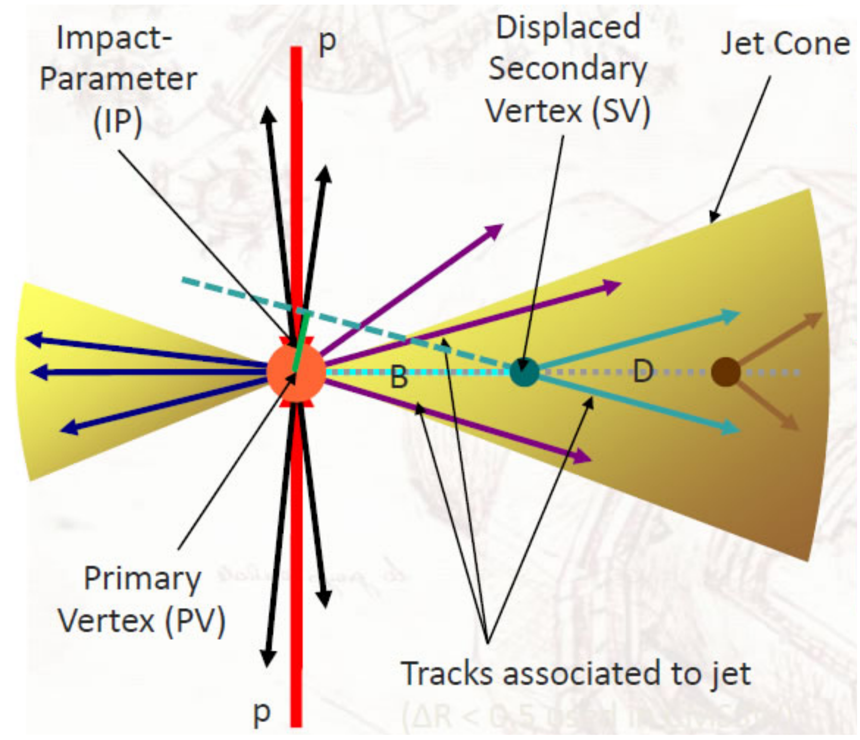
Heavy Flavor Jet Tagging in CMS

Soureek Mitra, on behalf of the CMS collaboration

Karlsruhe Institute of Technology

Motivation:

Hadronic jets originating from heavy flavor (b/c) quarks arise often in the studies involving the top quark, Higgs boson etc. Identification of these jets is therefore extremely important in the LHC physics program.



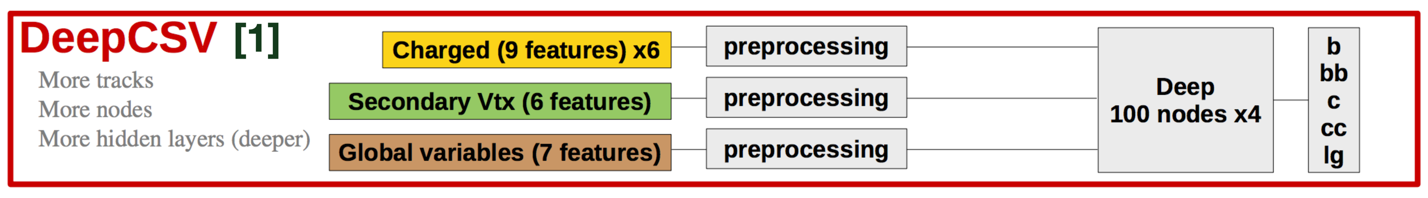
Typical features of heavy flavor jets:

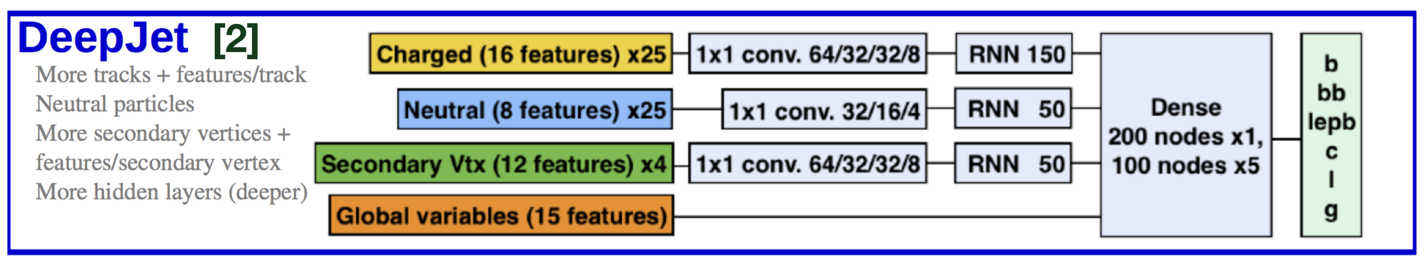
- Contain secondary vertices (SV) due to b(c)-hadron decays ☞ $m \approx 5.3\ (1.9)$ GeV, $\tau \approx 1.5\ (1.0)$ ps
- High track multiplicity with high impact parameter (IP) w.r.t primary vertex
- Presence of soft leptons inside jets due to semileptonic decay modes ☞ $\mathscr{B}(b \to \mu X) \approx 20\,\%$

Tagging typically relies on:

- Track info. ☞ IP2D, IP3D, track multiplicity etc.
- SV info. ☞ m_{SV}, SV flight distance etc.
- Particle flow hadron and soft lepton candidate info.
- Some combination of the above ☞ Neural Network

Tagging Algorithms:

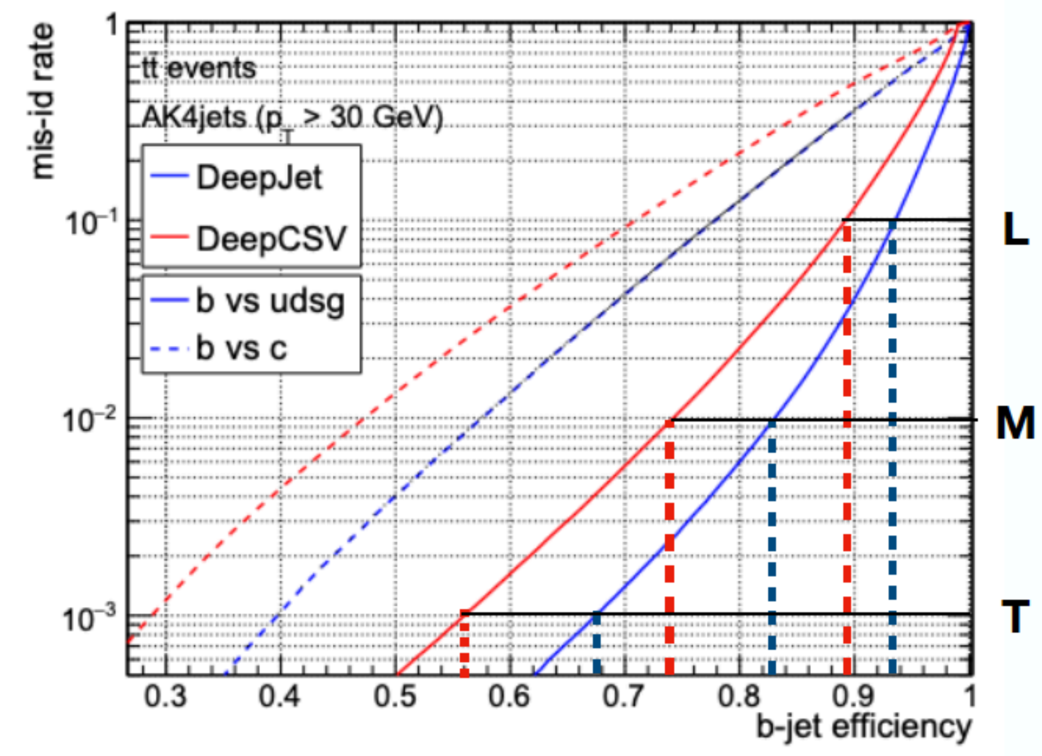




$$P(BvsAll) = P(b) + P(bb) + P(lepb)$$

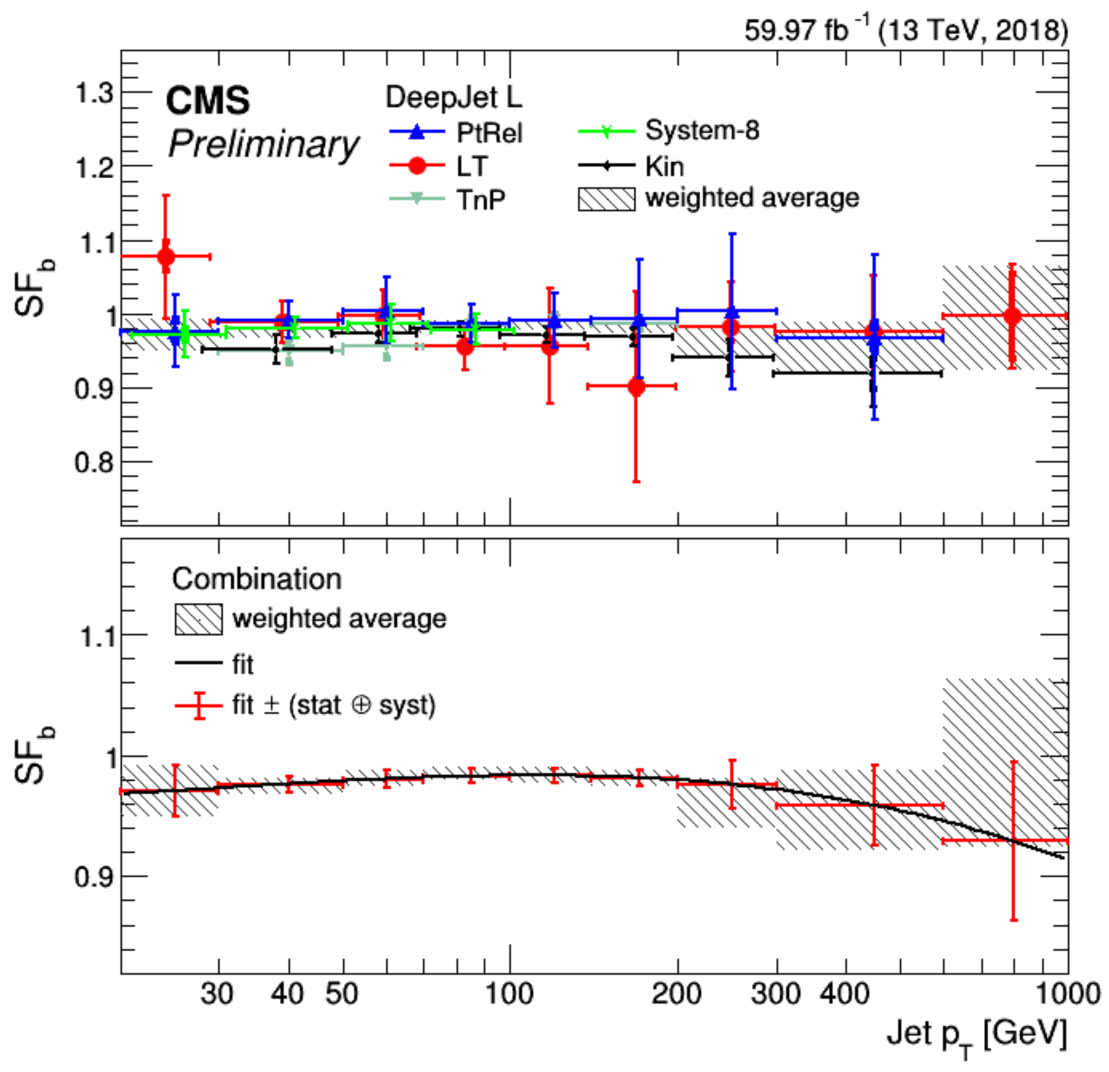
$$P(CvsB) = \frac{P(c)}{P(b) + P(bb) + P(lepb) + P(c)}\ ;\ P(CvsL) = \frac{P(c)}{P(uds) + P(g) + P(c)}$$

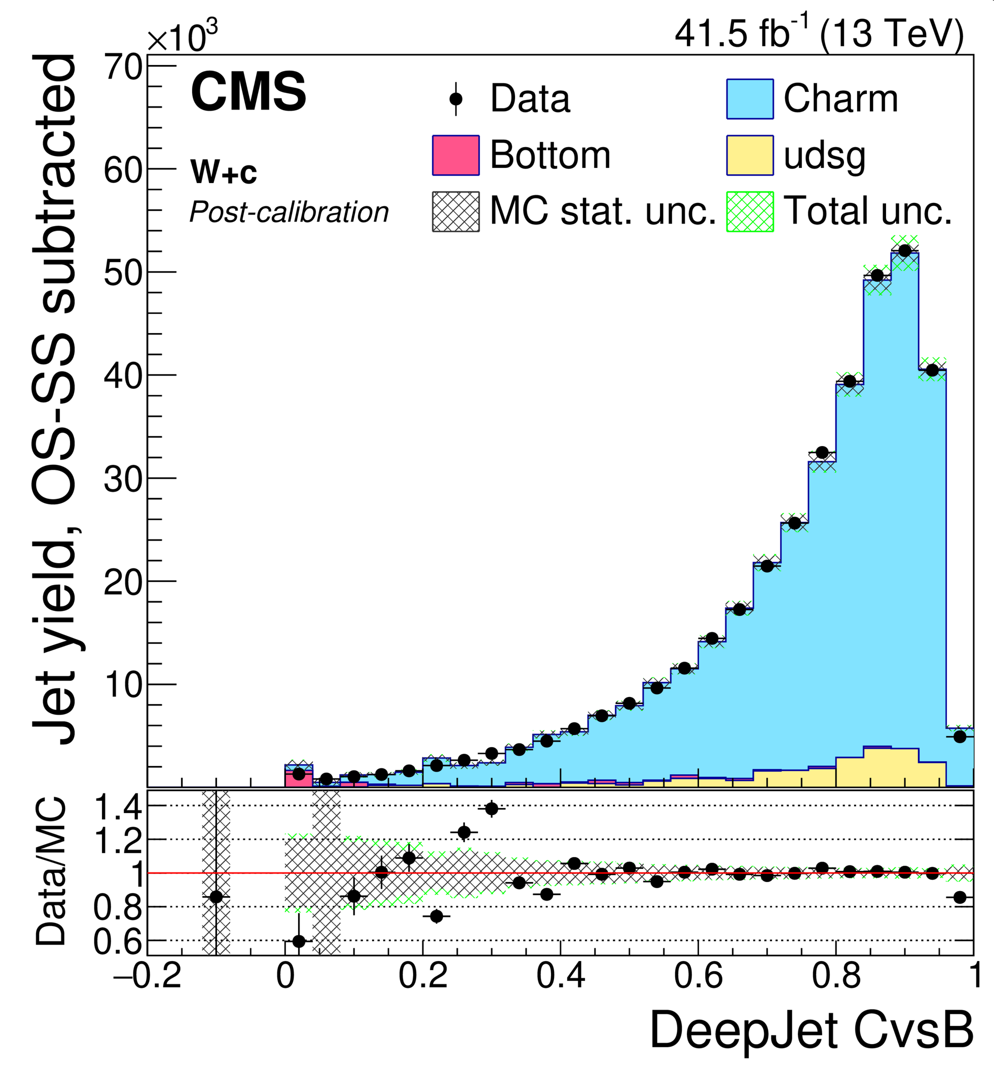
- Truth label obtained from *hadron flavor* definition
- Jets originating from pileup vertices are excluded from the training sample
- Trained on a mixture of simulated QCD multijet and top pair ($t\bar{t}$) events, while performance evaluated in simulated $t\bar{t}$ events
- Performance is typically shown via receiver operator characteristic (ROC) curves
- Working points (WP) based on *light* (udsg) mis-Id. rates ☞ **L**oose (10%), **M**edium (1%), **T**ight (0.1%)



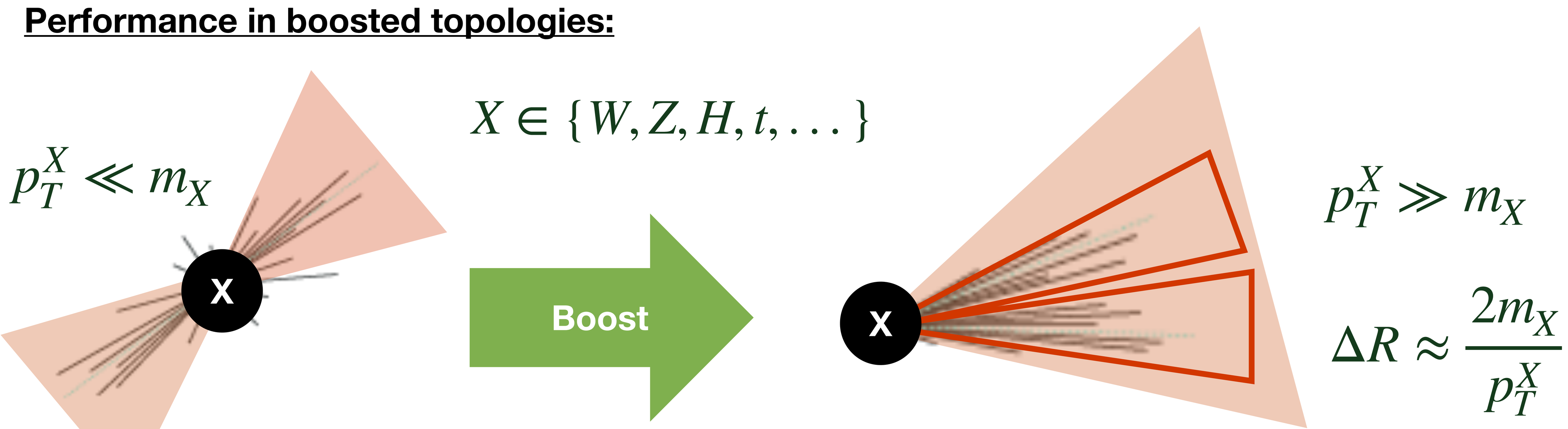
Performance in data with AK4 jets:

- $\varepsilon_f^{MC} = \dfrac{N_f^{Tagged}}{N_f^{Total}}\ ,\ \varepsilon_f^{Data} = SF_f \times \varepsilon_f^{MC}$
- Calibration of b-, c-, and *light*-jets performed using different methods based on QCD multijet, $t\bar{t}$, Drell-Yan (DY), and W+c events **[3,4]**
- Scale factors (SF) estimated at different WPs as well as for the *entire discriminant shape* for each flavor
- Calibration for c-tagging discriminants ($CvsL$ & $CvsB$) performed in c-enriched (W+c), b-enriched ($t\bar{t}\ \ell$ + jets), and *light*-enriched (DY, QCD multijet) regions

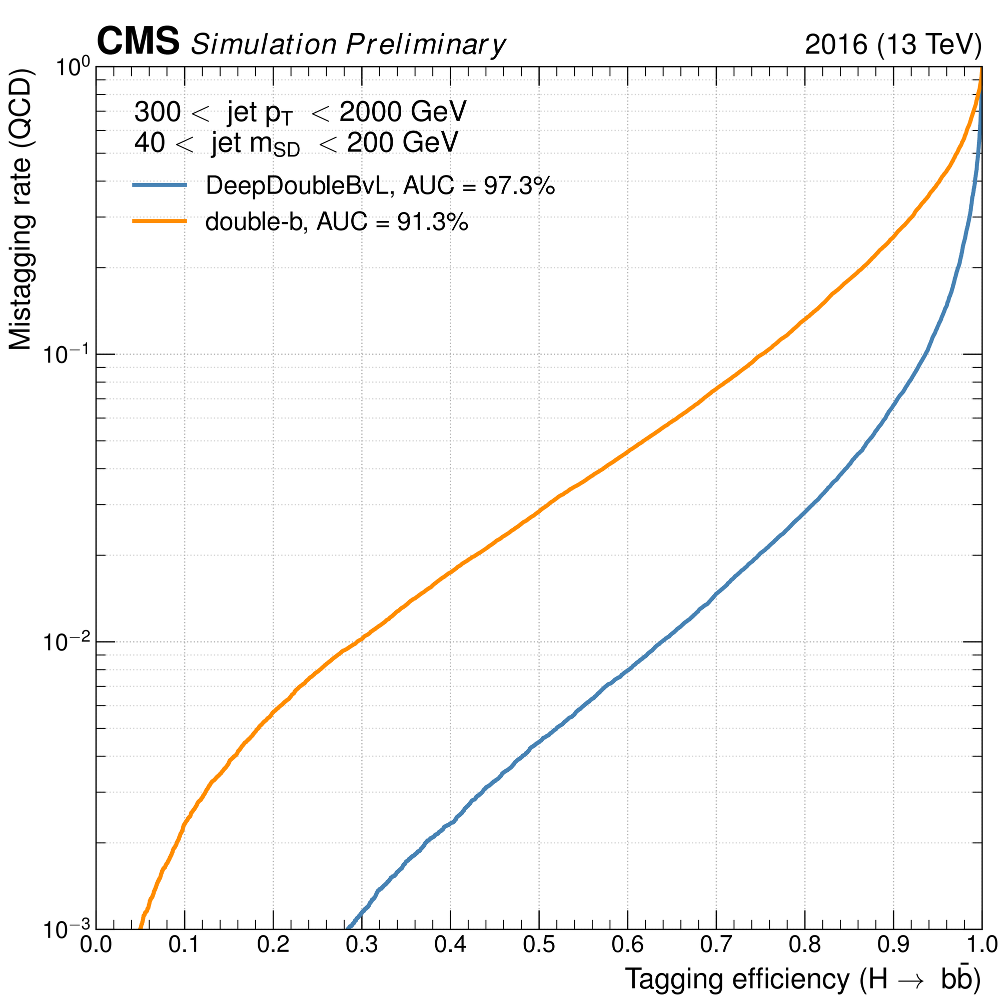


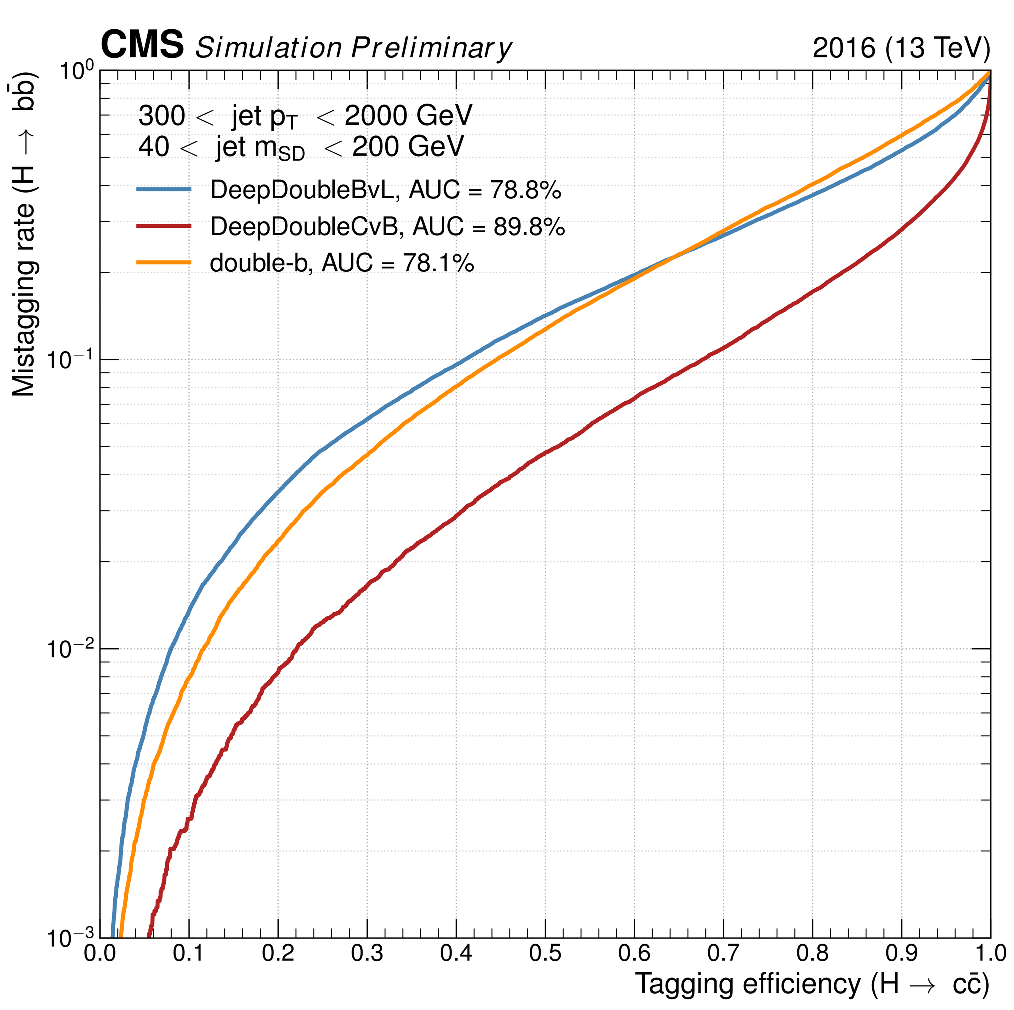


Performance in boosted topologies:



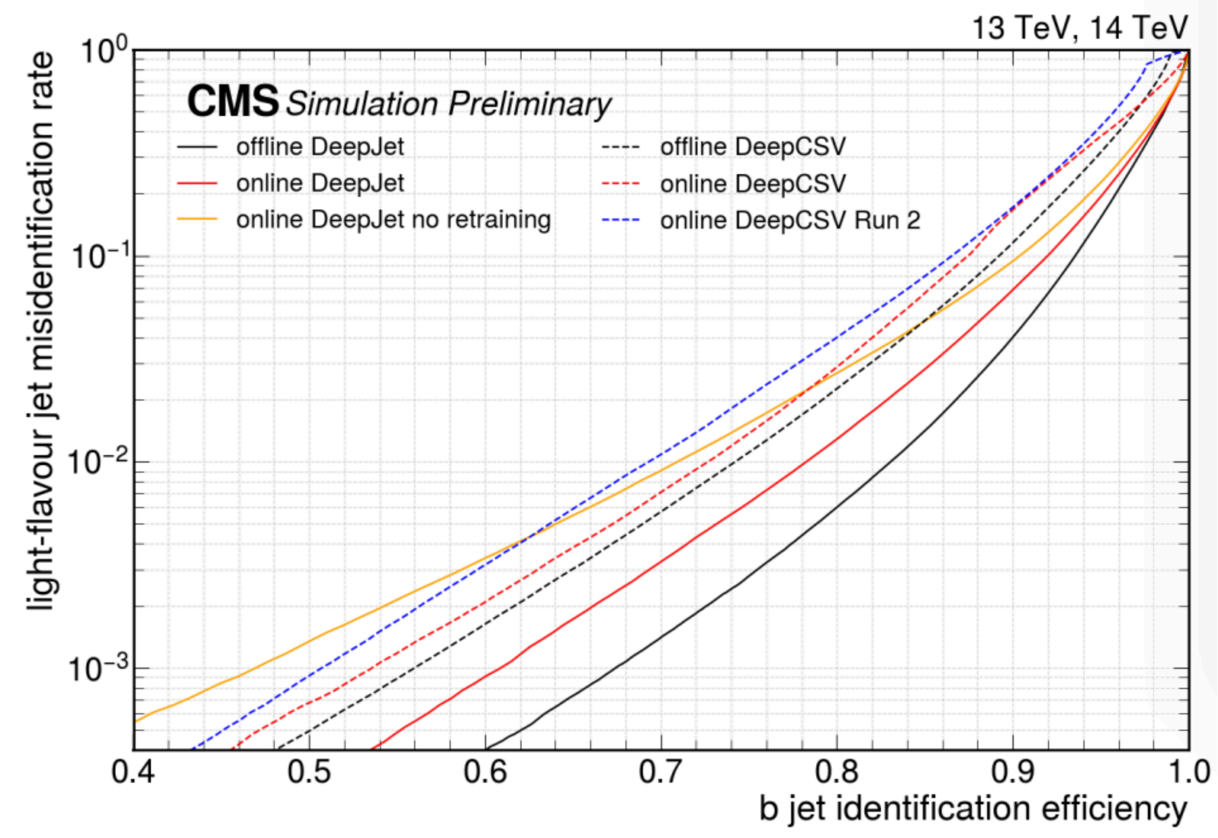
- Performance comparison of the **double-b** and **DeepDoubleX** algorithms for boosted jet ($q\bar{q}$) tagging are shown by studying the mis-Id. rate of QCD jets as a function of the tagging efficiency **[5]**
- ROC curves are obtained from a mixture of simulated QCD multijet and $H \to b\bar{b}(c\bar{c})$ events





Performance @ HLT:

- Dedicated training with HLT-level inputs shows better performance w.r.t evaluation of the offline training model at HLT **[6]**
- Already implemented for Run3



References:

[1] CMS Collaboration, *"Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV"*, JINST 13 (2018) P05011

[2] Emil Bols *et. al.*, *"Jet Flavour Classification Using DeepJet"*, JINST 15 (2020) P12012

[3] CMS Collaboration, *"B-tagging performance of the CMS Legacy dataset 2018"*, CMS DP-2021/004

[4] CMS Collaboration, *"A new calibration method for charm jet identification validated with proton-proton collision events at 13 TeV"*, JINST 17 (2022) P03014

[5] CMS Collaboration, *"Performance of Deep Tagging Algorithms for Boosted Double Quark Jet Topology in Proton-Proton Collisions at 13 TeV with the Phase-0 CMS Detector"*, CMS DP-2018/046

[6] CMS Collaboration, *"Expected Performance of Run-3 HLT b-quark jet identification"*, CMS DN-2022/015