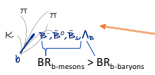
Jet flavour tagging using Deep Learning in the CMS experiment

Anna Stakia (CERN)
on behalf of the CMS Collaboration

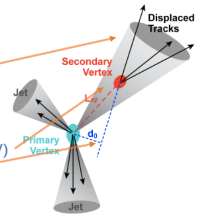
1. Motivation

The identification of jets originating from heavy flavour quarks (b/c 'tagging') is crucial in various studies at the Large Hadron Collider, in both New Physics searches and Standard Model processes.

2. Key features of b jets



b quarks typically hadronise in B hadrons forming the b jets



b hadrons: distinctive features with detectable particularities:

- Long lifetime ⟹ large displacement of Secondary Vertex (SV)
- Mass of 5–10 GeV ⟹ large decay angle ⟹ large d_0
- Possible decay to leptons [BR ≈ 42%]
- ~5 charged tracks per decay

3. b and c taggers

CSVv2: 'Combined* SV, version 2' [updated CSV for Run2]

- Neural Network (NN) instead of likelihood ratio
- Input: info on tracks & SV (='combined')

cMVAv2: 'combined* Multivariate Algorithm, version 2'

- Boosted Decision Tree (BDT)
- Input: info on tracks, SV (combined) & soft leptons in the jet as the output of the taggers:
  - → Jet Probability (JP), Jet B Probability (JBP)
  - → CSVv2
  - → Soft Electron (SE), Soft Muon (SM)

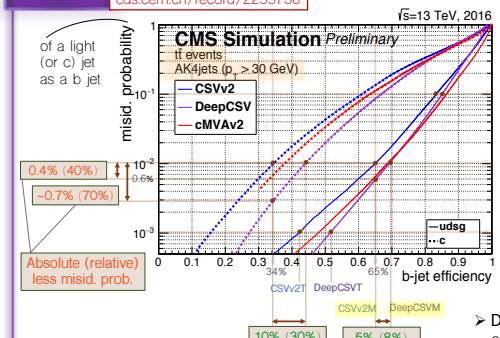
ctagger: c jet identification algorithm against b (CvsB) and light (CvsL) jets

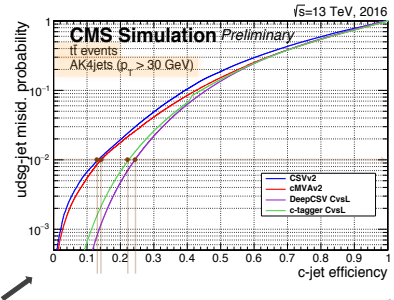
- Gradient Boosting Classifier (GBC)
- Input: info on tracks, SV (combined) & soft leptons in the jet

Same input as CSVv2 but based on more tracks, deep NN, multi-classification

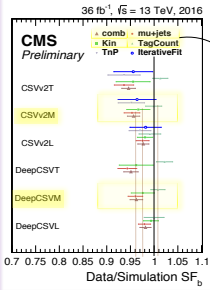
DeepCSV

CMS DP 2017-005
cds.cern.ch/record/2255736





- DeepCSV significantly better than CSVv2 in ~every b jet efficiency value, vs both light and c jet misid. (as b jet) prob.
- DeepCSV efficiency equal or better than (i) cMVAv2 and (ii) ctagger (in most cases), despite not containing lepton info. (ii) also shows DeepCSV as a multi-classifier

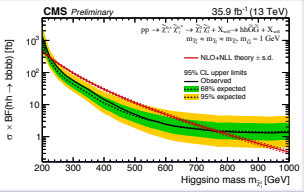


Methods for the b tagging efficiency measurement

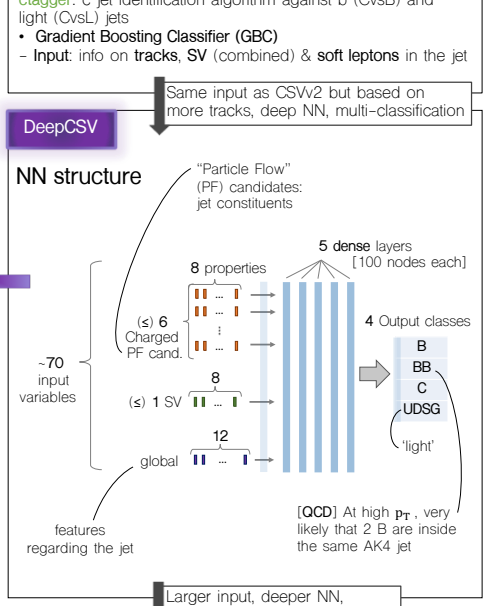
- Similar Scale Factors between CSVv2 and DeepCSV ⇒ better performance for DeepCSV in data too

- DeepCSV already been applied and improved sensitivity in the analysis on the search for: $\tilde{H} \to h\,(\to b\bar{b}) + |\vec{p}_T^{\,miss}|$, 4 b jets in the final state

CMS PAS SUS-16-044
cds.cern.ch/record/2256648



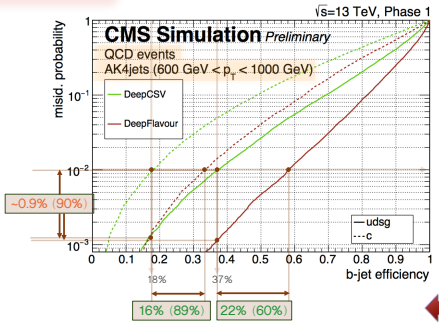
DeepCSV — NN structure



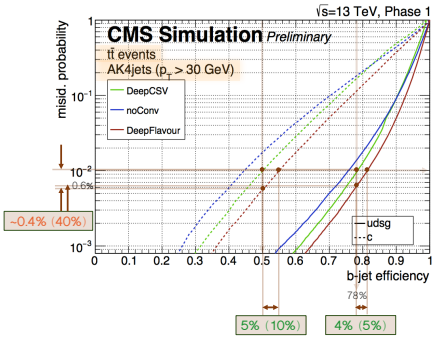
Larger input, deeper NN, +convolutional, recurrent layers

DeepFlavour

CMS DP 2017-013
cds.cern.ch/record/2263802



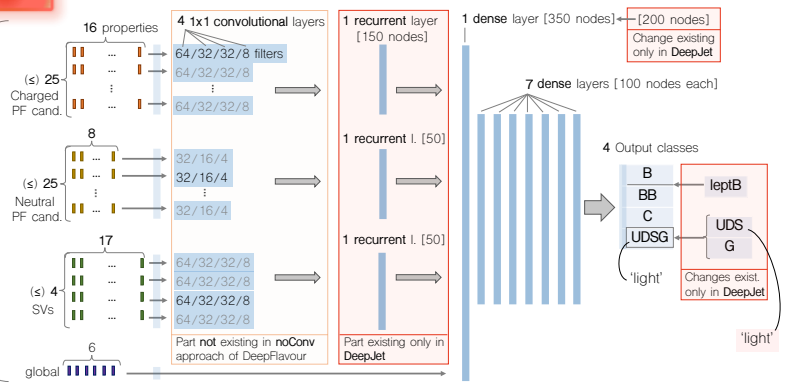
- Significant gain in efficiency of DeepFlavour w.r.t DeepCSV in all p_T regions, especially in high p_T
  ⇒ can lead to improved sensitivity for analyses with highly energetic b jets in the final state



- Removing the convolutional layers degrades performance, even with larger input and deeper NN
- Improvement w.r.t. DeepCSV comes from both larger input and a better NN model (i.e. exploiting input structure)

DeepFlavour — DeepJet

NN structure



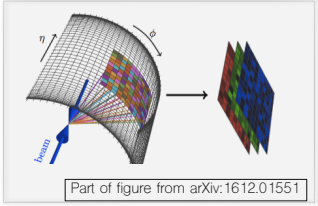
DeepJet

CMS DP 2017-027
cds.cern.ch/record/2275226

NN approaches compared to DeepJet in Quark/Gluon discrimination [Quark:'light']:

recurrent [inspired by arXiv:1702.00748]: Input: relative $p_T/\eta/\varphi$, and PUPPI (Pileup Per Particle Identification) weight (arXiv:1407.6013) of charged (fed to one recurrent layer of 100 nodes – LSTM) and neutral PF cand. (fed to identical layer). LSTM output then merged with global* variables in a dense NN [1 layer of 200 nodes, 5 of 100].

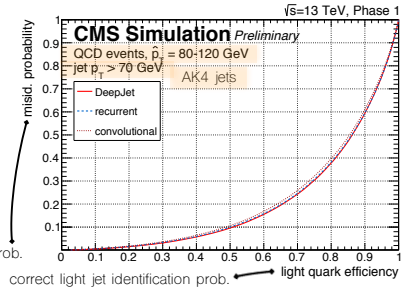
convolutional: Jet treated as image on $\eta - \varphi$ plane. 3rd dimension – color: (separate study for charged and neutral PF cand.) relative p_T and particle multiplicity within the pixel. This input info fed to a CNN (Convolutional NN) [as in arXiv:1612.01551], then merged with the global* variables in a dense layer of 128 nodes.



Part of figure from arXiv:1612.01551

*Jet variables:
jet p_T/η,
#(charged PF cand.),
#(neutral PF cand.),
#(SV within the jet)

[Jets originating from gluon splitting to bb or cc not considered as gluon jets]



- Similar performance among DeepJet and 2 alternative approaches

This Report is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 675440