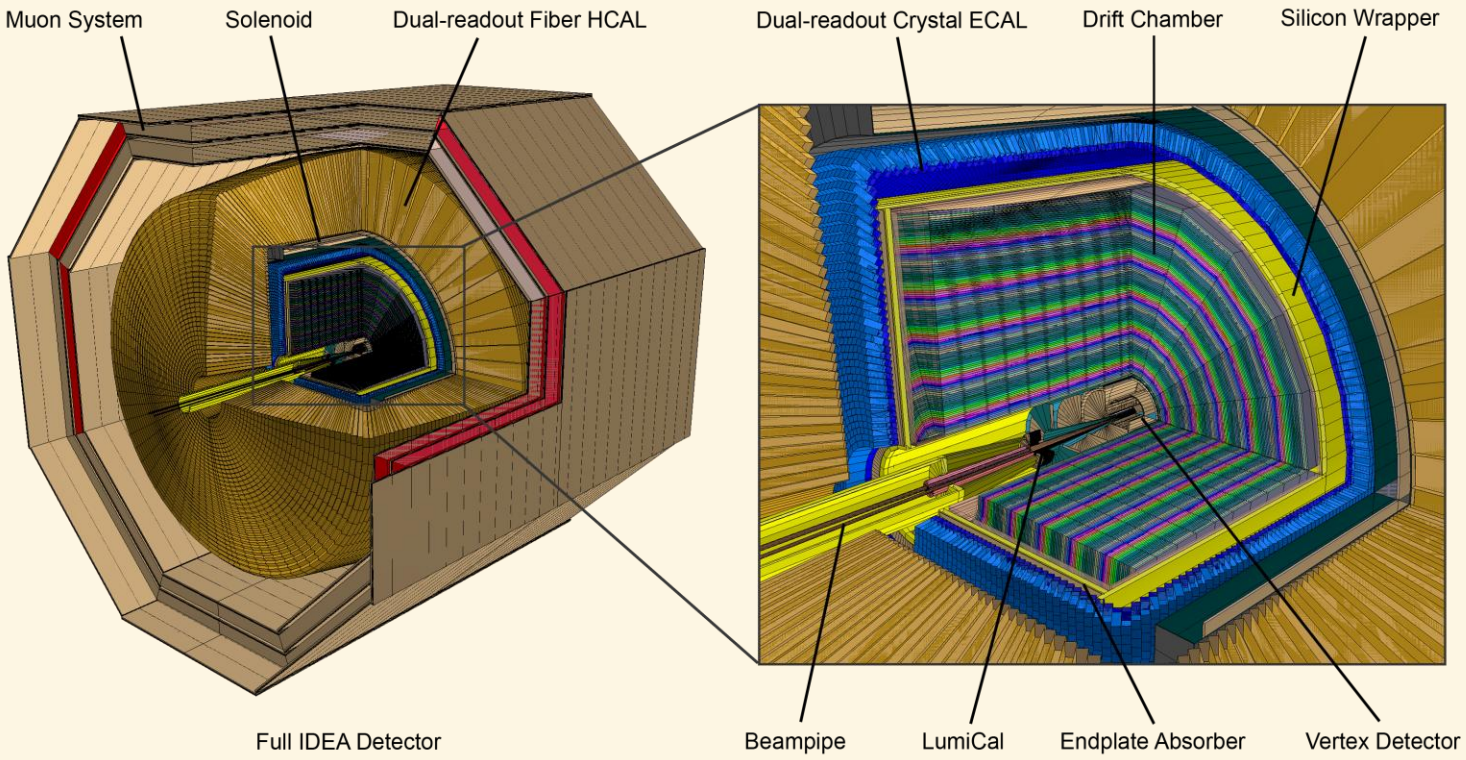
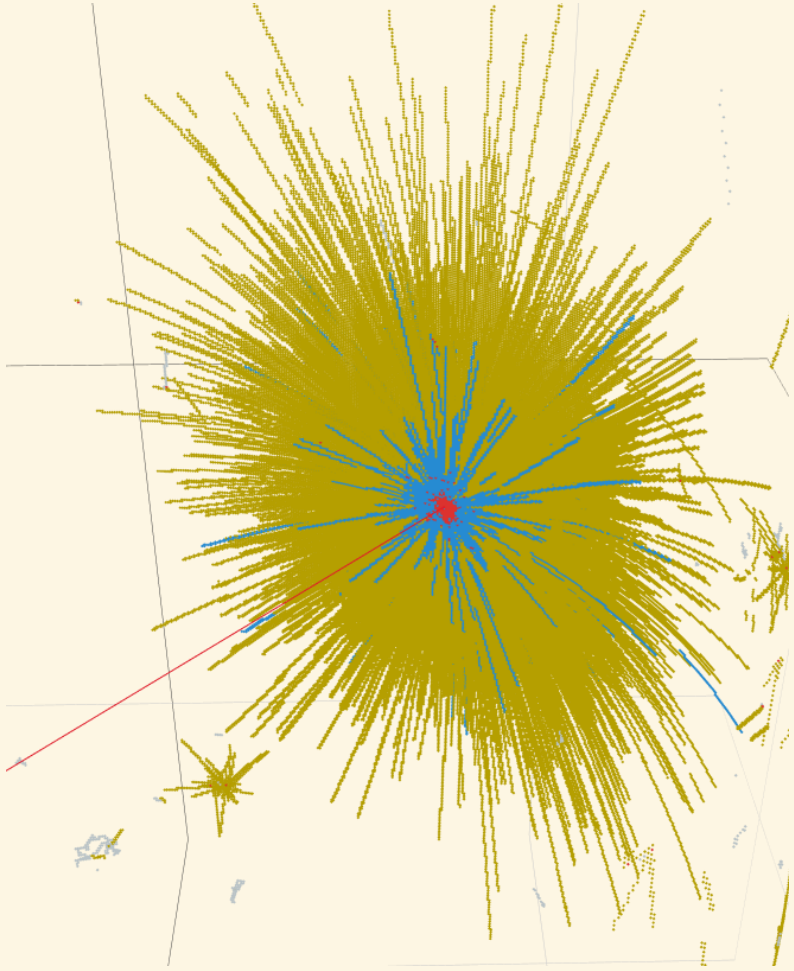
# Detector-Aware Anomaly Detection in Future Colliders with Synthetic Representations

[2505.05664]



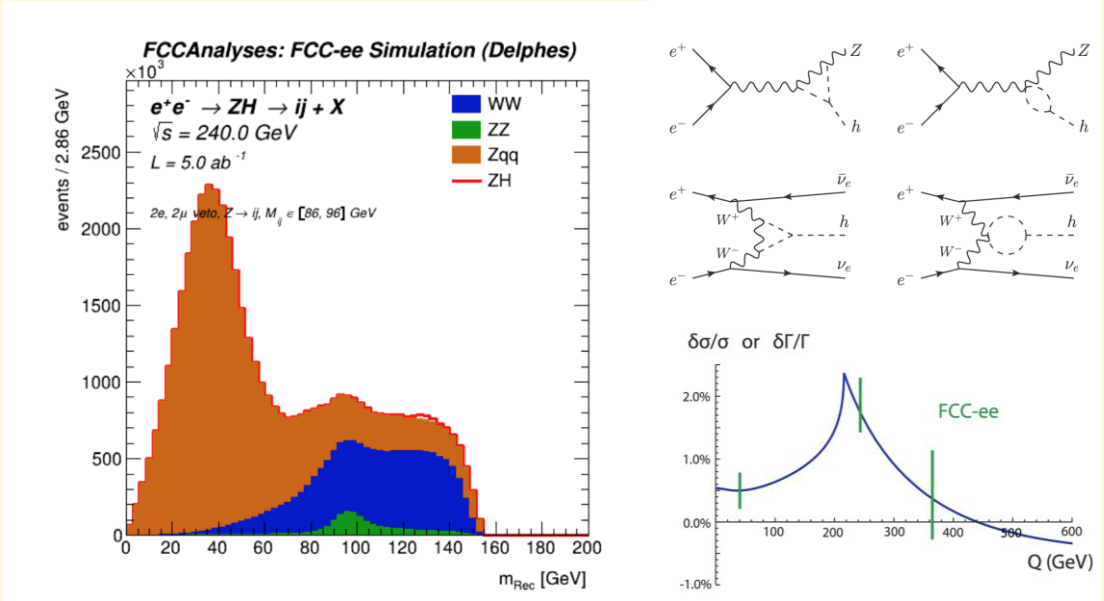
Wonyong Chung
Princeton University

June 2025

# The Next Lepton Collider

**Physics at 240 GeV**

- Tera-Z: Lepton universality
- HHH from loop corrections to the HZ cross section
- Z(qq)H apparently dominates
- Hinges on ability of detector to reduce ZZ background

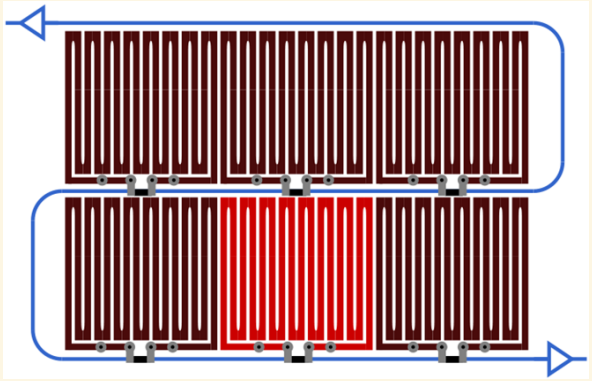


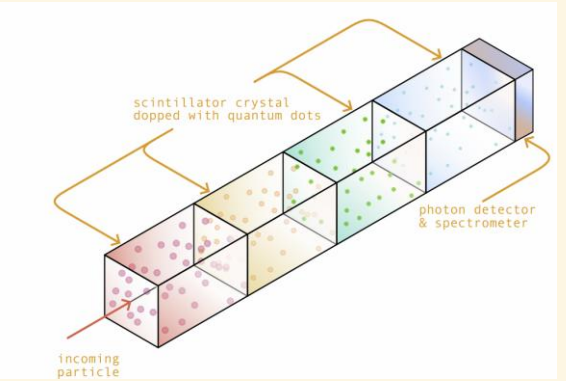

**Detector design in the AI/ML era**

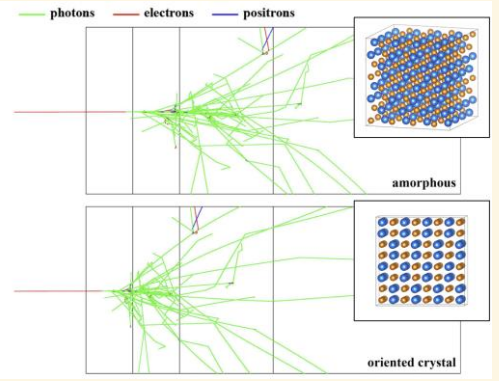
- Today: modular, flexible simulation
- Unified, top-down view of geometry and data schemas
- Hot-swappable subdetectors
- Triggerless DAQs
- Real-time inference on ASICs
- Picosecond timing
- 1:1 reconstruction

**New technologies**

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



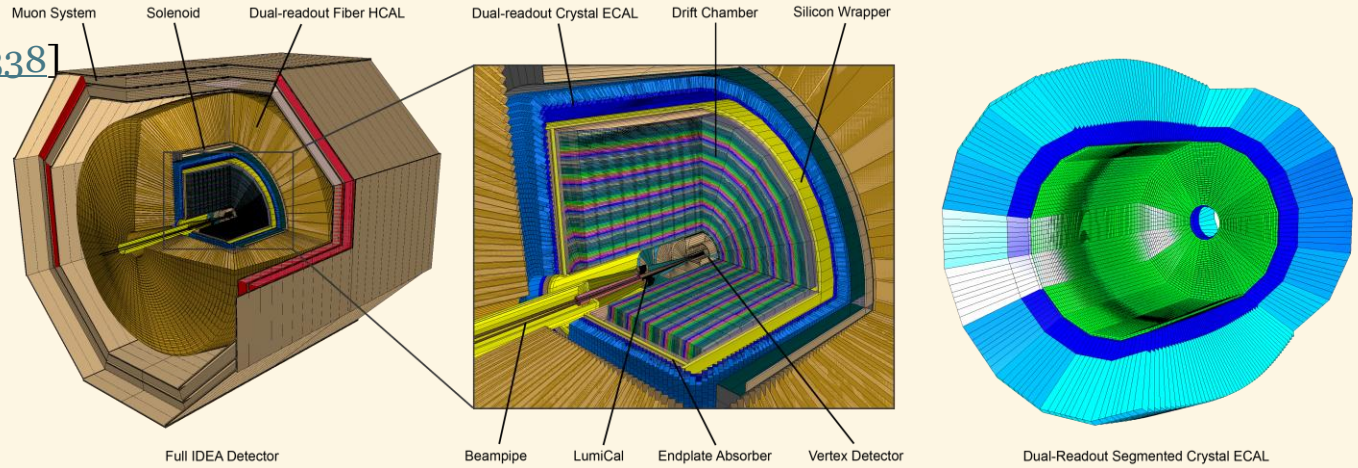


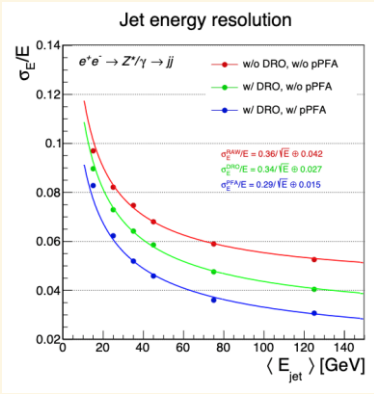



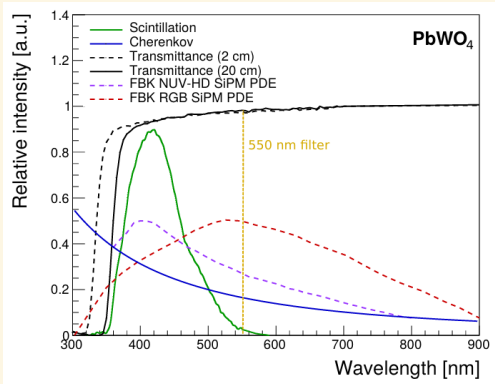
# “Differentiable” Full Simulation

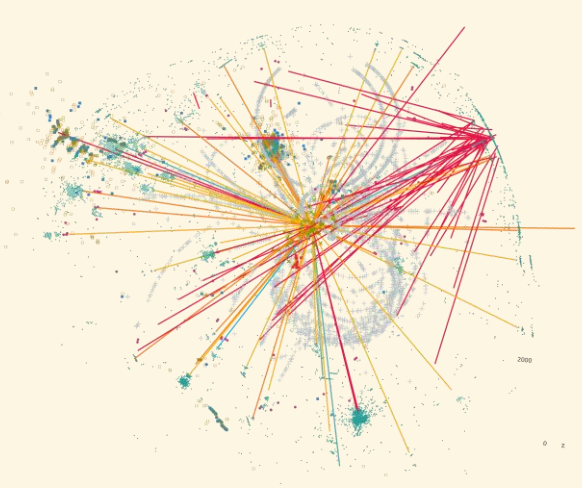
- **Segmented crystal ECAL** + precision timing layer [2008.00338]
  - Dual-readout (optical photons, separate scintillation/Cerenkov signals)
  - Added to IDEA detector baseline design [2502.21223]
  - Included in FCC Feasibility Study [2505.00272]

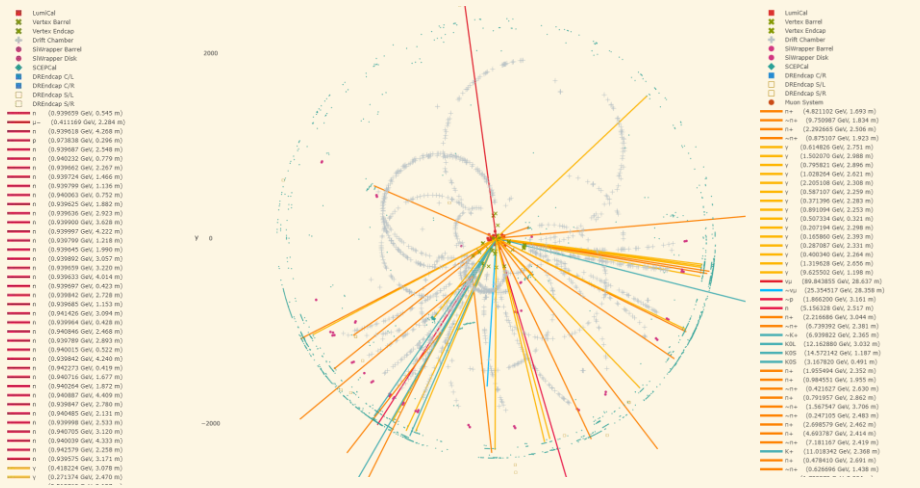
- Written from scratch in dd4hep/key4hep [2408.11027]
- Fully reconfigurable geometry, sensitive actions
  - “Differentiable” – not in-situ, but same pipeline effect
- https://github.com/wonyongc/SCEPCal

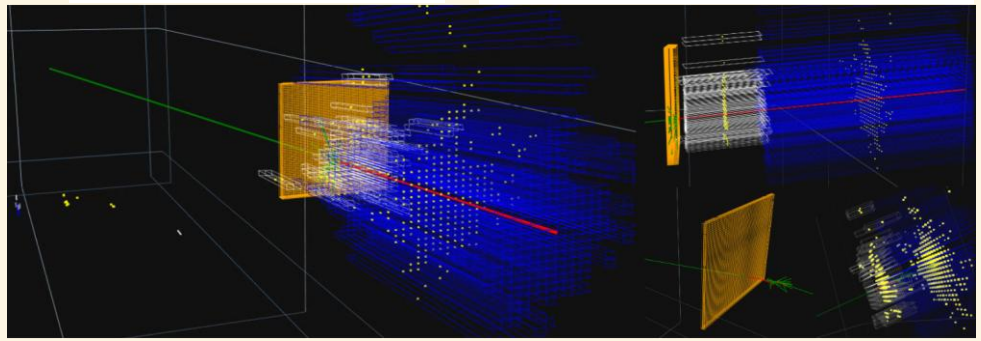




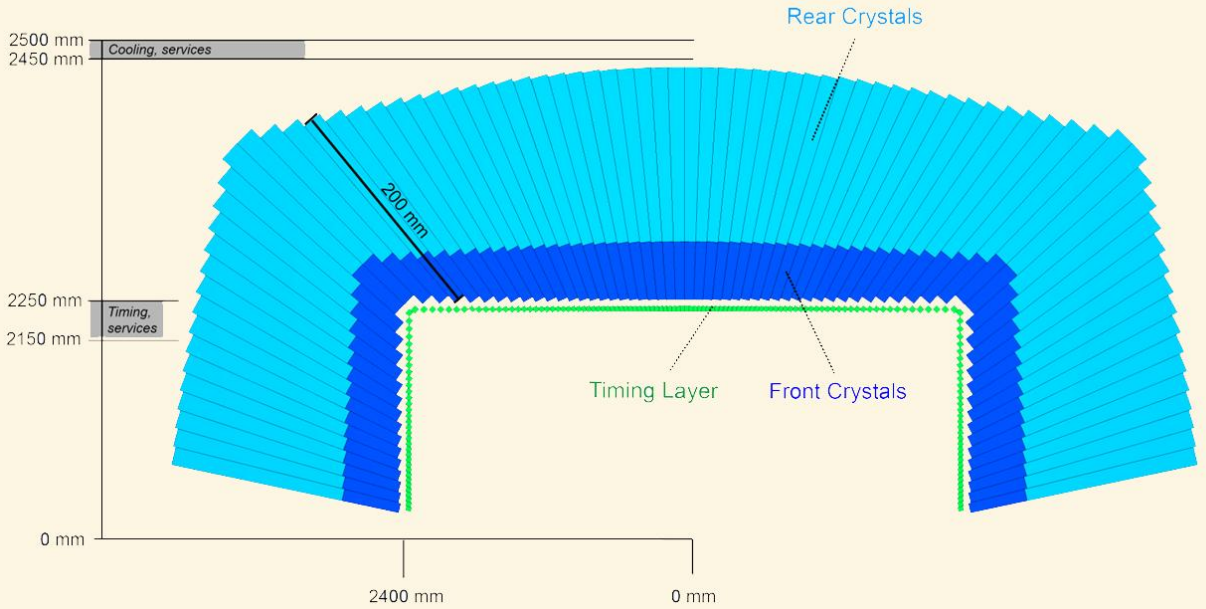


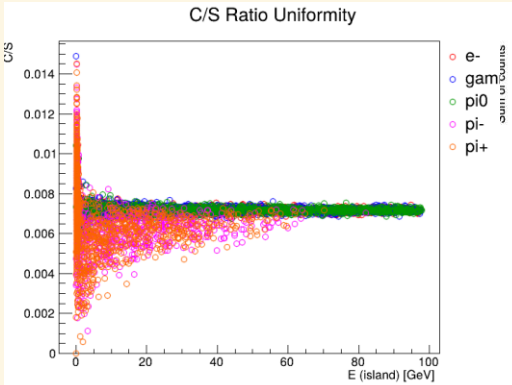
# Segmented Dual-Readout Calorimetry

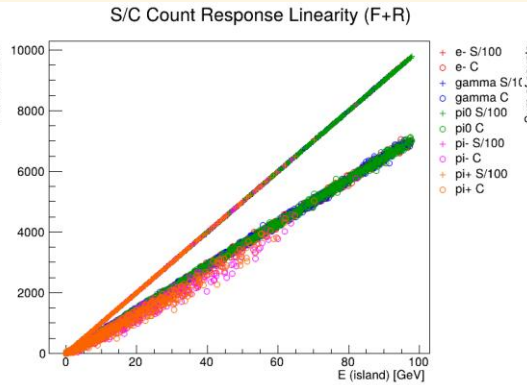
- **Calibrations:** 0-100 GeV e-, gamma, pi0, pi+, pi-
- **Technique:** Detect scintillation/Cerenkov light separately to mitigate event-by-event fluctuations in hadronic showers
- **Procedure:**
  - Calibrate on known EM/hadronic physics processors
  - Obtain the S/C response scaling factors
  - Determine EM fraction event-by-event

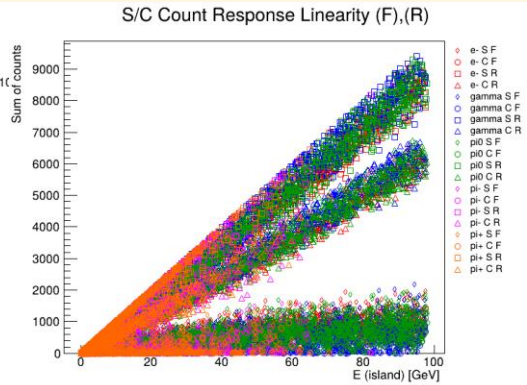
$$\begin{cases} S = E\left[\mathrm{f_{EM}} + \frac{1}{(e/h)|_S}(1-\mathrm{f_{EM}})\right] \\ C = E\left[\mathrm{f_{EM}} + \frac{1}{(e/h)|_C}(1-\mathrm{f_{EM}})\right] \end{cases}$$

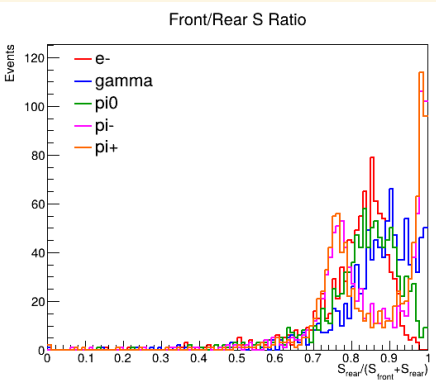
- **Segmentation** enhances separation power

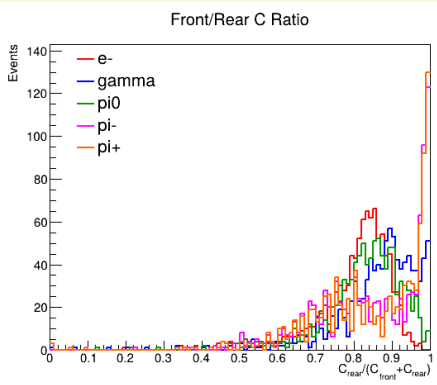


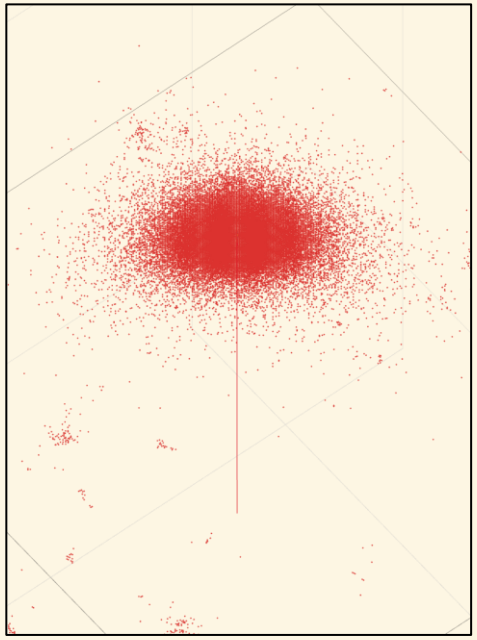



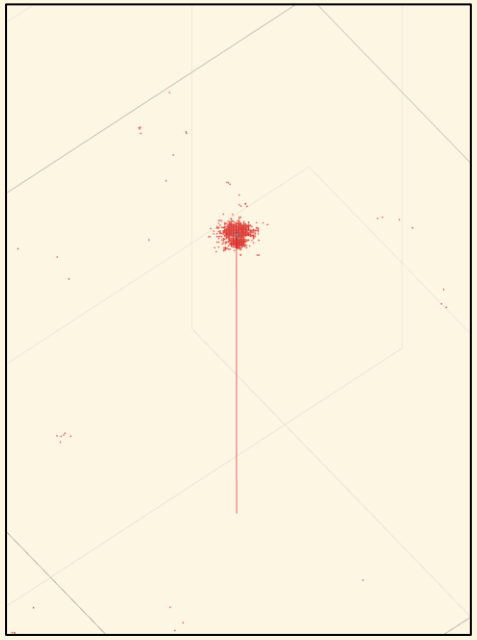
# Synthetic Representations of Detector Response

- Original work – before proceeding with traditional dual-readout analysis and particle flow – let's see what else we can do with this simulation

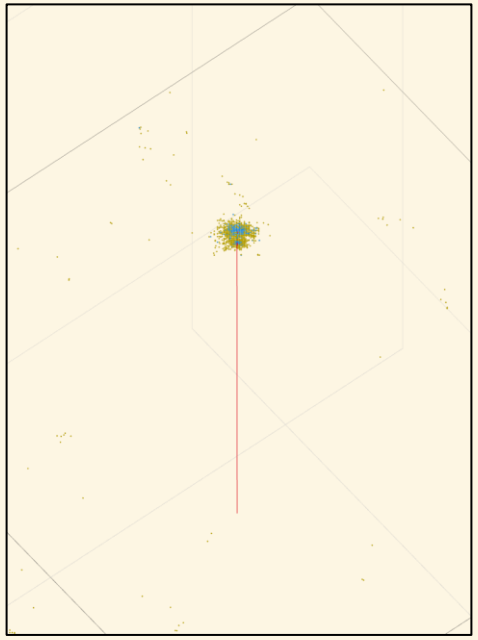
- Typical approach is to save hits based on **energy deposit** threshold (usually 1kev) at **step-level**
- For dual-readout, interested in optical photons, so apply **wavelength cuts** (300-600nm) at **track-level**
  - Save all hits for optical photons, even if energy deposit is zero – **simulated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



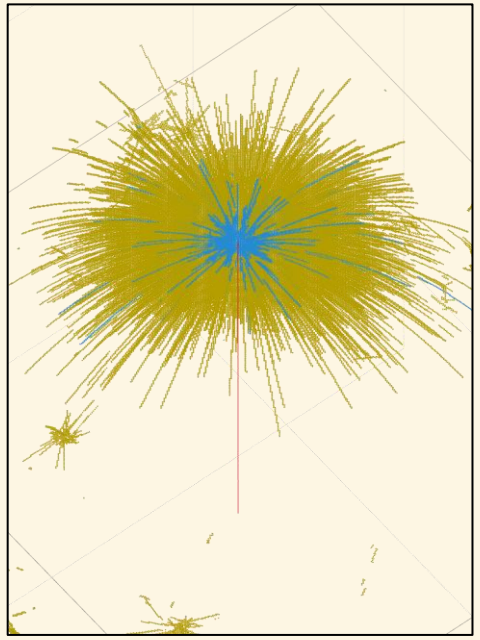
edep 0



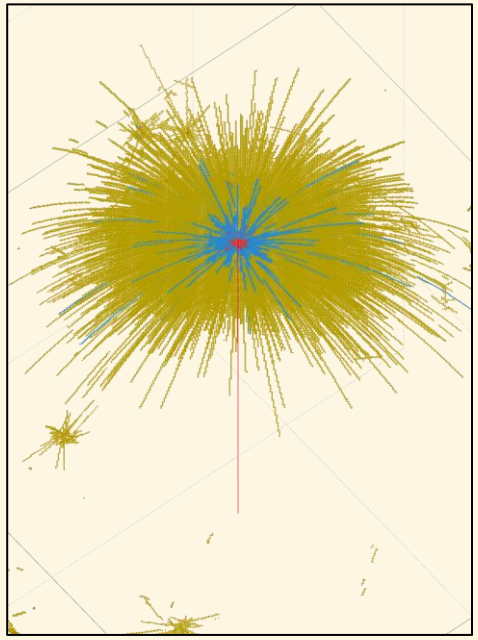
edep 1 keV



S/C counts



S/C tracks

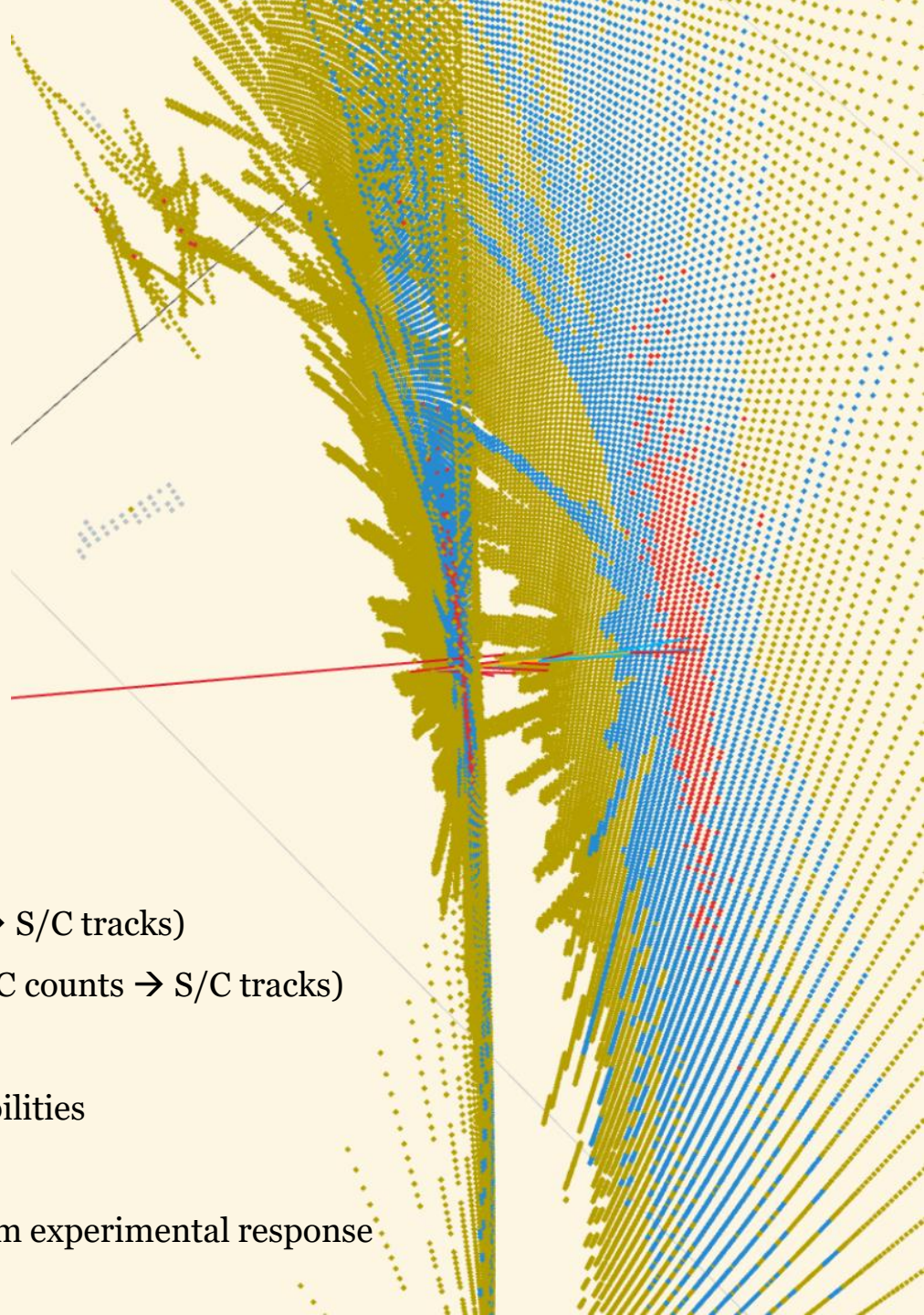


S/C tracks+edep 1keV

# Representation Bridging in Reconstruction Domains

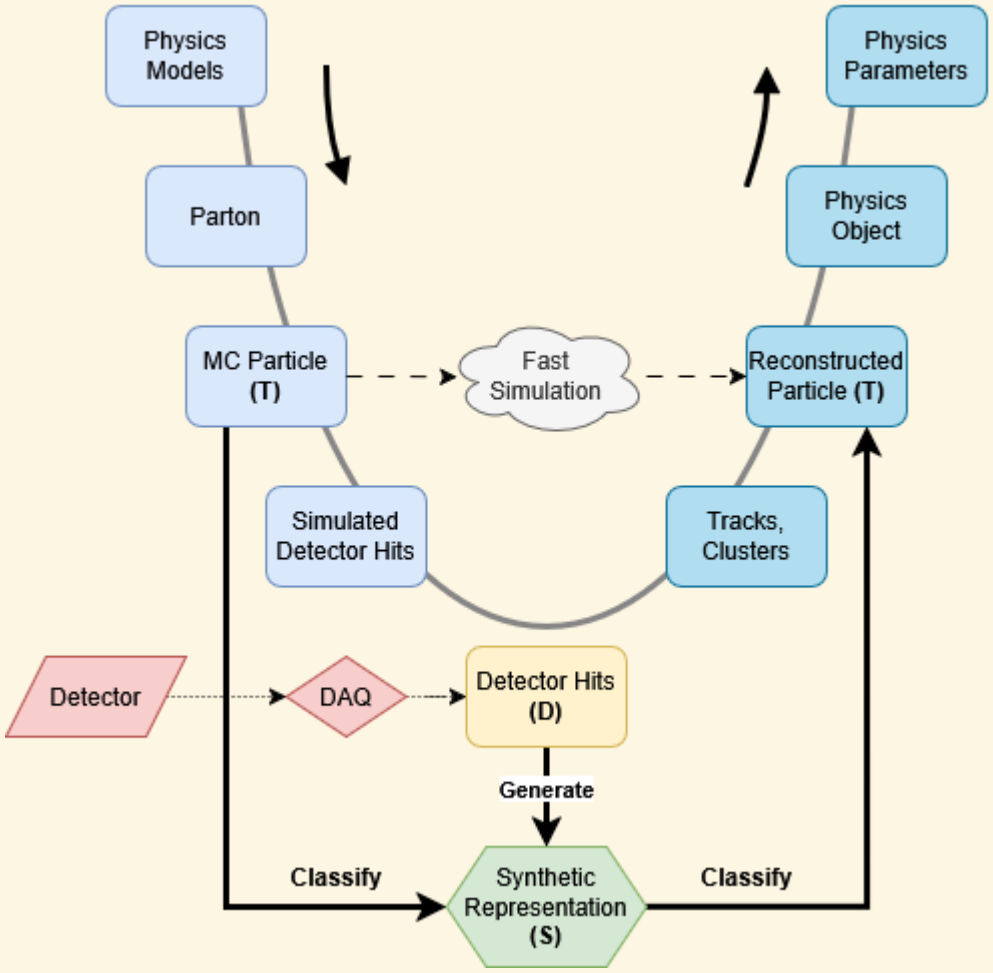
- S/C track hits are a form of **synthetic data**
  - *Unphysical* – won't ever see them in a real detector
  - *But not entirely unphysical* – are **representations of a physical process**

- 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 → S/C tracks)
  - Use a *generative* ML process to transform signals *upward* in dimensionality (S/C counts → S/C tracks)
  - *Classify* the transformed signals back down to MC (S/C tracks → 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



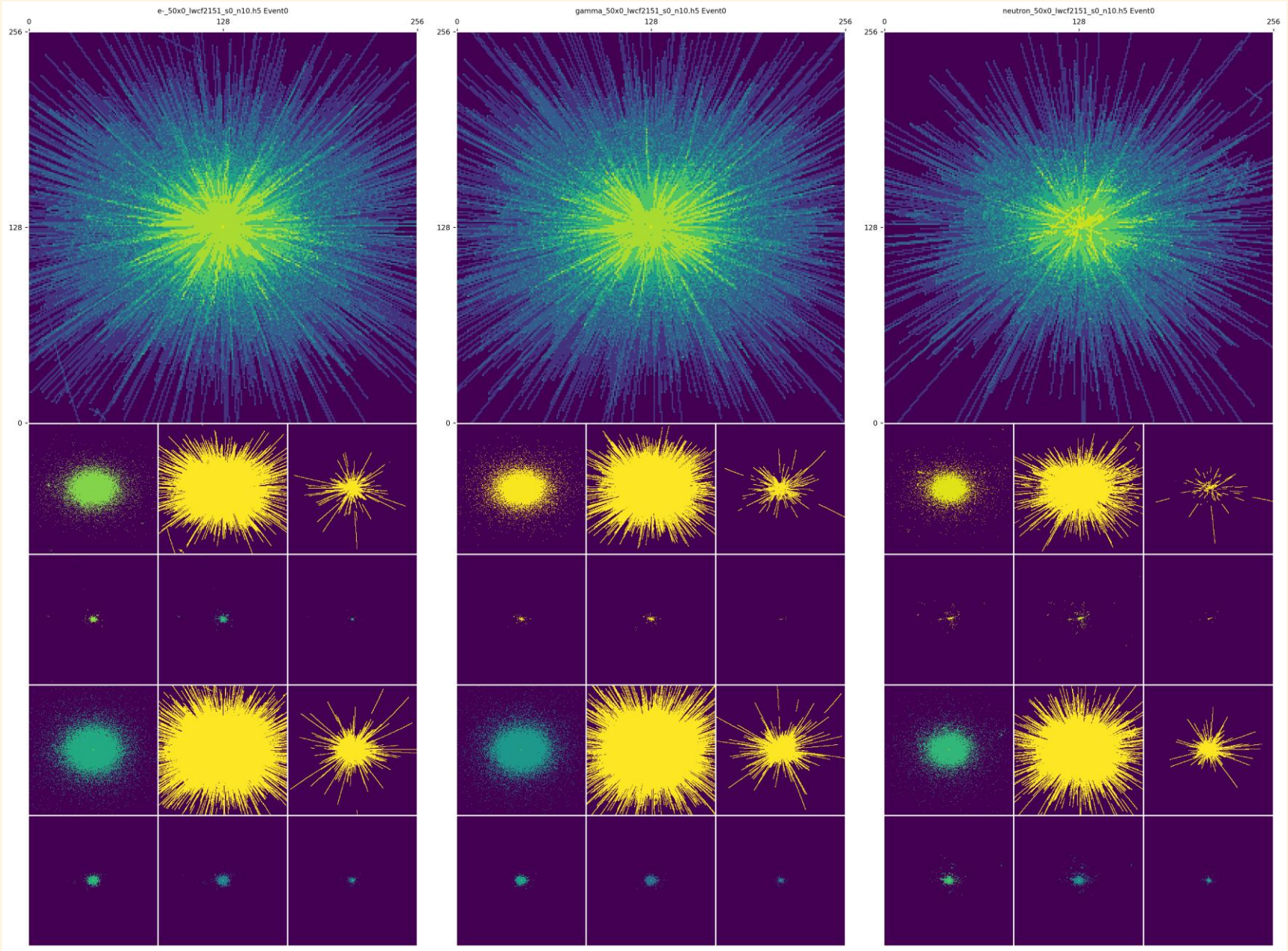
# A First Implementation

- **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 |



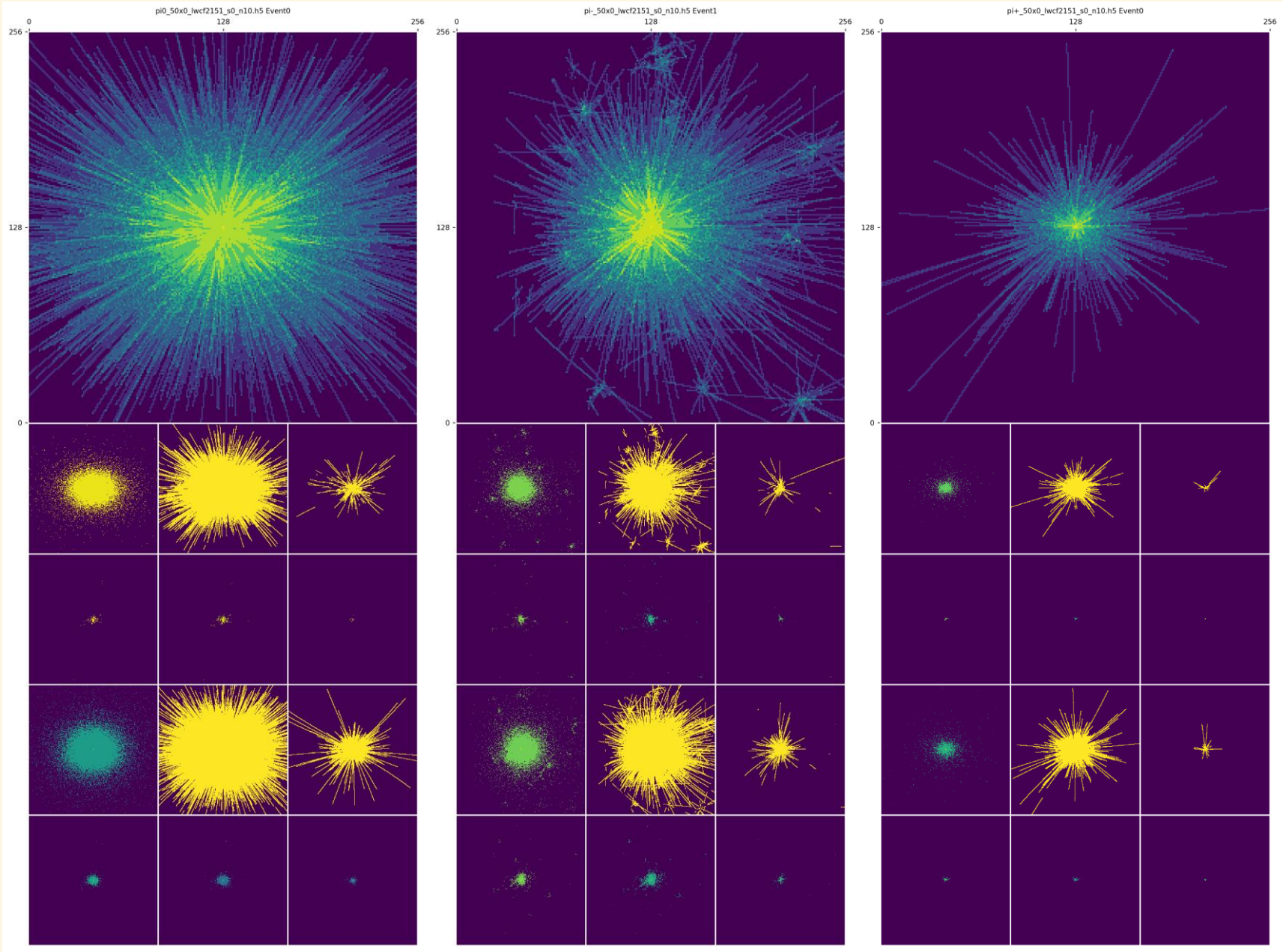

electron
50 GeV

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 |




$\pi^0$
50 GeV

$\pi^-$
50 GeV

$\pi^+$
50 GeV

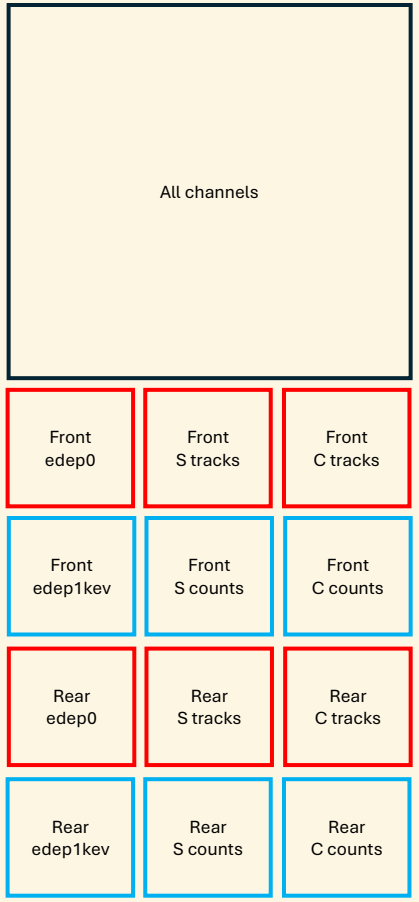
# Inference: First Look

- Tuning hyperparameters is the challenge
- The Cerenkov signal gets resolved first – more sparse
- Scintillation signal is chipped away at – suggests attenuation may be advantageous

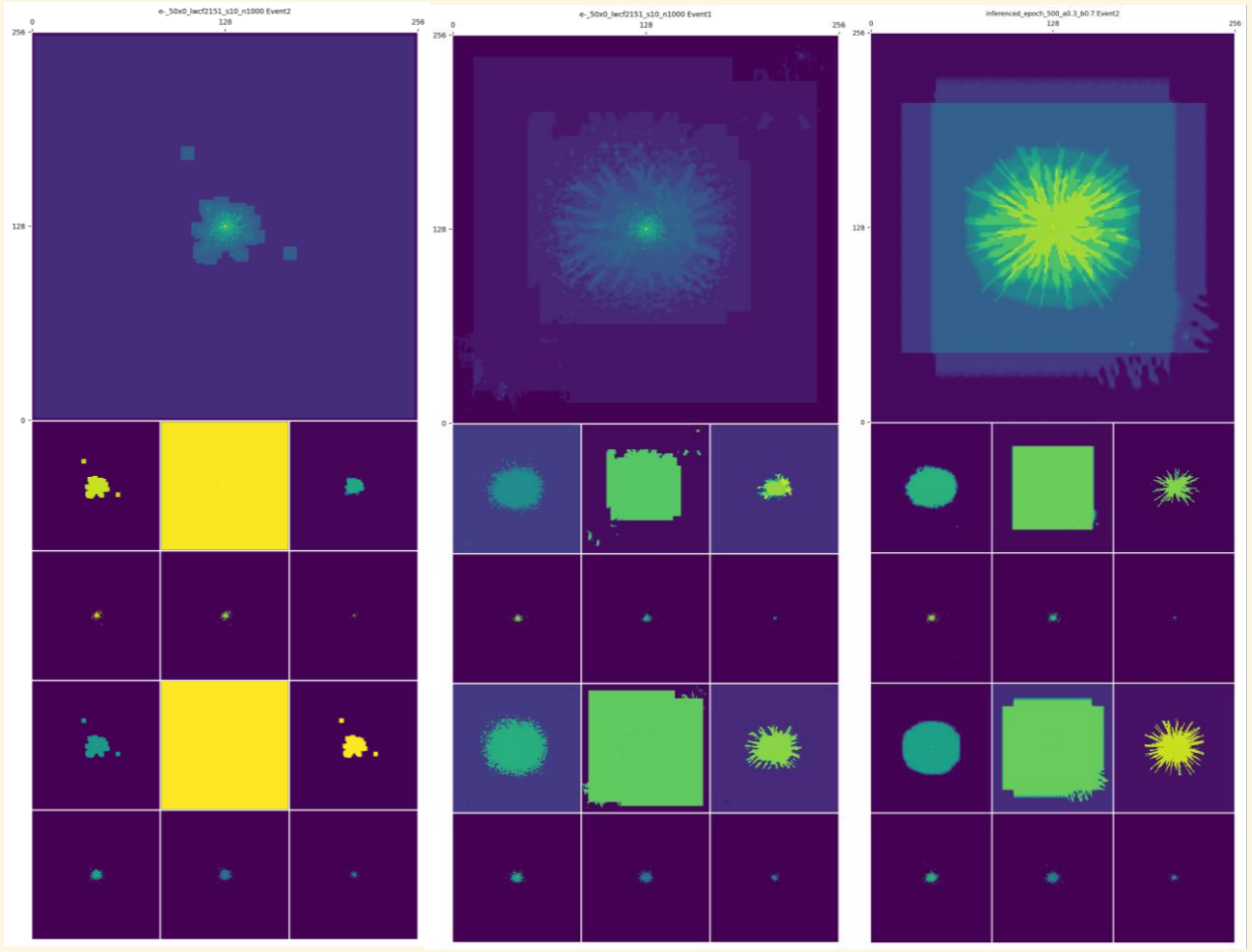
- Interpretation: effectively **machine learning the dual-readout correction**

- Example of a synthetic ML process rooted in a physical process
- Direct interpretability/explainability

- **Hypothesis:** Anomalous signals more likely to be 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/pi0 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, …