Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Reconstruction of multiple calorimetric clusters in the LHCf experiment with machine learning techniques

Giuseppe Piparo[1,2], on behalf of the LHCf collaboration

1. Istituto Nazionale di Fisica Nucleare (INFN), Catania section
2. University of Catania, Department of Physics and Astronomy

42nd International Conference on High Energy Physics (ICHEP 2024), Prague (CZE), Jul 17-24, 2024

INFN Catania Istituto Nazionale di Fisica Nucleare

WHY TO IMPROVE LHCF RECONSTRUCTION WITH ML?

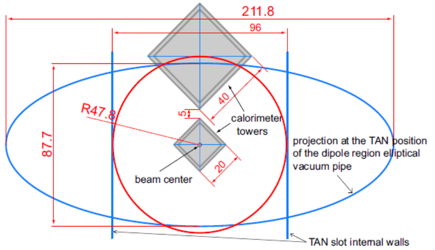
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

The LHCf Experiment at LHC

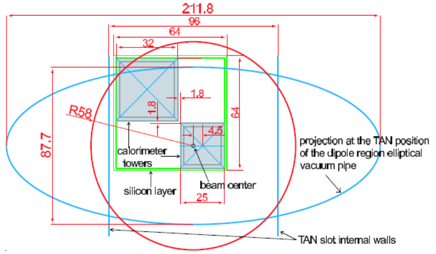
- A **unique experiment** designed to measure neutral particle production in the **forward pseudorapidity region**.
- Composed by two **sampling and immaging** calorimeters (ARM1 & ARM2), located at about ± 141 m from the LHC Interaction Point 1 (IP1).

Transversal view

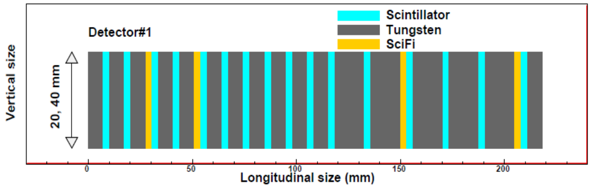
ARM1



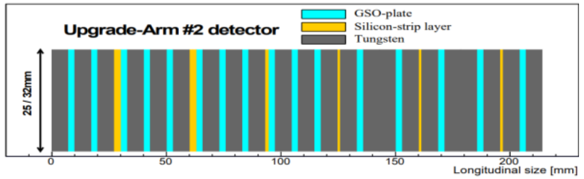
ARM2

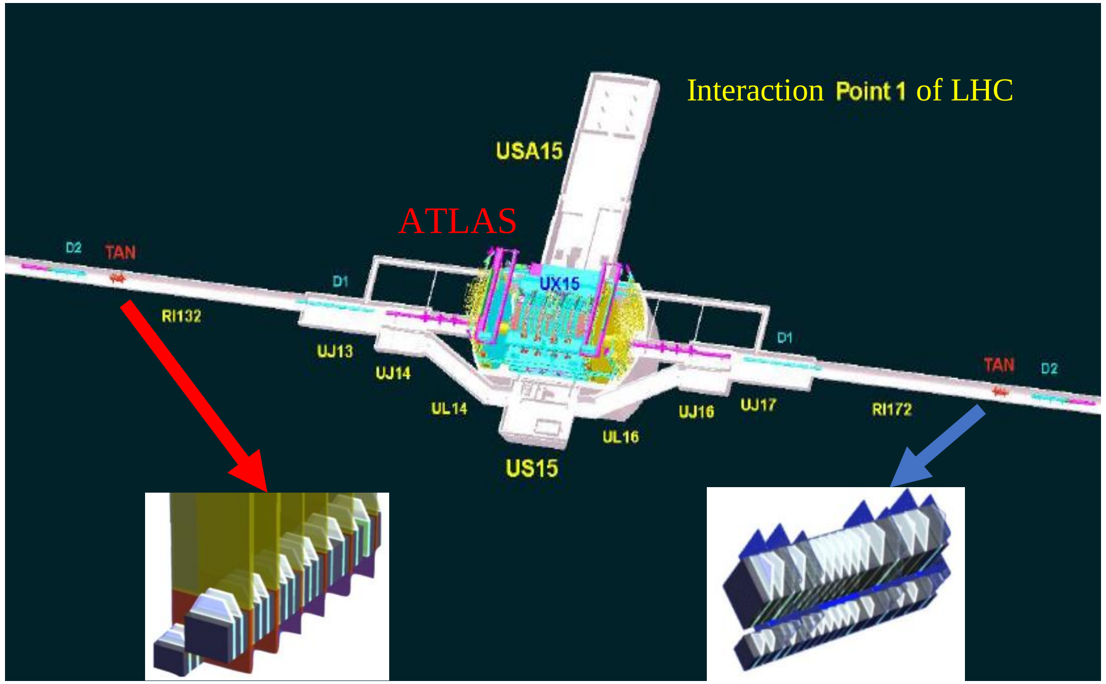


Longitudinal view



$44\ X_0$ and $1.6\ \lambda_I$ deep





LHCF-ARM2

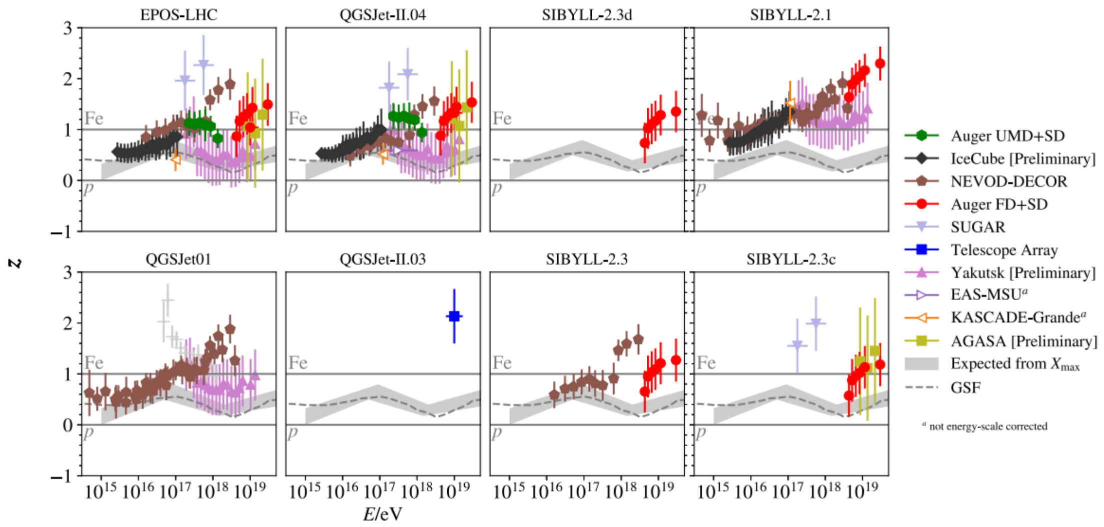
- **Energy resolution** <5% for photons and 35-40% for neutrons
- Tracking with 4 XY **silicon microstrips** layers.
- **Position resolution** $\approx 40\ \mu m$.

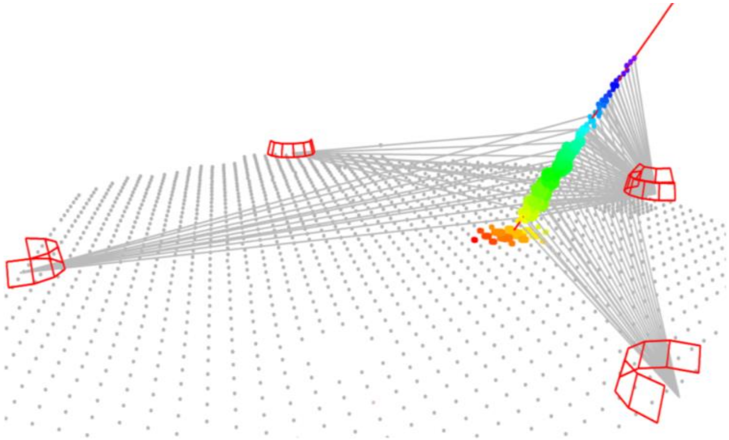
LHCF-ARM1

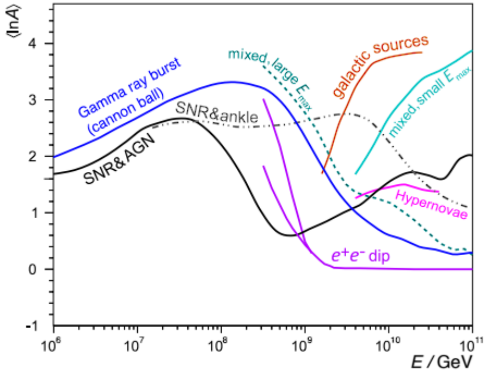
- Same Energy resolution as ARM2.
- Tracking with 4 GSO scintillator layers.
- **Position resolution** $\approx 200\ \mu m$.

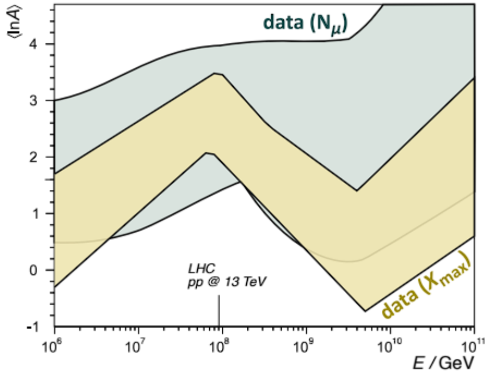
Experimental purpose of LHCf

- The main **hadronic interaction models** (HIM) (like QGSJET, SIBYLL or EPOS) suffer of large discrepancy due to limited understanding of the **soft QCD processes**.
- This is reflected on **large uncertainties** induced in the results of the ground-based cosmic rays experiments, due to the dependency of **air shower modeling** on HIM.
- LHCf provides neutral particles' energy and momentum distributions in the **forward region** to test and calibrate the models.









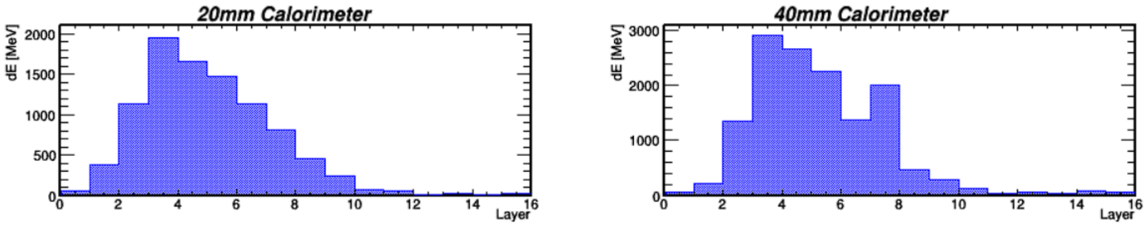
ASS 367, 1 (2022)

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Reconstruction of multiple calorimetric clusters in LHCf

- The total energy is reconstructed by summing the **calibrated energy releases** in the GSO layers. This sum is then converted to the total energy using a function derived from Monte Carlo simulations.

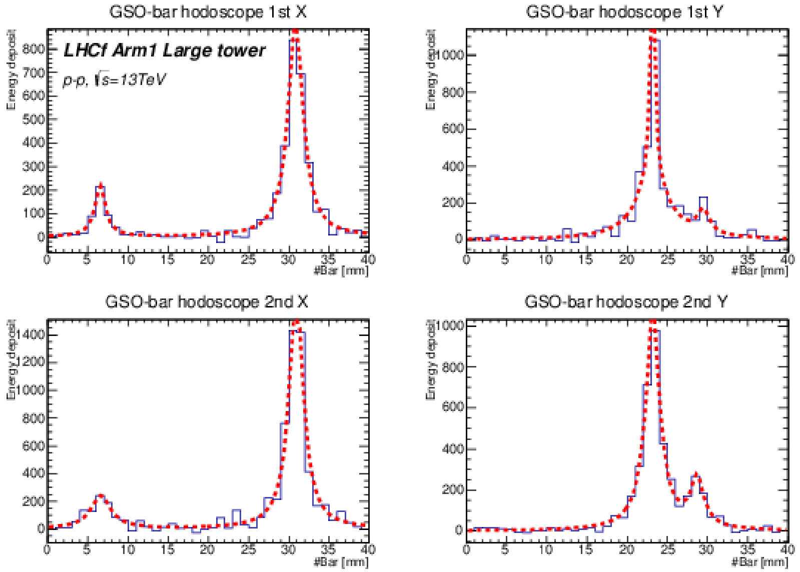
Energy releases



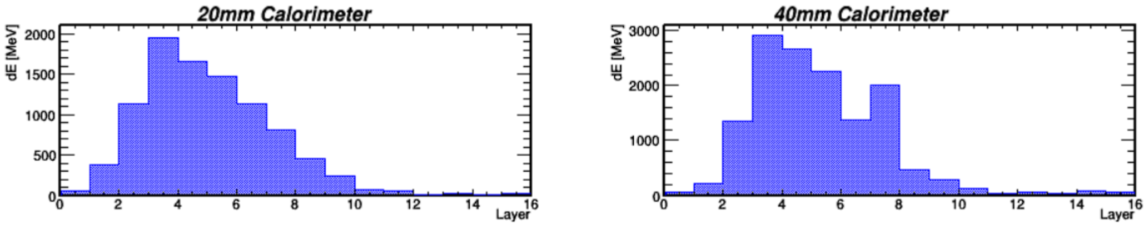
Reconstruction of multiple calorimetric clusters in LHCf

- The total energy is reconstructed by summing the **calibrated energy releases** in the GSO layers. This sum is then converted to the total energy using a function derived from Monte Carlo simulations.
- To reconstruct the position of hitting particles, we use the **transversal profile** of tracking detectors, by finding the peaks using TSpectrum and fitting them with a **3-component Lorentzian function for each peak**.
- The current energy-sharing method uses the **ratio of peak heights** for each particle to share the energy between the events.

Transversal profile with 2 clusters



Energy releases



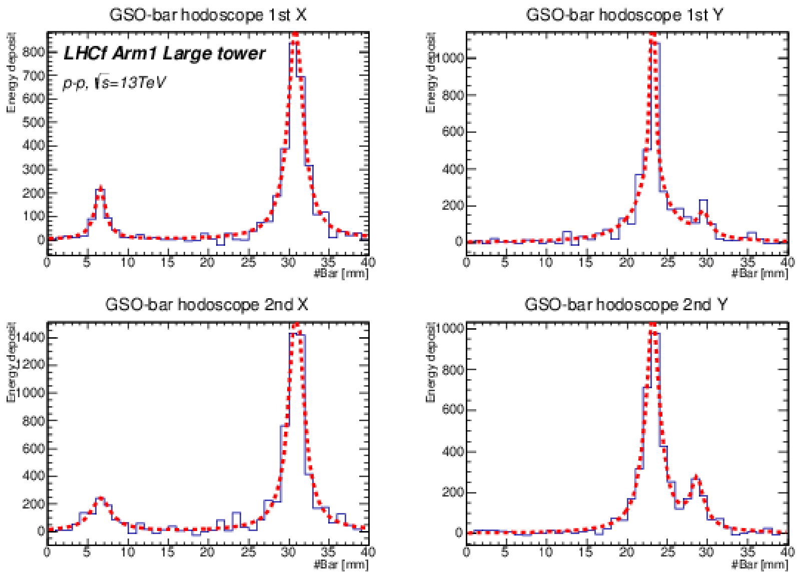
3-components Lorentzian function

$$f(x) = p_0 \left[\frac{p_2}{\frac{(x-p_1)^2}{p_3} + p_3} + \frac{p_4}{\frac{(x-p_1)^2}{p_5} + p_5} + \frac{1 - p_2 - p_4}{\frac{(x-p_1)^2}{p_6} + p_6} \right]$$

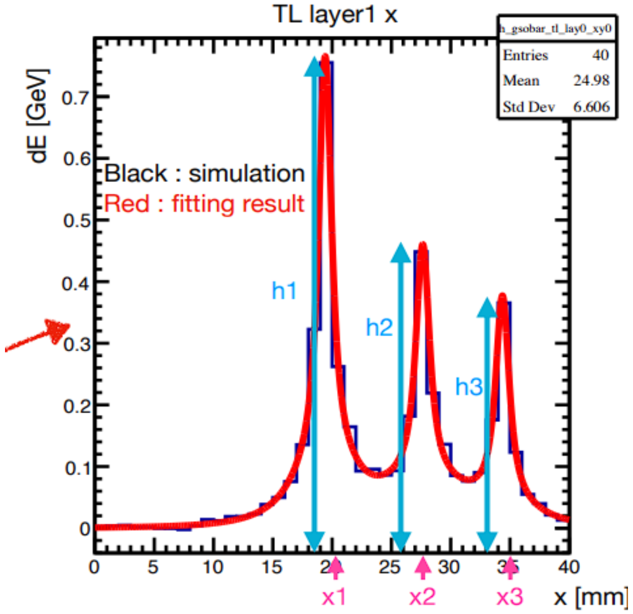
Reconstruction of multiple calorimetric clusters in LHCf

- The total energy is reconstructed by summing the **calibrated energy releases** in the GSO layers. This sum is then converted to the total energy using a function derived from Monte Carlo simulations.
- To reconstruct the position of hitting particles, we use the **transversal profile** of tracking detectors, by finding the peaks using TSpectrum and fitting them with a **3-component Lorentzian function for each peak**.
- The current energy-sharing method uses the **ratio of peak heights** for each particle to share the energy between the events.
- This method was initially developed to determine the energy when **two photons** hit the same detector tower. We are now extending this approach to handle cases involving **three or four photons** and to account for **the presence of neutrons**.

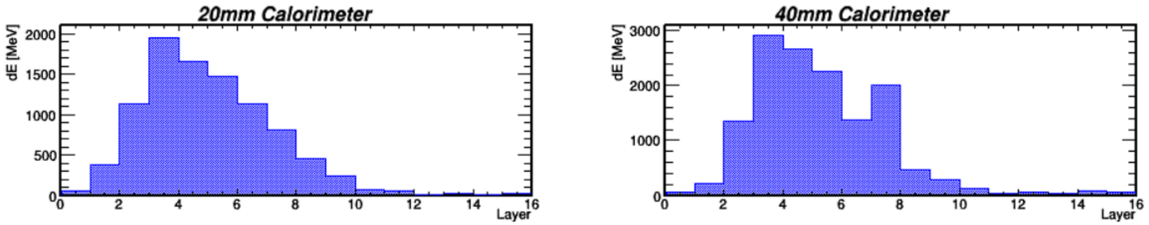
Transversal profile with 2 clusters



Transversal profile with 3 clusters



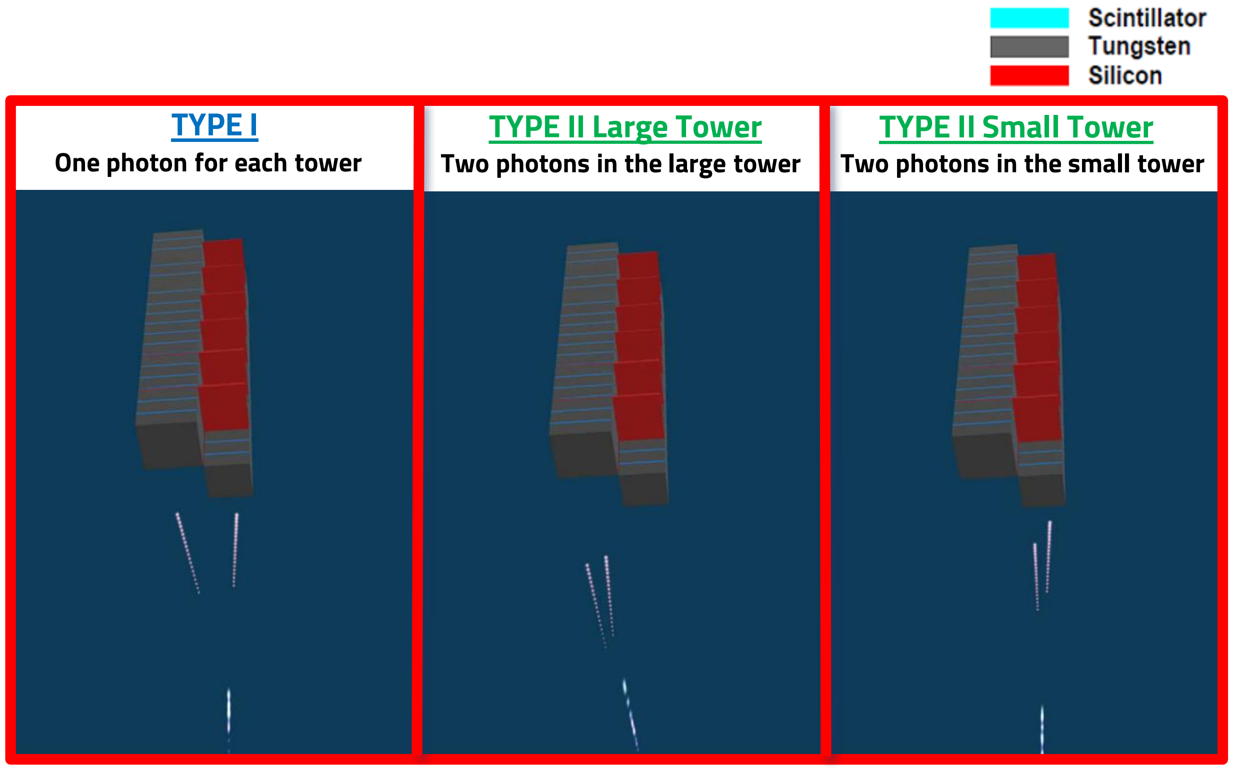
Energy releases



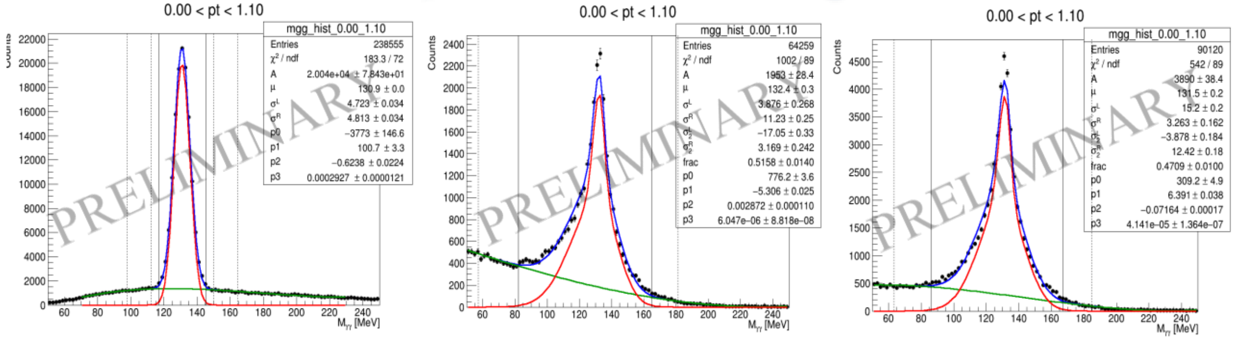
3-components Lorentzian function

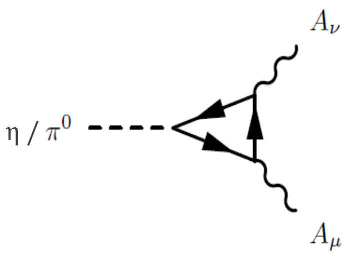
$$f(x) = p_0 \left[\frac{p_2}{\frac{(x-p_1)^2}{p_3} + p_3} + \frac{p_4}{\frac{(x-p_1)^2}{p_5} + p_5} + \frac{1 - p_2 - p_4}{\frac{(x-p_1)^2}{p_6} + p_6} \right]$$

Motivation I: Type II π^0 and η analysis



- The analyses of π^0 and η are two of the **main physics targets** of the LHCf experiment.
- They are identified by their decay into two photons through the analysis of **invariant mass spectra**.
- These two photons can either hit different detector towers (**Type I events**) or both enter the same tower (**Type II events**).
- **Type II events** analysis is less precise due to the complexity of reconstructing **two photons in a single tower.**

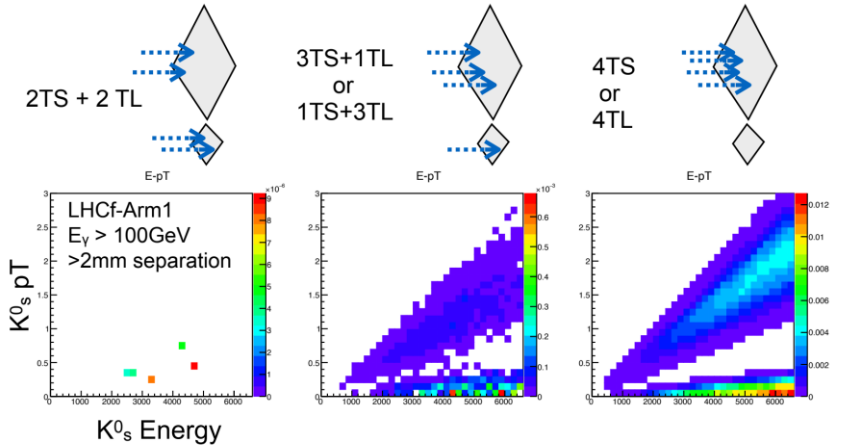


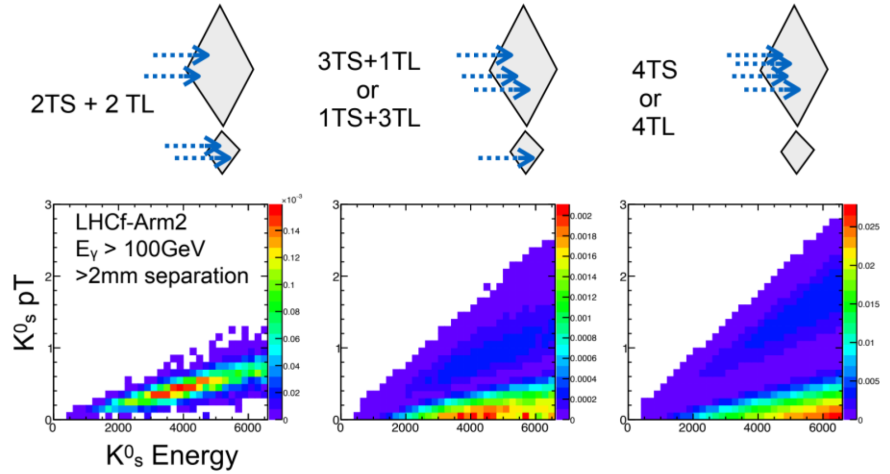


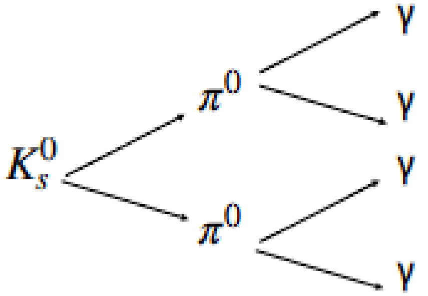
- Both particles decay mainly into **two photons:**
  - $\eta/\pi^0 \rightarrow \gamma\gamma$
- Branching ratio in the case of π^0 is about **98.82%**.
- In the case of η is about **39.36%**.

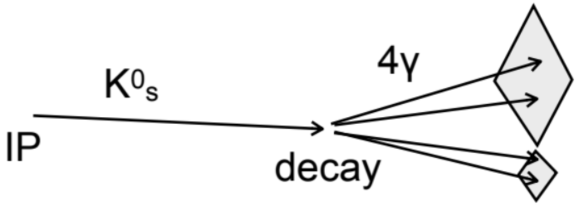
Motivation II: K^0 analysis

- K_S^0 measurements are critical for inferring the production of charged K mesons in EAS, which are significant sources of **TeV-PeV atmospheric neutrinos,** and to understanding **strange quark** forward production.
- Understanding K_S^0 production helps improve **models of hadronic interactions** in cosmic ray showers.
- K_S^0 events often involve **multiple calorimetric hits**, requiring the reconstruction of three or four particles.



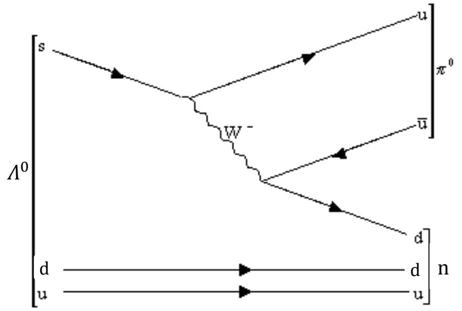


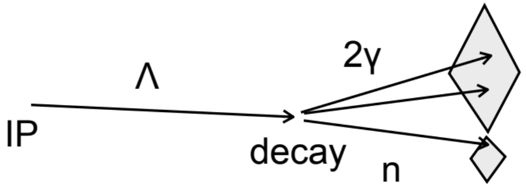


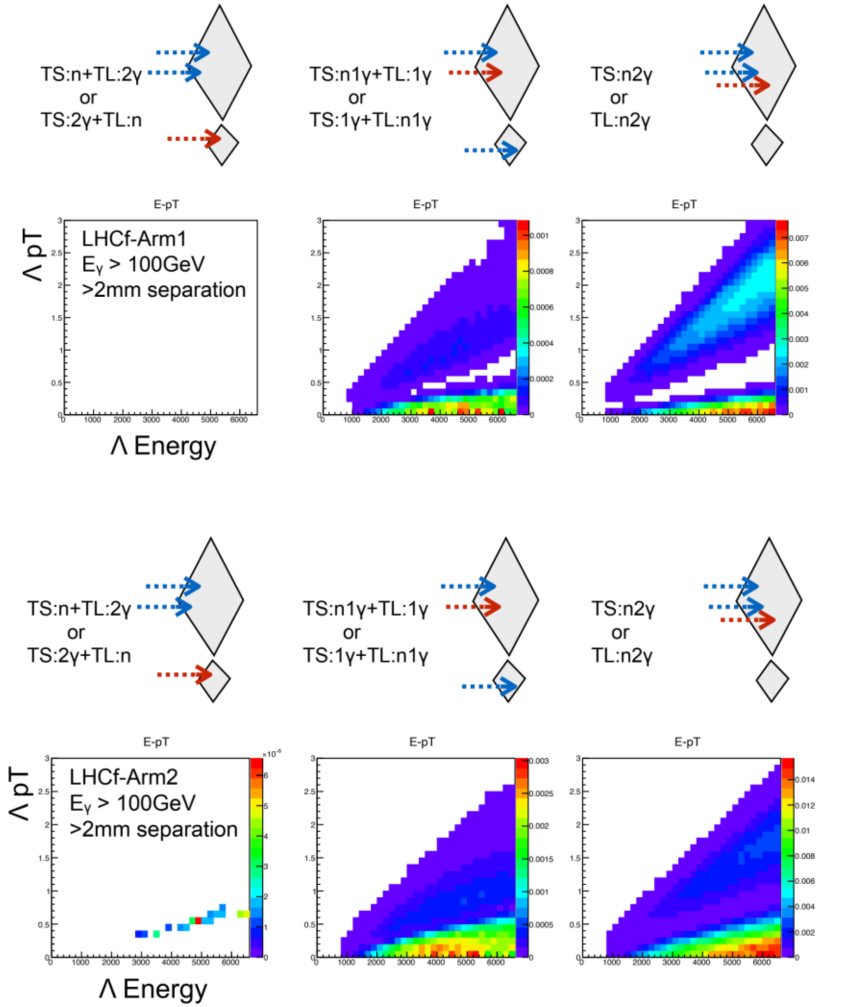


Motivation III: Λ^0 analysis

- Λ^0 baryons are key to understanding **forward strange particle production** and **hadronization** in high-energy collisions.
- Accurate Λ^0 measurements refine models of strange quark behavior and improve QCD process predictions.
- Λ^0 decays produce **complex event topologies**, involving mainly one or two photons and a neutron in a single tower.
- Effective reconstruction requires **distinguishing between different particle types** and precisely determining their energies and positions, which is challenging due to **overlapping signals** and varied interaction characteristics.







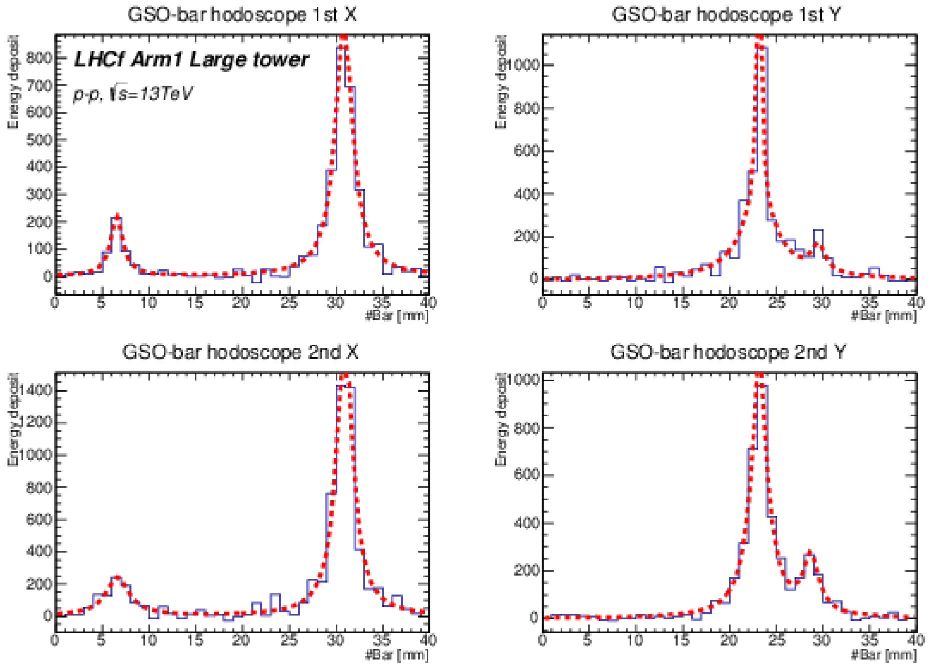
DEVELOPMENT STATUS

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Machine Learning in the LHCf multi-hit reconstruction pipeline

- There are **several steps** of the LHCf multihit-hit reconstruction that can be improved using **ML methods**:
  - **Peak identification** and inference on the **number of hits** (actually performed with TSpectrum).
  - **Position reconstruction** (actually performed by fitting the identified peaks).
  - **Energy sharing** (actually performed using the ratio of peak heights in the layer with the highest energy release).
- **In this first stage we are focusing on the last point, in the case of two photons hitting the same tower of the LHCf-Arm2 detector.**
- This will permit **to improve the analysis** of π^0 and η and to find the best models and methods to analyze events with 3 or 4 photons or with the presence of a neutron for K_s^0 and Λ^0 analysis.

Transversal profile with 2 clusters

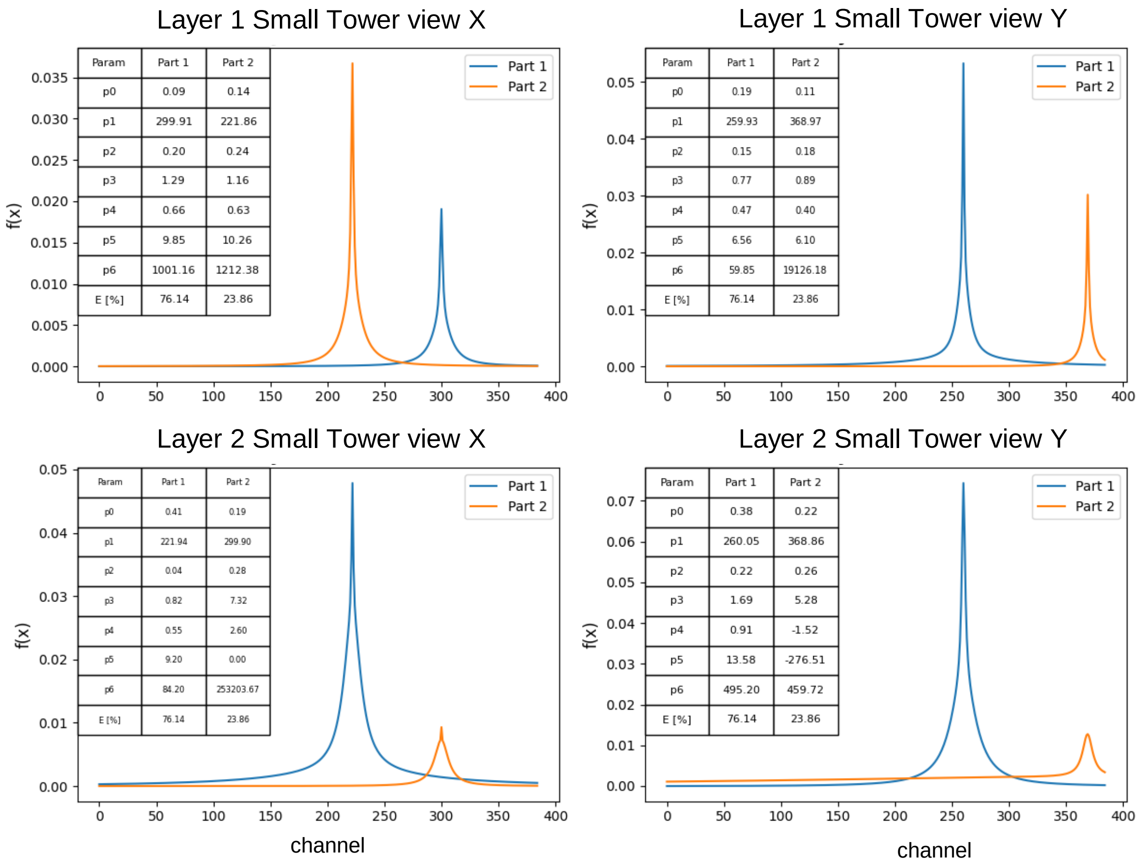


3-components Lorentzian function

$$f(x) = p_0 \left[\frac{p_2}{\frac{(x-p_1)^2}{p_3} + p_3} + \frac{p_4}{\frac{(x-p_1)^2}{p_5} + p_5} + \frac{1 - p_2 - p_4}{\frac{(x-p_1)^2}{p_6} + p_6} \right]$$

Dataset preparation

- To make inferences on energies of two-hit events we used as input for the **the fit results of the silicon transverse profile** in the first 4 layers.
- The dataset was obtained by a **full MC simulation** (collision generation, transport and interaction with the detector) of p-p collisions at $\sqrt{s}$=13 TeV, using as generator the model **QGSJETII-04**.
- Different models were constructed and trained for **each tower**.
- In particular, the 7 fit parameters for **each particle,** for **each view** (x and y) for the first two **silicon plane pairs** were used as input variables (56 input variables).



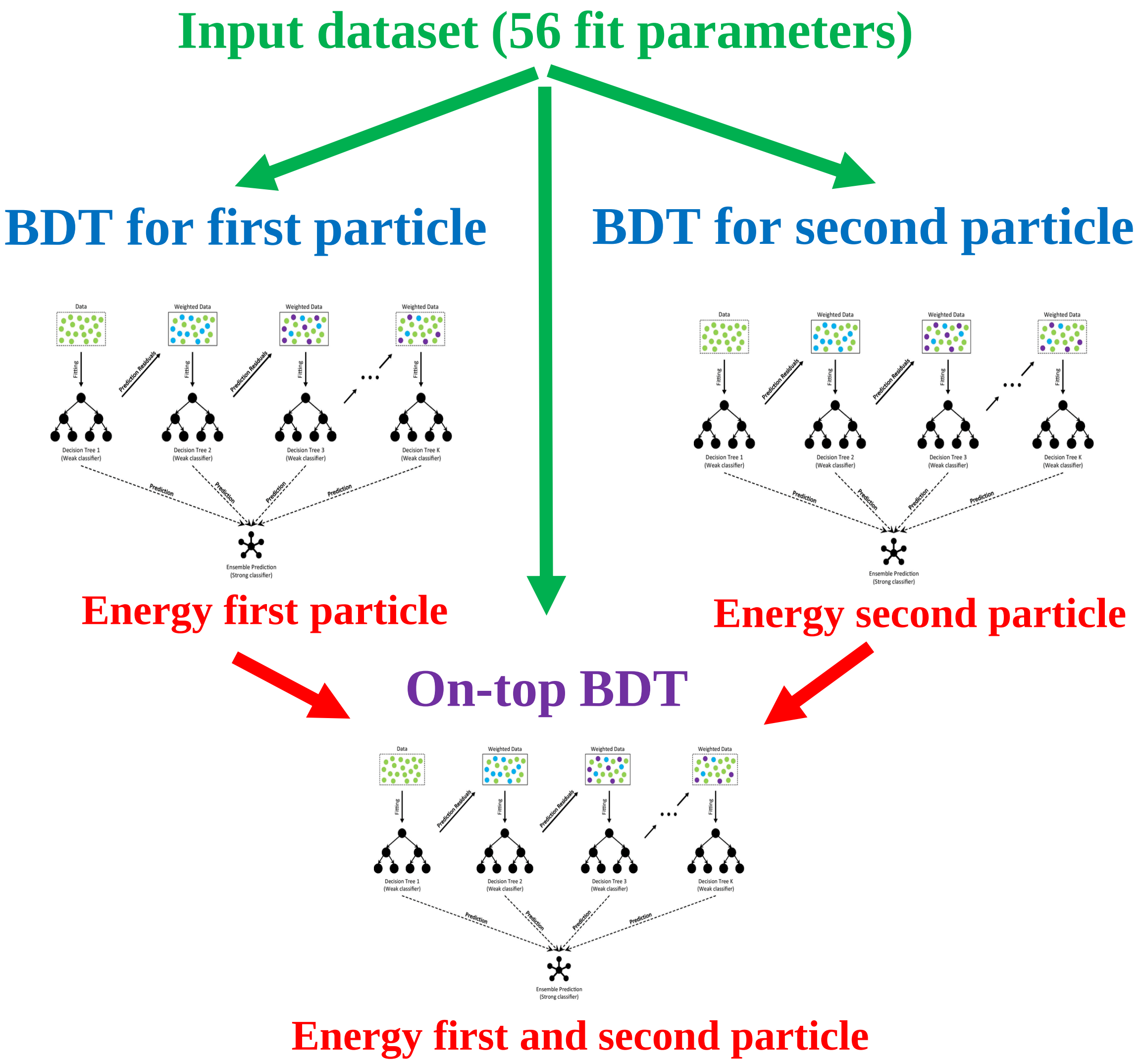
3-components Lorentzian function

$$f(x) = p_0 \left[\frac{p_2}{\frac{(x-p_1)^2}{p_3} + p_3} + \frac{p_4}{\frac{(x-p_1)^2}{p_5} + p_5} + \frac{1-p_2-p_4}{\frac{(x-p_1)^2}{p_6} + p_6} \right]$$

Tested models

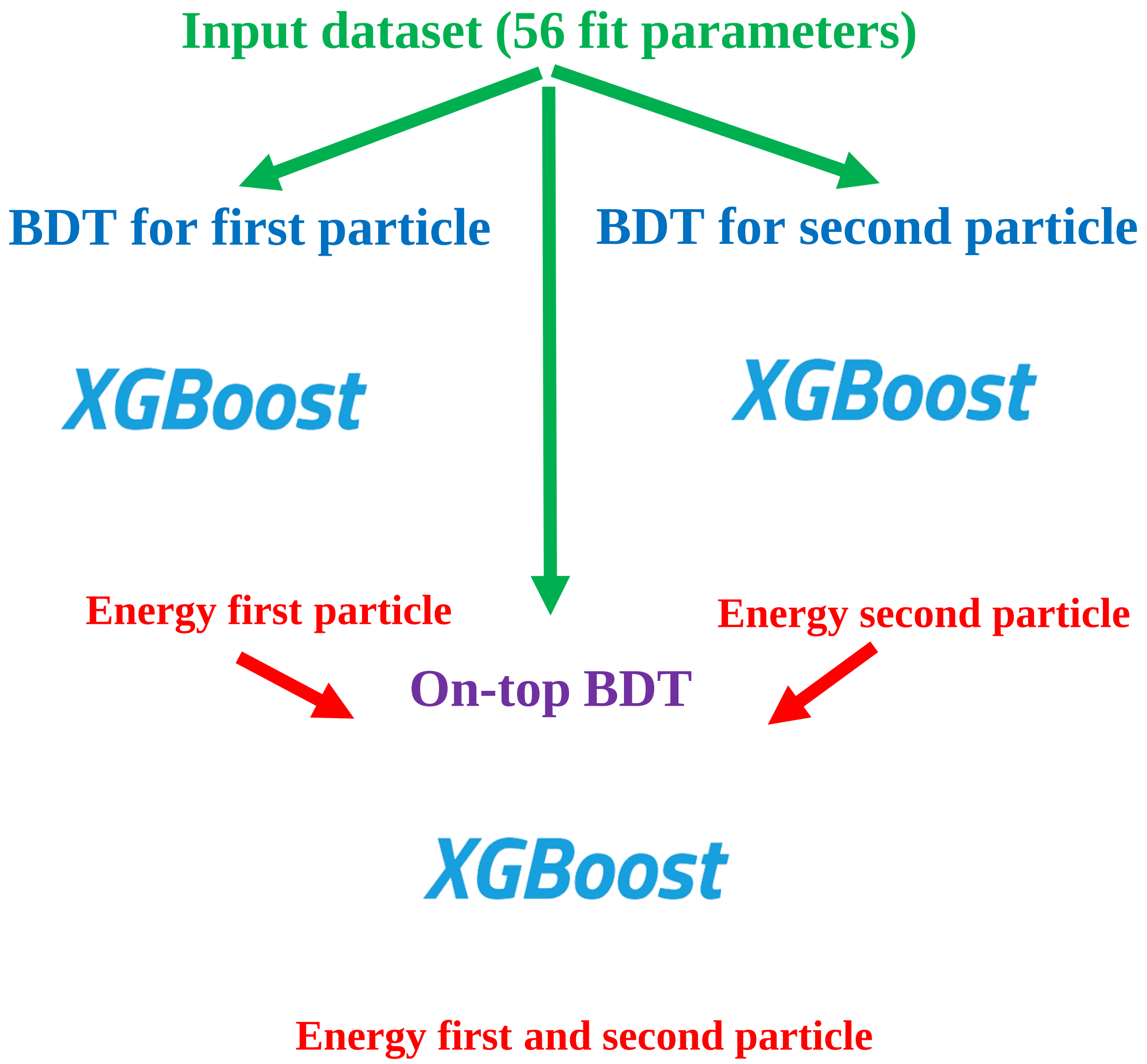
- We tested two similar approaches, both based on **gradient-boosting decision trees (BDTs)**.

- Two **ensemble models** were constructed, consisting of **two BDTs on a first level** to make inferences on the single energy of each of the two photons, and an **on-top BDT** which, based on the predictions of the previous two combined with the input dataset, would make inference on the energy of both particles.



Tested models

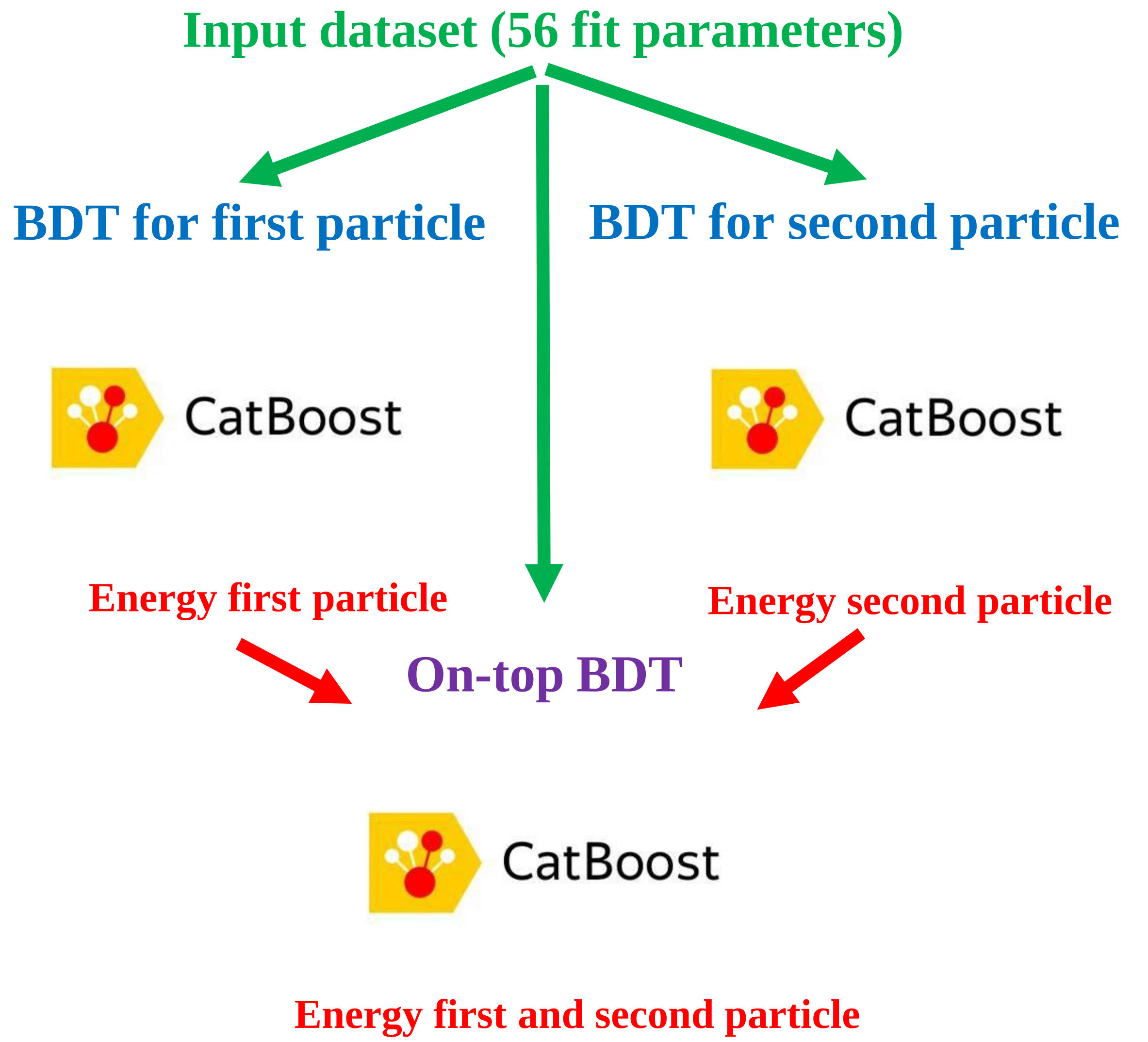
- We tested two similar approaches, both based on **gradient-boosting decision trees (BDTs)**.
- Two **ensemble models** were constructed, consisting of **two BDTs on a first level** to make inferences on the single energy of each of the two photons, and an **on-top BDT** which, based on the predictions of the previous two combined with the input dataset, would make inference on the energy of both particles.
- In particular, the two libraries used for the test are **XGBoost** and **CatBoost**, which were initialized with similar values of the hyperparameters (e.g. **300k weak learners** and **depth of each tree equal to three**).
- The **RMSE** was used as a metric for the evaluation.



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Tested models

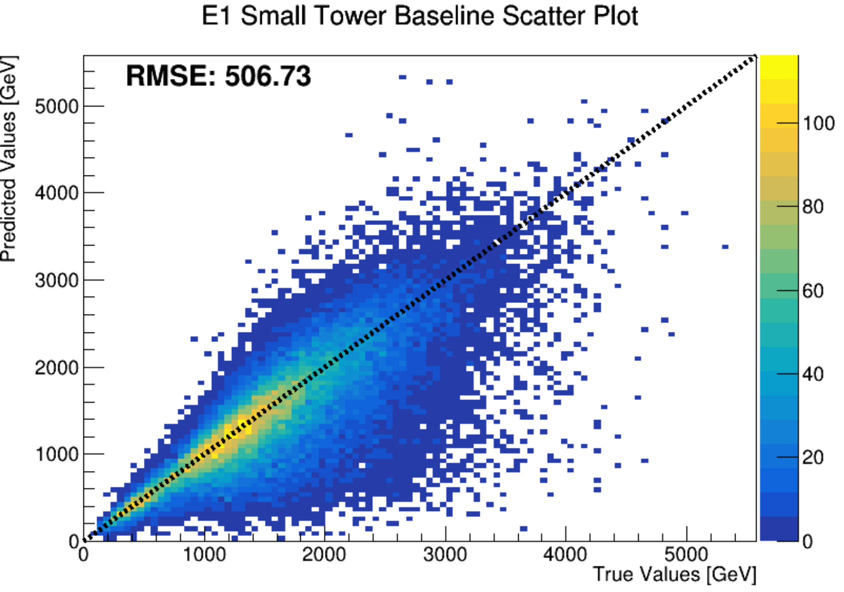
- We tested two similar approaches, both based on **gradient-boosting decision trees (BDTs)**.
- Two **ensemble models** were constructed, consisting of **two BDTs on a first level** to make inferences on the single energy of each of the two photons, and an **on-top BDT** which, based on the predictions of the previous two combined with the input dataset, would make inference on the energy of both particles.
- In particular, the two libraries used for the test are **XGBoost** and **CatBoost**, which were initialized with similar values of the hyperparameters (e.g. **300k weak learners** and **depth of each tree equal to three**).
- The **RMSE** was used as a metric for the evaluation.

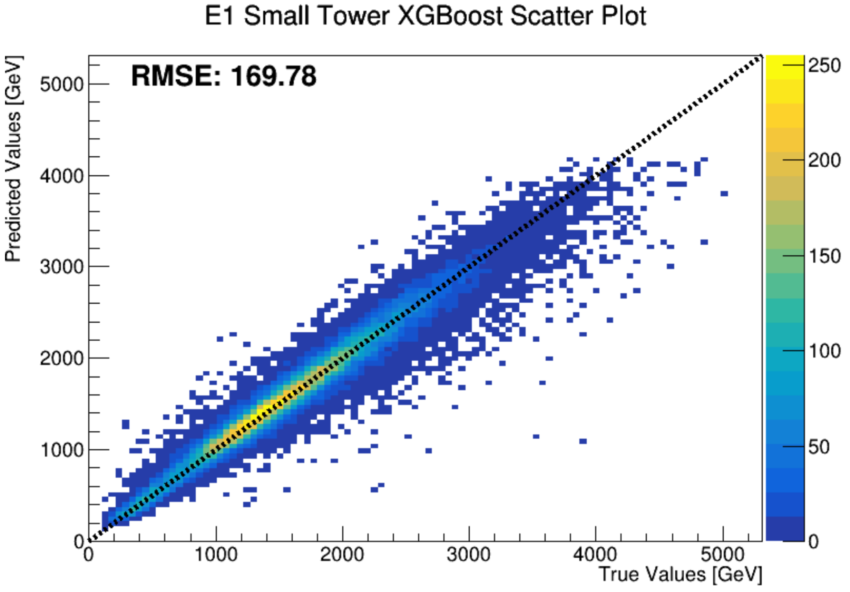


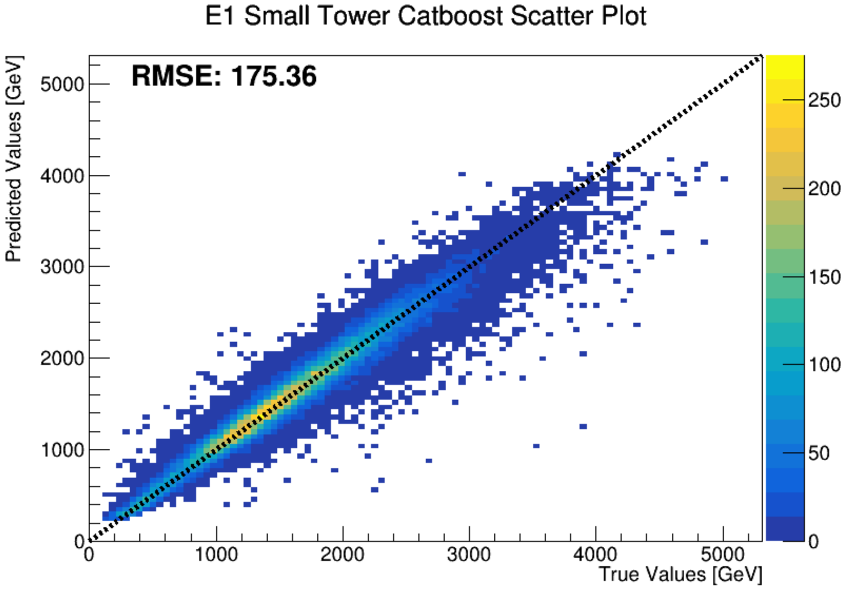
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Results of first particle energy in Small Tower

Train events = 93k
Test events = 40k

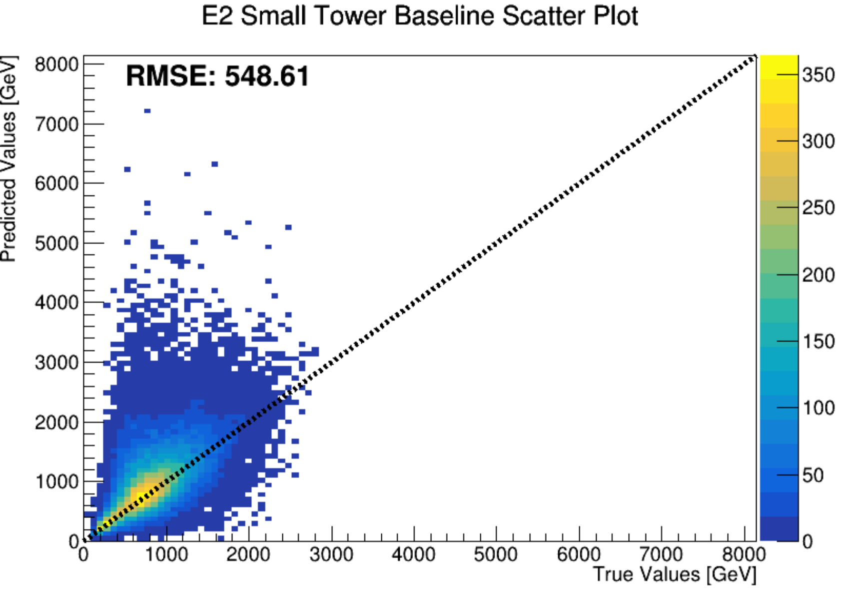


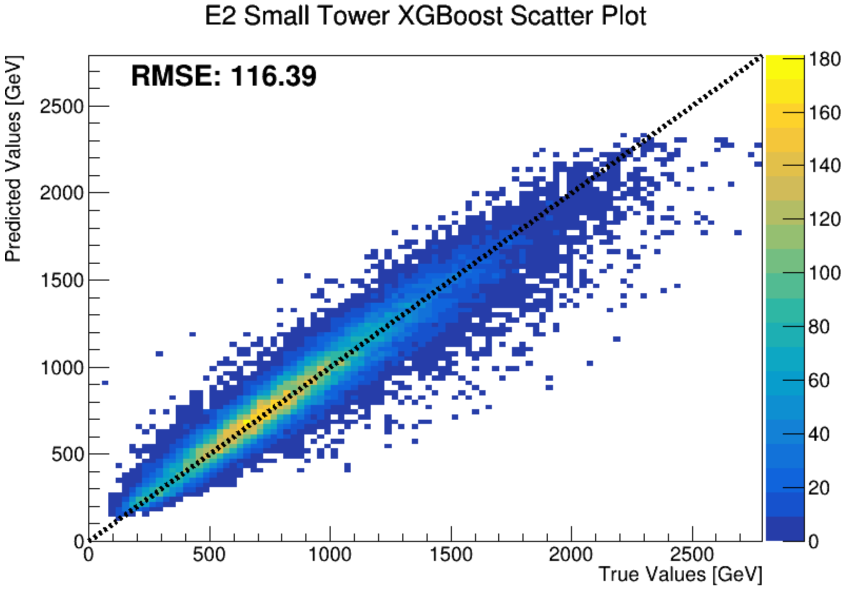


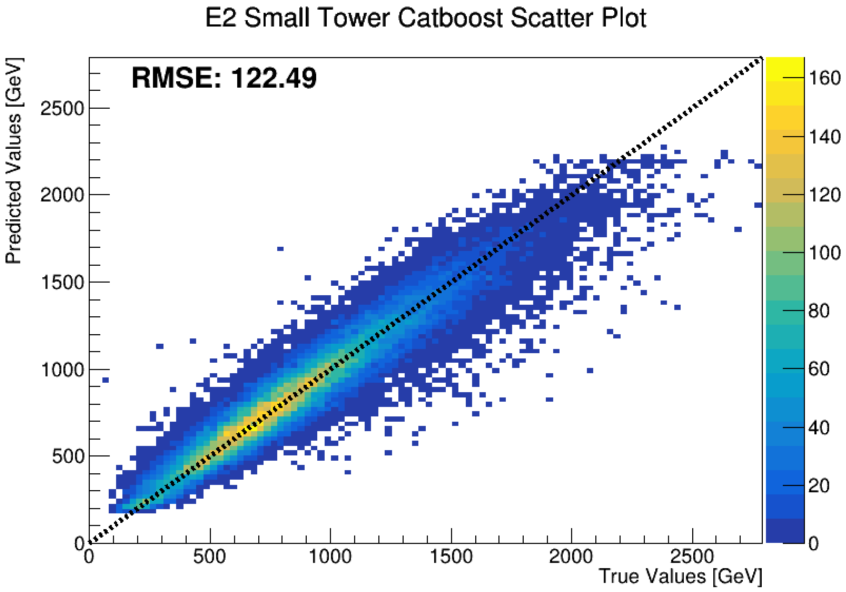


Results of second particle energy in Small Tower

Train events = 93k
Test events = 40k



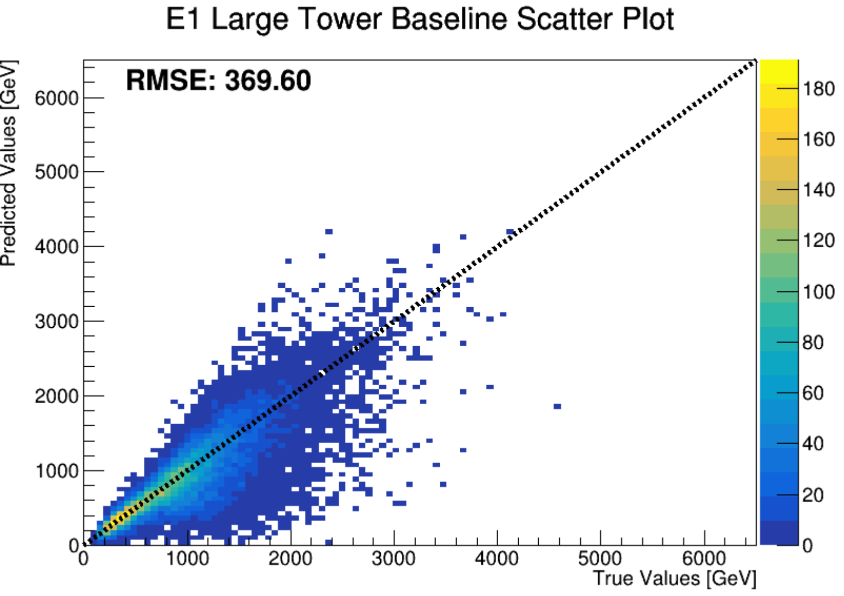


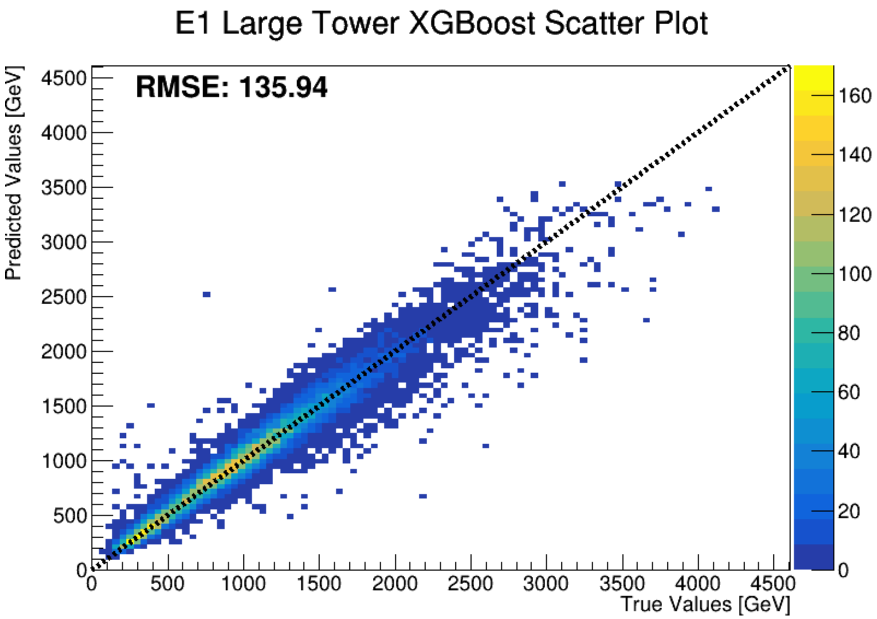


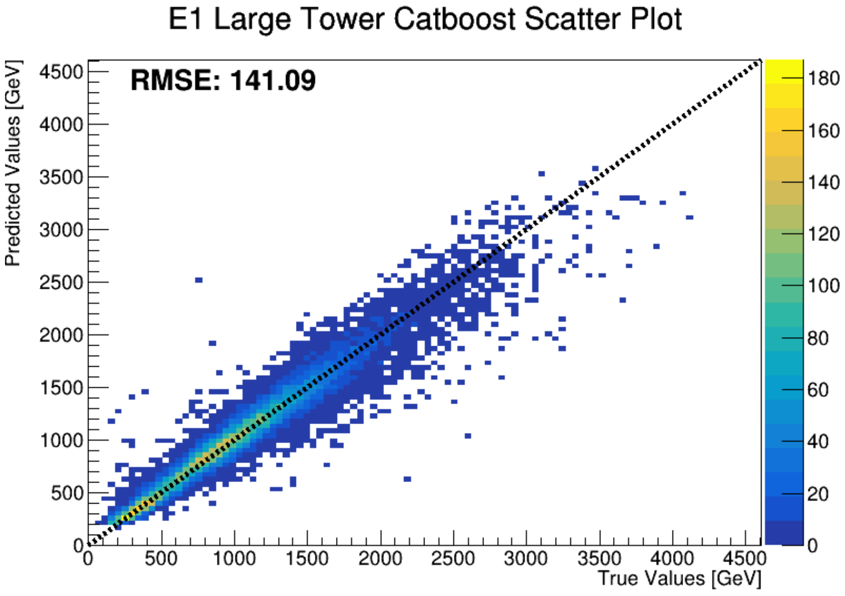
Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Results of first particle energy in Large Tower

Train events = 42k
Test events = 18k



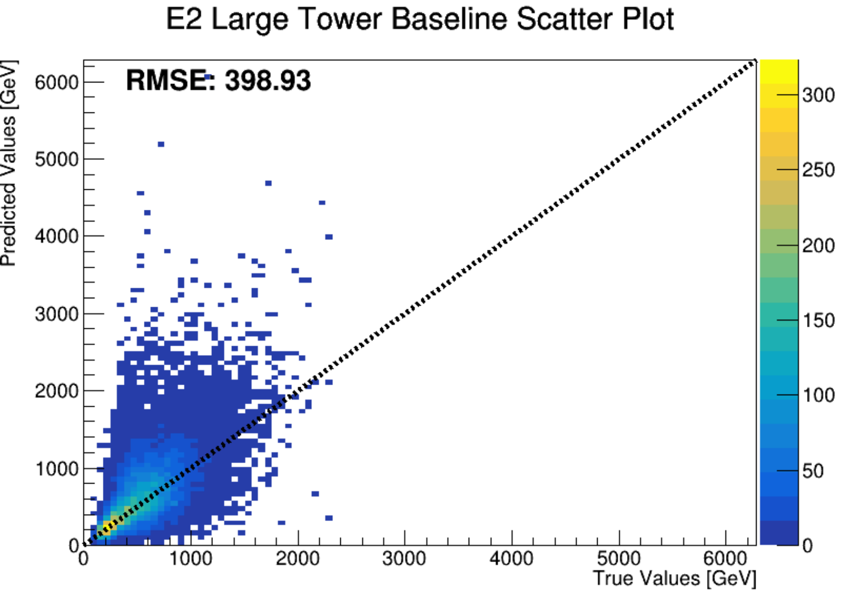


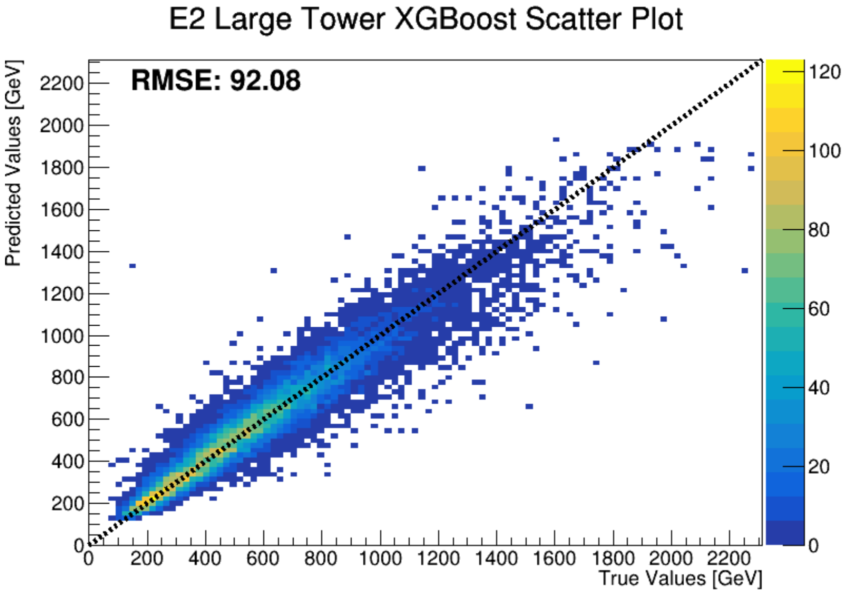


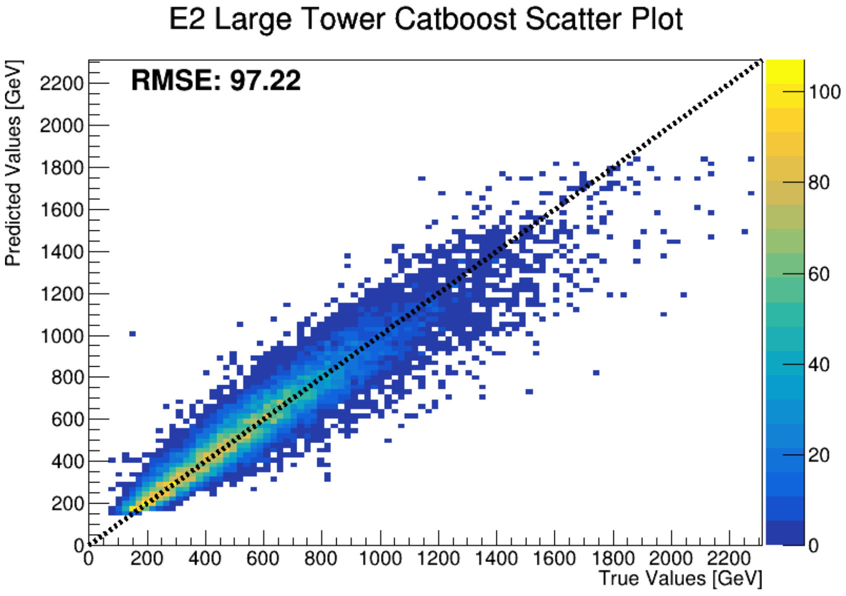
Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA | ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Results of second particle energy in Large Tower

Train events = 42k
Test events = 18k



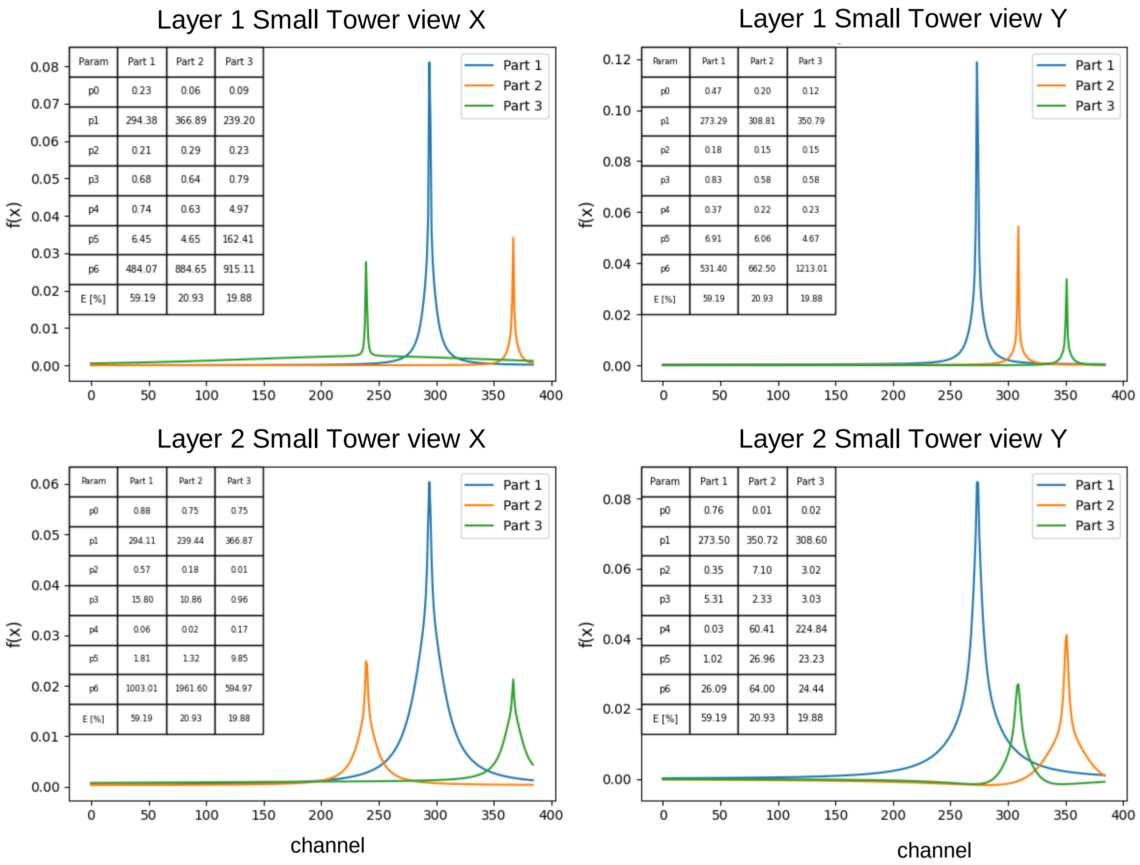




FUTURE PROSPECTS

Prospects: extend to more than two hits

- Current methods are primarily designed for events with up to **two particles** hitting the same detector tower.
- Extending these techniques to handle more complex scenarios with **three or four hits** is crucial for comprehensive multi-hit event reconstruction.
- Increased number of hits leads to significant signal overlap, making accurate deconvolution more challenging.
- We performed preliminary **tests** using a 3-hits dataset, but with low statistics (O(1k) events).



Layer 1 Small Tower view X

Param	Part 1	Part 2	Part 3
p0	0.23	0.06	0.09
p1	294.38	366.89	239.20
p2	0.21	0.29	0.23
p3	0.68	0.64	0.79
p4	0.74	0.63	4.97
p5	6.45	4.65	162.41
p6	484.07	884.65	915.11
E [%]	59.19	20.93	19.88

Layer 1 Small Tower view Y

Param	Part 1	Part 2	Part 3
p0	0.47	0.20	0.12
p1	273.29	308.81	350.79
p2	0.18	0.15	0.15
p3	0.83	0.58	0.58
p4	0.37	0.22	0.23
p5	6.91	6.06	4.67
p6	531.40	662.50	1213.01
E [%]	59.19	20.93	19.88

Layer 2 Small Tower view X

Param	Part 1	Part 2	Part 3
p0	0.88	0.75	0.75
p1	294.11	239.44	366.87
p2	0.57	0.18	0.01
p3	15.80	10.86	0.96
p4	0.06	0.02	0.17
p5	1.81	1.32	9.85
p6	1003.01	1961.60	594.97
E [%]	59.19	20.93	19.88

Layer 2 Small Tower view Y

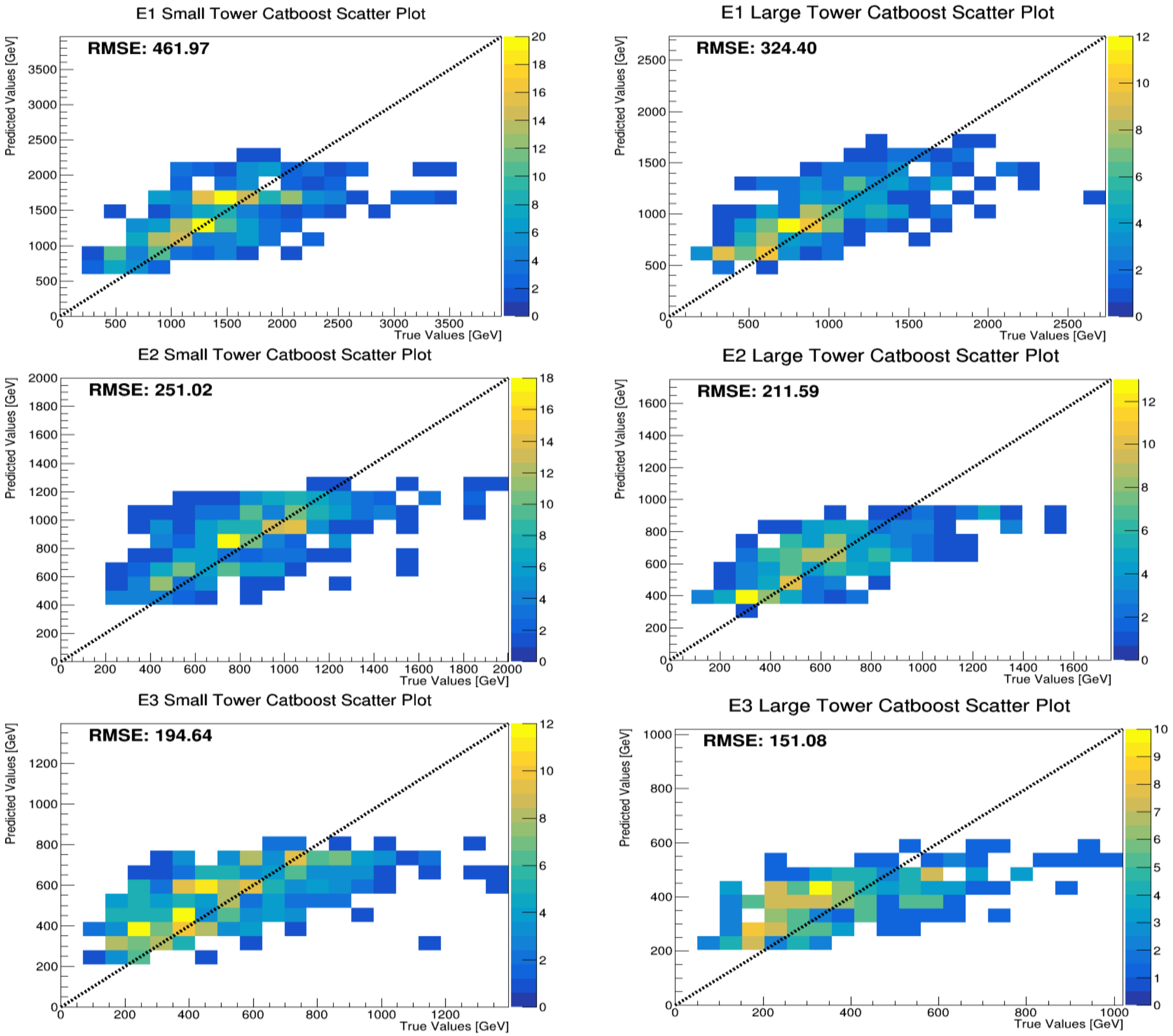
Param	Part 1	Part 2	Part 3
p0	0.76	0.01	0.02
p1	273.50	350.72	308.60
p2	0.35	7.10	3.02
p3	5.31	2.33	3.03
p4	0.03	60.41	224.84
p5	1.02	26.96	23.23
p6	26.09	64.00	24.44
E [%]	59.19	20.93	19.88

3-components Lorentzian function

$$f(x) = p_0 \left[\frac{p_2}{\frac{(x-p_1)^2}{p_3} + p_3} + \frac{p_4}{\frac{(x-p_1)^2}{p_5} + p_5} + \frac{1-p_2-p_4}{\frac{(x-p_1)^2}{p_6} + p_6} \right]$$

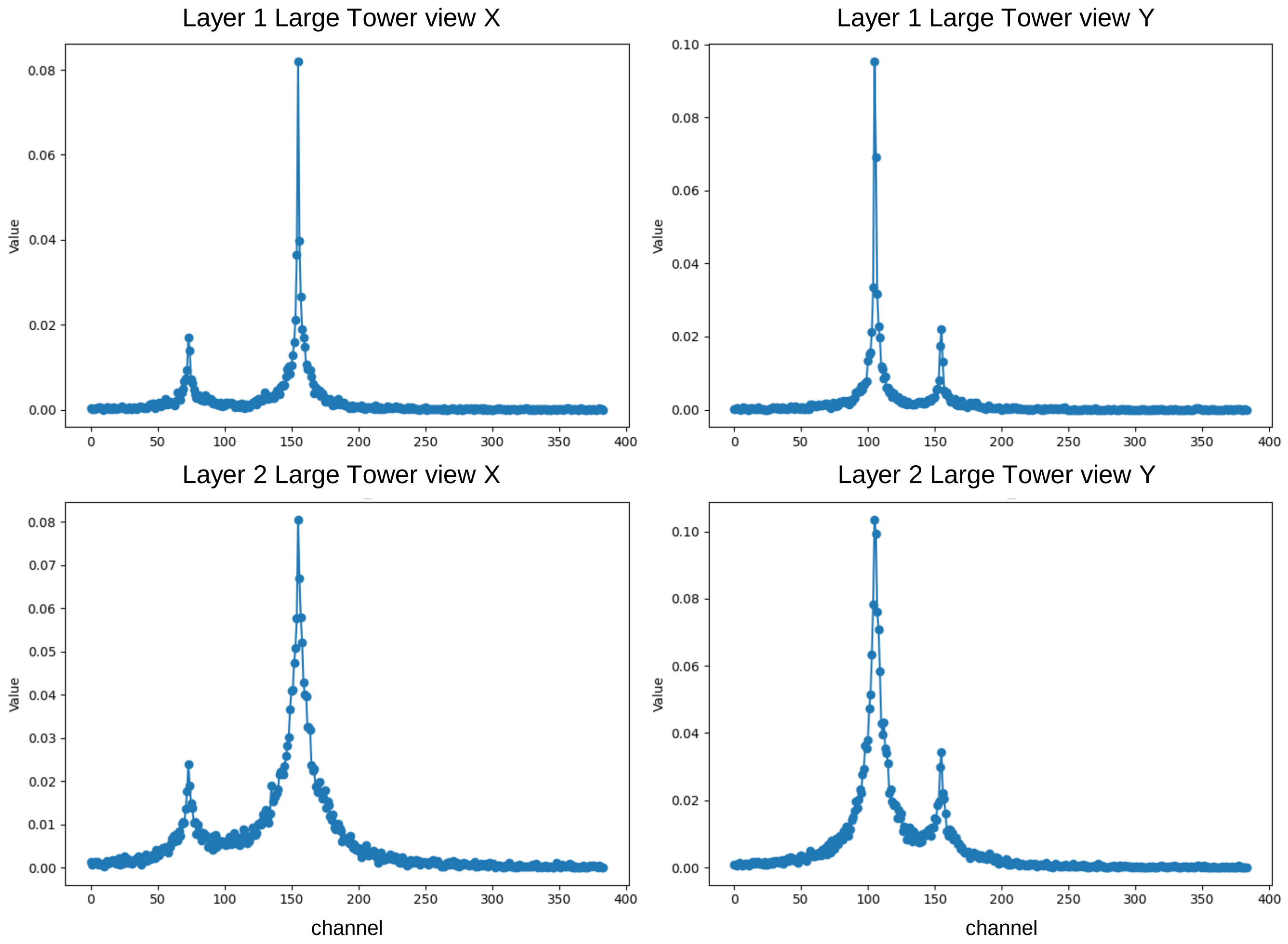
Prospects: extend to more than two hits

- Current methods are primarily designed for events with up to **two particles** hitting the same detector tower.
- Extending these techniques to handle more complex scenarios with **three or four hits** is crucial for comprehensive multi-hit event reconstruction.
- Increased number of hits leads to significant signal overlap, making accurate deconvolution more challenging.
- We performed preliminary **tests** using a 3-hits dataset, but with low statistics (O(1k) events).
- First results are encouraging, but more statistics is needed.



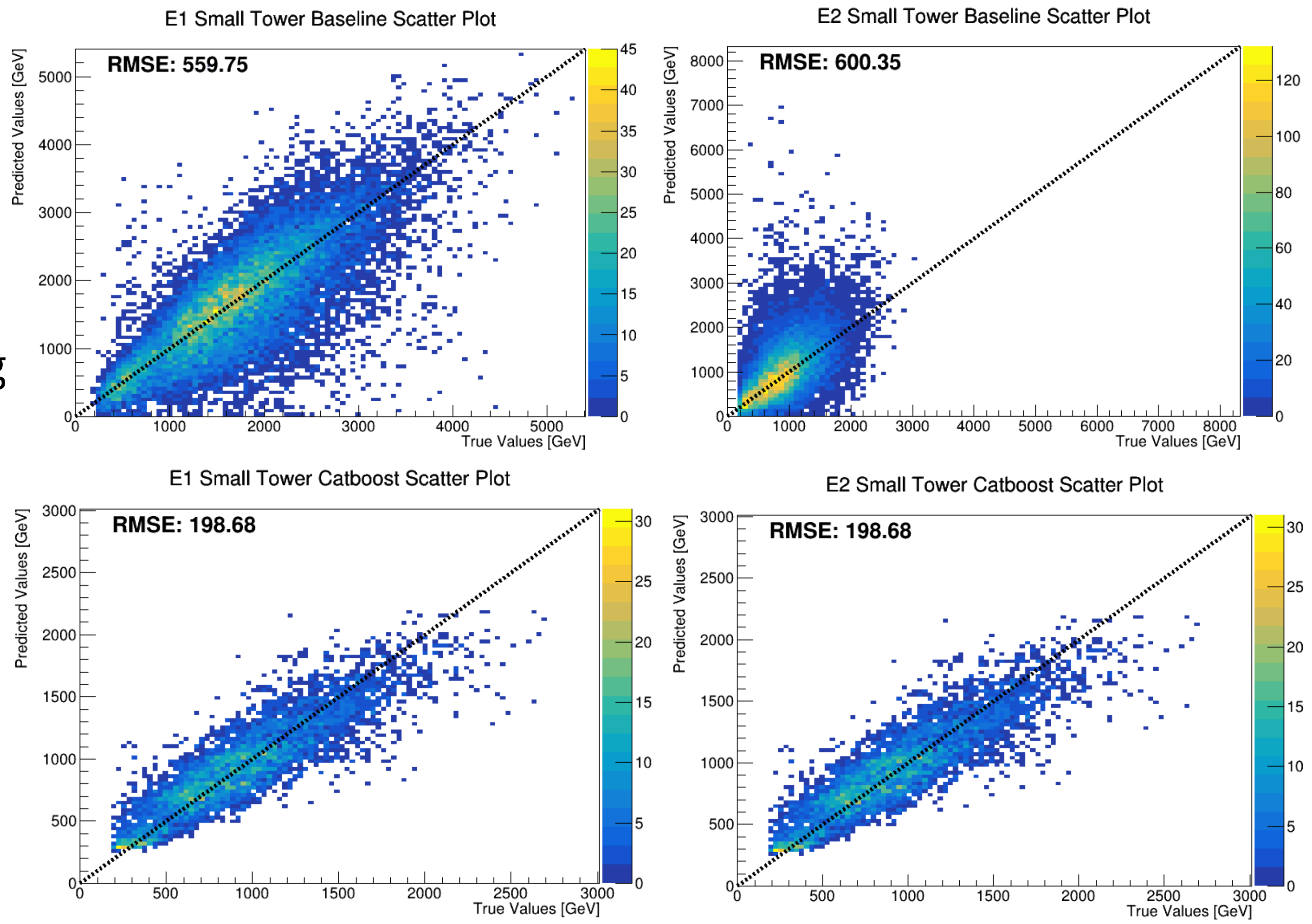
Prospects: use of the raw transverse distribution of position detectors

- We are exploring the use of **raw transverse distribution data** from position detectors, moving away from relying solely on fit parameters.
- This approach leverages the raw signals directly from the detectors to potentially enhance the **accuracy of particle reconstruction**.
- Test perfomed on a different dataset, created using the same methodology but with a new simulation of p-p collisions at $\sqrt{s}$=13 TeV (QGSJET II-04 as generator).



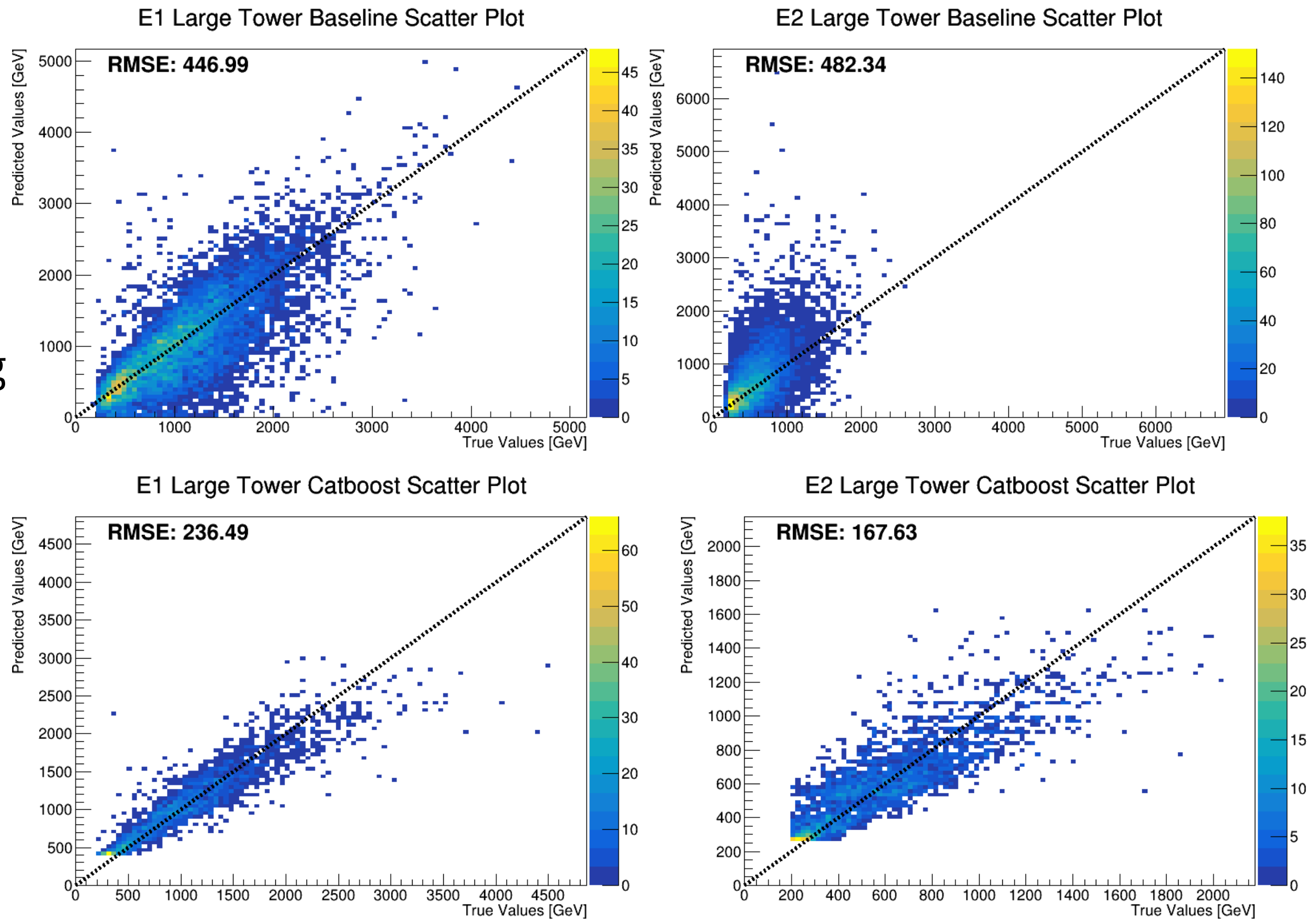
Prospects: use of the raw transverse distribution of position detectors

- We are exploring the use of **raw transverse distribution data** from position detectors, moving away from relying solely on fit parameters.
- This approach leverages the raw signals directly from the detectors to potentially enhance the **accuracy of particle reconstruction**.
- Test perfomed on a different dataset, created using the same methodology but with a new simulation of p-p collisions at $\sqrt{s}$=13 TeV (QGSJET II-04 as generator).
- **Encouraging Outcomes**: Initial results from are positive, showing improved accuracy in particle reconstruction.
- Only O(20k) events, improvements are expected by increasing the statistics.



Prospects: use of the raw transverse distribution of position detectors

- We are exploring the use of **raw transverse distribution data** from position detectors, moving away from relying solely on fit parameters.
- This approach leverages the raw signals directly from the detectors to potentially enhance the **accuracy of particle reconstruction**.
- Test perfomed on a different dataset, created using the same methodology but with a new simulation of p-p collisions at $\sqrt{s}$=13 TeV (QGSJET II-04 as generator).
- **Encouraging Outcomes**: Initial results from are positive, showing improved accuracy in particle reconstruction.
- Only O(20k) events, improvements are expected by increasing the statistics.



Summary

- The LHCf experiment measures neutral particle production in the forward pseudorapidity region to improve hadronic interaction models used in the cosmic rays field.
- Machine learning can **enhance the reconstruction of calorimetric clusters**, particularly for complex events involving multiple particles.
- Focus areas include energy sharing and position reconstruction, improving π^0 and η event analyses and opening new physics channels such as K_s^0 and Λ^0.
- **Two-Hit Events:** Improved accuracy in energy reconstruction using ensemble methods based on CatBoost and XGBoost.
- **Three-Hit Events:** Very preliminary results indicate potential for handling more complex scenarios, though additional data and refinement are needed.
- **Four hit studies, the possible presence of a neutron and improved position reconstruction using ML are foreseen, probably using raw transverse distribution data.**

Thanks for the attention!

This work is partially supported by ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union – NextGenerationEU

Giuseppe Piparo[1,2], on behalf of the LHCf collaboration

1. Istituto Nazionale di Fisica Nucleare (INFN), Catania section
2. University of Catania, Department of Physics and Astronomy

42nd International Conference on High Energy Physics (ICHEP 2024), Prague (CZE), Jul 17-24, 2024