Boosted Decision Tree Model for LEGEND

Henry Nachman

PIRE-GEMADARC Collaboration Meeting

Academica Sinica – Taipei, Taiwan

02 June 2023

Outline

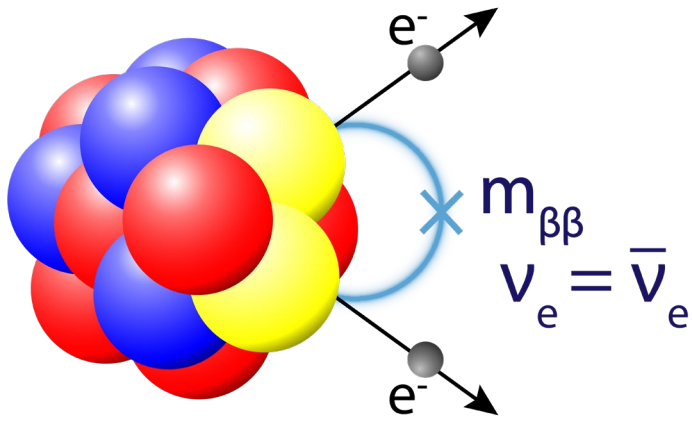
1. Brief Introduction –
   I. What is $0\nu\beta\beta$?
   II. What is LEGEND?
2. Parameter Extraction
   I. Waveform Selection
   II. Pulse Shape Parameter
3. BDT Model Training and Validation
4. Results, Discussion, and Conclusion

Neutrinoless Double Beta Decay - $0\nu\beta\beta$

$$(A, Z) \rightarrow (A, Z+2) + e^- + e^-$$

- Exchange of virtual light neutrinos results in no net neutrinos from this process
- Thus, net matter production

- If this process exists, it would be extremely rare – orders of magnitude longer than the age of the universe.

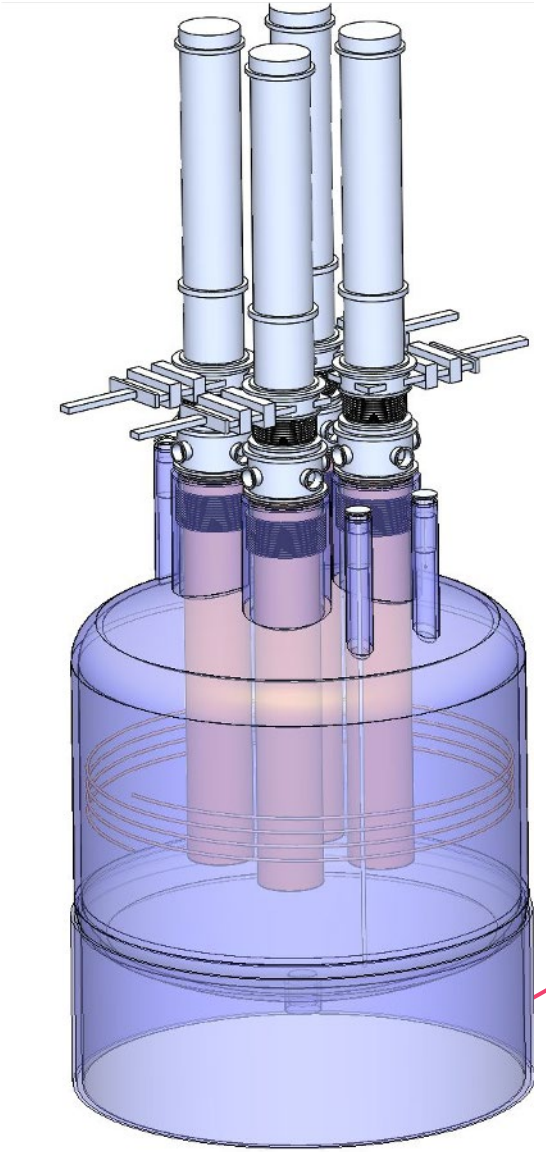


LEGEND

- LEGEND is a next generation collaborative experiment searching for $0\nu\beta\beta$ using high purity germanium detectors.

- Currently in the 200kg phase, LEGEND plans to use ~1000kg of ^{76}Ge to search for this elusive process.

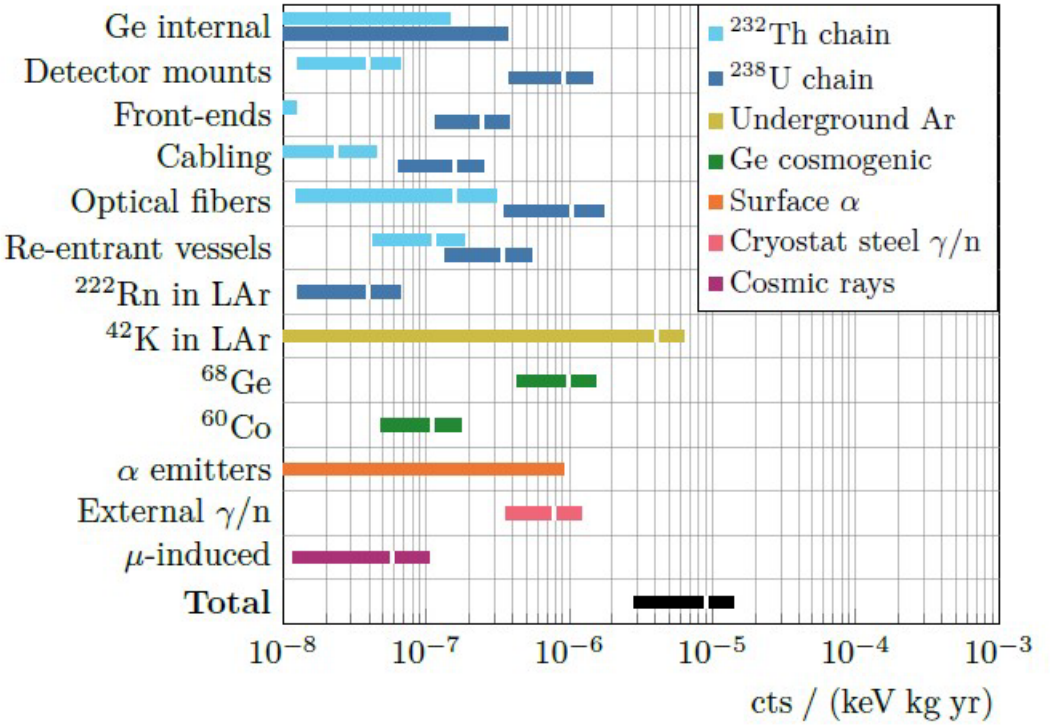
- Innovation for this high mass and low-background experiment will allow LEGEND to achieve a discovery potential at a half-life of 10^{28}years



Background in LEGEND-1000

- To achieve its discovery potential L-1000 strives for extremely low background.
- Only ~3 $0\nu\beta\beta$ signals would constitute a 3σ discovery

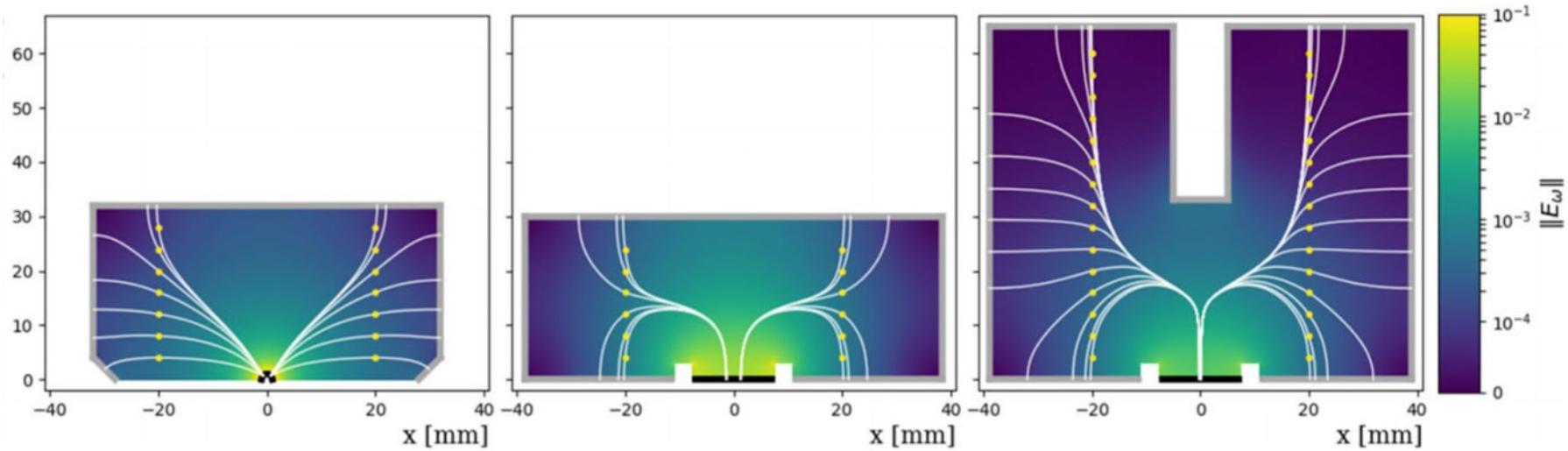
- Many background reduction methods:
  - Low background materials
  - Active veto – liquid argon scintillator
  - Analysis based background discrimination
    - Boosted Decision Tree (BDT) Model



Expected sources, and levels of backgrounds in L-1000

Inverted Coaxial Point Contact (ICPC) Detectors

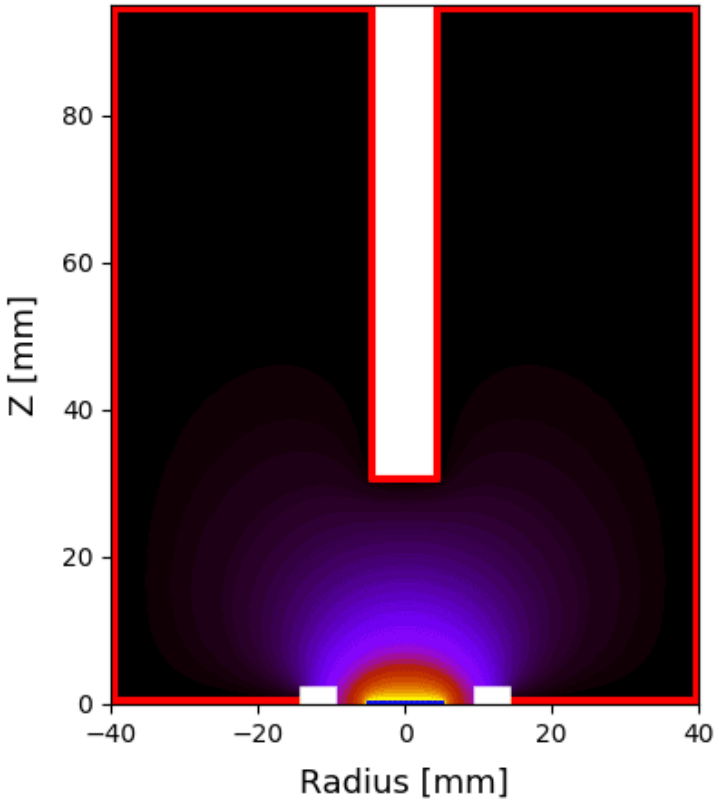
MJD (PPC) GERDA (BEGe) LEGEND (ICPC)



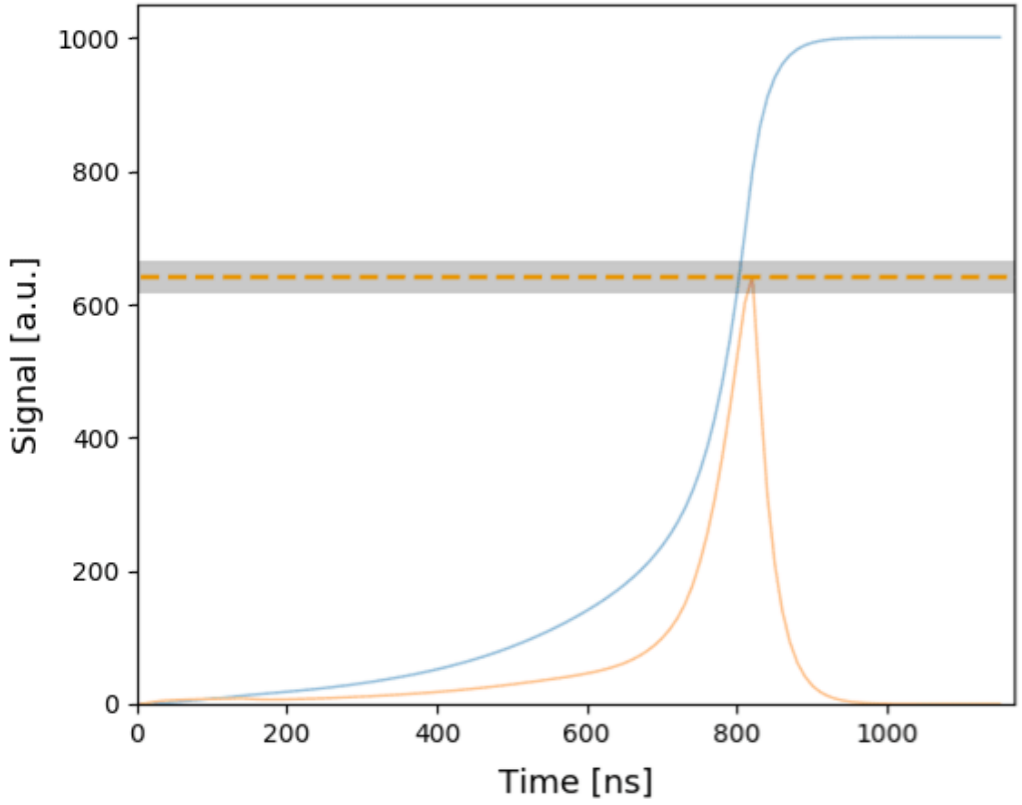
Larger size of ICPCs makes many pulse shape characteristic correlations (discussed later) more prominent

Single-Site Events

Weighting Potential and Charge Drift

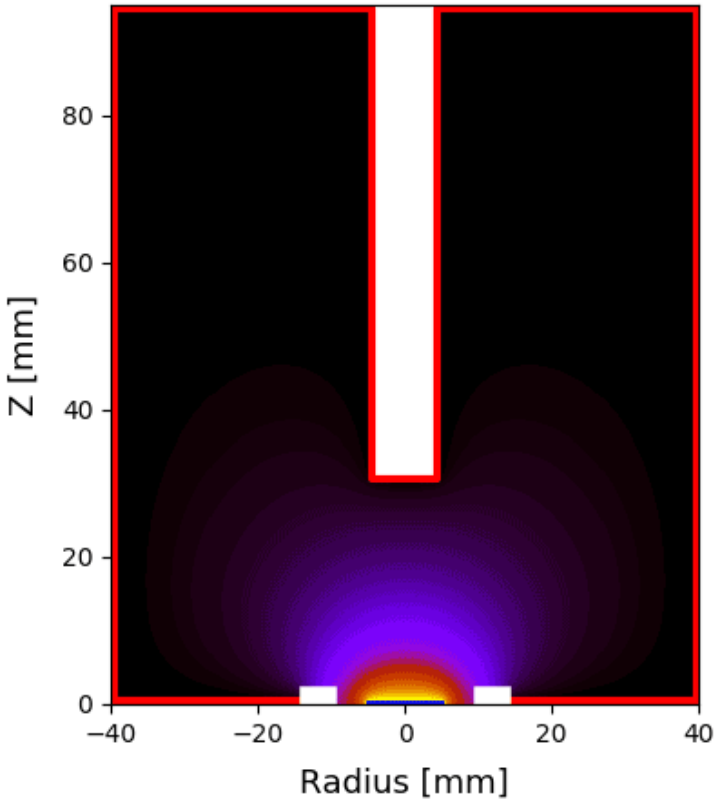


Generated Signal

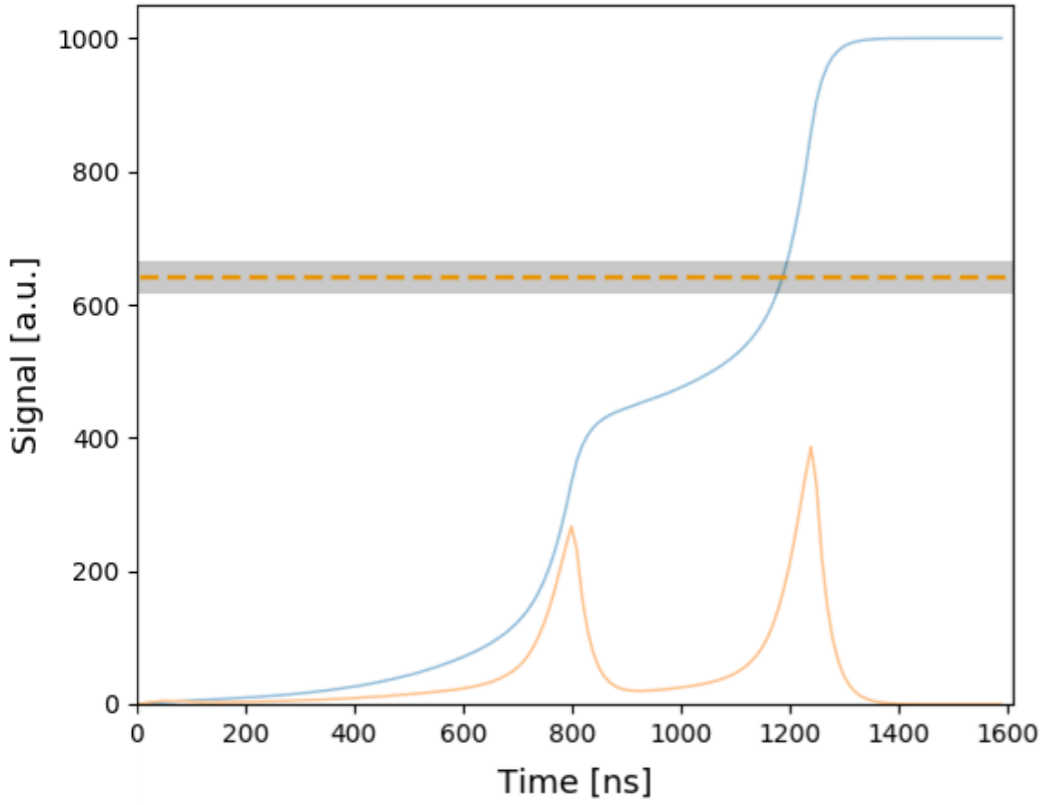


Multi-Site Events

Weighting Potential and Charge Drift



Generated Signal



Pipeline Goals -

1. Develop a Machine Learning (ML) software tool that can reject multisite events using only the raw waveform data.

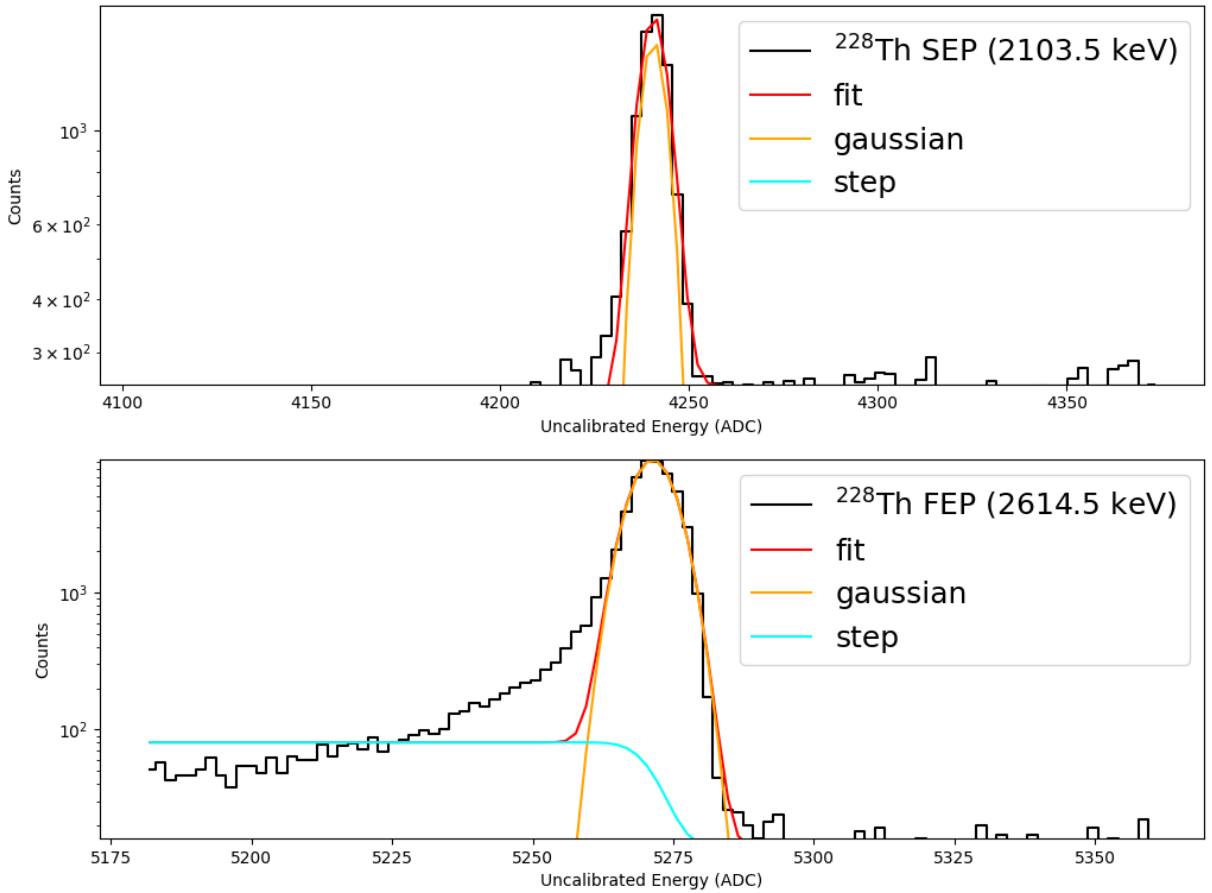
2. Create a fast (real time) analysis tool for efficient detector characterization analysis.

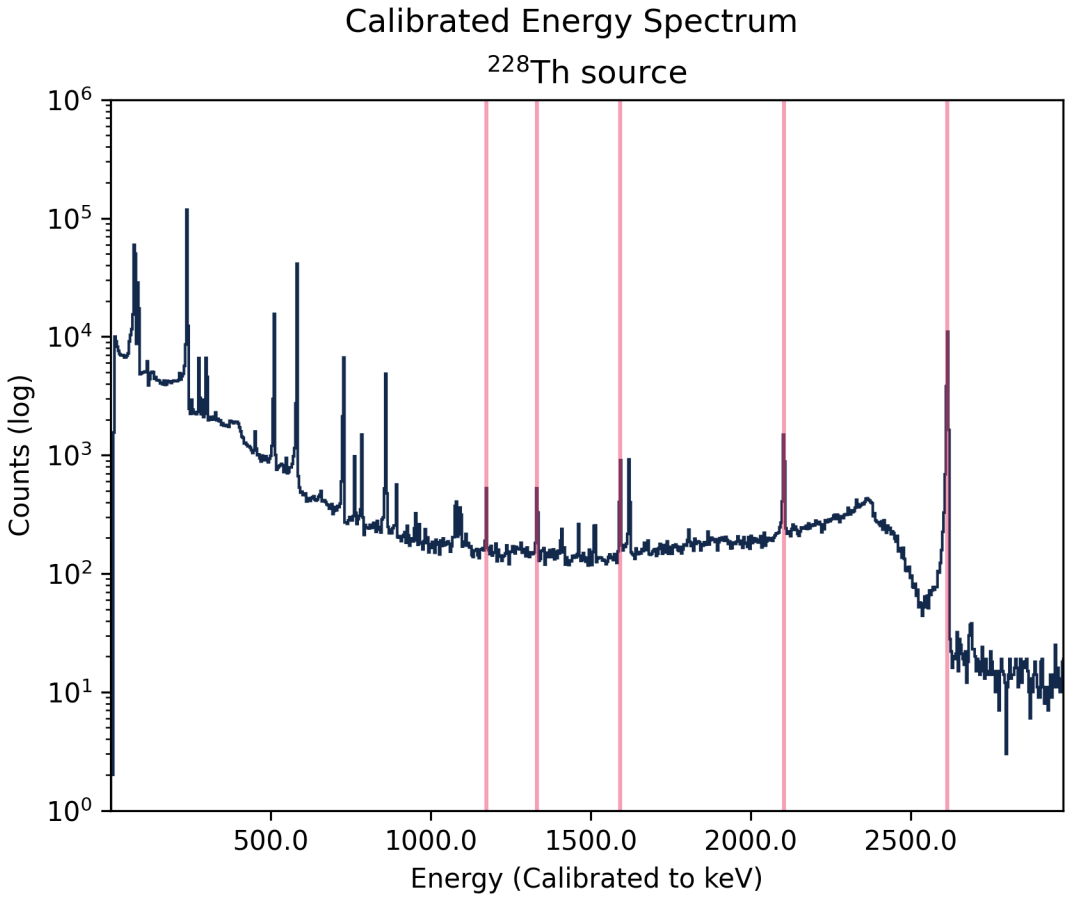
3. Explore and develop explainability tools to learn from and inform physics from the ML model.

Pipeline Steps -

1. Calibration and event selection
2. Pulse Shape Parameter extraction
3. Boosted Decision Tree (BDT) training and validation

Spectrum Calibration





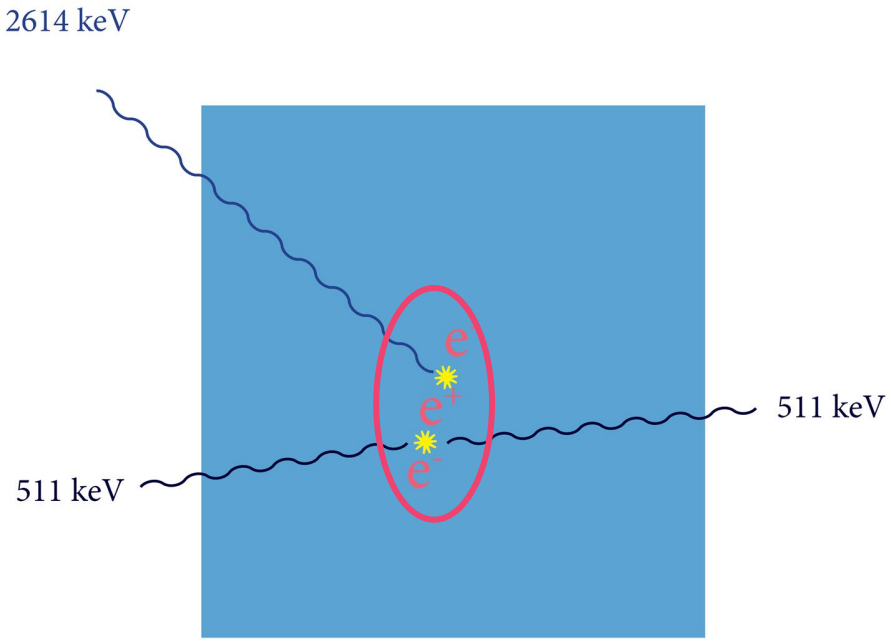
Proxy Events

- To train a model to classify events as single-site (SS) or multi-site (MS) we need large samples of each

- Of course - $0\nu\beta\beta$ signals are not exactly at our disposal.

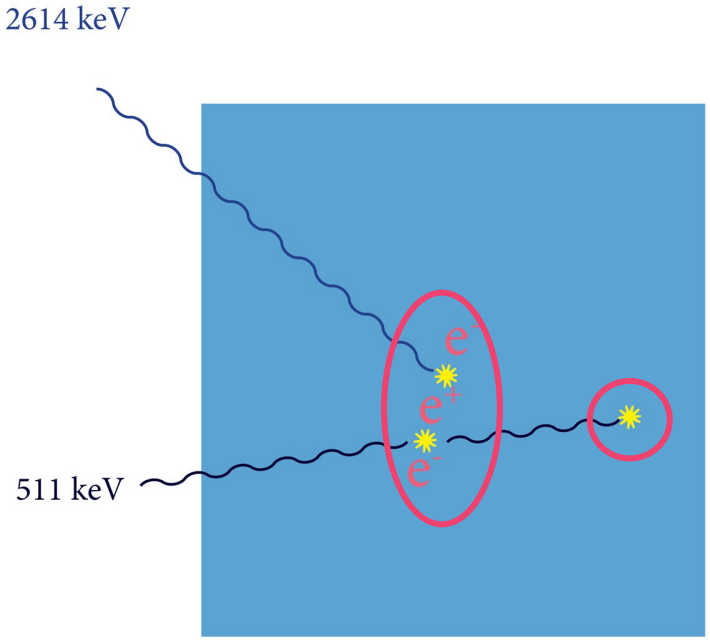
- Use data from the Th-228 calibration as proxy to represent the types of signals we would expect to see.
  - Double-Escape Peak (DEP) for Single-Site/$0\nu\beta\beta$-like
  - Single-Escape Peak (SEP) for Multi-Site/background-like

Single-Site Proxy Event



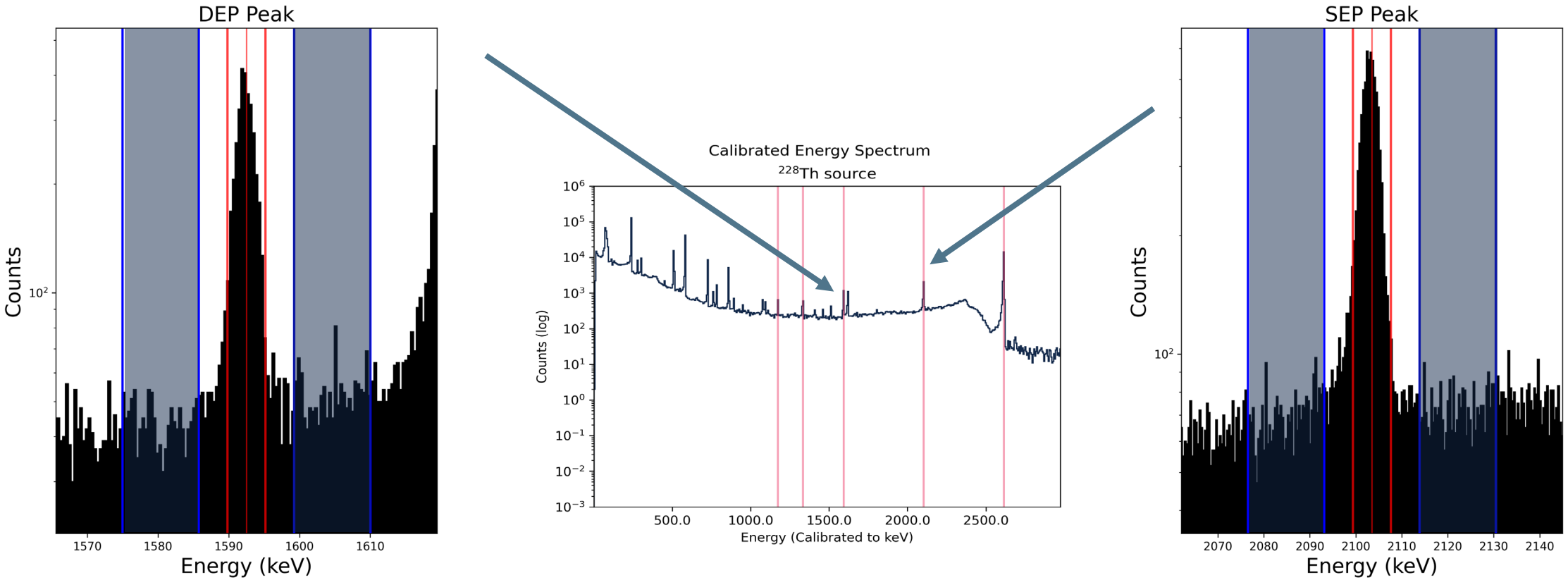
Double Escape
Total Collected Energy:
2614-511-511 = 1592 keV

Multi-Site Proxy Event



Single Escape
Total Collected Energy:
2614-511 = 2103 keV

Event Extraction

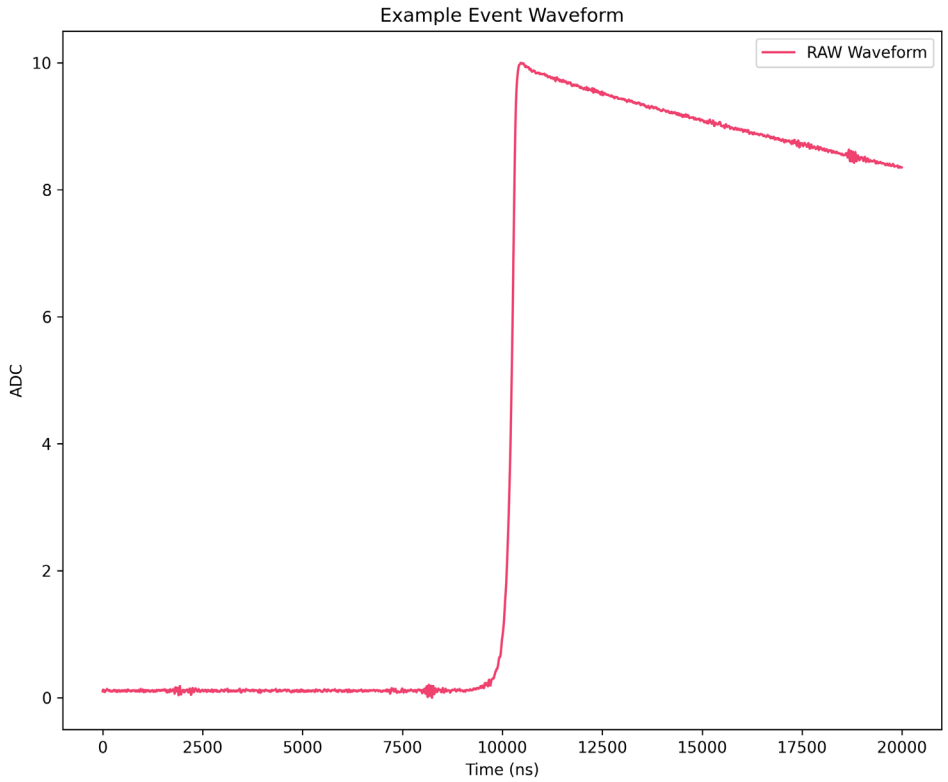


Pulse Shape Parameter

```
    Returns:
        - y_out: mean
    """
    [tau1, tau2, f] = popt
    y_out = np.zeros(wfArray.shape[0])
    max_amp = 10

    for wf in range(wfArray.shape[0]):
        wf_in = wfArray[wf,:]
        wf_in = (wf_in-wf_in.min())/(wf_in.max()-wf_in.min())
        wf_out = np.zeros(len(wf_in))
        # Defines the constant terms
        const1 = 1 / tau1
        const2 = 1 / tau2
        # Defines the exponential terms
        exp1 = np.exp(-1 / tau1)
        exp2 = np.exp(-1 / tau2)

        frac = f
        # Sets initial of output to same as input
        wf_out[0] = wf_in[0]
        e1 = e2 = wf_in[0]
        e3 = 0

        for i in range(1, len(wf_in), 1): # Iterates
            e1 += wf_in[i] - e2 + e2*const1
            e3 += wf_in[i] - e2 - e3*const2
            e2  = wf_in[i]
            wf_out[i] = e1 - frac*e3

        maxInd = np.argmax(wf_in) # Index of max
        pre_wf = wf_in[:maxInd] # waveform before ma

        peak = np.max(wf_in)
        peakInd = wf_in.argmax()
        val97 = 0.97*peak
        diffArr = np.absolute(wf_in - val97)
        closestFromPeak = diffArr[peakInd:].argmin()
        ind97 = peakInd+closestFromPeak
        y_out[wf] = (np.std(wf_out[(ind97):]))

    return np.mean(y_out)

##################################################

def dp0Vis(popt, wfArray):
    [tau1, tau2, f] = popt
    wfInAdj = np.zeros(wfArray.shape)
    wfCorr = np.zeros(wfArray.shape)
    max_amp = 10
    trash = []
    for wf in tqdm(range(wfArr
                         desc="
                         col
        wf_in = wf
        if wf
```

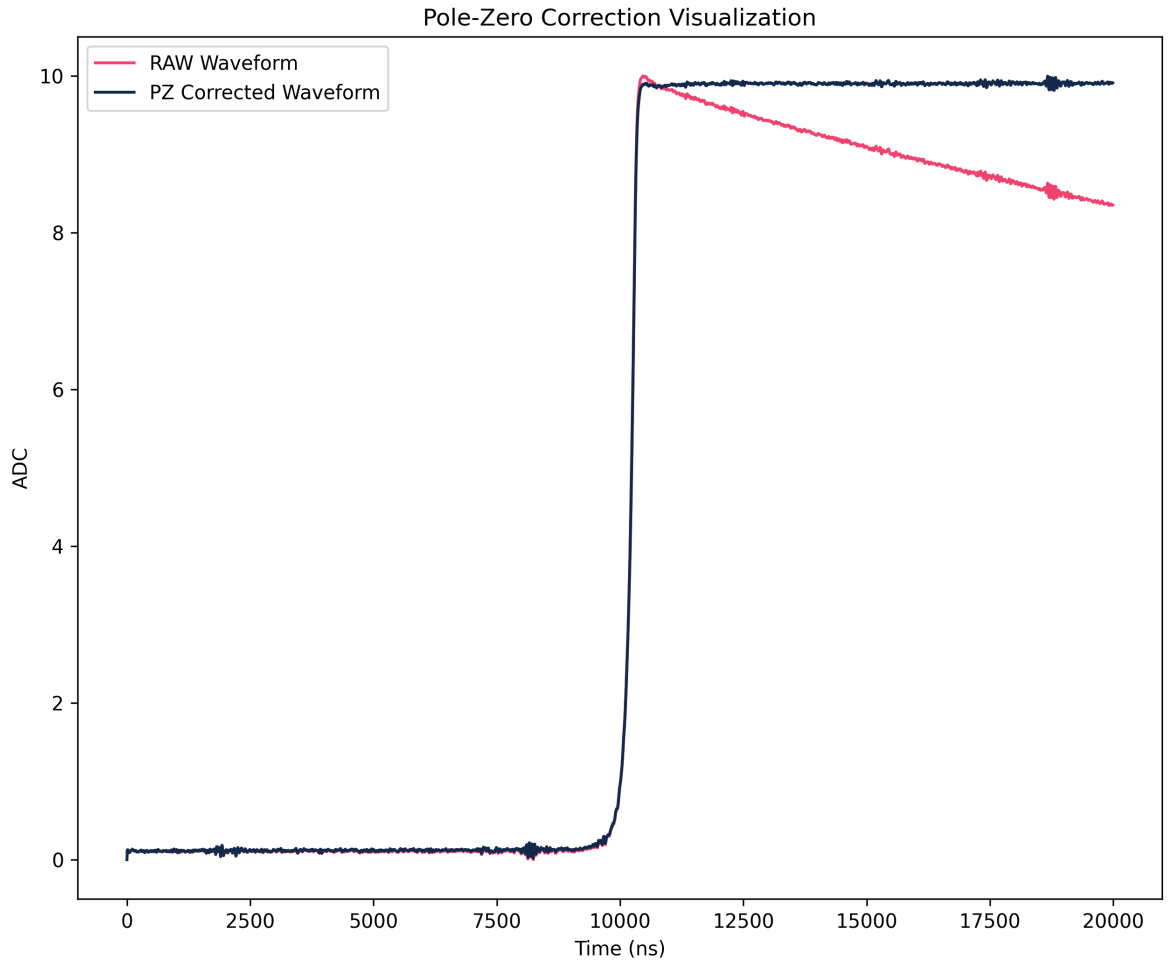
What are Pulse Shape Parameters?

- Lots of information about a detection event can be gleamed from the shape of the event waveform.
- Many of these parameters were developed for analysis of previous Ge detector experiments, namely the MAJORANA DEMONSTRATOR.
- PSD Extracted in this pipeline:
  - Current Amplitude ⊟ Energy (A/E)
  - Delayed Charge Recovery (DCR)
  - Late Charge (LQ80)
  - Drift Time 10%, 50%, 99.9% (TDRIFT)

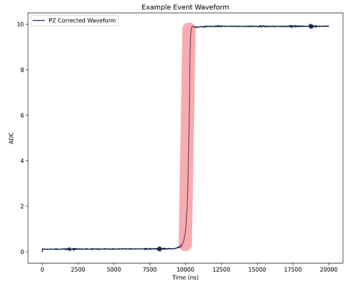


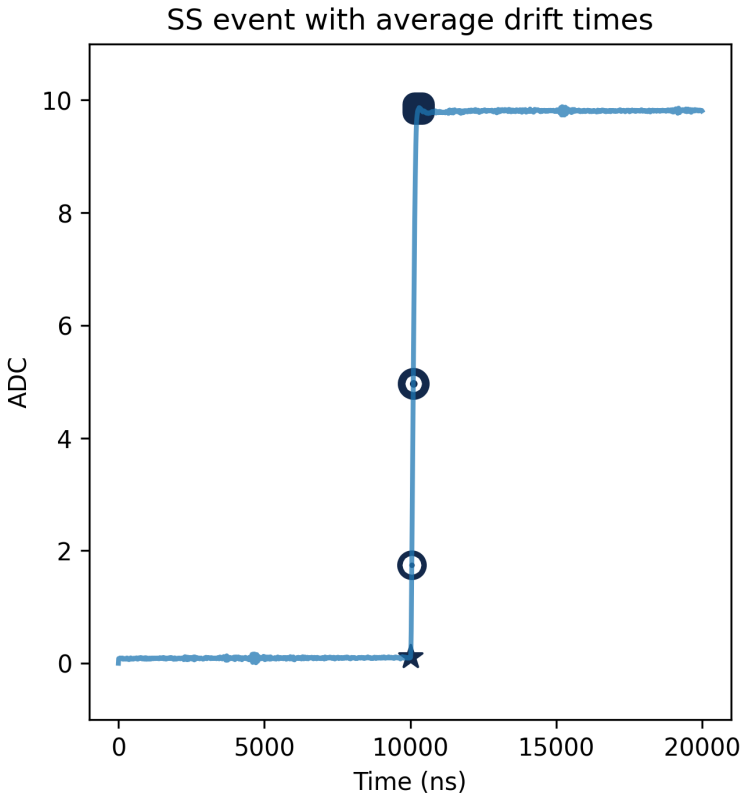
Pole-Zero Correction

- Raw waveforms exhibit an RC decay in the tail from the electronic circuit used to read out charge from the detector
- Can be fit with a two-term exponential decay function and corrected

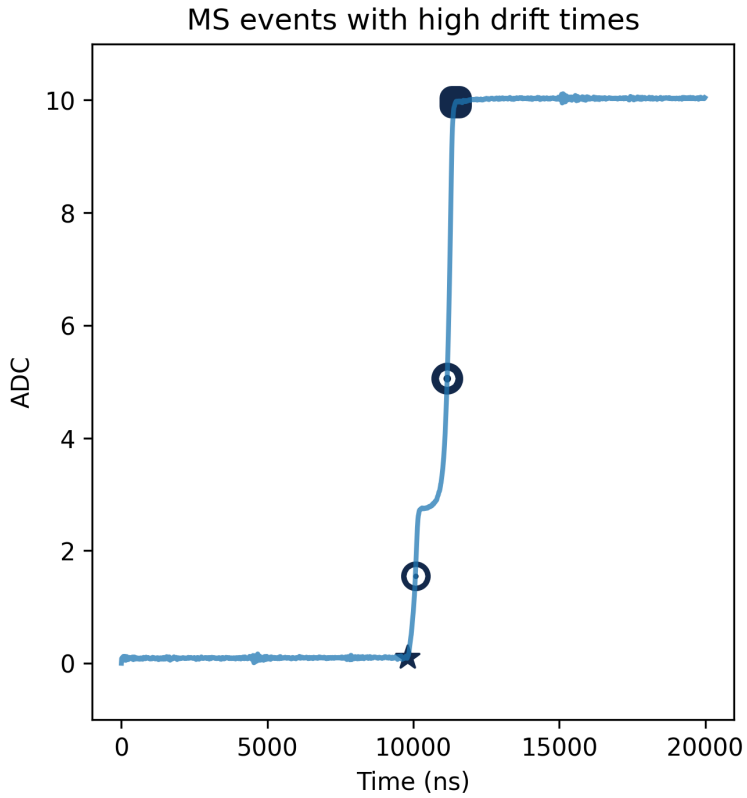


Drift Time

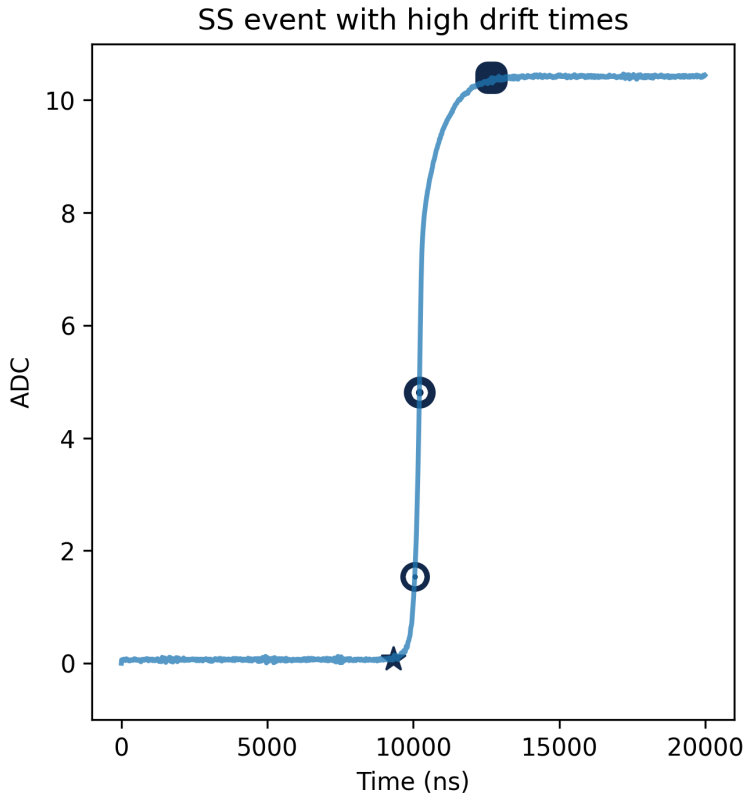




Start of Rise
tdrift (99.9%) -- 281.4 ns
tdrift50 (50%) -- 90.0 ns
tdrift10 (10%) -- 37.1 ns

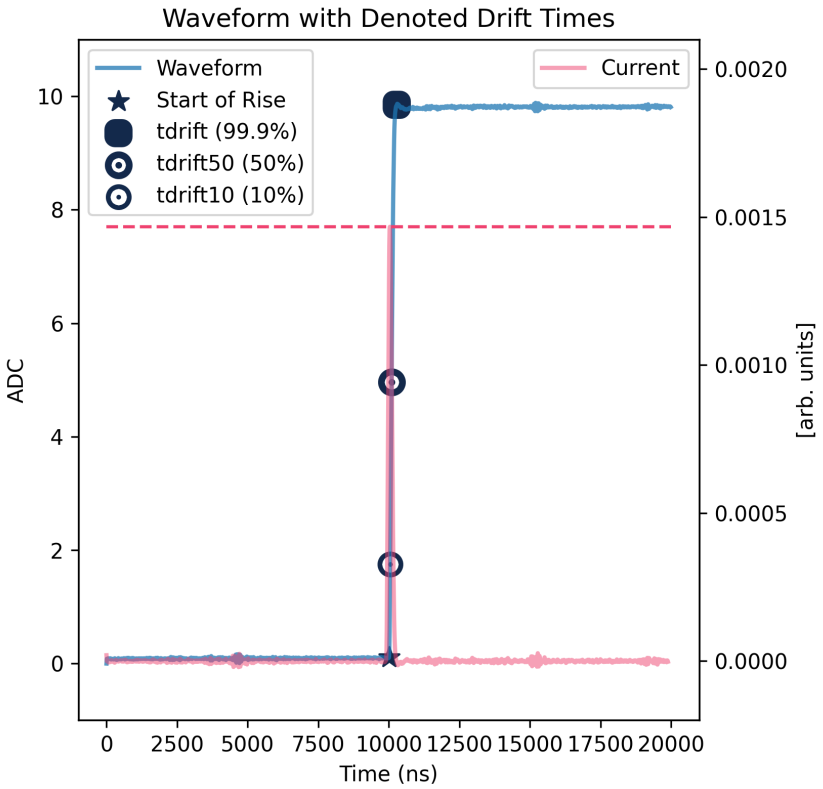


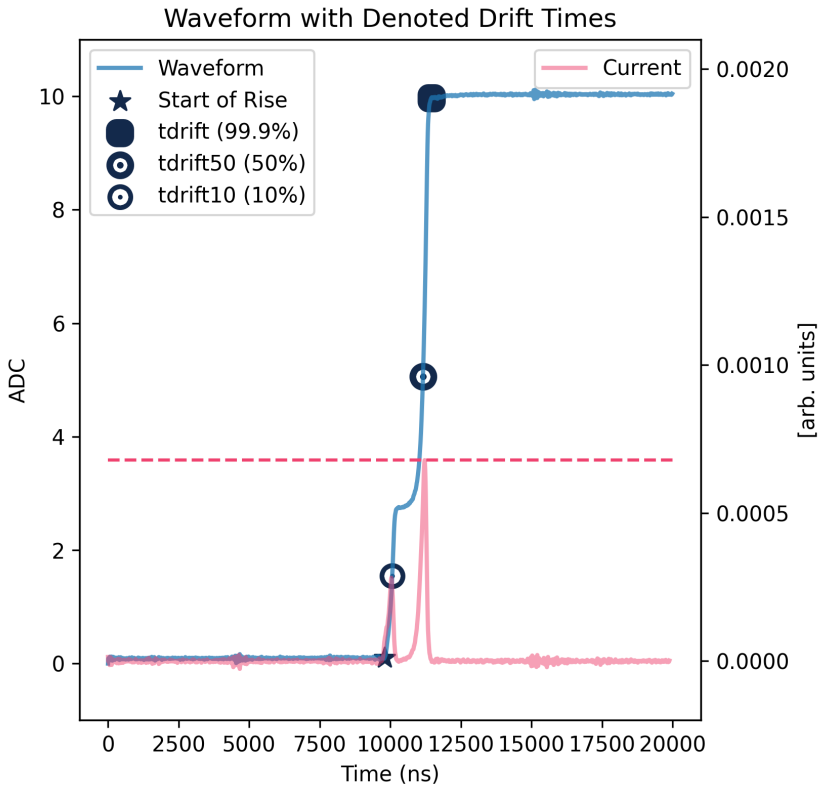
Start of Rise
tdrift (99.9%) -- 1661.4 ns
tdrift50 (50%) -- 1359.8 ns
tdrift10 (10%) -- 271.4 ns

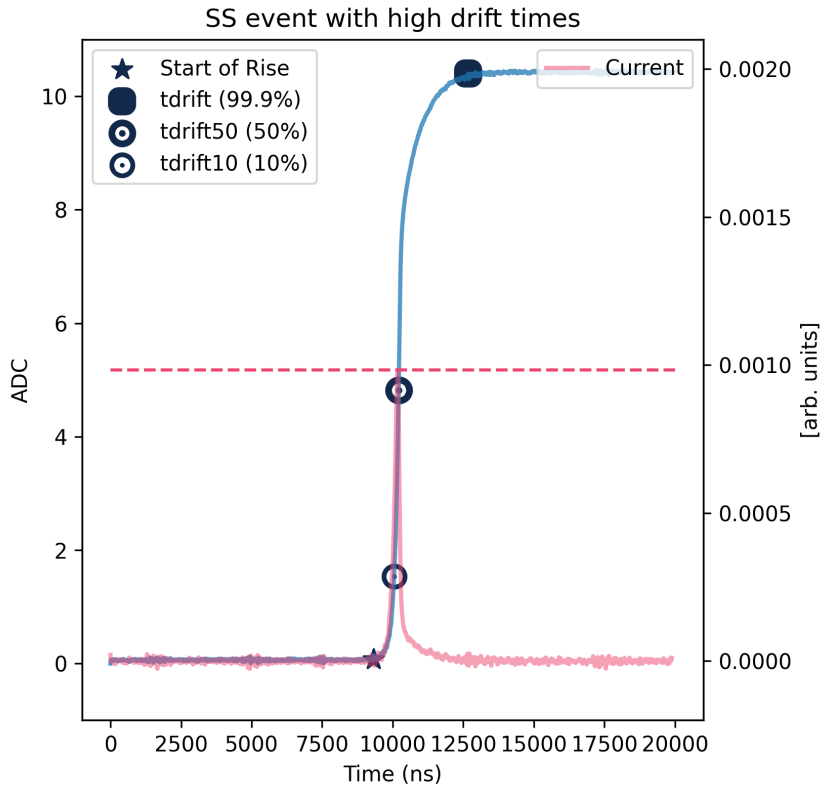


Start of Rise
tdrift (99.9%) -- 313.0 ns
tdrift50 (50%) -- 109.4 ns
tdrift10 (10%) -- 39.3 ns

Current Amplitude

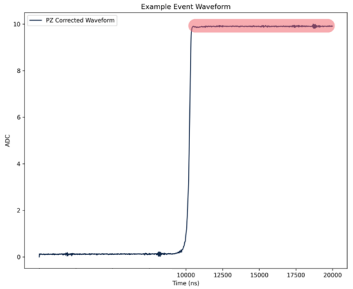


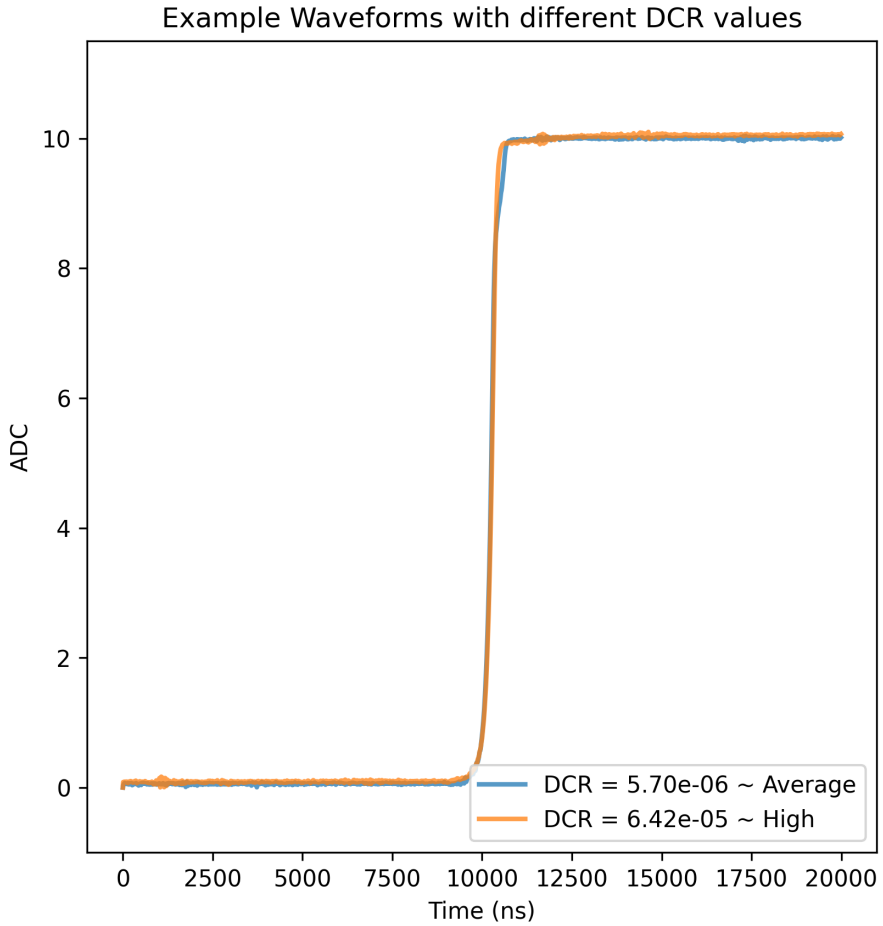


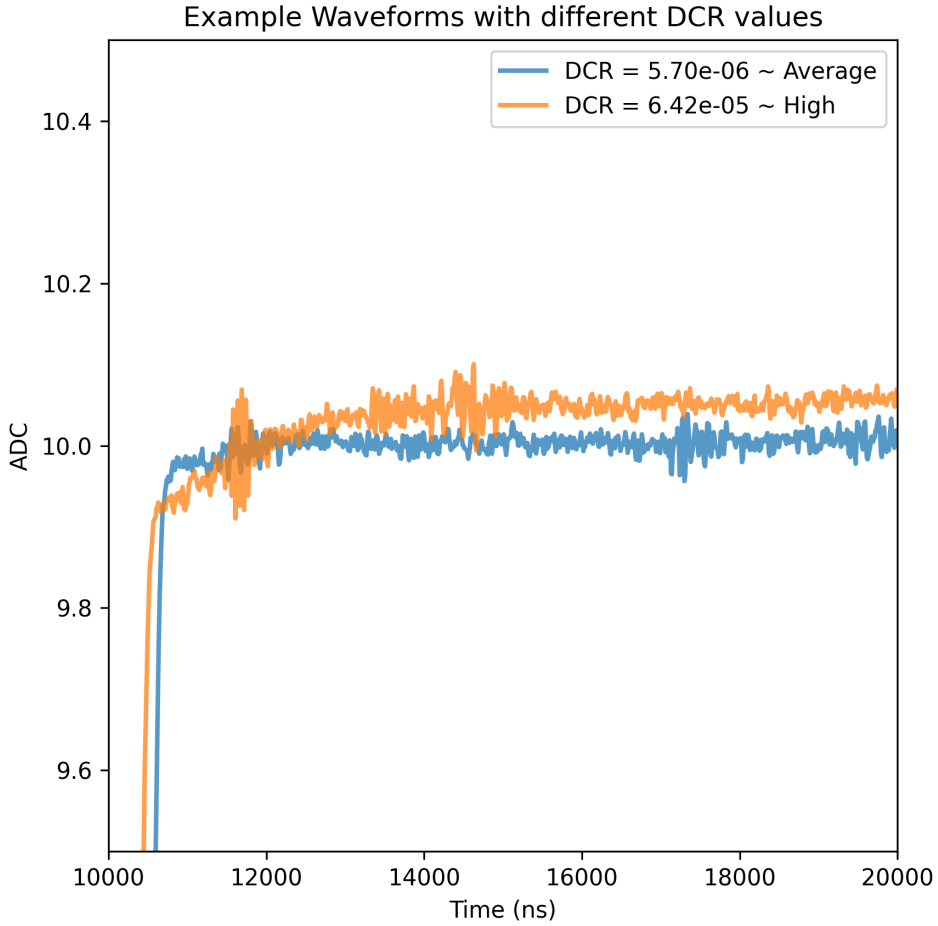


- Divide by DAQ energy to get : A/E

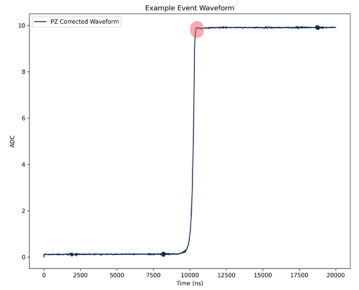
Tail Slope – Delayed Charge Recovery



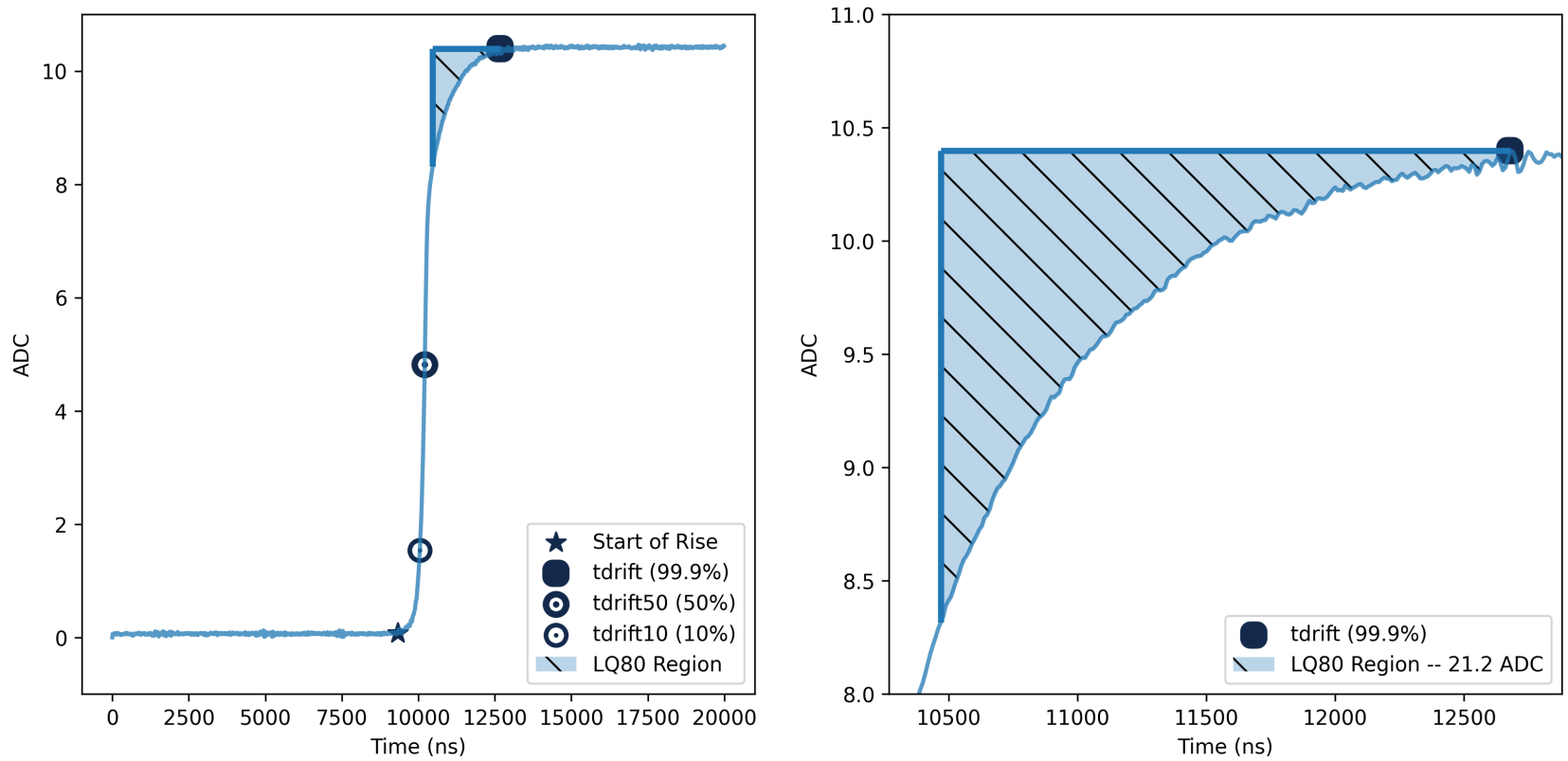




Late Charge



High LQ Event



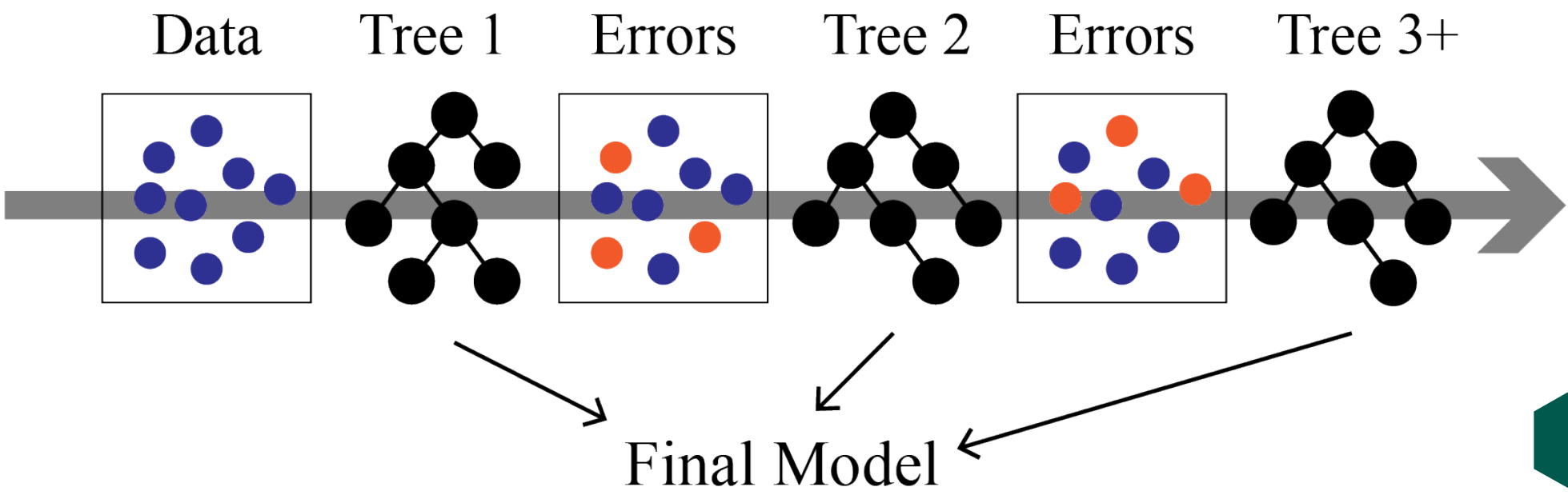
Parameter Correlations

- Traditional Analysis:
  - Each pulse shape parameter is considered independently with little to no consideration for multi-parameter correlation.
  - Requires fine-tuning of each individual parameter (time consuming).

- Machine Learning Analysis:
  - Simultaneously analyzes all parameters taking into account their correlations.
  - Does not require fine-tuning of individual parameters (fast analysis, ~30 sec training time)

- Motivates the use of a Boosted Decision Tree machine learning model which can train on many different parameters simultaneously.

Boosted Decision Tree



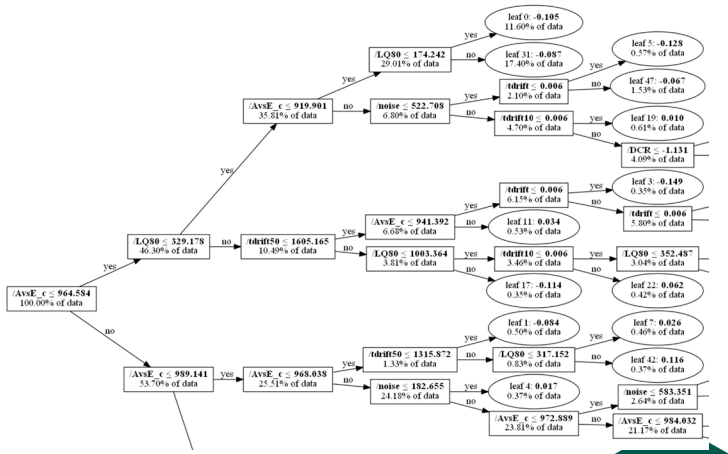
What is a Boosted Decision Tree Model

- Supervised machine learning model –
  - Machine Learning – Computer algorithms that learn from large datasets to fit models to observed patterns.
  - Supervised = labeled. Essentially guiding the machine to specified classifications

- Decision Tree
  - Network of Boolean (T/F) decisions through which events are classified

- Boosting –
  - Growing multiple trees in a series, each fitting to the residuals of the previous.

Fig: Screenshot of part of one trained tree.



Distribution Matching

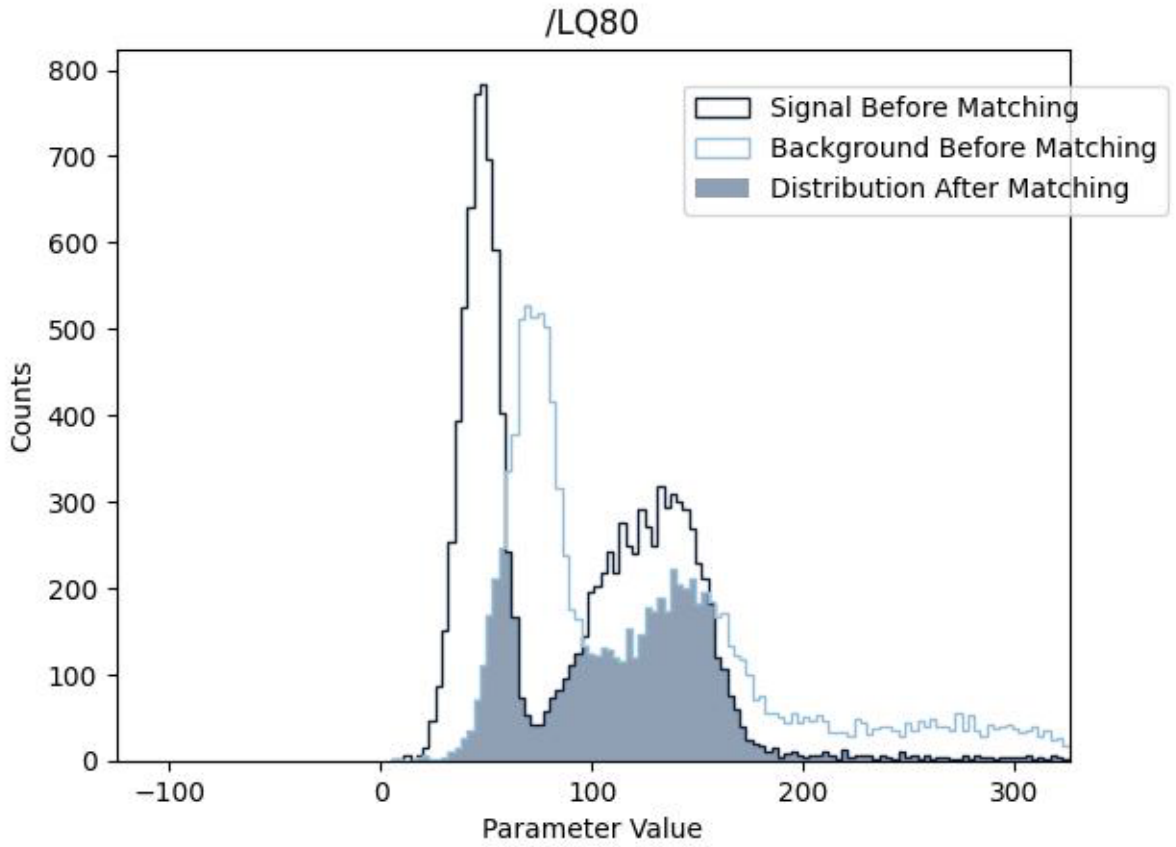
- The DEP and SEP datasets have different distributions of parameters which need to be remedied to avoid unintentional bias:
  1. **DATASET SIZE** – if the SEP (Multi-Site) has many more events than DEP, the model will have a statistical bias towards classifying things as multi-site.
  2. Need to remove **ENERGY DEPENDENCE** from parameters. Our datasets come from two peaks of different energy – not a property of single-site and multi-site events. Otherwise, model will simply train on energy and will be unrealistically good at classifying events.
  3. Encourages the model to investigate the correlations between parameters, rather than just differences in distributions of single parameters.

How to Match Distributions

1. Bin the data from both sets (single-site & multi-site)
2. Keep only the data that is consistent between both sets

- Drastically cuts the totally number of events available for training.

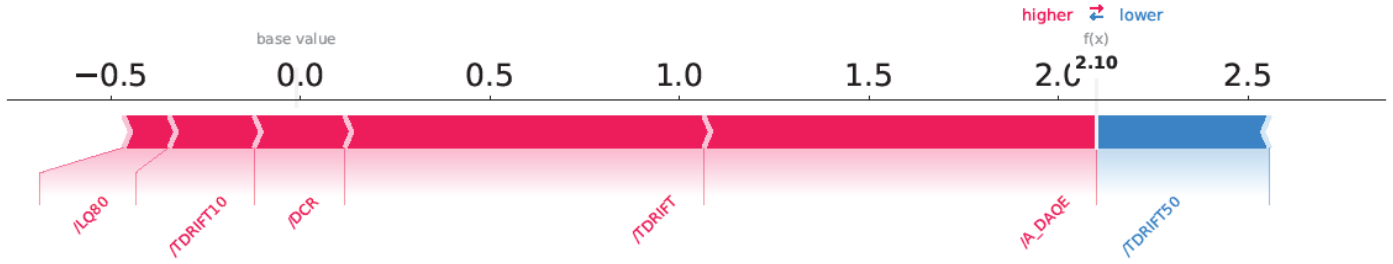
- Don't want to match A/E parameter as dividing by E should remove Energy dependence and other distribution differences are likely signatures of multi-sitedness



Model Explainability - SHAP

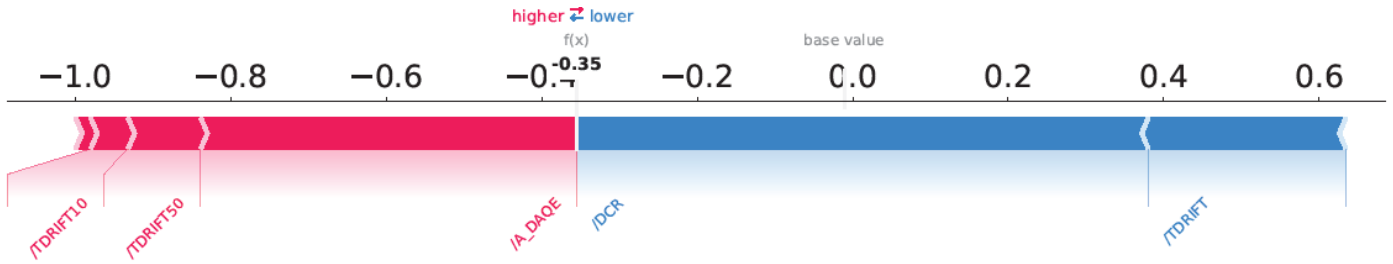
- Explain how the machine learned from the data – rather than blindly trusting the results
- SHAP – SHapley Additive exPlanations – game theory principles to assign each parameter a value based on its influence over the model's classification, for each event.

Single-site



Multi-site

Single-site



Multi-site

External Dependencies

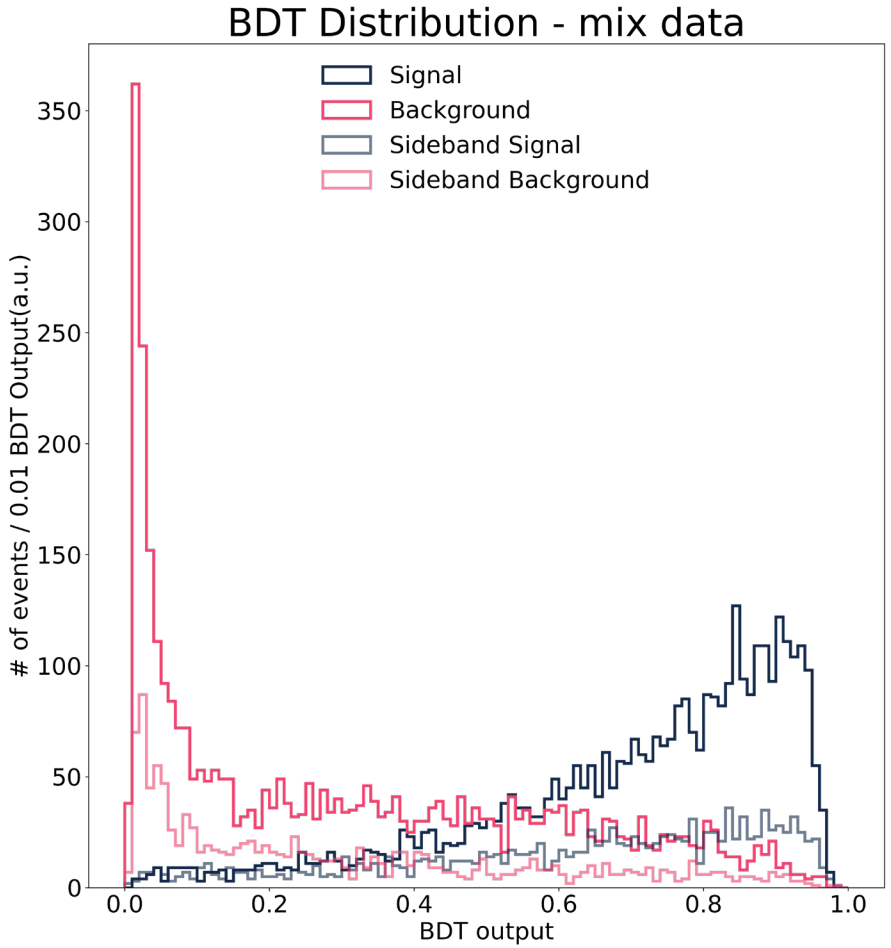
My pipeline makes use of external software frameworks including:

- PYGAMA – for peak fitting
  - Developed by LEGEND collaboration
- LightGBM – for model training
  - Guolin Ke et al.
- SHAP – for Shapley explainability study
  - Scott Lundberg, Su-In Lee

BDT Distribution

Mix Data	Pre-Matching	Post-Matching
SEP (Multi-Site)	35420	3899
DEP (Single-Site)	14730	3899
Trees	47	

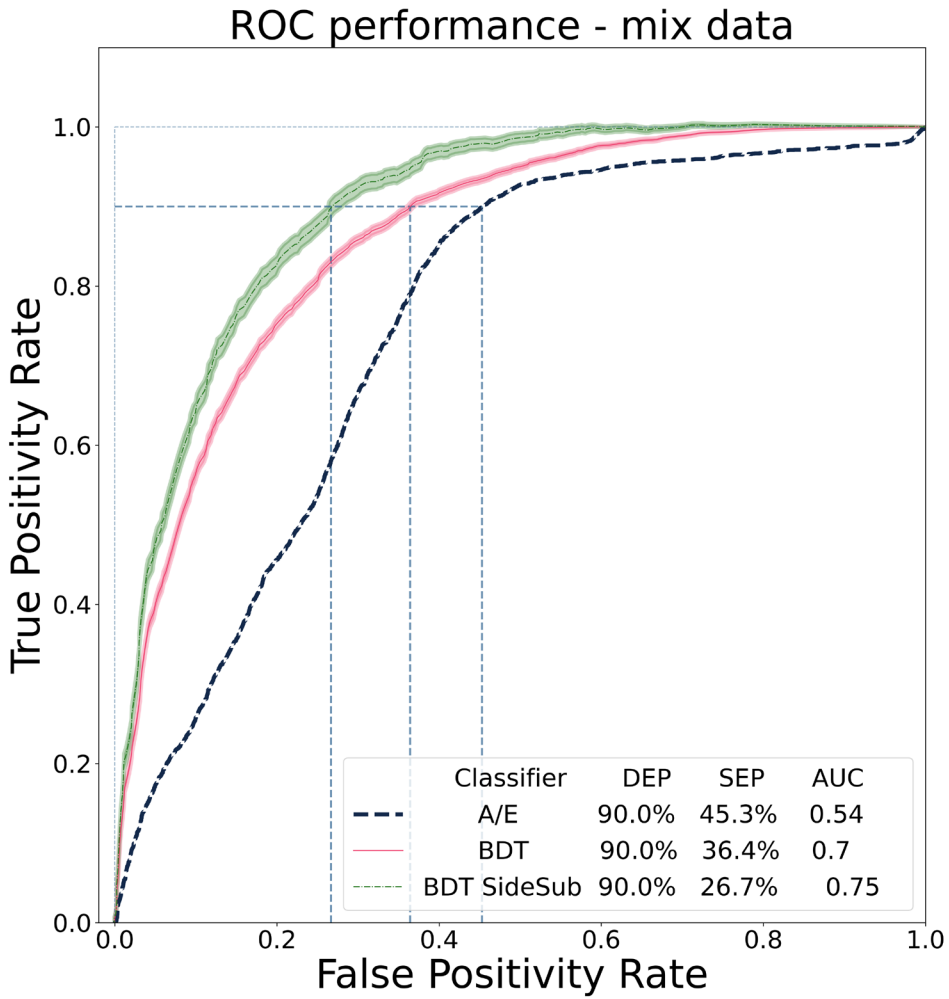
- Model validated on 30% of training data.



ROC – mix data

- Measure of binary classifier – Perfect classifier has an AUC = 1

- Note: the A/E measure shown here is without drift time correction or other methods used in the traditional LEGEND analysis

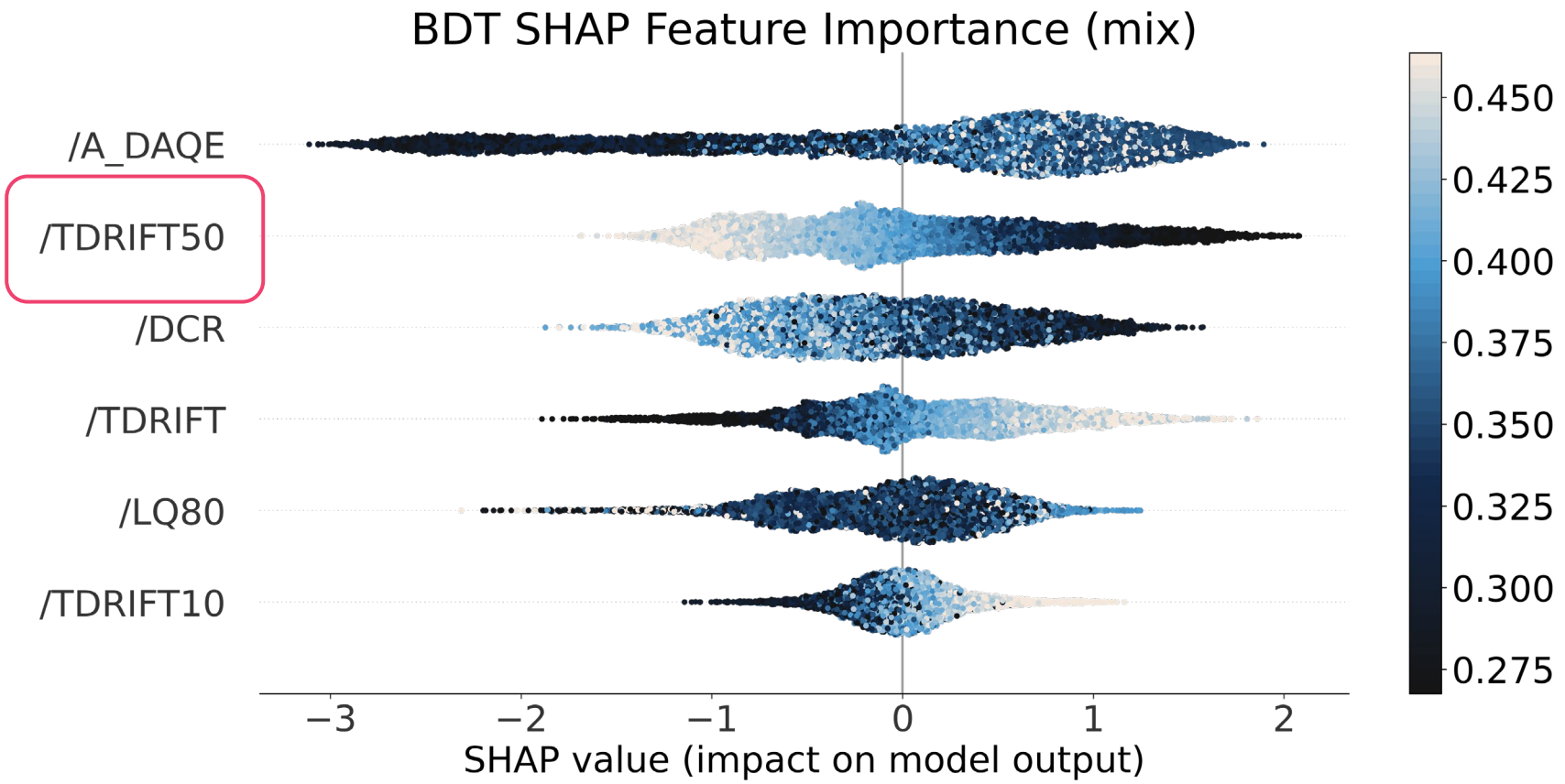


SHAP Feature Importance - Mix

- Made by plotting Shapely values for every event.
- Color of point represents the ultimate classification.

- "Important" features have a dumbbell shape

- Notice the anti-correlation between
  - TDRIFT50:A_DAQE
  - DCR:A_DAQE



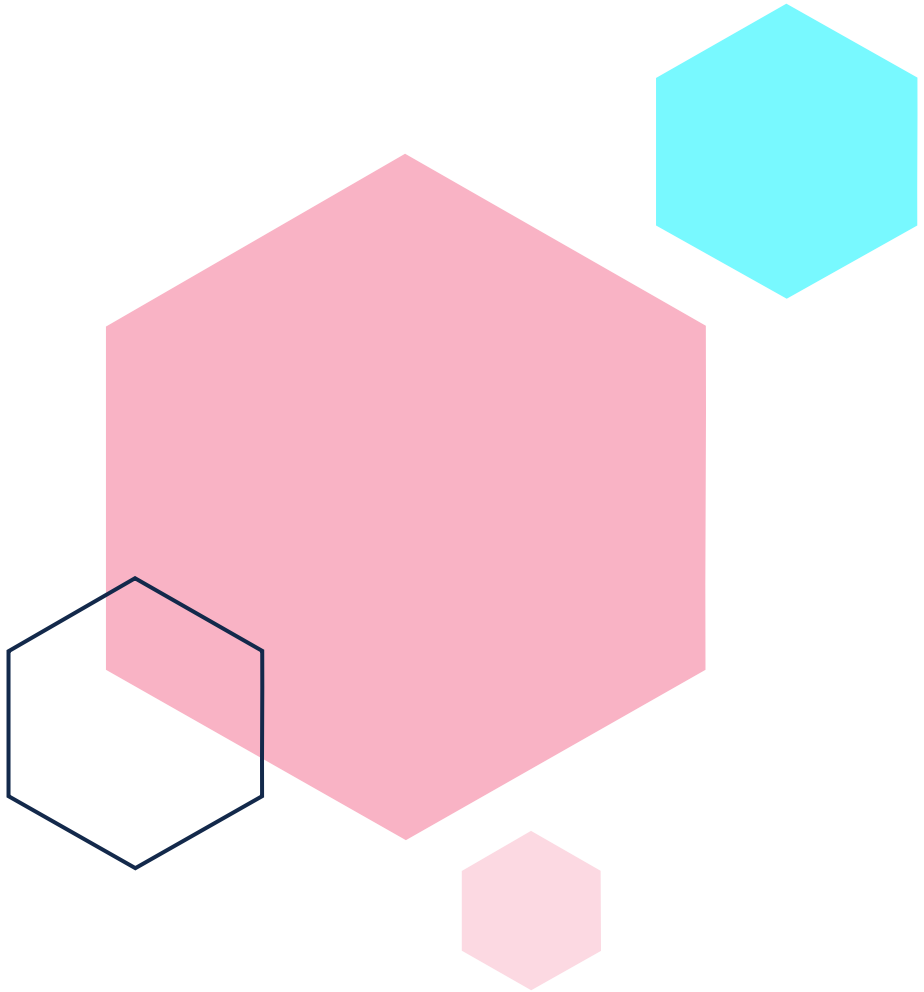
Discussion and Conclusion

- BDT does outperform (raw) A/E discriminator – but likely does not rival fully-tuned A/E in traditional analysis

- Fast analysis tool – Whole pipeline can be run in a matter of minutes from raw data to trained BDT model.
  - Particularly important as LEGEND ramps up in detector characterization with ~250 detectors to be analyzed in a short time frame.

Next Steps

- Multi-detector support.
- Further explainability studies.
- Incorporation of Solid State Detectors pulse shape simulation data.



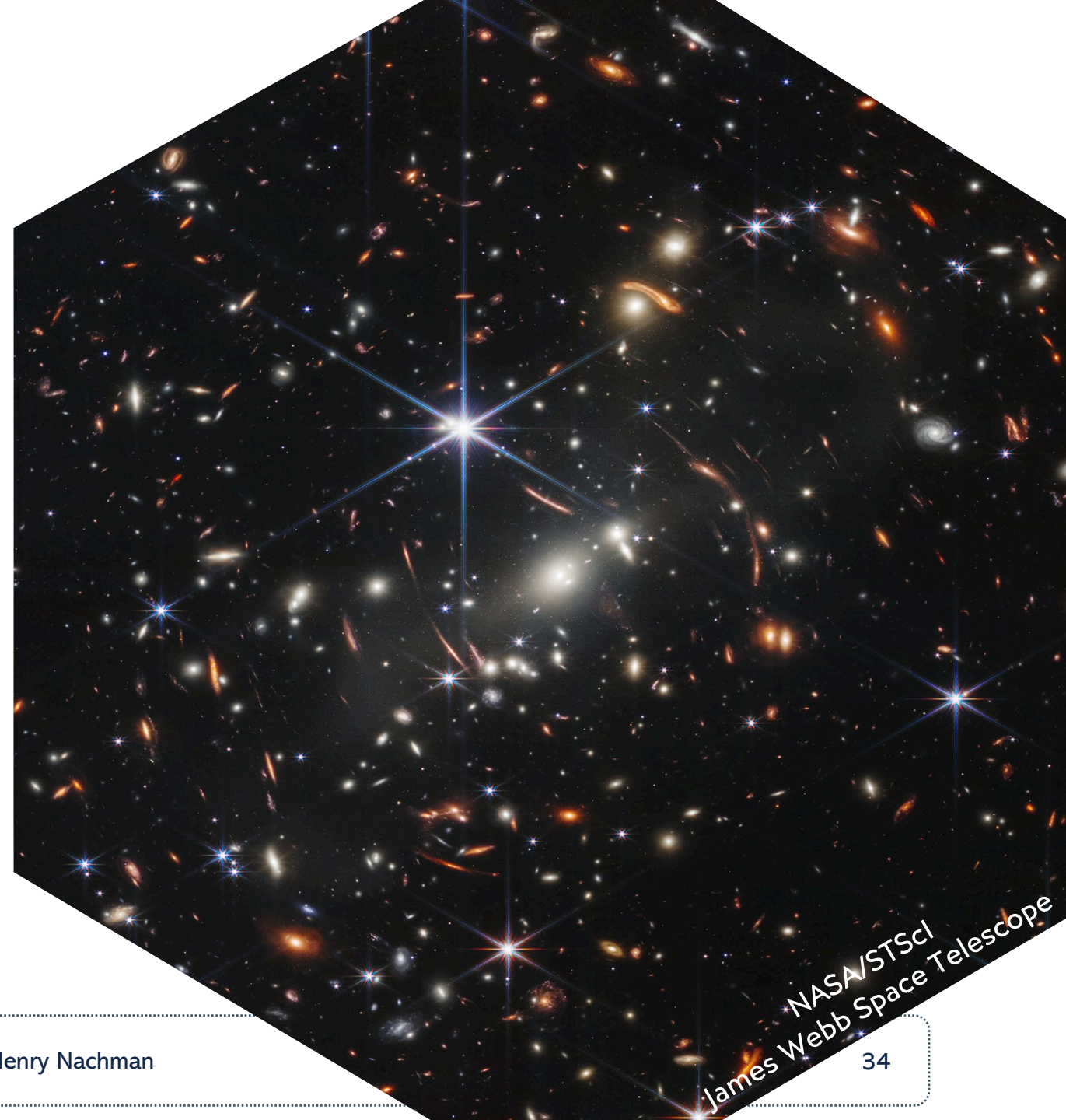
Questions?

Correspondence:

Henry.Nachman@unc.edu

Extra Slides

Why is the Universe Matter Dominated?



Standard Model of Particle Physics

- Within the Standard Model – total baryon and lepton number is conserved.
- For every matter particle, we expect to see an antimatter particle
- There must be a process that breaks the balance matter and antimatter to account for observed asymmetry

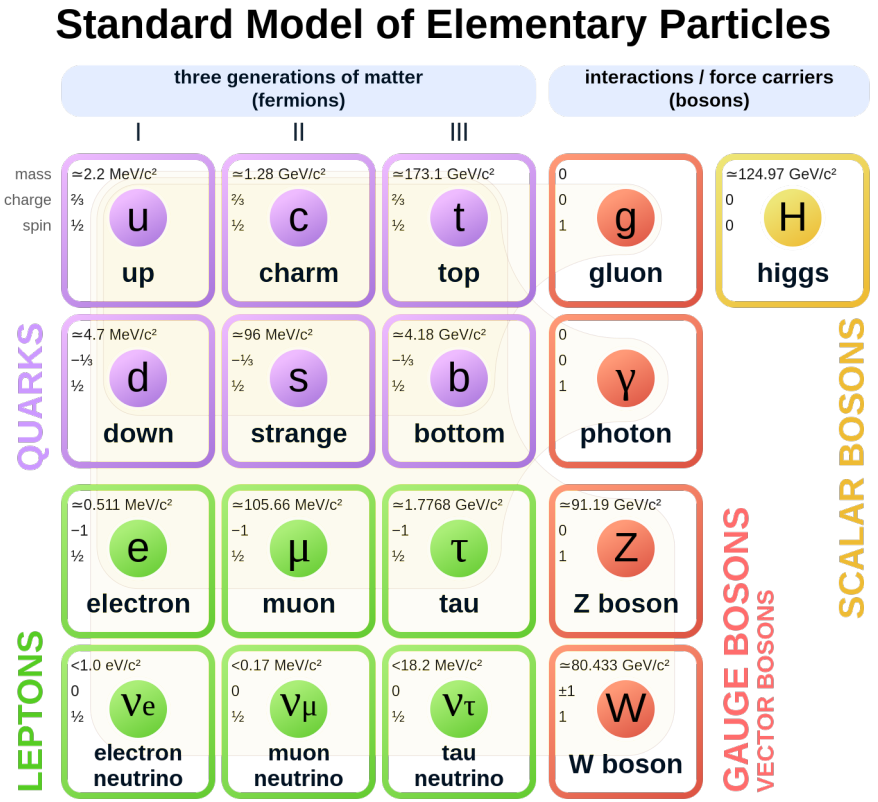
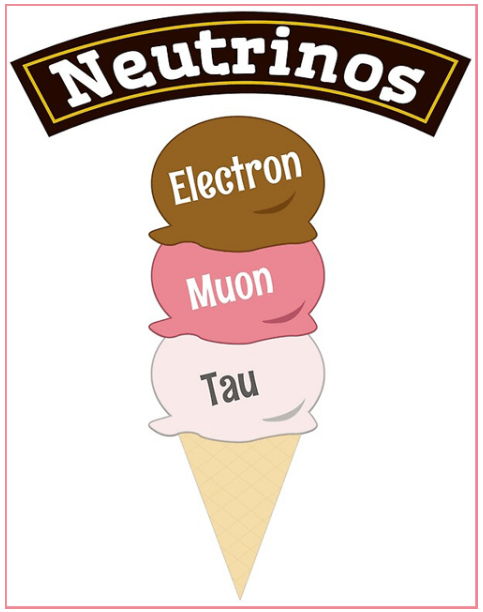


Image Source: Wikipedia

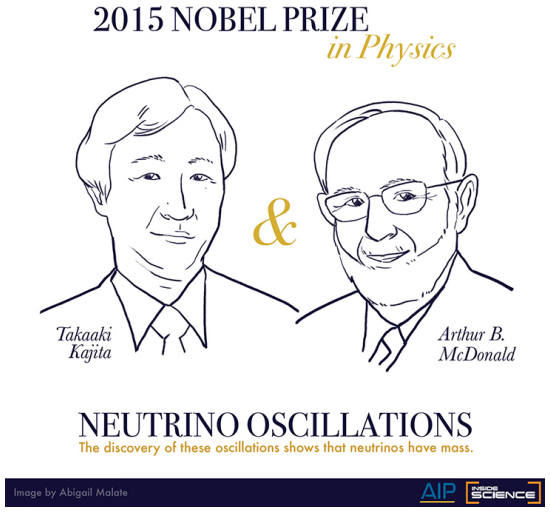
Enter Neutrino – ν

- Neutrinos are:
  - Neutral subatomic particles
  - Fermions (half-integer spin)
  - Come in 3 flavors



Source : Michael McClellan

- Importantly
  - Observations of neutrino oscillations show that neutrinos have **MASS.**
  - Discovery for which Takaaki Kajita & Art McDonald received the 2015 Nobel Prize in Physics.



Source: American Institute of Physics

Majorana Particle

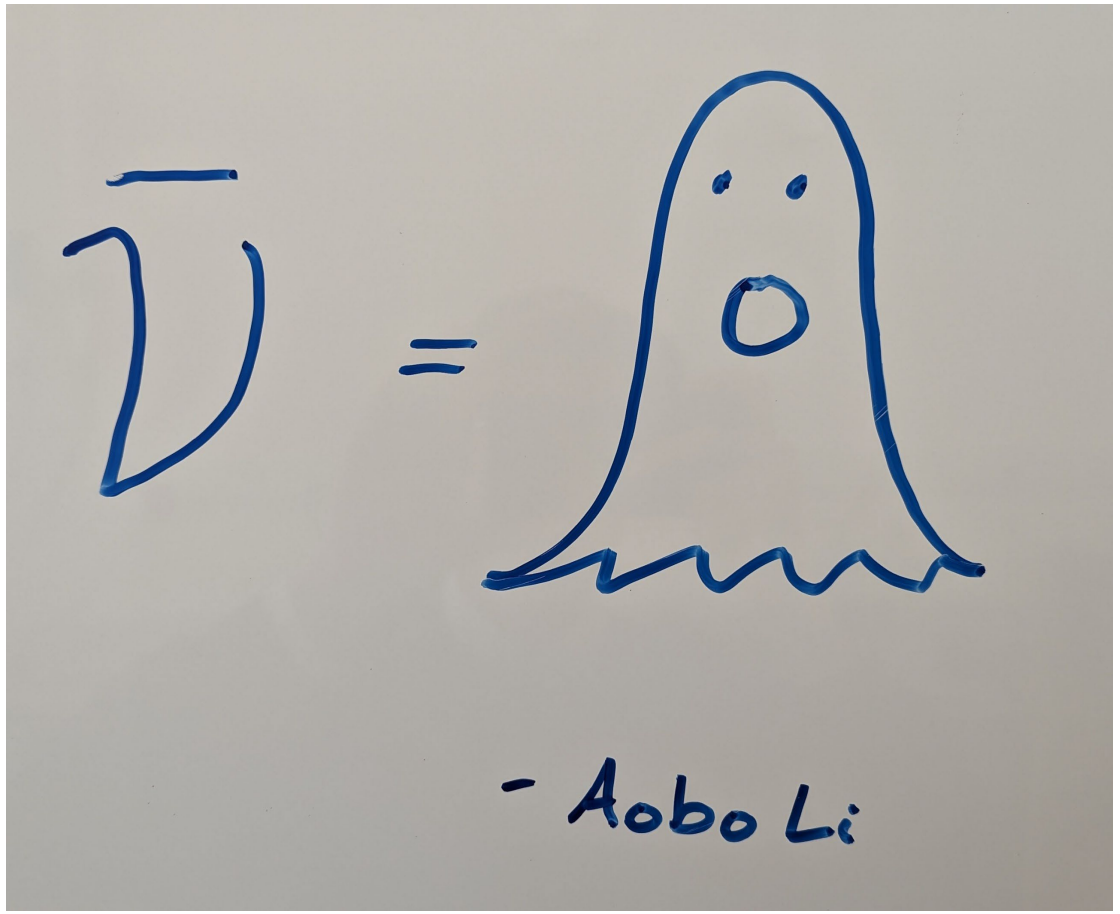
- These characteristics of neutrinos suggest it could be a Majorana Fermion – or its own antiparticle.

$$\nu = \bar{\nu}$$

- If the neutrino is Majorana, this could allow for baryogenesis (from neutrino leptogenesis)
- One possible explanation for matter – antimatter asymmetry.

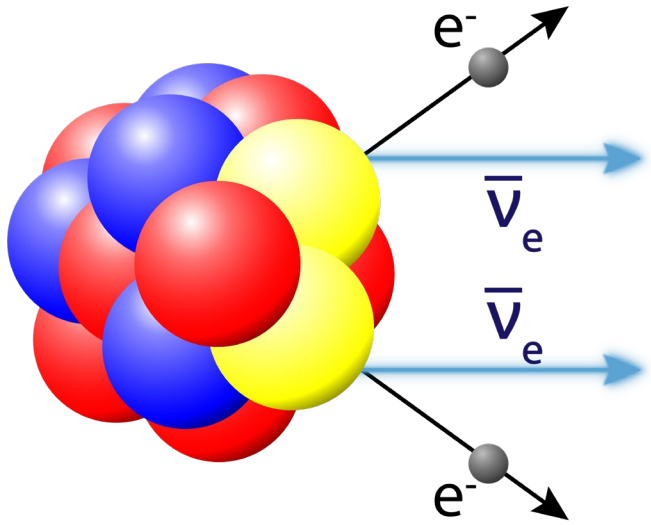
Ghosthunting for Neutrinos

- Lack of charge, and extremely low interaction cross-section make direct neutrino measurements very difficult.
- Instead – we can use the theorized process known as *Neutrinoless Double Beta Decay* to probe Majorana nature of neutrinos.
- Observing $0\nu\beta\beta$ would prove that neutrinos are Majorana.



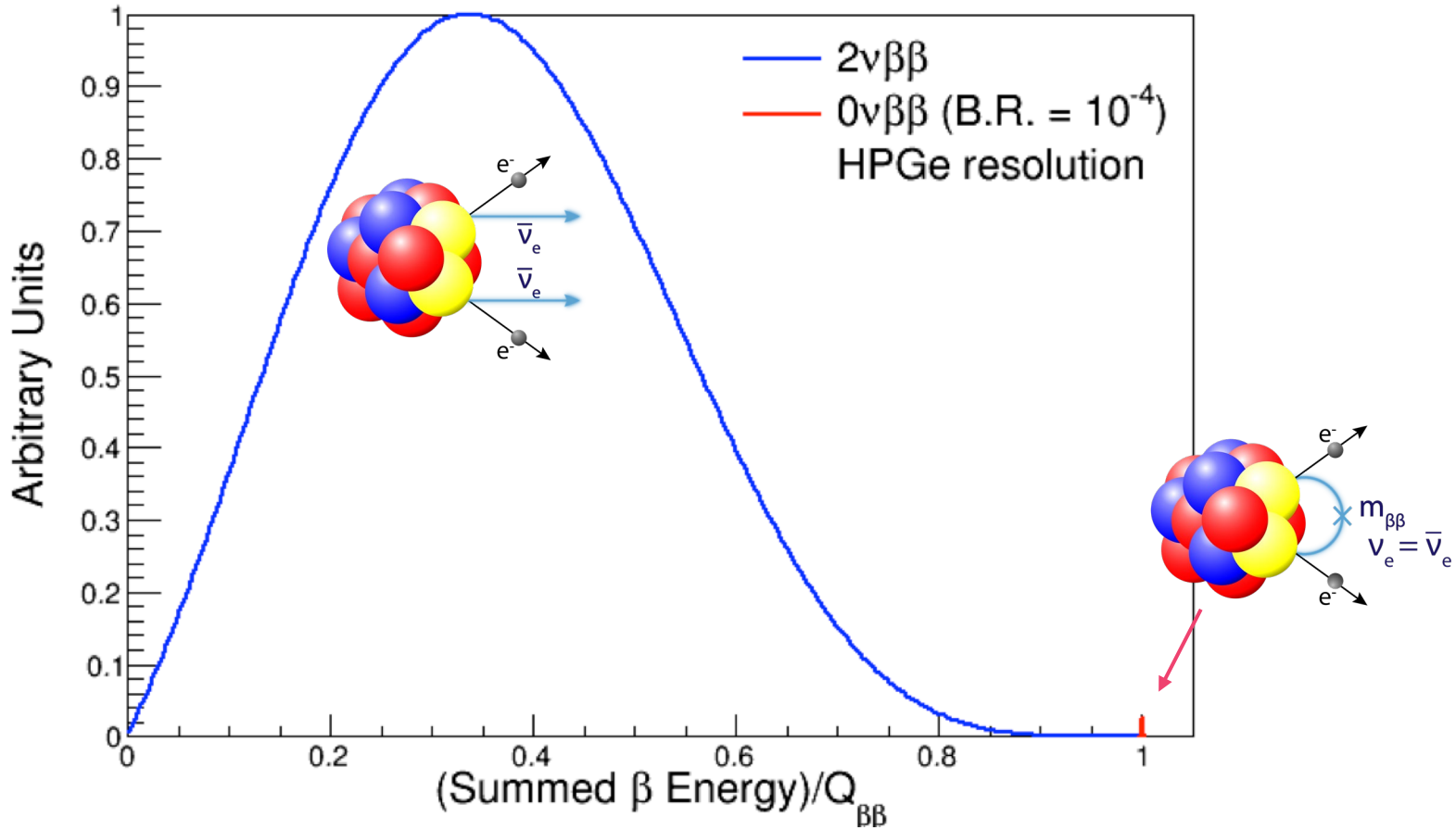
Standard Model – Double Beta Decay

$$(A, Z) \rightarrow (A, Z+2) + e^- + e^- + \bar{\nu}_e + \bar{\nu}_e$$



Observed in 11 different naturally occurring isotopes.

Among them: (Spoiler Alert) ^{76}Ge


2νββ
0νββ (B.R. = 10^-4)
HPGe resolution
Arbitrary Units
(Summed β Energy)/Q_ββ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
e-
ν̄e
ν̄e
e-
e-
e-
mββ
νe = ν̄e

Large Enriched Germanium Experiment for Neutrinoless ββ Decay

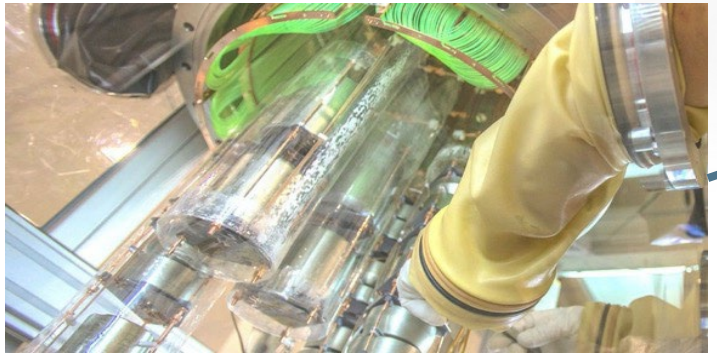
Henry Nachman

RUSSIAN FOUNDATION FOR BASIC RESEARCH

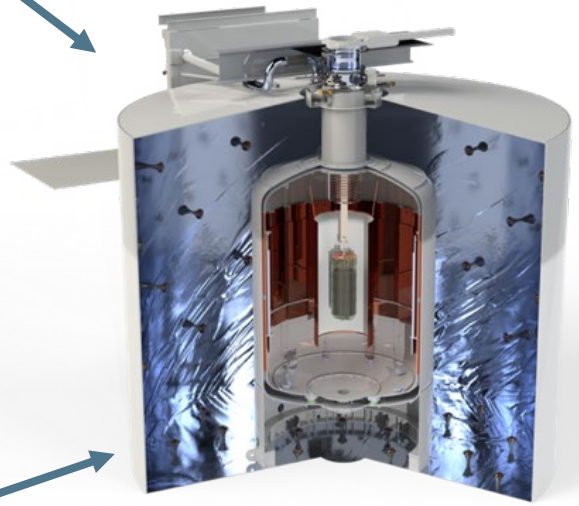
Istituto Nazionale di Fisica Nucleare

Innovation toward LEGEND-1000

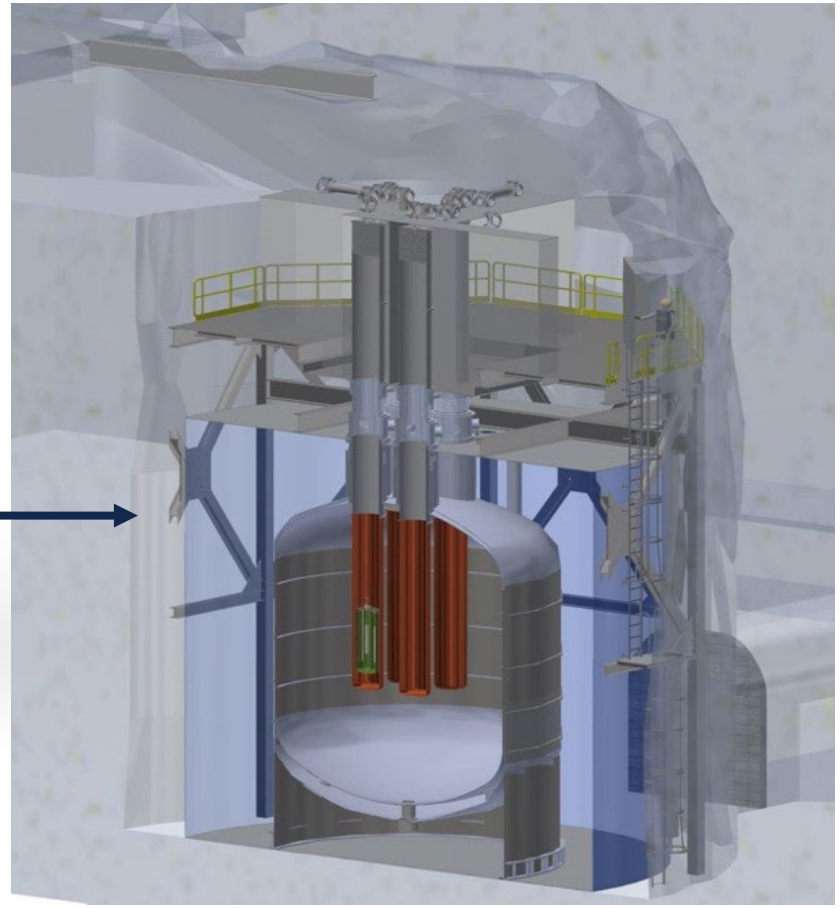
MAJORANA DEMONSTRATOR (MJD) : PPCs, low noise electronics



GERDA : LAr veto, water shield



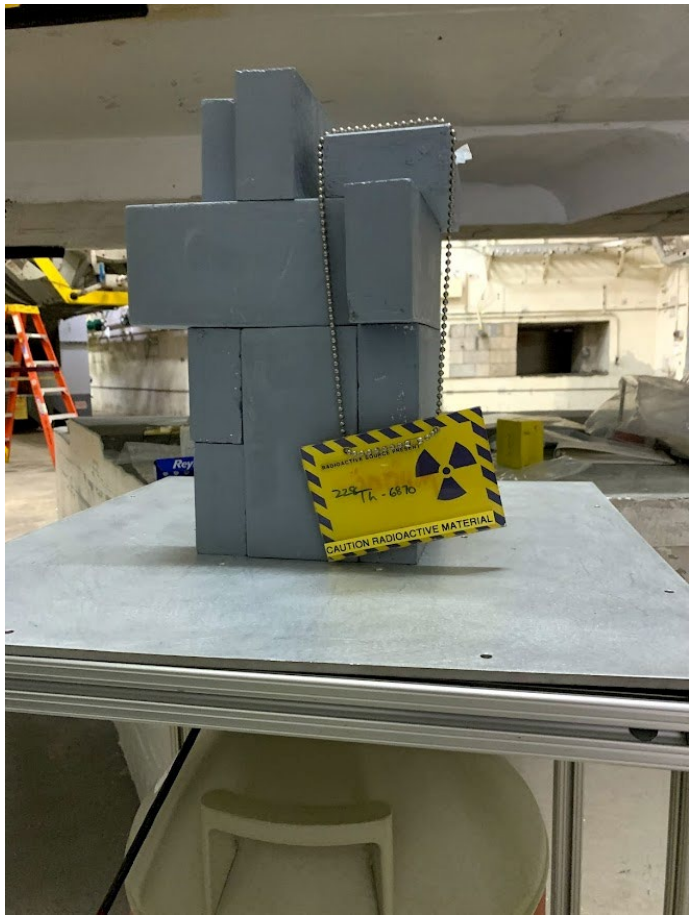
LEGEND-200 : Now taking data



LEGEND-1000 : Conceptual design development continuing

Characterization

- Prior to installation, detectors undergo tests to confirm appropriate energy resolution, pulse shape reconstruction and other qualities

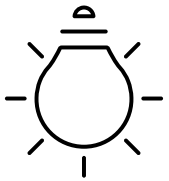


Characterization Data

- During characterization: detectors are biased (given a voltage gradient) and take data with a Th-228 source.

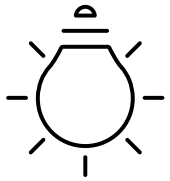
- Data for this model is taken over two different runs:
  - Top Source
  - Side Source

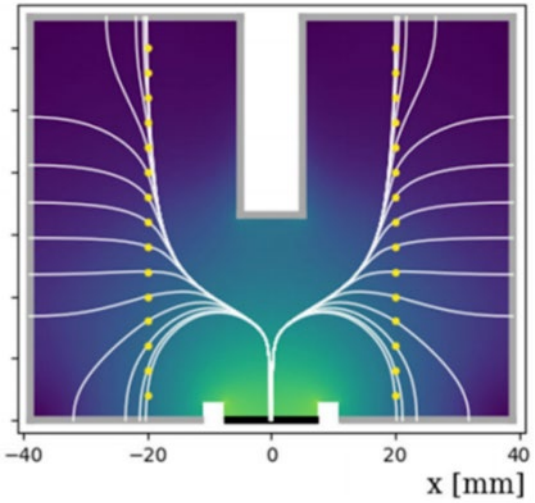
- All data used in this model from a single ORTEC ICPC detector (V01387A)
- Characterization performed by Morgan Clark at Oak Ridge National Lab on June 17, 2021.



Top Source

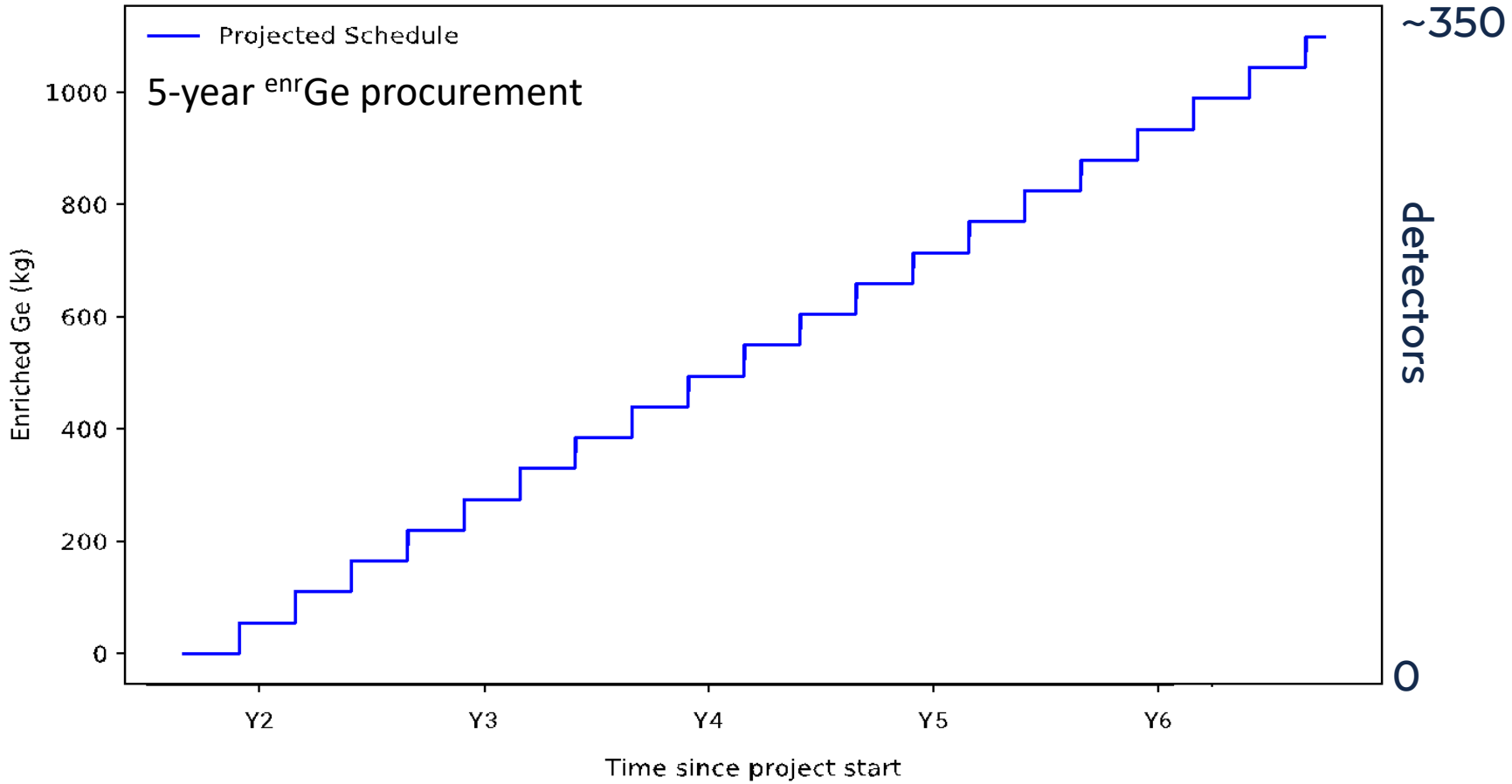
Side Source



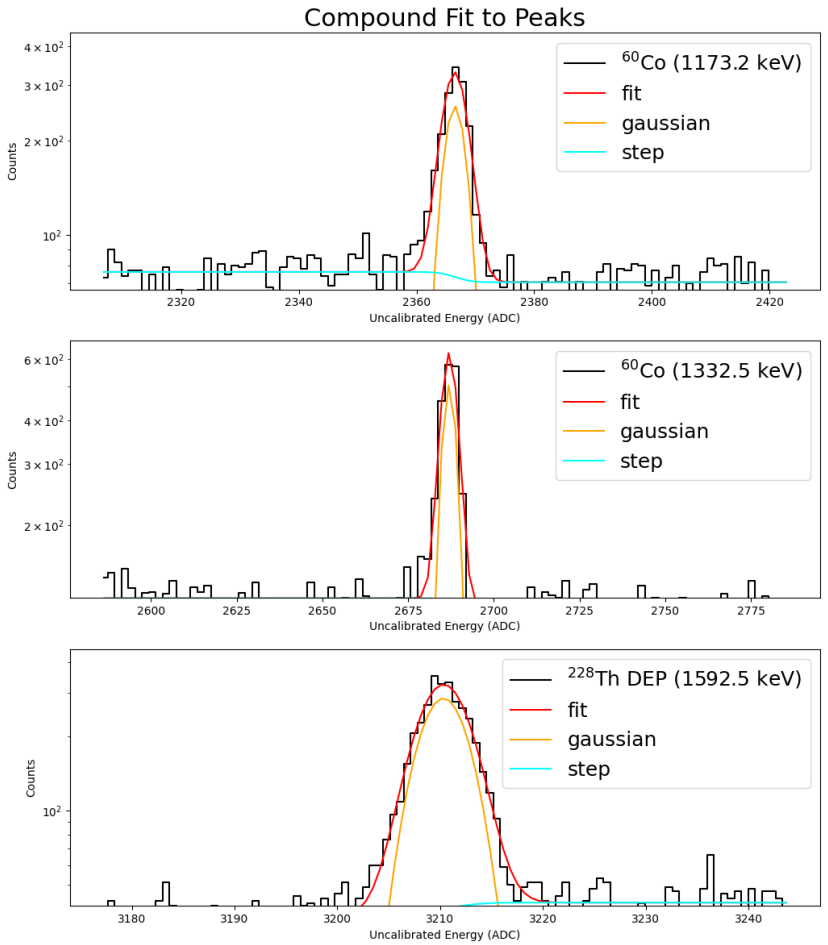


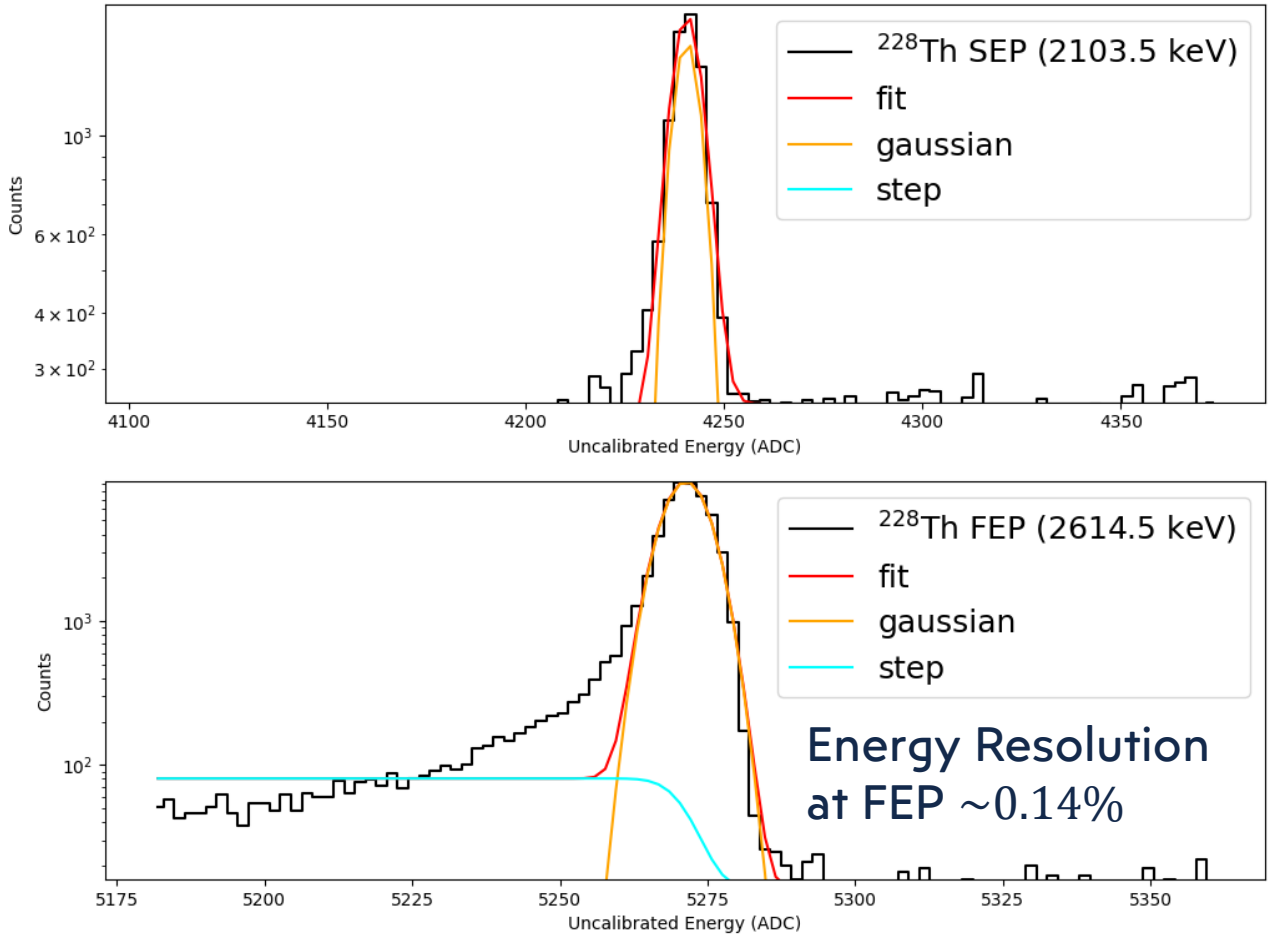
Scaling up for L-1000

- As the LEGEND experiment prepares for L-1000 many detectors will need to be efficiently characterized
- At peak ~ 2 detectors/week
- Motivates a need for a fast analysis tool
  - Such as the BDT model I developed.

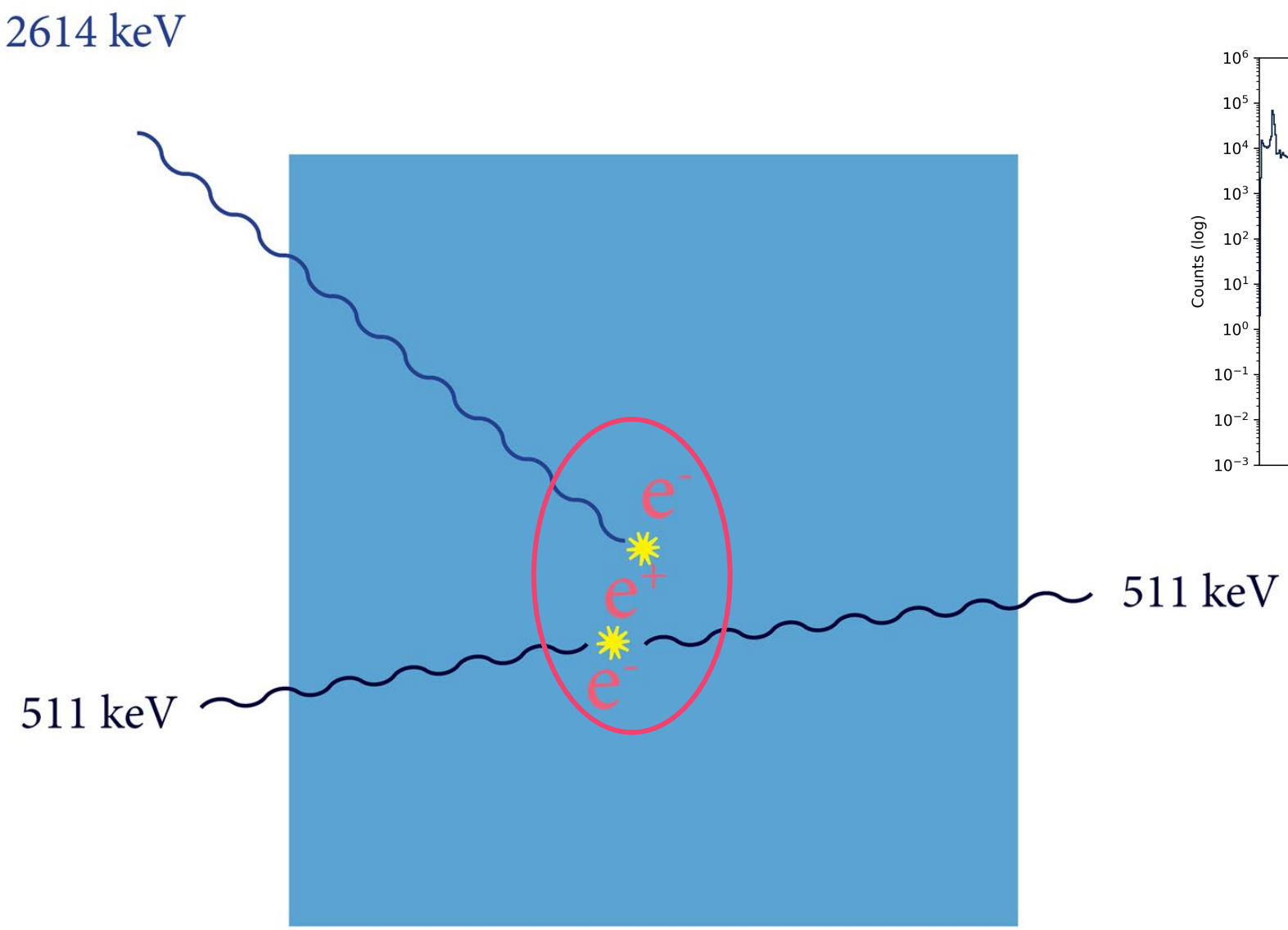


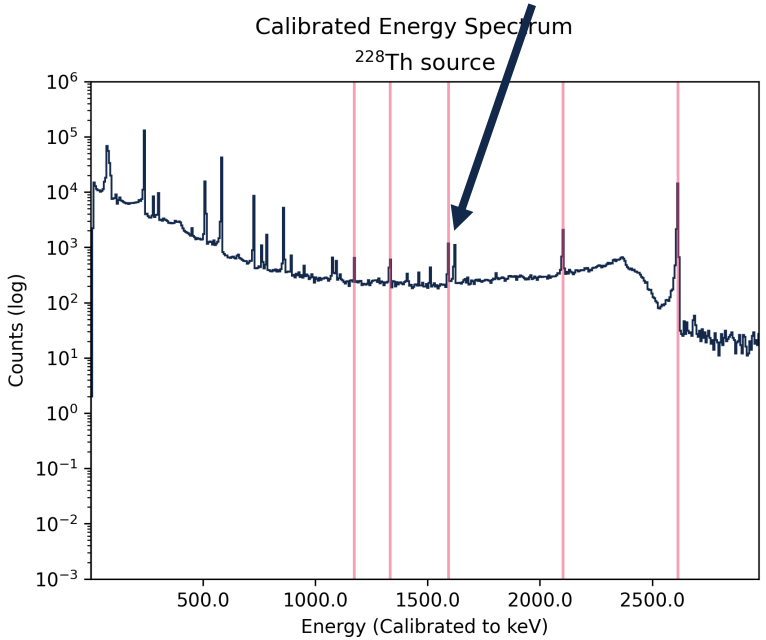
Fit to calibration peaks





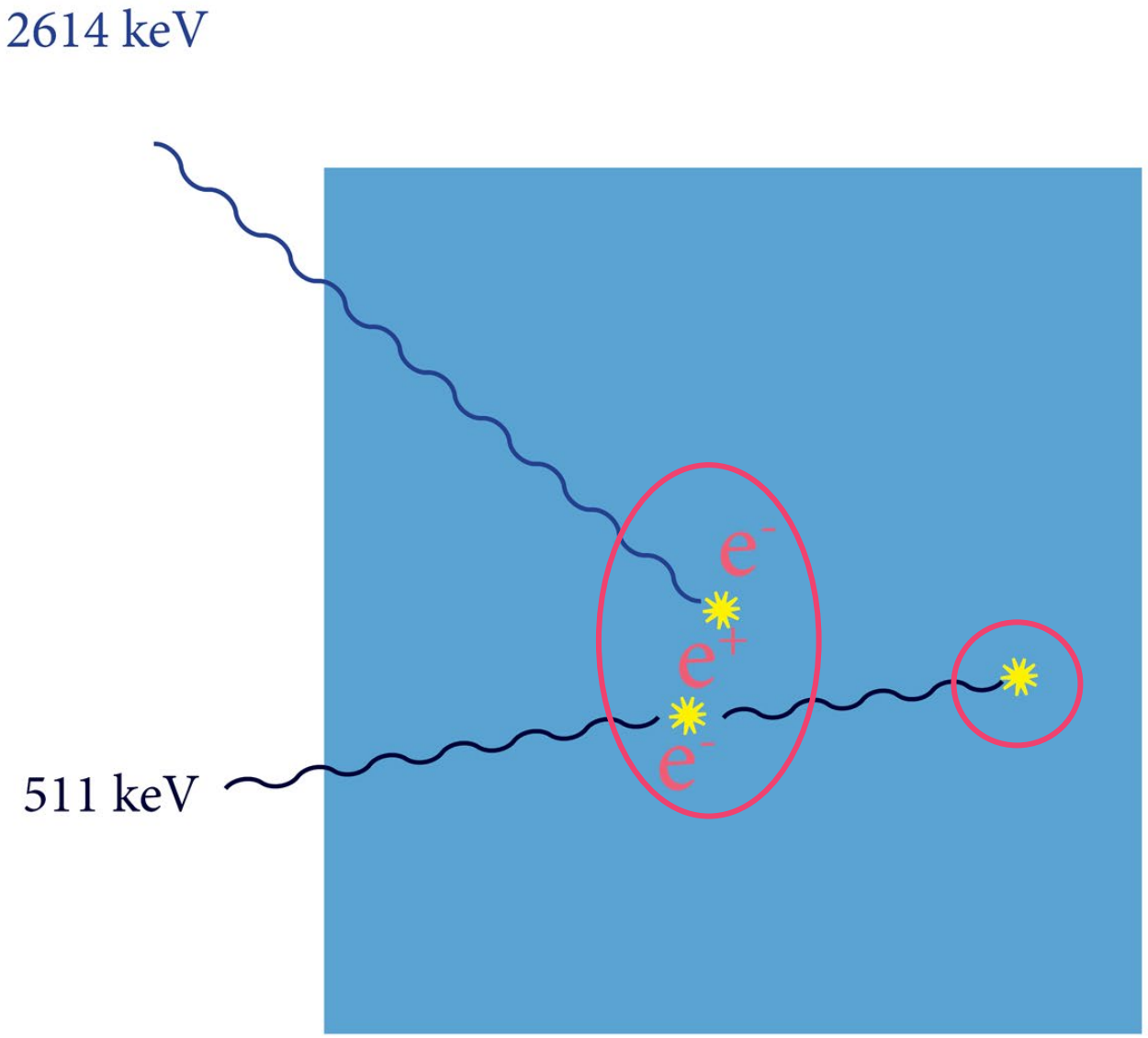
Single-Site Proxy Event

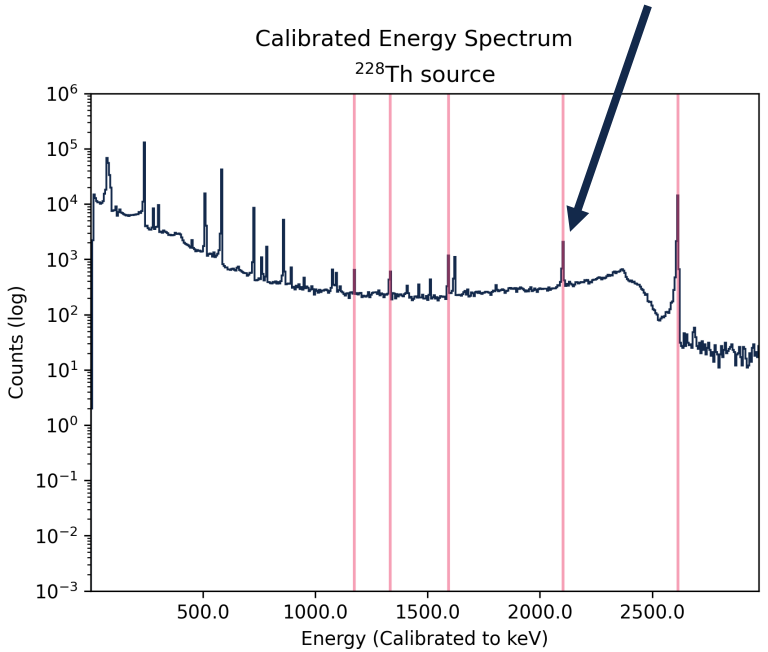




Double Escape
Total Collected Energy:
2614-511-511 = 1592 keV

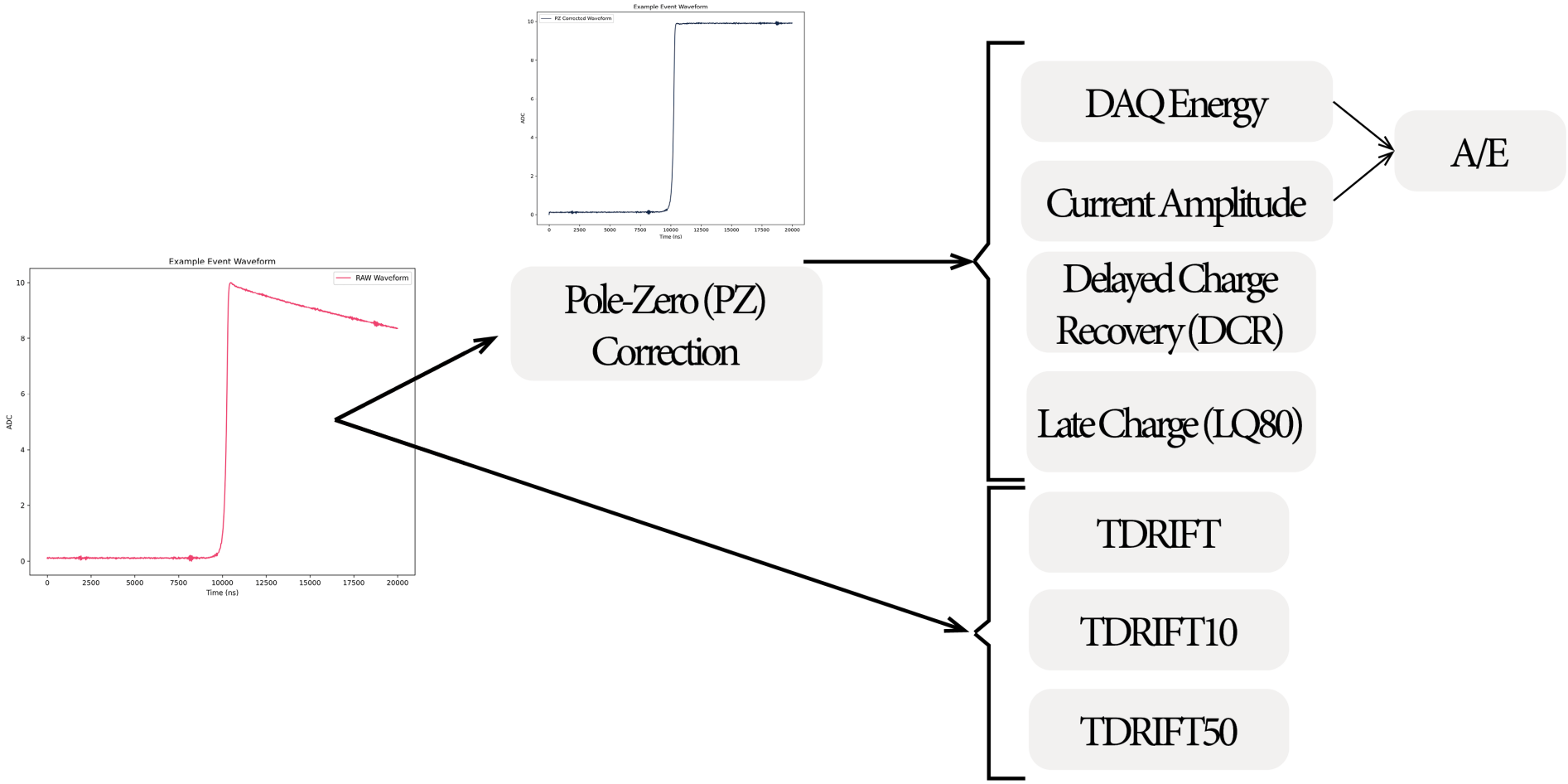
Multi-Site Proxy Event



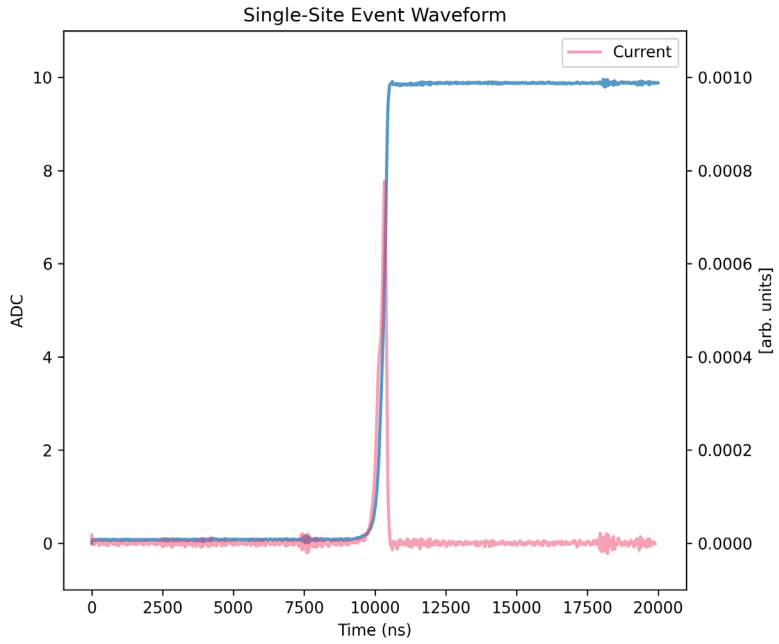


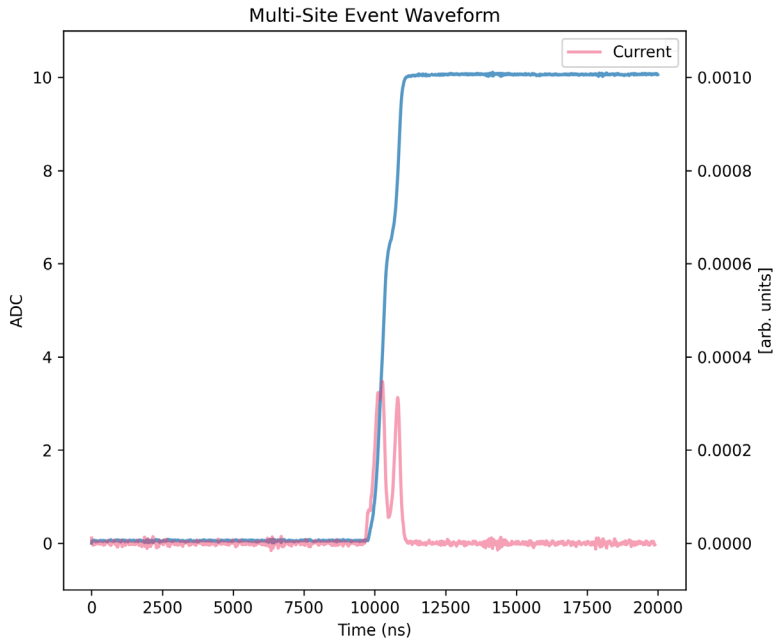
Single Escape
Total Collected Energy:
2614-511 = 2103 keV

What Parameters are Extracted?



Current Amplitude

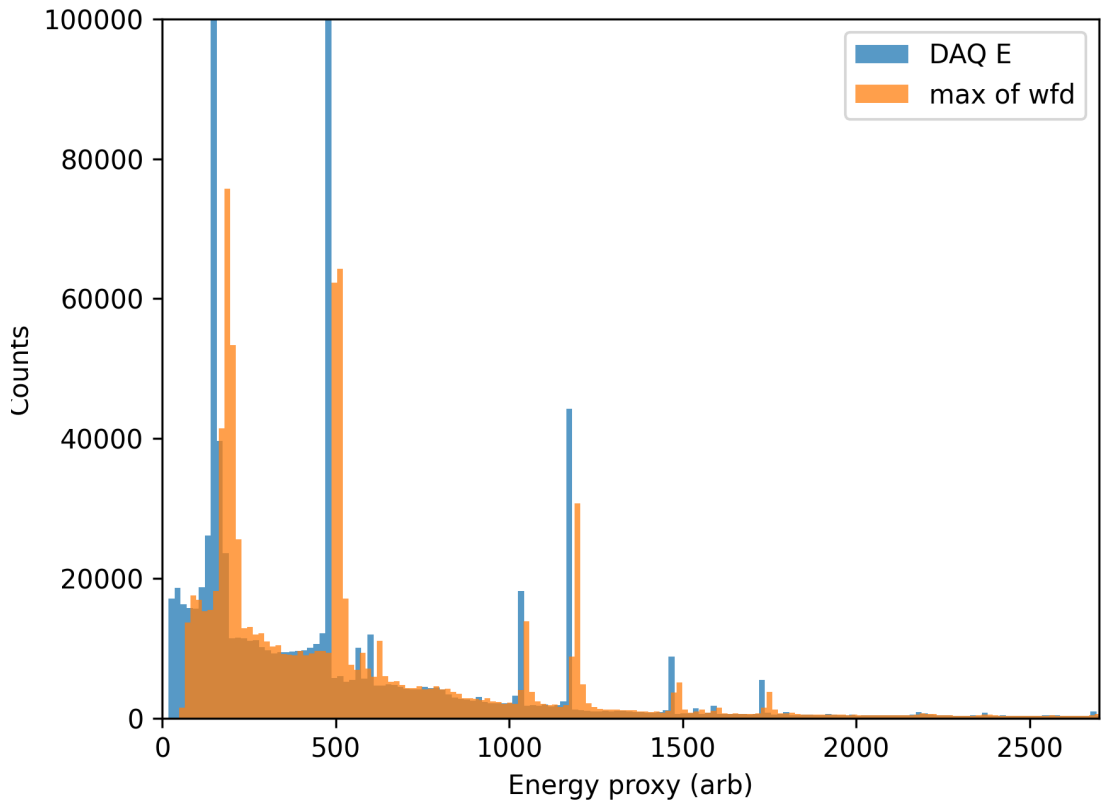




Energy

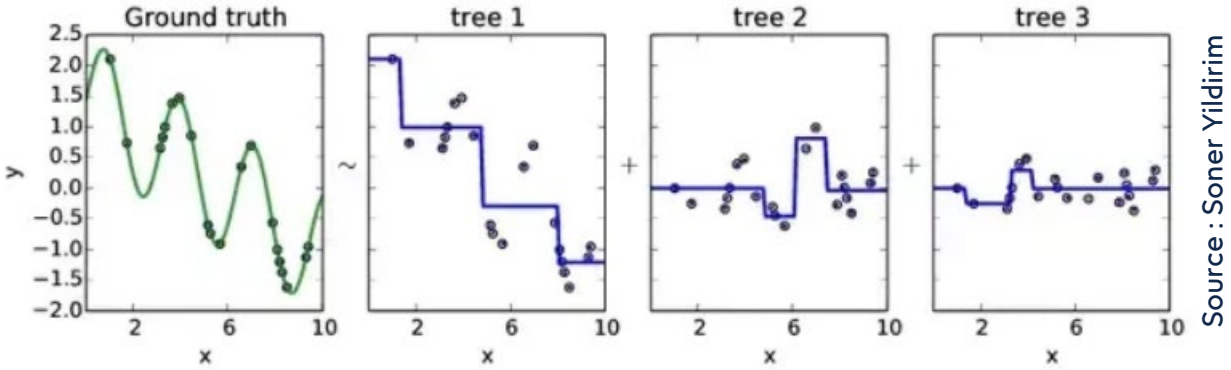
- Three different methods for estimating the energy of an event were developed
  - Energy from data acquisition (DAQ) system – simple trapezoidal filter
  - Offline trapezoidal filter
  - Maximum of the waveform

- Ultimately – DAQ Energy had a comparable distribution to the other methods and was chosen for its simplicity in integration into the pipeline

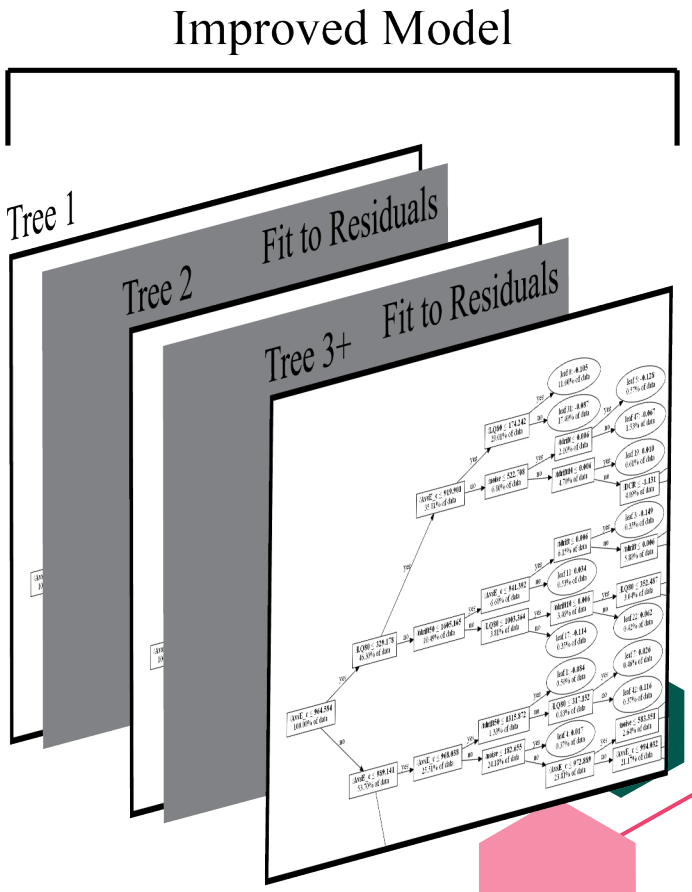


What makes it Boosted?

- A single decision tree model is a “poor learner” on its own.
- Boosting - growing multiple trees in a series, each fitting to the residuals of the previous.
- Validation Metric – Binary Log Loss function
  - Minimizing residuals of mischaracterized events



Source : Soner Yildirim



Model Settings

Hyperparameter Name	Value
Number of Iterations	Early Stopping: 10 (max 2000)
Validation Metric	Binary Log-loss
Learning Rate	0.0744
Number of Leaves	73
Bagging Frequency	62
Minimum data in Leaf	26
Max Bin	542
Drop Rate	0.330
Min Gain to Split	0.536
Boosting Algorithm	Gradient-based One-Side Sampling (GOSS)

Build trees until the metric does not improve for 10 iterations

Training

- 2 models : 1 with only top data, 1 with a mix of top and side data (with data augmentation)
- Data is split into training : validation subsets with 70:30 ratio.

Top Only	Pre-Matching	Post-Matching
SEP Size	17710	2062
DEP Size	7365	2062
Trees	76	

Mix Data	Pre-Matching	Post-Matching
SEP Size	35420	3899
DEP Size	14730	3899
Trees	47	

Data for Training

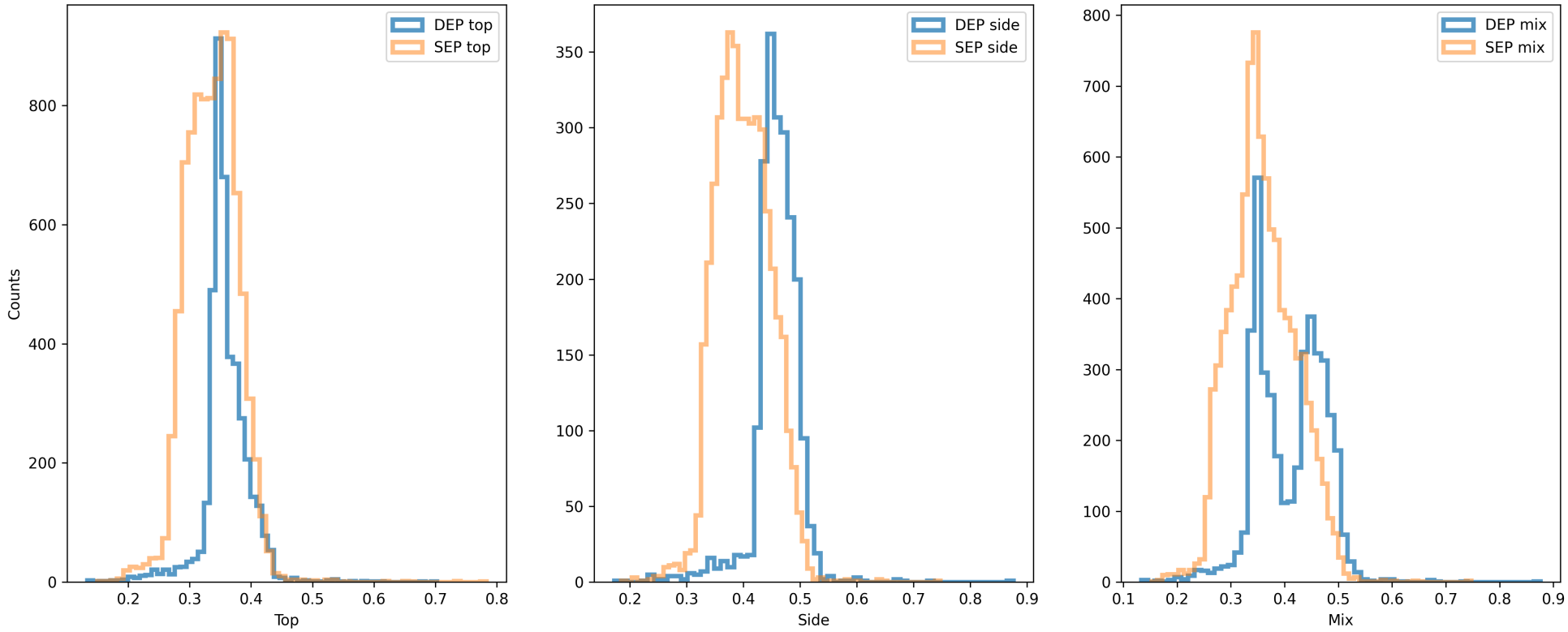
- Multi-Site Proxy : Single Escape Peak (SEP)
- Single-Site Proxy : Double Escape Peak (DEP)

- 2 Datasets : Top Source (Flood) Data & Side Source Data

Dataset	# Events in SEP	# Events in DEP
Top Source	25300	10521
Side Source	4712	2097

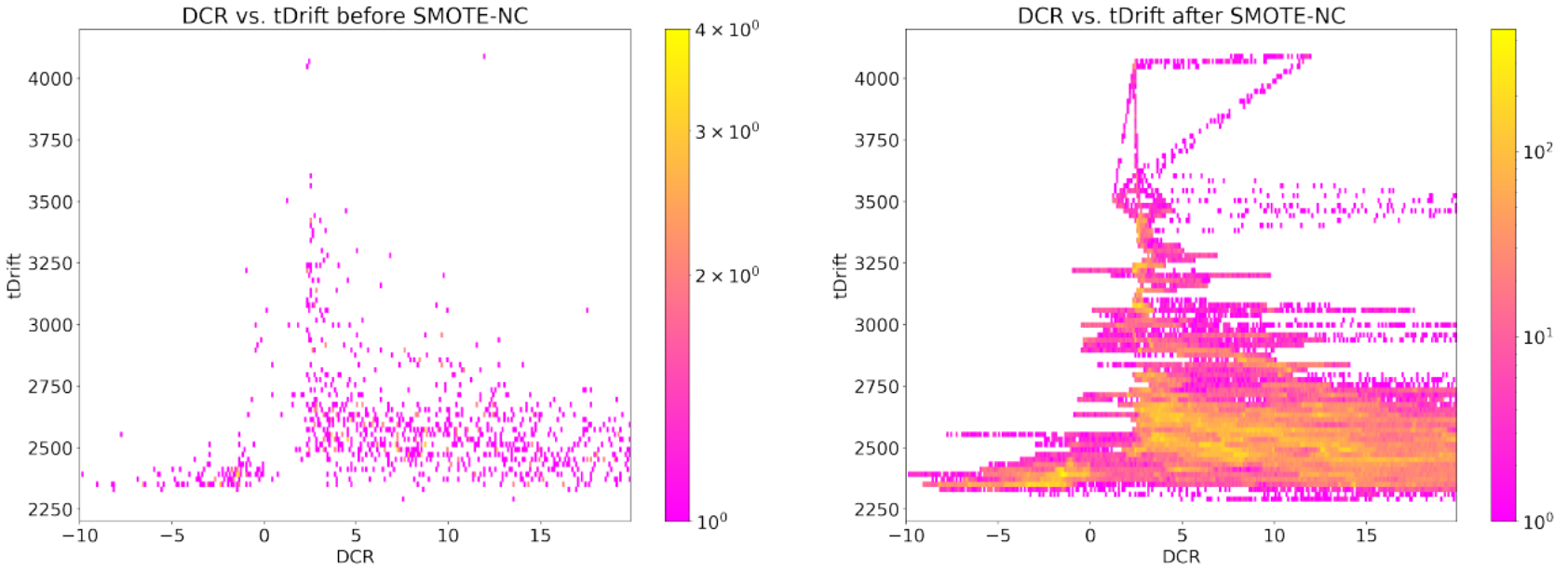
Top and Side Source Parameter Distribution

A/E Distributions



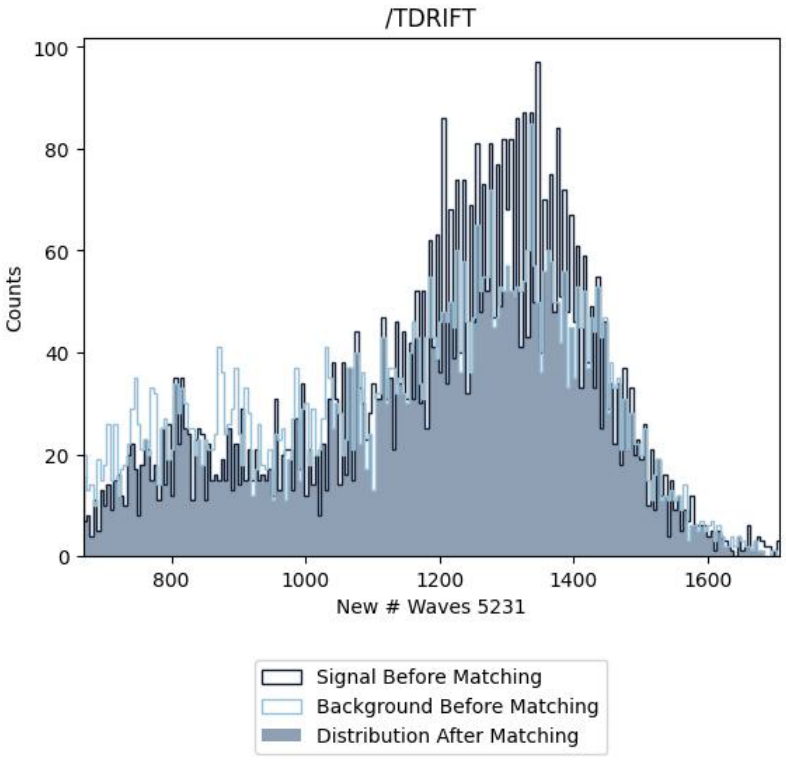
Data Augmentation

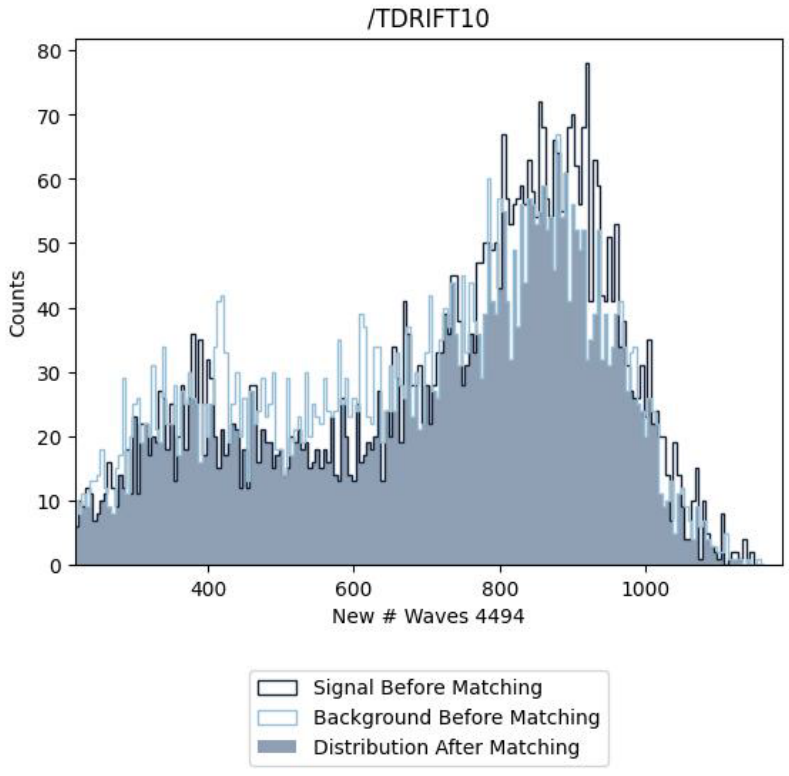
- Much more data exists for top source – vastly differently sized datasets can result in unintentional bias in the model.
- SMOTE-NC: Artificially synthesize new data points by interpolating between existing points in 6-dimensional parameter space.

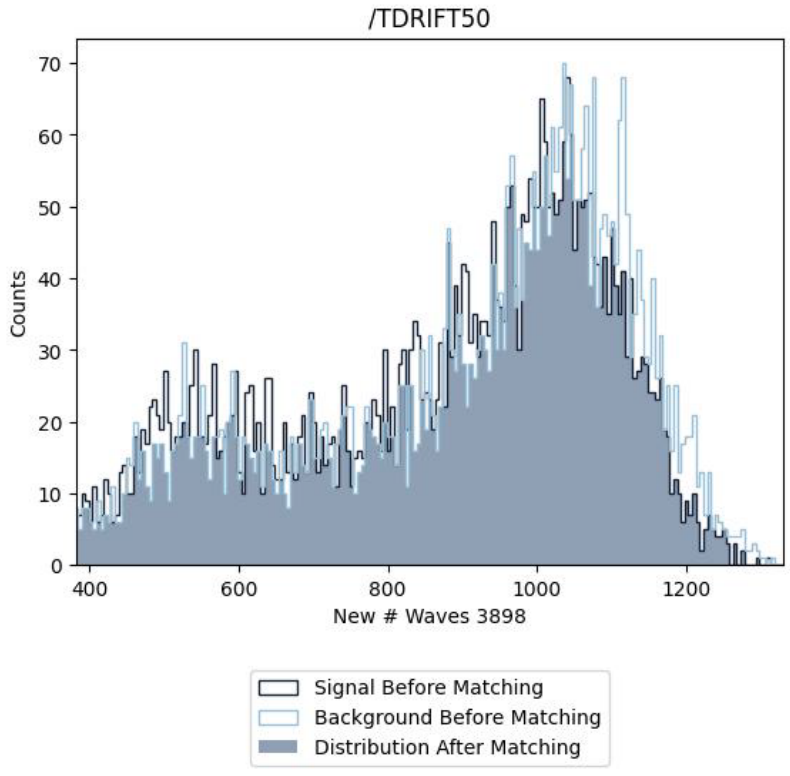


Note: this plot does not represent the data used in this model – it merely visualizes how SMOTE-NC works.

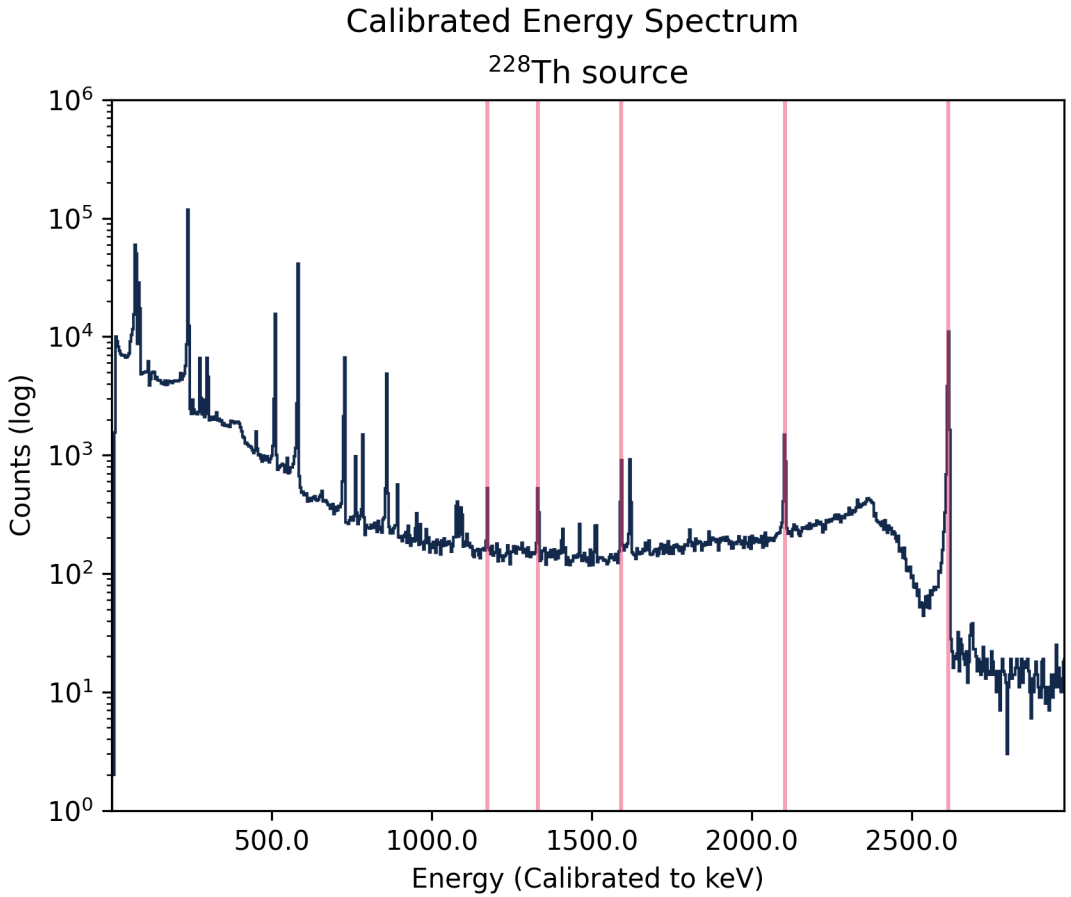
Source : A.Li, J.Gruszko

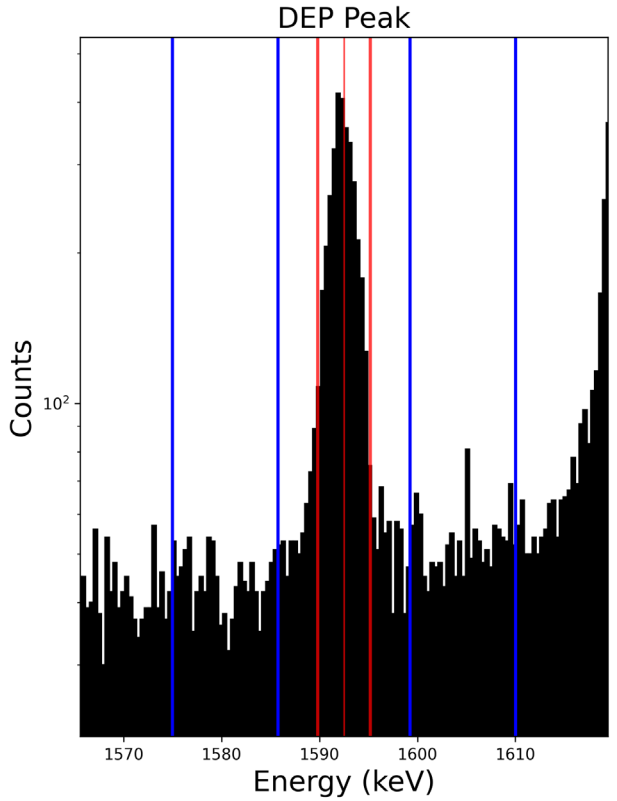

/TDRIFT
Counts
New # Waves 5231
Signal Before Matching
Background Before Matching
Distribution After Matching


/TDRIFT10
Counts
New # Waves 4494
Signal Before Matching
Background Before Matching
Distribution After Matching

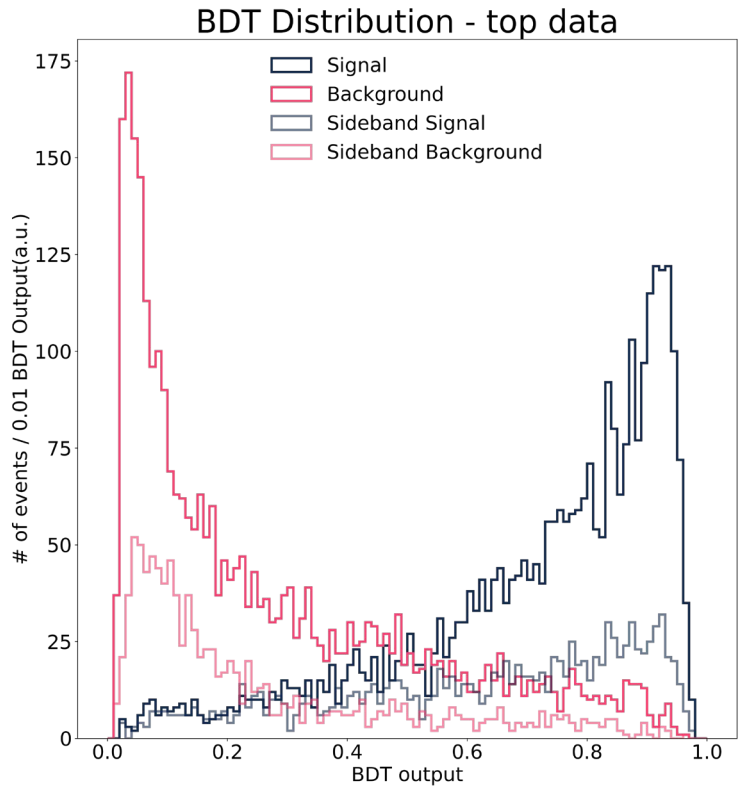

/TDRIFT50
Counts
New # Waves 3898
Signal Before Matching
Background Before Matching
Distribution After Matching

Background Events & Sideband Subtraction





BDT Distribution – Top data



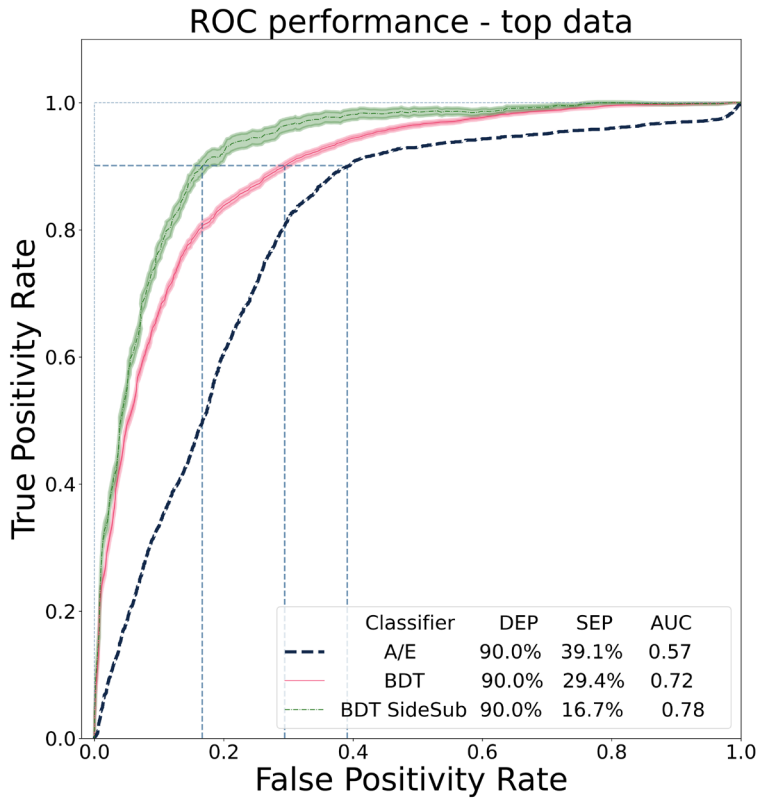
Receiver Operator Curve – Top data

Visualization of binary classification

- True Positivity Rate (rate of events correctly classified as SS) vs False Positivity Rate (falsely classified as SS)

- Perfect classifier would have a step function with an AUC = 1
- Random classifier would be a diagonal line with AUC = 0.5

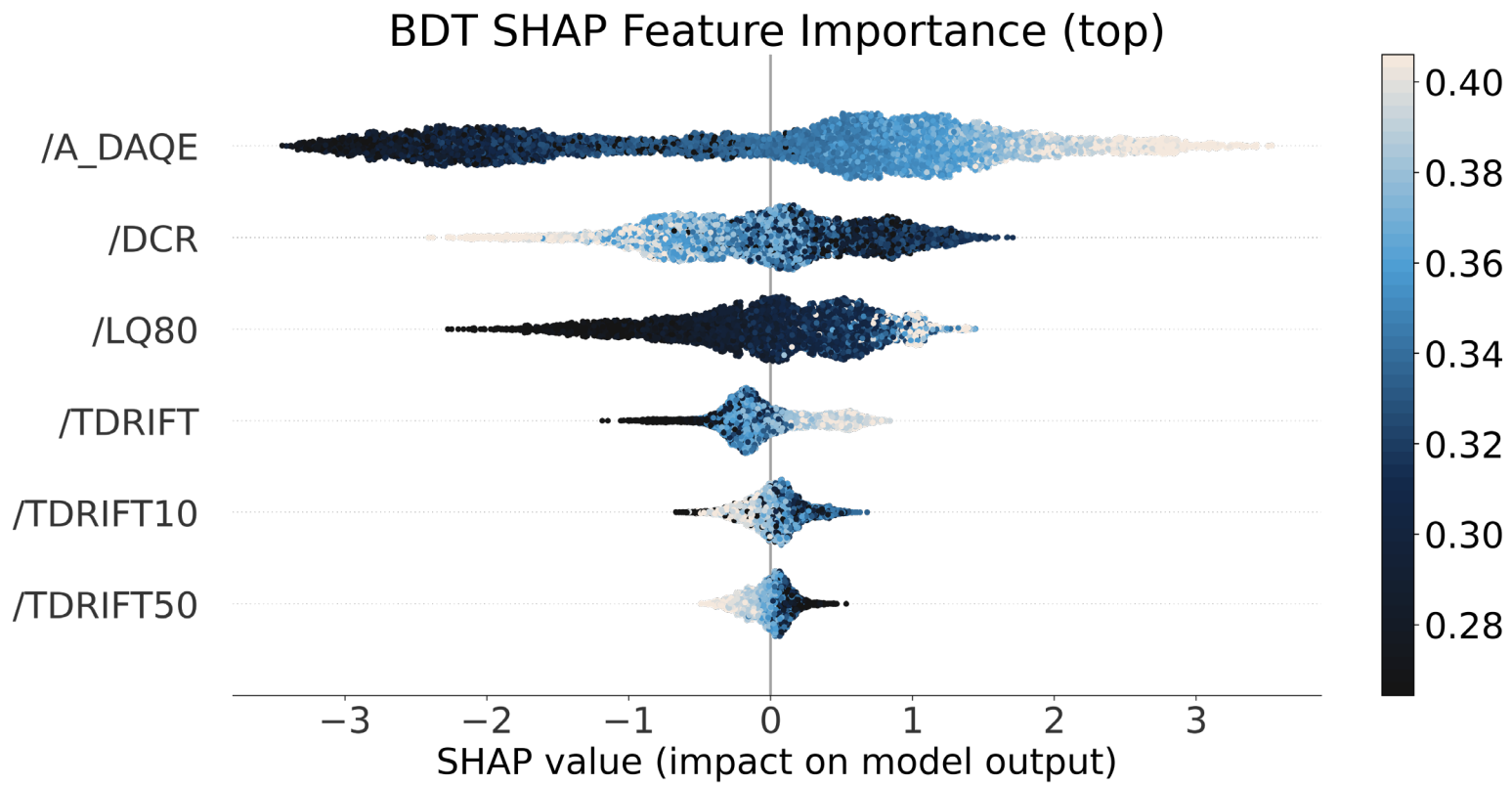
Note: the A/E measure shown here is without drift time correction or other methods used in the traditional analysis



SHAP Feature Importance - Top

- Made by plotting Shapely values for every event.
- Color of point represents the ultimate classification.

- "Important" features have a dumbbell shape



Next Steps

1. Multi-detector support – **In Progress**
   a. Increases total data yield
   b. Helps eliminate possible bias from “quirky” detectors
2. New (and improved) parameters
3. Explainability expansion
   a. Multivariate correlation studies
4. Other ML models – particularly those that look at the whole waveform, rather than just the provided parameters.

Avenues for Publication

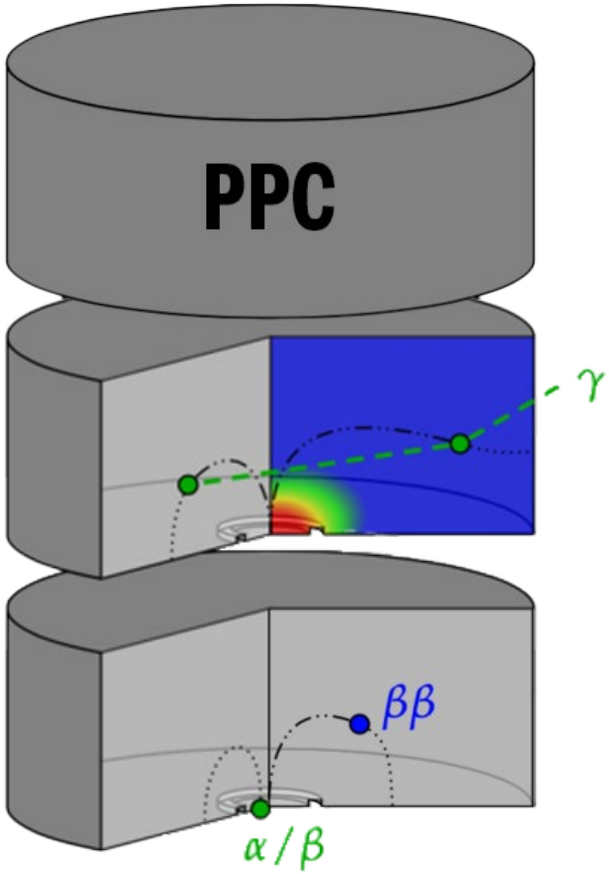
- Start with an internal LEGEND technical document.
  - Let the larger collaboration know about the work I am doing and provide an opportunity for guidance, and feedback

1. Depending on this package's use in future characterization work – possible incorporation into a paper on LEGEND's detector characterization efforts.
2. Contingent on reliable results on further explainability development – particularly multi-variate analysis – possible standalone ML paper

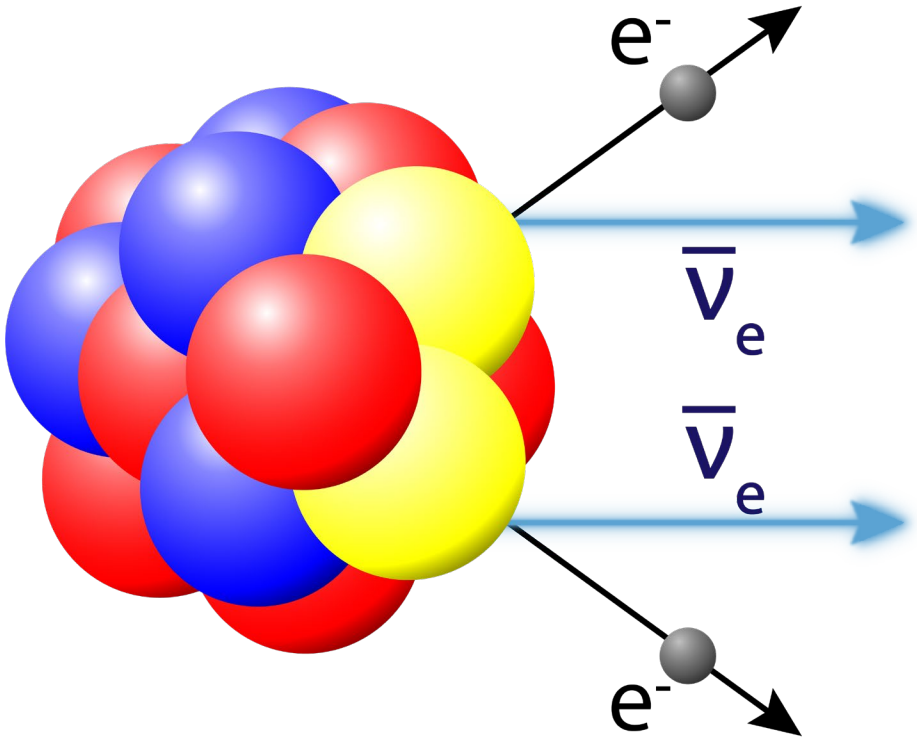
Binary Log Loss

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

Binary Cross-Entropy / Log Loss


PPC
γ
ββ
α / β

$2\nu\beta\beta$



$0\nu\beta\beta$