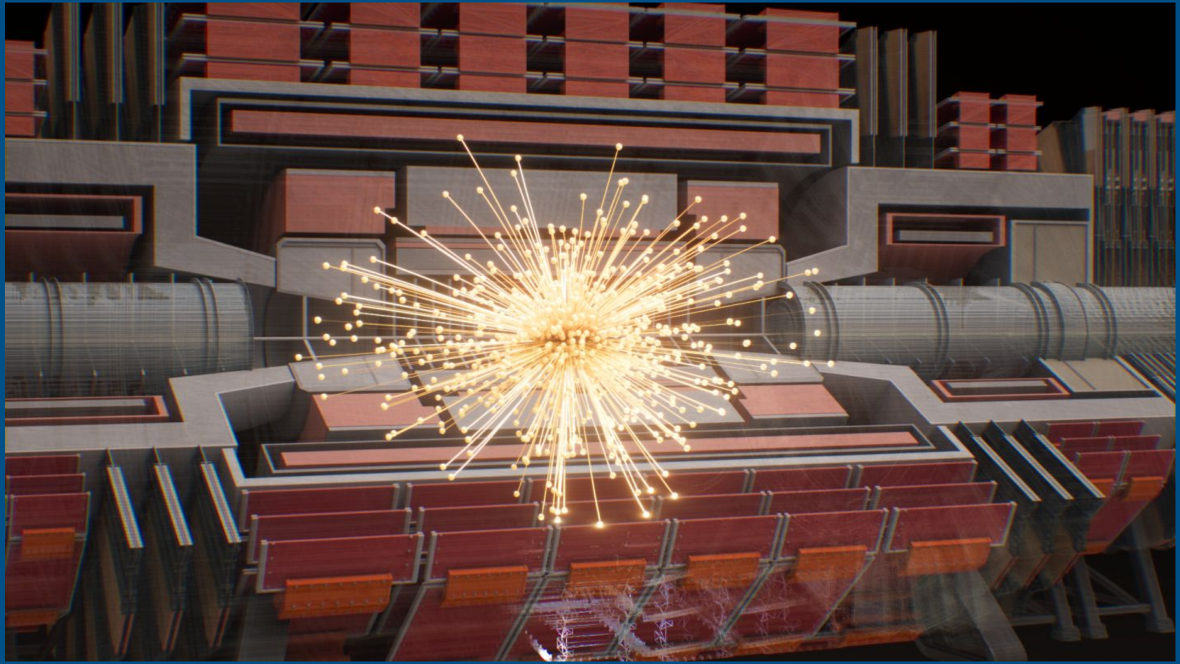
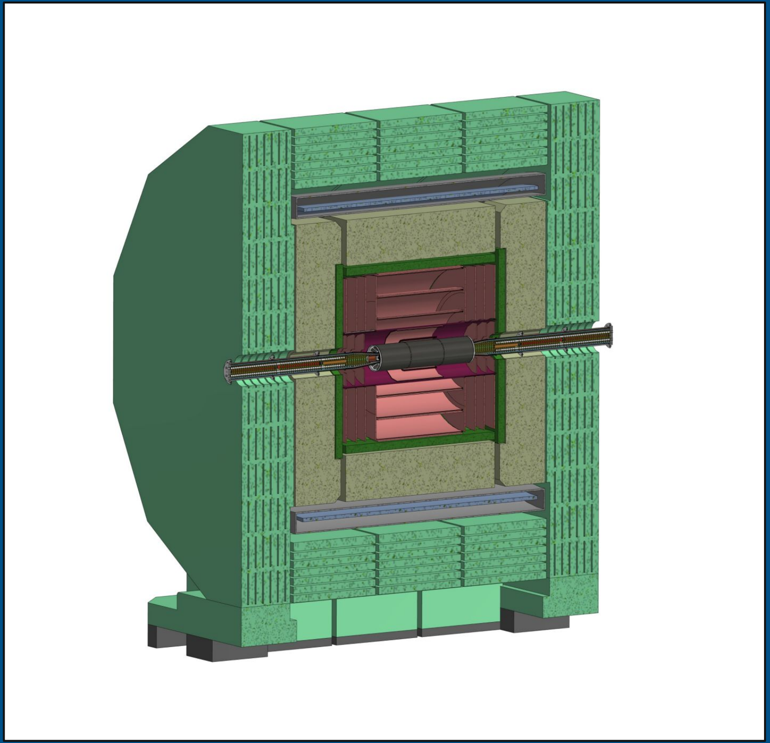
Cluster Geometry and Classification in FCC Vertex Detector



Emmett Forrestel | Summer 2025
Brown University, CERN

FCCee Context & Motivation

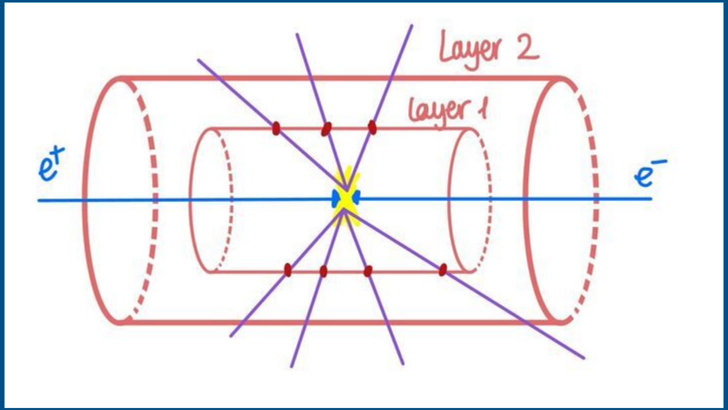
- The Future Circular Collider (FCCee) is a proposed e^+e^- collider operating at multiple energy stages, with the Z-pole expected to deliver unprecedented rates.
- The vertex detector must achieve exceptional spatial resolution and fast readout, both aided by suppression of beam-induced background.
- Beam background is most problematic, degrading vertexing precision and tracking algorithms, in the innermost layer where occupancies are highest.

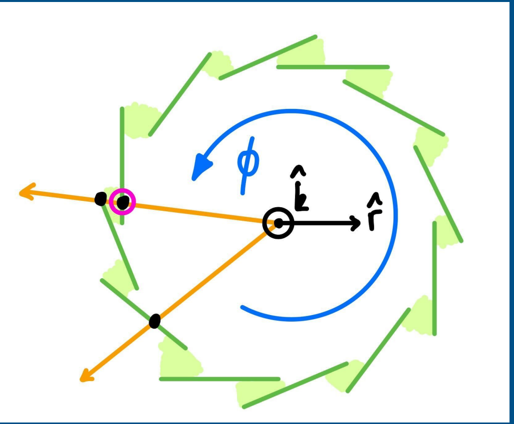


**CLD Cross Section

Project Focus

- Focused on hit clusters — groups of pixel hits deposited by a single Monte Carlo particle in Layer 1 of the CLD vertex detector.
- The detector was discretized into 25 μm × 25 μm bins, matching the Arcadia-MD3 pixel pitch, to reproduce the true readout granularity.
- This talk will follow my research path from:
  1. Identifying discriminative cluster features between signal and beam background.
  2. Exploration of intermodule crosstalk.
  3. Data preprocessing and classification algorithm development.

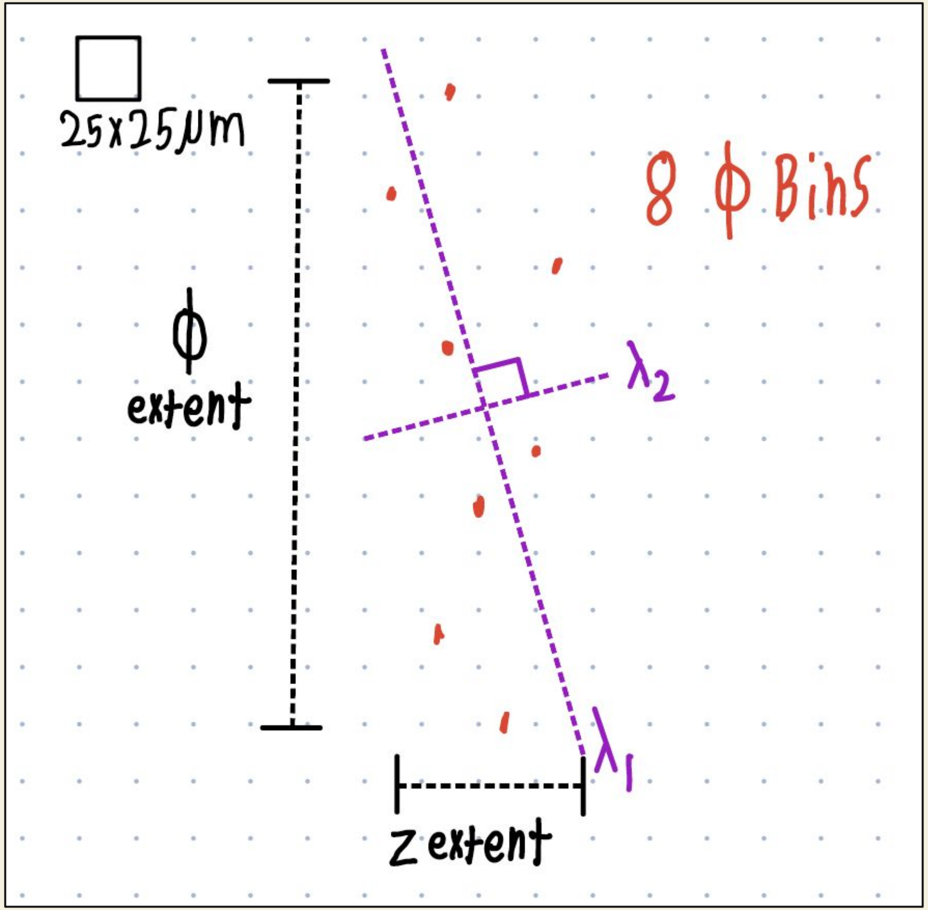




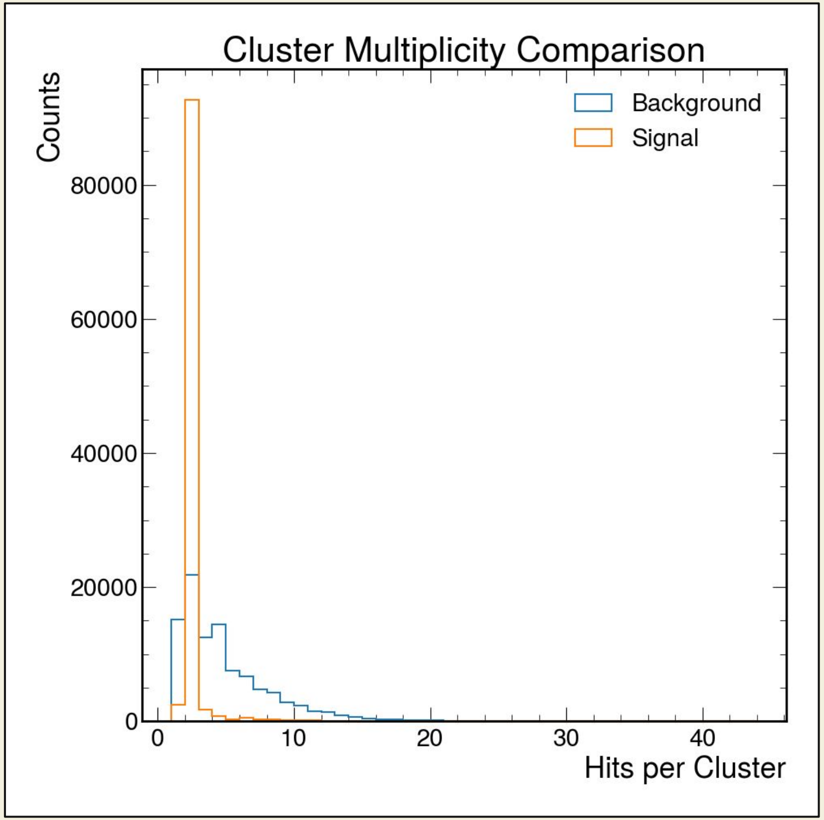
Background vs. Signal Feature Analysis

Cluster Metrics

- Six geometric and structural descriptors were extracted from each cluster:
  a. Energy Deposited
  b. Multiplicity
  c. cos(ϴ) of energy barycenter
  d. Φ Span
  e. Z Span
  f. PCA Elongation

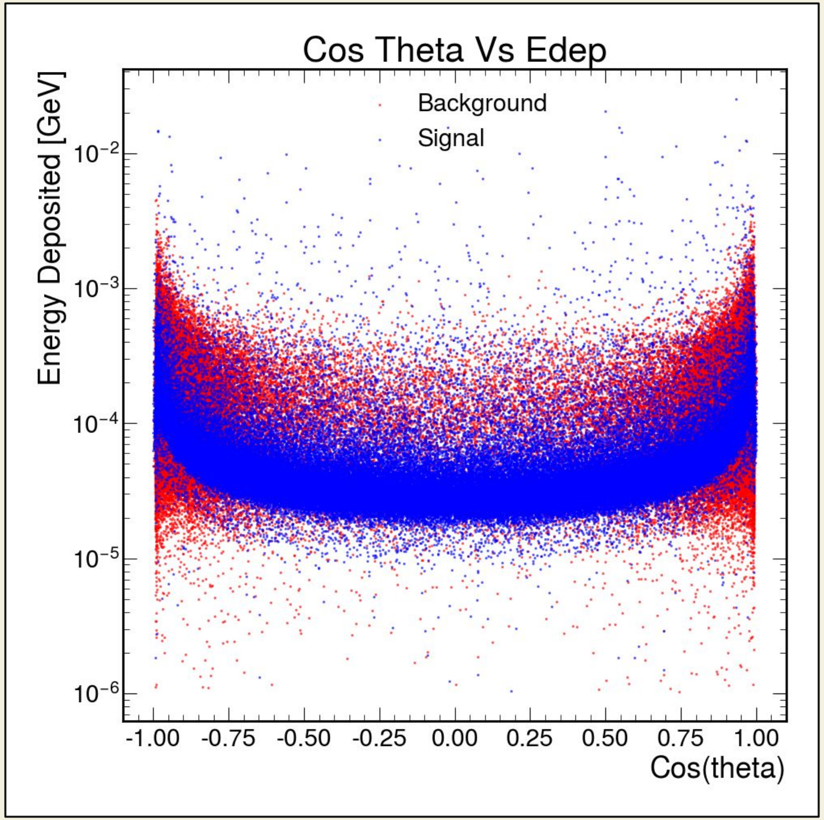


Cluster Multiplicity Comparison



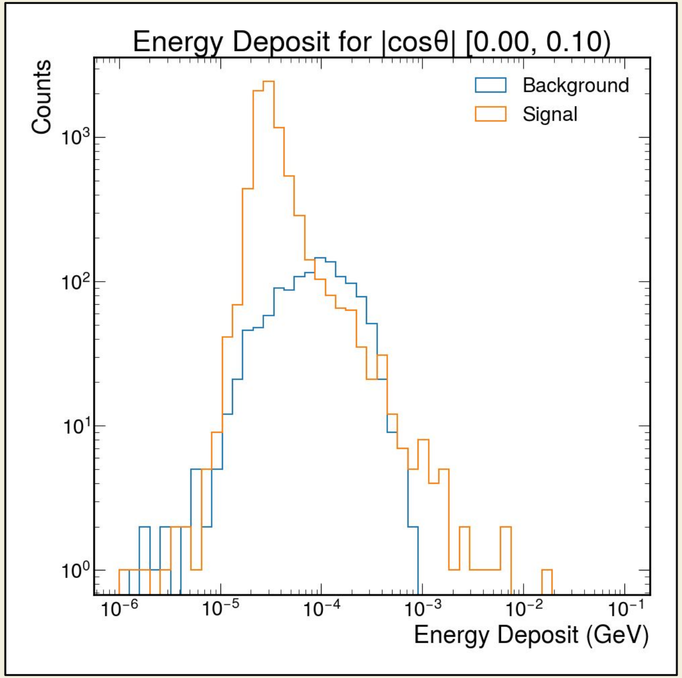
- ~ 100,000 clusters analyzed for both background and signal.
- Displays the number of hits each MC particle produces on the first layer of the vertex detector.
- Signal greatly peaks at two, while background is more dispersed.

Cos(Θ) vs. Energy Deposited

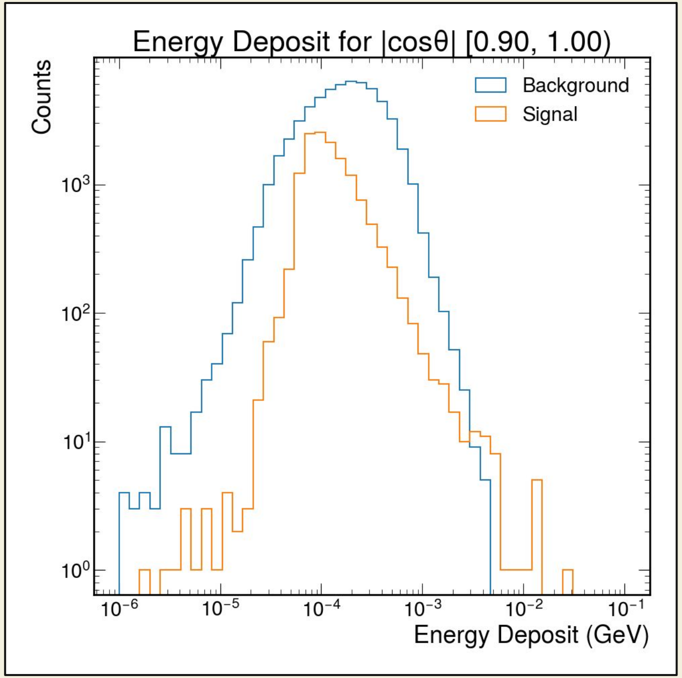


- Both signal and background and signal show higher energy deposits as |cos(Θ)|-> 1.
- Signal follows a much stronger association with cos(Θ), whereas background is more stochastic.

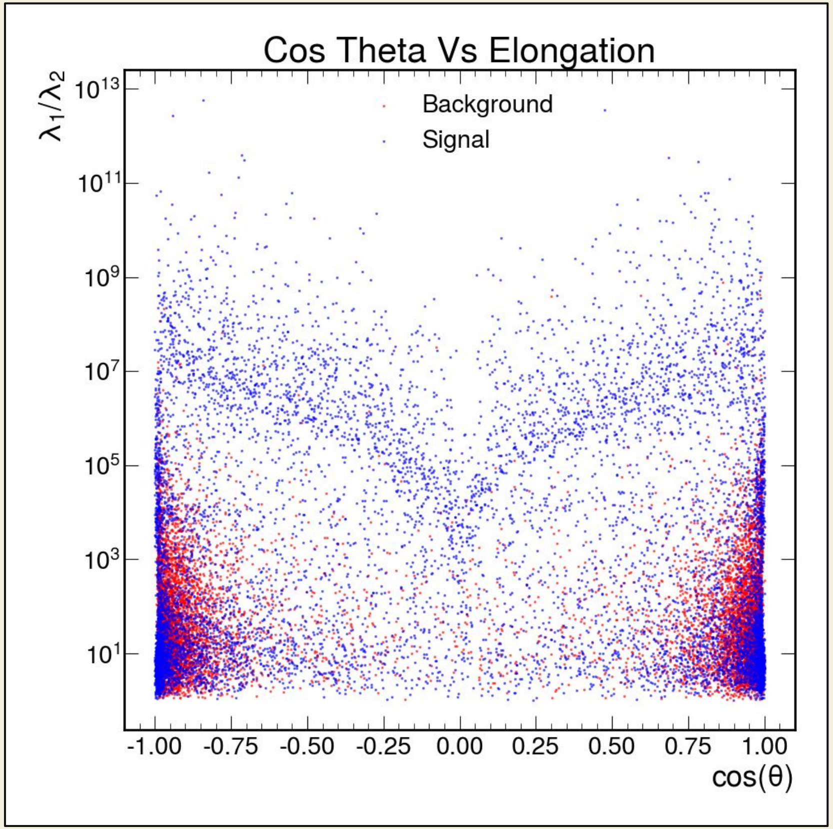
Energy deposited vs. Cos(Θ)



- As |cos(Θ)|-> 1, particles encounter more material, and background's higher prevalence here, explains its higher average energy deposit.

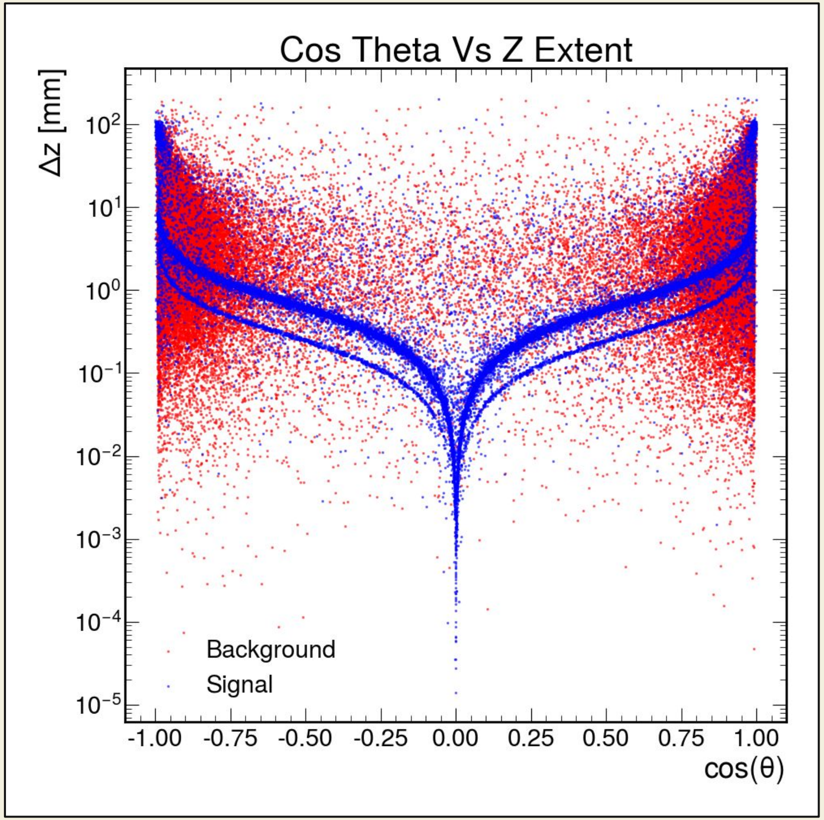


Elongation vs. Cos(Θ)



- With clusters of three or more hits, log(elongation) demonstrates a strong correlation with cos(Θ).
- Likely, lower energy clusters remain less linear, even at high theta, whereas higher energy signal clusters track very linearly, especially at high theta.

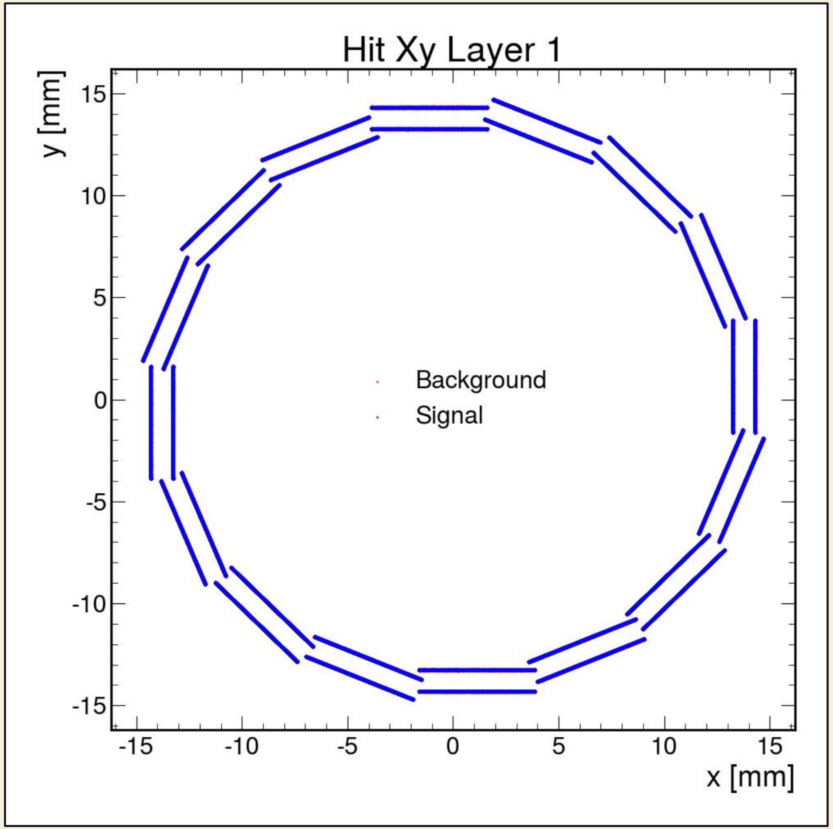
Z Extent vs. Cos(Θ)



- When plotting log(z_extent) (Z difference between max and min z in a cluster) against cos(Θ), signal clusters follow a very strong relationship.
- This relationship stands to be further investigated, especially the double bands in signal plot.

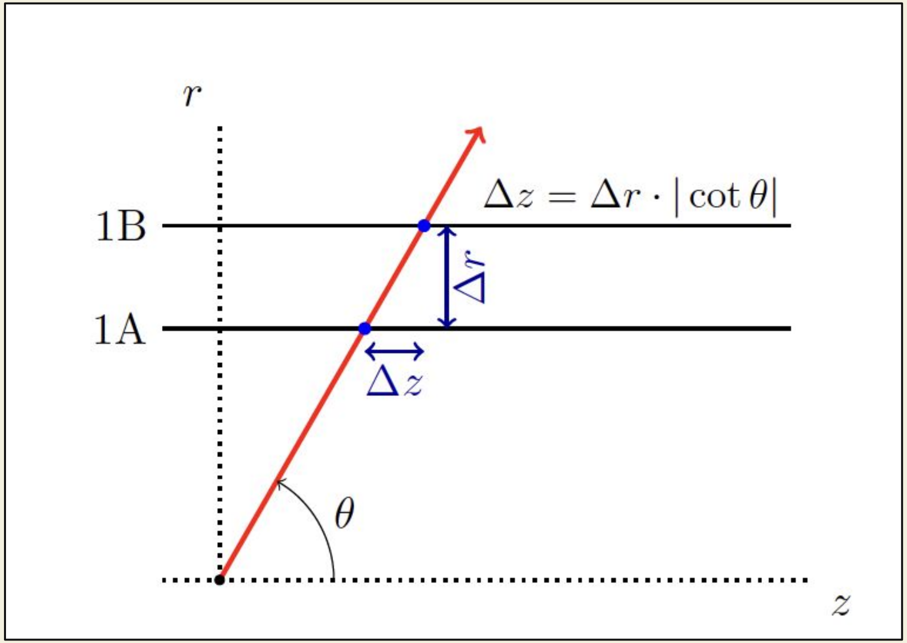
Intermodule Crosstalk Possibilities and Implications

Layers 1A and 1B



- My suspicion was that the double banded structure in Z Extent came from different radial traversals between clusters.
- When examining CLD structure a mechanism for these differing radii becomes apparent–we see two readout layers.
- The outer I refer to as, layer 1B, the inner as layer 1A.

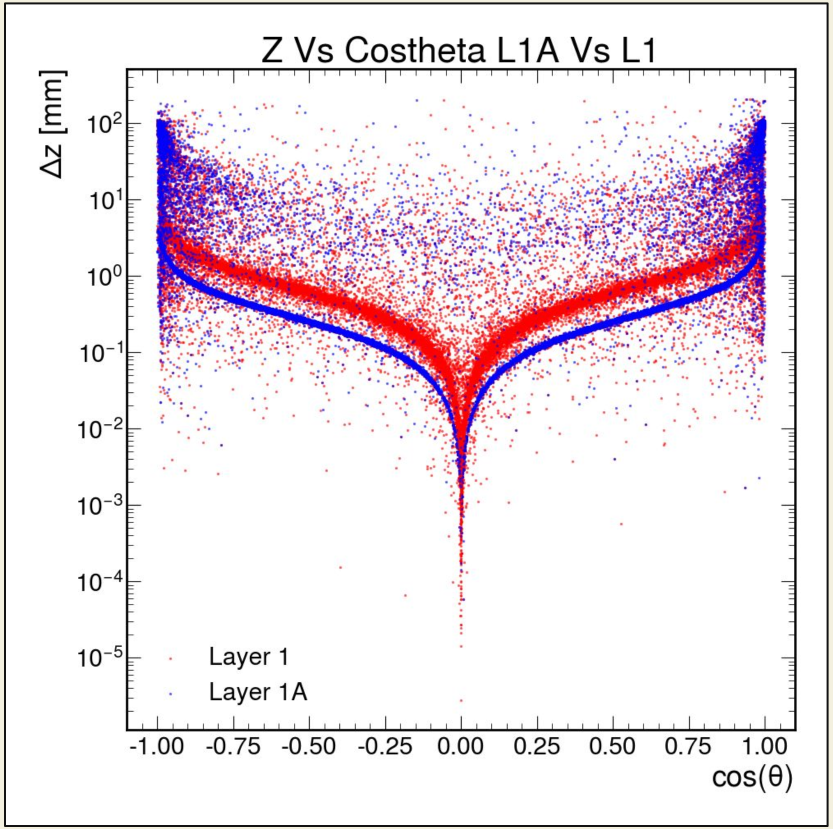
Closed Form Relationship



- This double band relationship comes directly from the geometry of the innermost layers.
- When exploring the geometry, the double band structure arises from: $\ln(\Delta z) = \ln(\Delta r) + \ln|\cot(\Theta)|$

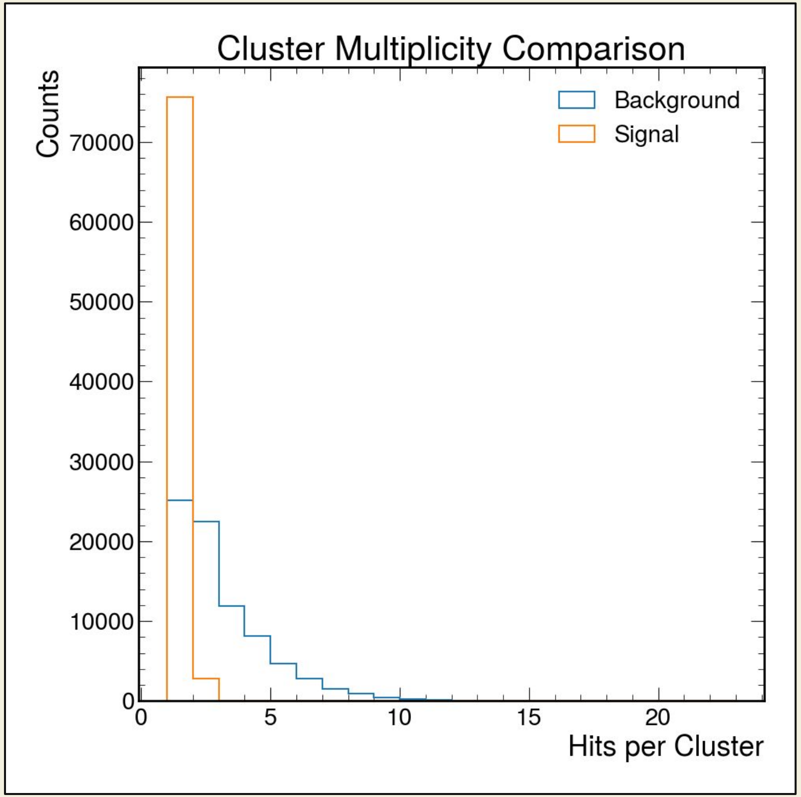
With two different Δr measurements for each band.

Cross Layer Δz vs. Cos(Θ)



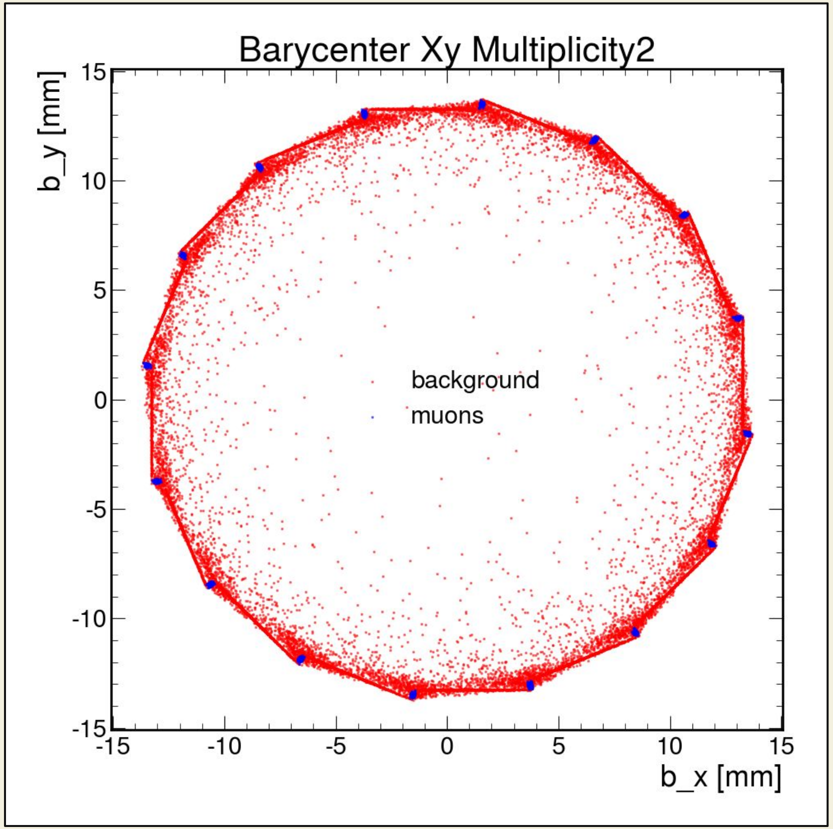
- The cause of this double band relationship is hits being dispersed over two layers, both categorized as layer 1.
- The upper band was created by a cluster spanning 1A and 1B.
- Lower band was populated by clusters with hits merely in the same layer.

Cluster Multiplicity Layer 1A



- This plot displays all the hits shown on the true innermost layer, 1A.
- Just a small number of signal clusters still have two hits.
- Z Extent requires at least two hits to calculate, so this Z Extent vs. Cos pattern represents only this small fraction of signal with multiplicity 2.

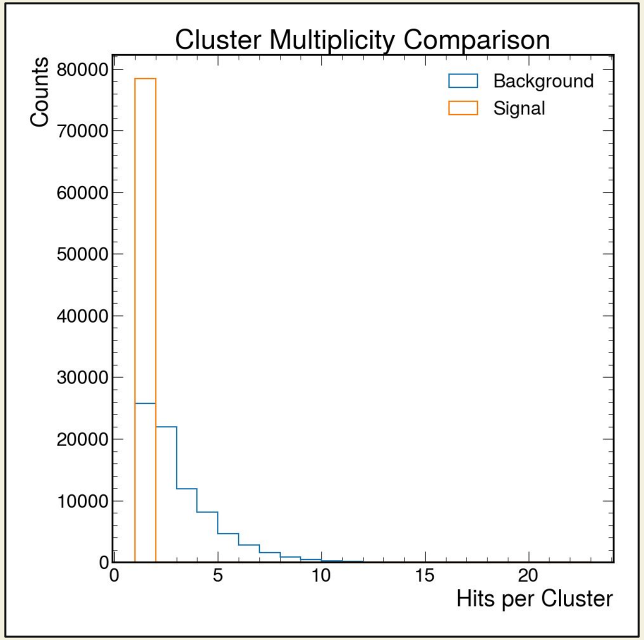
Barycenter, Multiplicity 2



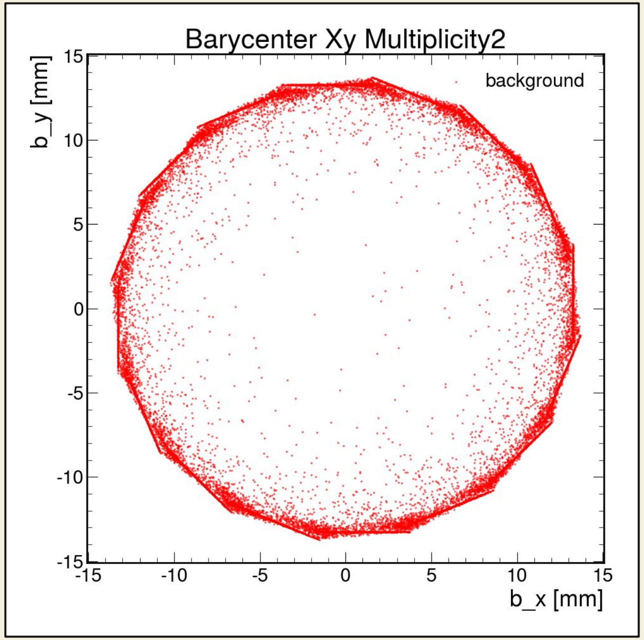
- After restricting, my analysis to 1A, I plotted the barycenters of multiplicity 2 clusters.
- I supposed that signal clusters with nonzero Z Extent only occur in module overlap regions.
- For background, however, they are quite distributed, indicating some curling and many hits from the same particle.

Merging Module Overlaps

Merged Cluster Multiplicity

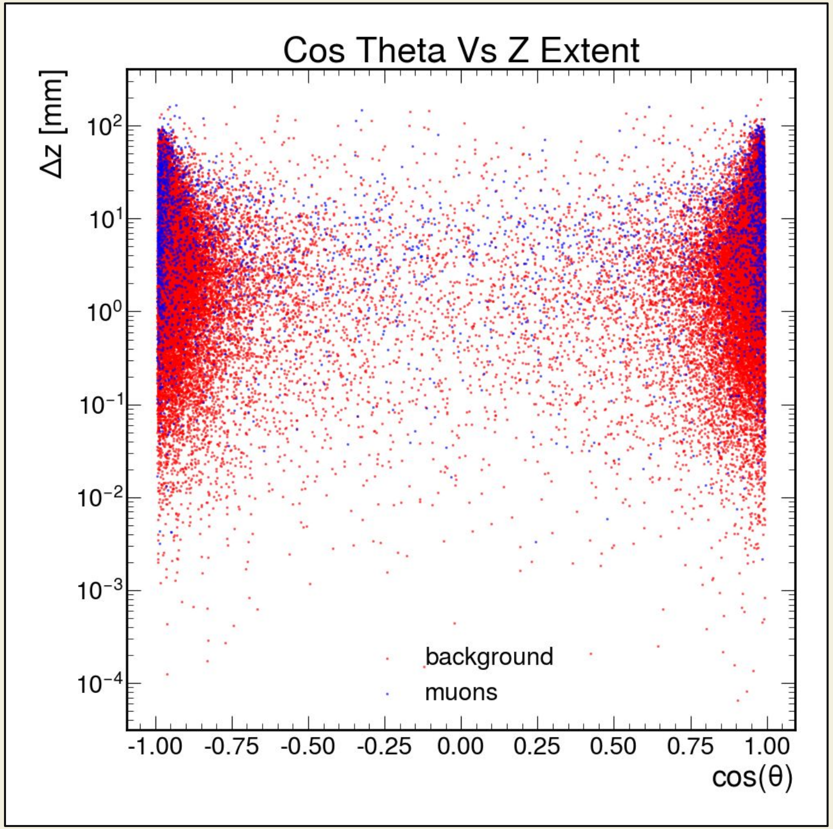


- To resolve module overlap artifacts, I developed an algorithm to merge any two hits on overlapping sections.
- It keeps the innermost hit and averages energy deposited.



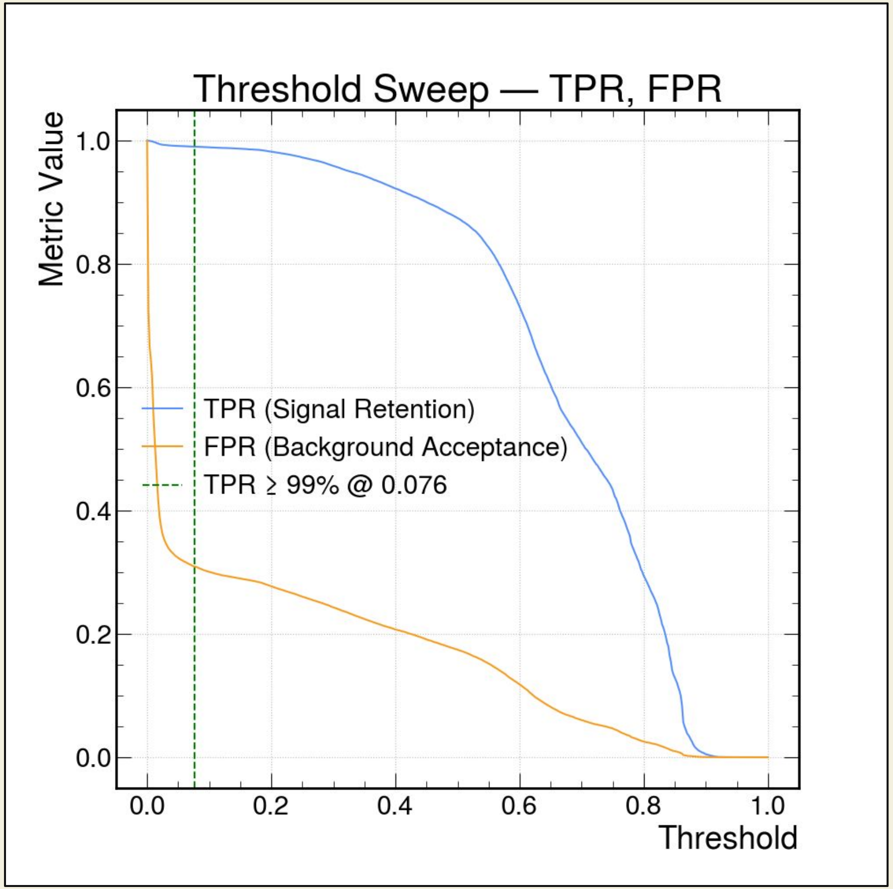
Conservative Module Level Separation

Conservative Separation Testing



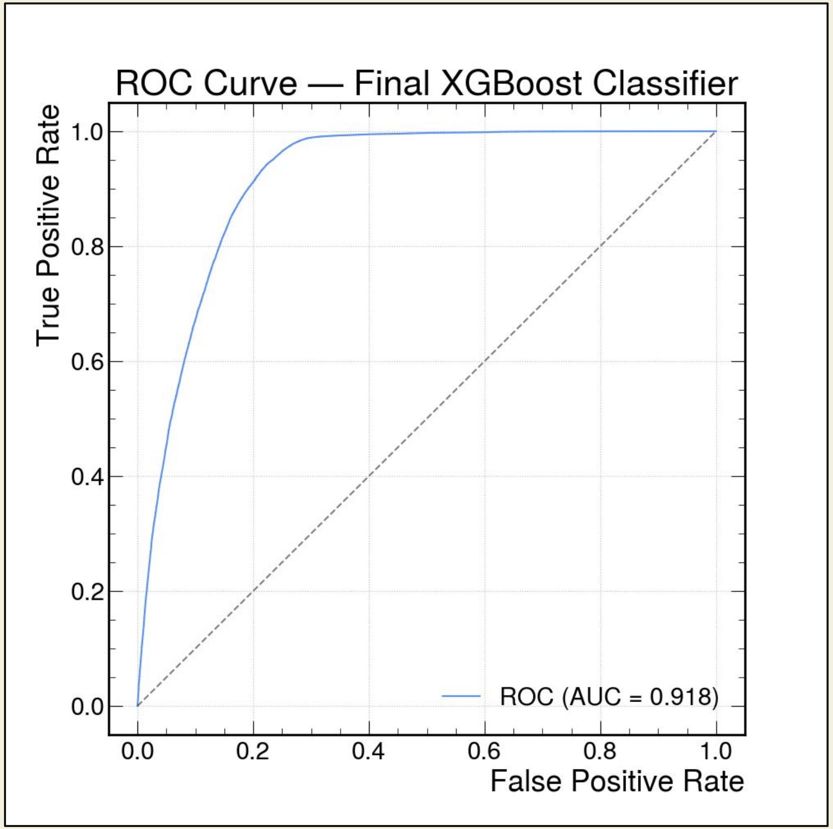
- Seeking to separate Z->qq and muons from beam background, while preserving at least 99% of signal clusters.
- More conservative approach, only considering module level readout on layer 1A, thus no Δz relationship.
- Using gradient boosted decision tree with input features including: $\cos(\theta)$, edep, Δz, etc.

Method



- Grid search across hyper parameters for XGBoost, settling on: max depth = 8, learning rate = 0.1, positive weight = 0.5.
- Swept TPR/FPR thresholds to achieve ≥ 99% signal retention and maximal background rejection.
- Final TPR/FPR threshold 0.0762.

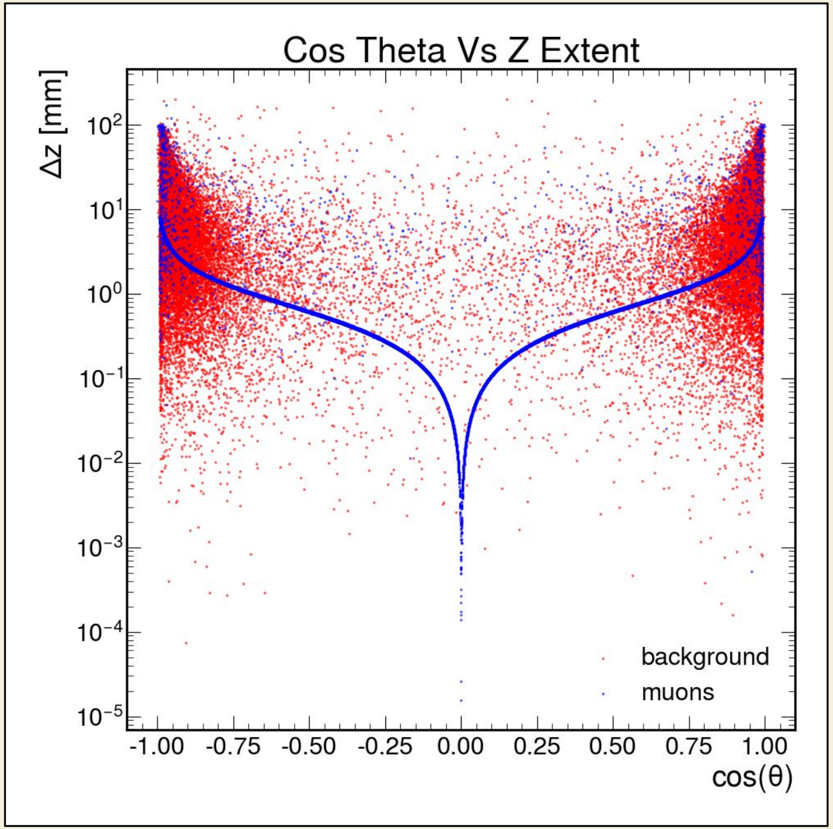
Final Conservative Classifier



- Final decision rule becomes:
  $\hat{y} = \mathbb{1}[\ P(\text{signal} \mid x) \geq 0.0762\ ]$
- Final classifier results in:
  a. Signal retention = 99.0%
  b. Background rejection: 69.1%
  c. ROC AUC: 0.9177
  d. Accuracy: 84.0%
  e. Precision (signal): 76.3%

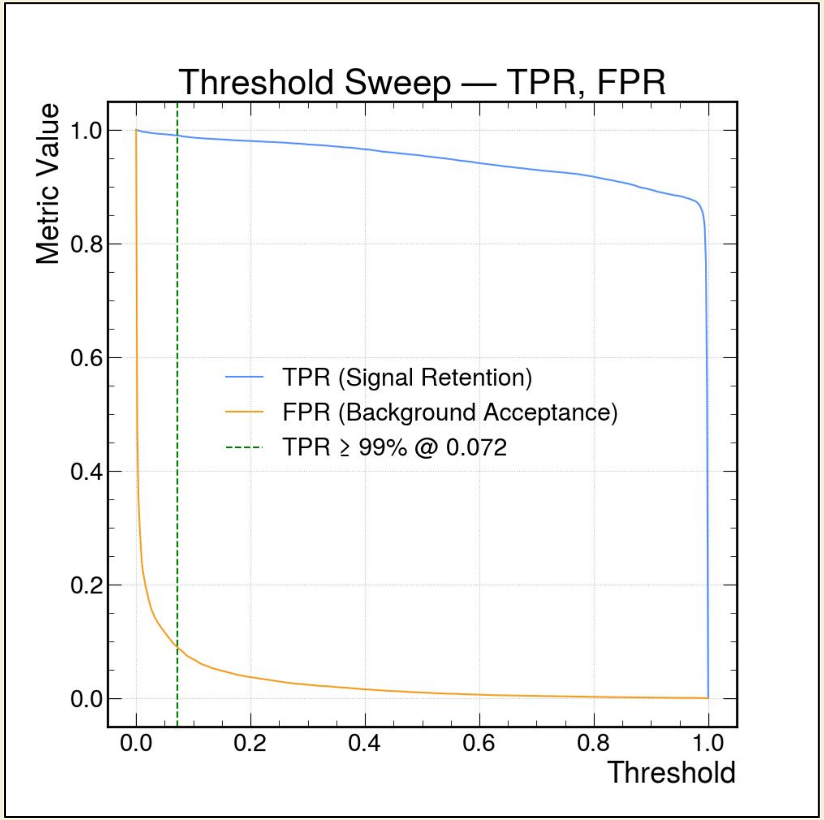
Optimistic Separation

Optimistic Separation Testing



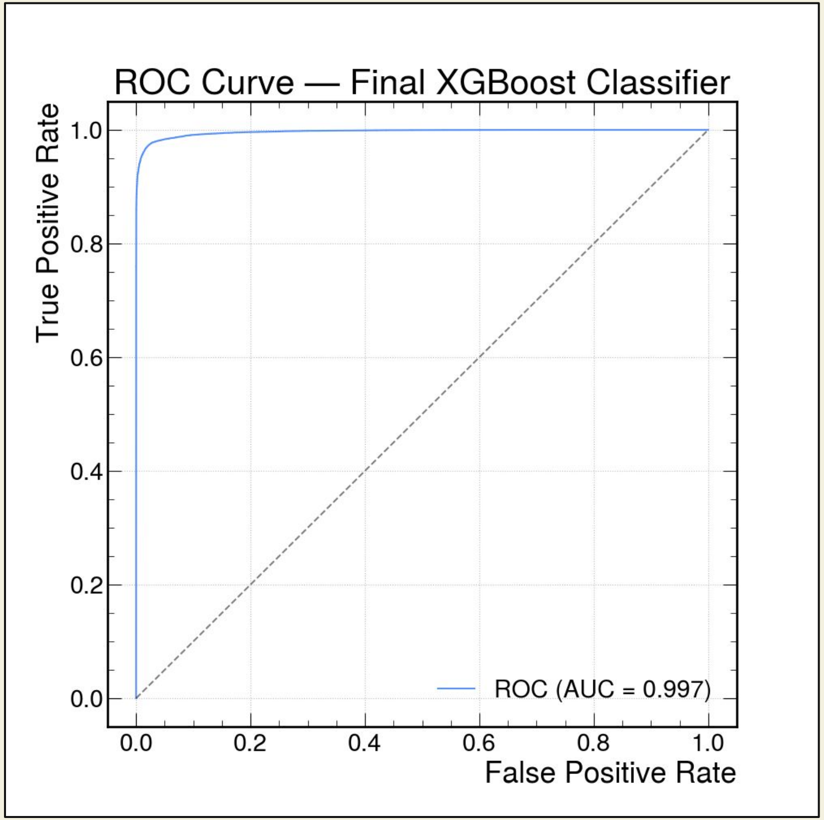
- Considers possibility of crosstalk between module pair on layers on 1A and 1B.
- With this crosstalk enabled: $\ln(\Delta z) = \ln(\Delta r * |\cot(\Theta)|)$ available to classifier (shown for muons & background).
- Using gradient boosted decision tree with input features including: $\cos(\theta)$, edep, $\log(\Delta z)$, etc.

Method



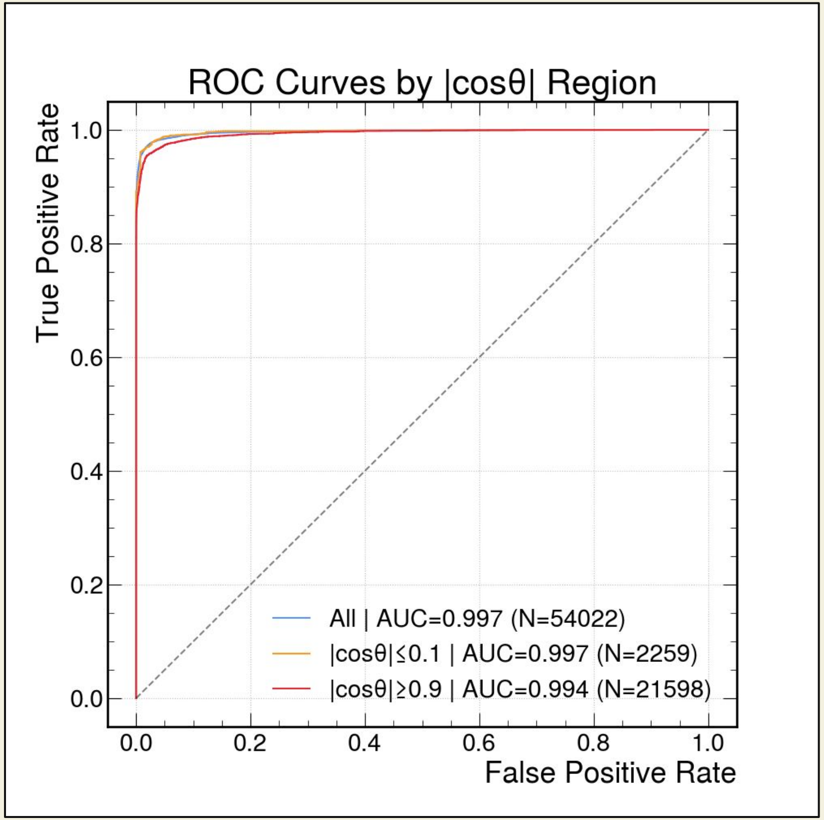
- Grid search across hyper parameters for XGBoost, settling on: max depth = 10, learning rate = 0.1, positive weight = 0.25.
- Swept TPR/FPR thresholds to achieve ≥ 99% signal retention and maximal background rejection.
- Final TPR/FPR threshold 0.072.

Final Optimistic Classifier



- Decision rule becomes:
  $\hat{y} = \mathbb{1}[\, P(\text{signal} \mid x) \geq 0.072]$
- Final classifier gives:
  a. Signal retention = 99.0%
  b. Background rejection: 91.0%
  c. ROC AUC: 0.9965
  d. Accuracy: 95.0%
  e. Precision (signal): 92.0%

|cosθ| Region Analysis



- Performance is fairly stable across |cosθ| regions.
- In |cosθ| ≤ 0.1 region, AUC = 0.9969 (N=2259).
- In |cosθ| ≥ 0.9 subset, AUC = 0.9940 (N=21598).
- Overall, performance slightly deteriorates at high |cosθ|, but classification retains effectiveness even in forward regions.

Summary and Key Contributions

- Explained the two-band structure in $\ln(\Delta z)$ vs $\cos\theta$ as a consequence of module overlap geometry.
- Designed an overlap-aware merging algorithm that restores geometric consistency and removes artificial patterns.
- Characterized muon and $Z \to q\bar{q}$ clusters at the single-module level, reflecting conservative readout assumptions.
- Developed XGBoost-based classifiers under both conservative and optimistic assumptions, achieving strong signal retention and background rejection.