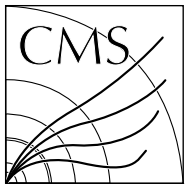
Available on CMS information server

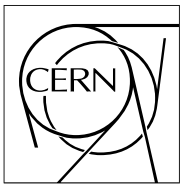
**CMS DP -2017/005**



**The Compact Muon Solenoid Experiment**

CMS Performance Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland



**13 March 2017**

# Heavy flavor identification at CMS with deep neural networks

The CMS Collaboration

**Abstract**

At the Large Hadron Collider, the identification of jets originating from heavy flavour quarks (b or c-tagging) is important for searches for new physics and for measurements of standard model processes. A variety of b-tagging algorithms has been developed by CMS to select b-quark jets based on variables such as the impact parameters of the charged-particle tracks, the properties of reconstructed decay vertices, and the presence or absence of a lepton, or combinations thereof. These algorithms heavily rely on machine learning tools and are thus natural candidates for advanced tools like deep neural networks. A new algorithm, DeepCSV, uses a deep neural network. The input is the same set of observables used by the existing CSVv2 b-tagger, with the extension that it uses information of more tracks. Also, the training strategy was adapted and about 50 million jets are used for the training of the deep neural network. The new DeepCSV algorithm outperforms the CSVv2 tagger, with an absolute b-tagging efficiency improvement of about 4% for a misidentification probability for light-flavour jets of 1%. In addition, DeepCSV is a multiclassifier simultaneously trained for c-tagging. For c-tagging DeepCSV outperforms the other taggers in CMS.

# DeepCSV

## The CMS collaboration

**CMS-POG-CONVENERS-BTAG@CERN.CH**

At the Large Hadron Collider, the identification of jets originating from heavy flavour quarks (b or c-tagging) is important for searches for new physics and for measurements of standard model processes. A variety of b-tagging algorithms has been developed by CMS to select b-quark jets based on variables such as the impact parameters of the charged-particle tracks, the properties of reconstructed decay vertices, and the presence or absence of a lepton, or combinations thereof. These algorithms heavily rely on machine learning tools and are thus natural candidates for advanced tools like deep neural networks. A new algorithm, DeepCSV, uses a deep neural network. The input is the same set of observables used by the existing CSVv2 b-tagger, with the extension that it uses information of more tracks. Also, the training strategy was adapted and about 50 million jets are used for the training of the deep neural network. The new DeepCSV algorithm outperforms the CSVv2 tagger, with an absolute b-tagging efficiency improvement of about 4% for a misidentification probability for light-flavour jets of 1%. In addition, DeepCSV is a multiclassifier simultaneously trained for c-tagging. For c-tagging DeepCSV outperforms the other taggers in CMS.

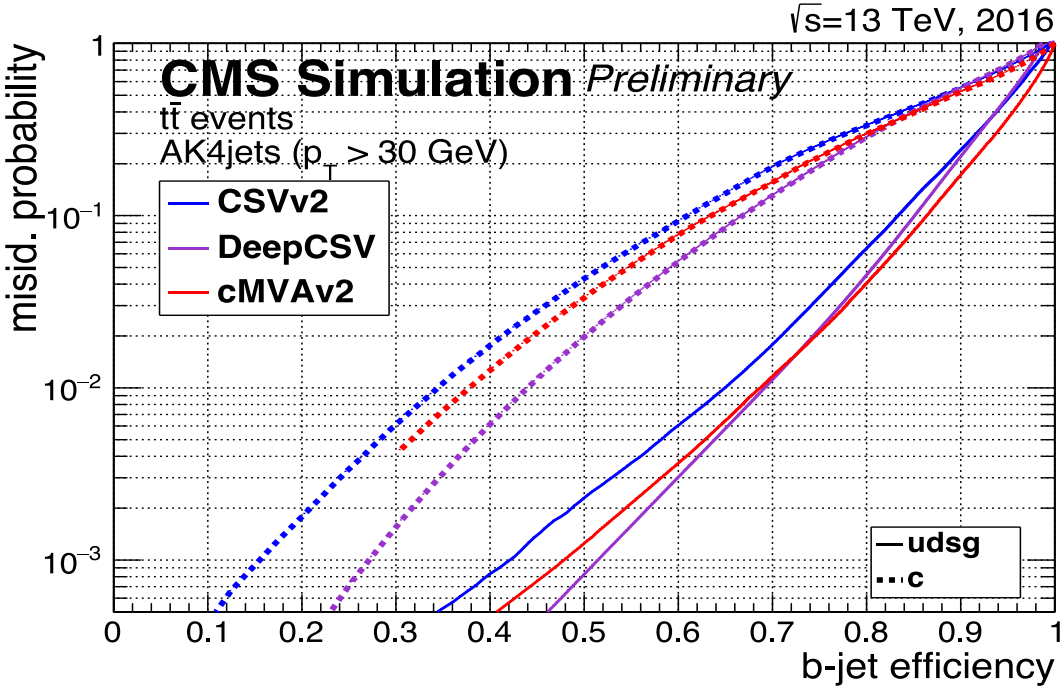
# Glossary: CMS flavour taggers

**CSVv2**: Combined Secondary Vertex version 2 algorithm, based on secondary vertex and track-based lifetime information, it is an updated version of the CSV algorithm used in Run 1 combining the variables with a neural network instead of a likelihood ratio and the secondary vertex information is obtained with the Inclusive Vertex Finder algorithm.
CSVv2L, CSVv2M, CSVv2T: CSVv2 algorithm at the loose, medium, tight operating points, defined as the values of the discriminator cut for which the rate for misidentifying a light jet as a b jet is 10%, 1%, and 0.1%, respectively.
**DeepCSV**: a new algorithm based on the same set of observables used by the CSVv2 b-tagger, with a simple extension to use more charged particle tracks. This algorithm is based on a deep neural network, with four hidden layer (i.e. six layers altogether) of a width of 100 nodes each.
DeepCSVL, DeepCSVM, DeepCSVT: DeepCSV algorithm at the loose, medium, tight operating points, defined as the values of the discriminator cut for which the rate for misidentifying a light jet as a b jet is 10%, 1%, and 0.1%, respectively.
**c-tagger**: a c jet identification algorithm exploiting properties related to displaced tracks, secondary vertices, and soft leptons inside the jets. The training of the classifiers is performed using a Gradient Boosting Classifier (GBC). Two separate GBCs are provided, one for discriminating c jets from light jets (CvsL) and one for discriminating c jets from b jets (CvsB).
**cMVAv2**: combined Multi Variate Algorithm version 2, using a Boosted Decision Tree taking as input the different algorithm outputs of CSVv2, a variant of CSVv2 using another vertex reconstruction, Jet Probability (JP), Jet B Probability (JBP), Soft Electron (SE) and Soft Muon (SM) taggers.

# Glossary: Flavour tagger performance measurements in data

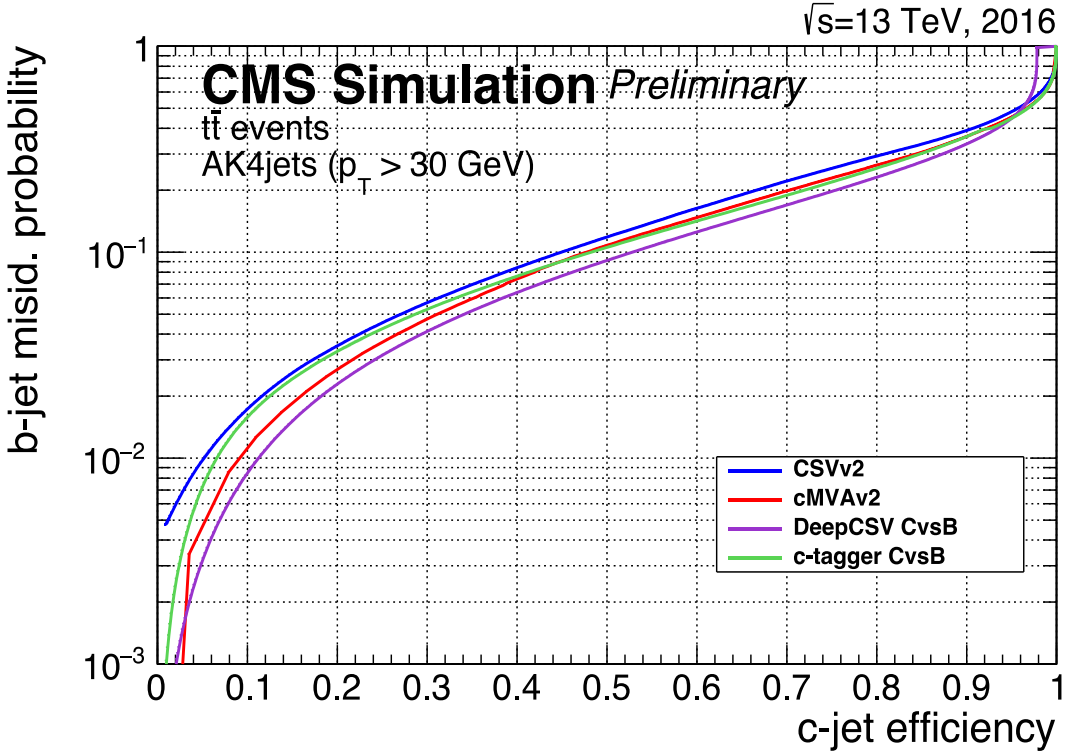
**mu+jets**: Measured b-tagging efficiency in multijet events with a muon, based on the combination of the results from different measurements, obtained using the PtRel, the LT and the System8 methods.
**PtRel**: Method for the measurement of the b-tagging efficiency in multijet events based on the transverse momenta of muons w.r.t. the jet axis.
System8: Method for the measurement of the b-tagging efficiency in multijet events with a muon, solving a system of 8 equations.
**LT**: Lifetime Tagging method for the measurement of the b-tagging efficiency in multijet events, based on template fits to the JP distributions.
Kin: Method for the measurement of the b-tagging efficiency in ttbar events in the dileptonic channel, based on a template fit to an MVA discriminator combining kinematic variables.
**TagCount**: Method for the measurement of the b-tagging efficiency in ttbar events in the dileptonic channel. The b-tagging efficiency is obtained by counting the number of events with two b-tagged jets in the selected sample of events.
**TnP**: Method for the measurement of the b-tagging efficiency in ttbar events in the semileptonic channel. The b-tagging efficiency is measured with a tag and probe method (TnP). As a tagging requirement, the CSVv2M requirement is applied to either the b-jet on the hadronic or leptonic side, while the b-jet from the other side is used as probe.
IterativeFit: Method for the measurement of the b-tagging efficiency in ttbar events in the dileptonic channel. This method is based on the calibration of the full b-tagging discriminator shape.

# ROCs for b-tagging



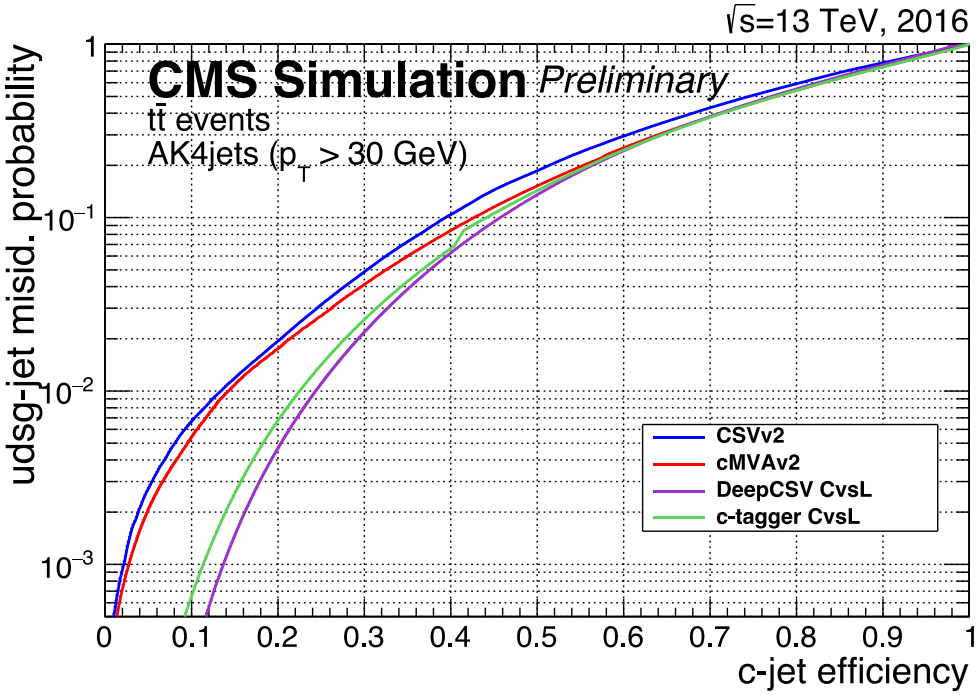

Performance of the b jet identification efficiency algorithms demonstrating the probability for non-b jets to be misidentified as b jet as a function of the efficiency to correctly identify b jets. The curves are obtained on simulated ttbar events using jets within tracker acceptance with $p_T$>30 GeV, b jets from gluon splitting to a pair of b quarks are considered as b jets. The lines shown are for CSVv2, DeepCSV, and cMVAv2. cMVAv2 uses also the information from the soft leptons inside jets, while CSVv2, DeepCSV do not. The performance in this figure serves as an illustration since the b jet identification efficiency depends on the pT and η distribution of the jets in the topology as well as the amount of b jets from gluon splitting in the sample.

# ROC for c vs b



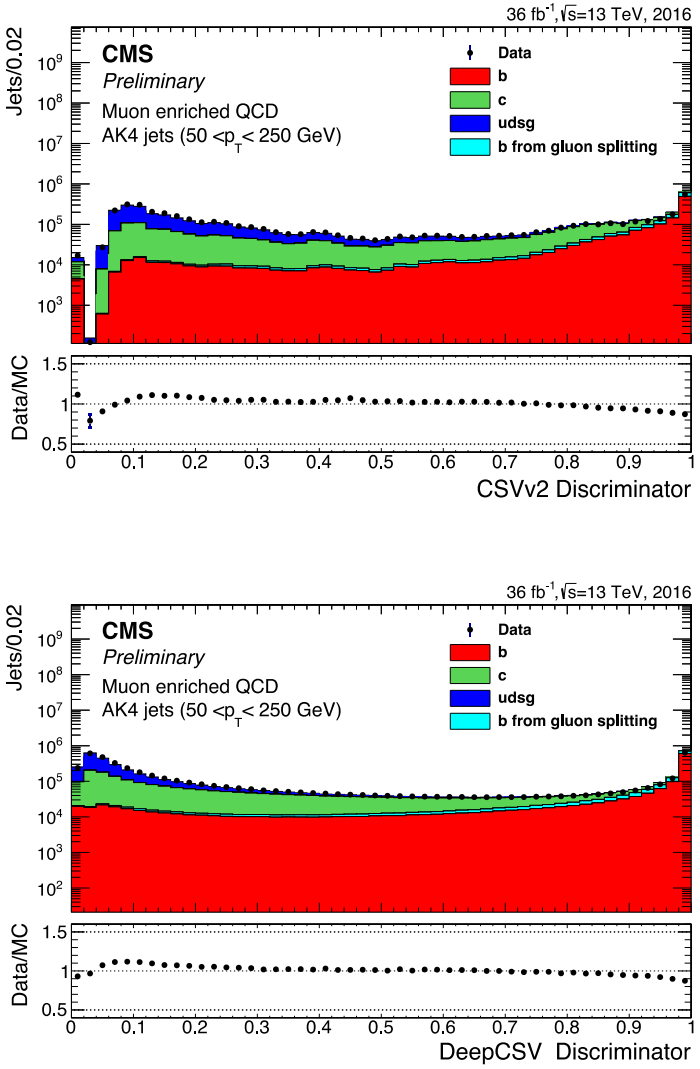

Performance of the c jet identification efficiency algorithms demonstrating the probability for b jets to be misidentified as c jet as a function of the efficiency to correctly identify c jets. The curves are obtained on simulated ttbar events using jets within tracker acceptance with $p_T$>30 GeV , b jets from gluon splitting to a pair of b quarks are considered as b jets. The lines shown are for CSVv2, DeepCSV CvsB, c-tagger CvsB and cMVAv2. cMVAv2 and the c-tagger use also the information from the soft leptons inside jets, while CSVv2, DeepCSV do not.

# ROC c vs. light



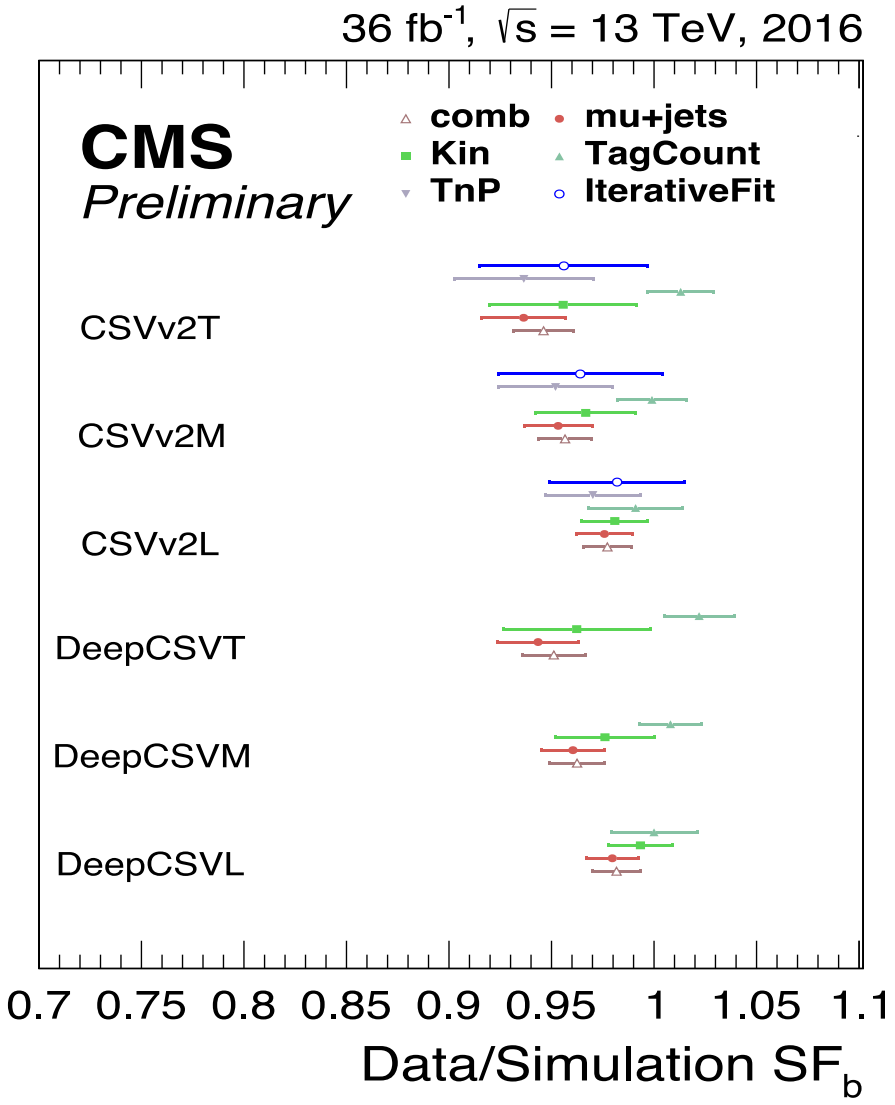

Performance of the c jet identification efficiency algorithms demonstrating the probability for light jets to be misidentified as c jet as a function of the efficiency to correctly identify c jets. The curves are obtained on simulated ttbar events using jets within tracker acceptance with $p_T$>30 GeV , b jets from gluon splitting to a pair of b quarks are considered as b jets. The lines shown are for CSVv2, DeepCSV CvsL, c-tagger CvsL and cMVAv2. cMVAv2 and the c-tagger use also the information from the soft leptons inside jets, while CSVv2, DeepCSV do not. The irregularity observed in the ROC curve of the c-tagger is caused by a sharp feature in the discriminator distribution due to jets without any selected tracks.

# Discriminator distributions



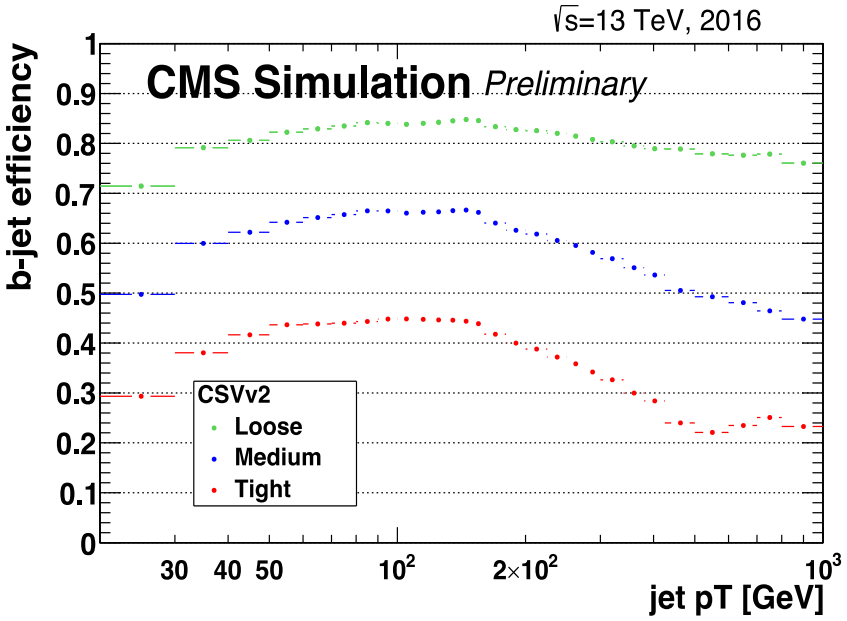

Distribution of the CSVv2 (top) and DeepCSV (bottom) discriminators for ak4 jets in a muon enriched jet sample. The markers correspond to the data. The stacked, coloured histograms indicate the contributions of the different jet flavours in the simulation. Simulated events involving gluon splitting to b quarks (“b from gluon splitting”) are indicated separately from the other b quark production processes (“b”). The distributions from the simulation have been scaled to match the observed number of entries in data. The last bins of the histograms contain all entries above the histogram range. The underflow bin is included in the first bin.

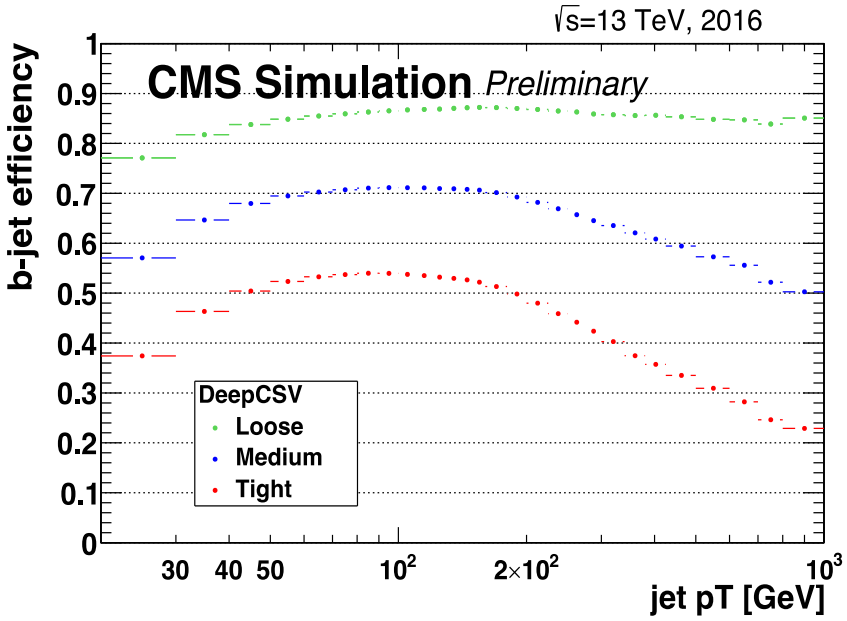
# SFs for CSVv2 and DeepCSV




Comparison between the scale factors measured by different methods in ttbar events (Kin, TagCount, TnP, IterativeFit), the combined scale factors obtained from the muon enriched sample (mu+jets), and the combined scale factors obtained from ttbar and muon enriched samples (comb). The "comb" combined scale factors are based on the mu+jets, Kin and TnP measurements for CSVv2 and on the mu+jets and Kin measurements for DeepCSV. The scale factors measured in the muon enriched sample are averaged over the observed pT spectrum of the b jets from ttbar decays. For the IterativeFit method a cumulative scale factor for jets with $p_T$ above 30 GeV is extracted to allow a comparison.

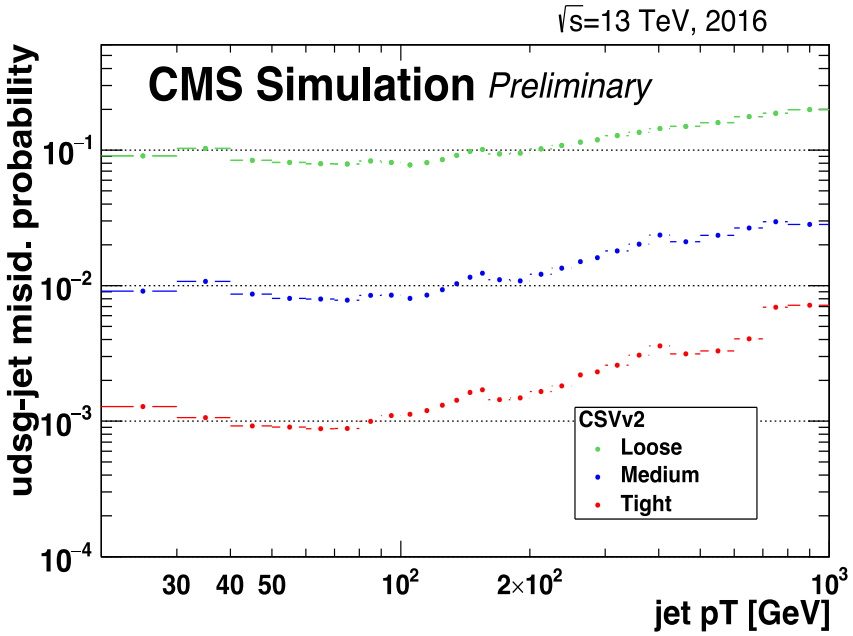
# b-jet efficiency as a function of $p_T$

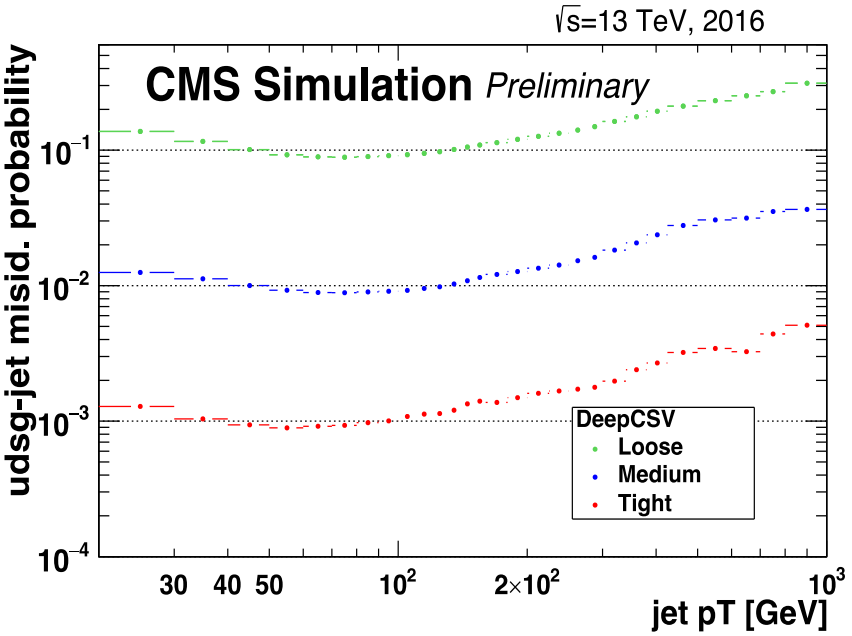



b-jet efficiency as a function of the jet transverse momentum for the CSVv2 and DeepCSV algorithms. These efficiencies are obtained on simulated ttbar events using jets within tracker acceptance with $p_T$>30 GeV. The last bin includes the overflow.

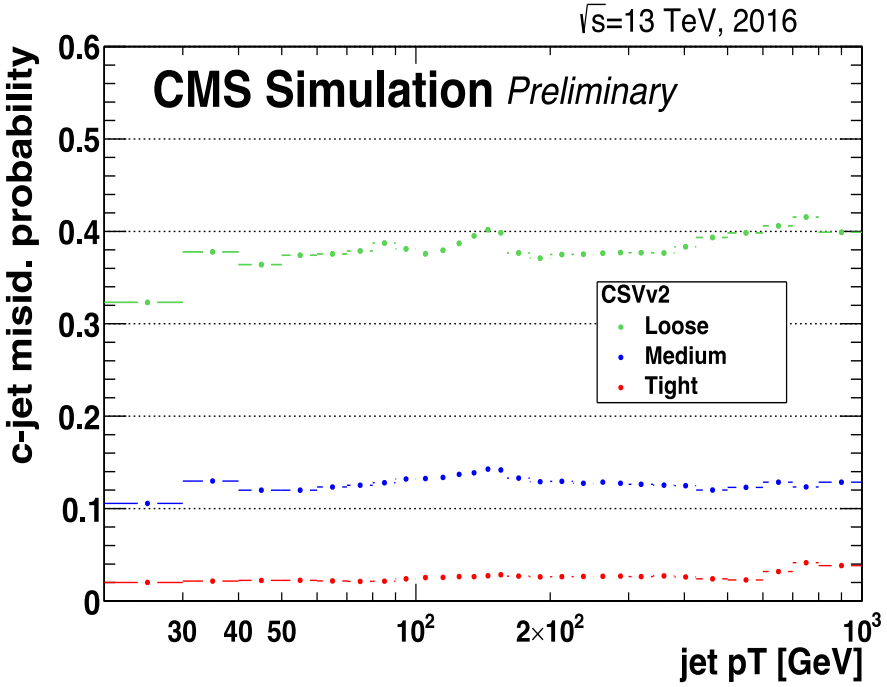
# Light-jet misid prob. as a function of $p_T$

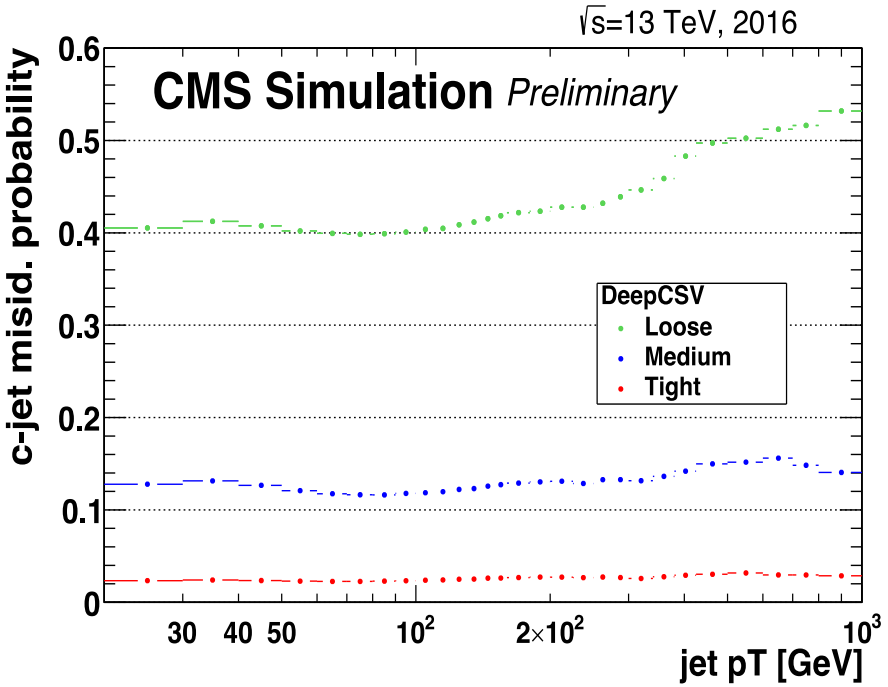



Light jet efficiency as a function of the jet transverse momentum for the CSVv2 and DeepCSV algorithms. These efficiencies are obtained on simulated ttbar events using jets within tracker acceptance with pT>30 GeV. The last bin includes the overflow.

# c-jet misid prob. as a function of $p_T$




c-jet efficiency as a function of the jet transverse momentum for the CSVv2 and DeepCSV algorithms. These efficiencies are obtained on simulated ttbar events using jets within tracker acceptance with pT>30 GeV. The last bin includes the overflow.