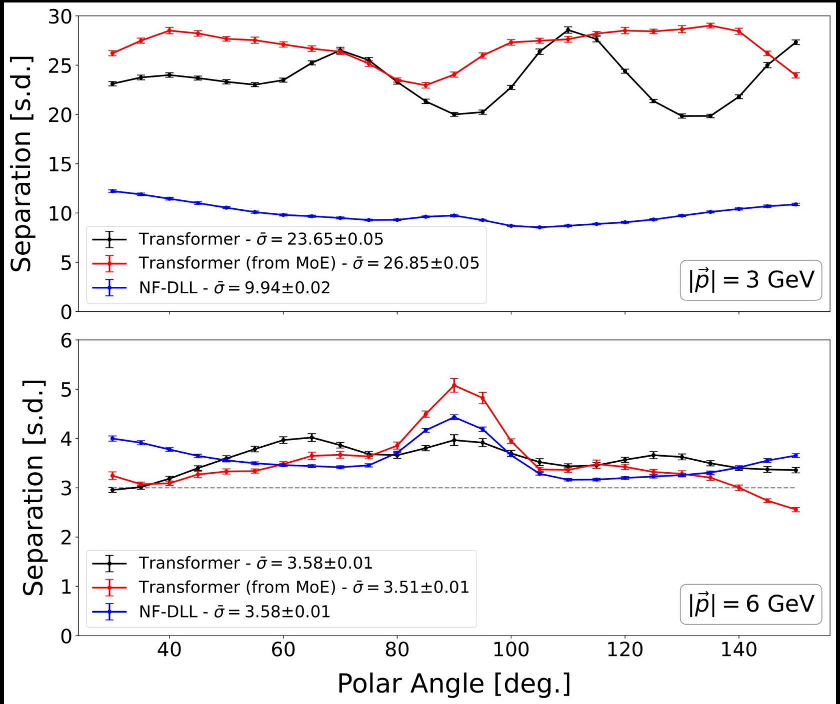


Foundation Model - PID

- 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







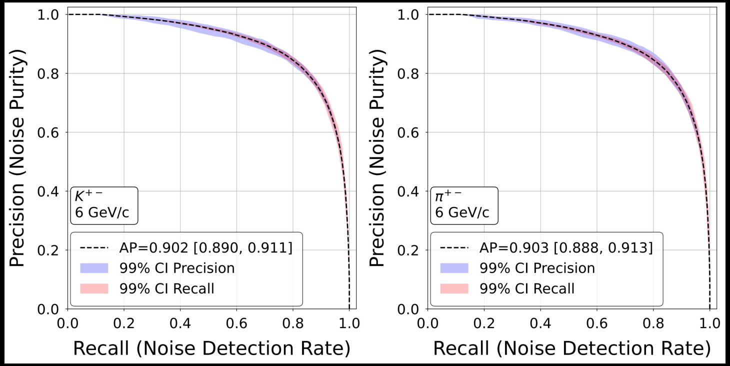
Fine-tuned models are initialized with weights from a separately trained generative model (e.g., trained on π or K);
“from MoE” models are initialized using weights from a multi-expert architecture with 4 experts

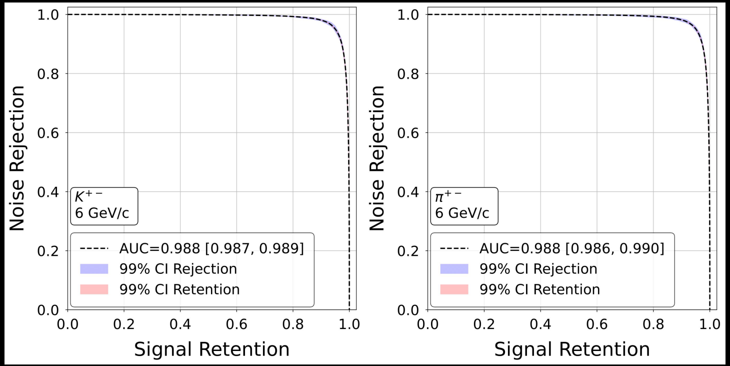
(hpDIRC standalone sim)

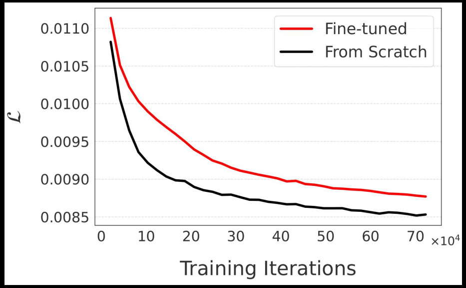
Foundation Model - Noise Filter

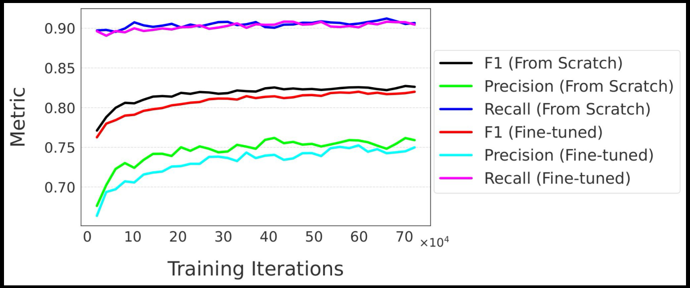
Noise filtering (proof of principle)

- 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)

Conclusions

Conclusions

- Simulation
  - Huge speedup over Geant4 (~100× faster on CPU for tracks; ~1000× on GPU for full PDF)
  - Usable by any user without GPU for track generation; GPU recommended for PID (see next point)
  - Can support novel hybrid PDF-based reconstruction methods (e.g. time-imaging) with PDFs generated on-the-fly
- Particle Identification
  - DL methods (e.g., SWIN transformer) outperform benchmark Geometric LUT (GlueX results; hpDIRC preliminary)
  - Compute-wise, LUT is cheap — but DL approaches are also efficient. If PID matches or outperforms LUT and is validated on real data, DL methods could increasingly replace classical ones (see also opportunities)
- Foundation Model
  - Unified architecture: bulk of model remains identical, only final layers differ
  - More computationally intensive than traditional approaches → GPU required
  - PID remains fast and cost-effective
  - Other tasks such as noise filtering possible
- Opportunities
  - FM generalization to other experiments and to other Cherenkov detectors (and beyond Cherenkov)
  - Deep learning enables learning detector response if directly trained on real data, reducing reliance on detailed simulation and improving realism for PID and other tasks such as alignment, calibration

Backup