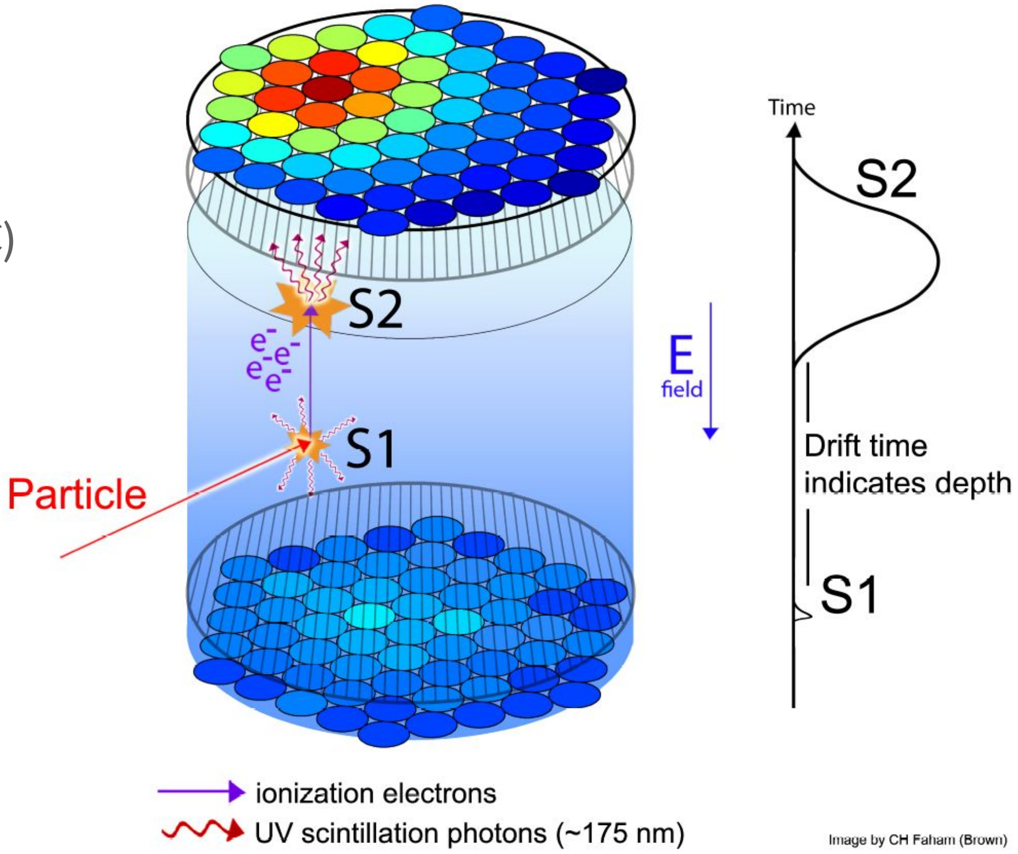
Machine Learning in LUX

Scott Kravitz
Lawrence Berkeley National Lab
On behalf of the LUX collaboration
10/29/2019

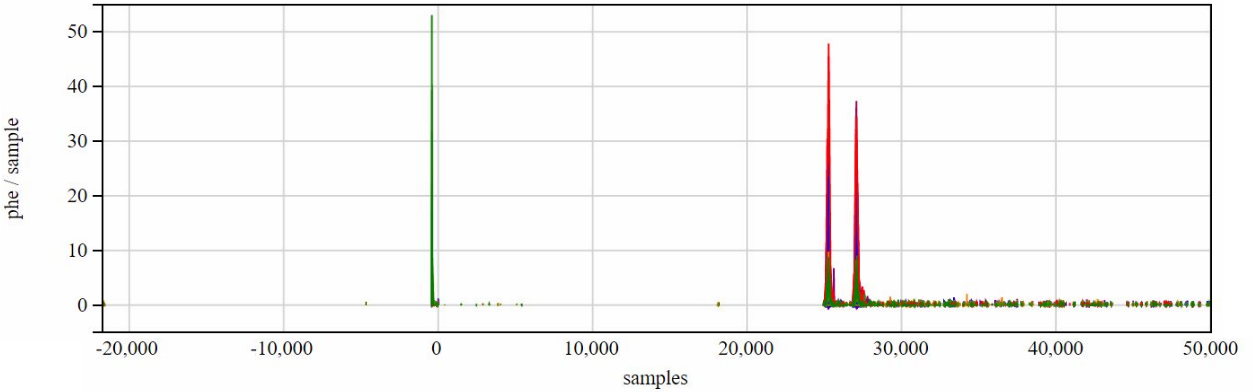
LUX Intro

- Two-phase liquid xenon time-projection chamber (LXe TPC)
- Data taking 2013-2016
- Raw data: waveform per PMT
- Typical reconstructed info (for each scatter):
    - **S1** (prompt scintillation) total area
    - **S2** (ionization signal) total area
    - **X, Y** position (from S2 PMT hit pattern)
    - **Z** (from Δt between S1 and S2)



LUX Machine Learning

- Projects:
  - **SS/MS discrimination: data-driven**, **CNN** applied to co-added tritium sum waveforms
  - **S2-only analysis: data-driven**, uses parametrized pulse shape info fed to a **BDT**
  - **LIP search:** uses **BDT** to optimize cuts for separating simulated LIP signal from background data
  - **Gamma-X:** uses **BDT** to optimize cuts for rejecting a pathological background (gamma-X events)
- **Common theme: use information not available in standard RQs / search, especially related to pulse shape**

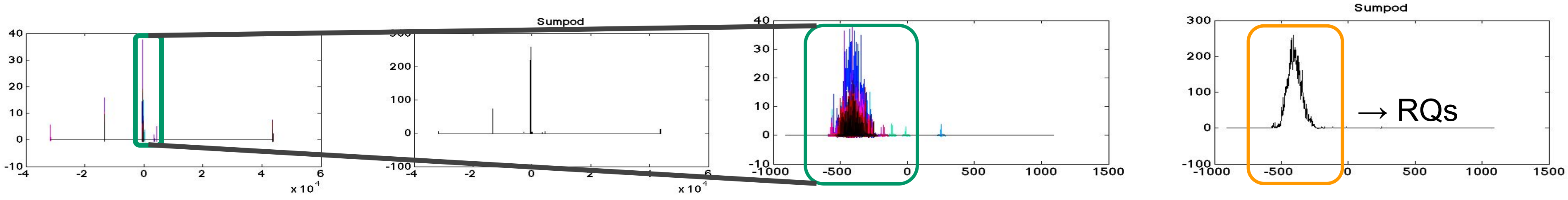


SS/MS Distinction w/ CNNs

Samuel Chan

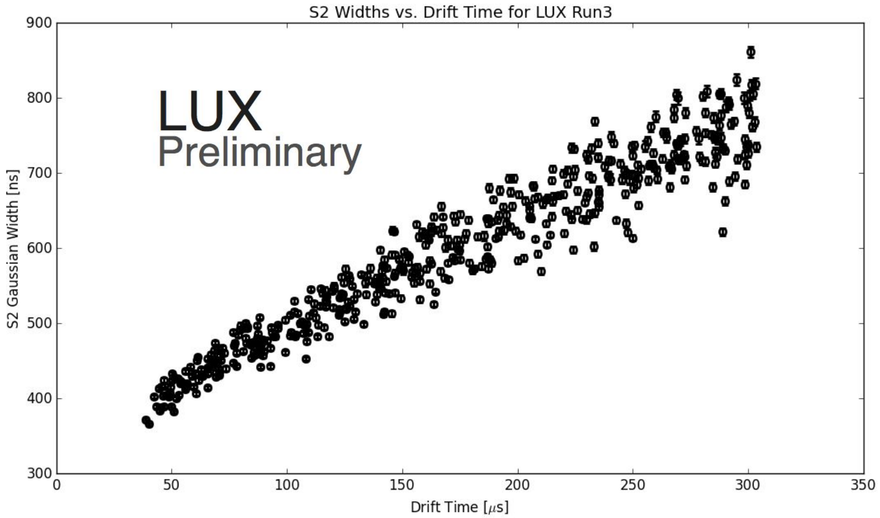
Improved S2 Separation with CNNs

- **Data processing loses info** - pulse shape, hit pattern etc.
- Double scatter classification is imperfect - merged S2s
- **Machine learning (ML) can recover info from the raw PMT traces**
- **One application:** convolutional neural network (CNN) image processing technique applied to summed waveforms for single vs. double scatter classification



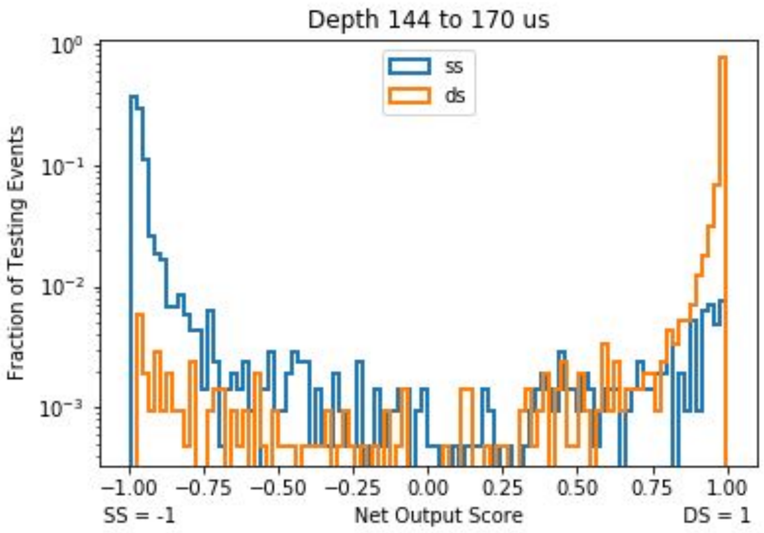
Pushing the Limits of Z Separation

- S2 widths vary with drift time due to e⁻ cloud expansion (plot from Greg Rischbieter, generated with 32 keV Kr S2 in NEST)
- **S2 width at 32 keV in the middle of LUX detector is ~0.5 us wide**
- Can we distinguish S2s with drift separations down to 0.1 us?

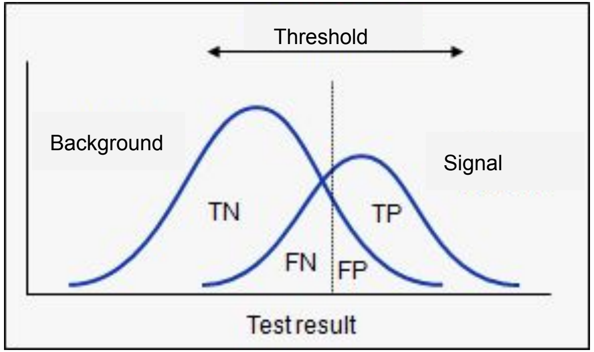


Tritium Data Training and Testing

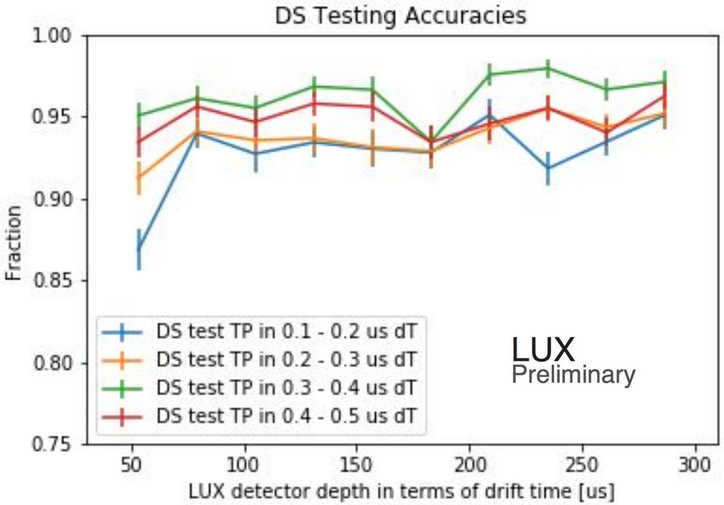
- **Data-driven training:**
  - **Single scatters from LUX tritium calibration (beta decay, electron recoil)**
  - **Double scatters constructed by co-adding the same single scatter event (same energy and PMT hit pattern) w/ a random time difference dT between [0.1, 0.5 us]**
- 120k events of each type were divided into **10 drift time (depth) bins** for separate training - variation in width < 0.05 us (see previous slide).
- Fed to a CNN using Keras
- Example CNN output for center drift bin at right

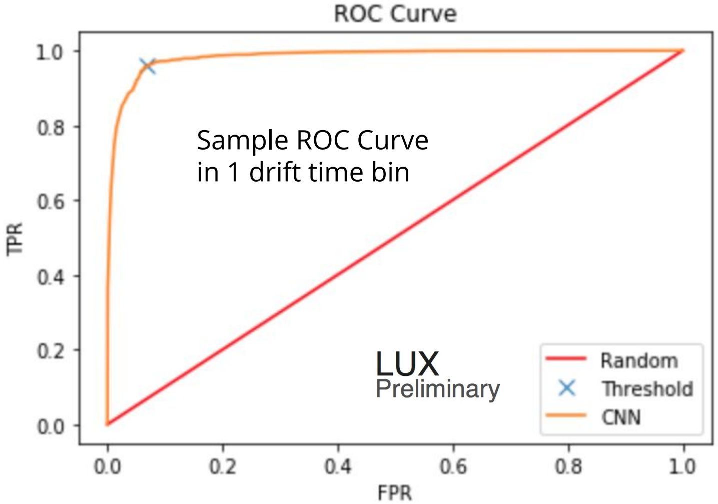


Tritium Data Accuracies



- TP: true positive ; FP: false positive ; FN: false negative
- Threshold set to separate SS from DS
- Left - DS true positive in different drift time bins
- **ROC curve shows how much better the net does compared to random guessing**





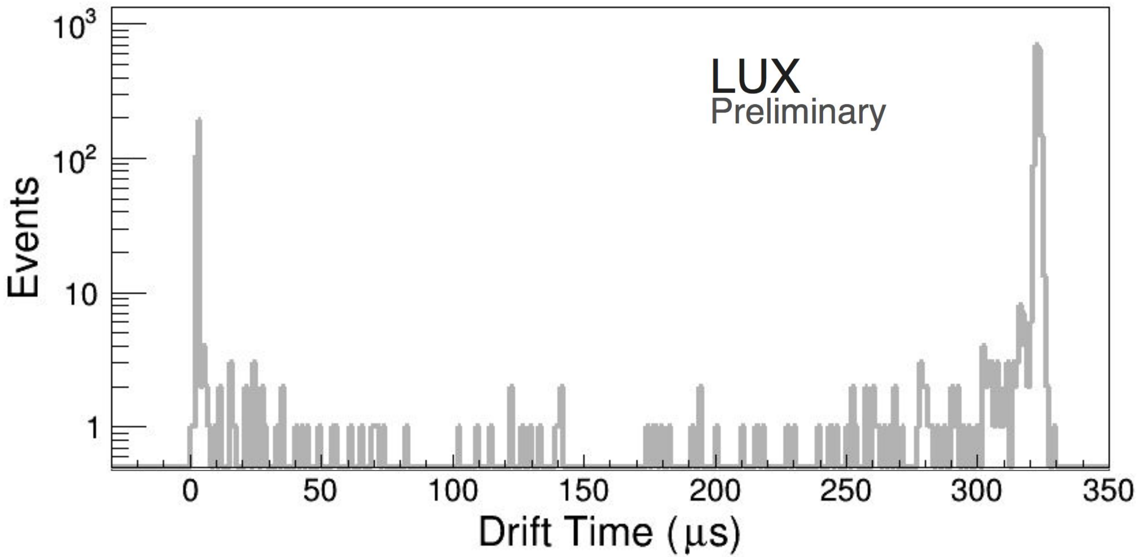
S2-only Analysis

Kelsey C Oliver-Mallory

S2-only Background

- Considering events w/o S1 lowers threshold, improves low-mass WIMP sensitivity
- But, S2-only bkg rate at 4 e- threshold is >> rate in fiducial (with drift time cut)
- Hypothesis: betas from the gate and cathode
- **Can use events w/ S1 + S2 to learn about S2-only events w/ same S2 area**

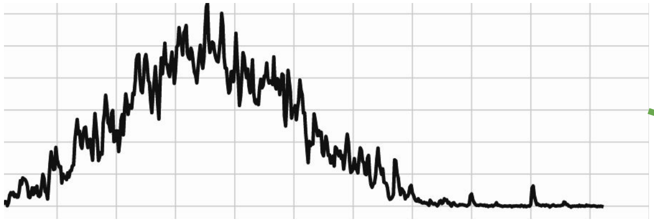
  (data-driven training)



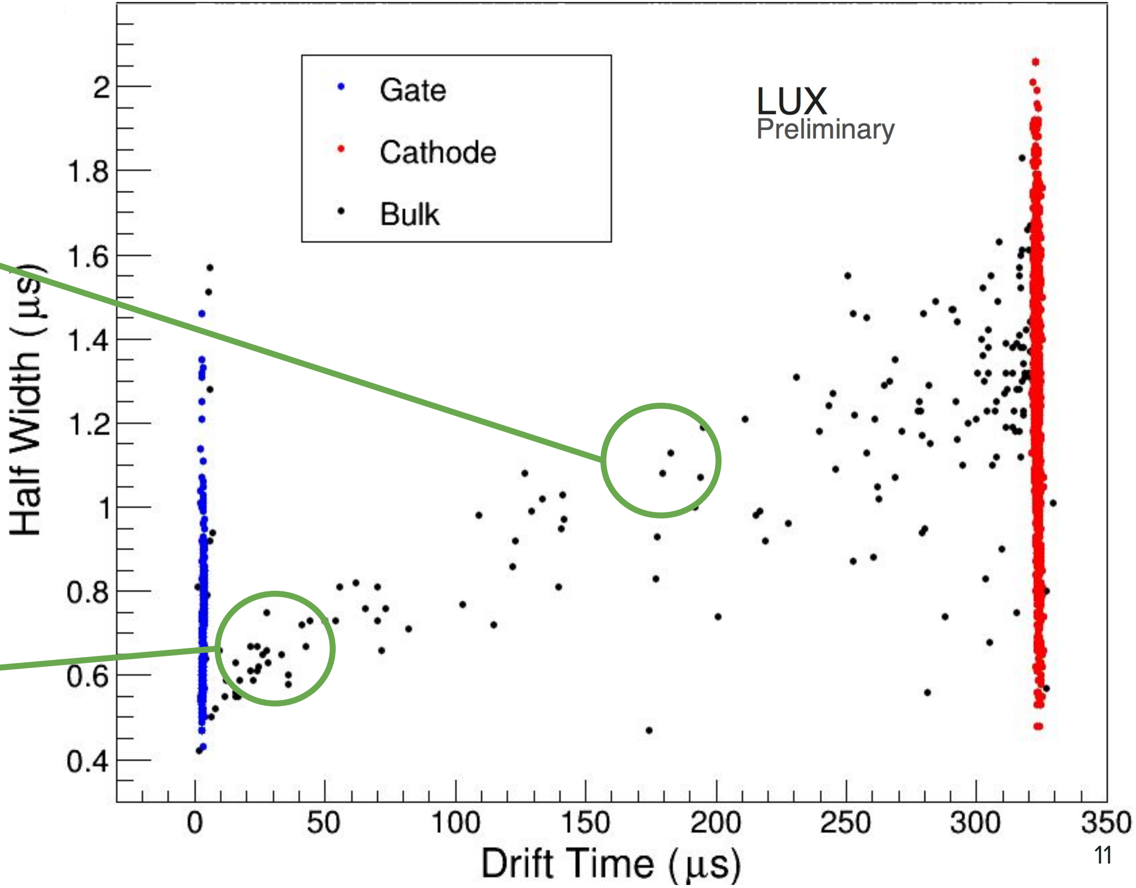
S1 + S2 events in the WIMP search region

Cut Events with Diffusion

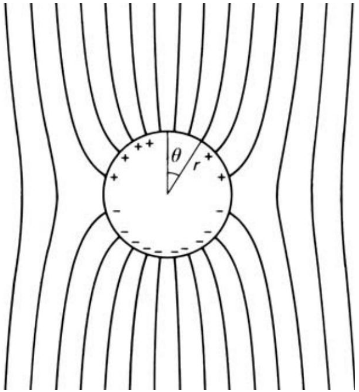
Wide and gaussian



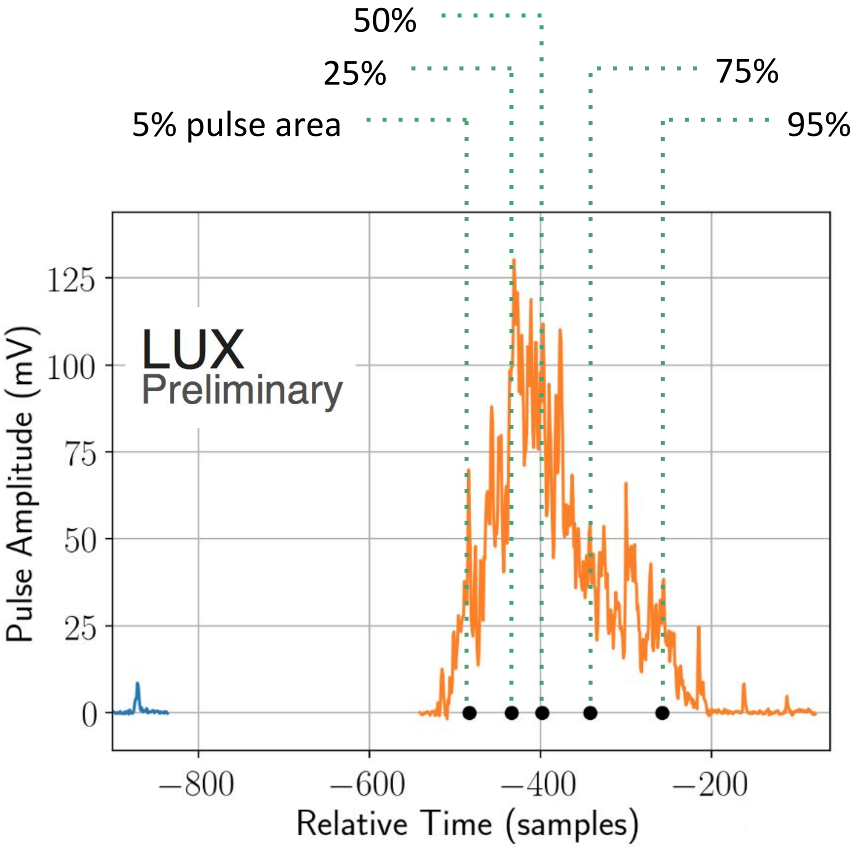
Thin and square



Pulse Shape with Machine Learning

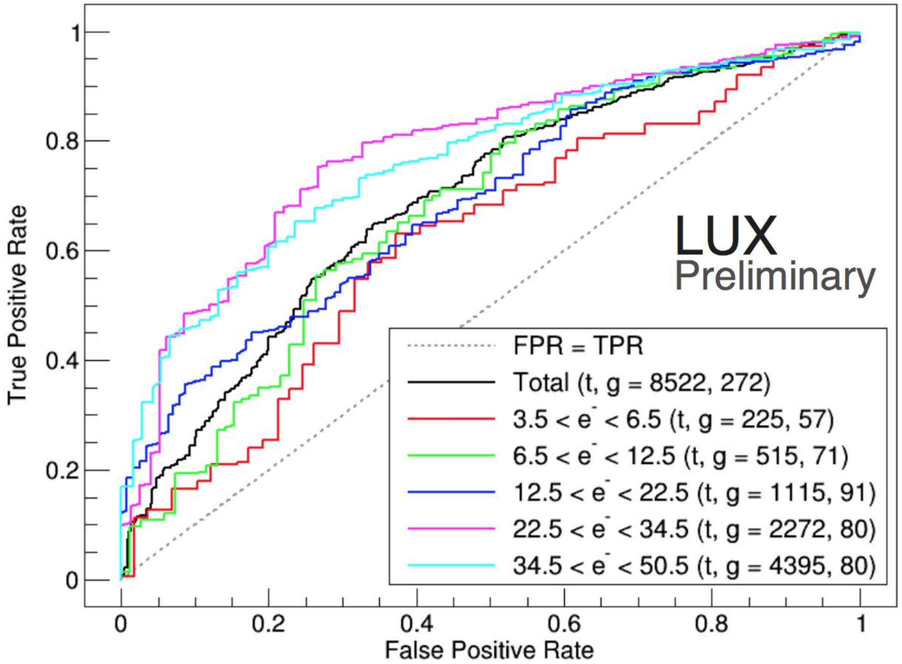


- Pulse width alone fails due to distortion of E-field near wires
- Makes grid events more spiky
- Parameterize pulse shape w/ e.g. area fractional timings
- Tritium events for bulk, else bkg at gate/cathode -> BDT

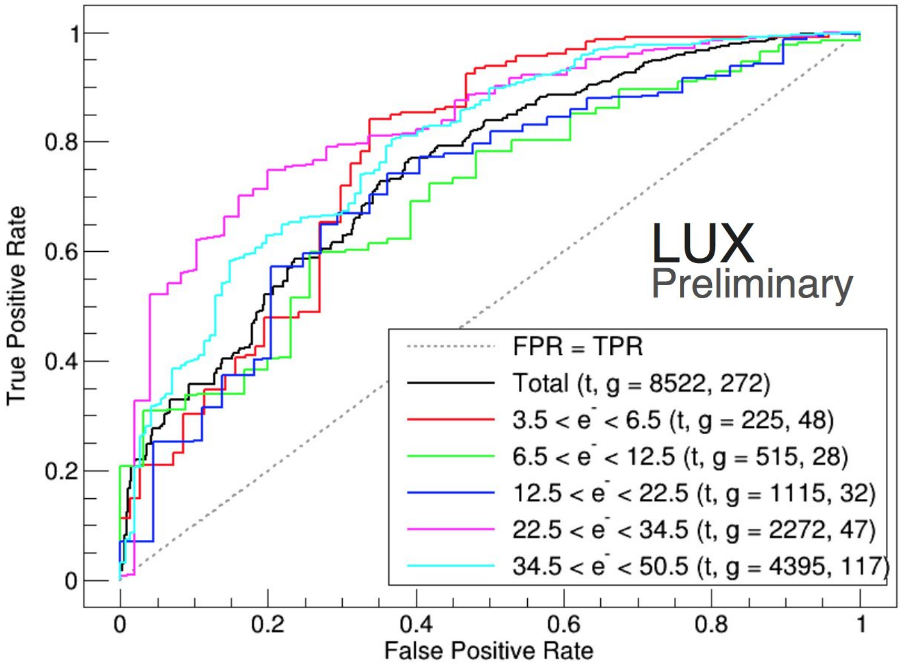


Receiver Operator Characteristic

Cathode



Gate



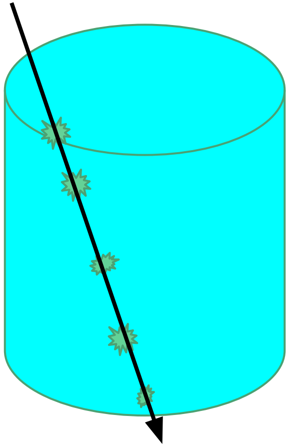
Grid events distinguishable from bulk events at few e- level!

Other ML Work

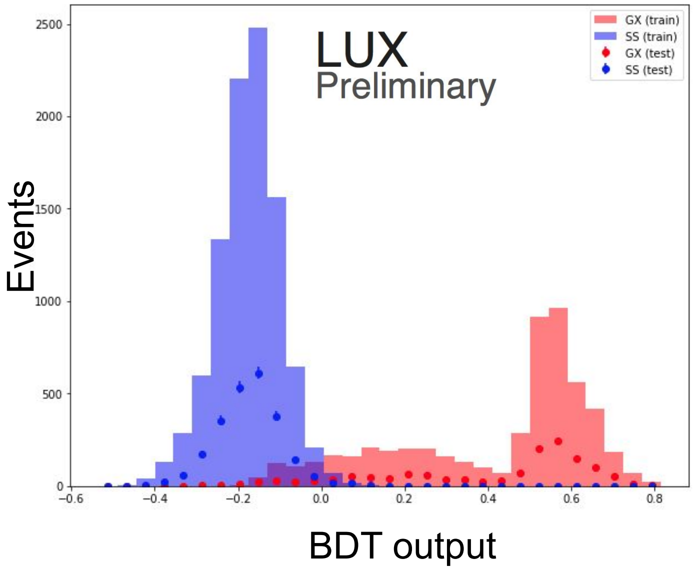
Peter Rossiter, Paul Terman, Nick Carrara

Other projects

- Lightly-ionizing particles search:
  - BDT automates and optimizes cuts
  - Improved efficiency over manual cuts



- Gamma-x backgrounds:
  - Pathological background removal using BDT



- Profile likelihood ratio:
  - Collapsing PLR dimensions into 1D using NN
  - Improved speed while preserving correlations between variables

Summary / Needs and Challenges

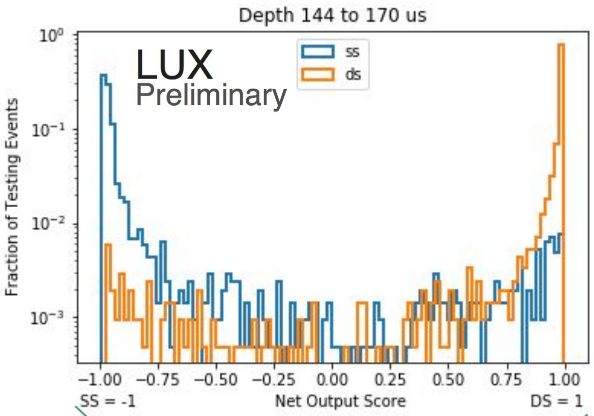
- LUX projects demonstrate improved performance over standard techniques in a broad range of domains, primarily using fairly simple ML algorithms
- Common source of new information is **timing/pulse shape**
- **Goal:** inclusion of more and lower-level information (deep learning)
- **Challenge:** getting suitably-large, reliable, and detailed training datasets to enable this (both LUX and LZ)
- **Techniques** for moving away from simulation dependence (generic):
    - Pivoting (sims only; reduce reliance on uncertain quantities)
    - Domain adversarial training (part sims, part unlabelled data)
    - Training using impure or unlabeled data from calibrations (fully data-driven)

Backup Slides

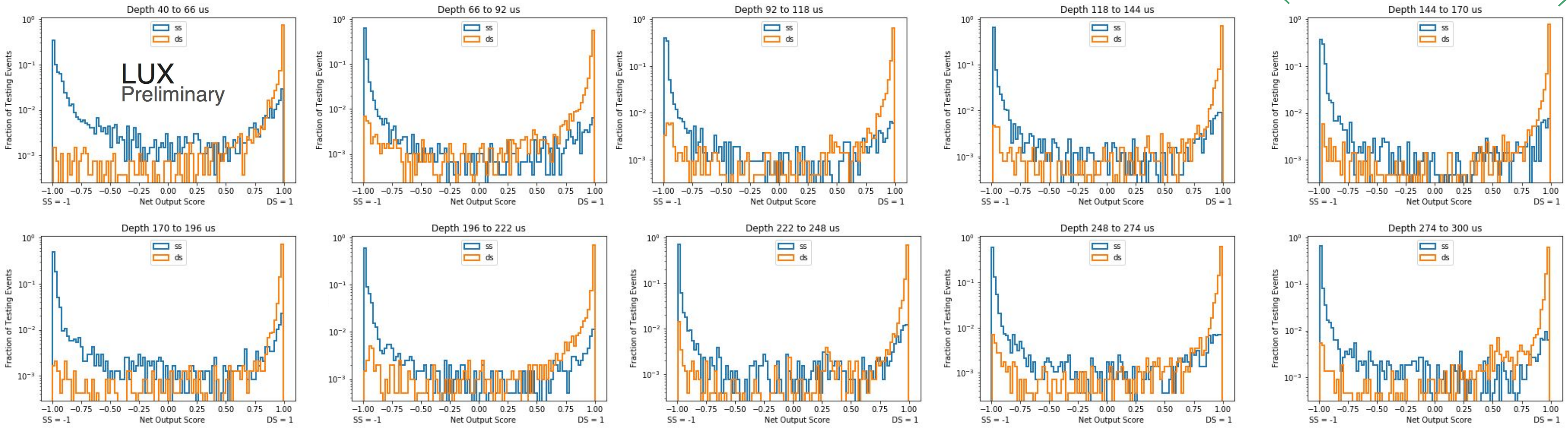
SS/MS Distinction w/ CNNs

Samuel Chan

Tritium Data Training and Testing



- Nets trained in **10 drift time bins**
- Tested using 2k SS and 2k DS events (same construction as training set)
- Good SS (blue) / DS (orange) separation seen in all bins

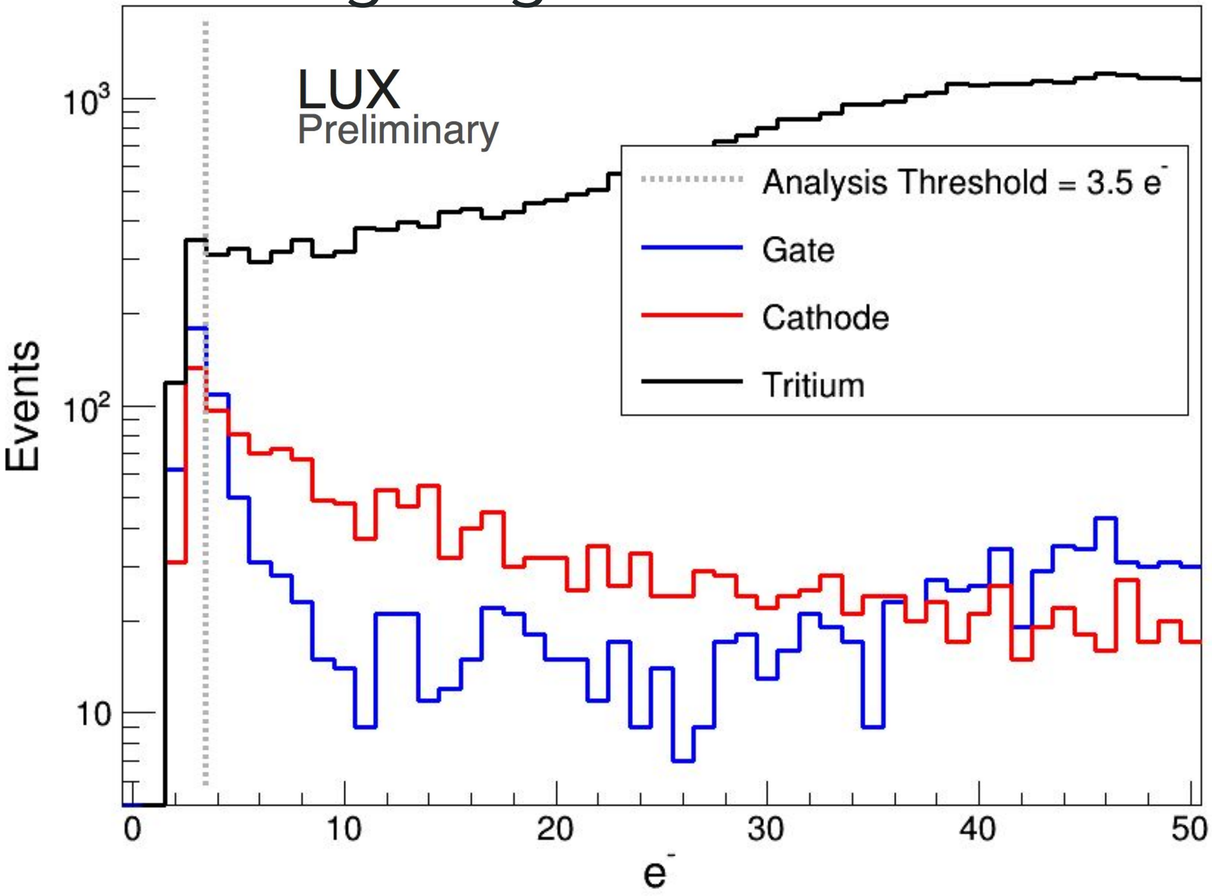


S2-only Analysis

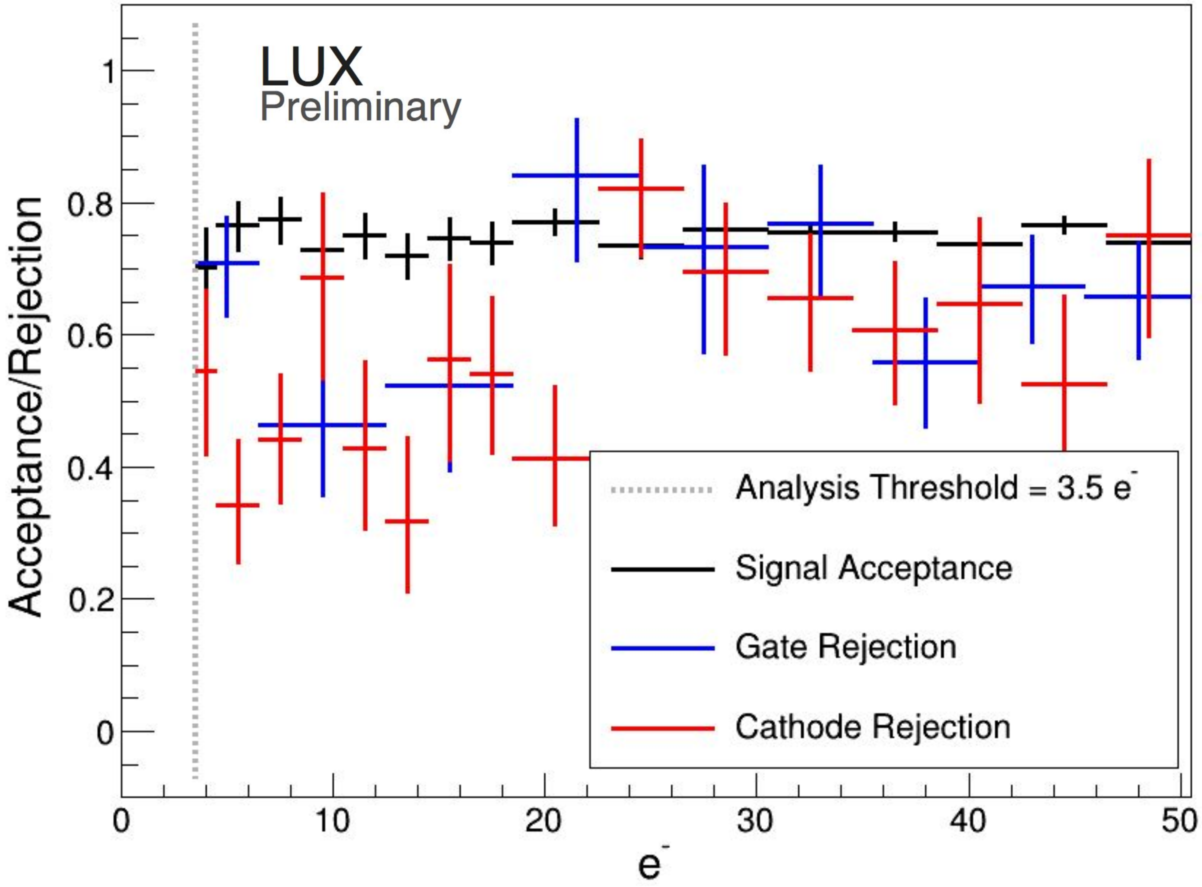
Kelsey C Oliver-Mallory

S2 spectrum, before re-weighting

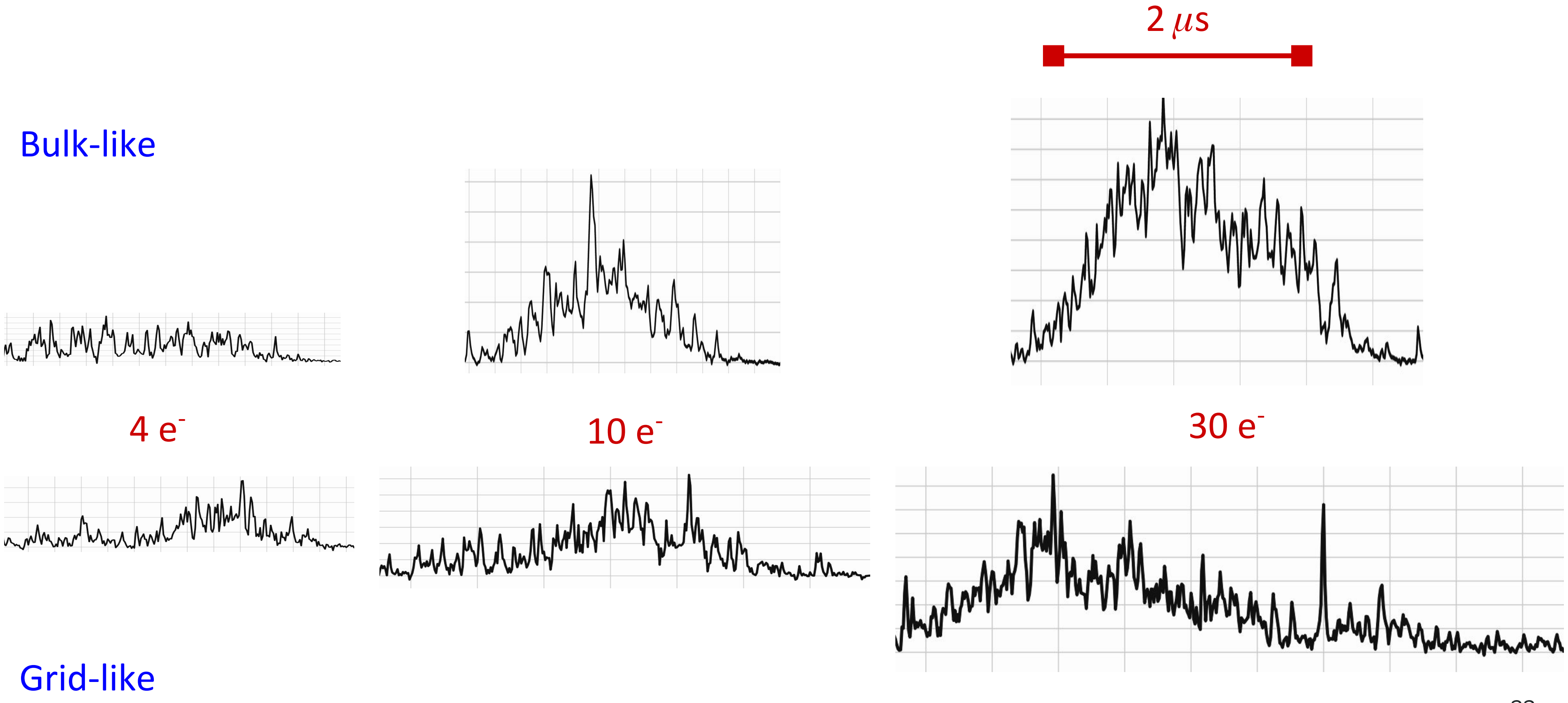
- Perfect opportunity to use machine learning
- Many parameters that quantify the S2 shape
- Accurate training/testing datasets



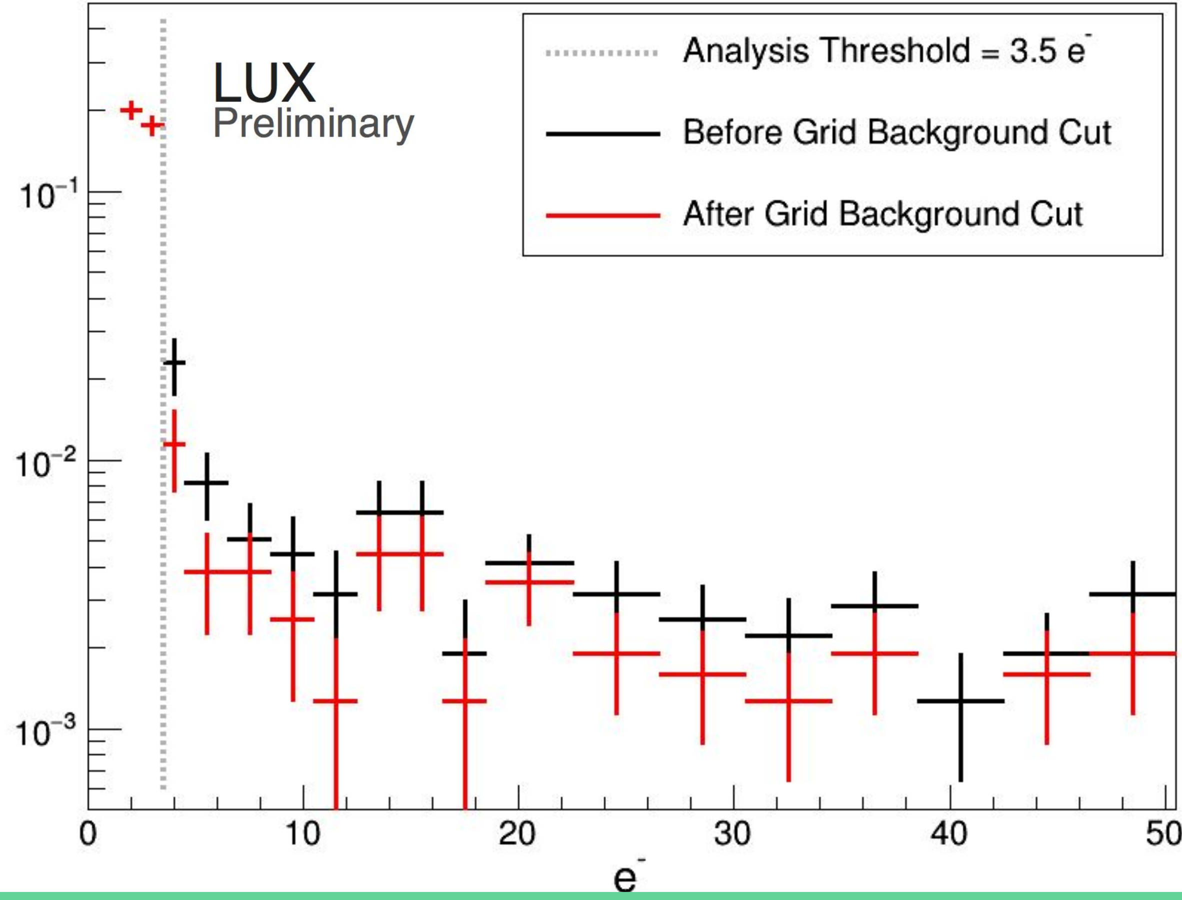
Acceptance/Rejection



Classification Results



Energy Spectrum of Background Data

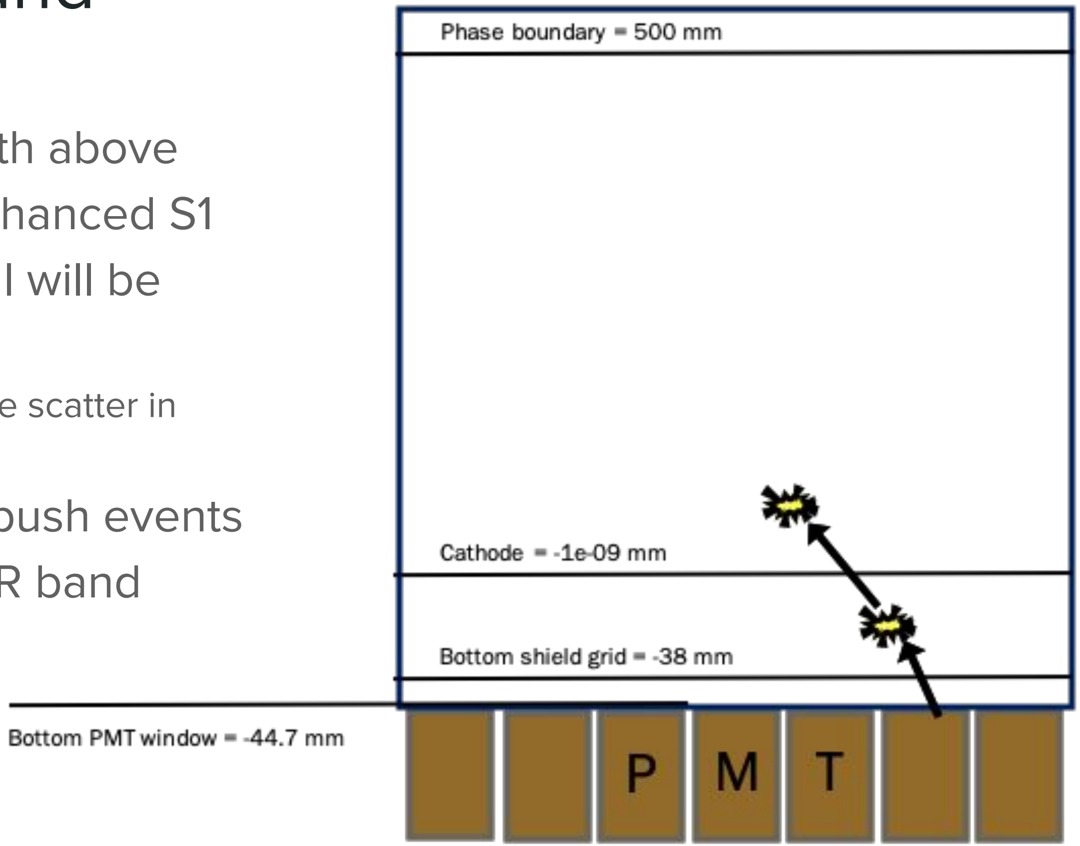


Gamma-X Rejection

Peter Rossiter

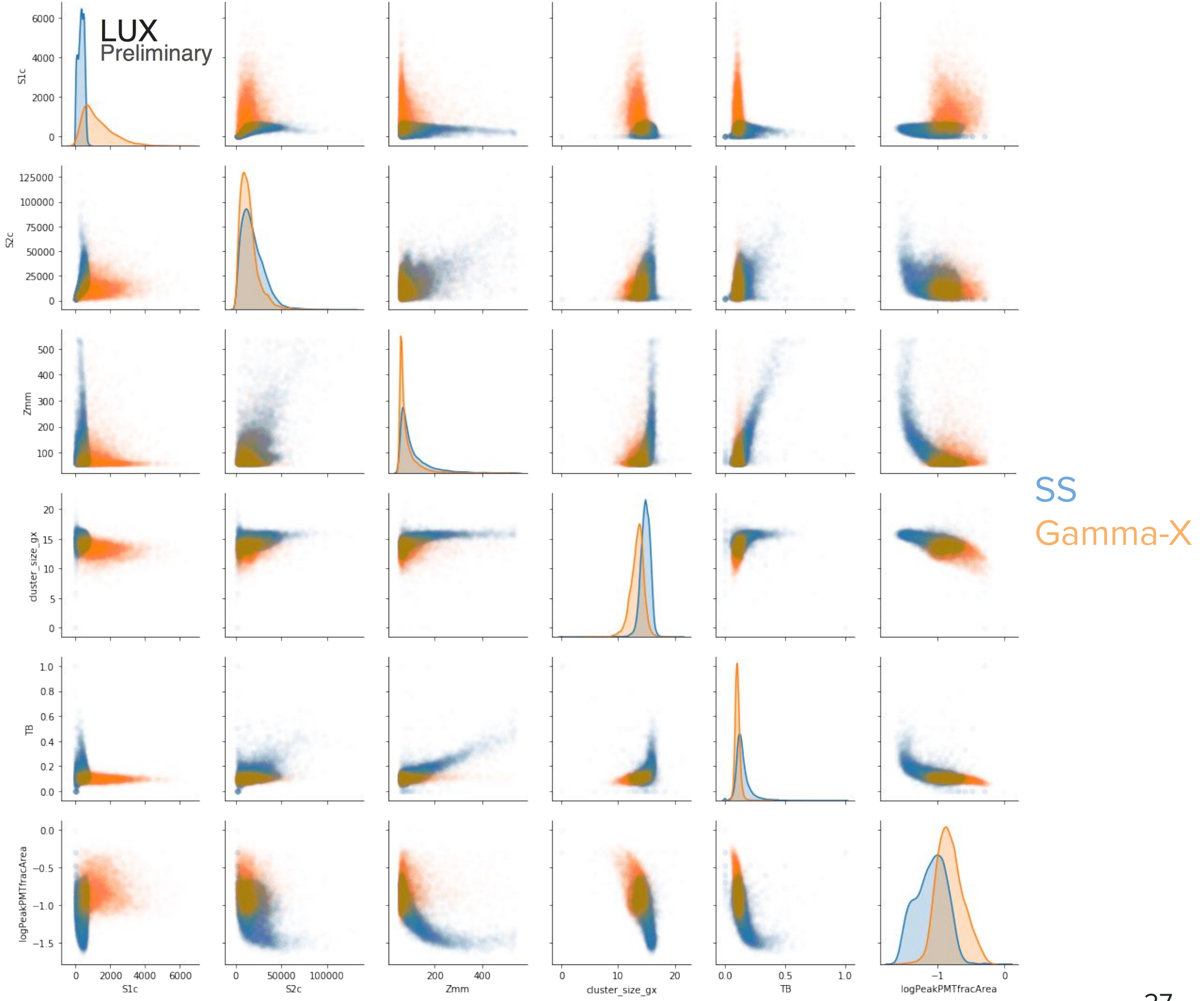
Overview of background

- If a particle scatters twice, both above and below the cathode an enhanced S1 signal relative to the S2 signal will be observed
  - Since only the S2 signal from the scatter in above the cathode is seen
- The reduced S2/S1 ratio can push events out of the ER band into the NR band



Analysis Variables

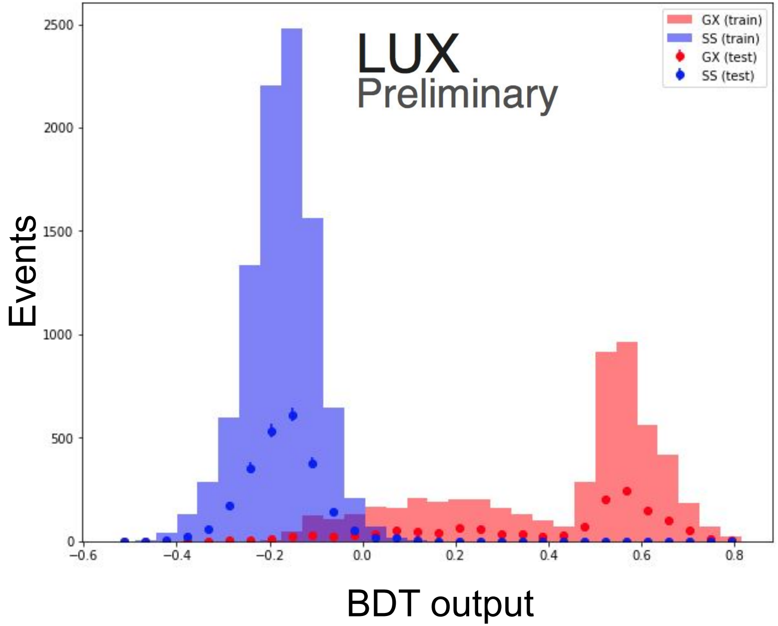
- S1 area
- S2 area
- Z position
- S1 hit pattern RMS
- Top-bottom asymmetry
- S1 max PMT area



Gamma-X BDT cut

- Combines several weak predictors into a single strong predictor
- Built to distinguish simulated gamma-X events from bottom PMT array from simulated SS events from bottom PMT array
- High degree of separation achieved

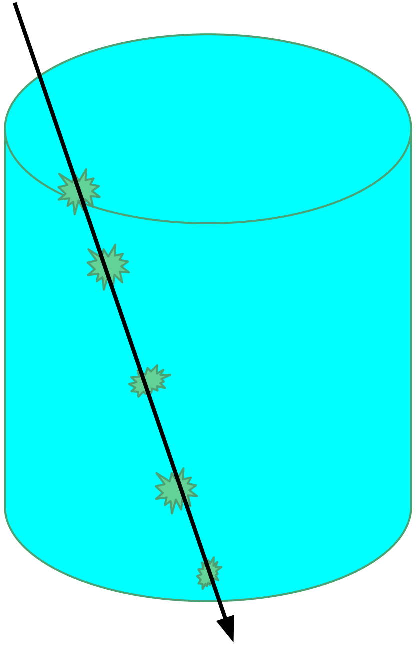
	Precision	Recall	F1-Score
Single Scatter	0.95	0.98	0.97
Gamma-X	0.97	0.92	0.94
Area under ROC curve for test data = 0.9782			



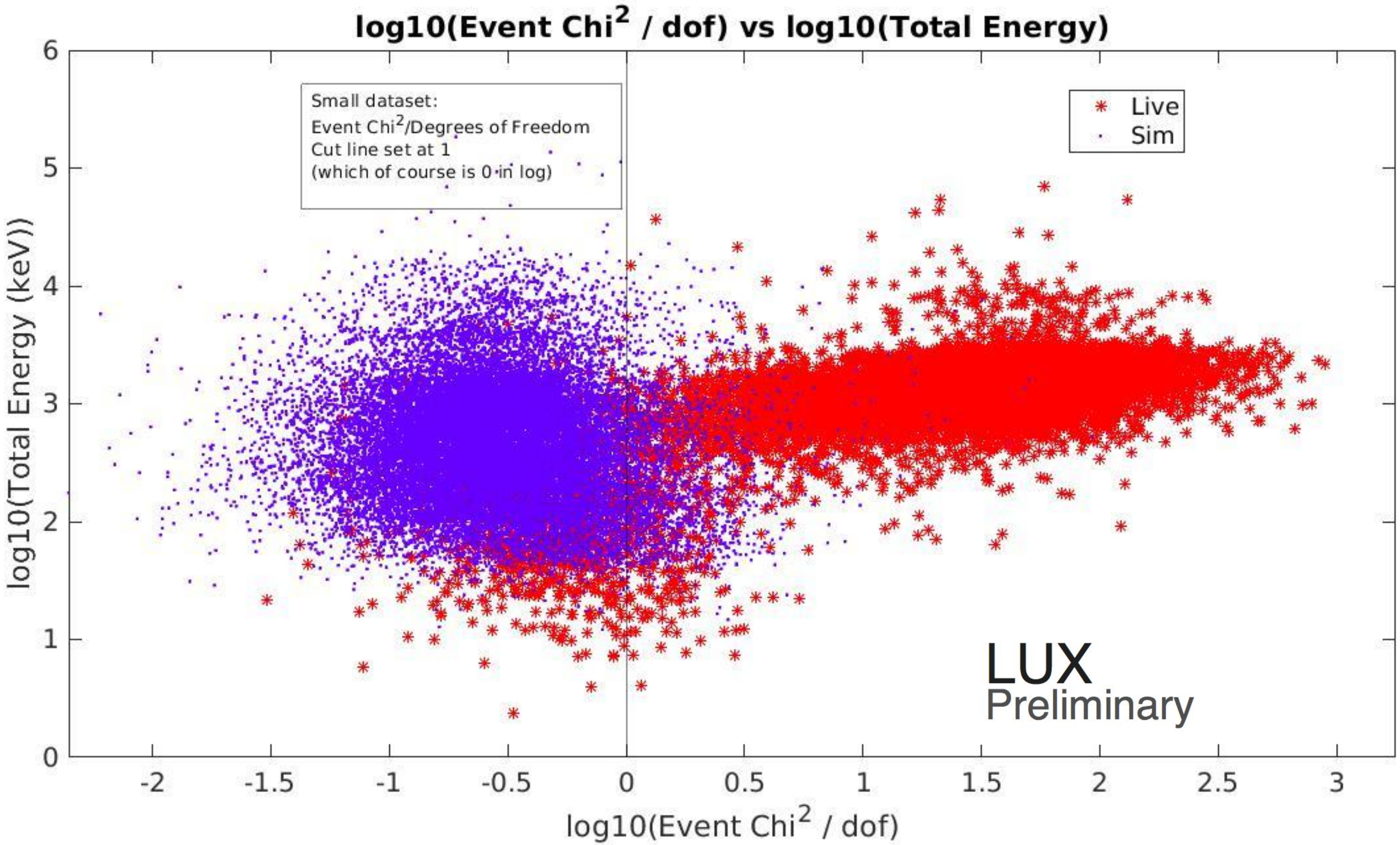
LIP Analysis

Paul Terman

Lightly-ionizing Particles



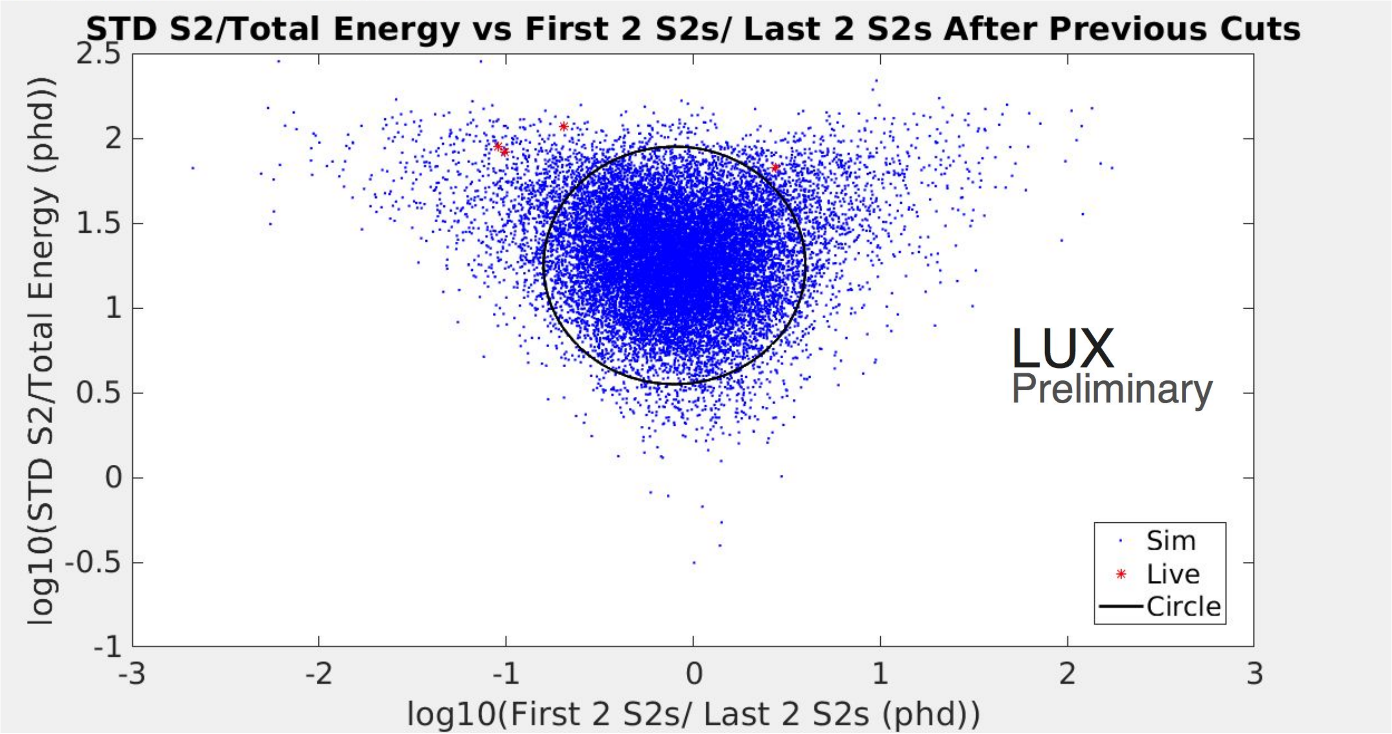
- **LIPs are hypothetical particles with fractional charge e*f, f<1**
- Cosmogenic - come from hemisphere above detector
- Signature is many hits along a line through the full detector
- Several variables developed for this search, including:
  - χ^2 of fit to a line (minimum of 5 scatters)
  - Angle of line from vertical
  - Track length over distance between first/last scatters
  - Standard deviation of pulse areas
  - ...
- Series of manual cuts considered
- **BDT automates and optimizes this process**
- Allows addition of further inputs without greatly affecting complexity



Fit multiple scatters to a line, check goodness of fit

Sim = simulated LIP signal

DOF = Number of fit points - 4
(4 is due to 3D line)



- Series of manual cuts considered
- **BDT automates and optimizes this process**
- Allows addition of further inputs without greatly affecting complexity

BDT Results

- Trained using:
  - LIP sims
  - Sample of bkg data
- Manual cuts and BDT both tuned for 0 background (training and testing)
- Efficiency of BDT noticeably exceeds that of manual cuts