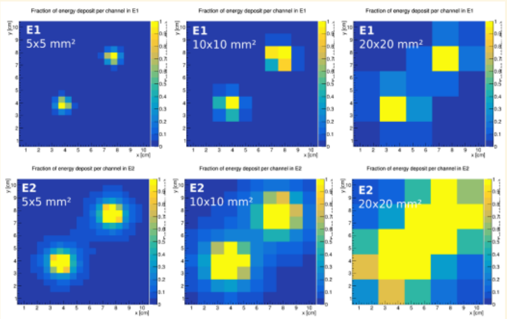
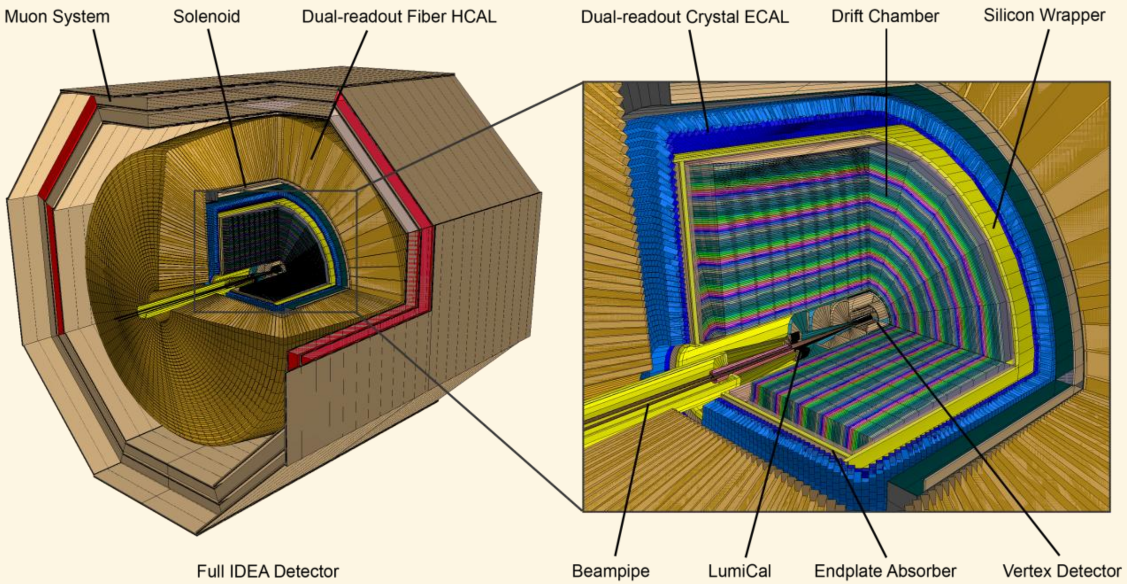
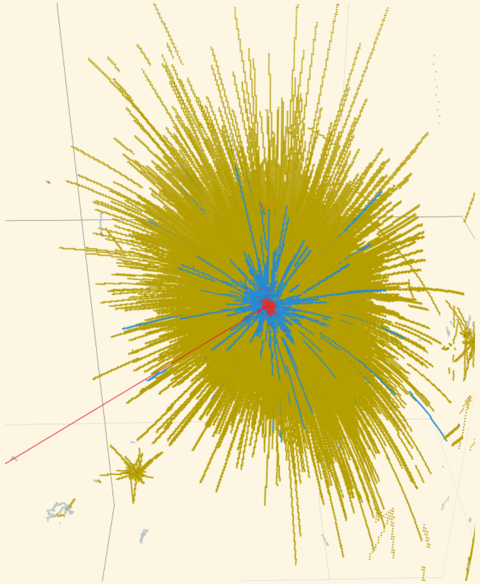
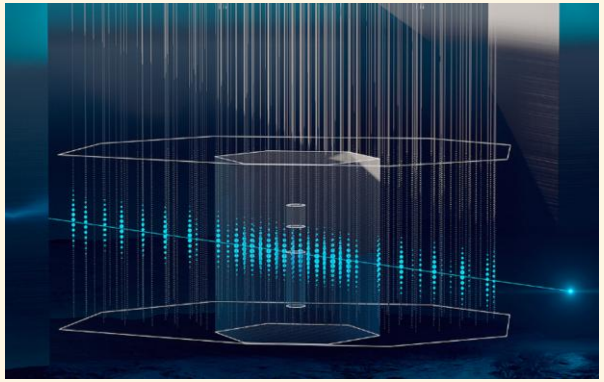
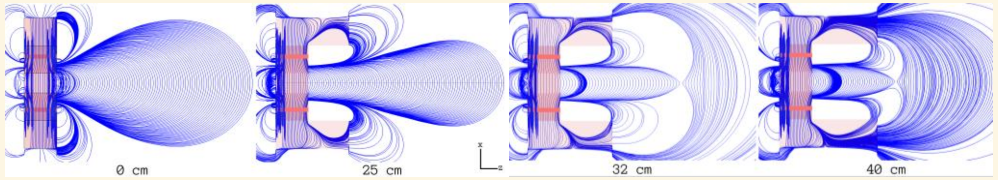
Phenomenological Detector Design

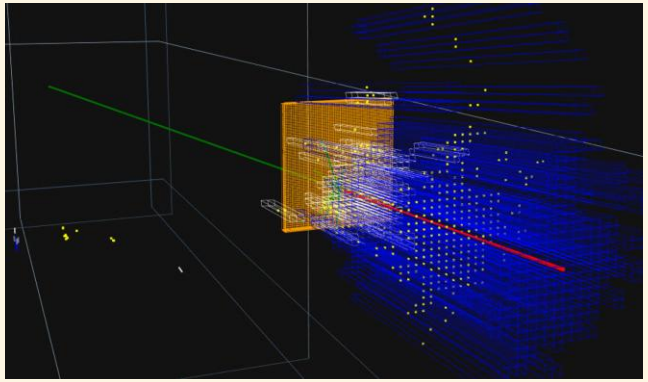












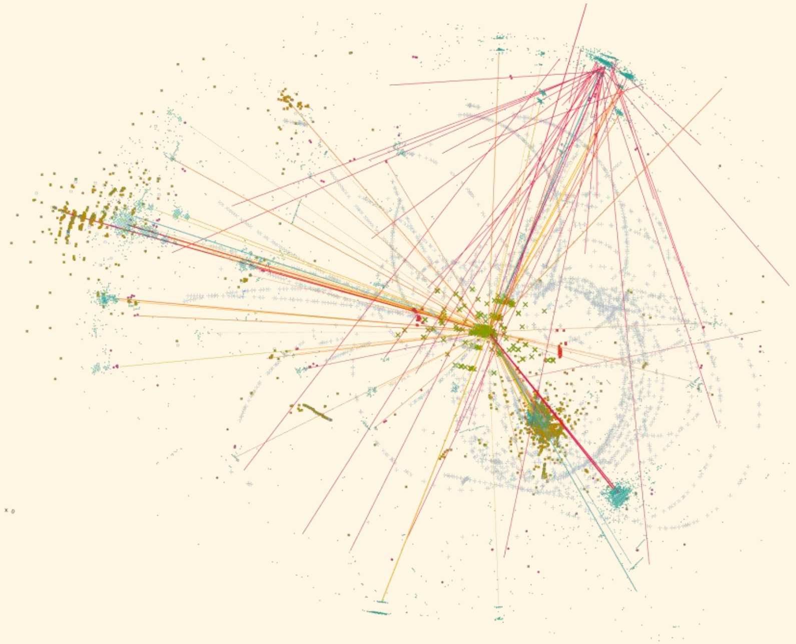
Wonyong Chung
Princeton University

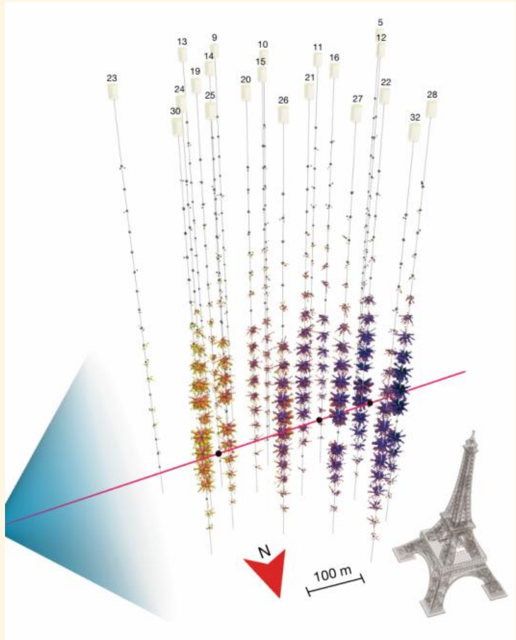
June 2026

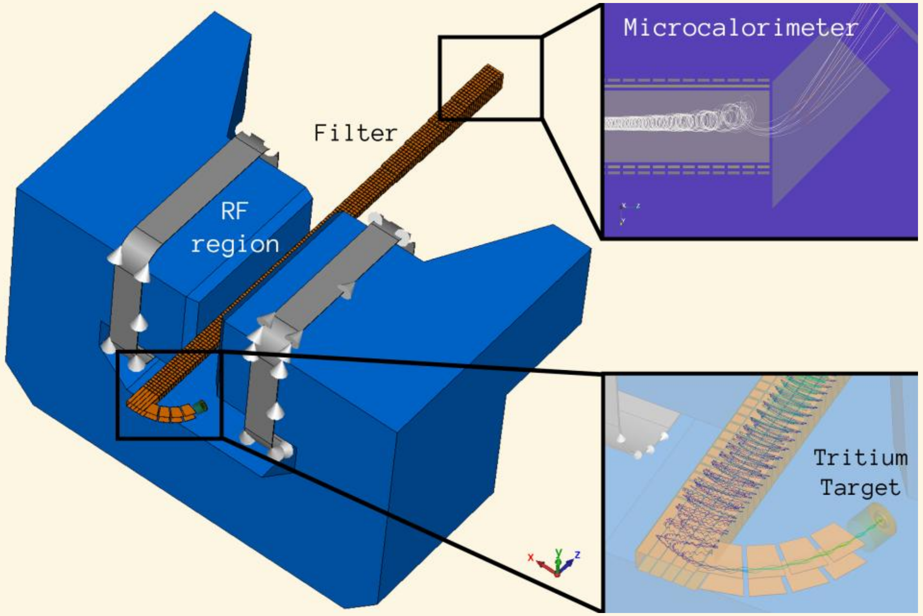
Preface – Detector Contexts and Constraints

- My perspective is framed around three types of experiments:

- Calorimetry in collider physics (dual-readout crystal calorimetry in IDEA)
- Open-water high-energy neutrino observatories (KM3NeT)
- Tritium beta-decay neutrino mass measurement and CvB detection (PTOLEMY)

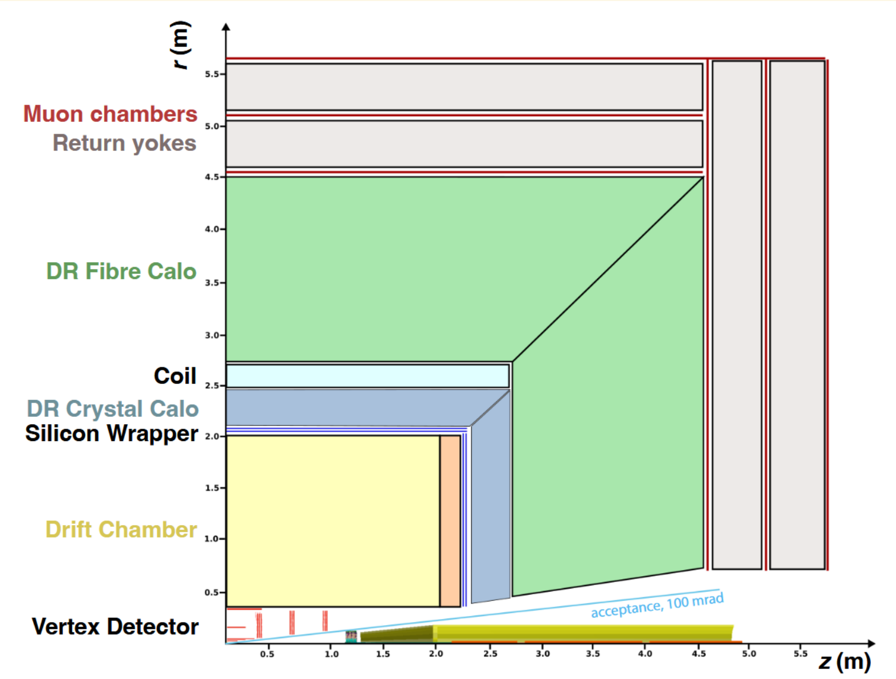


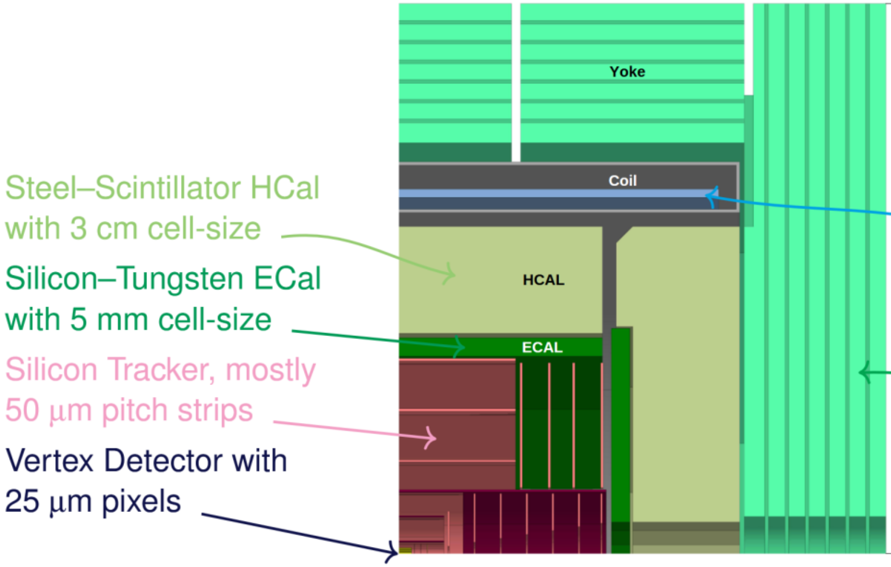




Calorimetry in Collider Physics

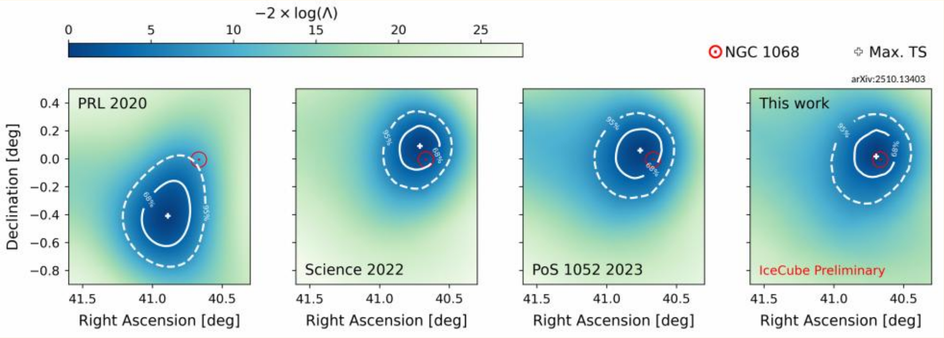
- Main split: homogeneous vs. sampling
- Well studied, cost is an outsize design factor, along with:
  - Accelerator environment (lepton vs. hadron)
  - Event rate / energy scale / physics goals
- Foundational constraints even before a detector is considered
- Modern PFA-oriented designs emphasize subsystem roles (tracking/pid, resolution, event filtering)
- Hard 'walls': neutral hadrons/confusion term, leakage, etc.
- Motivates a question: Is there a theoretical limit of information extraction based on material budget and properties?

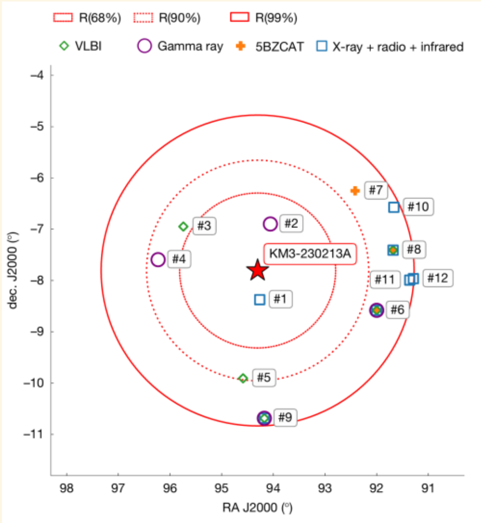


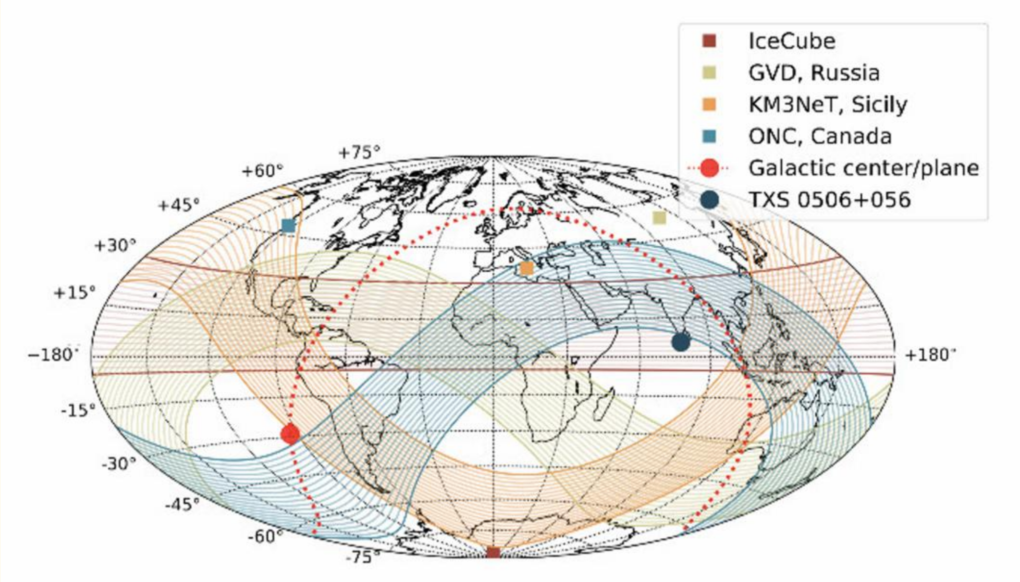


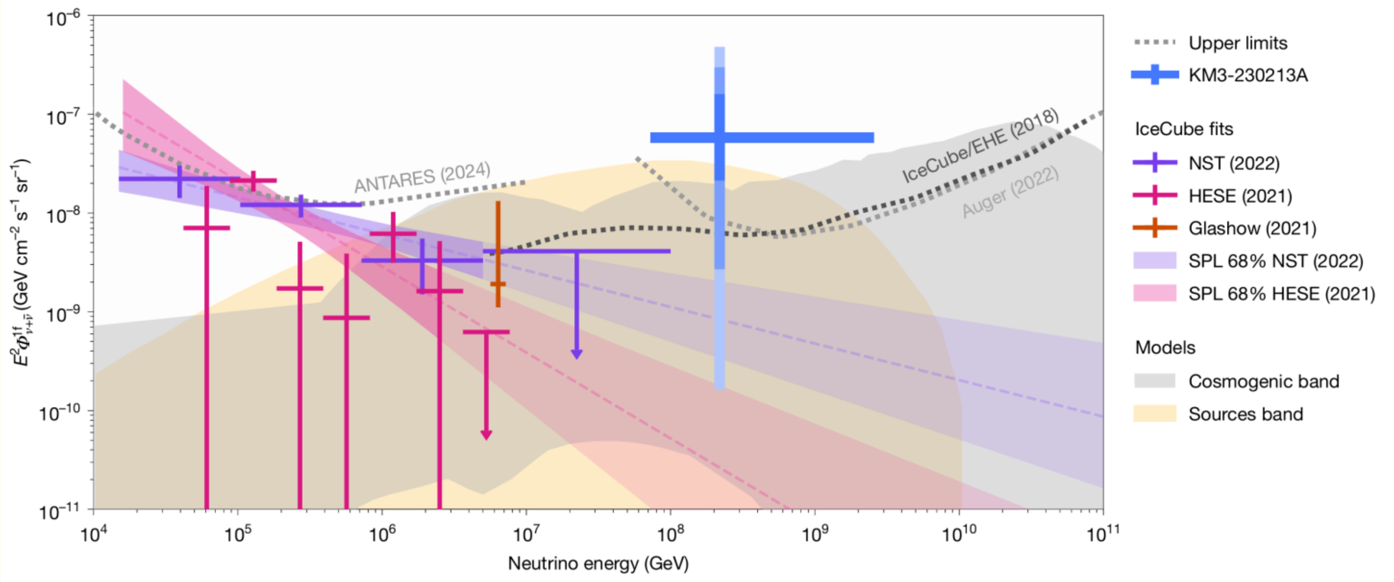
Neutrino Observatories

- Fundamentally different regime
  - Do not control events or energy scales
  - Detector acceptance/aperture vary depending on geography
- Open parameters make design very difficult





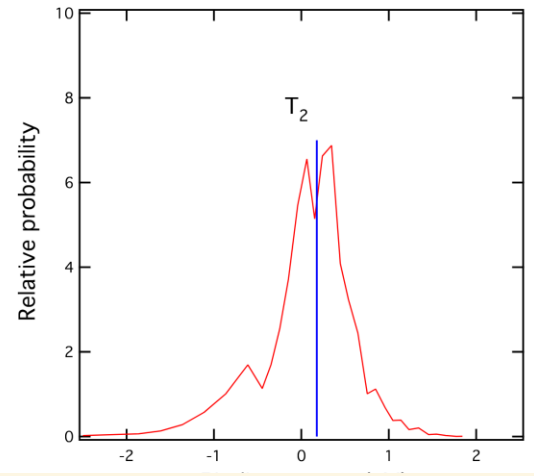


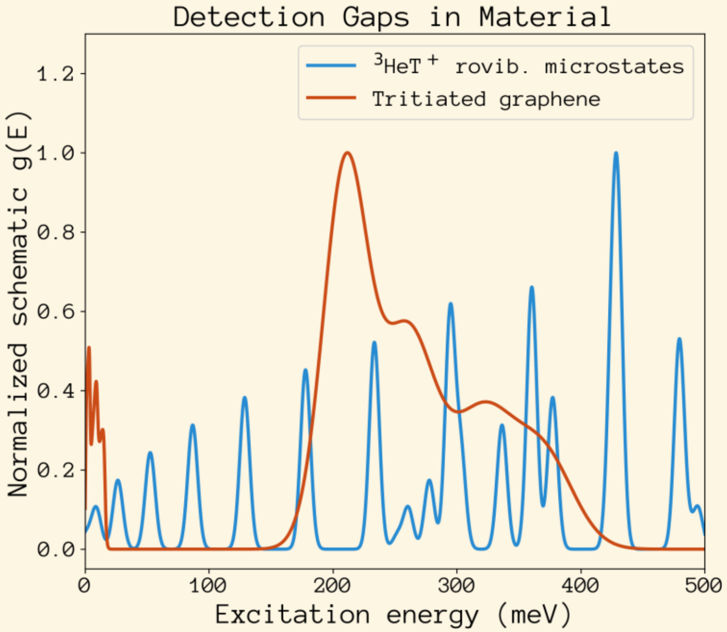


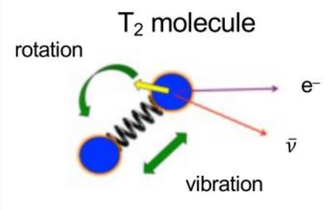
Neutrino Mass and CvB

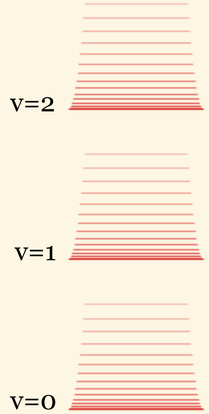
- Tritium-beta decay experiments (Mainz, Troitsk, KATRIN, Project 8) semi-standardized but approaching quasi-degenerate mass regime where ordering matters
- Final state distributions become key theory systematic, each final state has its own endpoint
- Atomic tritium on graphene shifts to "source as detector"
- Instead of minimizing FSD effects, design the material FSD to have a gap in m_β range
- Not searching for a shift buried under background
- Searching for mismatch between predictive spectral model and measured endpoint behavior
- <u>Natural interpretation of any discrepancy in the clean window is the neutrino mass</u>
- Joining the likes of scintillators (Stokes shift), silicon detectors (band gap), ionization detectors, bolometers, etc…

$$^3\mathrm{H}_2 \rightarrow {}^3\mathrm{HeT}^+ + e^- + \bar{\nu}_e$$

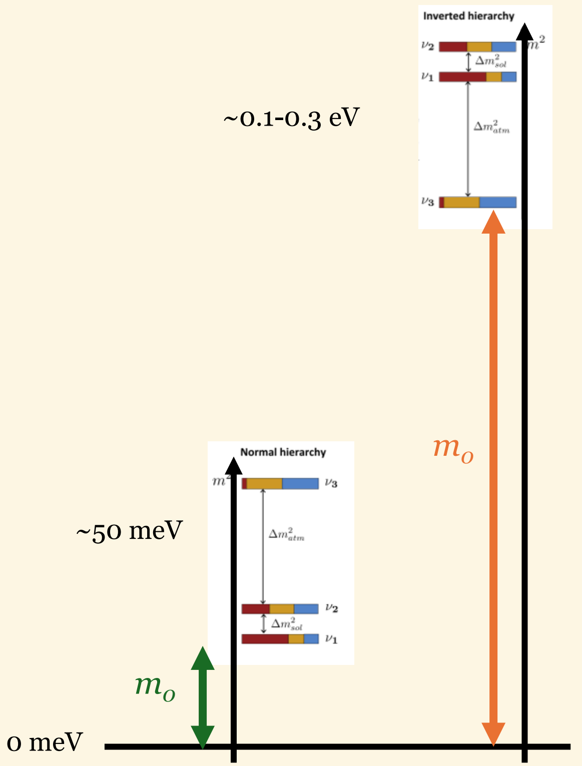








$$m_1 \approx m_2 \approx m_3 \gg \sqrt{|\Delta m^2_{31}|}$$

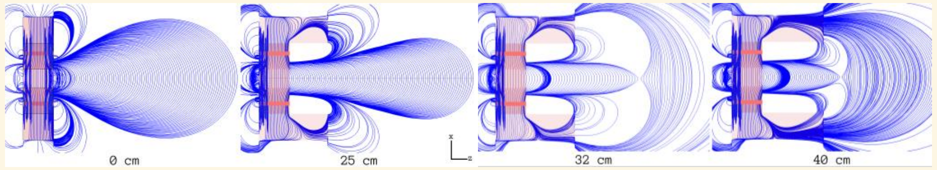


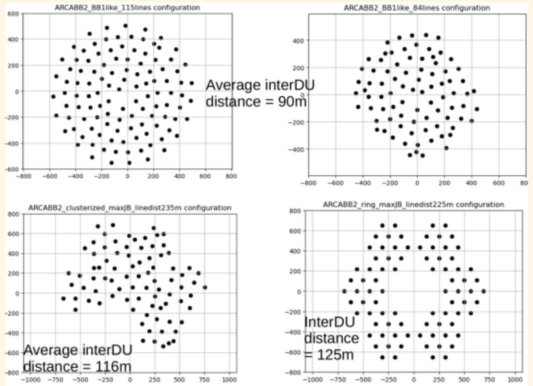
The Common Threads

- Seemingly disparate examples
- In the end, all concerned with identifying the signature of physical observables that result from fundamental interactions in matter – beyond “more resolution, more precision”

- **Signal to noise/background**
  - Statistics (increase signal over background)
  - Purity (decrease background below signal)
- Includes material/sensor properties incl. QE – effectively a cross-section on signal
- Colliders
  - Statistics controllable
  - Accept losses in material/sensor purity
- Neutrino observatories
  - Ice vs. water material effect on purity
- Neutrino mass
  - Atomic vs. molecular tritium

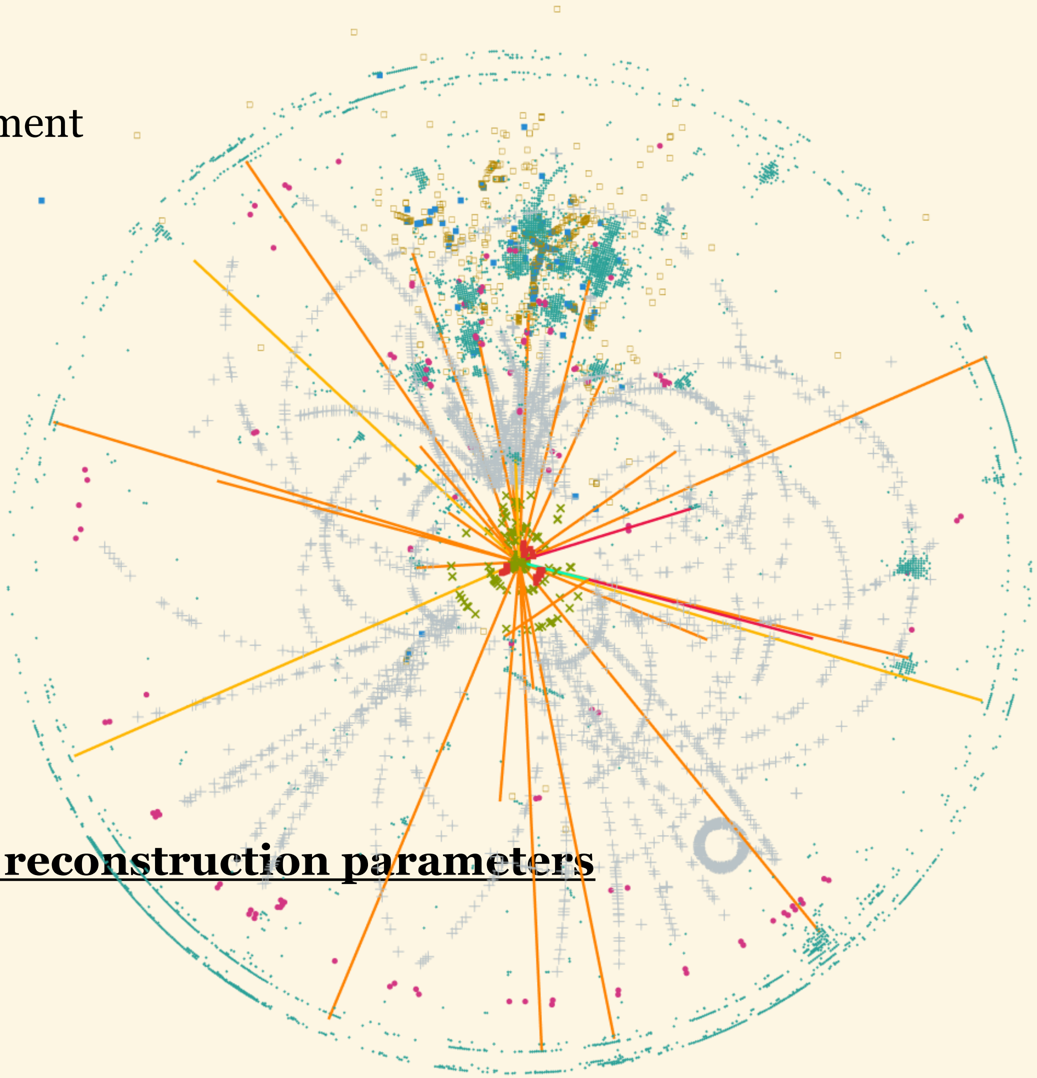
- **Geometry**
  - “Let the geometry do the work for you”
- Combination of lattice of material and sensor nodes that perturb+measure physics process
  - Configuration yields resolution and sensitivity as emergent properties in representation space





Some Recent Work at SLAC with Collider Simulations

- Want to resolve/map out some network effect → high statistics
- Colliders: sweet spot of geometry/external constraints with local movement
- Full simulations allow to hone in on what parameters actually matter



- **Strongly geometry-dependent objectives:**
  - Di-jet mass resolution $Z \to q\bar{q}$
  - Light jet background
  - Photon energy/angular resolution
  - Jet resolution vs polar angle
  - b-tagging (tracker dominated)
  - Missing-energy tails in $ZH \to \nu\bar{\nu}+X$
- <u>**Want to be able to systematically vary detector geometry and reconstruction parameters**</u>

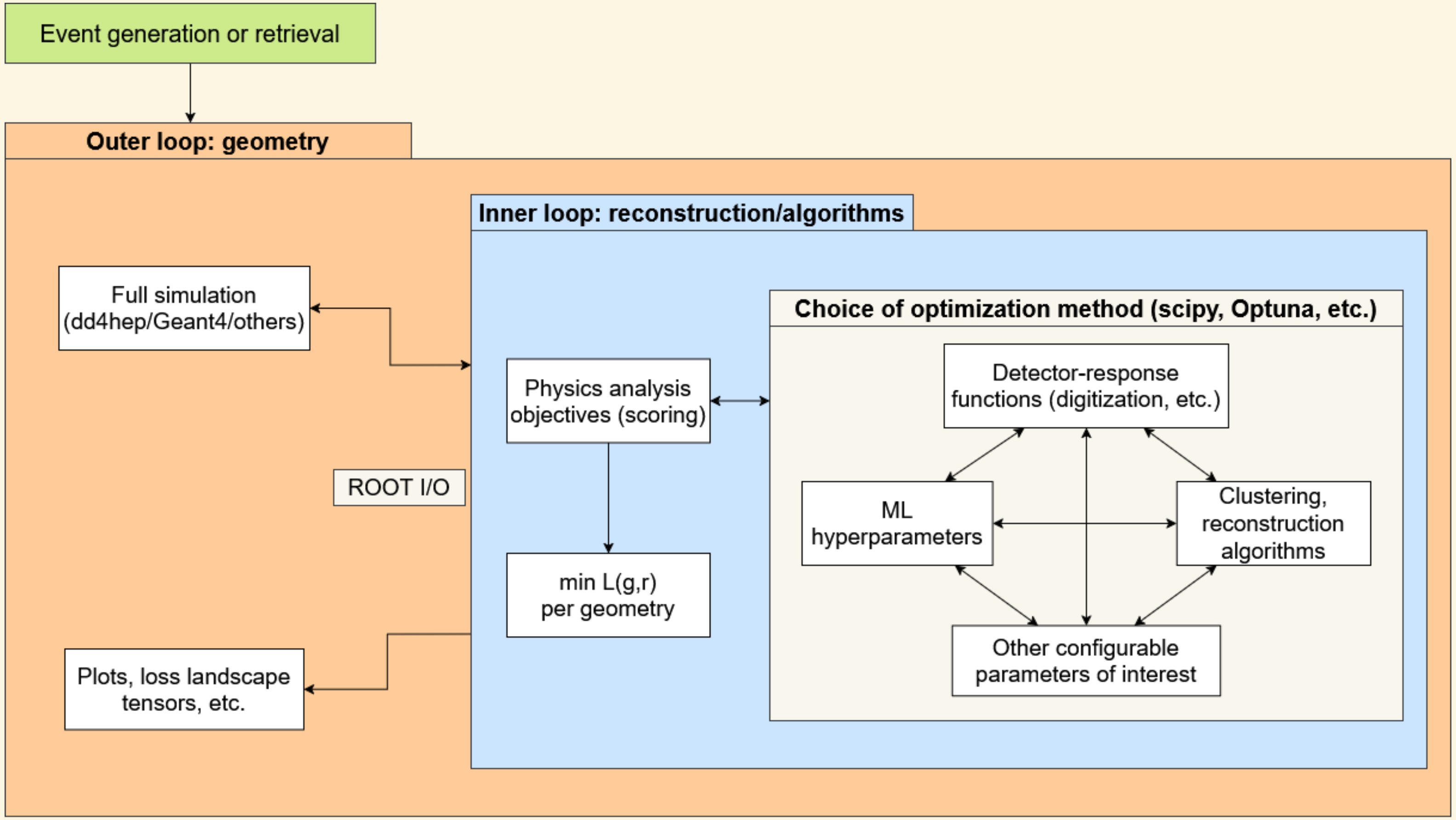
Precision physics → detector performance drivers

- Run points: Z(91 GeV), WW(161 GeV), ZH(240–250 GeV), $t\bar{t}$(~365 GeV)
- Tera-Z (O(6×10^{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

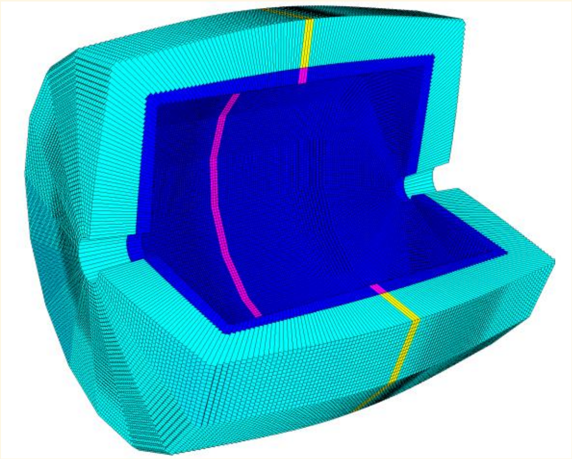
Bilevel Optimization

- **Scan the phase space (g,r)** of (detector geometry parameters, reconstruction algorithm parameters) against a physics objective

- **Pick** a detector, provide an algorithm+score metric, specify the optimization goal, physics process to simulate
- **Pipeline** generates/retrieves events, runs the varied simulations and algorithm templates, optimizes the specified parameters against the physics objective
- **Output:** The n-dimensional hypersurface for the objective (tensor of the loss landscape **L(g,r)**)

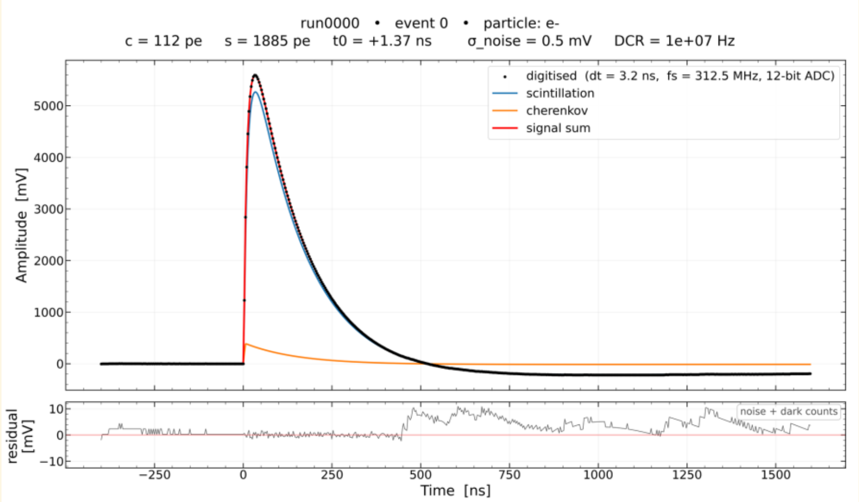


First Examples with DR Crystal ECAL



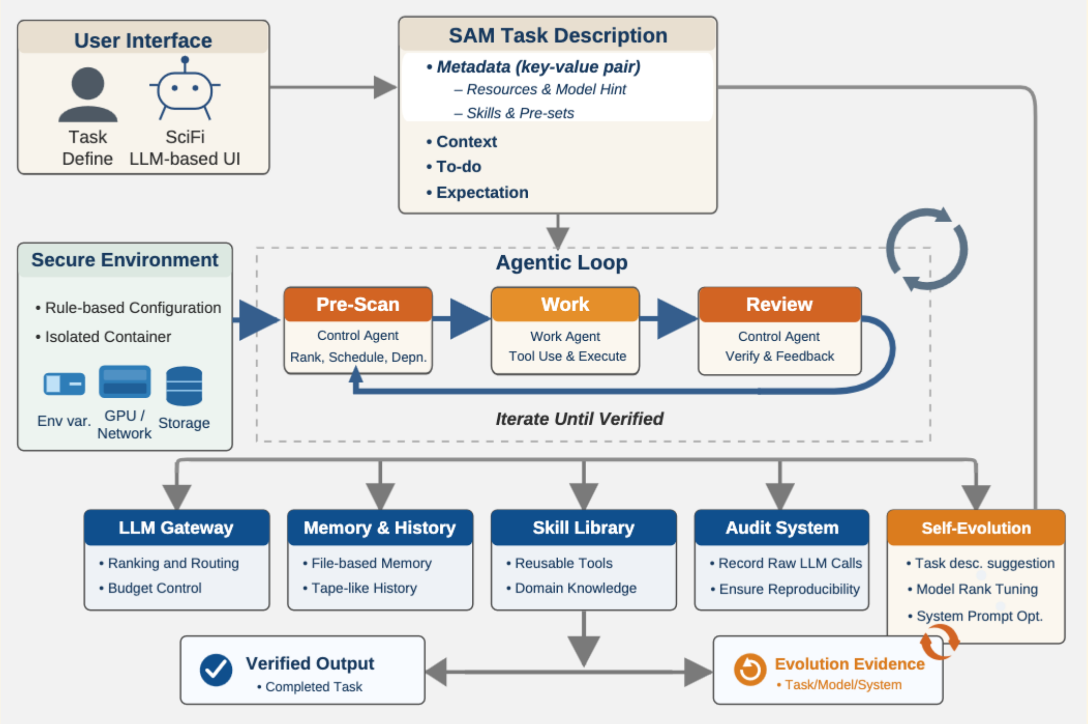
- **Detector:** IDEA dual-readout segmented crystal ECAL
- **Outer (N-param geometry):**
  - **Crystal width:** 1 cm → 2 cm, 1 mm steps, **Front crystal length:** 3 cm → 7 cm, 5 mm steps
  - **Projective offset:** 0 cm → 10 cm, 1 cm steps
  - **Events:** 1-20 GeV electrons in barrel region, N=~1000
- **Inner (N-param clustering):**
  - Simple center-of-gravity clustering based on shower radius, hit energy threshold
- **Inner (N-param digitization)**
  - Analog single-photon SiPM templates computed per S/C readout channel
  - Waveform scaled to hit photon counts, shifted by arrival time, dark counts/noise/DC baseline added, sampled at configurable rate/window
  - Signal mapped to configurable ADC dynamic range/bit depth
- **Simple physics score:**
  - Assume per-cell noise adds in quadrature, maximize signal-to-noise:

$$\text{score} = \left\langle \frac{E_{\text{cluster}}}{\sigma_0 \sqrt{N_{\text{hits}}}} \right\rangle_{\text{events}}$$



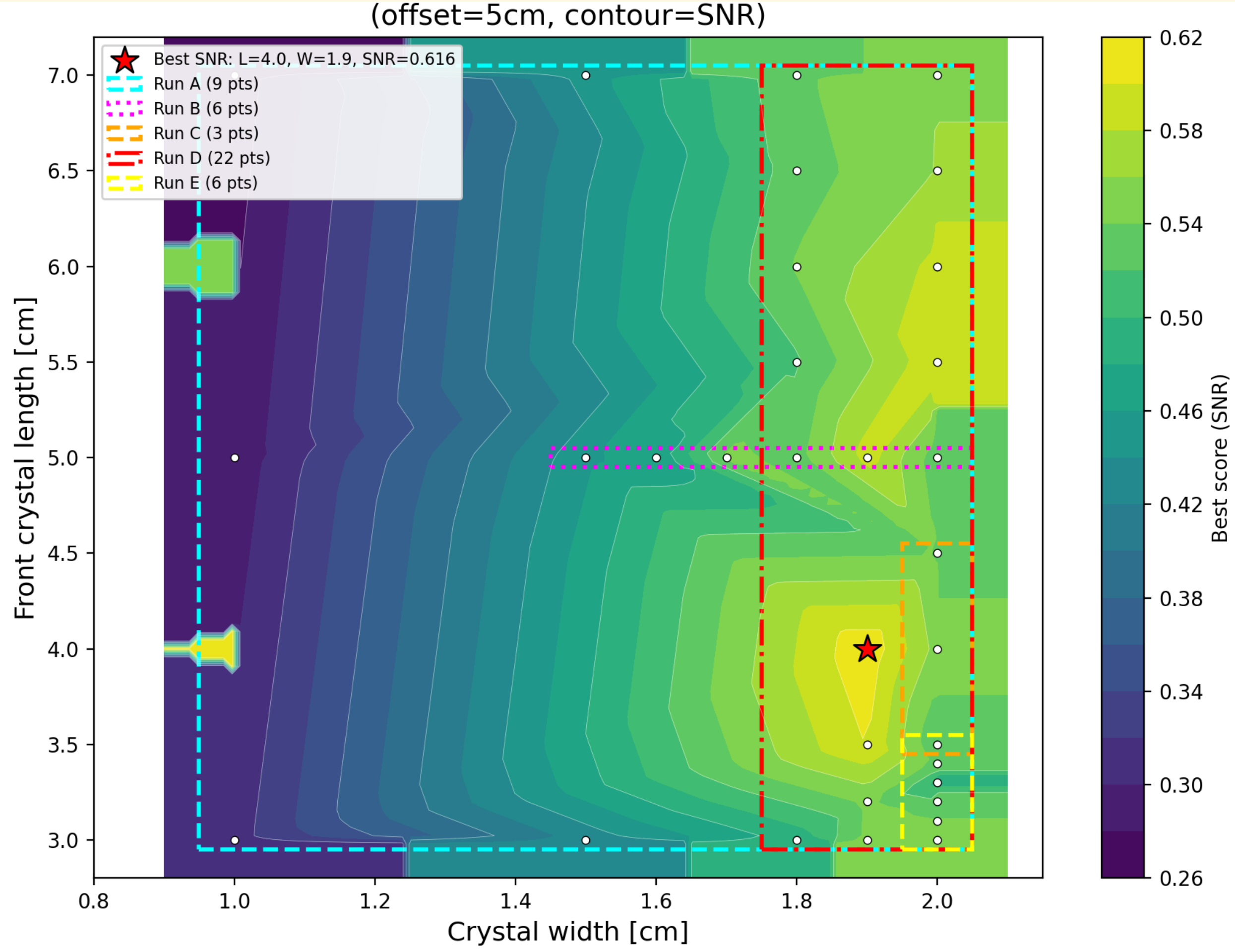
Driving with Agentic AI

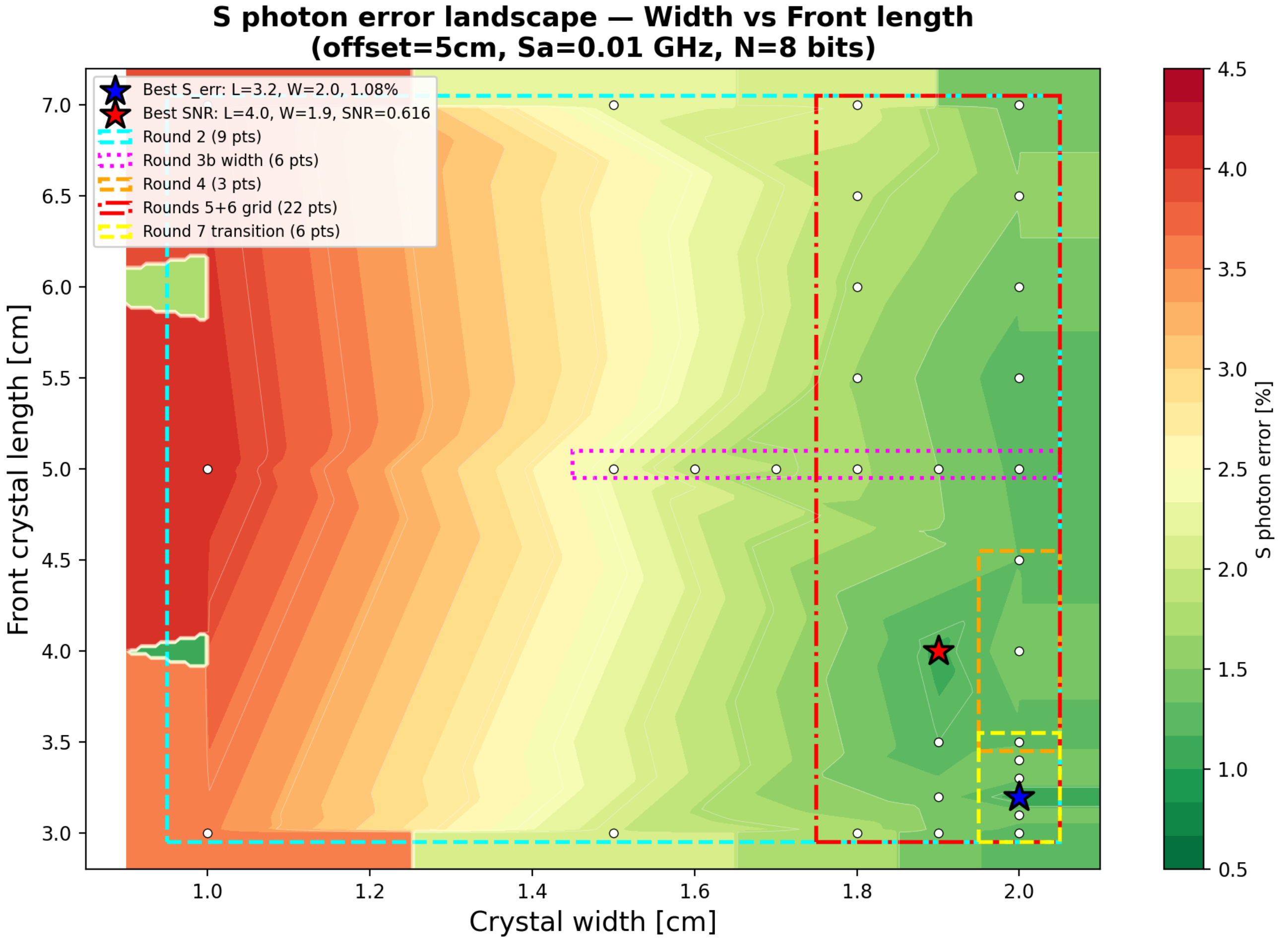
- **SciFi:** A Safe, Lightweight, User-Friendly, and Fully Autonomous Agentic AI Workflow for Scientific Applications
    - [arxiv:2604.13180]
    - [github:qibin2020/scifi]
- Write a task spec (context, todo, expect)
- System runs closed loop: Prescan → Agent loop → Independent review
- Different model verifies 'done' claim against 'expect' specification
- Agent analyzes and suggests next iteration steps/variables
    - Identified key (crystal width/length) vs. "nuisance" parameters (projective offset, E threshold, cluster radius)
    - Learned de/coupling between digi vs. non-digi params: broke down 11-param pool of iterations into stages: crystal width/length against SNR with nuisances, then ADC bit depth/sampling rate against S_photon error and ADC cost
    - Scan ranges, step sizes, event statistics all proposed by agent



First Examples with DR Crystal ECAL

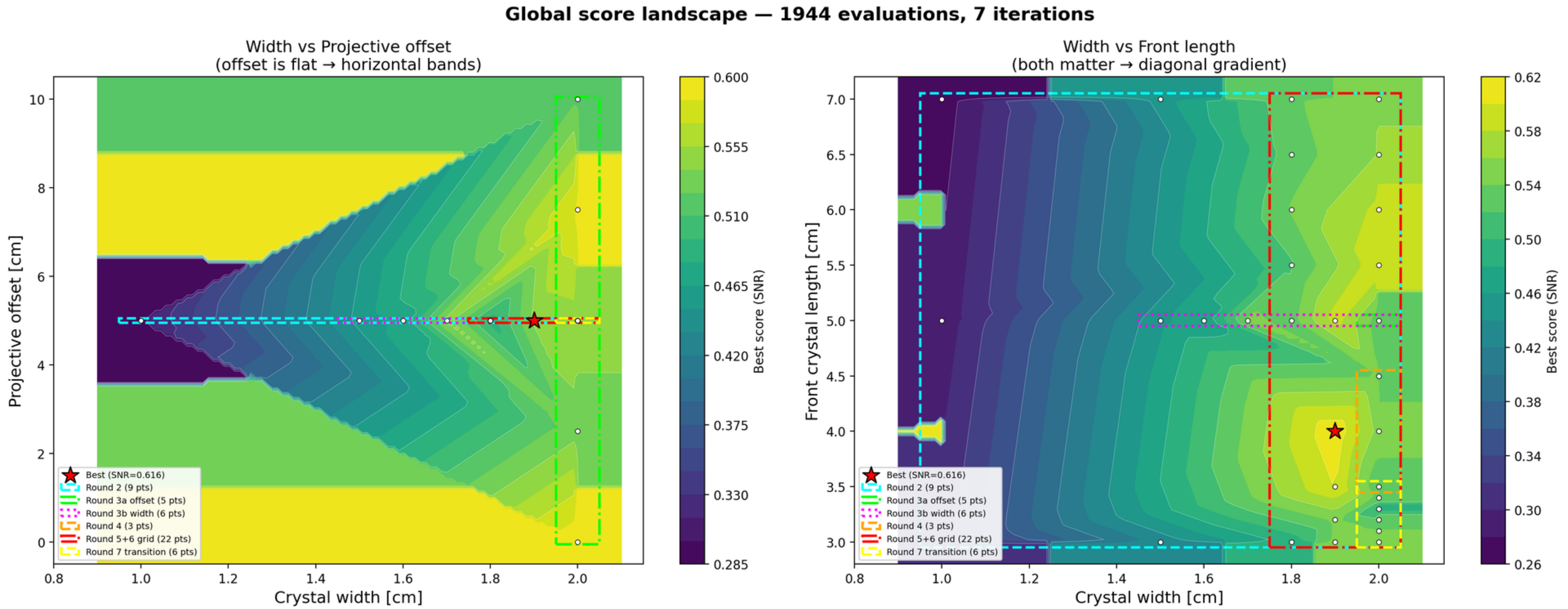
Crystal width vs. Front crystal length





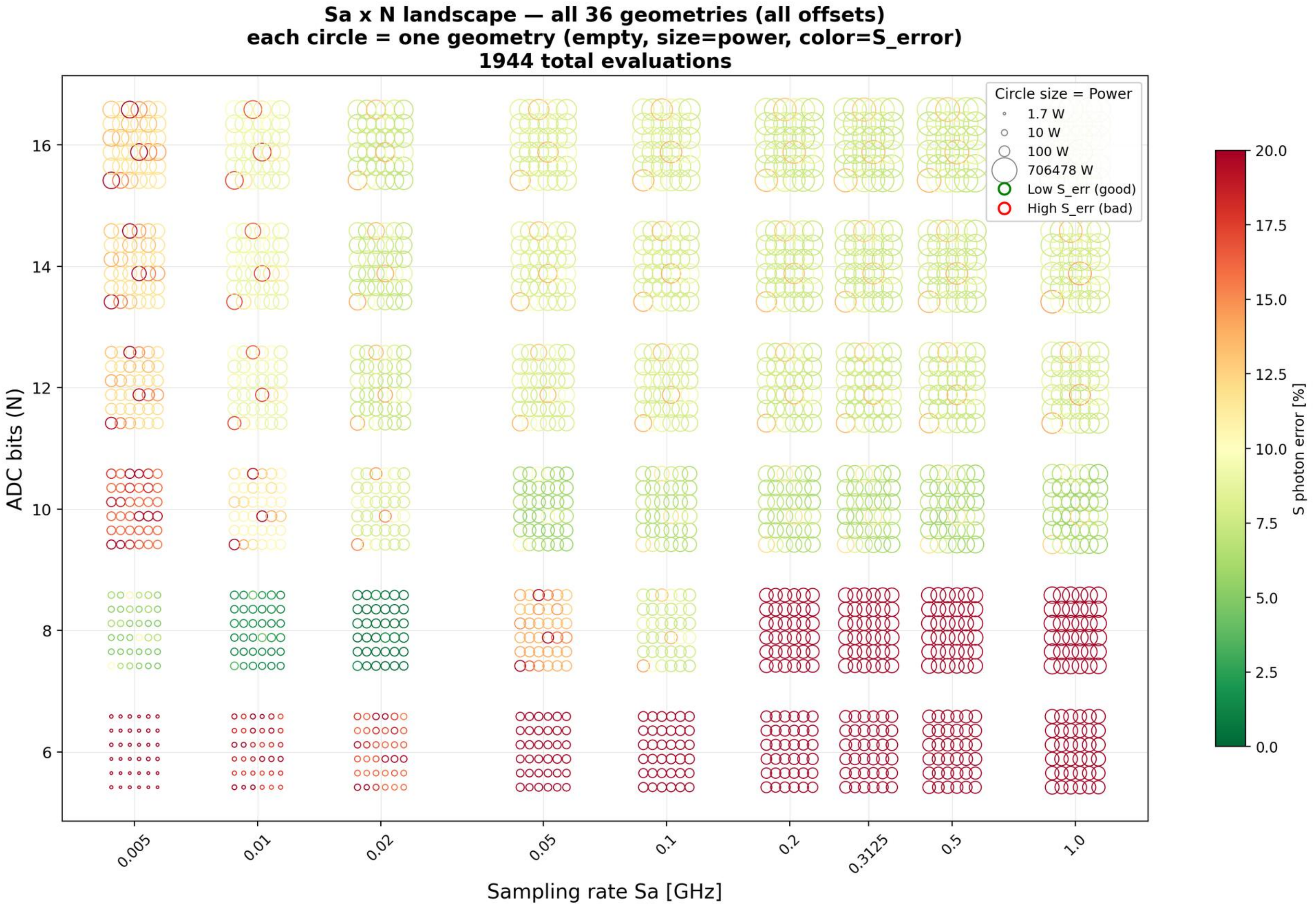
First Examples with DR Crystal ECAL

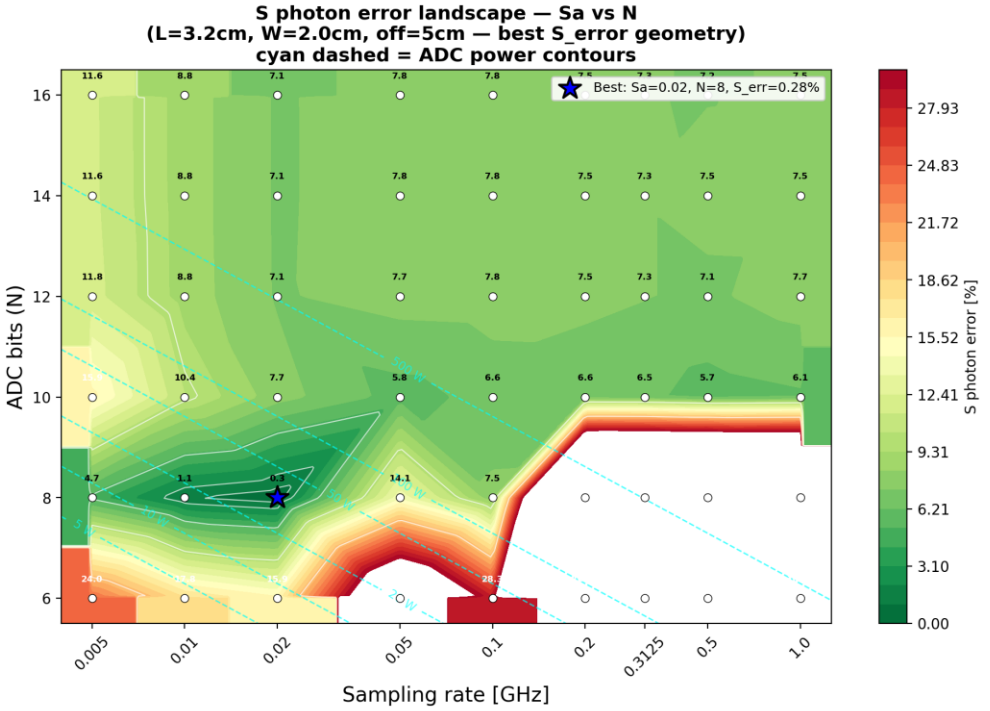
Crystal width vs. Projective offset

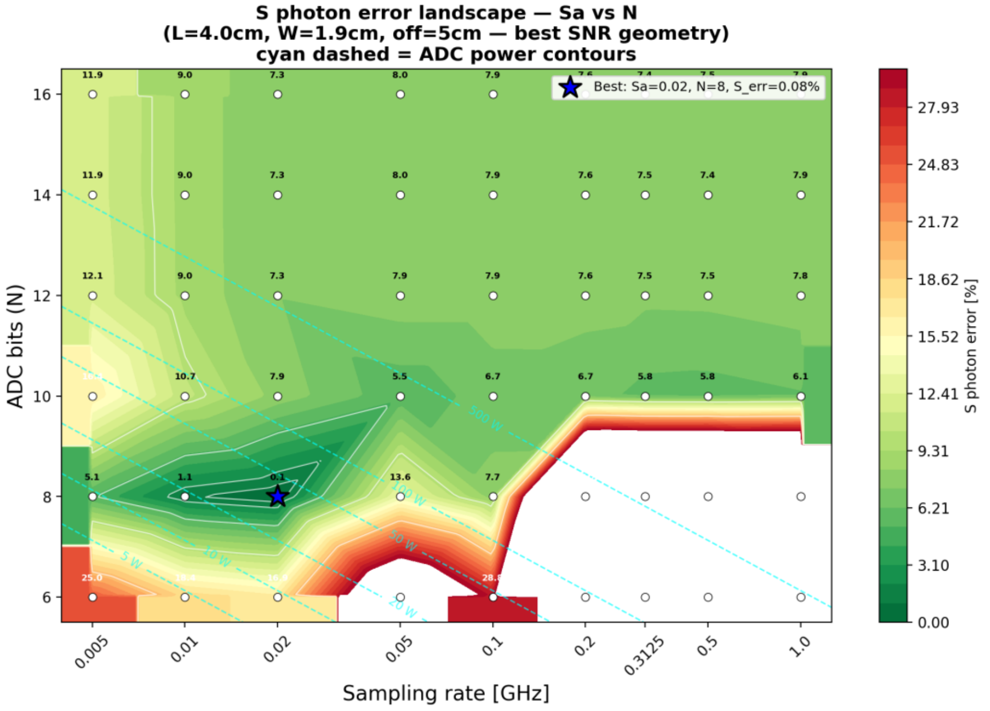


First Examples with DR Crystal ECAL

Sampling rate vs. ADC bits







Conclusions

- Human-input still needed for ~PhD+ physics-aware steering
- Statistical sampling methods of agents surprisingly good
- Efficient decoupling and splitting of multi-parameter spaces
- Huge savings in compute and labor
- Lowering iterative barriers produces new planes of data for human insight (a genuine value-add)
- Poster at Stanford Physics and AI conference
- Detailed ECAL-only study in prep for journal submission
- Full-IDEA FCC-ee analysis during summer targeting HHH sensitivity
- **V**ertically **I**ntegrated **B**ilevel Geometry, **R**econstruction, **a**nd **T**DAQ **O**ptimizer **(VIBRATO)**
- **Links**
  - [arxiv:2604.21804]
  - [github.com:wonyongc/bilevel_det_opt]
  - [arxiv:2604.13180]
  - [github:qibin2020/scifi]

arXiv:2604.21804v1 [physics.ins-det] 23 Apr 2026

Phenomenological Detector Design and Optimization in Vertically-Integrated Differentiable Full Simulations with Agentic-AI

Wonyong Chung
Princeton University
Jadwin Hall, Washington Road, Princeton, NJ 08544
wonyongc@princeton.edu

Qibin Liu
SLAC National Accelerator Laboratory
2575 Sand Hill Road, Menlo Park, CA 94025
qibin@slac.stanford.edu

Liangyu Wu
Department of Physics
Stanford University
450 Jane Stanford Way, Stanford, CA 94305
liangyu.wu@stanford.edu

Julia Gonski
SLAC National Accelerator Laboratory
2575 Sand Hill Road, Menlo Park, CA 94025
jgonski@slac.stanford.edu

Abstract

We present the first implementation of AI agents into the design and optimization of detectors in high-energy physics experiments via a bilevel optimization framework that vertically integrates detector geometry, front-end digitization, and high-level reconstruction algorithm parameters in differentiable full simulations. Using the example of a dual-readout, segmented crystal EM calorimeter with a baseline resolution of $3\%/\sqrt{E}$, we investigate the capabilities and value propositions of AI agents in the identification and reduction of key detector parameters and in the nonlinear traversal of a given detector design's full parameter space. We find that LLM-based reasoning models today, without being given additional experiment-specific context, are able to effectively execute complex workflows and proactively suggest generic but relevant avenues for further study or improvement. Here, we demonstrate an AI agent's ability to use the workflow to simultaneously optimize a representative subset of vertically integrated detector parameters: crystal granularity and length, number of ADC bits and sampling rate, and center-of-gravity hit-clustering radius. We find that effective integration of agents into the complex workflows of frontier areas of research not only significantly reduces labor and compute, but opens up efficient avenues for computational validation of first-principles design choices. While the ability to make autonomous leaps of physics-motivated judgment or insight is not demonstrated in this work, this study defines the current frontier of experimental design methods in high-energy physics.

Precision physics → detector performance drivers

- Run
- Tera
- √s di
  taggi
- Heav
- HZ r
- Elect
- ALP/
- 30 m
- 2 T m
- Cont
- Parti

HC
n (
L, (
es
ow
is

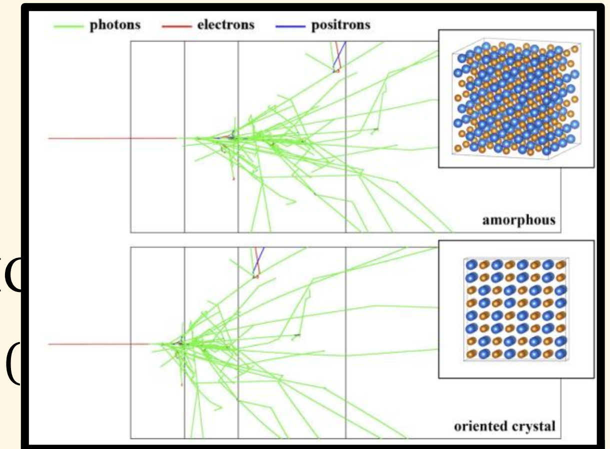
Detector design in the AI/ML era

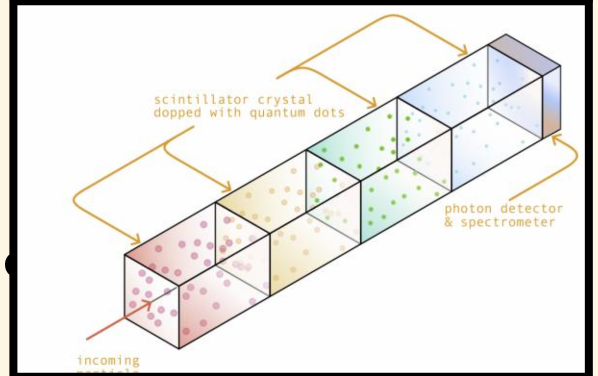
- Today: modular, flexible simulation
- Unified, top-down view of geometry and data schemas
- Hot-swappable subdetectors
- Triggerless DAQs
- Real-time inference on ASICs
- Picosecond timing
- **1:1 reconstruction**
  - **Motivates: Information-based detector perspective**

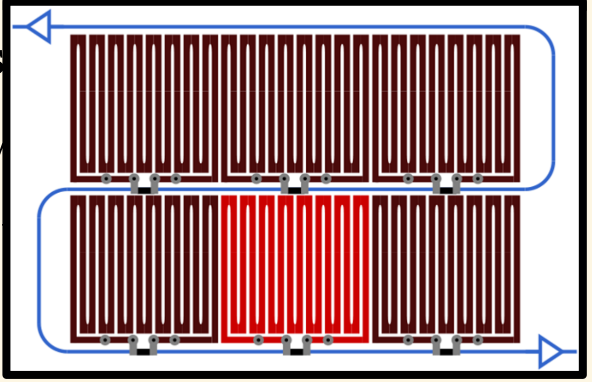
New technologies

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

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

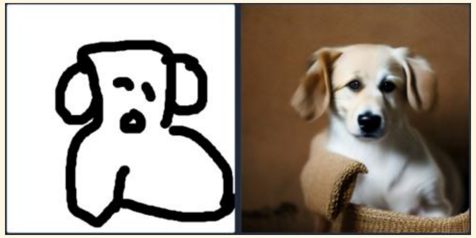




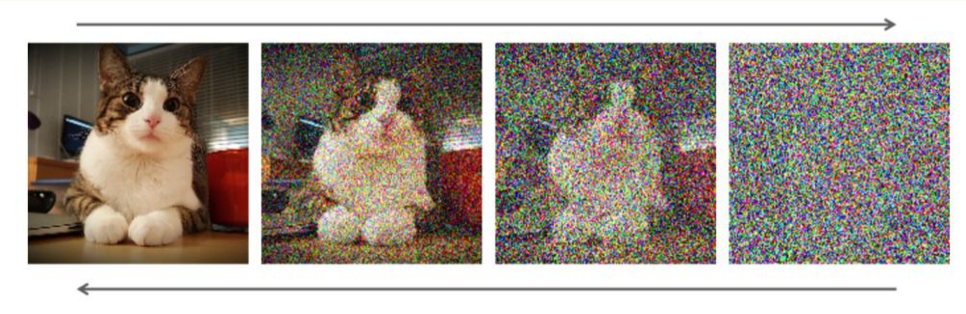


Pick the right neural network for the detector

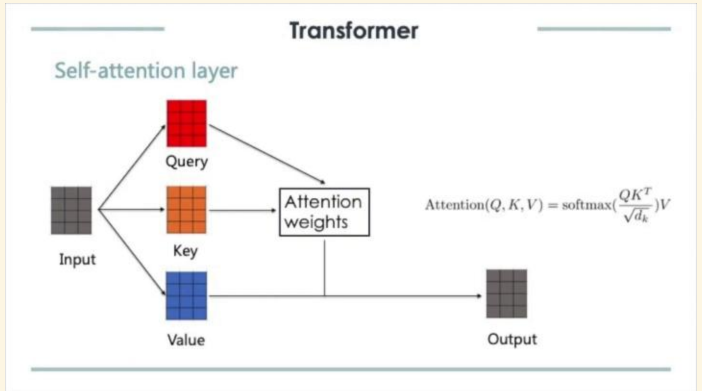
StableDiffusion + ControlNet



- **Tile-based sampling calorimeters:**
  - Diffusion models – **inherently designed to reconstruct from noise**
  - ControlNet/Inpainting – can use **tracks as a guide**



  - Reconstructing from noise – potential to integrate systematics into NNs



- **Multi-segmented/homogeneous calorimeters:**
  - Transformers – **designed to maintain long range context in sequential data, i.e. NLP**
  - Use **tracks for attention, segmentation and timing provide sequence**

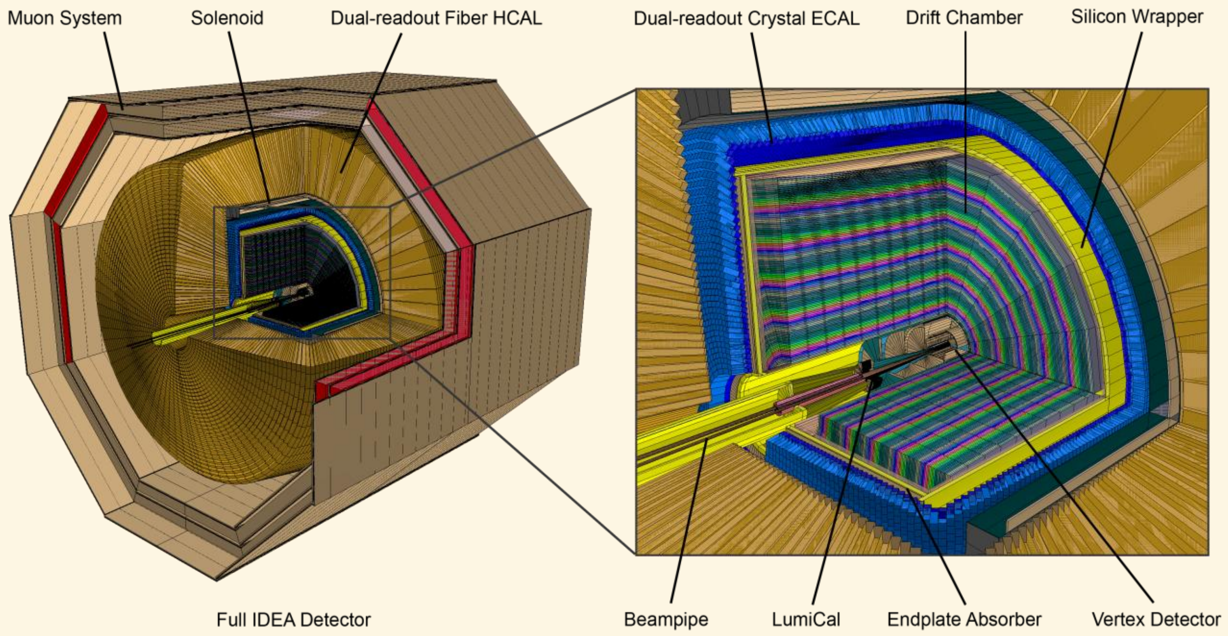
`The quick brown fox jumped over the lazy dog`

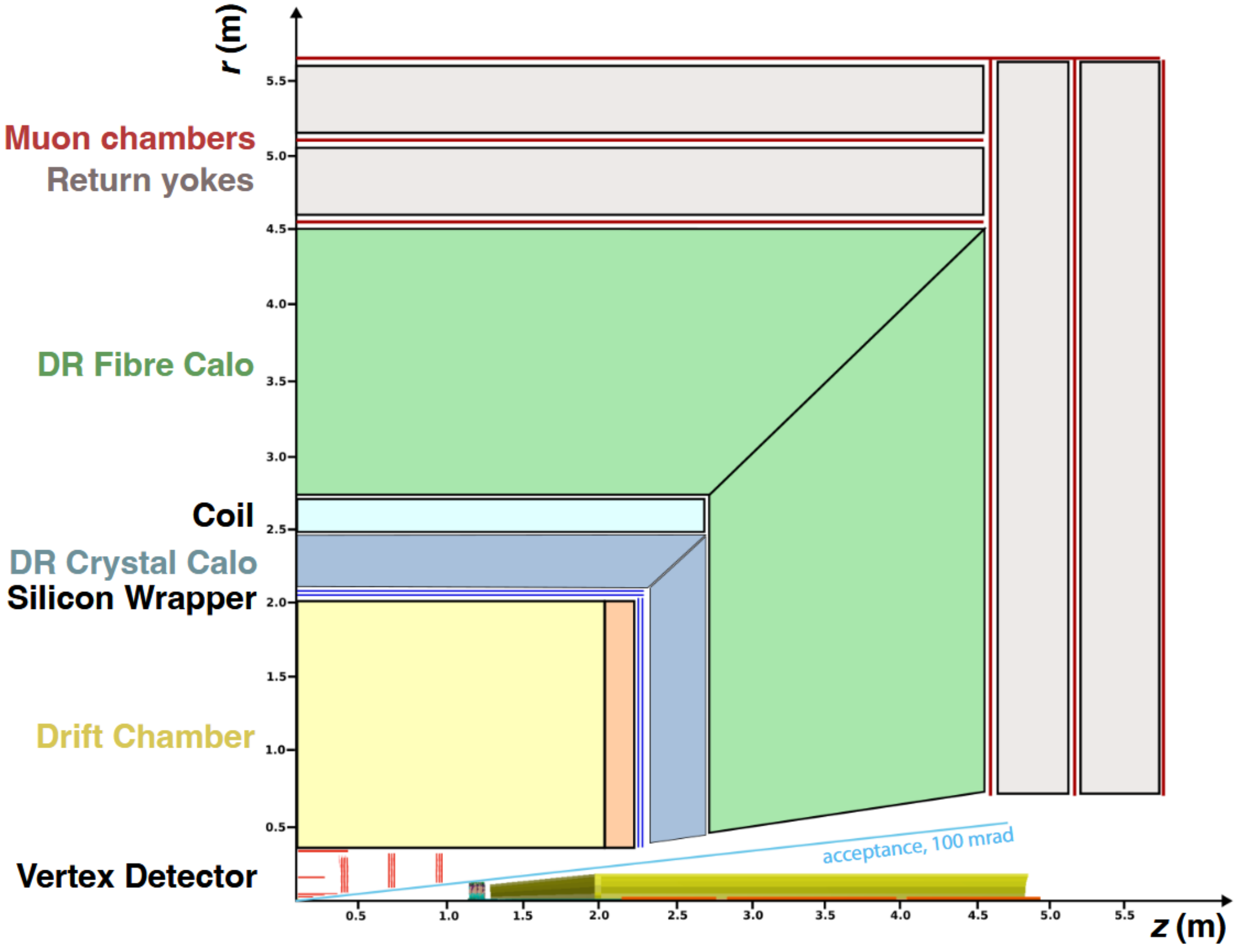
  - GNNs are generally short range and generally use distance, arguably same kind of assumptions as rule-based PFA

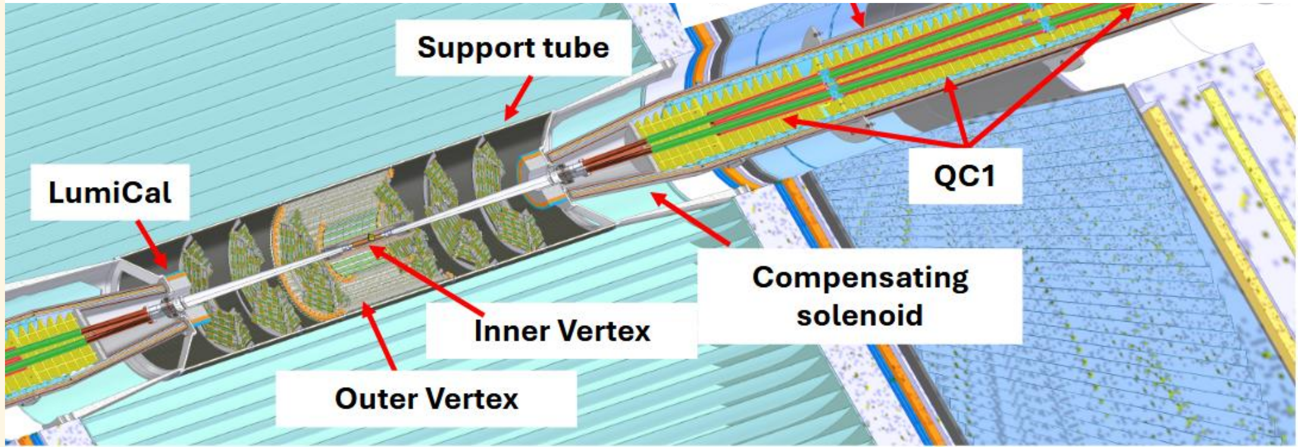
- **Mixed/"Semi-Homogeneous"**:
  - Generative Adversarial Networks – Diffusion model in sampling portion, transformer/GNN in homogeneous, directly leverage the detector design

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





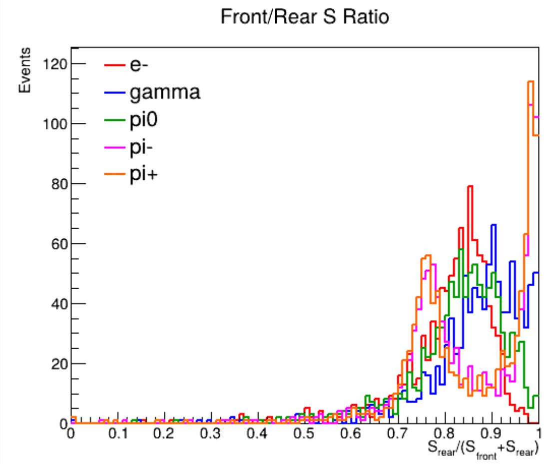


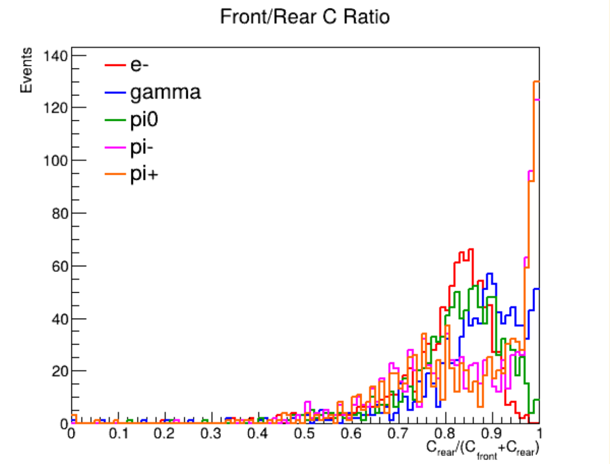
Dual-Readout Calorimetry

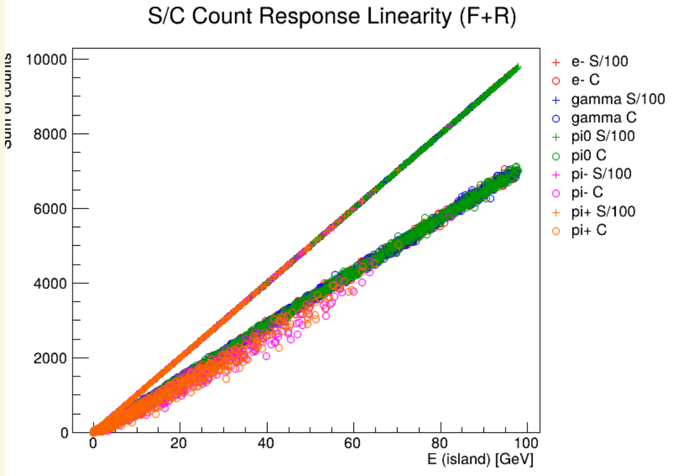
- **Calibrations:** 0-100 GeV e-, gamma, pi0, pi+, pi-
- **Technique:** Detect scintillation/Cerenkov light separately to mitigate event-by-event fluctuations in hadronic showers
- **Procedure:**
  - Calibrate on known EM/hadronic physics processes
  - Obtain the S/C response scaling factors
  - Determine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

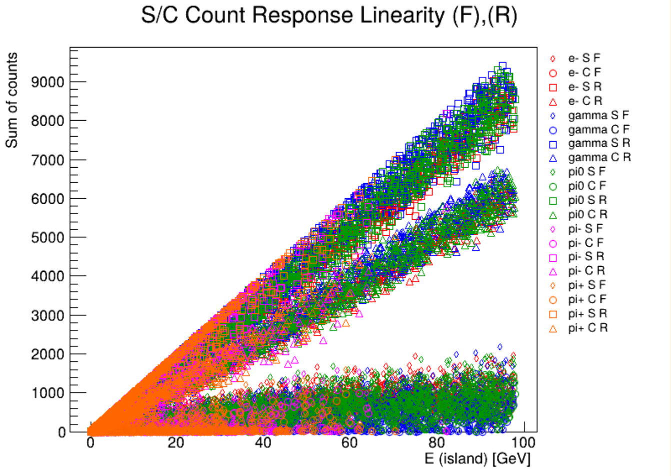
$$\begin{cases} S = E\left[f_{EM} + \frac{1}{(e/h)|_S}(1 - f_{EM})\right] \\ C = E\left[f_{EM} + \frac{1}{(e/h)|_C}(1 - f_{EM})\right] \end{cases}$$

- **Segmentation** enhances separation power





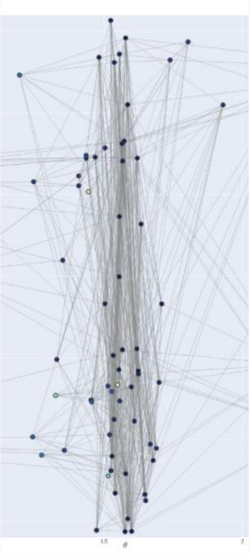




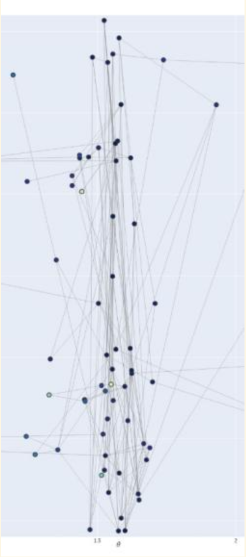
Dual-Readout Particle Flow

- 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**

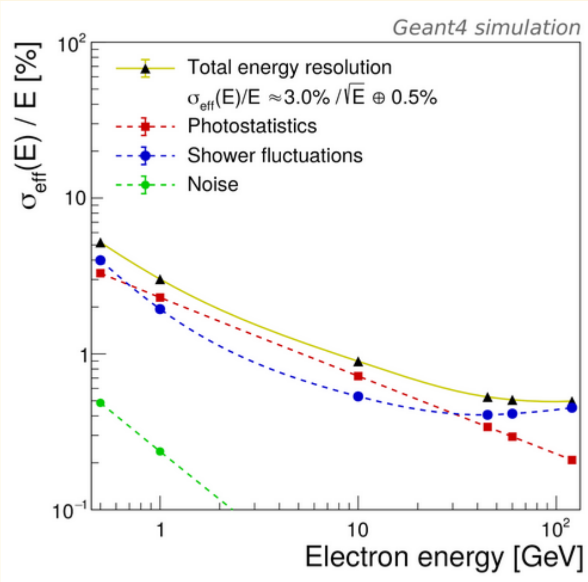
15%/√E

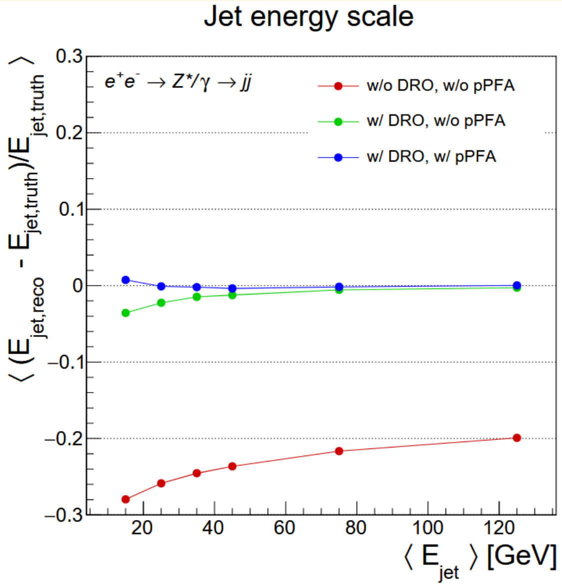


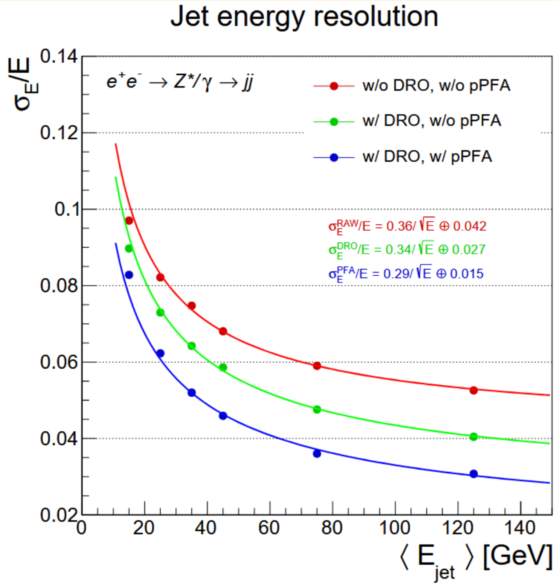
3%/√E

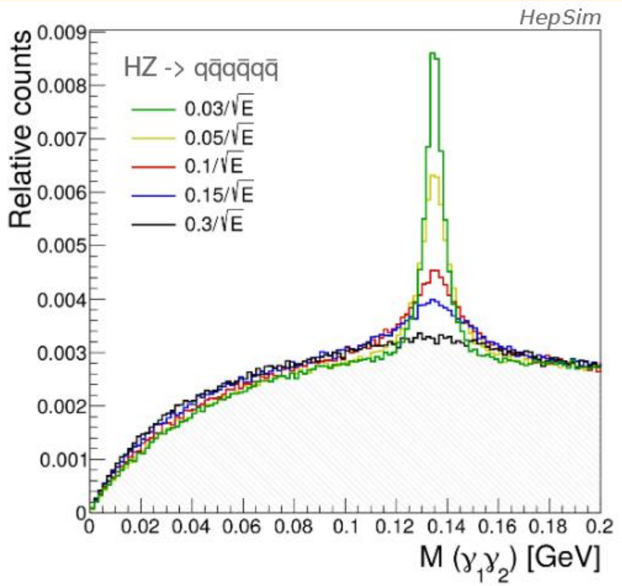


[2008.00338 , 2202.01474]



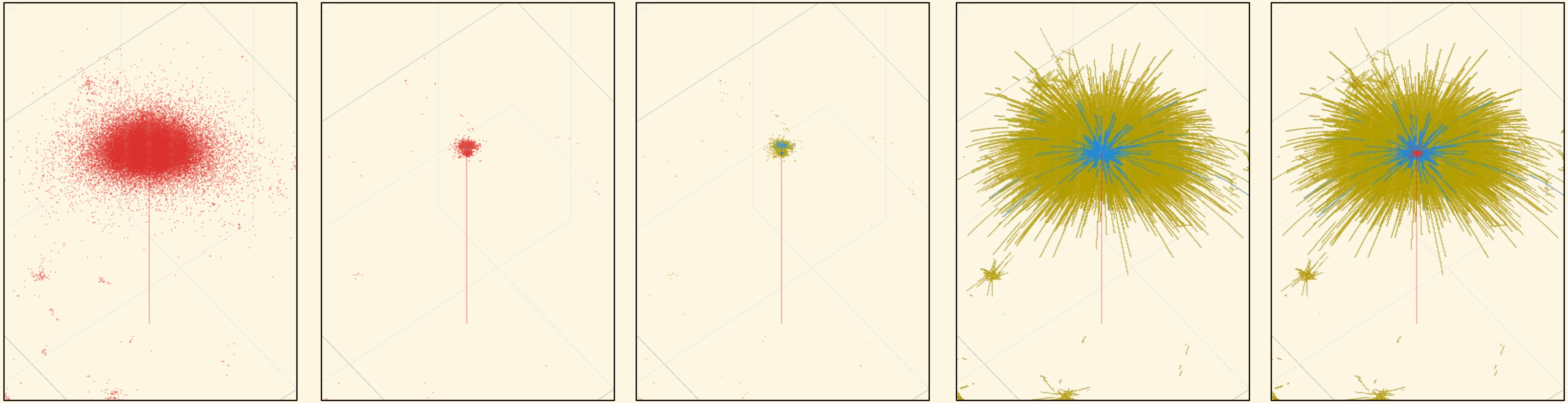






Synthetic Representations of Detector Response

- 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:



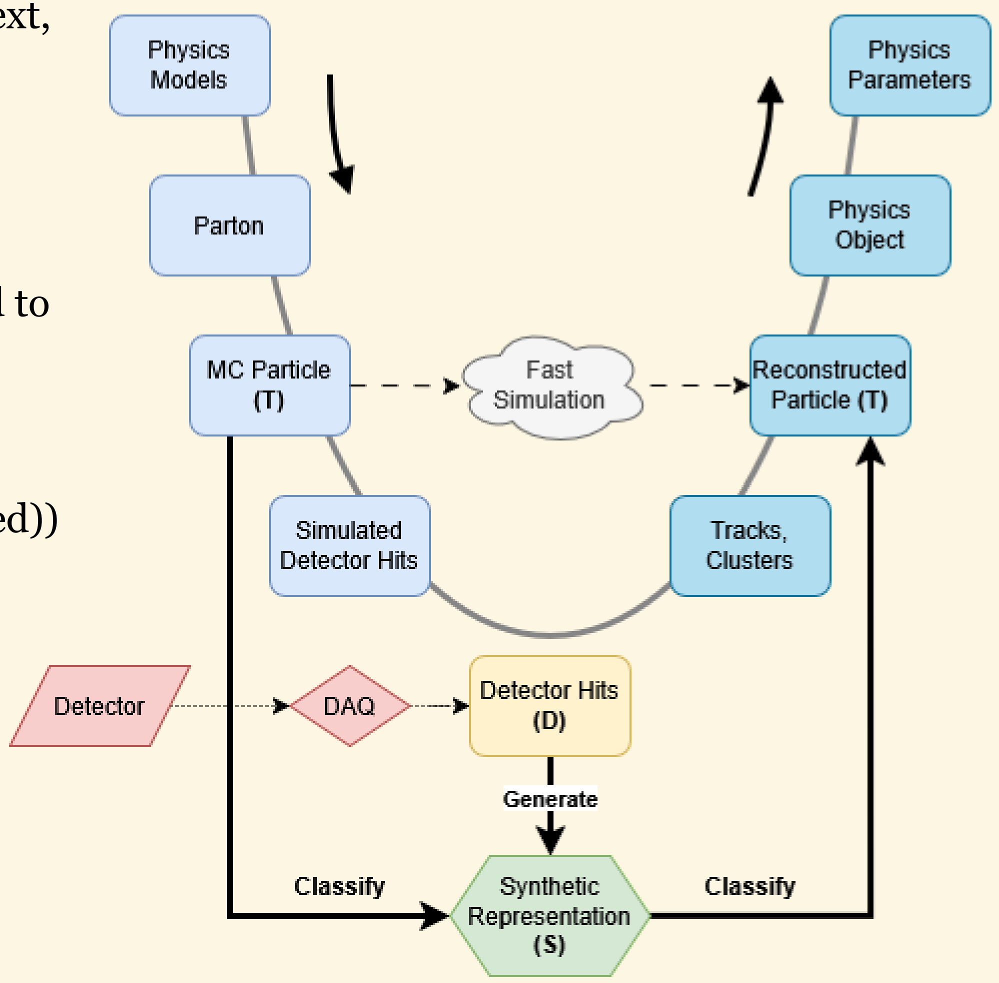
edep 0 | edep 1 keV | S/C counts | S/C tracks | S/C tracks+edep 1keV

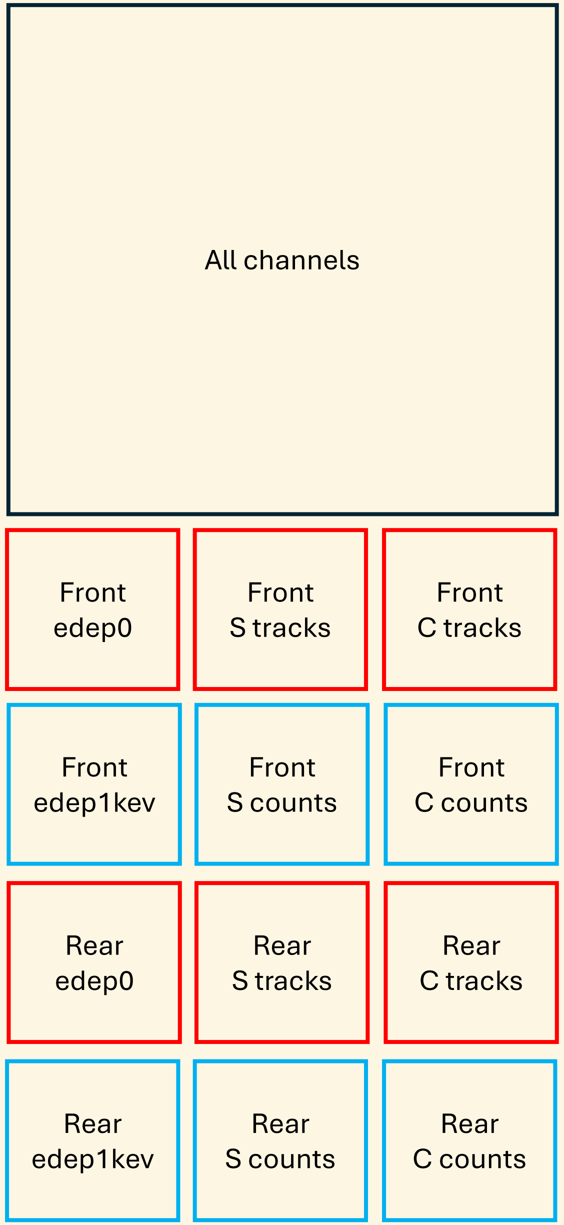
Representation Bridging in Reconstruction Domains

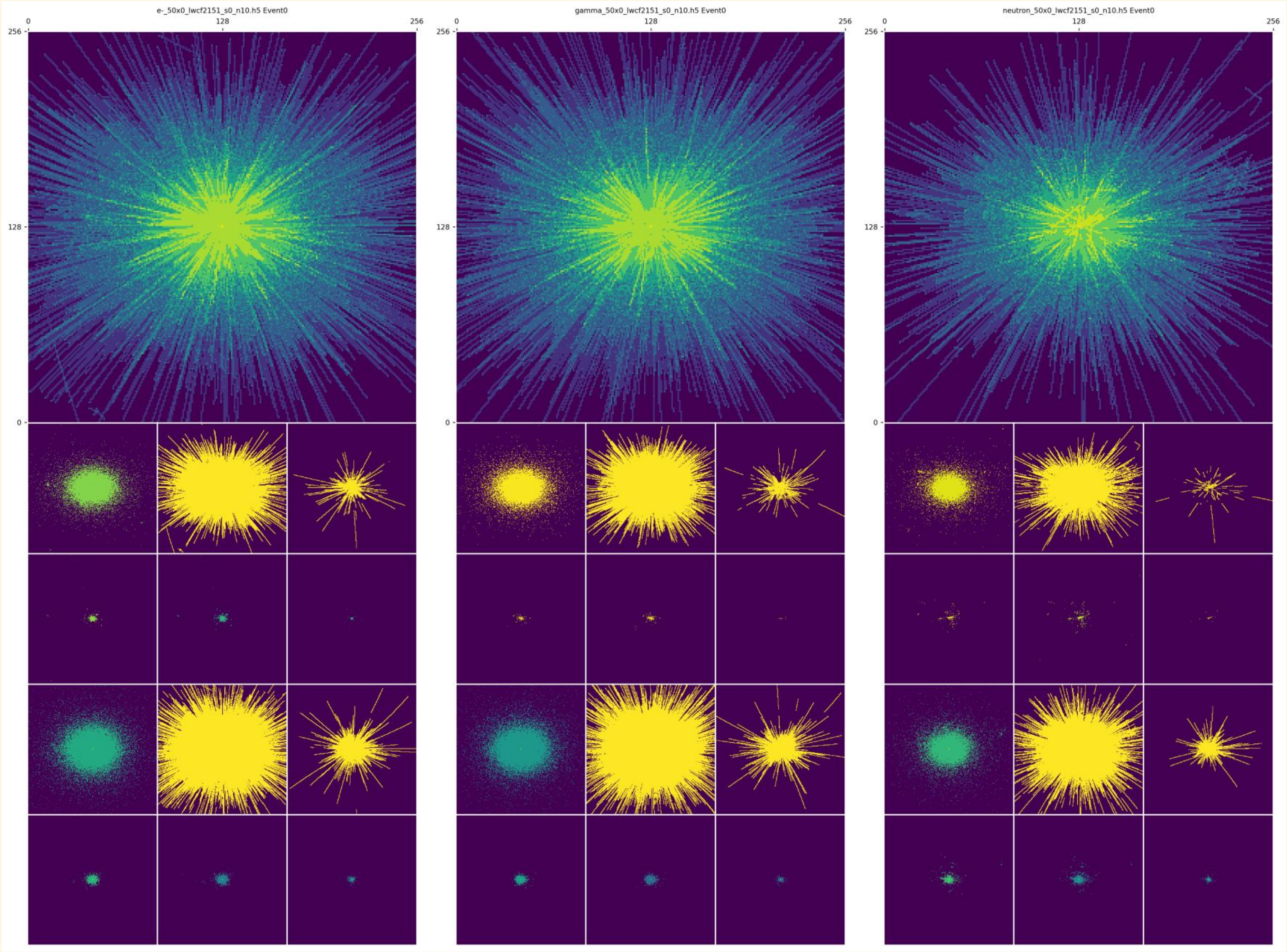
- S/C track hits are a form of **synthetic data**
  - *Unphysical* – won't ever see them in a real detector
  - *But not entirely unphysical* – are **representations of a physical process**
- Question: Can synthetic data be used in a meaningful way in reconstruction?
- Is there a need? *Yes* - a known 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 → 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



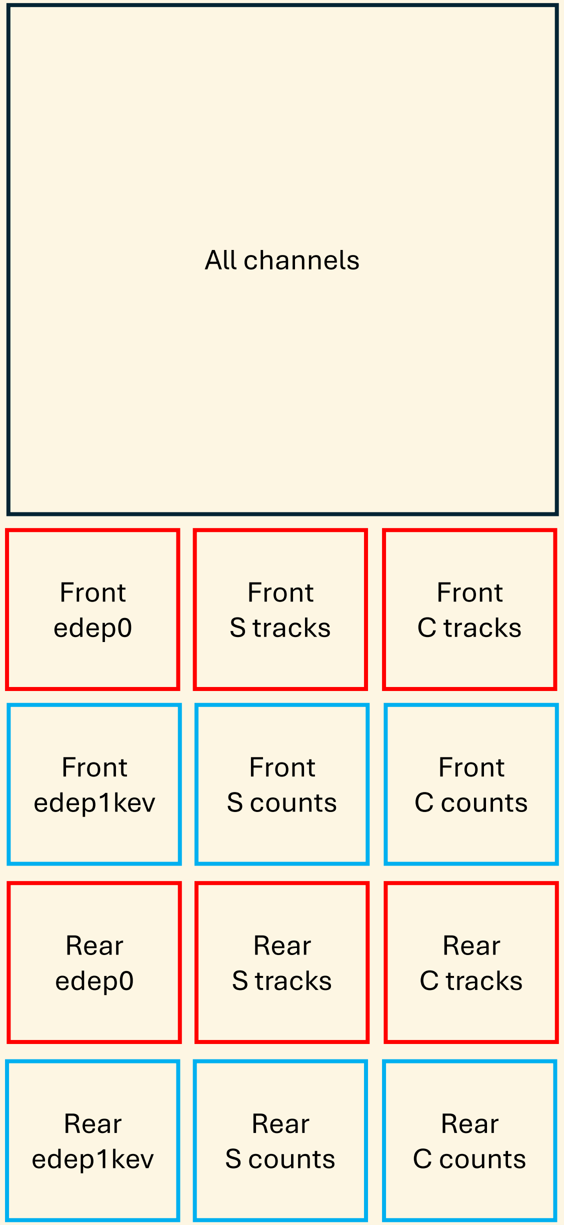


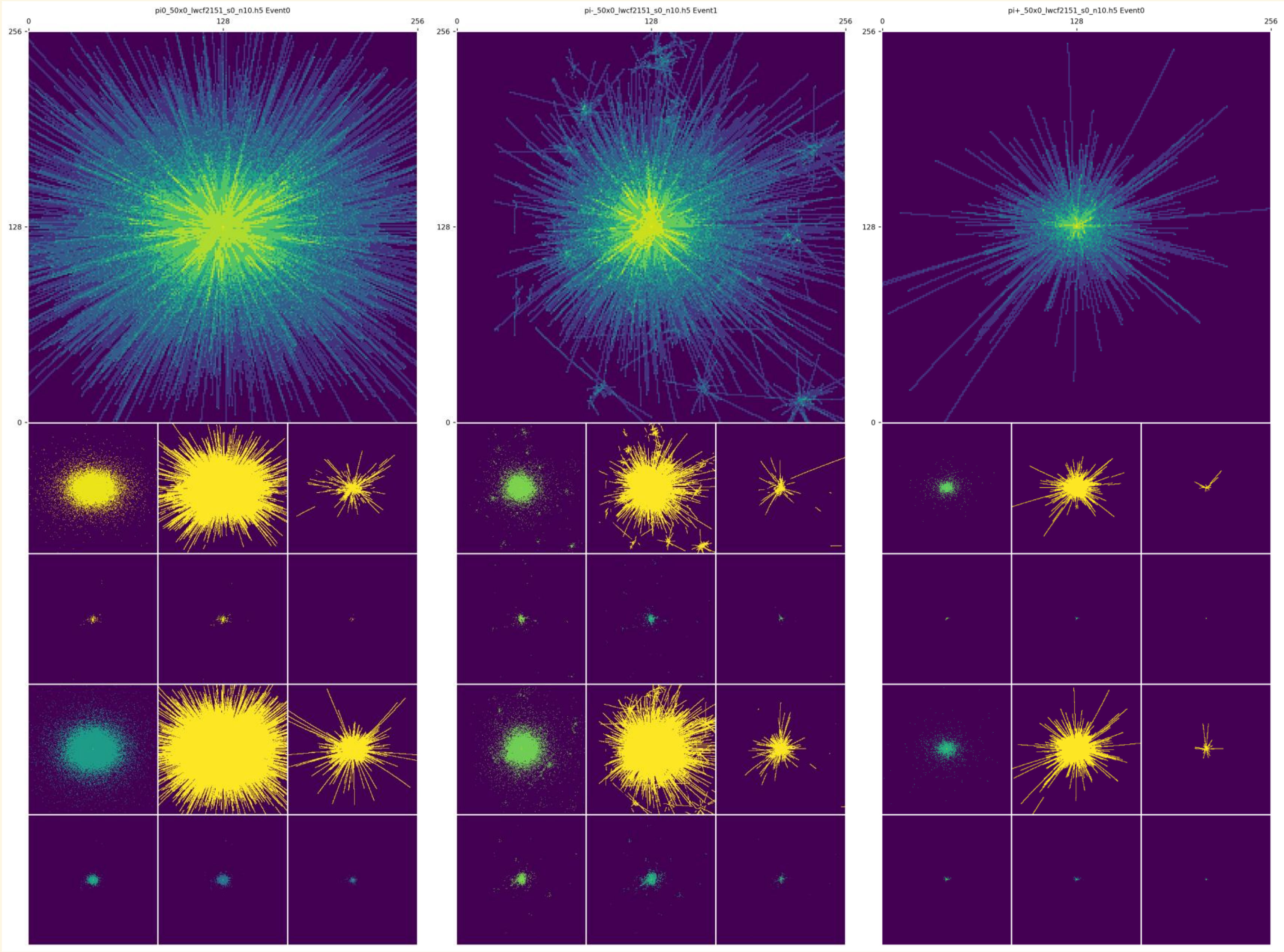


electron
50 GeV

gamma
50 GeV

neutron
50 GeV





π^0
50 GeV

π^-
50 GeV

π^+
50 GeV

Inference: First Look

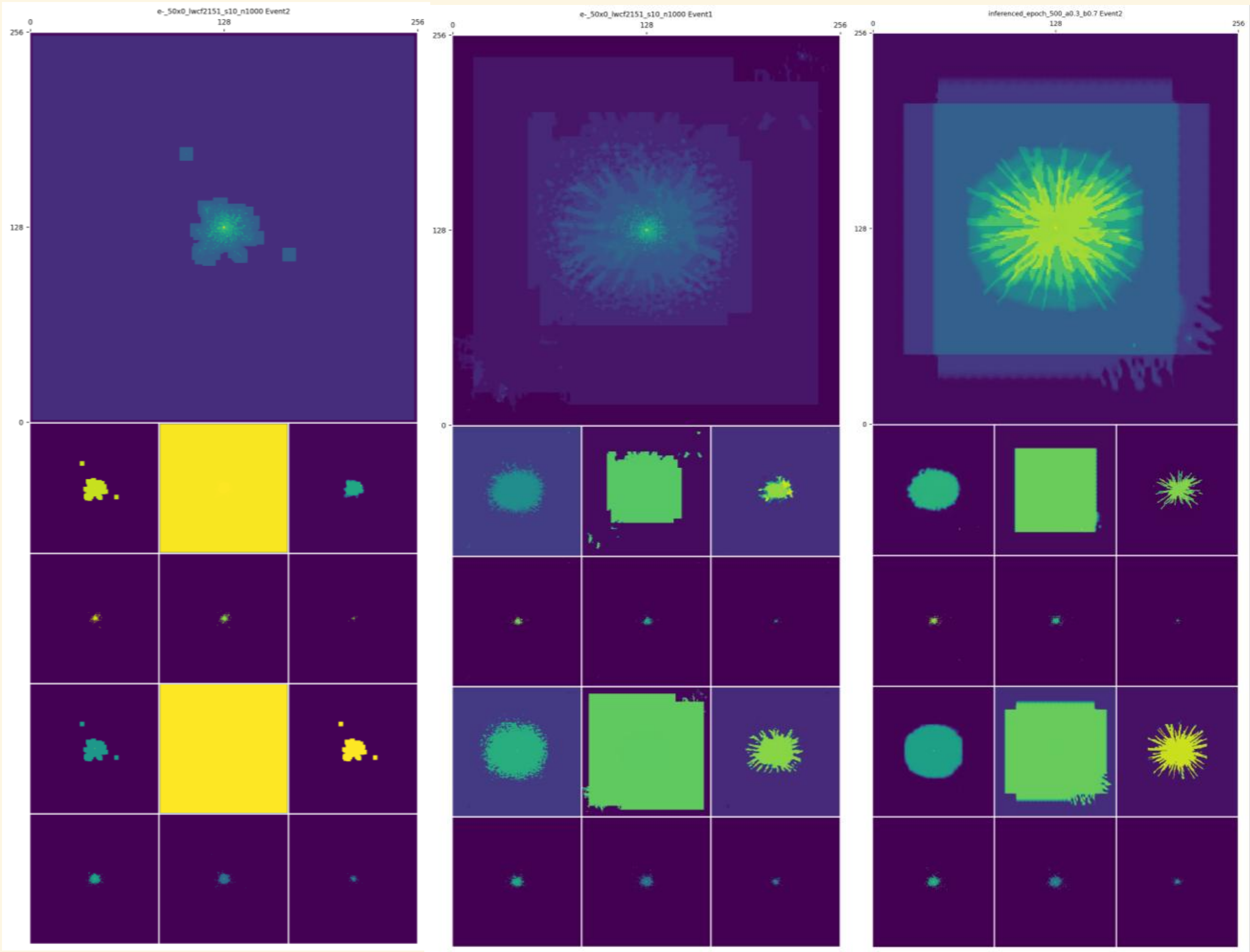
- Tuning hyperparameters is the challenge
- The Cerenkov signal gets resolved first – more sparse
- Scintillation signal is chipped away at

- 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/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, …