6/30/2026

CLAS12 ML PARTICLE IDENTIFICATION

STUDENT FLASH TALK

COOPER BELL
SIP Student

MARIA ŻUREK
Mentor

U.S. DEPARTMENT of ENERGY

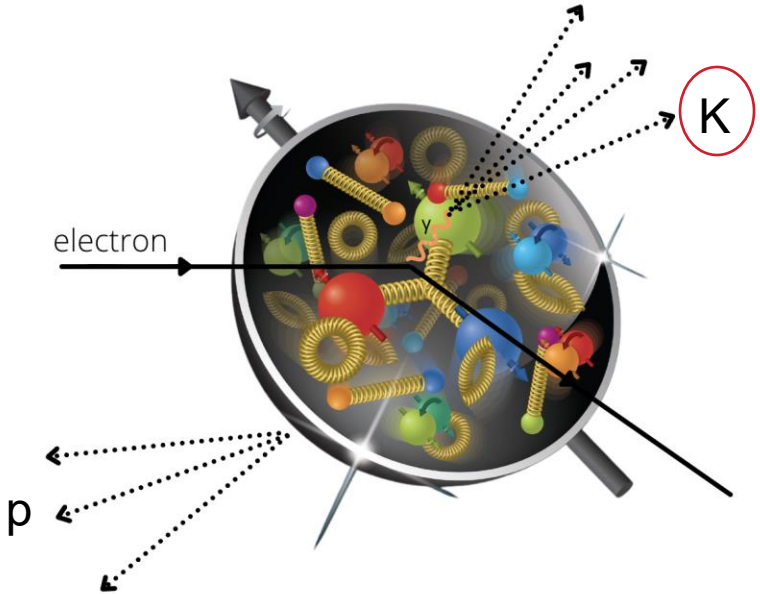
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

DUQUESNE UNIVERSITY

Argonne NATIONAL LABORATORY

PHYSICS OBJECTIVE AND ANALYSIS CHALLENGE

- **Process we want to study:** Semi-Inclusive Deep Inelastic scattering process with proton and kaon in the final state ep → epKX
  - Sensitive to fracture functions - conditional probability to find a quark inside a nucleon fragmenting into a final hadron.



Kaons give us access to the strange quark

Analysis challenge:
Increasing momentum → π/K signatures become more similar

- More pions misidentified as kaons
- Lower kaon sample purity

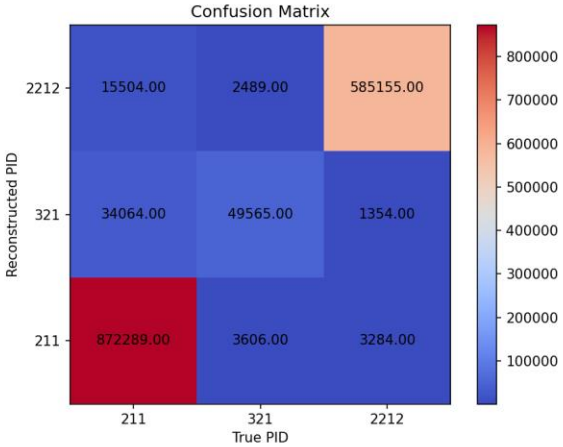
OBJECTIVE AND ANALYSIS CHALLENGE

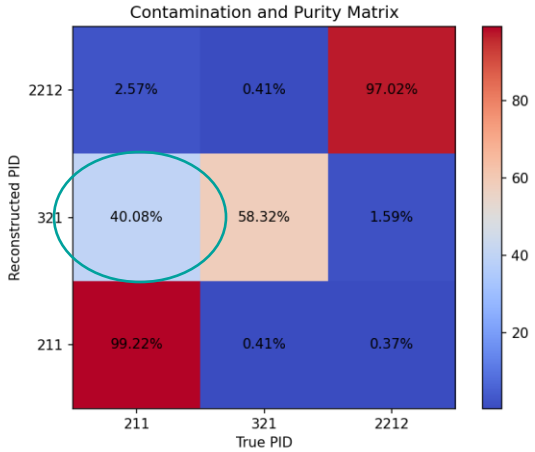
Goal: Produce a simple machine learning pipeline with an emphasis on K+ PID and compare with baseline performance.

- Compute Baseline purity, contamination, efficiency and Mis-ID
- Perform variable audit of Monte-Carlo versus experimental data to decide what to train the model on
- Train a binary classifier (gradient boosted decision tree) on large monte-carlo sample of EventBuilder reconstructed K
- Compare performance to initial baselines
- Compare performance of different classifiers and train multi-class classifier for higher momentum K
- Benchmark performance with data driven method (RICH and exclusive process studies)

QUANTIFY THE PROBLEM: CONTAMINATION

Contamination - fraction of non-kaons reconstructed as kaons.



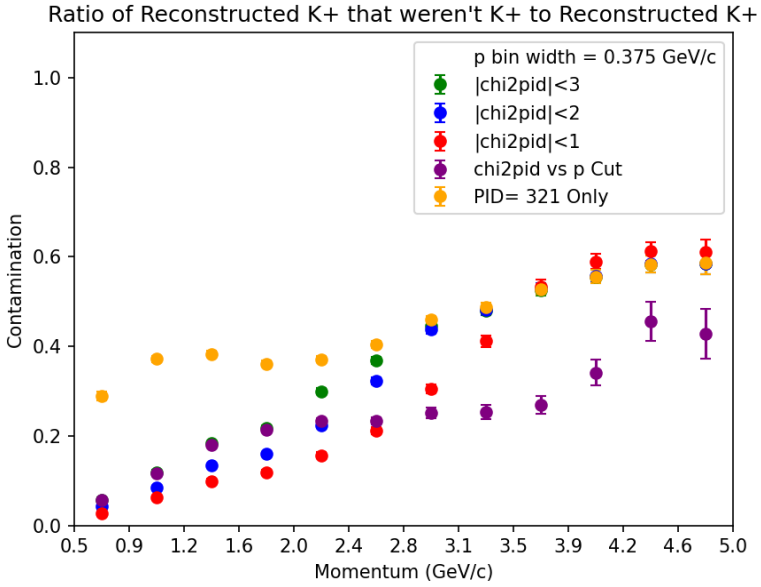


- MC Sample:
  - clasdis with for RGA fall 2018, Q2 > 2 GeV2, W > 2 GeV, y < 0.75, Mx(eKX)>1.6, Mx(eπX)>1.5, Mx(epX)>1.0 cuts on reconstructed level
  - Particle in Forward Detector only, with fiducial and vertex cuts applied
  - All reconstructed particles with matched true MC particles
- Large amounts of reconstructed kaons are true pions
- Although there are kaons reconstructed as pions, there are not nearly as many as the above.

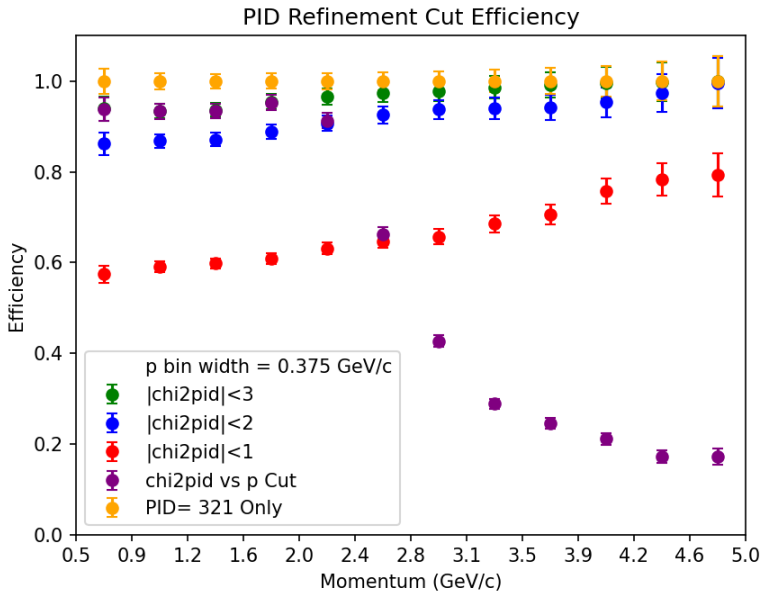
BASELINE PID PERFORMANCE: CHI2PID CUTS

$$\Delta t_i = t_0 - \left[t_{FTOF} - \frac{L}{\beta_i(p)} \right], \quad i = \pi/K/p/d/...$$

chi2pid is Signed-N_σ from nominal timing based on σ per FTOF-Paddle



ε_refine(cut) = N(true K passing chi2 cut) / N(true K, EB-K+, fiducial, vertex, phase space cut)

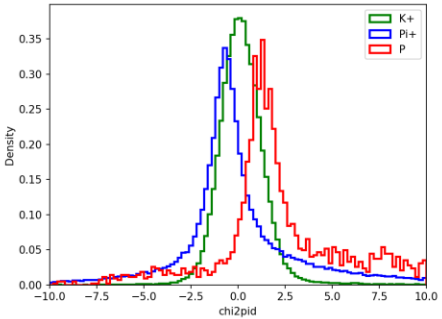


- Overall, the momentum-dependent cut developed for RGA data performs best, however with reduced efficiency at higher momentum.

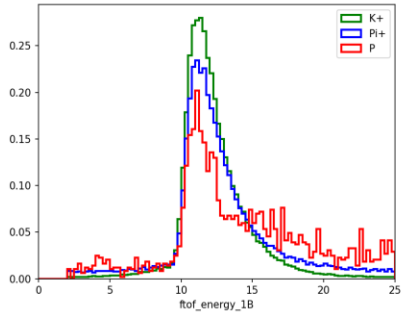
HOW TRAINING VARIABLES WERE CHOSEN

- For now, we focused on ftof, calorimetry, and htcc/ltcc variables available in DSTs
  - Energy, ToF and pathlength from FTOF 1A, 1B, 2; EC IN, OUT, PCAL and nphe HTCC and LTCC, chi2pid and beta
- **Series of statistical tests performed in (p, Θ)** bins to compare data and MC agreement
  - Wasserstien Norm, Population Stability Index, Max Local Residue, Kolmogorov-Smirnov and Chi2
- Variables chosen: chi2pid, beta, ecin_energy, ecin_path, ecin_time, ftof_energy_1B, ftof_path_1B, ftof_time_1B, ftof_energy_1B, ftof_path_1B

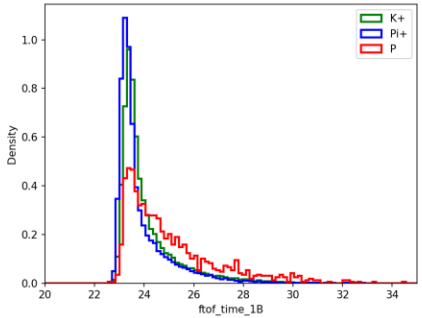
chi2pid



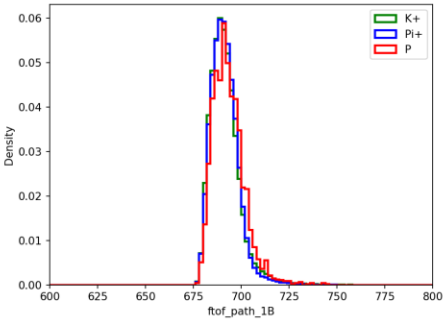
FTOF 1B energy



FTOF 1B tof



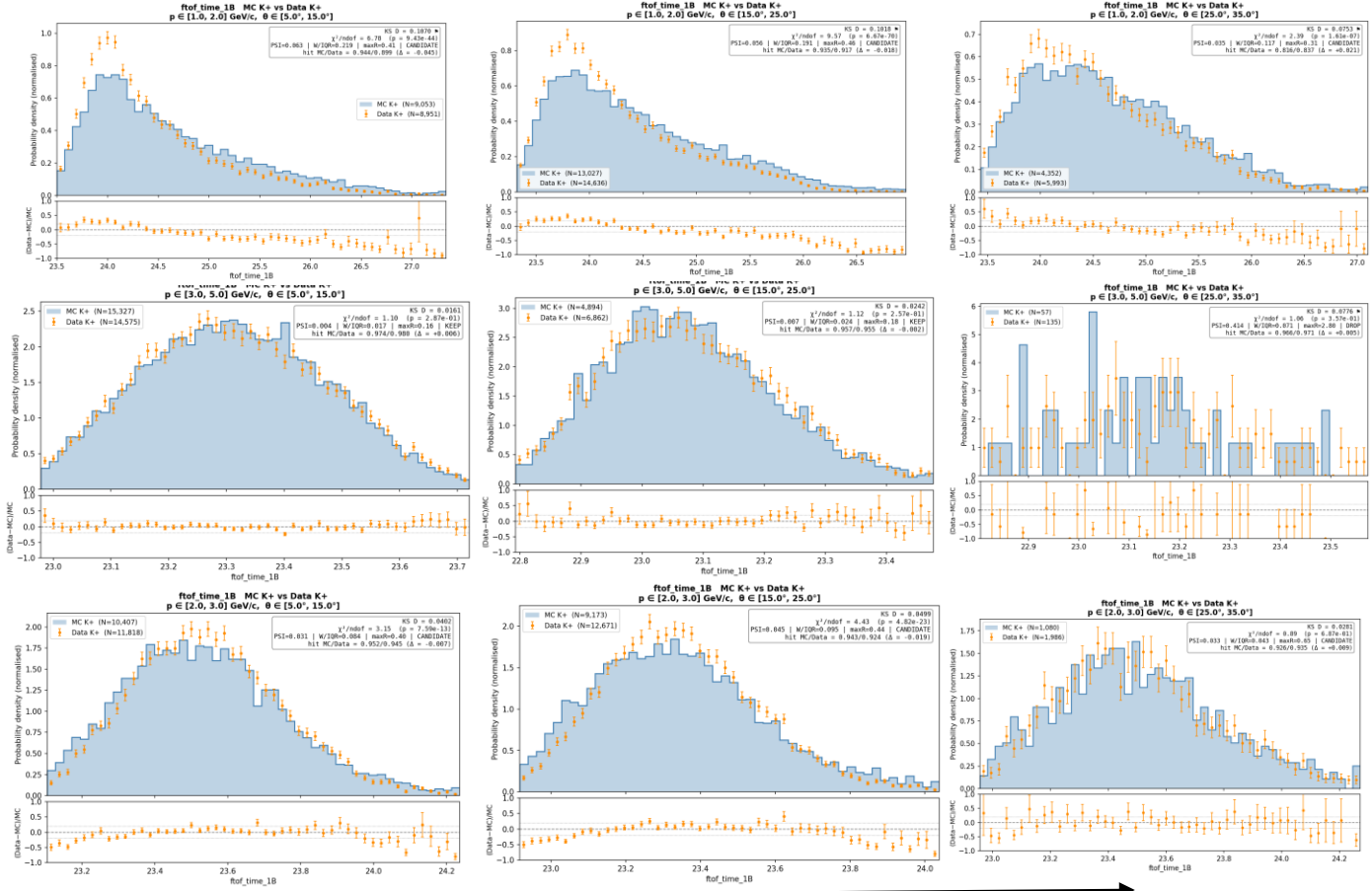
FTOF 1B path



Example contribution from true K, misidentified p and π+ in K+ reconstructed sample

VARIABLE AUDIT EXAMPLE – FTOF 1B TOF

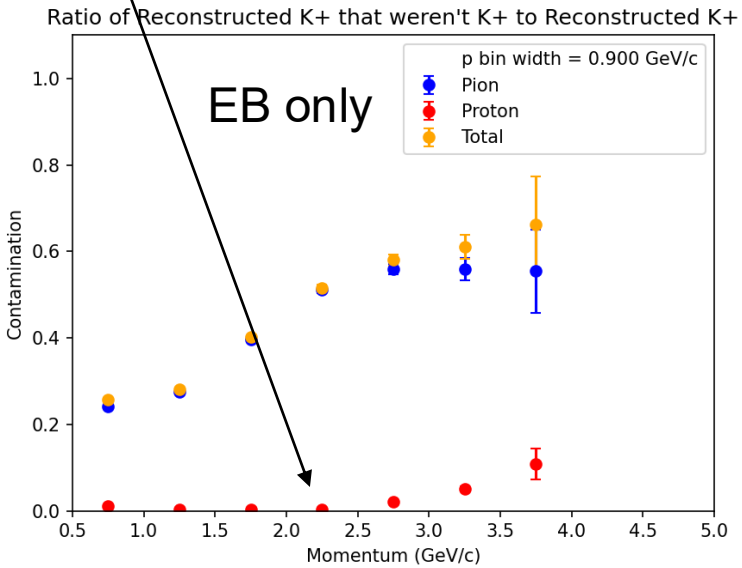
Increasing Momentum



Increasing Theta

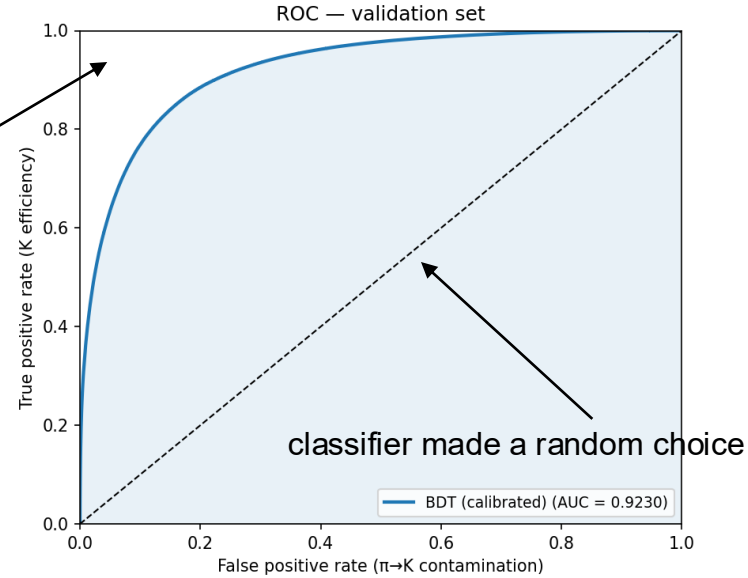
BDT PRELIMINARY RESULTS

- Binary classifier in momentum range of 0.5-3.2 GeV/c was chosen due to low contribution of protons based on EB only studies.

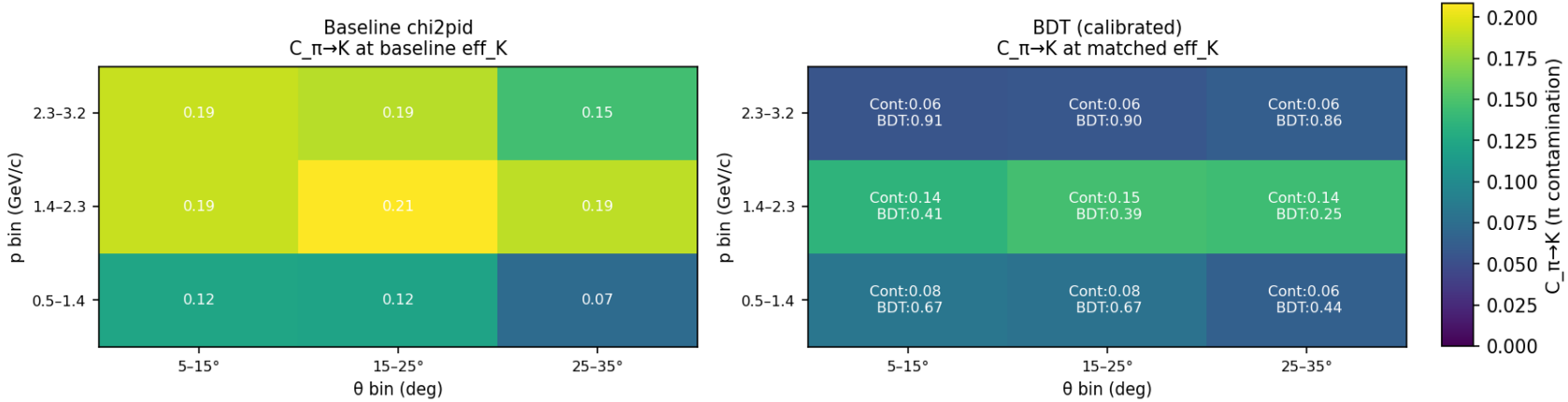


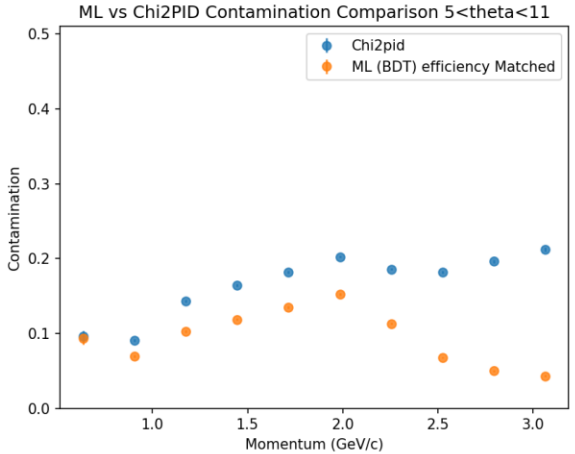
- LightGMB BDT, using 200 trees with a max depth of 6, learning rate 0.05
- Trained on 300 Monte-Carlo hipo files, using the detector variables found in DSTs
- After splitting files we trained on ~16M events

The closer to the left top corner, the better performanc e (AUC)



PRELIMINARY RESULTS





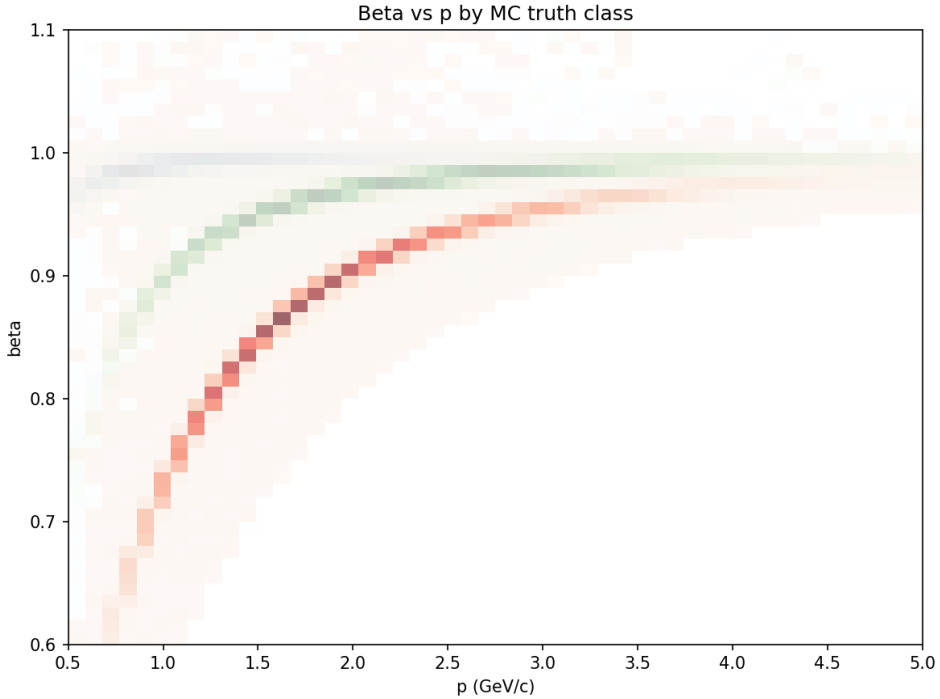
- Initial results show improvement in contaminations with the chi2pid p-dependent cut at the same (matched) efficiency.
- After 2 GeV the efficiency of the chi2pid cut drops, and the BDT performance at such low efficiency is much better than the baseline.

WHAT'S NEXT

- Threshold tunning for each bin to maximize the signal-to-noise ratio: $N_K / \sqrt{(N_K + N_\pi)}$
- Detector data will be passed through the optimized classifier
  - Cross check contamination studies possibly with $ep \rightarrow e\pi^+n$
- Tweak parameters to obtain better results/stronger models test other ML methods such as Neural Networks and move to the π/K/p classification at higher momentum
- Package and polish complete pipeline with instructions on how to use

COMMENTS/QUESTIONS

BETA VS P (TRUE ID CUT)

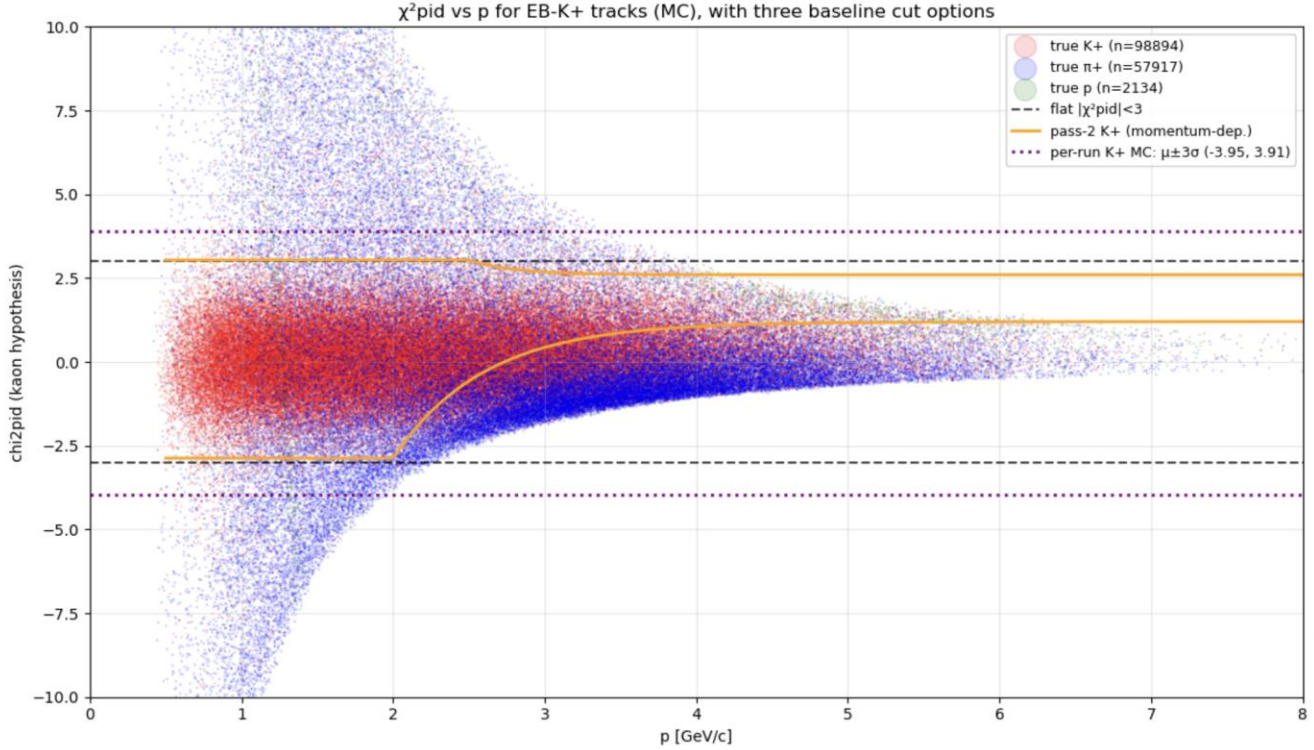


Blue -> Pion
Green -> Kaon

Red_>Proton

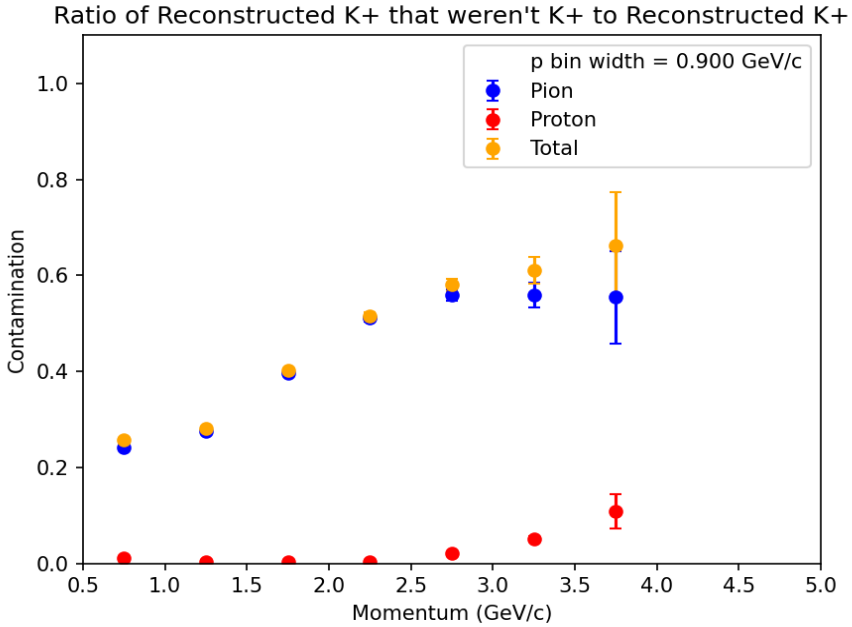
Pions and Kaons become indistinguishable at high momentum, making chi2pid methods produce lesser purity.

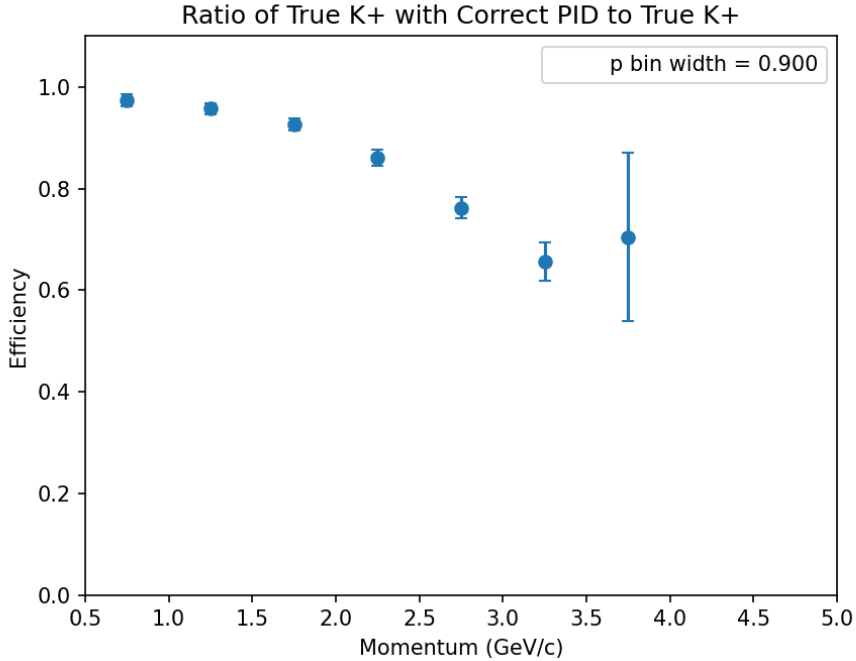
CHI2PID MOMENTUM BASED CUT



BASELINE PID PERFORMANCE WITH EB

- Main contamination comes from pions (20% - 50%)
- Proton contamination negligible for <3.2 GeV/c
  - This range was chosen **for the binary classifier studies**

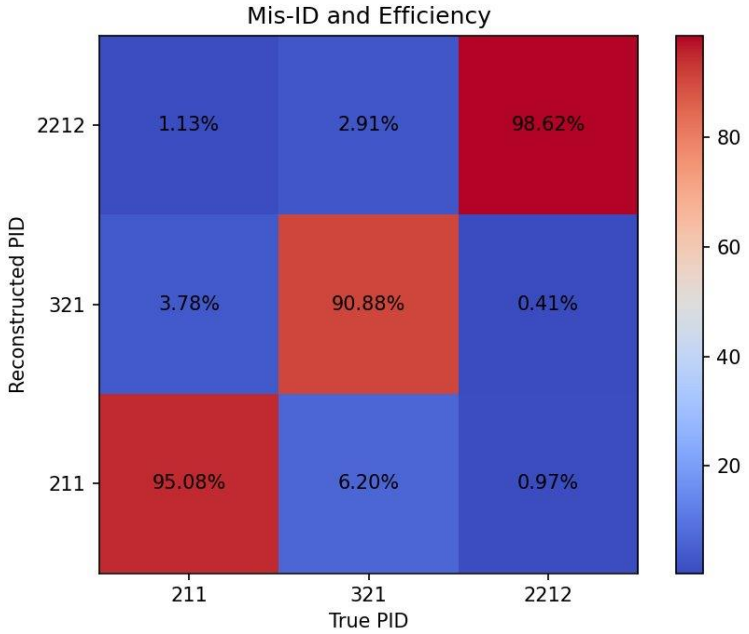




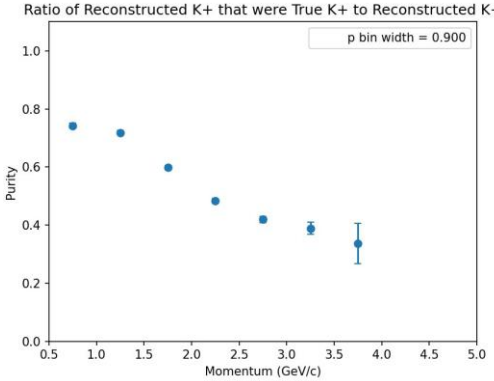
RAW EFFICIENCY /MIS ID MATRIX

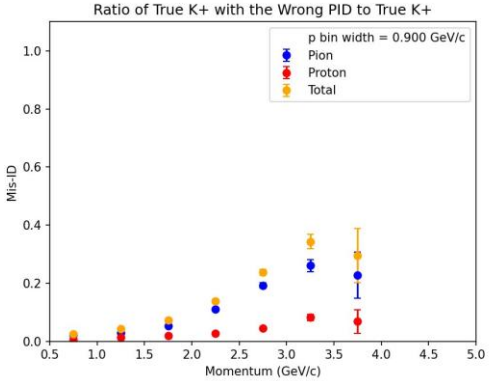
Uses PID cuts and base kinematic cuts.

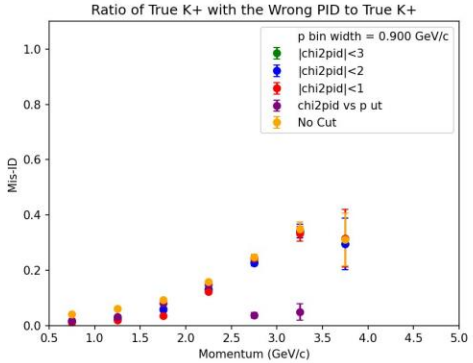
With no Strict filtering, K+ efficiencies are fine, but pion contamination was high

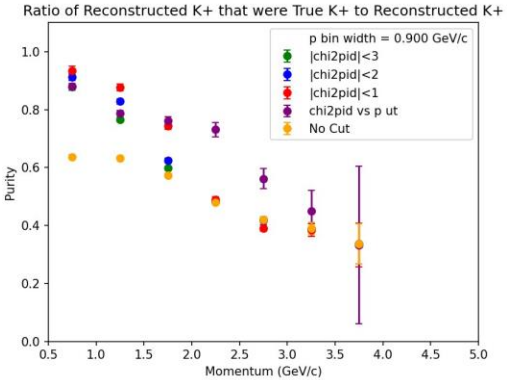


RAW CONTAMINATION, EFFICIENCY, POST CHI2PID





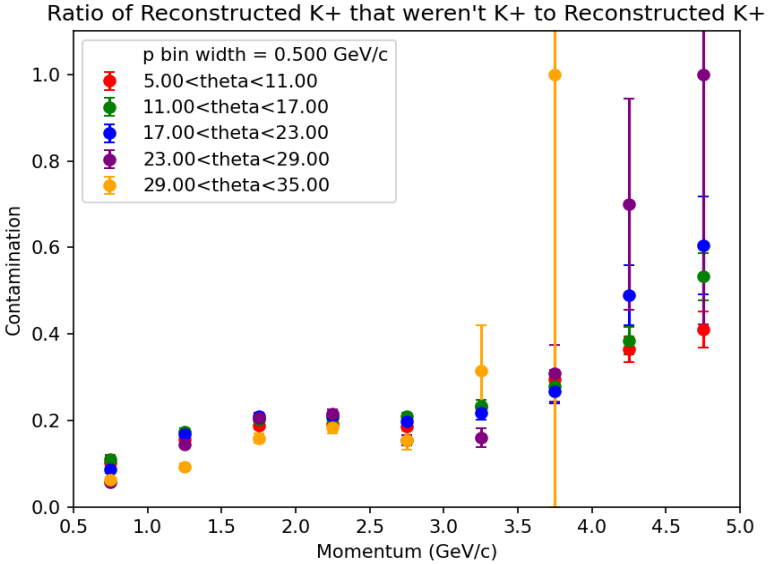




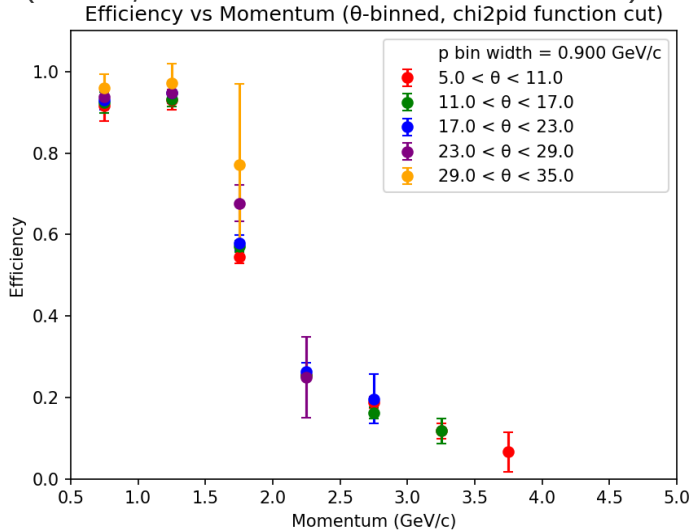
PID PERFORMANCE IN THETA BINS

Small theta-dependence observed

These plots use the Momentum-dependent chi2pid cut



ε_refine(cut) = N(true K passing chi2 cut) / N(true K, EB-K+ + fiducial + vertex cut)



Theta bins performed similarly under the momentum based chi2pid cut, apart from poor statistics in higher theta ranges.

STATISTICAL TESTS LIST

- Wasserstien Norm (Earth Mover's Distance) measures the minimum "cost" of transforming one probability distribution into another, where the cost is the amount of probability mass moved multiplied by the distance it is moved.

$$W(P,Q) = \int_{-\infty}^{\infty} |F_P(x) - F_Q(x)| \, dx$$

  - where FP(x) and FQ(x) are the cumulative distribution functions of the two distributions.
- Population Stability Index (PSI) -> A binned statistical measure of the difference between two probability distributions, computed by comparing the fraction of events in each bin.

$$\mathrm{PSI} = \sum_{i=1}^{N} (P_i - Q_i) \ln\left(\frac{P_i}{Q_i}\right)$$

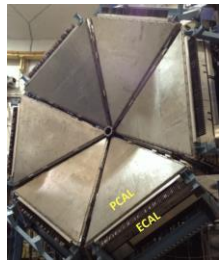
  - Pi = fraction of events in bin ii of the reference distribution,
  - Qi = fraction of events in bin ii of the comparison distribution.

The PSI is zero when the two distributions are identical and increases as they become more different.

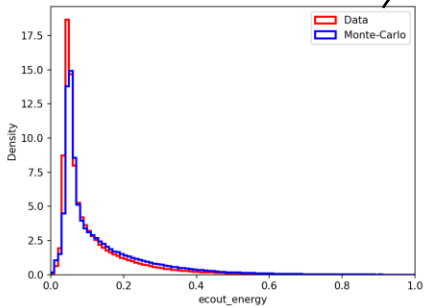
- Max Local Residue -> Computes the relative discrepancy between datasets with same binning as PSI. Returns the maximum value that occurs.
- Chi2 -> Tells whether two histograms are consistent from being drawn from a distribution. Formally : $\chi^2 = \Sigma\_i \ (d\_i - m\_i)^2 / (\sigma\_d\_i^2 + \sigma\_m\_i^2)$
- Kolmogorov-Smirnov -> Finds the maximum absolute difference between two CDFs, formally: D = sup_x | CDF_data(x) - CDF_MC(x) |

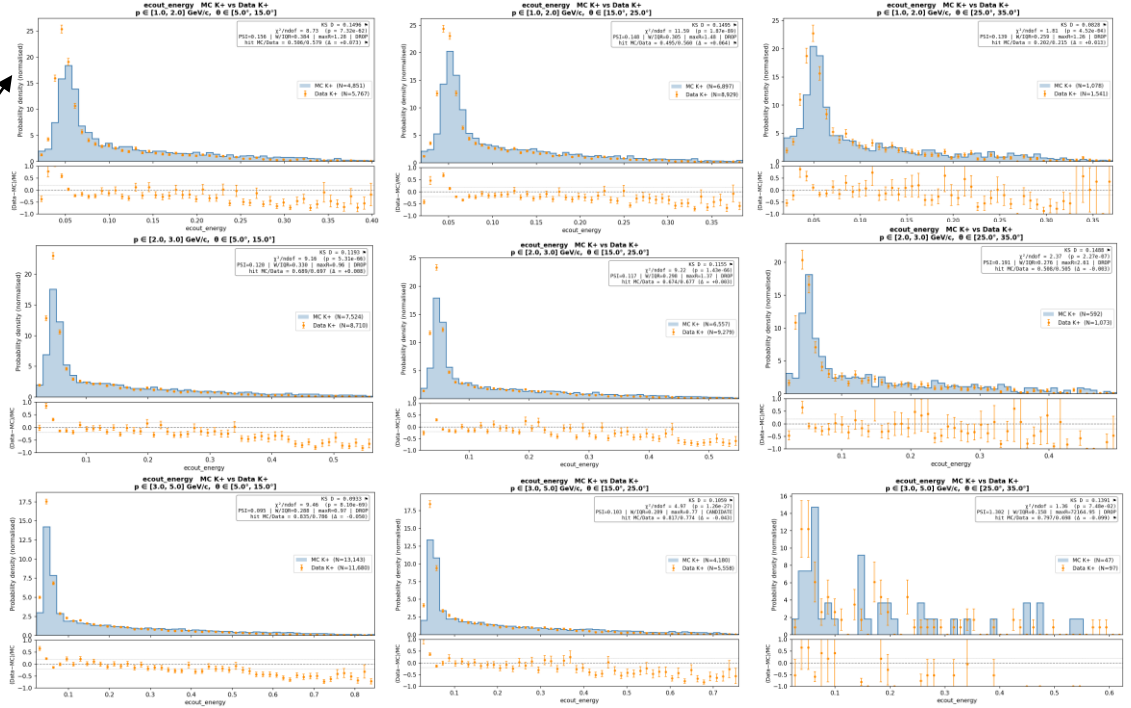
VARIABLE AUDITS EXAMPLE

- Example of a Tier 3 variable, ecout_energy, Metrics performed poorly:



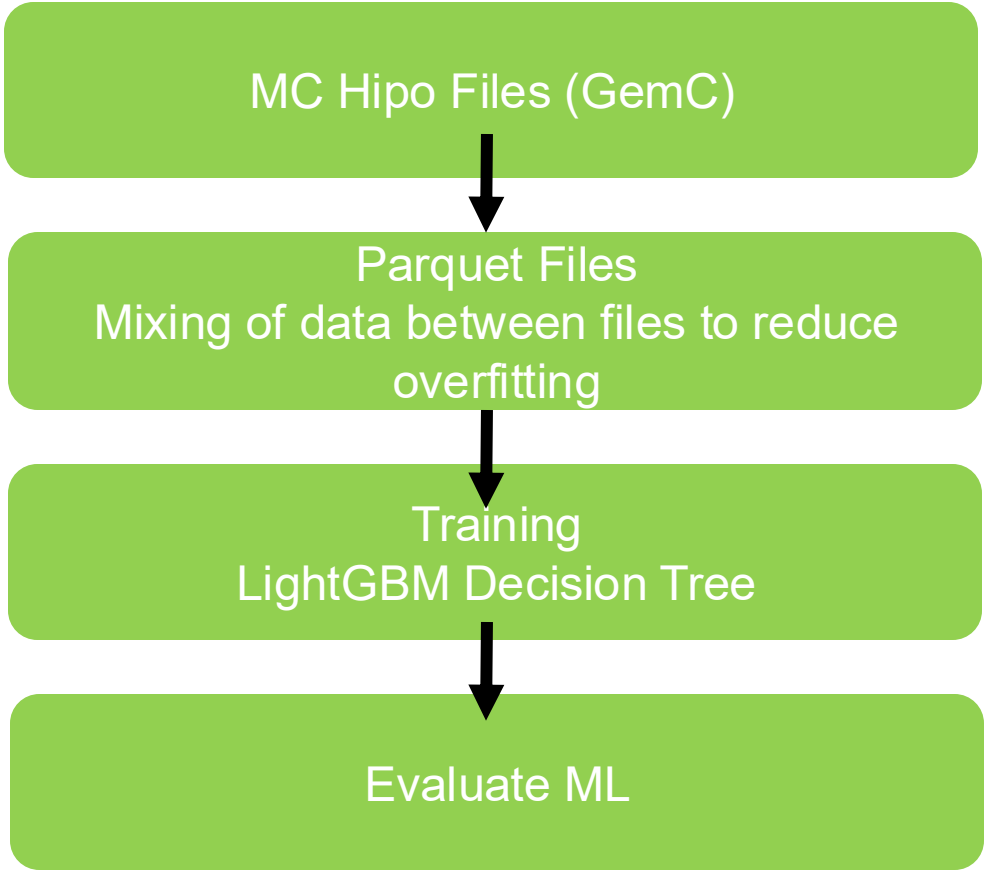
Integrated





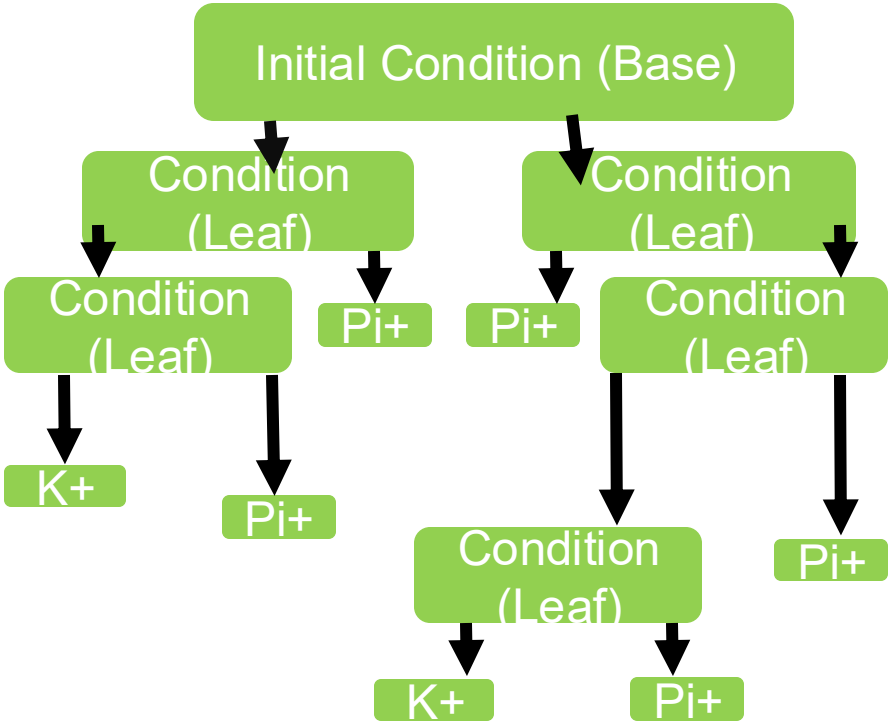
PIPELINE SUMMARY

- The pipeline for producing the ML is as follows.
- Start with the raw MC hipo files (GemC)
- This data is then parsed into three parquet files for training and validation
- The ML is then trained using LightGBM from the parquet files
- A script is then run to compare the identification with the baseline



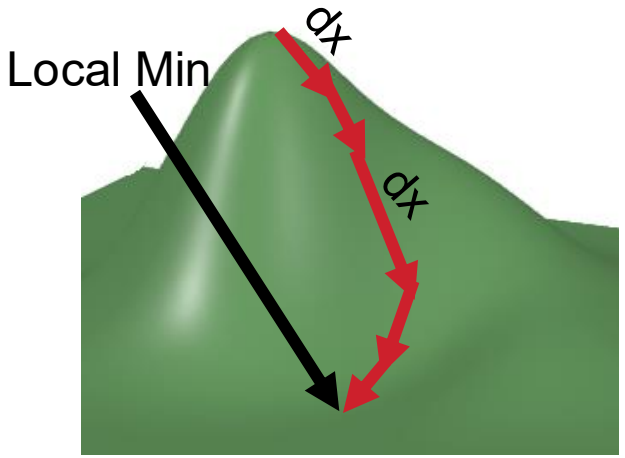
GRADIANT BOOSTED DECISION TREE: THE TREE

- A Decision tree is a structure of if then statements that terminate in a classification. In ML the parameters of each if-then are tuned based on training it off of data
- Several of these trees are run in parrallel, which helps fight overfitting, compared to having one deep tree.
- Decision Trees, after training, will run much faster than Neural Networks, making them a prime candidate for particle identification.



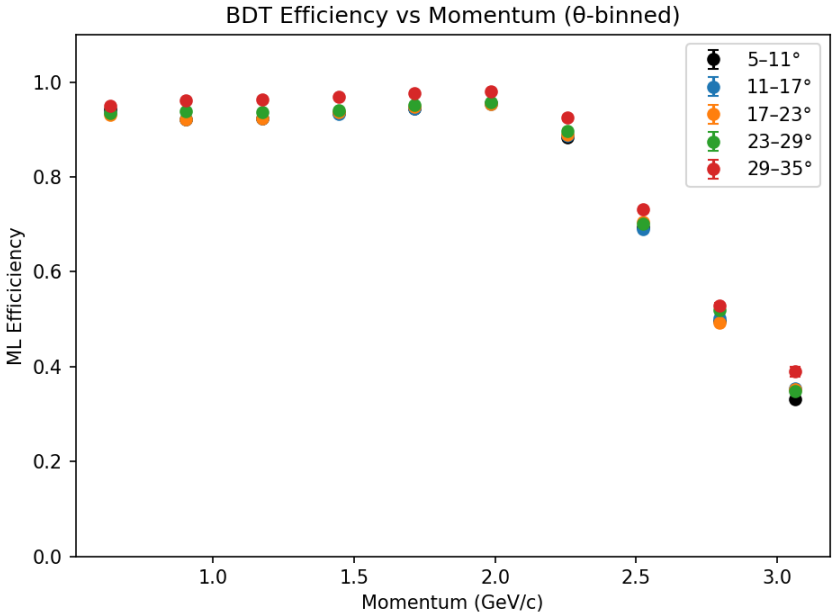
GRADIANT BOOSTED DECISION TREE: GRADIENT BOOSTING

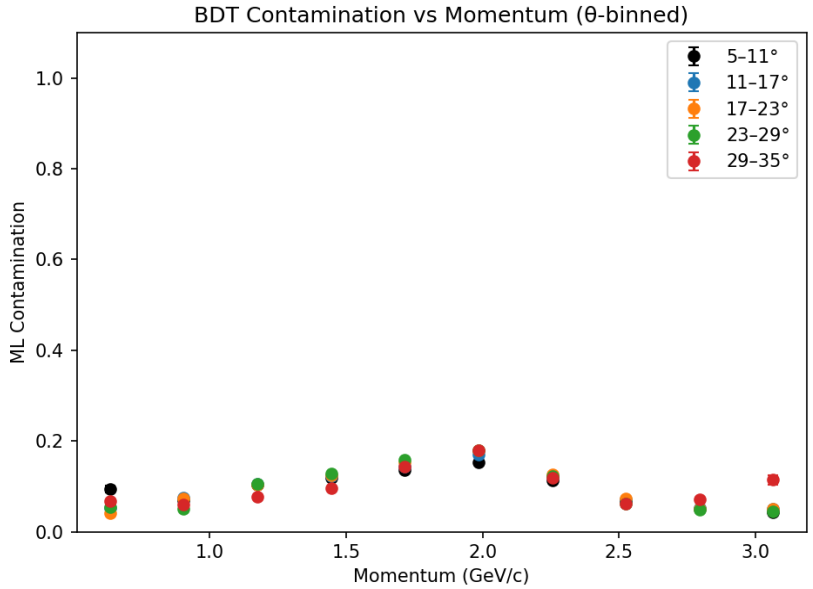
- Gradient Boosting is the primary means in which machines learn. For any ML there is a "cost" function that is defined, which measures how inaccurate the ML is. As a result, minimizing the cost function gives results that closely match the truth (It learned!).
- This minima is found in a high dimensional space, To find how to step the direction can be found by the negative gradient of the Cost function. This is then used to make a tiny step. Iterating this process, gradient descent, will aproach a local minimum.



$$dx_\mu = -d\lambda \frac{\partial}{\partial x^\mu} Cost(x^\mu)$$

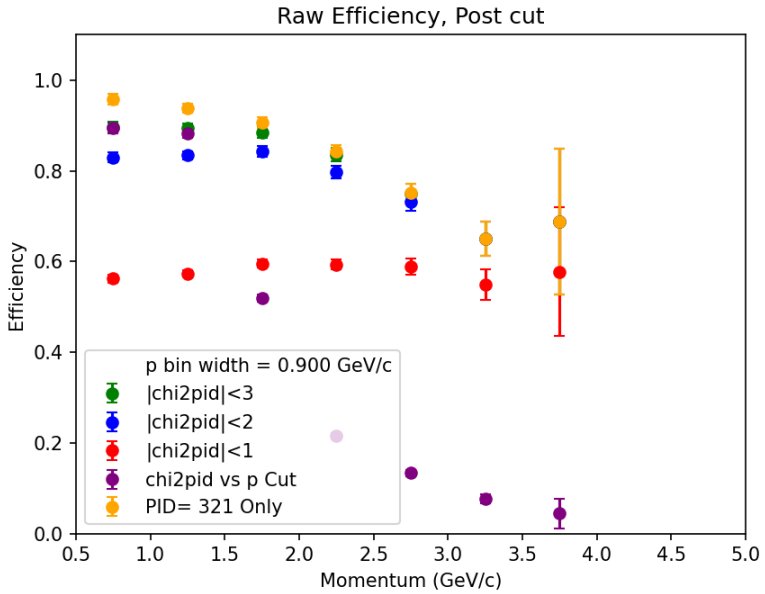
PRELIMINARY RESULTS CONT.





EB PID + CHI2PID CUT EFFICIENCY PLOT

ε_refine(cut) = N(true K passing chi2 cut + EB-K+) / N(true K, fiducial, vertex, phase space cut)



VARIABLE AUDIT TERMINOLOGY

- Teir 1: Variables that passed the audit and can be used without issue
- Teir 2: Variables that failed the automatic audit, but passed upon manual review
- Teir 3: Variables that failed the audit, but still matched within certain ranges, or had the right general trend.

VARIABLE AUDIT RESULTS

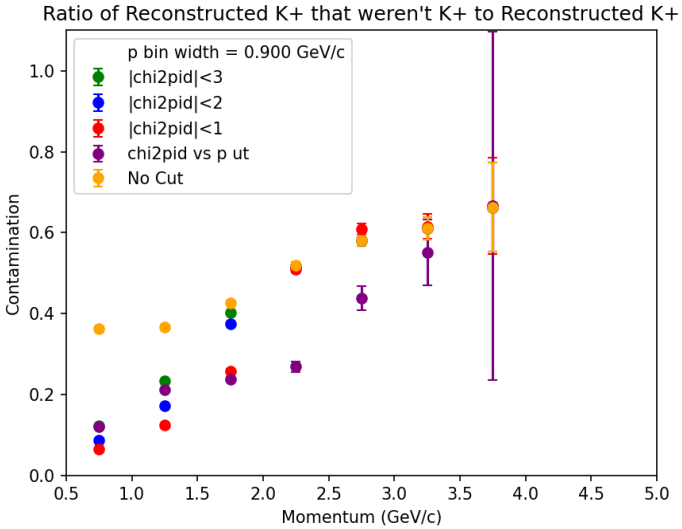
- Tier 1: ftof_energy_1B, ftof_time_1B, ftot_path_1A, ftof_path_1B, ecin_path, ecout_path, pcal_path
- Tier 2: ftof_energy_1A, ecin_time, nphe_htcc, nphe_ltcc, ecin_energy
- Tier 3: ftof_time_1A, pcal_time, pcal_energy, ecout_energy, ecout_time

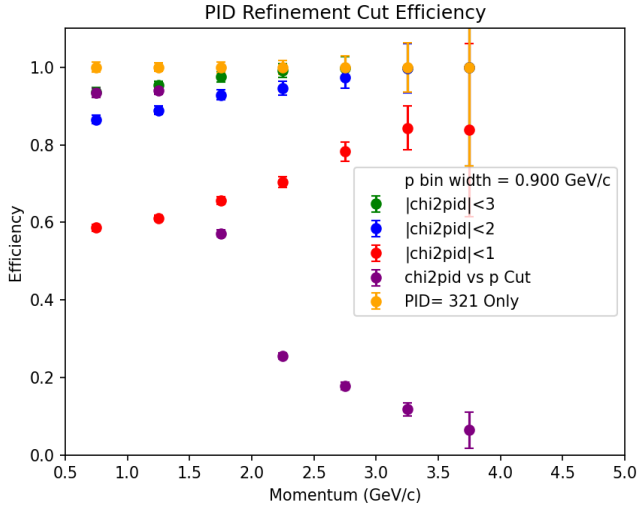
- Note: Kinematic Variables, Q2, Mx, theta, ect are not used in training.

BASELINE PID PERFORMANCE: CHI2PID CUTS

$$\Delta t_i = t_0 - \left[t_{FTOF} - \frac{L}{\beta_i(p)} \right], \quad i = \pi/K/p/d/...$$

ε_refine(cut) = N(true K passing chi2 cut) / N(true K, EB-K+, fiducial, vertex, phase space cut)





- Overall, the momentum-dependent cut developed for RGA data performs best, however with reduced efficiency at higher momentum.