ML for anomaly detection and background discrimination in LZ

Maris Arthurs *on behalf of LZ collaboration*

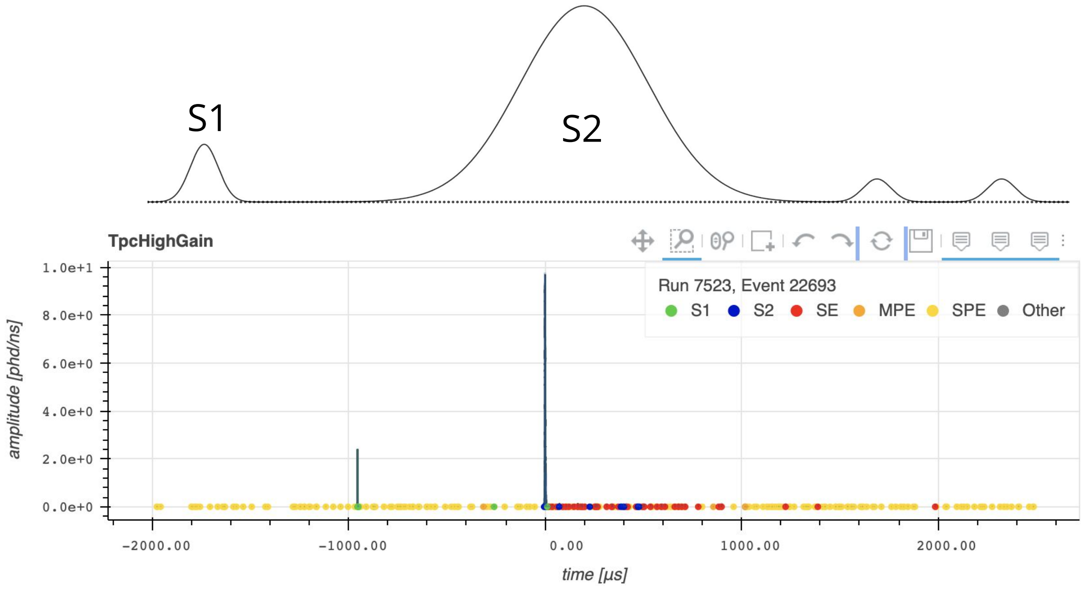
CoSSURF 2024

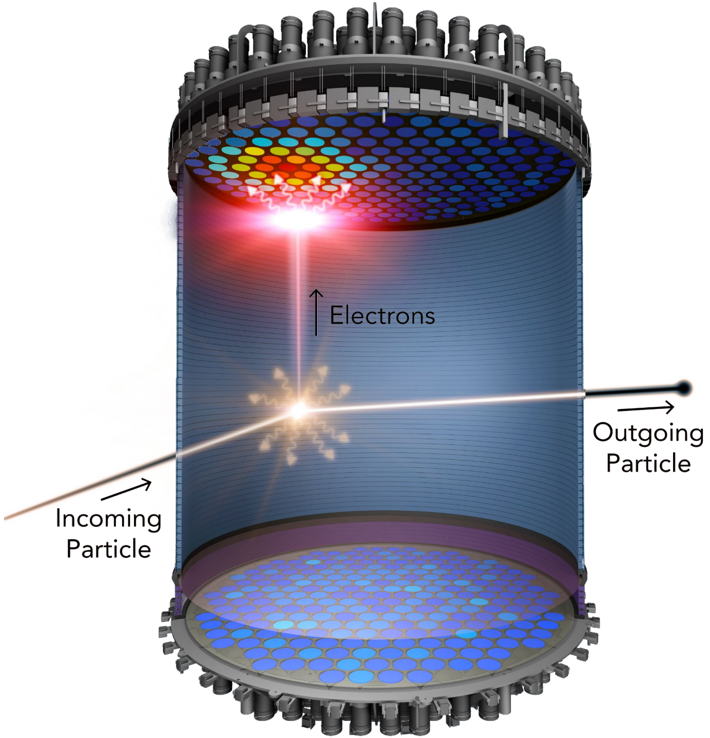
May 15, 2024

U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

The LZ Detector



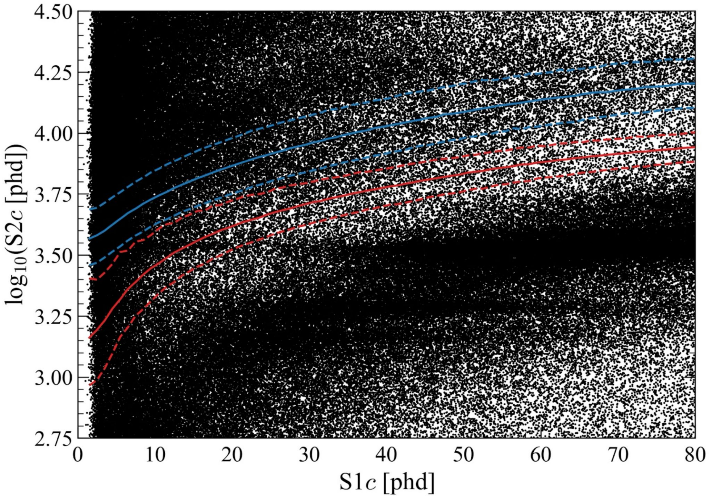


Dual Phase TPC Detector

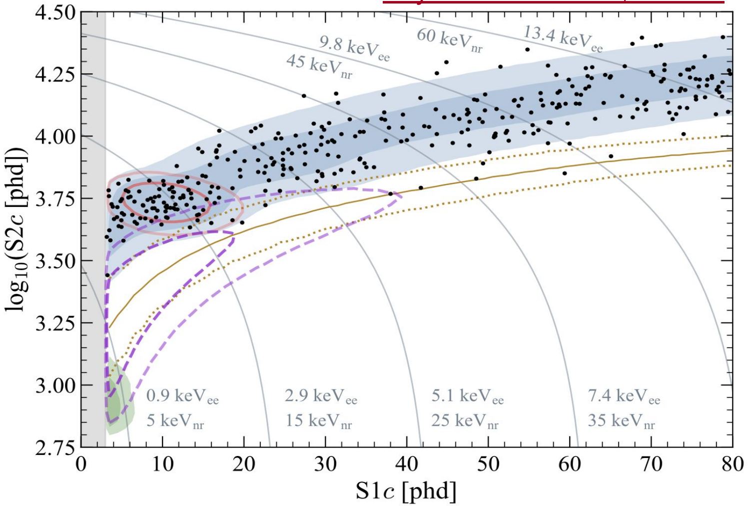
- Primary scintillation (light) → S1
- Secondary scintillation (from charge) → S2
- Radial position from top PMT array S2 pattern
- Z position determined from the drift time

Event Selection

looking for needle in a haystack



SR1 results - Phys. Rev. Lett. **131**, 041002



- Single Scatter Event Selection
- Temporal Uniformity Cuts (analysis hold-off)
- Spatial Uniformity Cut
- Accidental coincidence of S1 and S2 pulses
- Skin and OD veto cuts

Using ML to find anomalous data and discriminate backgrounds

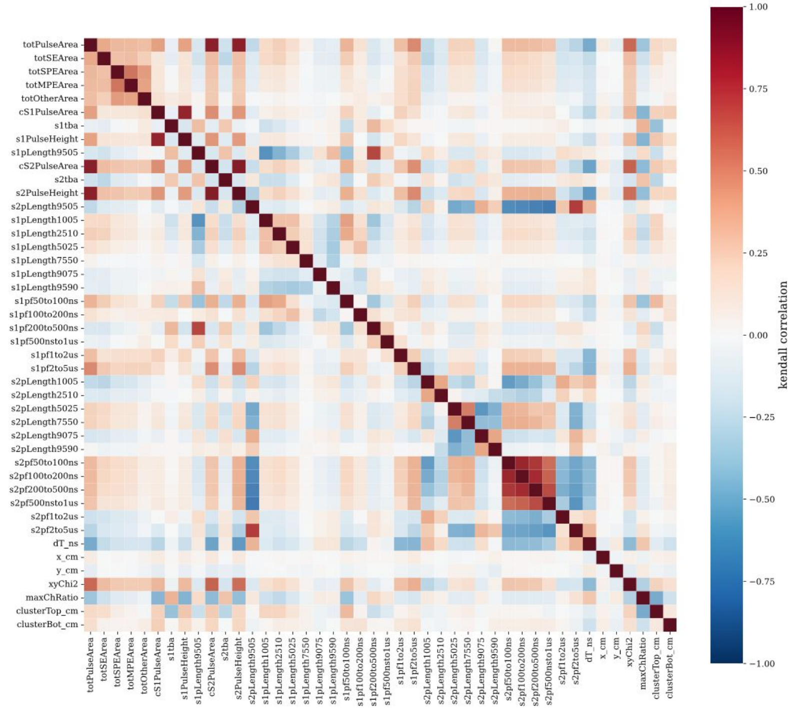
SLAC

1. **Anomaly finding with Dimensional Reduction**
    - Map *N dimensional* (~30 features) data to *2D representation*
    - **Why? –** Outliers in multidimensional feature spaces are difficult to detect visually.
    - **Goal –** quickly identify and study (*not cut*) outlier events → *tune the data reconstruction software, investigate anomalous populations*
    - **UMAP**: Uniform Manifold Approximation and Projection & **tSNE**: T-distributed Stochastic Neighbor Embedding → nonlinear dimensionality reduction method: tend to preserve *local* structure as well as *global* structure
2. **Anomalous S2 waveforms with autoencoders**
    - **Why? –** low-level waveforms capture anomalous features that are washed out during reconstruction process. S2 waveforms have a lot of relevant physical information
    - **Goal –** identify anomalous S2 pulses, has the potential of resolving overlapping multiple scatters! work in progress
3. **Discrimination of multiple scatter single ionisation (MSSI) background**
    - A boosted decision tree model was successfully used in LZ EFT studies to reduce a problematic background

Multidimensional Data

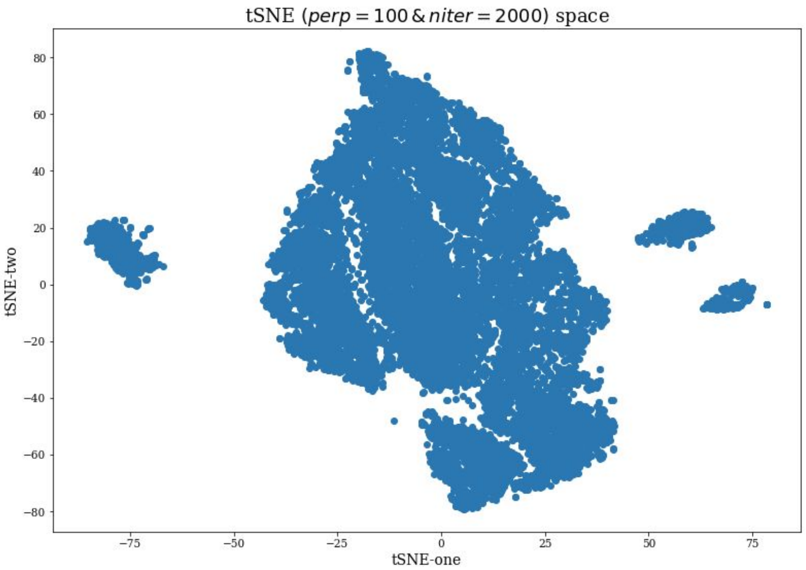
Hybrid data containing both pulse level and event level information

- **Event level Features**
  - Total area of all pulses in an event
  - Total area of different types of pulses in an event: single electrons (SE), single photo electrons (SPE)...
- **S1 Pulse Features:**
  - S1 pulse length and pulse area
  - S1 pulse shape features
- **S2 Pulses Features:**
  - S2 pulse length and pulse area
  - S2 pulse shape features
- **Other Features:**
  - S1 and S2 top bottom asymmetry (TBA)
  - Drift Time
  - XY information
  - S1 top array and bottom array cluster size
  - …

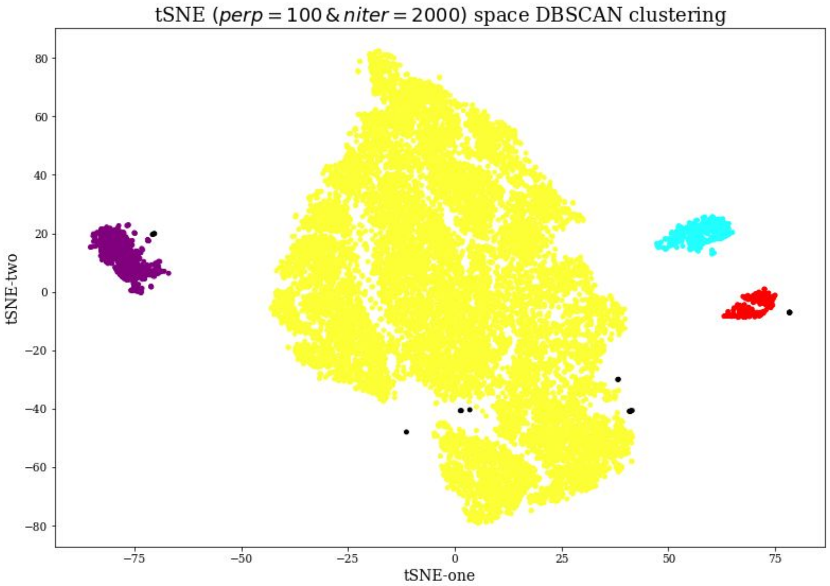


- Features are preprocessed to reduce cross correlation

Clustering in tSNE Space of Simulated Data

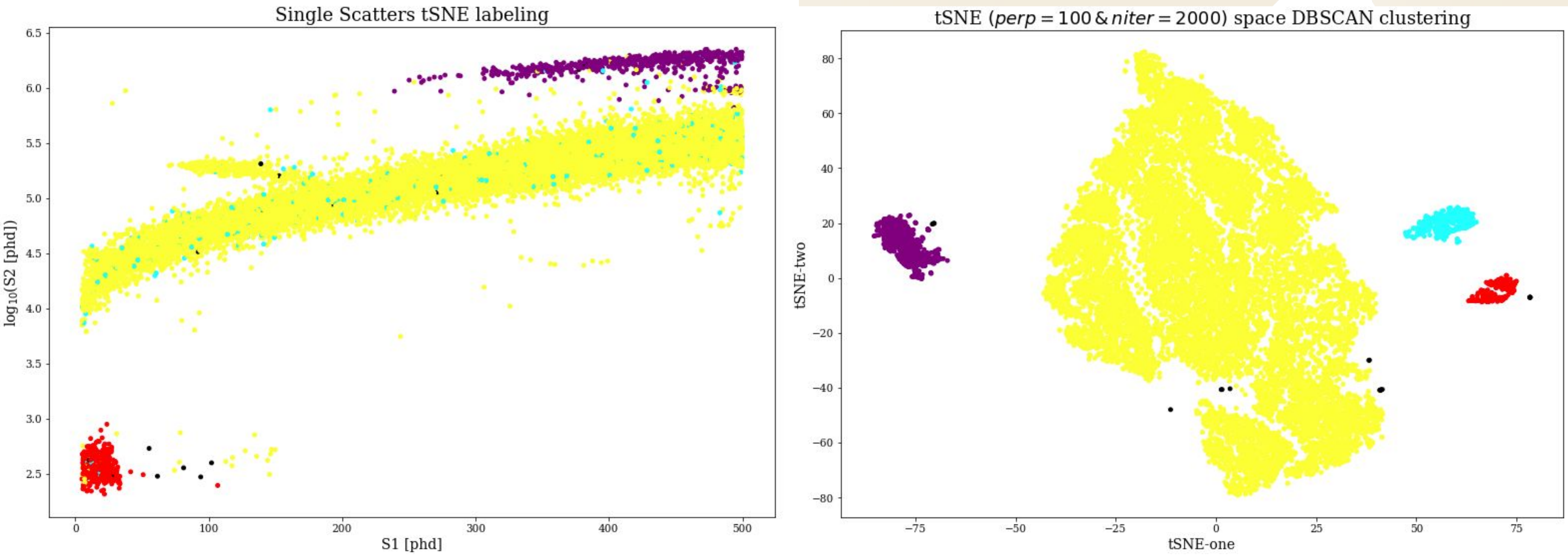


DBSCAN



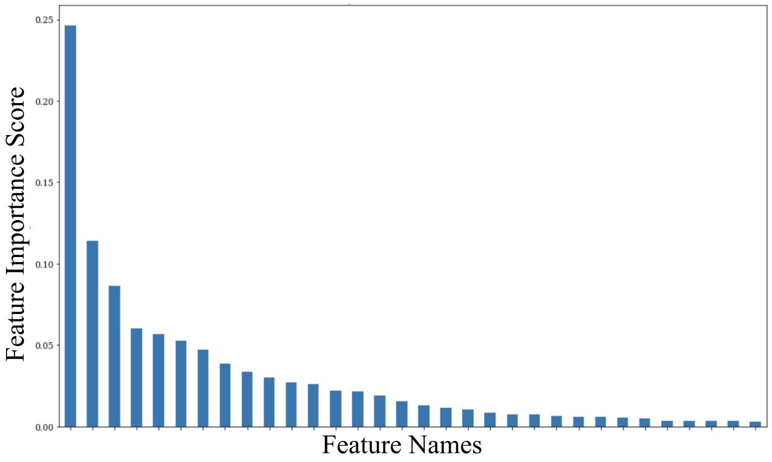
- Density based clusterization in tSNE space
- Black events classified as noise → null cluster

Clustering in tSNE Space of Simulated Data

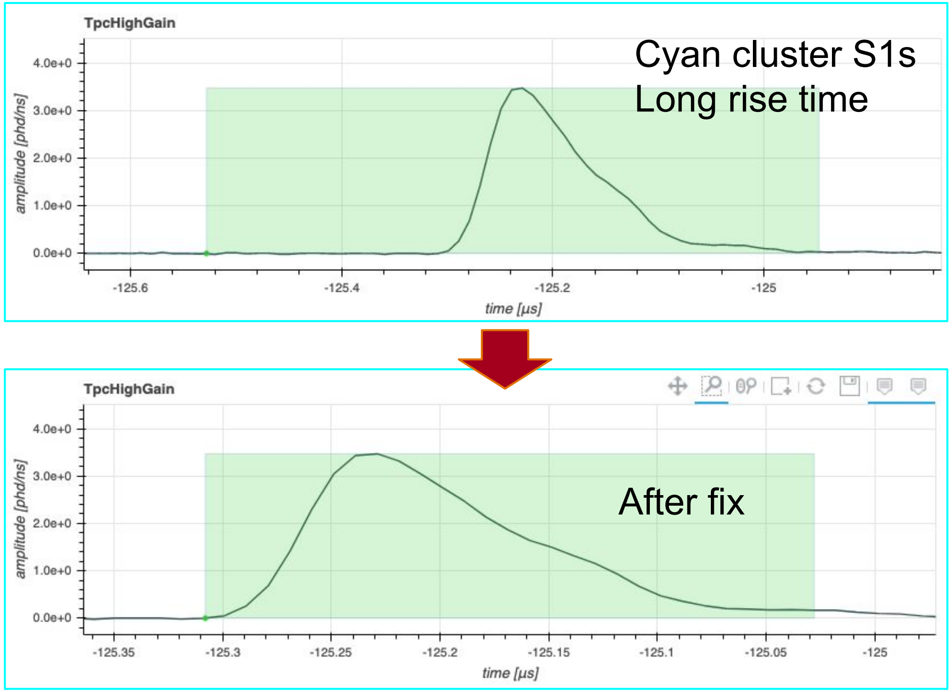


- Red population → Instrumental background
- Purple population → Simulation bug
- Cyan population → Reconstruction bug
- Yellow population → dominant cluster

Feature Importances of Clusters

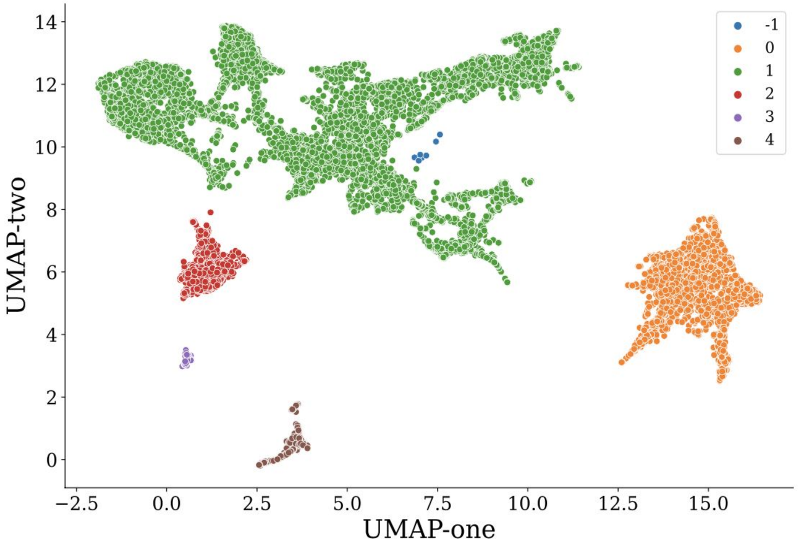


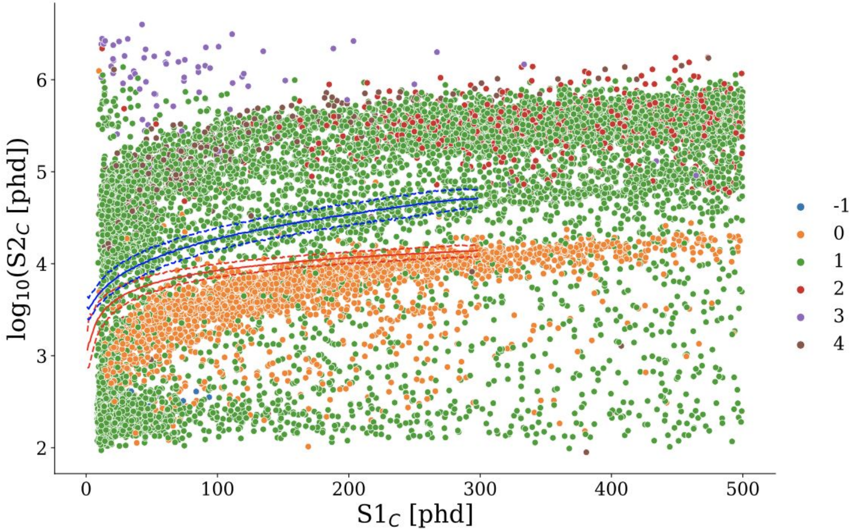
- Identify feature importances of clusters using RandomForestClassifier
  - What makes a cluster different?
  - Which parameter space is effective to make cuts?
- Improved our data reconstruction process



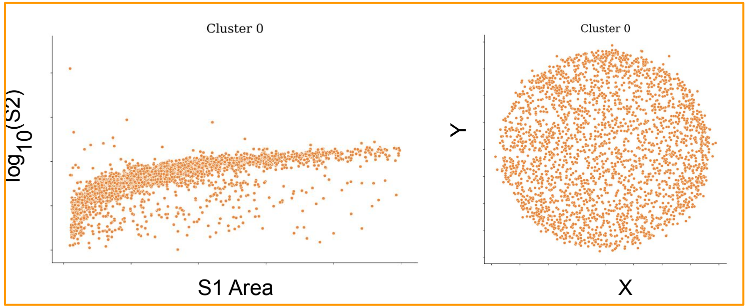
Early successes

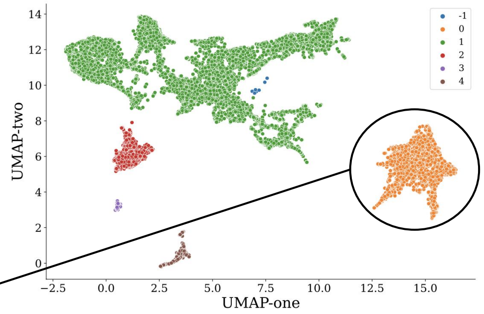
Application on First Science (SR1) Data



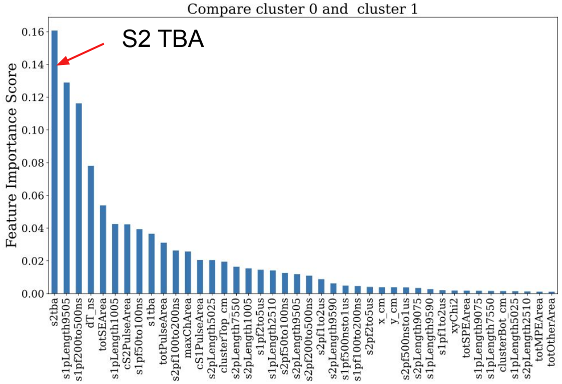


Over-Anode Gas Events





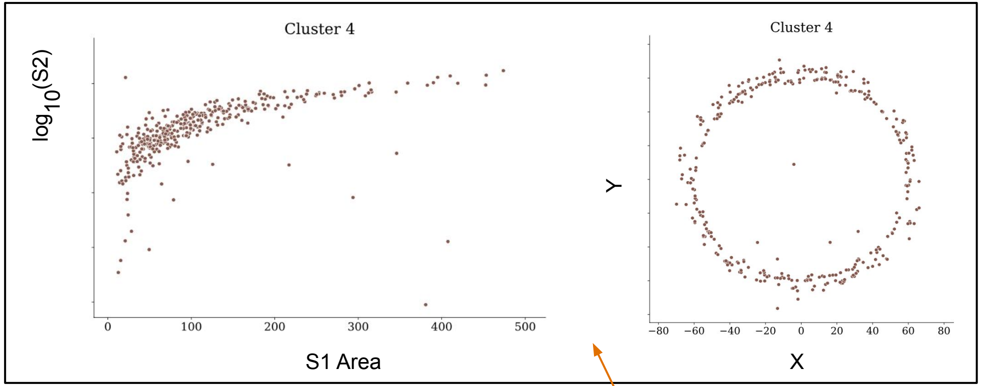
Comparing orange to dominant green population



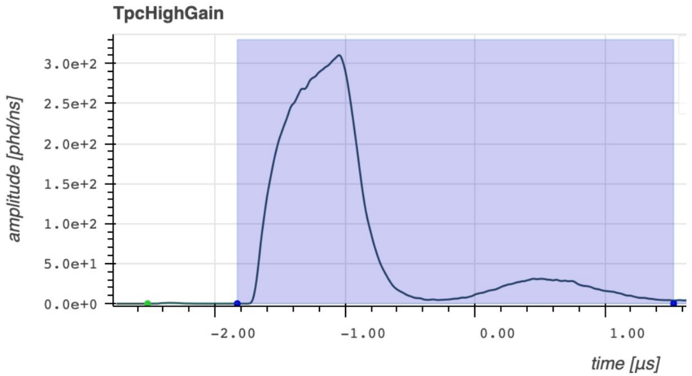
Cluster 0 (orange) found to be gas events originating above the anode → *detector effect*

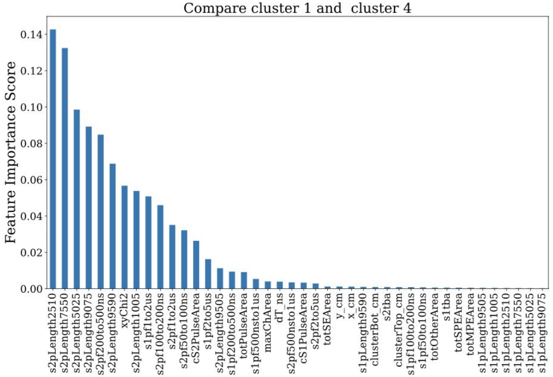
- *<u>Feature importances tool</u> allows identification of relevant RQ where this population can be further investigated*
- Found to have large S2 TBA → helped to develop conventional cuts to discriminate this background

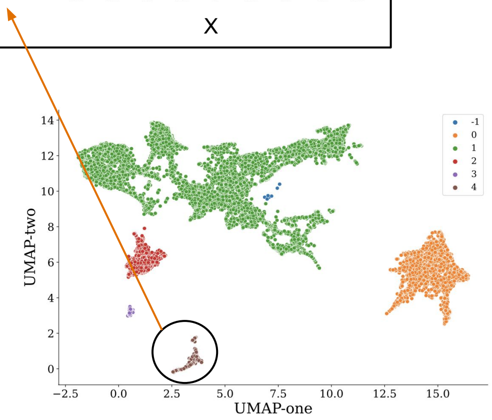
Gate Photoionization Events



- Photoionization of the wire electrodes from S2 light
- Originates near the walls in extraction region
- *Detector effect*







Using ML to find anomalous data and discriminate backgrounds

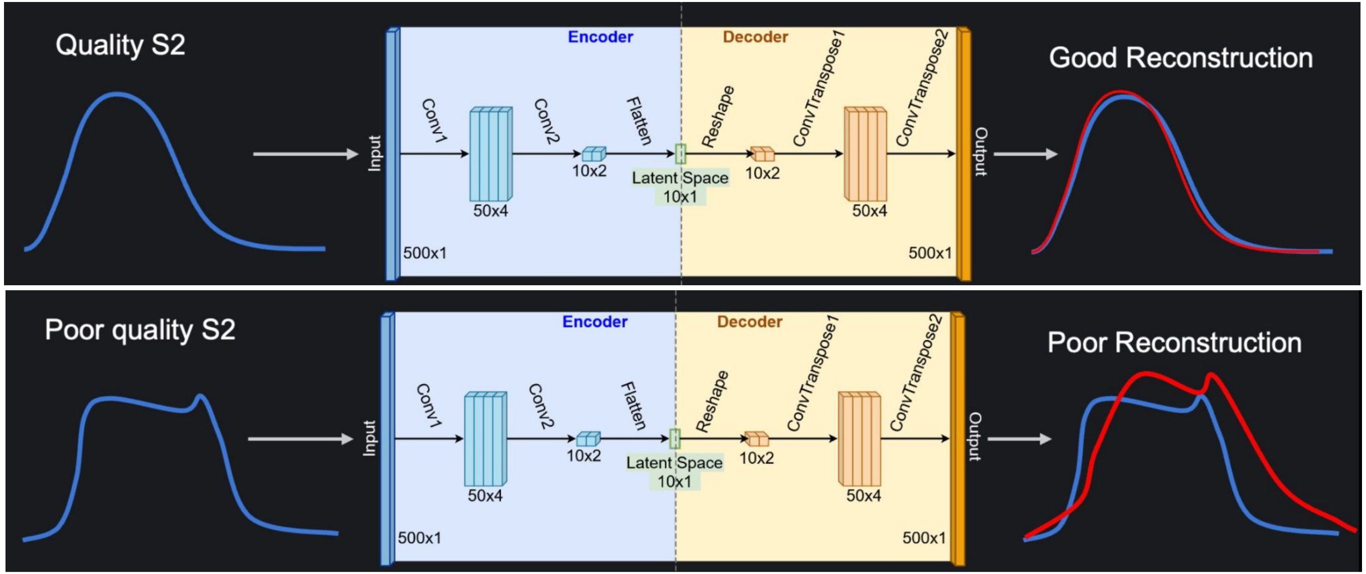
SLAC

1. **Anomaly finding with Dimensional Reduction**
    - Map *N dimensional* (~30 features) data to *2D representation*
    - **Why? –** Outliers in multidimensional feature spaces are difficult to detect visually.
    - **Goal –** quickly identify and study (*not cut*) outlier events → *tune the data reconstruction software, investigate anomalous backgrounds*
    - **UMAP**: Uniform Manifold Approximation and Projection & **tSNE**: T-distributed Stochastic Neighbor Embedding → nonlinear dimensionality reduction method: tend to preserve *local* structure as well as *global* structure
2. **Anomalous S2 waveforms with autoencoders**
    - **Why? –** low-level waveforms capture anomalous features that are washed out during reconstruction process. S2 waveforms have a lot of relevant physical information
    - **Goal –** identify anomalous S2 pulses, has the potential of resolving overlapping multiple scatters! work in progress
3. **Discrimination of multiple scatter single ionisation (MSSI) background**
    - A boosted decision tree model was successfully used in LZ EFT studies to reduce a problematic background

S2 Waveform Anomalies with Autoencoders

Work by Tyler Anderson

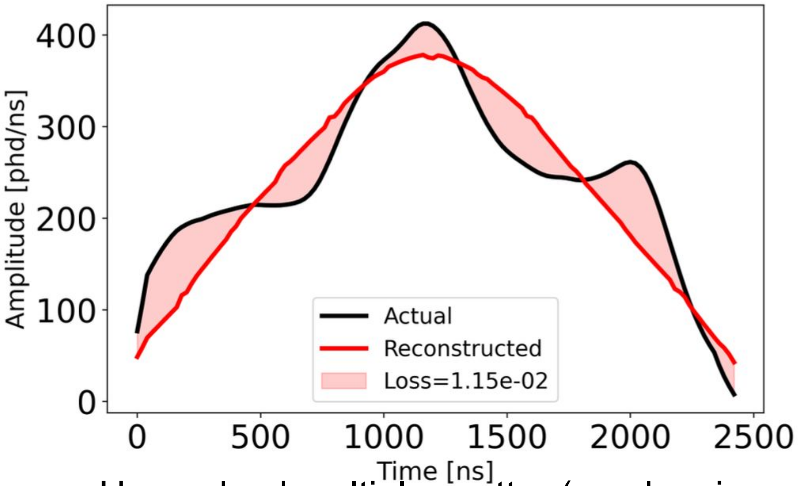
- Using signal waveforms allows to capture abnormal features washed out during reconstruction
- S2 waveforms encode a lot of relevant information due to electron cloud movement in drift region

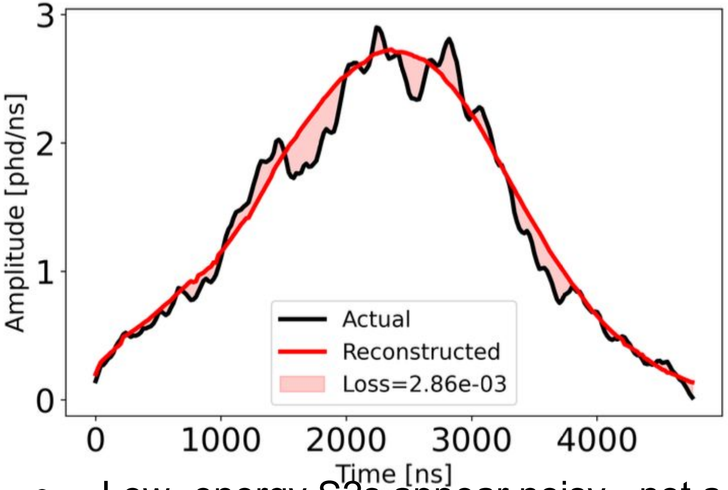


- A neutral network: trained to recreate their inputs
  - Encoder: compress the input to lower dimensional latent space
  - Decoder: convert from latent space back into original input
  - Reconstruction error can be a proxy for how different an input looks from training set.

Anomalous S2s found by the Autoencoder

Work by Tyler Anderson





- Unresolved multiple scatter (overlapping multiple S2s)
- potential background for DM searches as it would be classified as single scatter

- Low- energy S2s appear noisy - not a problem!

Using ML to find anomalous data and discriminate backgrounds

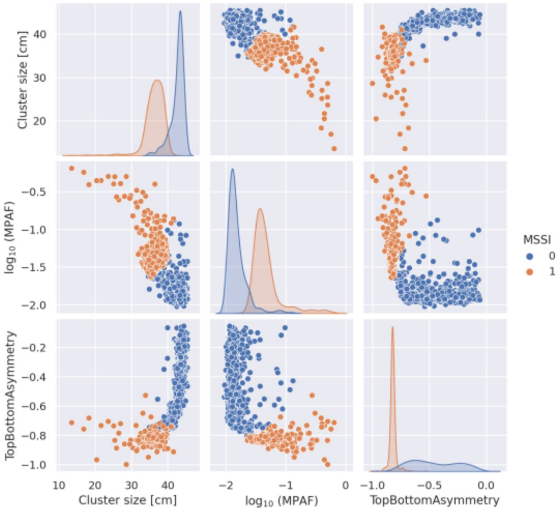
SLAC

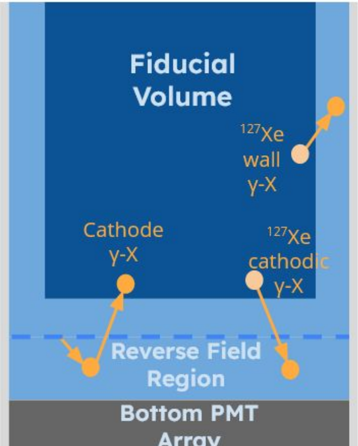
1. **Anomaly finding with Dimensional Reduction**
   - Map *N dimensional* (~30 features) data to *2D representation*
   - **Why? –** Outliers in multidimensional feature spaces are difficult to detect visually.
   - **Goal –** quickly identify and study (*not cut*) outlier events → *tune the data reconstruction software, investigate anomalous backgrounds*
   - **UMAP**: Uniform Manifold Approximation and Projection & **tSNE**: T-distributed Stochastic Neighbor Embedding → nonlinear dimensionality reduction method: tend to preserve *local* structure as well as *global* structure
2. **Anomalous S2 waveforms with autoencoders**
   - **Why? –** low-level waveforms capture anomalous features that are washed out during reconstruction process. S2 waveforms have a lot of relevant physical information
   - **Goal –** identify anomalous S2 pulses, has the potential of resolving overlapping multiple scatters! work in progress
3. **Discrimination of multiple scatter single ionisation (MSSI) background**
   - A boosted decision tree model was successfully used in LZ EFT studies to reduce a problematic background

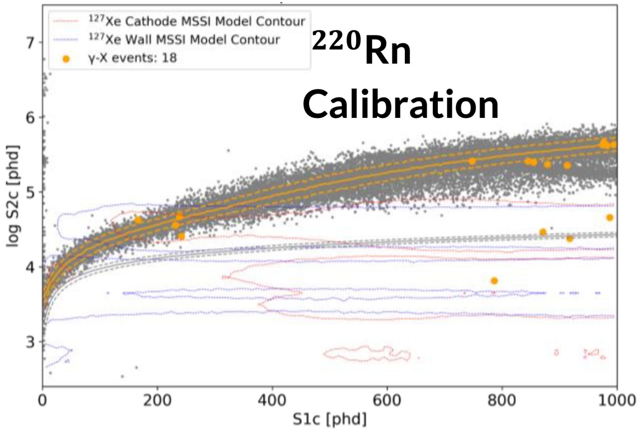
Multiple Scatter Single Ionisation: γ-X Events

Work by Chamindu Amarasinghe et.al

- **γ-X Events**: Multiple S1-contributing scatters, one S2-contributing scatter
  - WIMP-like ER Pathologies that can mimic DM signal
- **Boosted Decision Tree** trained on simulated data and tested on calibration and side band datasets
  - reduced quantities used in classification: *S1 cluster size, Max Peak Area Fraction, Top Bottom Asymmetry, S1 Area, log(S2 Area), Radius, Drift Time*







$\downarrow$ P, $\rightarrow$ T	SS	γ-X
SS	99.997 ± 0.005	0.4 ± 1.2
γ-X	0.003 ± 0.005	99.6 ± 1.2

First constraint EFT couplings -
arXiv:2312.02030

Summary

- We Demonstrate the utility of a **general purpose anomaly finder** based on unsupervised dimensionality reduction → tSNE, UMAP
    - **Interpretability** (Feature importances) →Using Random Forest Classifier
    - Applications: Investigating anomalous events, Data quality checks (simulations, real), Tuning of reconstruction and classification algorithms
- Application of Autoencoders on S2 waveforms has the potential to identifying rare waveform pathologies such as the unresolved multiple scatters. work in progress
- A boosted decision tree algorithm was successfully applied in EFT WIMP search study to discriminate the problematic multiple scatter single ionisation (MSSI) background

LZ (LUX-ZEPLIN) Collaboration, 38 Institutions

@lzdarkmatter

https://lz.lbl.gov/

250 scientists, engineers, and technical staff

- Black Hills State University
- Brookhaven National Laboratory
- Brown University
- Center for Underground Physics
- Edinburgh University
- Fermi National Accelerator Lab.
- Imperial College London
- King's College London
- Lawrence Berkeley National Lab.
- Lawrence Livermore National Lab.
- LIP Coimbra
- Northwestern University
- Pennsylvania State University
- Royal Holloway University of London
- SLAC National Accelerator Lab.
- South Dakota School of Mines & Tech
- South Dakota Science & Technology Authority
- STFC Rutherford Appleton Lab.
- Texas A&M University
- University of Albany, SUNY
- University of Alabama
- University of Bristol
- University College London
- University of California Berkeley
- University of California Davis
- University of California Los Angeles
- University of California Santa Barbara
- University of Liverpool
- University of Maryland
- University of Massachusetts, Amherst
- University of Michigan
- University of Oxford
- University of Rochester
- University of Sheffield
- University of Sydney
- University of Texas at Austin
- University of Wisconsin, Madison
- University of Zürich

US Europe Asia Oceania

Science and Technology Facilities Council

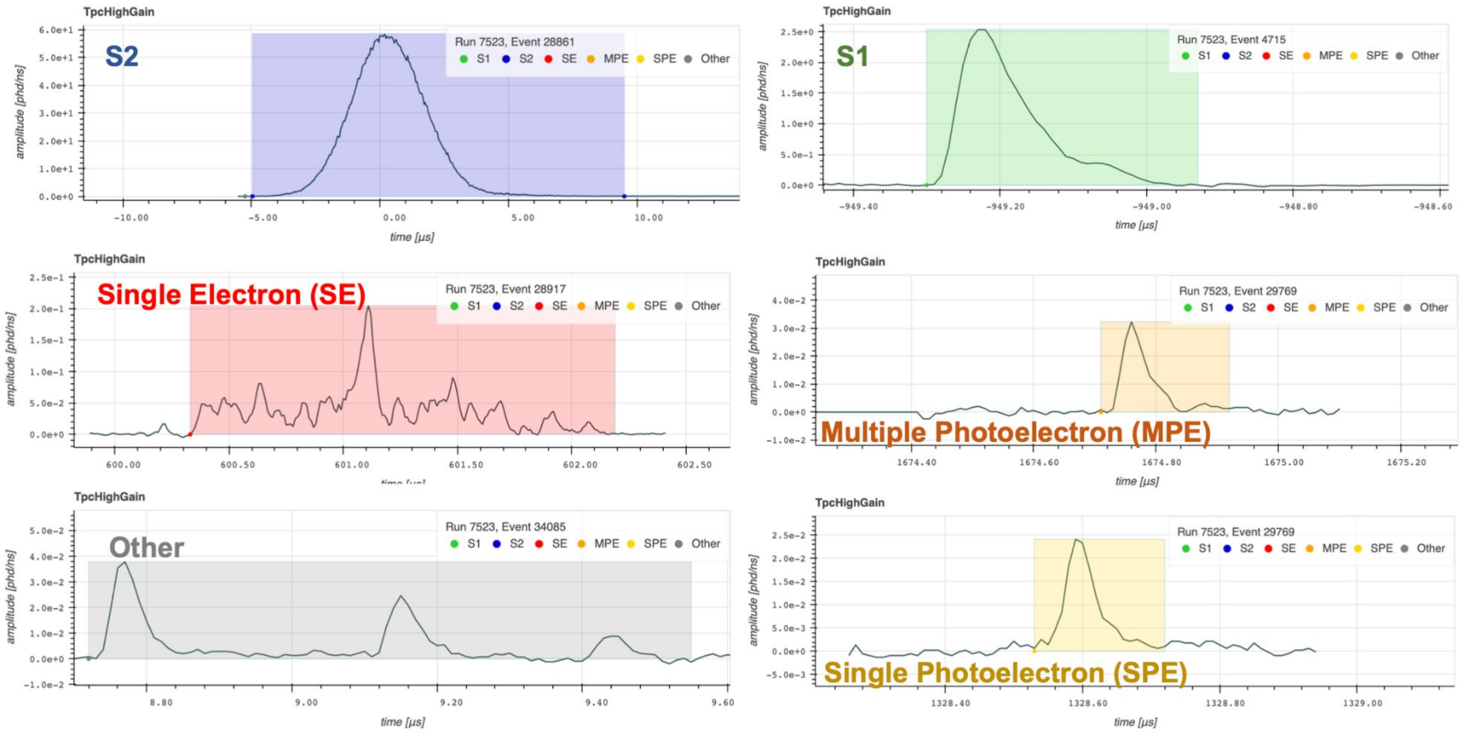
FCT Fundação para a Ciência e a Tecnologia

Sanford Underground Research Facility
South Dakota Science and Technology Authority

Institute for Basic Science

Thanks to our sponsors and participating institutions!

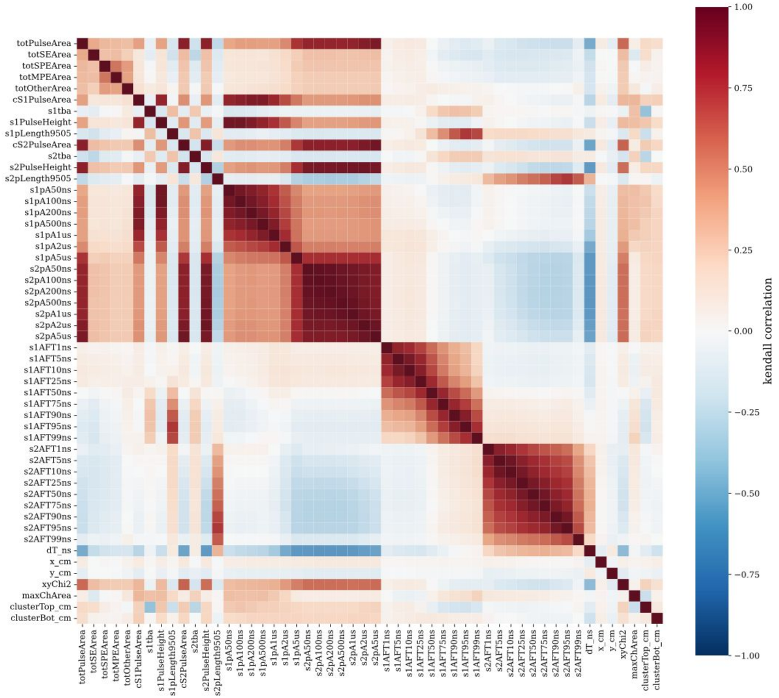
Types of Pulse Waveform



Multidimensional Data

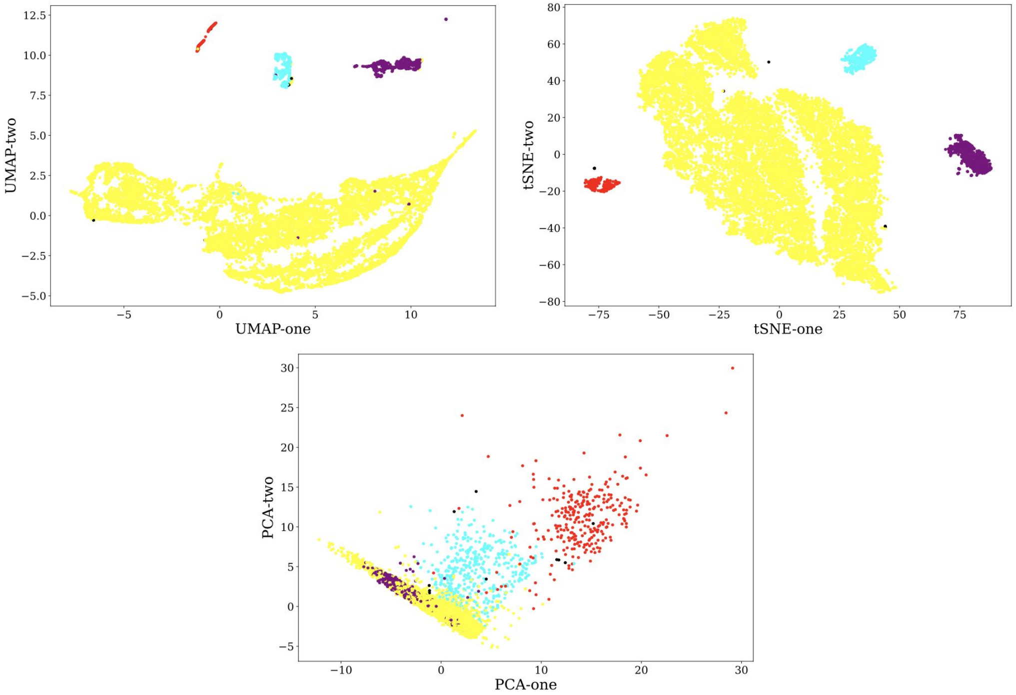
Hybrid data containing both pulse level and event level information

- **Event level Features**
  - Total area of all pulses in an event
  - Total area of different types of pulses in an event: single electrons (SE), single photo electrons (SPE)...
- **S1 Pulse Features:**
  - S1 pulse length and pulse area
  - S1 pulse shape features
- **S2 Pulses Features:**
  - S2 pulse length and pulse area
  - S2 pulse shape features
- **Other Features:**
  - S1 and S2 top bottom asymmetry (TBA)
  - Drift Time
  - XY information
  - S1 top array and bottom array cluster size
  - …



- Features are preprocessed to reduce cross correlation

Comparison of UMAP, tSNE, and PCA



- Red population → Instrumental background
- Purple population → Simulation bug
- Cyan population → Reconstruction bug