

# Open-data Tutorial for ATLAS ML jet-substructure

Robert Les, Michigan State University  
on behalf of the ATLAS Collaboration

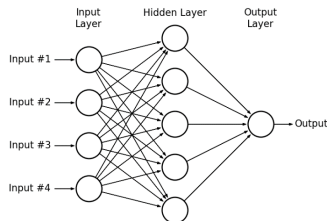
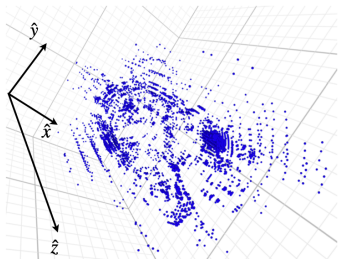
COMETA, Amsterdam  
Oct 1, 2024

# Introduction

We are well into the ML era of HEP!

ML used in all levels:

- Classification for low-level object definition
- Regression for calibration
- Density estimation for backgrounds
- Analysis specific signal/background separation
- Likelihood estimation and hypothesis testing



This tutorial will provide instructions on how to access and use the new ATLAS open-data for substructure studies of top vs QCD jets

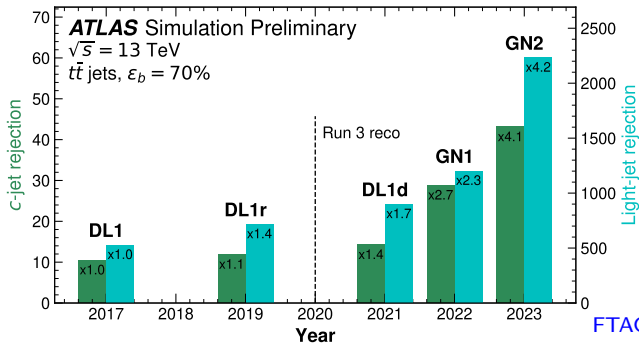
- [Tutorial Github](#)
- [Open-data](#)
- Paper: [JINST 19 \(2024\) P08018](#)

Let's first explain the motivation for this specific-case open-dataset

# Historical Perspective on Jet-Tagger Progress

The trend for boosted  $W/Z/t/h \rightarrow R = 1.0$  jet identification (jet-tagging):

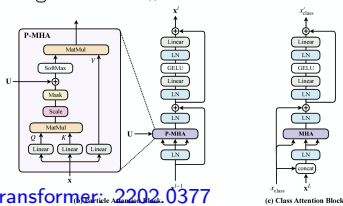
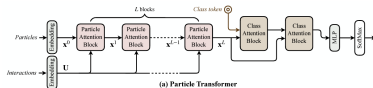
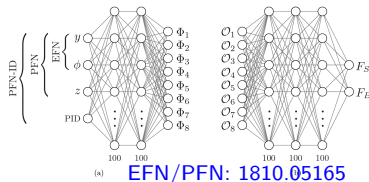
- 1) High-level taggers on jet substructure observables
  - Robust and easy to interpret at analysis/theory level
- 2) Machine learning taggers on jet-substructure
  - Non-trivially combine several observables for better discrimination
- 3) Machine learning taggers on low-level inputs
  - Lose interpretativeness for performance



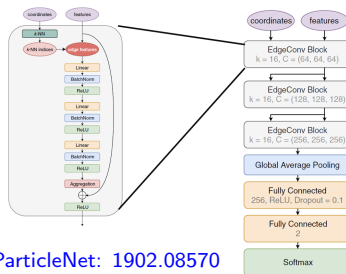
# ML-era: Architectures

Machine-learning/pheno community is developing faster than we can test on data!

- Normal dense neural networks
- **ResNet**: CNN Architecture representing jet as image
- **Energy/Particle Flow networks (EFN/PFN)**: General decomposition of IRC-safe observables
- **ParticleNet**: Graph network on point cloud
- **ParticleTransformer**: Transformer
- **GN2X**: Transformer with auxiliary tasks
- **LundNet**: Graph on declustering history
- **PELICAN**: Lorentz invariant network

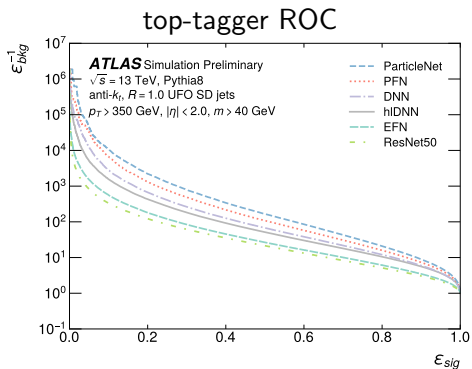


**ParticleTransformer: 2202.0377**



Constituent based **top-tagger** / **W-tagger** outperform high-level features ones:

- Provide network the lowest level information available: the jet constituents themselves
- Factor 2-3 improvement!
- ResNet/EFN under-perform w.r.t **theoretical performance**
- Real simulation studies important!



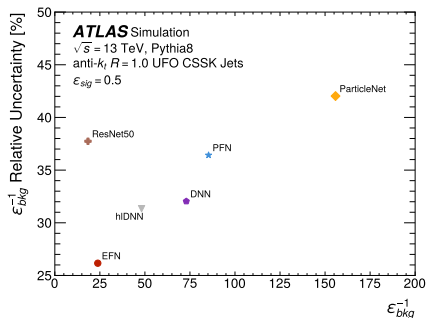
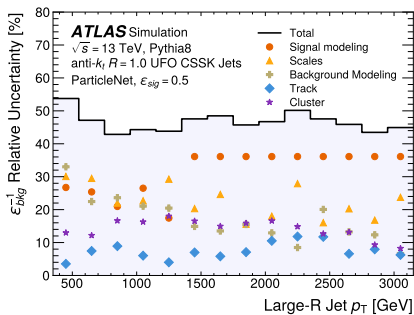
**W-tagging summary table**

Model	AUC	ACC	$\epsilon_{bkg}^{-1}$ @ $\epsilon_{sig} = 0.5$	$\epsilon_{bkg}^{-1}$ @ $\epsilon_{sig} = 0.8$	# Params	Inference Time
EFN	0.920	0.835	35.1	7.95	56.73k	0.065 ms
PFN	0.931	0.853	44.7	9.50	57.13k	0.11 ms
ParticleNet	0.933	0.826	46.2	9.76	366.16k	0.36 ms
ParticleTransformer	0.951	0.880	77.9	14.6	2.14M	0.28 ms

But as we provide lower-level information the taggers can start to learn features more specific to generator/training dataset

New results evaluating **approximate** bottom-up experimental uncertainties and theory uncertainties on top-taggers

- Better taggers have higher uncertainties by almost factor 2
- Want to develop ways to break this trend



ATLAS recently provided an [open data set](#) for the larger community to help tackle this problem

- Evaluate new R&D architectures with realistic simulation
- Explore newer topic of tagger “resilience”



Datasets features:

- 100 million top and QCD jets training+testing
  - **Full ATLAS detector and pile-up simulation!**
  - Jet constituents provided for each
- **Alternative samples to evaluate uncertainties**
  - Approximate experimental uncertainties on the constituents
  - Alternative signal/background generators for theory uncertainties

Each data-set has

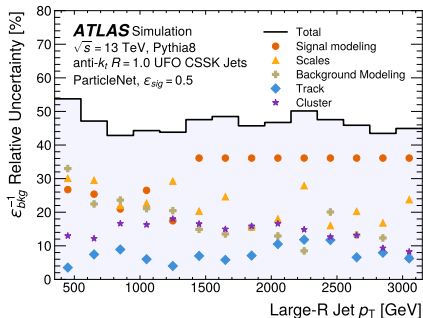
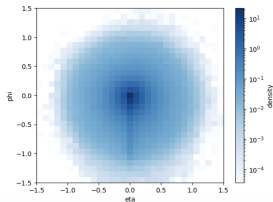
- The four vectors of constituent particles
- 15 high level summary quantities evaluated on the jet
- The four vector of the whole jet
- A training weight (nominal only)
- PYTHIA shower weights (nominal only)
- A signal vs background label

Aims to:

- Give you the info needed to re-produce the results at [JINST 19 \(2024\) P08018](#) with the provided [open data](#)
- Accessing the data
- Inspecting/pre-processing the data
- Basic neural net training example
- Take  $\sim 1$  hour
- Expect PhD student level physics/coding knowledge

```
In [13]: #make a 2D density histogram of the average jet
plt.hist2d(sig_eta, sig_phi, range=[[-1.5,1.5],[1.5,1.5]], bins=[31,31], cmap="Blues", density=True, norm=plt.colors.L
plt.xlabel('eta')
plt.ylabel('phi')
plt.colorbar(label="density")
```

Out[13]: <matplotlib.colorbar.Colorbar at 0x113139dab>





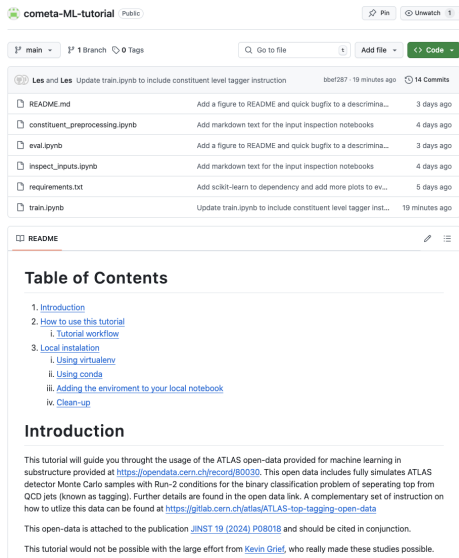
# This tutorial

## Github repository provided

- Jupyter notebooks provided
  - inspect\_inputs.ipynb: Access the open data
  - constituent\_preprocessing.ipynb: Pre-processing the data
  - train.ipynb: Basic NN training
  - eval.ipynb: NN evaluation including uncertainties

## How to run the notebooks

- Basic setup for an environment in conda or virtualenv
  - Mainly pytorch and jupyter
- If you don't want to download anything can also use [Binder](#)
  - Provide the github link to binder and run an interactive web session
  - Note that much slower than running local



The screenshot shows the GitHub repository page for 'cometa-ML-tutorial'. The repository is public and has 1 branch and 0 tags. The file list includes: README.md, constituent\_preprocessing.ipynb, eval.ipynb, inspect\_inputs.ipynb, requirements.txt, and train.ipynb. The README file is selected, showing a 'Table of Contents' with the following items:

1. [Introduction](#)
2. [How to use this tutorial](#)
  - i. [Tutorial workflow](#)
3. [Local installation](#)
  - i. [Using virtualenv](#)
  - ii. [Using conda](#)
  - iii. [Adding the environment to your local notebook](#)
  - iv. [Clean-up](#)

The 'Introduction' section of the README states: 'This tutorial will guide you through the usage of the ATLAS open-data provided for machine learning in substructure provided at <https://opendata.cern.ch/record/80030>. This open data includes fully simulated ATLAS detector Monte Carlo samples with Run-2 conditions for the binary classification problem of separating top from QCD jets (known as tagging). Further details are found in the open data link. A complementary set of instruction on how to utilize this data can be found at <https://gitlab.cern.ch/atlas/ATLAS-top-tagging-open-data>'.

It also notes: 'This open-data is attached to the publication [JINST 19 \(2024\) P08018](#) and should be cited in conjunction. This tutorial would not be possible with the large effort from [Kevin Grief](#), who really made these studies possible.'