Geometric Background Suppression in FCC-ee Vertex Detector

Emmett Forrestel[1]

[1]Brown University / CERN

June–August 2025

Abstract

This report summarizes my Summer 2025 research at CERN, centered on the analysis and classification of hit clusters in the first layer of the CLD (CLIC-Like Detector) vertex detector for the Future Circular Collider (FCC-ee). We studied the structural differences between beam background and physics ($Z \to q\bar{q}$, etc.) clusters in Geant4 simulations, leveraging them to guide the design of XGBoost-based machine learning pipelines for real-time background suppression. These methods culminated in classification algorithms enabling real-time suppression of beam background clusters at the chip level—contributing toward faster, more focused event reconstruction in the high-occupancy FCC-ee environment.

1. Introduction

The **Future Circular Collider** is a proposed high-luminosity electron–positron collider designed to operate at multiple energy stages, with the Z-pole run expected to deliver unprecedented event statistics. The vertex detector in such a collider must achieve exceptional spatial resolution and fast readout, both of which are aided by the suppression of beam-induced background. This challenge is especially pronounced in the innermost tracking layer (Layer 1) of the **CLD (CLIC-Like Detector)**, where occupancies are highest.

My research focused on the analysis and classification of **hit clusters**—groups of pixel hits deposited by a single Monte Carlo (MC) particle. These clusters were constructed by discretizing the vertex detector into 25 µm × 25 µm bins, matching the pixel pitch of the **Arcadia-MD3** chip used in the detector design, and assigning simulated hits to these cells. This approach reproduces the granularity of the actual CLD readout, ensuring that the analysis reflects the information and constraints present in real detector operation. The goal was to identify geometric features distinguishing **signal** ($Z \to q\bar{q}$) from **beam induced background**.

This study rests on the hypothesis that signal and background clusters—while overlapping in energy deposits—may differ in their geometric footprints due to different trajectories and magnetic interactions. For instance, lower-energy beam background particles may curl more tightly in the magnetic field, creating wide ϕ-distributed clusters, while signal particles originating from the interaction point form linear tracks along their trajectory–resulting in distinct cluster structures.

To quantify these shapes, the following descriptors were extracted from each cluster:

- **Energy Deposited (edep)**: The summed energy deposition of a cluster.
- **Multiplicity**: The count of hits in a given cluster.
- **Cos(θ)**: Cosine of the angle the energy barycenter of the cluster makes with the beam axis.
- **ϕ rows**: The number of distinct pixel rows (aligned in ϕ) the cluster spans.
- **Z extent**: The spread of the cluster

along the beam axis, defined as the difference between the maximum and minimum z positions of the hits.

- **PCA elongation**: The ratio of the eigenvalues from a Principal Component Analysis (PCA) of the cluster hit positions, describing its elongation along a dominant axis.

The analysis of these features revealed distinct signal and background trends, brought to light systematic artifacts from overlapping sensor regions, and informed training of lightweight machine learning classifiers such as XGBoost and SVM. The key objective was identifying highly discriminatory, low-complexity metrics useful in the real-time suppression of beam background clusters before full event reconstruction at the module or chip level.

Such early suppression would reduce readout rates and power consumption, both critical for detector lifespan, minimizing material budgets, and enabling passive cooling—optimal constraints for vertex detectors at the FCC-ee.

2. Cluster Metrics

A set of geometric and structural descriptors was extracted from each hit cluster to capture its shape, spread, and distribution within the vertex detector. These metrics, described below, form the basis for signal–background discrimination in later classification efforts.

2.1 Energy Deposited

As defined previously–the summed energy deposit of hits associated with a cluster. Signal deposits less average energy with tighter distribution, whereas background has a higher average energy deposit and has wider spread. Both signal and background overlap significantly in energy deposit.

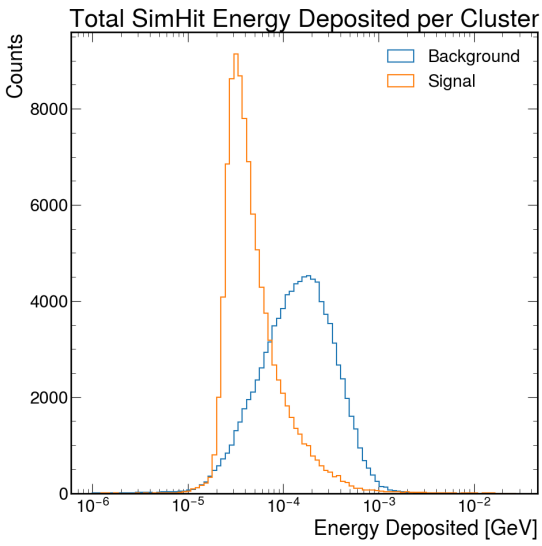

Figure 1: Energy deposited per cluster for $Z \to q\bar{q}$ and beam background.

2.2 Cluster Multiplicity

Defined as the number of hits per cluster. Strong discriminator: signal peaks sharply at 2; background is broader.

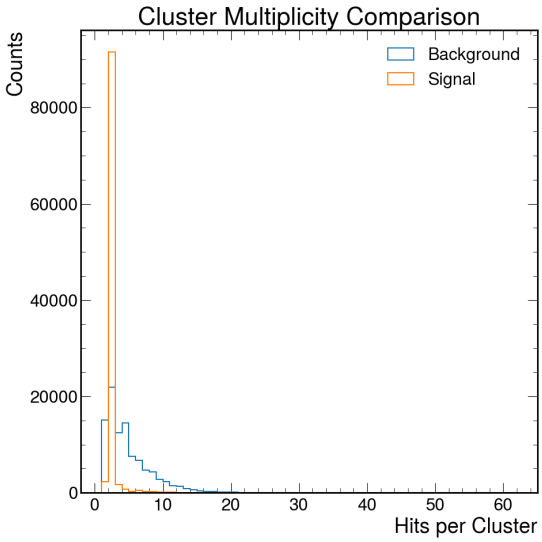

Figure 2: Cluster multiplicity distribution for signal and background.

2.3 Cosine of Polar Angle ($\cos\theta$)

Cosine of the angle between beam axis (z axis) and cluster energy barycenter.

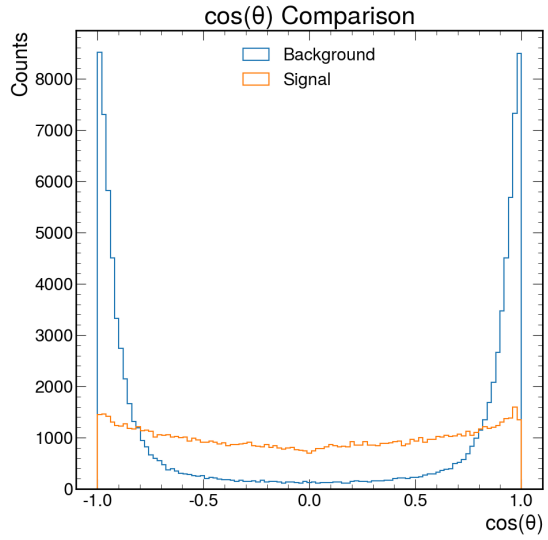

Figure 3: Distribution of $\cos\theta$ among clusters. Background is very forward, with signal more evenly distributed across theta.

2.4 ϕ Rows Hit

Count of 25 μm sensor bins activated, reflecting number of readout rows at detector granularity.

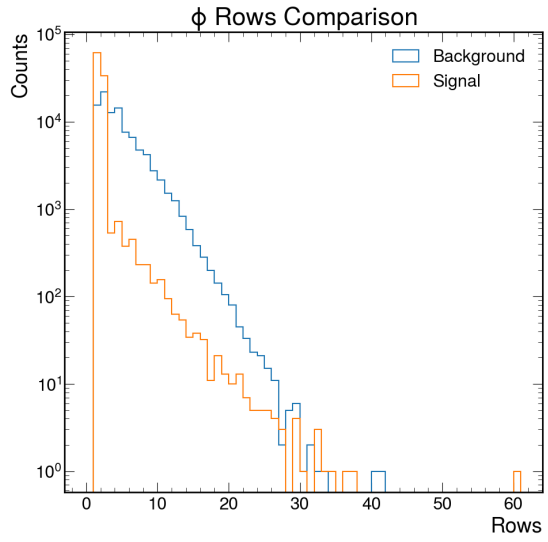

Figure 4: Displays count of ϕ rows hit. Signal sharply peaks at one or two rows activated, indicating little or no curling. Background tends towards higher spread, consistent with low-energy magnetic field induced curling.

2.5 Z Extent

Measures longitudinal cluster span along the beam axis, very sensitive to incident angle. It is observed that signal clusters tend to be more concentrated in both ΔZ and ϕ.

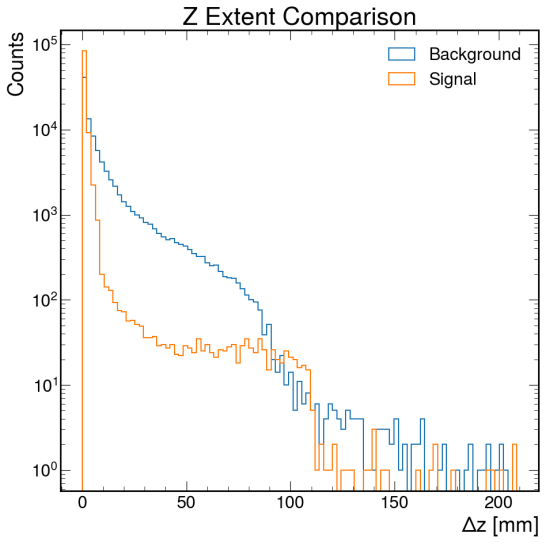

Figure 5: Displays count of ΔZ. Signal dramatically peaks at minimal ΔZ, consistent with very linear, low multiplicity tracks. Following the theme, background clusters spans more in ΔZ.

2.6 PCA Elongation

Computed as λ_1/λ_2, the ratio of principal to secondary eigenvalues from the ϕ–z covariance matrix. High ratios indicate linear clusters, whereas values near one stem from isotropic shapes. These metric can be understood as a compact, rotation-invariant shape summary.

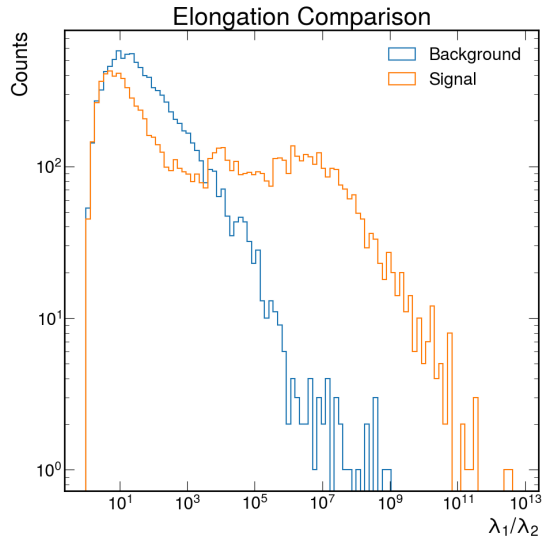

Figure 6: PCA elongation distributions. Background cluster tend to be more isotropic consistent with low energy magnetic curling. Signal clusters dominate at very high elongation, indicating nearly linear clusters.

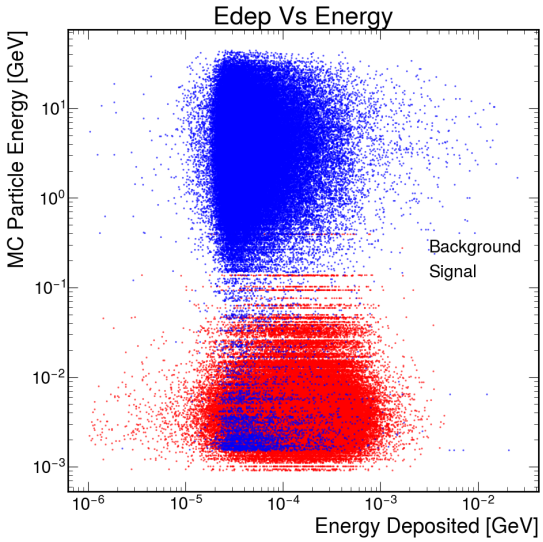

Figure 7: Total deposited energy vs. MC particle energy.

3. Feature Correlation Analysis

In this section, we examine key pairwise correlations between cluster features. These relationships reveal the structural trends and kinematic origins within signal and background data, guiding the development of classification models.

3.1 Deposited Energy vs. MC Particle Energy

MC Energy refers to the true relativistic energy of the Monte Carlo particle responsible for producing a given cluster. Signal population is bifurcated into a higher energy main population and lower energy residual. Beam background is composed entirely of a low energy population. Minimal correlation is displayed between the two metrics.

3.2 Deposited Energy vs. z Extent

Our hypothesis was that clusters with higher Δz, correspond to a lower incidence angle and higher distance traveled within silicon sensors, and consequently would deposit more energy. This hypothesis is consistent with the relationship demonstrated.

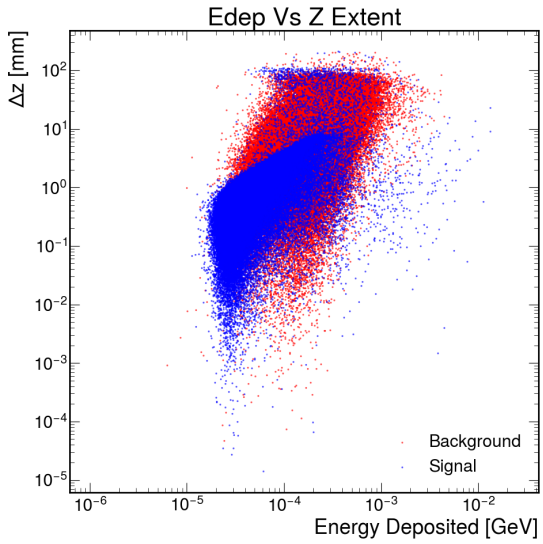

Figure 8: Deposited energy vs. Δz of clusters. Both signal and background clusters exhibit a positive relationship between z extent and energy deposited, with signal following a tighter distribution and background more dispersed.

3.3 PCA Elongation vs. z Extent

Signal population is again bifurcated into low elongation, high Δz clusters, and a positively correlated band of high elongation and moderate z extent clusters, this may indicate two distinct populations produced within $Z \to q\bar{q}$ simulation events. Background clusters are consistently low elongation, and distributed across z extent.

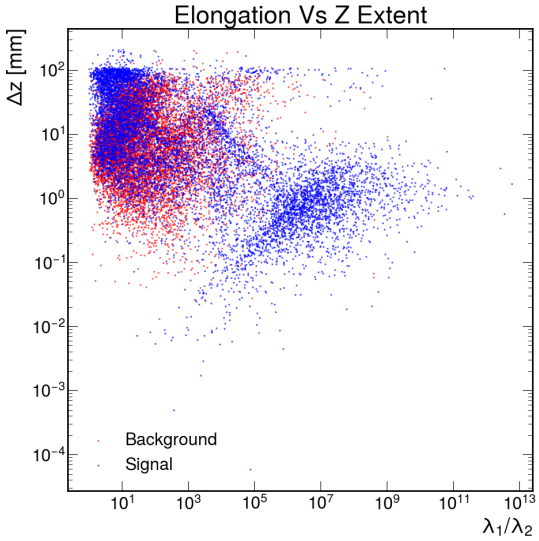

Figure 9: PCA elongation vs. cluster z extent. Signal splits into two populations, one completely distinct in a high-elongation region.

3.4 $\cos\theta$ vs. Energy Deposited

Similarly to the energy deposit Δz relationship, higher $|\cos\theta|$ corresponds to lower incidence angles, leading to more detector material passed through and higher energy deposits. Signal exhibits a stronger relationship than background, which is more diffuse and stochastic.

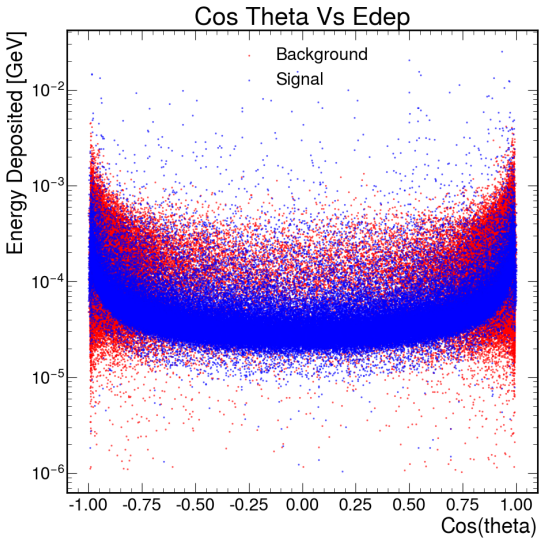

Figure 10: Energy deposited vs. $\cos\theta$ for signal and background clusters. Signal demonstrates a clear positive relationship between angle angle and energy deposited.

3.5 $\cos\theta$ vs. Multiplicity

Similarly to previous relationships, more forward particles have lower incidence angles and pass through more material, resulting in a higher number of pixel activations, particularly in background clusters.

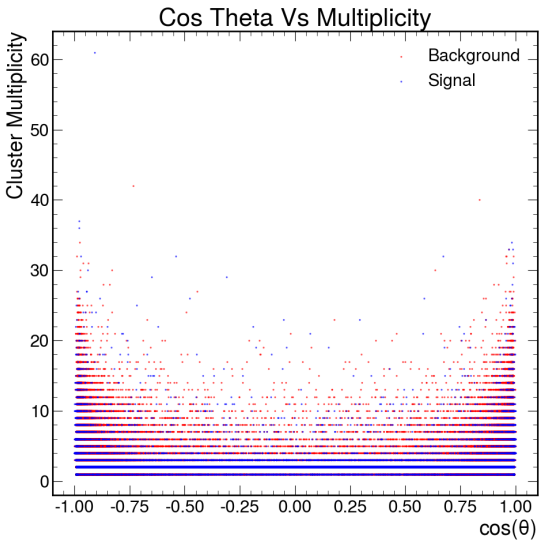

Figure 11: Multiplicity versus $\cos\theta$. Signal clusters exhibit multiplicity of two most frequently, agreeing with earlier plots, while both increase at high $\cos\theta$.

3.6 $\cos\theta$ vs. PCA Elongation

At high $|\cos\theta|$, signal clusters are more elongated, reflecting near-linear tracks entering at steep angles. Background clusters are more isotropic and less dependent on θ. Again, signal clusters form two distinct populations, one with a strong, positive relationship between $|\cos\theta|$ and elongation, the other occupying high $|\cos\theta|$ low elongation regions, similar to beam background.

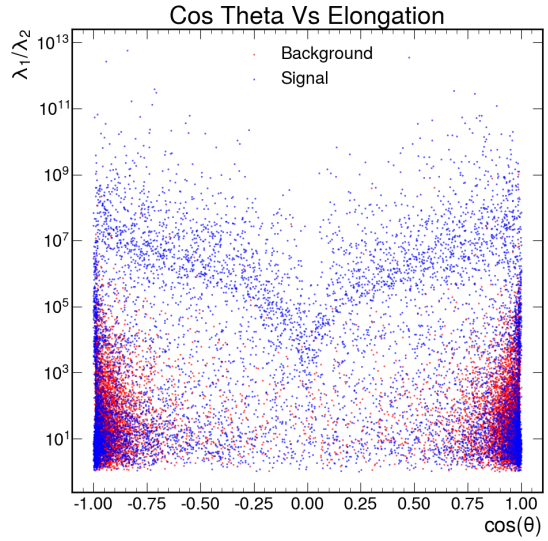

Figure 12: PCA elongation (λ_1/λ_2) versus $\cos\theta$. One of the two relationships signal clusters follow forms a very distinct high-elongation population completely uninhabited by beam background.

3.7 $\cos\theta$ vs. z Extent

For particles originating at the interaction point (IP) (almost all signal clusters do), as $\cos\theta$ increases, clusters tend to graze the sensor more and leave longer tracks in z. Background appears scattered, as it does not necessarily stem from the IP. However, signal forms a very strong double banded structure which warranted further investigation.

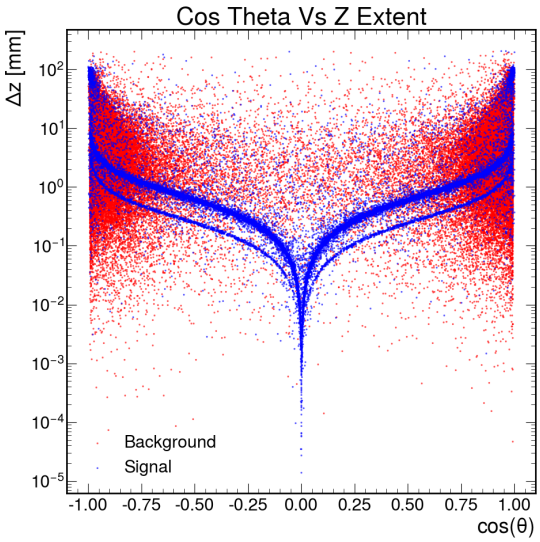

Figure 13: Longitudinal cluster extent (Δz) versus $\cos\theta$. Far and away the strongest relationship exhibited between any two metrics is the highly compelling double banded signal structure.

4. Overlap Region Analysis

The appearance of two distinct bands in the $\ln(\Delta z)$ vs. $\cos\theta$ distribution prompted a deeper investigation into the geometric origins of hit clusters in the innermost layers of the vertex detector. For a particle passing through two sensor layers separated by a radial distance Δr, the geometric relation (depicted in the following diagram):

$$\Delta z = \Delta r \cdot |\cot\theta|$$

naturally leading to the log form:

$$\ln(\Delta z) = \ln(\Delta r |\cot\theta|).$$

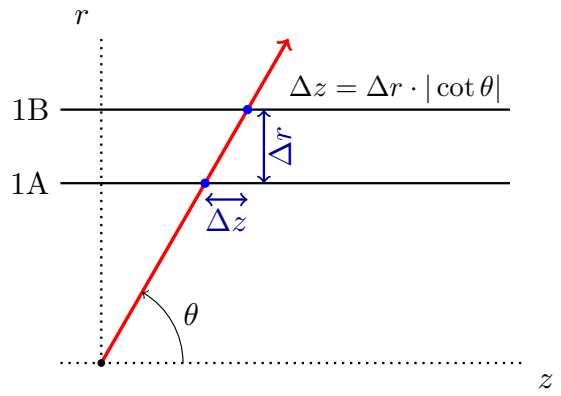

Figure 14: Geometric relation between radial and longitudinal hit separations for a particle crossing layers 1A and 1B, assuming a linear track stemming from the interaction point.

Two distinct bands appear, each consistent with this formula but implying two different values of Δr, indicating two sources of radial separation among cluster hits. This becomes apparent when visualizing the detector in the x–y plane. As shown in Fig. 15, Layer 1 consists of two concentric sublayers: Layer 1A and Layer 1B.

Hits from both 1A and 1B contribute to clusters with large radial separation Δr, explaining the upper band in $\ln(\Delta z)$ vs. $\cos\theta$. But the lower band stems from clusters whose hits are confined to a much smaller Δr—within Layer 1A itself.

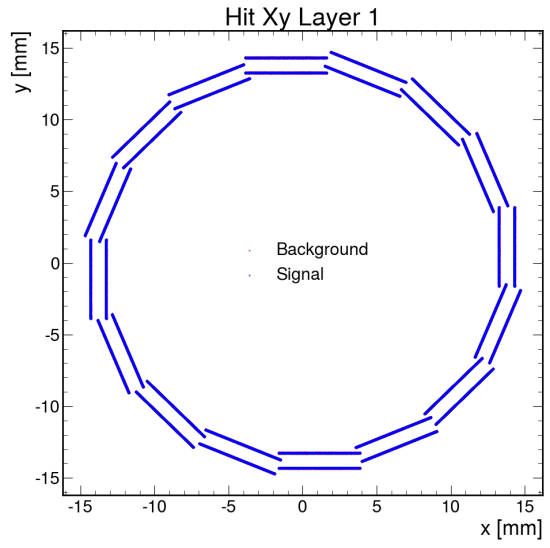

Figure 15: x–y distribution of all hits in Layer 1, showing two distinct rings: 1A (inner) and 1B (outer).

To investigate the lower band further, we focus on muon particle gun clusters in Layer 1A with exactly two hits, rather than $Z \to q\bar{q}$ clusters. These events are clean and unambiguous. When plotting their spatial distribution, a clear pattern emerges: two-mulitiplicity muon clusters in 1A only occur in narrow zones corresponding to overlaps regions between adjacent sensor modules.

This suggested that the lower $\ln(\Delta z)$ vs. $\cos\theta$ band originated from these overlap regions, rather than clusters spanning layers 1A and 1B, and consequently would not be useful on a detector level readout.

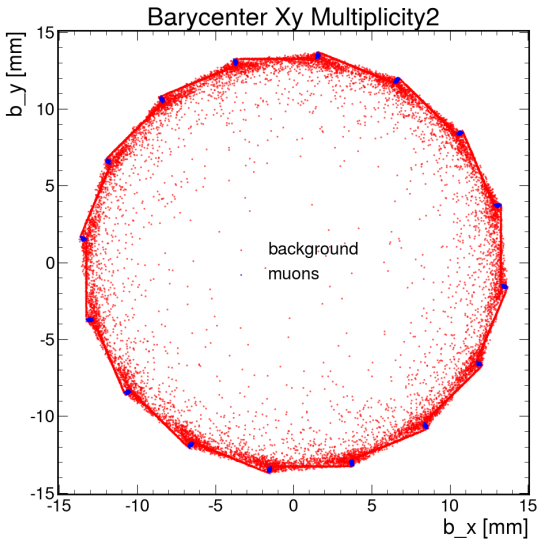

Figure 16: Energy barycenters of muons and beam background with multiplicity two; used to define merging regions. Note, all barycenters lie in convex hull of detector, and thus are valid.

4.1 Merging Algorithm for Overlap Regions

To rectify this overlap, a merging algorithm was developed. Overlap zones were defined geometrically based on known module tiling in ϕ, and clusters with multiple hits falling within the same overlap region were merged. The innermost hit was preserved, and its energy deposit was defined as the average energy deposit of its hits in the overlap zone.

After applying the merging algorithm, the lower log-linear band in $\ln(\Delta z)$ vs. $\cos\theta$ vanished entirely. Only the upper band re-

mained—corresponding to clusters formed by hits in both Layer 1A and Layer 1B, with large Δr. This confirmed that the lower band was an artifact of sensor overlap in 1A and that our geometric merging strategy correctly removed it.

If readout is, in fact, possible across layers 1A and 1B, this metric would serve an elegant and highly discriminatory classifier input–stemming from intrinsic differences in origins of beam background and physics signal.

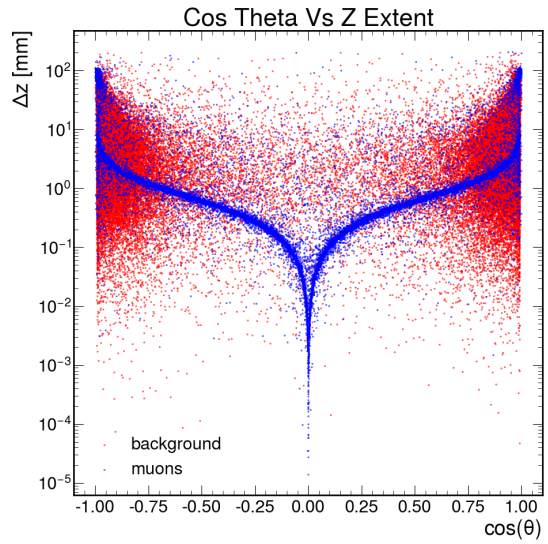

Figure 17: Post-merging $\cos\theta$ vs. Δz distribution for muon clusters in Layer 1A. The lower band is suppressed.

5. Layer 1A Cluster Characterization

To characterize cluster geometry under conservative hardware constraints, we restrict our analysis to readout Layer 1A on a per-module basis. This reflects the most conservative and localized idea of module-level readout: clusters are formed only from hits within a single module on Layer 1A, with no merging across module boundaries and no dependence on other sublayers.

This choice avoids complications due to sensor overlaps and radial separation between Layers 1A and 1B, which were shown in previous sections to introduce structured Δz bands and artificial multiplicities. By focusing on localized, intra-module clusters, we aim to characterize the intrinsic structure of signal events as they would appear to an isolated module in the vertex detector.

Two distinct datasets are used:

- A muon particle gun sample, providing an idealized single-particle baseline.
- $Z \rightarrow q\bar{q}$ events, reflecting realistic hadronic signal cluster behavior.

5.1 Muon Particle Gun

Muons are minimal-ionizing particles with clean trajectories and no hadronic fragmentation. They provide a reliable reference for studying cluster geometry and detector response in isolation.

5.1.1 Multiplicity Distribution

Most muon clusters contain almost exclusively 1 hit. This reflects the narrow traversal of a single particle through the sensor volume.

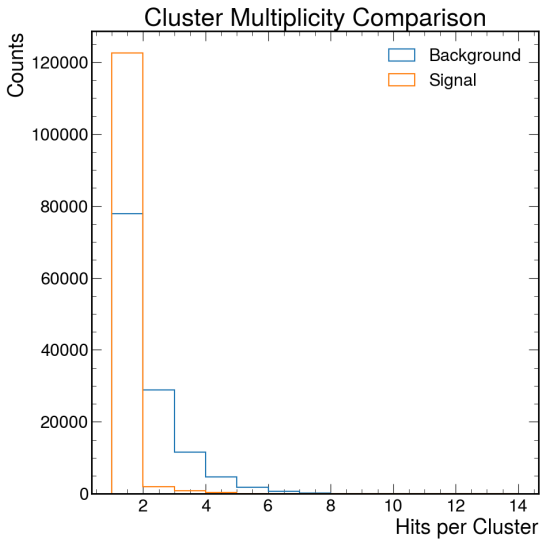

Figure 18: Multiplicity distribution of muon clusters in Layer 1A. Muons sharply peak at multiplicity one, with signal more dispersed.

5.1.2 Deposited Energy vs. MC Energy

Muon samples exhibit two clear population, one of 50 GeV muons, and the others of low energy $< 10^{-1}$ GeV residuals. Muon events are shown in isolation to make these two populations more apparent.

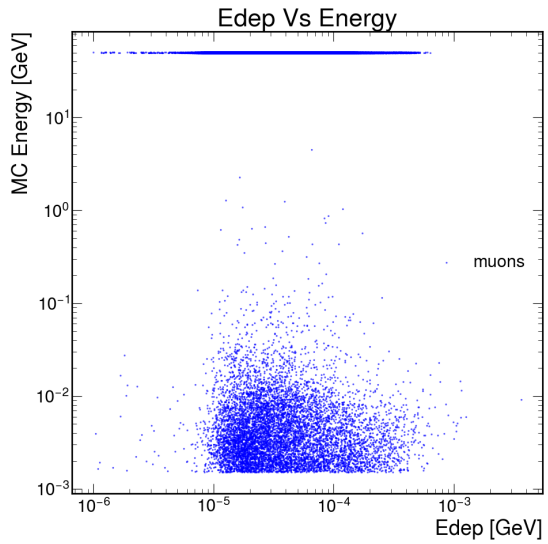

Figure 19: Deposited energy vs. MC energy for muon clusters in Layer 1A. Two distinct populations are shown.

5.1.3 $\cos\theta$ vs. Deposited Energy

More forward muons ($|\cos\theta| \to 1$) deposit more energy, as they traverse more silicon at a shallower angle–consistent with earlier findings.

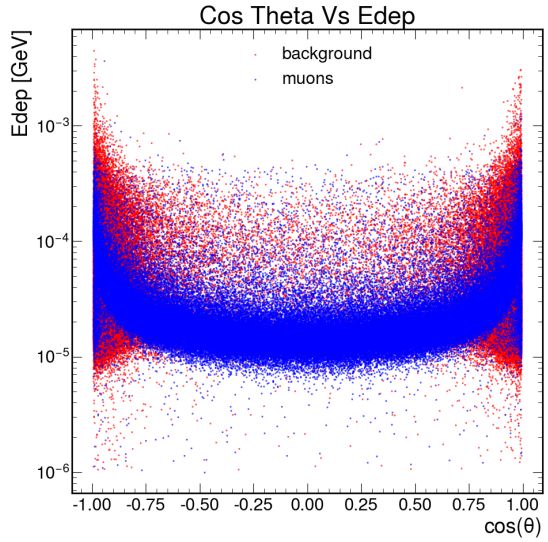

Figure 20: $\cos\theta$ vs. deposited energy for muon and beam background clusters.

5.1.4 z Extent

Muon clusters dramatically peak at near zero z extent, with background being much more diffuse.

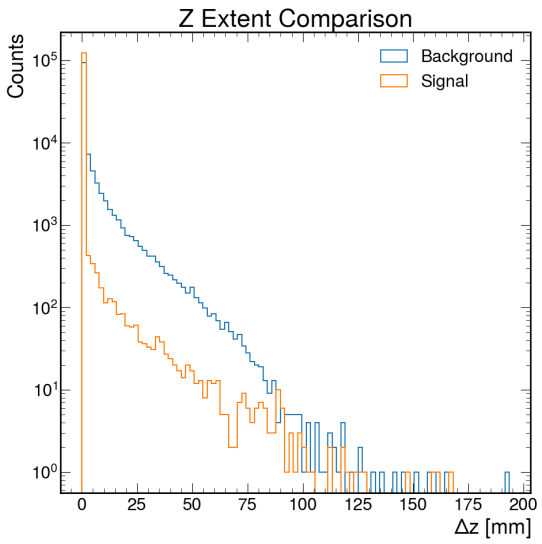

Figure 21: Distribution of z extent for muon and background clusters in Layer 1A.

5.1.5 $\cos\theta$ vs. z Extent

Unlike the previous analyses that spanned both Layers 1A and 1B, or between modules in 1A, no $\ln(\Delta z)$ vs. $\cos\theta$ relationship appears here. This is expected: in per-module Layer 1A readout, all hits occur at the same radius, and no radial separation Δr exists to support the original geometric relationship. Further, very few muons even have two hits and thus have a z extent of zero.

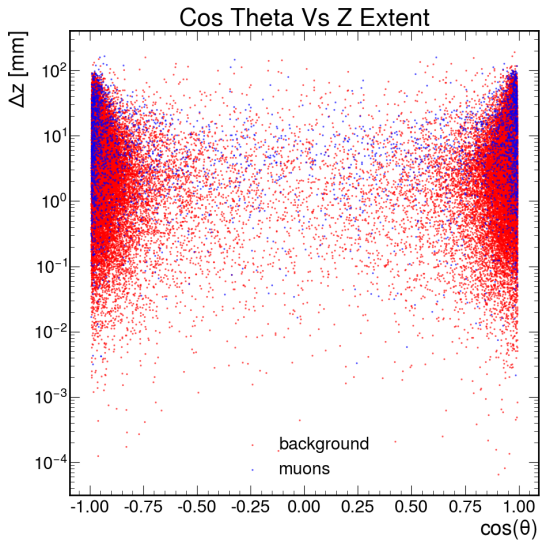

Figure 22: $\cos\theta$ vs. z extent for clusters in Layer 1A.

5.2 $Z \to q\bar{q}$ Signal

Clusters from $Z \to q\bar{q}$ events and muons exhibit very similar geometries (Δz, multiplicity, energy deposited, etc.), indicating that classification between physics and background would hold across multiple physics channels.

5.2.1 MC Energy vs. Deposited Energy

Unlike muons, signal particles vary significantly in energy. Again two distinct population are shown in $Z \to q\bar{q}$ events, now understood as true hadronic events in one population and low-energy residual in the other.

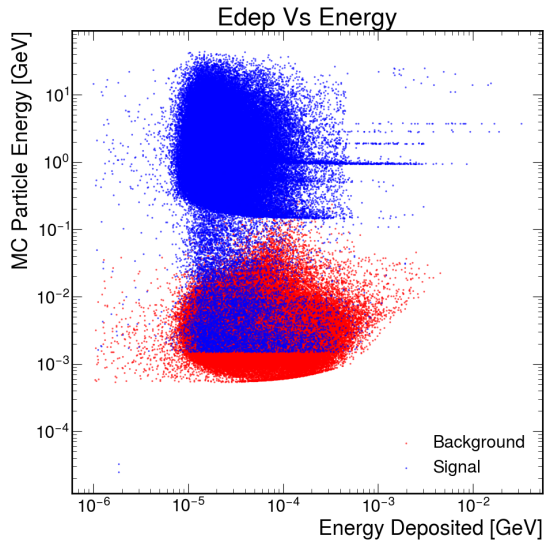

Figure 23: Deposited energy vs. MC energy for signal clusters in $Z \to q\bar{q}$ and beam background events in layer 1A. Residual and hadronic population evident in signal.

5.2.2 Multiplicity Distribution

Signal clusters have nearly the same multiplicity distributions as muons, again reflecting single particle traversal through individual sensor volumes.

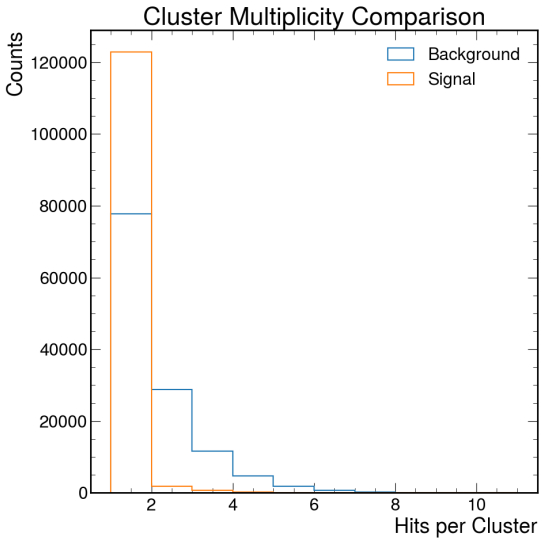

Figure 24: Multiplicity distribution of signal and beam background in Layer 1A.

5.2.3 $\cos\theta$ vs. Deposited Energy

Signal clusters also show increased energy deposition at high $|\cos\theta|$, though with more variability. Background follows a much weaker relationship.

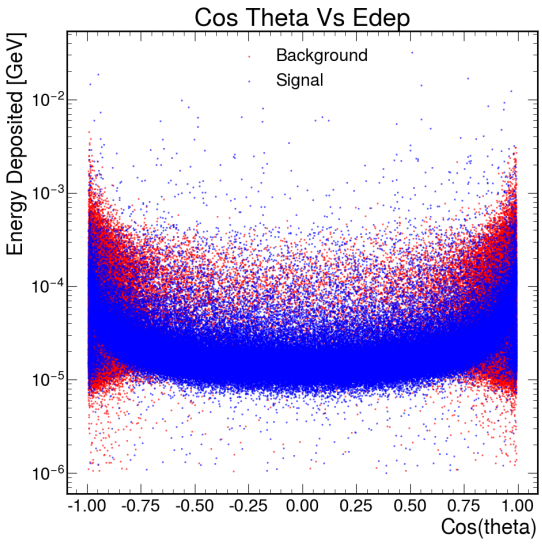

Figure 25: Distribution of $\cos\theta$ vs. deposited energy for layer 1A clusters.

5.2.4 z Extent

Signal clusters exhibit a very similar distribution in z extent compared to muons, again with background dominating at non-zero Δz extent.

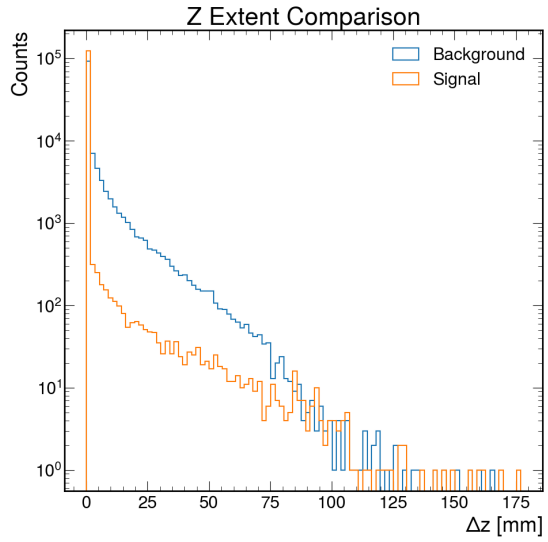

Figure 26: Distribution of z extent for $Z \to q\bar{q}$ and beam background events in Layer 1A.

5.2.5 $\cos\theta$ vs. z Extent

Again, no $\ln(\Delta z)$ structure is observed, since all clusters are confined to a single module without radial traversal. Background is more numerous, as very few signal particles create even two hits.

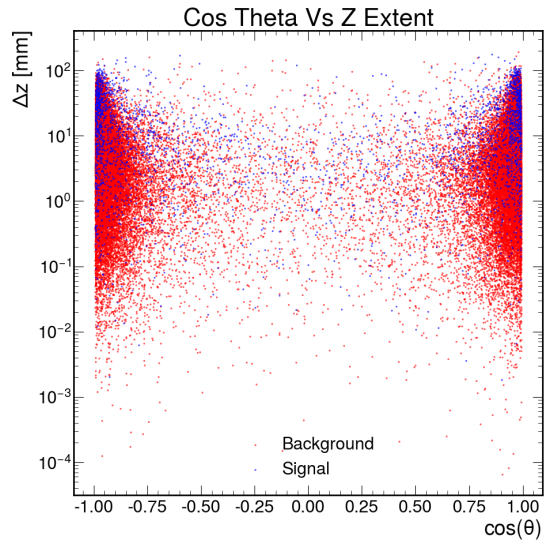

Figure 27: Distribution of $\cos\theta$ vs. z extent for $Z \to q\bar{q}$ and beam background events in layer 1A.

6. Classification Pipelines

6.1 Preprocessing and Residual Reassignment

Before training classifiers, we implemented a preprocessing step to reassign previously discussed low-energy residual clusters—present in both the muon and $Z \to q\bar{q}$ signal files—to the background dataset. These clusters are not removed from analysis but are excluded from the signal class to avoid corrupting the learning target with background-like features and preserving irrelevant physics.

This residual population is composed predominantly of soft photons and electrons, with PDG IDs:

- 22 (photon)
- 11 (electron)
- −11 (positron)

and energies below 0.01 GeV. These particles are not part of the primary physics process under study and instead arise from secondary interactions or detector effects. Notably, they are the sole constituents of the beam background dataset.

To maintain the integrity of the classification task, we apply the following rule:

> Clusters associated with MC particles satisfying `PDG ID` $\in \{-11, 11, 22\}$ and `MC Energy` $<$ 0.01 GeV are reassigned to the background class.

This reassignment ensures that clusters used as training signal truly reflect the physics of interest—rather than incidental noise.

Figure 28 shows the MC energy spectrum for particles in the beam background dataset, which is dominated by low-energy $e^{\pm}$ and γ.

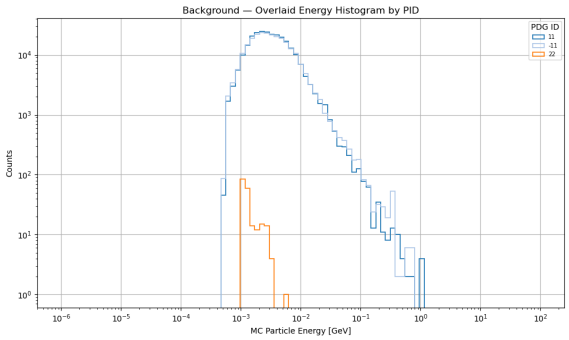

Figure 28: Energy distribution of MC particles in beam background files, stratified by PDG ID. The background is composed entirely of sub-1 GeV e^+, e^-, and γ.

The same PID–energy profile appears in the muon gun dataset, seen in Figure 29. These low-energy secondaries are not representative of primary physics and are reassigned accordingly.

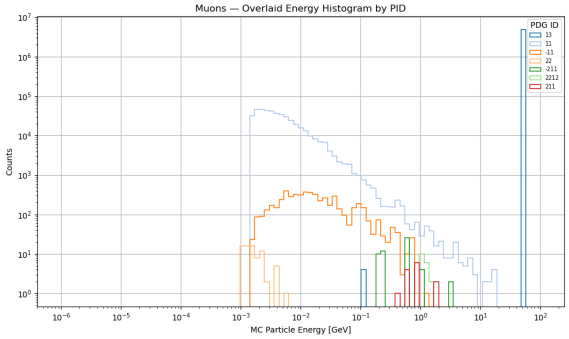

Figure 29: MC particle energy in muon files, stratified by PDG ID. A low-energy residual component very similar to that of beam background is visible.

Similarly, the $Z \to q\bar{q}$ events contain a population of soft $e^{\pm}$ and γ, shown in Figure 30. These are also reassigned, preventing contamination of signal from beam-background-like effects.

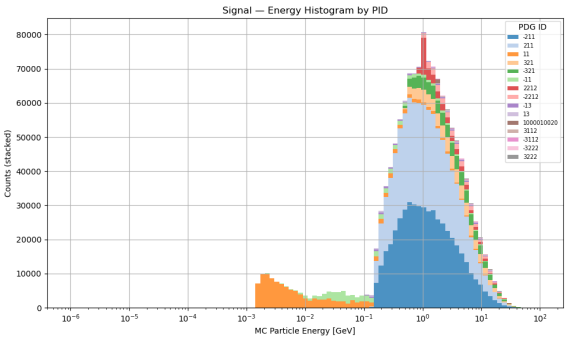

Figure 30: Stacked histogram of MC particle energy in $Z \to q\bar{q}$ files. The same soft $e^{\pm}$ and γ tail is present and reassigned to background.

Reassignment Summary:

- Reassigns clusters from low-energy $e^{\pm}$ and γ to background across all datasets.
- Improves training set quality by isolating true signal (muons, hadrons).
- Prevents classifier leakage from simulation artifacts or detector secondaries.
- Preserves all clusters while ensuring consistent labeling across sources.

6.2 Conservative Model: Layer 1A Module-Level Readout

This model assumes strict modular independence: each Layer 1A module is read out in isolation, with no crosstalk between neighboring sensors. As a result, only features computable within an individual module are used in classification. These weak assumptions about hardware intercommunication represent an implementable baseline for real-time beam background suppression. It provides a lower bound on expected performance against which more advanced readout schemes can be compared

Input Features:

- Multiplicity – a count of the number of hits left by an MC particle.
- Energy Deposited – total energy deposited in the cluster.
- $\cos\theta$ – angular direction of the cluster barycenter with respect to the beam .
- ϕ rows – a measure of the extent of the cluster in the azimuthal pixel direction.
- z Extent – a measure of the extent of the cluster along the beam axis.

- PCA Elongation – geometric anisotropy of the cluster in the ϕ–z plane.

Final Training Parameters:

- Maximum tree depth: 8
- Learning rate: 0.1
- Positive class weight: 0.5

Final Decision Rule:

$$\hat{y} = \mathbb{1}\left[P(\text{signal} \mid x) \geq 0.0762\right]$$

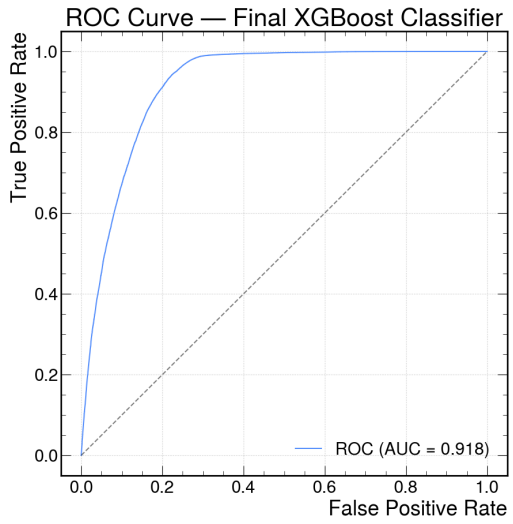

Figure 31: ROC curve for the conservative model. The displayed moderate AUC reflects the use of intra-module features only.

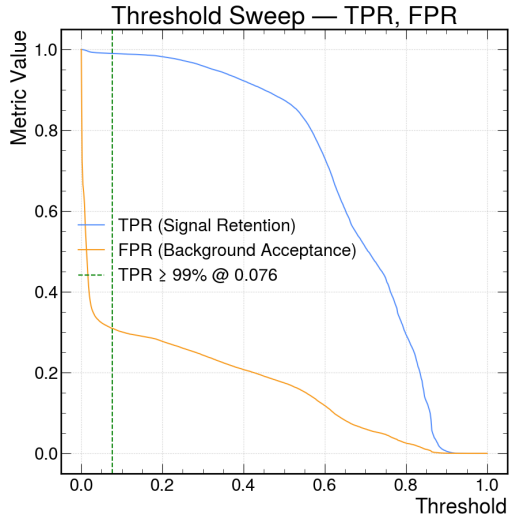

Figure 32: Threshold sweep for the conservative classifier showing signal retention vs. background rejection. The 99% operating point is marked.

Performance:

- Signal retention: 99.0%
- Background rejection: 69.1%
- ROC AUC: 0.9177
- Accuracy: 84.0%
- Precision: 76.3%

6.3 Optimistic Model: Layer 1A/1B Crosstalk-Enhanced Readout

This model assumes a more sophisticated readout system allowing crosstalk between layers 1A and 1B. This assumption enhances the discriminatory use of measured geometric features, notably enabling the strong Δz, $\cos\theta$ relationship.

Input Features:

- Multiplicity – count of hits left by an MC particle.
- Energy Deposited – total cluster energy deposit.
- $\cos\theta$ – angle between cluster barycenter and beam.
- ϕ rows – metric for cluster extent in the ϕ direction.
- z Extent – measure of the cluster along z beam axis.
- PCA Elongation – descriptor of cluster shape in ϕ–z plane.

Training Parameters:

- Maximum tree depth: 10
- Learning rate: 0.1
- Positive class weight: 0.25

Final Decision Rule:

$$\hat{y} = \mathbb{1}\left[P(\text{signal} \mid x) \geq 0.072\right]$$

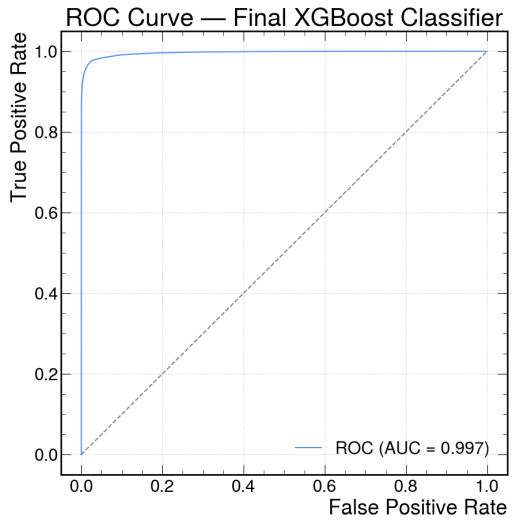

Figure 33: ROC curve for the optimistic model. 1A/1B crosstalk significantly improves separability, by exploiting the geometry of origination differences between signal and beam background.

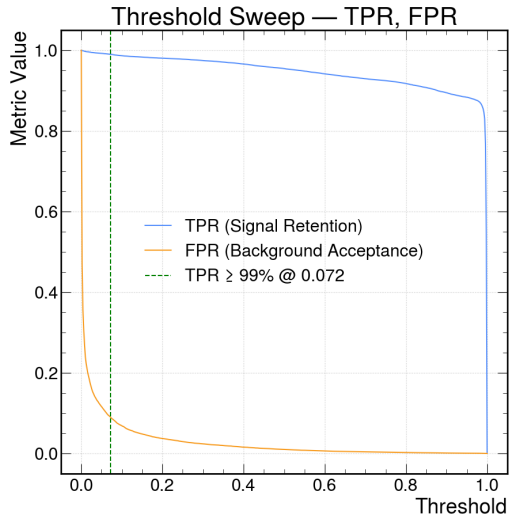

Figure 34: Threshold sweep for the optimistic classifier. Much stronger background rejection is achieved at the same signal retention.

Performance:

- Signal retention: 99.0%
- Background rejection: 91.0%
- ROC AUC: 0.9965
- Accuracy: 95.0%
- Precision: 92.0%

This optimistic scenario illustrates the potential performance gains from spatial correlation and overlap-aware features, albeit with assumptions about readout that may prove infeasible in practice.

7. Conclusions and Future Work

This study represents a detailed investigation into beam background and physics clusters in the first layer of the CLD vertex detector of the FCC-ee. We constructed and interpreted cluster structures, explained detector-specific effects due to module overlaps, and implemented effective signal–background discrimination pipelines based on realistic readout assumptions.

Key Contributions:

- Explained a symmetrical two-band structure in $\ln(\Delta z)$ vs. $\cos\theta$ through module overlap geometry.
- Designed an overlap-aware merging algorithm that restores geometric consistency and suppresses artificial geometric structures.
- Characterized muon and $Z \to q\bar{q}$ cluster properties at the single-module level to reflect conservative readout assumptions.
- Introduced a residual filtering procedure that isolates relevant physics signal by removing beam background-like secondaries with low MC energy.
- Developed XGBoost classifiers under both conservative and optimistic readout assumptions, achieving strong signal retention and background rejection.

Next Steps:

- Evaluate the impact of classification on downstream physics observables—such as invariant mass distributions—ensuring the preservation of physics signal integrity.
- Generalize the approach to other physics processes at the Z pole.

In short–this study establishes an initial understanding of hit clusters in the FCC-ee vertex detector and lays a framework in which real-time, module-level suppression of beam background can be performed.