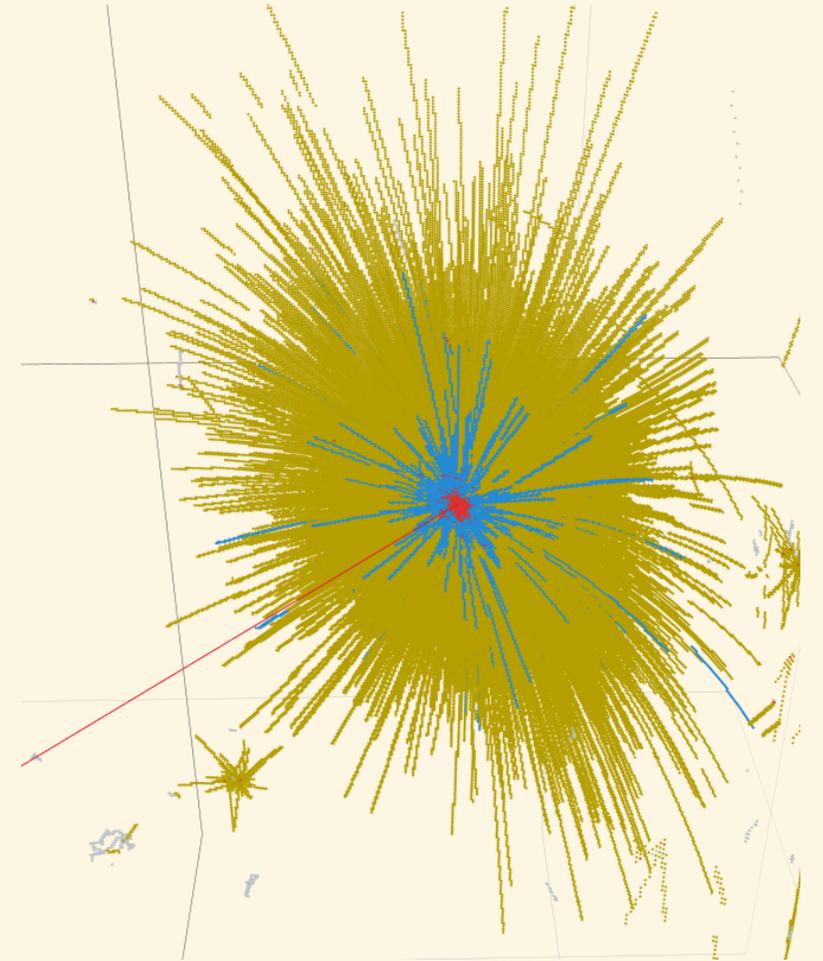
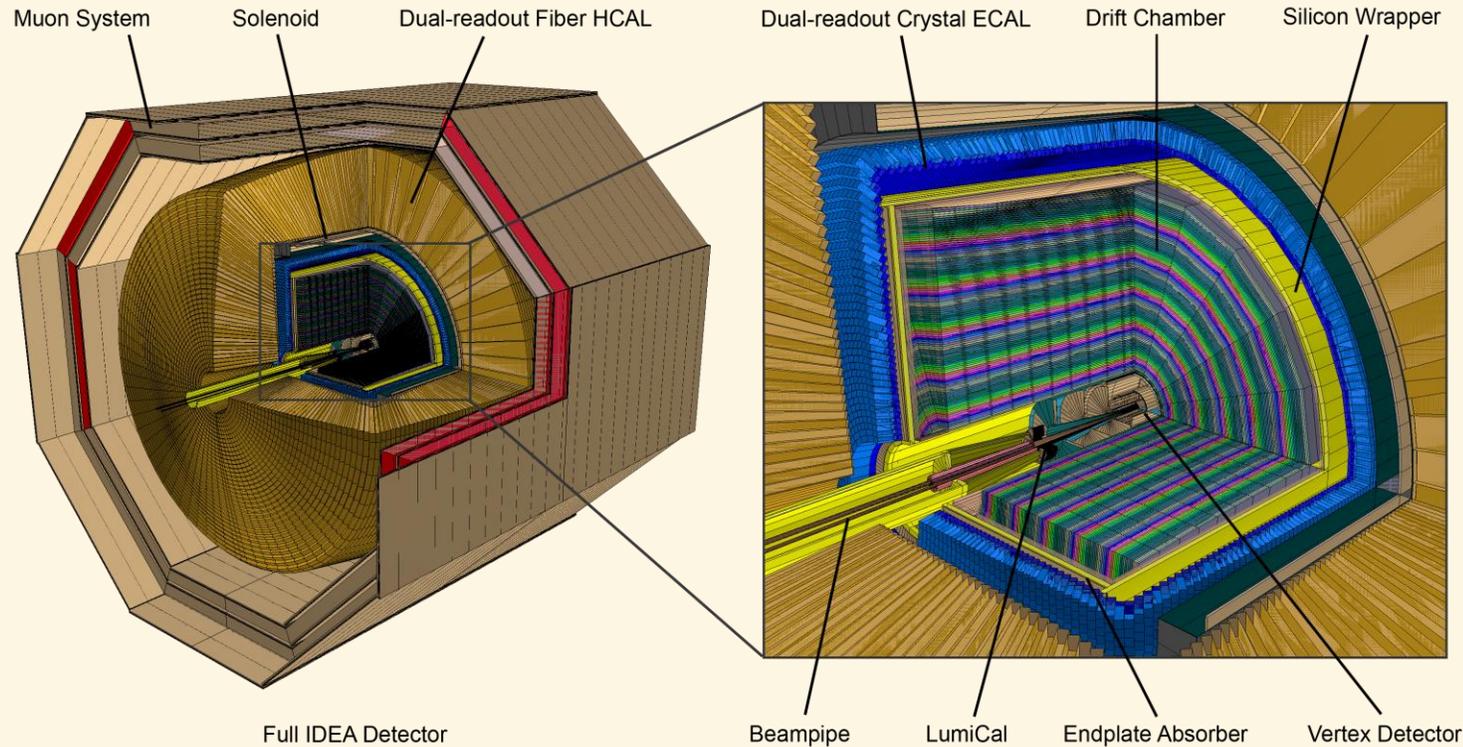


Synthetic Training and Representation Bridging in Reconstruction Domains

[\[2505.05664\]](#)



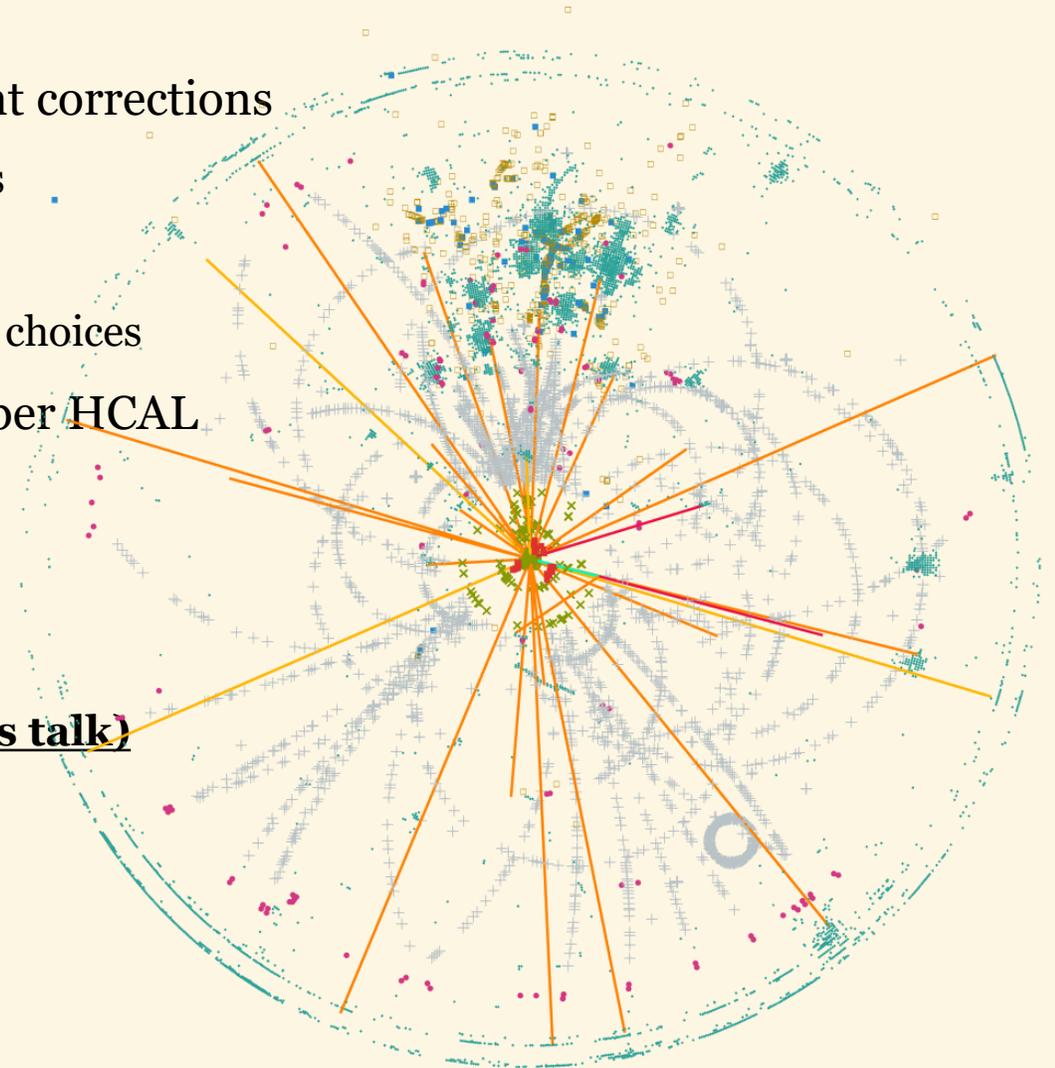
Wonyong Chung
Princeton University

August 2025
Lepton-Photon Symposium

IDEA for FCC-ee: Precision by Design

Precision physics \rightarrow detector performance drivers \rightarrow design choices

- Beyond resolution: design for systematic control and per-event corrections
 - Z, ZH – systematics and reconstruction are the limit, not statistics
 - Accelerator/MDI rewards linearity, timing, calibration control
 - CLD, IDEA, ALLEGRO illustrate calorimetry-driven architectural choices
- Hybrid dual-readout calorimetry: segmented crystal ECAL, fiber HCAL
 - Qualitatively new observables, event-by-event compensation
- Full simulation in key4hep, digitization in progress
 - Now at IRIS-HEP: Bilevel optimization framework
 - **Blue-sky AI/ML with synthetic data/representations (this talk)**
- Performance headlines:
 - $\sim 3\text{-}4\%$ at >50 GeV jets
 - Hadronic $\sim 26\%/\sqrt{E}$
 - EM $\sim 3\%/\sqrt{E}$



Precision physics → detector performance drivers

- Run points: Z(91 GeV), WW(161 GeV), ZH(240–250 GeV), tt(~ 365 GeV)
- Tera-Z ($O(6 \times 10^{12})$ Z) → 100-200 kHz events, comparable to L1 rates at HL-LHC
- \sqrt{s} dimuon kinematics, ISR/FSR control → superb tracking, angular resolution (~ 0.1 mrad), photon tagging/recovery
- Heavy-flavor program: soft photons → ECAL resolution, π^0 separation
- HZ recoil mass → bremsstrahlung recovery, transparent MDI+tracker
- Electron Yukawa at Higgs pole via monochromatisation → $\delta\sqrt{s}$ vs. integrated L, demands stable EM scales
- ALP/LLP, τ /rare-B signatures → photon angular/energy precision
- 30 mrad crossing-angle → global acceptance calibrations, mechanical tolerances, endcap geometry
- 2 T magnetic field limit to minimize emittance → larger tracking volume, shallow ECAL inside coil
- Continuous ~ 50 MHz crossings, front-end pile-in/time-walk → high S/N, precision timing
- Particle-flow quality → detector segmentation

Precision physics → detector performance drivers

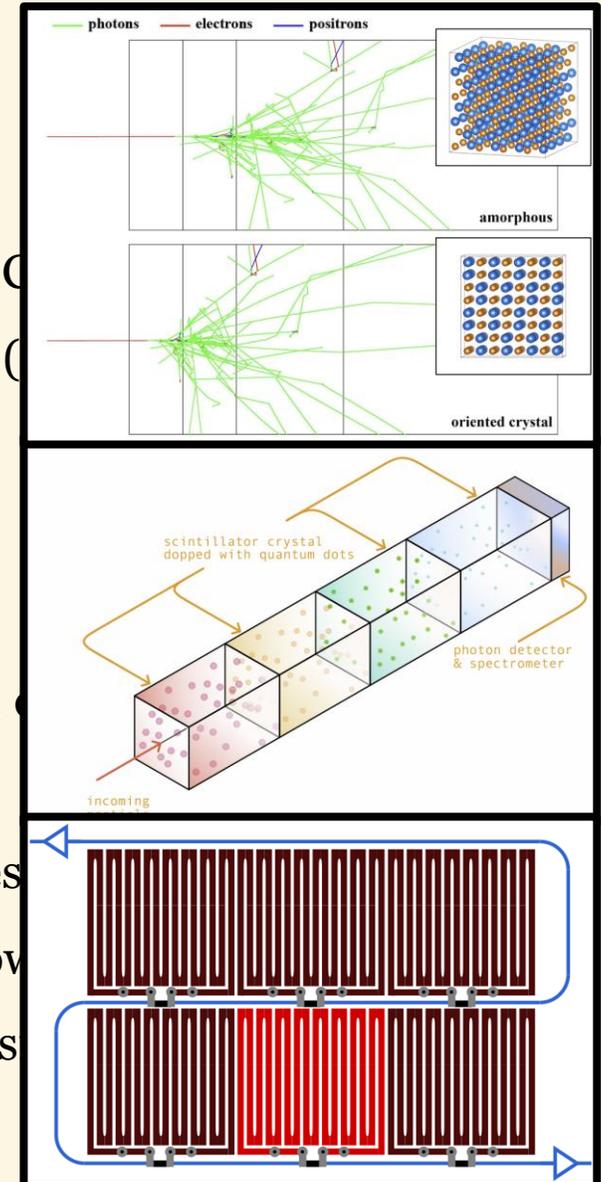
Detector design in the AI/ML era

- Today: modular, flexible simulation
- Unified, top-down view of geometry and data schemas
- Hot-swappable subdetectors
- Triggerless DAQs
- Real-time inference on ASICs
- Picosecond timing
- **1:1 reconstruction**
 - **Motivates: Information-based detector perspective**

New technologies

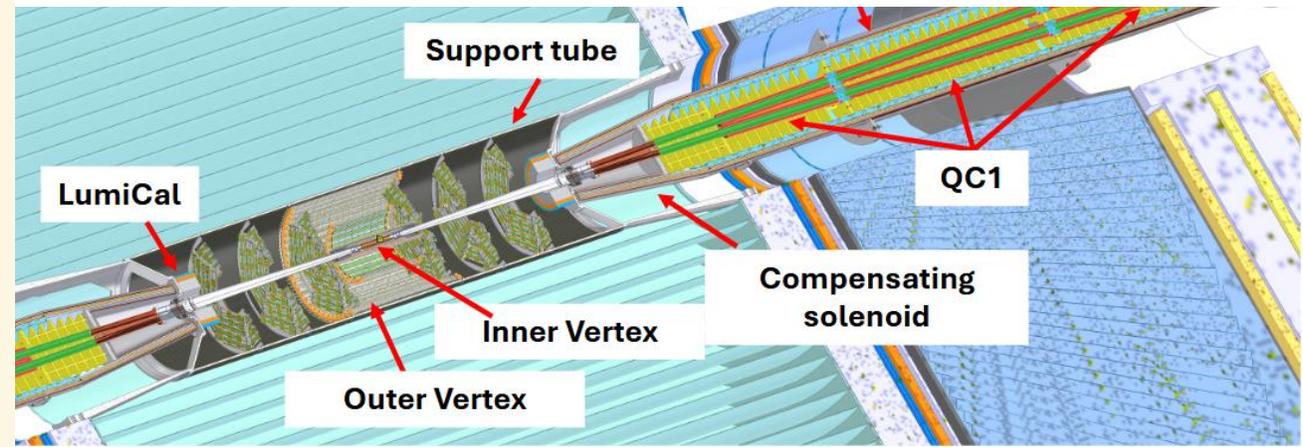
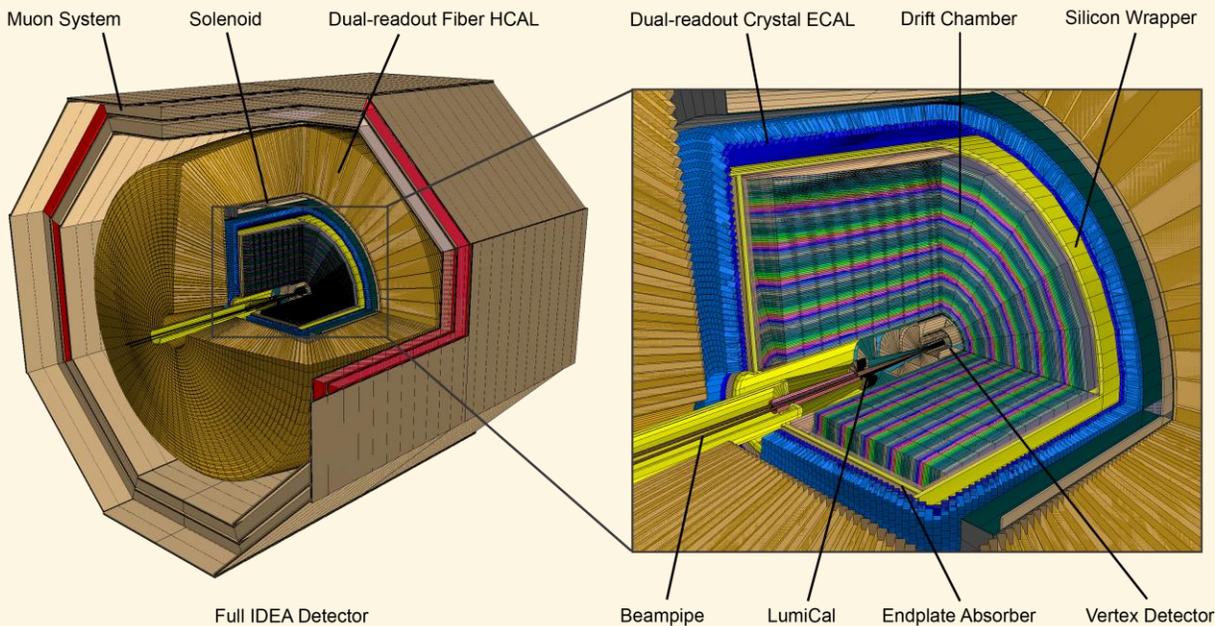
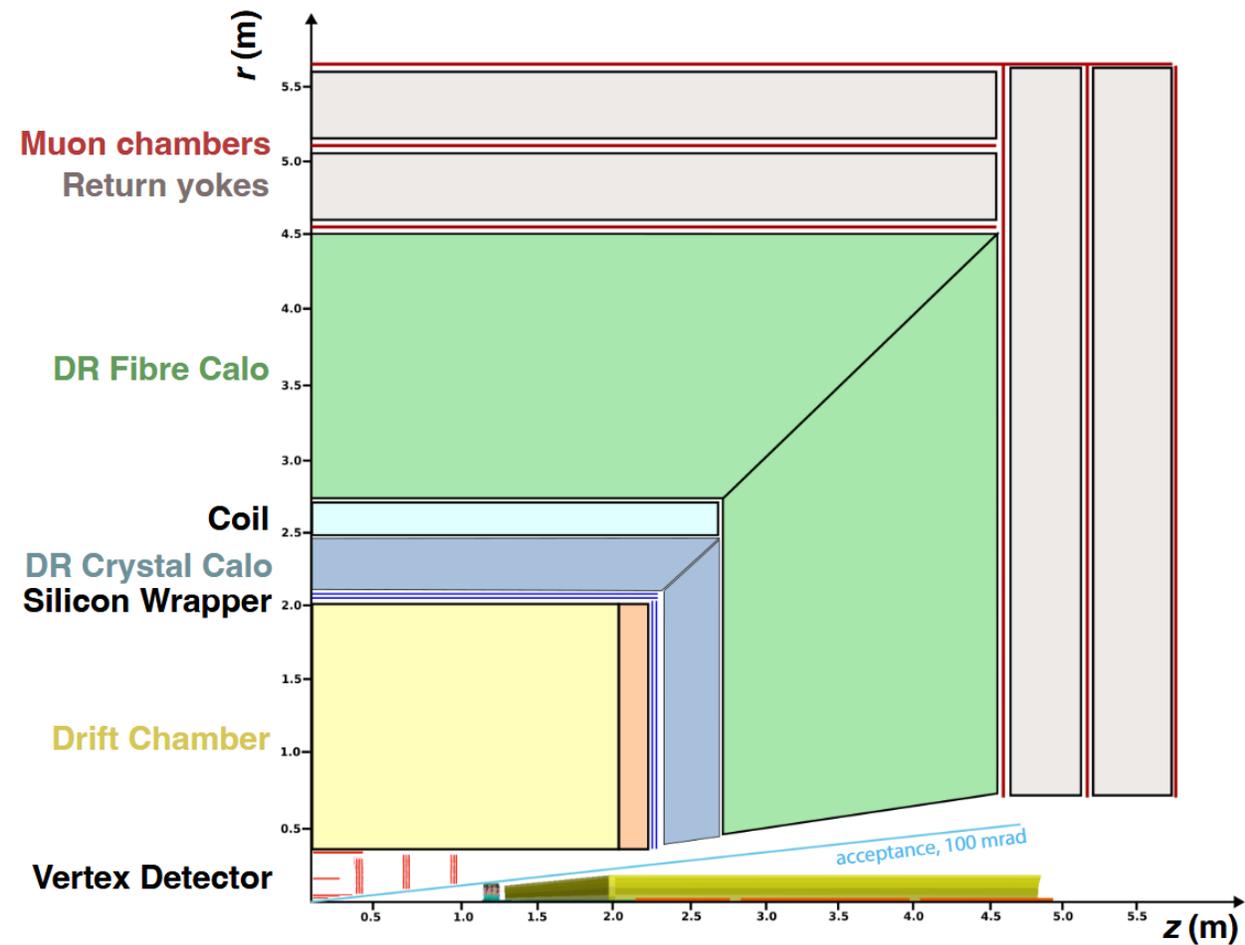
- Lattice-oriented crystals
- Chromatic calorimetry
- SNSPDs, etc.

Can simulation-to-reality squeeze more information from detectors?



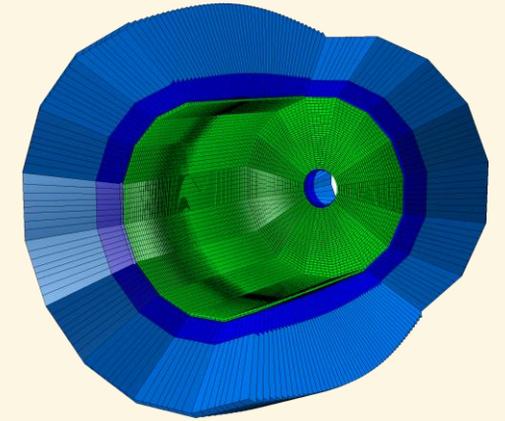
IDEA current baseline

- Vertex, ultra-light drift chamber, silicon wrapper
- Dual-readout segmented crystal ECAL
- Thin, low-mass superconducting solenoid
- Dual-readout fiber HCAL
- μ -RWELL muon system in return yoke
- LumiCal in forward region

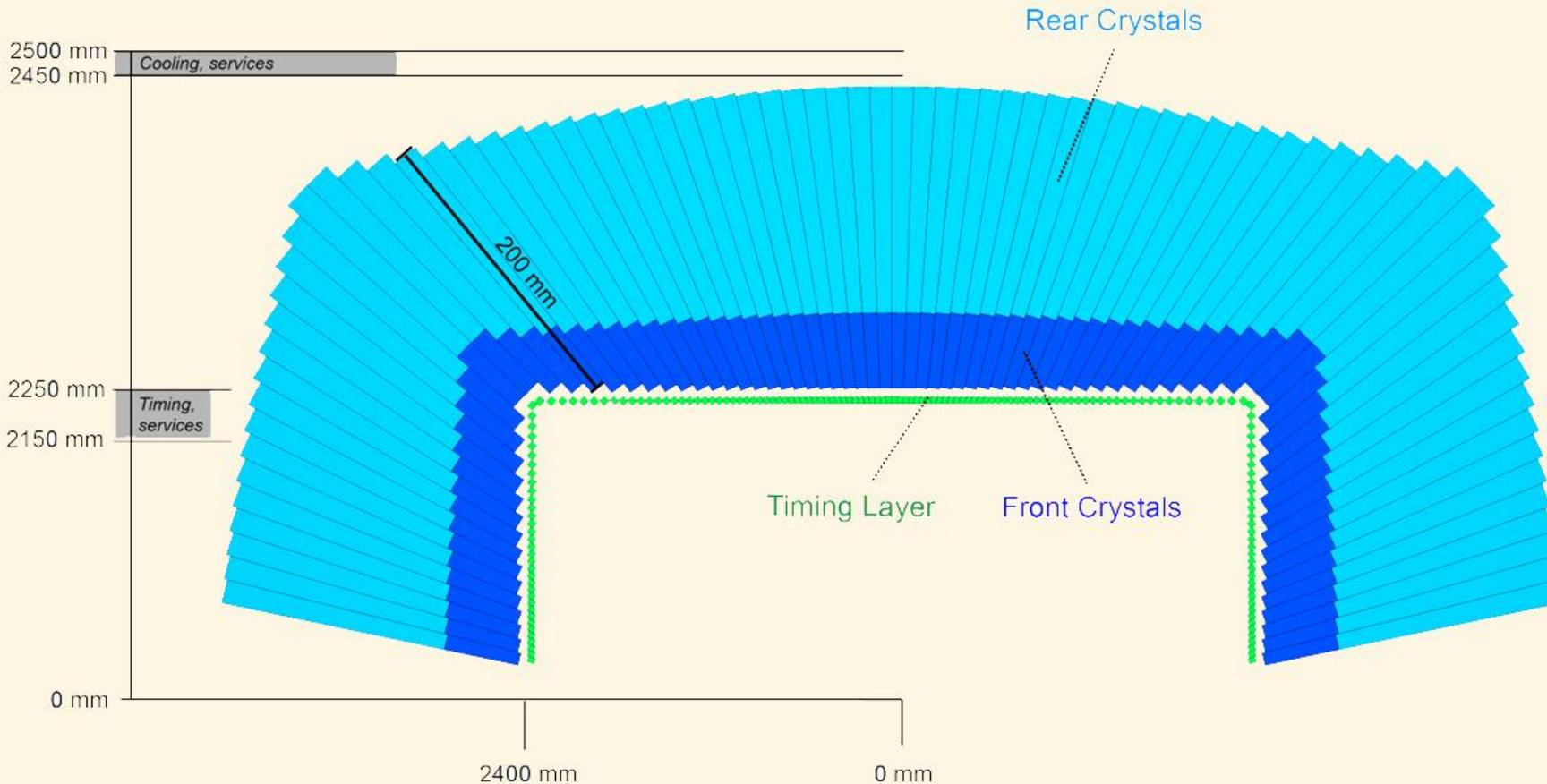
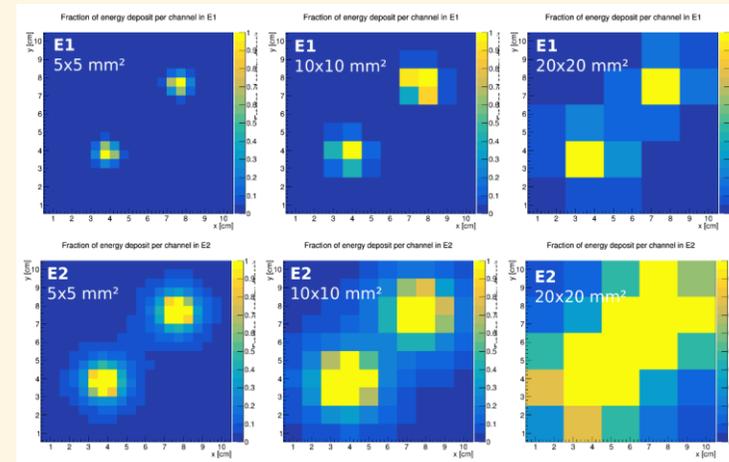
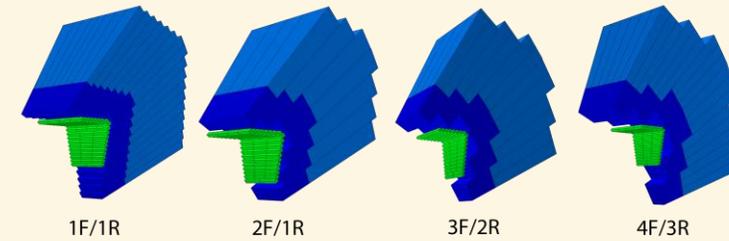


Segmented Crystal ECAL

- Baseline: PbWO_4 crystals $6 X_0$ (Front) + $16 X_0$ (Rear) = $22 X_0$
- 10×10 mm transverse granularity (barrel, endcap)
- $\sim 1 X_0$ fast-scintillating precision, projective timing layer (e.g. LYSO)
- Full simulation in key4hep, fully reconfigurable/differentiable geometry



Different Front/Rear Crystal Tower Divisions



Dual-Readout Calorimetry

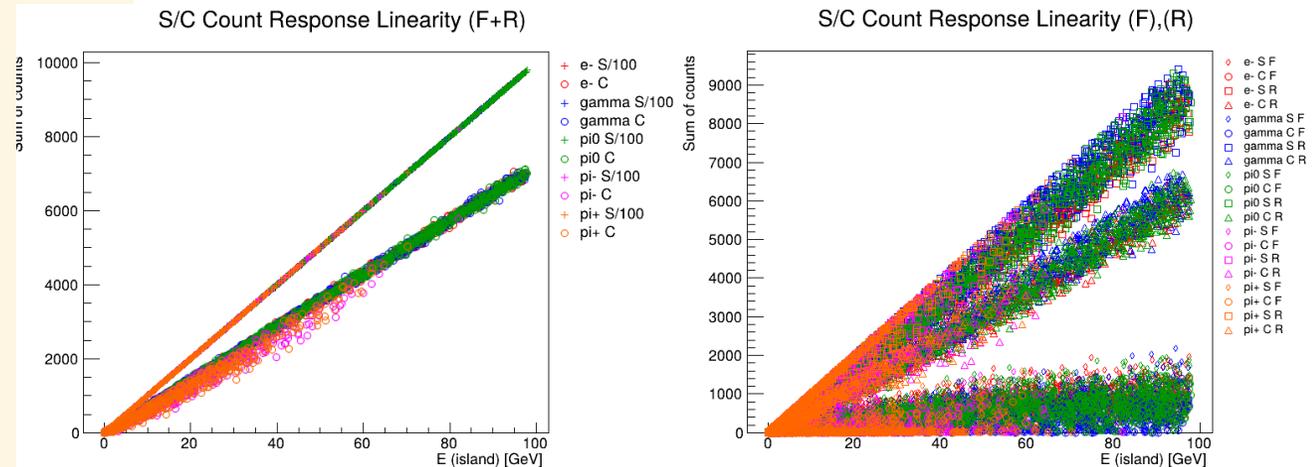
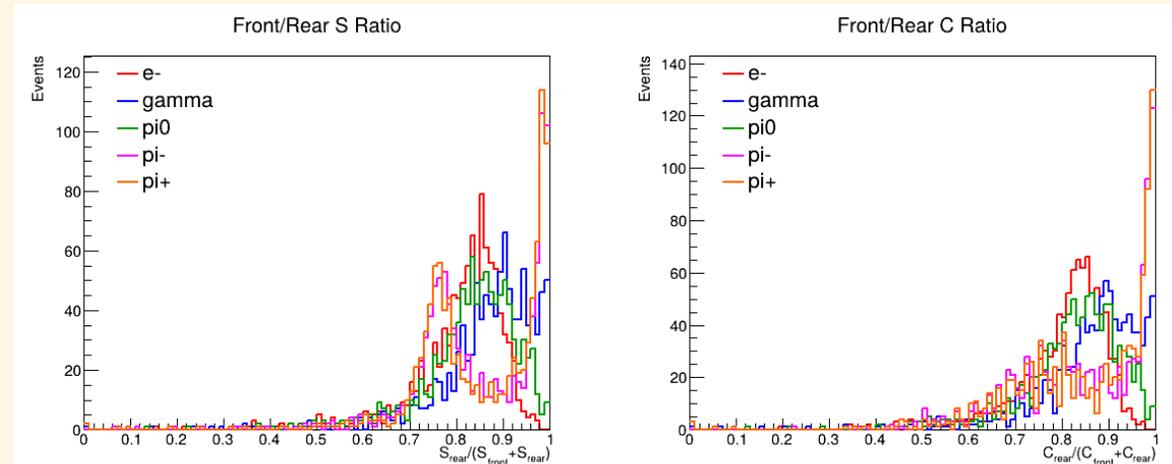
- **Calibrations:** 0-100 GeV e-, gamma, pi0, pi+, pi-
- **Technique:** Detect scintillation/Cerenkov light separately to mitigate event-by-event

fluctuations in hadronic showers

- **Procedure:**
 - Calibrate on known EM/hadronic physics processes
 - Obtain the S/C response scaling factors
 - Determine EM fraction event-by-event

$$\begin{cases} S = E \left[f_{EM} + \frac{1}{(e/h)|_S} (1 - f_{EM}) \right] \\ C = E \left[f_{EM} + \frac{1}{(e/h)|_C} (1 - f_{EM}) \right] \end{cases}$$

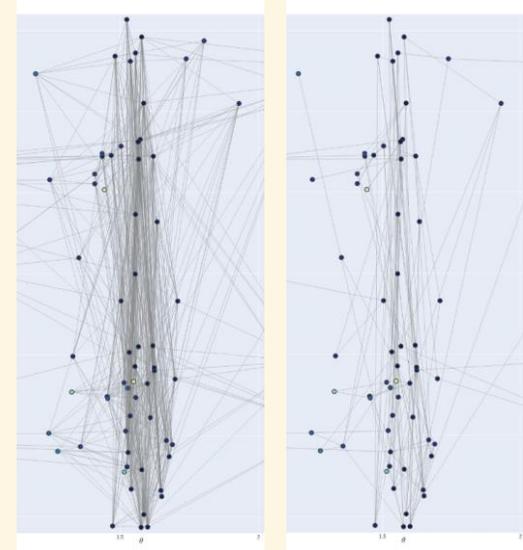
- **Segmentation** enhances separation power



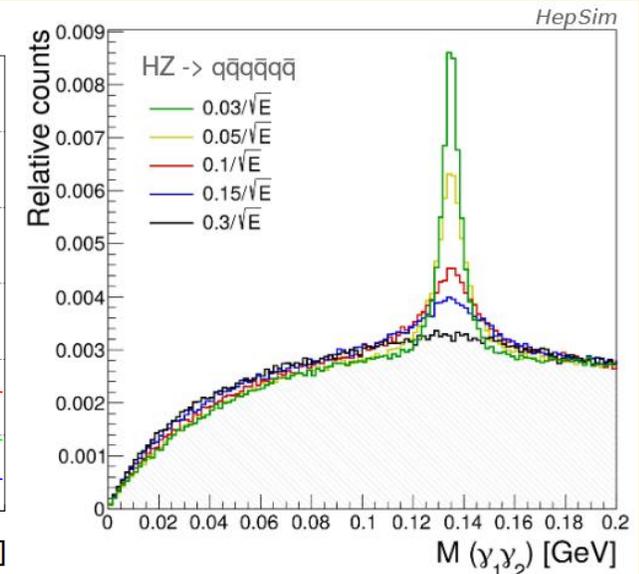
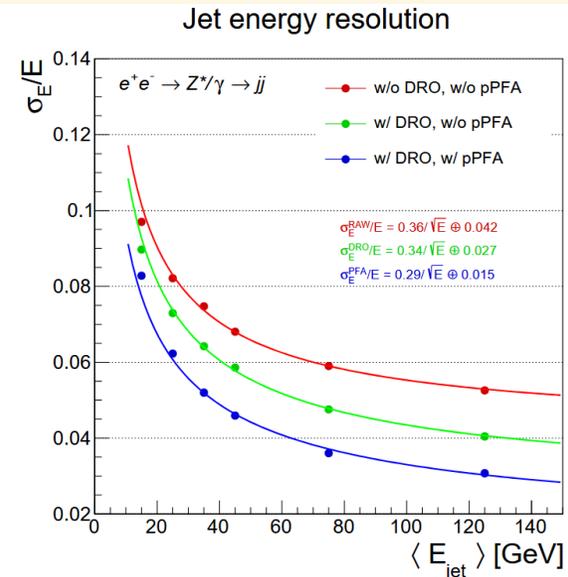
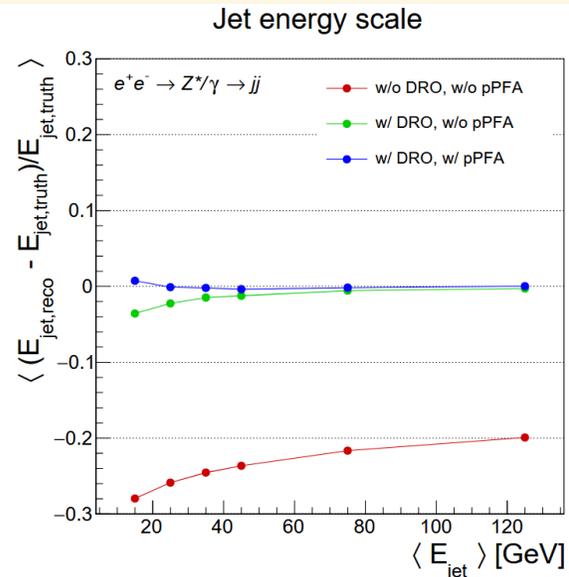
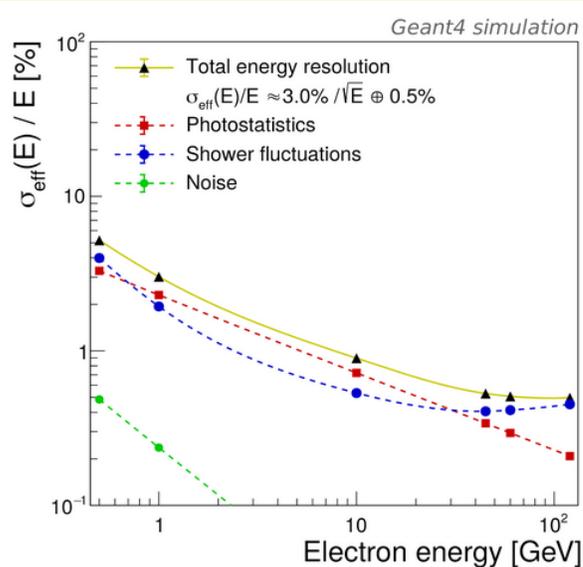
Dual-Readout Particle Flow

- Pandora/Arbor favor shower imaging, longitudinal segmentation → track-cluster topology
- DR-PF: compact showers, fine transverse segmentation, moderate longitudinal segmentation → emphasize energy, timing
- Detector-specific algorithms: e.g. combinatorial π^0 merging
- These studies will continue
- **High S/N, linearity** → **ripe for AI/ML**

15%/√E 3%/√E

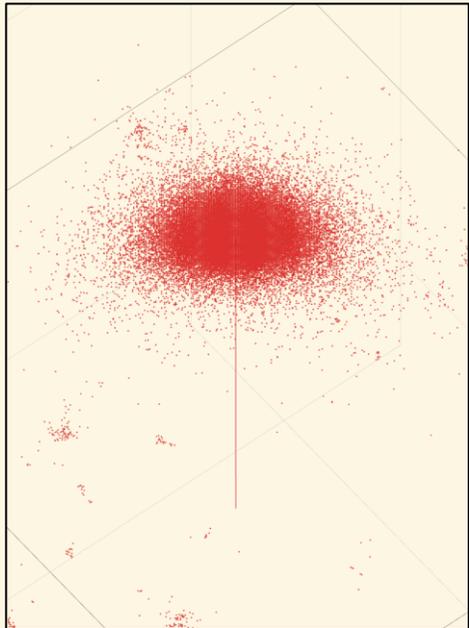


[[2008.00338](#), [2202.01474](#)]

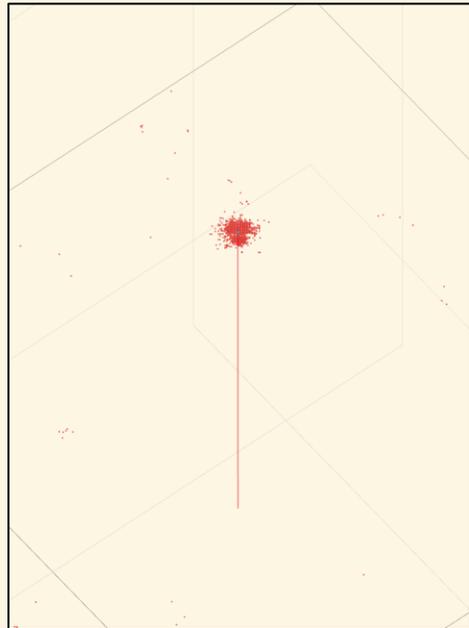


Synthetic Representations of Detector Response

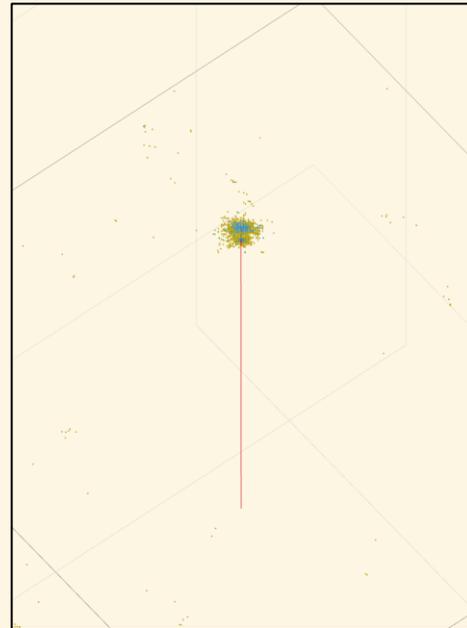
- Typical approach is to save hits based on **energy deposit** threshold (usually 1keV) at **step-level**
- For dual-readout (optical photons), apply **wavelength cuts** (300-600nm) at **track-level**
 - Save all hits for optical photons, even if energy deposit is zero – **simulated observables**
- Consider a 50 GeV electron:



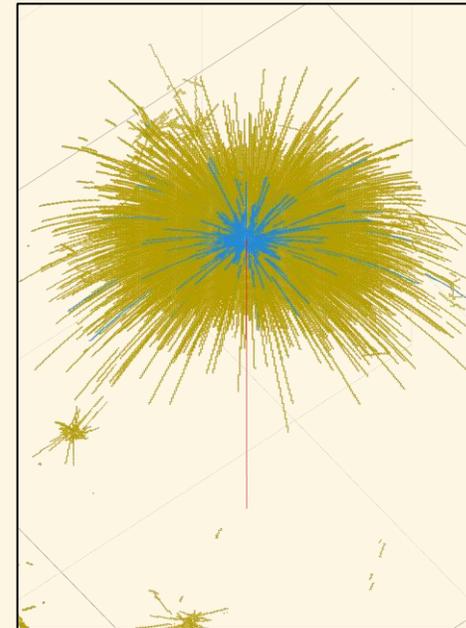
edep 0



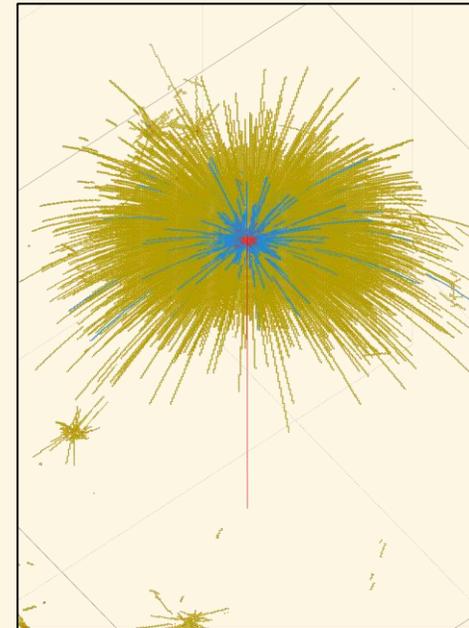
edep 1 keV



S/C counts



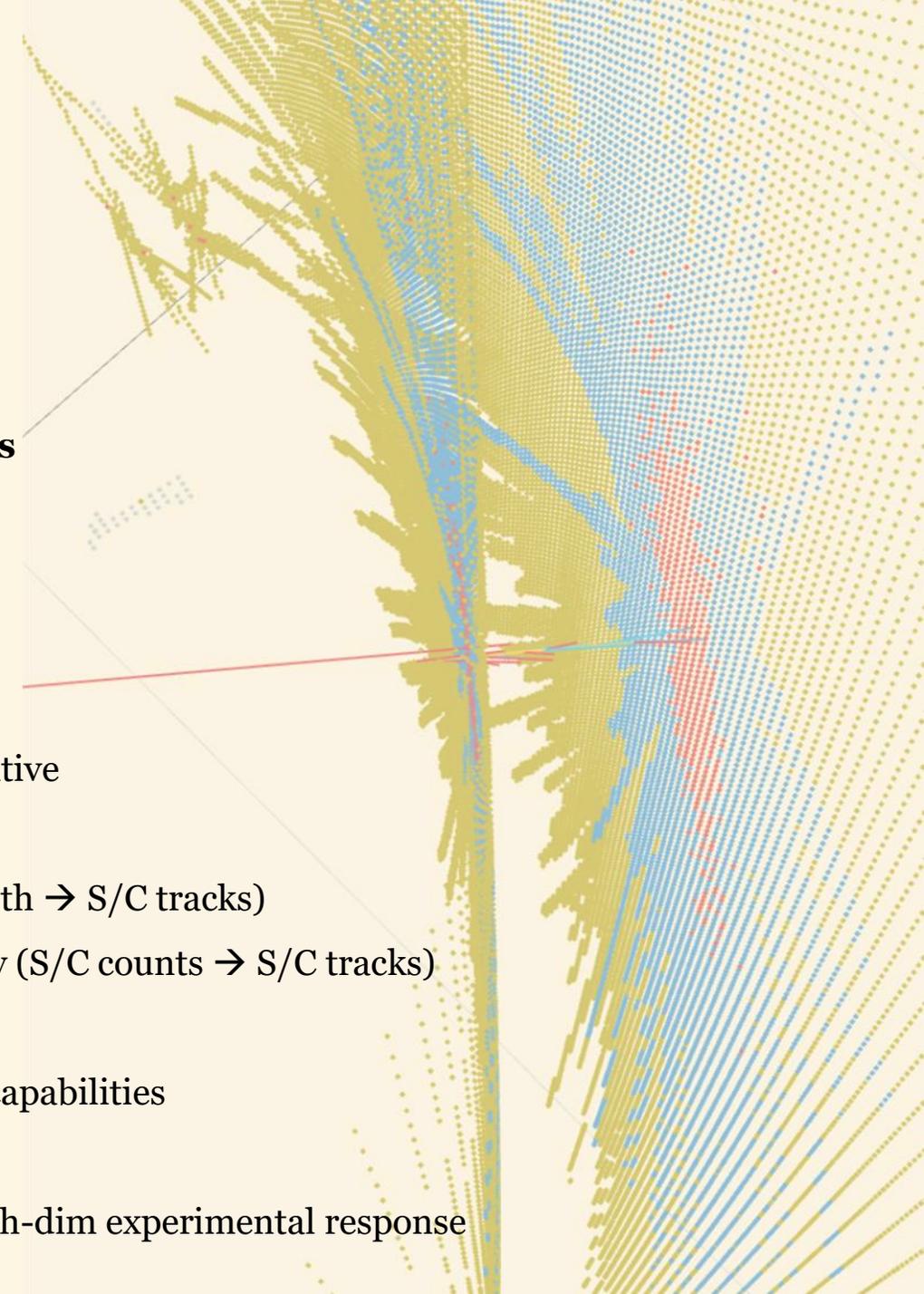
S/C tracks



S/C tracks+edep 1keV

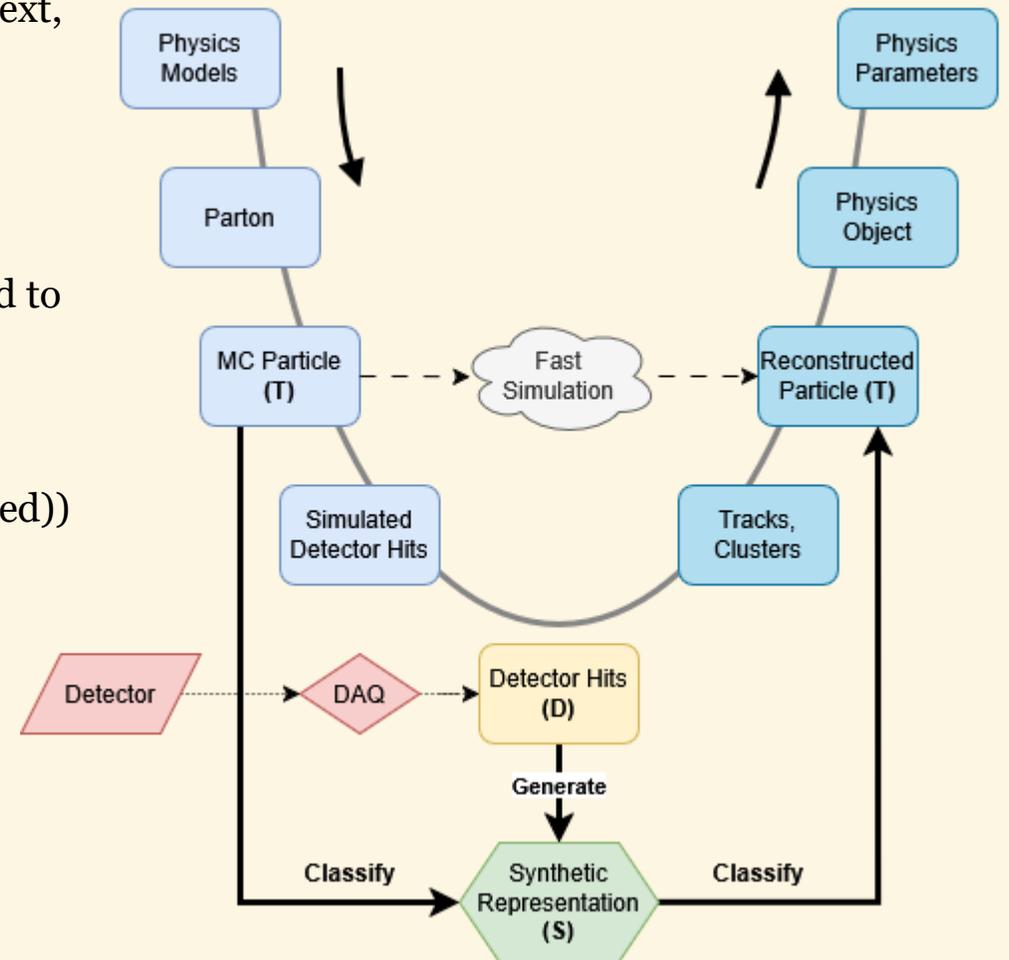
Representation Bridging in Reconstruction Domains

- S/C track hits are a form of **synthetic data**
 - *Unphysical* – won't ever see them in a real detector
 - *But not entirely unphysical* – are **representations of a physical process**
- Question: Can synthetic data be used in a meaningful way in reconstruction?
- Is there a need? *Yes* - a known problem: **domain bridging**
 - MC truth is low-dimensional - particle ID and momentum
 - Detector hits are high-dimensional - many, many hits
 - Compressing the phase space of detector hits to MC truth is highly degenerative
- **Idea:** Flip the problem so that truth is higher-dimensional than signal
 - *Classify* MC truth into the space of high-dimensional synthetic data (MC truth \rightarrow S/C tracks)
 - Use a *generative* ML process to transform signals *upward* in dimensionality (S/C counts \rightarrow S/C tracks)
 - *Classify* the transformed signals back down to MC (S/C tracks \rightarrow MC truth)
- **Crux:** Transform MC truth into the representation space of detector hardware capabilities
 - Grounded in known physical processes available only in full simulation
- **Broadly applicable** to any scenario mapping low-dim simulation-labels to high-dim experimental response



A First Implementation

- **Detector simulation chain is a U-Net structure – a type of CNN**
 - Expand features going down, pick a bottleneck to aggregate global context, reconstruct back up merging encoded features at matching scales
- **Image preparation:**
 - Collect island of hits around highest-E hit (tree of direct neighbors)
 - Encode 12 channels of image (next slide) (log-weighted S/C counts used to normalize dynamic range – compresses small signals)
- **First implementation: a 3-level U-Net**
 - 2 copies of each image: full and masked (zero-out synthetic channels (red))
 - Use a weighted L1+SSIM loss (absolute pixel difference + structural correlations)
 - Scan hyperparameters (batch size, learning rate+scheduler)
 - Images are highly similar, so test with N=1000
 - Train for various epochs
 - Run inference to generate images



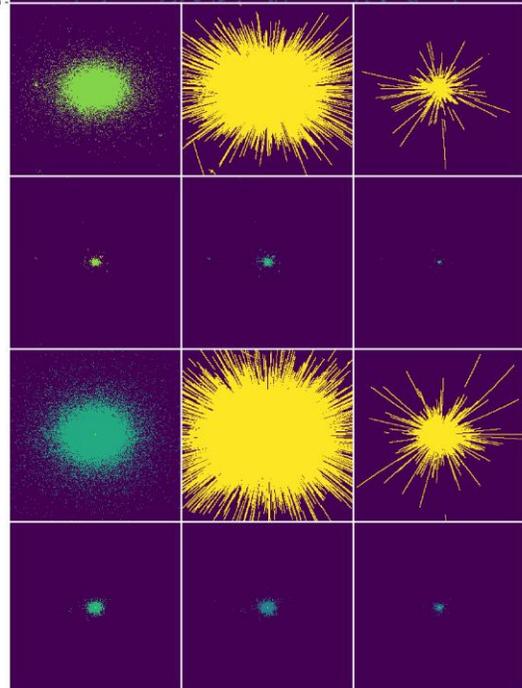
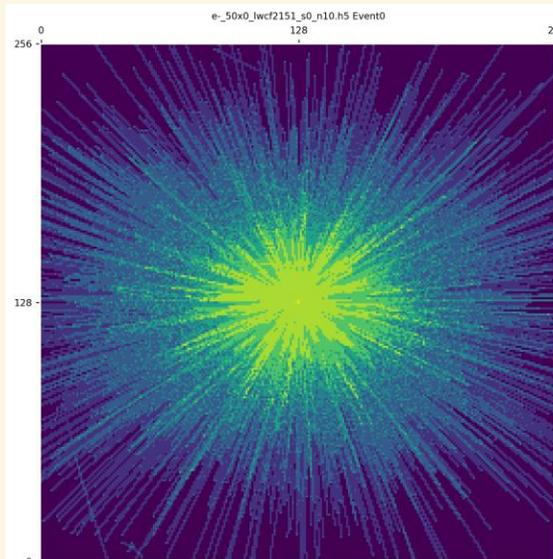
All channels

Front edep0 Front S tracks Front C tracks

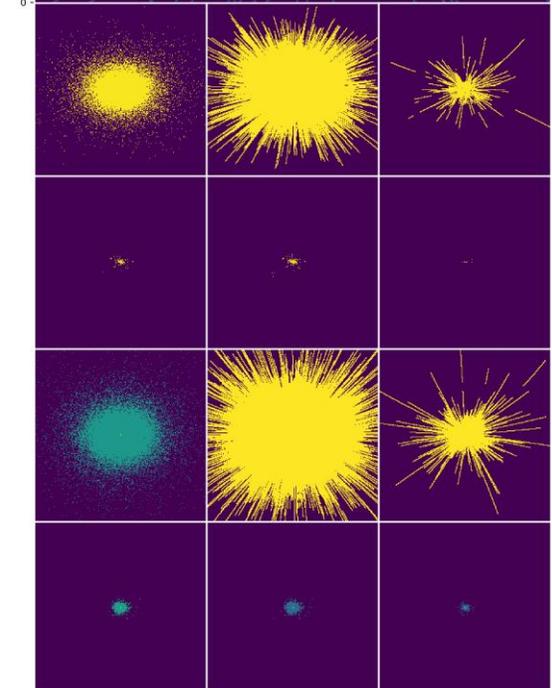
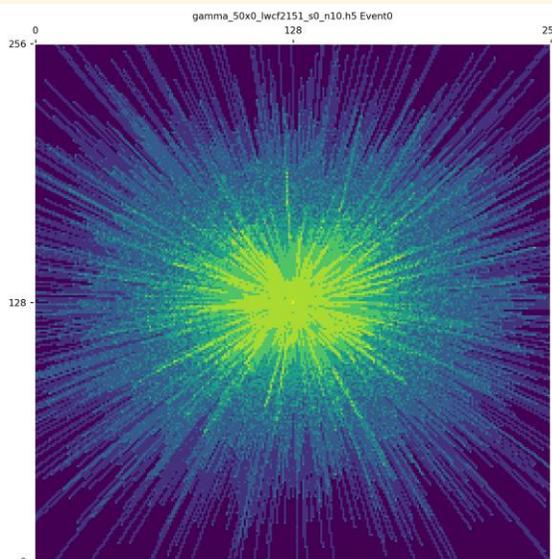
Front edep1kev Front S counts Front C counts

Rear edep0 Rear S tracks Rear C tracks

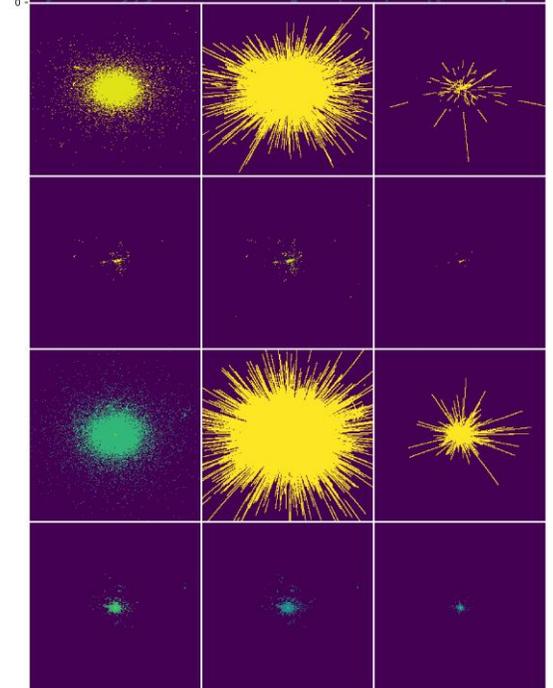
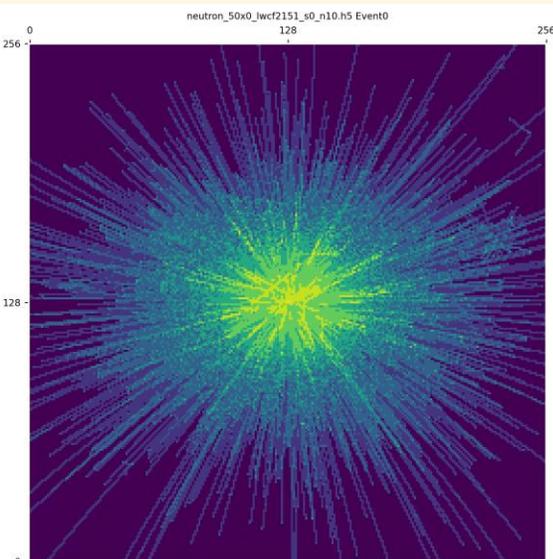
Rear edep1kev Rear S counts Rear C counts



electron
50 GeV



gamma
50 GeV



neutron
50 GeV

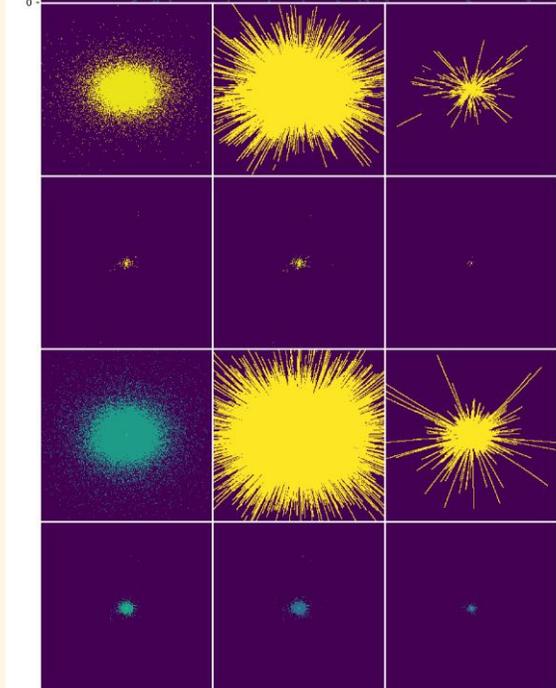
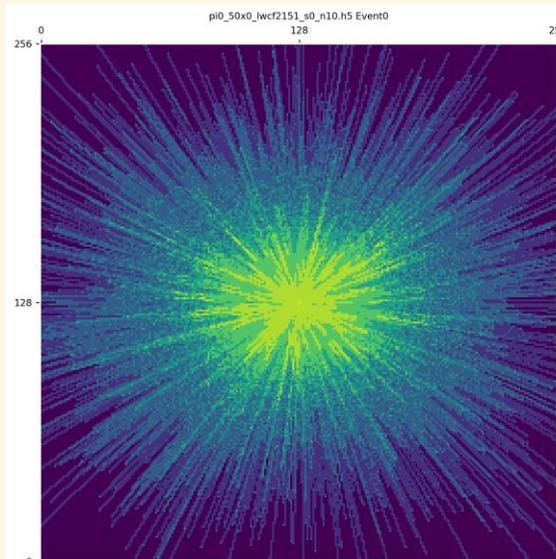
All channels

Front edep0 Front S tracks Front C tracks

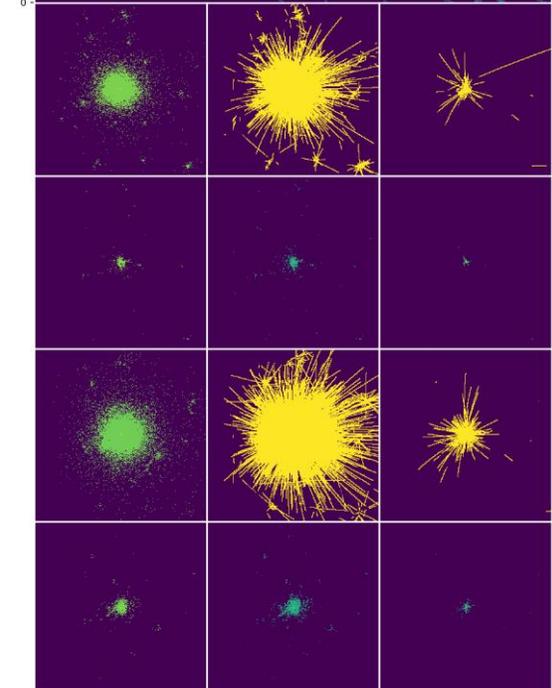
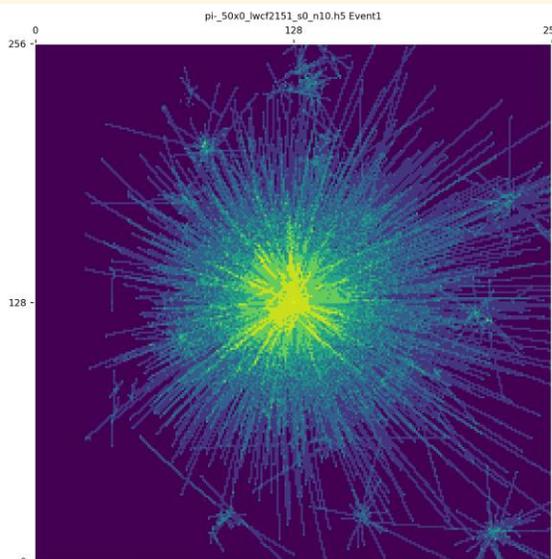
Front edep1kev Front S counts Front C counts

Rear edep0 Rear S tracks Rear C tracks

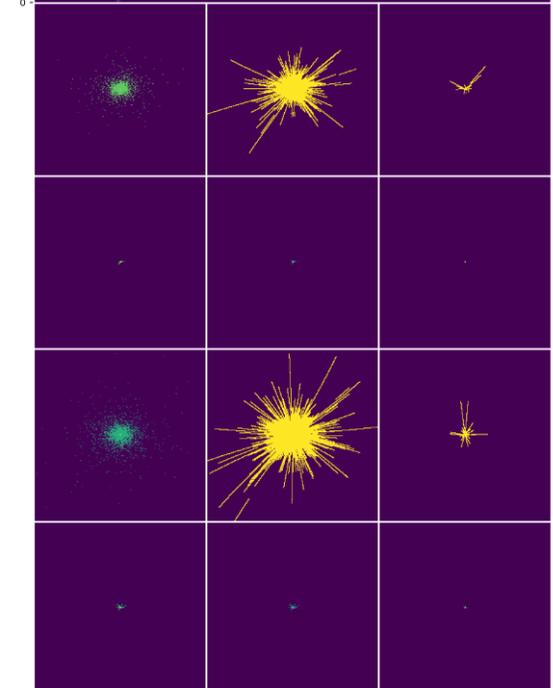
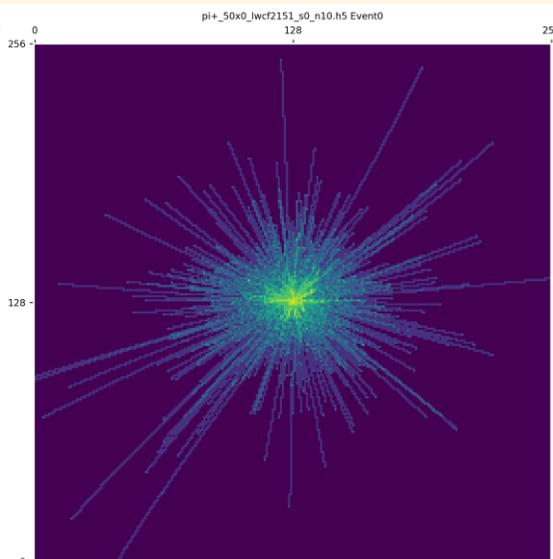
Rear edep1kev Rear S counts Rear C counts



π^0
50 GeV



π^-
50 GeV

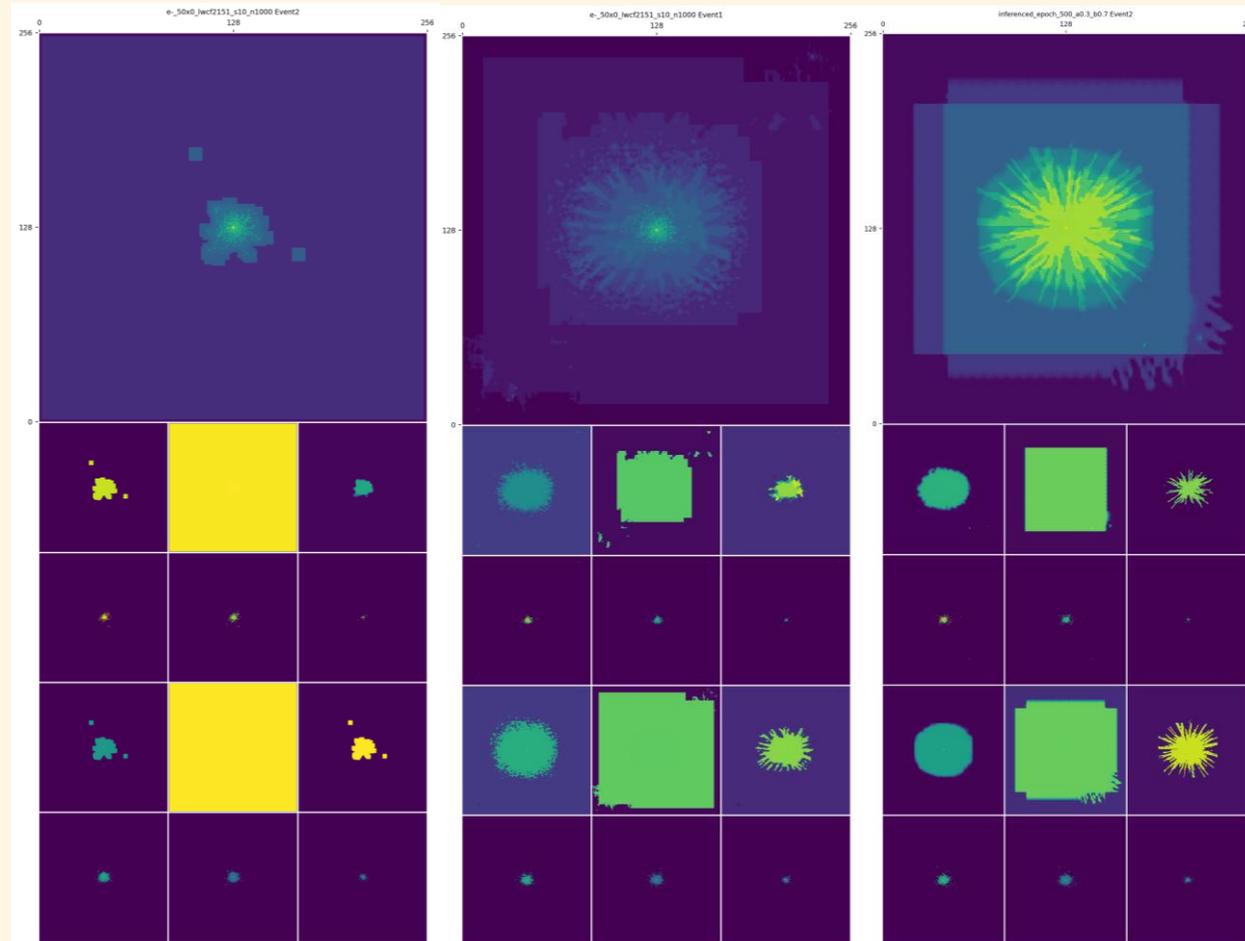


π^+
50 GeV

Inference: First Look

- Tuning hyperparameters is the challenge
- The Cerenkov signal gets resolved first – more sparse
- Scintillation signal is chipped away at
- Interpretation: effectively **machine learning the dual-readout correction**
- Example of a synthetic ML process rooted in a physical process
- Direct interpretability/explainability
- Anomalous signals – possibility to be more physical?

50 GeV electron inference



~10 epochs
(batch size 4)

~500 epochs
(batch size 8)

~500 epochs
(batch size 4)

Physical Interpretations of ML Models

- Detector simulations and ML ultimately express **programmed** stochastic processes and theories – random number generation, quantum interactions, etc.
- By linking these processes to a synthetic detector response/simulated observables, can they be surfaced to the real world?
- Intrinsic e/gamma/pio separation in ECAL could be studied, however with tracks, charged particle identification approaches almost 100% anyway
- More interesting question is whether this method can add an **ECAL handle on neutral hadron identification**
- Next steps:
 - Full classifier chain, more sophisticated generative model (latent diffusion), multi-particle final states, ...

