ALERT AI Particle Identification

CLAS Collaboration Meeting

March 10th, 2026

Uditha Weerasinghe

Mississippi State University

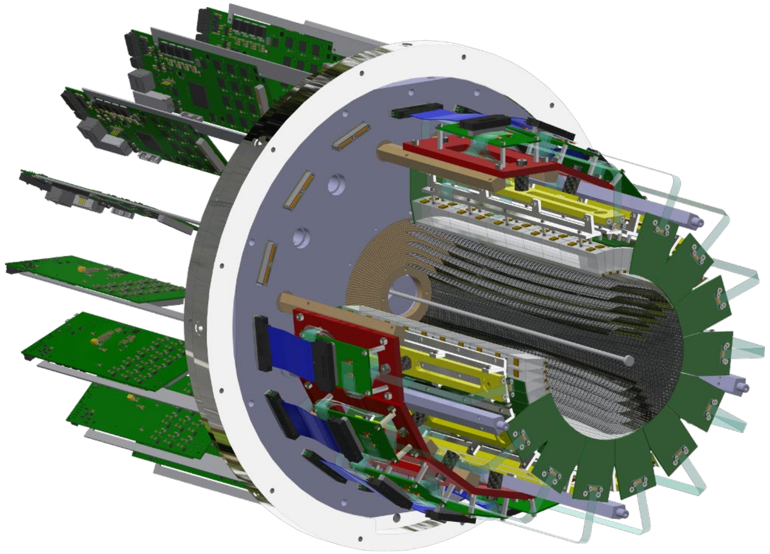
Jefferson Lab

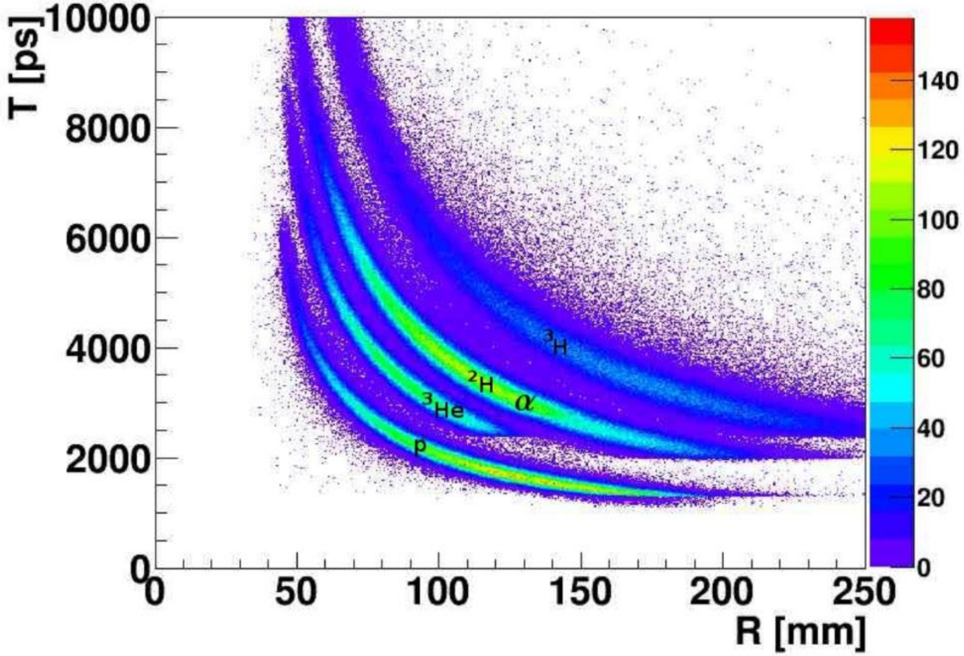
Outline

- Particle Identification in ALERT
- A Two-Stage AI PID Framework
  - Pre-KF PID
    - Features
    - Training
    - Performance
  - Post-KF PID
    - Features and Training
    - Different Approaches
- Integration to coatjava
- Simulation vs Real Data
- Ongoing and Planned Improvements
- Summary and Outlook

Particle Identification in ALERT

- ALERT detects recoil nuclear fragments:
  - ^{1}H, ^{2}H, ^{3}H
  - ^{3}He, ^{4}He
- Proposal assumptions:
  - ALERT Time of Flight (ATOF) resolves most species
  - ALERT Hyperbolic Drift Chamber (AHDC) resolves complementary combinations
- In practice:
  - Significant overlap between H and He isotopes
  - Traditional cut-based PID struggles in key kinematic regions





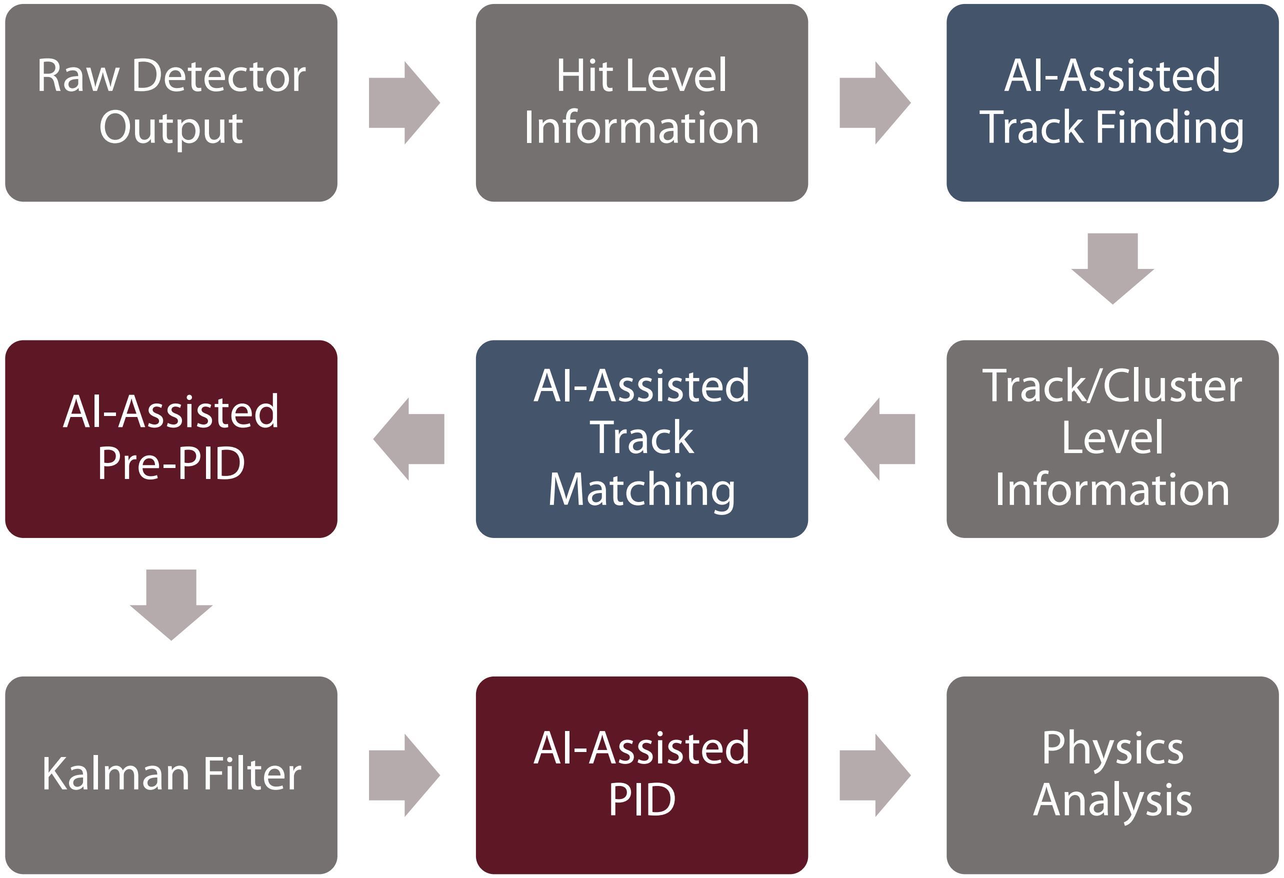
A Two-Stage AI PID Framework

PrePID (Implemented)

- Uses AHDC track + matched ATOF hit
- Provides fast probabilistic classification
- Intended to assist Kalman Filter (KF)

Post-KF PID (Developed, not yet integrated)

- Uses KF-refined kinematics
- Final physics-level classification
- Improved accuracy



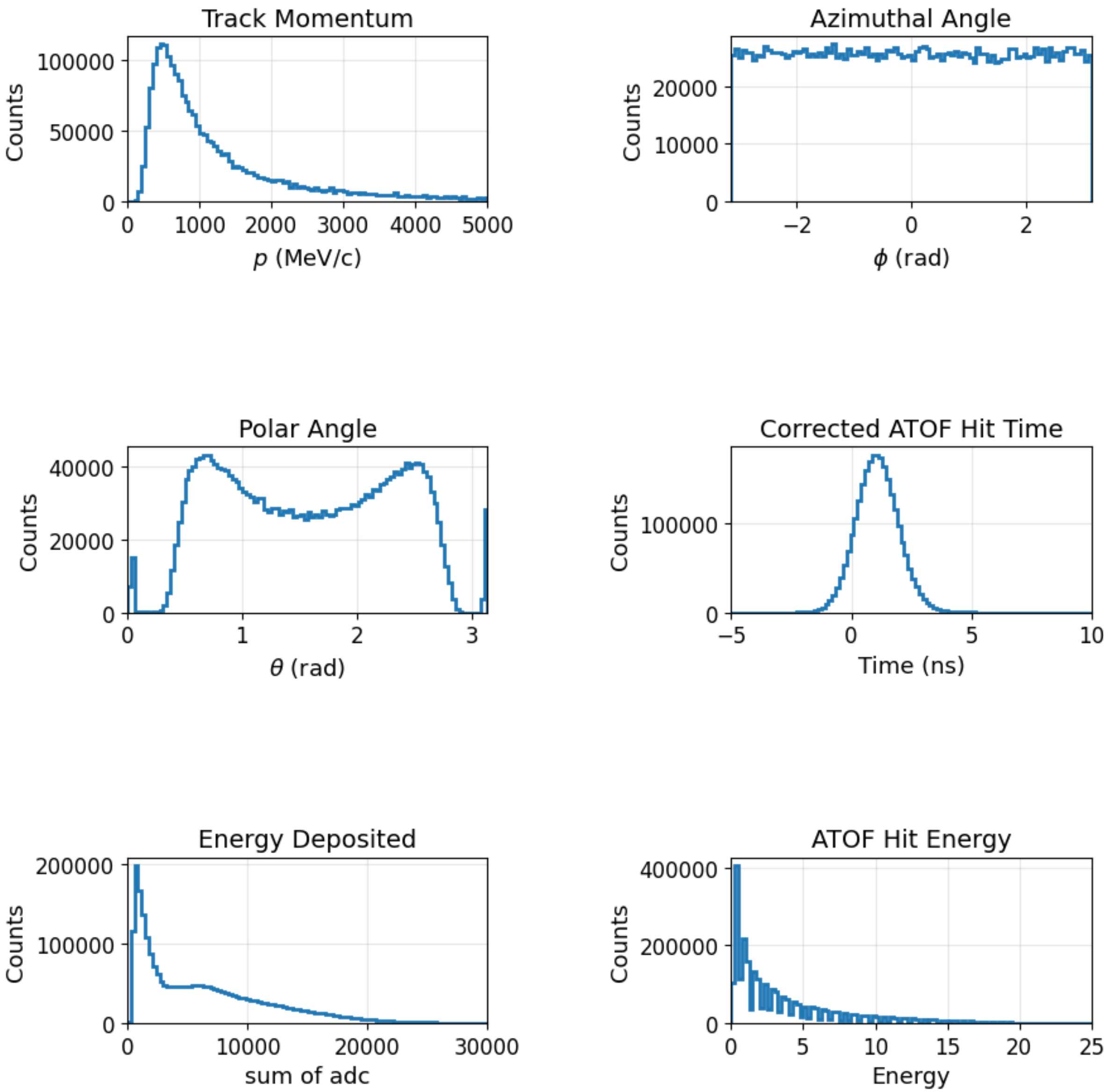
Pre-KF PID: Features

AHDC::track (10 features)

- x, y, z
- p_x, p_y, p_z
- n_{hits}
- sum(adc)
- χ^2
- sum_residuals

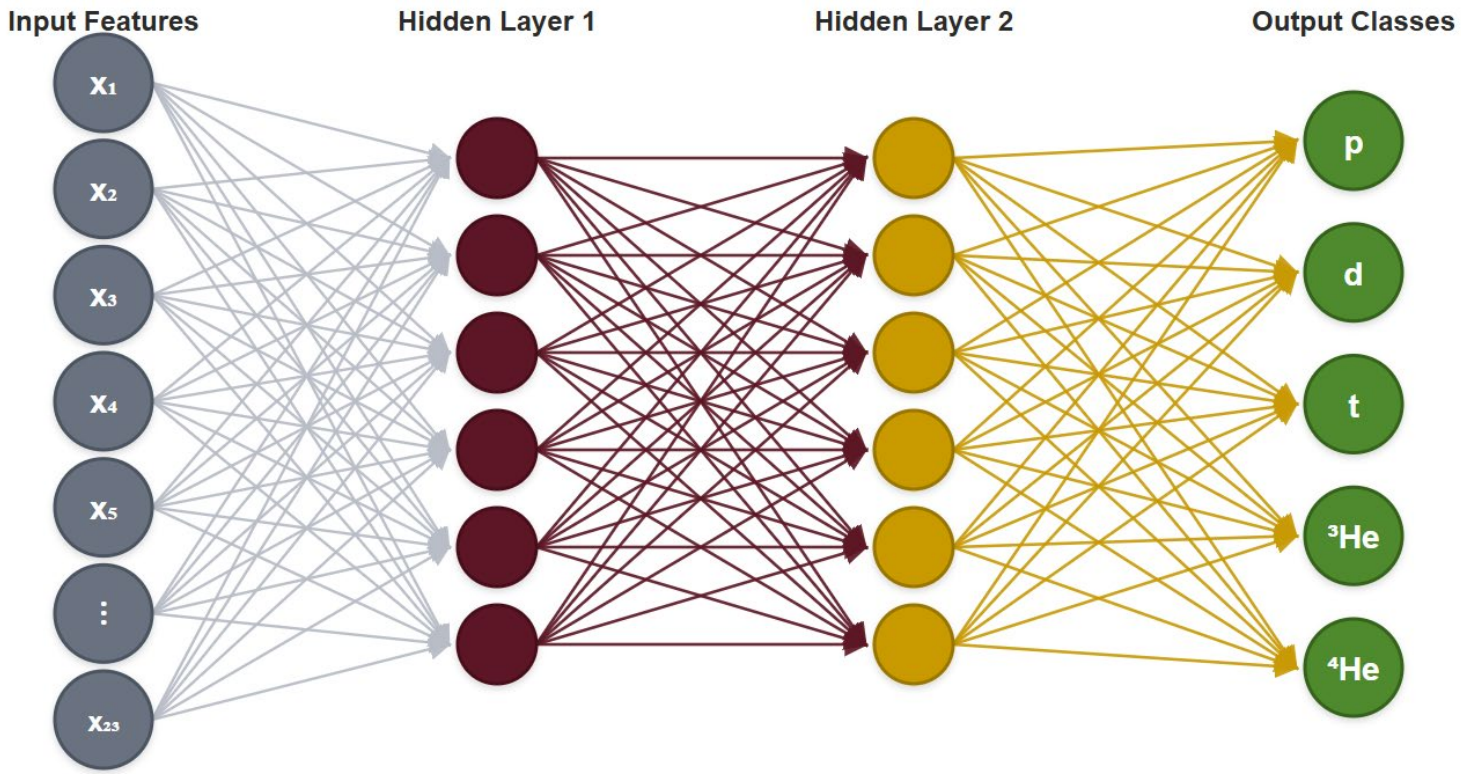
ATOF::hits (5 features)

- time
- x, y, z
- energy



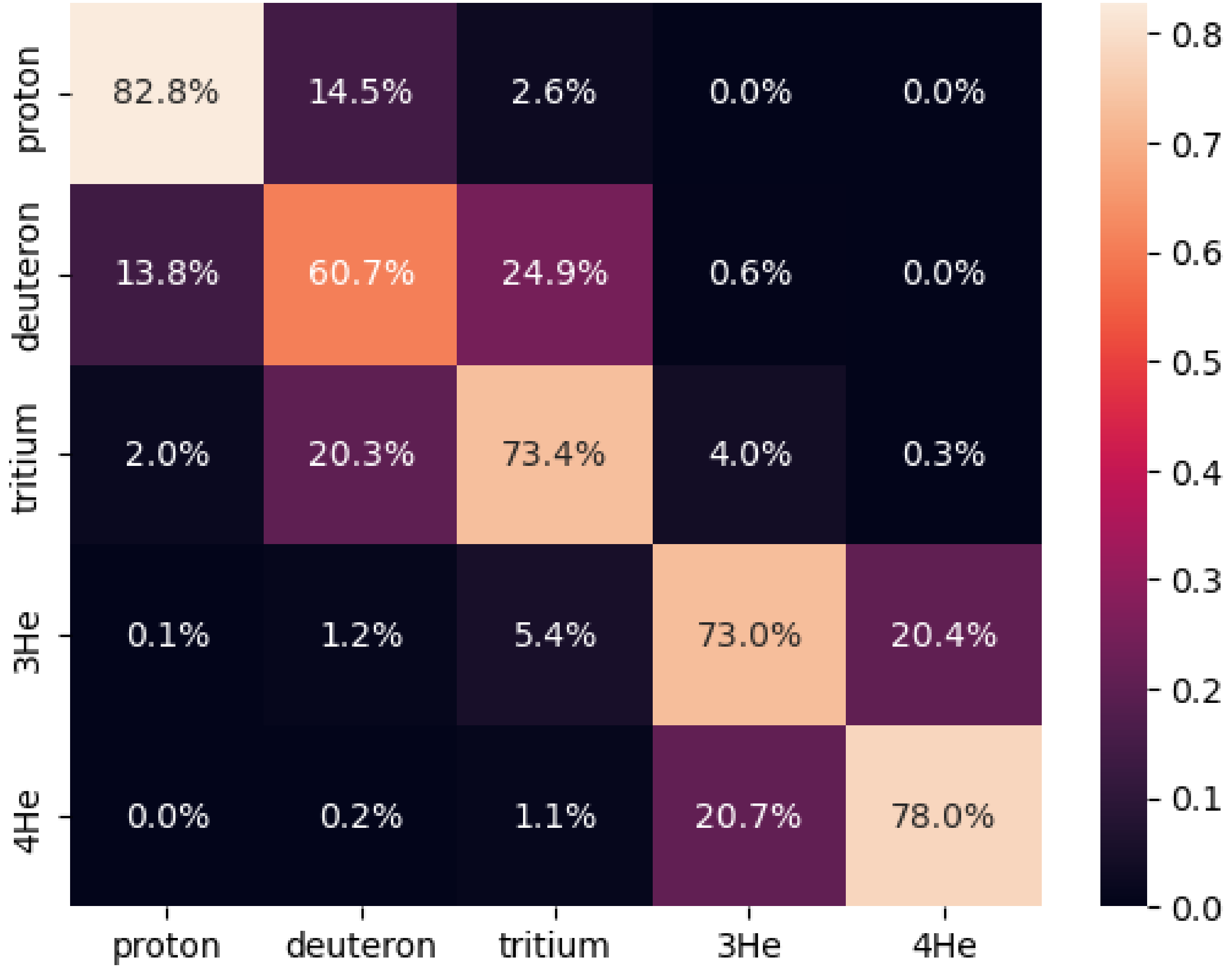
Pre-KF PID: Training

- A training dataset of 5M reconstructed tracks was prepared, with 12% containing matched ATOF hits
- The dataset was class-balanced across five particle species[^{1}H, ^{2}H, ^{3}H, ^{3}He, ^{4}He] and split into training and test samples
- A Multi-Layer Perceptron (MLP) was trained to predict the particle-type probability for each track–hit combination.



Pre-KF PID: Performance

- Overall classification accuracy is about 73% on simulated data.
- Strong proton identification where the model reliably identifies protons with above 80% accuracy.
- Weakest classification is the deuteron which shows the lowest identification performance.
- Primary confusion mode is the misclassification mainly occuring between neighboring mass species.



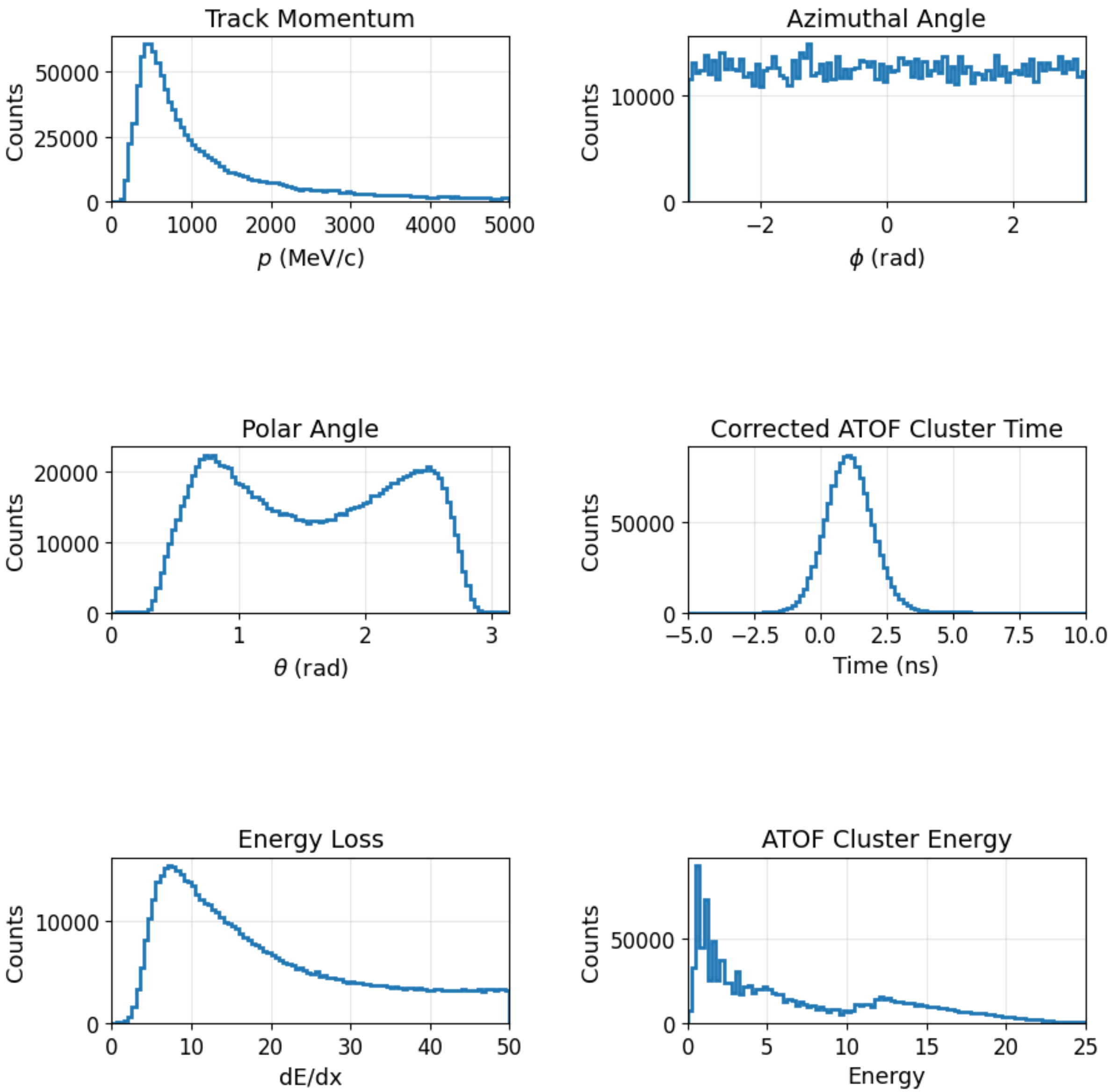
Post-KF PID: Features and Training

AHDC::kftrack (13 features)

- x, y, z
- p_x, p_y, p_z
- n_{hits}, sum(adc)
- path, dE/dx
- p_{drift}, χ^2
- sum_residuals

ATOF::clusters (9 features)

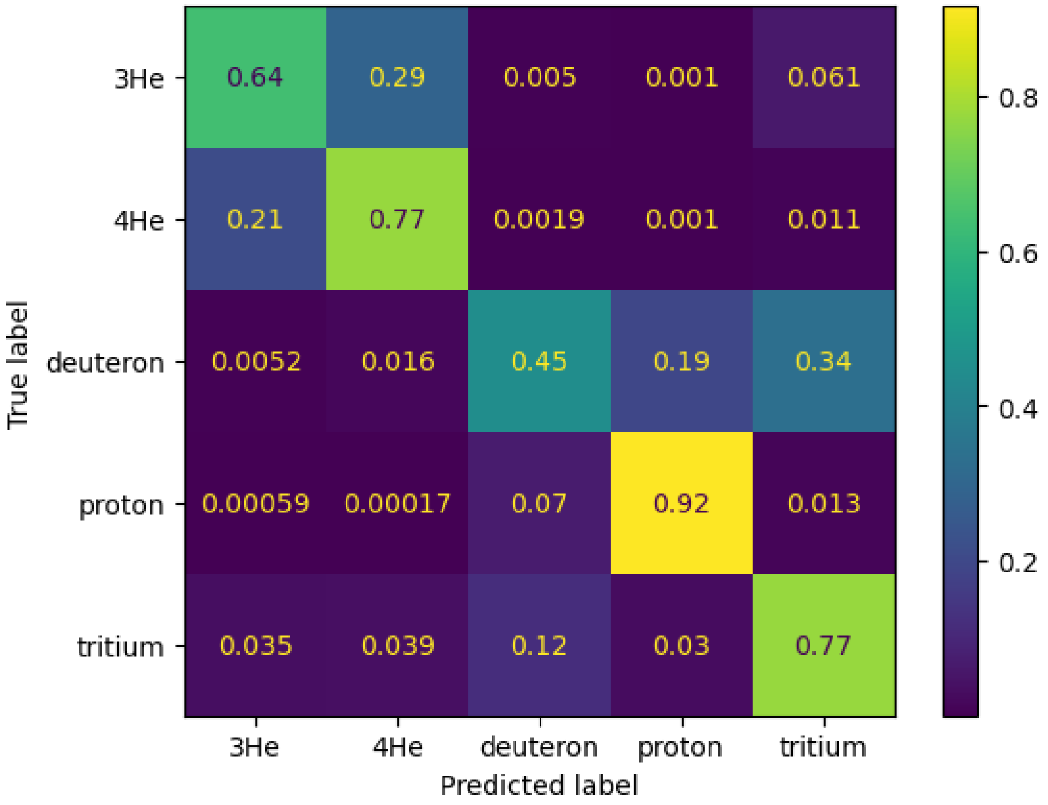
- n_bar, n_wedge
- time
- x, y, z
- energy
- pathlength, inpathlength



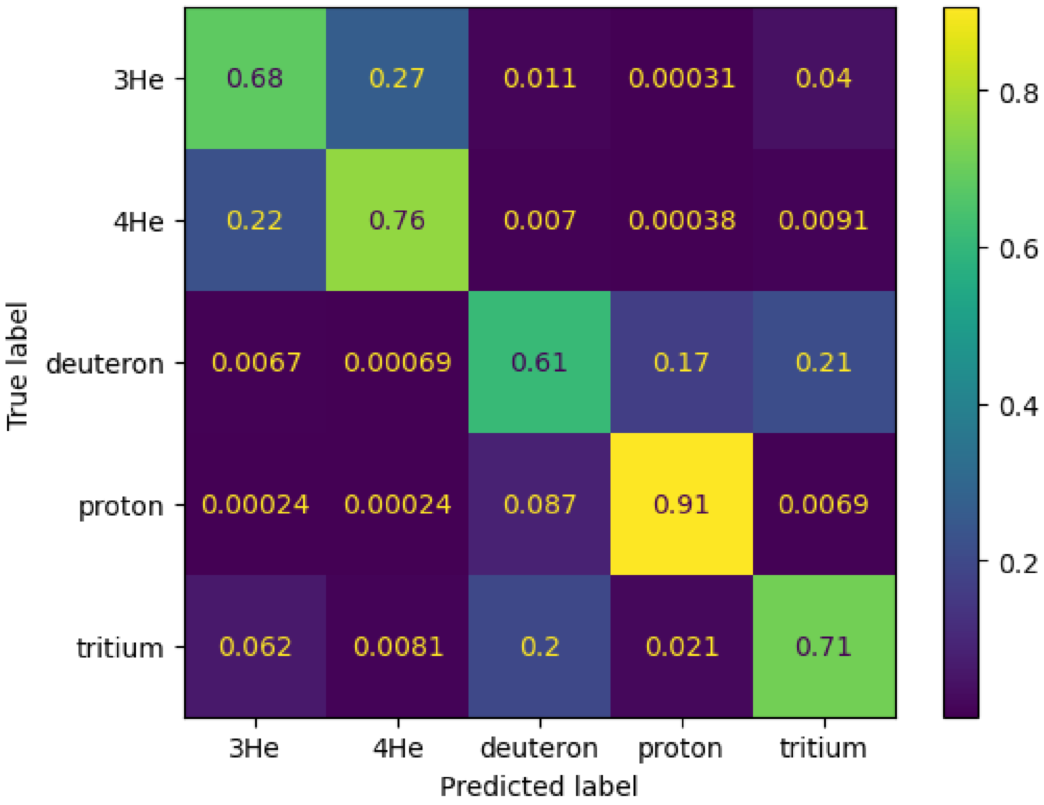
Post-KF PID: Different Approaches

- Transforming non-regular variables (e.g., p_x, p_y, p_z) into physically meaningful coordinates (p, θ, ϕ) aligns the inputs with the detector geometry and underlying kinematics
- This method reducing correlations and making patterns easier for the model to learn.

Using p_x, p_y only (dropping p_z)



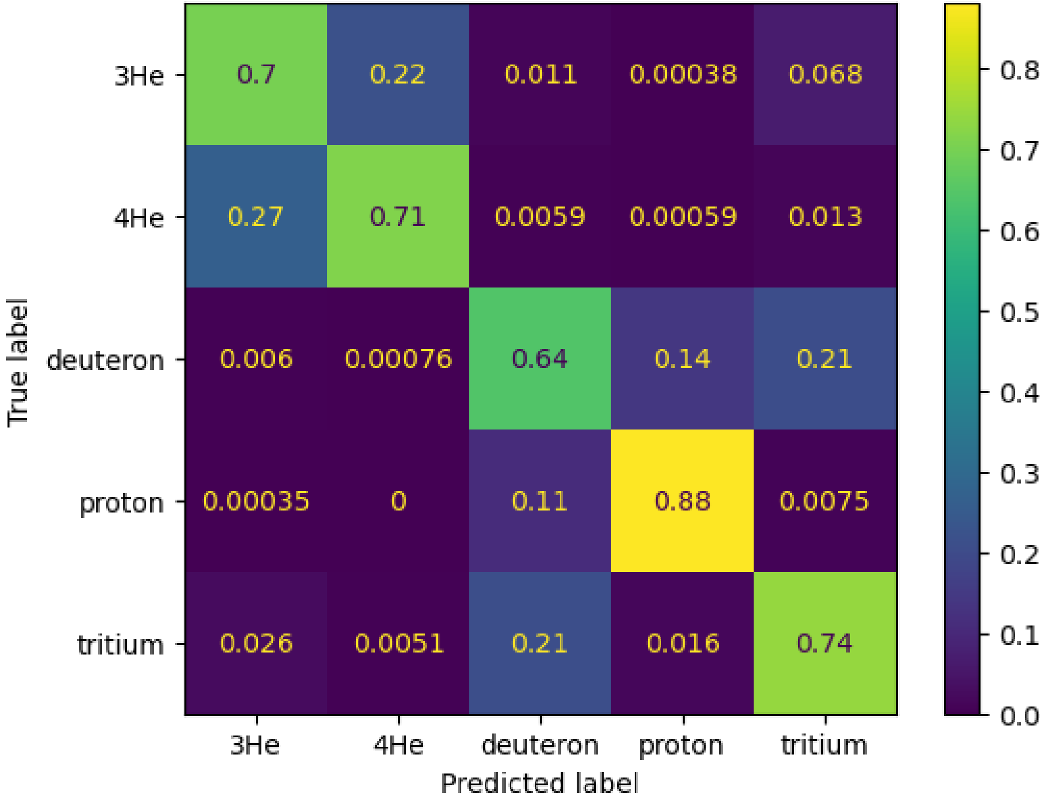
Using P, ϕ instead of p_x, p_y



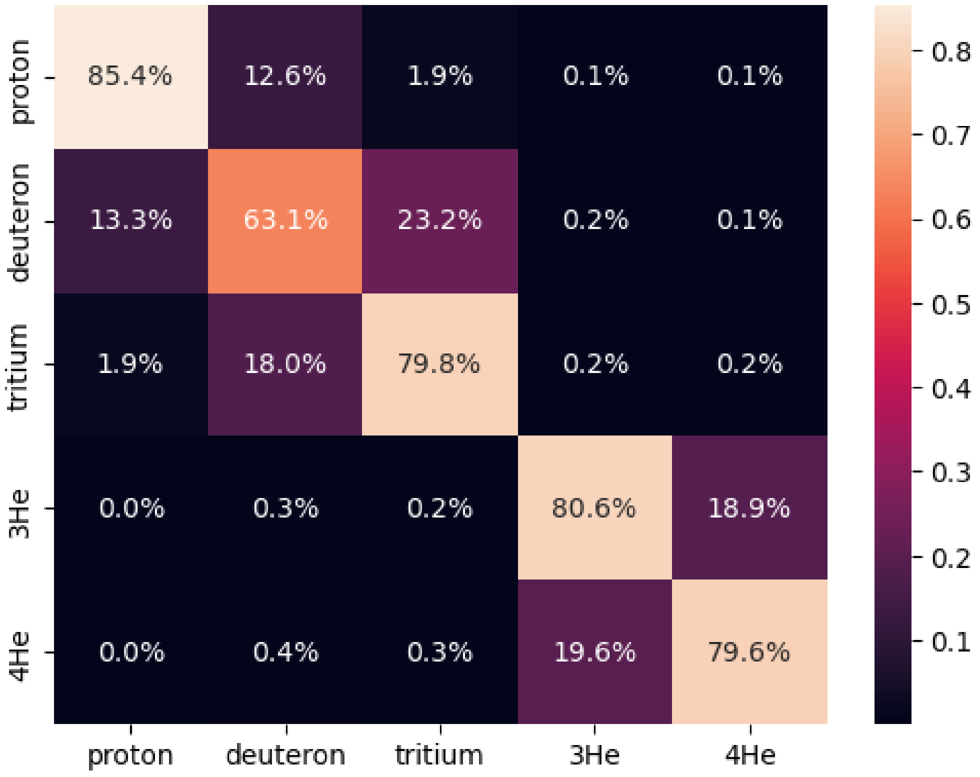
Post-KF PID: Different Approaches

- Starting from the MLP base, several other classification models were tested
- Gradient Boosted Trees (GBT) achieved slightly better accuracy than the MLP, particularly improving the weakest class separations
- The training times were significantly faster for GBT compared to MLP

MLP Classifier



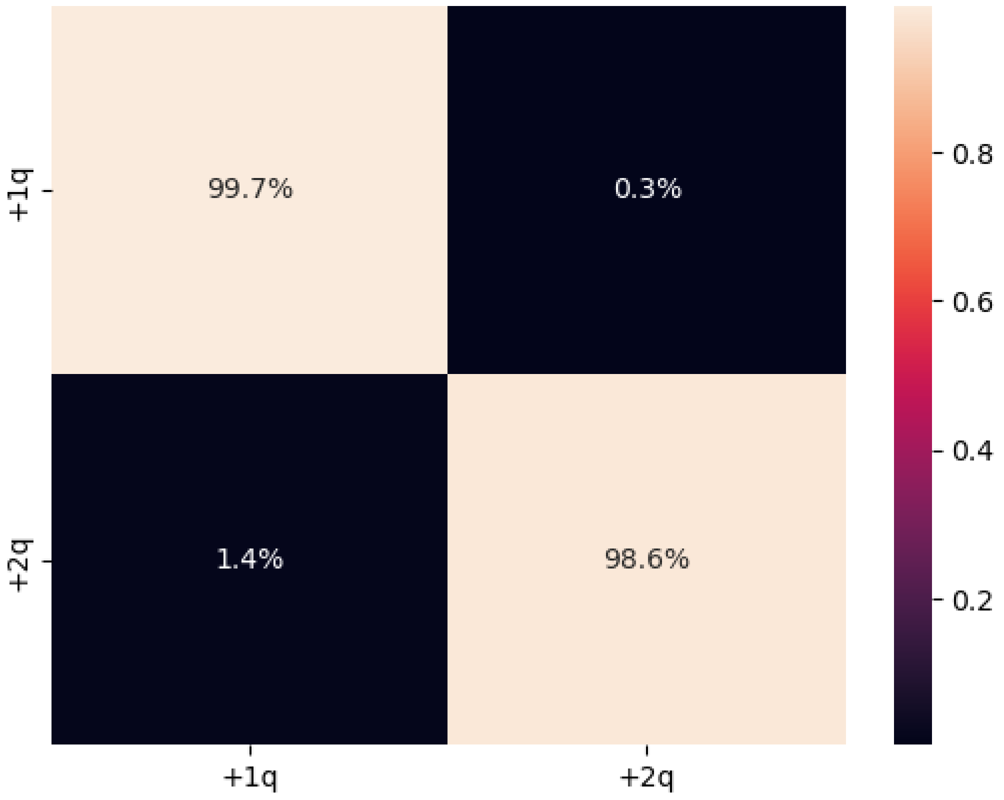
GBT Classifier



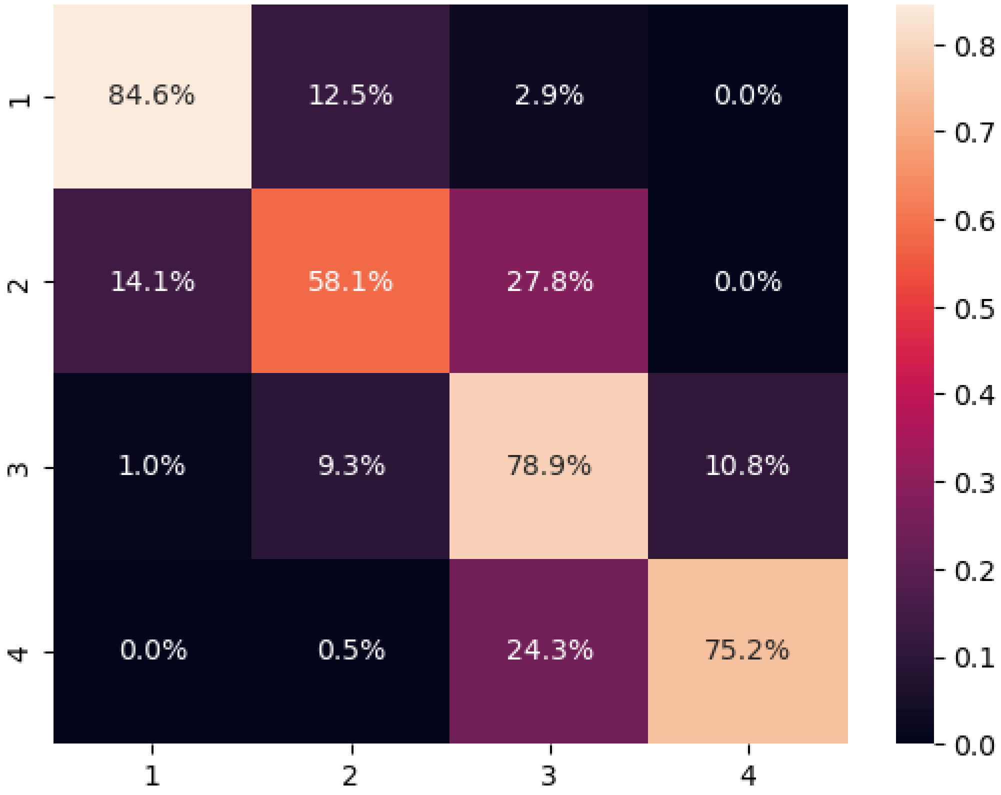
Post-KF PID: Different Approaches

- Based on a suggestion, an attempt was made to use two models to predict charge and mass separately
- Results in increased error and unrealistic predictions (`q=1 & m=4` or `q=2 & m=2`)
- A custom loss-function could make it work, but is not worth the effort

Predict charge only



Predict mass only



Integration to coatjava

PrePID:

- Fully implemented in coatjava (see pull request #1112)
- Inference is made per matched AHDC track – ATOF hit pair
- Output is written to `ALERT::ai:pid` bank:
  - `trackid` – matched AHDC track
  - `clusterid` – matched ATOF ~~cluster~~ hit
  - predicted class
  - class probabilities
- Designed to feed KF mass hypothesis

Post-KF:

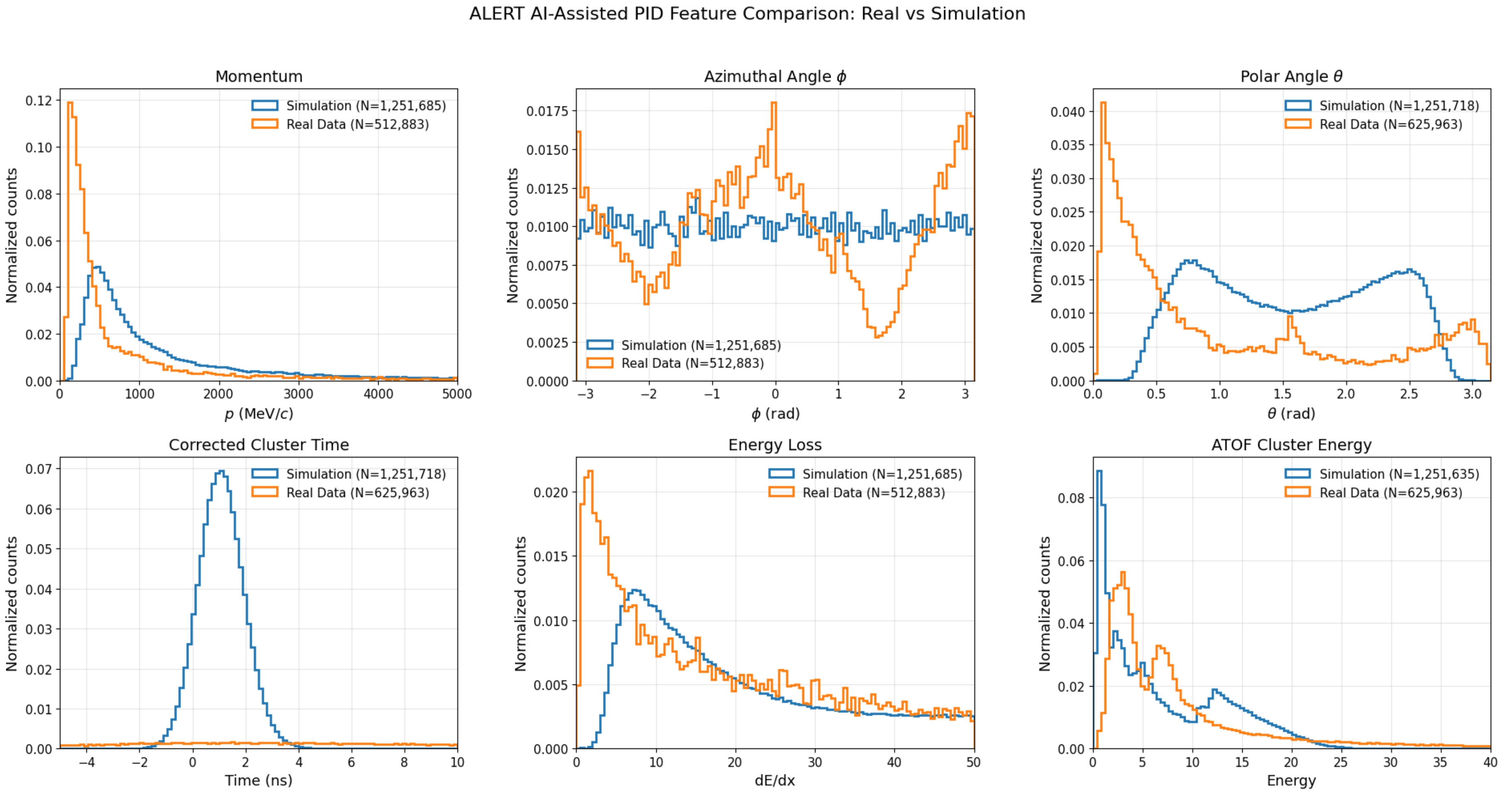
- Architecture prepared
- Awaiting KF update for integration

```
 position for [ALERT::ai:prepid]
* NODE * group =  23000, item =
        trackid :           1
      clusterid :           1
         prepid :          47
          p2212 :      0.0010
            p45 :      0.0290
            p46 :      0.2742
            p47 :      0.4855
            p49 :      0.2103


Choose (n=next,p=previous, q=qu
 position for [MC::Particle] ==
* NODE * group =     40, item =
            pid :          47
             px :     -0.8340
             py :     -0.8274
             pz :      0.3051
             vx :      0.0000
             vy :      0.0000
             vz :    -11.5744
             vt :    124.0000
```

Simulation vs Real Data

- Excellent performance in simulation
- Performance degrades on real data
- Likely causes:
  - Detector mismodeling
  - Resolution differences
  - Calibration effects



Ongoing and Planned Improvements

- **Different AI models:** Testing and comparing different types of classification models in addition to the currently used MLP and GBT models
- **Hyperparameter optimization:** Systematic tuning of model hyperparameters using Optuna to improve classification performance and training stability.
- **Feature engineering studies:** Continued exploration of physics-motivated feature transformations to improve model learning and separation power.
- **Domain adaptation:** Investigating strategies to improve model robustness when applied to real detector data, reducing simulation–data discrepancies.
- **Systematic stability tests:** Evaluating model performance across different kinematic regions and detector conditions to ensure stable behavior.
- **Extended inference capability:** Modifying the Pre-PID model to produce particle predictions even when no matched ATOF hit is present, allowing the model to assist track reconstruction in more cases.

Summary and Outlook

- Two-stage AI PID framework was developed to identify nuclear recoil fragments in ALERT
- PrePID model is currently implemented in coatjava and can be accessed by collaborators
- Post-KF PID demonstrates ~79% accuracy for simulated data
- Charge separation is nearly optimal (97%), but mass separation needs improvement
- Real-data performance is under active study
- Future improvements are in development

Thank You!