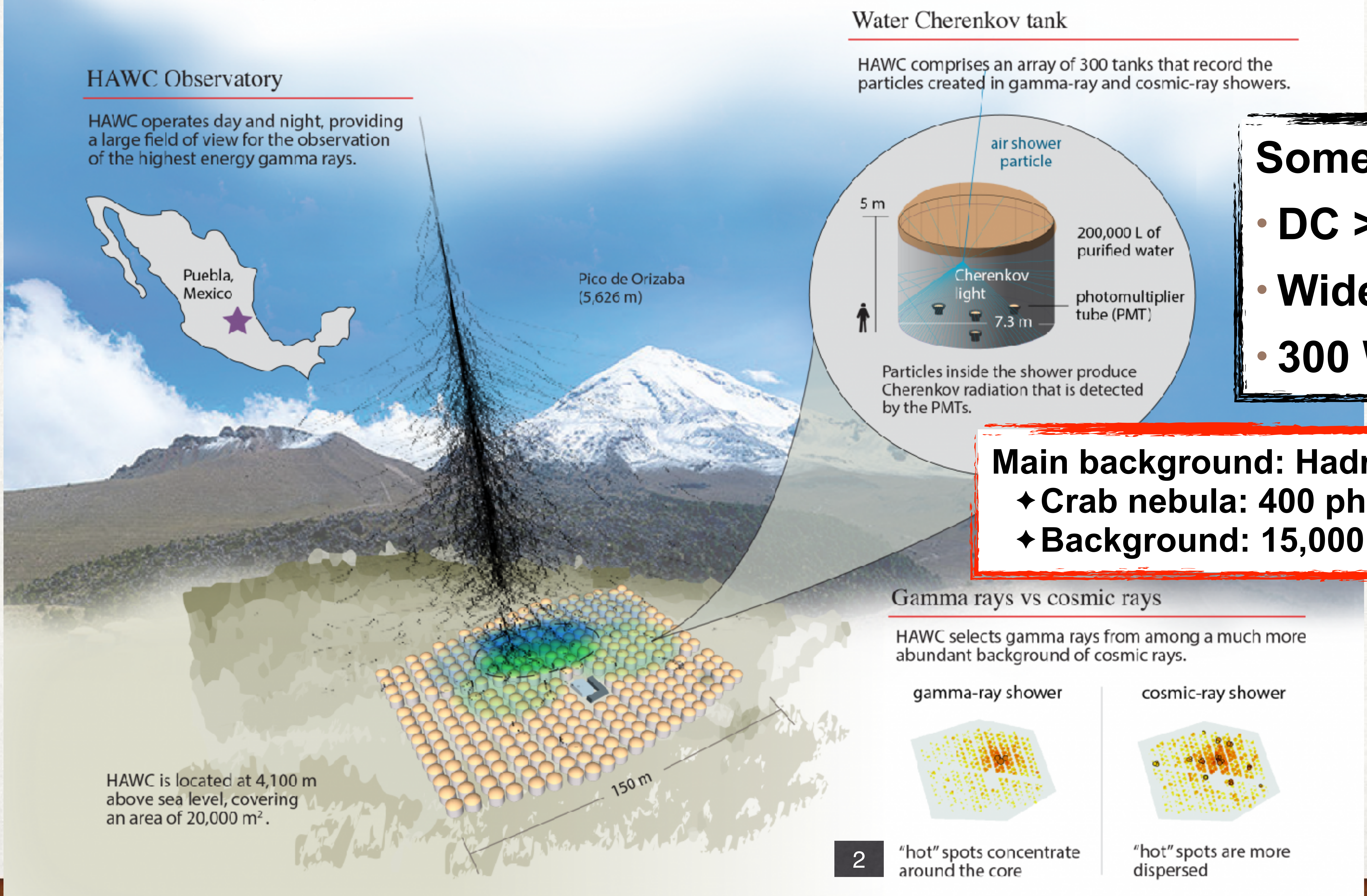
Techniques in HAWC Observatory for classification of the showers detected

10th International Workshop on Astronomy and Relativistic Astrophysics
September 8th, 2022

T. Capistrán (IA-UNAM), K. L. Fan (UMD), J.T. Linnemann (MSU), I. Torres (INAOE),
P. Saz (HKU) and for HAWC Collaboration

High Altitude Water Cherenkov (HAWC)



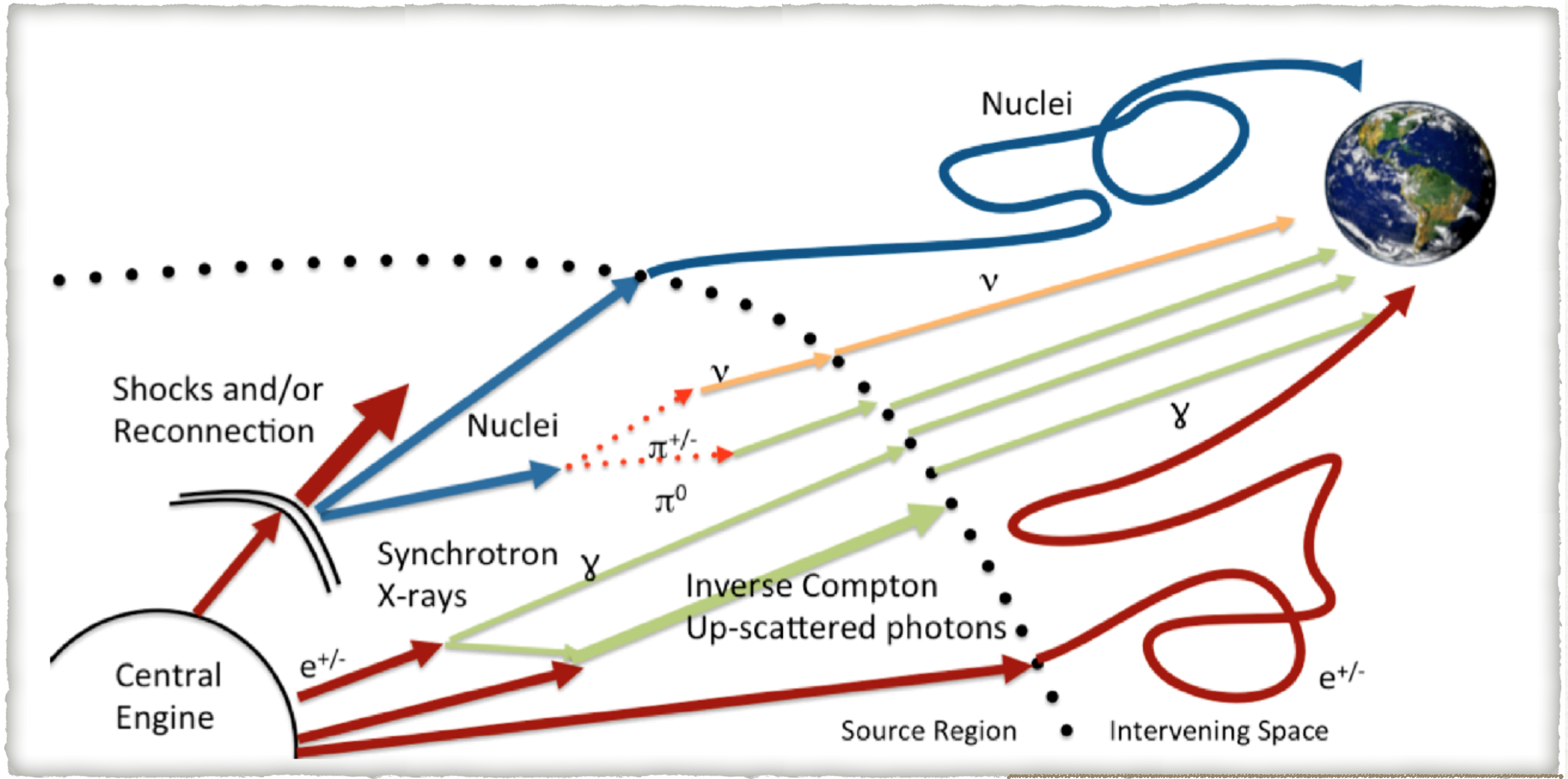
Some characteristics:

- **DC > 95%**
- **Wide F.O.V**
- **300 WCD ->1200PMT**

Main background: Hadronic cosmic ray

- ✦ **Crab nebula: 400 photons/day**
- ✦ **Background: 15,000 cosmic ray/second**

Crucial issue in HAWC observatory



Pretz, J. (2016), https://doi.org/10.22323/1.236.0025

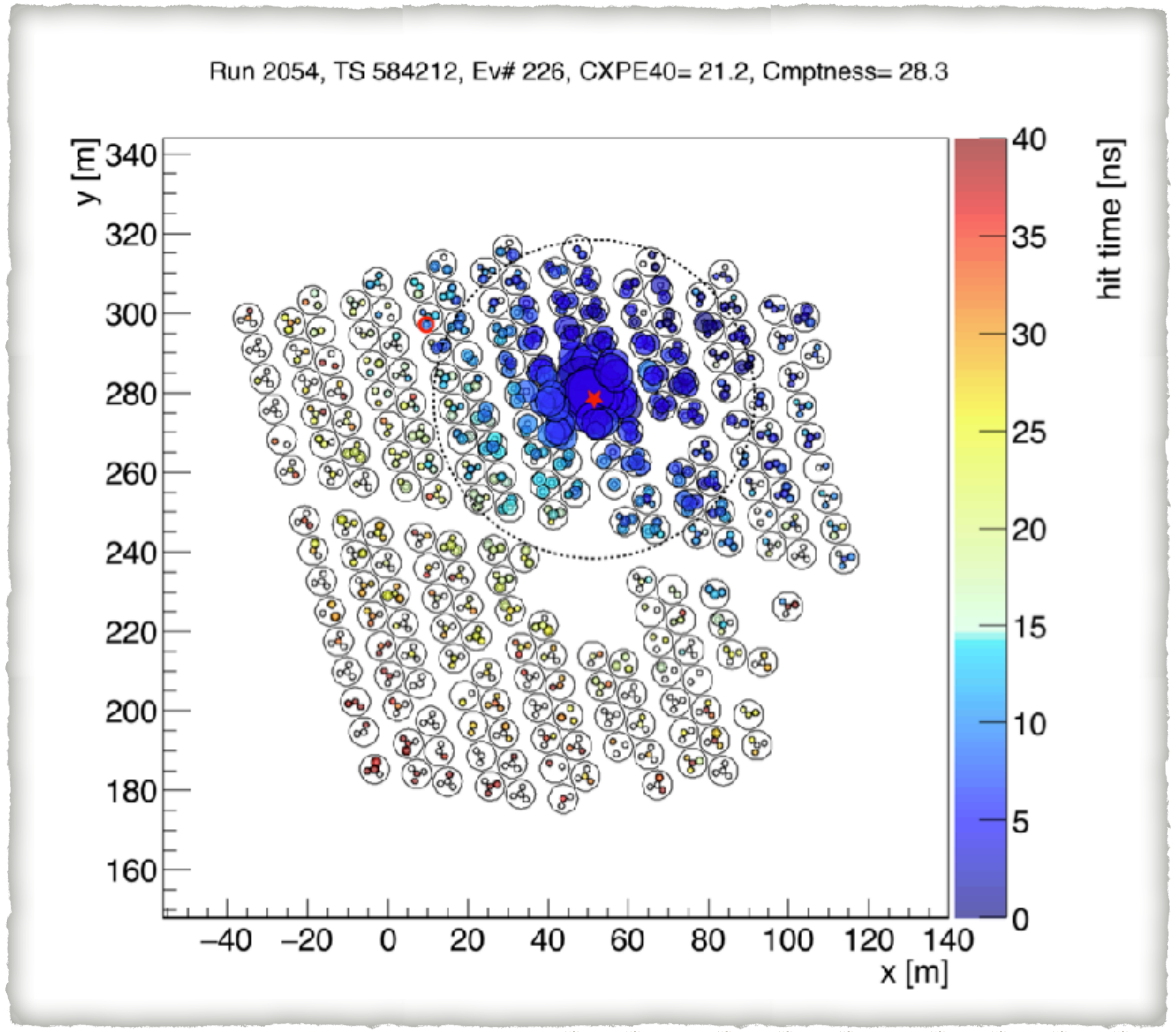
CRs are deflected by magnetic field

They are ~99.9% of the particles that arrive at the Earth

How to distinguish them?

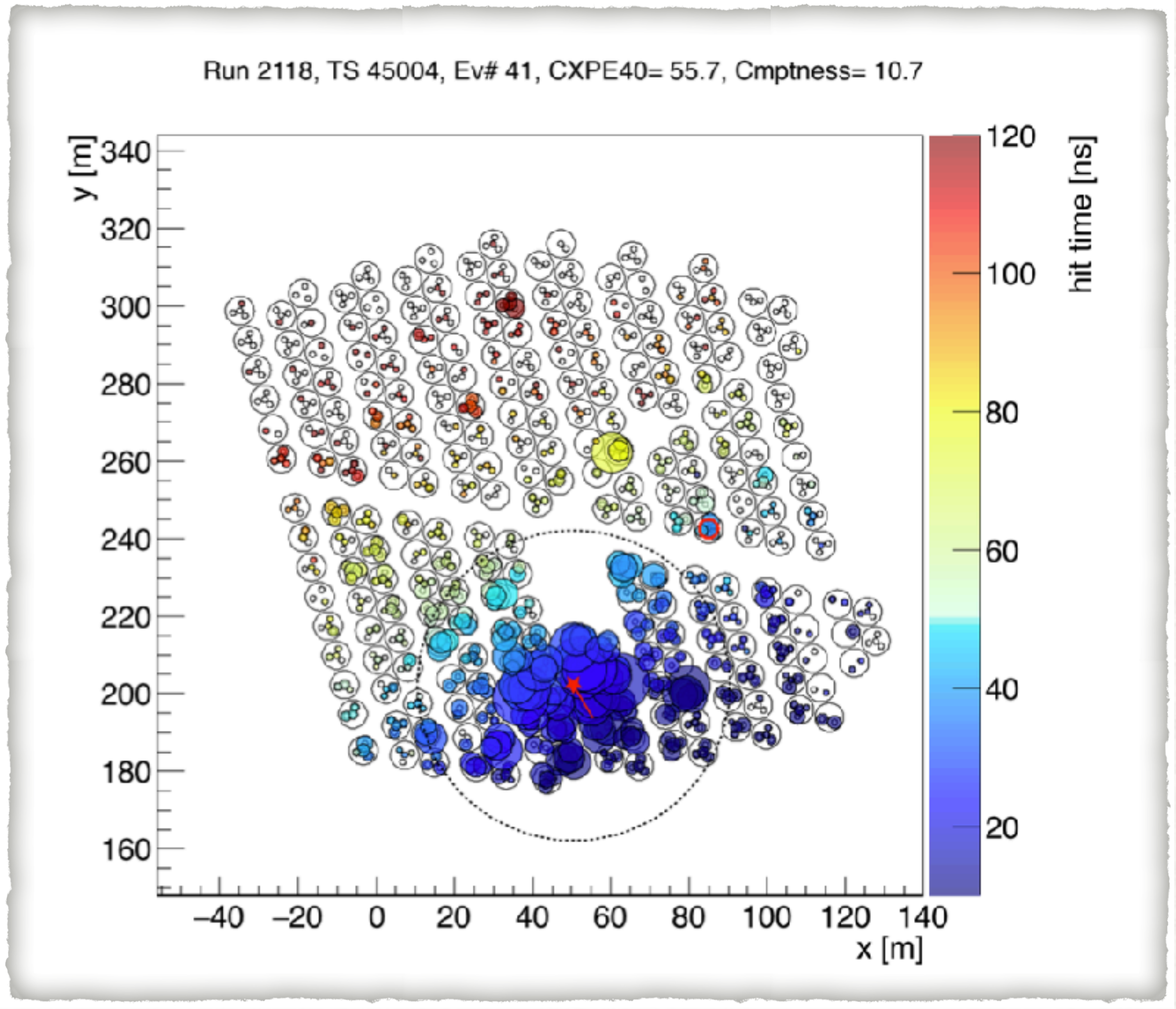
Looking at their differences

Compact



Pretz, J. (2016), https://doi.org/10.22323/1.236.0025

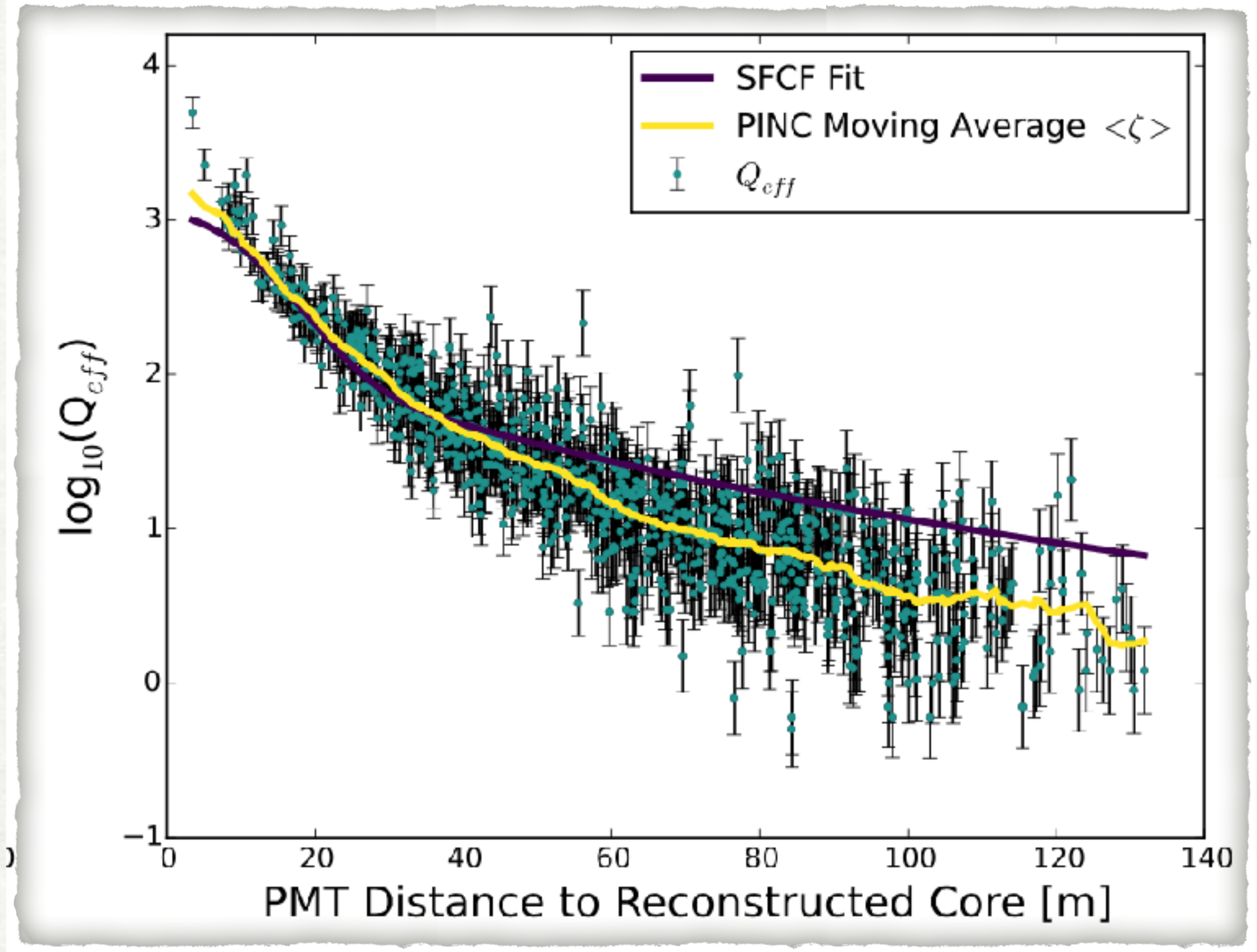
Diffuse



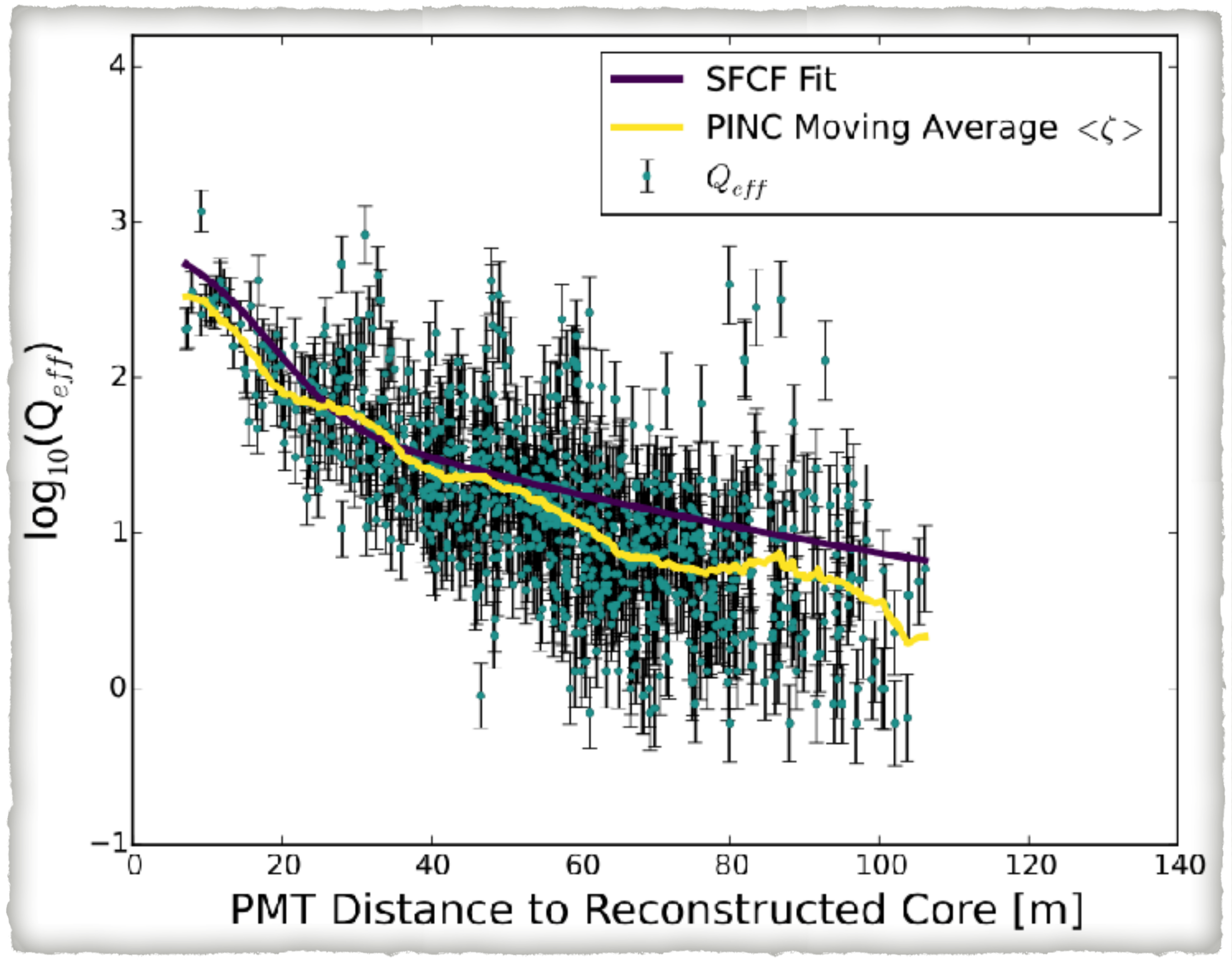
Pretz, J. (2016), https://doi.org/10.22323/1.236.0025

Potencial parameters

- LIC
- disMax
- LDFChi2
- LDFAmp
- PINC



Photon from Crab Nebula



Cosmic Ray

Abeysekara, A. U., et al (2017), https://doi.org/10.3847/1538-4357/aa7555

$$LIC = \log_{10} \frac{1}{compactness} = \log_{10} \frac{CxPE_{40}}{nHit}$$

$$NKG = A\ \rho^{s-3}\ (1+\rho)^{s-4.5}$$

$$PINC = \frac{1}{N} \sum_{i=0}^{N} \frac{[\log_{10}(q_i) - \langle \log_{10}(q_i) \rangle]^2}{\sigma^2}$$

Bins

- 10 fractional hit bin ($\mathcal{B}$).
- 12 energy bins (ebin).

$\mathcal{B}$	Range (%)	*ebin*
0	4.4–6.7	2.50
1	6.7–10.5	2.75
2	10.5–16.2	3.00
3	16.2–24.7	3.25
4	24.7–35.6	3.50
5	35.6–48.5	3.75
6	48.5–61.8	4.00
7	61.8–74.0	4.25
8	74.0–84.0	4.50
9	84.0–100.0	4.75
		5.00
		5.25

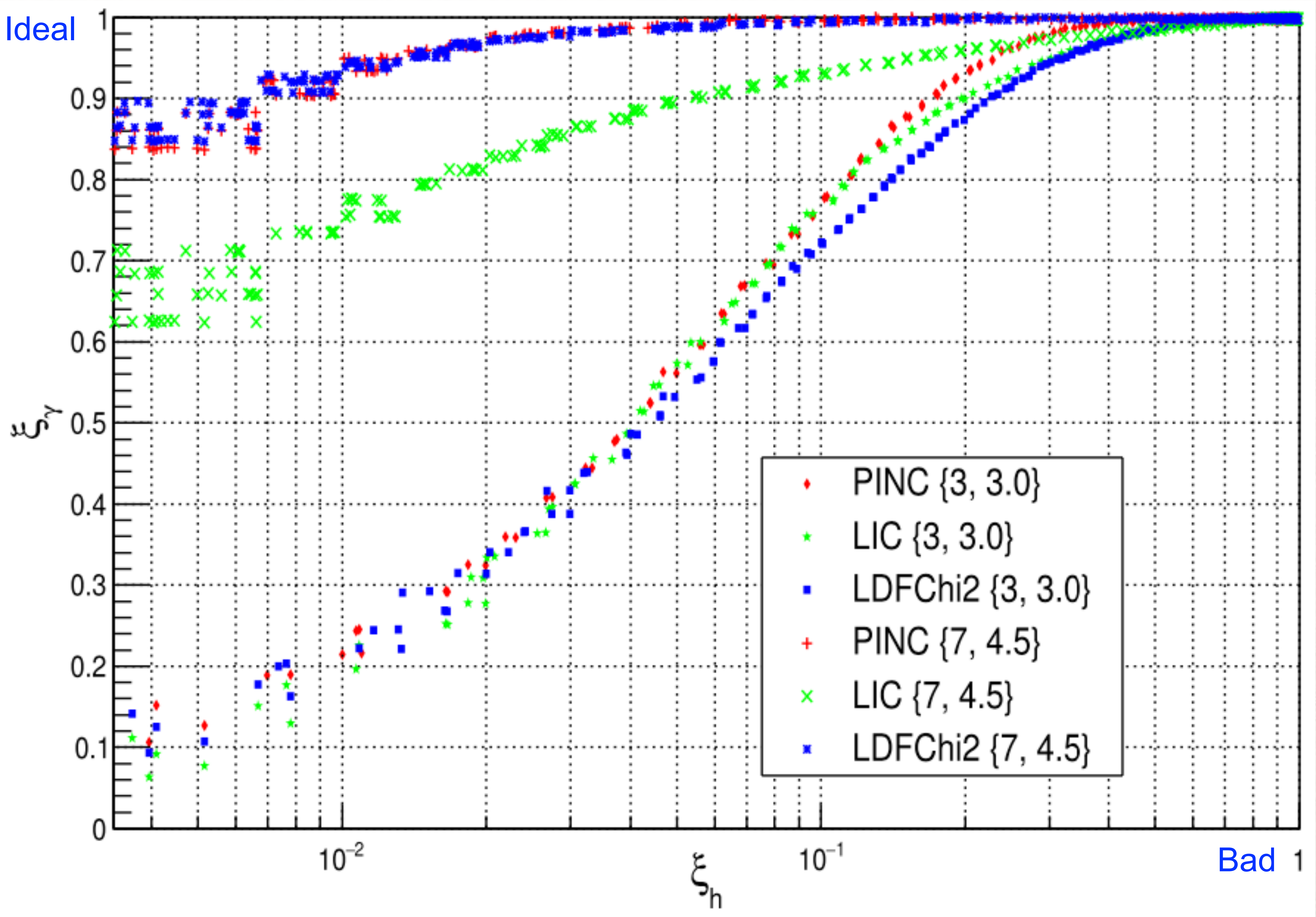
Note: Some bins are not used because they have poor statistic and they add more noise than signal. So, in this work, 67 bins are used.

How to verify their performance

Receiver Operating Characteristic (ROC) Curves

ξ_γ Gamma efficiency

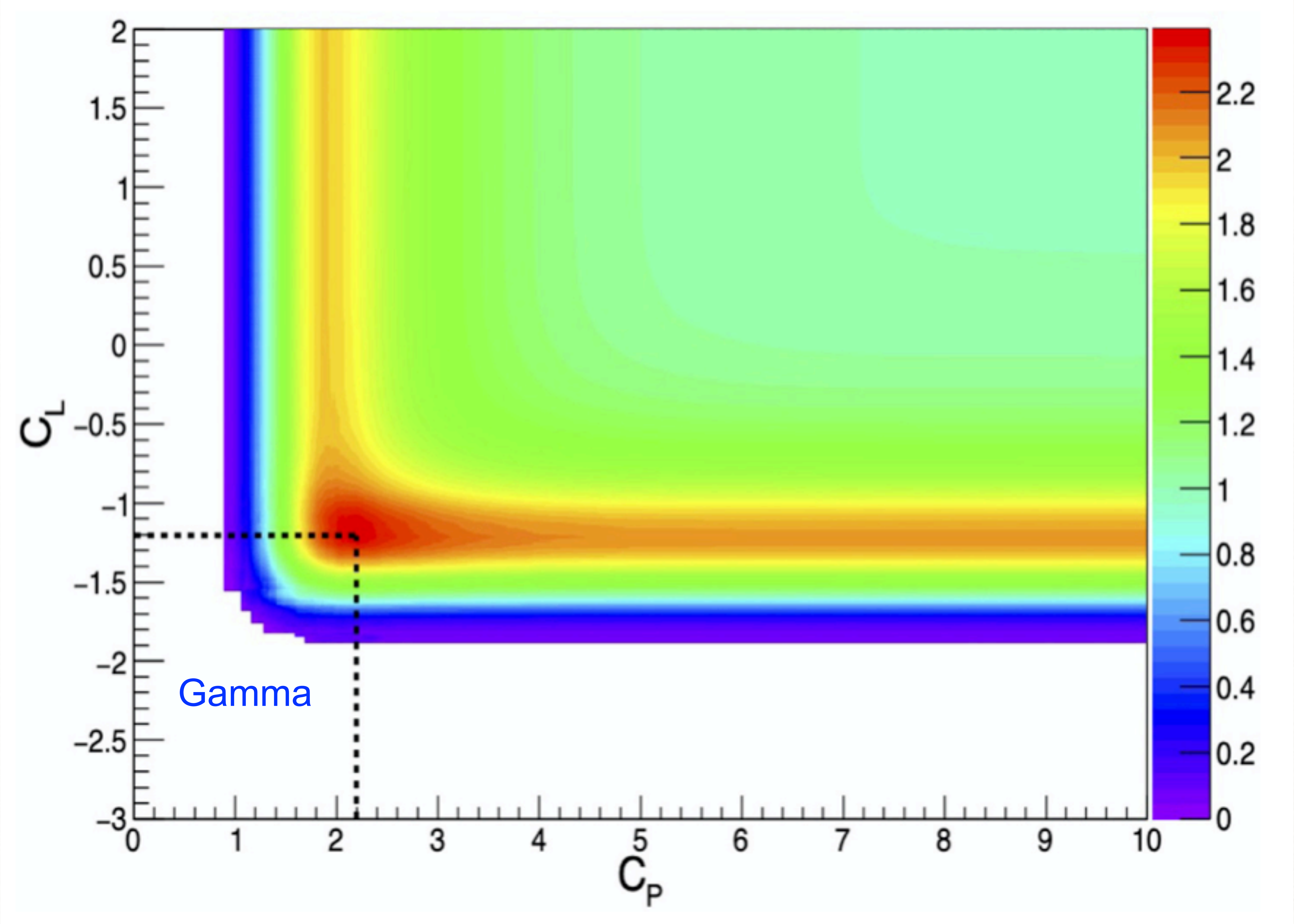
ξ_h Hadron misidentification rate

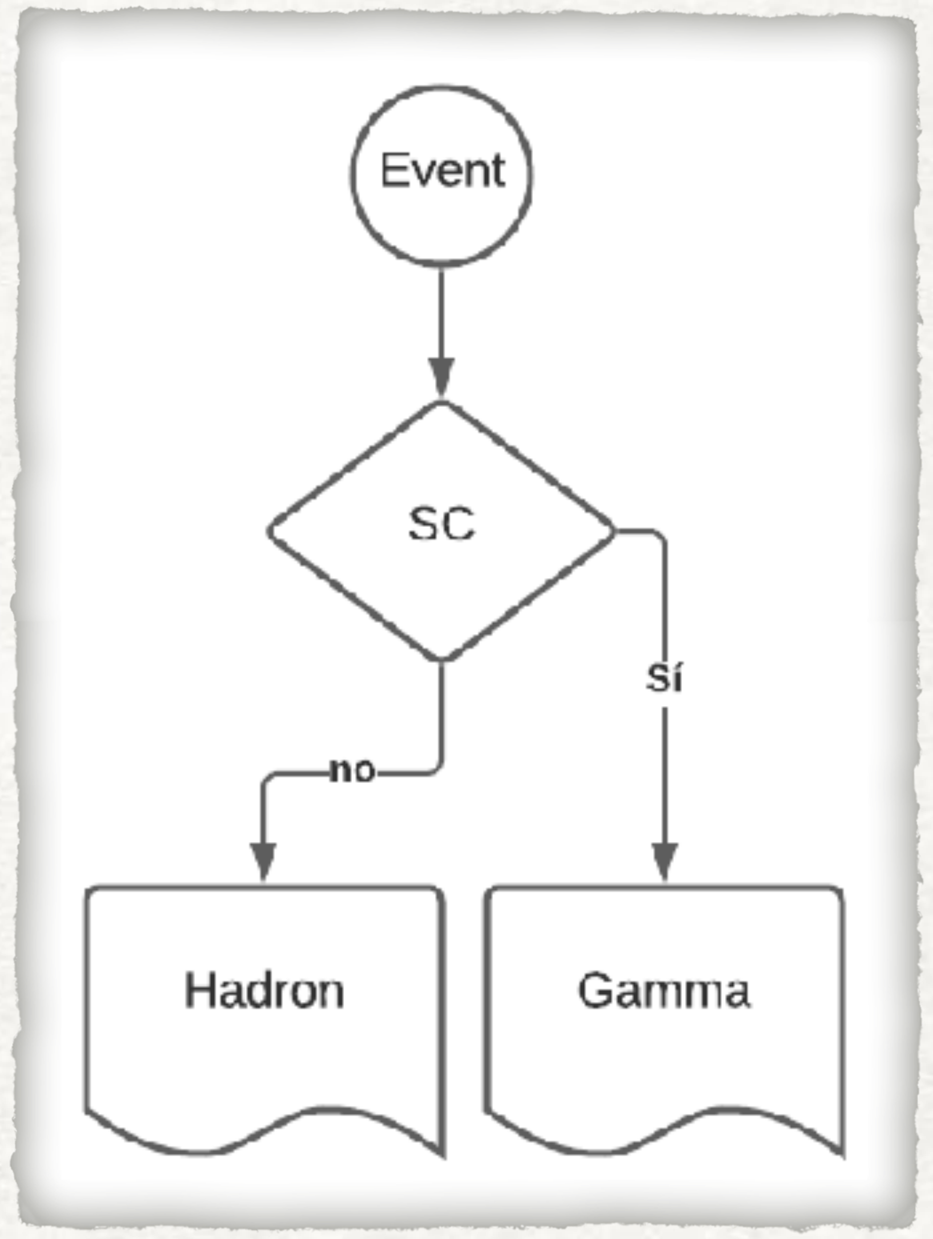


Alfaro, R. (2022), https://doi.org/10.1016/j.nima.2022.166984

Gamma/hadron separation models

- **Standard Cut (SC):** $(LIC < C_L)\ \&\ (PINC < C_P)$





$$Q = \frac{\xi_\gamma}{\sqrt{\xi_h}}$$

ξ_γ Gamma efficiency

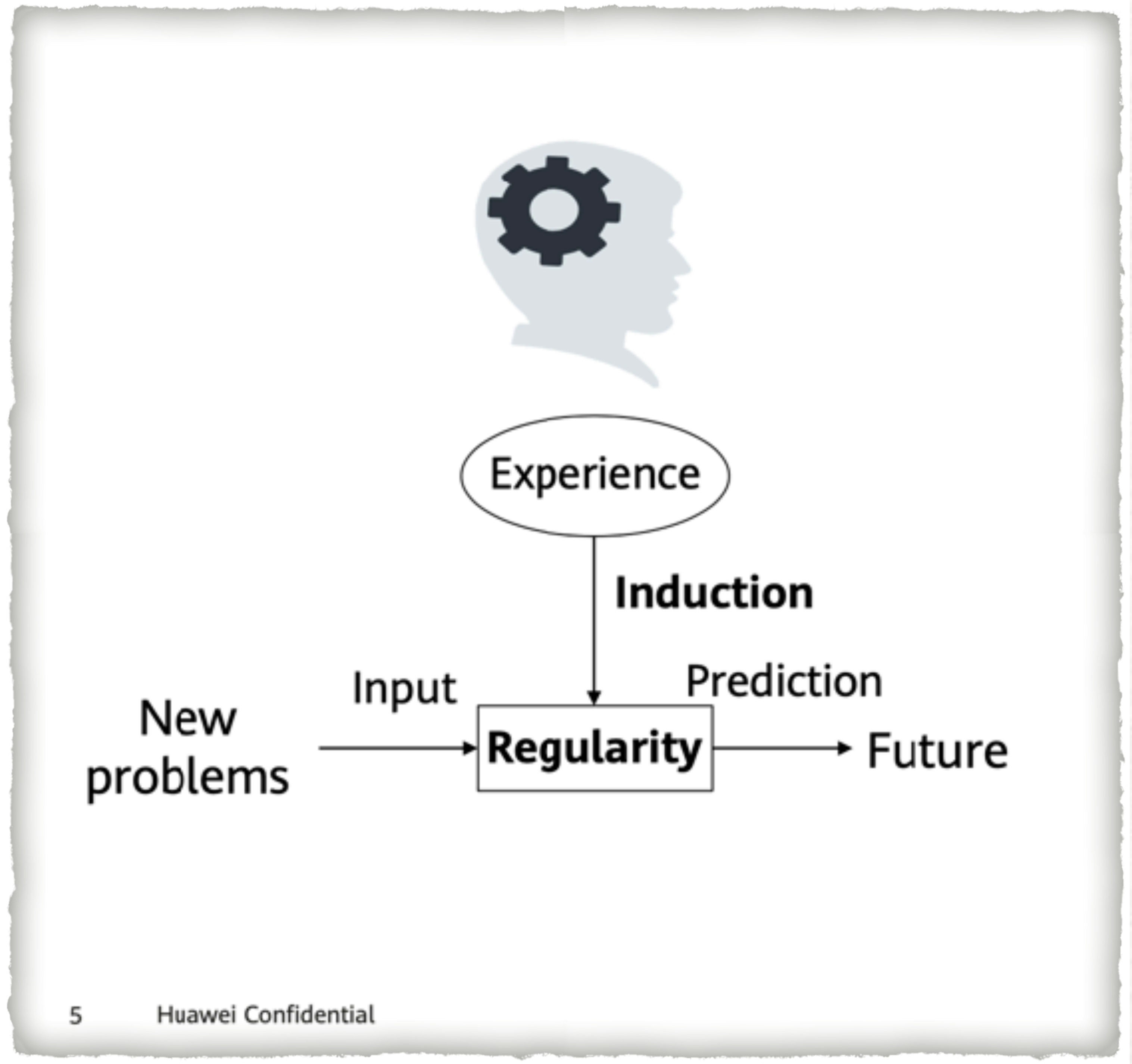
ξ_h Hadron misidentification rate

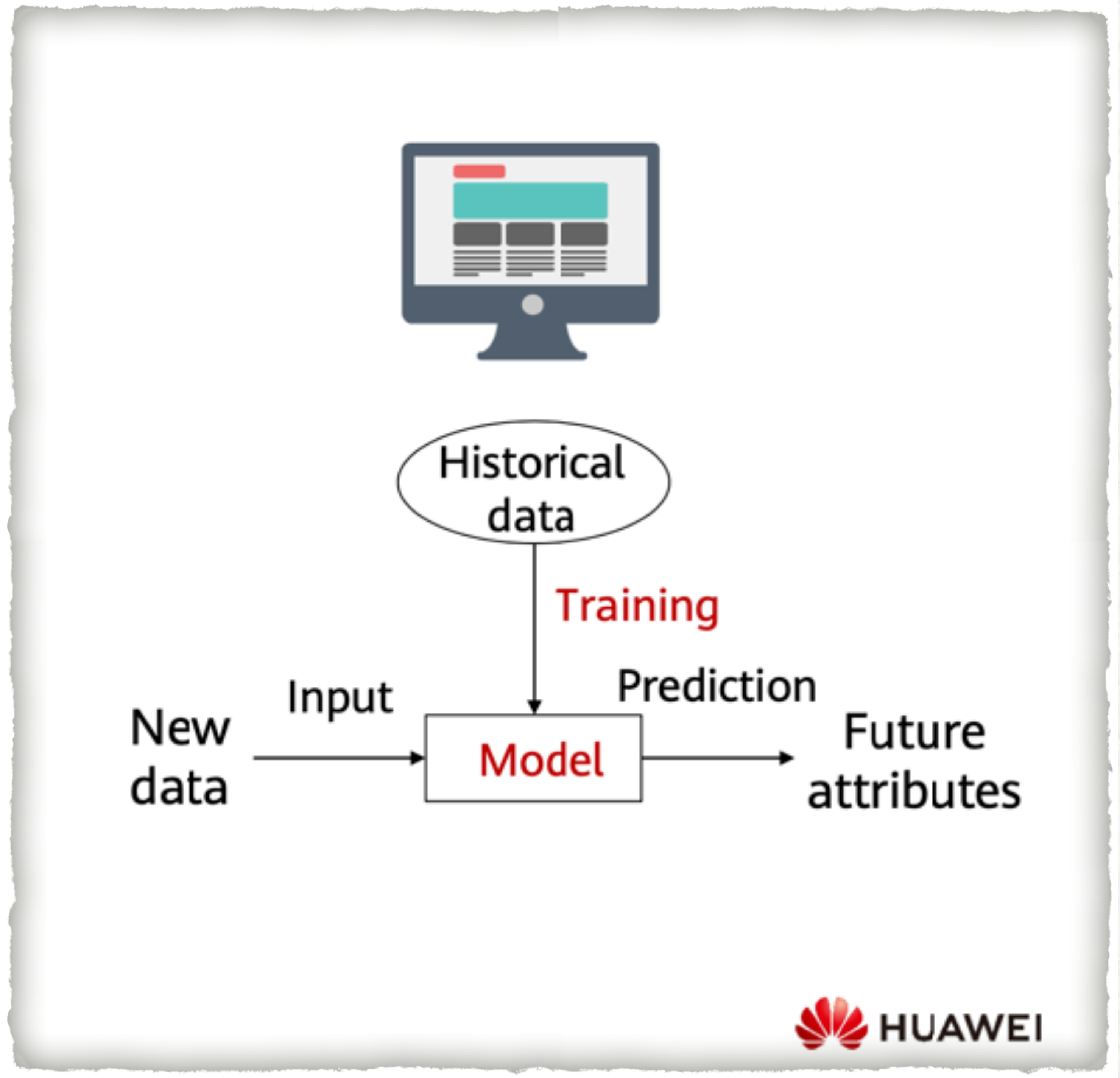
Alfaro, R. (2022), https://doi.org/10.1016/j.nima.2022.166984

Gamma/hadron separation models

- Machine Learning Technique (MLT):
  - Learn from data
  - A complex model is built
- "Mathematical view"

$$R^n \to R^m$$



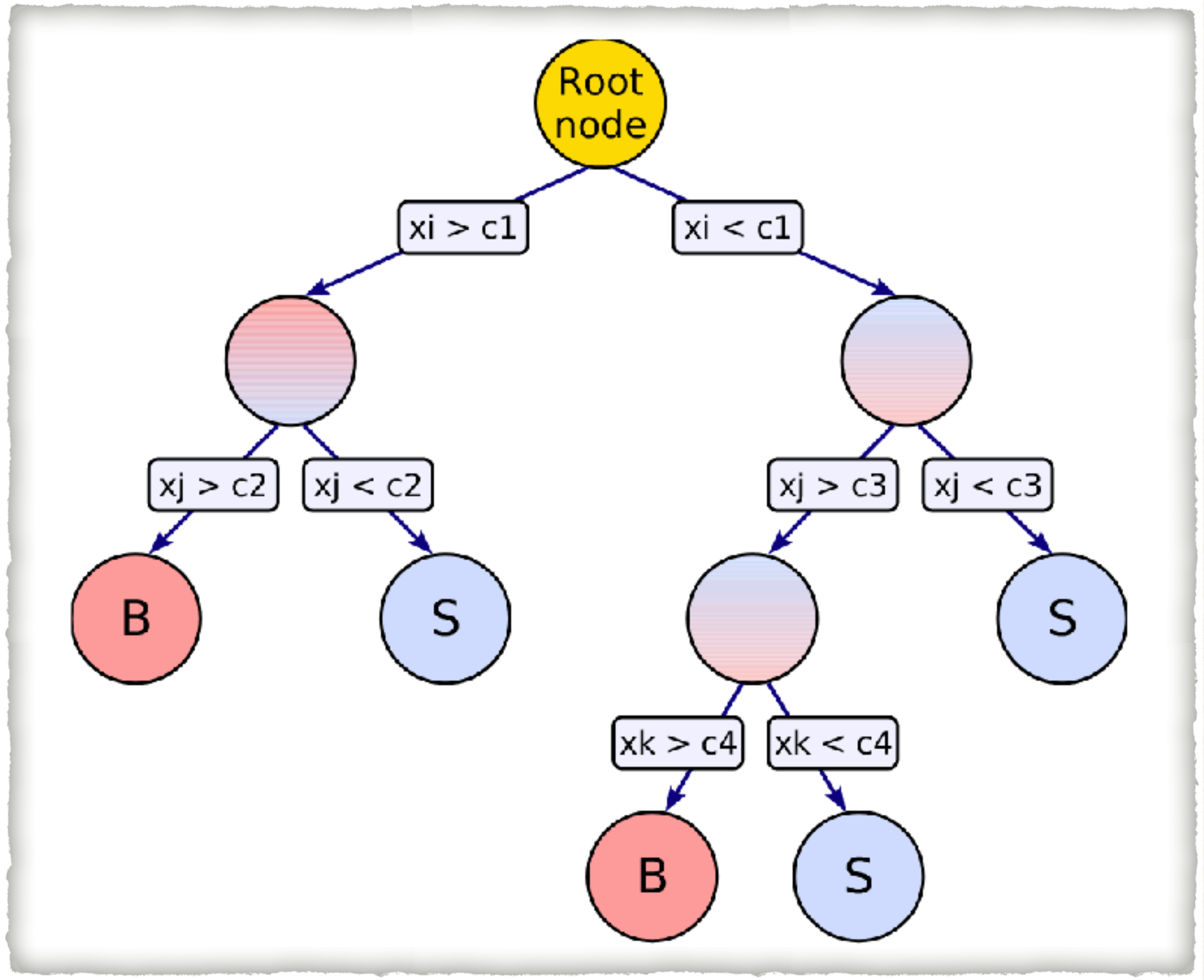


Taken from https://ilearningx.huawei.com/portal/courses/HuaweiX+EBG2020CCHW1100087/about

Building the model

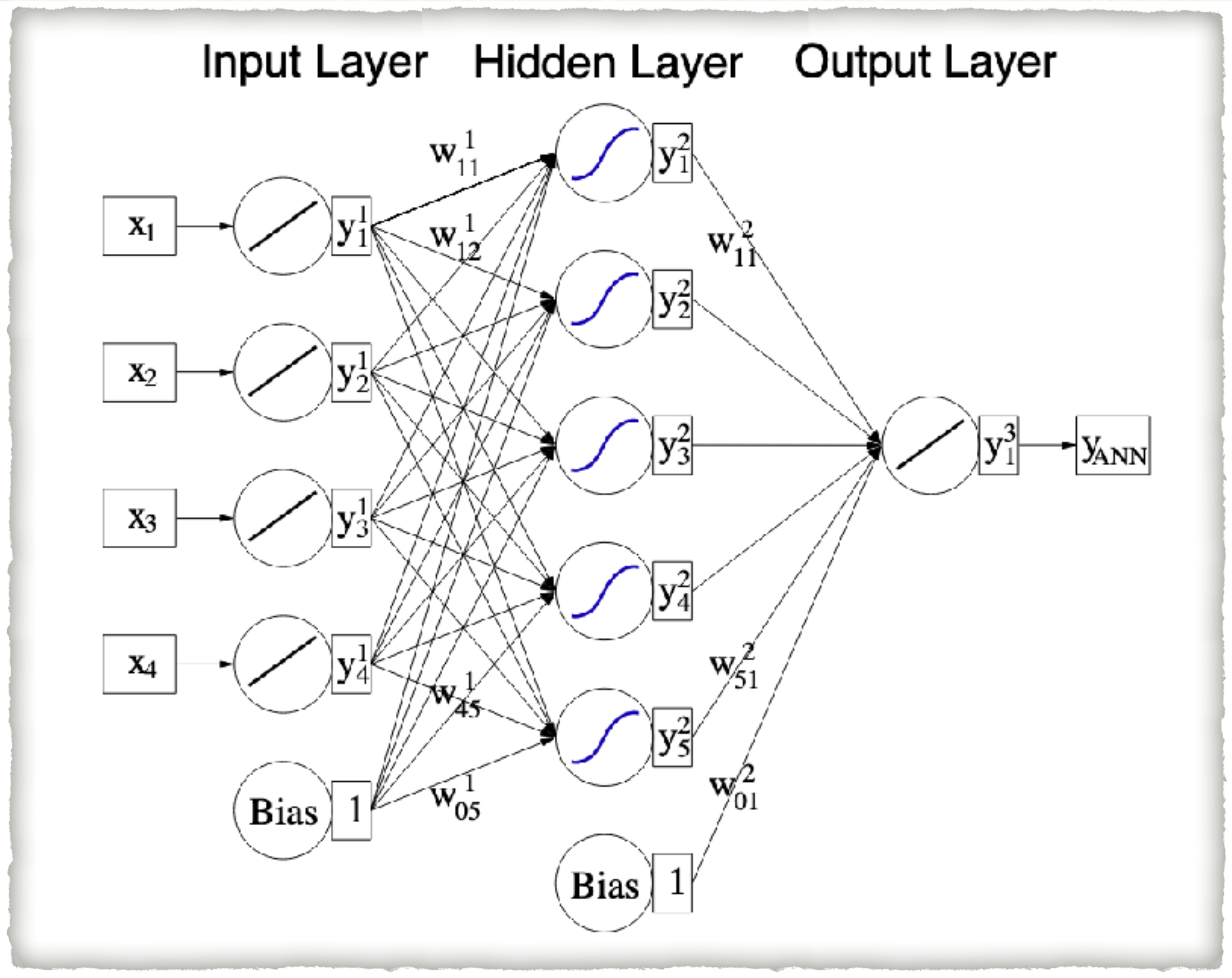
- Machine Learning Technique (MLT):

Boosted Decision Tree



Speckmayer, P., et al (2010) https://doi.org/10.1088/1742-6596/219/3/032057

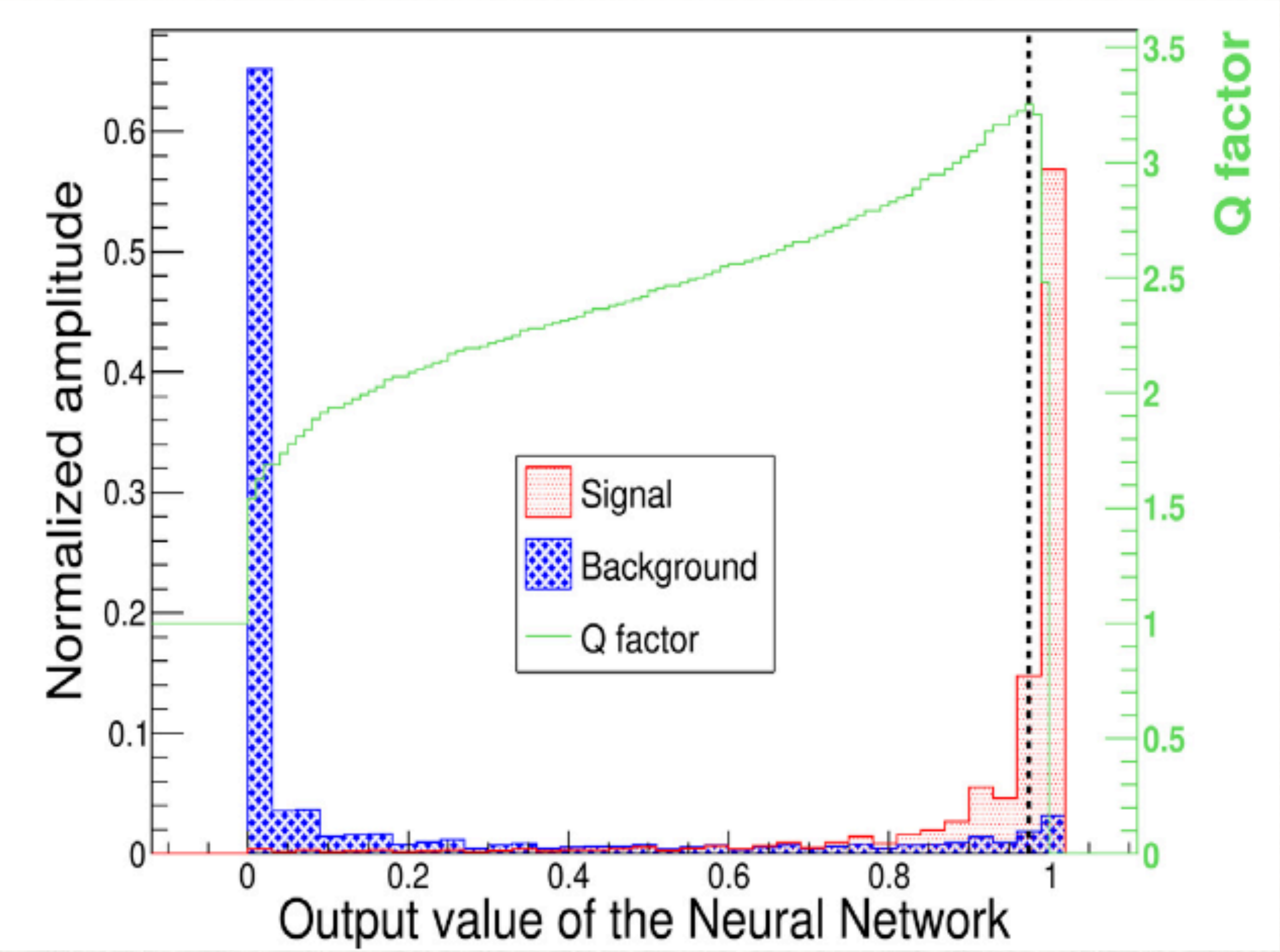
Neural Networks



Speckmayer, P., et al (2010) https://doi.org/10.1088/1742-6596/219/3/032057

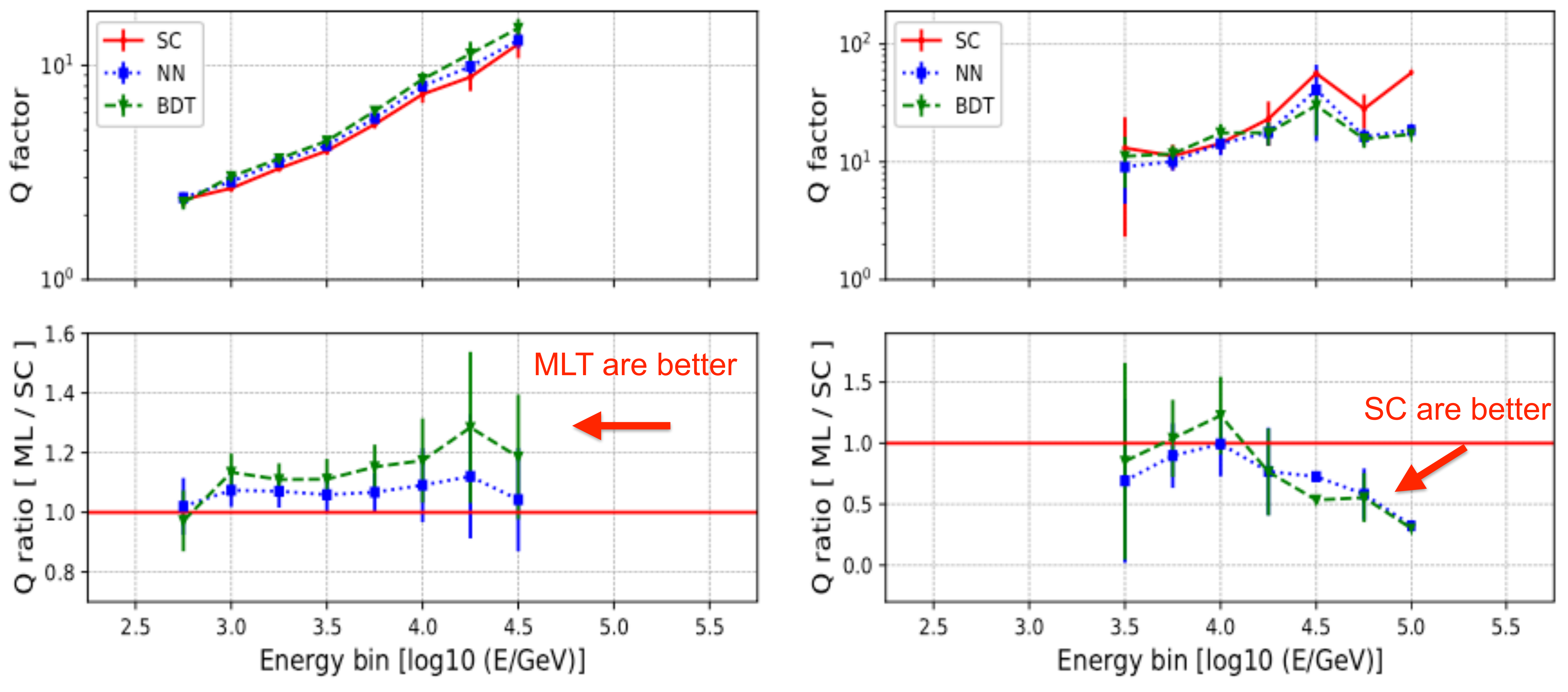
Building the model

- Machine Learning Technique (MLT):



Alfaro, R. (2022), https://doi.org/10.1016/j.nima.2022.166984

Results: Testing on MC data

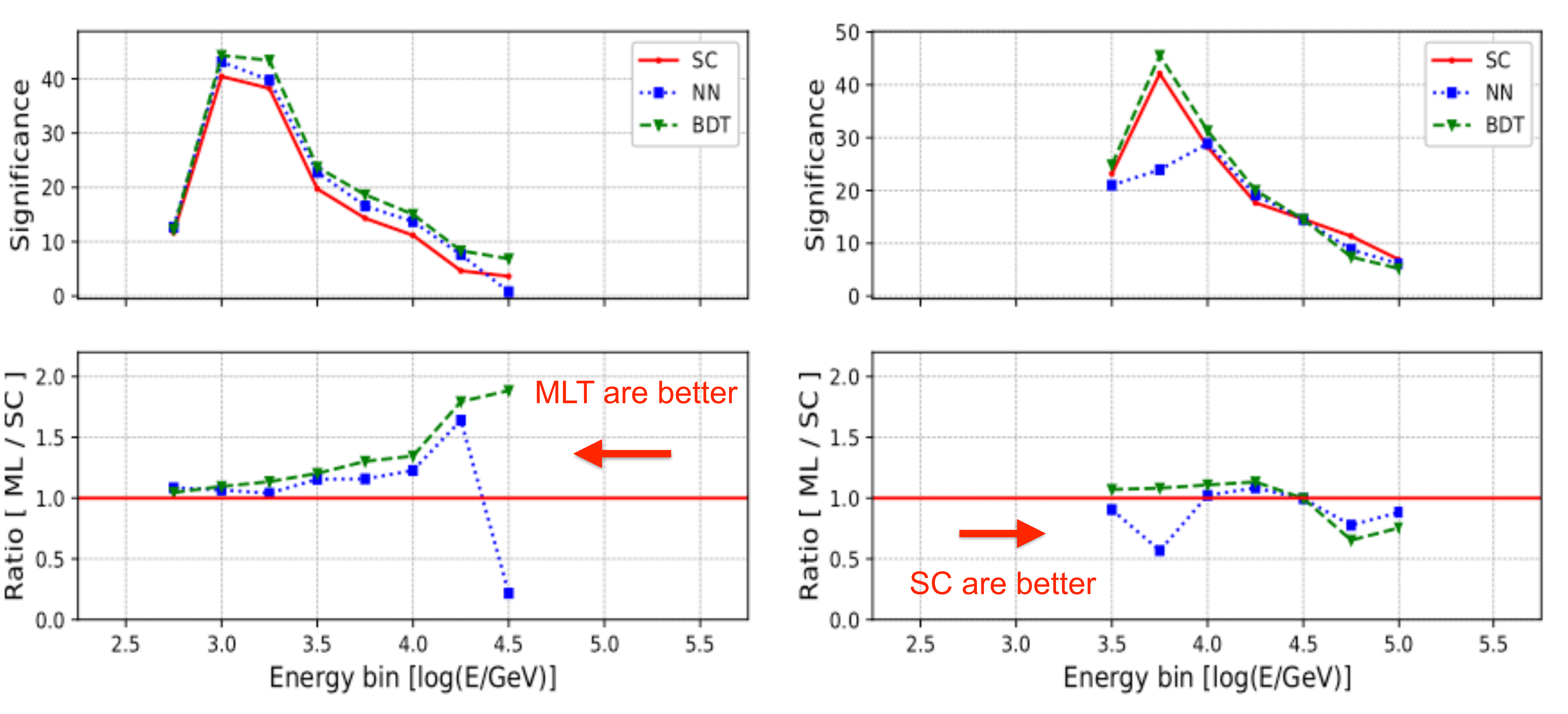


(a) $\mathcal{B} = 3$ (b) $\mathcal{B} = 6$

Alfaro, R. (2022), https://doi.org/10.1016/j.nima.2022.166984

MLTs reports better performance on the first six $\mathcal{B}$ bins, while the SC is better for the rest bins

Results: Testing on Real data (Crab Nebula)

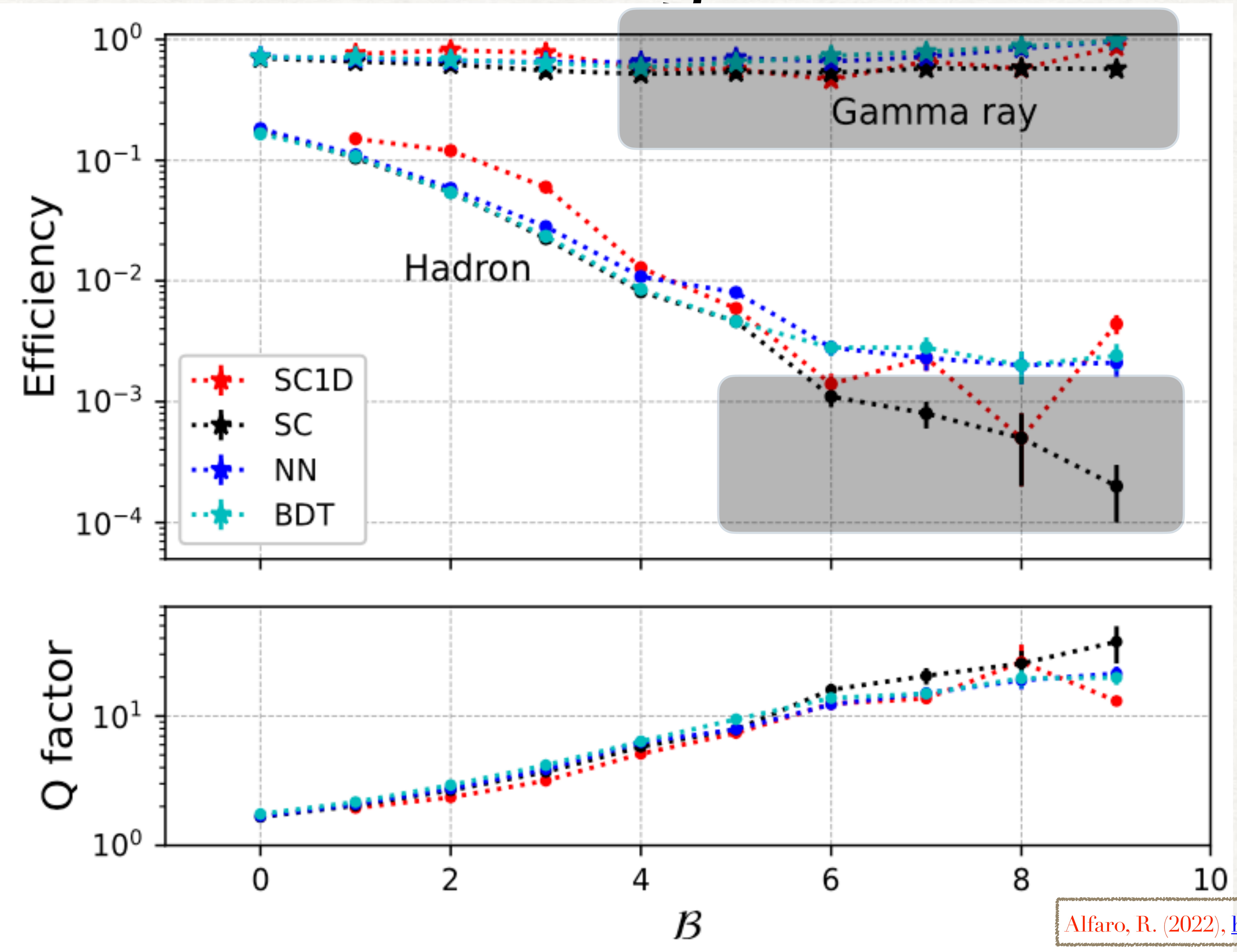


(a) $\mathcal{B}$=3 (b) $\mathcal{B} = 6$

Alfaro, R. (2022), https://doi.org/10.1016/j.nima.2022.166984

MLTs show a real improvement over the SC, however in $\mathcal{B} = 6$ all models has similar results and at high energies bins the SC is the best.

Results: Testing on MC data



MLT are better

SC are better

Alfaro, R. (2022), https://doi.org/10.1016/j.nima.2022.166984

The 2D models generally have greater predicted Q factor

Results: Testing on Real data

- Crab Nebula

$\mathcal{B}$	Significance				Difference in % between				
	SC1D	SC	NN	BDT	SC & SC1D	NN & SC1D	BDT & SC1D	NN & SC	BDT & SC
0	–	15.2	14.7	16.0	–	–	–	−3	5
1	26.9	27.6	27.5	28.22	3	2	5	0	2
2	37.8	44.1	44.6	46.4	17	18	23	1	5
3	59.2	62.4	66.1	72.0	5	12	22	6	15
4	70.6	69.7	76.3	76.2	−1	8	8	10	9
5	67.3	71.3	69.7	80.1	6	4	19	−2	12
6	52.3	61.5	48.3	66.0	18	−8	26	−21	7
7	39.1	47.7	49.2	50.3	22	26	28	3	5
8	27.6	32.8	35.1	34.8	19	27	26	7	6
9	28.2	28.7	31.3	31.3	2	11	11	9	9
1–9	144.0	155.7	156.9	170.7	8	9	19	1	10
0–9	–	156.3	157.5	171.3	–	–	–	1	10

Generally, the 2D models provide better results than SC1D.

Adding $\mathcal{B} = 0$ gives only a slight improvement. This bin requires a different approach if a useful signal is to be extracted from it.

Results: Testing on Real data

- Markarian 421

$\mathcal{B}$	Significance				Difference in % between				
	SC1D	SC	NN	BDT	SC & SC1D	NN & SC1D	BDT & SC1D	NN & SC	BDT & SC
0	–	8.46	8.28	8.40	–	–	–	−2	−1
1	11.9	13.2	12.5	13.0	11	5	10	−5	−1
2	16.2	16.2	15.6	16.6	0	−4	2	−3	2
3	19.0	18.9	19.9	21.2	−1	4	11	5	12
4	21.6	19.5	21.9	20.7	−10	2	−4	12	6
5	16.5	15.0	15.5	17.6	−9	−6	7	4	18
6	9.7	9.3	8.4	11.0	−4	−13	13	−9	18
7	4.2	5.6	7.2	6.9	34	72	65	28	23
8	–	–	–	–	–	–	–	–	–
9	–	–	–	–	–	–	–	–	–
1–9	35.9	35.3	36.0	38.6	−2	0	8	2	10
0–9	–	36.0	36.6	39.3	–	–	–	2	9
Crab Improvements									
1–9					8	9	19	1	10

- Markarian 501

$\mathcal{B}$	Significance				Difference in % between				
	SC1D	SC	NN	BDT	SC & SC1D	NN & SC1D	BDT & SC1D	NN & SC	BDT & SC
0	–	–	–	–	–	–	–	–	–
1	3.4	3.8	4.2	4.6	12	25	36	11	21
2	4.5	2.9	3.1	3.7	−36	−32	−17	6	29
3	4.7	5.3	4.5	4.2	14	−5	−10	−16	−21
4	5.1	5.1	6.2	4.4	0	20	−14	20	−14
5	4.1	3.8	4.3	5.7	−9	4	38	15	51
6	3.8	5.0	2.0	5.7	31	−47	50	−59	14
7	1.6	2.2	2.5	2.9	43	60	85	12	30
8	2.6	2.7	2.3	2.9	3	−10	12	−13	8
9	–	–	–	–	–	–	–	–	–
1–9	10.3	10.6	10.2	11.9	4	0	16	−4	12
Crab Improvements									
1–9					8	9	19	1	10

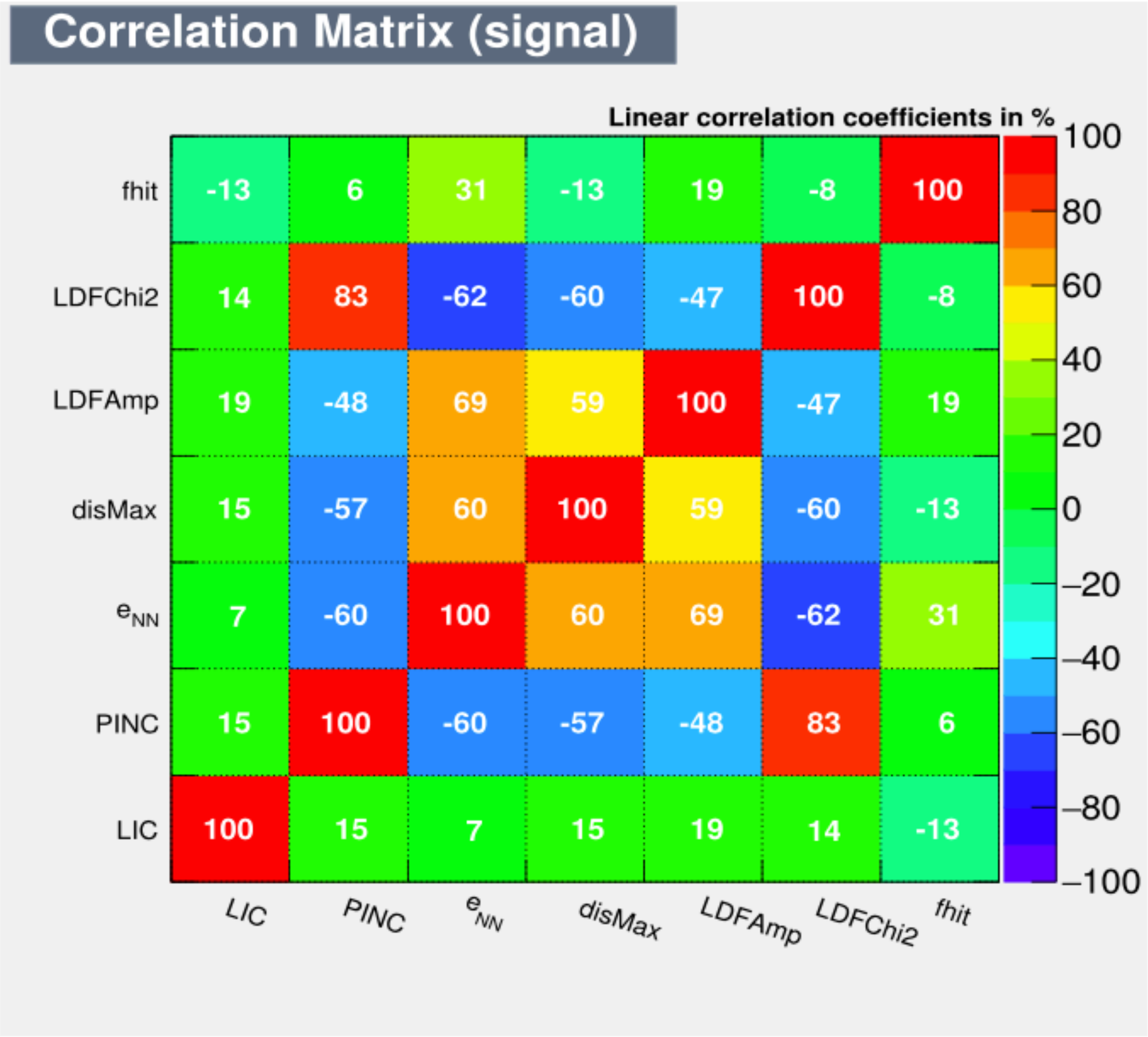
Summary

- Two new implementations were used to classify the shower detected by HAWC: BDT & NN.
- They are compared with the official HAWC technique to distinguish gamma from hadron using MC data and real data.
- The 2D models generally have greater predicted Q factor. MLT recognizes better the hadrons at low B bins and gammas at high B bins.
- Generally, the 2D models provide better results than SC1D using real data. The best model is BDT, It improved the results on the three sources.

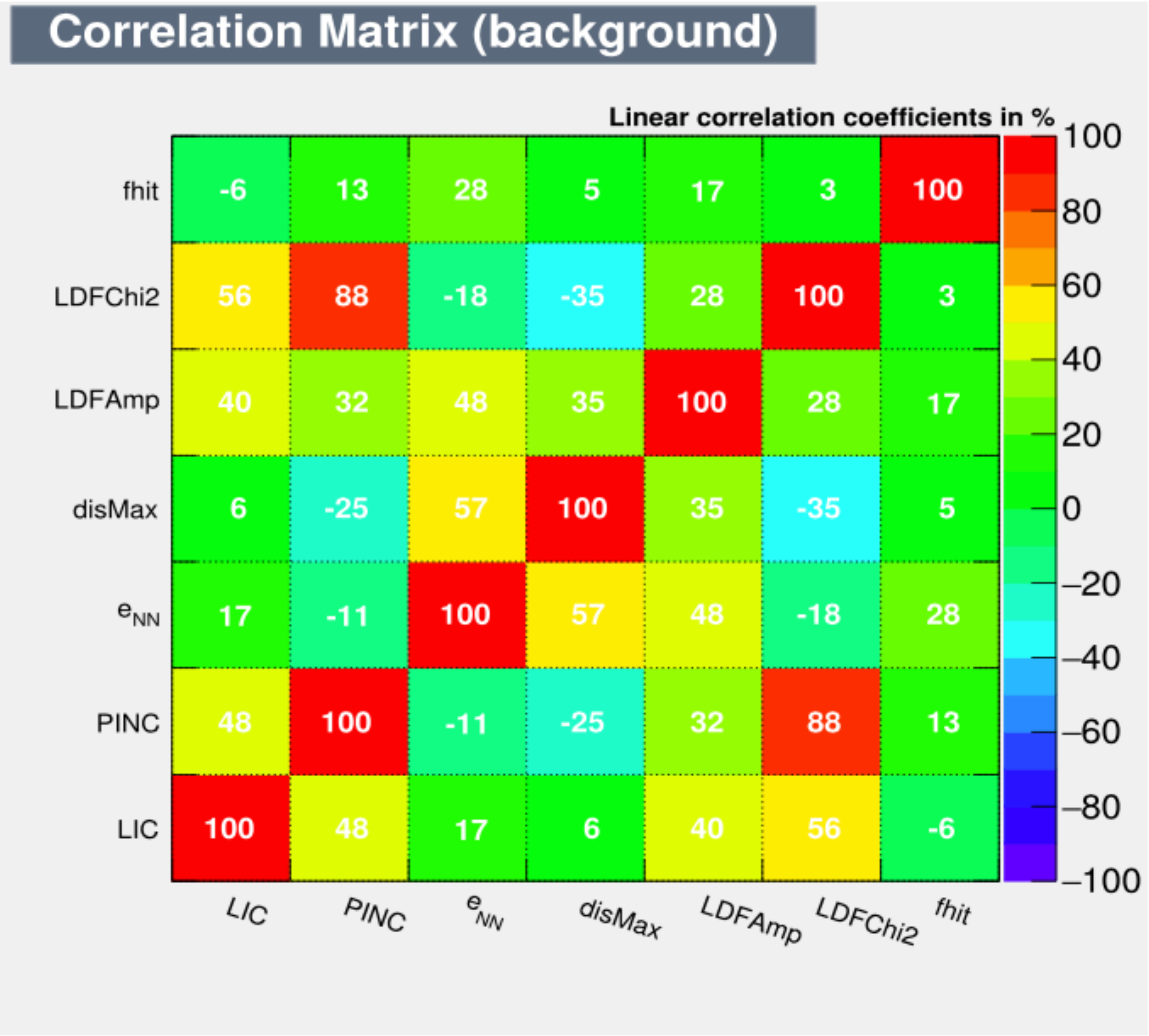
Thank you!

Backslides

Linear Correlation

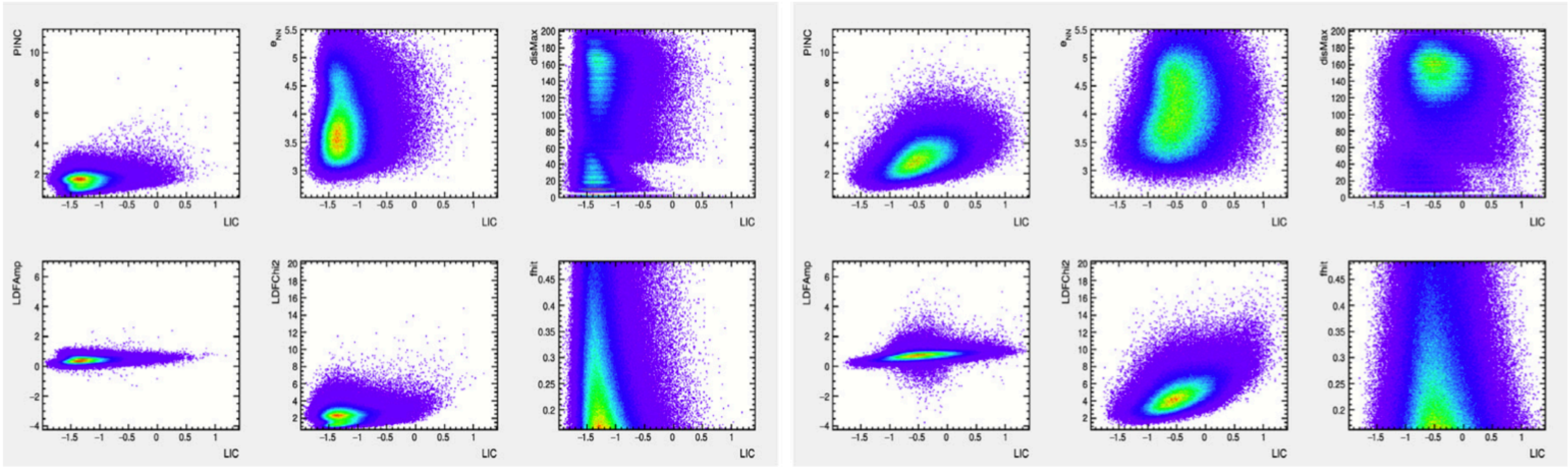


(a) Signal.



(b) Background

Correlations among variables



(a) Signal

(b) Background

Ranking

NN			BDT		
$\mathcal{B}$ 0–2	$\mathcal{B}$ 3–5	$\mathcal{B}$ 6–9	$\mathcal{B}$ 0–2	$\mathcal{B}$ 3–5	$\mathcal{B}$ 6-9
PINC	**PINC**	**PINC**	**LDFChi2**	**PINC**	**PINC**
LDFChi2	**LDFChi2**	**LDFChi2**	**LiC**	**LiC**	**LDFAmp**
fHit	**LiC**	**LDFAmp**	**PINC**	**LDFAmp**	**LDFChi2**
e_{NN}	disMax	*fHit*	***fHit***	**LDFChi2**	LiC
LiC	*fHit*	disMax	LDFAmp	*fHit*	*fHit*
disMax	e_{NN}	LiC	e_{NN}	disMax	disMax
LDFAmp	LDFAmp	e_{NN}	disMax	e_{NN}	e_{NN}