LAGO - INDICA

Nov 18 – 20, 2024

Machine Learning Pipeline for Particle Classification for the LAGO Water Cherenkov Detectors

T. Torres Peralta (1,2,3), M. G. Molina (1,2,3,4), H. Asorey (5), I. Sidelnik (3,6), A. J. Rubio-Montero (7), S. Dasso (3.8,9,10), R. Mayo-García (7), A. Taboada (12), L. Otiniano (11), for the LAGO Collaboration

1 Tucumán Space Weather Center (TSWC), Argentina,
2 Facultad de Ciencias Exactas y Tecnología (FACET-UNT), Argentina
3 Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina
4 Instituto Nazionale di Geofisica e Vulcanologia (INGV), Italy
5 Medical Physics Department, Centro Atómico Bariloche, Comisión Nacional de Energía Atómica (CNEA), Argentina
6 Departamento de física de neutrones, Centro Atómico Bariloche, Comisión Nacional de Energía Atómica (CNEA), Argentina
7 Centro de Investigaciones Energéticas Medioambientales y Tecnológicas (CIEMAT), Spain
8 Laboratorio Argentino de Meteorología del Espacio (LAMP), Argentina
9 Instituto de Astronomia y Física del Espacio (IAFE), Argentina
10Instituto de Astronomía y Física del Espacio (IAFE), Argentina
11 Comisión Nacional de Investigación y Desarrollo Aeroespacial (CONIDA), Peru
12 Instituto de Tecnologías en Detección y Astropartículas (ITeDA), Argentina

The LAGO Collaboration, see the complete list of authors and institutions at https://lagoproject.net/collab.html

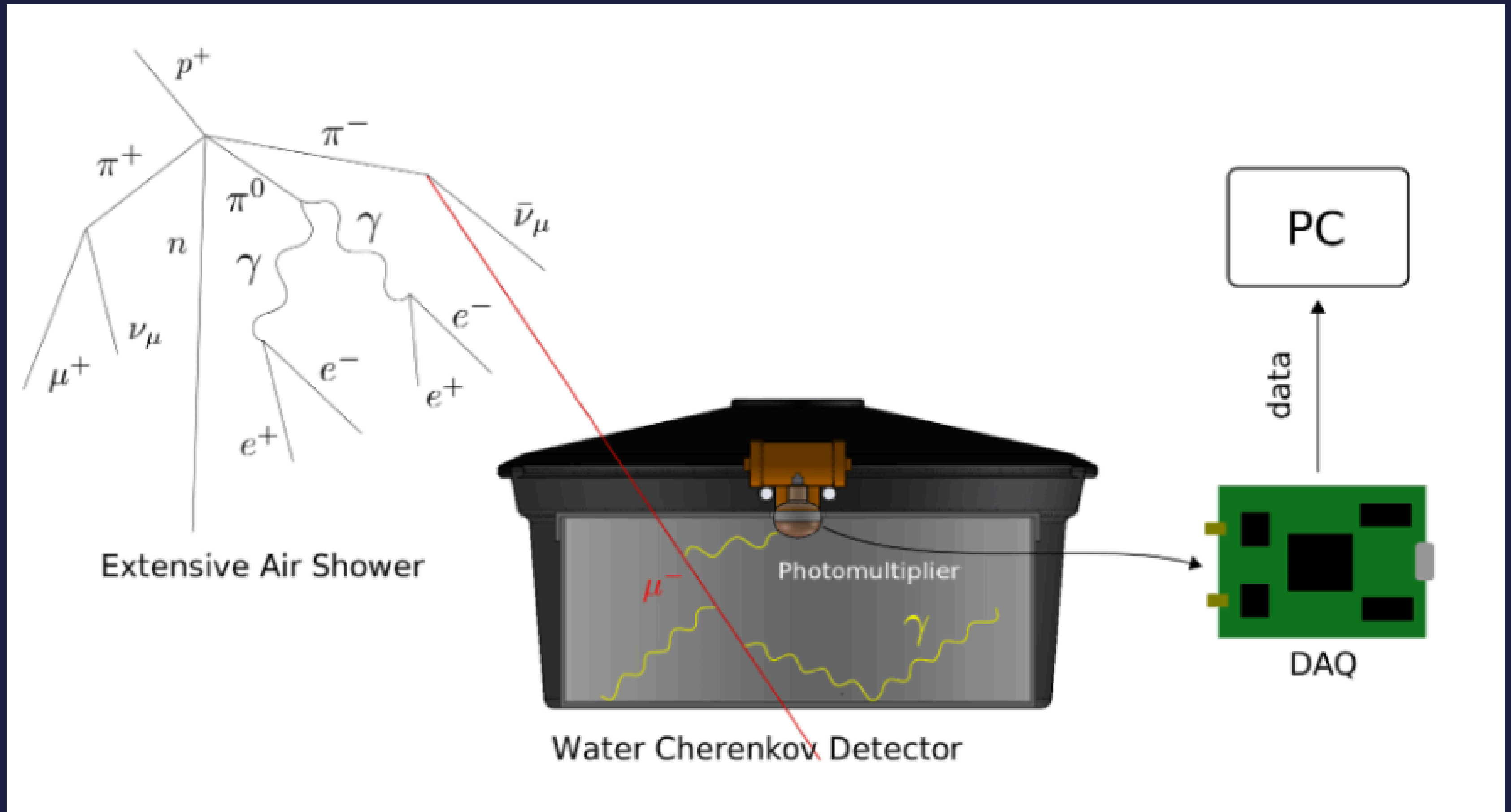
1. CONTEXT

LAGO Water Cherenkov Detectors

- Single, large-area photomultiplier tube as the primary sensor

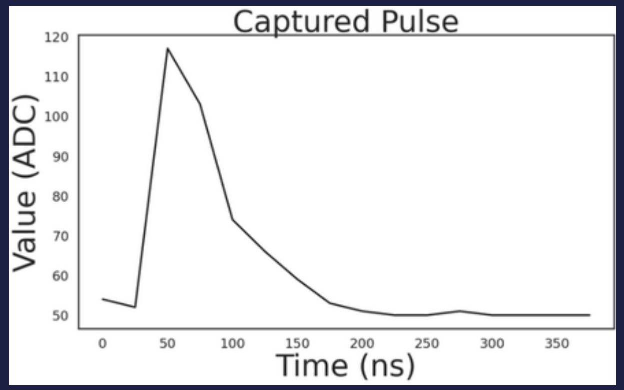
Datasets

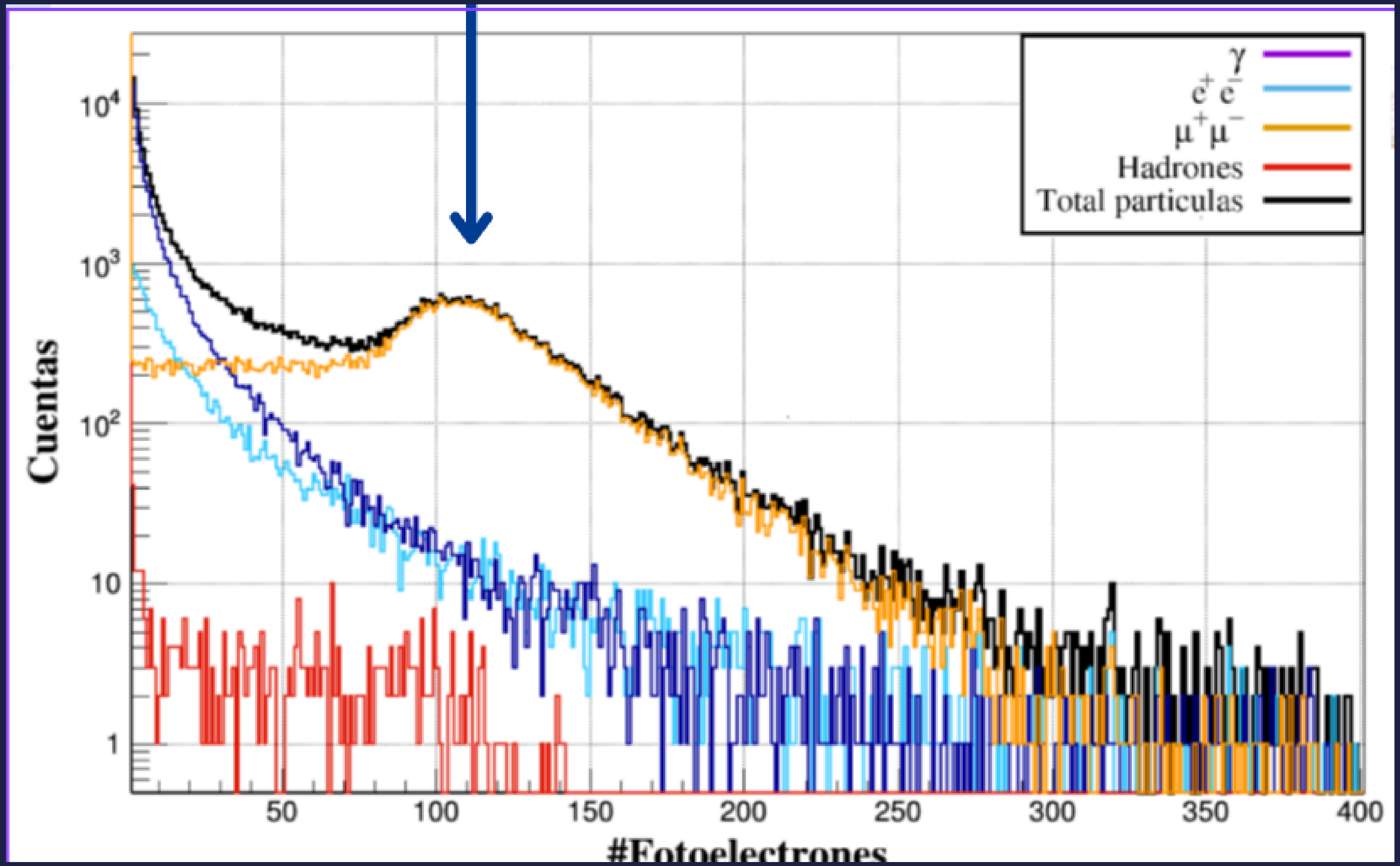
- Real: 24hs of data from March 2012, at LAGO site in Bariloche, Argentina
- Simulated: Combination of the outputs of the ARTI and Meiga simulation frameworks, simulated the expected WCD signals produced by the flux of secondary particles during 24hs at the LAGO site in Bariloche, Argentina, situated at 865 m above sea level.



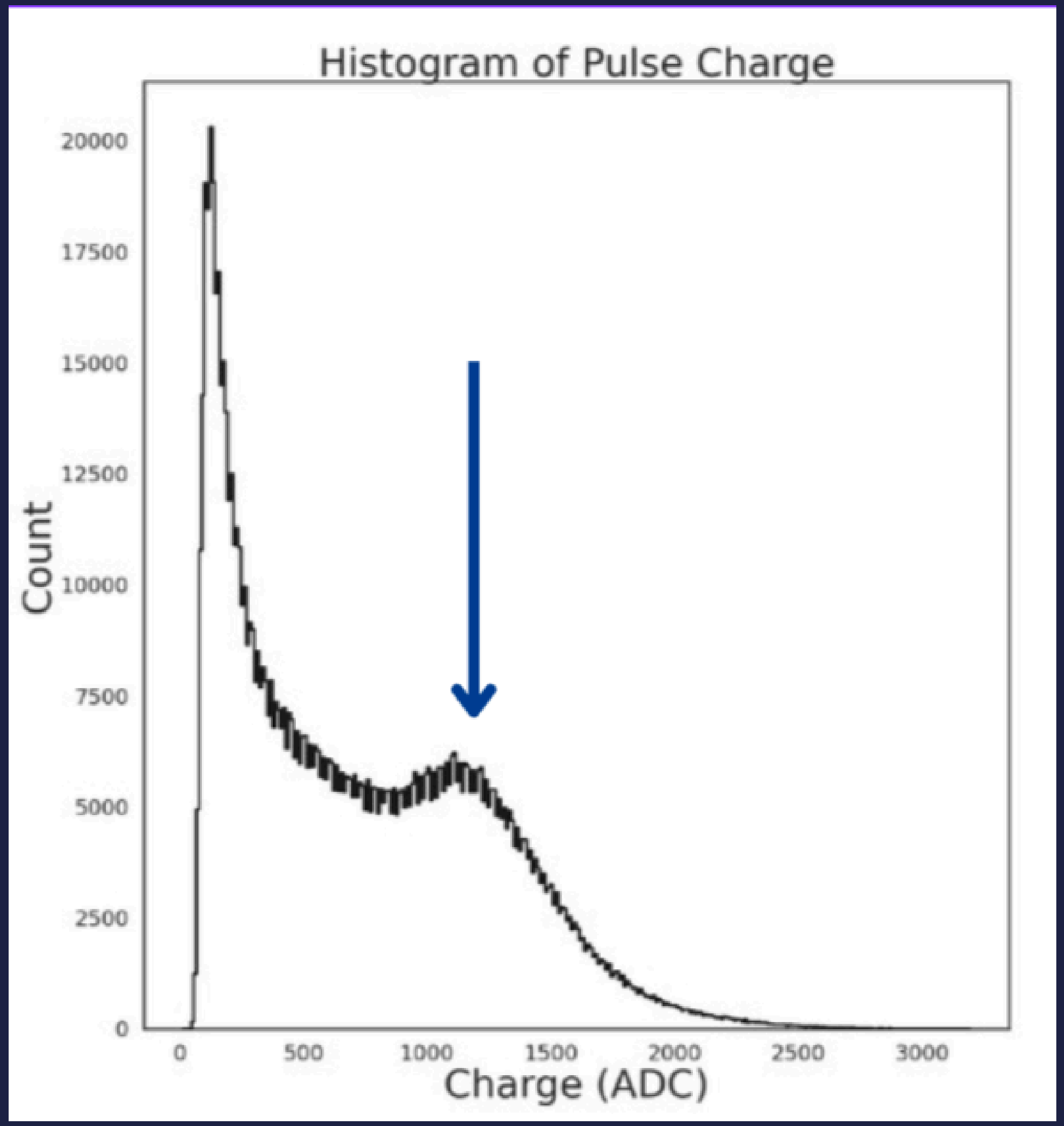
1. PROBLEM DESCRIPTION AND OBJECTIVE

- Water Cherenkov Detectors provide no direct way to discriminate between secondary particle contributions.
- We propose a machine learning pipeline using clustering for the classification of secondary particle contributions.





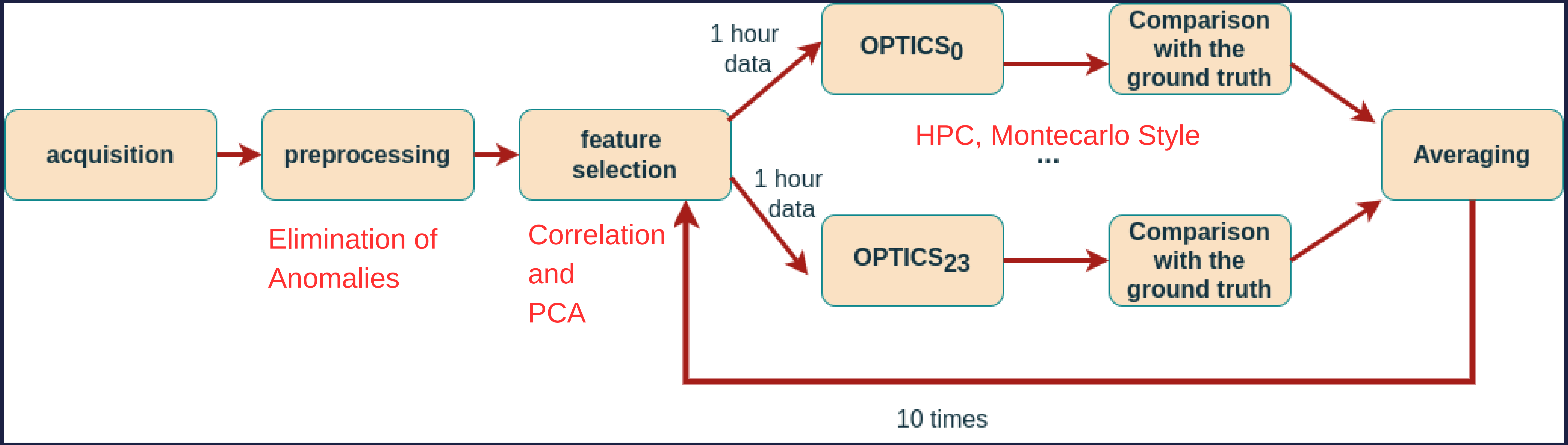
Simulated Data



Real Data

2. METHODOLOGY

We proposed a methodology based on data science where we use machine learning (ML) to implement a data-driven model and processing pipeline. The main protagonist of this pipeline is a hierarchical density-based unsupervised machine learning method for clustering pulses based on similarity patterns, called OPTICS (Ordering Points to Identify the Clustering Structure).



2. METHODOLOGY - PREPROCESSING

We applied two steps:

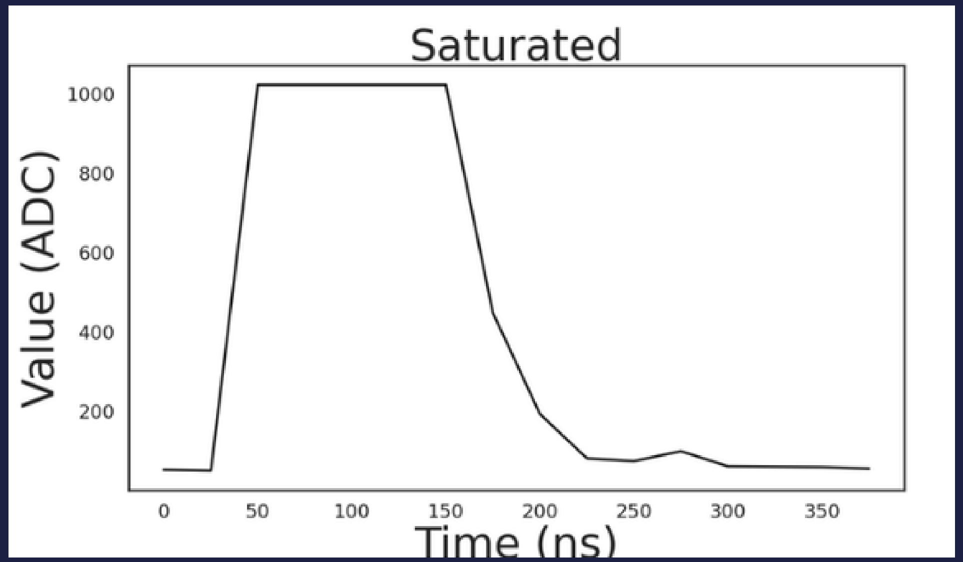
1. Filtering (actual & synthetic data) to remove anomalies and increase the data quality of the dataset.
2. Splitting of the synthetic data set into an input for OPTICS and a target/ground truth to later validate the results.

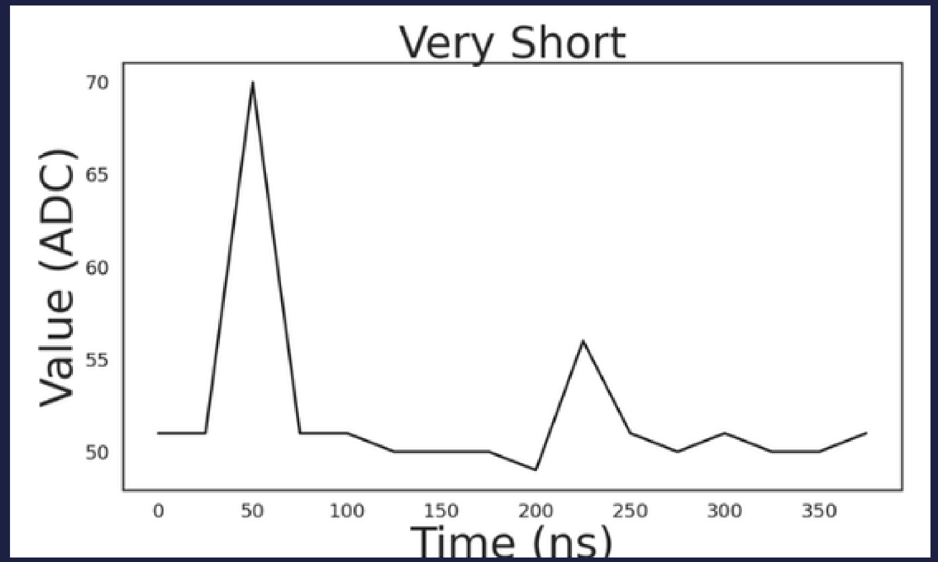
For real data:

- Saturated pulses
- Complex pulses (multiple peaks)
- Pulses with negative values
- Very short pulses

For simulated data:

- Particles that did not have enough energy to produce a photon.





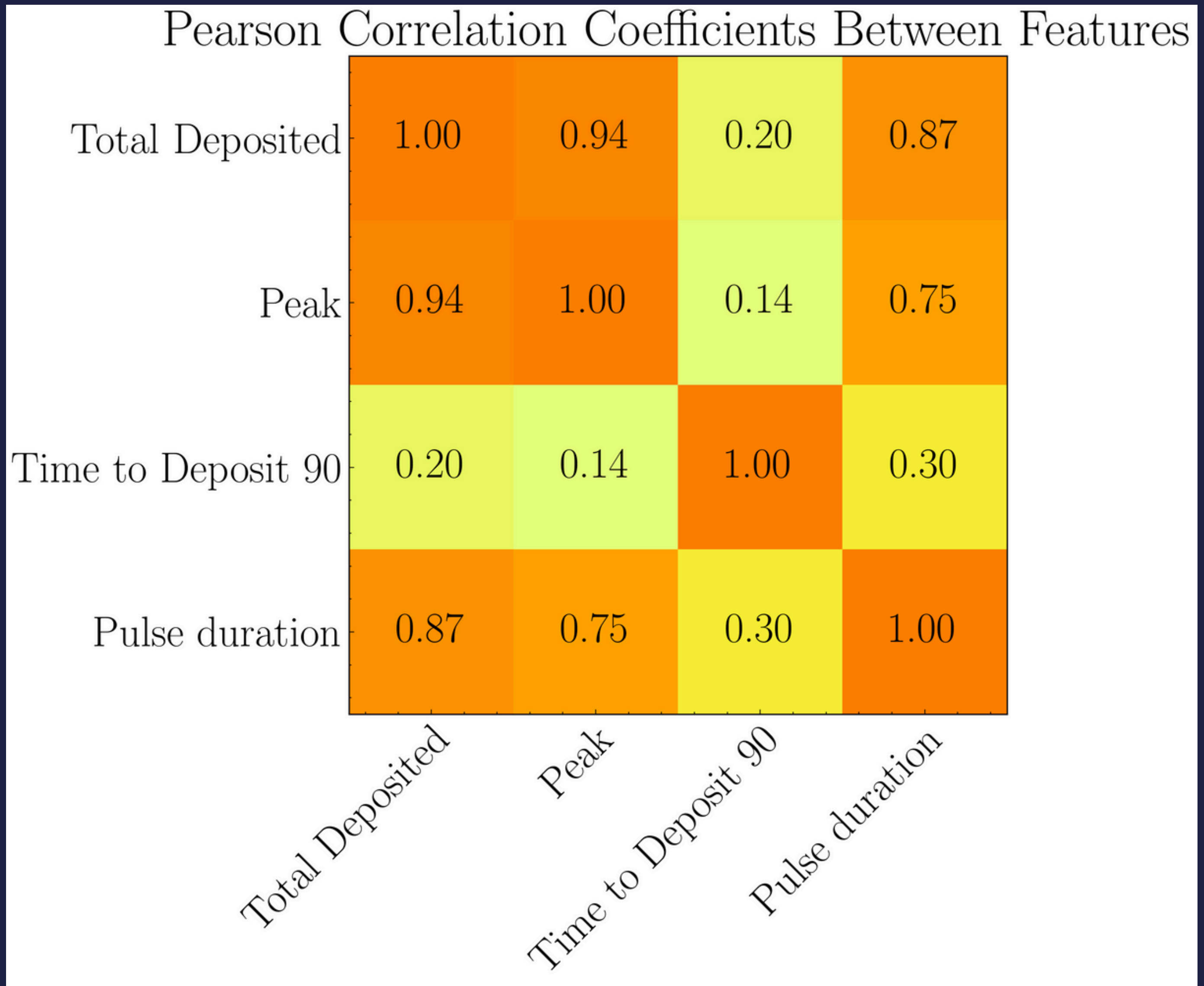
2. METHODOLOGY - FEATURE SELECTION

We used four initial features:

- Total Deposited Charge
- Peak of Pulse
- Time to Deposit 90% of Charge
- Pulse Duration

With these features, we applied a standard normalization and Principal Component Analysis (PCA) step to create a better behaved feature set for the machine learning (ML) algorithm.

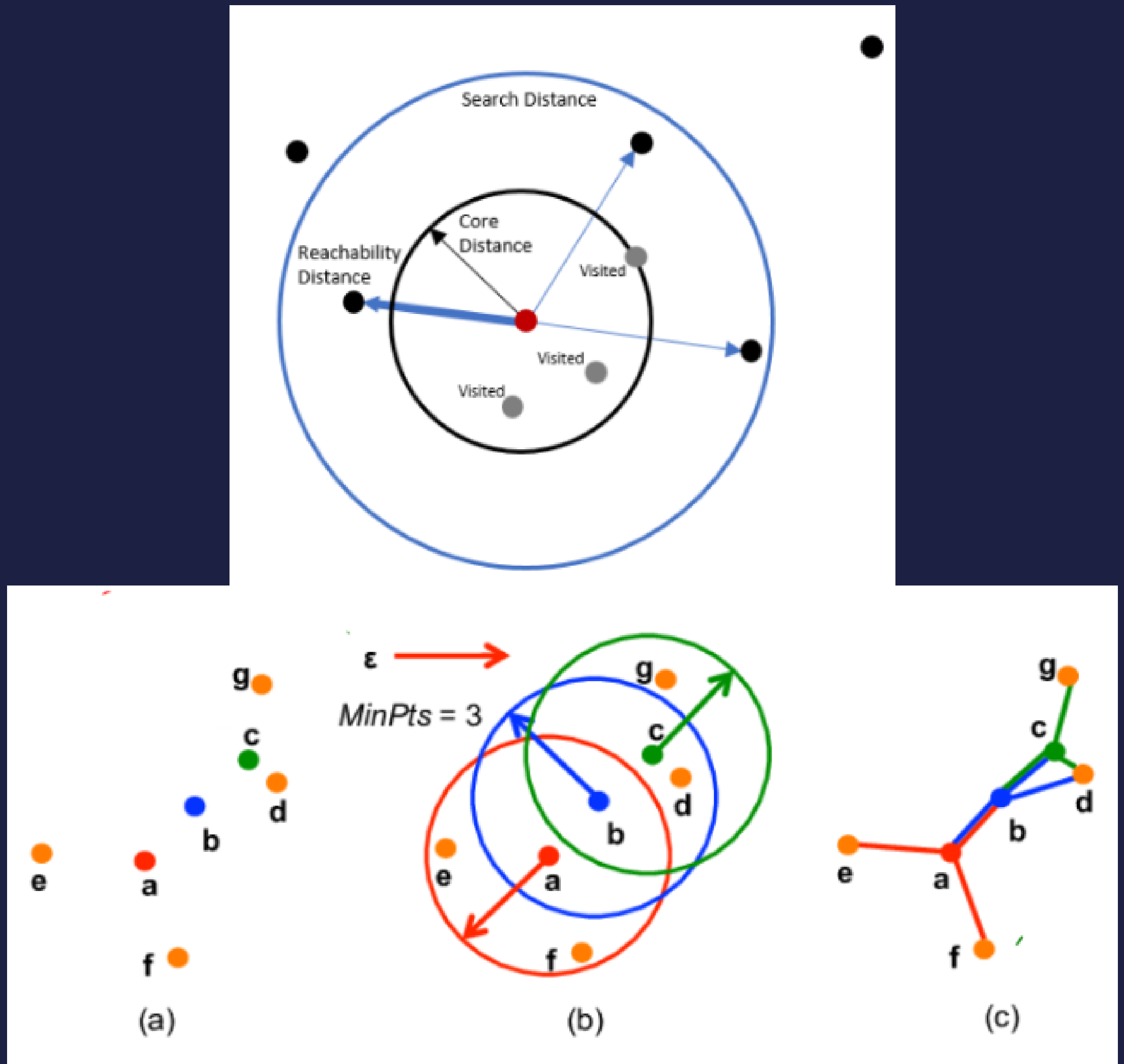
After analyzing the Pearson Correlation, it was found that Peak and Pulse duration features had high correlation with Total Deposited feature. After running the complete pipeline it was found that eliminating Peak produced better results.



Final Features	Description
Total Deposited	Total Photoelectrons (PE) that where deposited by the pulse, in total count of PE.
Time to Deposit 90%	Time the pulse took to depoite 90% of its PEs, in ns.
Pulse Duration	Duration of the pulse, in ns.

2. METHODOLOGY - OPTICS

Is a hierarchical density-based unsupervised machine learning method. It defines a reachability-distance, called epsilon, that is a minimum distance that describes cluster structure. One can then use a, or multiple, thershold(s) to define a cluster



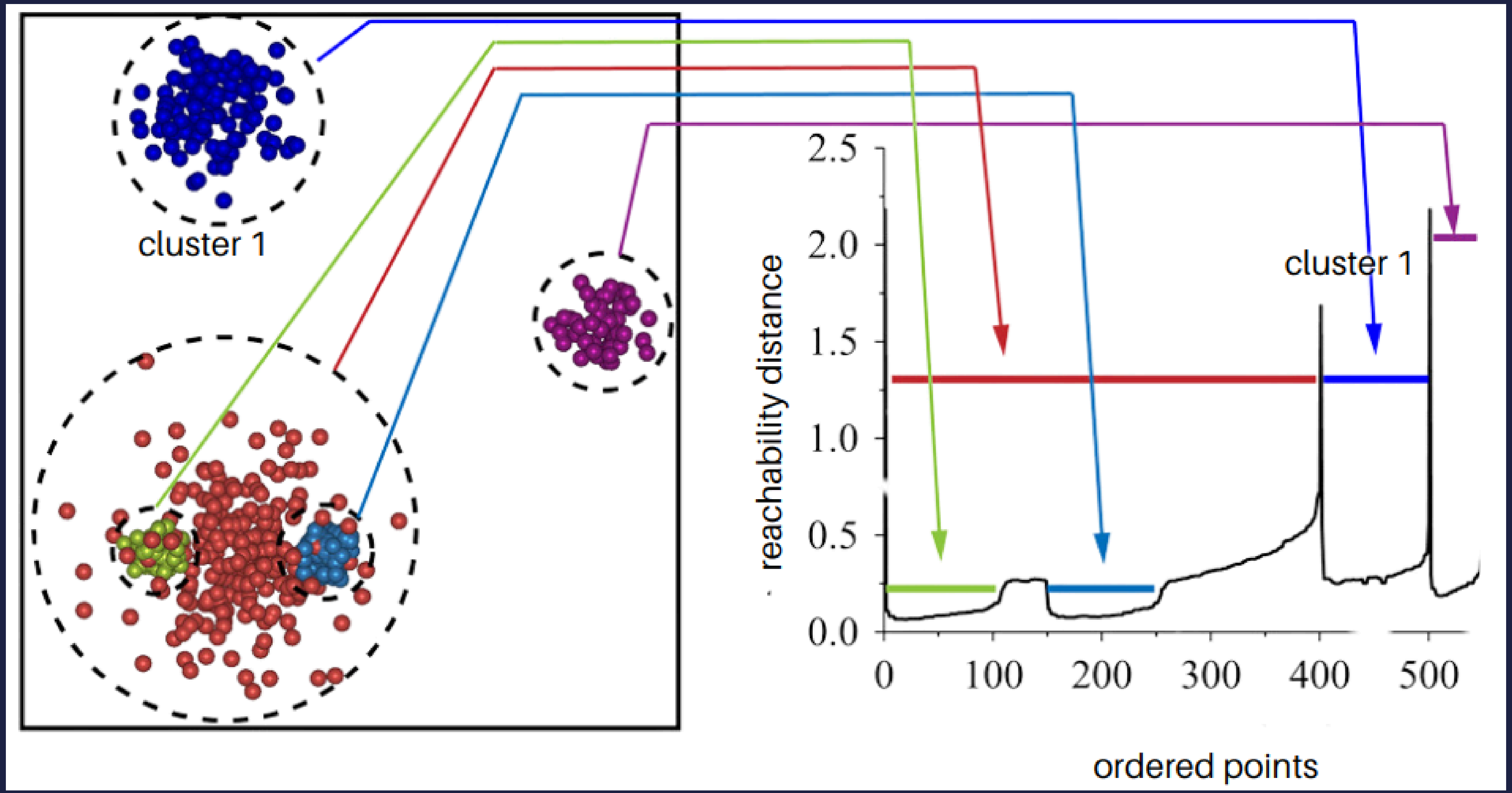
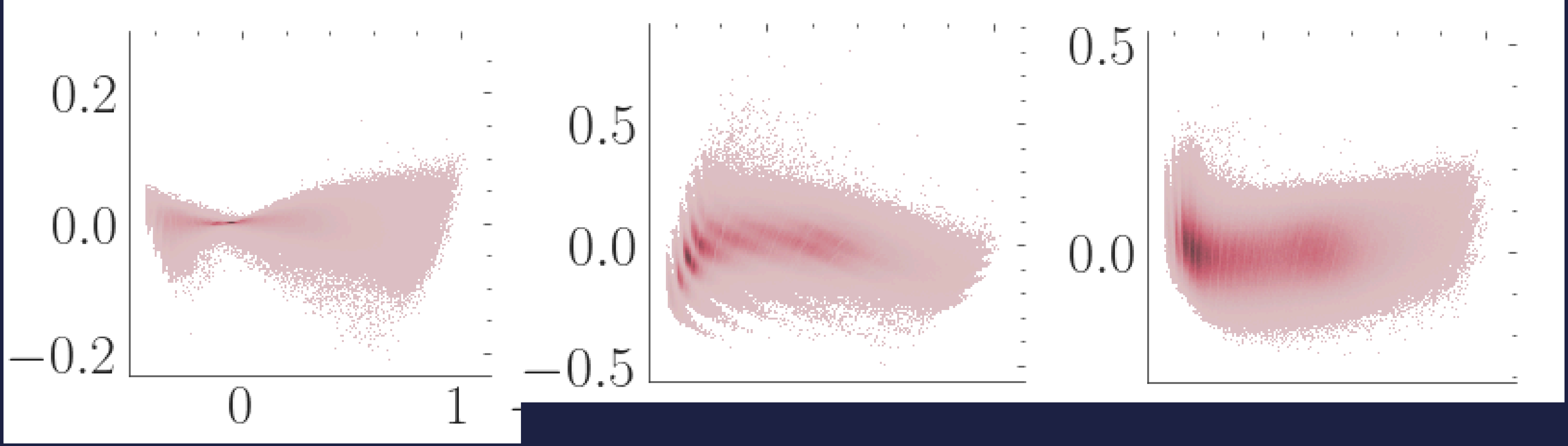


Figure adapted from Wang et al., 2019

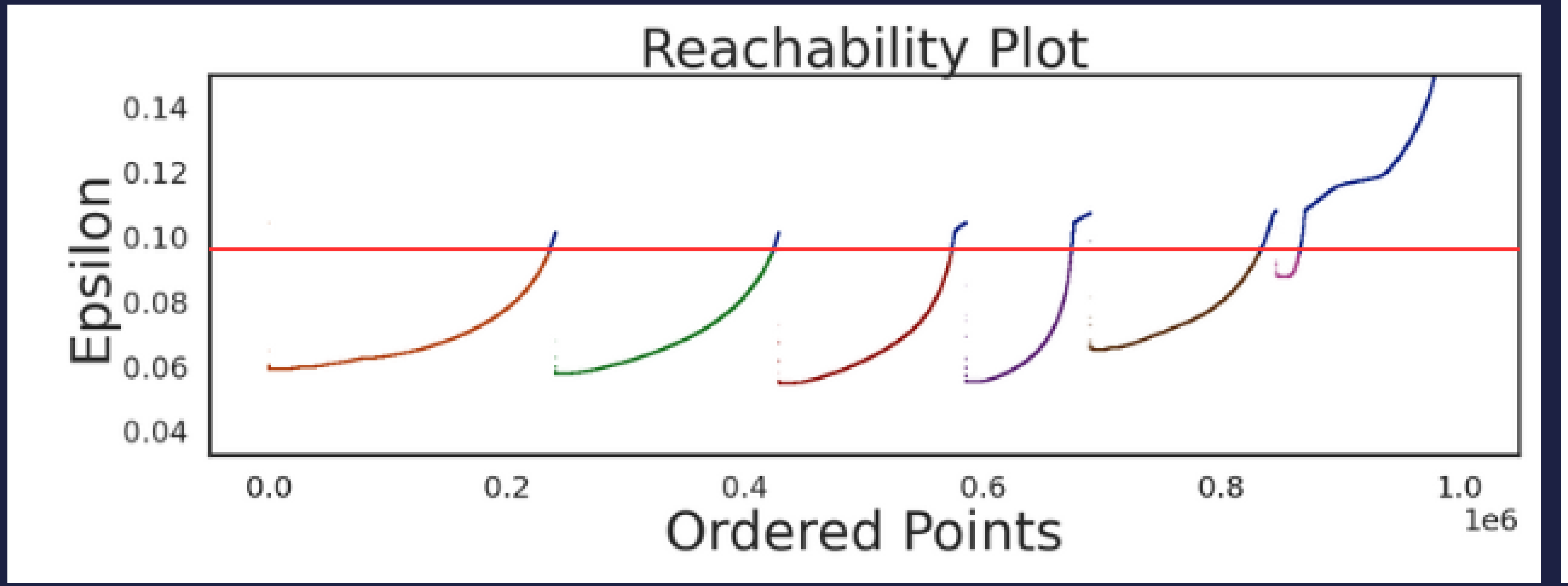
2. METHODOLOGY - OPTICS

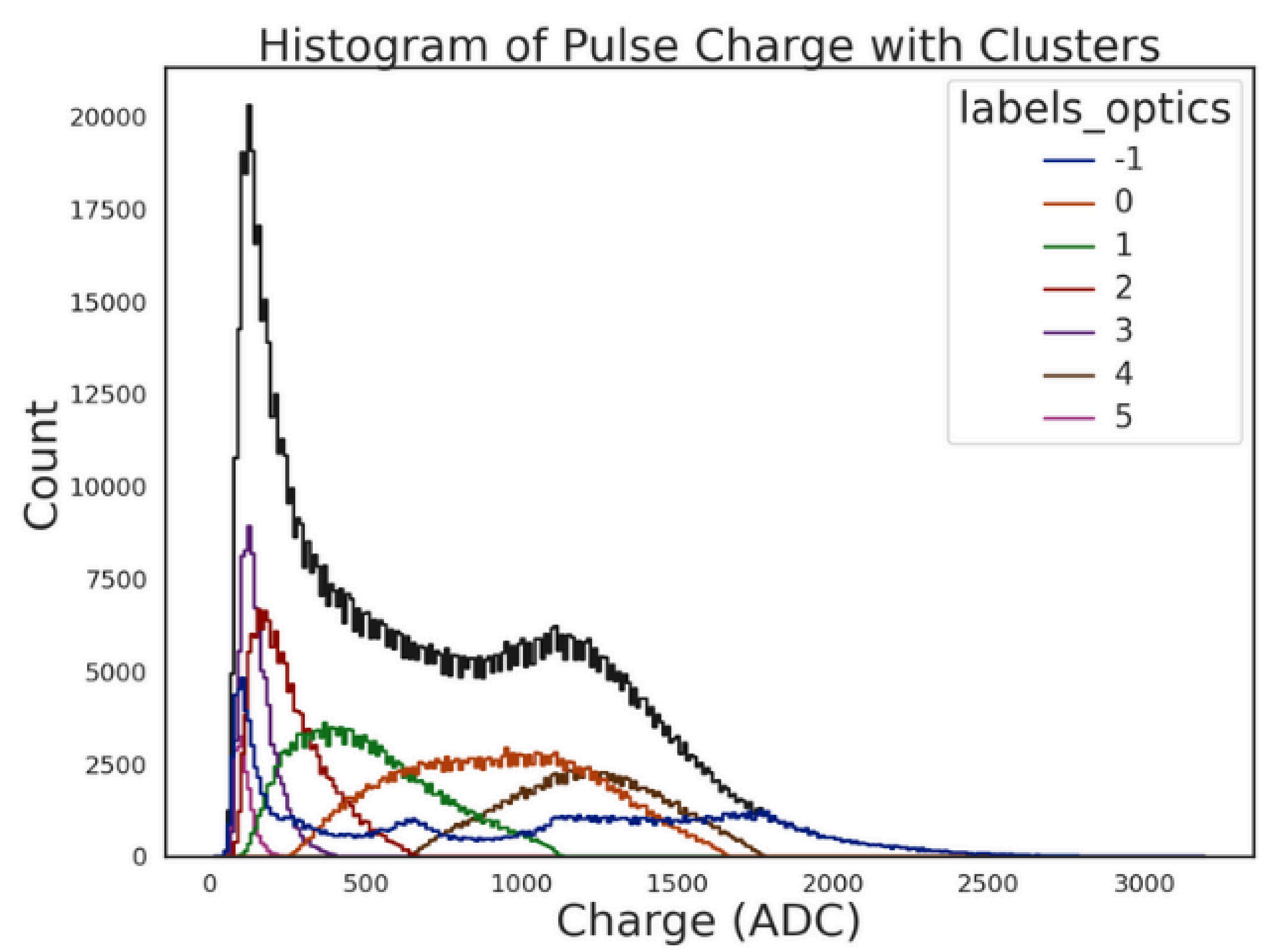
Exerpt of 2D projections of the PCA features created from the initial features used.



3. RESULTS: REAL DATA

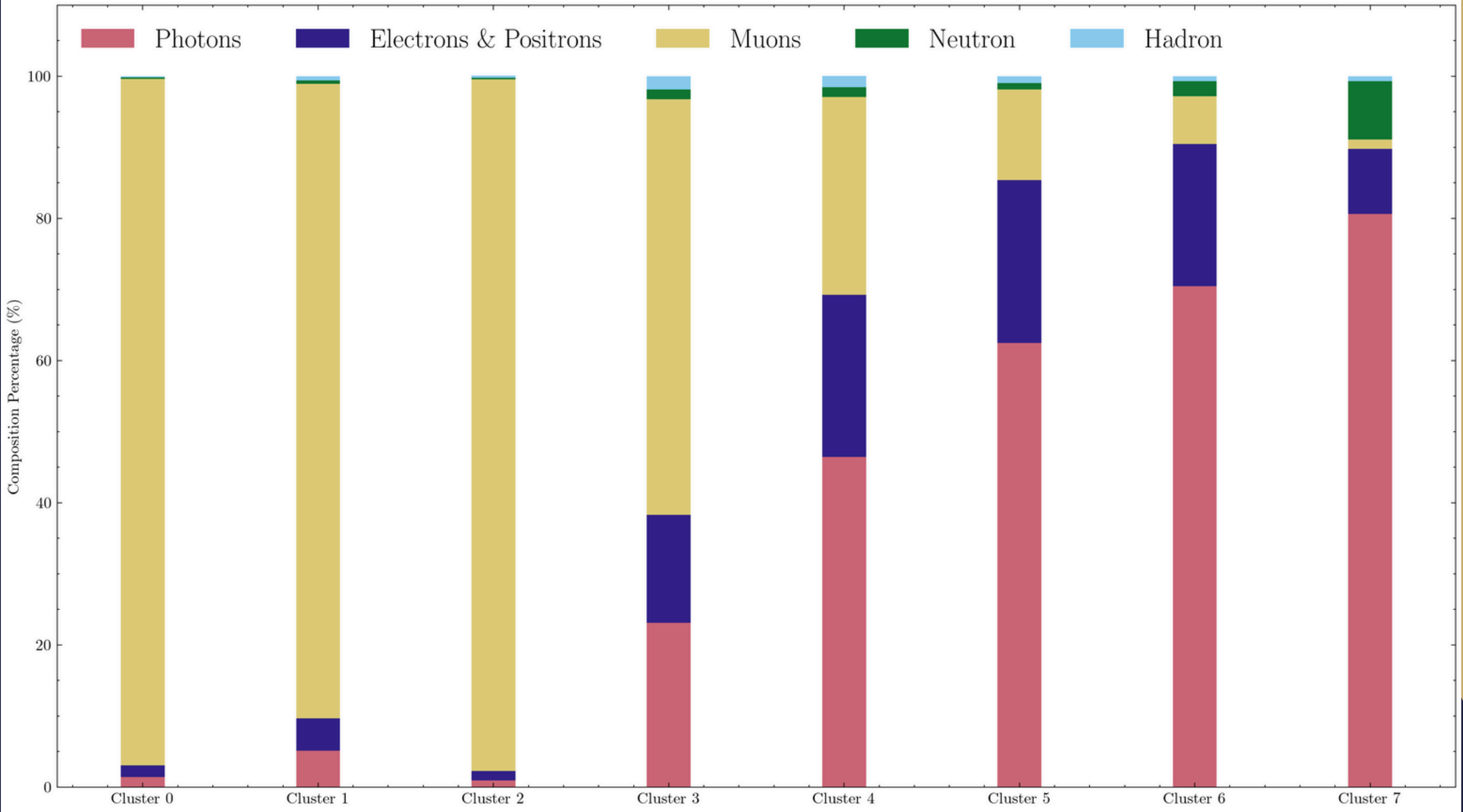
- Raw data from “Nahuelito” WCD site at Bariloche, Argentina. Total of 24hs of data between 13:00, 01 of March of 2012 and 12:00, 2 of March of 2012.
- ~39 million events preserved after preprocessing (~ 40%). Each hour contained ~1.6 million events.

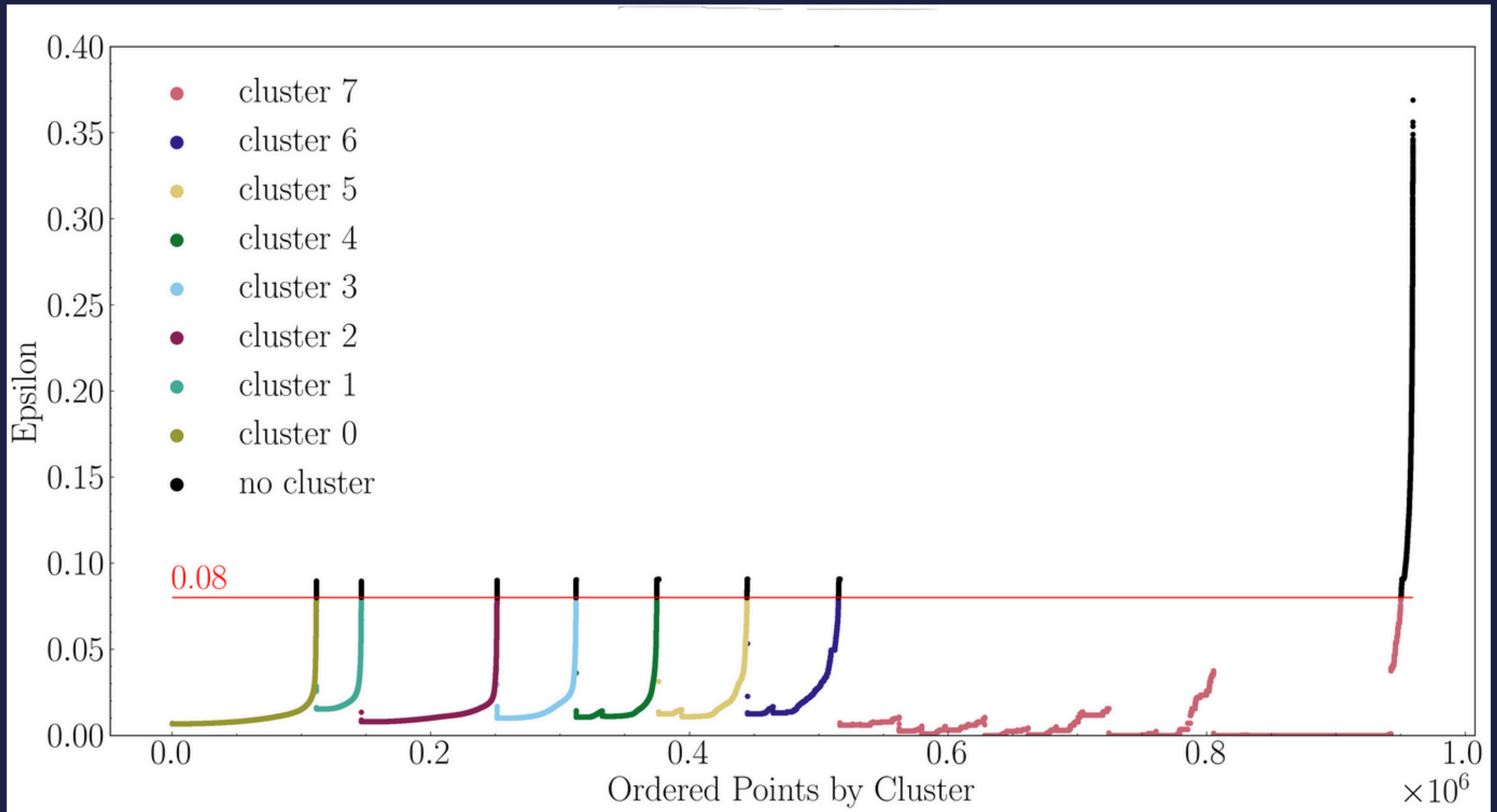


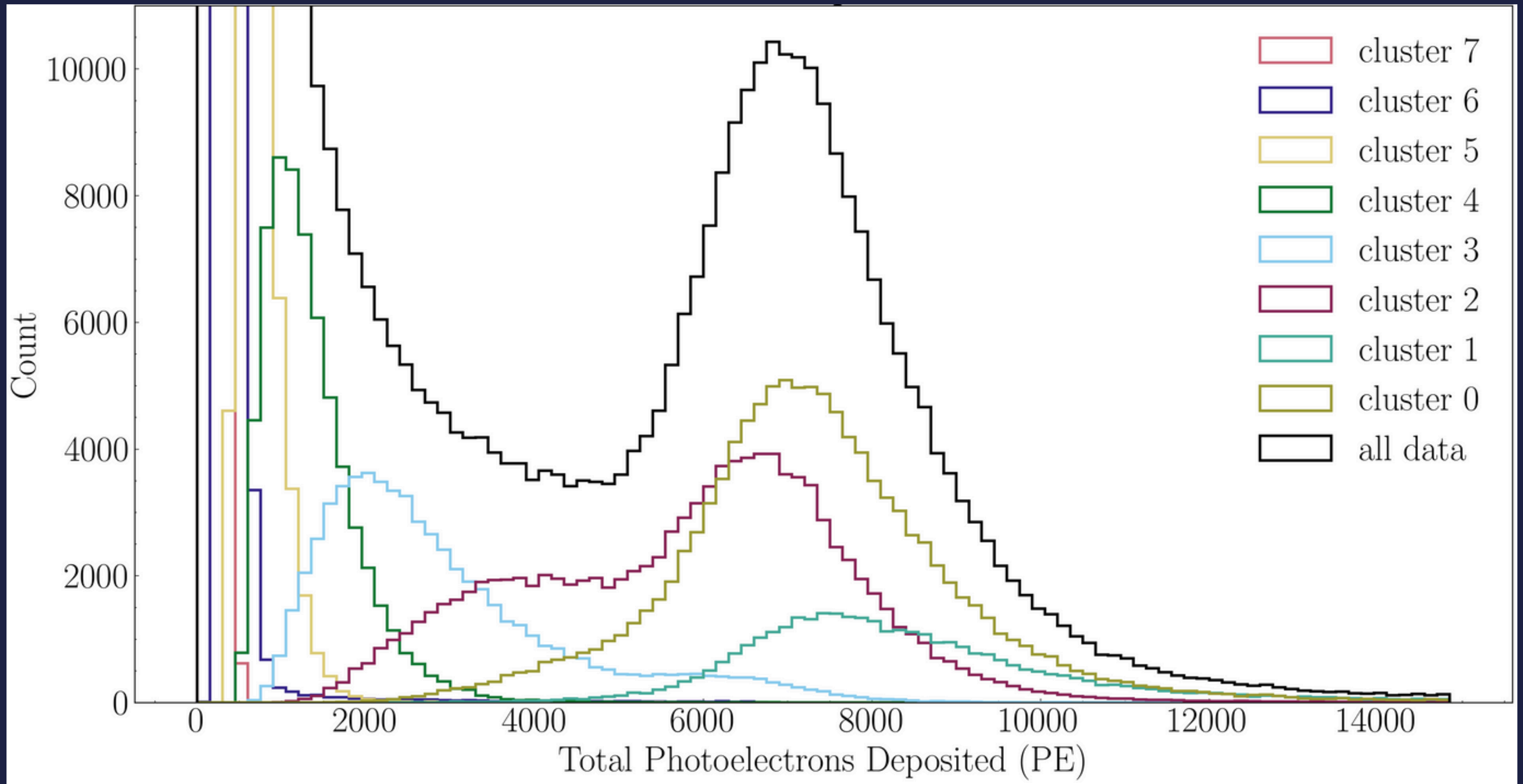


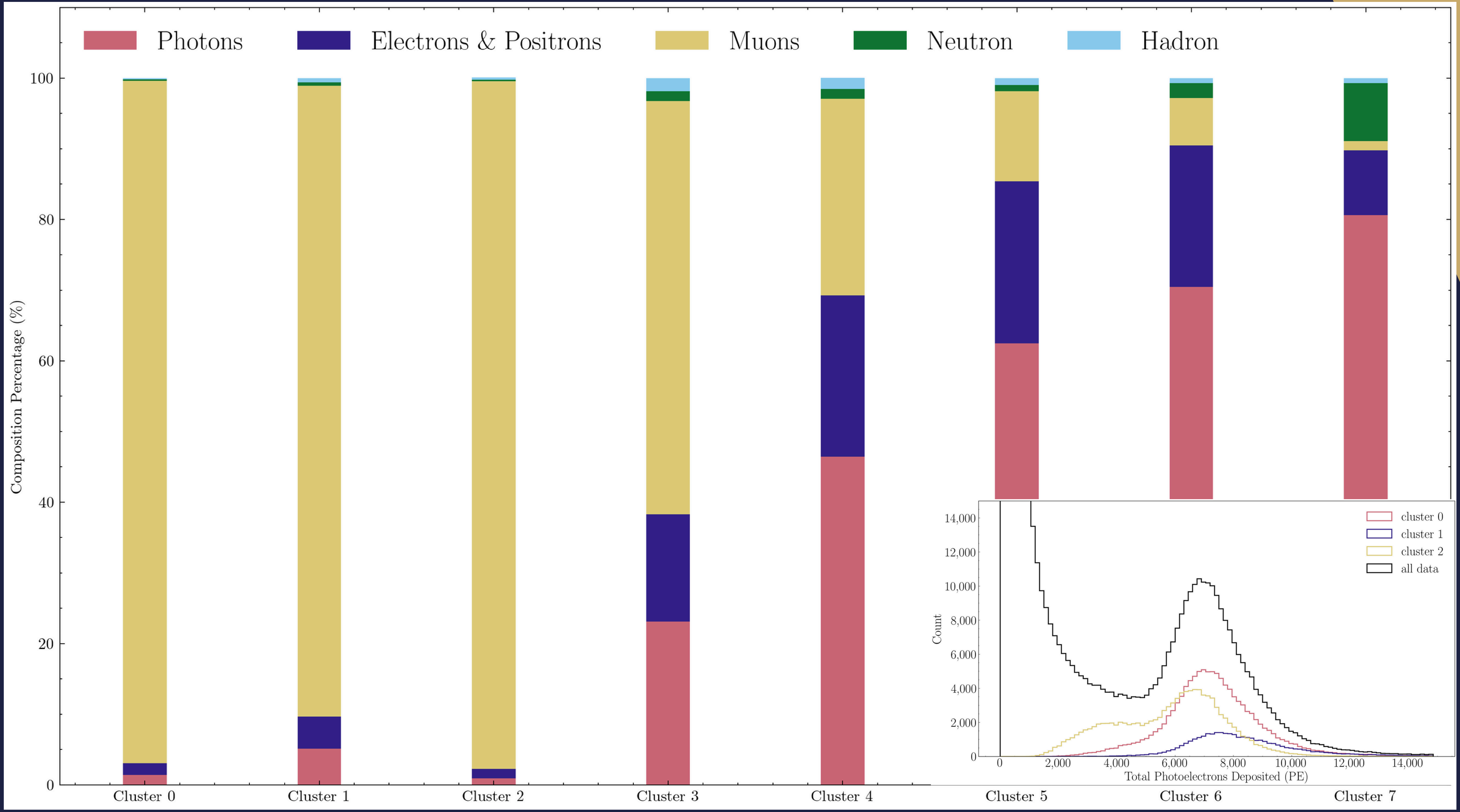
3. RESULTS: SIMULATED DATA

- 24 hours of synthetic data for spaceweather conditions on March of 2012 at “Nahuelito” WCD site at Bariloche, Argentina.
- After preprocessing there remained about ~24 million events .

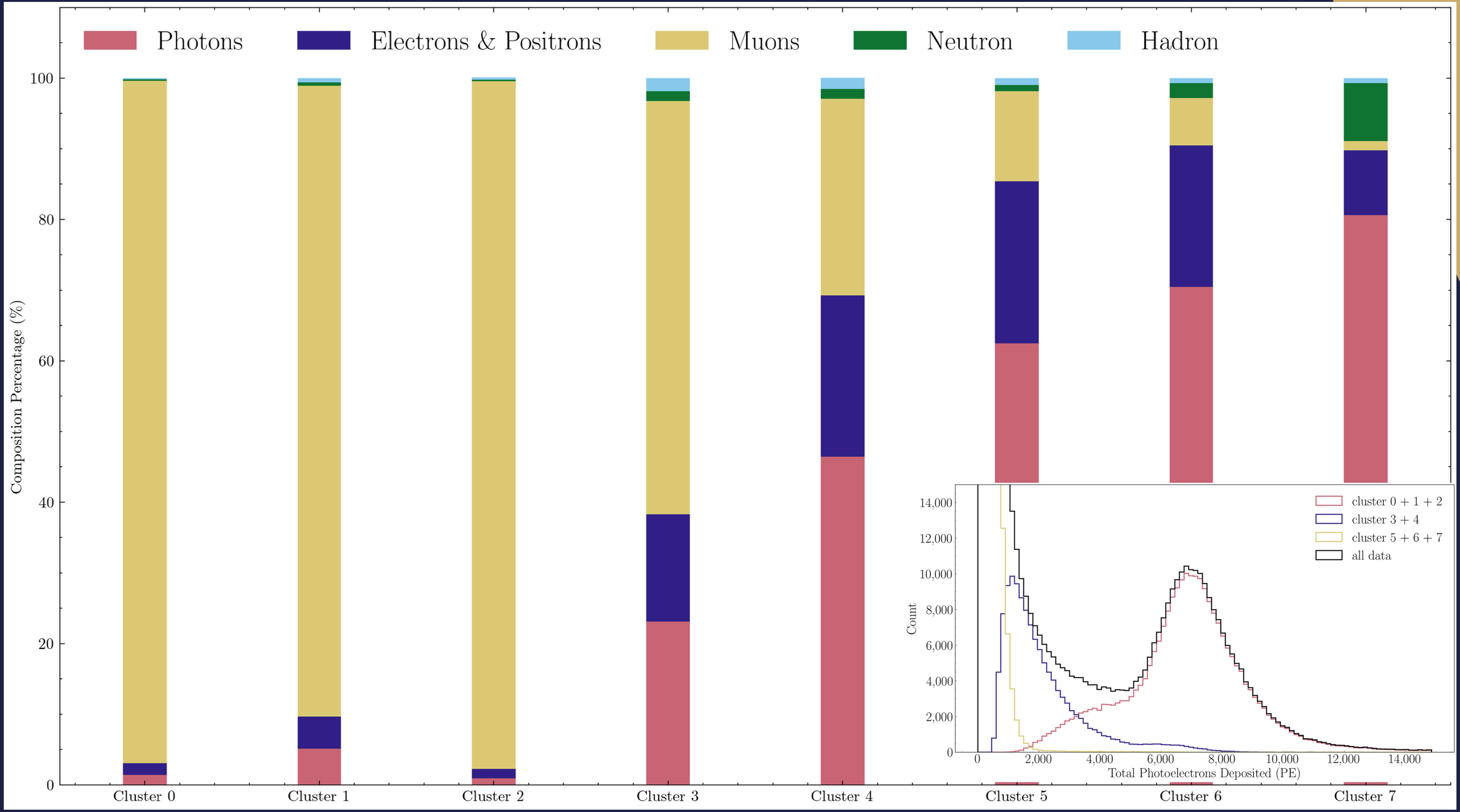








Photons
Electrons & Positrons
Muons
Neutron
Hadron
Composition Percentage (%)
0
20
40
60
80
100
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
cluster 0
cluster 1
cluster 2
all data
Count
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
Total Photoelectrons Deposited (PE)

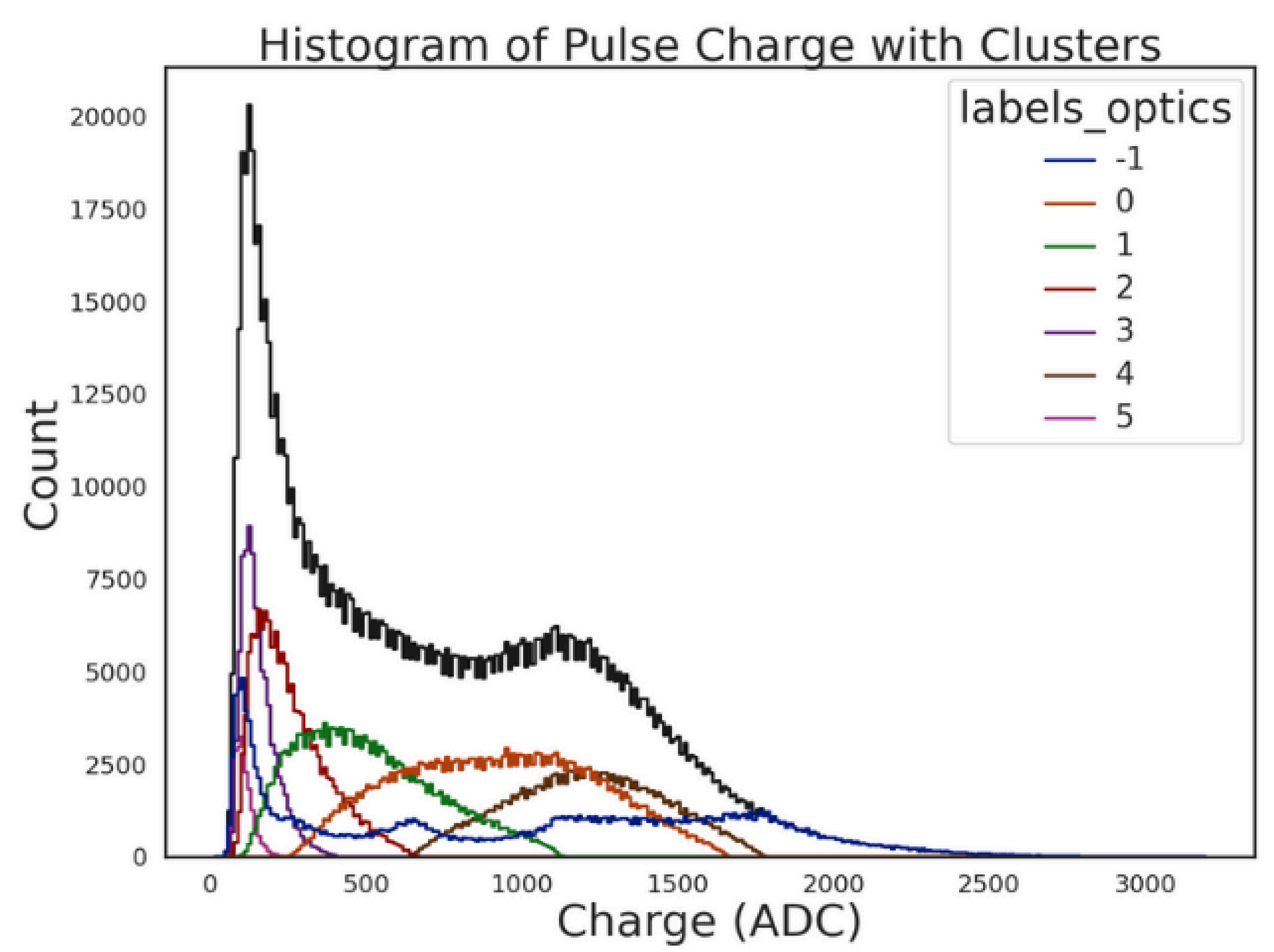


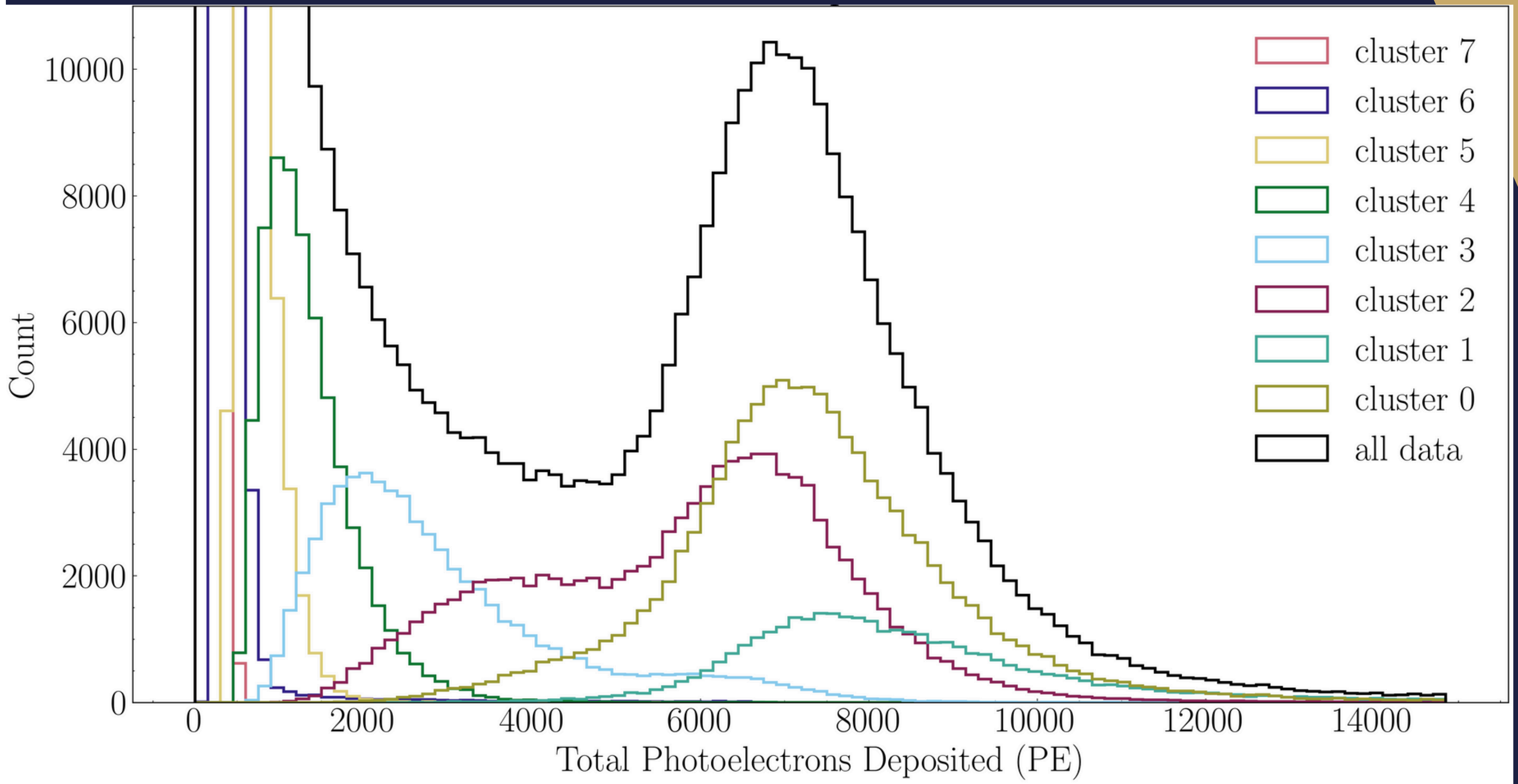
Photons
Electrons & Positrons
Muons
Neutron
Hadron
Composition Percentage (%)
0
20
40
60
80
100
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
cluster 0 + 1 + 2
cluster 3 + 4
cluster 5 + 6 + 7
all data
Count
Total Photoelectrons Deposited (PE)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000

3. RESULTS: VARIABILITY OF RESULTS

No.	Photons	Electrons & Positron	Muon	Neutron	Hadron
0	1.41% ± 0.12%	1.61% ± 0.11%	96.58% ± 0.24%	0.20% ± 0.02%	0.20% ± 0.02%
1	5.13% ± 0.34%	4.53% ± 0.25%	89.25% ± 0.59%	0.49% ± 0.03%	0.60% ± 0.02%
2	0.91% ± 0.15%	1.35% ± 0.15%	97.27% ± 0.30%	0.22% ± 0.02%	0.33% ± 0.02%
3	23.08% ± 0.96%	15.20% ± 0.55%	58.46% ± 1.41%	1.38% ± 0.07%	1.88% ± 0.08%
4	46.45% ± 0.87%	22.78% ± 0.44%	27.86% ± 1.15%	1.34% ± 0.05%	1.58% ± 0.08%
5	62.45% ± 0.51%	22.92% ± 0.25%	12.76% ± 0.67%	0.90% ± 0.03%	0.97% ± 0.07%
6	70.45% ± 0.43%	20.01% ± 0.25%	6.71% ± 0.37%	2.12% ± 0.15%	0.71% ± 0.03%
7	80.60% ± 0.61%	9.16% ± 0.07%	1.30% ± 0.06%	8.23% ± 0.61%	0.71% ± 0.03%

3. RESULTS: SIDE BY SIDE





4. FUTURE WORK

- More exploration of features
- Convert code into a production library
- Optimize code for HPC environment

Thank You!!!

Questions?

Contacto: ttorres@herrera.unt.edu