Automated proton track identification in MicroBooNE using gradient boosted decision trees

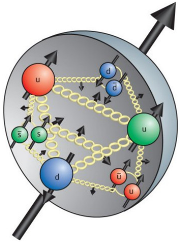
Katherine Woodruff
with
Vassili Papavassiliou Stephen Pate

ACAT 2017, Seattle, WA

August 24, 2017

Strange quark contribution to nucleon spin

The net spin of the proton is composed of contributions from its quarks and gluons



$$\frac{1}{2} = \frac{1}{2}\sum_q \Delta q + \Delta G + L_q + L_g$$

- $\sum_q \Delta q$ and ΔG are the contributions from the spin and L_q and L_g are the contributions from the angular momentum of the quarks and gluons
    - $\sum_q \Delta q = \Delta u + \Delta d + \Delta s$

We want to know the total contribution to the nucleon spin that comes from the spin of strange quarks and antiquarks (Δs)

$$\Delta s = (s^{\uparrow} + \bar{s}^{\uparrow}) - (s^{\downarrow} + \bar{s}^{\downarrow})$$

Δs was expected to be zero

- Found to be negative in polarized, charged-lepton, DIS
    - Assumes flavor SU(3) symmetry
    - Analyses give range $\Delta s = -0.08$ to -0.14 [1]
    - Measurements in semi-inclusive DIS gave results consistent with zero

[1] R. L. Jaffe and A. Manohar, Nucl. Phys. B337, 509 (1990).

Neutral-Current Elastic νp Scattering

Δs can be determined independently in neutral-current (NC) elastic scattering:

- NC elastic νp cross section:

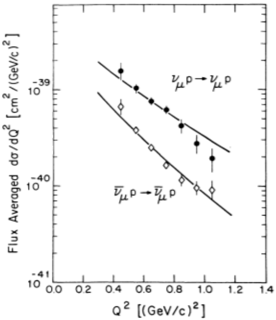
$$\left(\frac{d\sigma}{dQ^2}\right)_\nu^{NC} = \frac{G_F^2}{2\pi}\left[\frac{1}{2}y^2(G_M^{NC})^2 + \left(1 - y - \frac{M}{2E}y\right)\frac{(G_E^{NC})^2 + \frac{E}{2M}y(G_M^{NC})^2}{1 + \frac{E}{2M}y} + \left(\frac{1}{2}y^2 + 1 - y + \frac{M}{2E}y\right)(G_A^{NC})^2 + 2y\left(1 - \frac{1}{2}y\right)G_M^{NC}G_A^{NC}\right]$$

- G_E^{NC}, G_M^{NC}, G_A^{NC} are form factors representing the electric, magnetic, spin and distributions in the nucleon
- Can get net spin contribution from all three quarks from axial form factor when $Q^2 \to 0$

$$G_A^{NC}(Q^2 = 0) = -\Delta u + \Delta d + \Delta s$$

- $\Delta u - \Delta d$ has been determined in neutron decay

Neutrino-Based Experimental Measurements of Δs

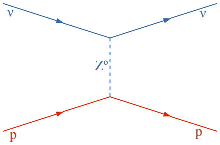


NC elastic measurement from the E734 neutrino scattering experiment at BNL

- Measured NC elastic $\nu - p$ interactions down to $Q^2 = 0.45$ GeV^2
- Found $-0.31 \leq \Delta s \leq -0.04$ [2]
  - Sensitive to choice of shape of form factor
- Much of uncertainty due to lack of data at low momentum transfer (Q^2)

NC elastic νp signal is a single, isolated proton

- Difficult to measure at low Q^2
- Kinematics determined by proton energy: $Q^2 = 2Tm_p$

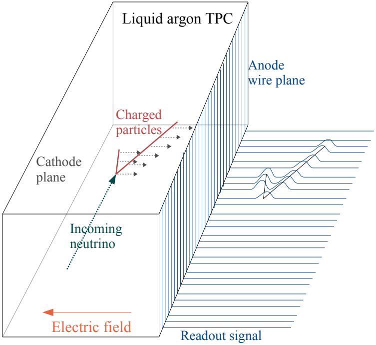


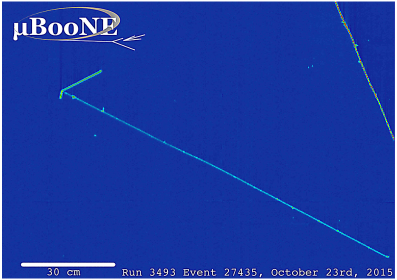
Need a dense, high-resolution detector
$\Rightarrow$ Liquid argon time projection chamber

[2] G.T. Garvey, W.C. Louis and D.H. White, Phys. Rev. C 48 (1993) 761.

Liquid Argon TPCs

- Large liquid argon target for neutrino interactions
- Charged particles produced in interaction ionize the argon
- Ionization electrons drift to anode wire plane due to electric field





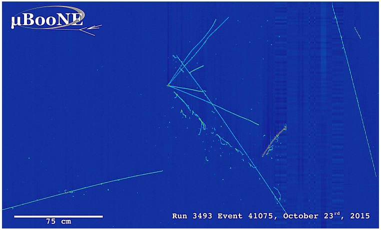
- Signal from electrons on wires is read out
- Reconstruct images of events

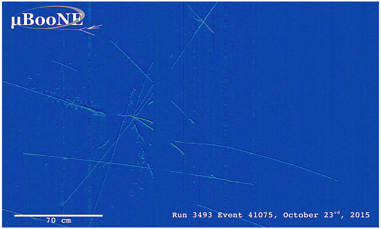
MicroBooNE LArTPC

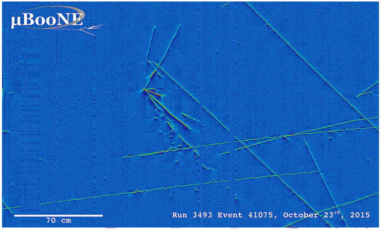
Photo: Fermilab

MicroBooNE detector:

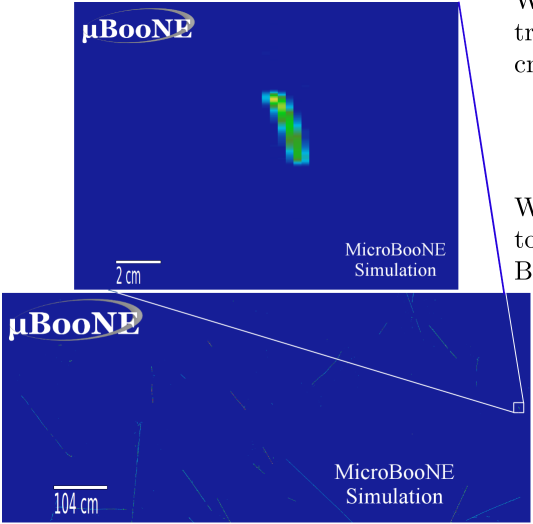
- 3 planes of ∼3000 wires each with 3mm spacing
- 10 x 2.5 x 2.3 m^3
- Each event is ∼30 MB file size
- Installed in detector hall summer 2015
- Two years of running:
  have collected 5.6e20 protons on target
  - ∼200,000 neutrino events







Neutral-Current Elastic νp events in MicroBooNE



Simulated 70 MeV proton from NC elastic event (Q^2=0.13 GeV^2)

We are able to detect protons that traverse as few as five wires (1.5 cm)

- Corresponds to a NC elastic interaction with $Q^2 = 0.08$ GeV^2

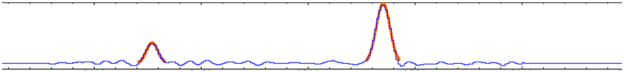
We expect 10,000 NC elastic proton events above during MicroBooNE's three year run

- Makes up ~5% of neutrino interactions in MicroBooNE
- Large cosmic background
- Need automated reconstruction and selection!
    - Hasn't been done before in a LArTPC

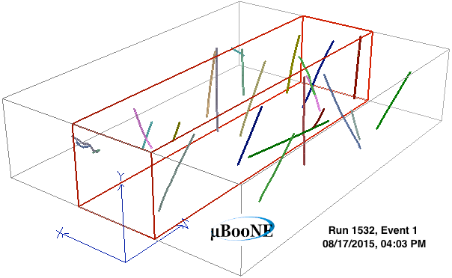
LArSoft Reconstructed Tracks

Reduce the problem by reconstructing track objects before identifying the particle and interaction type

1. Hit finding: Fit gaussians to de-noised waveform peaks



2. Track finding:
   - Combine hits from step (1) into tracks
   - Return set of reconstructed three-dimensional tracks



Have gone from 3 × 20 million pixels to ∼20 track objects without losing much information

- Big reduction in dimensionality!

Reconstructed track features

The reconstructed track objects contain information about each track that can be used to classify track type

- There are two main classification goals:
  1. Separate neutrino-induced tracks from cosmic-induced tracks
  2. Identify neutrino-induced particle type (proton, muon, etc.)

Example goal (1) features:

- Position — is it entering or near the top of the detector?
- Angle — how forward or downward going is the trajectory?

Example goal (2) features:

- Shape — how long, dense, or curvy is the track?
- Charge — charge deposited, how steep is the dE/dx curve?

None of these tell the whole story — we can use a machine learning algorithm to optimize selections in multiple dimensions at once

Boosted Decision Trees

Why trees?

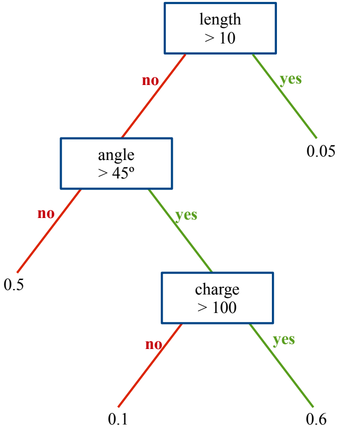
- Conceptually similar to traditional physics cuts
    - The feature space is easily interpretable/understandable
- Works with large datasets

Regression tree:

- A decision tree where each leaf contains a continuous outcome
- Each split made to maximize information gain or minimize loss function

Boosted trees:

- Ensemble method (many weak learners combined)
- Trees are created iteratively
- Each new tree trains based on the mis-classification of the previous trees



Boosting and the XGBoost[3] algorithm

Boosting (continued):

- The prediction is a sum of the output of each tree in the ensemble

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i)$$

- f_k represents the structure and weights of the kth tree

  The goal is to minimize the objective function:

$$\mathcal{L} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

- The loss function, $l(\hat{y}_i, y_i)$, measures the difference between a prediction ($\hat{y}_i$) the true label (y_i) of the ith sample
- The regularization term, $\sum_k \Omega(f_k)$, penalizes the complexity of the trees

[3] Tianqi Chen and Carlos Guestrin. 22nd SIGKDD Conference on Knowledge Discovery and Data Mining (2016) arXiv:1603.02754 (https://github.com/dmlc/xgboost)

Boosting and the XGBoost[3] algorithm

Gradient-Boosting:

- The loss function, l at tree t is

$$l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i))$$

- the difference between the true label (y_i) and the prediction of the existing ensemble ($\hat{y}_i^{(t-1)}$) plus the output of the new tree ($f_t(\mathbf{x}_i)$)
- To simplify the computation, use the second-order approximation:

$$l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) \approx l(y_i, \hat{y}_i^{(t-1)}) + \frac{\partial\, l(y_i, \hat{y}_i^{(t-1)})}{\partial\, \hat{y}^{(t-1)}} f_t(\mathbf{x}_i) + \frac{1}{2} \frac{\partial^2\, l(y_i, \hat{y}_i^{(t-1)})}{\partial^2\, \hat{y}^{(t-1)}} f_t^2(\mathbf{x}_i)$$

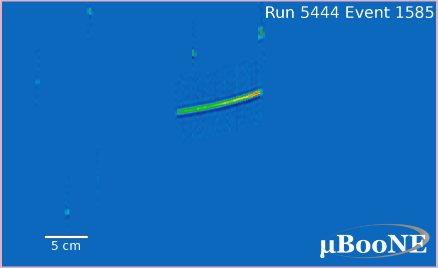
Now only need to compute the loss function and its derivatives once per iteration instead of for each split

[3] Tianqi Chen and Carlos Guestrin. 22nd SIGKDD Conference on Knowledge Discovery and Data Mining (2016) arXiv:1603.02754 (https://github.com/dmlc/xgboost)

MicroBooNE specifics

Using a multi-class classifier

- Classes: proton, muon, pion, electron/photon, and cosmic
  - Protons include both neutrino and cosmic induced simulated tracks
  - Muons, pions, electrons, and photons are neutrino induced like
  - Cosmics are any non-proton cosmic induced tracks
  - Classifies each track independently
  - Returns five probabilities per track

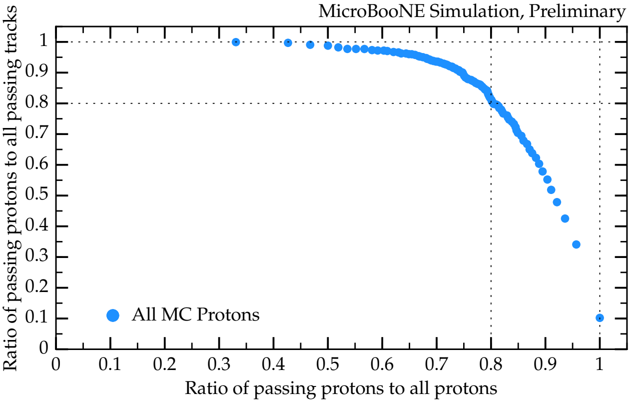


$$
\begin{aligned}
P(\mathrm{p}) &= \mathbf{0.9789} \\
P(\mu^{\pm}) &= 0.0012 \\
P(\pi^{\pm}) &= 0.0067 \\
P(e^{\pm}/\gamma) &= 0.0075 \\
P(\mathrm{cosmic}) &= 0.0058
\end{aligned}
$$

- Each track is a set of 20 reconstructed track features
  - Described on slide 9
- Trained on simulated neutrino and cosmic events

Performance on simulated protons

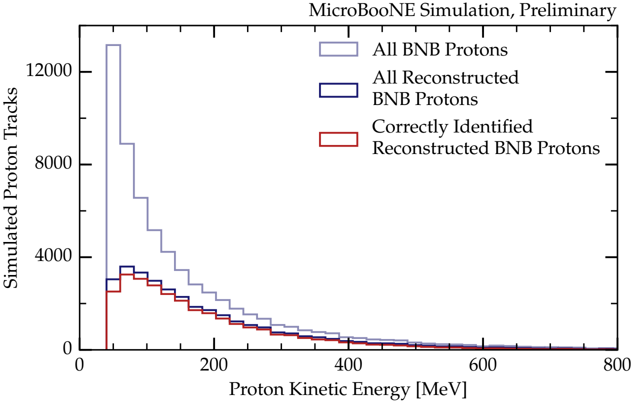
Tested the classifier on Monte Carlo simulated neutrino and cosmic events in MicroBooNE



- Showing the efficiency vs. purity of the selection on all protons in the simulation
- The different points represent cuts on different values of proton probability

Performance on simulated protons

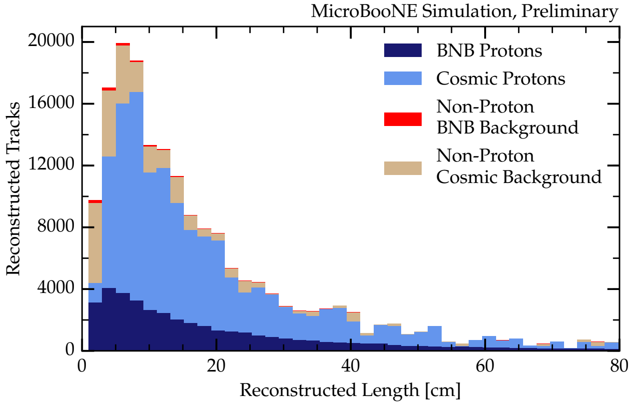
Tested the classifier on Monte Carlo simulated neutrino and cosmic events in MicroBooNE



- Showing the number of simulated neutrino-induced ("BNB") protons **generated**, **reconstructed**, and **classified correctly**
- A proton probability of greater than 50% is considered classified as a proton

Performance on simulated protons

Tested the classifier on Monte Carlo simulated neutrino and cosmic events in MicroBooNE



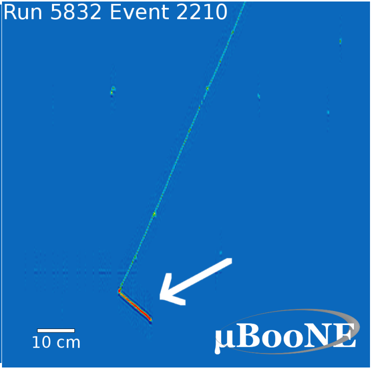
- Showing the different simulated track types classified as protons
- The blues are protons and the others are mis-classified backgrounds
- A proton probability of greater than 50% is considered classified as a proton

Example events from data

- Can select events by requiring that reconstructed tracks are indentified as specific particle types

NC candidate

CC candidate



- Isolated track classified as proton

- Tracks classified as proton and muon

Conclusion

- Can determine the net spin of the strange quarks in the proton through neutral-current elastic νp scattering
- MicroBooNE can measure low Q^2 neutral-current elastic neutrino-proton events
  - The signal is a single short proton track
- Can reconstruct track objects to reduce the dimensionality of the classification problem
  - From 30MB events to tracks with 20 features
- Can accurately classify particle types using gradient-boosted decision trees

Thank you!