Shared Data and Algorithms for Deep Learning in Fundamental Physics

Erik Buhmann*, Erika Garutti, Gregor Kasieczka

with

Lisa Benato, Jonas Glombitza, Martin Erdmann, Peter Fackeldey, Nikolai Hartmann, William Korcari, Thomas Kuhr, Jan Steinheimer, Horst Stöcker, Tilman Plehn, Kai Zhou

arXiv:2107.00656

GitHub: pd4ml

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

Bundesministerium für Bildung und Forschung

ErUM-Data IDT

* erik.buhmann@uni-hamburg.de

ErUM-Data IDT Collaboration

- IDT = Innovative Digital Technologies for Research on Universe and Matter (ErUM)
- German science project involving 17 partners from universities and research centres
- Tackling the challenge of increasing data rates and volumes
- Exploit modern technologies for the development of experiment overarching solutions

Bundesministerium für Bildung und Forschung

ErUM-Data IDT

https://www.erum-data-idt.de

Challenges of Data Science in Physics

- Need for abundant training data for Machine Learning models
- Various data structures such as low-level or high-level observables, images, etc.
- Machine learning architecture designs often experiment specific
- Difficulties to communicate physics specific challenges to the wider computer science community

—> Domain and experiment-specific data representations are inefficient:

- Different groups replicate similar developments
  - Or: *Shared Data and Algorithms for Deep Learning in Fundamental Physics*

Shared Data and Algorithms for Deep Learning

A. Shared Data:

- Collection of various open datasets
- Easy-to-use: Import data with one line of Python code

Available
via pd4ml package

B. Shared Algorithms:

- Provide a expert-tuned reference model for each dataset
- Implements some basic dataset specific preprocessing
- Includes overarching ML models performing well on all datasets
  - Showing that in particular a Graph Networks can be applied to almost any data structure

The pd4ml Python Package

Physics Data for Machine Learning (pd4ml) library:

Available
via pd4ml package

- Easy-to-use library to work with all the provided datasets
- Data are packed in a convenient format (as numpy arrays)
- Contains implementation of ML models and preprocessing routines

```
from pd4ml import TopTagging

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

Shared Data (current status)

Five datasets currently available (looking forward to further contributions):

	Origin	Task	Examples	Structure	Dimension	Reference model
Top Tagging Landscape	Simulation	Classification	2M	Four vectors	200 particles, 4 features/particle	GraphNet
Smart Backgrounds	Simulation	Classification	280k	Decay Graph	100 particles, 9 features/particle	GraphNet
Spinodal or Not	Simulation	Classification	29k	2D Histogram	20x20 histogram of pion spectra	CNN
EoS	Simulation	Classification	200k	2D Histogram	24x24 histogram of pion spectra	CNN
Air Showers	Simulation	Regression	100k	81 1D Traces	81 stations, 80 signals + timing	RNN

Shared Algorithms (current status)

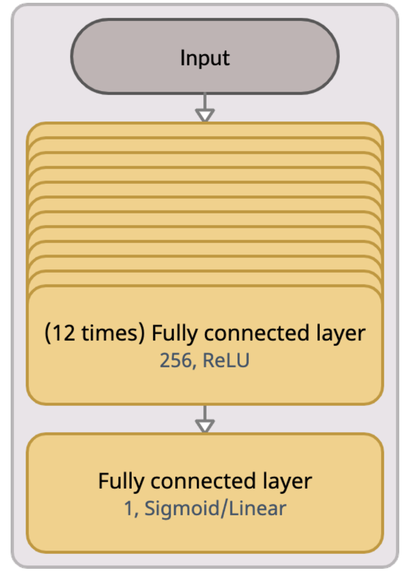
- Each datasets comes with a Reference model as an expert-tuned ML architecture

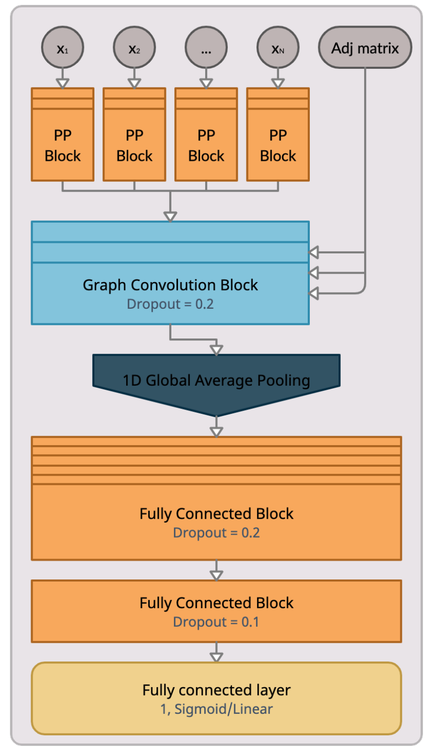
- Additionally we provide <u>two shared ML models</u> working on all datasets:

  1. Fully Connected Network (standard Neural Network, 12 layers à 256 nodes, Sigmoid or Linear output node)

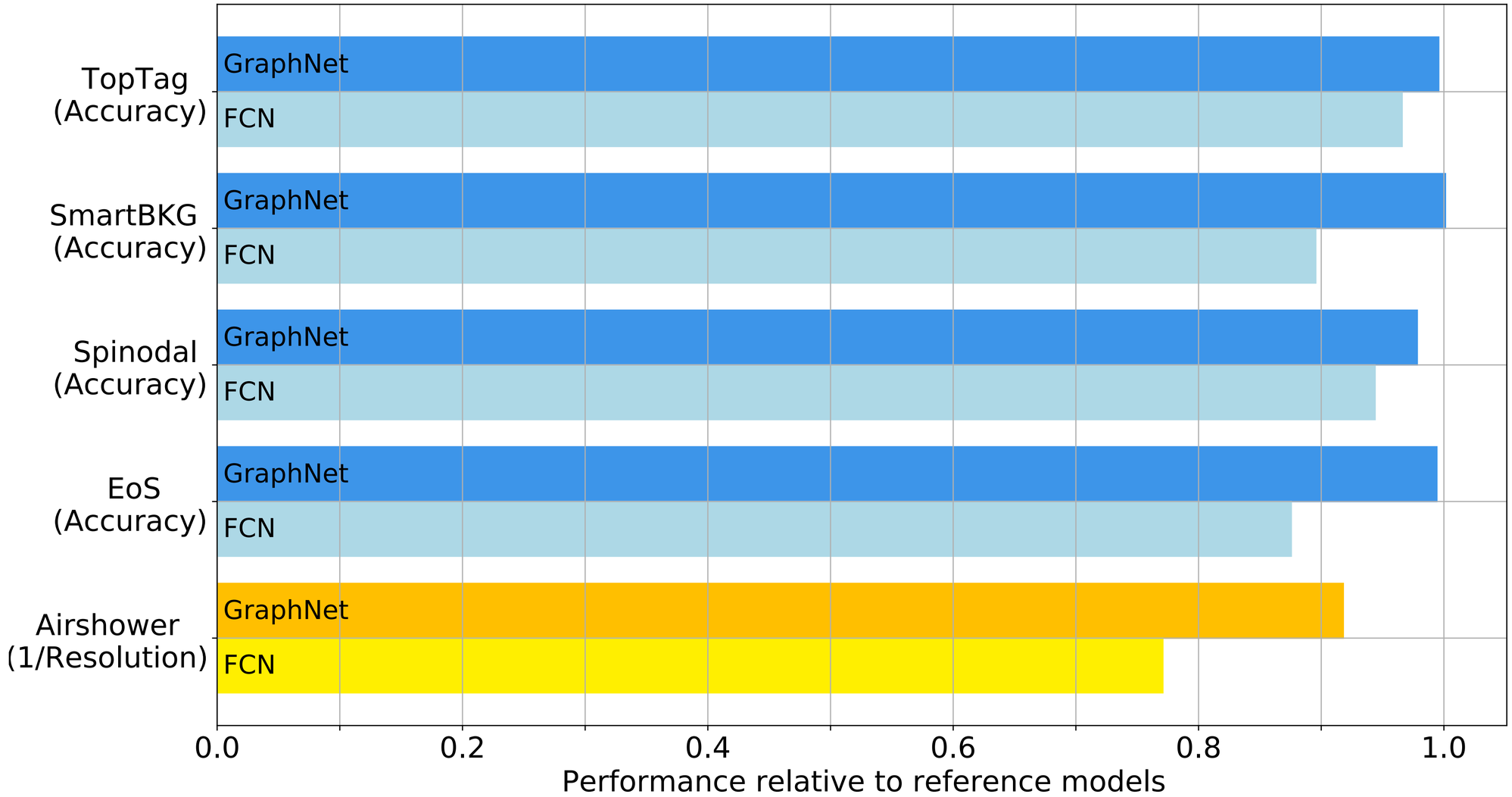
  2. Graph Convolution Network (using an adjacency matrix calculated from data to take spatial relations into account)

- Dataset-specific preprocessing is included (& adjacency matrix creation)





Performance comparison of shared models

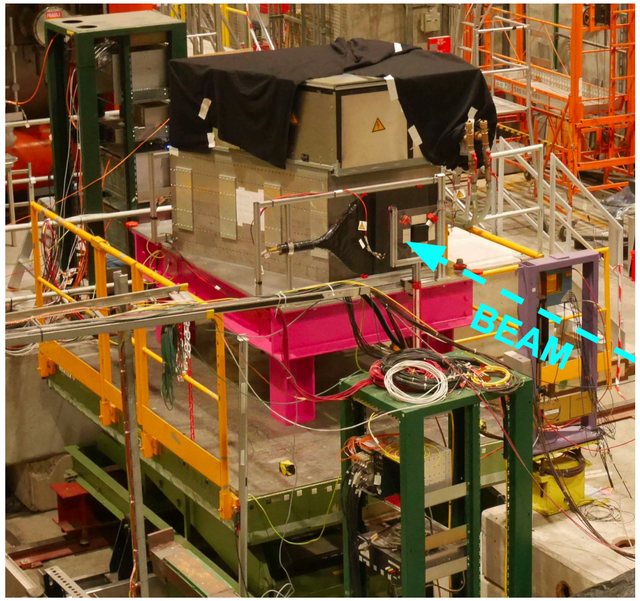


Potential CALICE contribution

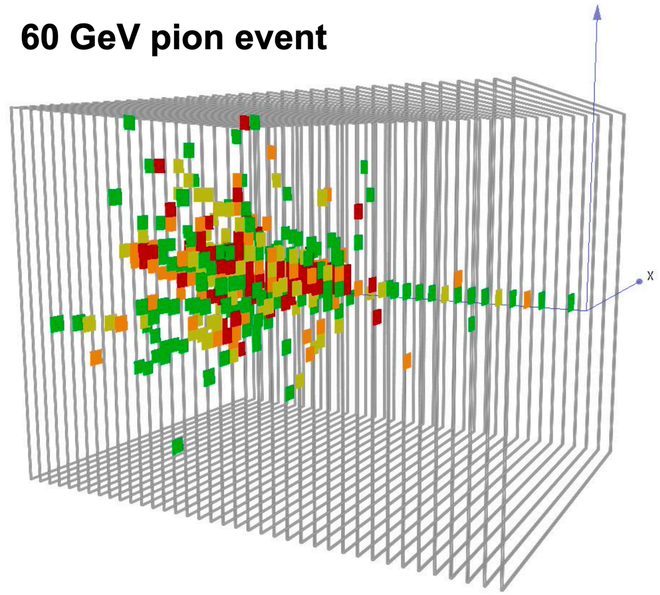
CALICE open data?

- CALICE has many exciting datasets & an increasing amount of data science contributions
- I.e. experimental testbeam data (2018: AHCAL technological prototype at SPS)
  - Quite unusual in the space of data science: Labeled ‘real’ data
  - Attractive dataset for data scientists
- Low level calorimeter event information could be made openly available
  - AHCAL hit information: hit_X, hit_Y, hit_Z, energy, time
  - (Potentially calculated high-level event variables as well)
- Potential ML tasks: Regression (energy reconstruction)

AHCAL technological prototype @ SPS in 2018

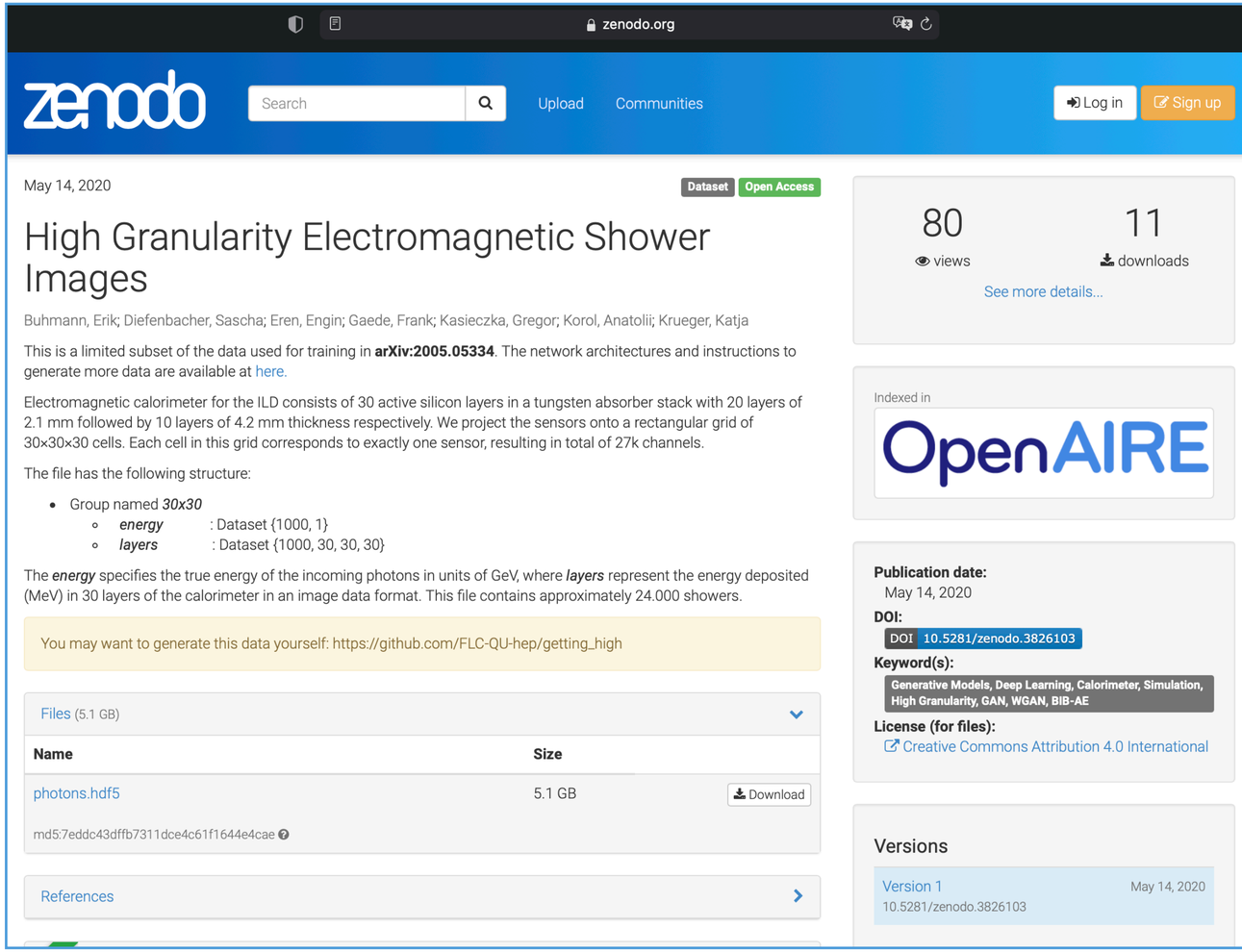


60 GeV pion event



Open Data on Zenodo

- Open science platform for research data hosted by CERN since 2013
- Used by scientists of many domains
- Can include multiple files for one dataset (i.e. root files and numpy files) and allows for version control
- Citable via Digital Object Identifier (DOI)
- Independent of our pd4ml library (but pd4ml can download from Zenodo)



https://doi.org/10.5281/zenodo.3826103

CALICE open data for PD4ML

A CALICE contribution could include:

- Experimental AHCAL test beam runs from the June 2018 at SPS
- 10-80 GeV pion showers in AHCAL
    - With event selection & calibration from corresponding publication
- Showers in the form of low-level hit information, labels with beam energy
    - Upload to Zenodo as root files and as numpy arrays
      (for the pd4ml Python package)
- Reference ML model performing an exemplary task
  (i.e. energy reconstruction)

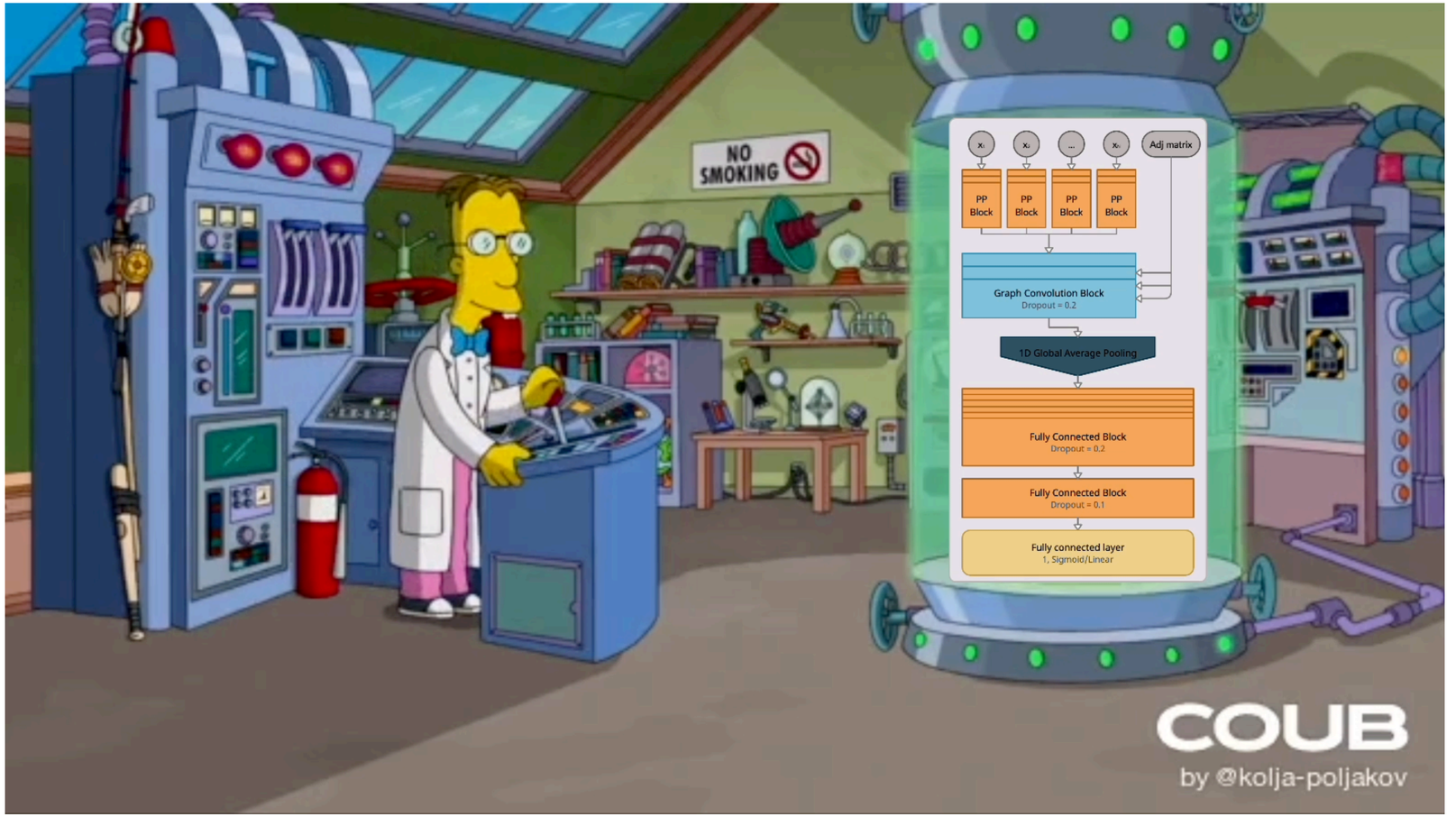
Conclusions

- The pd4ml package allows easy loading of fundamental physics datasets
- Provides space for model comparisons and cross-disciplinary benchmarking
- Includes shared ML models well performing over multiple datasets
- Looking forward to expanding the pd4ml library further

—> CALICE contribution possible via open data on Zenodo

arXiv:2107.00656 **GitHub: pd4ml**

Import your datasets and/or do your experiments with the pd4ml git repository

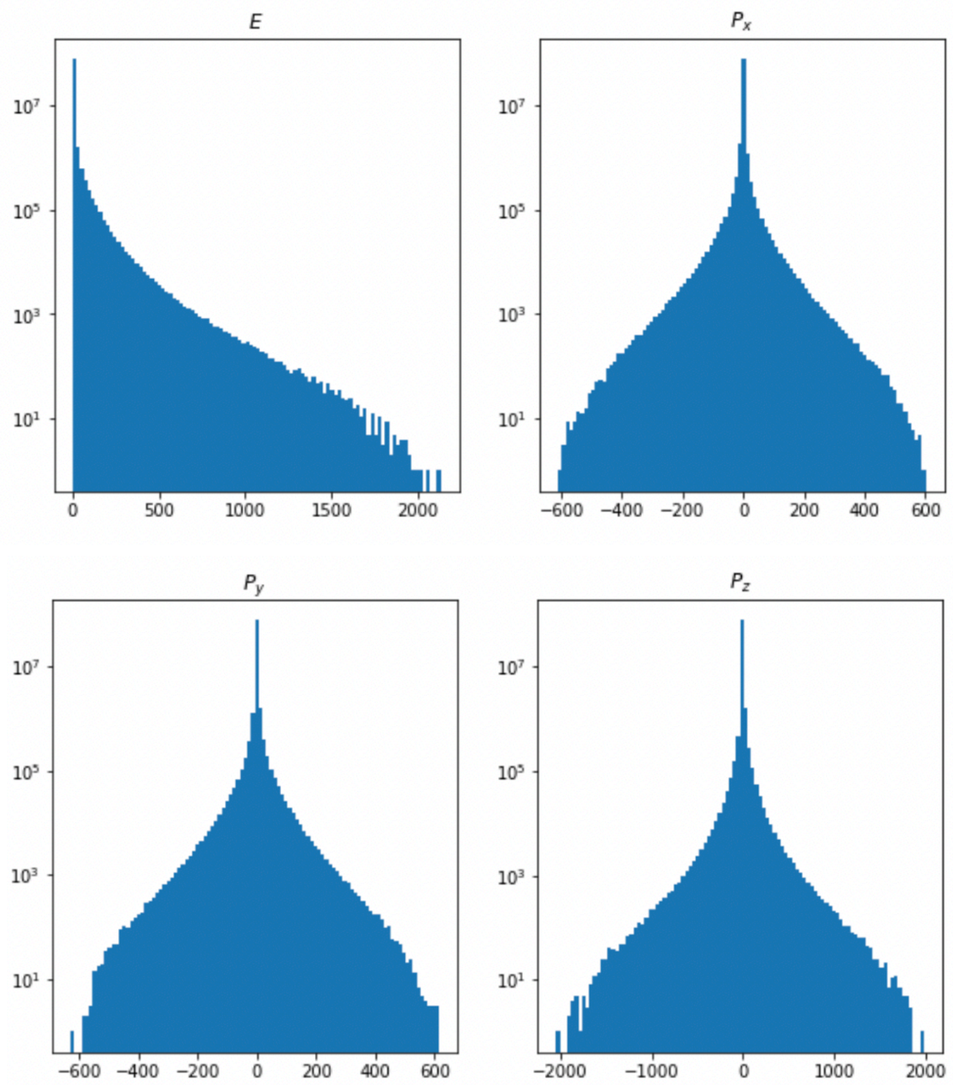


Bonus

The Datasets

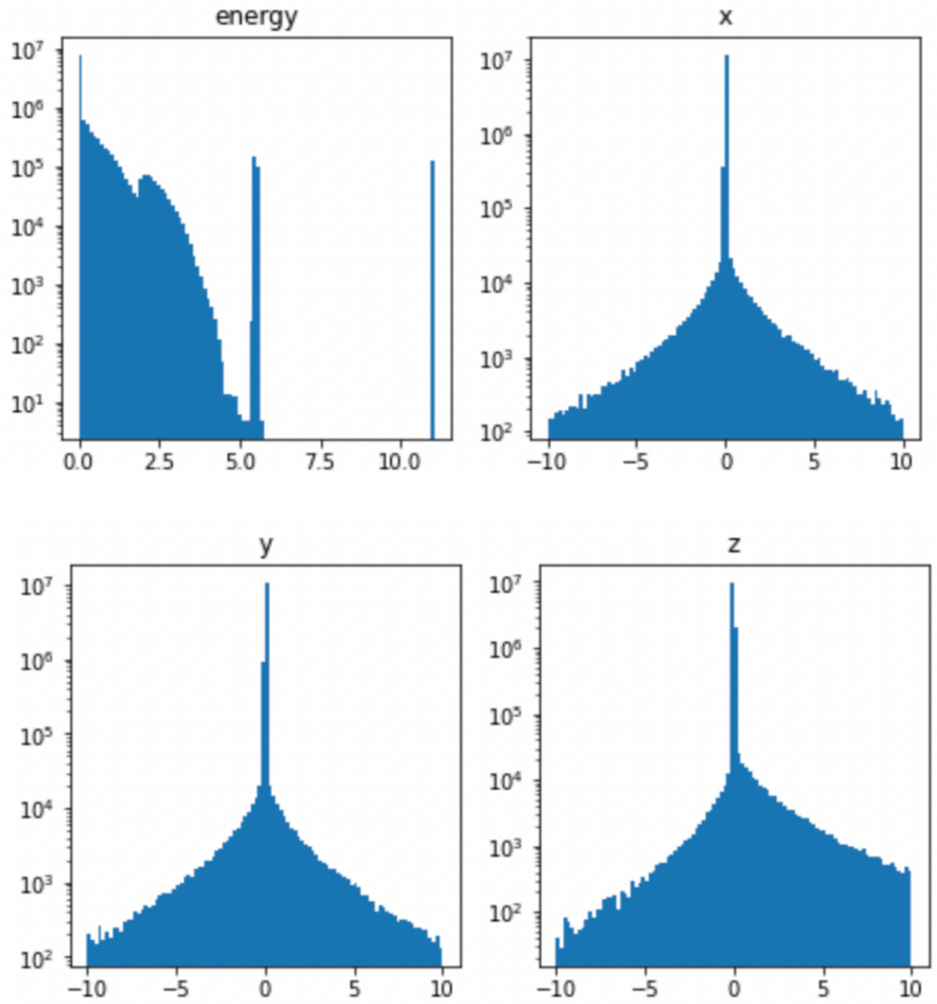
The Datasets: Top Tagging

- 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: ParticleNet (1902.08570)
  - Graph Neural Network acting on the unordered set of jet constituents



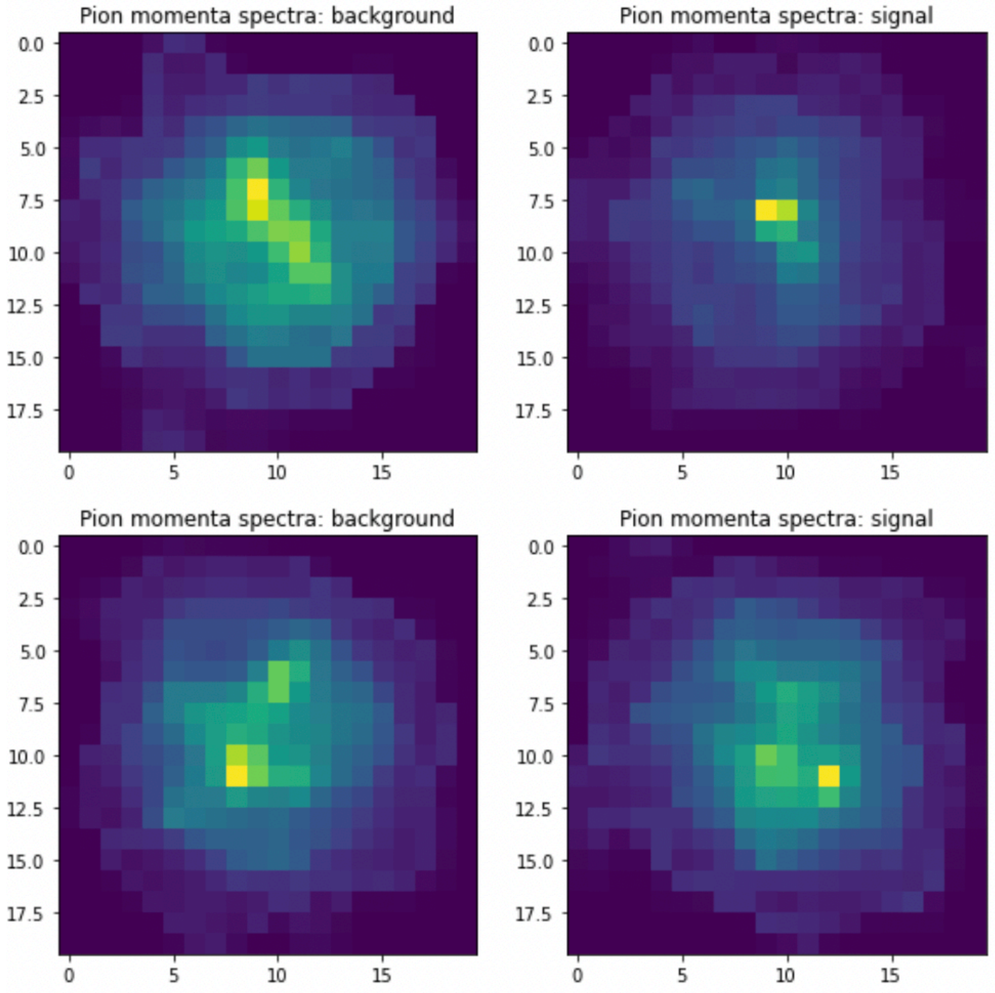
The Datasets: smartBKG (Belle II)

- 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 characterised by 9 features:
  - Production time, E, x, y, z px, py, pz, PID
- Indices of mother particles are used to create adjacency matrix network input
- Reference model: Graph Convolutional Network



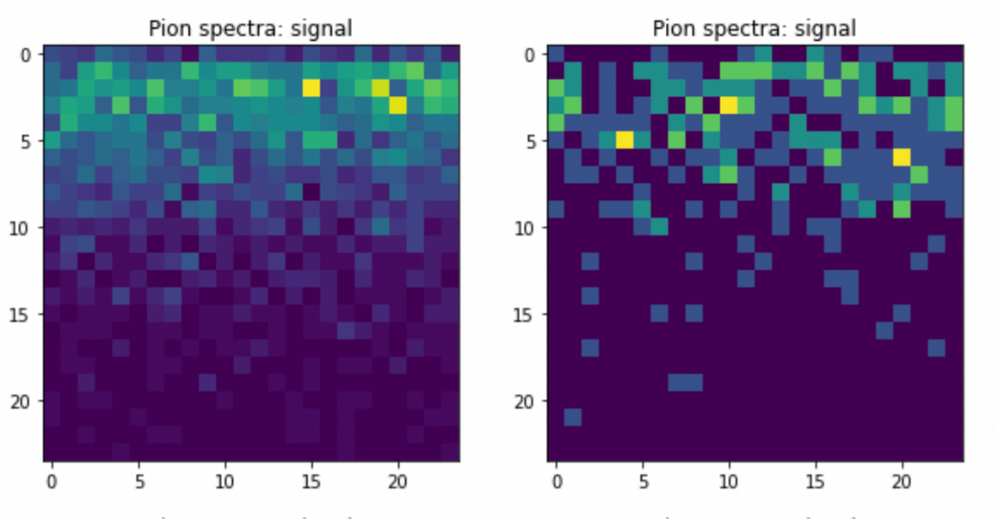
The Datasets: Spinodal

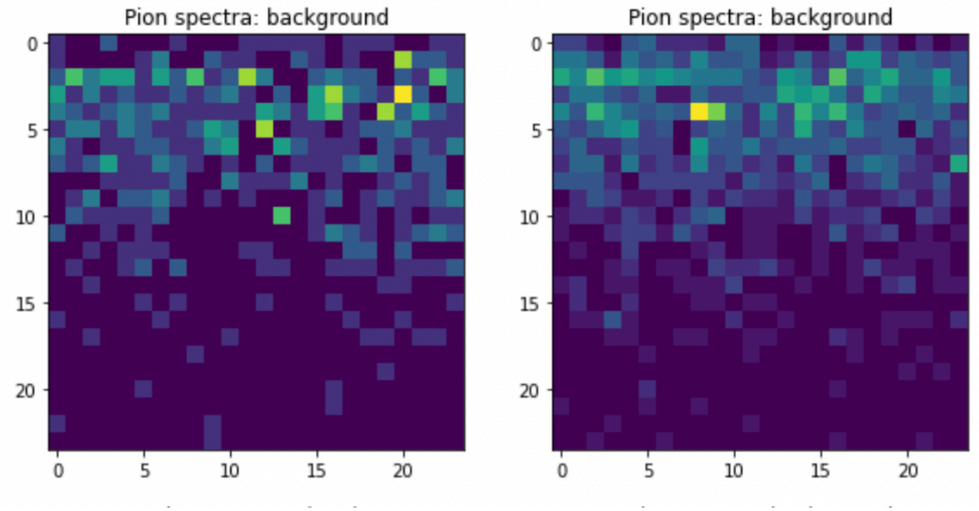
- Classify the nature of QCD phase transitions in heavy ion collisions at the Compressed Baryonic Matter (CBM) experiment
- Dataset is composed of 29,000 2D histograms describing pion momenta
- Reference model: Convolutional Neural Network (1906.06562)



The Datasets: EoS

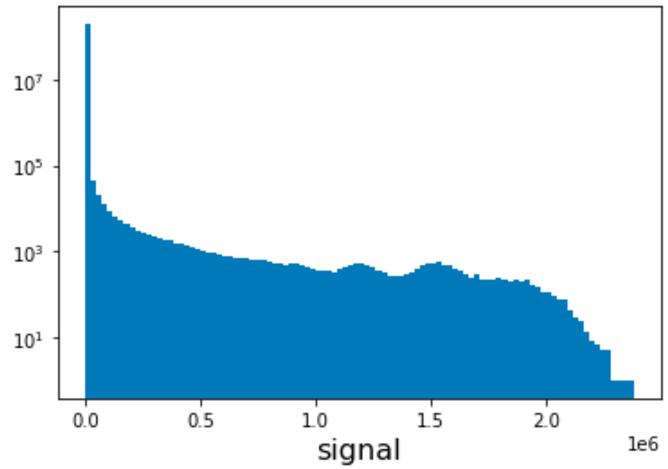
- Classify the QCD transition nature in heavy-ion collisions from the final state pion spectra
- 2 equation of state: cross-over EOSL or 1st order EOSQ
- 180,000 2D histograms of pion spectra
- Data simulated with different parameters for the test set
- Reference Model: Convolutional Neural Network (1910.11530)

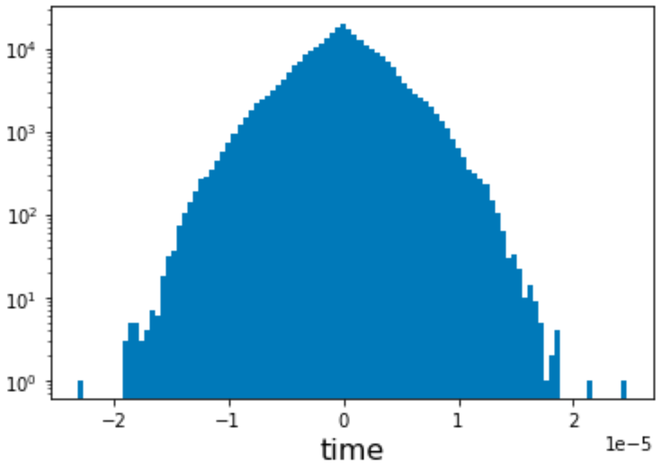


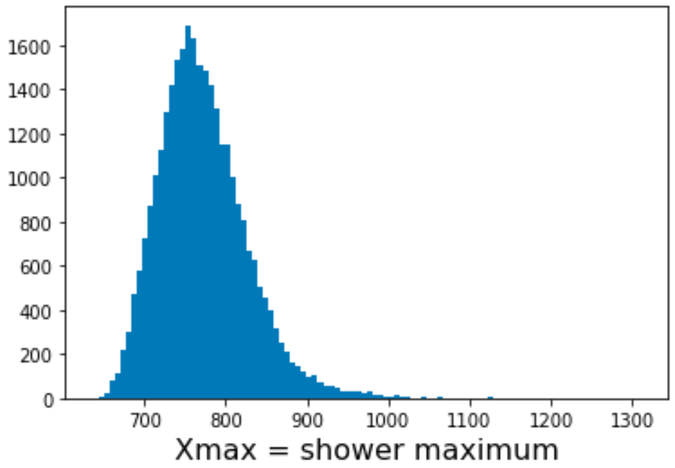


The Datasets: Cosmic-ray induced Air Showers

- Regression task: predict the shower maximum
- 100,000 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



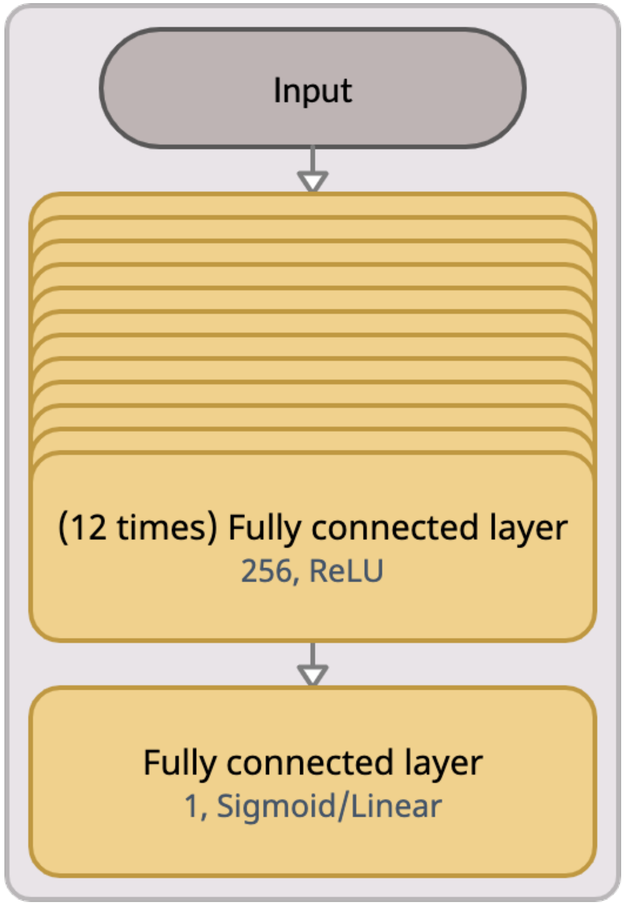




The Models

The Fully Connected Network

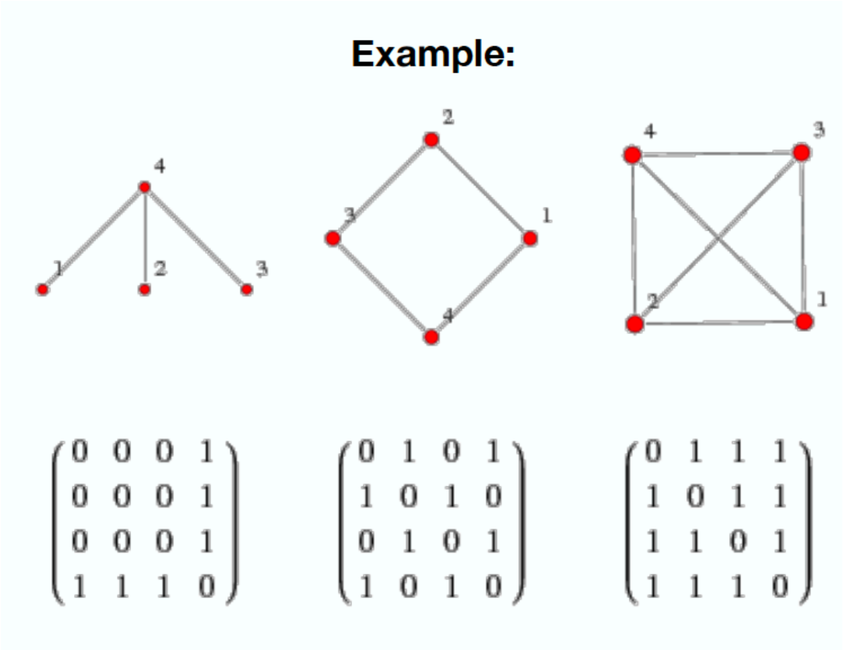
- TensorFlow implementation:
  - Number of inputs changes depending on the dataset
  - Dense layers with 256 nodes each
  - Output layer changes depending on the task:
  - Batch size: 256
  - Loss: BCE or MSE
  - Epochs: 300
  - Learning rate 0.001 with Adam optimizer
- Keras callbacks:
  - Reduce on plateau with patience 8 epochs
  - Early stopping with patience 15 epochs

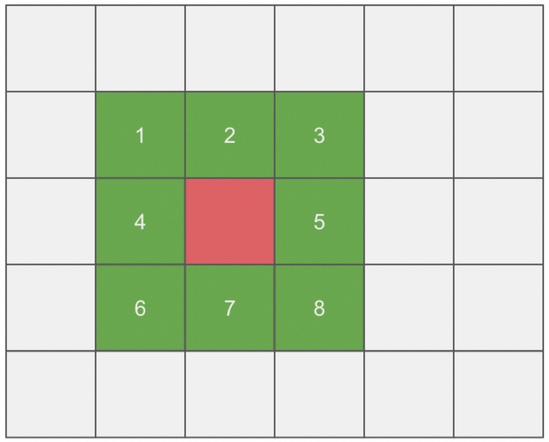


The Graph Network: adjacency matrices

Creating adjacency matrices from the datasets:

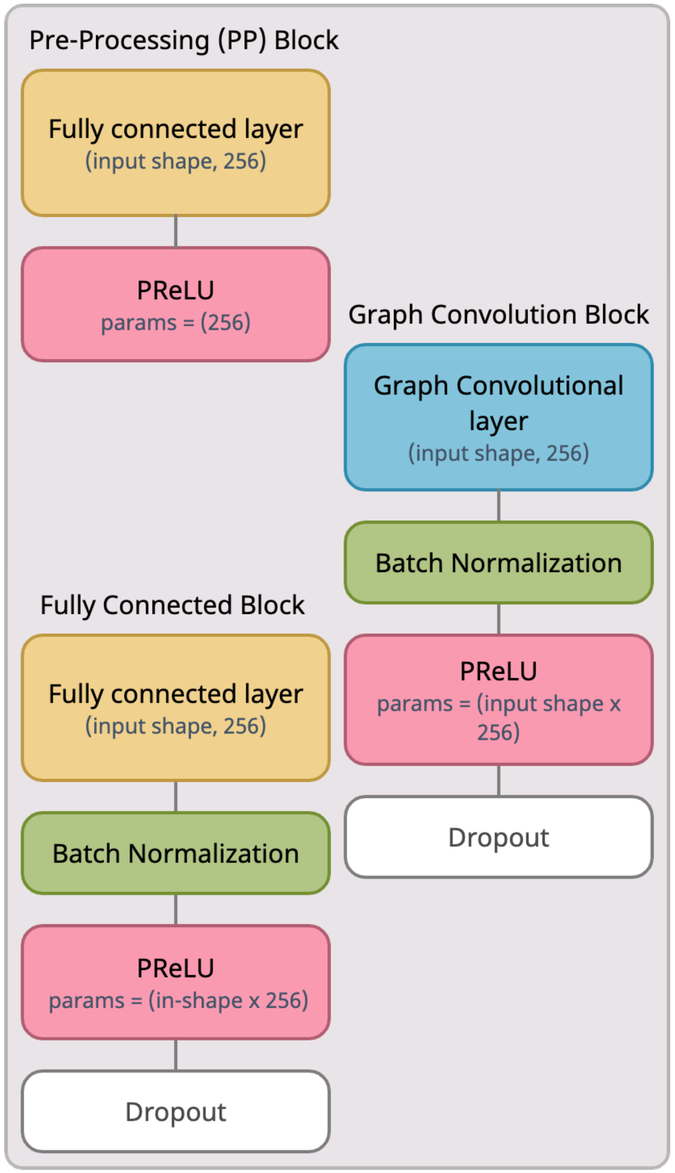
1. TopTagging (jet constituents):
   kNN clustering of particles per events (k = 7)
2. Spinodal & EOS (images):
   8-connected neighboring pixels
3. Belle (jet constituents):
   Matrix with event history via mother & daughter particles (same as Reference Model)
4. Airshowers (signal bins & timing of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





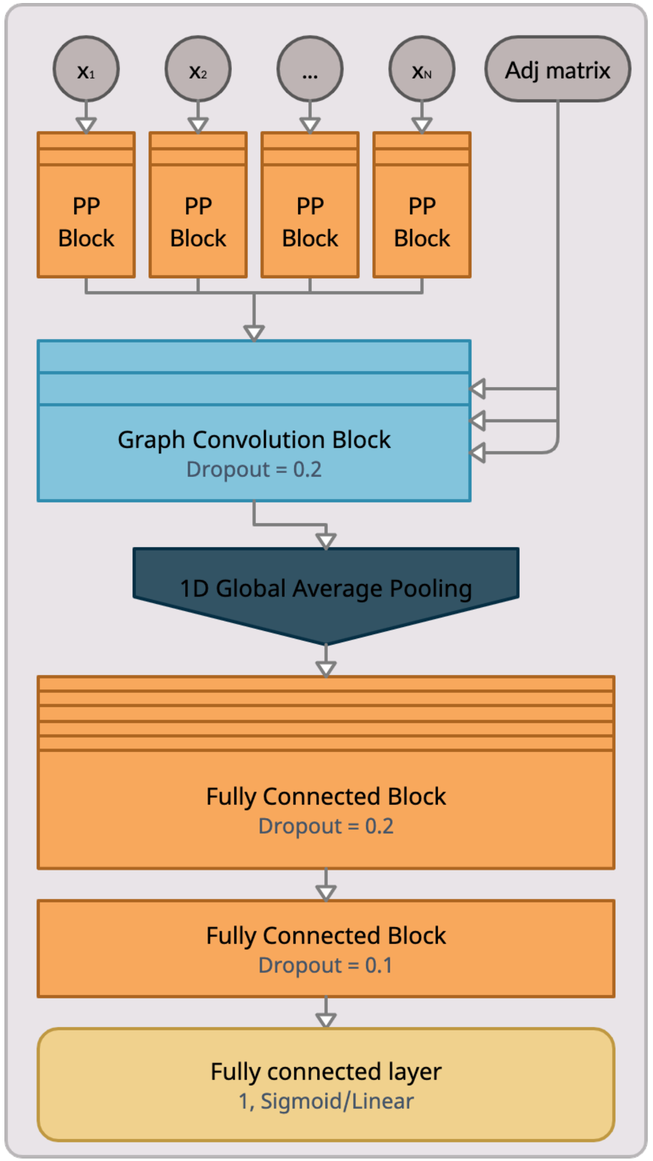
The Graph Network

- Three blocks composing the GraphNet:
  - Pre-Processing Block:
    dense layer followed by a PReLU layer
  - Fully Connected Block:
    a fully connected layer followed by a batchNorm layer, PReLU and Dropout
  - Graph Convolution Block:
    A graph convolutional layer followed by a batchNorm layer, PReLU and Dropout



The Graph Network

- TensorFlow implementation:
  - Dense layers with 256 nodes each
  - **Graph Convolution layers (1609.02907) with 256 nodes each**
  - GlobalAvgPool1D
  - Output layer changes depending on the task
  - Batch size: 32
  - Loss: BCE or MSE
  - Epochs: 400
  - Learning rate 0.0001 with Adam optimizer
- Keras callbacks:
  - Reduce on plateau with patience 8 epochs
  - Early stopping with patience 50 epochs



Performance comparison

	Top Tagging		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	0.935 ± 0.001	0.983 ± 0	0.854 ± 0.004	0.916 ± 0.005	0.687 ± 0.005	0.766 ± 0.005	0.824 ± 0.001	0.903 ± 0	1185 ± 26	34.12 ± 0.47
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 ± 0.001	0.81 ± 0.001	1816 ± 41	42.6 ± 0.47

Shared Data (currently status)

Five datasets currently available (looking forward to further contributions):

1. “Top Tagging at the LHC”: 1902.09914
    1. Jet constituents (4-vectors) for Top vs QCD jet tagging; Reference: ParticleNet (Graph Network)
2. “Spinodal or not?” : 1906.06562
    1. Images (2d histograms) of pion spectra; Reference: CNN
3. “EOSL or EOSQ” : 1910.11530
    1. Images (2d histograms) of pion spectra; Reference: CNN [test set differs from training set]
4. SmartBKG dataset (Belle II - generated events passing downstream selection): link
    1. Decay graph of particles, Reference: GCN
5. Cosmic Airshowers: publication
    1. 81 1d traces (signals and timing) of air showers for shower maximum reconstruction, Reference: RNN