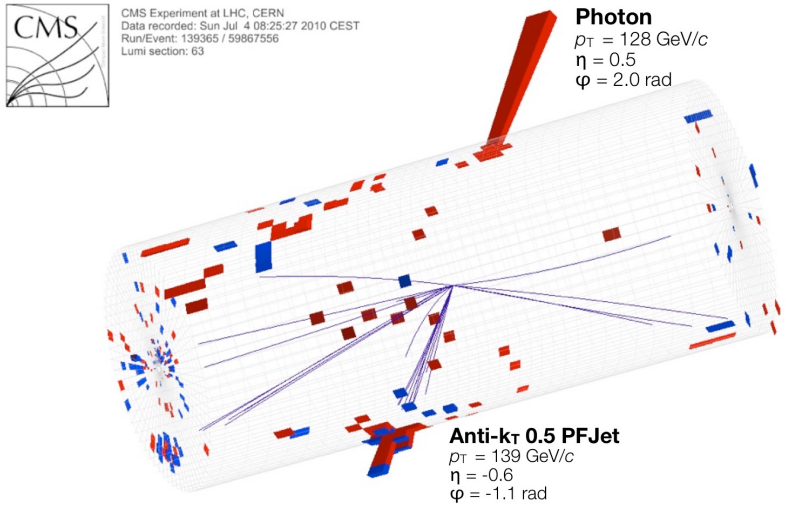
Hbb jet tagging with CMS open data. Data pipeline plan



DEPARTMENT OF PHYSICS AND ASTRONOMY

RICE UNIVERSITY

COLLIN ARBOUR AND IASON KROMMYDAS , 09/22/2022

The data

Sample with jet, track and secondary vertex properties for Hbb tagging ML studies HiggsToBBNTuple_HiggsToBB_QCD_RunII_13TeV_MC

Duarte, Javier

Description

The dataset consists of particle jets extracted from simulated proton-proton collision events at a center-of-mass energy of 13 TeV generated with Pythia 8. It has been produced for developing machine-learning algorithms to differentiate jets originating from a Higgs boson decaying to a bottom quark-antiquark pair (Hbb) from quark or gluon jets originating from quantum chromodynamic (QCD) multijet production.

The reconstructed jets are clustered using the anti-kT algorithm with R=0.8 from particle flow (PF) candidates (AK8 jets). The standard L1+L2+L3+residual jet energy corrections are applied to the jets and pileup contamination is mitigated using the charged hadron subtraction (CHS) algorithm. Features of the AK8 jets with transverse momentum pT > 200 GeV and pseudorapidity |η| < 2.4 are provided. Selected features of inclusive (both charged and neutral) PF candidates with pT > 0.95 GeV associated to the AK8 jet are provided. Additional features of charged PF candidates (formed primarily by a charged particle track) with pT > 0.95 GeV associated to the AK8 jet are also provided. Finally, additional features of reconstructed secondary vertices (SVs) associated to the AK8 jet (within ΔR < 0.8) are also provided.

http://opendata.cern.ch/record/12102
DOI:10.7483/OPENDATA.CMS.JGJX.MS7Q

- Data consist of simulated jet events with Pythia 8 and MadGraph.
- Labelled to be used for ML algorithms to differentiate $H \rightarrow b\,\bar{b}$ from regular QCD jets.
- Served as ROOT files.

Literature

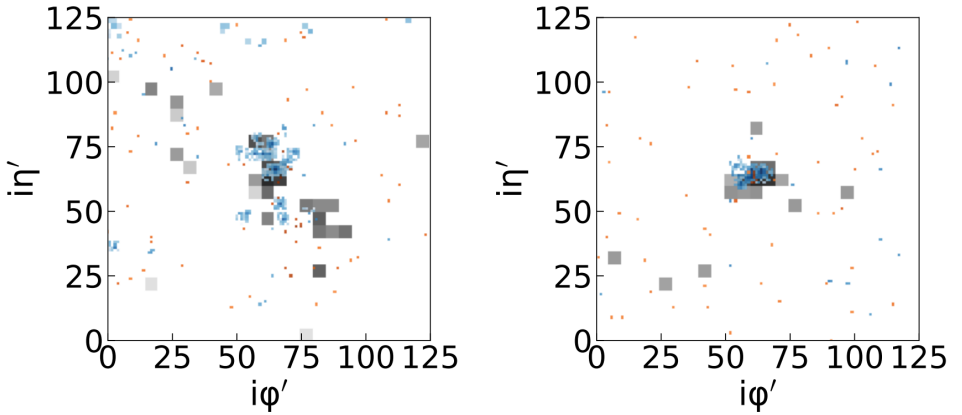
- This particular dataset hasn’t been used in the literature.
- Generally, a lot of research has been done on jet classification.
- b-tagging is very common in the literature.
- Hbb-tagging is slightly less common.
- Different algorithms and architectures used. We couldn’t find a full pipeline ready for use.
- We will not be copying any algorithms.
- We will be using our own methods and architectures.

Data wrangling

- Data are spread over 90 ROOT files. 200,000 events each. Roughly 100 GB in total.
- Each ROOT files contains a TTree with all 200,000 event features as TBranches.
- Each jet event has multiple features such as P_T, η, φ etc.
- Because each jet event has multiple tracks and subjets, some features are jagged arrays within those TBranches.
- We will work in the scikit-hep ecosystem, using tools such as uproot and Awkward Array to manipulate the data.
- We will use these tools to extract useful features from the data and bring them to the appropriate dimensions to be used by ML frameworks.
- Work needs to be done on how to treat the jagged arrays as features.

Data wrangling

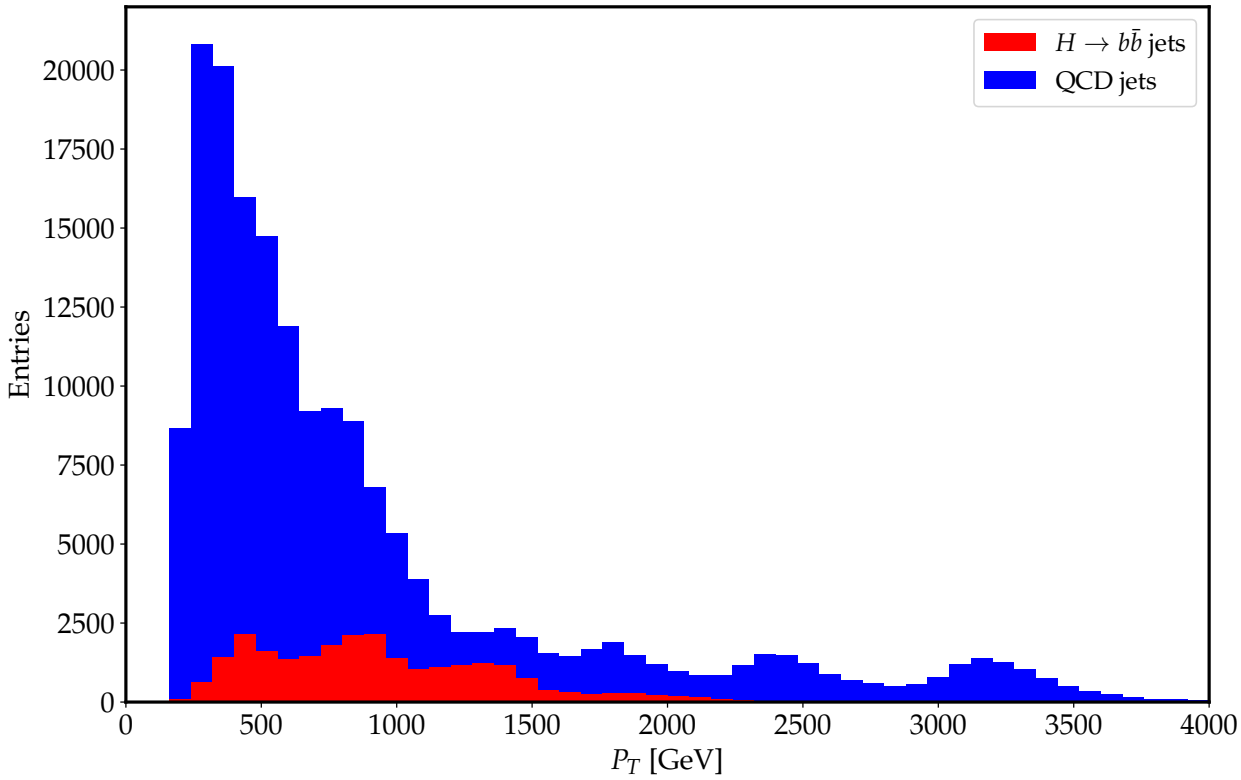
- We will also try to create jet images to use with image classification methods.
- This will most likely not be possible with the current state of the dataset.
- We will have to resort to using the raw MiniAOD files that were used to create this dataset that contain ECAL and HCAL hits.



(d) Composite jet image. Left: gluon jet, Right: quark jet.

Source: arXiv:1902.08276

Data exploration



- We firstly plan to use the usual data exploration methods such as: plotting histograms of the features, creating correlation matrices, standard scaling, PCA, eigendecomposition of the correlation matrix, Kullback–Leibler divergence.

Modelling

- We plan to use Deep Neural Networks and Recurrent Neural Networks.
- We also plan to use simpler binary classification methods such as BDTs and/or Linear, Logistic, Least-squares classifiers.
- If we also manage to create jet images, we will use Convolutional Neural Networks.

QCD-Aware Recursive Neural Networks for Jet Physics

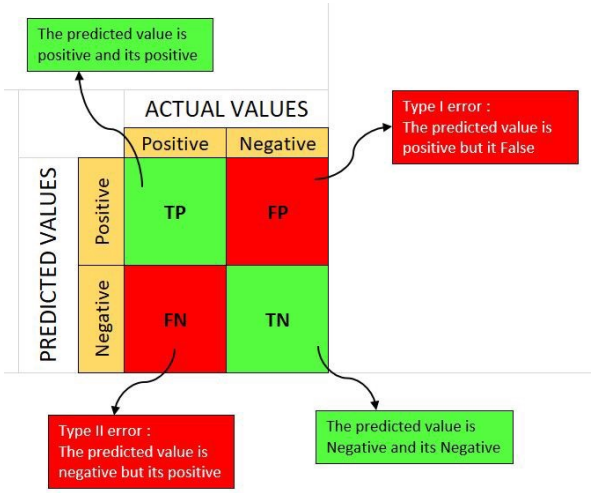
Gilles Louppe[a,b,1] **Kyunghyun Cho**[b] **Cyril Becot**[a,2] **Kyle Cranmer**[a,b]

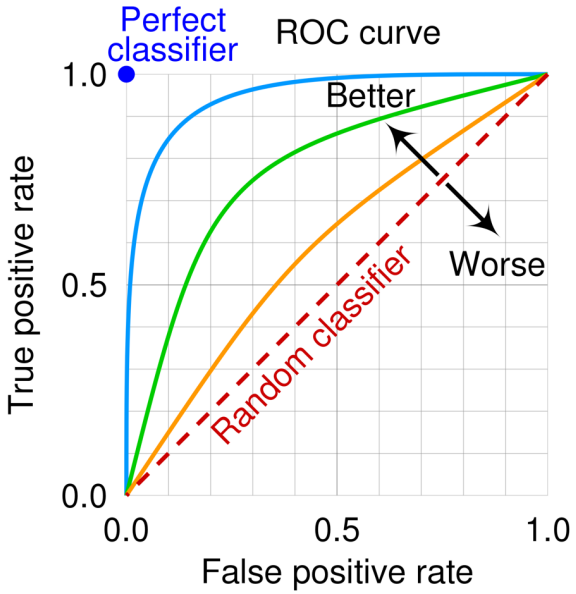
[a] *New York University, Center for Cosmology & Particle Physics, 726 Broadway, New York, NY*
[b] *New York University, Center for Data Science, 60 5th Ave., New York, NY*

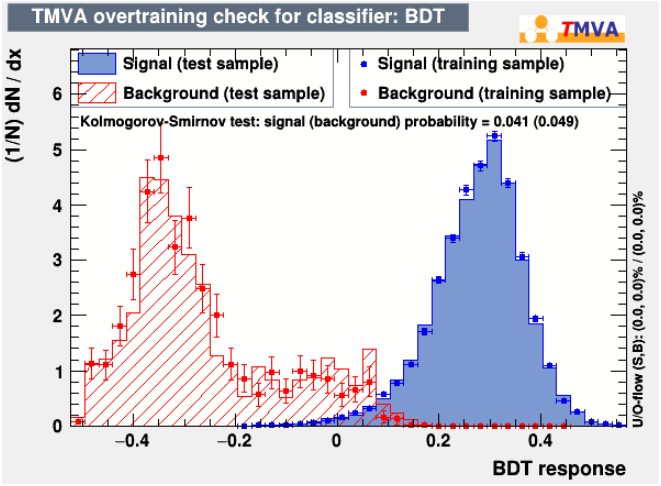
E-mail: g.louppe@uliege.be, kyunghyun.cho@nyu.edu, cyril.becot@cern.ch, kyle.cranmer@nyu.edu

Validation

- Accuracy
- Confusion Matrix
- ROC curves
- KS plot







Thank you!