Graph Neural Networks for Event Classification in ARCA

G. Vermarien, A. Domi

30/4/2021

RDF and CNN

RDF and CNN are both **well performing**, however, there is still space for optimisations:

1. RDF: robust, but it needs a very good knowledge of the variables to consider during the classification -> time consuming from the “human” side and carries the risk of not finding the best variables to exploit the full potential of the detector.

2. CNN: do not require input variables from reconstructions but use rigid pre-chosen spatial and temporal bins to model the detector. This means they should be trained also with MC simulations accounting for the physical movement of the DUs driven by sea currents (a huge amount of MC simulations needed).

Why focus on GNNs?

ADVANTAGES

- wrt RDF: Deep Learning methods such as GNNs do not require reconstruction information -> only hits information needed.
- wrt CNN: GNNs do **not** need rigid pre-chosen spatial and temporal bins to model the detector. They only rely on hits relations. -> **Potentially** more suited for a spatially dynamic, moving and rotating detector such as KM3NeT (to be tested and quantified!)

Classification in ARCA using GNNs

GOAL

Offline and Online Classification of:

1. Tau neutrinos. (**Gjis**)
2. Track vs shower.
3. Neutrino vs atmospheric muon.

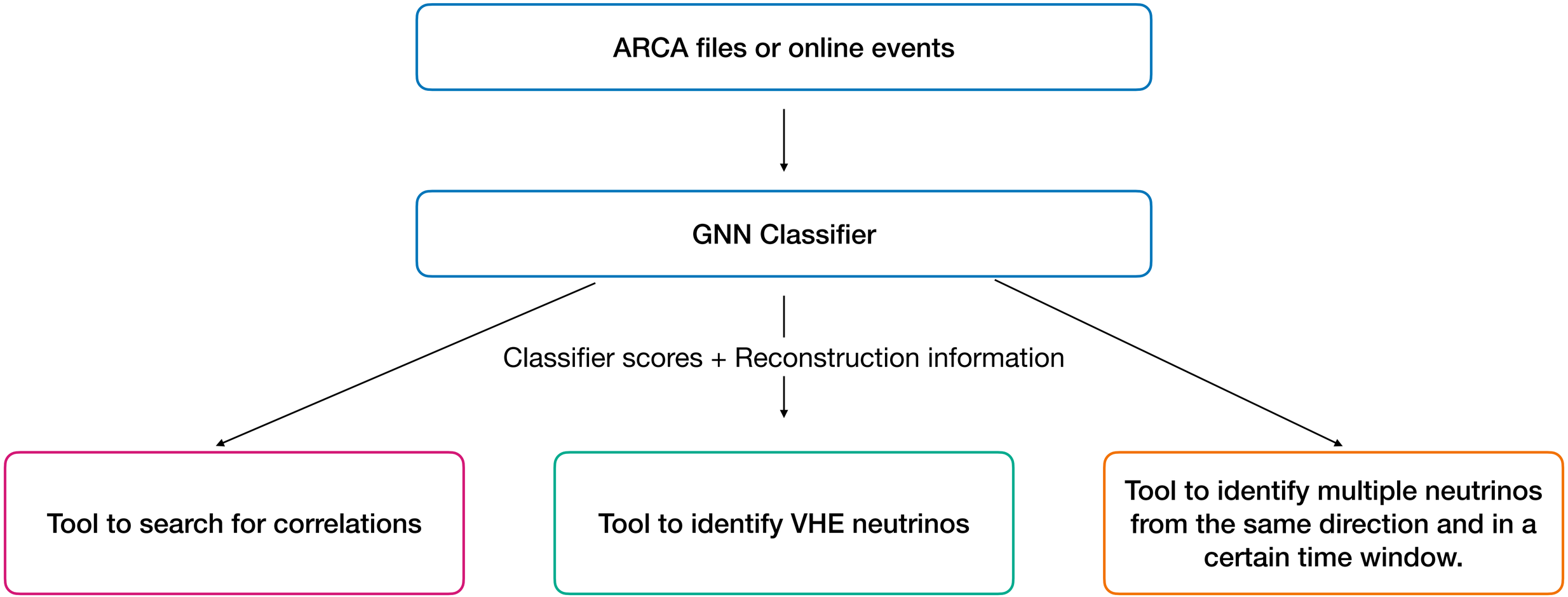
a) With a “good” neutrino sample we can look for correlations.

b) identify subsample of VHE neutrinos that can be used in the **follow-up** (expected rate: 2-3 events per month)

c) another goal for the **follow-up** is to identify multiple neutrinos coming from the same direction and in a given time window (for whatever energy or class). The time window can be few seconds in case of fast transients, up to few days in case of AGN flares.

Classification in ARCA using GNNs

APPROACH



GNN Classifier

TRAINING

- ARCA MC Files for training:

https://wiki.km3net.de/index.php/HE_Astrophysics/ARCA_High_Energy_Analysis#MC_production_v5

ARCA MC Files are the input of **OrcaSong** (https://ml.pages.km3net.de/OrcaSong/): preprocesses raw KM3NeT data for the use with deep neural networks, based on km3pipe.

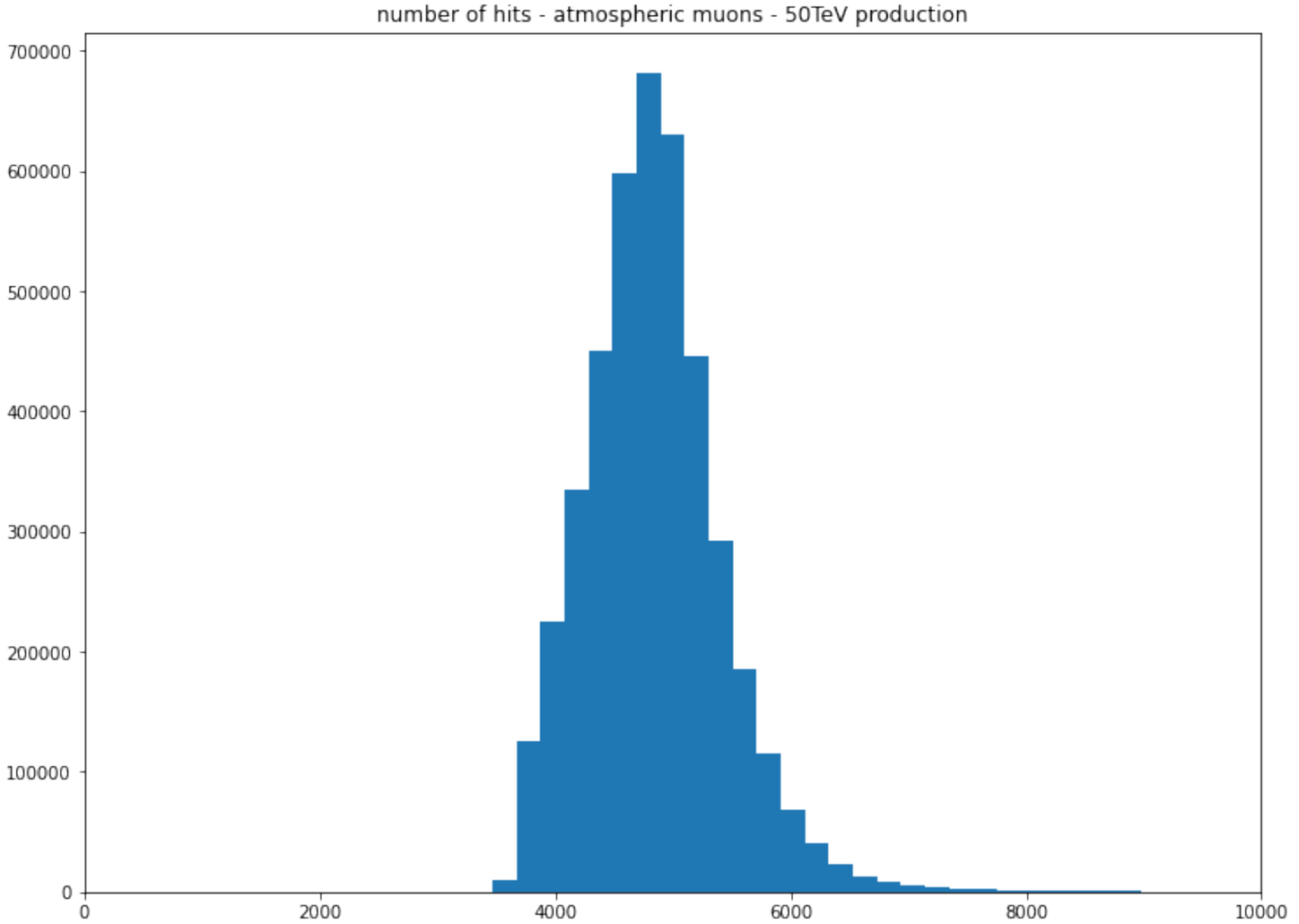
Number of nodes

- OrcaSong considers each hit as a node.
- When converting h5 files into ML format, we should give a maximum number of nodes -> reasonable maximum number of hits.
- Choose a reasonable value of maximum_nodes in ARCA:
  - In GNNs, operations scale as $nodes^2$.
  - Therefore, using the maximum_hits is not feasible.
  - Try to use as maximum_nodes the average number of hits of atmospheric muons.

Number of Nodes

Atmospheric Muons: 50 TeV production

Mean value = 4846 hits



I have chosen max_hits = 5000

GNN Performances

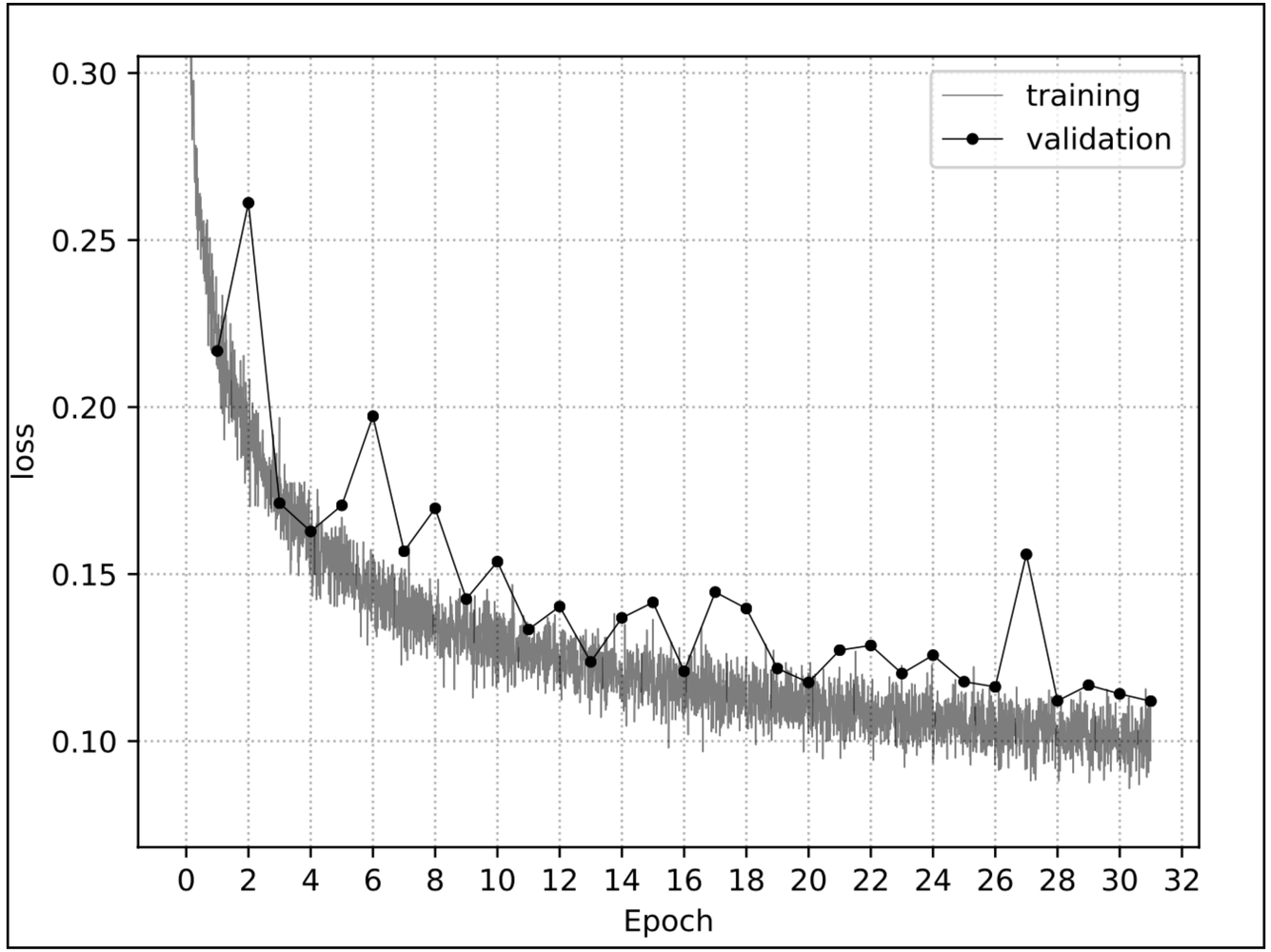
Training & Validation speed

- 1 epoch of training with 800k events (86GB file) takes **7 hours.**
- Validation time: ~30 min for 127k events -> **0.01s per event.**

⟶ **fast! no issues for online applications.**

GNN Performances - Training and Validation

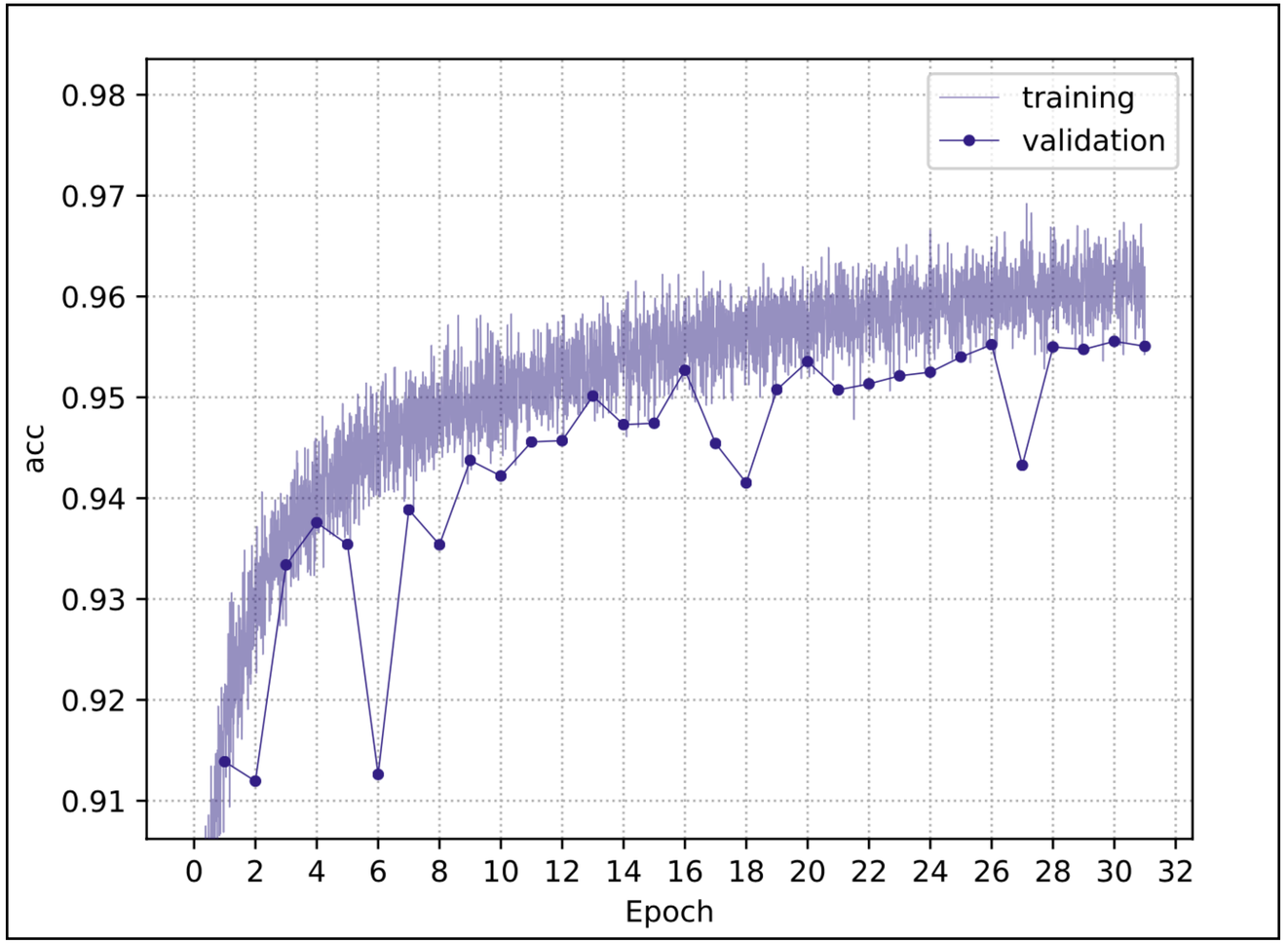
Neutrinos vs Atmospheric muons Classification



Loss function = unhappiness of the classification outcome.

GNN Performances - Training and Validation

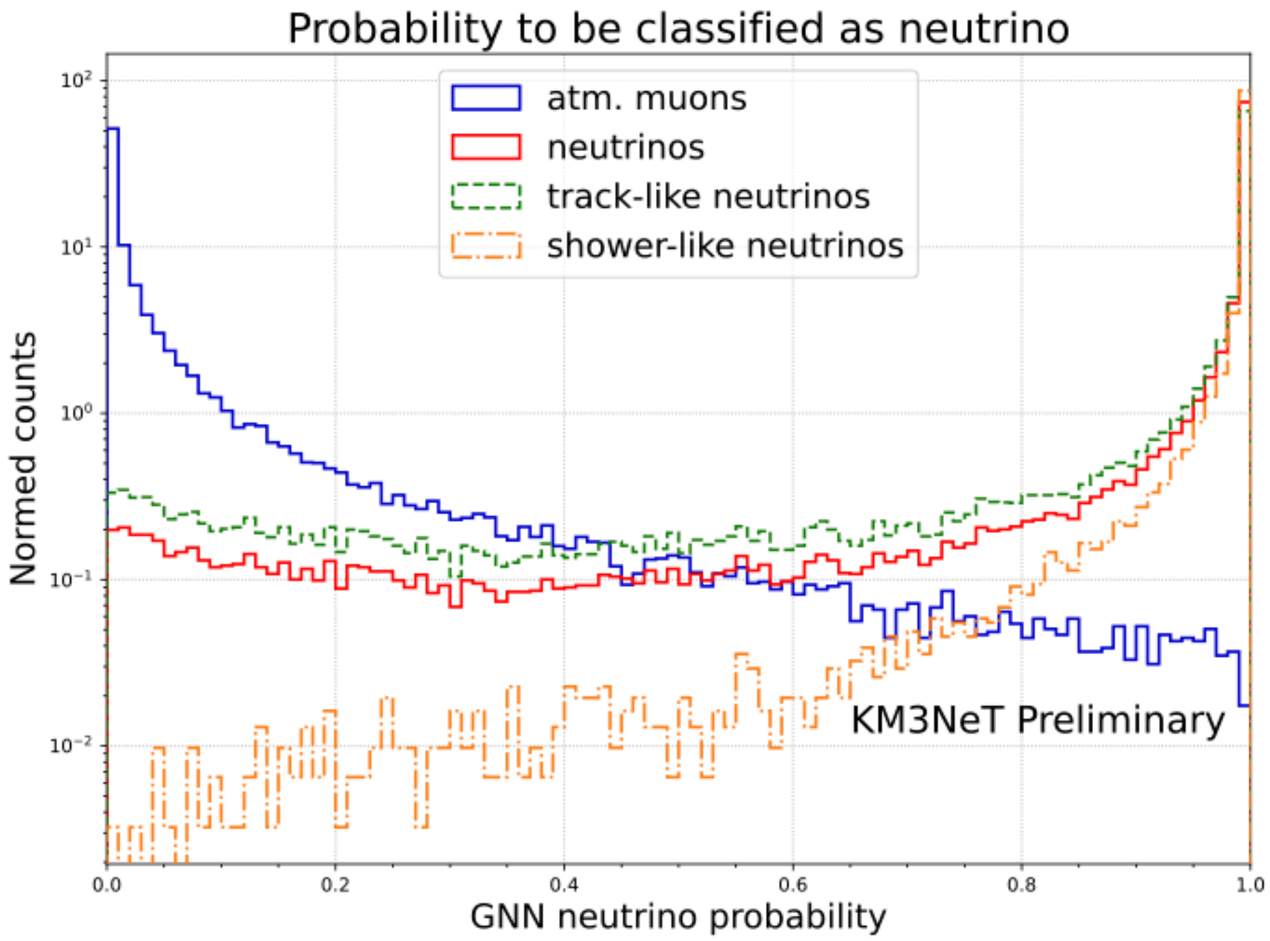
Neutrinos vs Atmospheric muons Classification



Accuracy = fraction of predictions that were correct.

GNN Performances

Neutrinos vs Atmospheric muons Classification

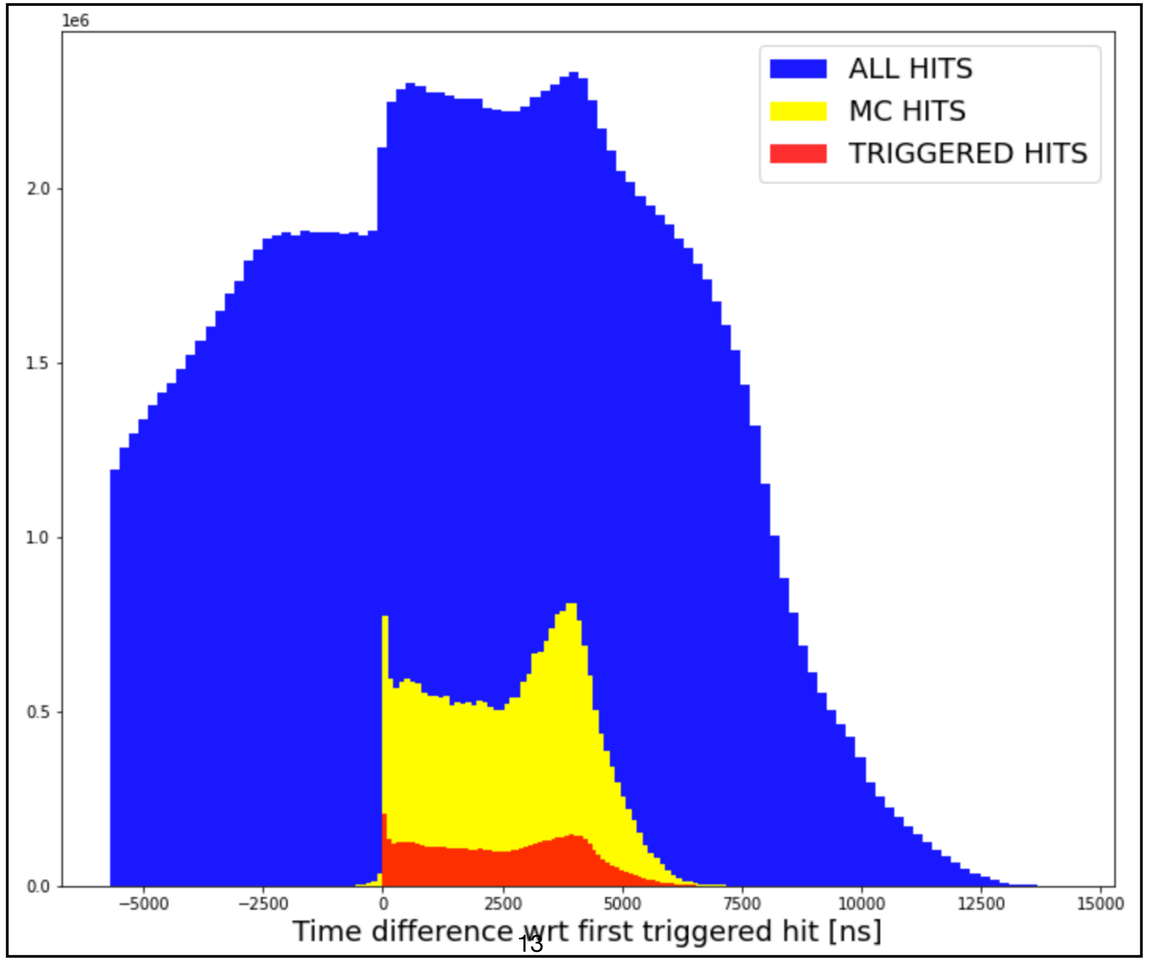


New ML files

- To reduce the disk space needed for ML files and to reduce the training time, we do not use anymore ALL hits but we select hits within a certain time window from the first triggered hit.

Chosen time window:

[-1000, +7500] ns

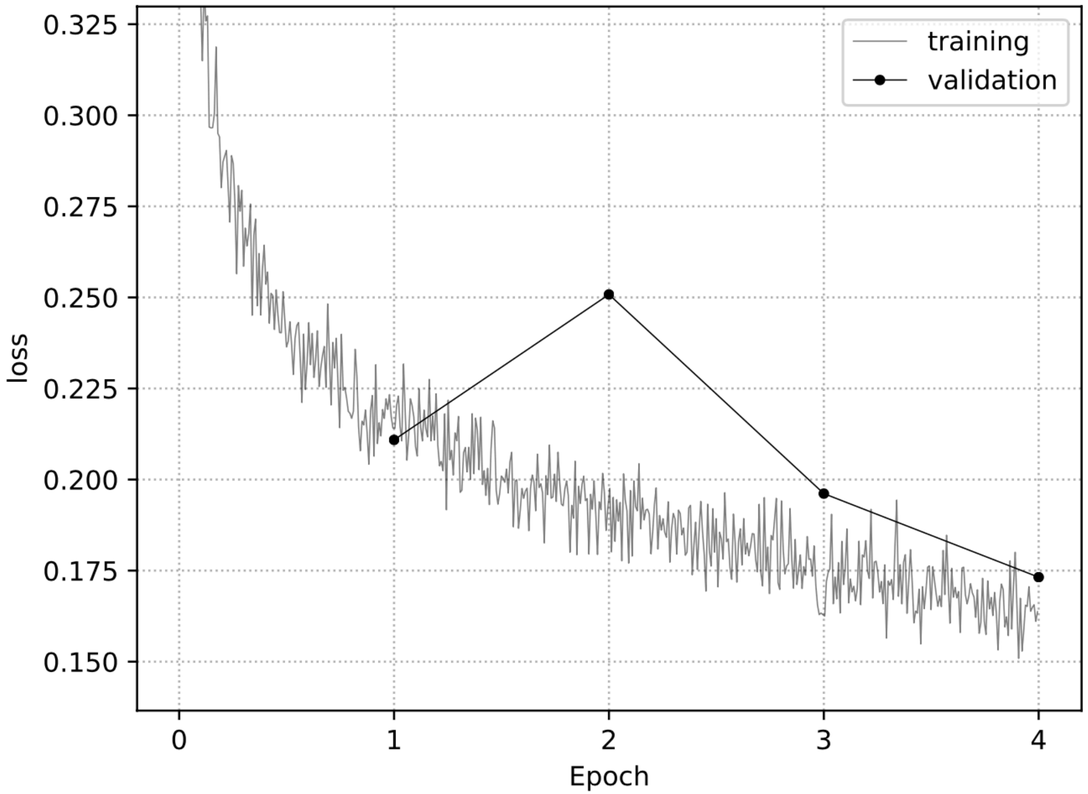


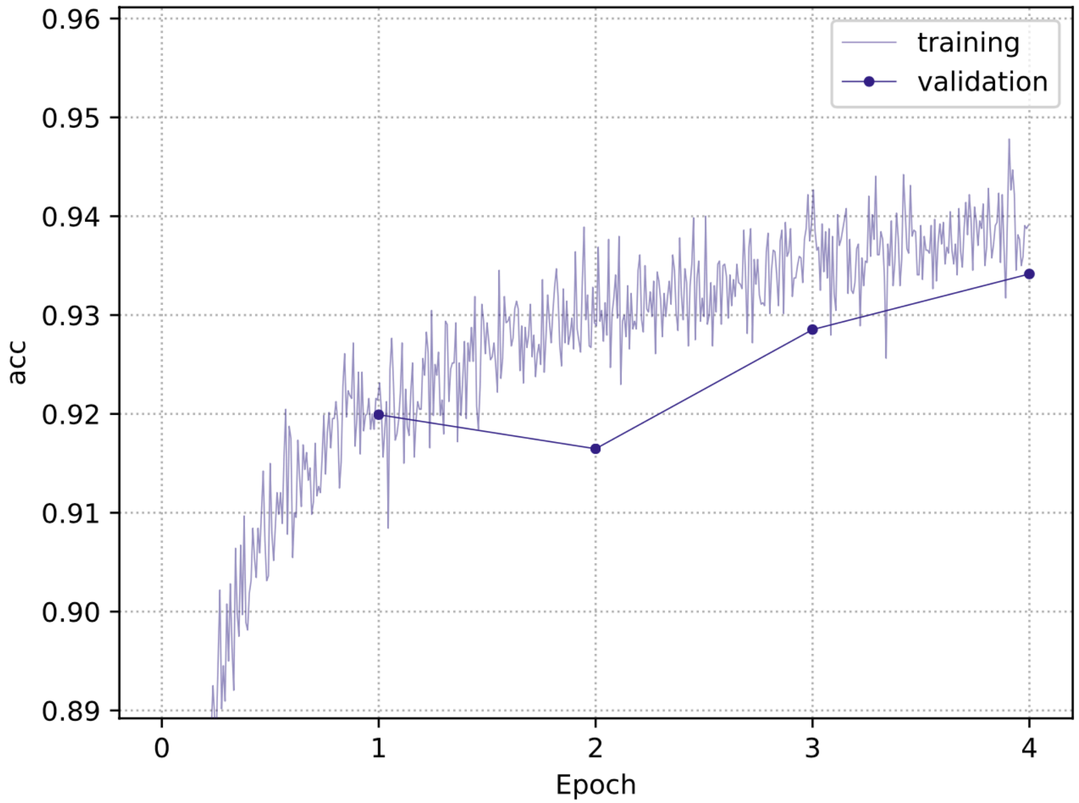
Factor 1.5 reduction in disk space.

GNN Training on FULL-ARCA v5 MC files

- I have re-started the training of the GNN with the goal “atmospheric muon vs neutrino” selection with 800k events (half muons, half neutrinos - all flavours) with the new ML files to see how the time window impacts the performances.
- New training file -> 58 GB (instead of 86 GB).
- The training is ongoing….

GNN Training on new ARCA v5 files



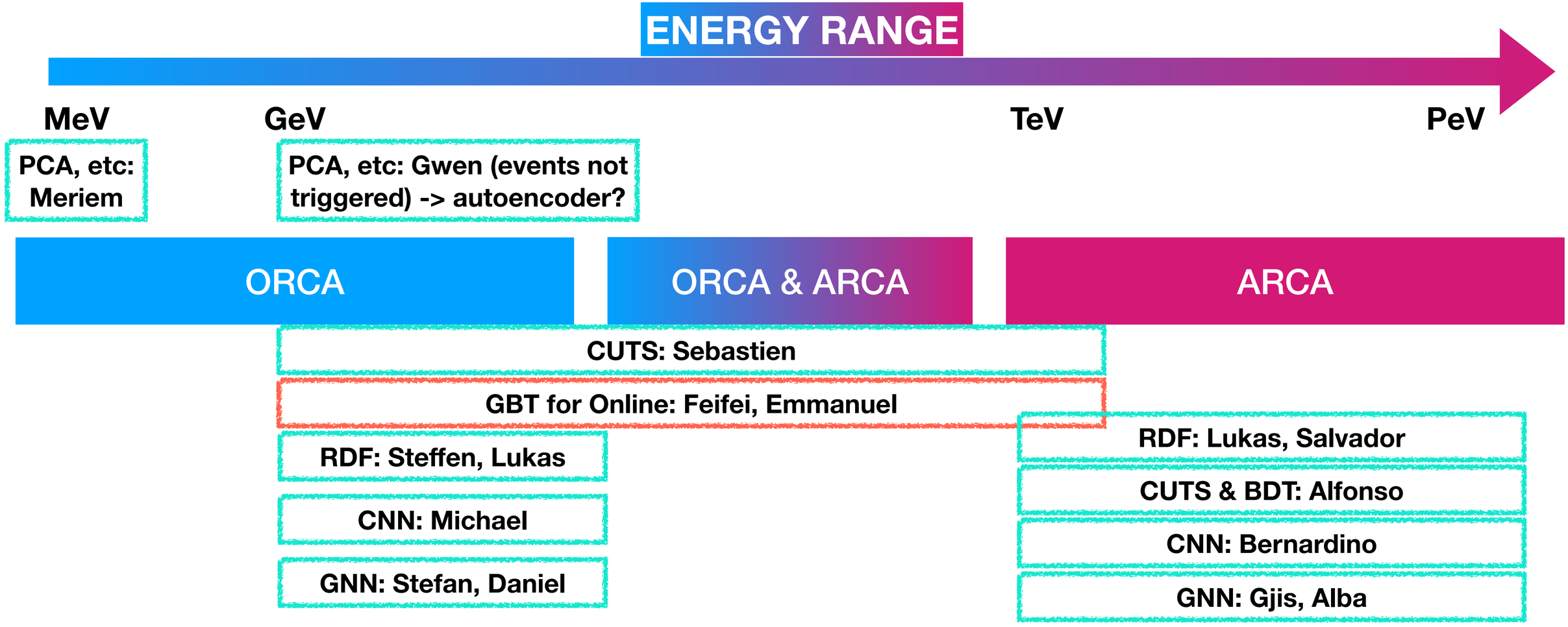


Coherent results as with the previous training (as expected).

Classification Benchmarking

GOAL

- Several Classification efforts within the collaboration:



Classification Benchmarking

GOAL

- Several Classification efforts within the collaboration.
- We should check which algorithms are more suited for a class of analyses in a specific energy range and for a specific detector configuration (full ORCA/ARCA and ORCA/ARCA-N DUs).
- For this, we need to have a fair comparison of the performances, which can be easily checked and reproducible for future developements.
- The goal of the classification benchmarking project is to create a set of official plots, that can be presented also at conferences/papers, with fair comparisons between algorithms and easy to crosscheck by the collaboration.

Classification Benchmarking

HOW

- New git project:

https://git.km3net.de/adomi/classification-benchmarking

- Steps:

1. Define a set of benchmark files, for each MC production, where to run the (already trained) classification algorithms and put them in Lyon:/sps/km3net/… .

2. Save the classification scores in simple csv files and include them in the git project.

3. Define which plots are useful for comparisons and for being shown at conferences and include them in jupyter notebooks (easy to crosscheck by the collaboration).

To Do

- Use only L1 or triggered hits -> further reduce disk space but possible loss of information.
- Start track/shower classification.
- Comparison of all classification algorithms.

Open tasks

- Quantify the impact of DUs movement in the classification score.
- Test other variables for nodes connections (ToT).
- Muon bundles.

Updates in view of VLVNT

- Talk contribution with title “**Graph Neural Networks for reconstruction and classification in KM3NeT**”.
- General contribution: ORCA (Stefan Reck & Daniel Guderian) + ARCA (Gjis Vermarien & Alba Domi).
- Today we will focus on the ARCA side.
- I re-converted all ARCA v5 files to allow to retrieve also aashower information (not needed for training but needed for making final plots - comparisons with other methods).