

MAJORANA DEMONSTRATOR Data Release and Data Challenges

Aobo Li

Halicioğlu Data Science Institute & Department of Physics
UC San Diego

04/12/2023

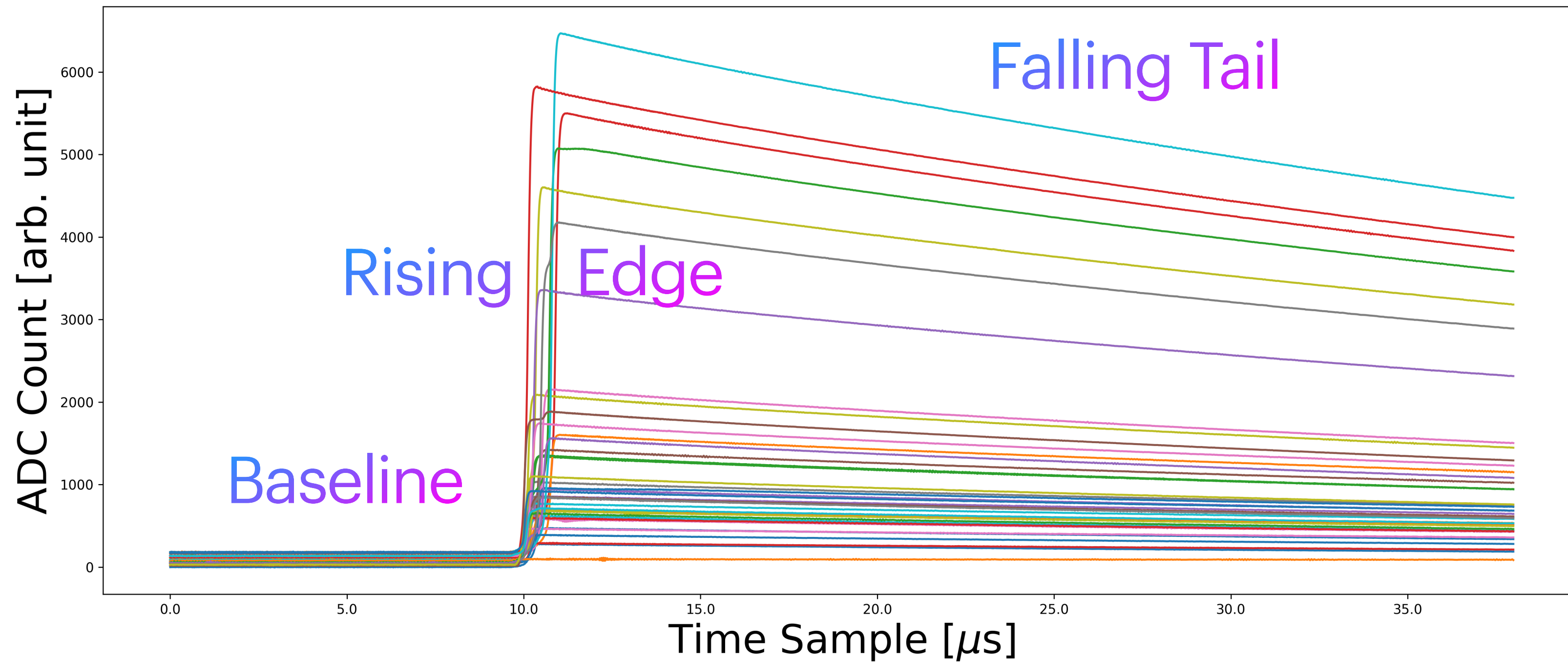
I am hiring PhD students & postdocs!
Please email: liaobo77@ucsd.edu



Motivation

- A key component of AI/ML success lies in the training datasets
- Particle & nuclear physics experiments generate substantial volumes of high-quality data
 - Complemented by robust labels derived from dedicated physics analysis
- Nelson Memo from OSTP requires “make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release”
- the MAJORANA DEMONSTRATOR experiment produces high-quality time series data from High-Purity Germanium (HPGe) detectors, accompanied by their analysis labels

Short Time Series Data



- HPGe detector waveform, simple but information rich
- Pulse shape contains physics information (up to many corrections):
 - Waveform amplitude is proportional to particle energy
 - Shape of the rising edge reflects interaction type
 - Falling tail reflect electronic response

Analysis Label

TABLE I: Description of information contained in each data point of this release.

Field	Description	Data Type	Note
<code>raw_waveform</code>	Detector Waveform	<code>array(size=(3800,) dtype=float)</code>	
<code>energy_label</code>	Analysis Label	<code>float</code>	
<code>psd_label_low_avse</code>	Analysis Label	<code>binary</code>	1 means accepted, 0 means rejected
<code>psd_label_high_avse</code>	Analysis Label	<code>binary</code>	1 means accepted, 0 means rejected
<code>psd_label_dcr</code>	Analysis Label	<code>binary</code>	1 means accepted, 0 means rejected
<code>psd_label_lq</code>	Analysis Label	<code>binary</code>	1 means accepted, 0 means rejected
<code>tp0</code>	Analysis Parameter	<code>integer</code>	Start of the rising edge
<code>detector</code>	Metadata	<code>integer</code>	unique ID for each detector
<code>run_number</code>	Metadata	<code>integer</code>	unique ID for each run
<code>id</code>	Metadata	<code>integer</code>	unique ID for each data point

Data saved as multiple
HDF5 files (2Gb each)

- Derived from the analysis of MAJORANA DEMONSTRATOR final result
 - Phys. Rev. Lett. **130**, 062501
- Run-by-run, detector-by-detector tuning
- Physics analysis is a “labeling process”
 - AI company hires people to label images/texts
 - Physics experiment educate PhD students/postdocs to “label” our data

Accessing Full Dataset

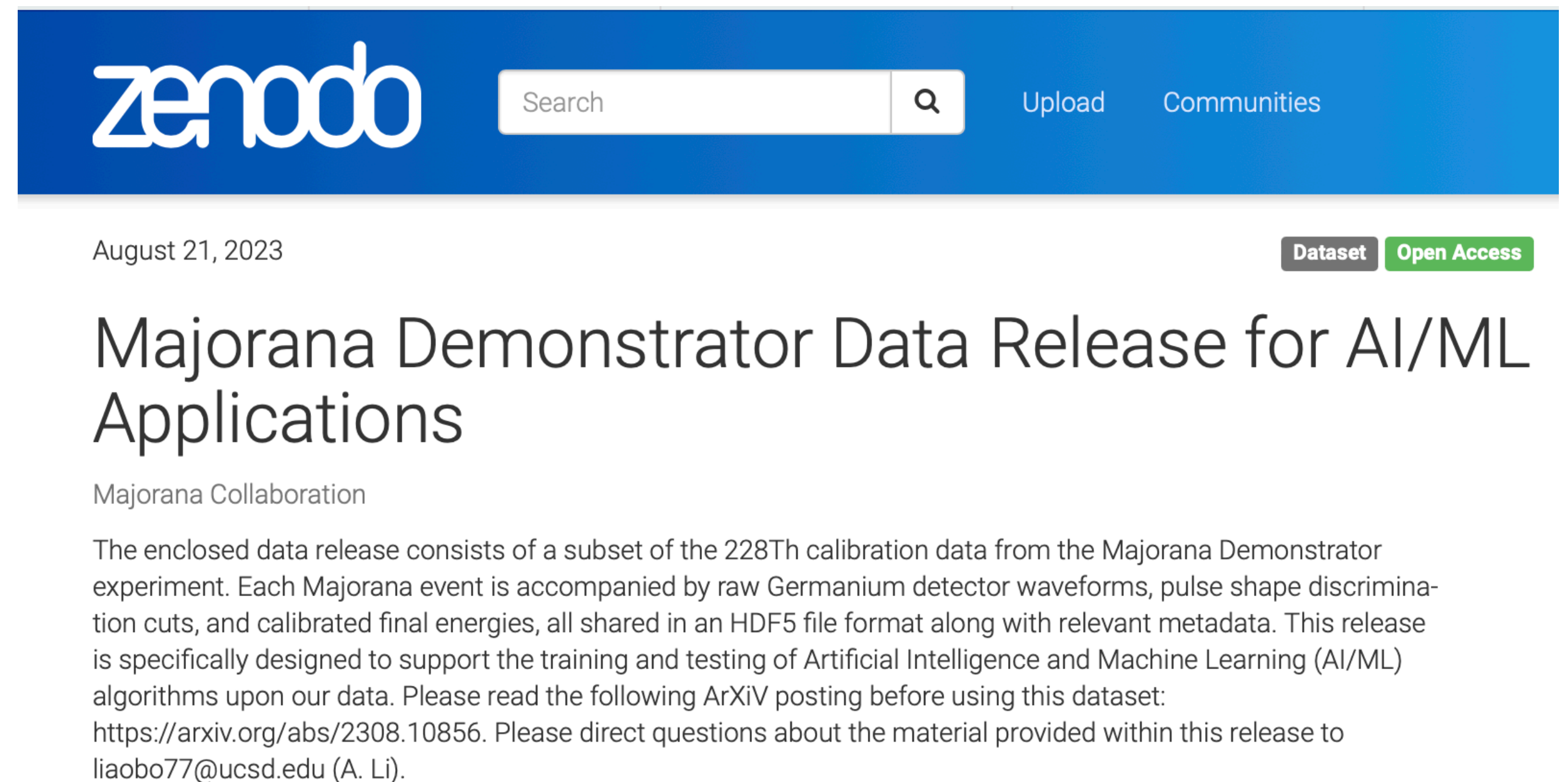
- The full dataset will be available on DataPlanet
 - UCSD-built dataverse system
- Physics dataverse access link:
<https://dataplanet.ucsd.edu/dataverse/physics>
- **Majorana Demonstrator** dataset under the physics dataverse

The screenshot shows the UC San Diego DataPlanet interface. The top navigation bar includes the logo, a search dropdown, and links for 'User Guide' and 'Support'. On the left, there are filters for 'Dataverses (3)', 'Datasets (2)', and 'Files (10)'. Below these are filters for 'Dataverse Category' (Department, Organization or Institution, Research Group), 'Publication Year' (2023), 'Publication Status' (Published, Unpublished, Draft), 'Author Name' (Nealey, Isaac), and 'Subject'. The main content area displays '1 to 5 of 5 Results'. The first result is 'Majorana Demonstrator Experiment' (Draft, Unpublished) dated Jul 31, 2023, with a description of the data release. The second result is 'Physics' (Unpublished) dated Jul 31, 2023, with a description of the dataverse content. The third result is 'TLS Panorama Demo' (Unpublished) dated Mar 6, 2023, with a description of the panoramic images.

**Not available now: DataPlanet under technical difficulties
Expected to come back by end of September.**

Accessing Partial Dataset

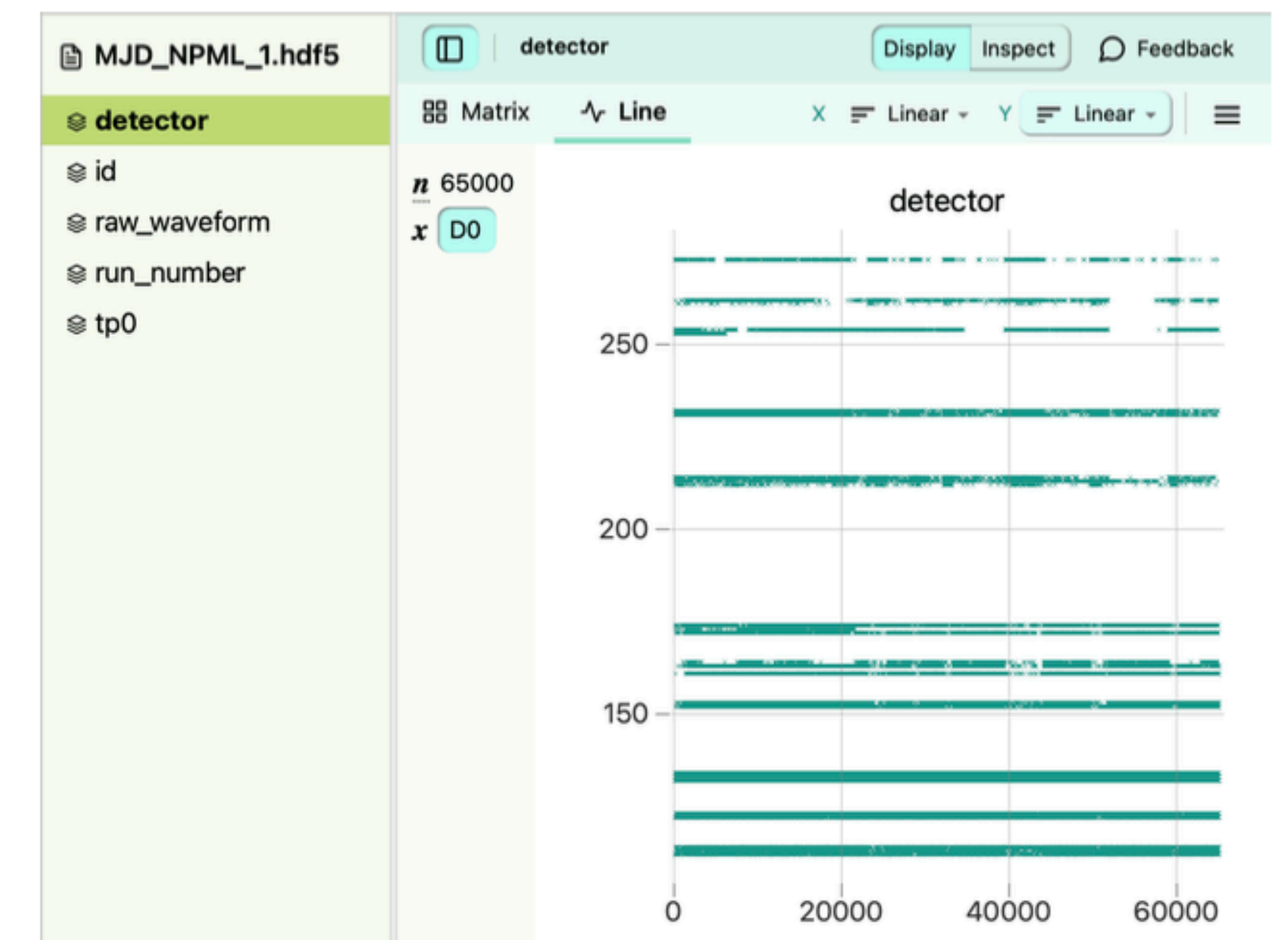
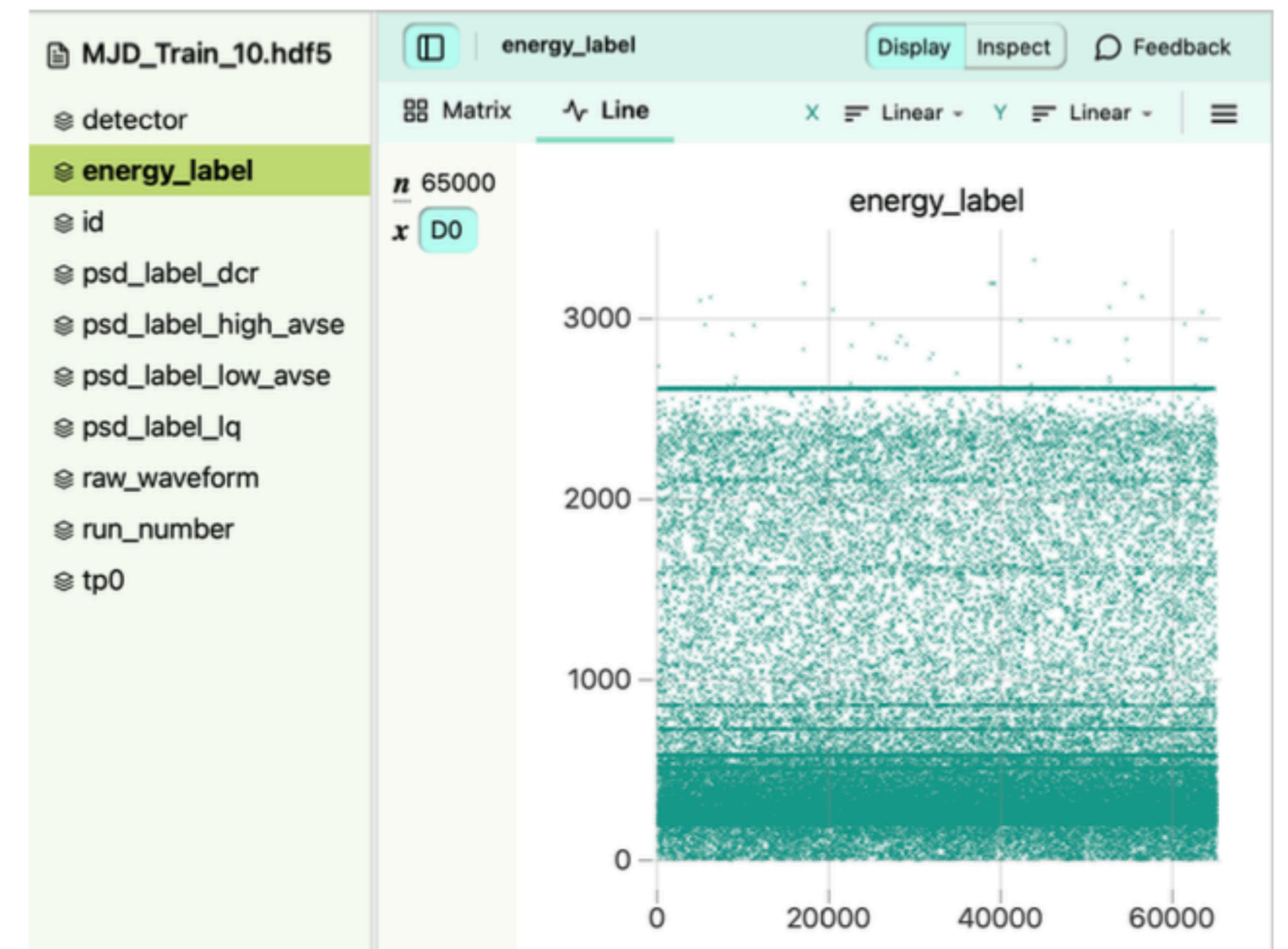
- Zenodo partial release of the same dataset
 - ~50% size of the full dataset
- Access: <https://doi.org/10.5281/zenodo.8257027>



The screenshot shows the Zenodo website interface. At the top, there is a blue header with the Zenodo logo, a search bar, and links for 'Upload' and 'Communities'. Below the header, the date 'August 21, 2023' is displayed on the left, and 'Dataset' and 'Open Access' tags are on the right. The main title of the dataset is 'Majorana Demonstrator Data Release for AI/ML Applications', followed by the author 'Majorana Collaboration'. The description states: 'The enclosed data release consists of a subset of the 228Th calibration data from the Majorana Demonstrator experiment. Each Majorana event is accompanied by raw Germanium detector waveforms, pulse shape discrimination cuts, and calibrated final energies, all shared in an HDF5 file format along with relevant metadata. This release is specifically designed to support the training and testing of Artificial Intelligence and Machine Learning (AI/ML) algorithms upon our data. Please read the following ArXiv posting before using this dataset: https://arxiv.org/abs/2308.10856. Please direct questions about the material provided within this release to liaobo77@ucsd.edu (A. Li).'

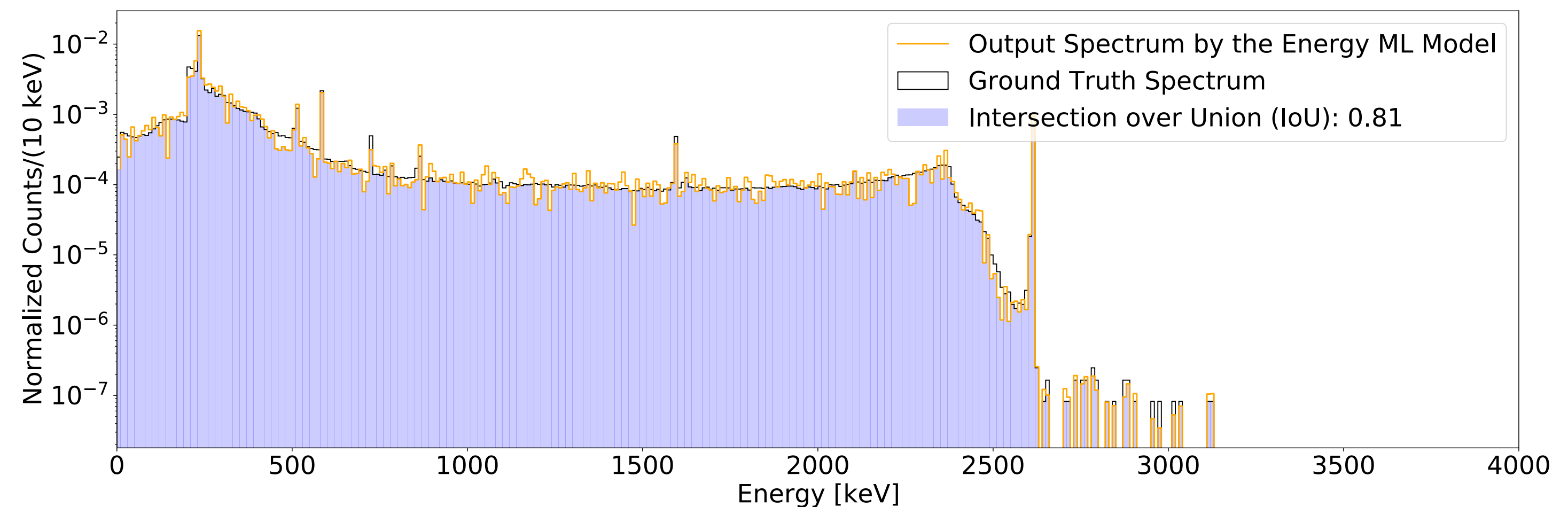
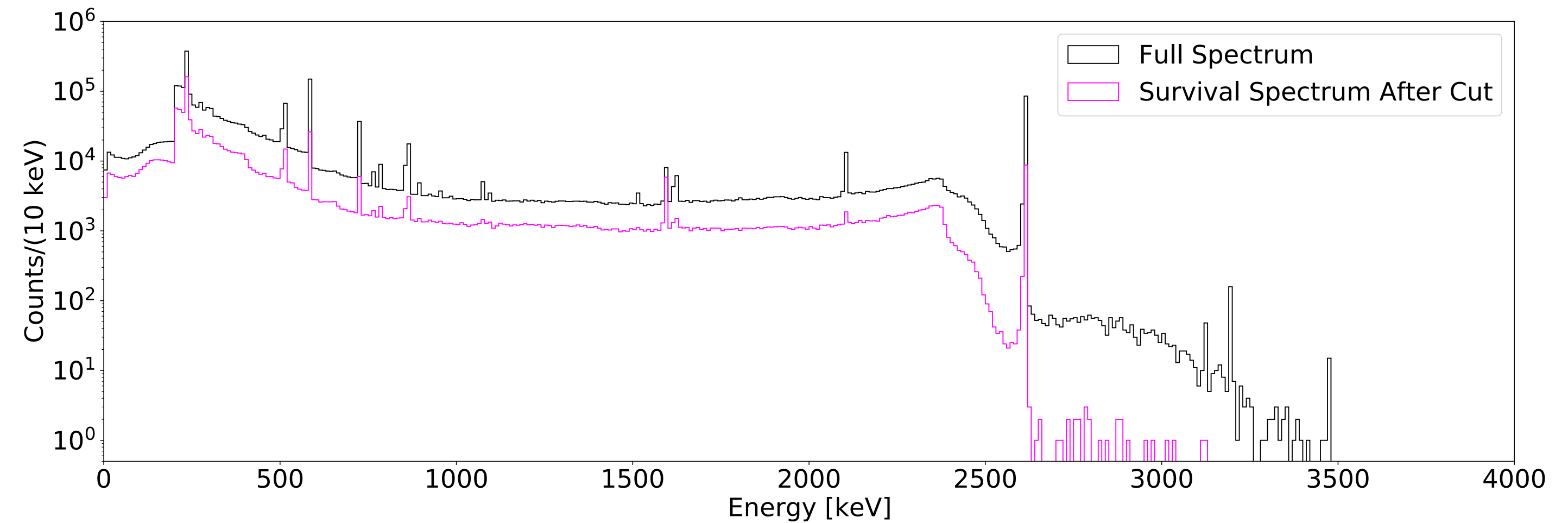
Majorana Data Challenge

- A small subset of the data release was converted to a dedicated NPML dataset
 - Contains all information except analysis labels
- Participants of the MAJORANA data challenge will build two models to reproduce these analysis labels



Majorana Data Challenge

- **Classification Model:** classify and collect “clean event” that passes all `psd_label_*`
- **Regression Model:** Reconstruct energy label of all clean events into an energy spectrum
- Evaluate performance on histogram IoU
- Challenges to build these model:
 - Long-tailed dataset
 - Multi-tasking
 - Information highly concentrated on the rising edge



Majorana Data Challenge

- Challenge opens until May 2024
- To participate, send the following to liaobo77@ucsd.edu

Dimension		Content		Note
id	0,1,2	...	159672,159673,159674	
predicted psd_label	0,1,0	...	1,0,1	1 only if all 4 predicted psd_labels are 1
predicted energy_label	3.25,1923.74,323.64	...	582.5, 938, 812.74	in keV

- Info listed in the table above over the NPML dataset (as numpy array)
- A short analysis paper (4-pages) describing your model
- The people with highest classification accuracy and histogram IoU will get:
 - A Plaque
 - Invited talk at NPML 2024
 - Award stipend (subject to funding)

Future NPML Challenges

- In the future, we plan to prepare & release multiple datasets for this type of challenges
- We will design a dedicated data release format and framework
- This MAJORANA DEMONSTRATOR dataset will merge into that framework when it comes

Important Information

- The information I presented today is summarized in <https://arxiv.org/abs/2308.10856>
- Access Zenodo partial dataset: <https://doi.org/10.5281/zenodo.8257027>
- Access DataPlanet full dataset (Available in October): <https://dataplanet.ucsd.edu/dataverse/physics>
- I'm hiring graduate students & postdocs, please email me at liaobo77@ucsd.edu