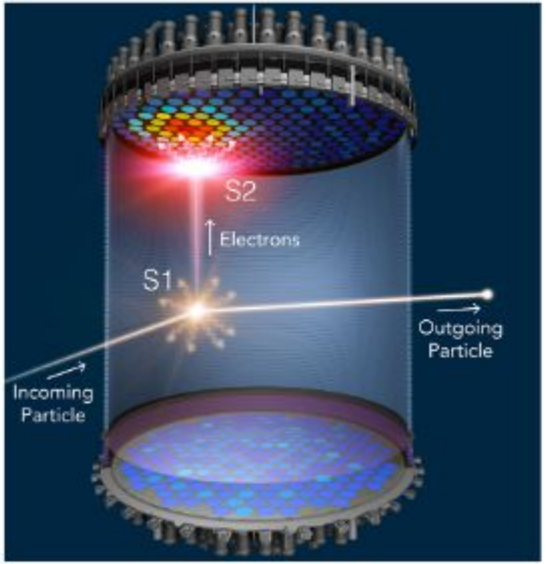
Classification of pulses in the LUX-ZEPLIN dark matter detector.

Inês Sequeira nº201828283321
Maria Moita nº2017249626

Description of the project

- Rare event observatory with an ultra-low background dual-phase xenon time projection chamber (TPC) designed for particle dark matter searches.



Signals considered in our classification:

- S1 - Scintillation signal
- S2 - Electroluminescence signals
- SE - single electron signals
- Other signals

Description of the data

1000000 pulses; 20 features

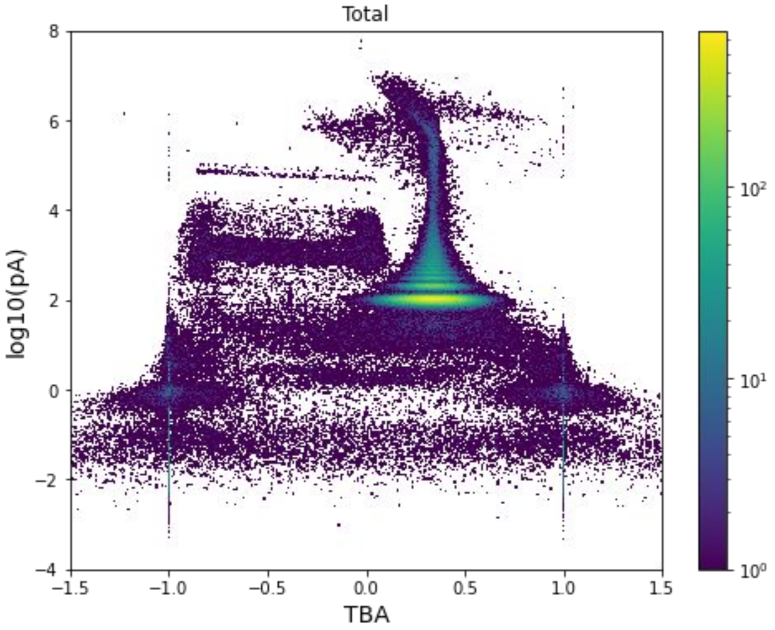
Most relevant:

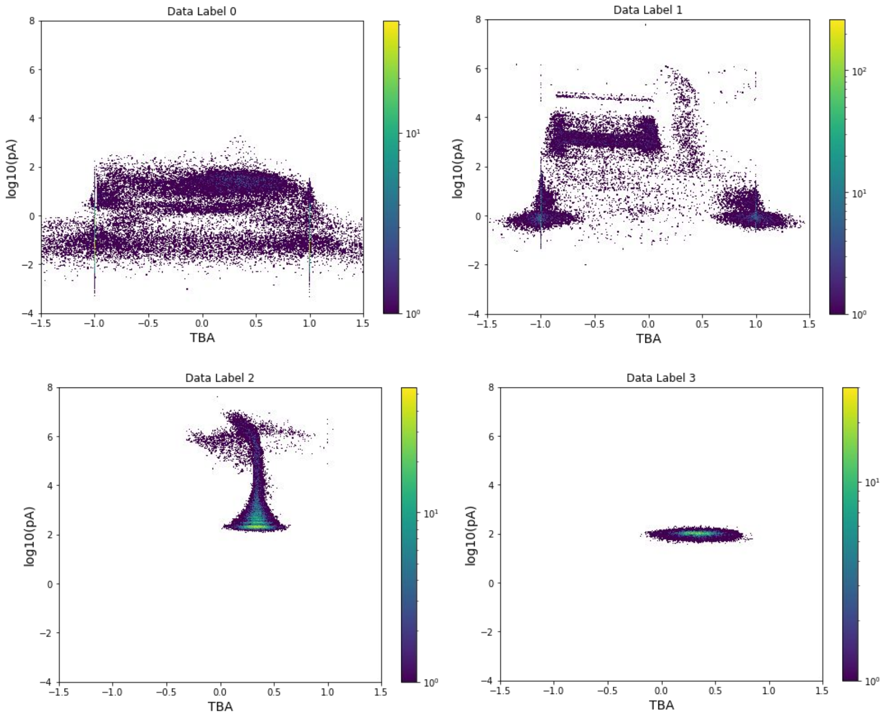
- **pA:** Total area of summed pod from pulse start to end
- **pH:** Max amplitude of summed pod within pulse
- **pL:** Length of the pulse
- **TBA:** Ratio of total area of top PMTs vs. total area of bottom PMTs
- **aft5:** Time at which summed pod reaches 5% of total area
- **pF100:** Fraction of summed pod area in 100ns window at start of pulse relative to total pulse area; window defined from aft5 to 100ns after
- **pF500:** Fraction of summed pod area in 500ns window at start of pulse relative to total pulse area; window defined from aft5 to 500ns after
- **pF5k:** Fraction of summed pod area in 5μs window at start of pulse relative to total pulse area; window defined from aft5 to 5μs after.

Data plots

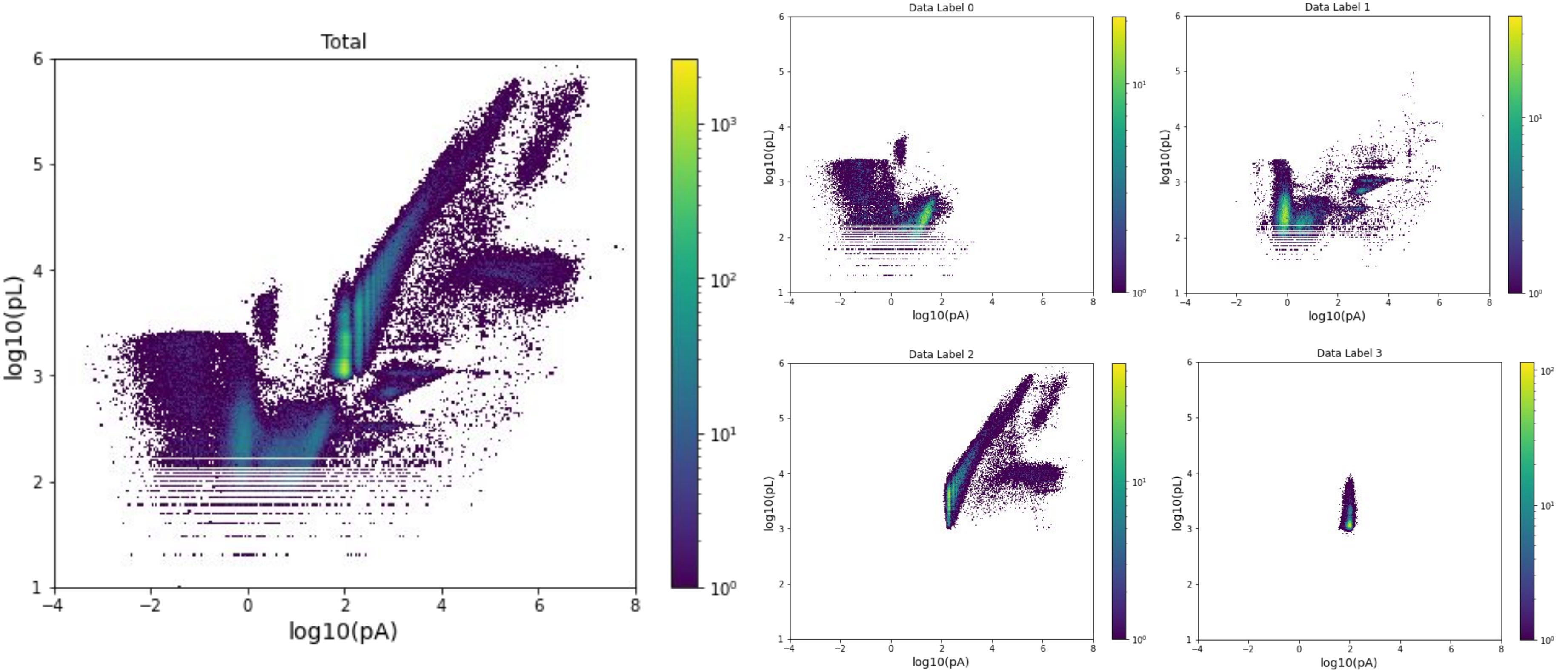
To get an idea of how the data behaved, several plot were done:

TBA vs log(pA)

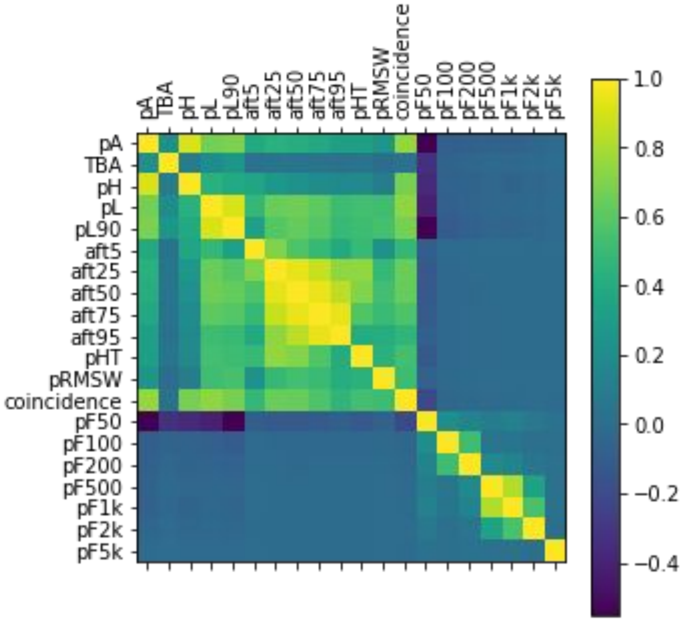




log(pA) vs log(pL)



Correlation Matrix before normalization



Normalization

Two distinct normalizations were done:

Quantile Transformer

Standard Scaler

Normalization

Two distinct normalizations were done:

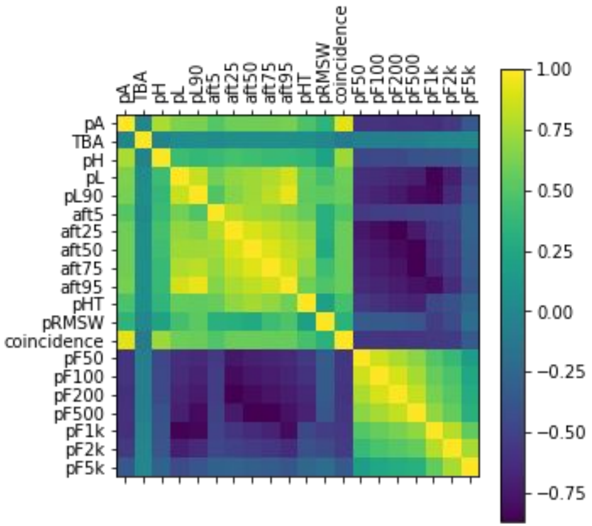
Quantile Transformer

This method transforms the features to follow a uniform or a normal distribution. Therefore, for a given feature, this transformation tends to spread out the most frequent values. It also reduces the impact of (marginal) outliers: this is therefore a robust preprocessing scheme.

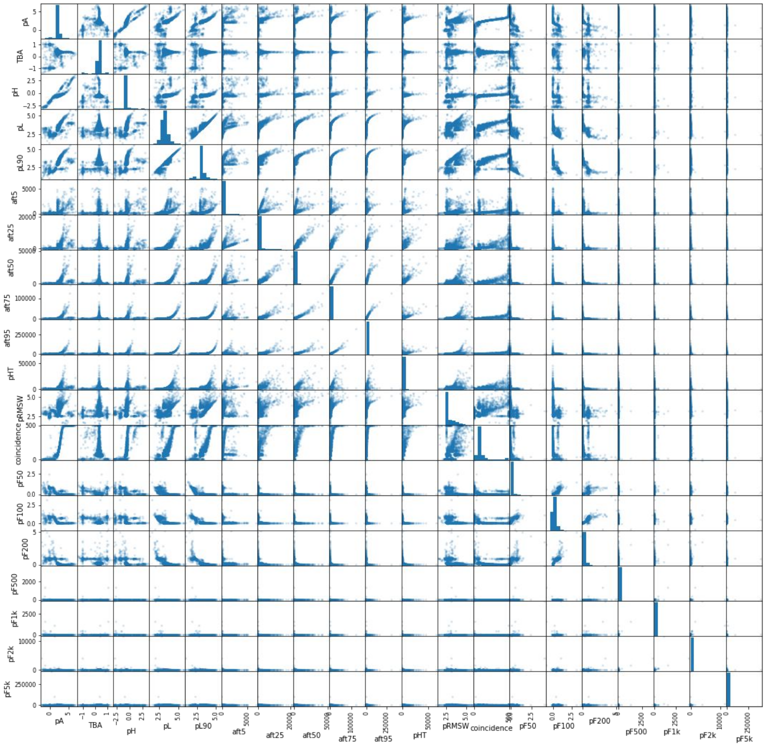
Standard Scaler

Quantile Transformer

Correlation Matrix

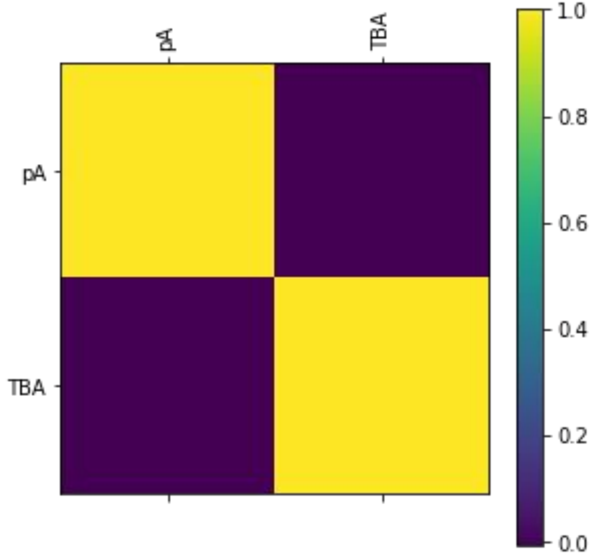


Scatter Matrix

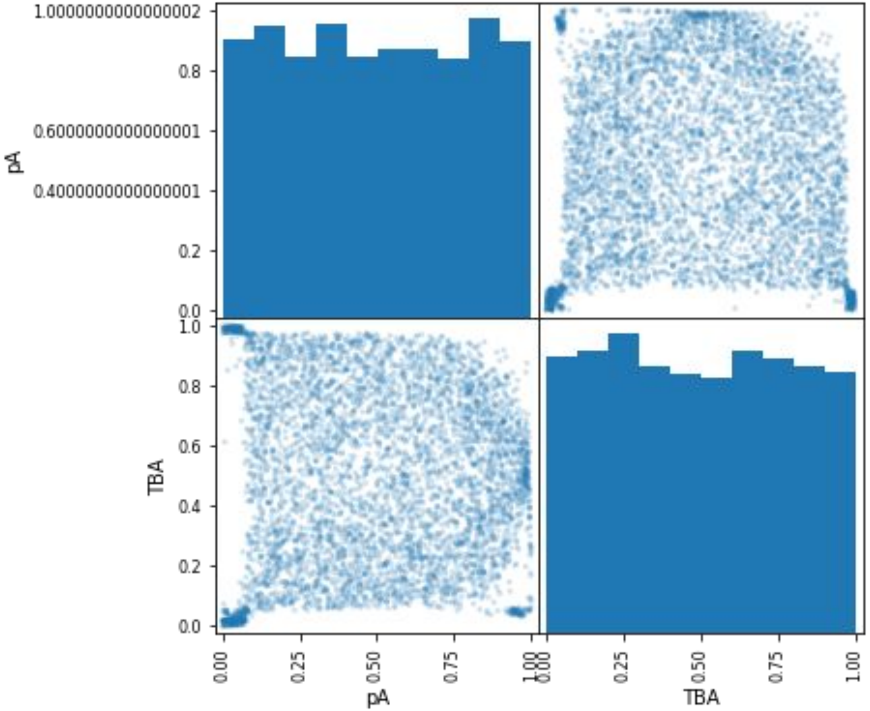


Quantile Transformer (threshold=0,5)

Correlation Matrix



Scatter Matrix



Normalization

Two distinct normalizations were done:

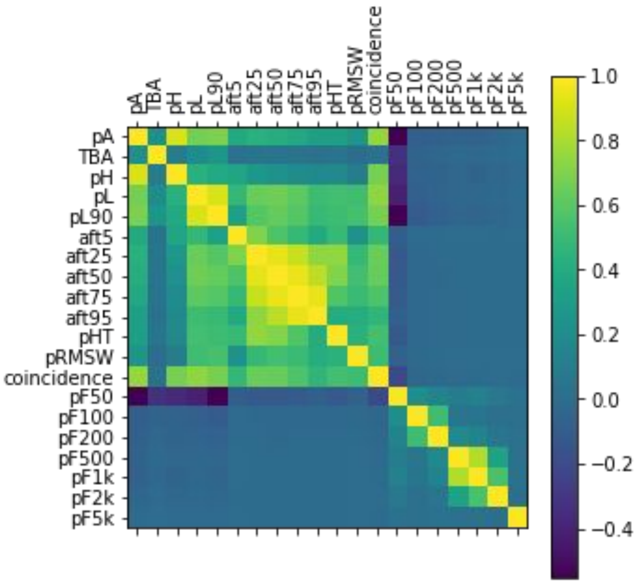
Quantile Transformer

Standard Scaler

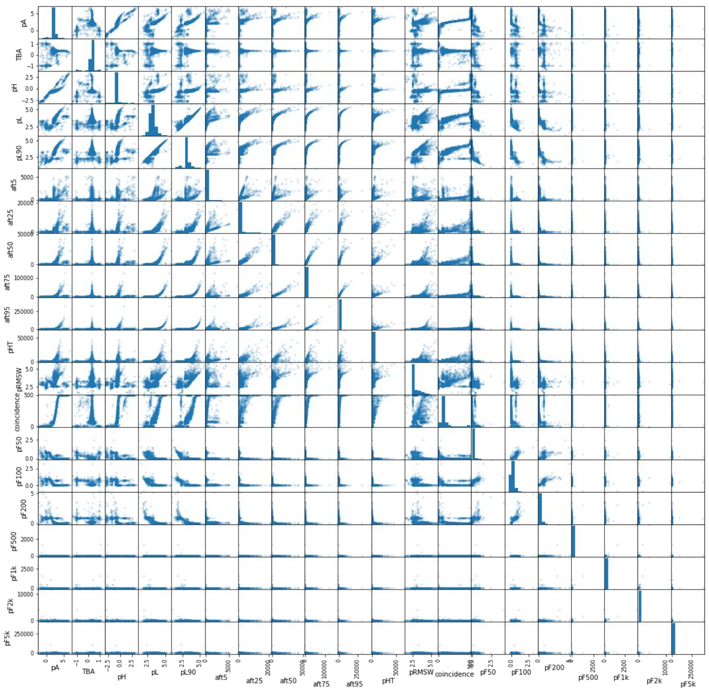
Standardize features by removing the mean and scaling to unit variance.
Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).

Standard Scaler

Correlation Matrix

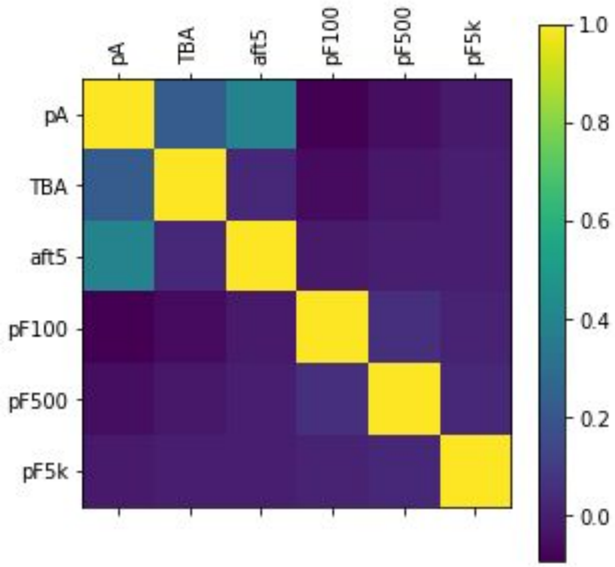


Scattter Matrix

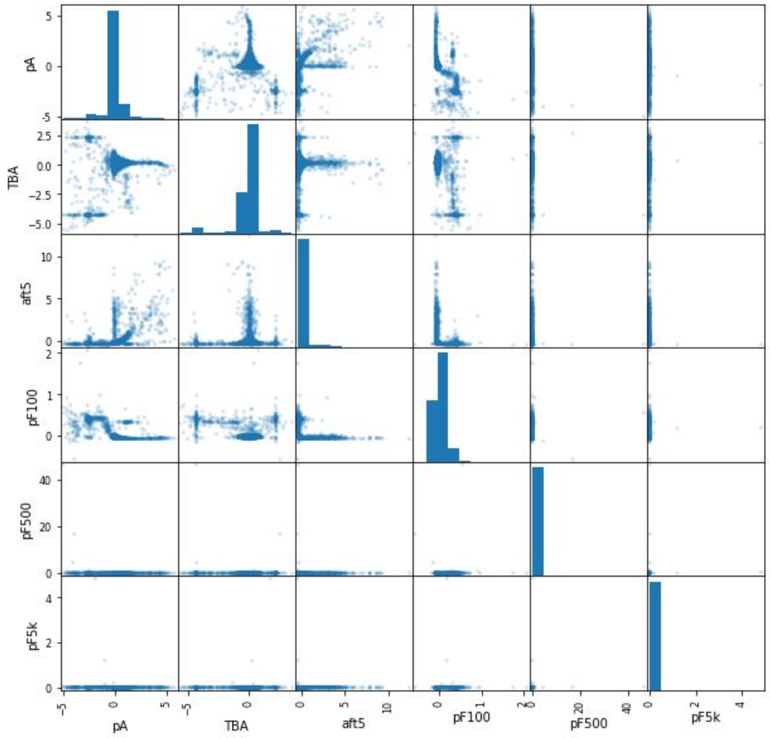


Standard Scaler (threshold=5,5)

Correlation Matrix



Scatter Matrix



Supervised learning

Random Forest Model

Results after the classification:

	precision	recall	f1-score	support
0	0.89	0.83	0.86	9191
1	0.86	0.91	0.88	10615
2	1.00	1.00	1.00	45858
3	1.00	1.00	1.00	134336
accuracy			0.99	200000
macro avg	0.94	0.93	0.93	200000
weighted avg	0.99	0.99	0.99	200000

Predicted Class	0	1	2	3
Actual Class				
0	7628	1546	0	17
1	955	9617	30	13
2	0	0	45837	21
3	4	0	39	134293

Supervised learning

Neural Network

Results after the classification:

	precision	recall	f1-score	support
0	0.70	0.49	0.57	9191
1	0.66	0.81	0.73	10615
2	0.93	0.92	0.92	45858
3	0.97	0.97	0.97	134336
accuracy			0.93	200000
macro avg	0.81	0.80	0.80	200000
weighted avg	0.93	0.93	0.93	200000

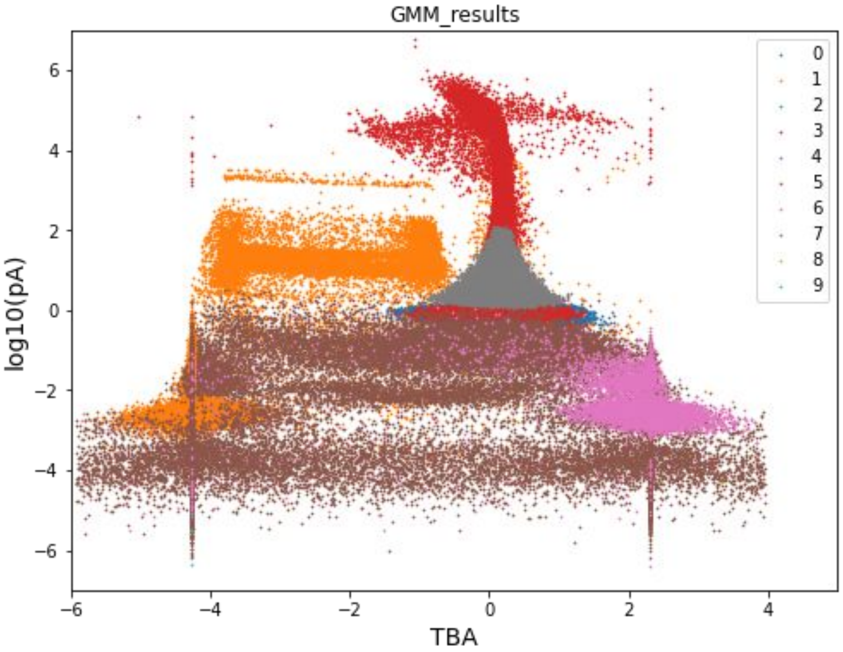
Predicted Class	0	1	2	3
Actual Class				
0	8033	1102	0	56
1	3815	6712	77	11
2	0	0	45814	44
3	5	0	51	134280

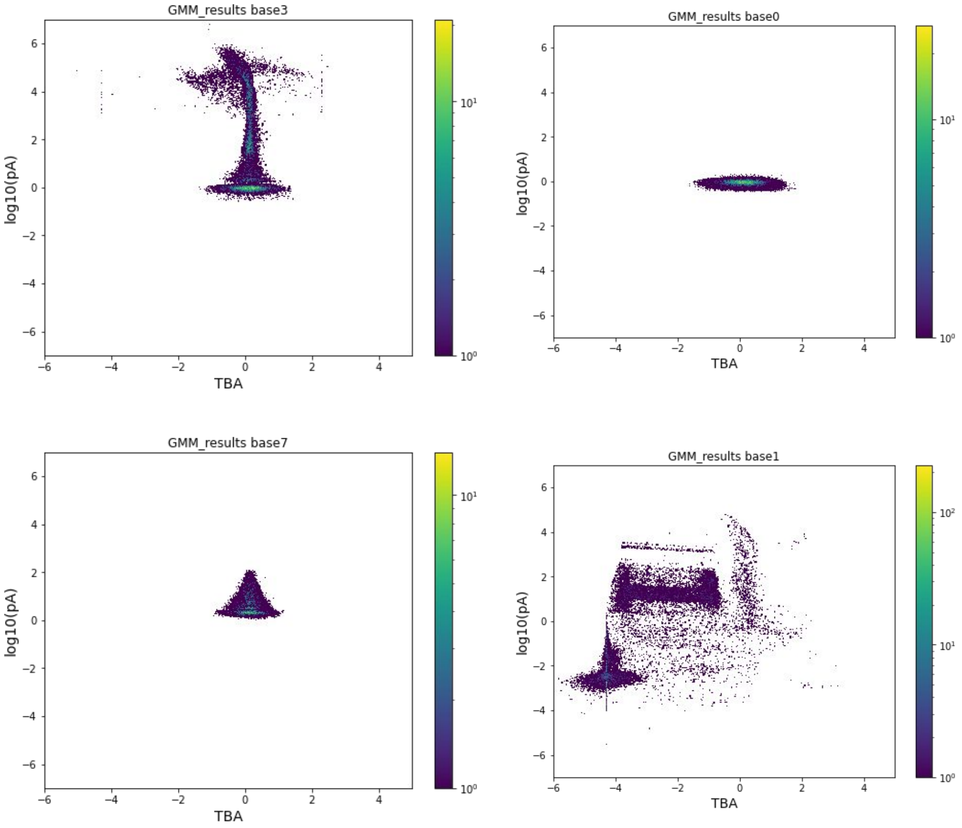
Unsupervised learning

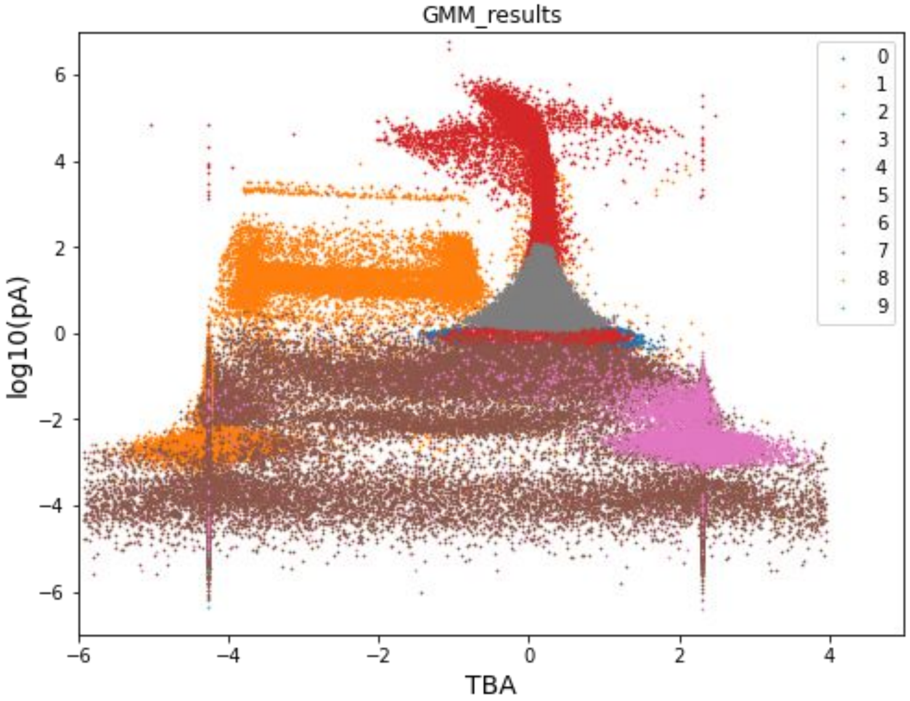
Gaussian Mixture Model

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

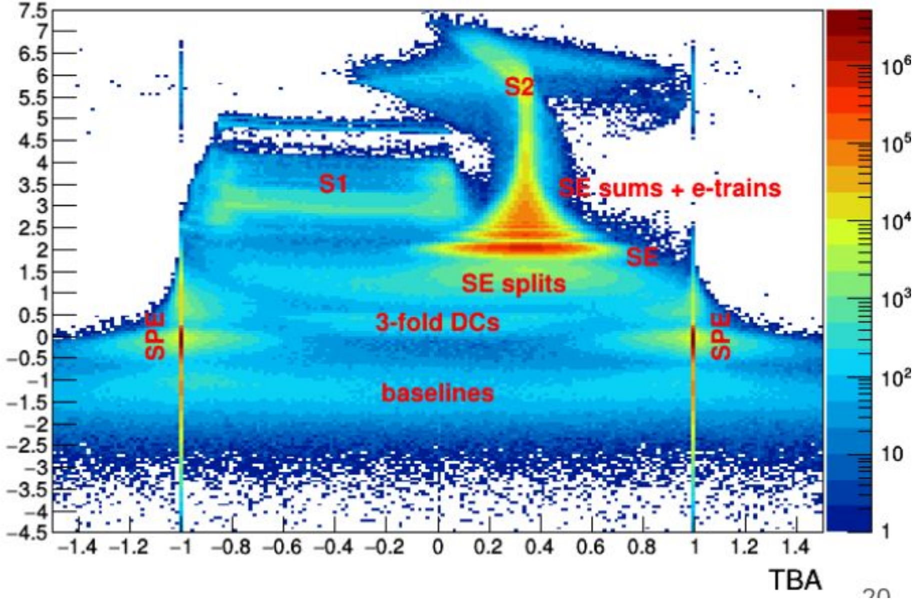
Clustering of the data



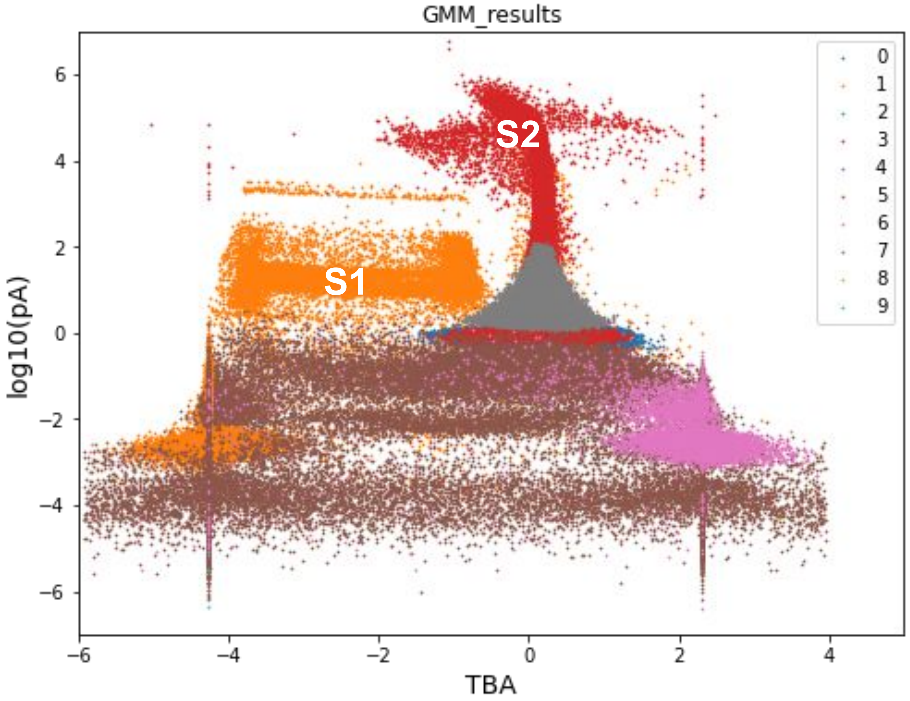


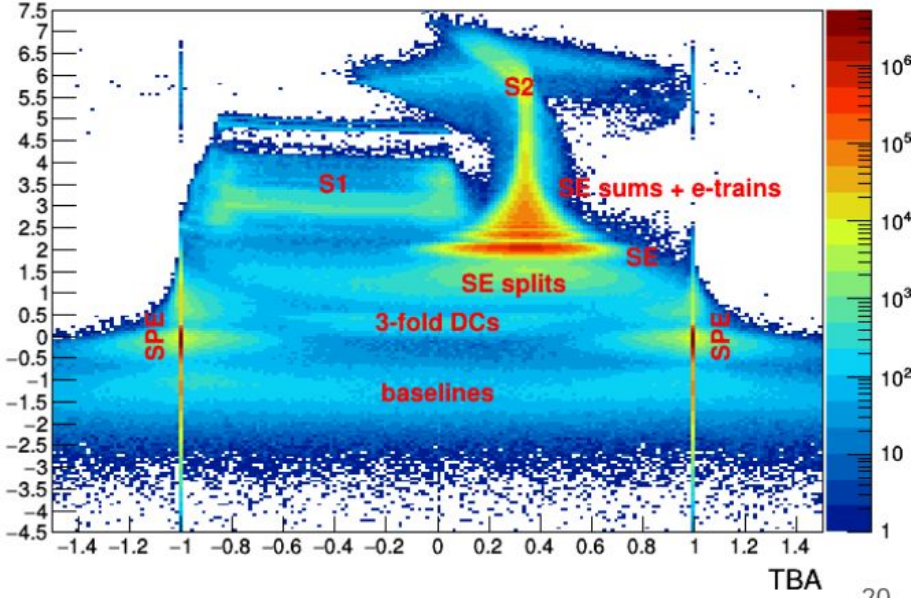


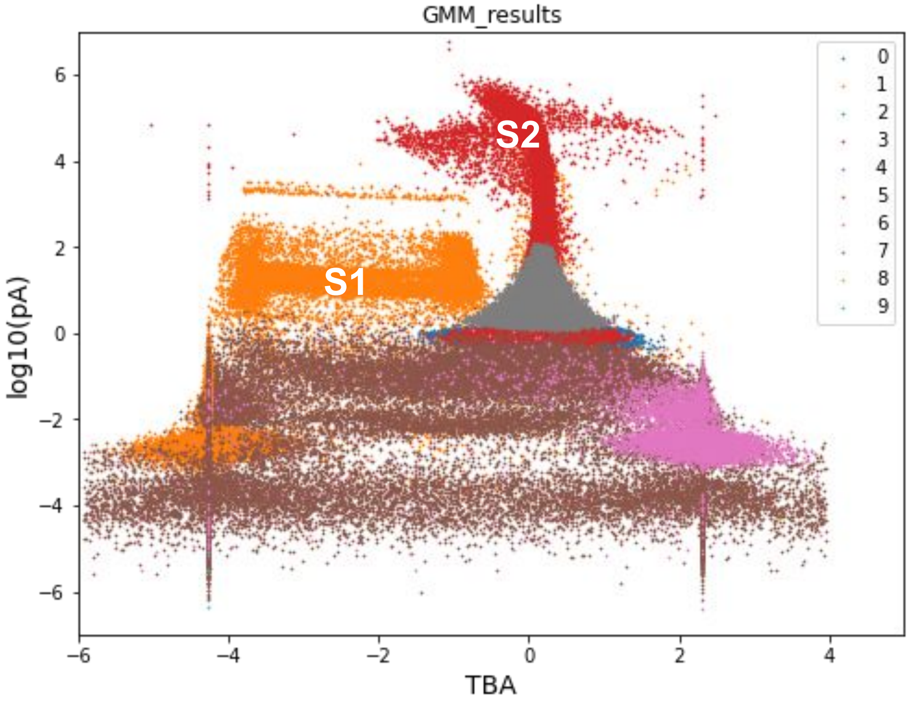
GMM_results
log10(pA)
TBA
0
1
2
3
4
5
6
7
8
9



S2
S1
SE sums + e-trains
SE
SE splits
3-fold DCs
SPE
SPE
baselines
TBA


GMM_results
S2
S1
log10(pA)
TBA
0
1
2
3
4
5
6
7
8
9


S2
S1
SE sums + e-trains
SE
SE splits
3-fold DCs
SPE
SPE
baselines
TBA



	0	1	2	3
0	0.0	0.0	0.0	643337.0
1	5.0	36536.0	0.0	0.0
2	1.0	0.0	0.0	0.0
3	0.0	67.0	34149.0	27030.0
4	1.0	0.0	0.0	0.0
5	45576.0	0.0	0.0	0.0
6	547.0	16495.0	1.0	0.0
7	0.0	0.0	196251.0	0.0
8	1.0	0.0	0.0	0.0
9	3.0	0.0	0.0	0.0

Thank you!