Updates on 1 GeV/c proton-argon inelastic cross-section analysis

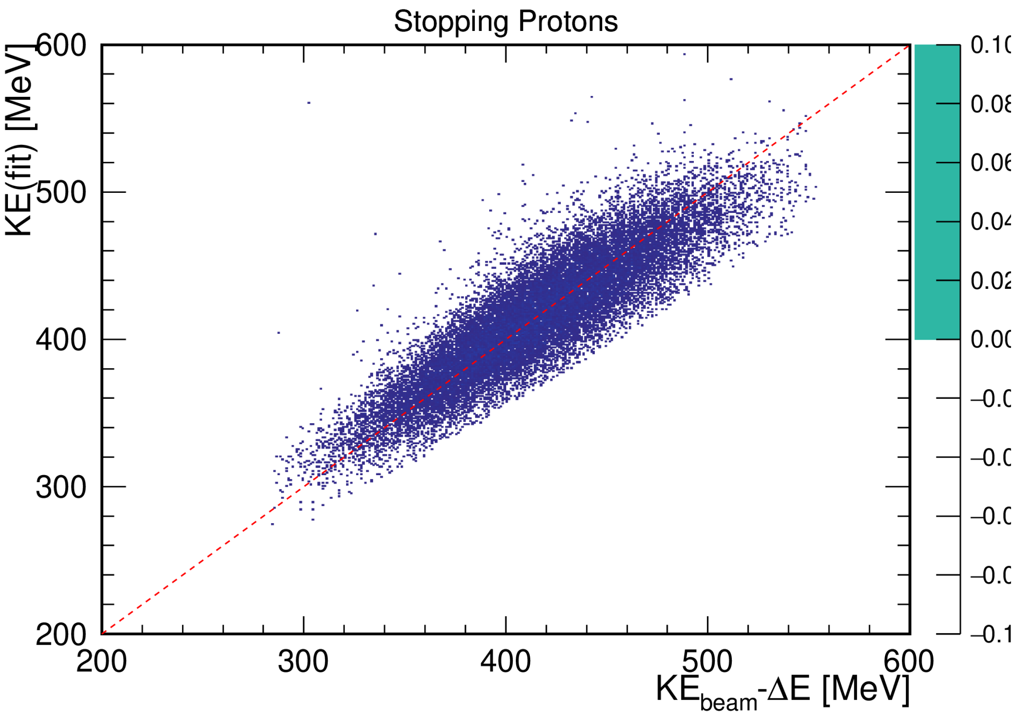
- KE Systematics update
- Improvement on signal event selection

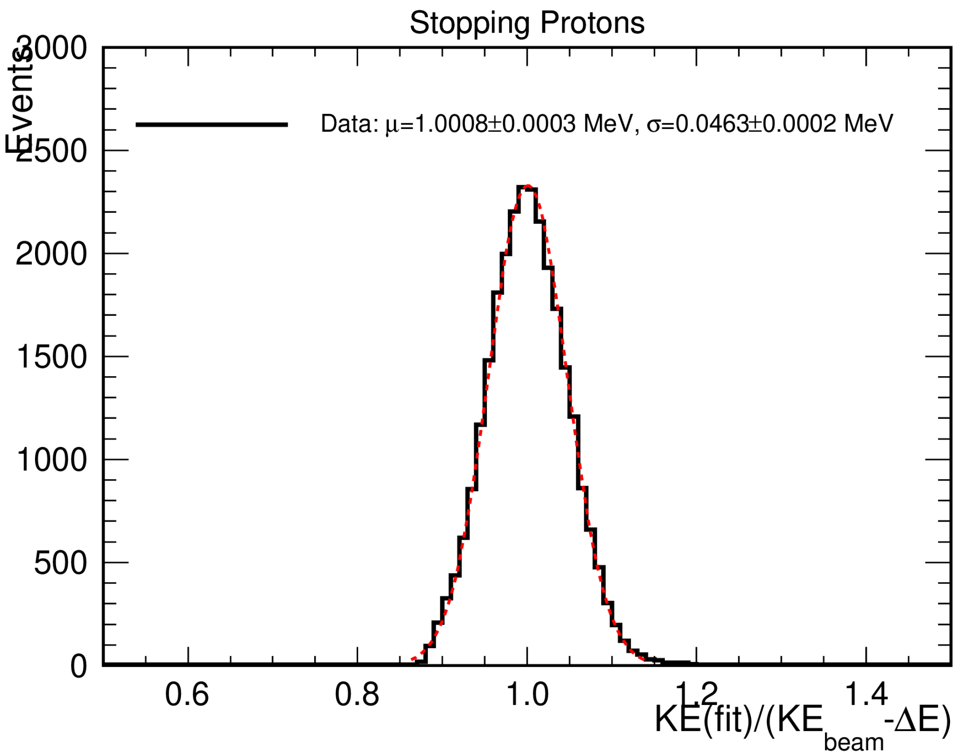
Heng-Ye Liao

Hadron analysis meeting

Sep 08, 2022

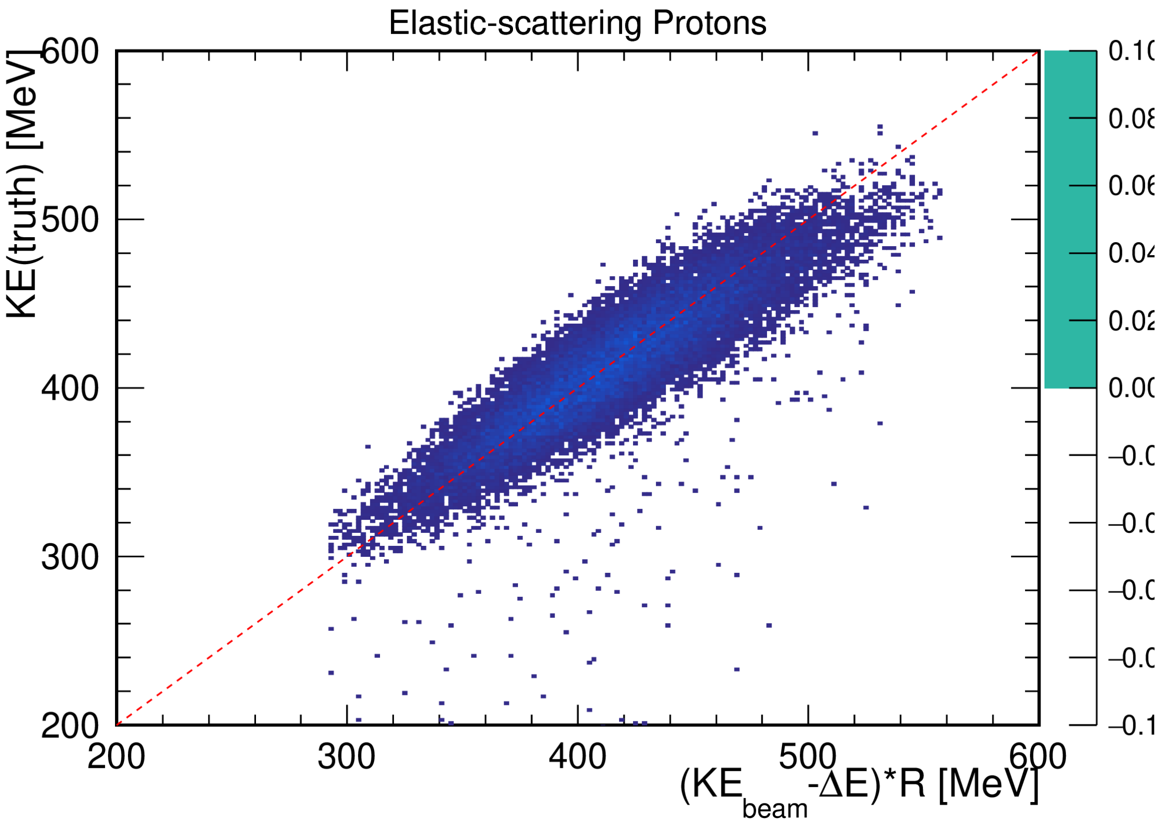
KE at TPC FF

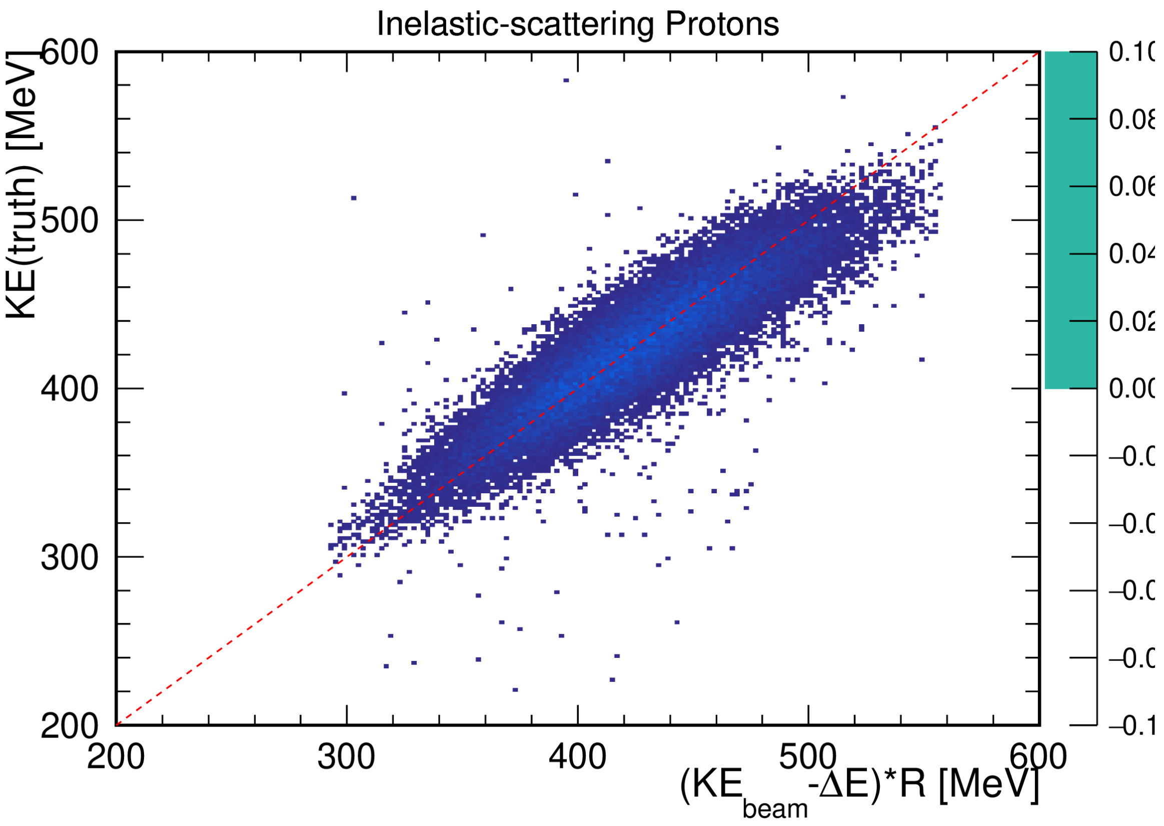




- Use stopping proton as standard candle to calibrate KE_{FF}
- Event-by-event correction at TPC FF
- Ratio between KE(fit) and KE_{beam}-ΔE around one showing that good energy reco.
  - ΔE is derived using the scanning method with KE(fit) on stopping protons

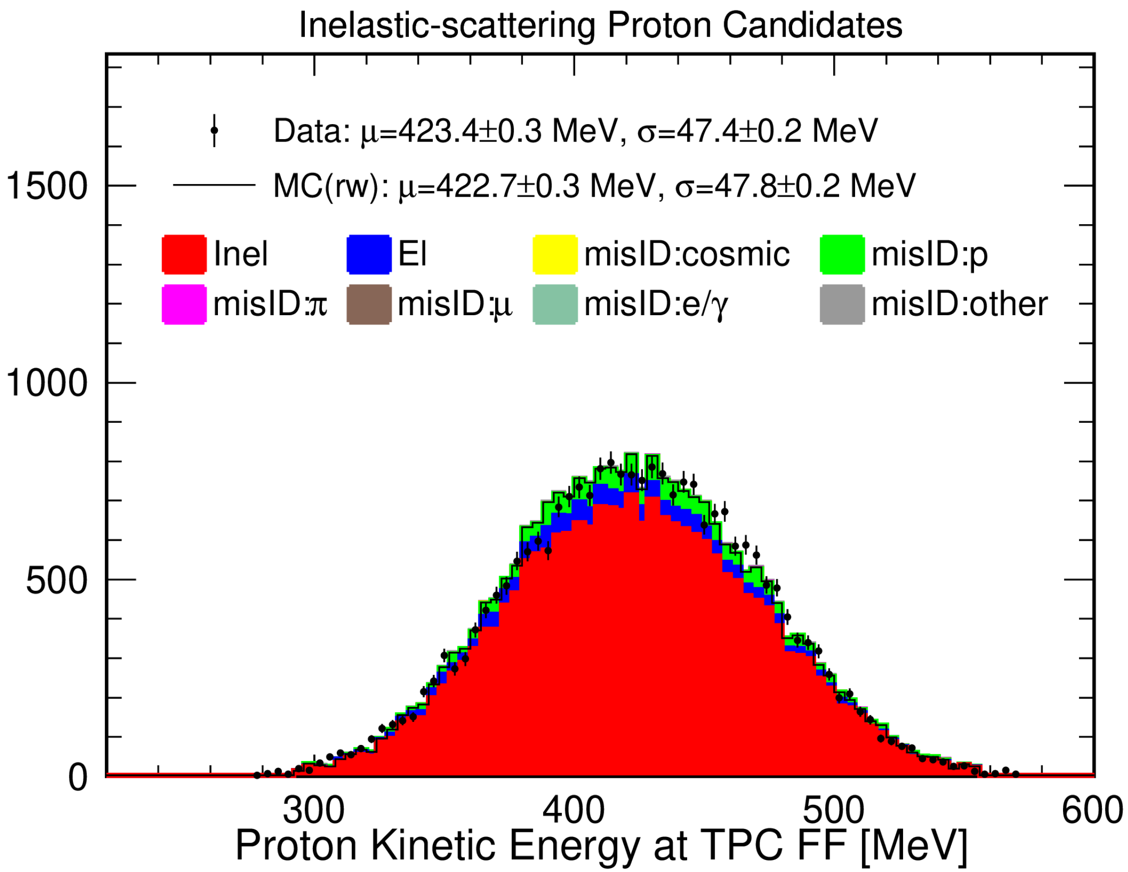
Reconstructed KE_{FF} after Ratio Correction

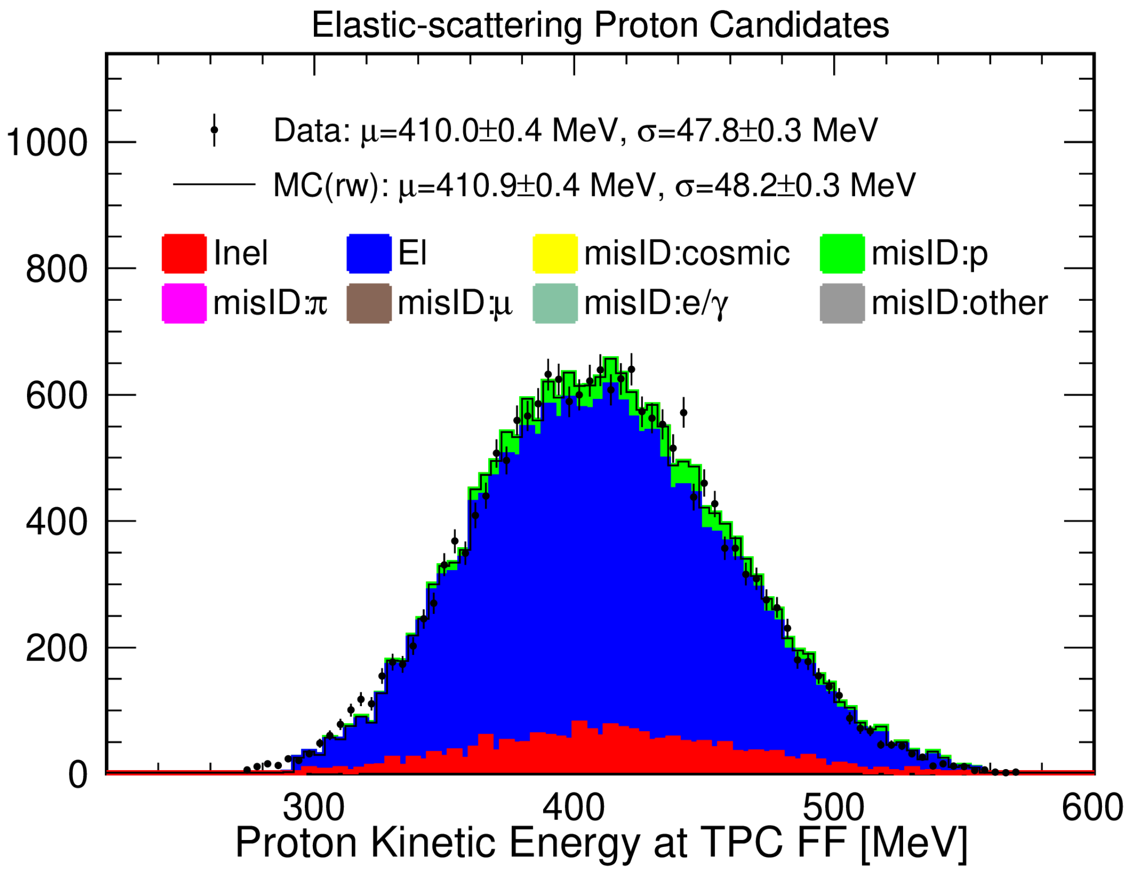




- Good KE_{FF}(reco) for both data and MC

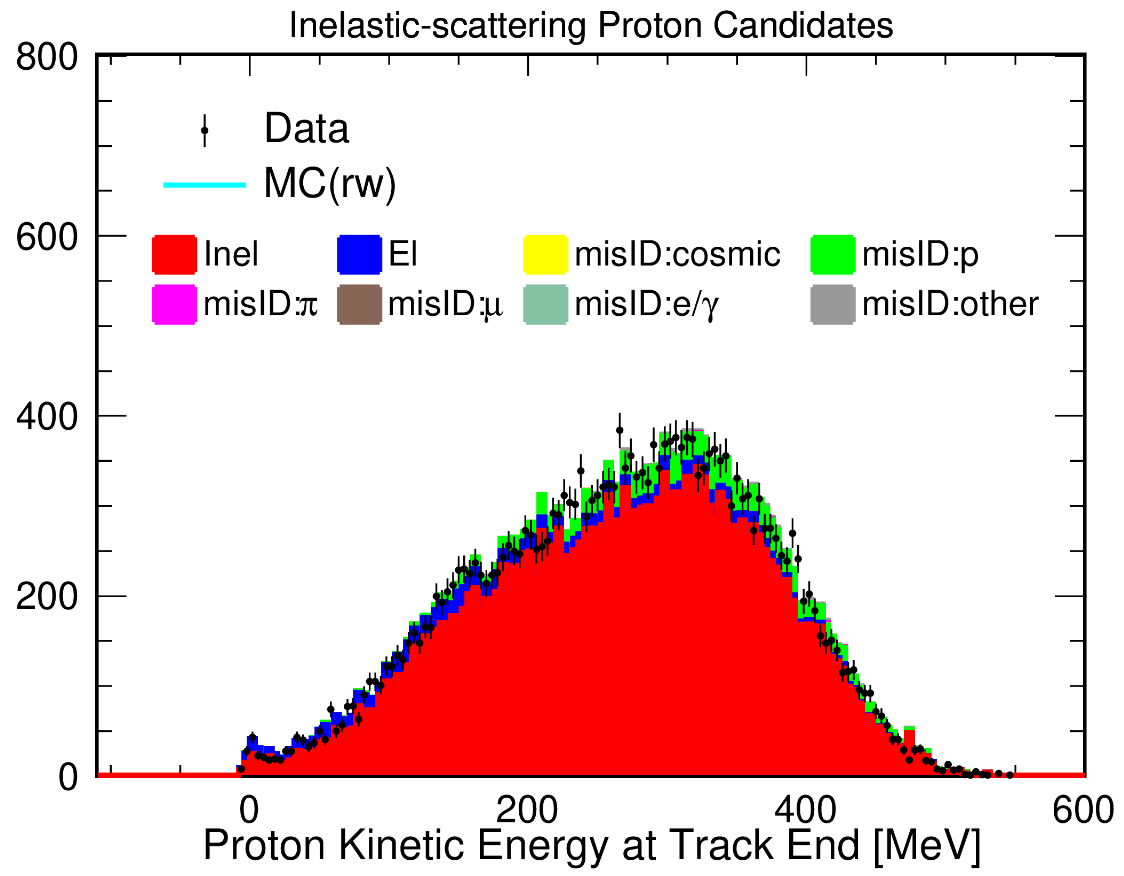
KE_{ff} with Const E-loss Assumption

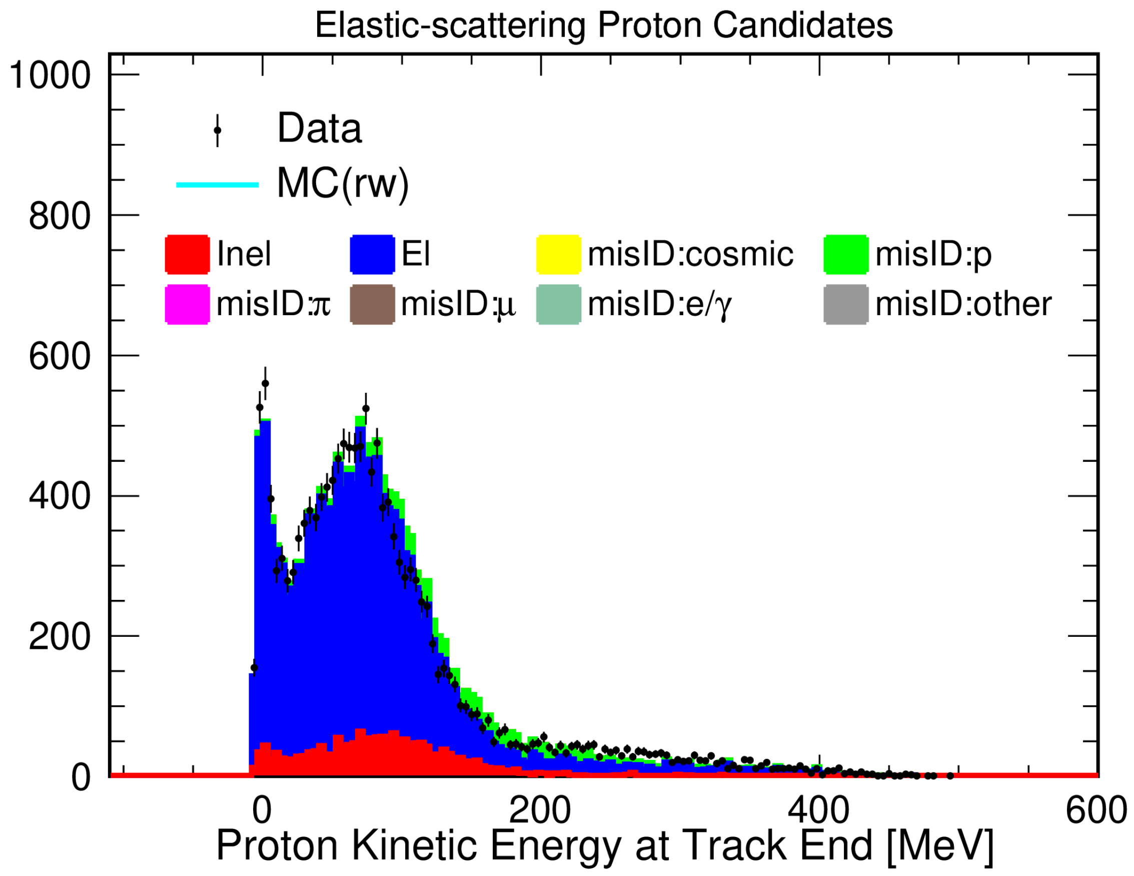




- $KE_{ff}=(KE_{beam}-\Delta E)*R$, R~1
- Good reconstruction at KE_{ff} for both data and MC

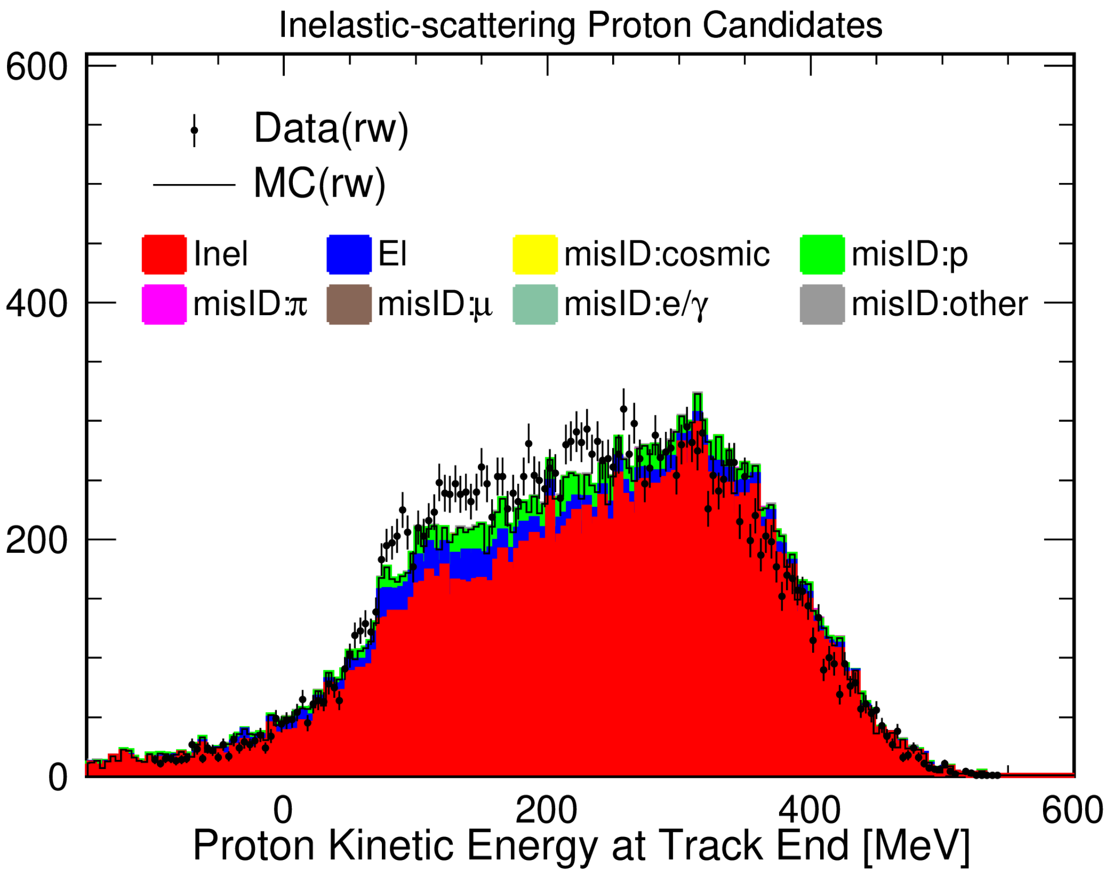
KE at Track End (Bethe-Bloch)

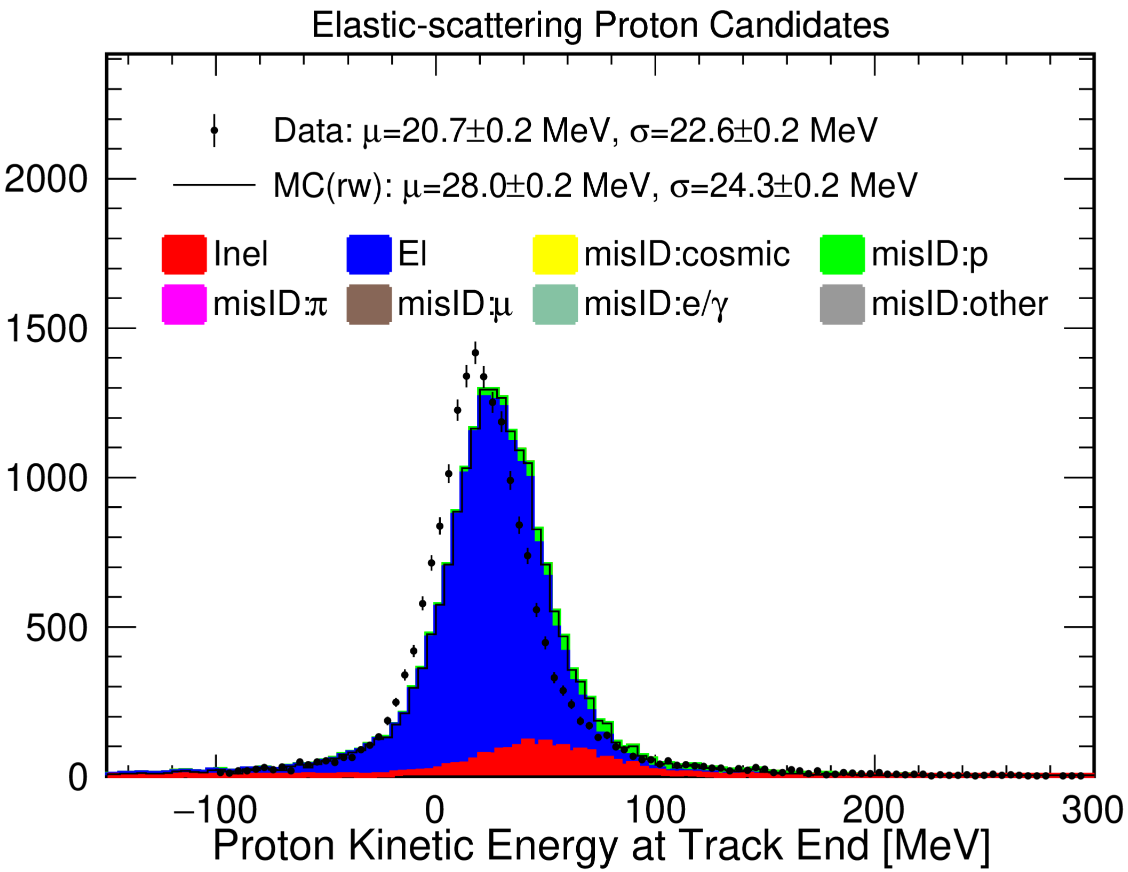




- Good reconstruction at KE_{end} for both data and MC!

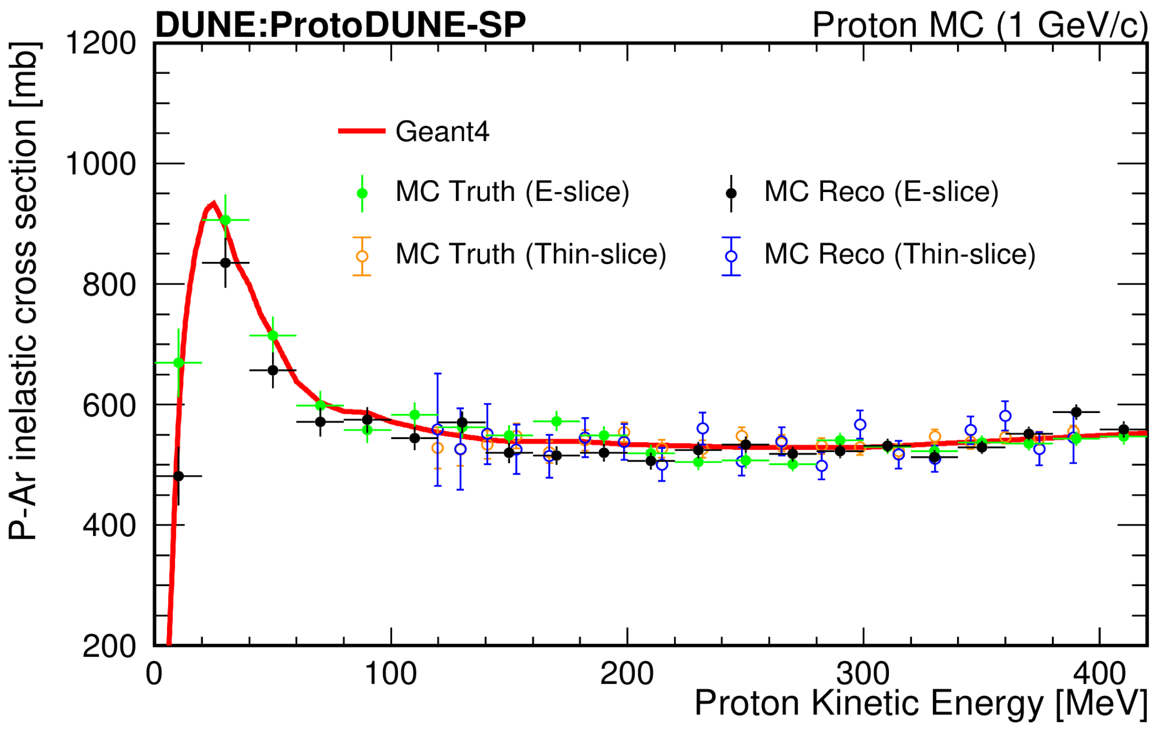
KE at Track End (Calorimetric Reconstruction)



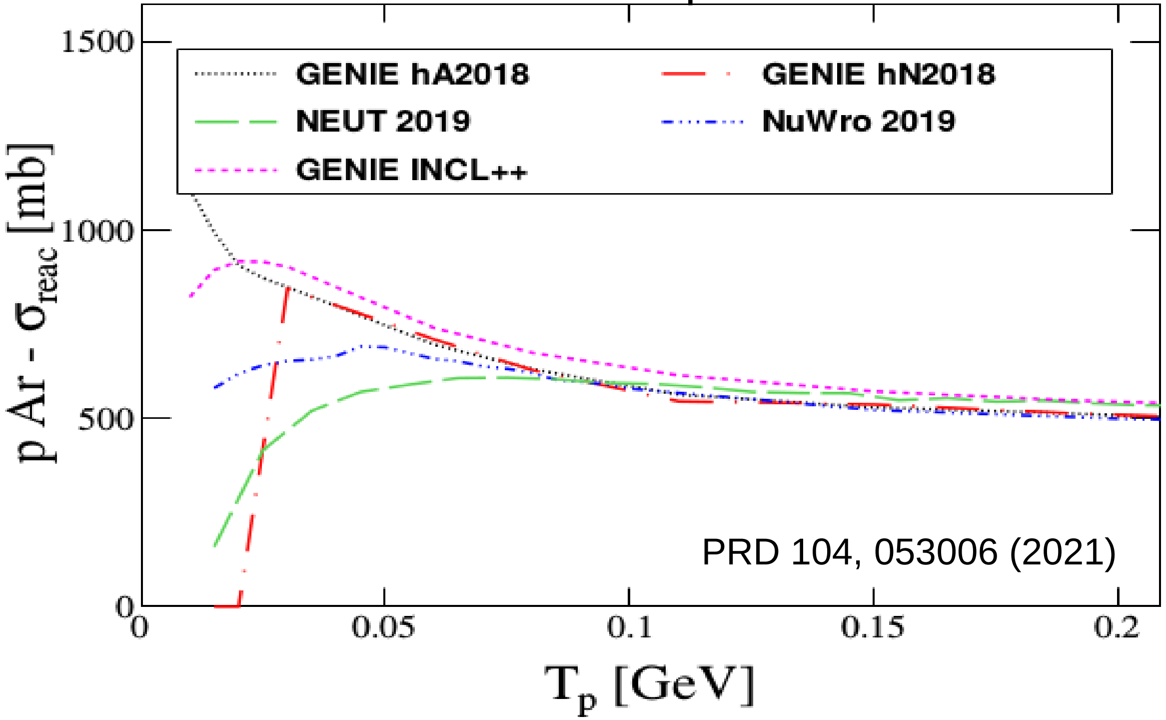


- Systematics between data and MC

Proton-Ar Inelastic Cross-section



Theoretical predictions

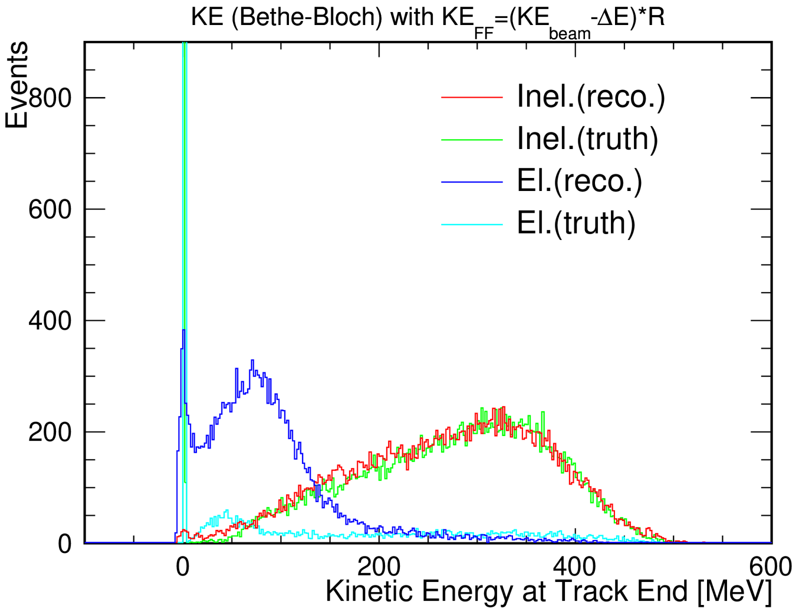


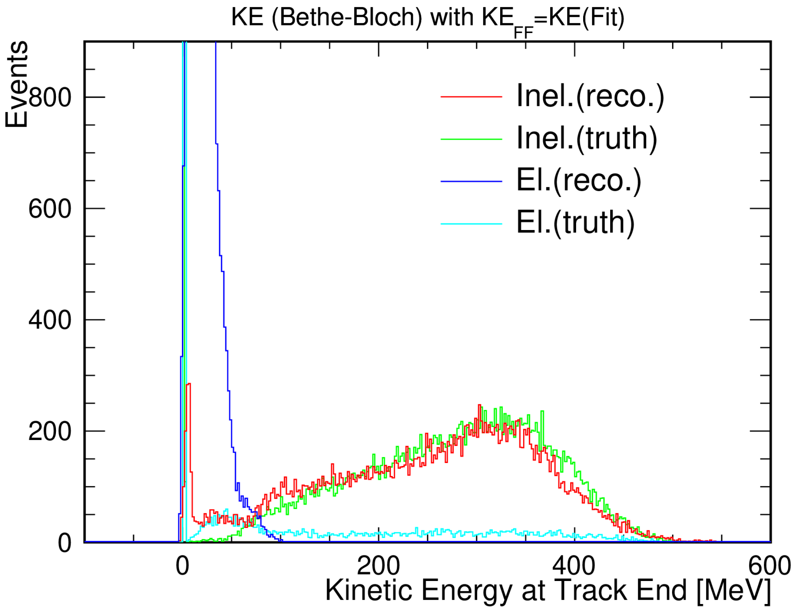
- Exciting Physics at low energy (KE<100 MeV)

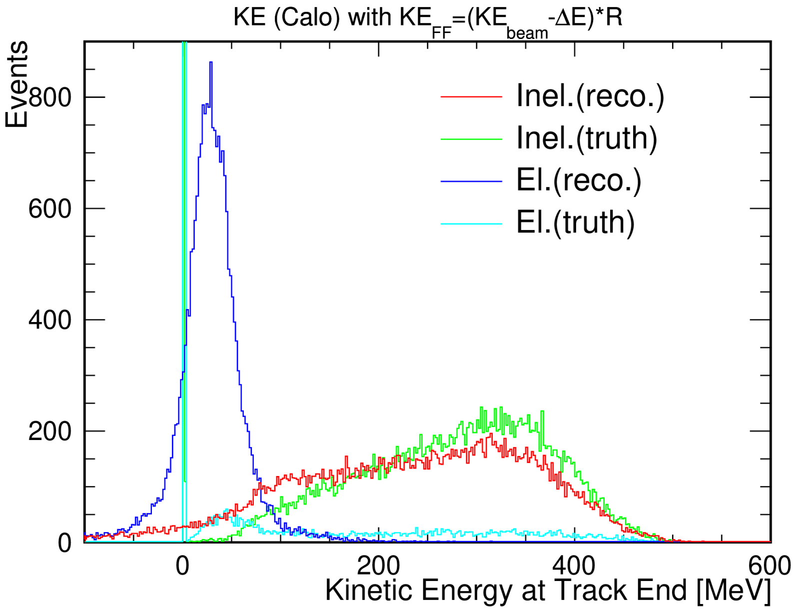
KE at Track End: Method Comparison

Bethe-Bloch

Calorimetry







Threshold=**140 MeV**

Best energy reco for inelastic-scattering protons
(reco shape=truth shape)

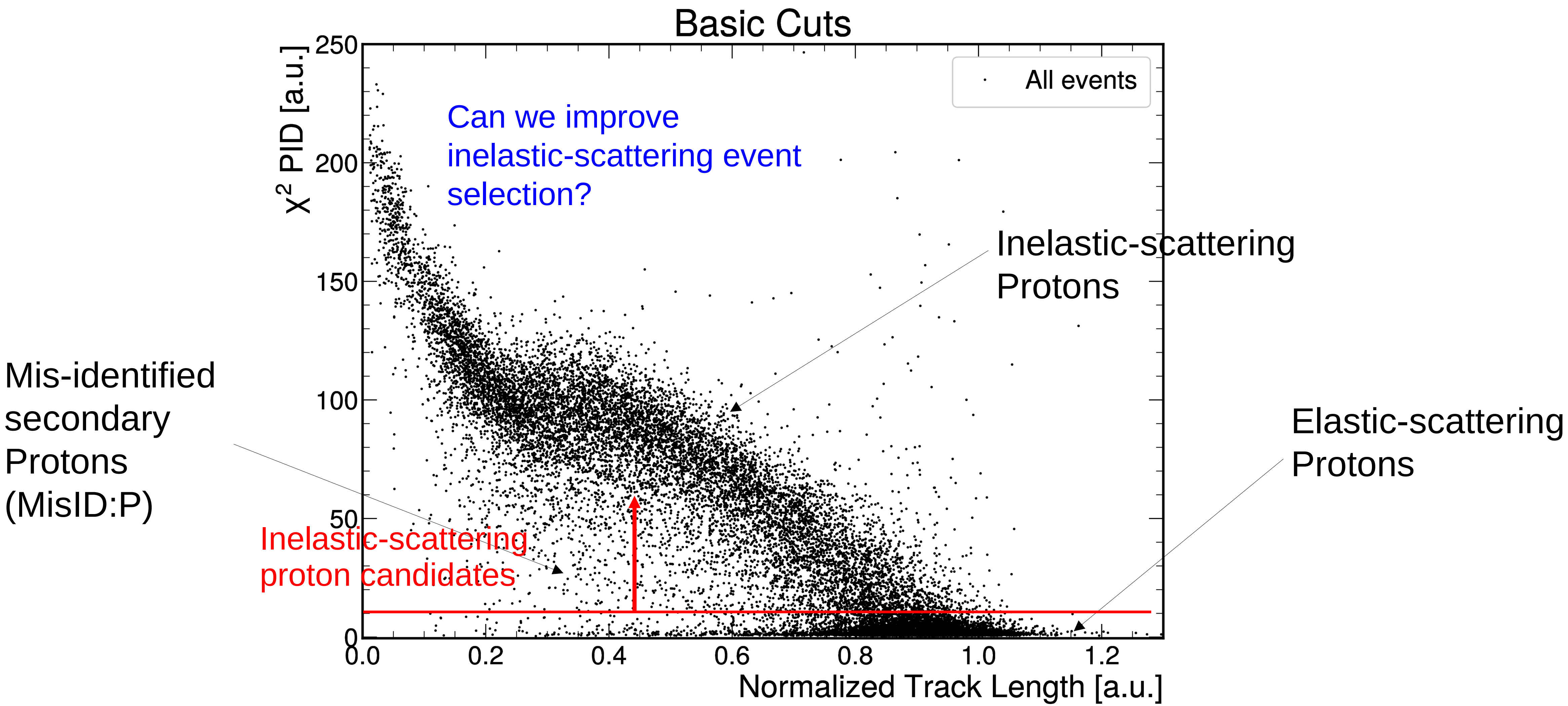
Threshold=**70 MeV**

Better energy threshold
Distorted energy spectrum for inelastic-scattering protons

Threshold=**70 MeV**

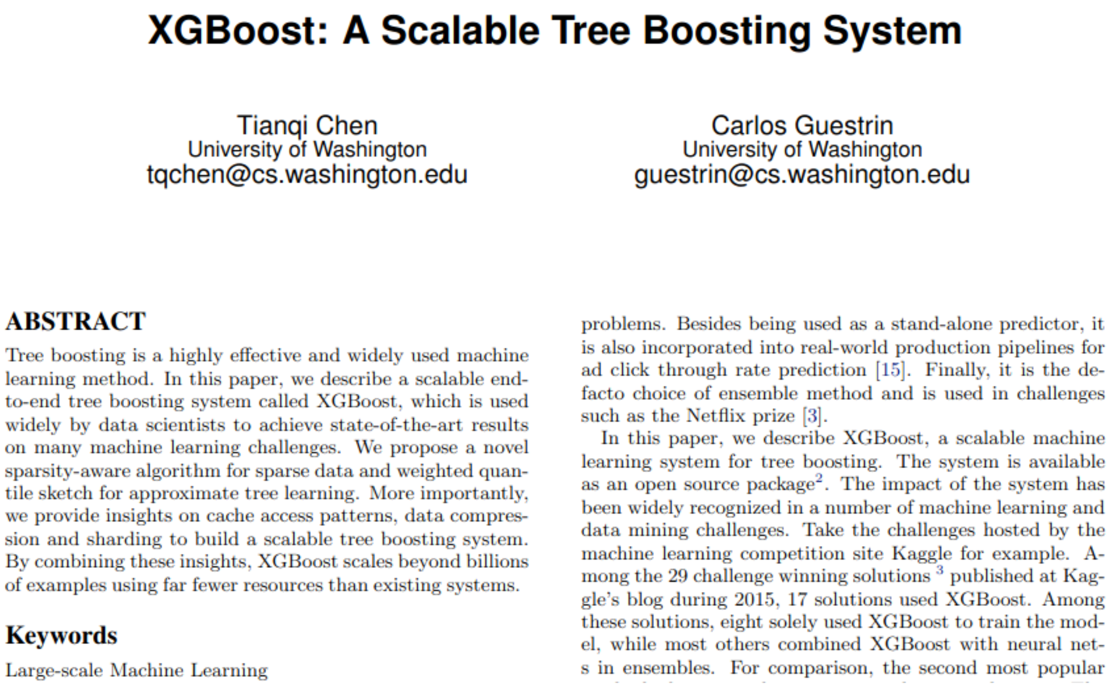
Better energy threshold
Distorted energy spectrum for inelastic-scattering protons

Inelastic-scattering Proton Event Selection



Inelastic Event Selection using XGBoost

- XGBoost: eXtreme Gradient Boosted trees (2016)
- Software package: https://xgboost.readthedocs.io/en/stable/



XGBoost: A Scalable Tree Boosting System

Tianqi Chen
University of Washington
tqchen@cs.washington.edu

Carlos Guestrin
University of Washington
guestrin@cs.washington.edu

ABSTRACT

Tree boosting is a highly effective and widely used machine learning method. In this paper, we describe a scalable end-to-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges. We propose a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. More importantly, we provide insights on cache access patterns, data compression and sharding to build a scalable tree boosting system. By combining these insights, XGBoost scales beyond billions of examples using far fewer resources than existing systems.

Keywords

Large-scale Machine Learning

problems. Besides being used as a stand-alone predictor, it is also incorporated into real-world production pipelines for ad click through rate prediction [15]. Finally, it is the de-facto choice of ensemble method and is used in challenges such as the Netflix prize [3].

In this paper, we describe XGBoost, a scalable machine learning system for tree boosting. The system is available as an open source package[2]. The impact of the system has been widely recognized in a number of machine learning and data mining challenges. Take the challenges hosted by the machine learning competition site Kaggle for example. Among the 29 challenge winning solutions [3] published at Kaggle's blog during 2015, 17 solutions used XGBoost. Among these solutions, eight solely used XGBoost to train the model, while most others combined XGBoost with neural nets in ensembles. For comparison, the second most popular

https://dl.acm.org/doi/pdf/10.1145/2939672.2939785

Question: Does the person like computer games?
Inputs: age, gender, occupation (i.e. features)

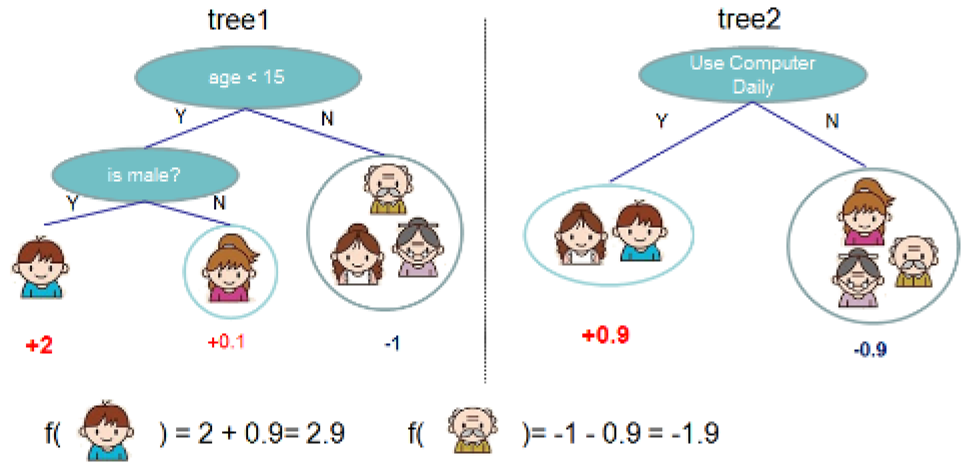
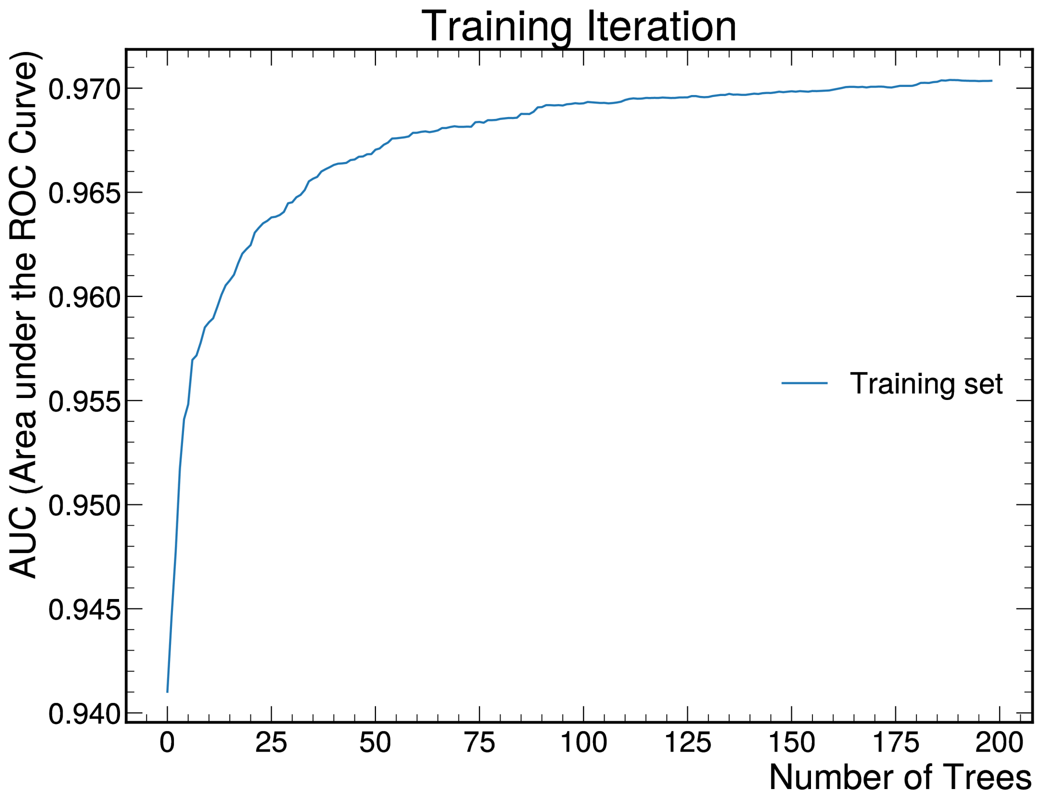


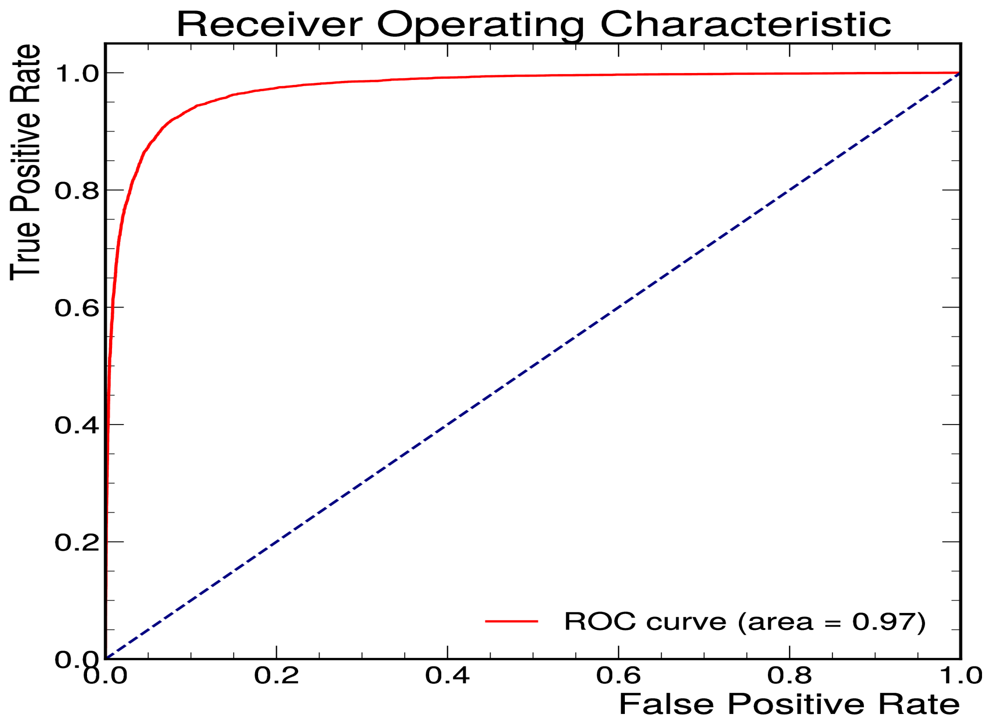
Figure 1: Tree Ensemble Model. The final prediction for a given example is the sum of predictions from each tree.

Feature Observables

- 9 features used in total:
  (1) PID: Chi^2 PID
  (2) ntrklen: Normalized track length
  (3) B: Impact parameter
      (3D distance between endpoint to the projected line fitted using the first 3 hits)
  (4) trklen: track length
  (5) calo: Σ(dE/dx*dx)
  (6) mediandedx: Median dE/dx
  (7) avcalo: Σ(dE/dx*dx)/track length (energy loss per distance)
  (8) endpointdedx: Averaged dE/dx of the last 3 hits
  (9) costheta: Angle between beam and TPC track

XGBoost: Training Process



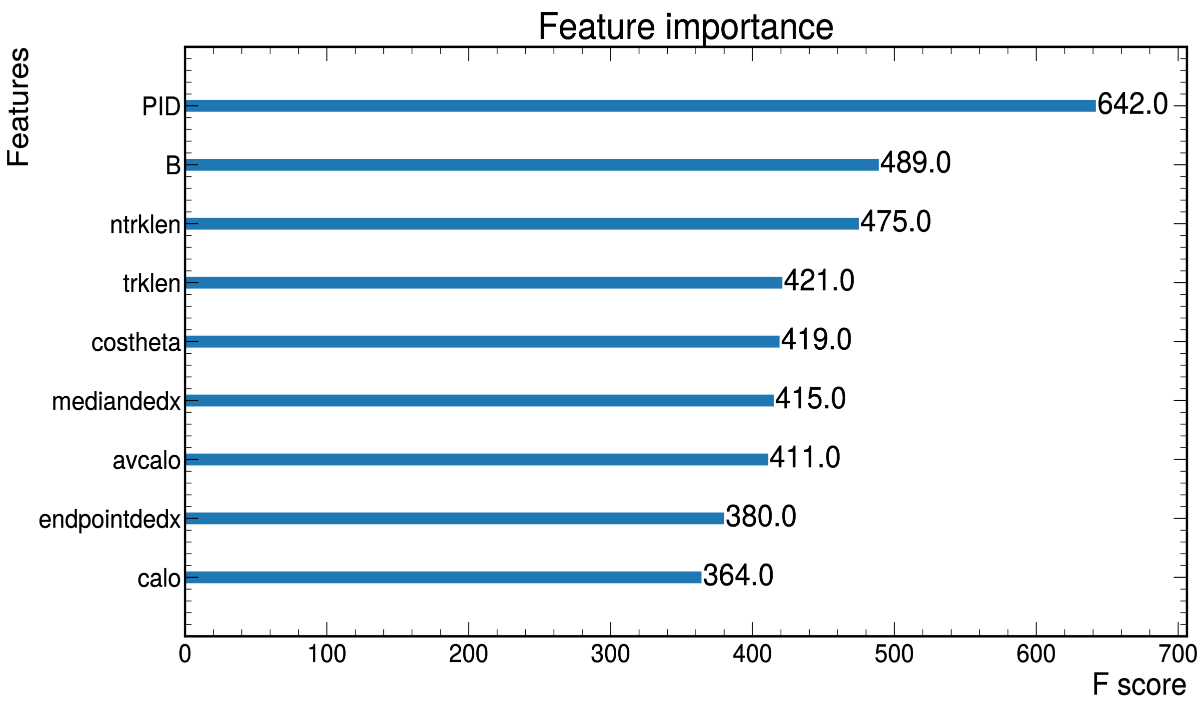


True positive (TP):A test result that correctly indicates the presence of a condition or characteristic

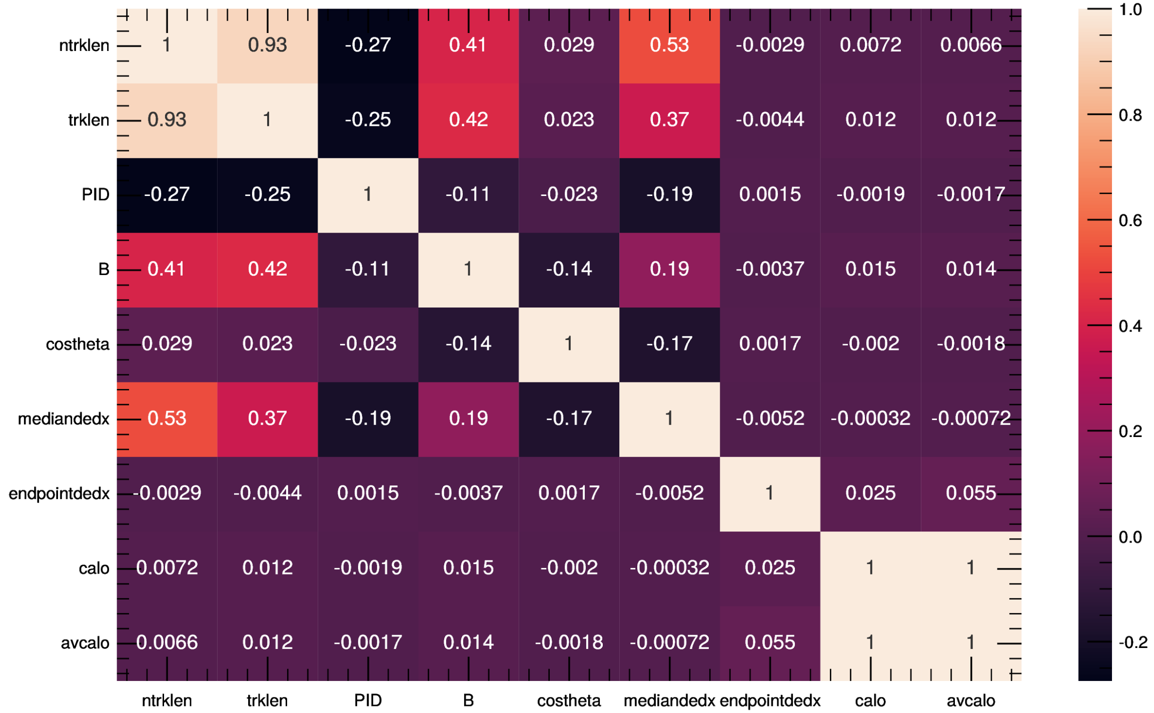
False positive (FP):A test result which wrongly indicates that a particular condition or attribute is present

- MC: 60% used for training; 40% for cross-validation
- AUC (Area under ROC) is used for evaluation of “distance” between reco and truth
- Less than 40 sec processing time using pre-built model

Feature Variables

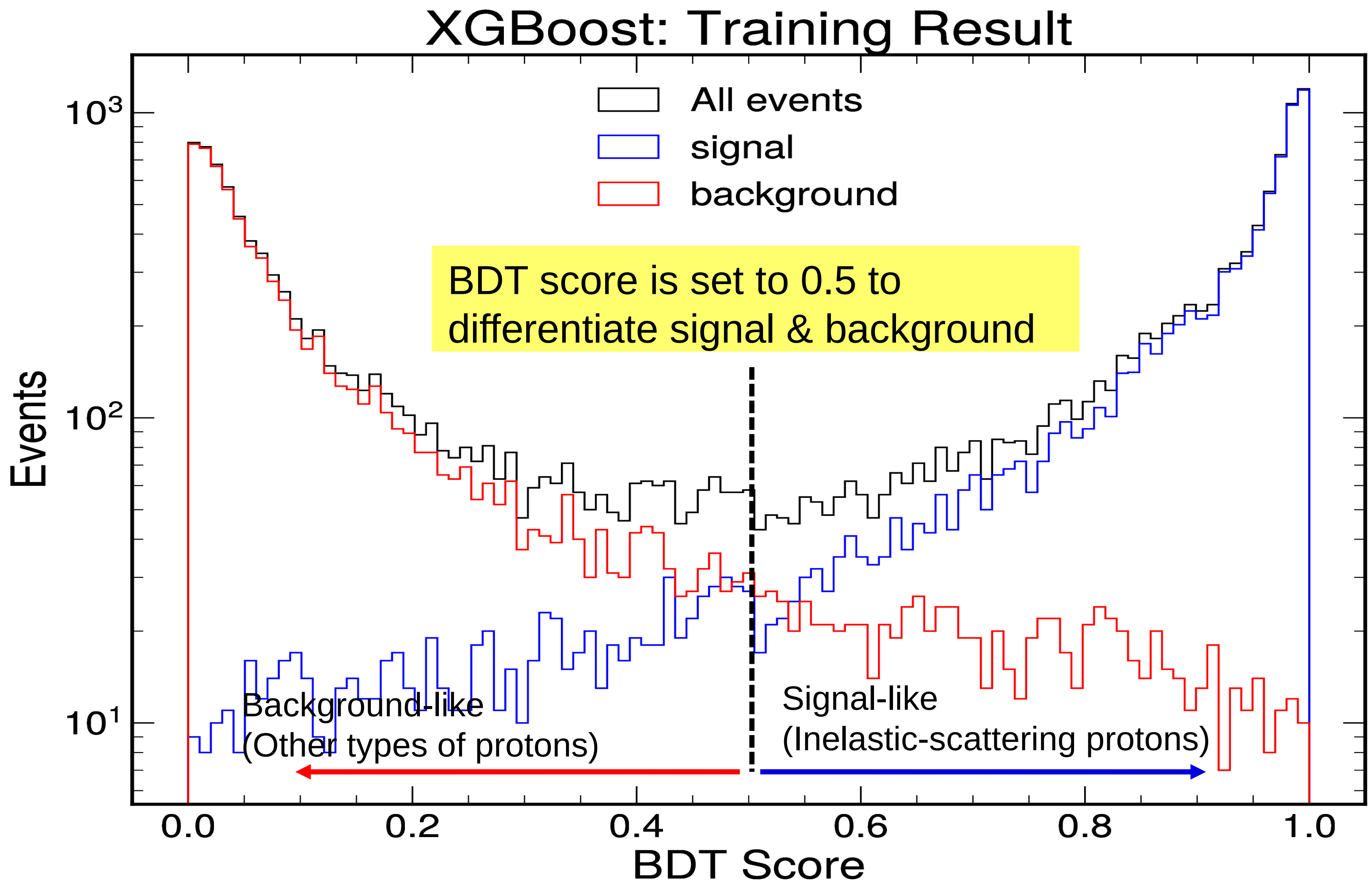


Correlation between different variables



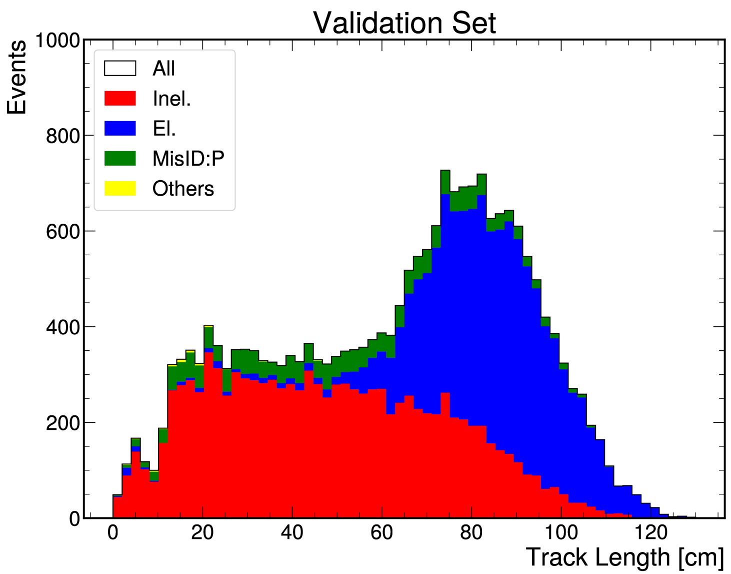
- F-score: A metric that sums up number of times each feature is split on
- PID is the most important feature
- Correlation matrix seems reasonable

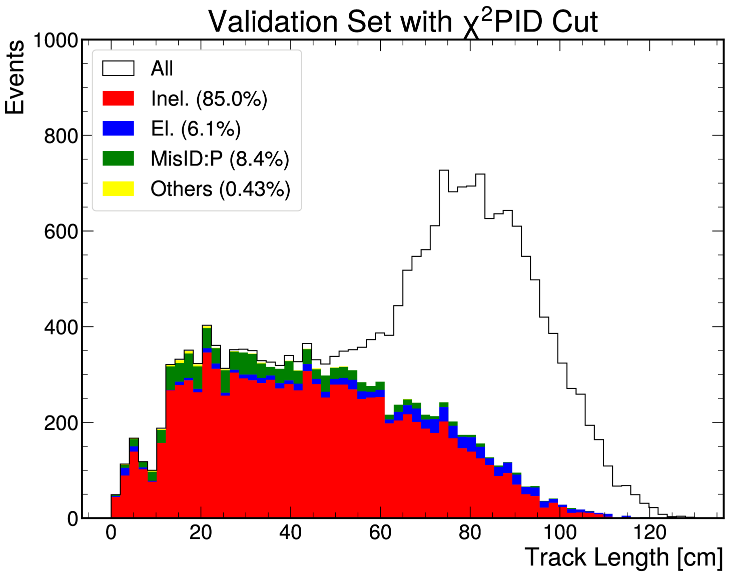
Training Result & Selection Cut

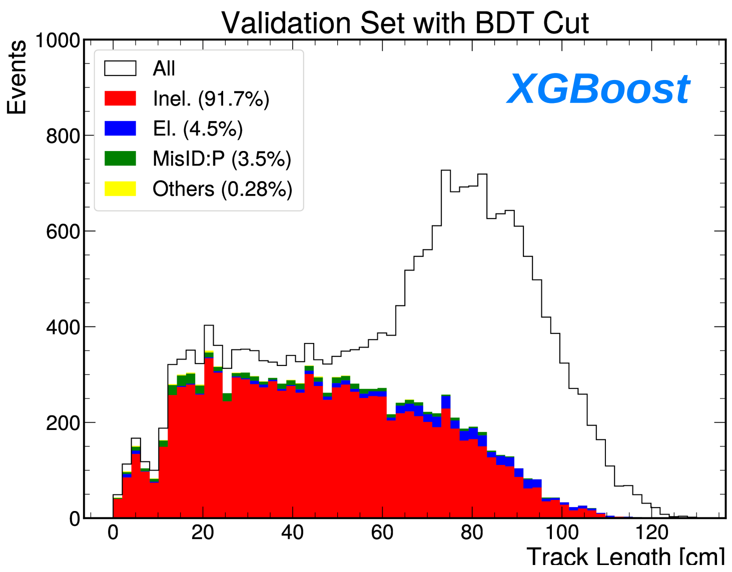


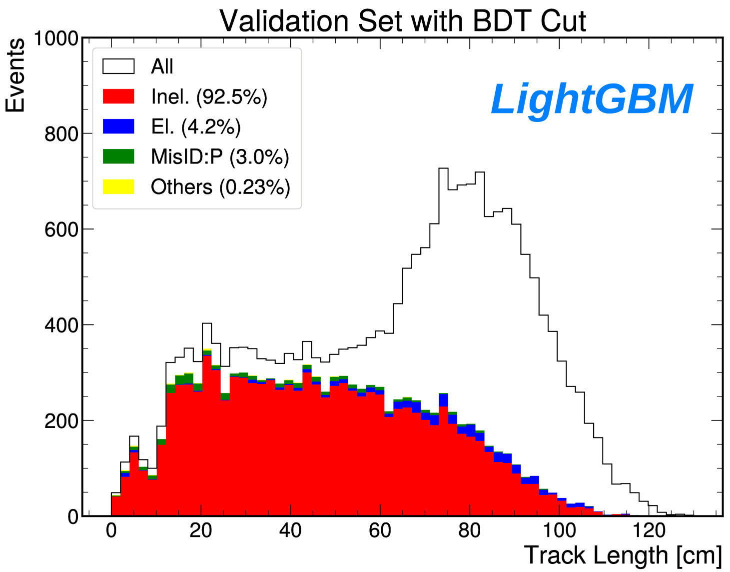
- Good separation between signal and background

Event Selection Cut









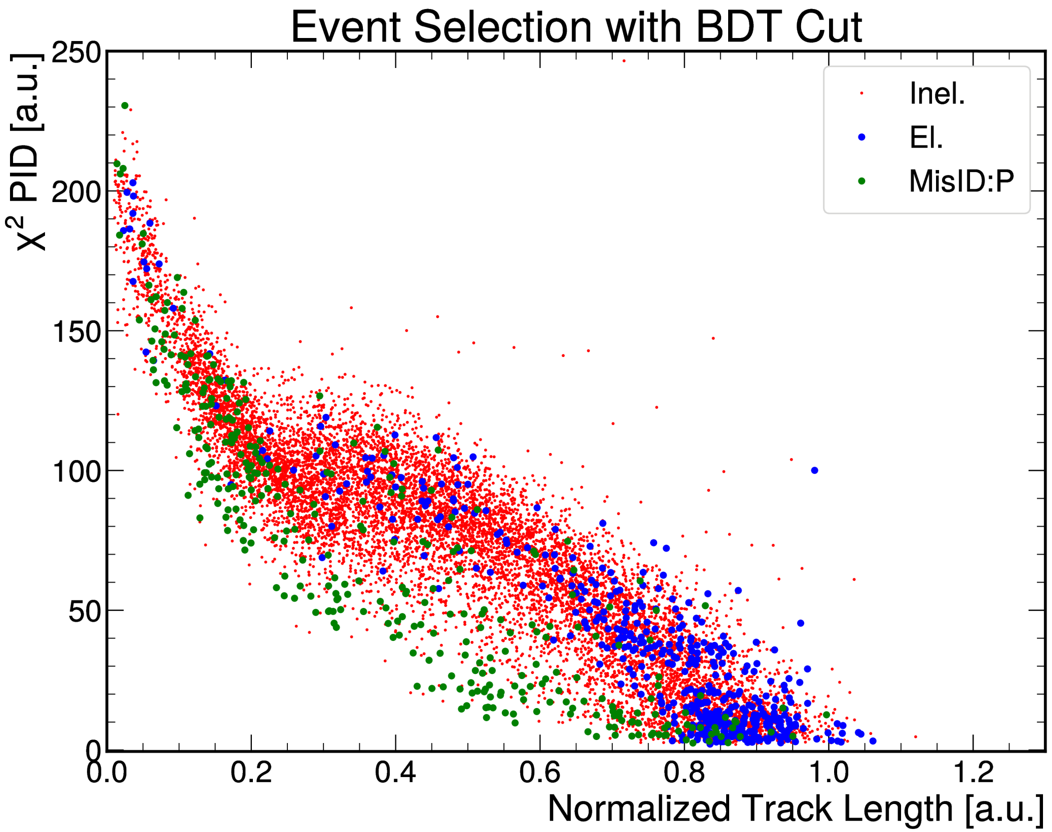
- Inel.: 8% improvement (93 % purity obtained) (4% MisID:P + 3 % El. background)

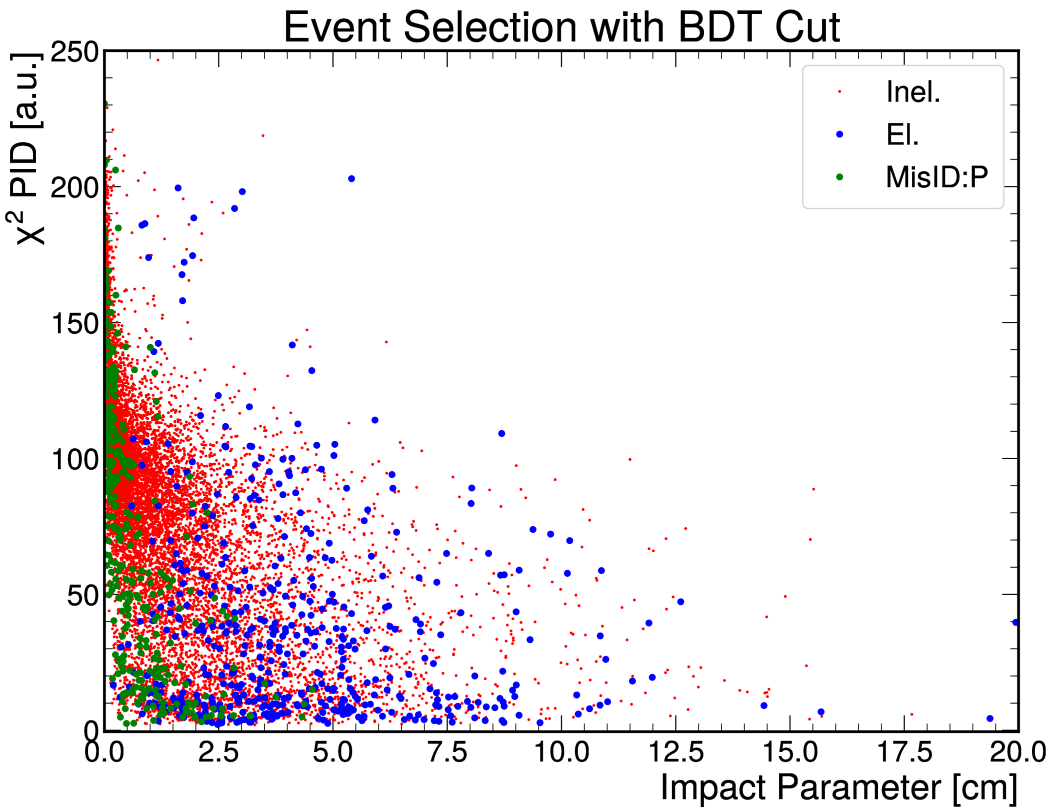
Summary

- New KE reconstruction mitigates KE systematics
- 93% purity obtained (8% improvement) on selecting inelastic-scattering protons using XGBoost and LightGBM
- Potential improvement on lowering energy threshold to 70 MeV using hypothetical residual fit

Backup

Event Selection: Signal & Background





- Will be hard to cut out remaining backgrounds using current observables
- Possible improvement including more energy-related observables (i.e. KE_{bb}, KE_{ff}, KE_{calo}, ...)