Selection of CRE in Fermi-LAT data with Unsupervised Learning techniques

Raffaella Bonino
University and INFN Torino
rbonino@to.infn.it

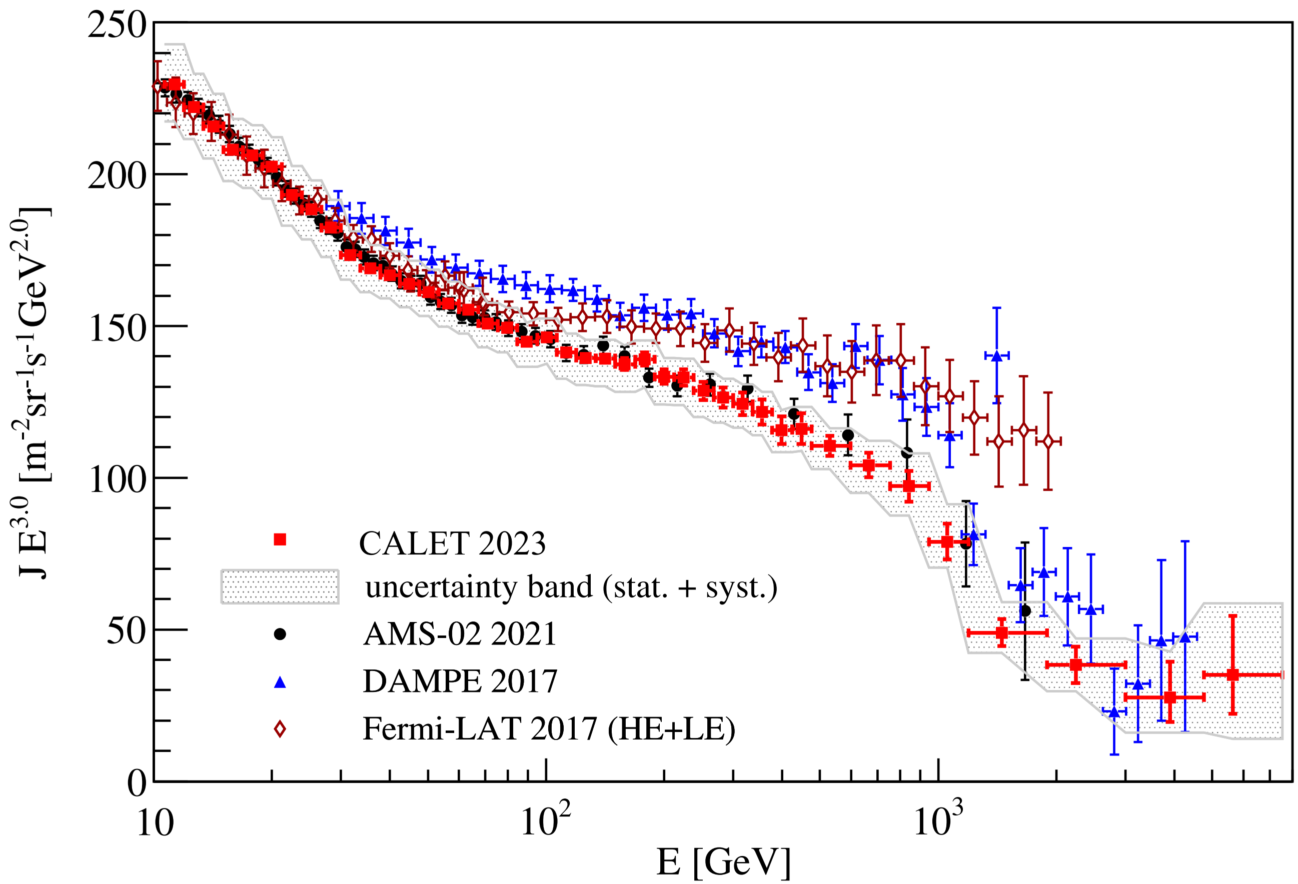
on behalf of the Fermi-LAT Collaboration

Outline

- **State-of-the-art Cosmic Ray Electrons direct measurements**
- **Fermi-LAT as CRE detector**
- **New analysis technique** based on Unsupervised Machine Learning
  - Method
  - Results (based on simulated data)
- **Conclusions**

CRE (e^+e^-) measurements

- CRE spectrum in the energy range 100 GeV - few TeVs can provide evidence of local **CRE sources** of astrophysical (SNR & PWN) or exotic origin (DM)

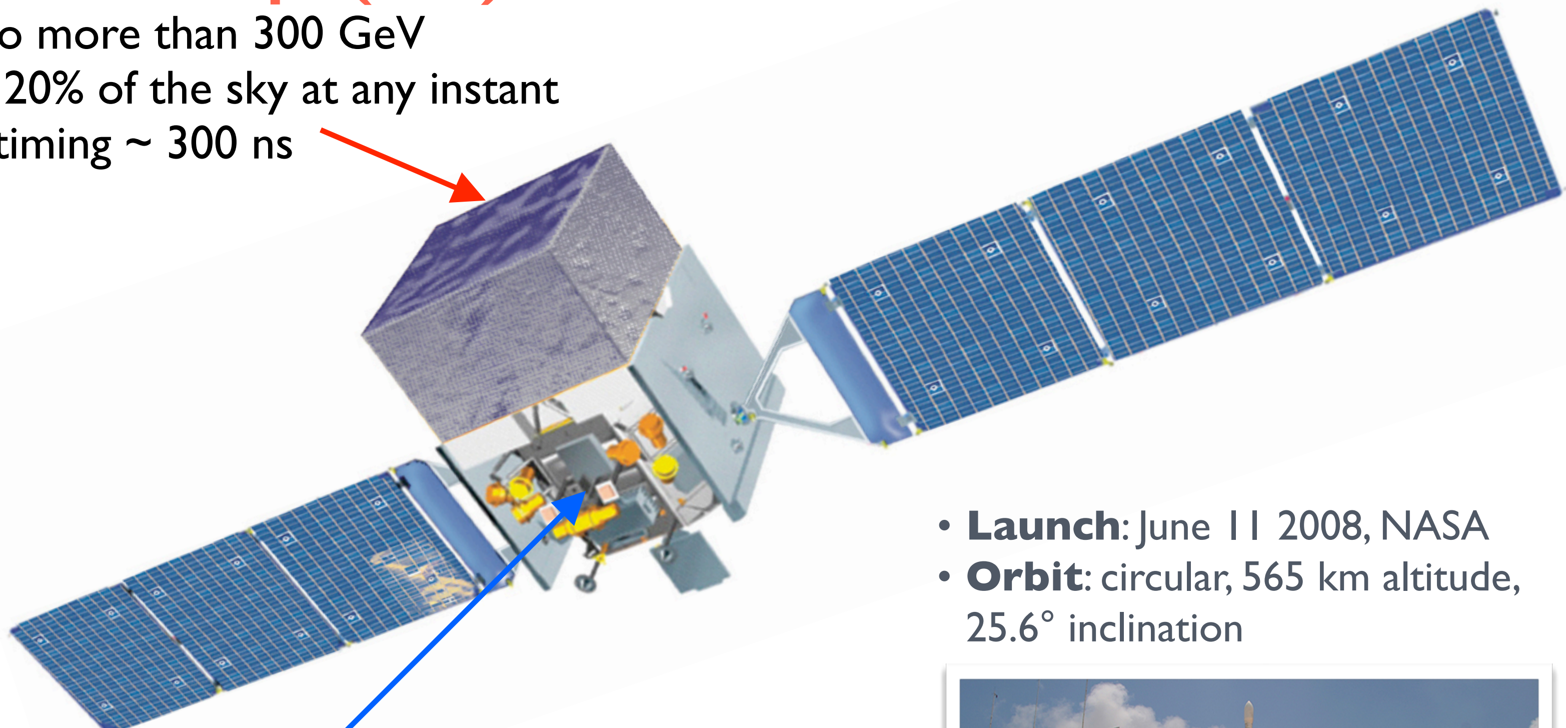


- **Significant differences** among some spectra, particularly **at higher energies** where uncertainties are more considerable

The Fermi observatory

Large Area Telescope (LAT):

- 20 MeV to more than 300 GeV
- observes 20% of the sky at any instant
- absolute timing ~ 300 ns



- **Launch**: June 11 2008, NASA
- **Orbit**: circular, 565 km altitude, 25.6° inclination

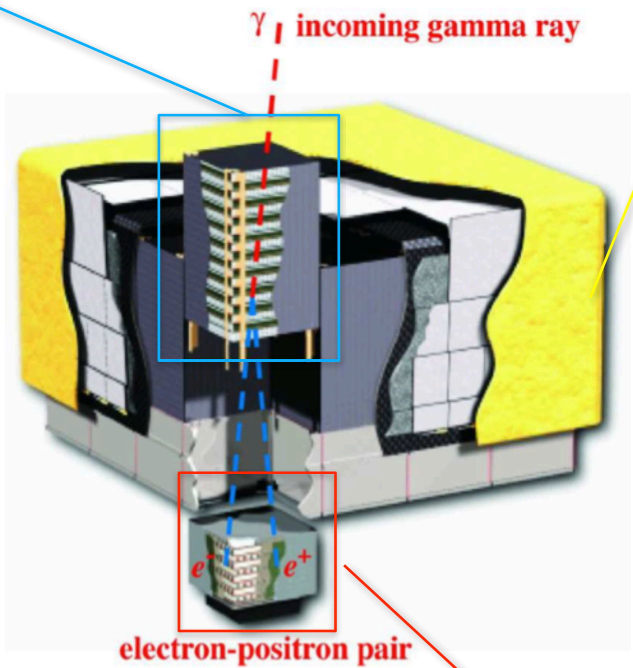
Gamma-ray Burst Monitor (GBM):

- 8 keV to 40 MeV
- observes entire unocculted sky
- absolute timing ~ 2μs
- compute burst location

The Fermi LAT detector

TRACKER-CONVERTER

- High precision tracking
- 18 x, y tracking planes: Silicon Strip Detector (73 m^2 of Si active area)
- 16 planes of tungsten conversion foils:
  - "FRONT" → first 12 "thin" layers
  - "BACK" → next 4 "thick" layers
  - ➡ 1.5 radiation lengths on-axis



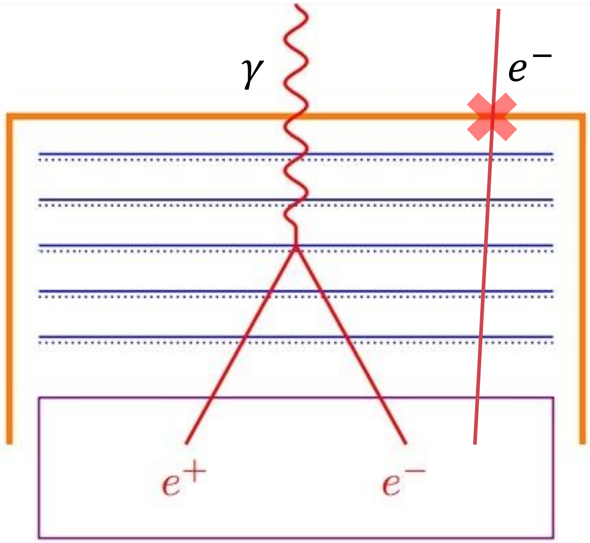
ANTICOINCIDENCE DETECTOR

- Charged-particle bkg rejection (with 0.9997 efficiency)
- Plastic scintillator, WLS fibers
- 89 segmented tiles to minimize self-veto at high E

CALORIMETER

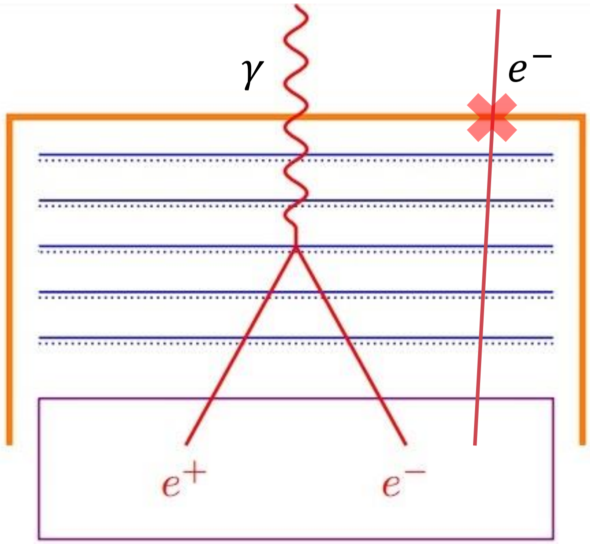
- energy deposition and shower development imaging
- 1536 CsI(Tl) crystals; hodoscopic
- 8.6 radiation lengths on-axis.

LAT as electron detector

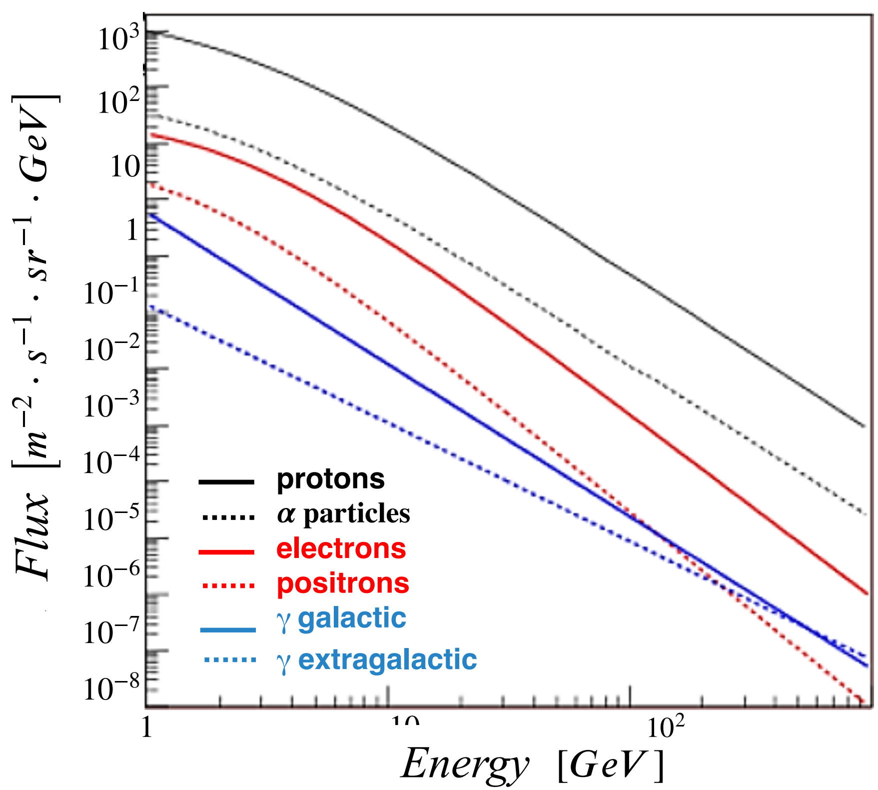


- LAT is designed for electromagnetic showers:
  - naturally including electrons
  - **event reconstruction** works also for e^+e^-

LAT as electron detector



- LAT is designed for electromagnetic showers:
  - naturally including electrons
  - **event reconstruction** works also for e^+e^-



- Electron identification requires dedicated **event selection**

Analysis methods

- **GOAL:** identify ***electrons and positrons*** out of *cosmic rays background* in Fermi-LAT data (and compute their energy spectrum)

Analysis methods

- **GOAL:** identify ***electrons and positrons*** out of *cosmic rays background* in Fermi-LAT data (and compute their energy spectrum)

- **Supervised Learning:**
  - Boosted Decision Trees: published in 2017
  - Neural Networks → similar results

  ☑ Supervised approach implies training on Monte Carlo simulations:
  → strong dependence on models and simulations quality
  → **sensitive to important systematic uncertainties or biases**

Analysis methods

- **GOAL:** identify ***electrons and positrons*** out of *cosmic rays background* in Fermi-LAT data (and compute their energy spectrum)

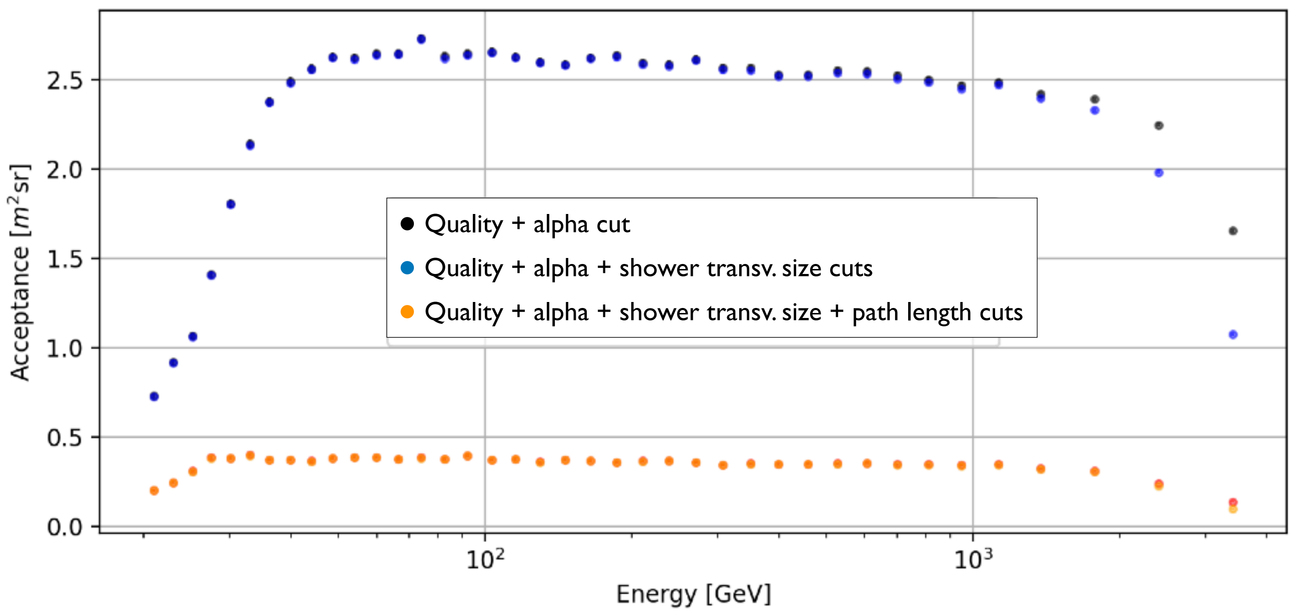
- **Supervised Learning:**
  - Boosted Decision Trees: published in 2017
  - Neural Networks → similar results
  - ☑ Supervised approach implies training on Monte Carlo simulations:
    → strong dependence on models and simulations quality
    → **sensitive to important systematic uncertainties or biases**

⇩

- **Unsupervised Learning:**
  - ✓ No labels and minimum human supervision:
    → independence of models / MC → **systematic uncertainties reduced**
  - ☑ ***Difficulty***: very different cluster sizes (i.e. background dominant wrt signal)
  - ☑ ***Potential drawback***: irreducible bkg (hadronic shower fluctuating into e.m. shower)

Data set

- **Data set:** *MC* e^+e^- and *p* with realistic flux ratio (spectral index -3 for e and -2.7 for p)
- **Energy range:** 70 GeV - 750 GeV
- **Cuts**:
  - basic quality cuts (trigger, filter, track found and minimal PSF quality)
  - remove alphas and heavier ions (residual contamination < few ‰ wrt protons)
  - cut on shower transverse size to reduce the proton background
  - path length > 12 X_0 in the CAL



Unsupervised Approach

1. **Selection of variables**

2. **Dimensionality reduction** with UMAP algorithm

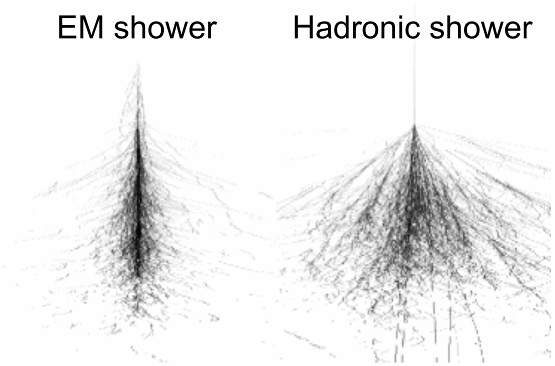
3. Agglomerative Hierarchical **Clustering**

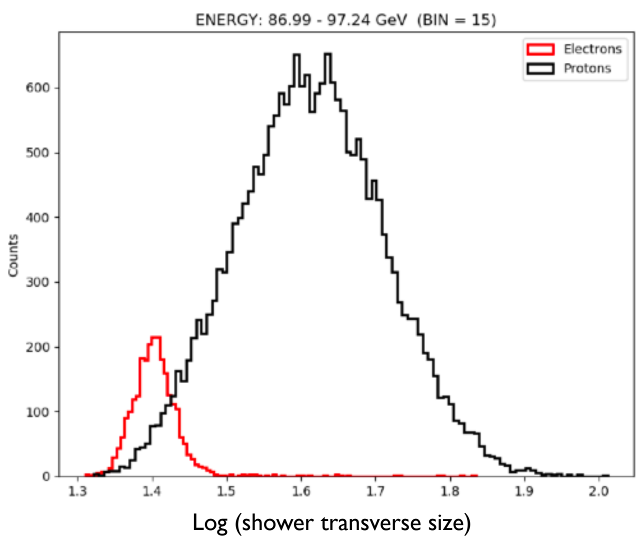
4. **Iterations**

Unsupervised Approach

1. Selection of variables

- highlight the differences in the shower topology between protons and electrons

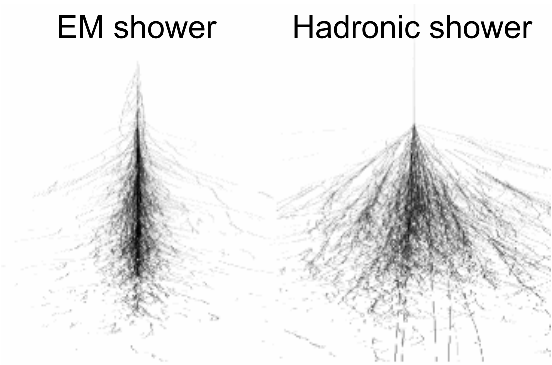


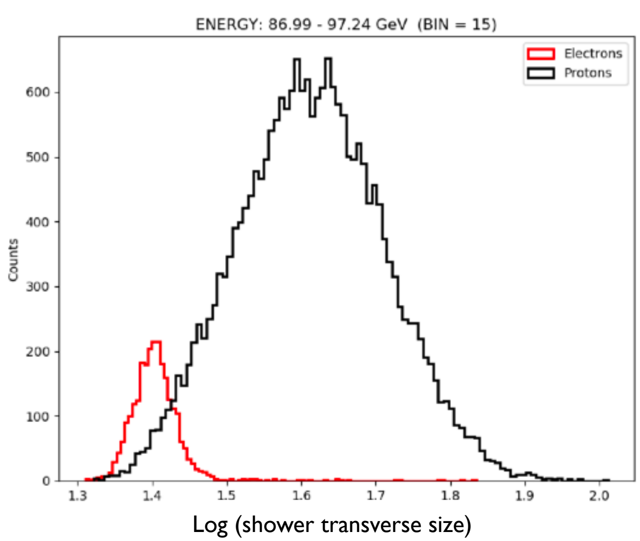


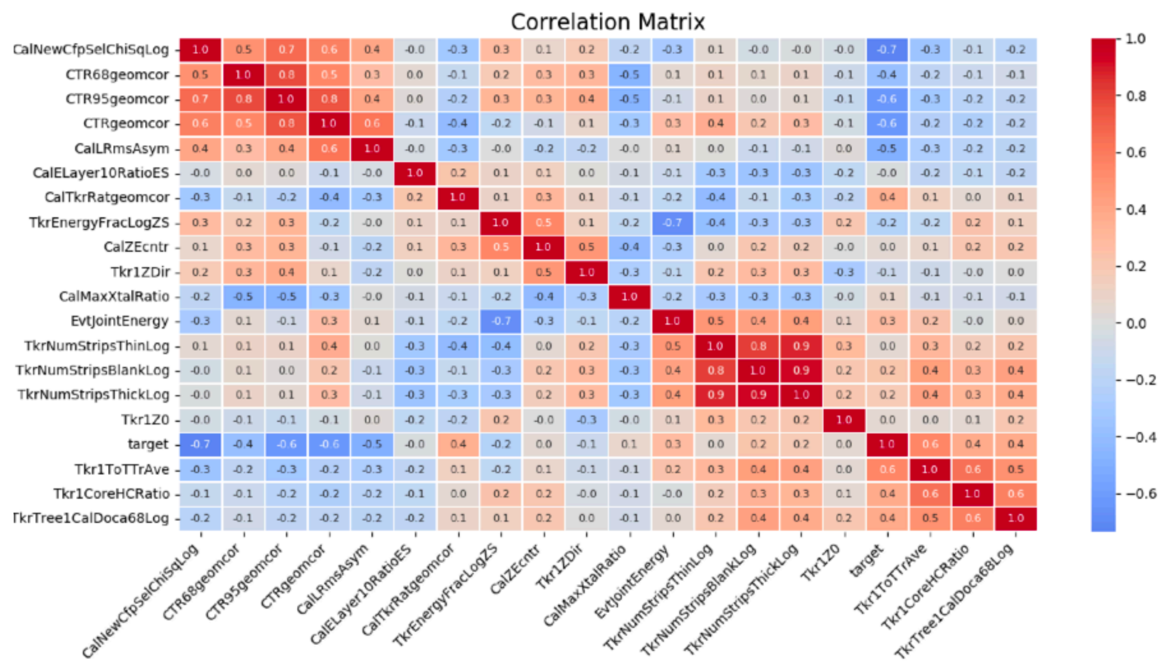
Unsupervised Approach

1. Selection of variables

- highlight the differences in the shower topology between protons and electrons
- avoid highly correlated variables







Unsupervised Approach

1. **Selection of variables**

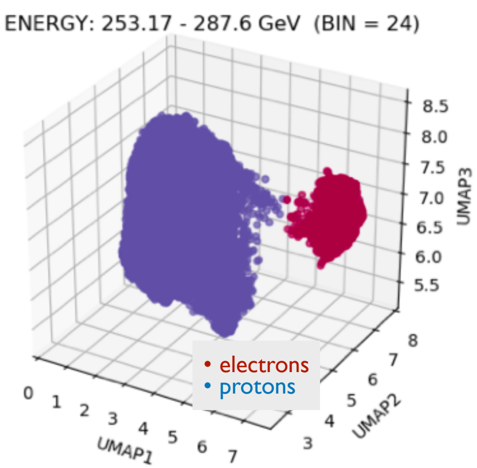
2. **Dimensionality reduction** with UMAP algorithm (10.21105/joss.00861)
   - clustering algorithms cannot easily handle variable spaces of high dimensions
     → map data onto lower dimensionality spaces, preserving relationships / patterns

Unsupervised Approach

1. **Selection of variables**

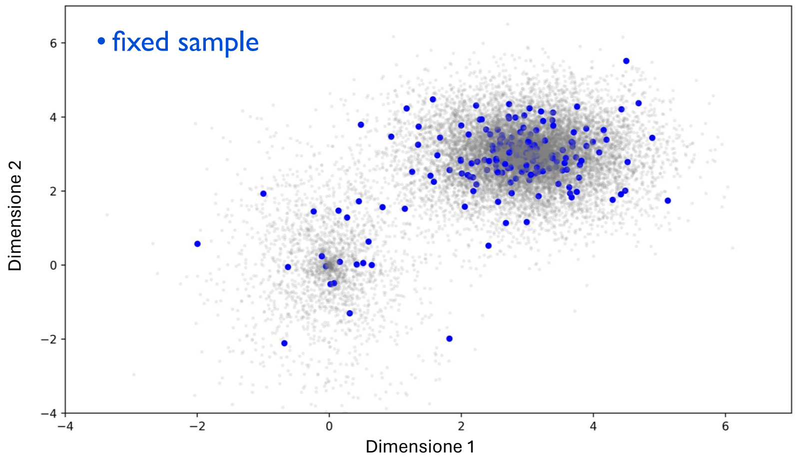
2. **Dimensionality reduction** with UMAP algorithm

3. Agglomerative Hierarchical **Clustering**
   - search for *similarity* in events
   - progressively finds the 2 closest differently labeled elements and tag them with the same label until only 2 labels remain → separate data into 2 clusters



Unsupervised Approach

1. **Selection of variables**
2. **Dimensionality reduction** with UMAP algorithm
3. Agglomerative Hierarchical **Clustering**
4. **Iterations**:
   - 10 000 data reprocessed x 20 times
   - In each of the 20 iterations:
     - 10% of data fixed + 90% changes



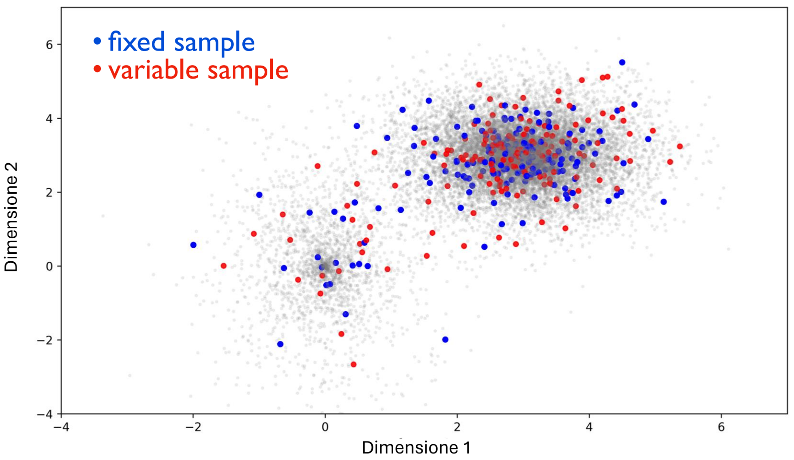
Unsupervised Approach

1. **Selection of variables**

2. **Dimensionality reduction** with UMAP algorithm

3. Agglomerative Hierarchical **Clustering**

4. **Iterations**:
   - 10 000 data reprocessed x 20 times
   - In each of the 20 iterations:
     - 10% of data fixed + 90% changes



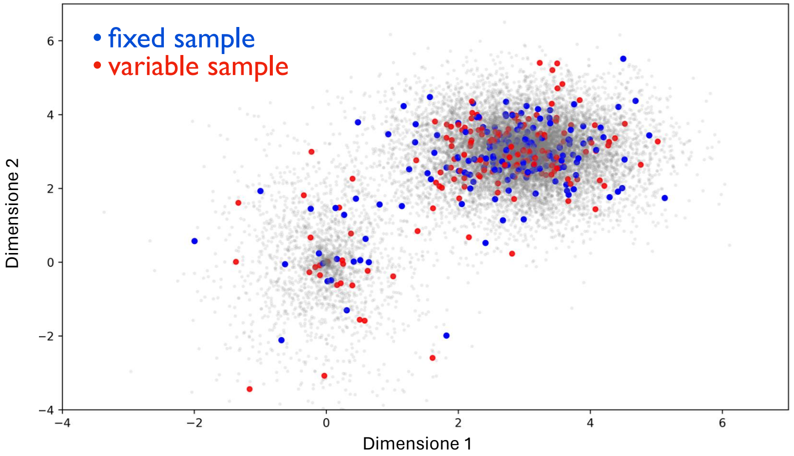
Unsupervised Approach

1. **Selection of variables**

2. **Dimensionality reduction** with UMAP algorithm

3. Agglomerative Hierarchical **Clustering**

4. **Iterations**:
    - 10 000 data reprocessed x 20 times
    - In each of the 20 iterations:
        - 10% of data fixed + 90% changes



Unsupervised Approach

1. **Selection of variables**

2. **Dimensionality reduction** with UMAP algorithm

3. Agglomerative Hierarchical **Clustering**

4. **Iterations**:
   - 10 000 data reprocessed x 20 times
   - In each of the 20 iterations:
     - 10% of data fixed + 90% changes
       → **electrons fraction is estimated**
     - all the outputs for dimensions between 3 and 10 are considered and averaged

Unsupervised Approach

1. **Selection of variables**

2. **Dimensionality reduction** with UMAP algorithm

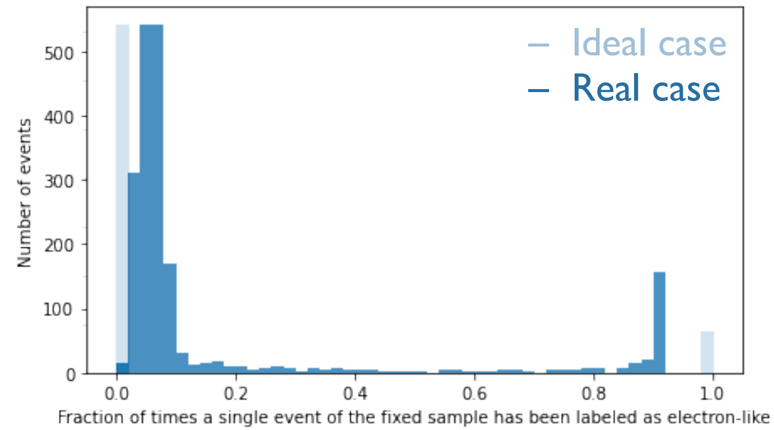
3. Agglomerative Hierarchical **Clustering**

4. **Iterations**:
   - 10 000 data reprocessed x 20 times
   - In each of the 20 iterations:
     - 10% of data fixed + 90% changes
       → **electrons fraction is estimated**
     - all the outputs for dimensions between 3 and 10 are considered and averaged
   - The entire procedure is repeated for:
     - all the considered energy bins
     - 3 different seeds (3 random selection of the fixed dataset)

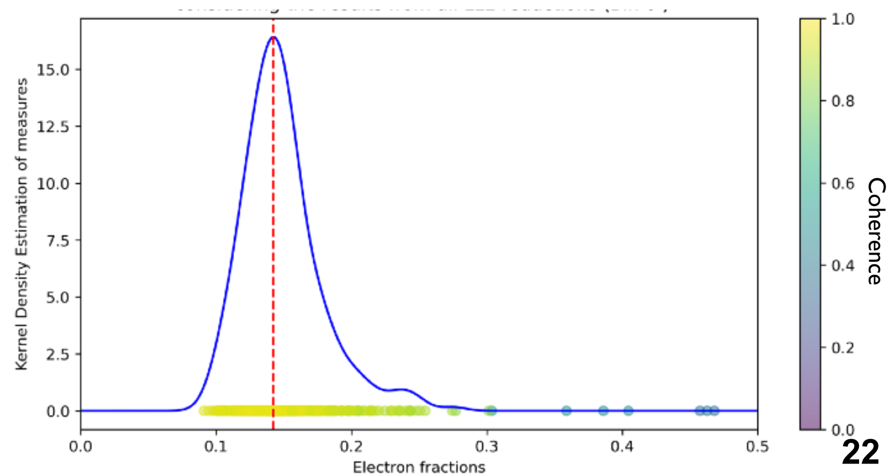
Figure of Merit: ***coherence***

- Evaluate the **self-coherence of the algorithm** in order to:
  - choose the optimal parameters (e.g. samples dimension, nb of iterations, …)
  - quantify the **reliability** of each electrons fraction estimation

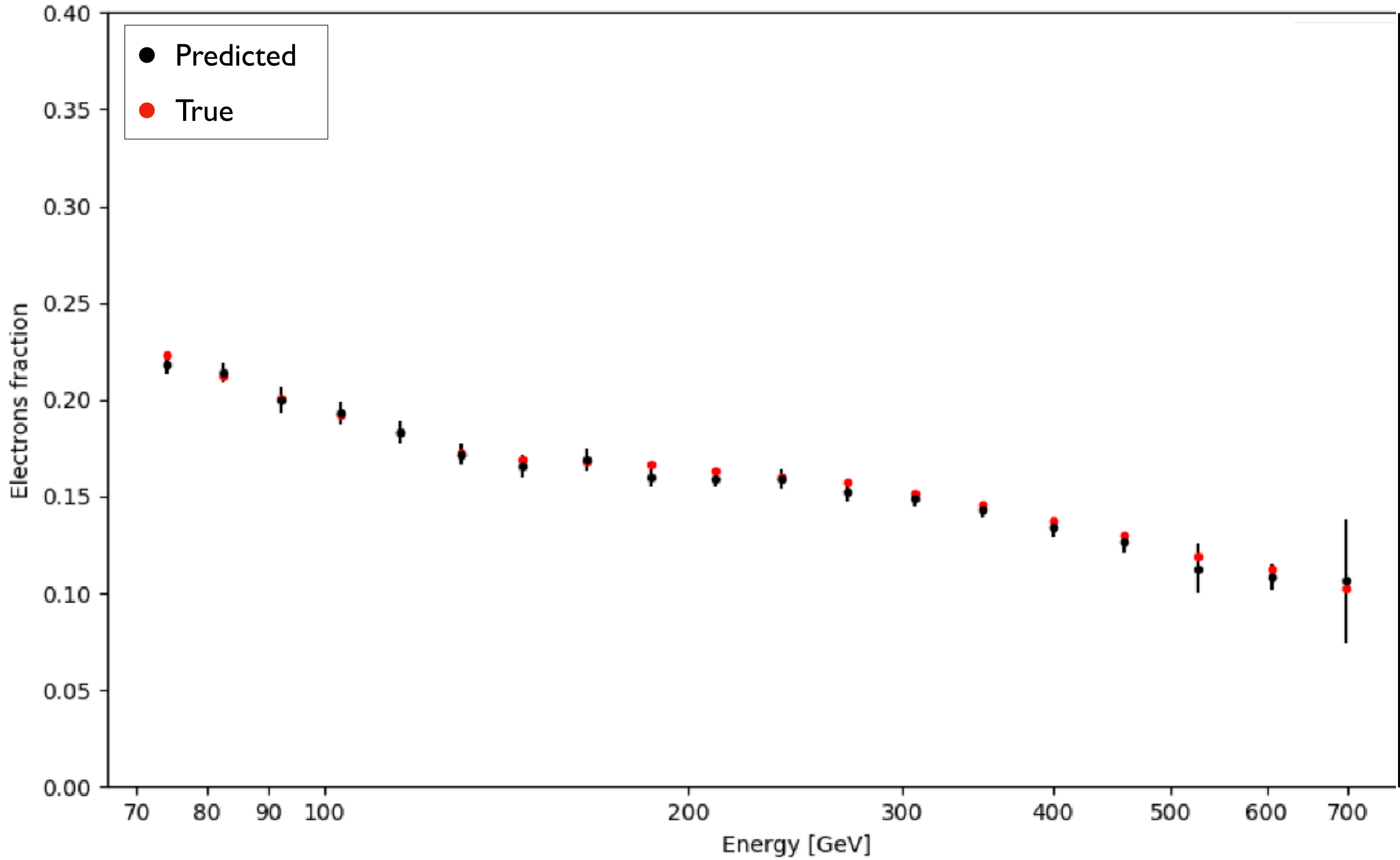
- *How?* By comparing labels assigned to fixed samples through the 20 iterations



- Used to estimate the most probable **e⁺e⁻ fraction** in each energy bin



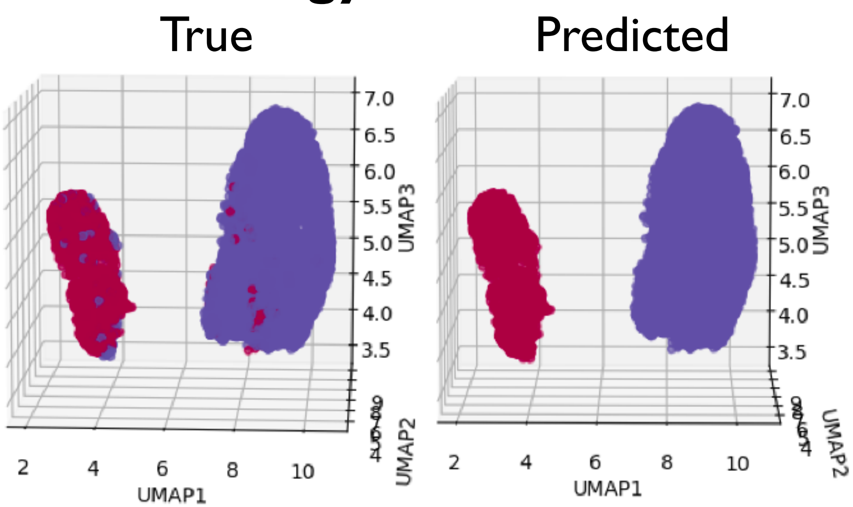
Results: MC electrons fraction



- ***Predicted*** electron fraction is always **compatible** with the ***true*** value
  → *irreducible background* doesn't seem to introduce an evident bias
- Results can be considered reliable up to 750 GeV (too few MC events above)

Results: MC electrons fraction

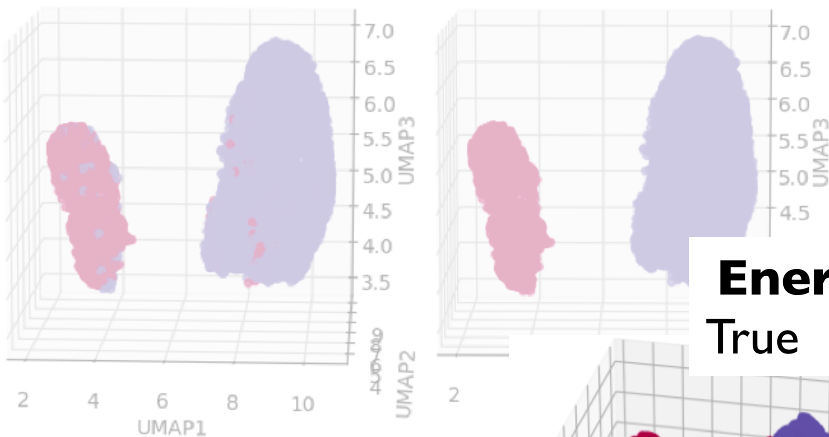
Energy: 70-78 GeV



Results: MC electrons fraction

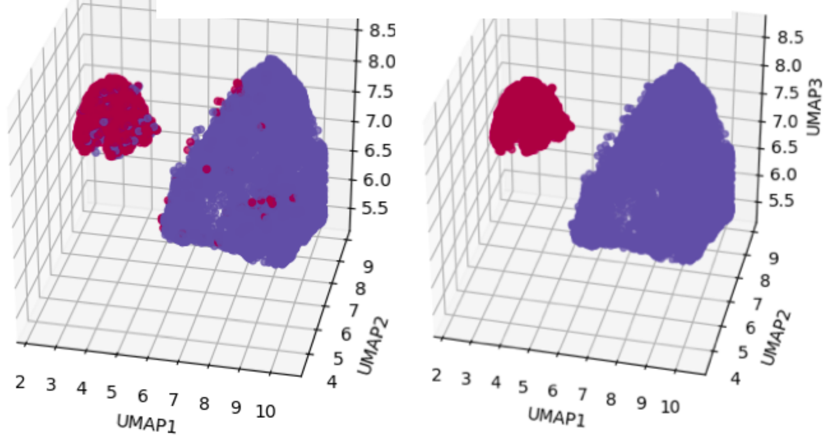
Energy: 70-78 GeV

True Predicted



Energy: 155-175 GeV

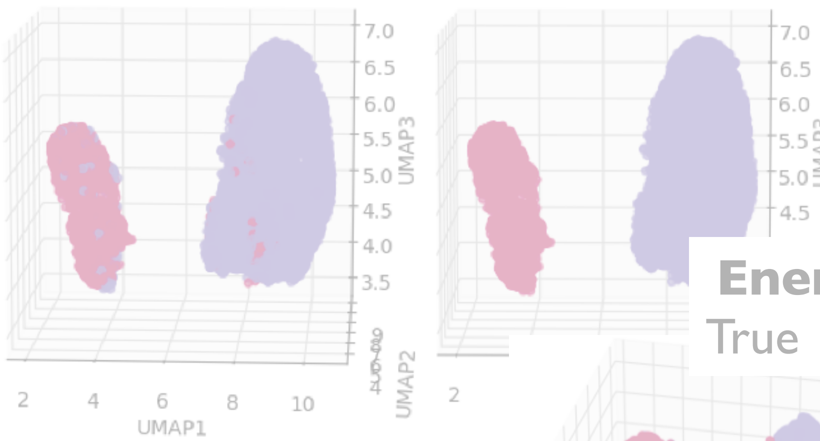
True Predicted



Results: MC electrons fraction

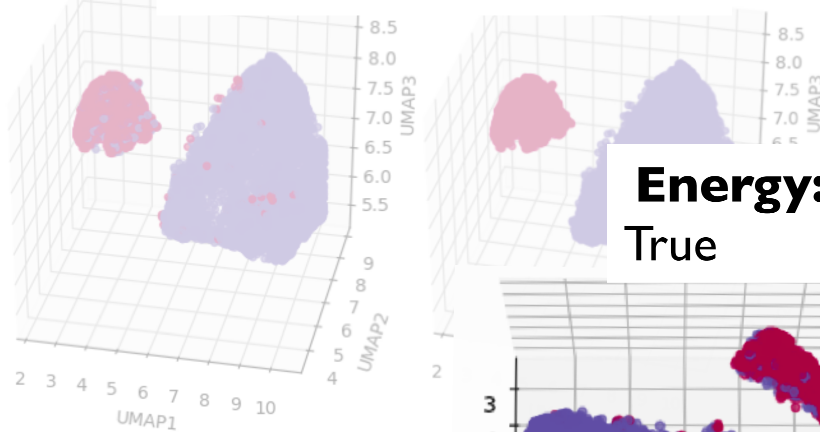
Energy: 70-78 GeV

True Predicted



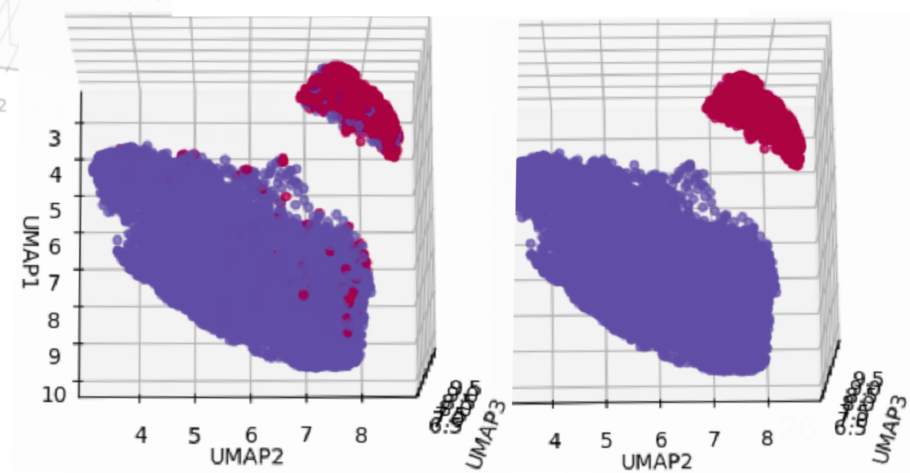
Energy: 155-175 GeV

True Predicted

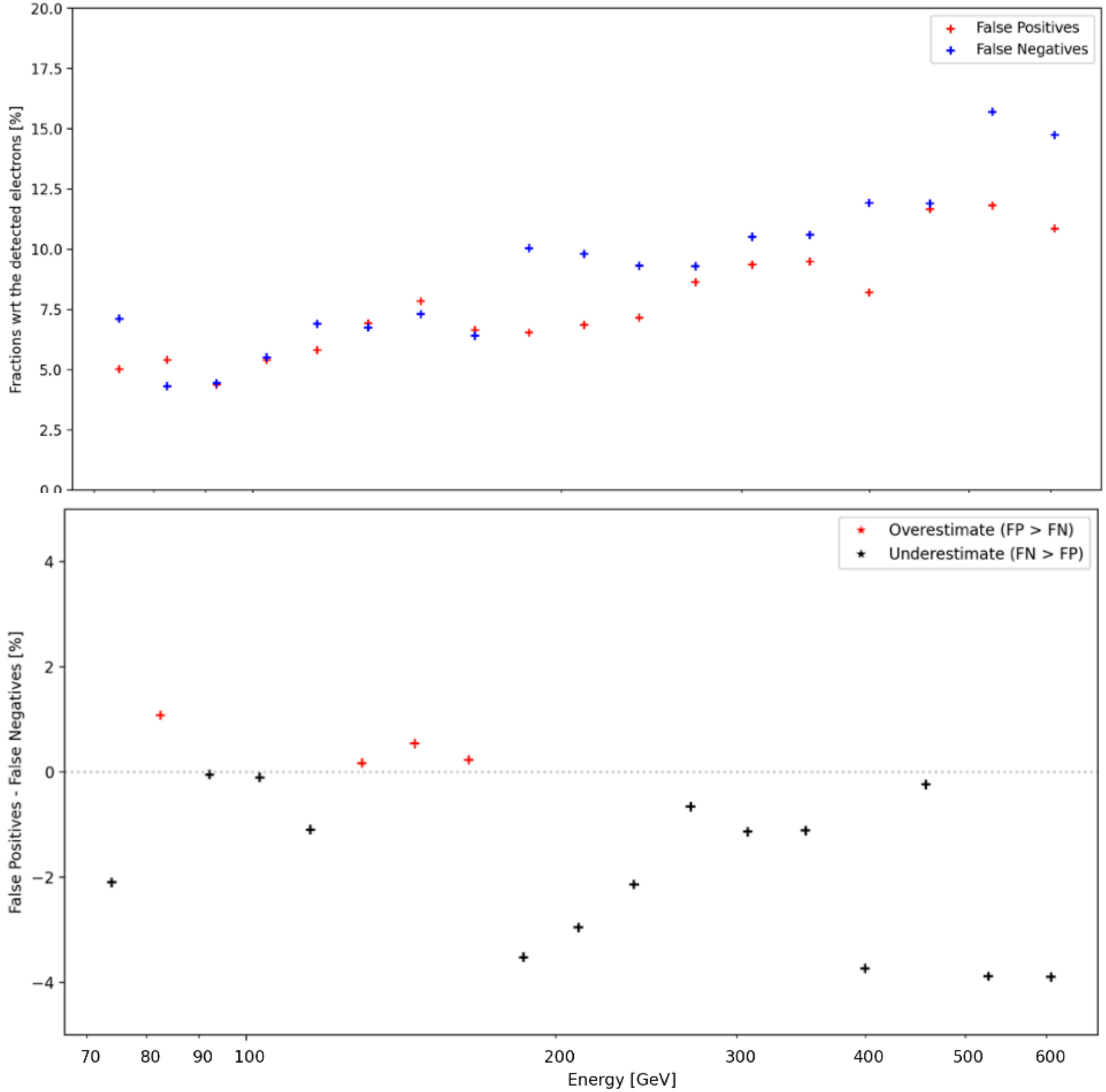


Energy: 288-327 GeV

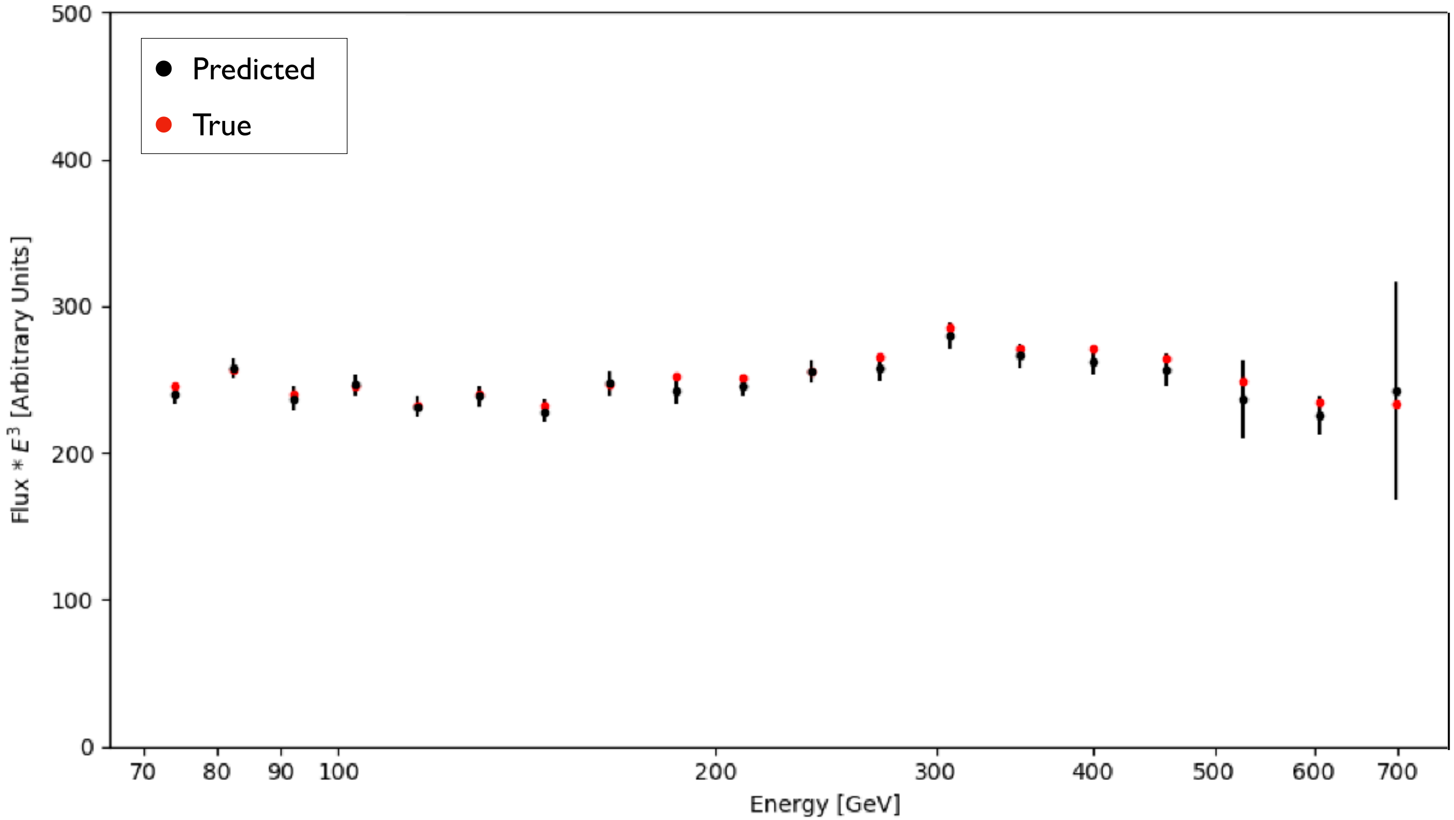
True Predicted



Results: MC electrons fraction



Results: MC CRE spectrum



- ***Predicted*** flux is always **within 1 σ** wrt ***true*** one (input spectral index = -3)
- Fluctuations are attributable to the acceptance

Conclusions

- An approach based on **Unsupervised Learning techniques** has been developed for the selection of CRE in Fermi-LAT

- **Feasibility study** successfully accomplished on Monte Carlo data set:
  - electron fraction correctly estimated (always compatible with true value)
  - false positives always lower than 12%

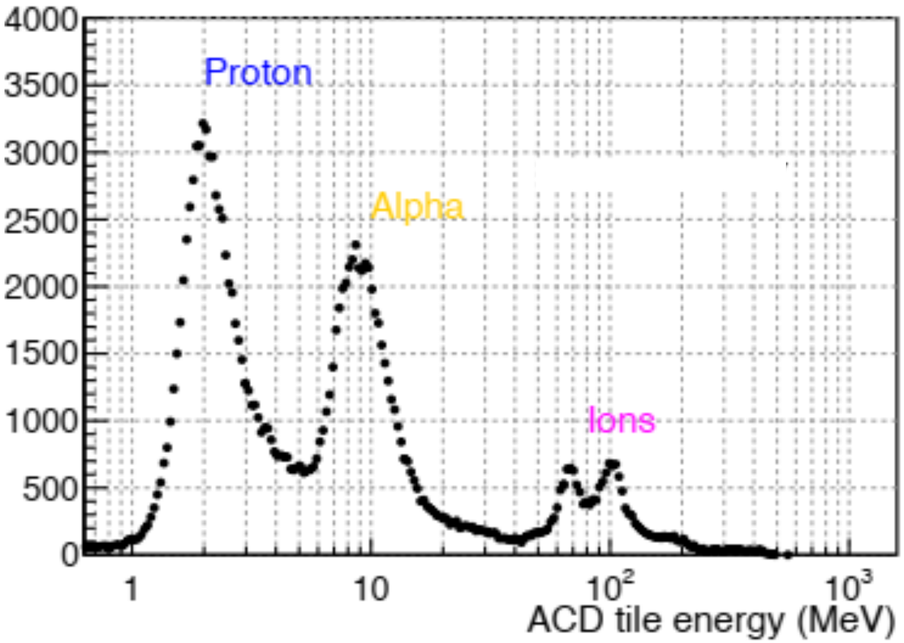
- The method will be **soon applied to experimental data**

- The method could be applied to **other science cases**

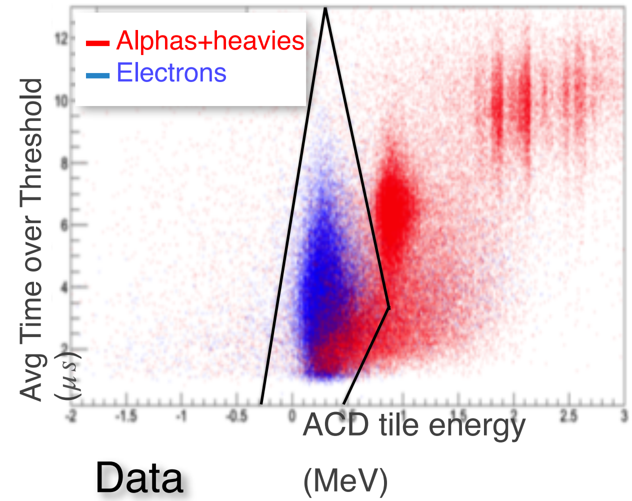
ACKNOWLEDGMENTS: *this is part of the project "SKYNET: Deep Learning for Astroparticle Physics", PRIN 2022 (CUP: D53D23002610006).*

Backup slides

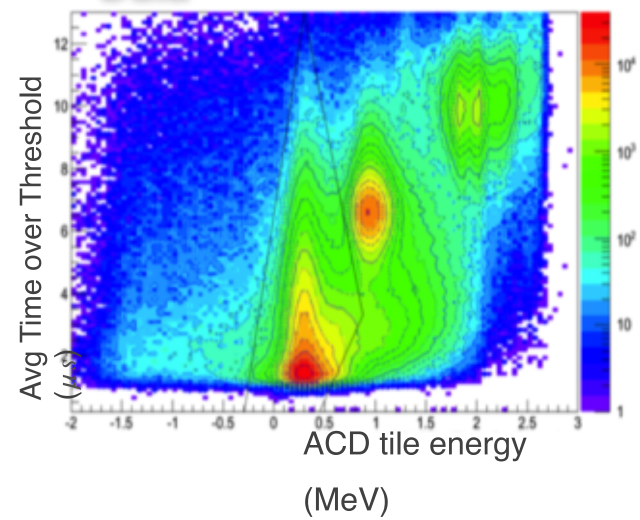
Alpha cut



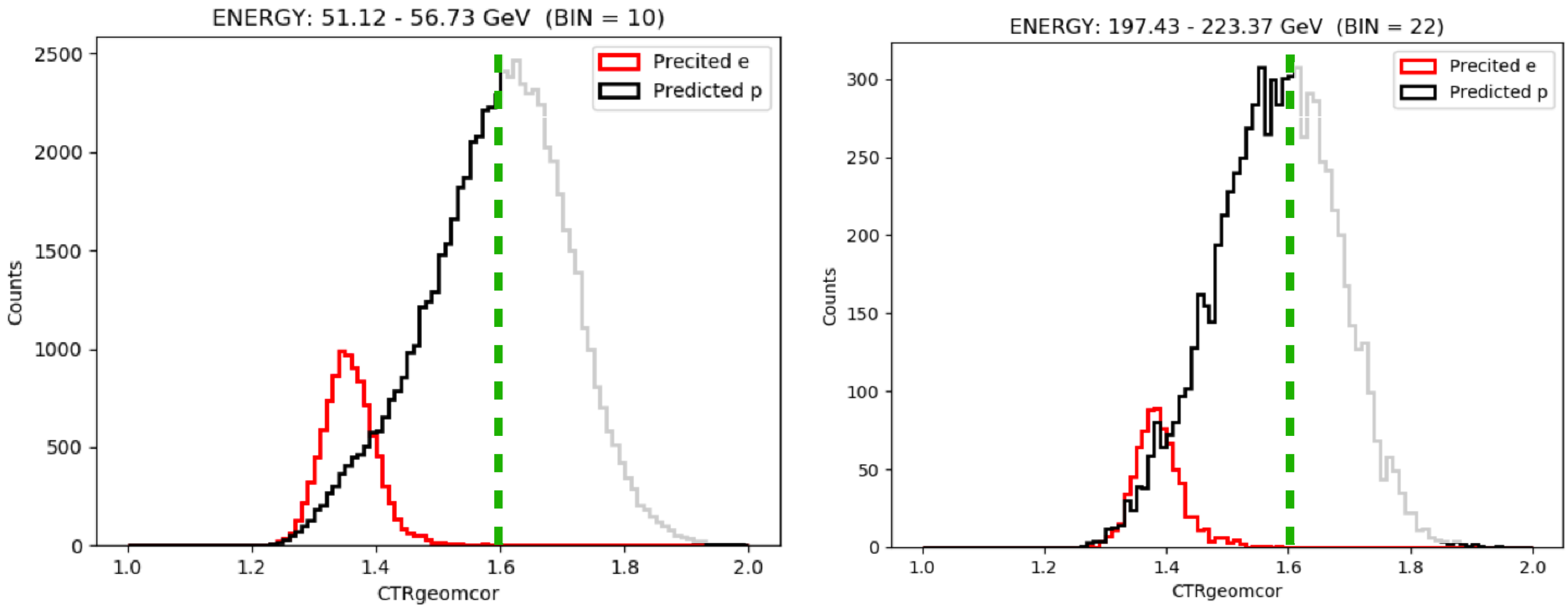
Simulation



Data



CTRgeomcor cut



- ~50% of total events are removed, but only a **negligible fraction of e^+e^- is affected** (between 0 and 0.5%), with not evident dependence on energy.
- Total fraction of electrons goes from 10 to 20%.
- Clustering performance seem to improve significantly.

Energy resolution

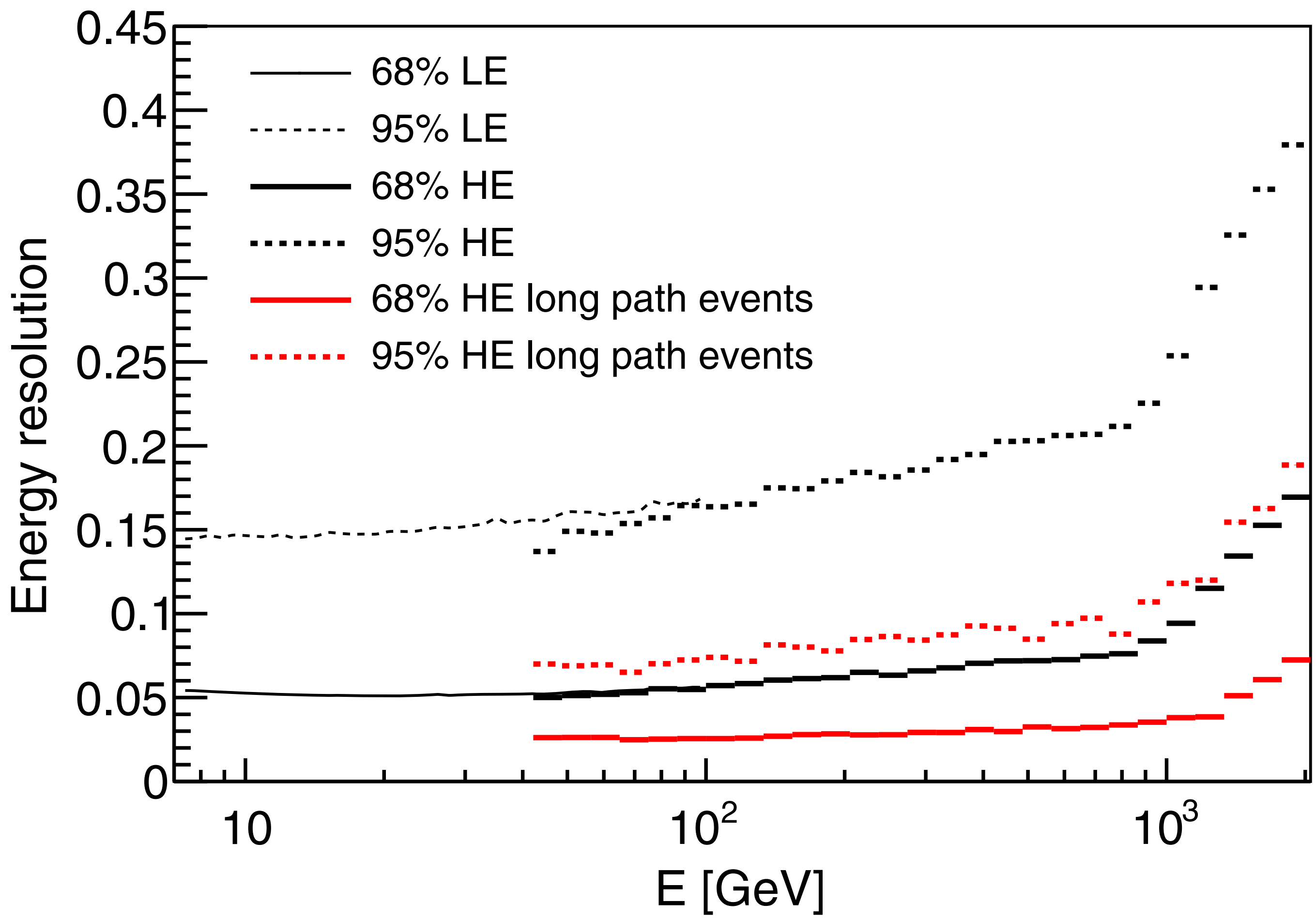
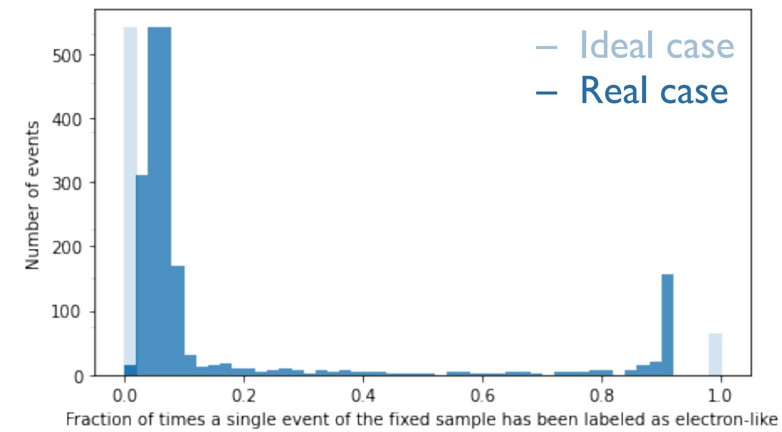


Figure of Merit: ***coherence***

- Evaluate the **self-coherence of the algorithm** in order to choose the optimal parameters (e.g. samples dimension, nb of iterations, …)

  - *How?* By comparing the labels assigned to the sample of 900 fixed events through the different iterations



  - Used to:
    - quantify the **reliability** of each electrons fraction estimation