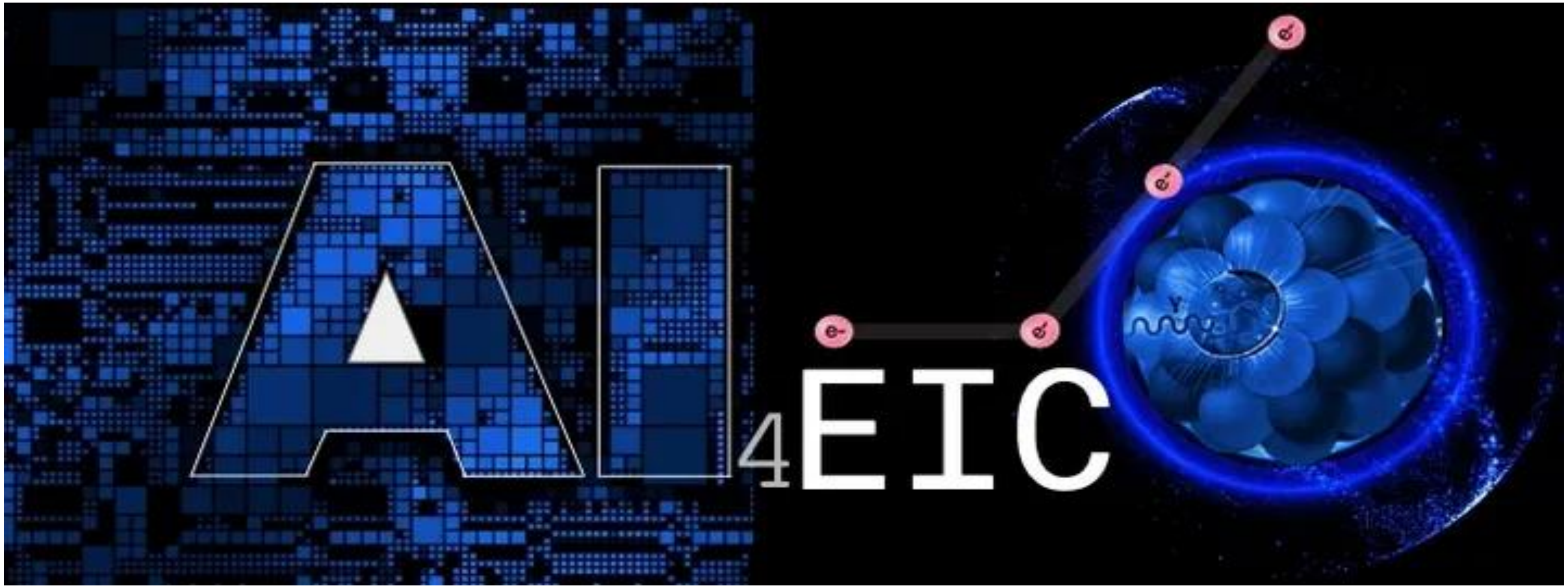
SRO AI/ML Models for Meson Structure Function Extraction



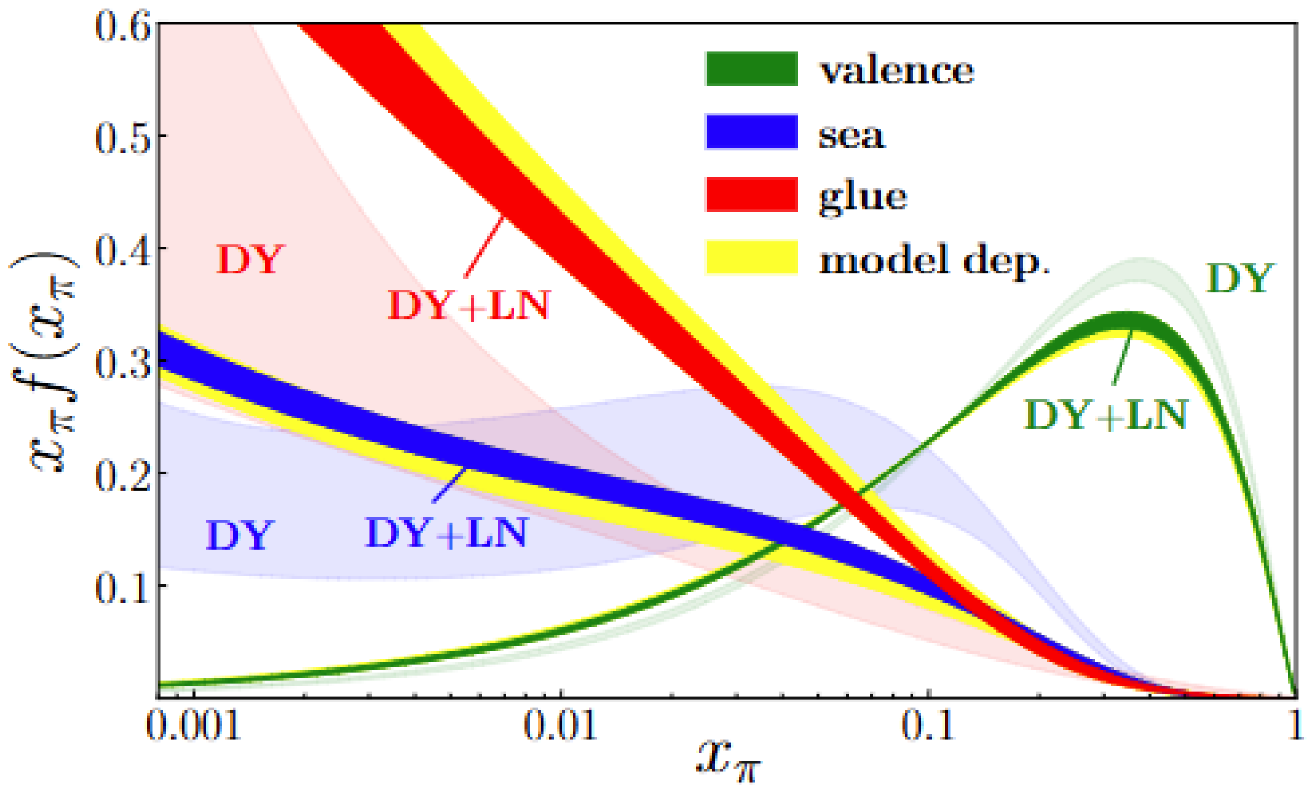
Multi-Method Ensemble for Anomaly Detection in EIC Reconstructions for SRO Frames with Background

Collaborators – Sandeep Shiraskar, Dominick Rizk, Tanja Horn, Dmitry Romanov, Baptiste Fraïsse, Avnish Singh.

THE CATHOLIC UNIVERSITY OF AMERICA

Jefferson Lab

Meson Structure Functions in QCD



Pion PDFs xfx at Q^2 = 10 GeV² from global QCD fit (DY + LN data), decomposing valence, sea, and gluon contributions analogous to F_2^π(x, Q^2) in meson DIS.

Reference : Barry, Sato, WM, C.-R. JiPRL 121, 152001 (2018)

EIC design is well suited for Sullivan Process, structure function measurements.

Mesons (e.g., pion) as quark-antiquark pairs reveal non-perturbative QCD dynamics

- $\boldsymbol{F_2(x, Q^2)}$:Structure function measuring quark/gluon momentum fractions in mesons via DIS
- $\boldsymbol{x}$**:** Bjorken fraction of meson's momentum carried by parton (0 to 1)
- $\boldsymbol{Q^2}$:Momentum transfer squared; probes resolution inside meson
- DGLAP Evolution: PDF scaling with Q^2 from gluon radiation and splitting
- EIC Endpoint (x ≈ 1): Shape functions link QCD to HQET for heavy mesons

Dataset

Reconstructed Output data, as documented in Background Mixed Samples

More Information: Background wiki

Event	signal	synrad	ebrems	etouschek	ecoloumb	p.b.gas
Event 1	●	● ● ● ●	●			
Event 2	●	● ● ● ●			●	●
Event 3	●	● ● ●		●		
Event 4	●	● ● ●	●			●
Event 5	●	● ● ● ●	●			

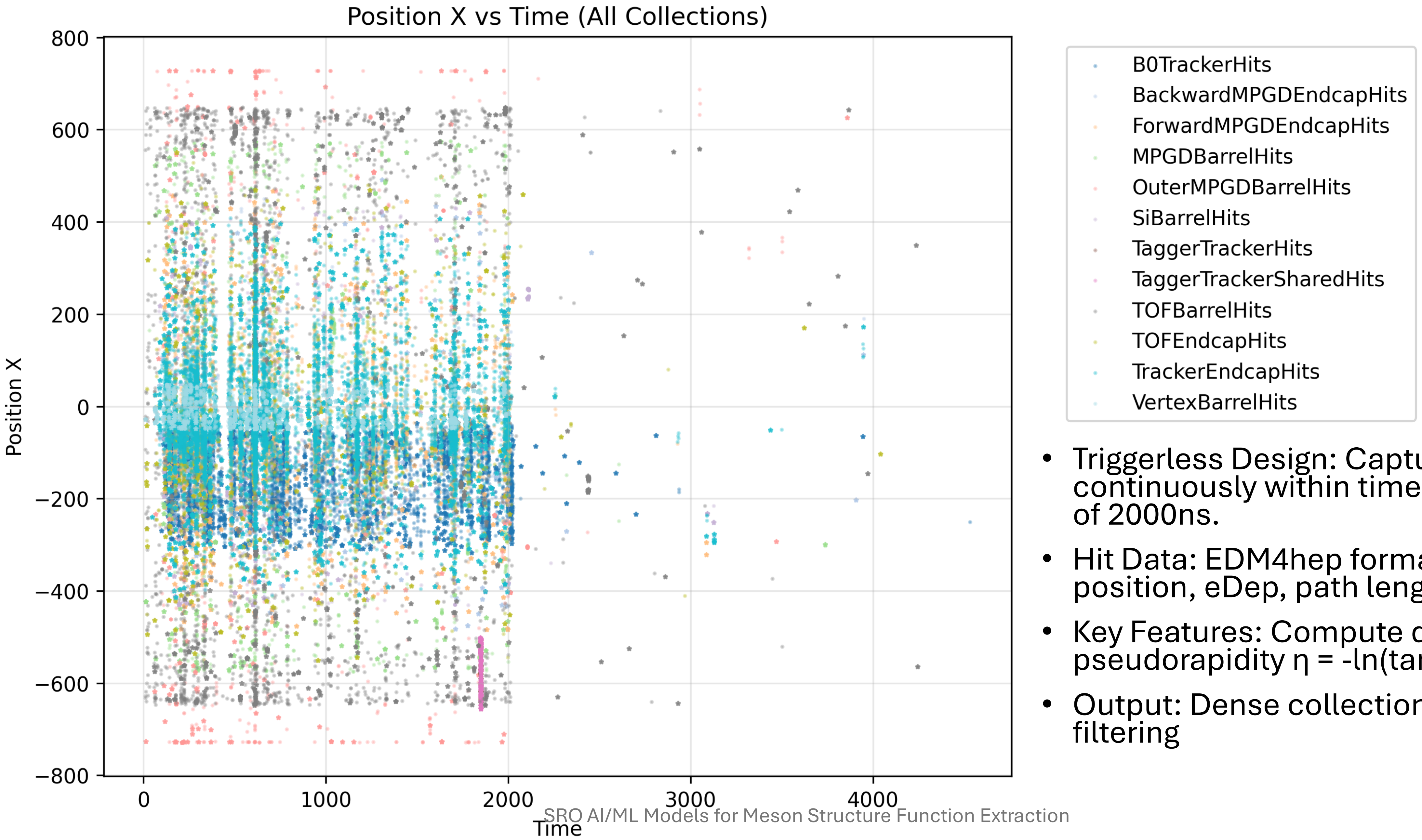
Signal frequency of 500 kHz, each event contains at least 1 signal contribution

Symbol	Process	Description	Sampling Frequency (kHz)	Status Code Shift
●	signal	DIS NC 18x275 Q^2>1 (Deep inelastic scattering neutral current)	500	0
●	synrad	Synchrotron Radiation	14000	2000
●	ebrems	Electron bremsstrahlung radiation	316.94	3000
●	etouschek	Electron Touschek scattering (intrabeam scattering)	1.3	4000
●	ecoulomb	Electron Coulomb scattering processes	0.72	5000
●	p.b.gas	Proton beam gas interactions	22.5	6000

The background contributions get allocated per event based on their sampling frequency.

Processes with less than 500 kHz sampling frequency are not guaranteed a contribution in every event.

Background Enriched Data in EIC

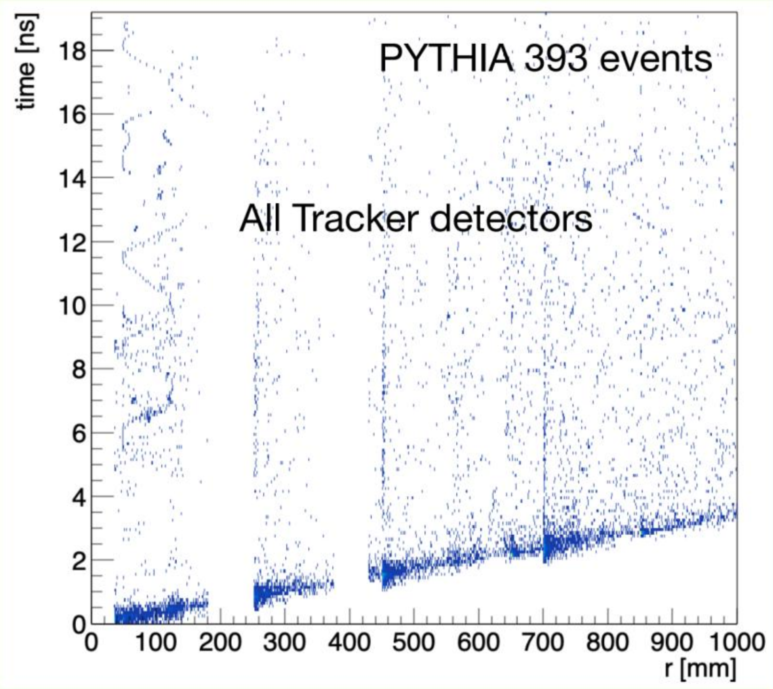


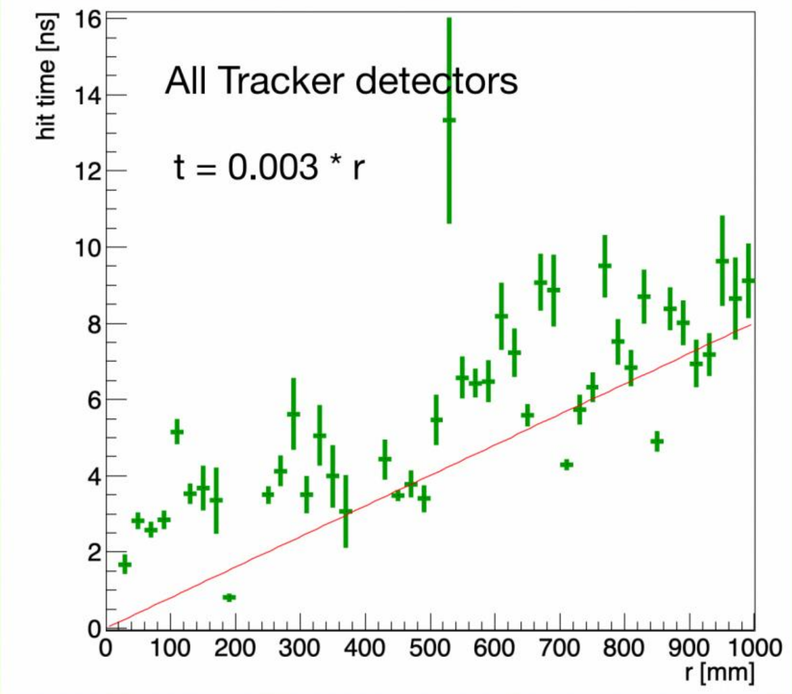
- Triggerless Design: Captures all events continuously within timestap intervals of 2000ns.
- Hit Data: EDM4hep format with position, eDep, path length, time
- Key Features: Compute dE/dx ratios, pseudorapidity $\eta = -\ln(\tan(\theta/2))$
- Output: Dense collections for ML filtering

Without Using Machine Learning

- One of the preTDR goals is software able to work with data frames

These plots show the relation between hit time vs hit position to alignment hit time.

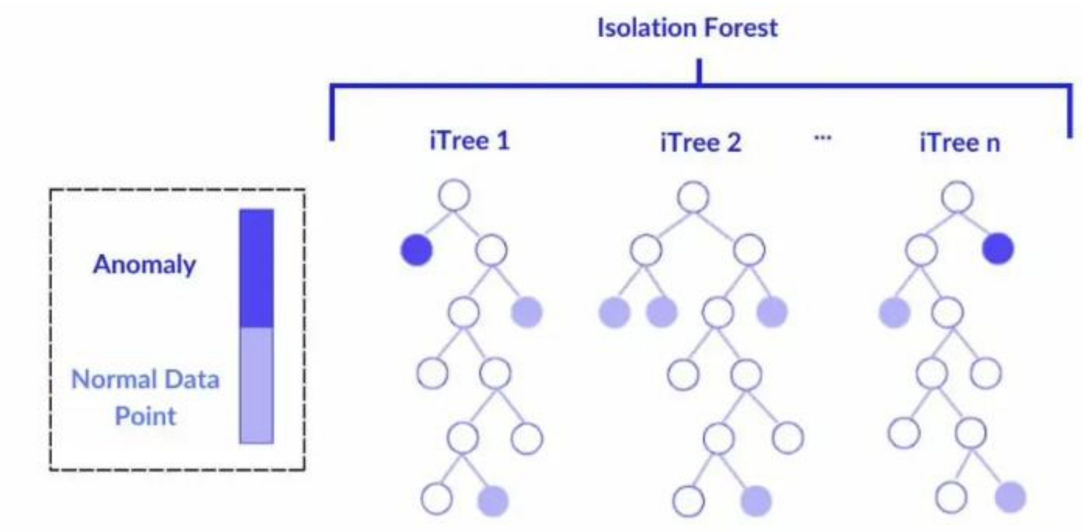




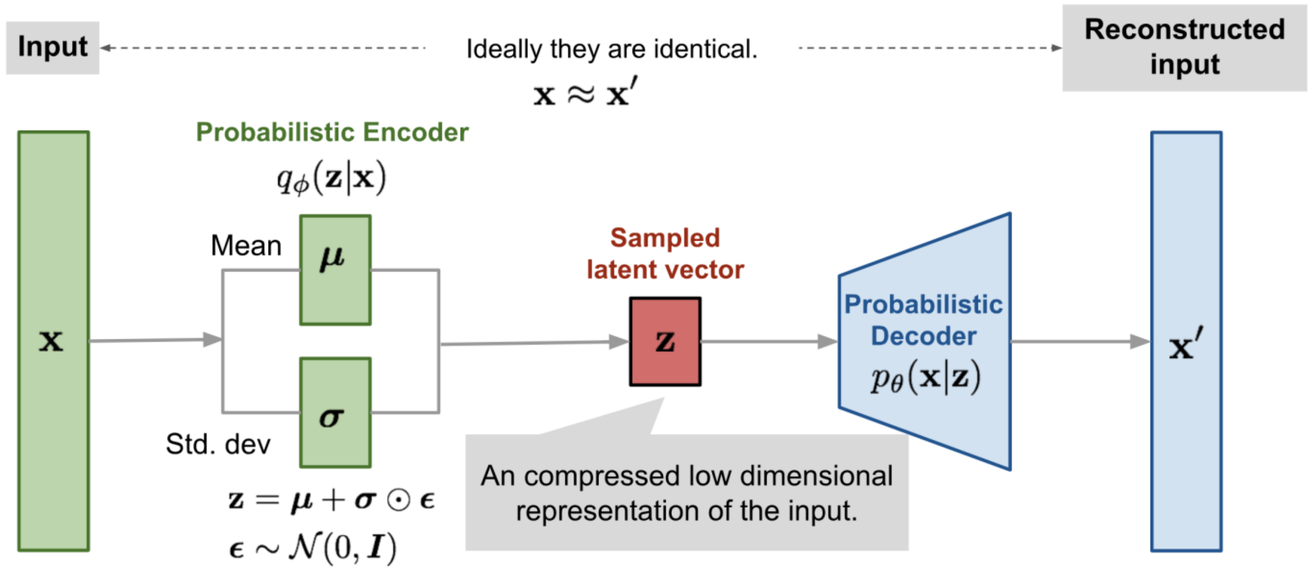
→A correlation between distance and hit time is observed.
However, since we are looking at MC hits, this result is expected (W/O detector response).

Reference : Takuya Kumaoka and Taku Gunji

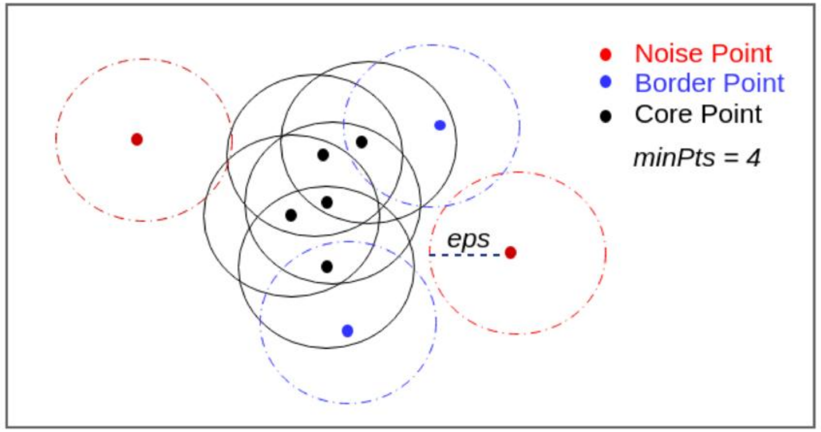
AI/ML for Anomaly Detection



Isolation Forest: Scores outliers by random feature splits



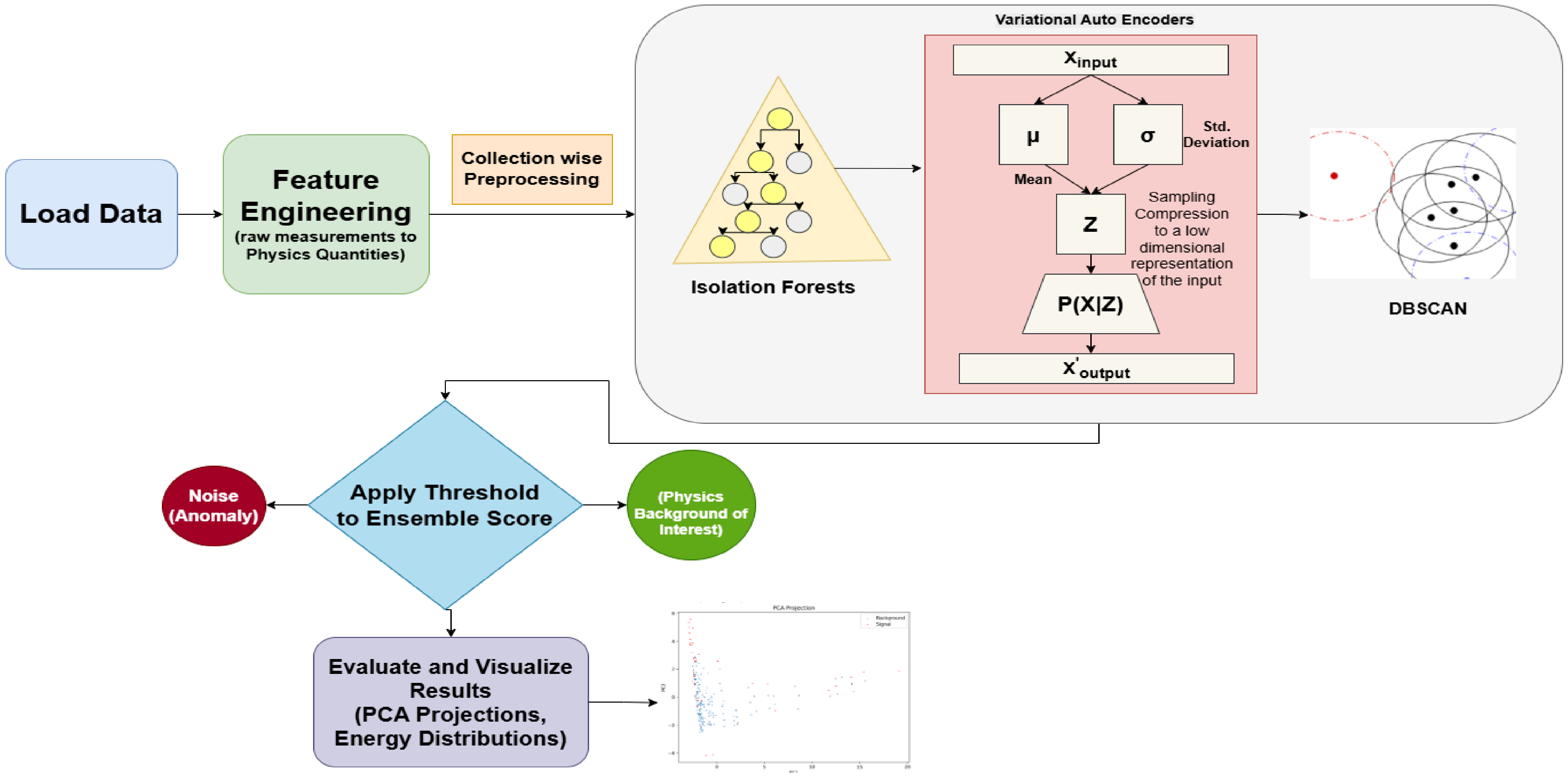
VAEs: Reconstruct hits, flag high MSE deviations



DBSCAN: Clusters dense signals vs. isolated noise

ML Pipeline for Anomalous Noise Detection

Schematic Illustration of the HEPSignalClassiferEnsemble



Feature Engineering

Table 1: Physics-Motivated Feature Engineering Summary

Feature Group	Quantity	Description / Formula
Kinematic Quantities	Total momentum magnitude	$\lvert p\rvert = \sqrt{p_x^2 + p_y^2 + p_z^2}$
	Transverse momentum	$p_T = \sqrt{p_x^2 + p_y^2}$, perpendicular to beam axis
	Momentum polar angle	$\theta_{\mathrm{momentum}} = \arccos(p_z/\lvert p\rvert)$, clipped to $[-1, 1]$ with zero-division handling
	Momentum azimuthal angle	$\phi_{\mathrm{momentum}} = \mathrm{atan2}(p_y, p_x)$
	Momentum z-ratio	Longitudinal fraction $p_z/\lvert p\rvert$
Spatial Quantities	Radial distance	$r = \sqrt{x^2 + y^2 + z^2}$, total distance from interaction point
	Cylindrical radius	$\rho = \sqrt{x^2 + y^2}$, distance from beam axis
	Position polar angle	$\theta_{\mathrm{position}} = \arccos(z/r)$, with clipping and zero-division handling
	Position azimuthal angle	$\phi_{\mathrm{position}} = \mathrm{atan2}(y, x)$
	Position z-ratio	Longitudinal fraction z/r
Particle ID Features	Specific energy loss	$dE/dx = eDep/pathLength$, normalized energy deposit per unit path length
	Pseudorapidity	$\eta = -\ln(\tan(\theta/2))$, standard detector acceptance variable
	Velocity	$v = pathLength/time$
	Beta factor	$\beta = v/c$, where $c = 299.792$ mm/ns
	Mass-squared estimator	$m^2 = \lvert p\rvert^2(1/\beta^2 - 1)$, from relativistic kinematics
Vertexing Features	Transverse impact parameter	$d_0 = \rho$, closest approach to beam axis
	Longitudinal impact parameter	$z_0 = z$ (implicit from vertex coordinates)

Calculate physics quantities (total momentum, energy loss rate, angles, etc.) instead of using raw numbers

Why? Generalize across detectors, standard in HEP ML

Training the Models : Transitioning to Anomalous Noise Detection

IF : Excels at global outliers in high-dim hit data, fast, distribution-free for noisy momenta/eDep

Ensemble of isolation trees (n_estimators=100, contamination=0.10, random_state=42)

OUTPUT:
Binary labels {-1 anomaly, +1 normal}; scores (lower = more isolated)

VAE : Learns complex normal patterns in kinematics/PID features—flags recon drifts via poor fits.

Encoder/Decoder: Input → 64 ReLU → 64 ReLU → 16D latent (μ, σ); z = μ + ε·σ; symmetric decode

Loss: MSE recon + KL divergence; Adam (lr=0.001), 50 epochs, full-batch

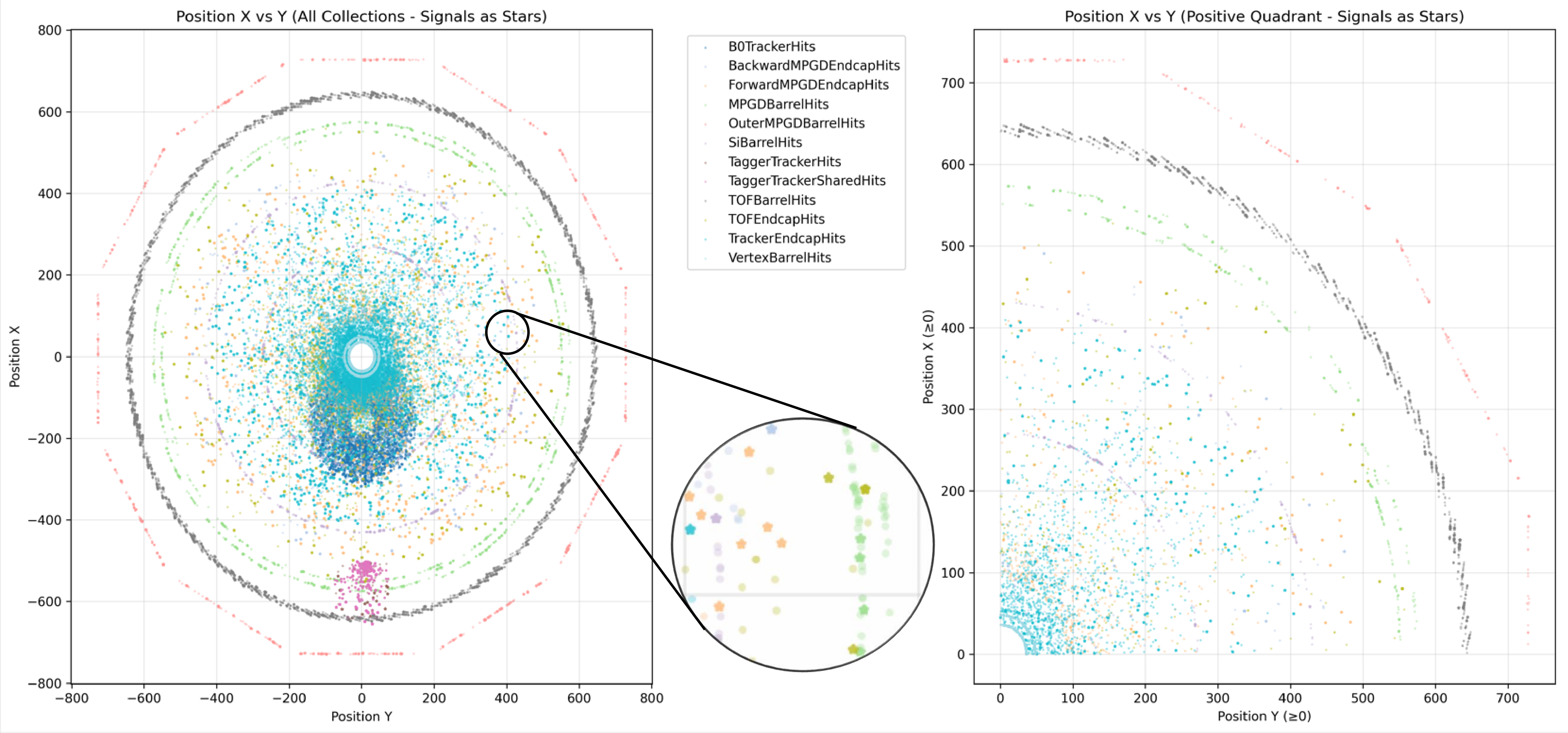
OUTPUT:
Anomaly: Recon error >90th percentile (~10% flagged)

DBSCAN : Identifies sparse spatial anomalies without fixed cluster count—ideal for geometry-tied hits.

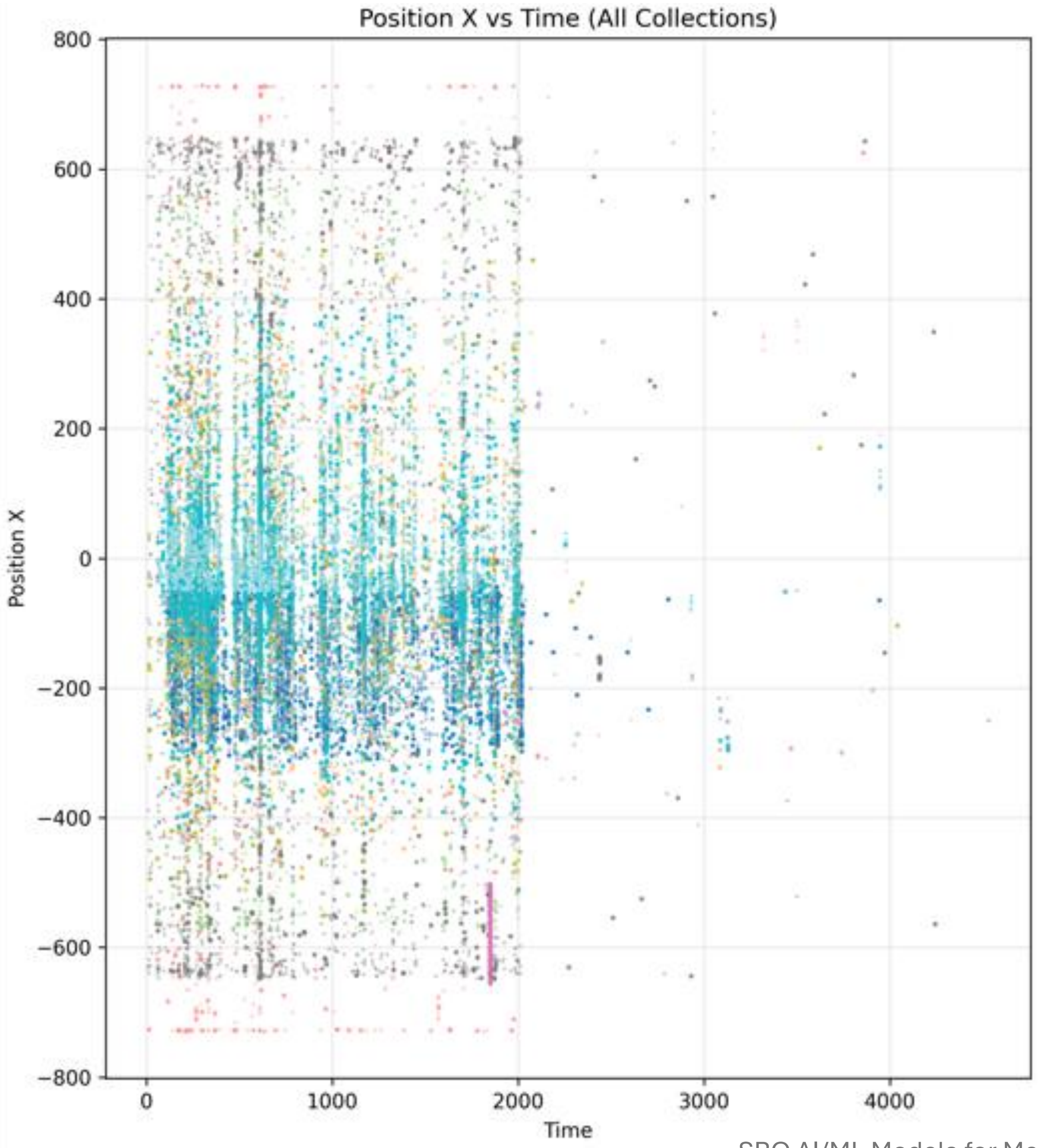
Density-based clustering (eps=0.5; min_samples adaptive to N)

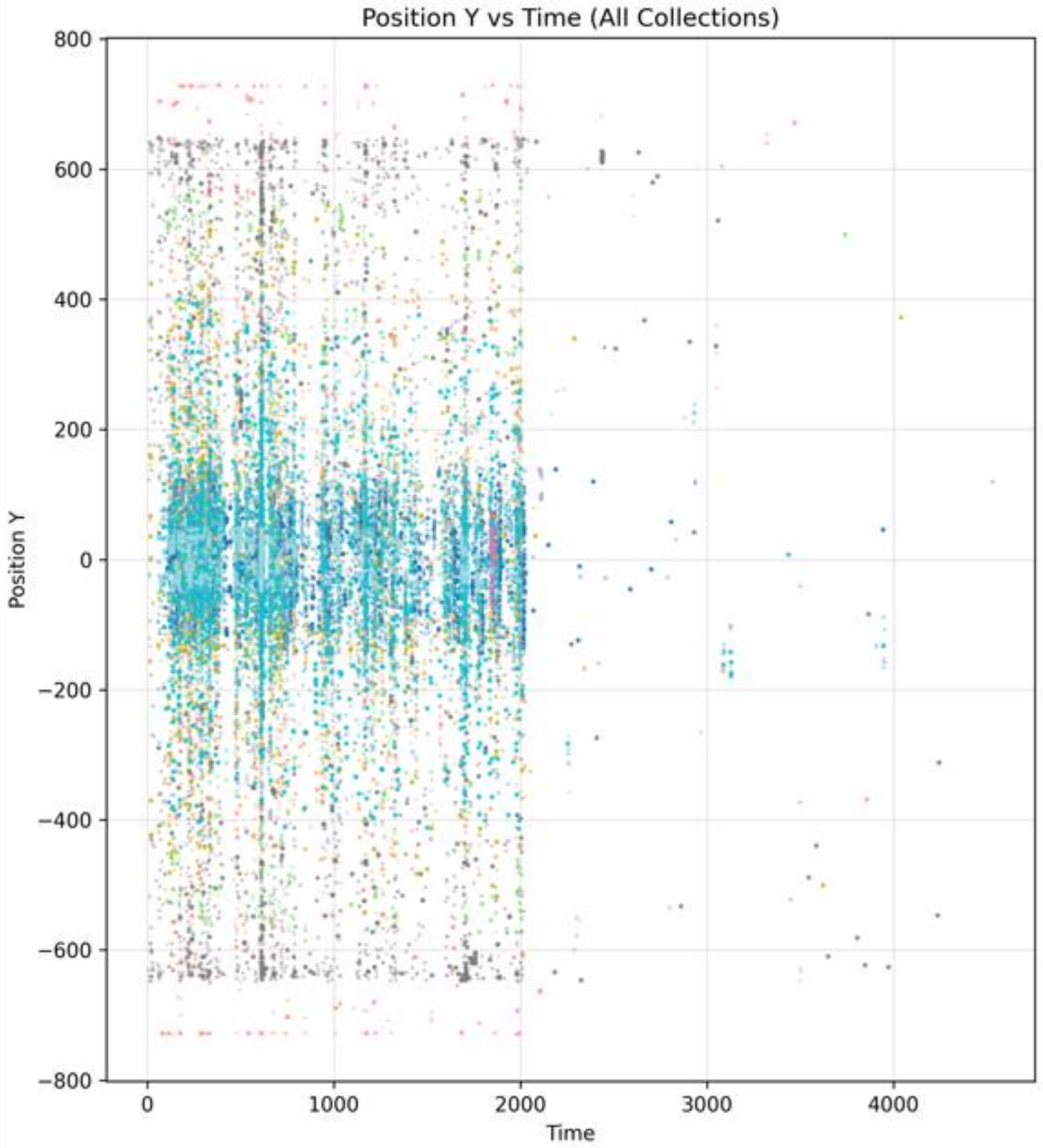
OUTPUT:
Labels {0 physics background, -1 noisy signals}; metrics (clusters, core points, noise count)

Results



Position X vs Y (All Collections - Signals as Stars)
Position X
Position Y
B0TrackerHits
BackwardMPGDEndcapHits
ForwardMPGDEndcapHits
MPGDBarrelHits
OuterMPGDBarrelHits
SiBarrelHits
TaggerTrackerHits
TaggerTrackerSharedHits
TOFBarrelHits
TOFEndcapHits
TrackerEndcapHits
VertexBarrelHits
Position X vs Y (Positive Quadrant - Signals as Stars)
Position X (≥0)
Position Y (≥0)


Position X vs Time (All Collections)
Position X
Time
800
600
400
200
0
−200
−400
−600
−800
0
1000
2000
3000
4000


Position Y vs Time (All Collections)
Position Y
Time
800
600
400
200
0
−200
−400
−600
−800
0
1000
2000
3000
4000

Interpretation of Results

Validation Metrics

- **VAE Reconstruction Error**: MSE between input hits and model reconstruction
- **Low MSE** = normal (reconstructed well); High MSE = anomaly (poor fit)

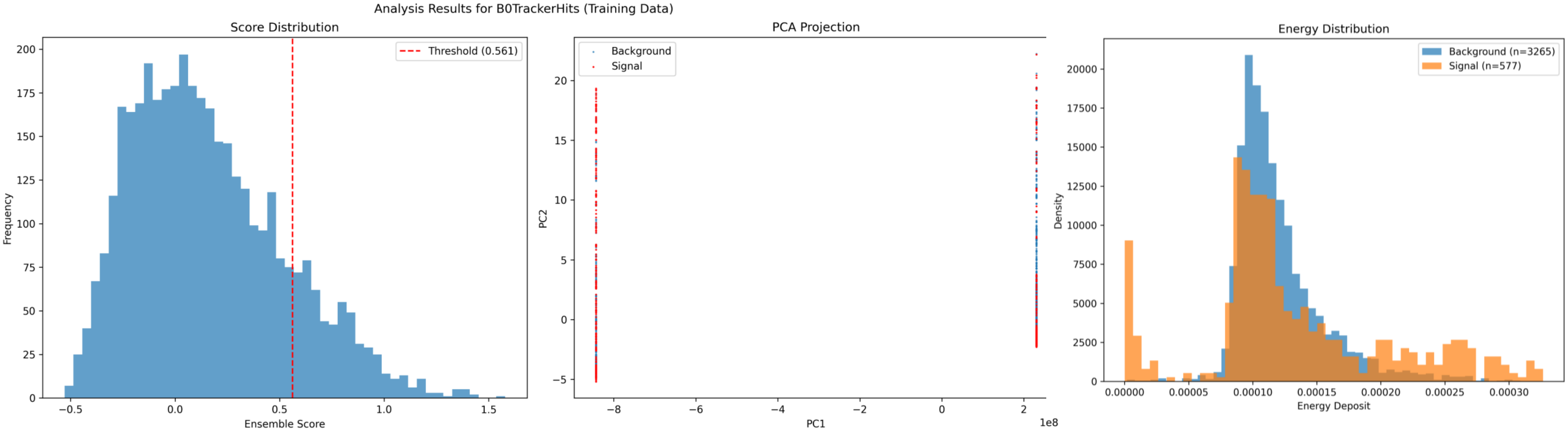
***Training Assumption**: Fit on majority "normal" data to learn baseline.

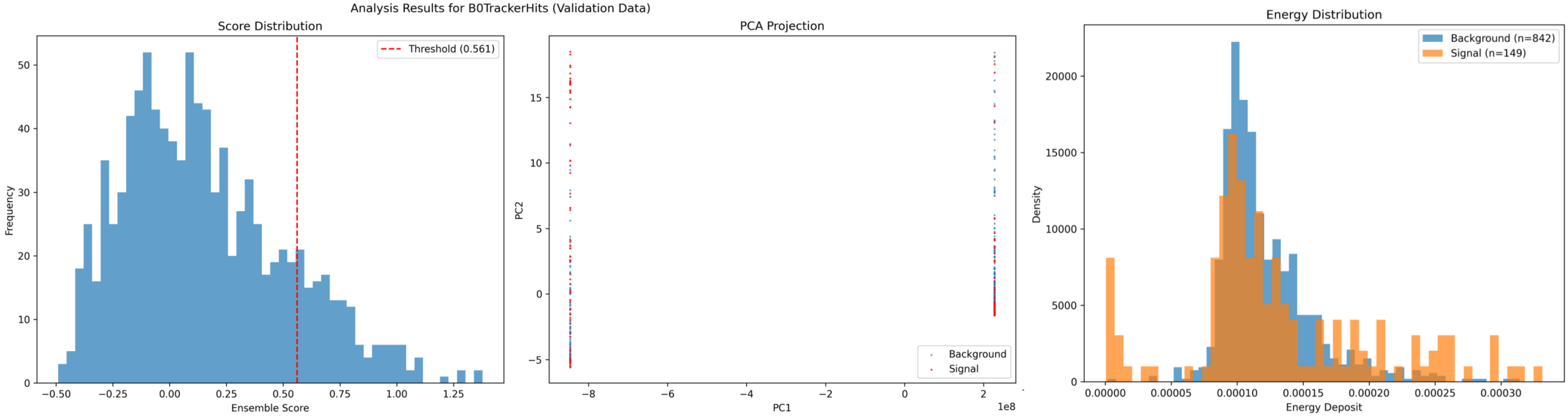
***Threshold Setting**: 90-95th percentile of training errors for flagging

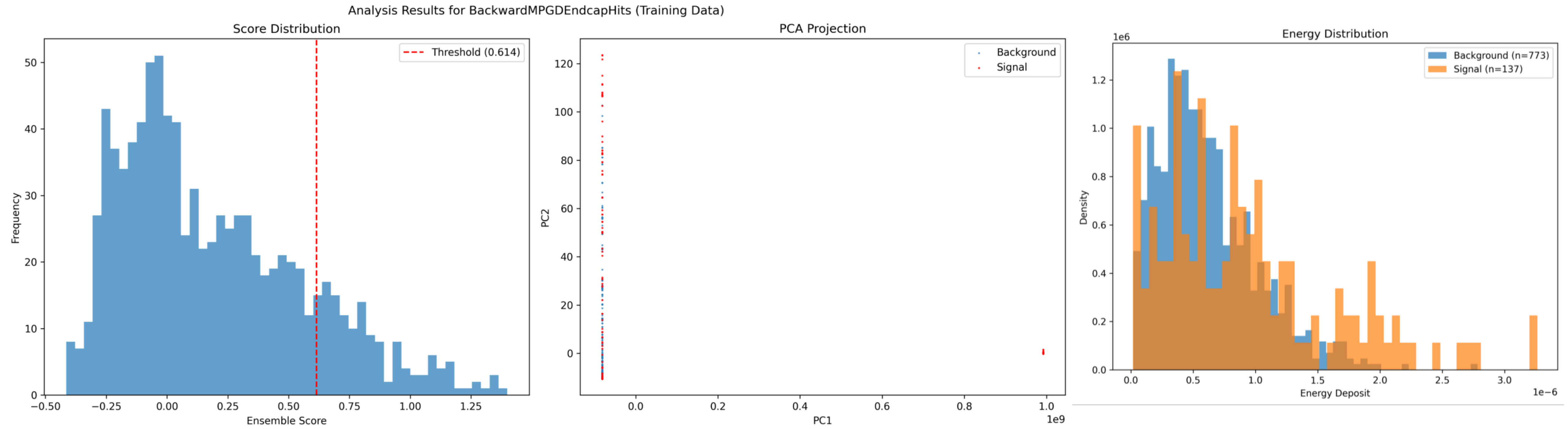
- **Validation Process: Apply to test set; compare anomaly scores across methods**

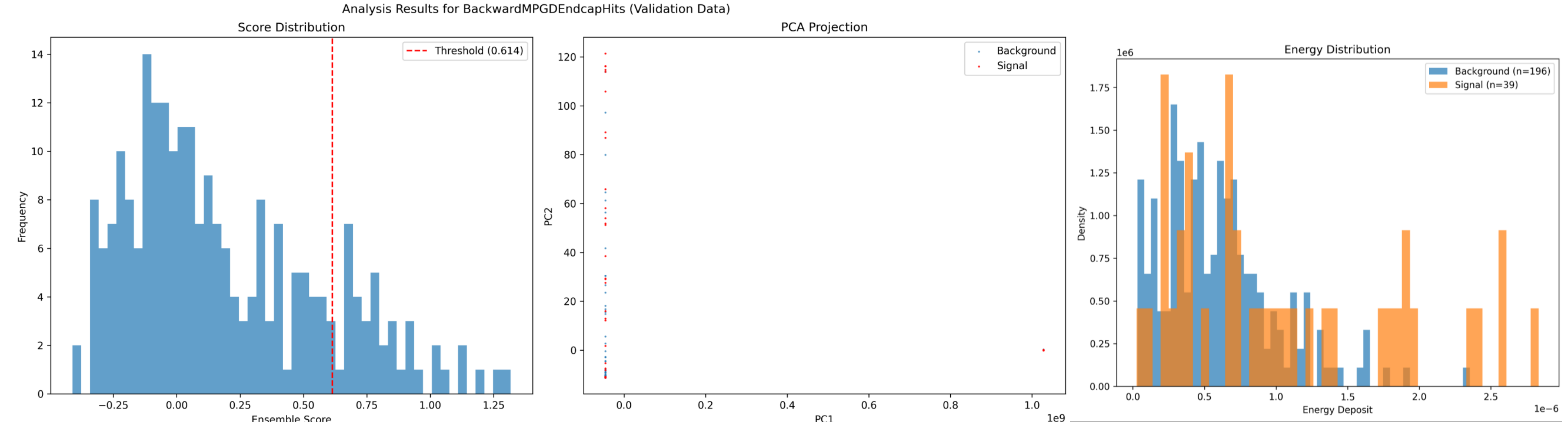
Pipeline Integration: MSE feeds into ensemble with IF/DBSCAN for final score

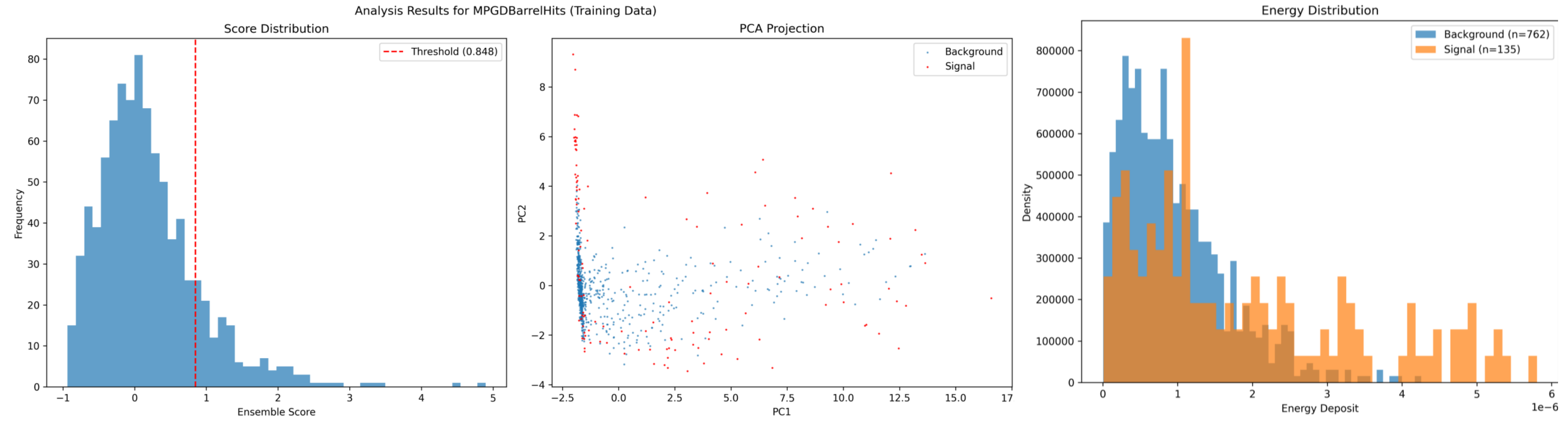
- **Score Distributions**: Histograms show ensemble score tails; threshold separates normal from anomaly
- **PCA Projections**: 2D plots cluster background vs. signal; outliers deviate from main distribution
- **Energy Checks (dE/dx):** anomalies show inconsistent energy loss
- ***Validation Logic: High anomaly score = low reconstruction error = flagged for ensemble***
- Pipeline Check: Compare predictions across collections for consistency

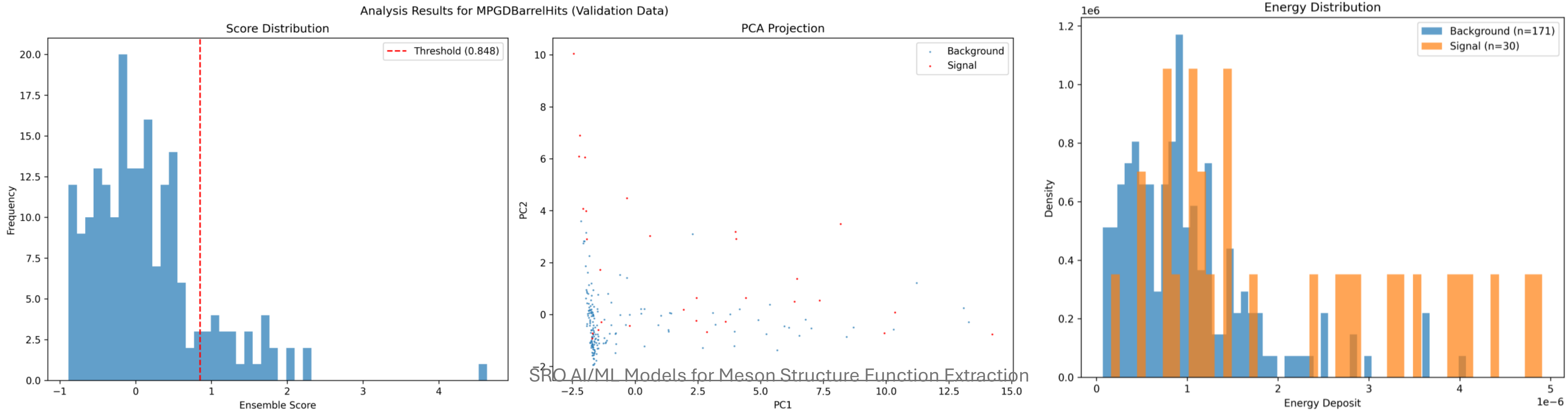

Analysis Results for B0TrackerHits (Training Data)
Score Distribution
Threshold (0.561)
Frequency
Ensemble Score
PCA Projection
Background
Signal
PC2
PC1
1e8
Energy Distribution
Background (n=3265)
Signal (n=577)
Density
Energy Deposit

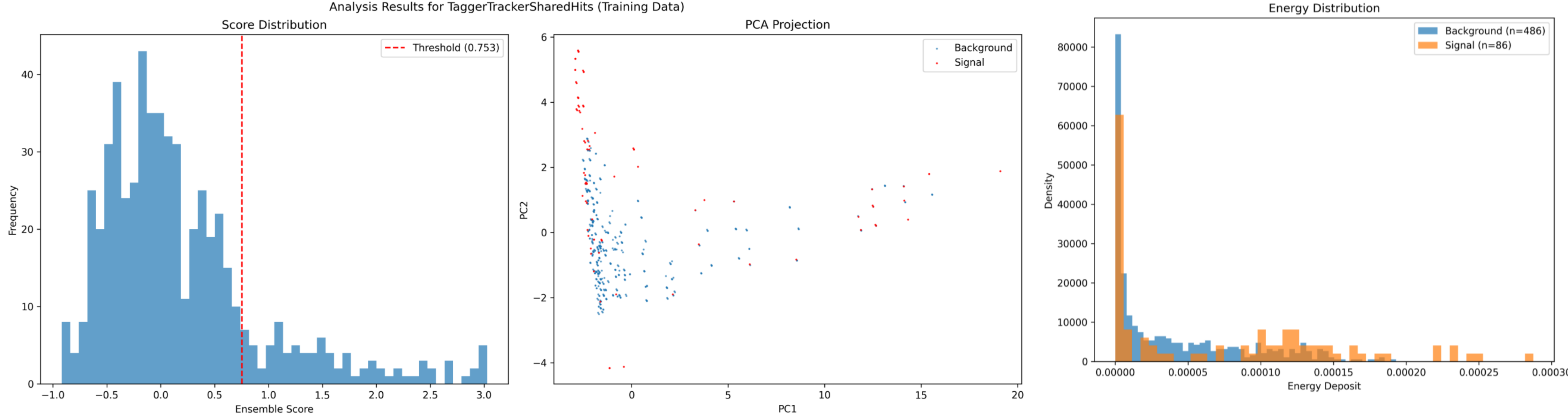

Analysis Results for B0TrackerHits (Validation Data)
Score Distribution
Threshold (0.561)
Frequency
Ensemble Score
PCA Projection
Background
Signal
PC2
PC1
1e8
Energy Distribution
Background (n=842)
Signal (n=149)
Density
Energy Deposit

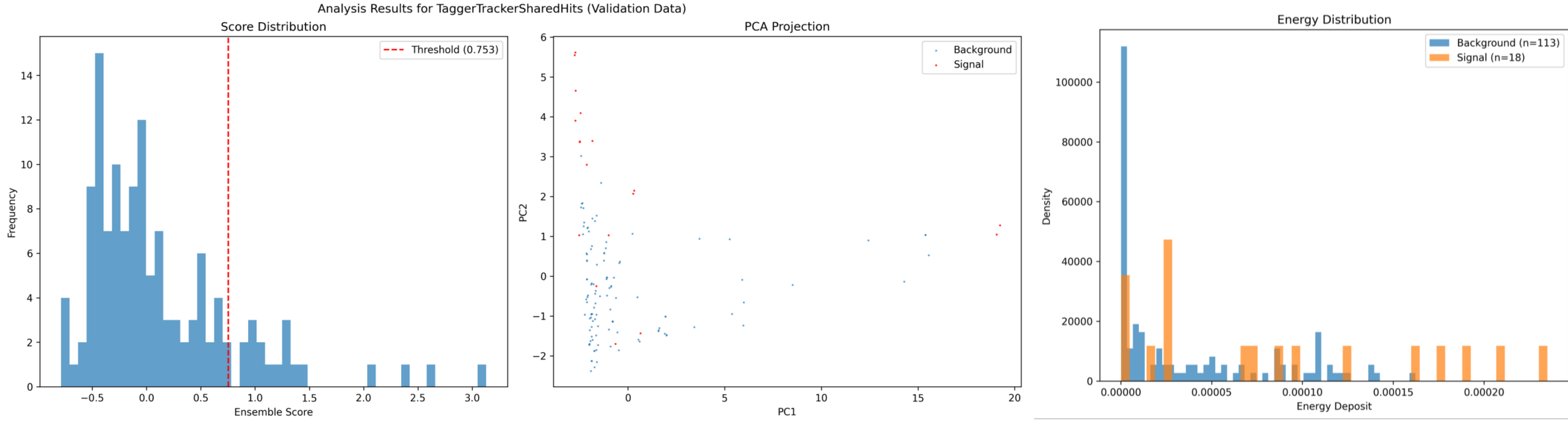

Analysis Results for BackwardMPGDEndcapHits (Training Data)
Score Distribution
Threshold (0.614)
Frequency
Ensemble Score
PCA Projection
Background
Signal
PC2
PC1
1e9
Energy Distribution
1e6
Background (n=773)
Signal (n=137)
Density
Energy Deposit
1e−6


Analysis Results for BackwardMPGDEndcapHits (Validation Data)
Score Distribution
Threshold (0.614)
Frequency
Ensemble Score
PCA Projection
Background
Signal
PC2
PC1
1e9
Energy Distribution
1e6
Background (n=196)
Signal (n=39)
Density
Energy Deposit
1e−6


Analysis Results for MPGDBarrelHits (Training Data)
Score Distribution
Threshold (0.848)
Frequency
Ensemble Score
PCA Projection
Background
Signal
PC1
PC2
Energy Distribution
Background (n=762)
Signal (n=135)
Density
Energy Deposit
1e−6


Analysis Results for MPGDBarrelHits (Validation Data)
Score Distribution
Threshold (0.848)
Frequency
Ensemble Score
PCA Projection
Background
Signal
PC1
PC2
Energy Distribution
1e6
Background (n=171)
Signal (n=30)
Density
Energy Deposit
1e−6


Analysis Results for TaggerTrackerSharedHits (Training Data)
Score Distribution
Threshold (0.753)
Frequency
Ensemble Score
PCA Projection
Background
Signal
PC1
PC2
Energy Distribution
Background (n=486)
Signal (n=86)
Density
Energy Deposit


Analysis Results for TaggerTrackerSharedHits (Validation Data)
Score Distribution
Threshold (0.753)
Frequency
Ensemble Score
PCA Projection
Background
Signal
PC1
PC2
Energy Distribution
Background (n=113)
Signal (n=18)
Density
Energy Deposit

Project Overview

AIM : showcase how AI-driven analysis of streaming readout, frame-based data can enhance physics event detection in noisy environments.

To achieve this, we are:

• Developing a robust ML pipeline tailored for efficient processing and analysis

This approach has the potential to allow for scalable AI workflows, paving the way for faster, more accurate insights in high-background experimental settings.

Conclusions

- ML Pipeline: Unsupervised models (IF, VAE, DBSCAN) effectively filter out anomalies in SRO hits
- Validation Success: Consistent signal (noise detection) rates across collections
- **Physics Impact: Clean data enables precise extraction for EIC meson studies**
- Ongoing: Integrate multi-task NN training with labeled data for supervised enhancement.
- **Promise: Robust anomaly detection supports streaming data EIC reconstruction, advancing QCD insights**
- Strengths of ML In Physics: Interpretable scores, spatial correlations.“

Thank you for your time and attention !