Joint Data and Algorithms for Deep Learning in Fundamental Physics

Lisa Benato, Erik Buhmann, Jonas Glombitza, Martin Erdmann, Peter Fackelday, Nikolai Hartmann, Gregor Kasieczka, William Korcari*, Thomas Kuhr, Jan Steinheimer, Horst Stöcker, Tilman Plehn, Kai Zhou

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

ErUM-Data IDT

Bundesministerium für Bildung und Forschung

*william.korcari@uni-hamburg.de

The project

- Collect different datasets from the ErUm data group;
- Implementation of ML models that perform reasonably well on all these datasets.

1. “Top Tagging at the LHC”: 1902.09914;
2. “Spinodal or not?” : 1906.06562;
3. “EOSL or EOSQ” : 1910.11530;
4. Cosmic Airshowers: publication;
5. SmartBKG dataset (Belle II - generated events passing downstream selection): link.

The erum_data_data package

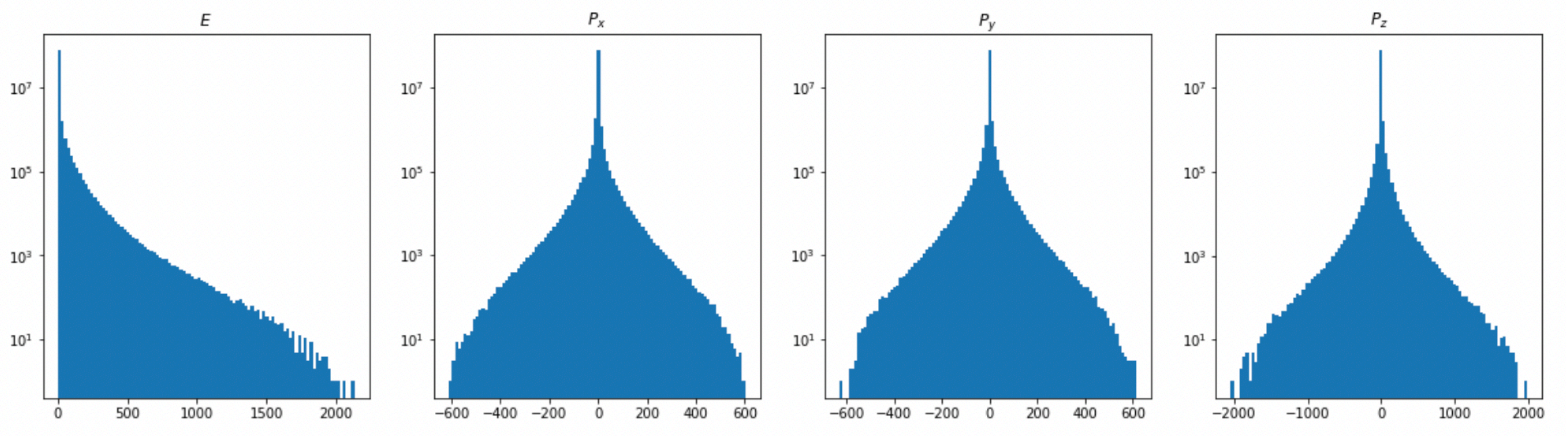
- Provides an easy-to-use library* to work with all the provided datasets;
- Data are packed in a convenient format (.npz);
- Flexible in the implementation of both the models and the preprocessing routines.

```python
from erum_data_data import TopTagging

# load non processed training and testing set
X_train, y_train = TopTagging.load('train', path = './datasets')
X_test, y_test = TopTagging.load('test', path = './datasets')

# load processed training and testing set
X_train, y_train = TopTagging.load_data('train', path = './datasets', graph = True)
X_test, y_test = TopTagging.load_data('test', path = './datasets', graph = True)
```

*erum_data_data git repository

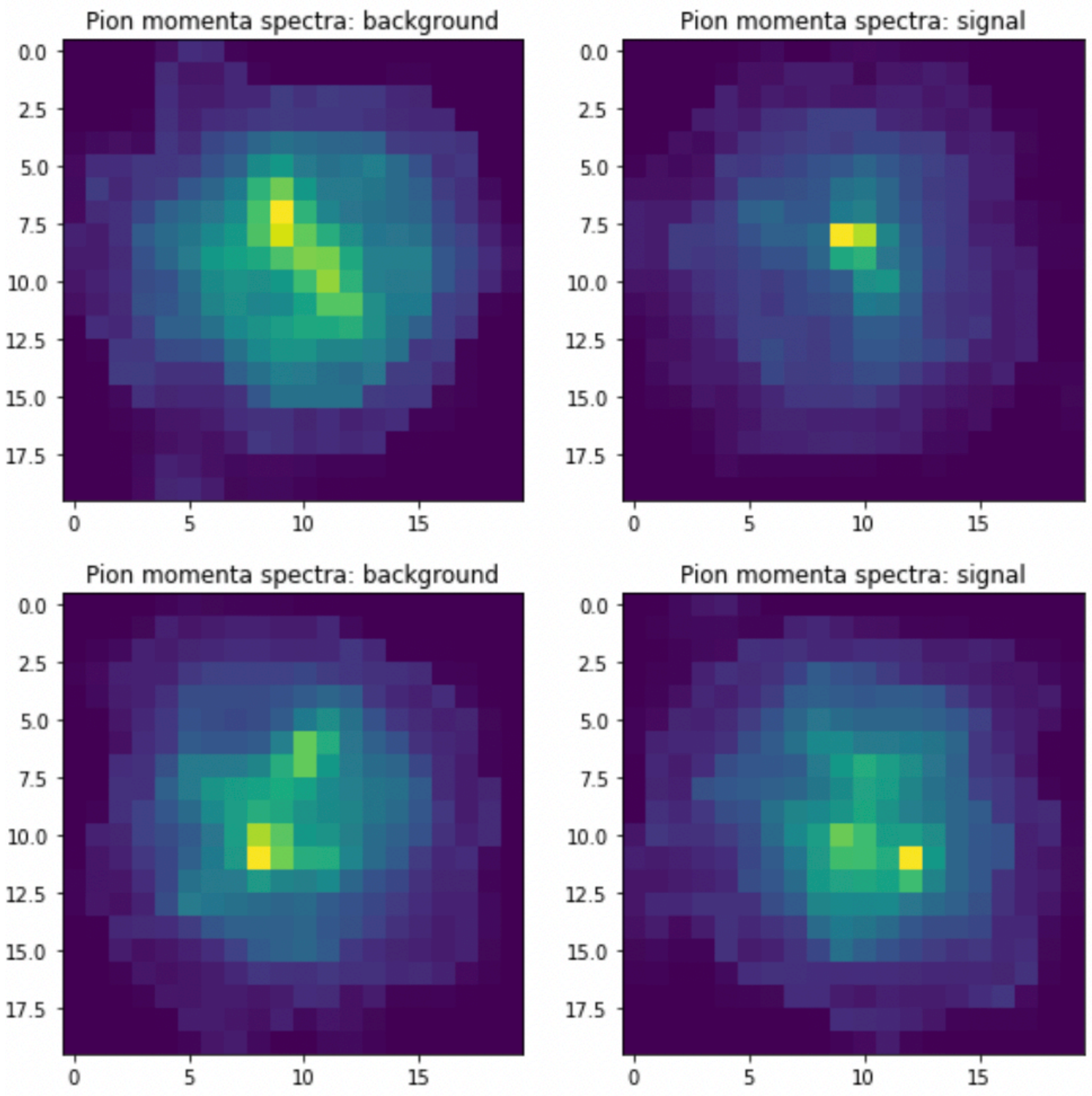
The datasets: Top tagging

- 14 TeV, hadronic tops for signal, QCD dijets background, Delphes ATLAS detector card with Pythia;
- The leading 200 jet constituent 4-momenta are stored, with zero-padding for jets with fewer than 200;
- Reference model: particle Net 1902.08570.



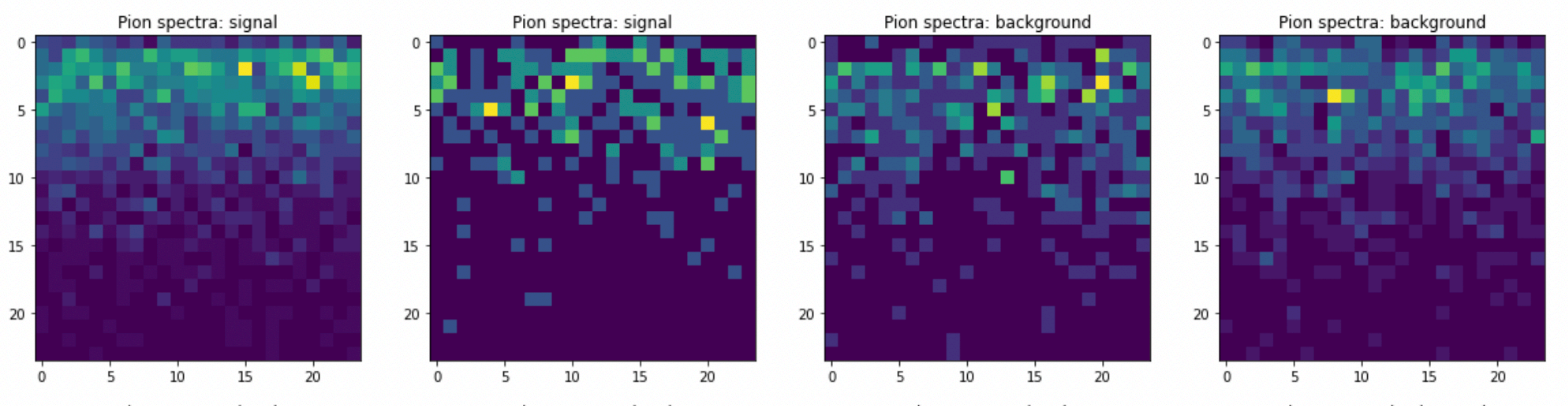
The datasets: Spinodal

- Classify the nature of QCD phase transitions in heavy ion collisions at the CBM experiment;
- Signals for b-associated with the phase transition can be found in the final momentum spectra of certain collisions;
- The dataset is composed of 29'000 2D histograms describing pion momenta.
- Reference model: Convolutional Neural Network as in 1906.06562



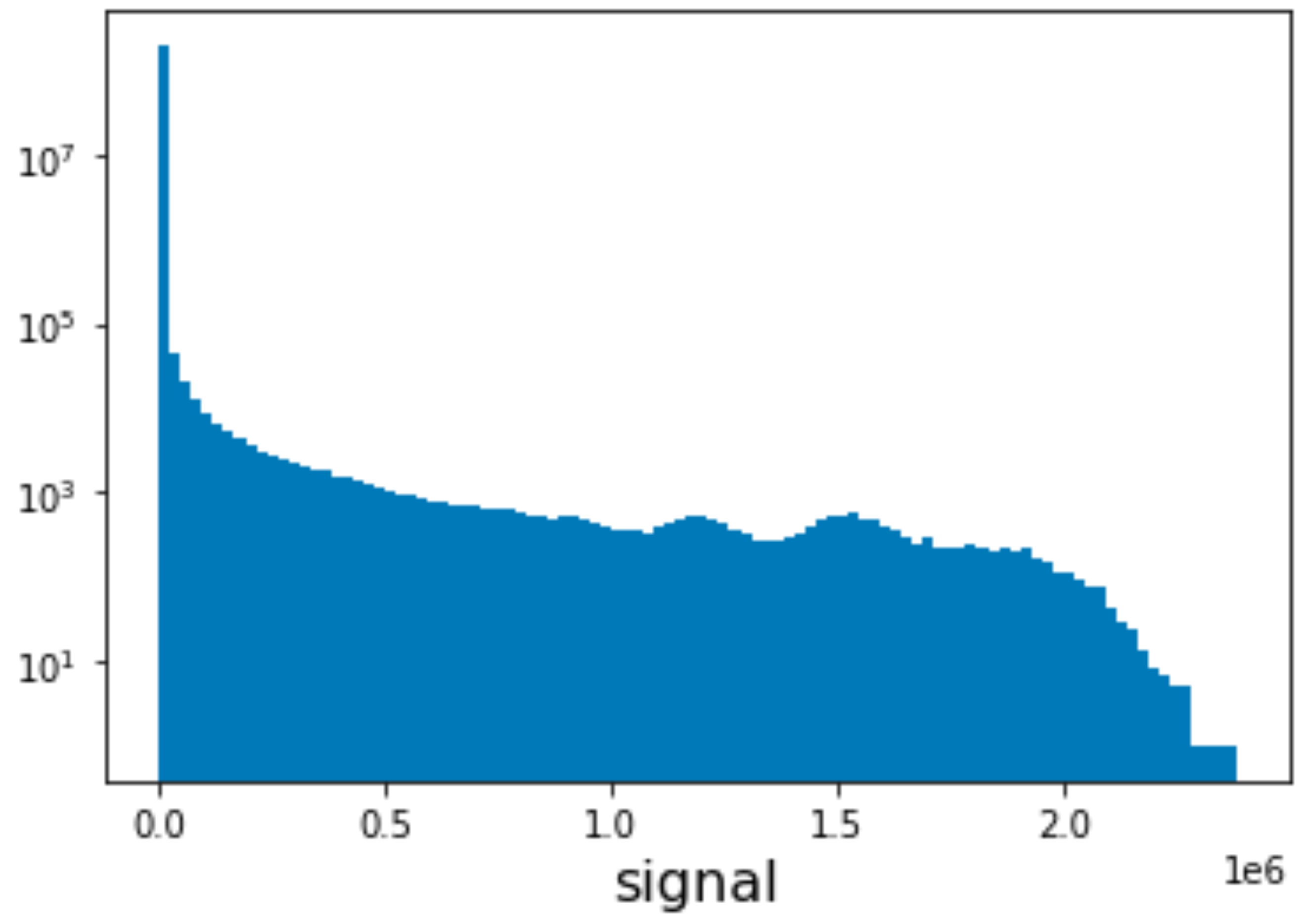
The datasets: EoS

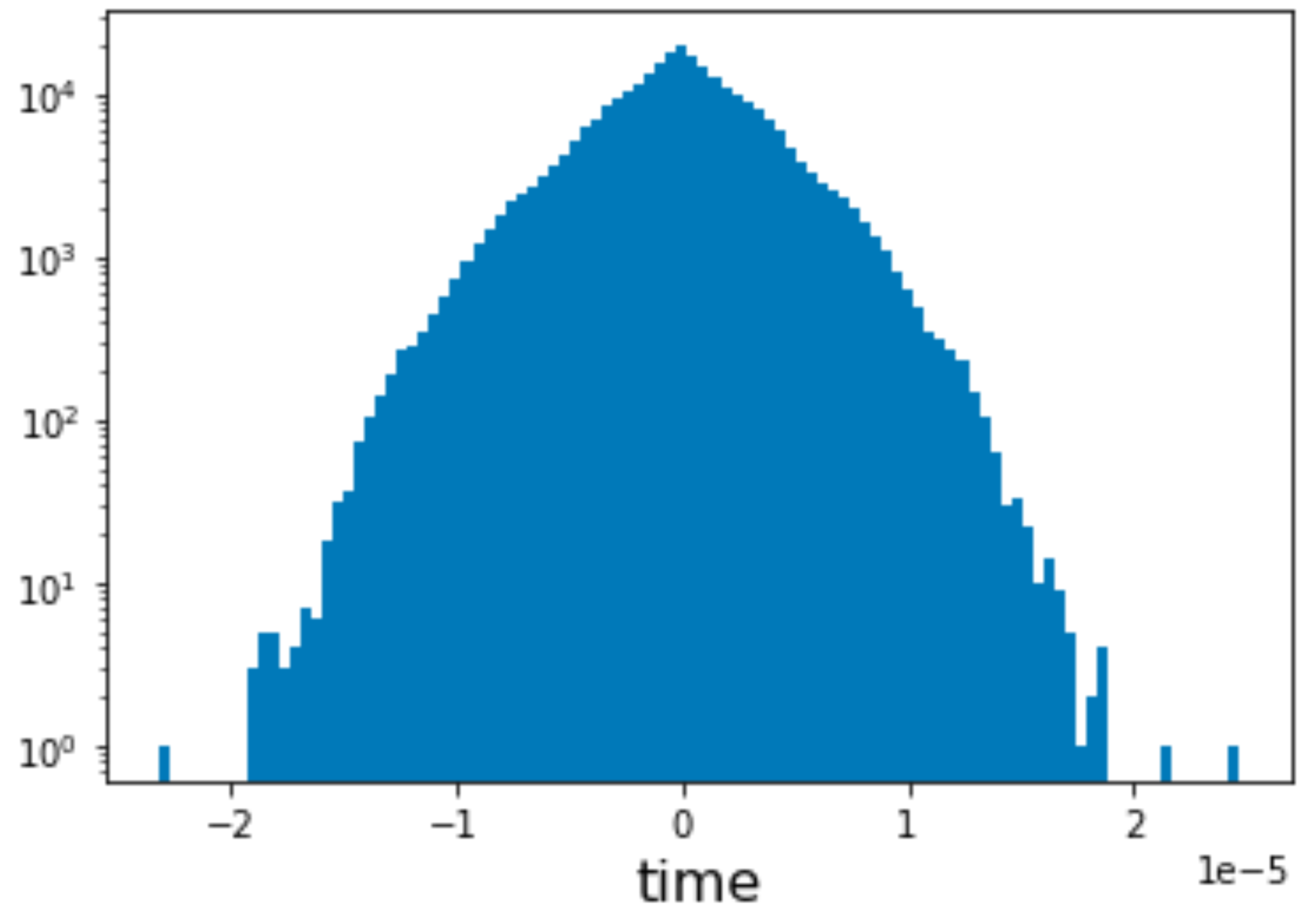
- Classify the QCD transition nature in heavy-ion collisions from the final state pion spectra;
- 2 equation of state: cross-over EOSL or 1st order EOSQ;
- Modeling for heavy-ion collisions by varying different physical parameters (collision energy, centrality, initial time, etc.);
- Data simulated with different parameters for the test set.
- Reference Model: Convolutional Neural Network as described in 1910.11530

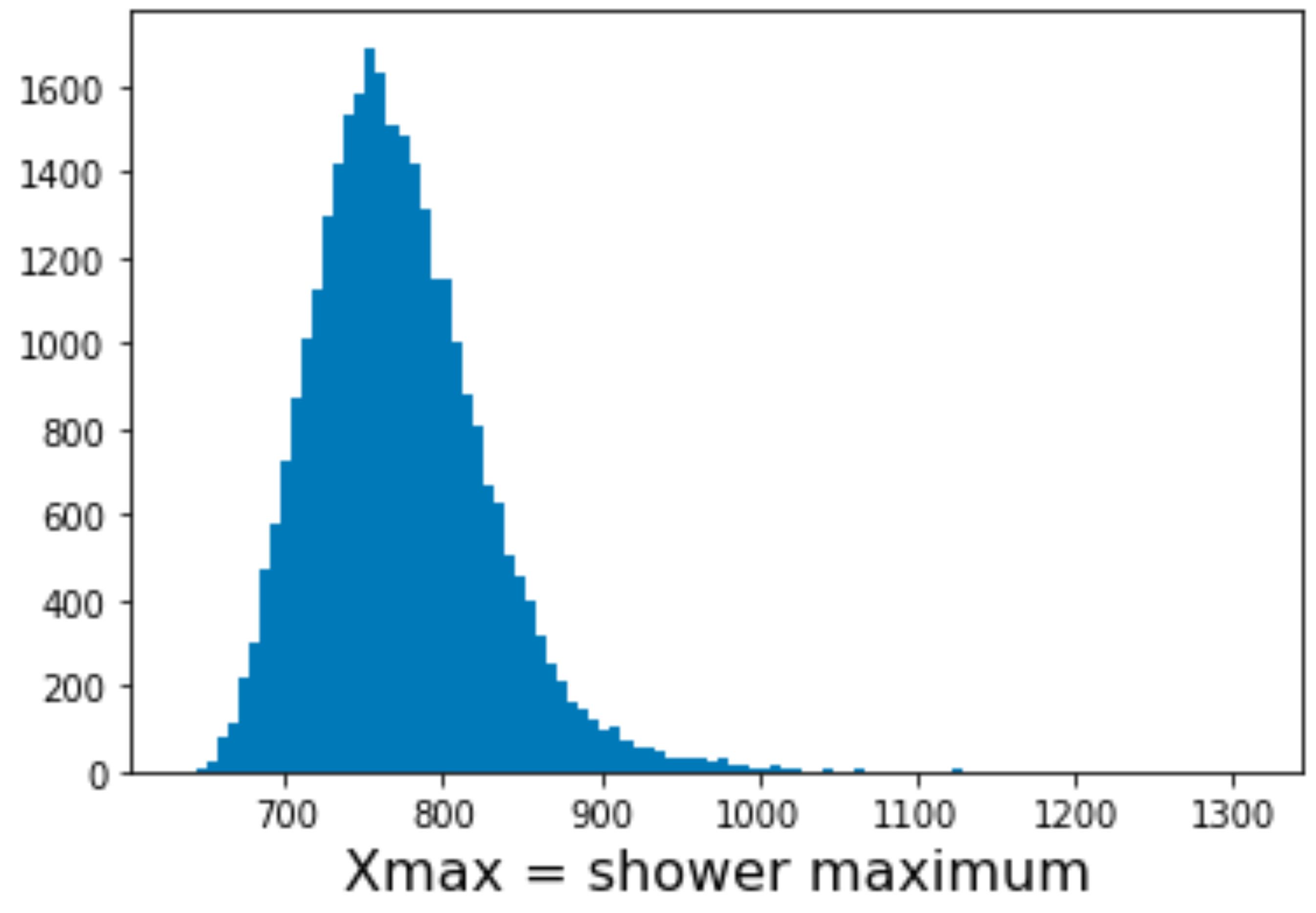


The datasets: Cosmic Airshowers

- Regression task: predict the shower maximum;
- 2 subsets;
- 70k events (airshowers);
- 81 ground detector stations disposed in a 9x9 grid;
- 80 measured signal bins (forming one signal trace per station);
- 1 starting time of the signal trace (arrival time of first particles at each station);
- Reference model: Residual Neural Network (see J. Glombitza talk at 9:40)

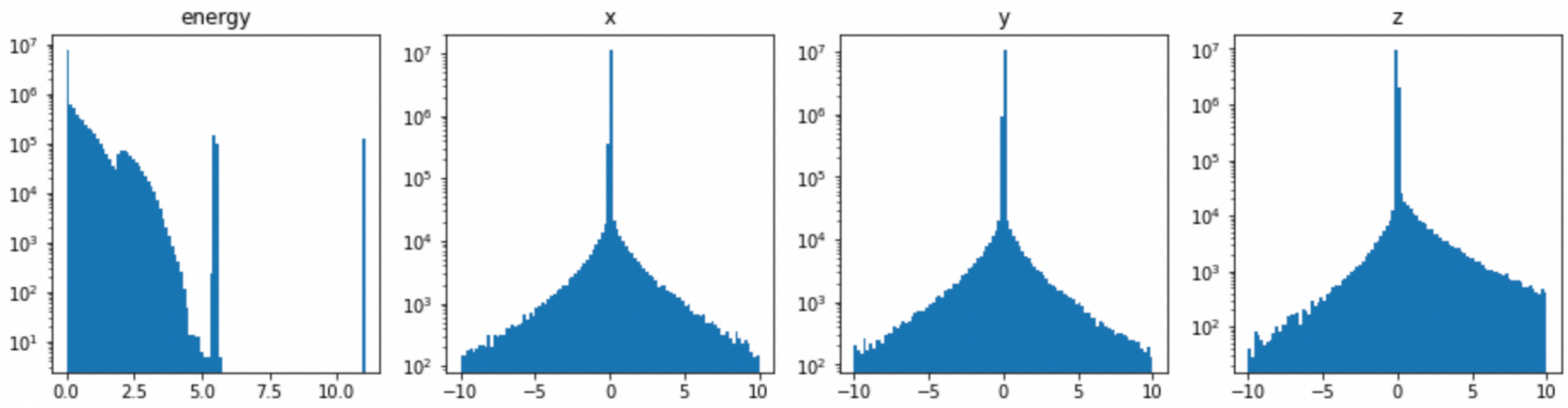






The datasets: smartBKG (Belle II)

- Simulated events with generator level information;
- Event passes (1) or fails (0) a selection that was applied after detector simulation and reconstruction;
- Total of 400k events, max. 100 particles per event characterized by 9 features:
    1. Production time, E, x, y, z px, py, pz, PID.
- PID corresponding to a unique PDG particle ID mapped to a continuous space;
- Indices of mother particles are used to create adjacency matrix for GCN.
- Reference model: Graph Convolutional Network.



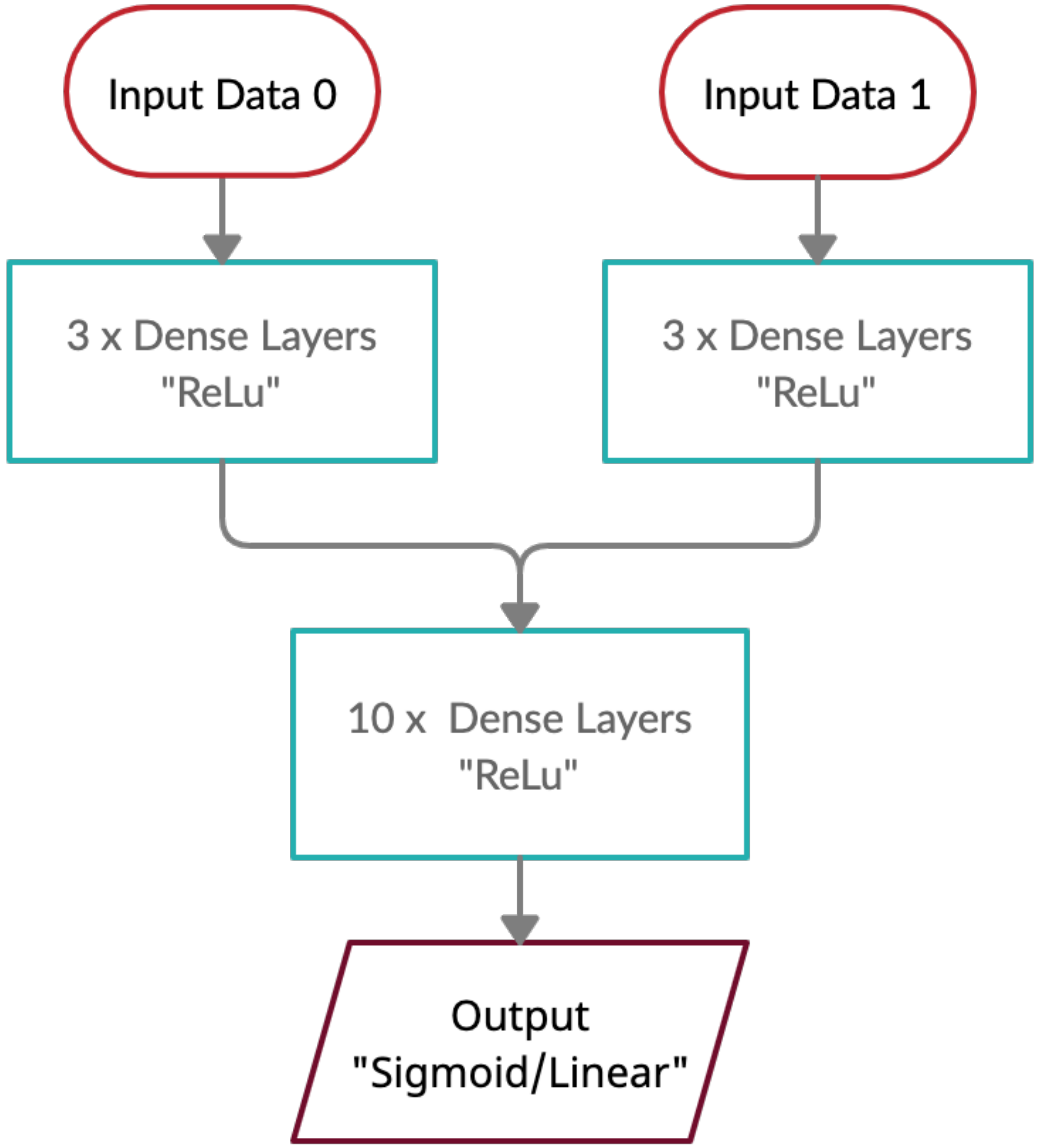
The Fully Connected Network model

TensorFlow implementation:

- Number of inputs changes depending on the dataset;
- Dense layers with 256 nodes each;
- Output layer changes depending on the task:
- Batch size: 256;
- Loss: BCE or MSE;
- Epochs: 300;
- Learning rate 0.001 with Adam optimizer.

Callbacks:

- Reduce on plateau with patience 8 epochs;
- Early stopping with patience 15 epochs.



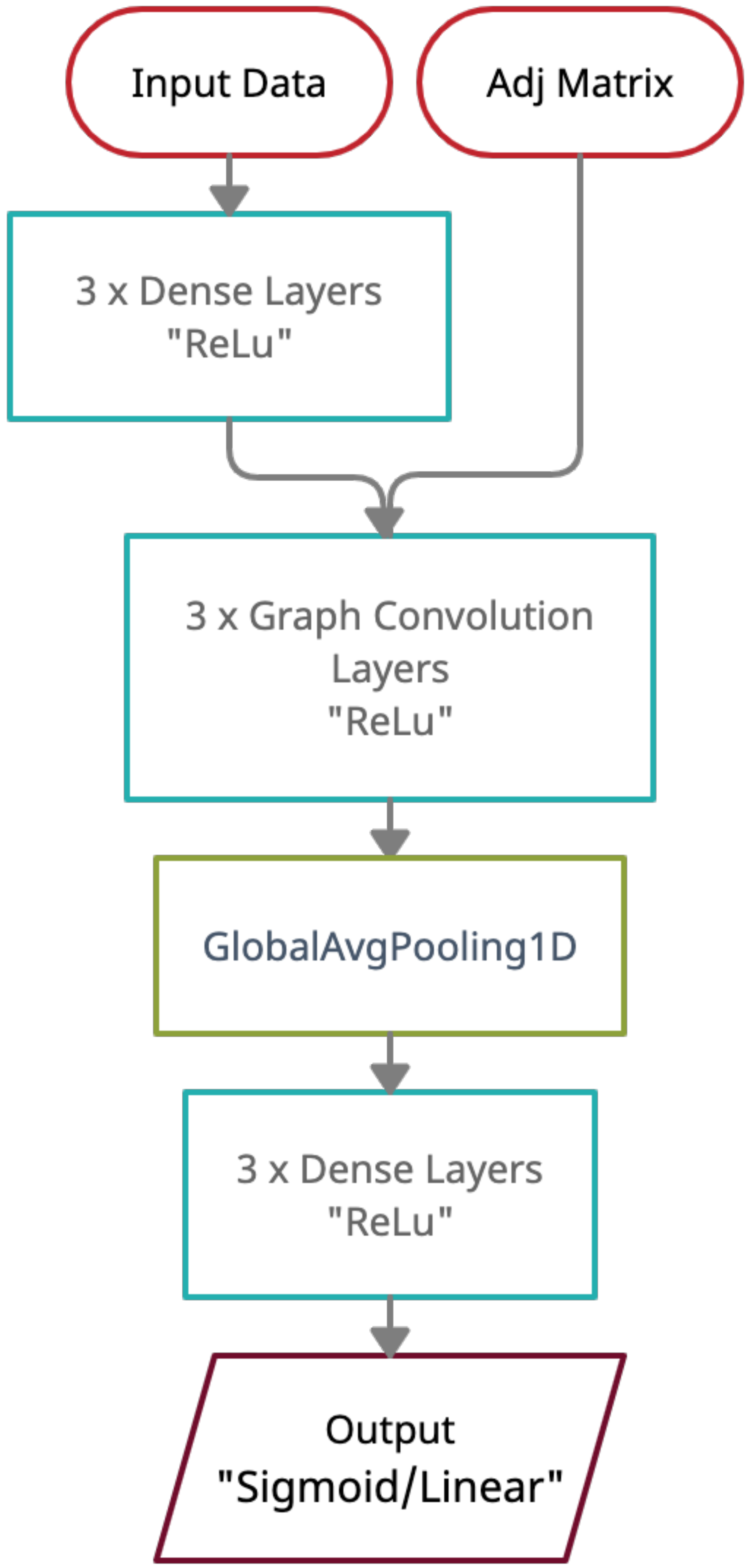
The Graph Network model

TensorFlow implementation:

- Dense layers with 256 nodes each;
- Graph Convolution layers with 256 nodes each;
- GlobalAvgPool1D;
- Output layer changes depending on the task:
- Batch size: 256;
- Loss: BCE or MSE;
- Epochs: 300;
- Learning rate 0.001 with Adam optimizer.

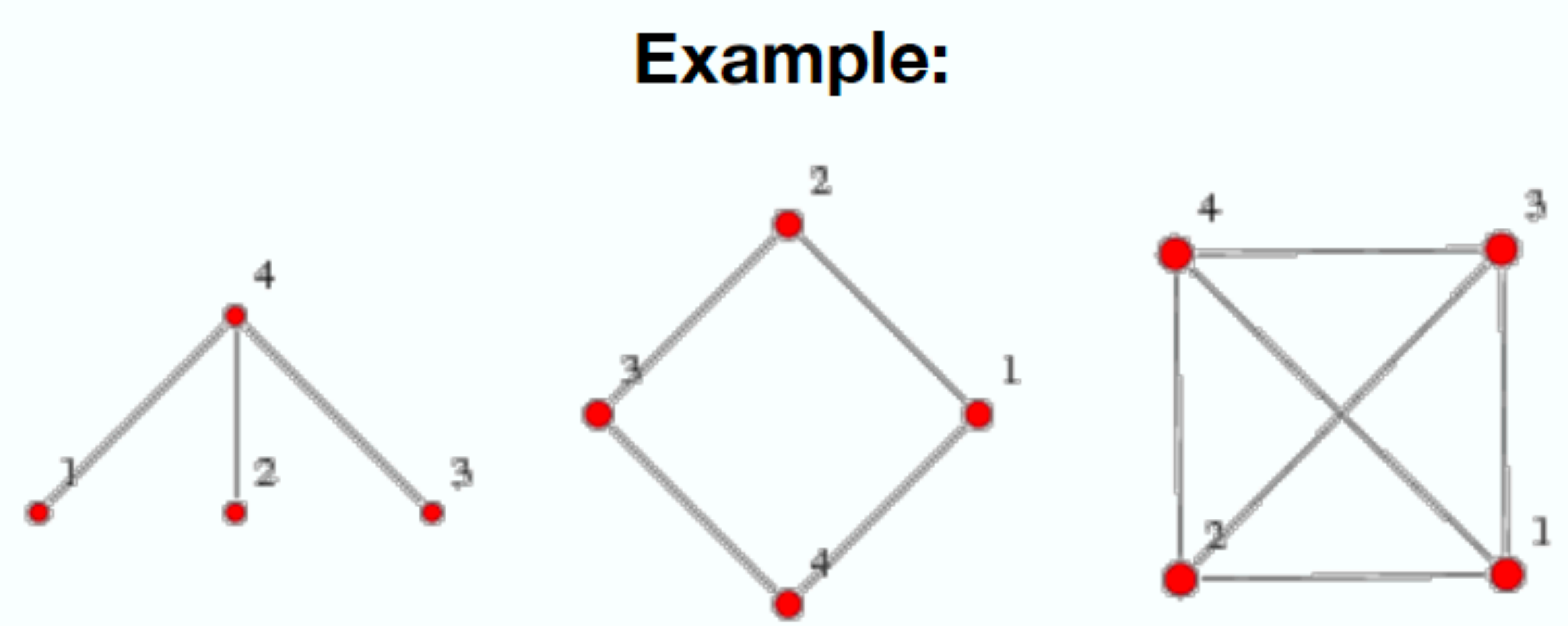
Callbacks:

- Reduce on plateau with patience 8 epochs;
- Early stopping with patience 15 epochs.



The Graph Network model: adjacency matrices

- TopTagging (jet constituents): kNN clustering of particles per events (k = 7);
- Spinodal & EOS (images): 8-connected neighboring pixels;
- Belle (jet constituents): Matrix with event history via mother & daughter particles (same as Reference Model);
- Airshowers (signal bins & timing of 81 ground stations): 8-connected neighboring stations (assumes rectangular 9x9 grid);



$$\begin{pmatrix}0&0&0&1\\0&0&0&1\\0&0&0&1\\1&1&1&0\end{pmatrix}\quad\begin{pmatrix}0&1&0&1\\1&0&1&0\\0&1&0&1\\1&0&1&0\end{pmatrix}\quad\begin{pmatrix}0&1&1&1\\1&0&1&1\\1&1&0&1\\1&1&1&0\end{pmatrix}$$

Performance comparison

	TopTag		Spinodal		EOS		smartBKG		Airshower	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	MSE	Resolution
Ref. Model	0.939 ± 0	0.985 ± 0	0.873 ± 0.004	0.925 ± 0.005	0.691 ± 0.005	0.788 ± 0.005	0.823 ± 0.001	0.906 ± 0.009	1000 ± 52	31.32 ± 0.75
Graph Model	0.933 ± 0	0.982 ± 0	0.871 ± 0.002	0.929 ± 0.001	0.673 ± 0.003	0.735 ± 0.005	0.822 ± 0.001	0.902 ± 0	1514 ± 274	38.69 ± 3.52
FCN Model	0.907 ± 0.001	0.968 ± 0.001	0.824 ± 0.001	0.883 ± 0.001	0.605 ± 0.019	0.739 ± 0.008	0.736 ± 0.004	0.812 ± 0.001	1528 ± 31	38.96 ± 0.34

Conclusions

- The edd package allows easy loading of Fundamental Physics datasets;
- Provides a space for model comparisons;
- Models that perform reasonably well on all these datasets can be build:
  1. FCN model: reasonable overall performance;
  2. GraphNet: performances comparable to the reference models.

Thank you

Backup slides

Performance comparison: EOS

EOS	Test set		Validation set	
	ACC	AUC	ACC	AUC
Reference	0.691 ± 0.005	0.788 ± 0.005	0.816 ± 0.002	0.9 ± 0
FCN	0.605 ± 0.019	0.739 ± 0.008	0.74 ± 0.009	0.827 ± 0.005
GraphNet	0.673 ± 0.003	0.735 ± 0.005	0.821 ± 0.001	0.906 ± 0.001