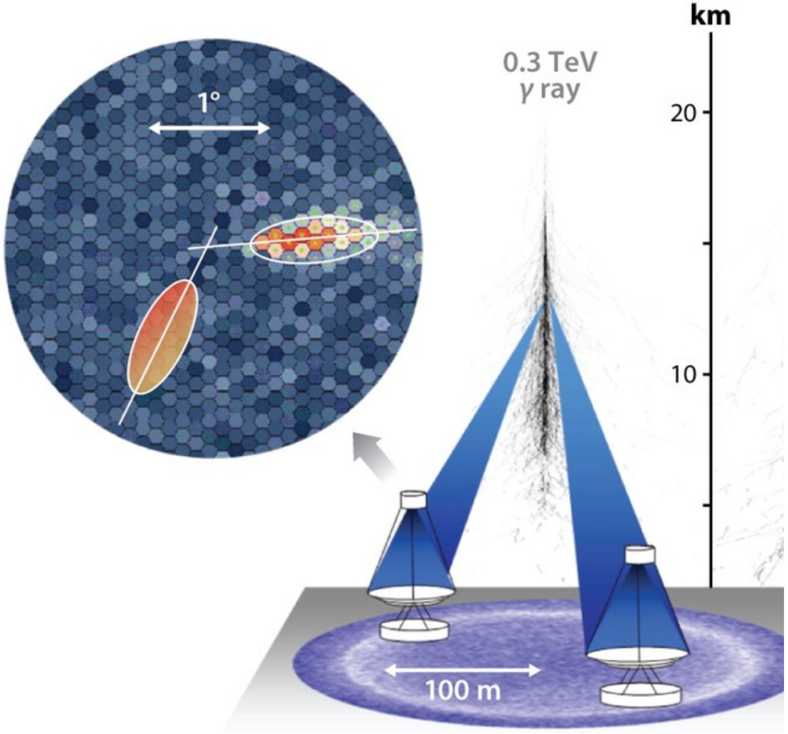
Application of Deep Learning methods to analysis of IACT data

Büchele, Matthias
(for the H.E.S.S. Deep Learning Task Group)
Aachen, 19.02.2019

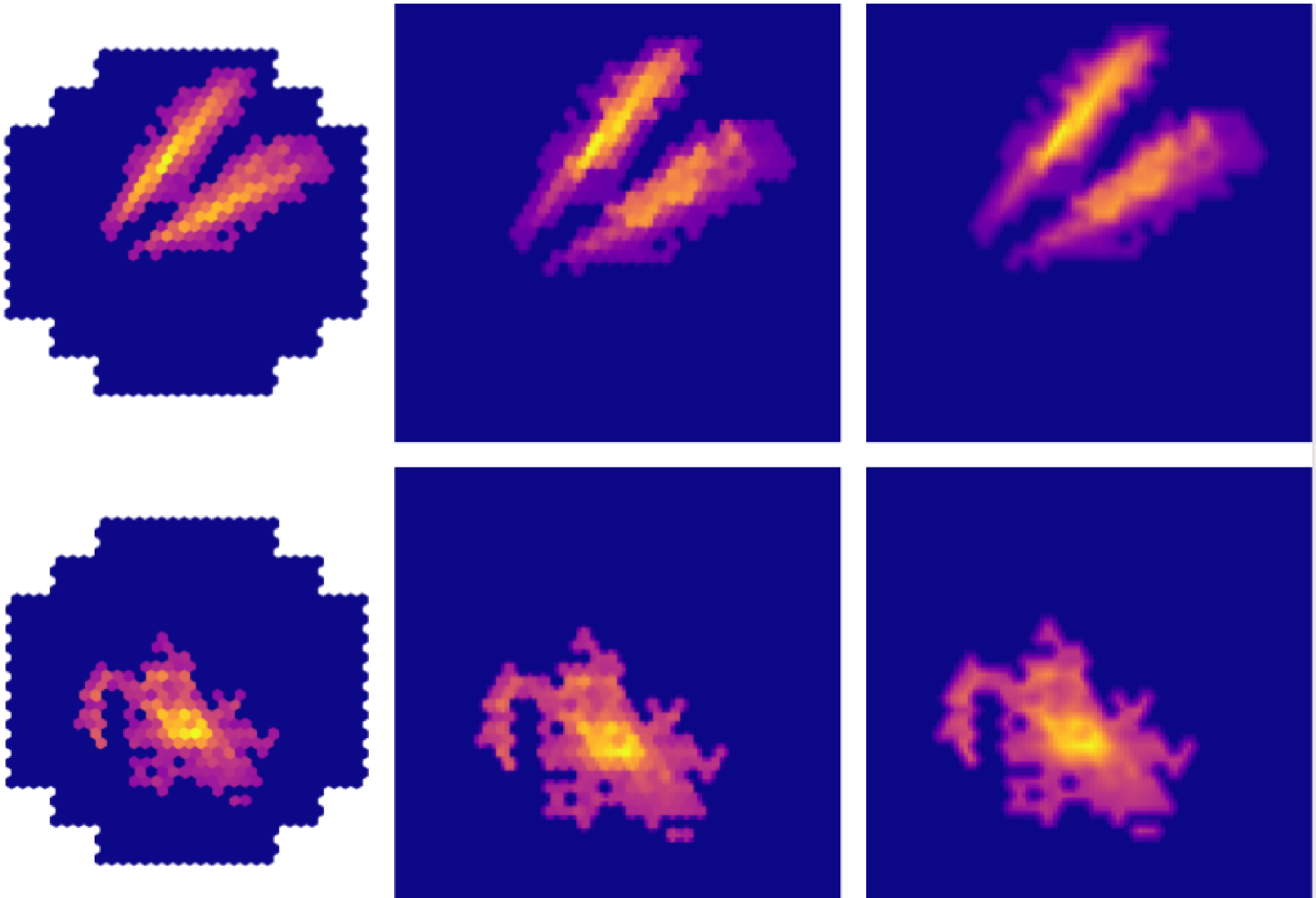
H.E.S.S.
FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS
HUMBOLDT-UNIVERSITÄT ZU BERLIN
Universität Potsdam

H.E.S.S. - our IACTs



Hinton JA, Hofmann W. 2009.
Annu. Rev. Astron. Astrophys. 47:523–65

Our Data & Architecture

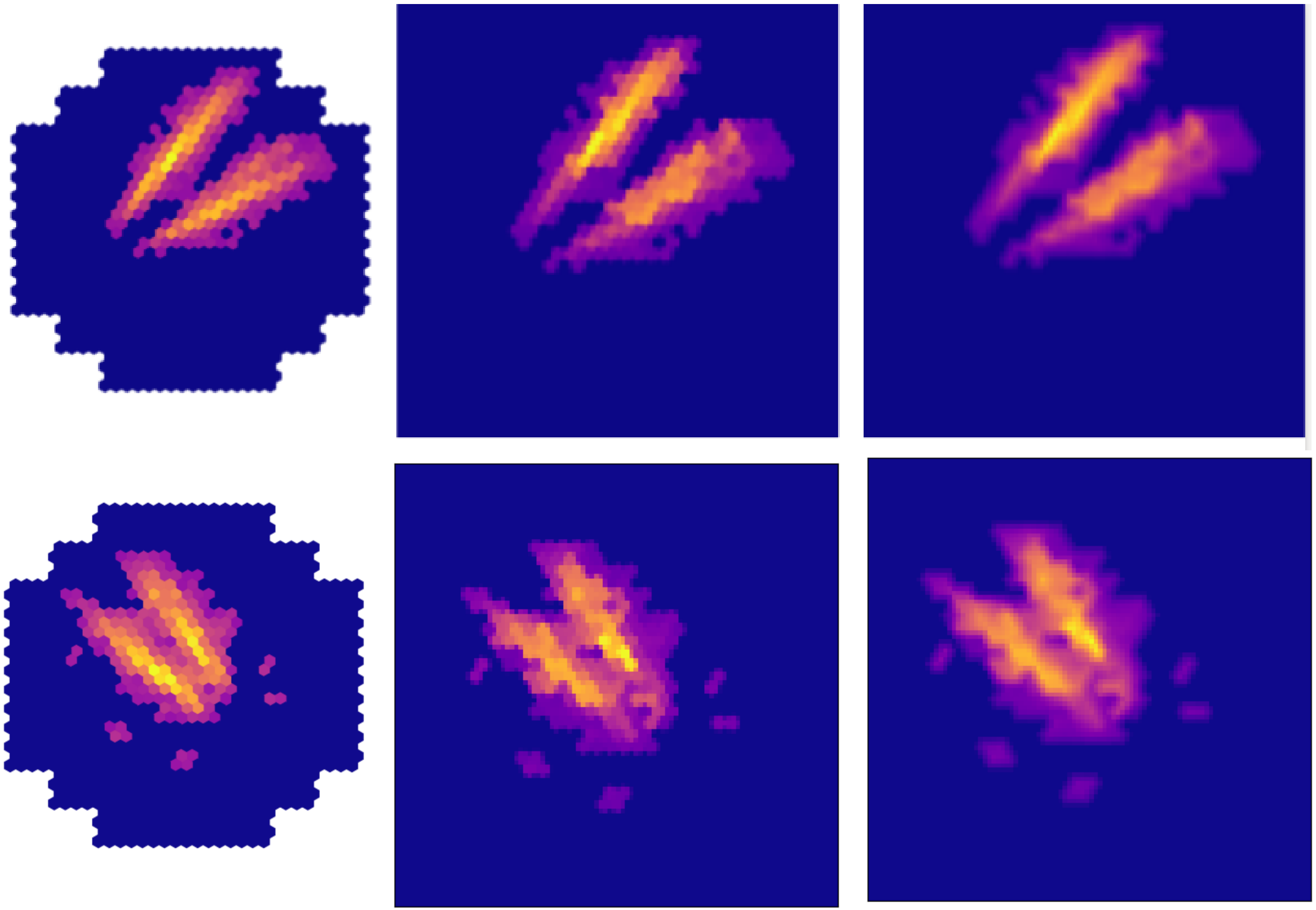


Gamma
(Signal)

In Data:
1 in 1000

Proton
(Background)

Our Data & Architecture



Gamma
(Signal)

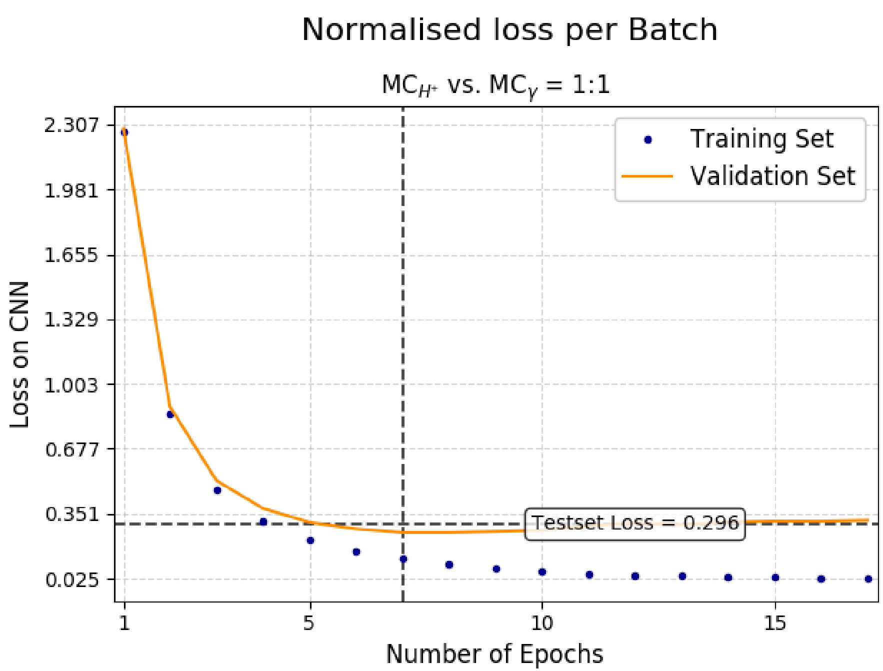
In Data:
1 in 1000

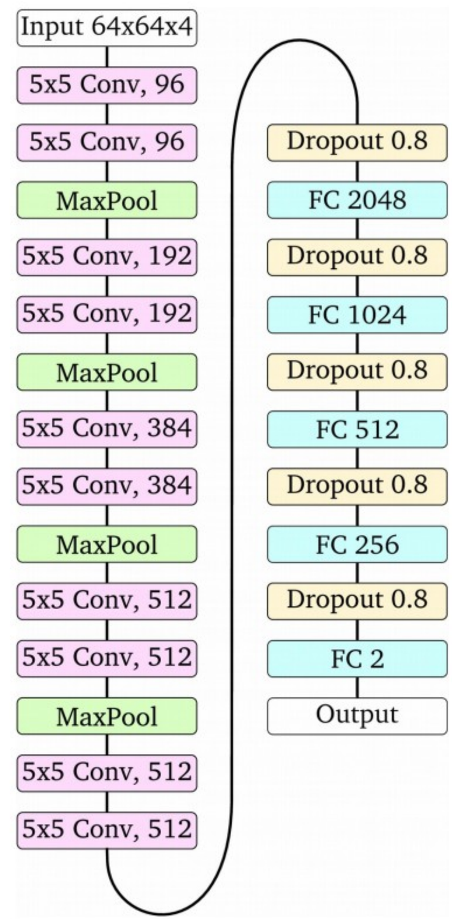
Proton
(Background)

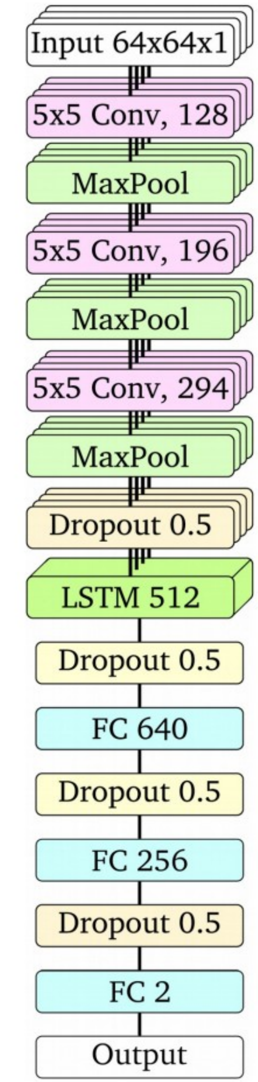
Our Data & Architecture

Most of the time:

non fancy.







Status on Tasks

- Classification (γ : Background) on Monte Carlos: Acc 96.1%
- Classification (γ : Background) on Data: better then best H.E.S.S.-Analysis

- Classification (γ : H : He : C : Si : Fe) on Monte Carlos: Acc ~60%
- Classification (γ : H : He : C : Si : Fe) on Data: w.i.p.

- Regression: Direction Reconstruction „workes“

  PSF (R_{68} @ 1 TeV) ~ 0.065° (MC), 0.102° (data)

  for comparison: Hillas ~ 0.100°, ImPACT ~ 0.050°

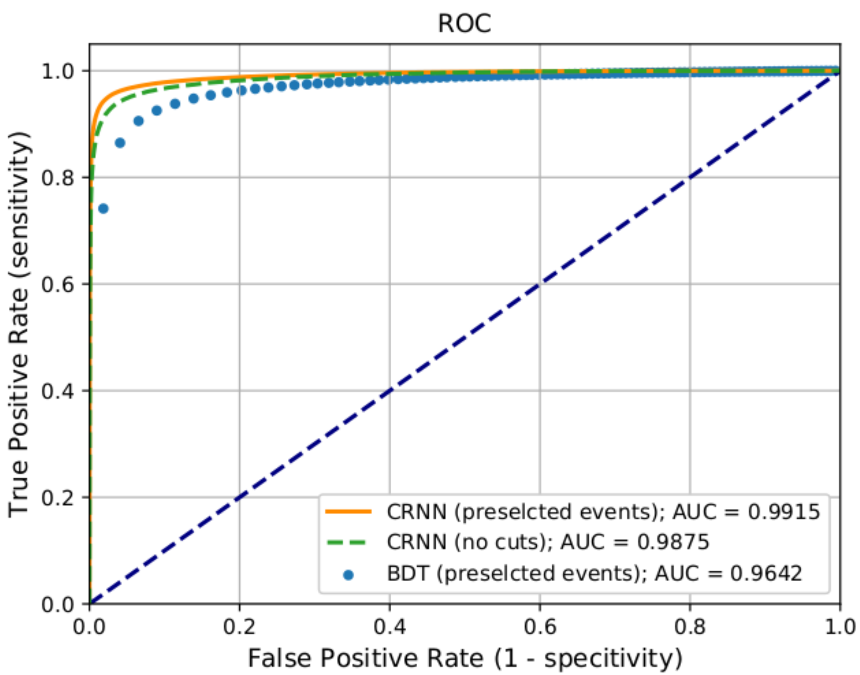
- Regression: Fe-Energy-Spectrum: Masters Thesis C. Hillig w.i.p.

So, as M. Erdmann mentioned: we had a paper

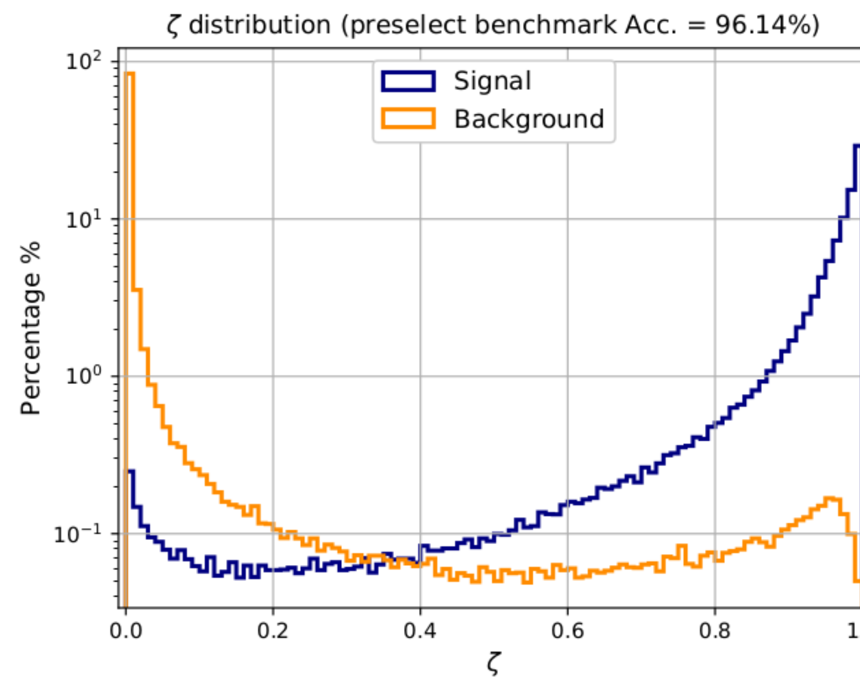
(For those who didn't find the time to read it, here are the main results)

* https://authors.elsevier.com/c/1Xy1J3Ix5tdddw

Paper results (Classification on MC)



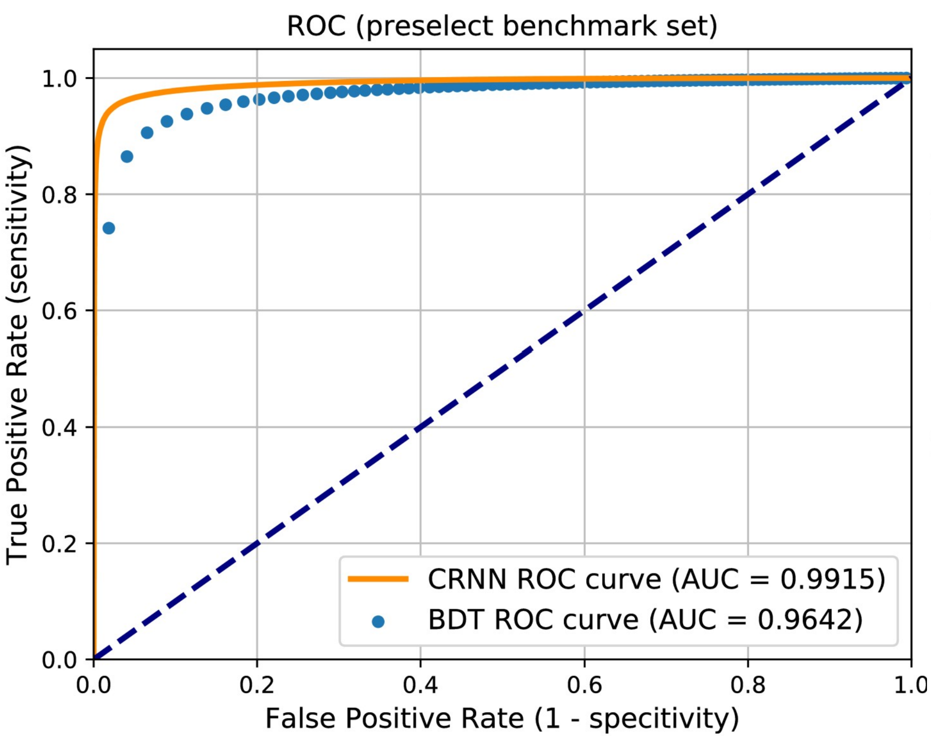
(a) ROC curves with matching AUC values.

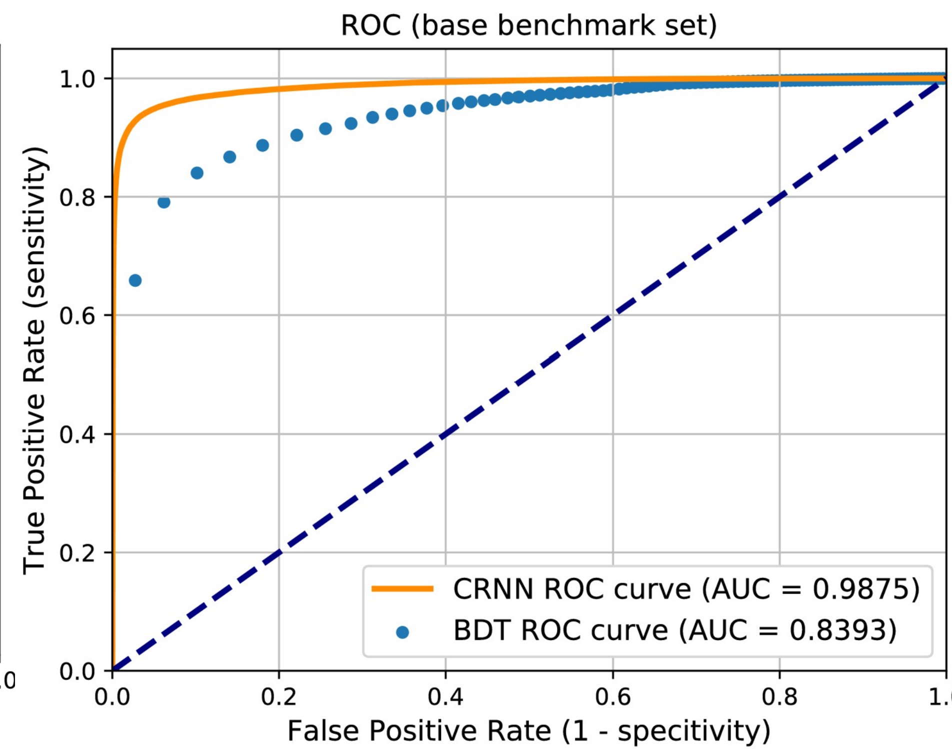


(b) CRNN ζ-score distribution for simulated signal and background events.

Figure 3: ROC curves for the CRNN classifier and the H.E.S.S. BDT classifier (left) and the ζ-score distribution for the CRNN classifier, obtained on the benchmark data-set with pre-selection cuts.

Paper results (Classification on MC)





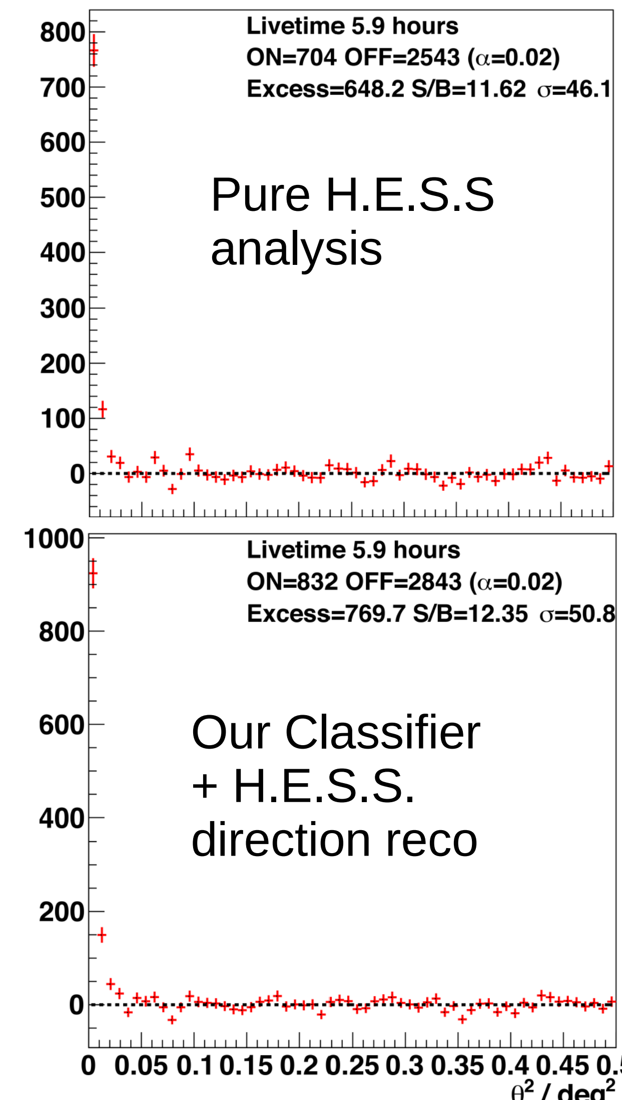
both with cuts

both without cuts

Paper results (Classification on data vs TMVA)

Configuration	N_{on}	αN_{off}	σ	S/B
ImPACT_BDT	704	55.8	46.1	11.6
ImPACT_CRNN	832	62.3	50.8	12.4

Table 1: Event statistics, significance as calculated by the method of [26] and signal to noise ratio of the 14 runs of PKS 2155-304 data, for two analyses: an ImPACT direction reconstruction with a BDT classifier (*top row*) and an ImPACT direction reconstruction with a CRNN classifier (*bottom row*).



Paper results (Direction on MC vs Hillas and ImPACT)

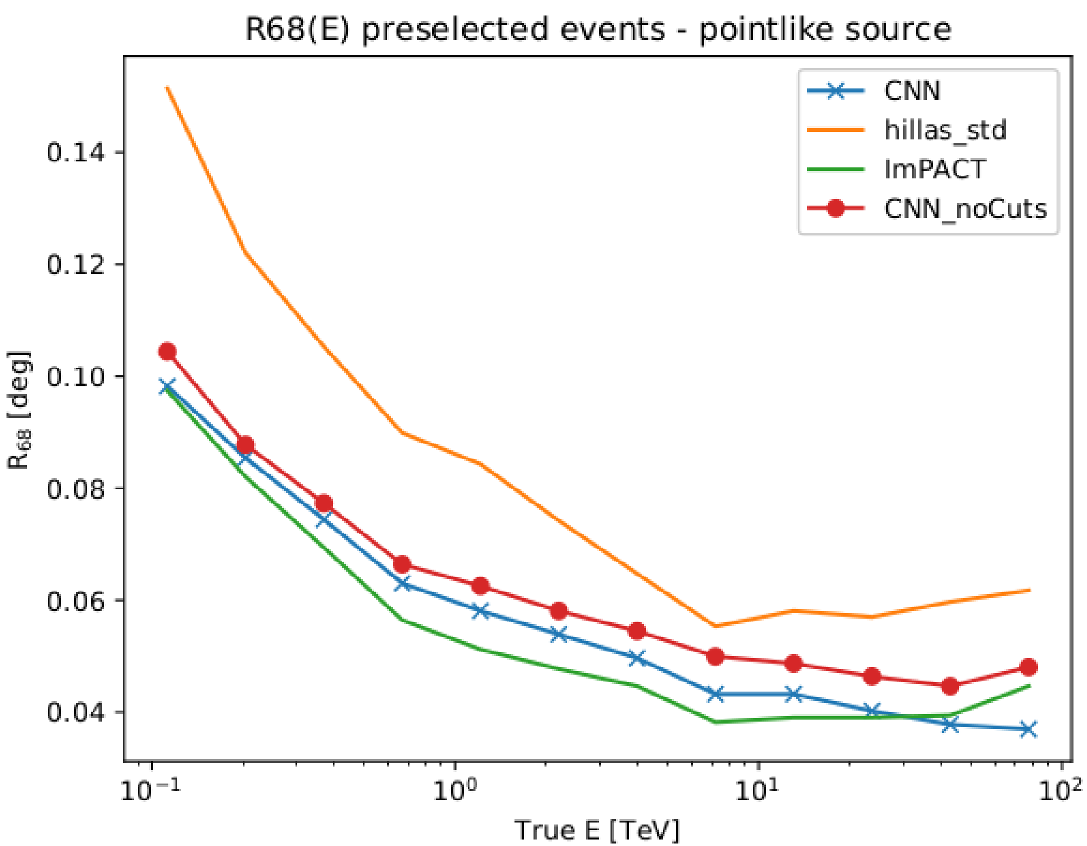


Figure 4: Angular resolution vs. true simulated energy at 20° zenith angle. The results of two CNN regressors are shown in comparison to the Hillas-based and ImPACT PSFs. The dotted curve refers to a regressor trained without applying pre-selection cuts to the training data, while the 'X'-decorated curve refers to a regressor trained on pre-selected events. All reconstructions are carried out on the same pre-selected benchmark set.

Paper results (Direction on data vs Hillas and ImPACT)

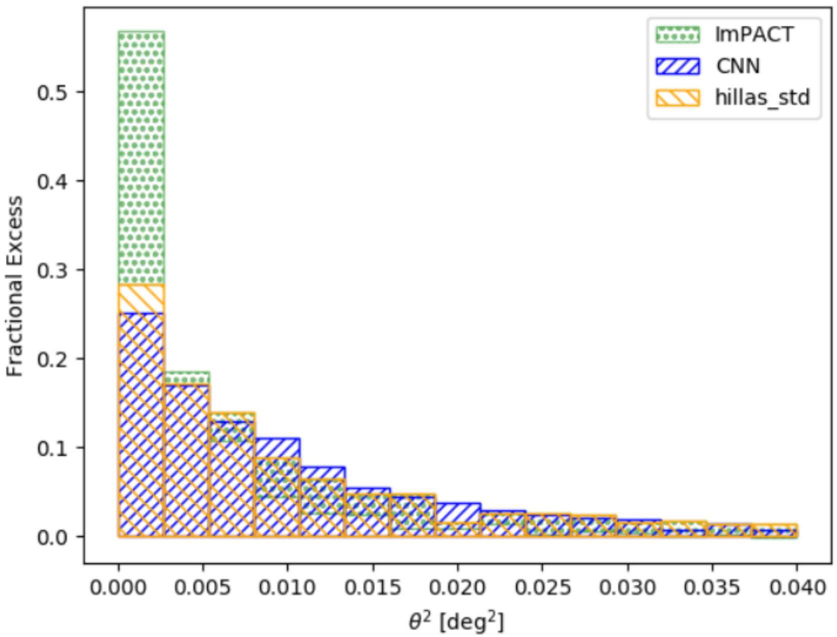


Figure 5: The squared angular distance θ^2 distribution for excess events from one flare observation of PKS 2155-304, using the Hillas, ImPACT and CNN direction reconstruction methods.

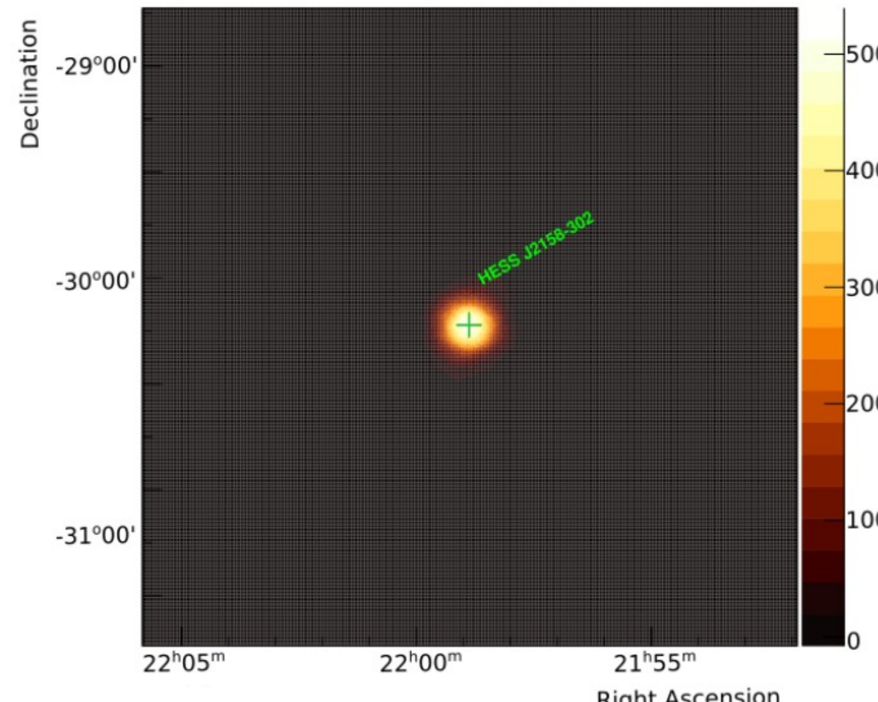


Figure 6: A two dimensional distribution of excess events observed in the direction of PKS 2155-304 for the one flare observation, using the CRNN classifier and CNN regressor.

Paper results (compressed)

- Classification works excellent on MC
- Classification works quite good on data (still better than TMVA)

- Regression works „okay-ish“ on MC
- Regression nothing to be prouf of on data

Open Question:
Why is the Monte Carlo performance **so** far from data performance?
And worse: not always correlated!

Example:
Classifyer 1 with architecture xy performs excellent on MCs
Classifyer 2 with architecture z performs well, but worse than 1 on the very same MC training-/validation- and test-sets

But on data:
2 is clearly better than 1 for no obvious reason…

(we just stumbled across it by chance)

This is a serious issue:

In order to calculate fluxes, one relies on MC based effective areas, which are affected by all three analysis tasks.
For example, the cut efficiency of the classifier in use directly affects the number of surviving signal events. Since a DL-based classifier acts differently on simulation versus observation data, the effective areas are not reliable when applied to observation events and the derived fluxes could be biased.

What we wanted to do:

- Investigate data-MC discrepancy and learn from it

  or even
- Circumvent the discrepancy without knowing the reason

  after that:
- Start working on hybrid analysis
- Start working on energy reconstruction
- Since Classification works best: try separating higher nuclei
- And of course: keep improving the full analysis chain

Investigation of MC-data discrepancy

- We trained a RNN Classifyer on:
  - 256k MC protons (Class label 0)

    vs.
  - 256k events from a PKS-2155 real data run (Class label 1)

  (yes ok, there are ~60 γ on-events in it, but who cares. It‘s basically only real protons)
- We tested that very Classifyer on:
  - MC protons vs. real protons (the validation for the training)
  - MC gamma vs. real protons
  - MC gamma vs. MC protons

Investigation of MC-data discrepancy

- Test on MC protons vs. real protons (the validation for the training)

  remember: training: MCP – realP testing: MCP – realP

Accuracy (how many correct classifications were made)	99.34%
Precision (how many Class 0 events are amongst the events classified as 0)	99.22%
Recall (how many Class 0 events were classified correctly)	99.46%
Specificity (how many Class 1 events were classified correctly)	99.22%

Ok.
So the proton simulations aren’t perfect…
We all knew that, right?!

But there is more...

Investigation of MC-data discrepancy

- Test on MC gammas vs. real protons

  remember: training: MCP – realP   testing: MCγ – realP

Accuracy (how many correct classifications were made)	99.48%
Precision (how many Class 0 events are amongst the events classified as 0)	99.23%
Recall (how many Class 0 events were classified correctly)	99.74%
Specificity (how many Class 1 events were classified correctly)	99.22%

The network never saw those during training!!!

Investigation of MC-data discrepancy

- Test on MC gammas vs. real protons

remember: training: MCP – realP testing: MCγ – realP

Accuracy (how many correct classifications were made)	99.48%
Precision (how many Class 0 events are amongst the events classified as 0)	99.23%
Recall (how many Class 0 events were classified correctly)	99.74%
Specificity (how many Class 1 events were classified correctly)	99.22%

The network never saw those during training!!!

Investigation of MC-data discrepancy

- Test on MC gammas vs. real protons

  remember: training: MCP – realP testing: MCγ – realP

This network learned perfectly how Monte Carlo Gammas look different to real protons by only learning from Protons.

→ There has to be a particle type independent “Monte Carlo Feature”

A	99.48%
P[illegible]	99.23%
Recall (how many Class 0 events were classified correctly)	99.74%
Specificity (how many Class 1 events were classified correctly)	99.22%

Investigation of MC-data discrepancy

- Test on MC gammas vs. MC protons

  remember: training: MCP – realP testing: MCγ – MCP

Accuracy (how many correct classifications were made)	50.11%
Precision (how many Class 0 events are amongst the events classified as 0)	50.06%
Recall (how many Class 0 events were classified correctly)	99.74%
Specificity (how many Class 1 events were classified correctly)	0.48%

H.E.S.S.

Investigation of MC-data discrepancy

Everything is Simulation, Nothing data

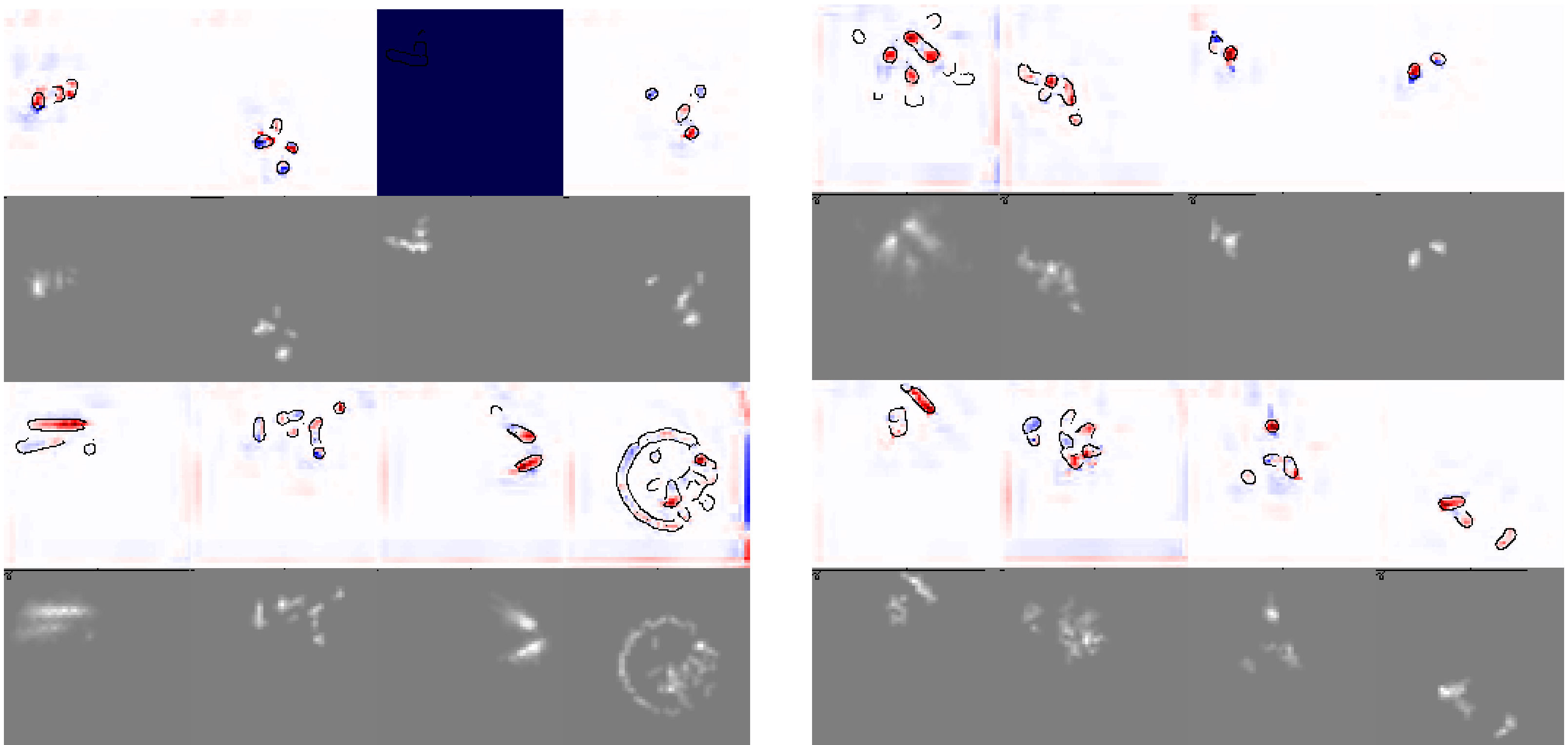
- Test on MC gammas vs. MC protons

remember: training: MCP – realP testing: MCγ – MCP

Accuracy (how many correct classifications were made)	50.11%
Precision (how many Class 0 events are amongst the events classified as 0)	50.06%
Recall (how many Class 0 events were classified correctly)	99.74%
Specificity (how many Class 1 events were classified correctly)	0.48%

Investigation of MC-data discrepancy

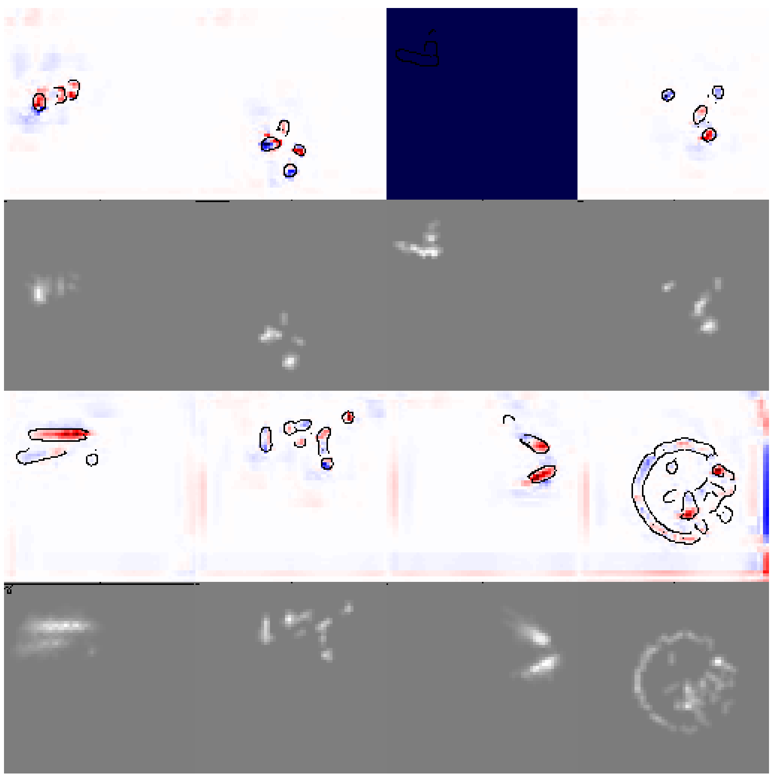
- Work in Progress: Layerwise Relevance Propagation (LRP*)

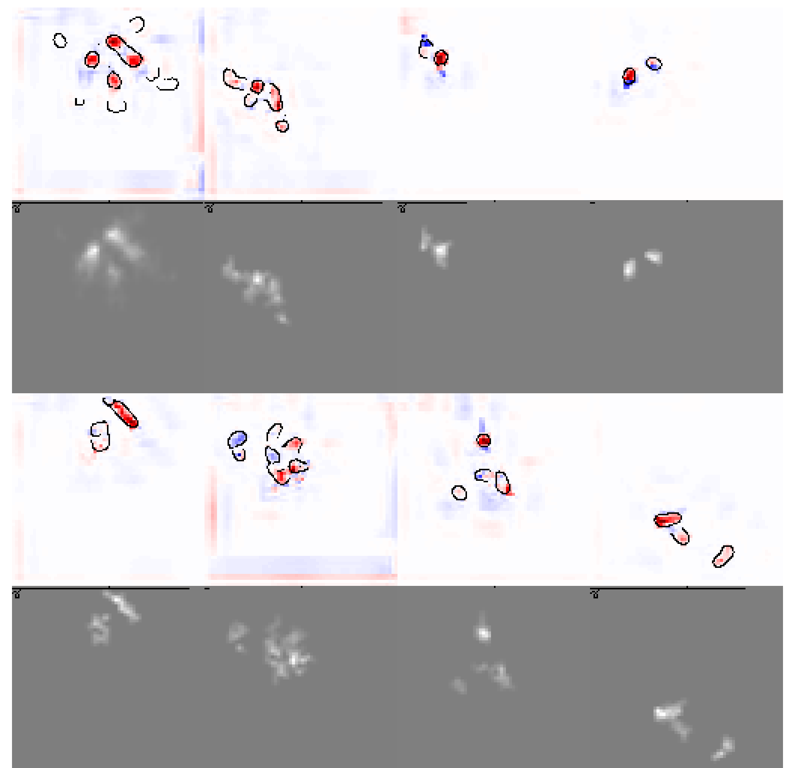


* https://www.itu.int/en/journal/001/Pages/05.aspx

Investigation of MC-data discrepancy

- Work in Progress: Layerwise Relevance Propagation (LRP*)
- These are all protons, but can you tell what is simulated and data?

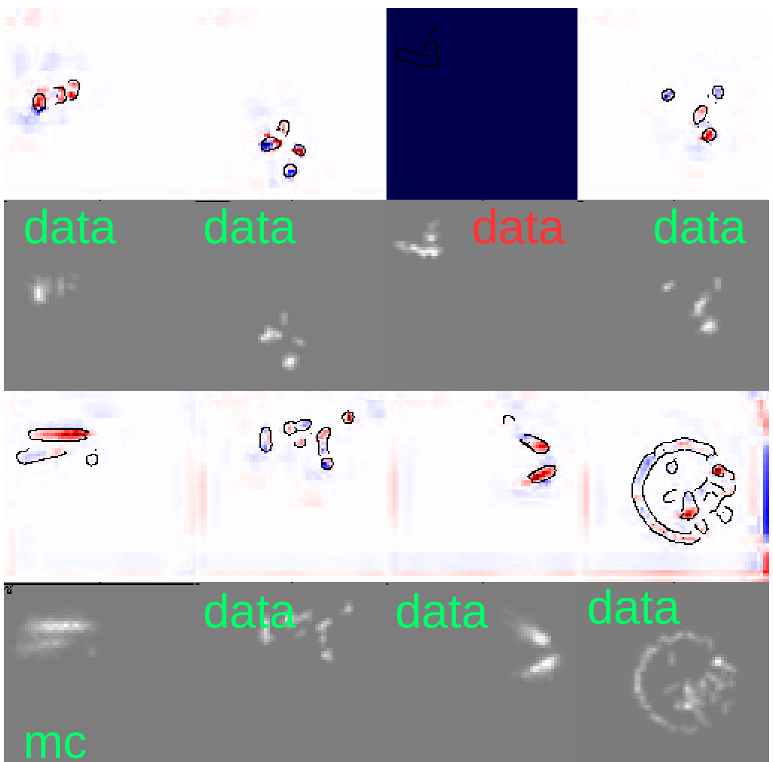


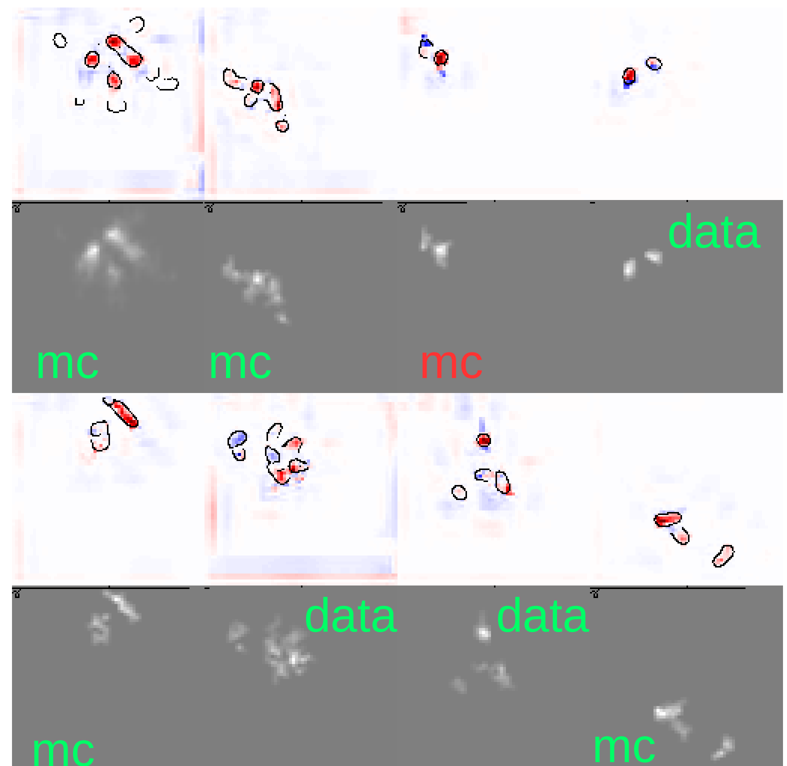


* https://www.itu.int/en/journal/001/Pages/05.aspx

Investigation of MC-data discrepancy

- Work in Progress: Layerwise Relevance Propagation (LRP*)
- These are all protons, but can you tell what is simulated and data?

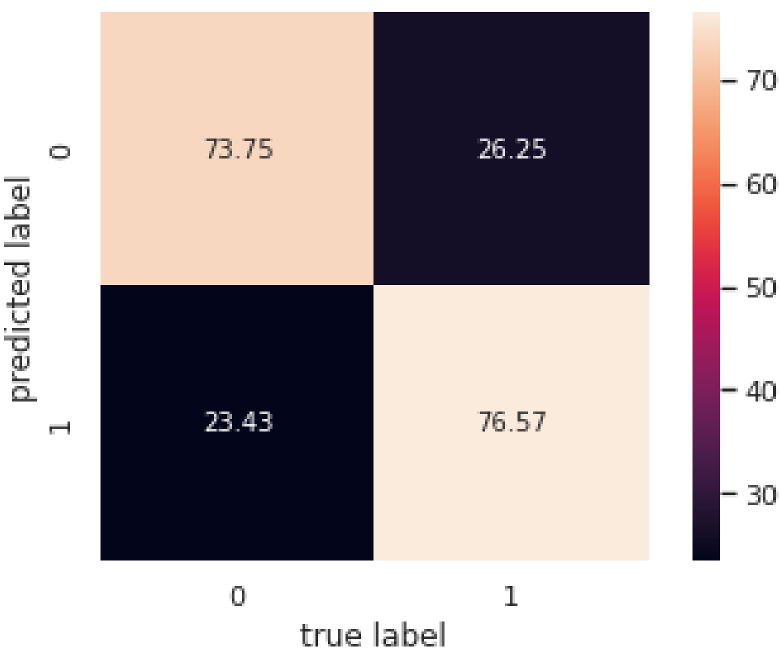




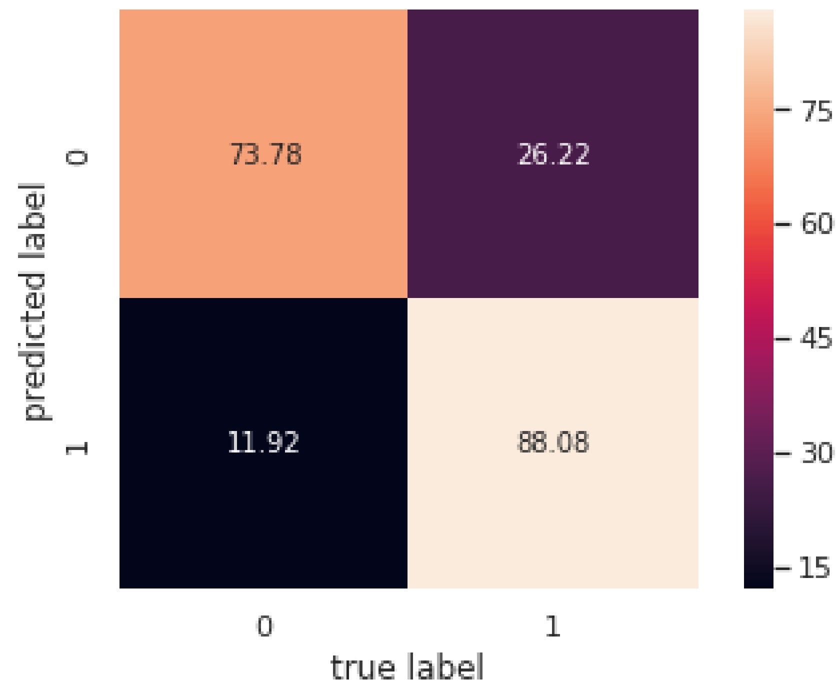
* https://www.itu.int/en/journal/001/Pages/05.aspx

Circumventing the discrepancy

- Using Auto-Encoder:
  - Results so far: okay-ish.



Boosted Decision Tree



Multi Layer Perceptron

Conclusions

- Task Group made (good) progress
- 1 Paper accepted, 1 more in pipeline
- Deep Learning Techniques can improve performace (Classification)
- MC – data discrepancy is the showstopper for further, deeper improvement of direction/energy at the moment

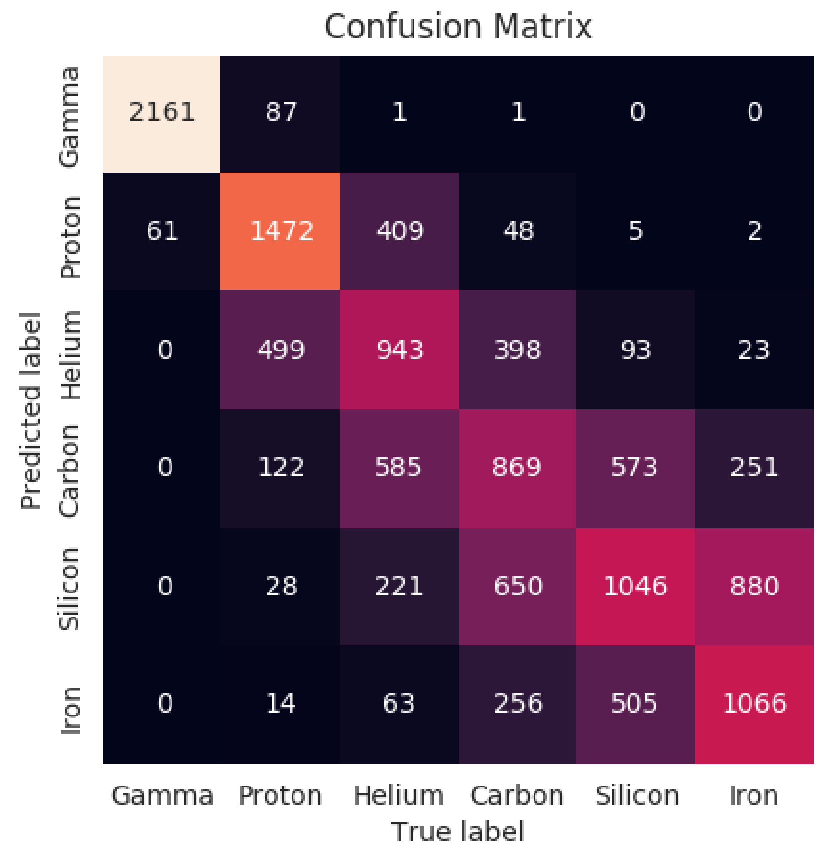
Outlook

- Solve the issues
- Do hybrid, energy, truncated events and heavy nuclei
- Accept the extraordinary capability of the standard analysis and help to improve it by only looking at events that do not pass regular cuts

NOT MY WORK, but still worth being mentioned :)

- Classification of heavy nuclei:

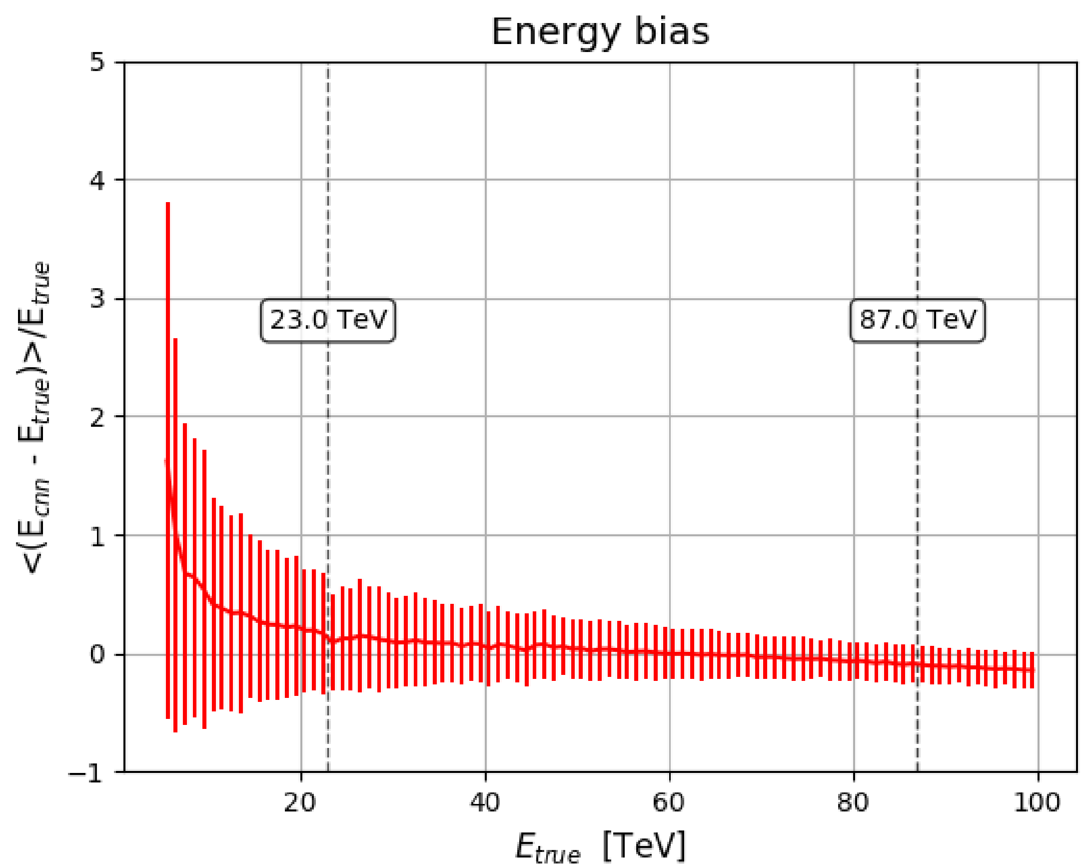
  (Christina)



NOT MY WORK, but still worth being mentioned :)

- Reconstruction heavy nuclei:

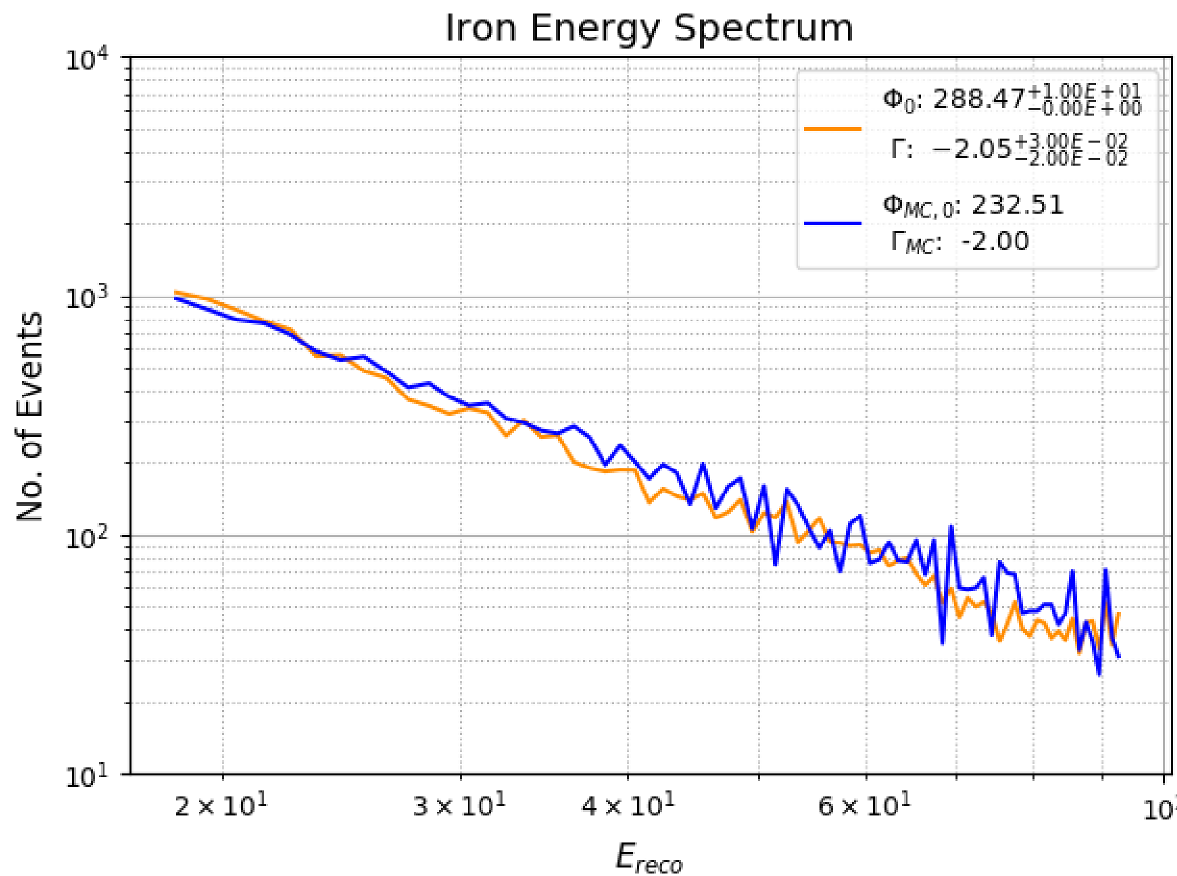
  (Christina)



NOT MY WORK, but still worth being mentioned :)

- Reconstruction heavy nuclei:

  (Christina)



Conclusions (again)

- Deep Learning is a interesing „tool“ for analyzing IACT data

- But: Sophisticated „standard“ analysis chains are hard to outperform

  → Don‘t try to beat them on on their home-base, but rather focus on regimes where standard analysis has no chance at all

  (truncated images, cosmic rays)

Thanks a lot.

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG