The AUTOGRAPH pipeline

***A**utomatic **U**nified **T**raining and **O**ptimization for **G**raph **R**ecognition and **A**nalysis with **P**ipeline **H**andling*

Greta Brianti(1), Roberto Iuppa(1), Marco Cristoforetti(2)

16th Topical Seminar on Innovative Particle and Radiation Detector

Siena, 25 – 29 September

(1) Università degli studi di Trento – TIFPA, (2) Fondazione Bruno Kessler – DSIP

16TH TOPICAL SEMINAR ON INNOVATIVE PARTICLE AND RADIATION DETECTORS (IPRD23)

Siena, 25 - 29 September 2023

FONDAZIONE BRUNO KESSLER

INFN Istituto Nazionale di Fisica Nucleare

UNIVERSITÀ DI TRENTO

b–tagging at collider

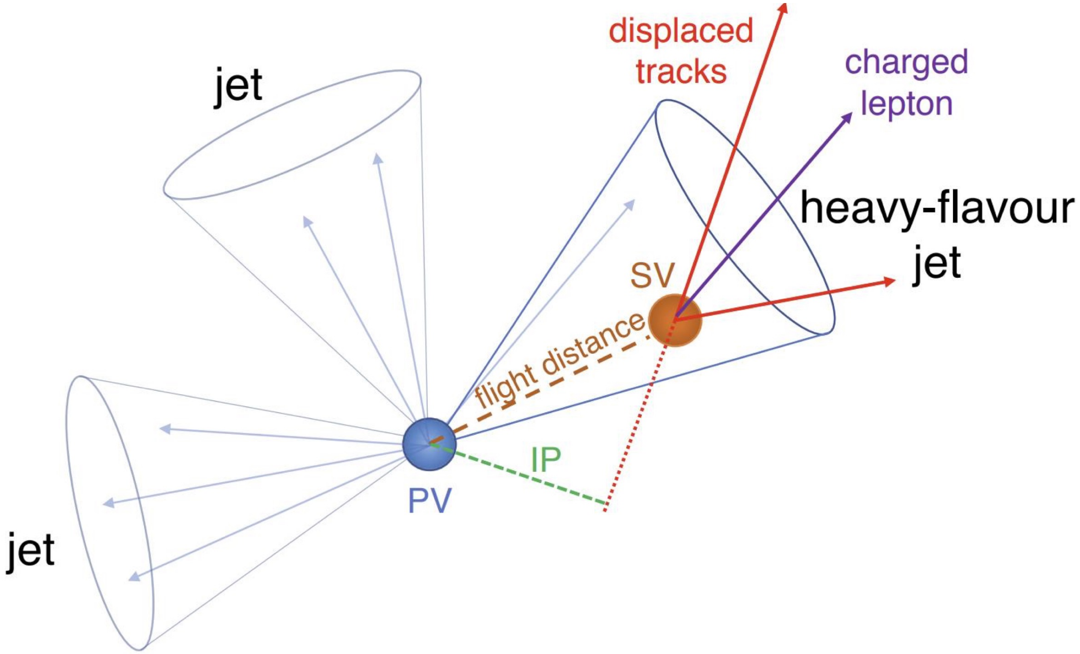
A jet is defined as a **collimated cone** of stable particles arising from fragmentation and **hadronization of a parton** after a collision.

B–hadrons

Bound states involving b–quark

– Unique jet features:

- Measurable lifetimes (~ 1.5 ps)
- Large track impact parameters
- Displaced secondary vertices and tertiary vertices

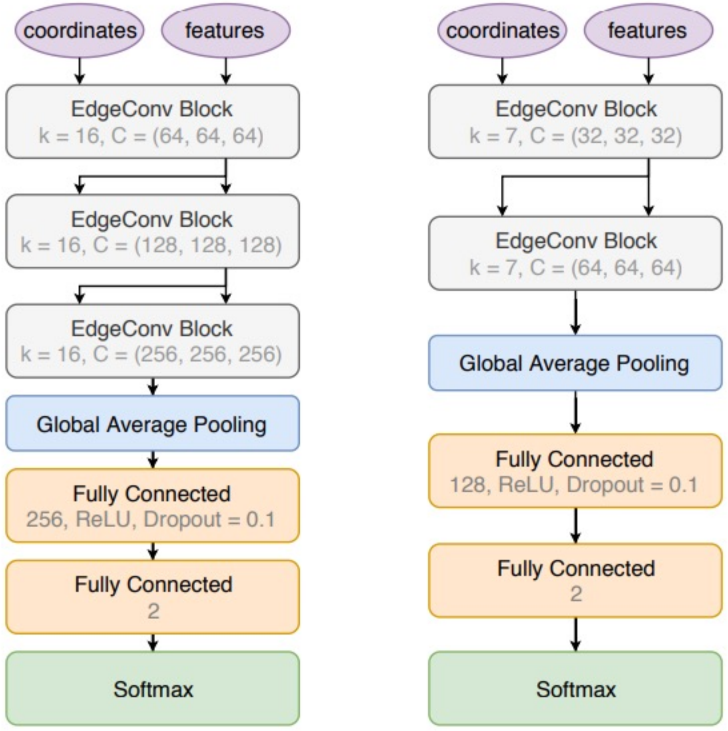


The flavour tagging is of particular importance for the study of the Standard Model (SM) Higgs boson and the top quark and additionally for several Beyond Standard Model (BSM) resonances.

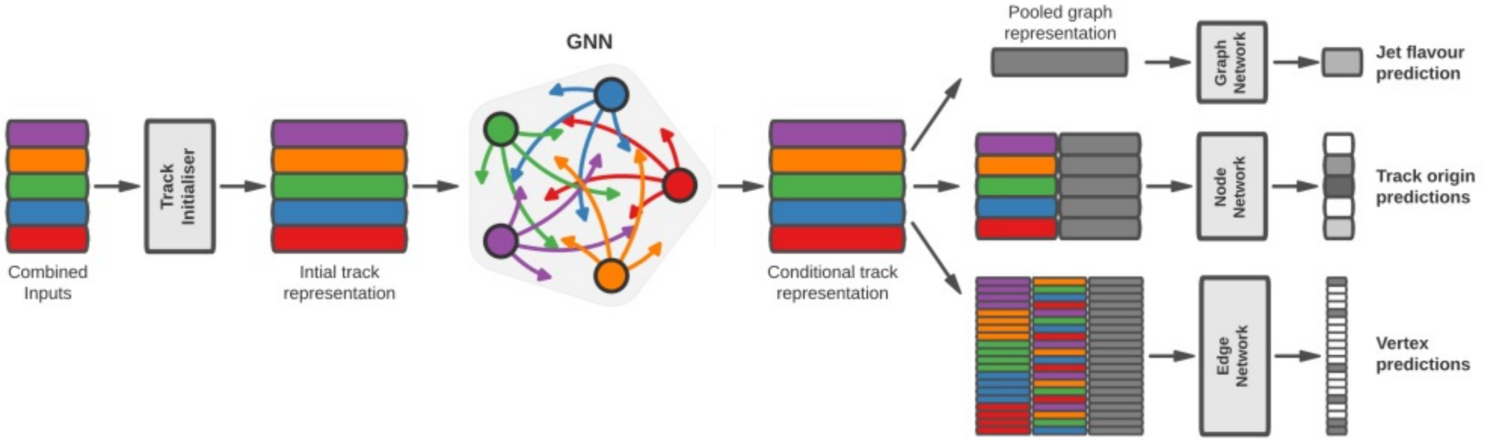
GNNs in HEP

CMS experiment – ParticleNet tagger [1]

ATLAS experiment – GN1 tagger [2]



(a) ParticleNet (b) ParticleNet-Lite



Graph Neural Networks (GNNs) are a machine–learning based tools which exploit the physical structure of the jet to identify the originating parton.

CMS and ATLAS, the LHC general purpose experiments, apply GNNs to flavour tagging. Moreover, GNN algorithms are applied to offline analysis for example as background/signal classificator. [3]

The AUTOGRAPH pipeline

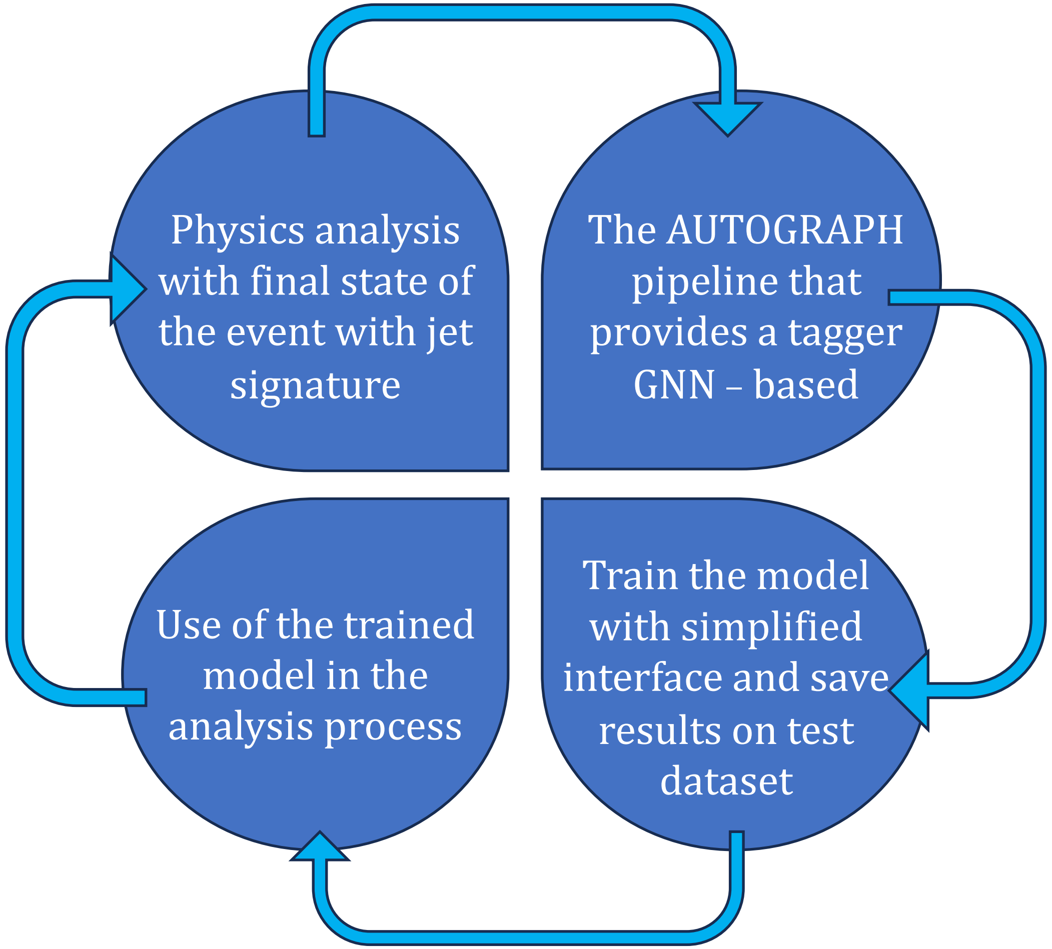
A tagging GNN–based method for all

Why?

The increasing use of GNNs needs a framework to accelerate the optimization and evaluation processes for the physicist that would like to implement a machine learning - based tagger in their analysis.

How?

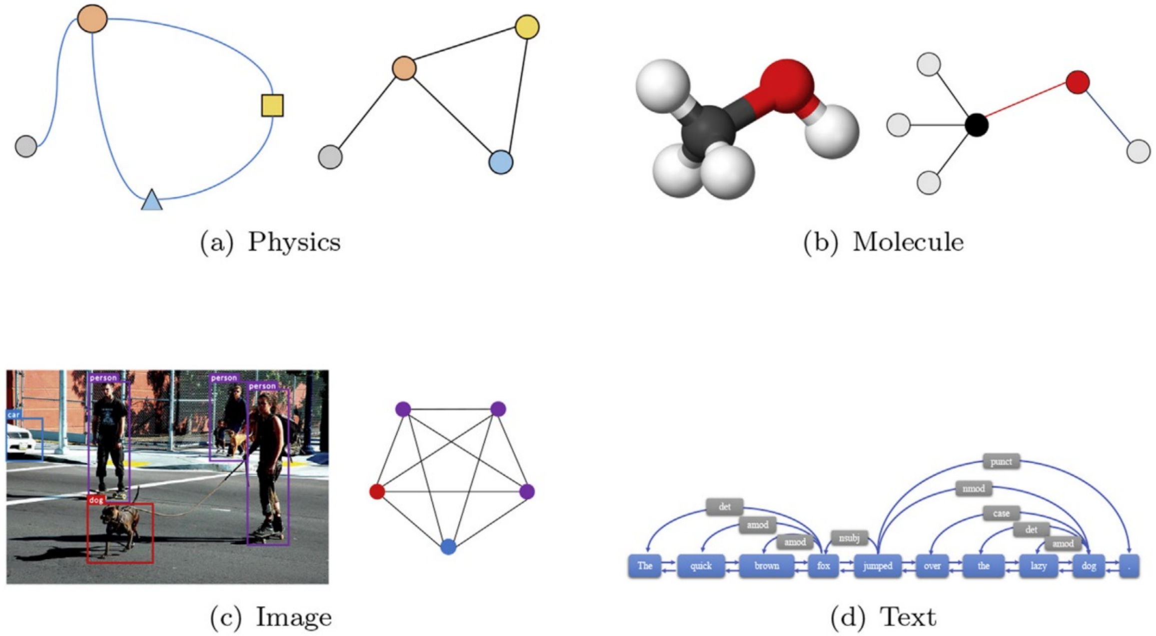
The user sets the configuration file which is provided in the pipeline, wherein he/she can choose the dataset setting, the network architecture and the training setting. Moreover, he/she can acquire additional physical information from the feature ranking or the discriminant plot production.



Graph Neural Networks (GNNs)

From the social network to a wide landscape of possible applications

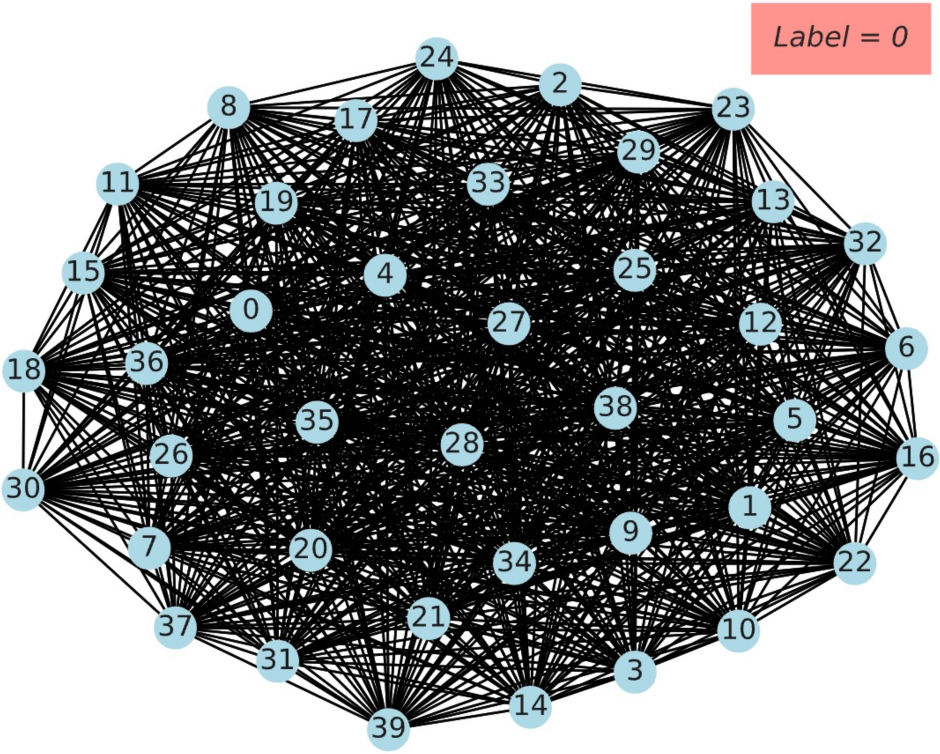
Graph of a social network



Other possible applications

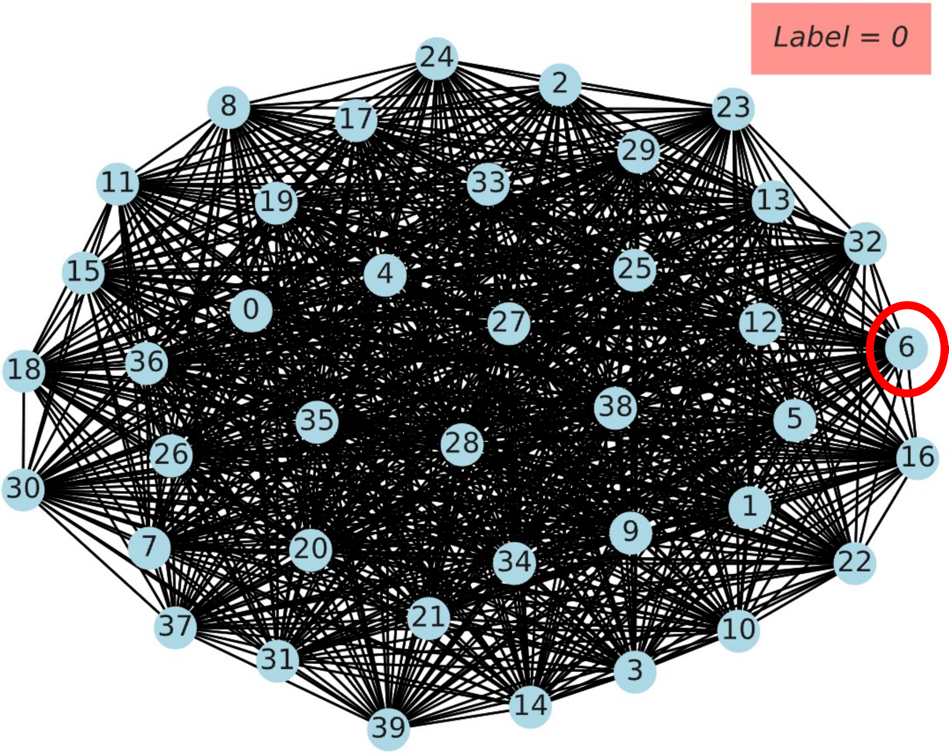
GNNs for b–tagging

A single jet can be represented as a GRAPH



GNNs for b–tagging

A single jet can be represented as a GRAPH

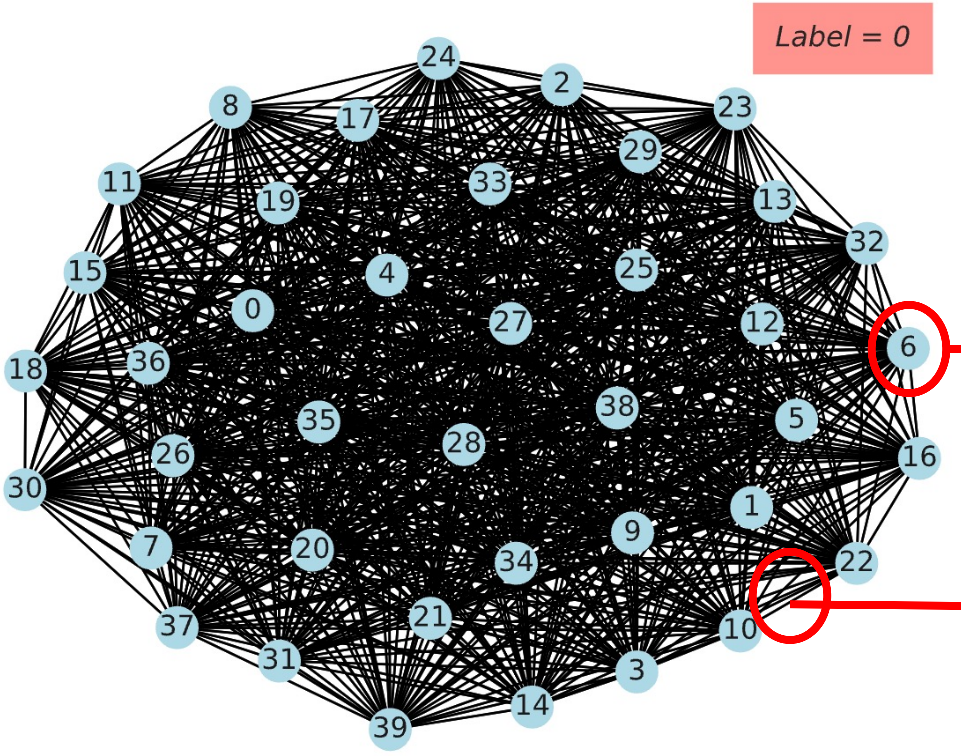


A single track associated to the jet is a node of the graph

GNNs for b–tagging

Topical Seminar on Innovative Particle and Radiation Detectors (IPRD23)
Siena, 25 - 29 September 2023

A single jet can be represented as a GRAPH

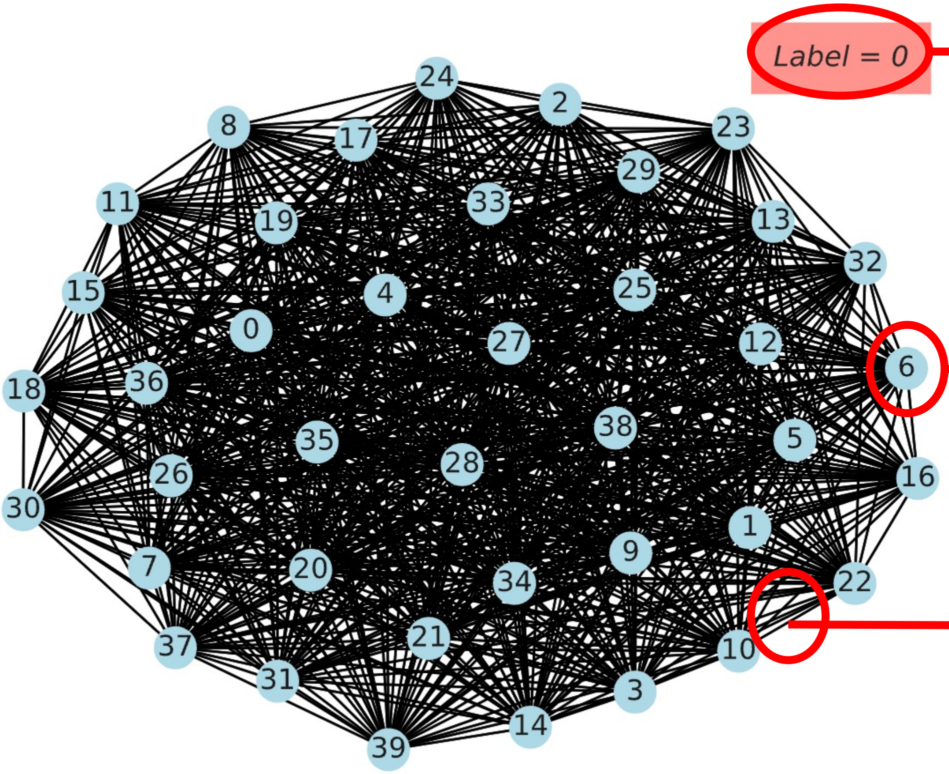


A single track associated to the jet is a node of the graph

Each track is linked with edges with all the other tracks (Fully connected graph)

GNNs for b–tagging

A single jet can be represented as a GRAPH



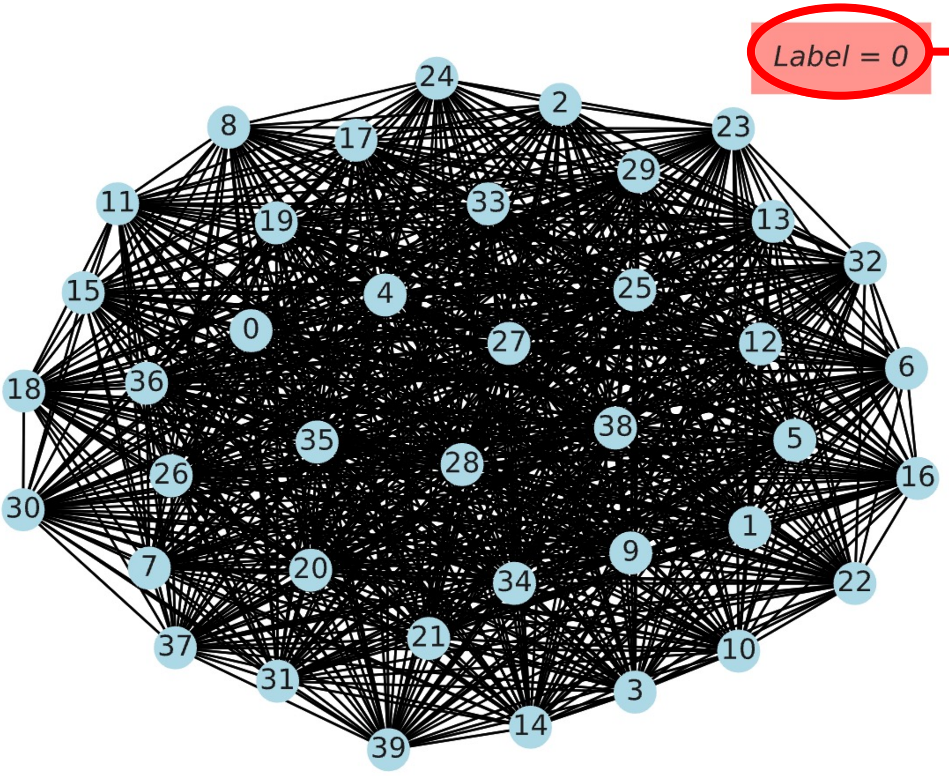
From the MC simulation, the graph is labelled with the truth flavour index (in this case a light jet)

A single track associated to the jet is a node of the graph

Each track is linked with edges with all the other tracks (Fully connected graph)

GNNs for b–tagging

A single jet can be represented as a GRAPH



From the MC simulation, the graph is labelled with the truth flavour index (in this case a light jet)

The network learns its parameters during the training with the MC dataset.

Once the network is trained its performance can be evaluated on data

The MC must represent the real data as realistically as possible.

Dataset pre–processing

The pipeline is designed to be compatible with input binary file, but a set of script to convert dataset from different formats to binary are provided.

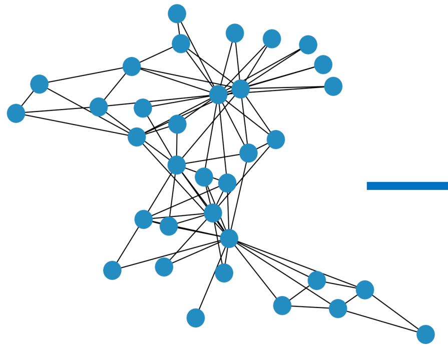
Dataset setting in the user interface

- Number of events to be processed from the binary file
- Train and test dataset fraction
- Number of nodes per graph that is the number of associated tracks to a single jet
- Number of variables per node that represents the tracks associated to the jet
- Number of global variables that corresponds to the jet features

First automated step: from n–tuples to graphs

INPUT

n–tuples of variables in which each line collects the jet and the associated tracks features in binary format

→  →

OUTPUT

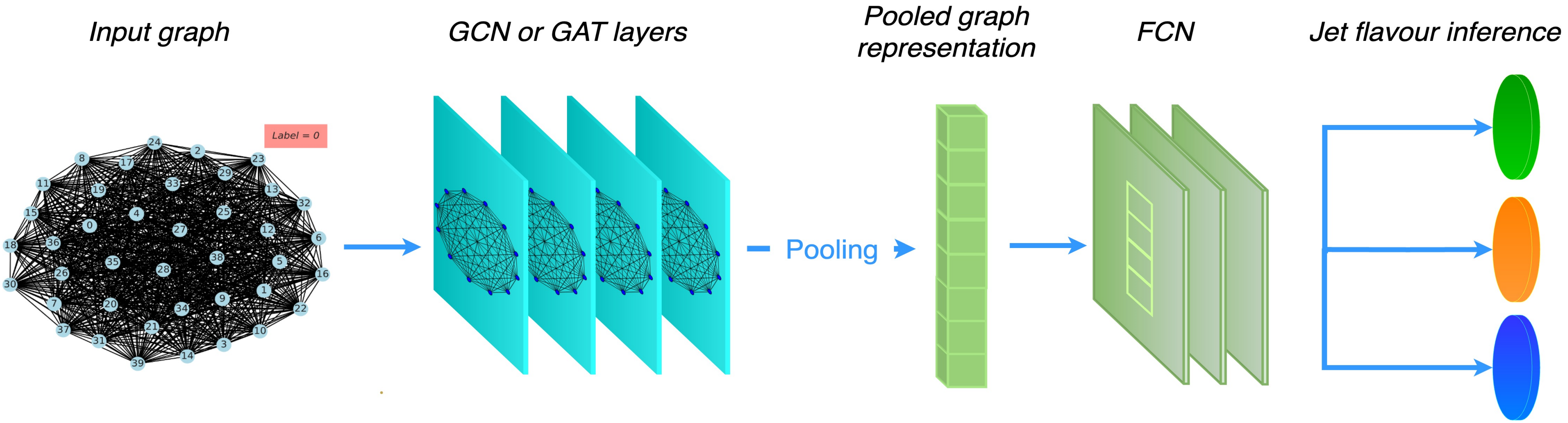
Training and validation datasets in Pytorch Geometric **`DataLoader`** format

Network architecture

Network setting in the user interface

- Graph layers type (Graph Convolutional Layer or Graph Attention Layer)
- Number of graph layers
- Number of hidden nodes per graph layer
- Pooling function
- Number of linear layers

Second automated step: network construction

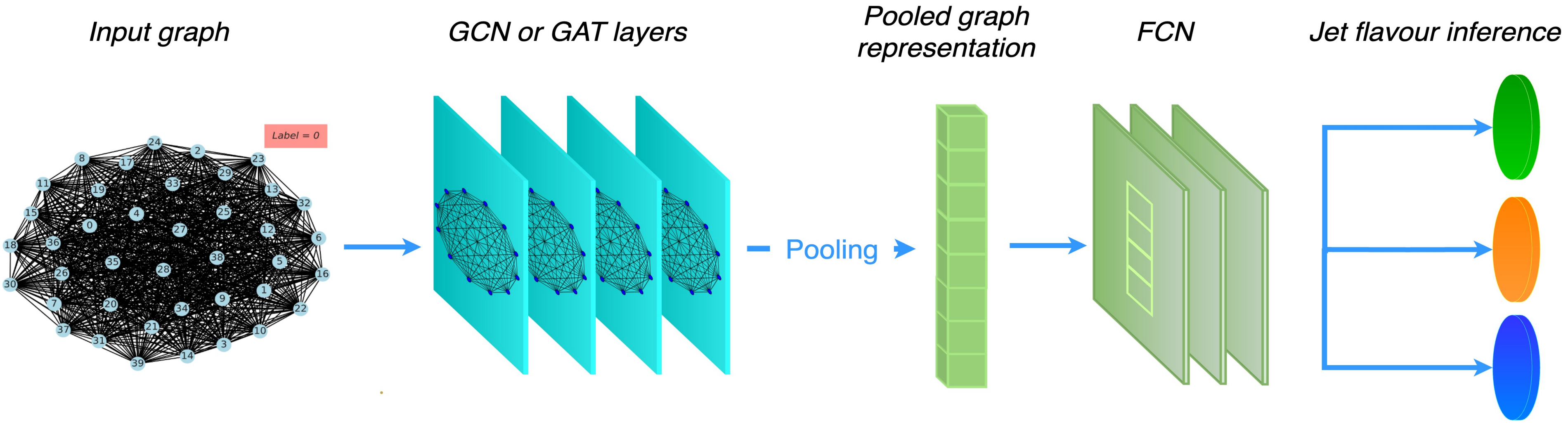


Network architecture

Network setting in the user interface

- Graph layers type (Graph Convolutional Layer or Graph Attention Layer)
- Number of graph layers
- Number of hidden nodes per graph layer
- Pooling function
- Number of linear layers

Second automated step: network construction

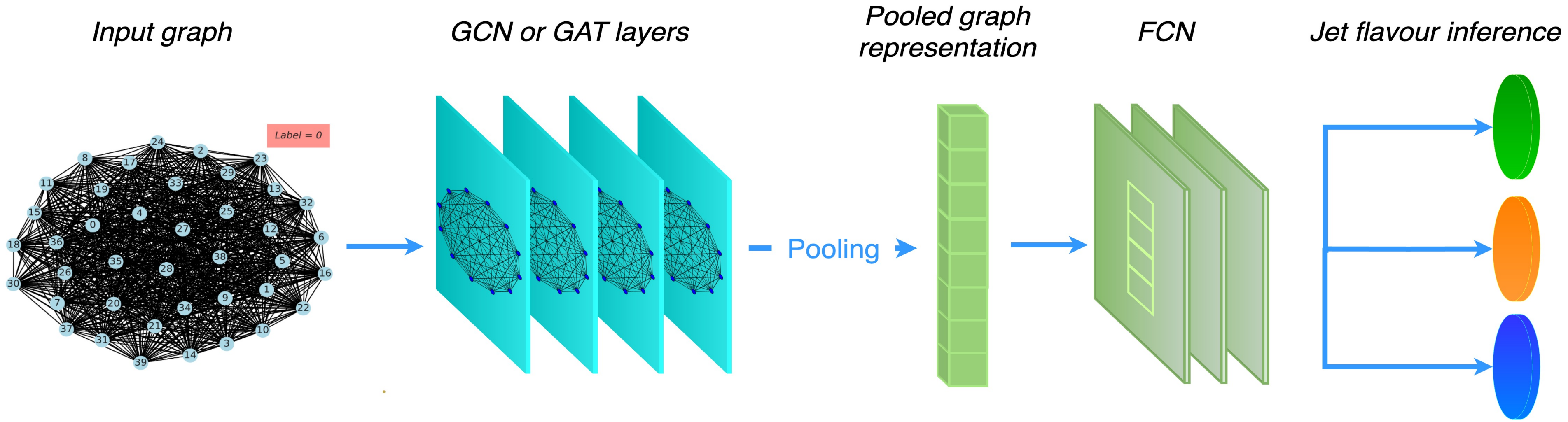


Network architecture

Network setting in the user interface

- Graph layers type (Graph Convolutional Layer or Graph Attention Layer)
- Number of graph layers
- Number of hidden nodes per graph layer
- Pooling function
- Number of linear layers

Second automated step: network construction



Training setting

Training and storing setting in the user interface

- Number of epochs which the network has to be trained
- Learning rate (the step size at each iteration while moving toward a minimum of a loss function)
- Batch size (number of samples processed before the model is updated)
- Save performance (output of network, discriminant values, tagging efficiency, background rejection)
- Plot production (variables to plot, plot setting)

Third automated step: training and performance evaluation

Training of the customized model

INPUT

- Datasets
- Network

→

OUTPUT

Network inference on training and validation dataset

→ Performance evaluation
Loss function, discriminant

→ Feature ranking

The AUTOGRAPH pipeline

Topical Seminar on Innovative Particle and Radiation Detectors (IPRD23)
Siena, 25 - 29 September 2023

A tagging GNN–based method for all

User interface

Configuration file

Dataset setting

```
– Input file
– Number of events
– Train dataset fraction
– Number of variables per node
– Number of global variables
– Number of nodes per graph
```

Network setting

```
– Graph layers type
– Number of layers
– Number of hidden nodes
– Pooling layer
```

Training setting

```
– Epochs
– Learning rate
– Batch size
```

Storing setting

```
– Save performance
– Plot production
– Output directory
```

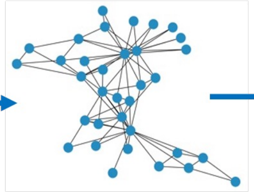
Automated steps

1. Pre-processing

Dataset setting

INPUT

n–tuples of variables in which each line collects the jet and the associated tracks features

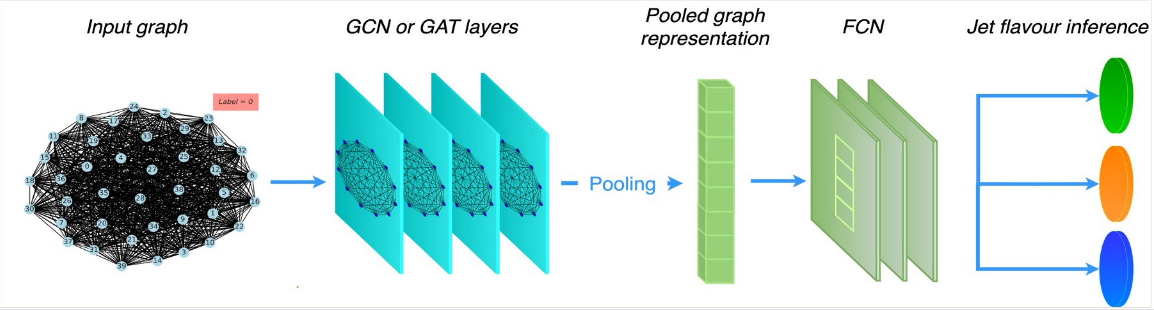


OUTPUT

Training and validation datasets in Pytorch Geometric **`DataLoader`** format

2. Network architecture

Network setting



3. Train and validation

Training setting

INPUT

– Datasets
– Network

OUTPUT

Network inference

Performance evaluation
Loss function, discriminant

Feature ranking

Simulated dataset

Madgraph5_aMC@NLO

Computations of cross sections, generation of hard events and matching with event generators.

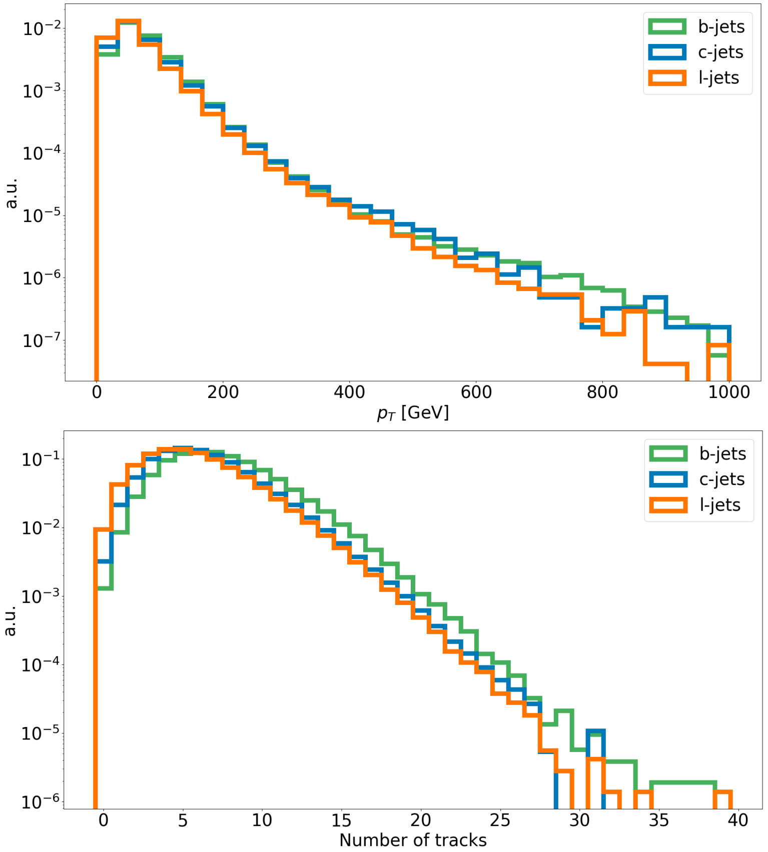
Pythia 8.302

A general-purpose Monte Carlo event generator. Generation of high-energy physics collision events and hadronization.

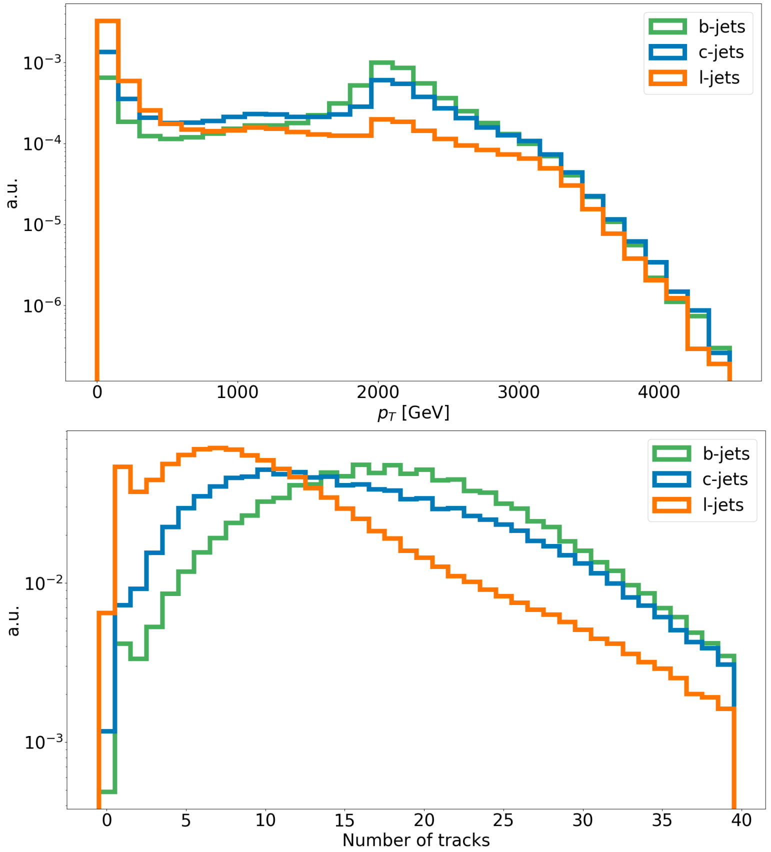
Delphes 3.5.0

Fast multipurpose detector response simulation. ATLAS detector card has been used for both the samples.

$t\bar{t}$ next-to-leading order



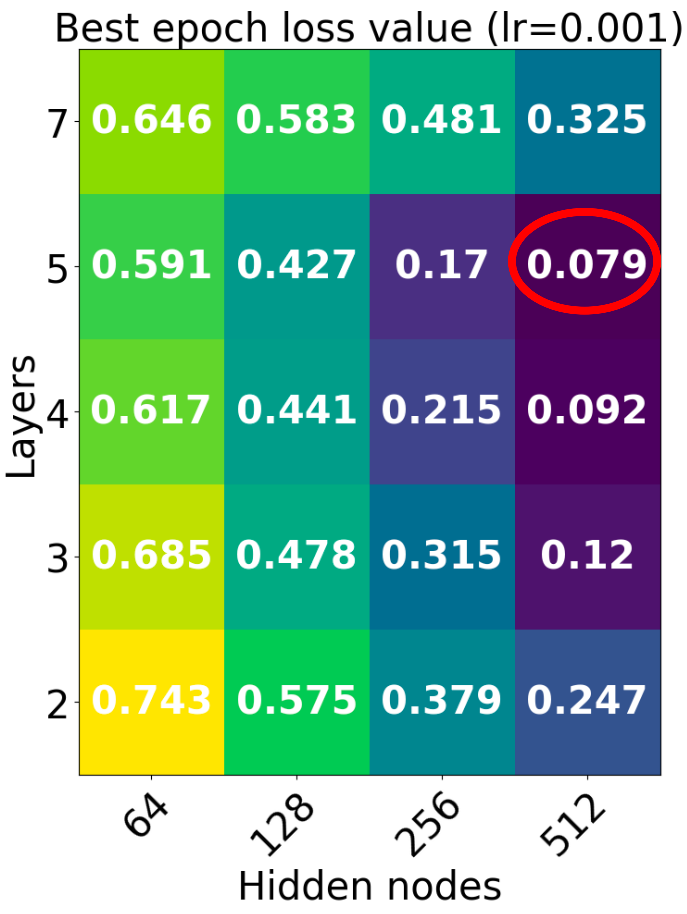
$Z'H$ leading order



Grid search

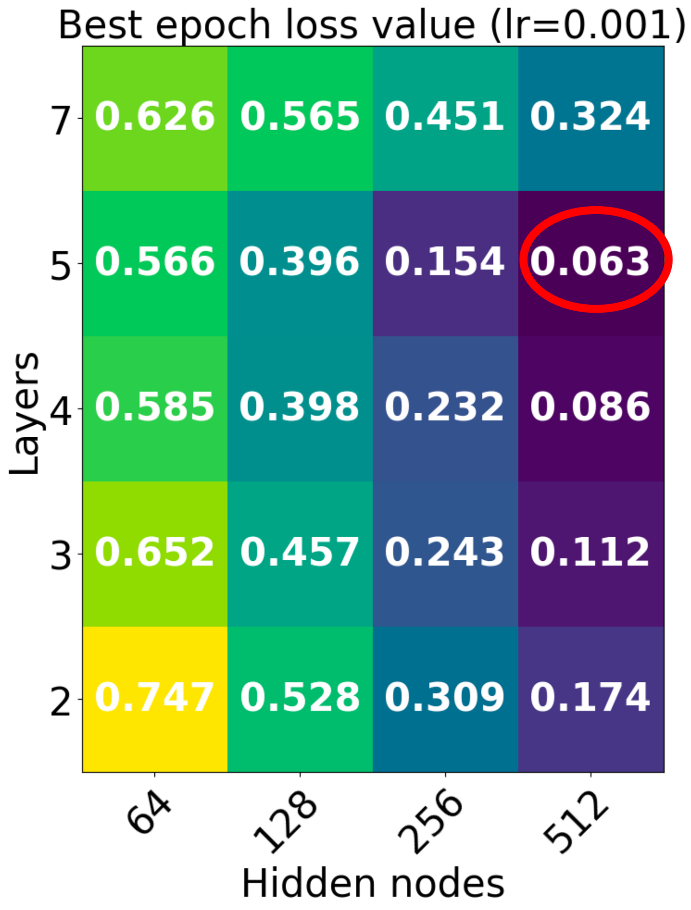
16TH Topical Seminar on Innovative Particle and Radiation Detectors (IPRD23)
Siena, 25 - 29 September 2023

$t\bar{t}$ dataset



Layers	64	128	256	512
7	0.646	0.583	0.481	0.325
5	0.591	0.427	0.17	0.079
4	0.617	0.441	0.215	0.092
3	0.685	0.478	0.315	0.12
2	0.743	0.575	0.379	0.247

Z'H dataset



Layers	64	128	256	512
7	0.626	0.565	0.451	0.324
5	0.566	0.396	0.154	0.063
4	0.585	0.398	0.232	0.086
3	0.652	0.457	0.243	0.112
2	0.747	0.528	0.309	0.174

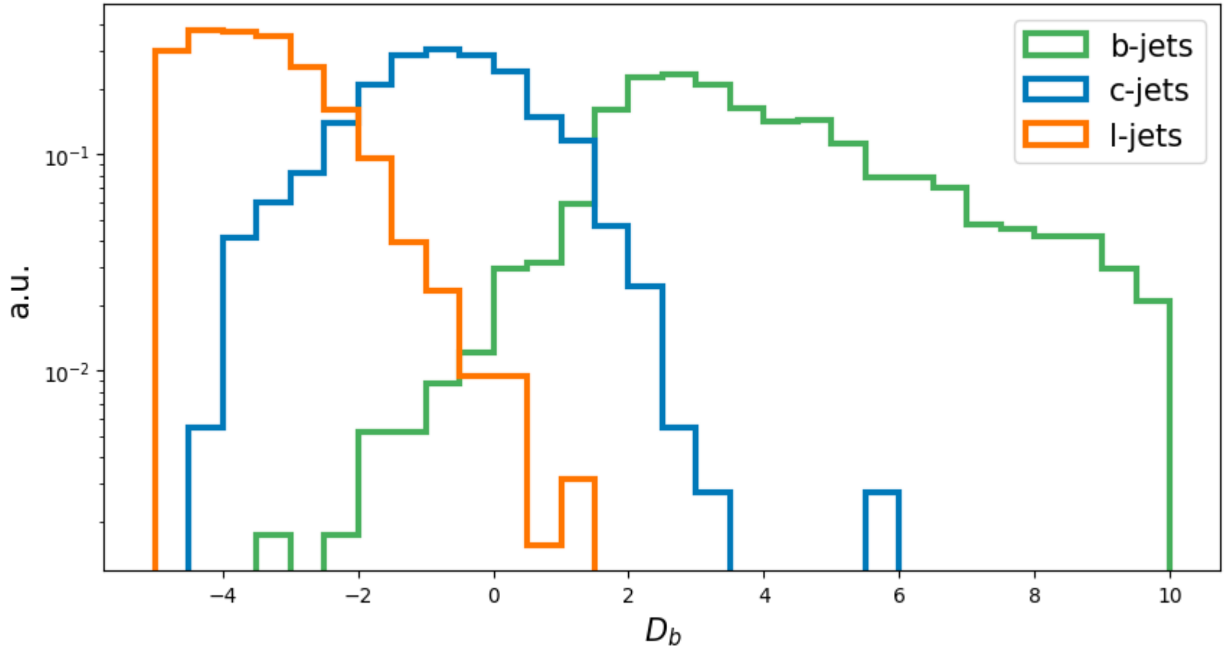
Each model has been trained for 500 epochs with the learning rate value fixed to 10^{-3} and a batch size of 500.

These results are obtained on a sample of 10'000 events.

For both the simulated datasets the best architecture corresponds to 5 Graph Convolutional Layers and 512 hidden nodes.

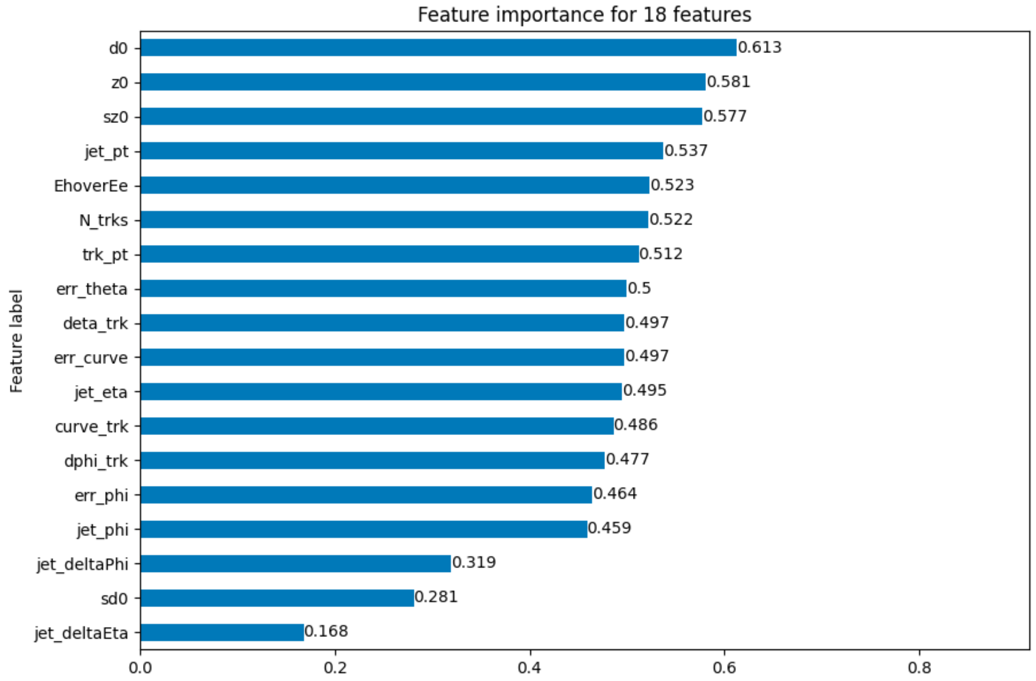
Results and feature ranking

Discriminant (Z'H sample)



$$D_b = \log\left(\frac{p_b}{p_c f_c + (1 - f_c) p_u}\right)$$

Feature ranking (Z'H sample)



Pytorch Explainer class [4]

Conclusion

- The GNNs are applied at general purpose experiments of LHC as flavour tagger and they play a significant role in identification of track and vertex characteristics.
- The AUTOGRAPH pipeline is designed to be applied to a wide range of analysis.
- The pipeline is easily customizable and it could provide additionalphysical information along with an advanced GNN–based flavour tagger.
- The user has the possibility to set the whole structure of the jet–graph representation, the features to use and the network architecture.

For the future:

- Improve and verify the ranking function.
- Apply the pipeline to a physics analysis.
- Add customizaible options (i. e. different number of nodes per layer).

Conclusion

- The GNNs are applied at general purpose experiments of LHC as flavour tagger and they play a significant role in identification of track and vertex characteristics.
- The AUTOGRAPH pipeline is designed to be applied to a wide range of analysis.
- The pipeline is easily customizable and it could provide additionalphysical information along with an advanced GNN–based flavour tagger.
- The user has the possibility to set the whole structure of the jet–graph representation, the features to use and the network architecture.

For the future:

- Improve and verify the ranking function.
- Apply the pipeline to a physics analysis.
- Add customizaible options (i. e. different number of nodes per layer).

Thank you for the attention!

References

[1] Jet Tagging via Particle Clouds - CMS

[2] ATL-PHYS-PUB-2022-027

[3] Graph Neural Networks at the Large Hadron Collider

[4] Pytorch Geometric - Explainer

Backup

Greta Brianti

109° SIF National Congress

Salerno – 11/15 september

TOPICAL SEMINAR ON INNOVATIVE PARTICLE AND RADIATION DETECTORS (IPRD23)

Siena, 25 - 29 September 2023

FONDAZIONE BRUNO KESSLER

INFN Istituto Nazionale di Fisica Nucleare

UNIVERSITÀ DI TRENTO