Graph Neural Networks for Photon Searches with the Underground Muon Detector of the Pierre Auger Observatory

ITeDA

KIT Karlsruher Institut für Technologie

OBSERVATORIO PIERRE AUGER

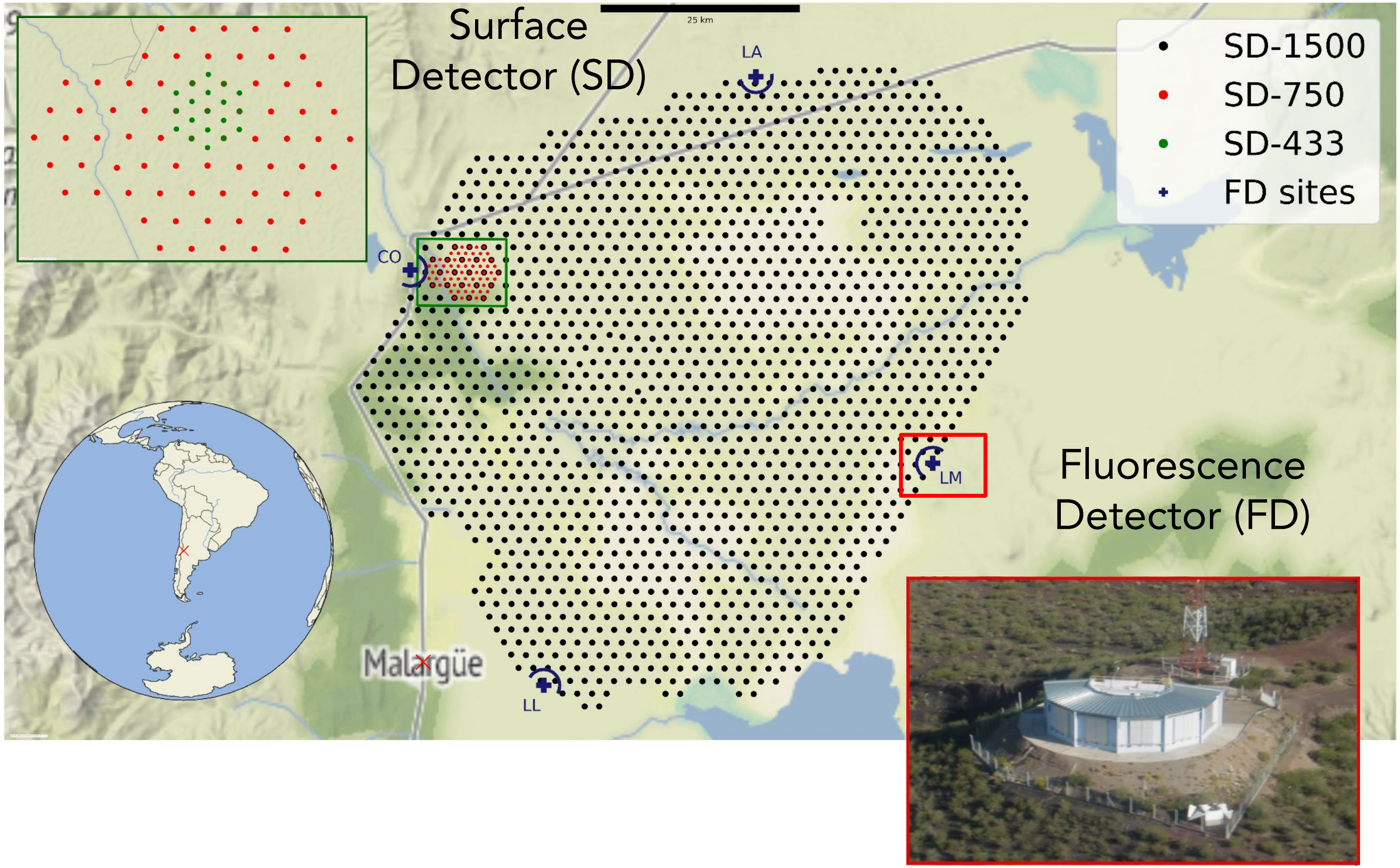
Ezequiel Rodriguez[1,2] for the Pierre Auger Collaboration[3]

1. Instituto de Tecnologías en Detección y Astropartículas, San Martín, Argentina
2. Karlsruher Institut für Technologie, Karlsruhe, Germany
3. Observatorio Pierre Auger, Malargüe, Argentina

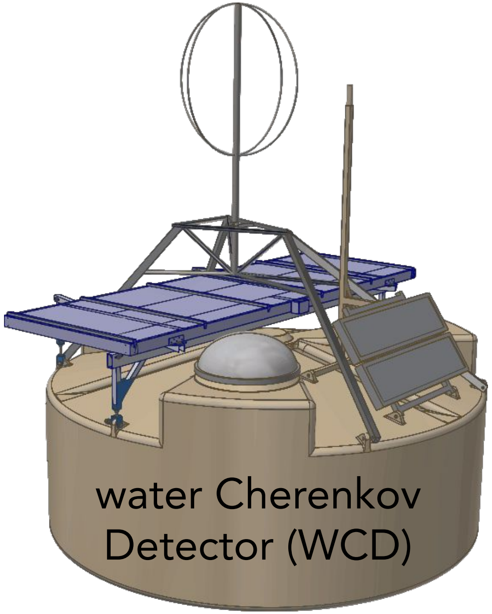
January 30, 2025

Workshop on Machine Learning for Analysis of High-Energy Cosmic Particles

The Pierre Auger Observatory



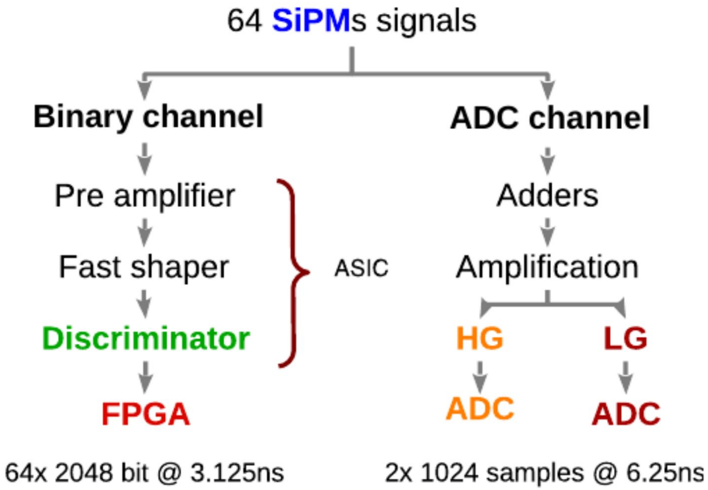
Upgraded SD station
(AugerPrime)

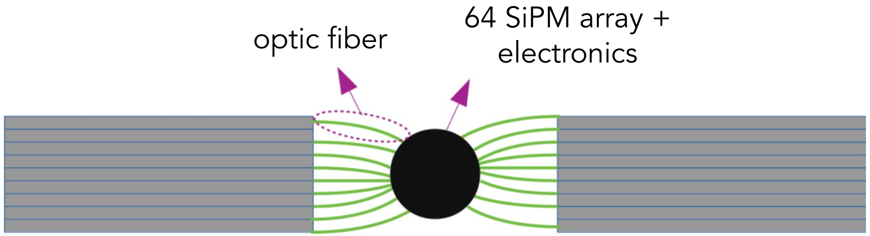


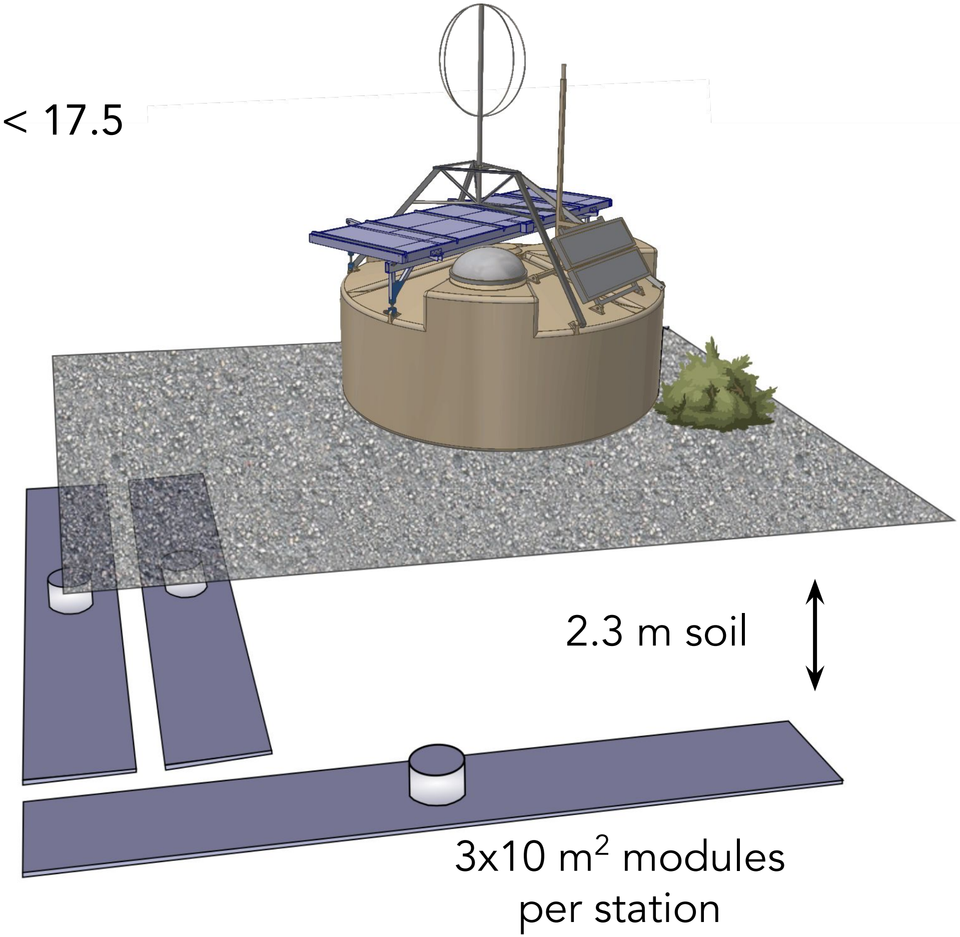
Each WCD has 3 Large
Photomultiplier Tubes (PMTs)

Underground Muon Detector (UMD)

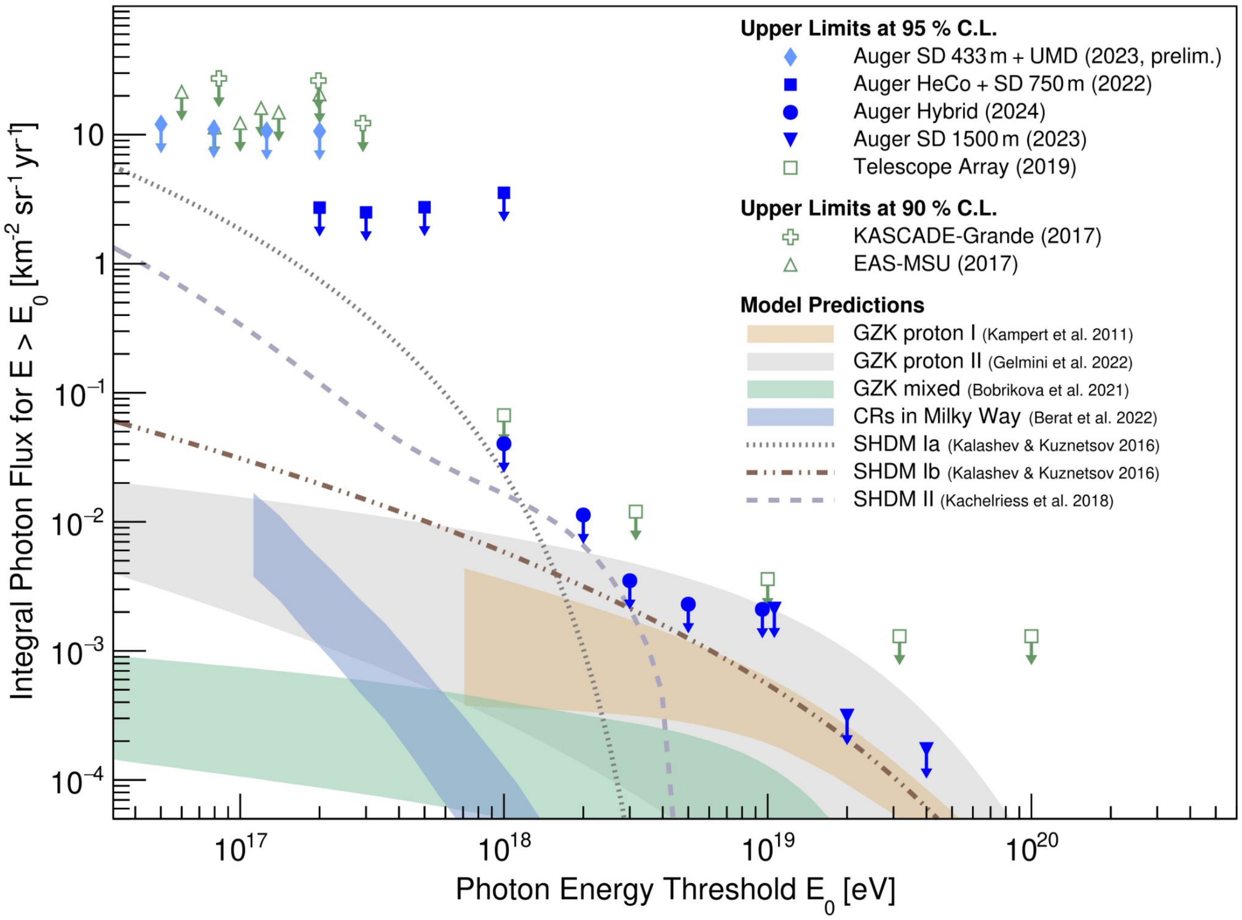
- Coupled to SD-750 and SD-433
- The energy range for this work is 16.5 < lg(E/eV) < 17.5







Photon searches at Auger



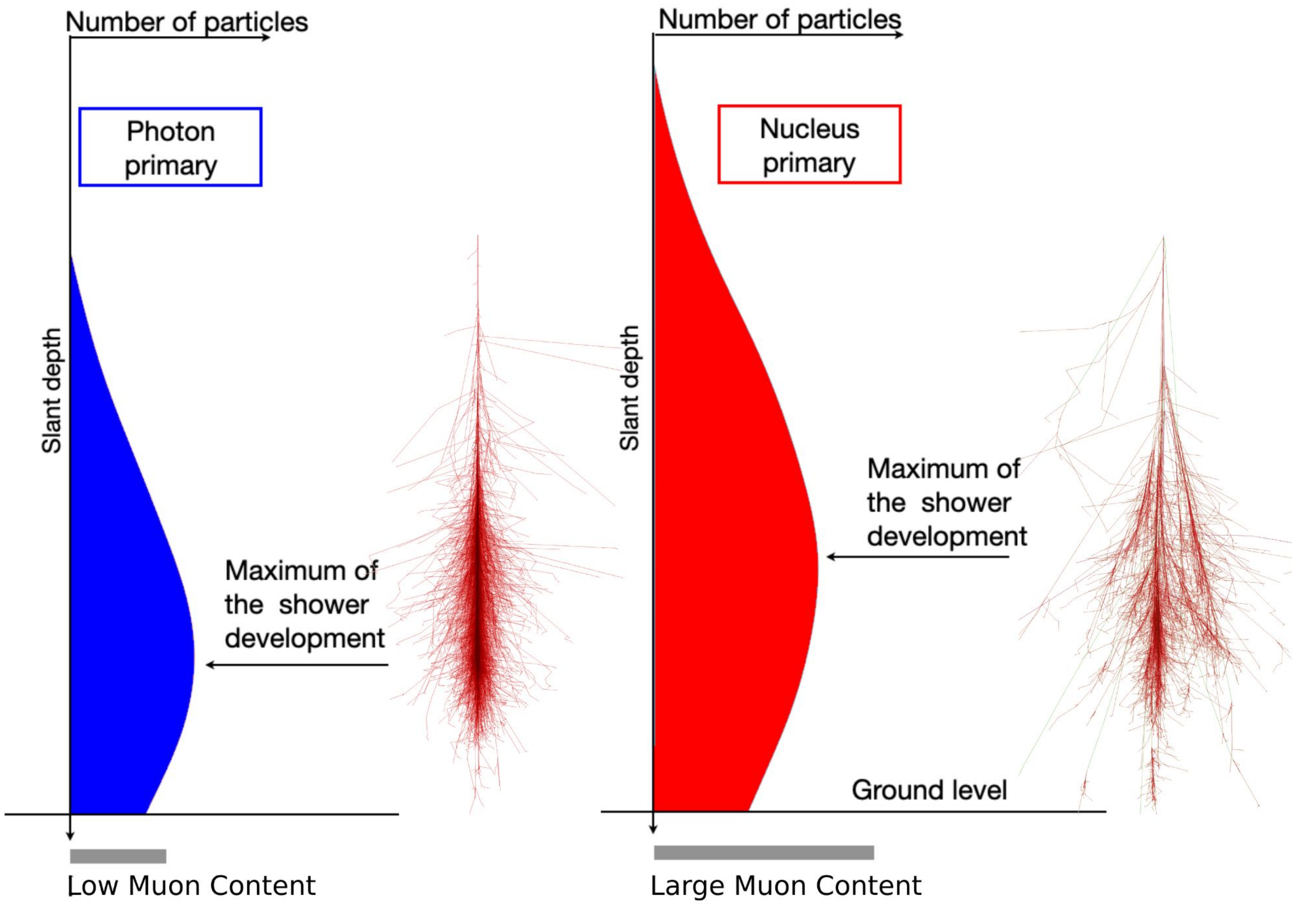
Produced in cosmic-ray acceleration sites (astrophysical fluxes), during cosmic-ray propagation (cosmogenic fluxes) or in the decay of dark matter particles, photons trace the local Universe.

All previous searches were conducted with traditional statistical methods.

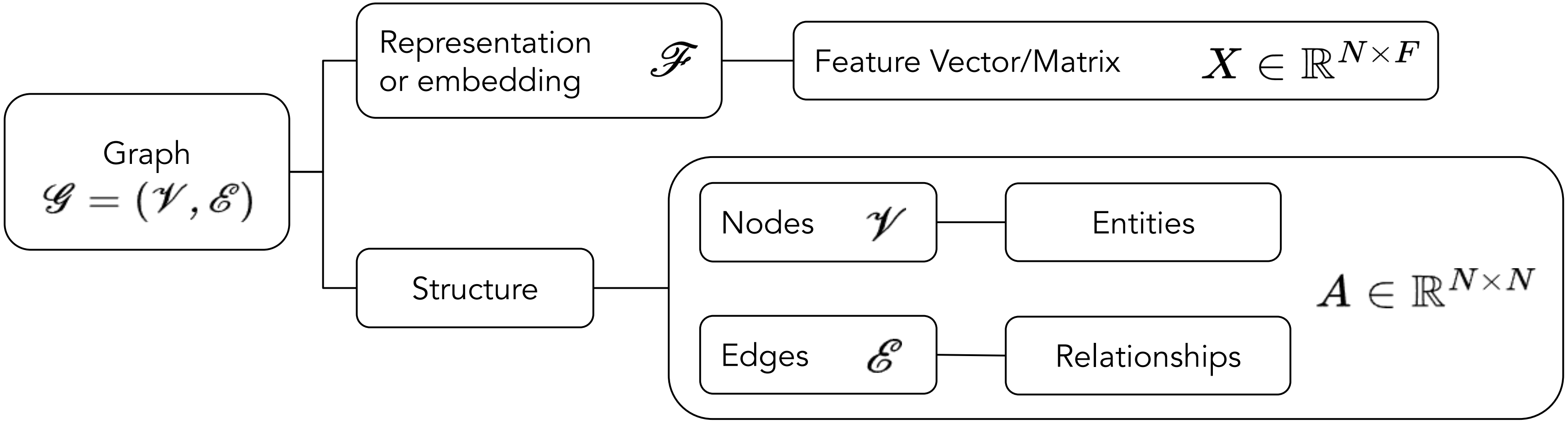
Photon-induced air showers

- Photon-Induced Showers:
  Low muon content, develop deeper in the atmosphere.

- Hadron-Induced Showers:
  Large muon content, develop shallower in the atmosphere.

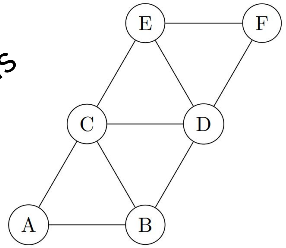


What are graphs in the context of Deep Learning?



Toy example:

1st neighbour connections



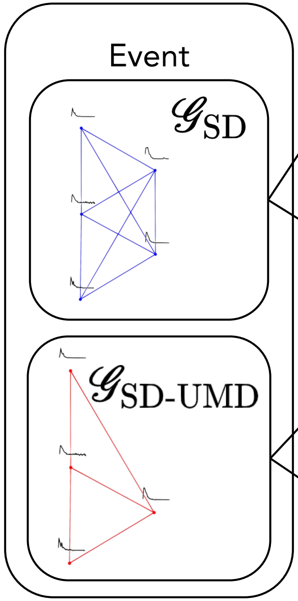
$\mathbf{A} =$

	A	B	C	D	E	F
A	0	1	1	0	0	0
B	1	0	1	1	0	0
C	1	1	0	1	1	0
D	0	1	1	0	1	1
E	0	0	1	1	0	1
F	0	0	0	1	1	0

$\mathbf{X} =$

Node	x	y	$\lg(S)$
A	-216.5	-374.98	-0.05
B	216.5	-374.98	-0.08
C	0	0	2.2
D	216.5	0	0.77
E	433	374.98	0.85
F	807.98	374.98	0.29

Event representation



$X_{\text{SD}} =$

Node	$\Delta x_{\text{hottest}}$	$\Delta y_{\text{hottest}}$	$\Delta z_{\text{hottest}}$	$\Delta t_{\text{hottest}}$	n_{PMTs}
$\text{SD}_{\text{hottest}}$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

up to 2nd neighbour connections

$X_{\text{traces}} =$

Node	t_1	t_2	$\cdots$	$t_{n_{\text{t-bins}}}$
$\text{SD}_{\text{hottest}}$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

More event representations in the backup slides!

$X_{\text{SD-UMD}} =$

Node	$\Delta x_{\text{hottest}}$	$\Delta y_{\text{hottest}}$	$\Delta z_{\text{hottest}}$	$\Delta t_{\text{hottest}}$	n_{PMTs}	ρ_μ	A_{eff}
$\text{SD}_{\text{hottest}}$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

- Some stations might be missing information from the UMD
- We need 2 graphs to account for missing values in UMD features
- We don't want imputation!

X_{SD} : 5 rows, 5 cols

$X_{\text{SD-UMD}}$: 4 rows, 7 cols

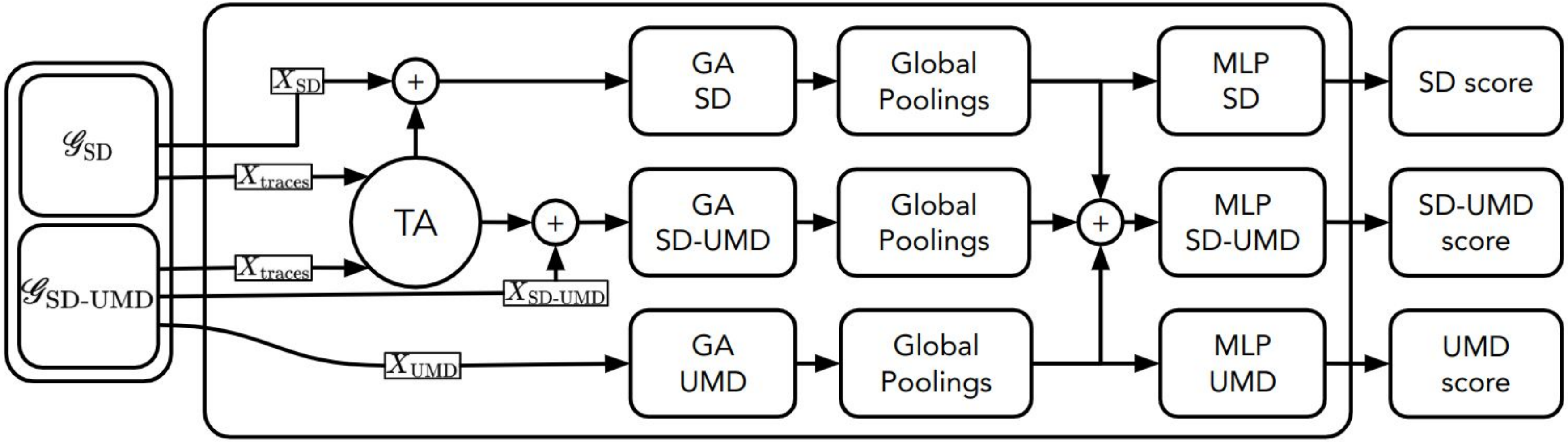
X_{traces} : 5 rows, $n_{\text{t-bins}}$ cols for $\mathscr{G}_{\text{SD}}$

X_{traces} : 4 rows, $n_{\text{t-bins}}$ cols for $\mathscr{G}_{\text{SD-MD}}$

Redundant information.

$$\mathscr{V}_{\text{SD-UMD}} \subseteq \mathscr{V}_{\text{SD}}$$
$$\mathscr{E}_{\text{SD-UMD}} \subseteq \mathscr{E}_{\text{SD}}$$
$$\mathscr{F}_{\text{SD}} \subseteq \mathscr{F}_{\text{SD-UMD}}$$

The network



TA

Trace Analyzer
3 layers of 1D convolutions

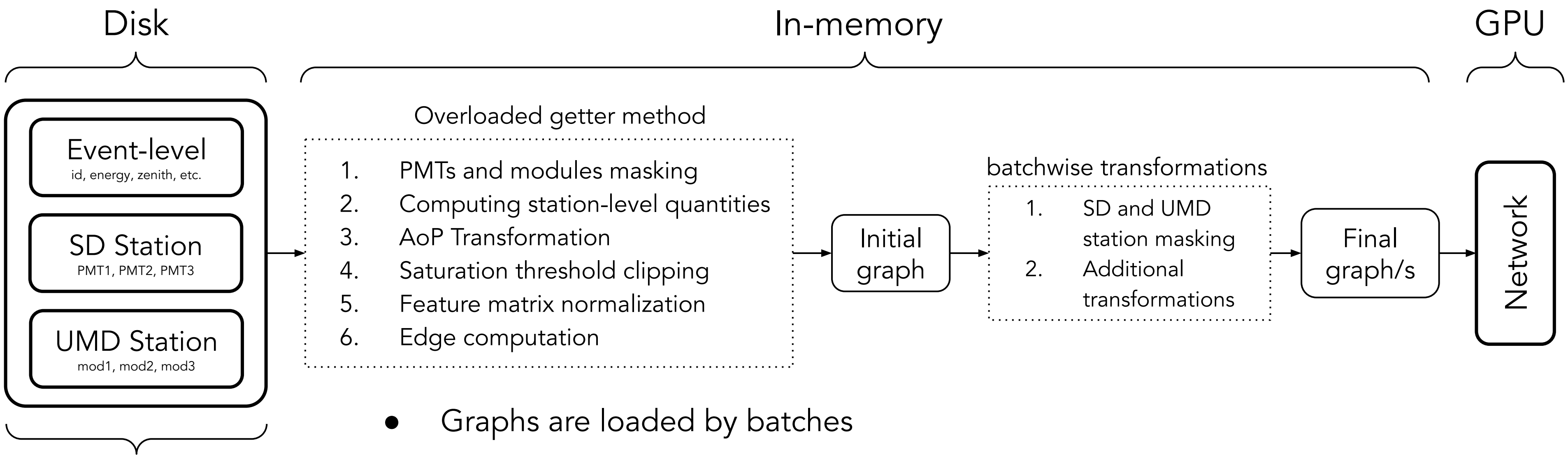
GA

Geometry Analyzer
3 layers of Graph Attention Layers (GATv2)

MLP

Multi-layer Perceptron
3 dense layers

Augmentation pipeline and training loop



- Pre-processed once
- Streamed from JSON files and stored in PyTorch format

- Graphs are loaded by batches
- All parts of the augmentation can be turned on and off
- Edge logic can be modified
- Graph transformations are applied in the training loop
  - Implementing transformations can be cumbersome
- Requires more memory than just reading from disk
- DataLoaders load the new batch in-memory as the current batch goes to GPU

Dataset and training

Models: EPOS-LHC; FLUKA INFN

Offline: v4r0p2-pre3

CORSIKA: 7.7xx

Energy range: lg(E/eV) ∈ (16.5-17.5)

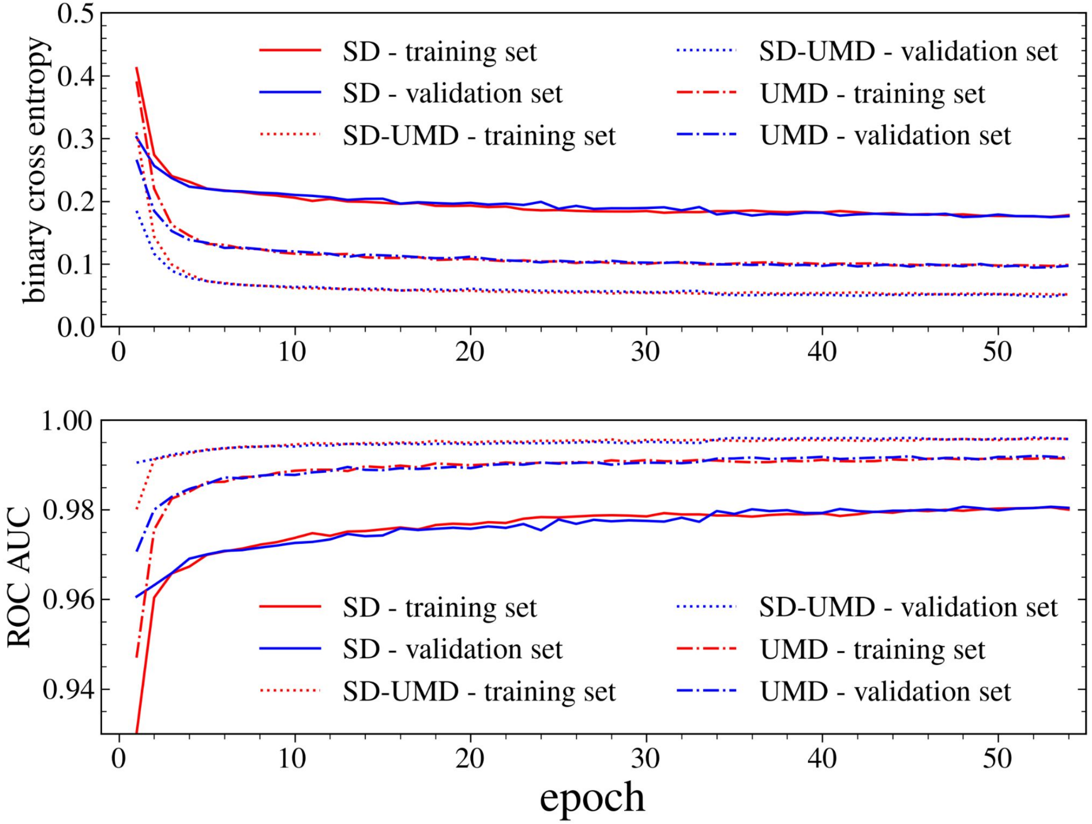
Zenith up to 45 deg

Training: 118k, Validation: 40k, Testing: 50k

Equal Mix of photons/protons

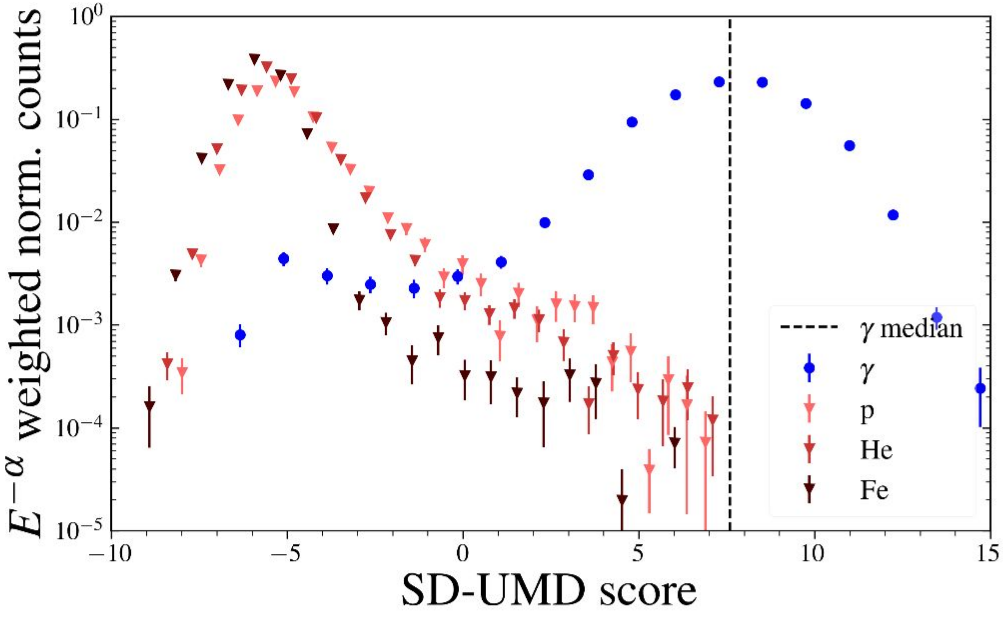
- Binary cross entropy as loss
  - global loss is the sum of the 3 losses
- We augment the dataset during training

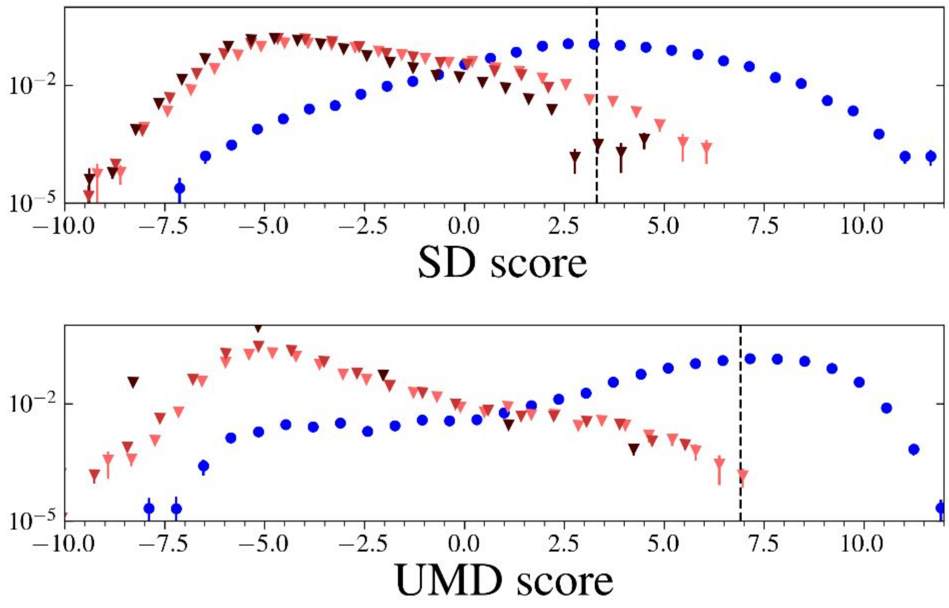
Check the backup slides for more information.



Performance

Photons are weighted by $\propto E^{-2}$ and hadrons $\propto E^{-3}$.





- Complete energy and zenith range (lg(E/eV) ∈ (16.5-17.5) & θ<45 deg)
- No background events after signal median → study of tail of distributions for bkg. estimation
- Separation improves with the primary mass

Scores vs MC quantities

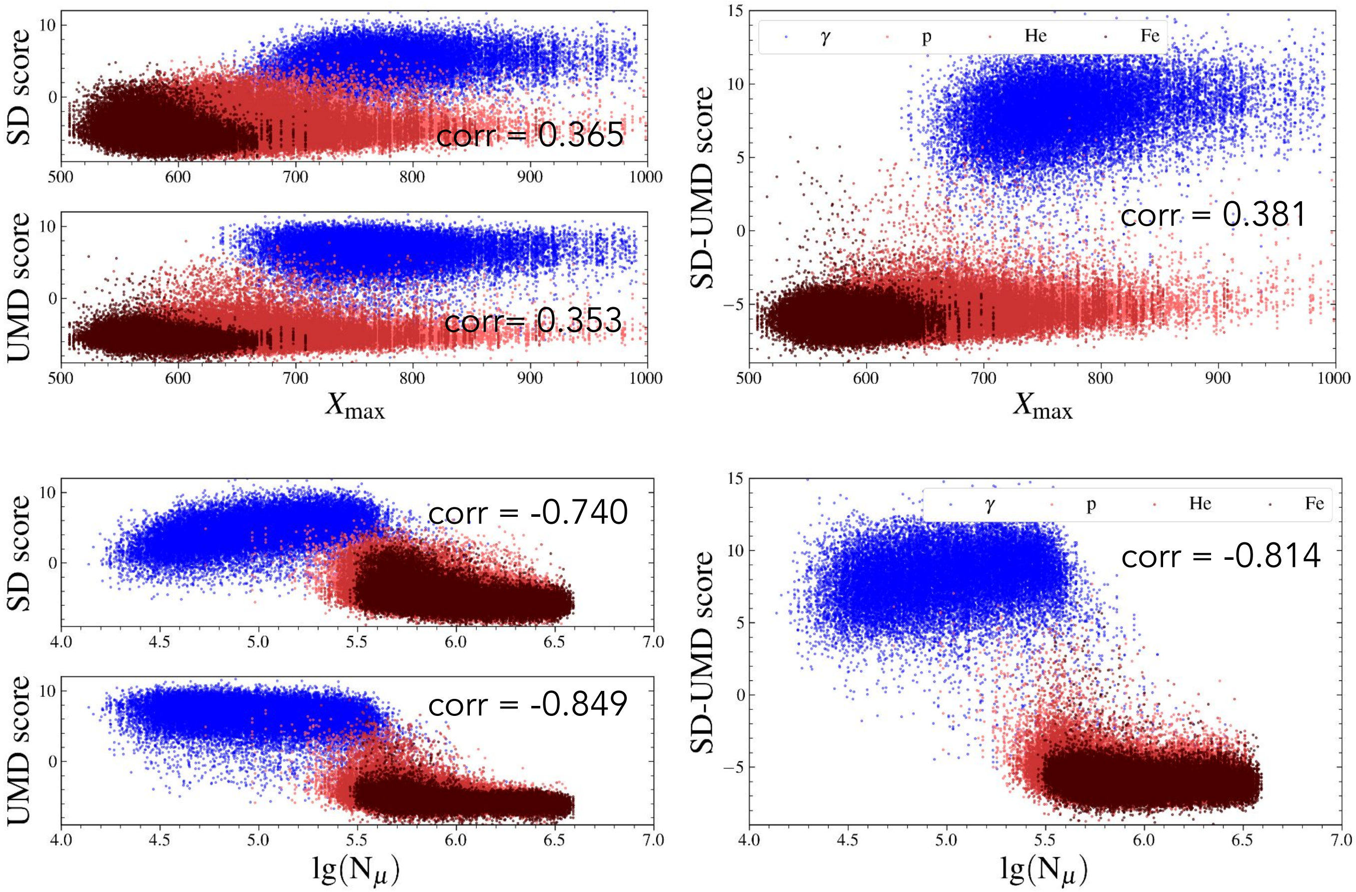
Best correlation with X_{max}

UMD-SD score

Best correlation with

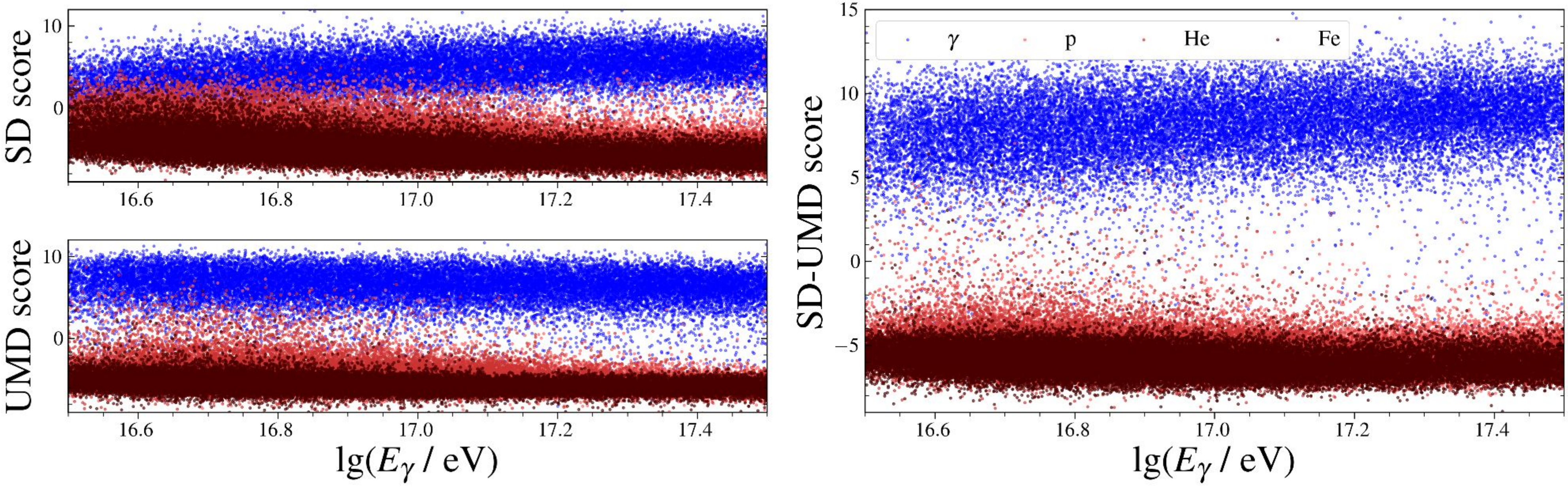
$\lg(N_\mu)$

UMD score



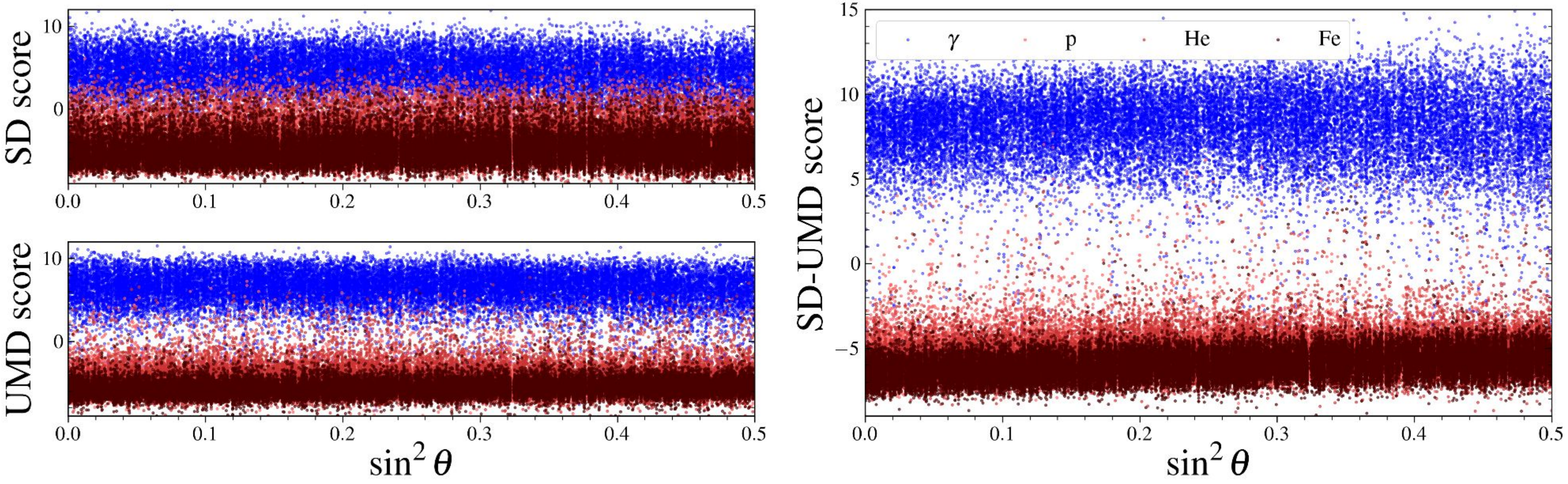
The score reflects physical air-shower observables

Scores vs reconstructed photon energy



Discrimination depends little on reconstructed energy E_γ

Scores vs reconstructed zenith angle



Discrimination depends little on reconstructed zenith angle

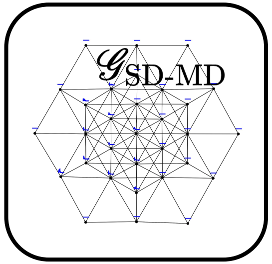
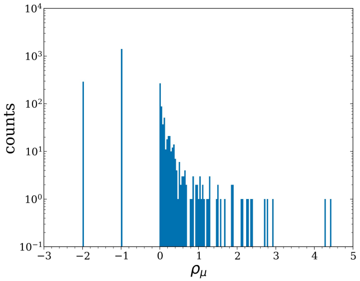
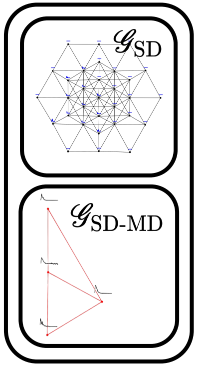
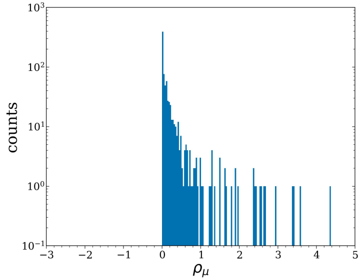
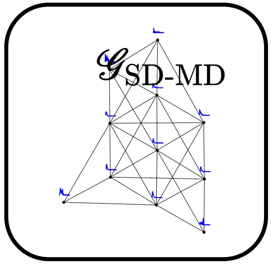
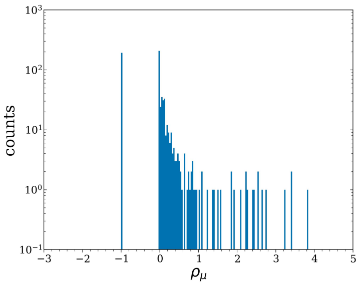
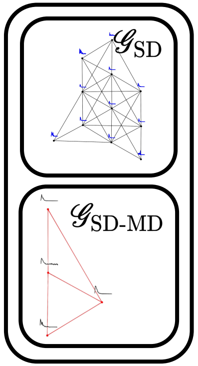
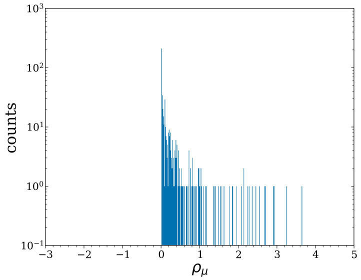
Summary

- Cosmic rays below 10^{17} eV can be measured with Auger SD+UMD 433 m array
- In the quest for photon detection at the 10^{16} eV decade, graphs allow for different types of representations
- Separating information inside the network allows for some level of introspection
- Networks are able to separate photons from hadrons
  - Cross-check with heavier primaries shows promising results
  - Separation depends little on reconstructed energy and zenith angle

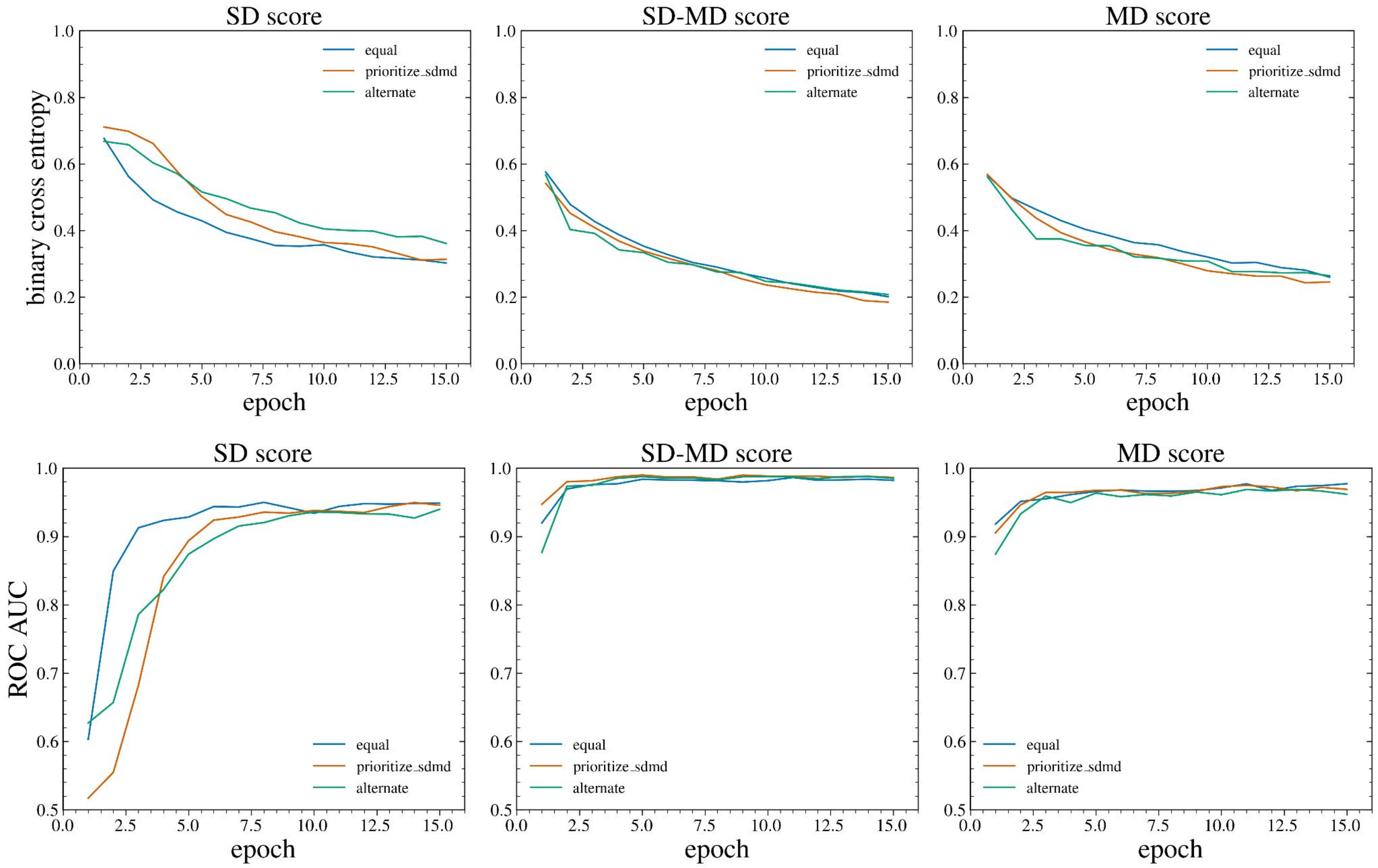
Backup

Event representation with homogeneous graphs

1 graph = 3 matrices (features, traces, edges)

	single input		dual input	
with non-triggered stations		 Min Max Scaling		 Standardization
without non-triggered stations		 Min Max Scaling		 Standardization

Learning curves for different training strategies



Test with reduced dataset

10000 training events

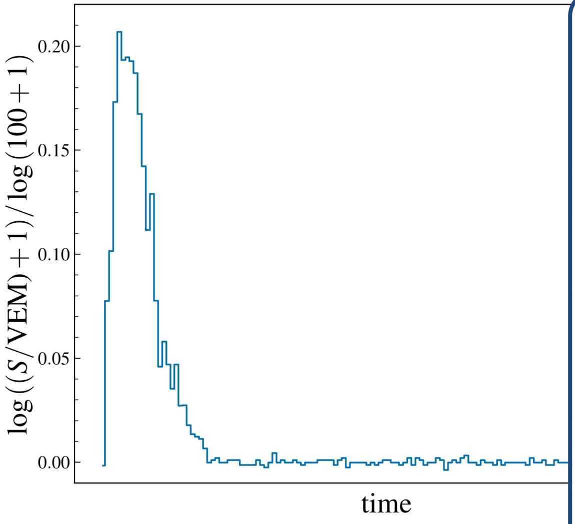
1000 validation events

Only 15 epochs

Trace Analyzer

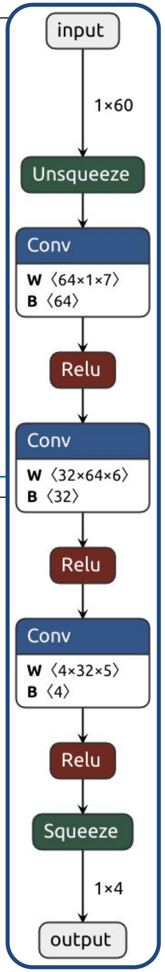
- First 60 bins of the averaged VEM trace
- Randomly masking 0, 1, or 2 PMTs
- Batch normalization helps A LOT
- Only 4 features as output → Thanks Fiona!
- One Trace Analyzer for the whole network

```
TraceAnalyzer(
  (conv_layers): Sequential(
    (0): Conv1d(1, 64, kernel_size=(7,), stride=(3,))
    (1): BatchNorm1d(64, eps=1e-05, momentum=0.1, affine=False, track_running_stats=True)
    (2): ReLU()
    (3): Conv1d(64, 32, kernel_size=(6,), stride=(3,))
    (4): BatchNorm1d(32, eps=1e-05, momentum=0.1, affine=False, track_running_stats=True)
    (5): ReLU()
    (6): Conv1d(32, 4, kernel_size=(5,), stride=(1,))
    (7): BatchNorm1d(4, eps=1e-05, momentum=0.1, affine=False, track_running_stats=True)
    (8): ReLU()
  )
)
```



NETRON

Visualization tool for NNs that supports many formats



Graph Convolutional Networks (GCN)

new representation

linear combination of features

$$H^{k+1} = \sigma(\underbrace{\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}} \overbrace{H^k W^k})$$

activation function

$$\tilde{A} = A + \mathbf{I}$$

$$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$$

learnable parameters

weights based on neighborhood and node degree

Conceptually: fancy linear combination

In practice: Forward Pass is just matrix multiplication

Geometry Analyzer using Graph ATtention

- Using 3 Graph Attention layers
- 3 attention heads per layer
- 3 Geometry Analyzers
- Batch normalization here too

P. Veličković et al. Graph attention networks. *arXiv preprint* [arXiv:1710.10903](https://arxiv.org/abs/1710.10903), 2017.

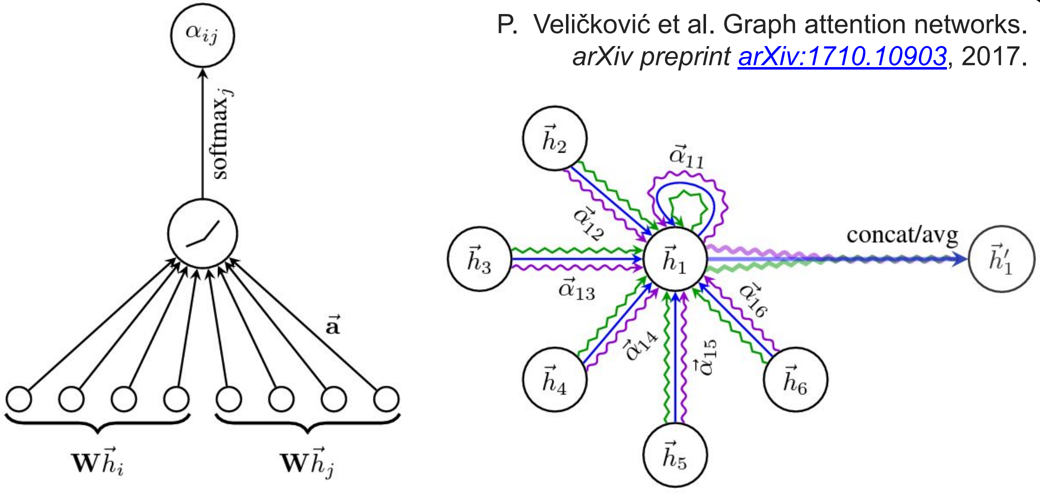
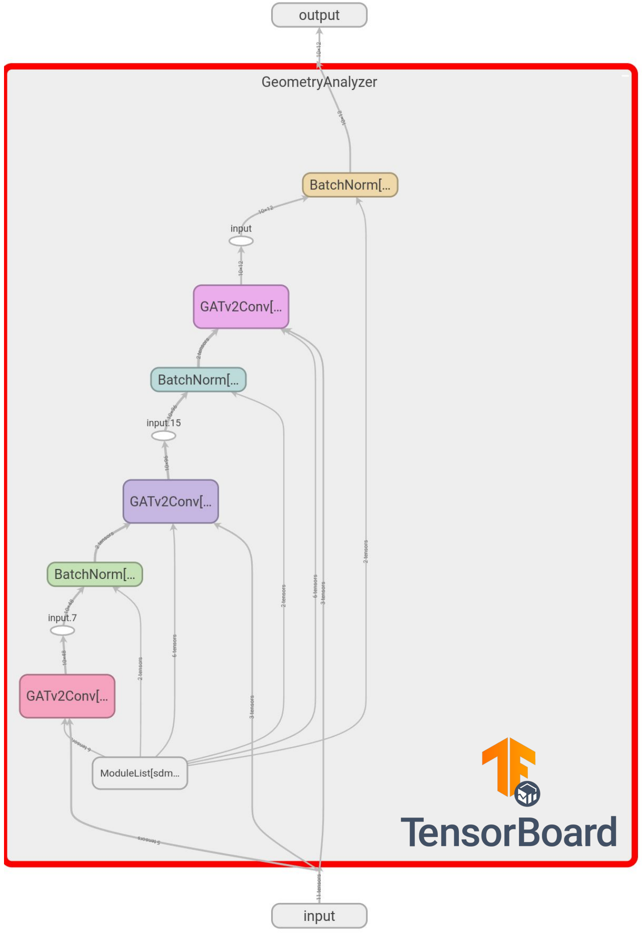


Figure 1: **Left:** The attention mechanism $a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j)$ employed by our model, parametrized by a weight vector $\vec{\mathbf{a}} \in \mathbb{R}^{2F'}$, applying a LeakyReLU activation. **Right:** An illustration of multi-head attention (with $K = 3$ heads) by node 1 on its neighborhood. Different arrow styles and colors denote independent attention computations. The aggregated features from each head are concatenated or averaged to obtain $\vec{h}'_1$.



Augmentation

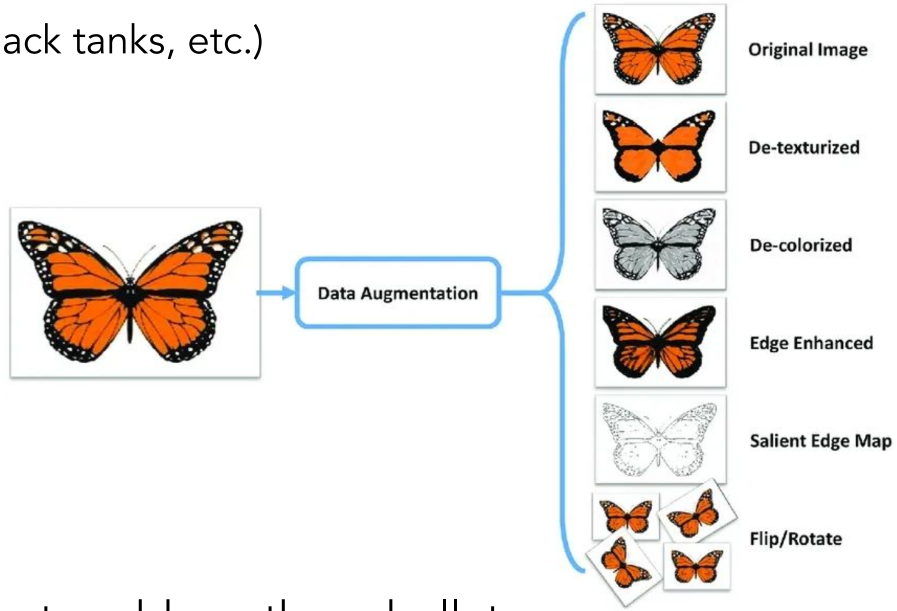
Generating more simulations from existing simulations.

- Increase the variability in the dataset
  - ~105.k photon and ~105.k divided in train/validation/test sets)
- Reproduce more faithfully the events in data
  - Simulations are ideal (saturation effects, ageing, black tanks, etc.)
- Stress in the training to get a more robust network

We should retain the physics!

We'll focus on:

- Failures at PMT and UMD module level
- Area over Peak (AoP)
- Saturation
- Failures at SD and UMD station level



We use the latest Phase I SD433-UMD test productions to address these bullets.