Flavour Tagging at CMS

Luca Scodellaro
Instituto de Física de Cantabria - CSIC
On behalf of the CMS Collaborations

CMS Heavy Flavour Tagging Workshop
April 11th-13th, 2018
Bruxelles (Belgium)

Outline

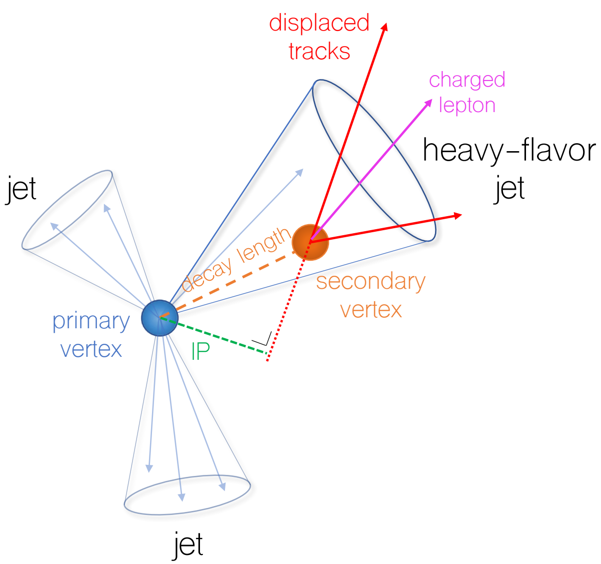
- Properties of heavy flavour jets
- Heavy Flavour tagging algorithms:
  - Identification of jets from bottom quarks
  - Identification of jets from charm quarks
  - Measurements of identification performance on data
- Identification of b jets in events with boosted topologies:
  - AK8 b tagging, subjet b tagging, double-b tagger
  - Performance measurements in boosted topologies
- Performance of b jet identification at trigger level
- Preparation for High-Luminosity LHC

Except where specified, all material presented is from:

arXiv:1712.07158 (submitted to JINST)

Properties of Heavy Flavour Jets

- Heavy hadrons from the hadronization of b (and c) quarks present special properties:
  - **Long litime:** their displaced decays results in tracks with large impact parameter (IP) and secondary vertices
  - **Large mass:** their decays products have a higher momentum relative to jet direction
  - **Semileptonic decays:** presence of soft muons or electrons in the jet



- Jet flavour assignment in simulated events:
  - Generated heavy hadrons used in jet clustering with momentum rescaled to a negligible value (ghost hadron)
  - Jet flavour assignment based on the presence of ghost b or c
  - Jets not matched to a gen jet (p_T>8 GeV) are treated as pileup

b Tagging Algorithms at CMS

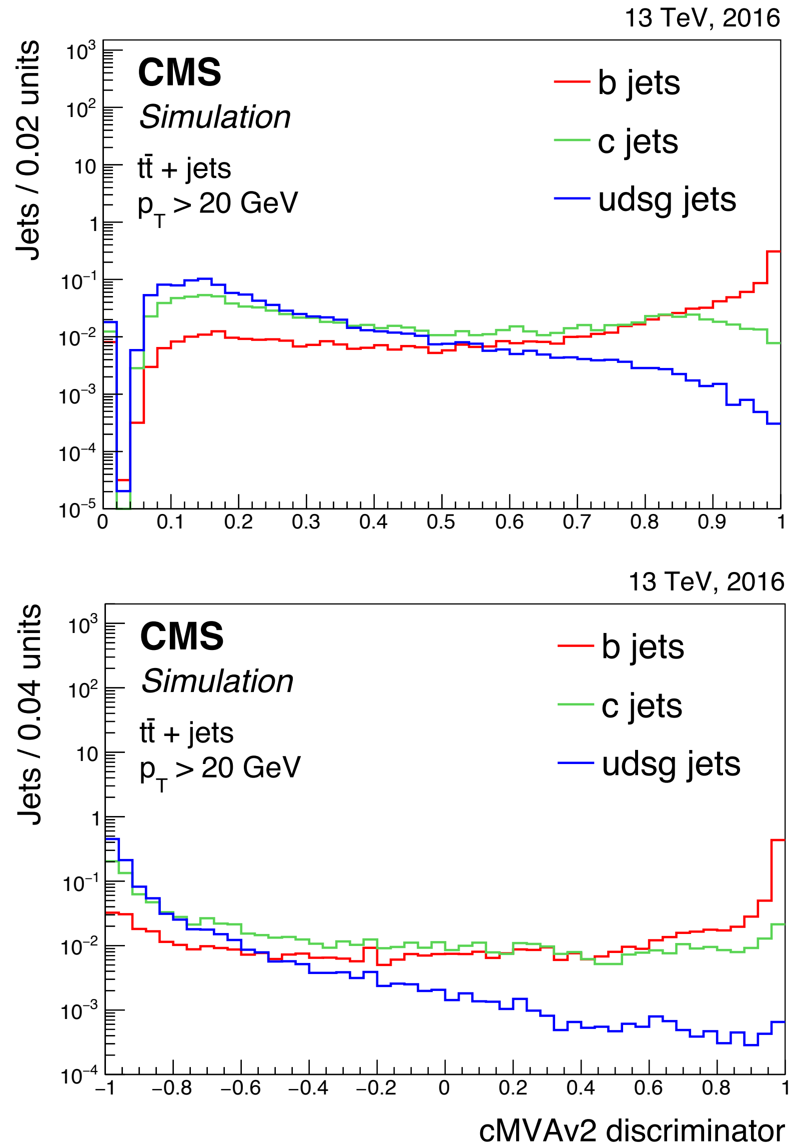
- Exploiting information from one or more b jet properties

b jet property	Algorithms
Tracks with large impact parameters (IP)	TCHP, TCHE, JP, JBP
Secondary vertices (SV)	SSVHP, SSVHE, Inclusive Vertex Finder
Soft leptons from semi-leptonic B decays	Soft Lepton Taggers
Multivariate combinations	CSV, CSVv2, cMVAv2, DeepCSV, DeepFlavour

- The algorithms provide a discriminant value for each jet

Multivariate Combinations

- The new CSVv2 is an evolution of the RunI CSV algorithm:
  - Neural network (NN) instead of likelihood ratio allows to combine more variables
  - Secondary vertices from the Inclusive Vertex Finder algorithm:
    - Fitting inclusively the tracks in the event, without prior association with the jets
- The cMVAv2 tagger combines the outputs from CSVv2, JP, JBP, and soft lepton taggers



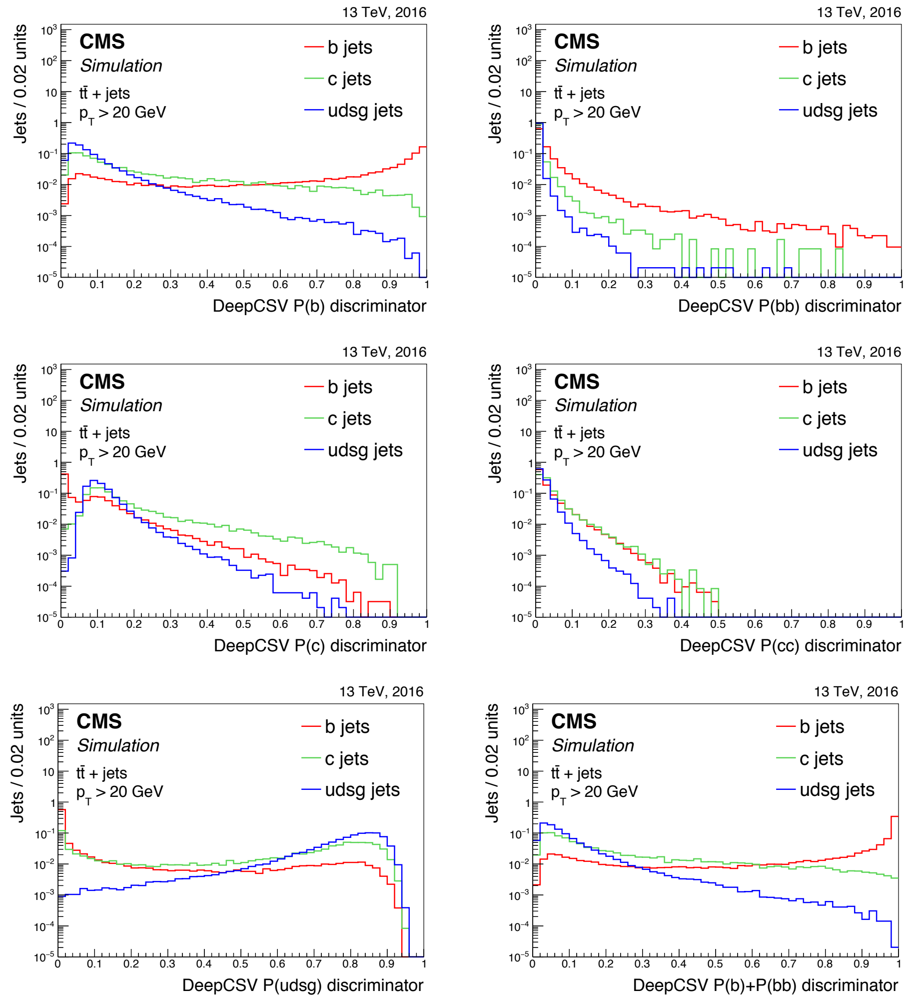
Deep Learning in b Tagging

- Use of more sophisticated neural network classes allows to better exploit the information available for b-tagging:
  - Can combine a large number of input features
  - Can handle more low-level information
  - Allows for multi-classification, providing an output probability for each jet flavor hypothesis

- DeepCSV: a new version of the CSVv2 tagger has been developed through the use of deep neural networks (DNN)
  - Four hidden layers of a width of 100 nodes each
  - Same track selection and input observables as CSVv2
  - However, first six most displaced tracks are used instead of first four tracks as for CSVv2

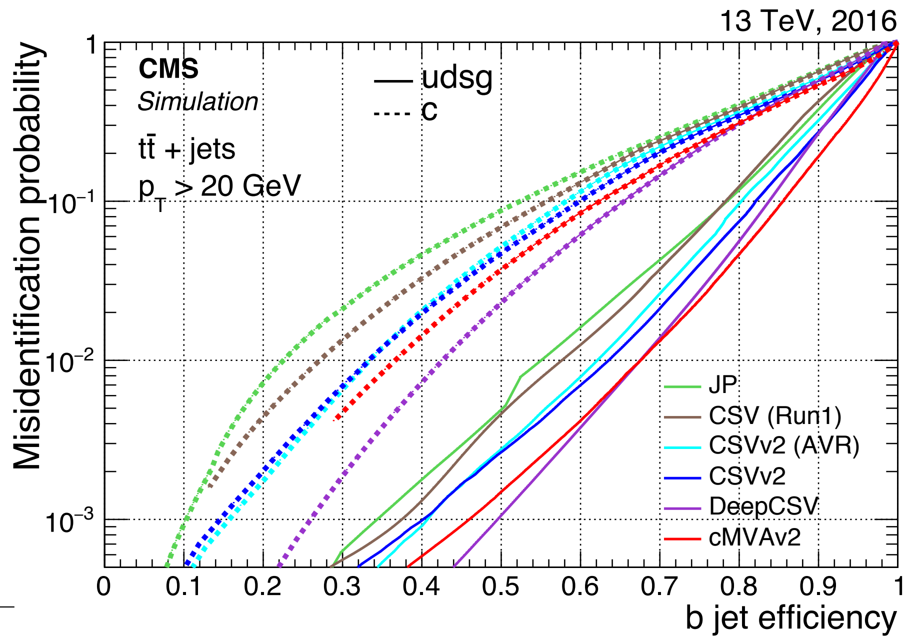
DeepCSV Tagger

- The output of the algorithm consists of a probability for each of the five classes of jets used in the training:
  - Jet contains exactly one or at least two b quarks
  - Jet contains exactly one or at least two c quarks
  - None of the above
- It has been shown that summing the probabilities of two classes is equivalent to doing a combined training:
  - DeepCSV(b+bb) = DeepCSV(b) + DeepCSV(bb)



Algorithm Performances in 2016

- Probability for non-b jets to be mis-identified as b jets, as a function of the b tagging efficiency

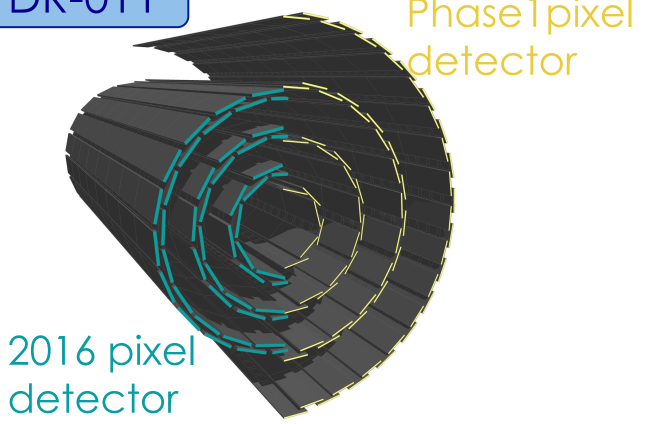


Tagger	Working point	ε_b (%)	ε_c (%)	ε_{udsg} (%)
Jet probability (JP)	JP L	78	37	9.6
	JP M	56	12	1.1
	JP T	36	3.3	0.1
Combined secondary vertex (CSVv2)	CSVv2 L	81	37	8.9
	CSVv2 M	63	12	0.9
	CSVv2 T	41	2.2	0.1
Combined MVA (cMVAv2)	cMVAv2 L	84	39	8.3
	cMVAv2 M	66	13	0.8
	cMVAv2 T	46	2.6	0.1
Deep combined secondary vertex (DeepCSV) $P(b) + P(bb)$	DeepCSV L	84	41	11
	DeepCSV M	68	12	1.1
	DeepCSV T	50	2.4	0.1

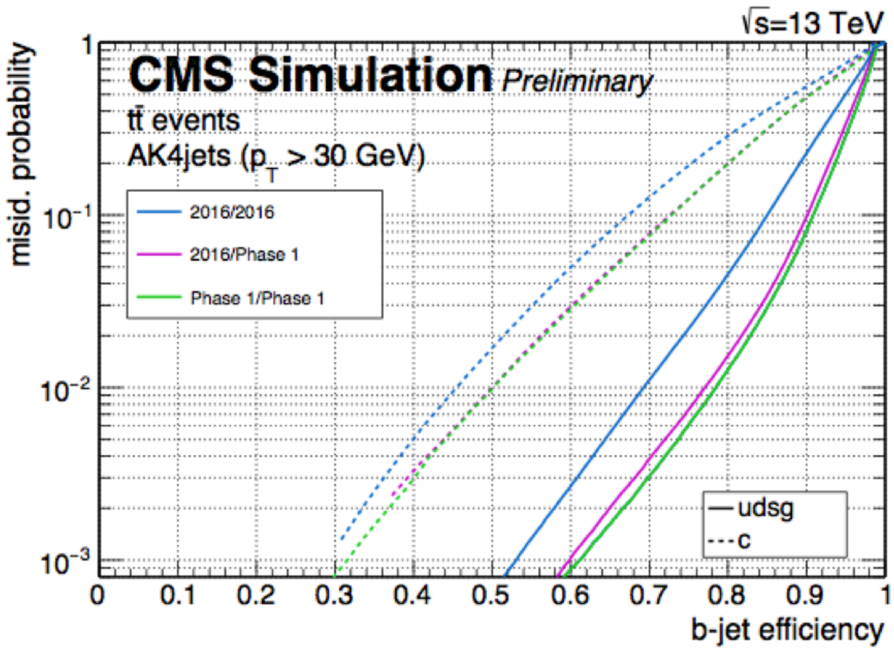
- Three working points defined as the cut on the discriminator value allowing to reduce the mis-identification probability for light jets to 10, 1 and 0.1%

Algorithm Performances in 2017

CMS-TDR-011



- The CMS Phase 1 upgrade included a new pixel detector with an additional layer, closer to the beam spot



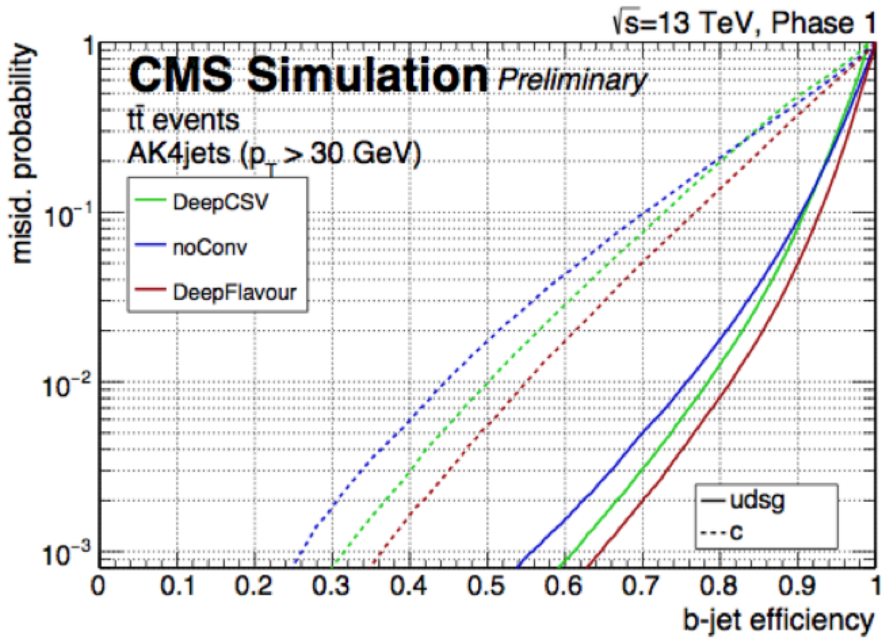
- Comparison of DeepCSV performance with 2016 detector, Phase 1 detector and 2016 training, and with Phase 1 detector and new dedicated training

CMS DP-2017/013

DeepFlavour Tagger

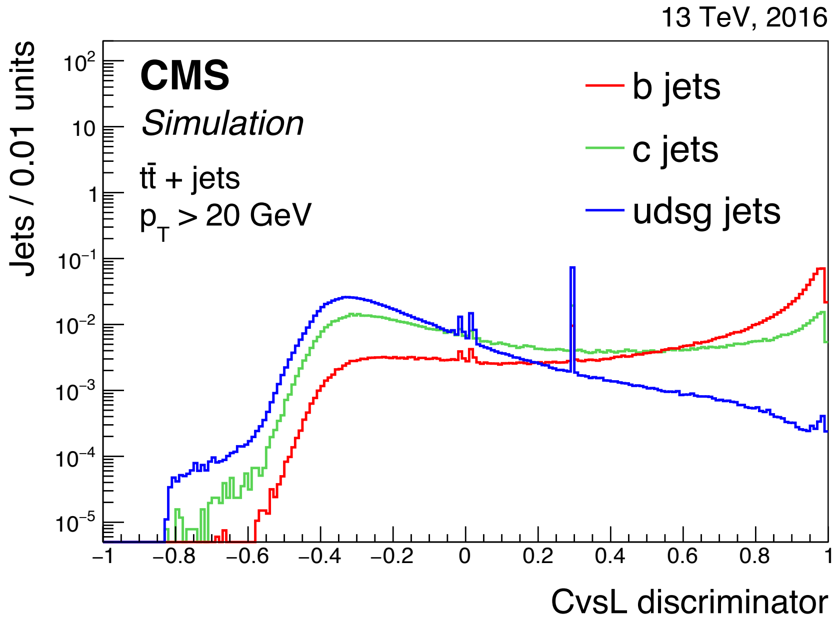
CMS DP-2017/013

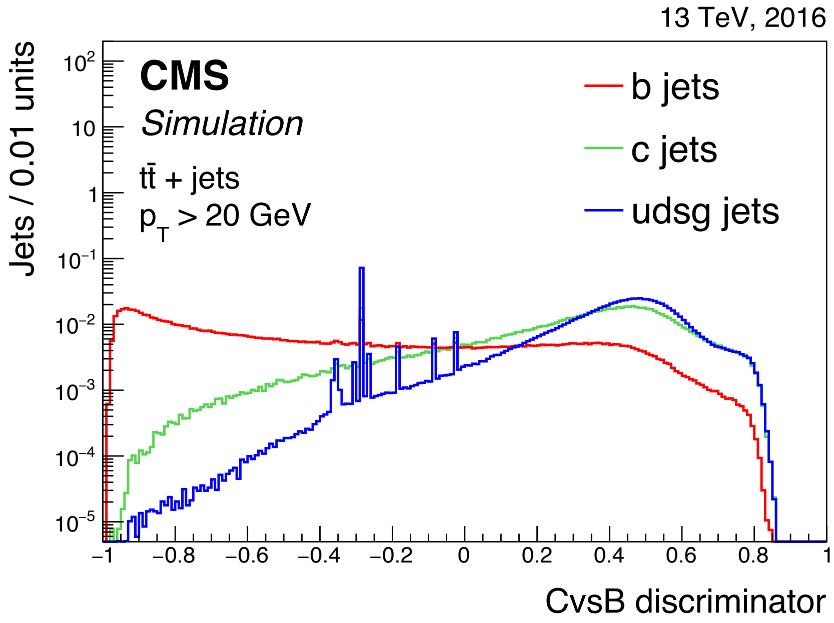
- Architecture of the DeepFlavour tagger:
  - No quality requirements applied to charged track selection
  - Using 16 (6) properties of charged (neutral) particle-flow jet constituents, and 17 properties of SVs associated to the jet
  - Properties of each category engineered by 1x1 convolutional layers
  - Output is merged to jet global properties and fed to a dense NN
- Expected performance:
  - *4% absolute improvement in b-tag efficiency for a mistag rate of 0.1% against DeepCSV*
- Extended to gluon vs quark discrimination (DeepJet):
  - CMS DP-2017/027



Identification of c Jets

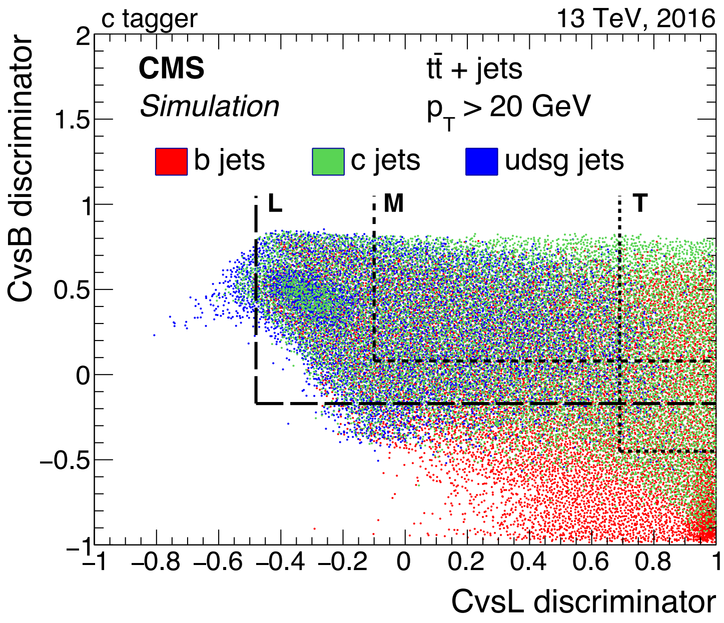
- The algorithm of c jet identification is based in similar input variables and jet vertex categories as defined in CSVv2
- In addition, the c-tagger exploits information from the soft lepton taggers to add more observables and jet categories
- A Gradient Boosting Classifiers (GBC) is used for two trainings to discriminate c jets against light (CvsL) and b (CvsB) jets

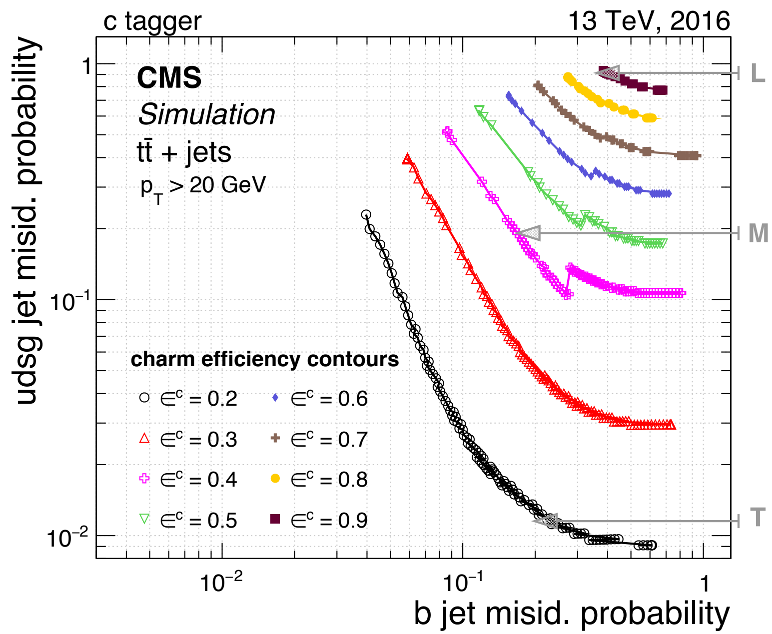




c Jet Tagging Performance

- Performance of the c-tagger can be studies by applying simultaneously thresholds on CvsL and CvsB to define curves of constant c-efficiency in the b vs light mis-id. probability plane





Working point	ε_c (%)	ε_b (%)	ε_{udsg} (%)
c tagger L	88	36	91
c tagger M	40	17	19
c tagger T	19	20	1.2

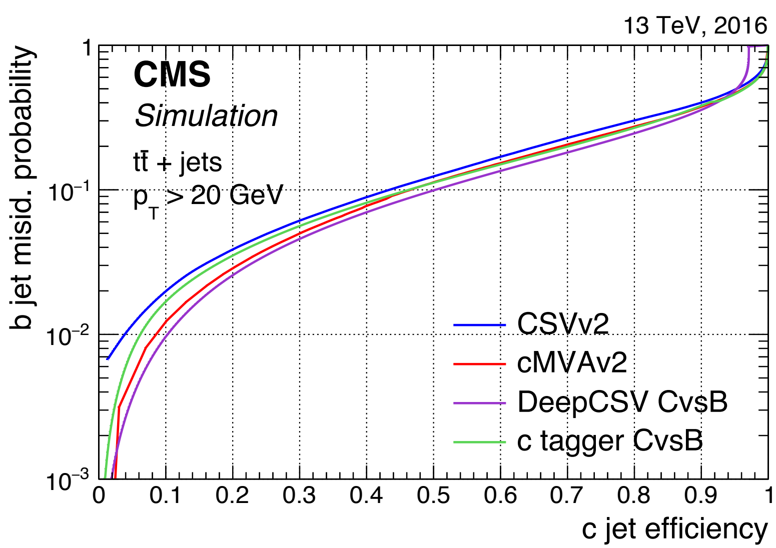
c Jet Tagging with Multiclassifiers

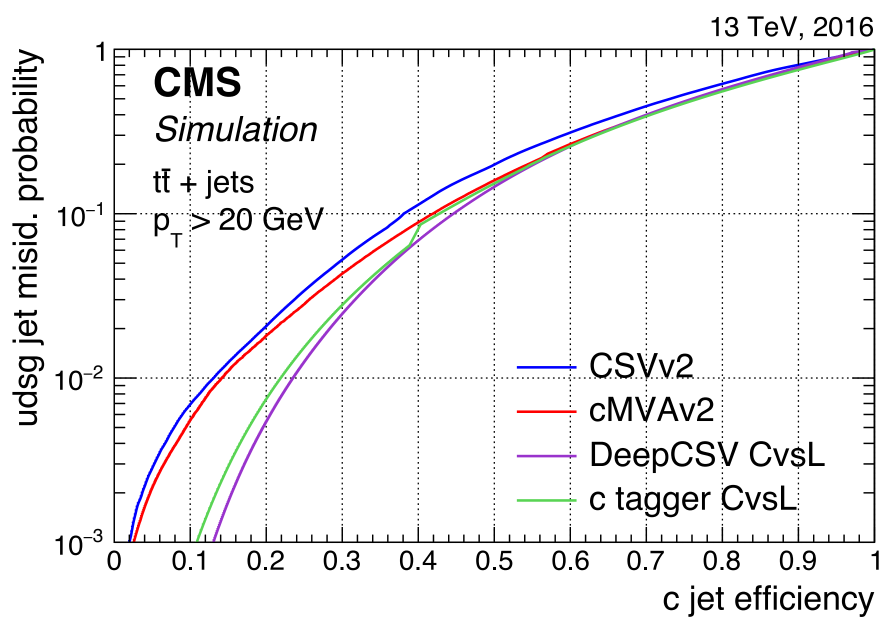
- c taggers can be built from DeepCSV outputs:

$$\text{DeepCSV CvsB} = \frac{P(\text{c}) + P(\text{cc})}{1 - P(\text{udsg})},$$

$$\text{DeepCSV CvsL} = \frac{P(\text{c}) + P(\text{cc})}{1 - (P(\text{b}) + P(\text{bb}))},$$

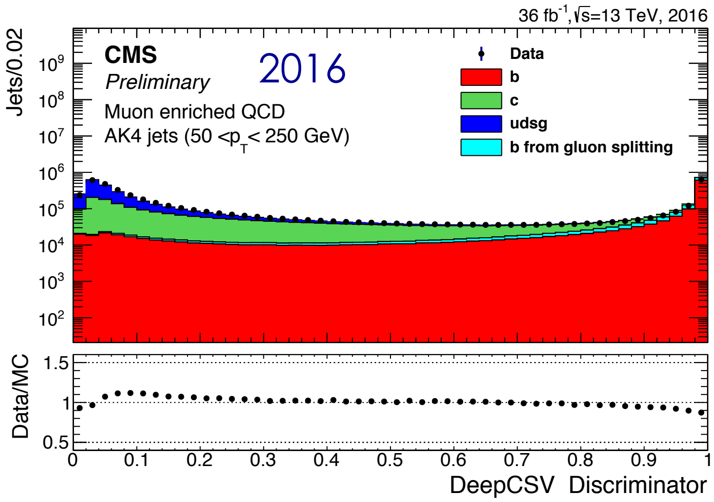
- DeepCSV is already outperforming the dedicated c-tagger

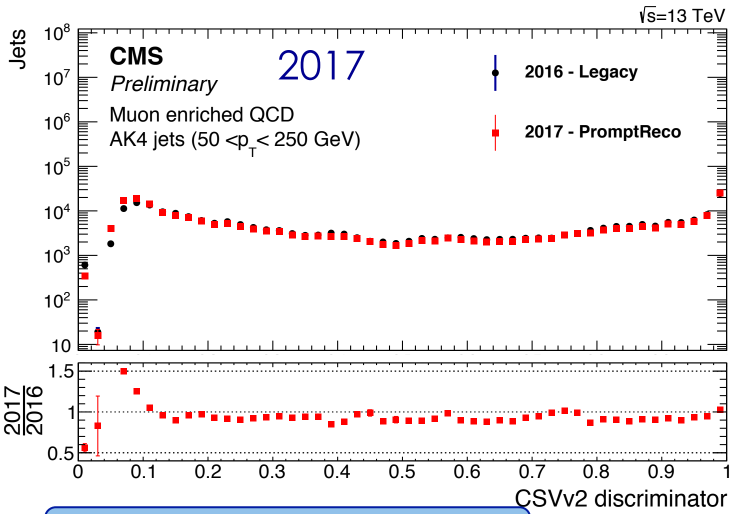




Performance In Data

- Samples with different jet flavor composition are exploited to commission the algorithms:
  - Inclusive jets from QCD processes
  - Jets from QCD with an embedded soft muon
  - Top pair production events
- To correct b-tagging efficiencies in physics analysis, data-to-MC scale factors (SFs) are computed for each operating point through data driven techniques

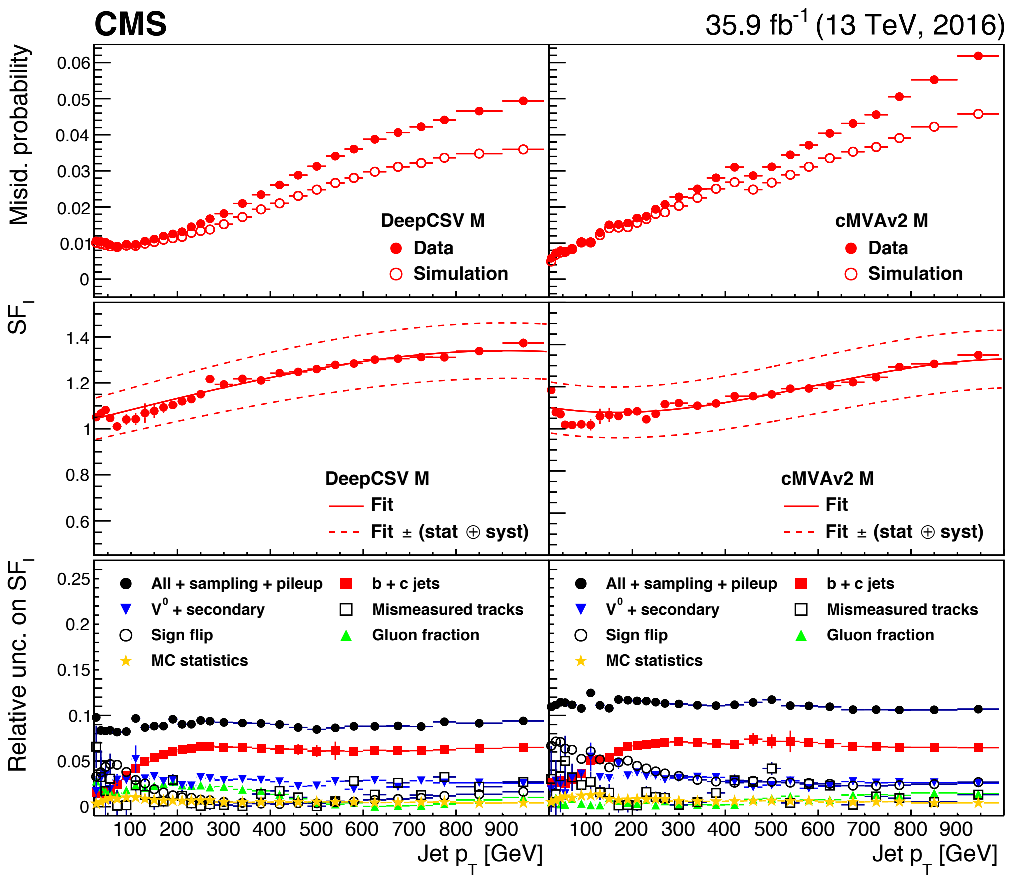




CMS DP-2017/037

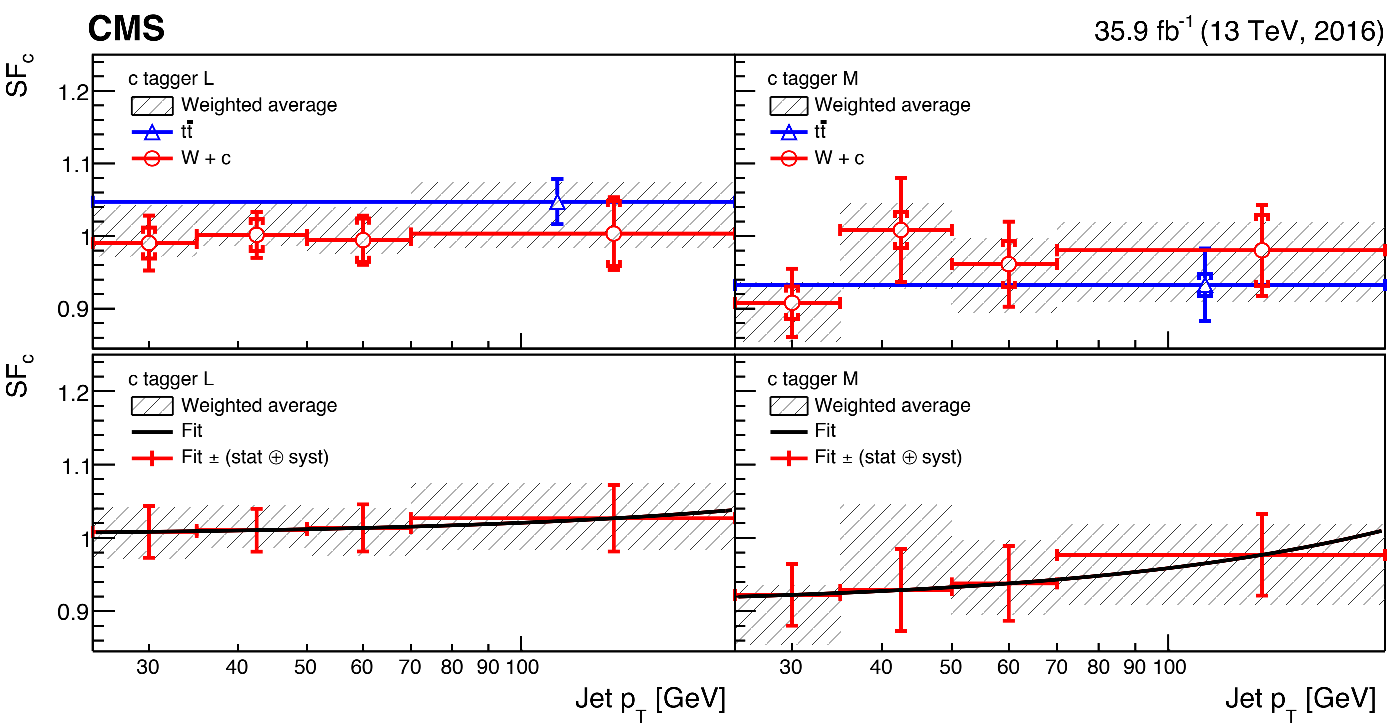
Scale Factors for light Jets

- SFs for light-jet mistag rate measured in inclusive jet samples:
  - Negative taggers built using only tracks with negative IP and SVs with negative flight distance
  - Negative tag rate from data corrected to positive mistag rate through a MC derived scale factor

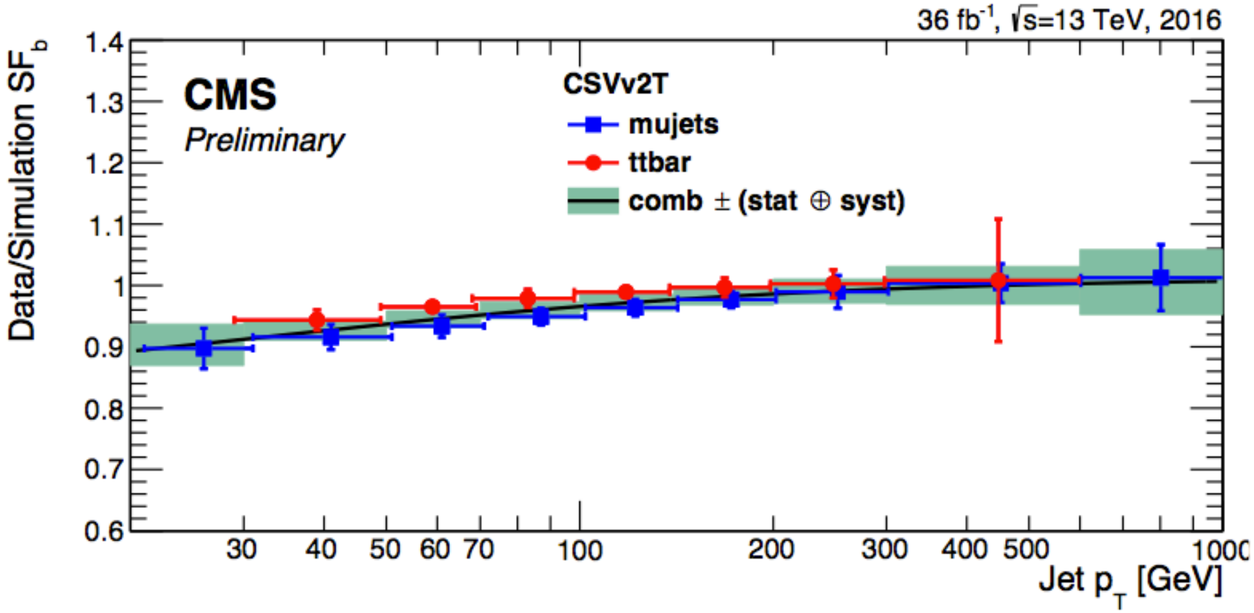


Scale Factors for c Jets

- CMS: performance measured in two c-jet enriched samples:
  - W+c->l ν c events, selected by requiring a soft muon in the c-jet
    - Background subtraction from events with same-sign leptons
  - ttbar events in lepton+jets final states
    - Fit to a mass discriminant λ_M to extract the c tagging efficiency



Scale Factors for b Jets

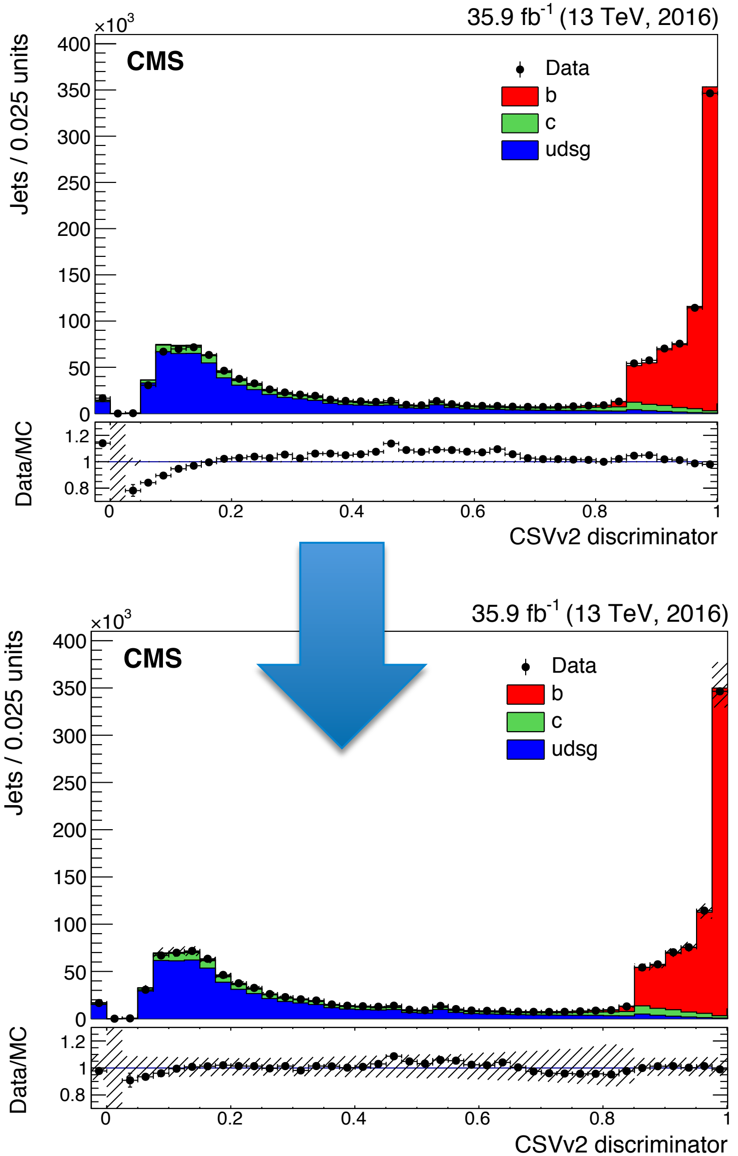


- Scale factor measurements in CMS exploits various methods:
  - QCD muon-eriched based: PtRel, System8, Lifetime Tagger
  - ttbar based: kinematic fits in dilepton and lepton+jets channels
- Single measurements are combined to reach the best precision

Shape Discriminator Corrections

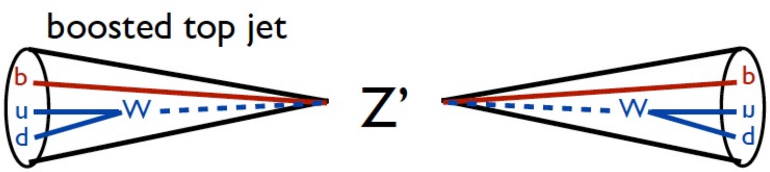
- Aiming to calibrate the whole b-tagging discriminator shape:
  - Designed for analyses that want to use the discriminator in a fit or a MVA rather then just select jets above a certain threshold

- Simultaneously determining reweighing factors for b and light jets by a iterative procedure in two different samples:
  - Dilepton ttbar events
  - Z -> ll events

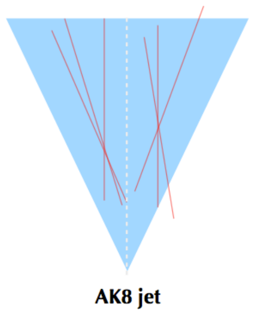


b-Tagging in Boosted Topologies

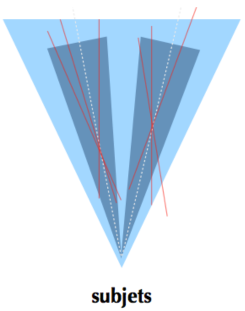
- In high energy collisions, particles decaying to b quarks can be produced with large momentum (boosted topology):
  - B decay products can overlap with particles from other jets
  - Important in many BSM searches



- Two approaches developed during Run1
- <u>AK8 b tagging:</u>
  - b tagging algorithms applied on all the tracks in the reconstructed AK8 jets
  - Relaxed criteria for assigning tracks and SV to jets

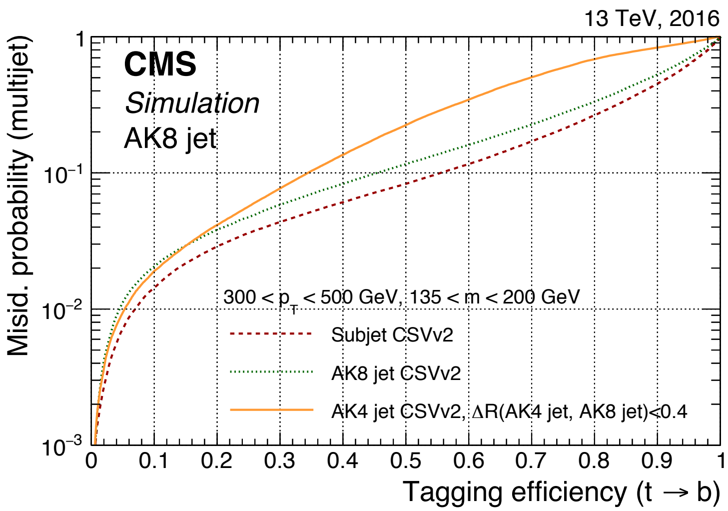


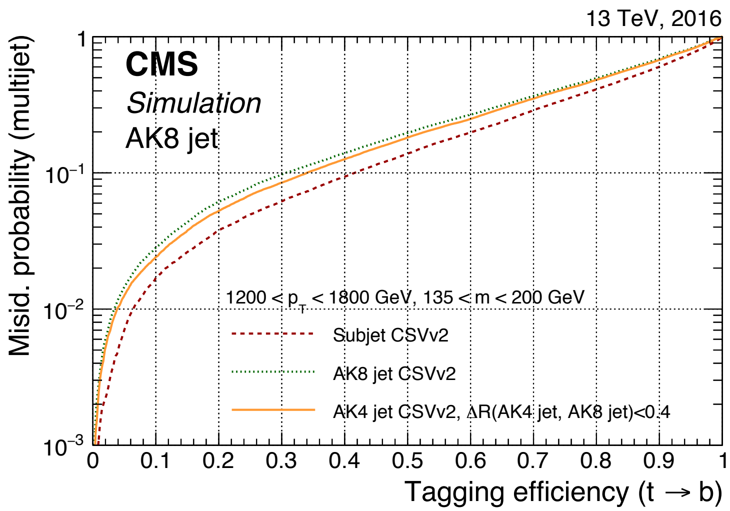
- <u>Subjet b tagging:</u>
  - Soft drop declustering to resolve jet substructure
  - Applying b tagging criteria on individual subjets



b Tagging in Boosted Topologies

- Subjet b tagging still baseline for boosted top quarks
  - CSVv2 algorithm used both for AK8 and subjet b tagging

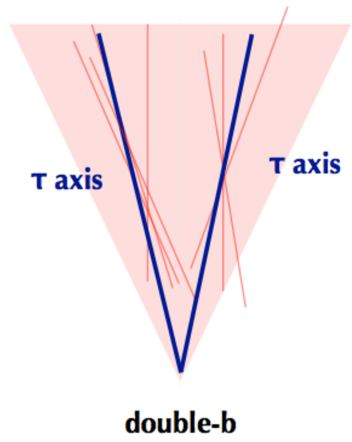




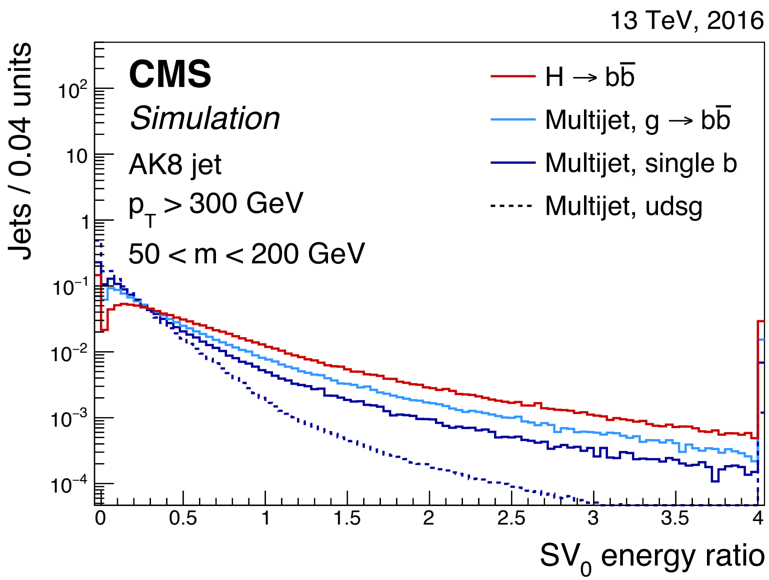
- Performance for H->bb identification are discussed in the next slides

Double-b Tagger

- In RunII, new dedicated algorithm aimed at tagging boosted decays to b pairs:
  - Exploiting not only the presence of two b in the jet, but also the correlations between their flight directions



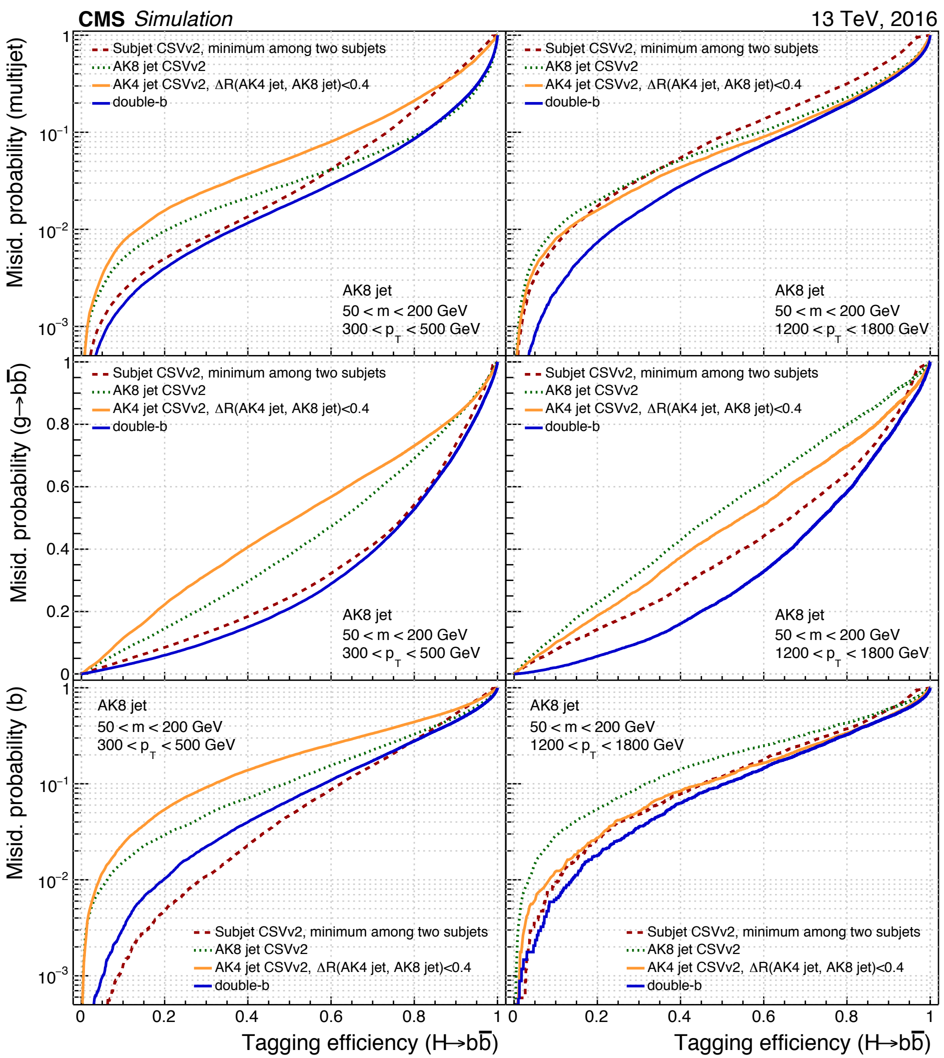
- N-subjettiness axes are used to associate tracks and vertex to the subjets, and to build the input observables



Performance for H->bb Tagging

- Double-b tagger outperforming AK8 and subjet b tagging for H->bb against multijet and g->bb backgrounds

- Subjet b tagging has better performance against single b jet background at low AK8 jet p_T, where the two subjets are very well separated

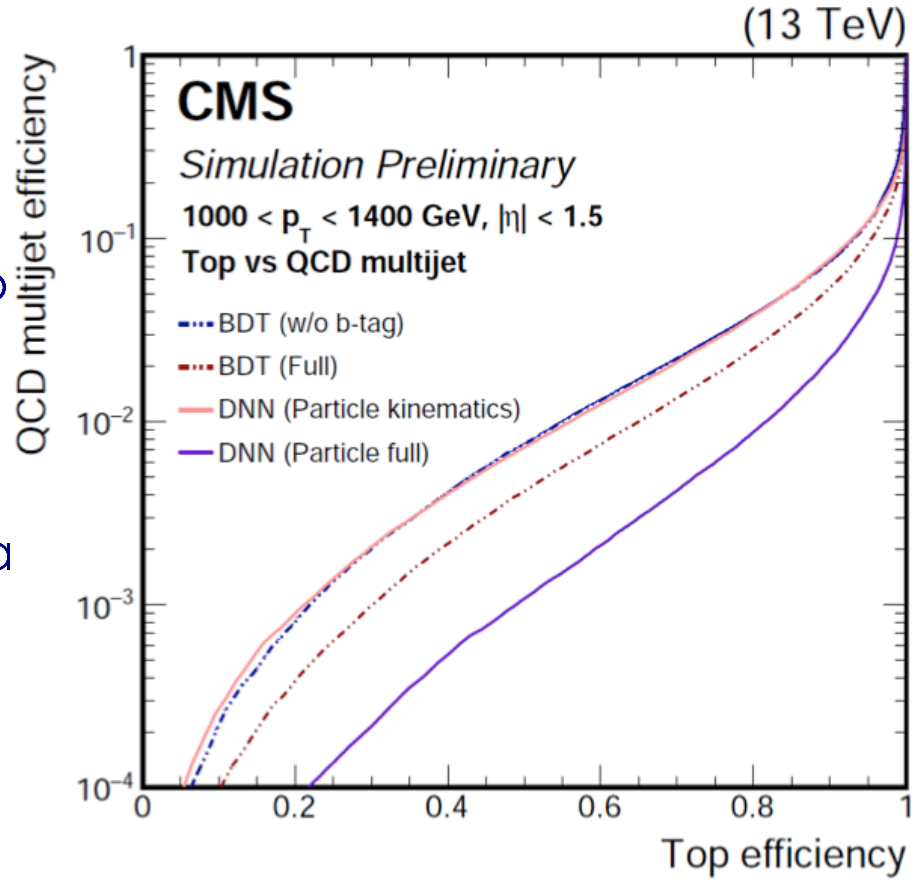


Boosted Jet Identification with DNN

- Identification of hadronically decaying boosted top quarks using deep neural networks

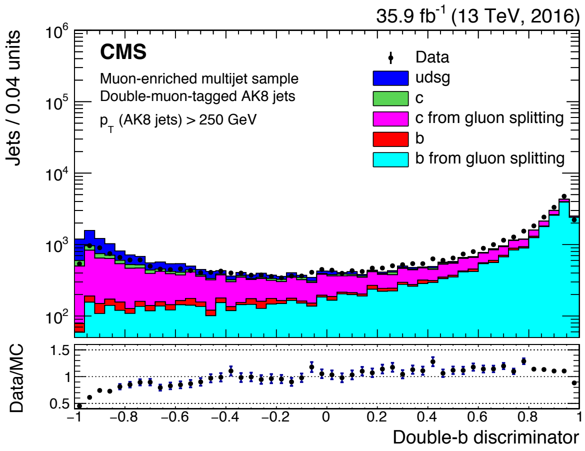
CMS-DP-2017/049

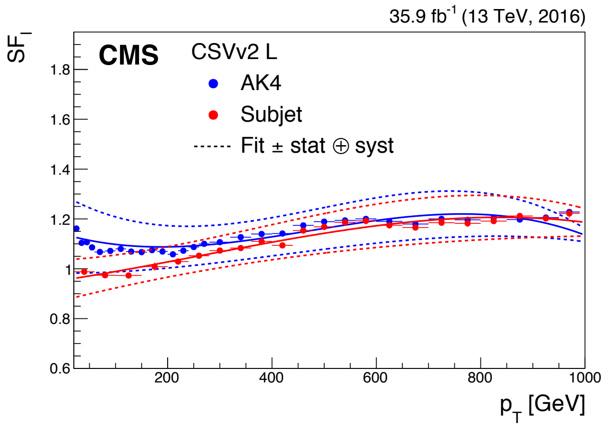
- Using an 1D convolutional NN on jet constituent particles:
  - Comparing a version using particle kinematic variables to a full version exploiting b tagging related information
  - Also comparing to a AK8 jet classification algorithm using a boosted decision tree, based only on jet observables

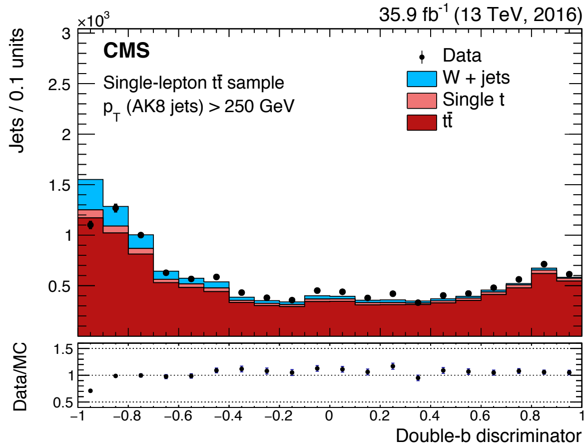


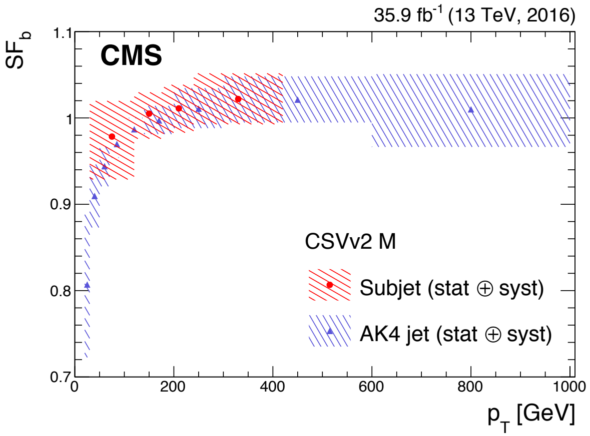
Boosted b Tagging Validation in Data

- Commissioning and measurements of efficiencies based on similar techniques as for AK4 jets in muon-enriched QCD events
  - Selecting AK8 jets with one (subjet b tagging) or two (double-b tagger) subjets containing a soft muon
- Misidentification probability also measured:
  - Subjet b tagging: inclusive jet data
  - Double-b tagger: boosted l+jets ttbar events



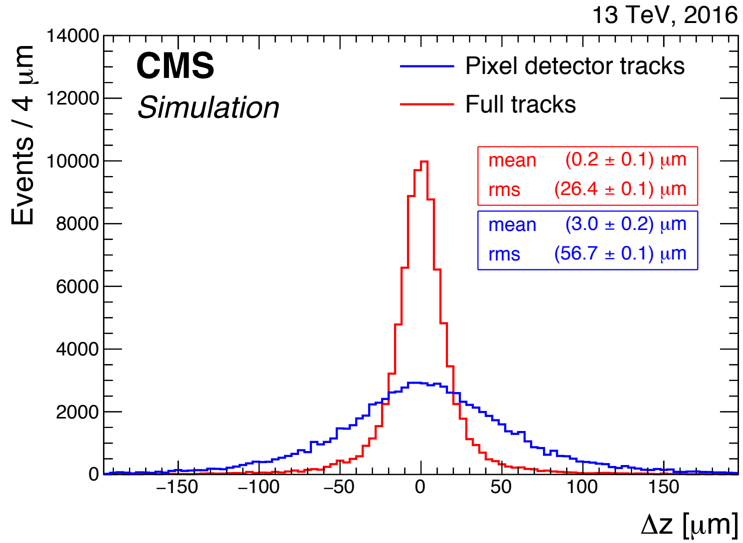


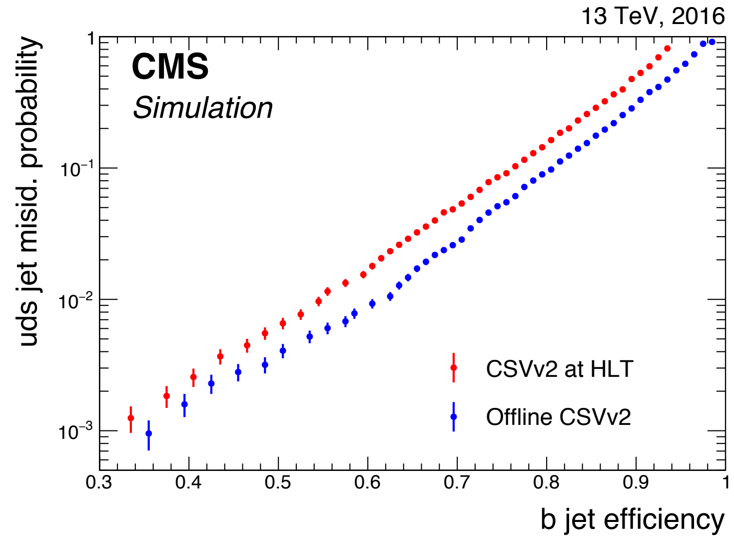




b Tagging at Trigger Level

- Primary vertex (PV) and track reconstruction at trigger level are done via an iterative procedure:
  - Estimate of the PV projecting the pixel hits along the jet direction
  - Regional pixel tracking
  - Pixel tracks used as seed for full track and PV reconstruction
- Tracks and PV are used as input for the CSVv2 algorithm
  - The performance of the online b tagging is compared to offline performance in simulated ttbar events (PU=35, $\Sigma\, p_T^{jet}$>250 GeV)

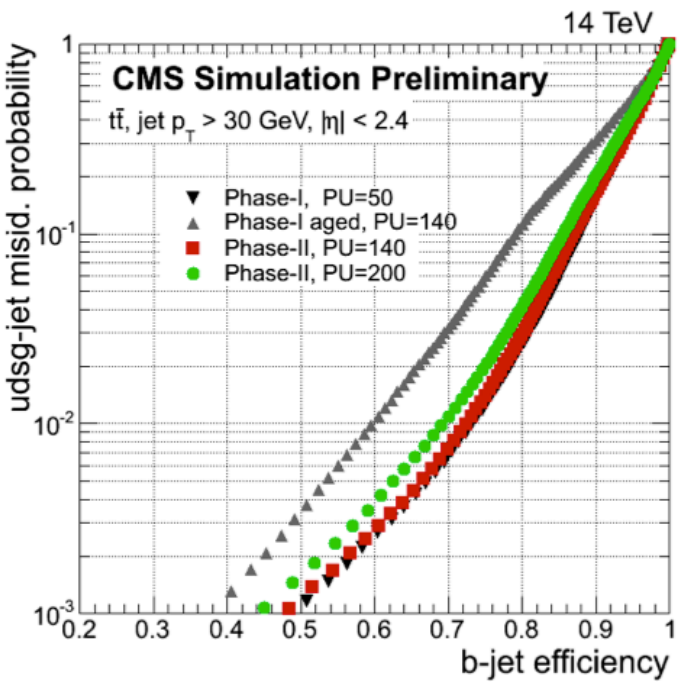


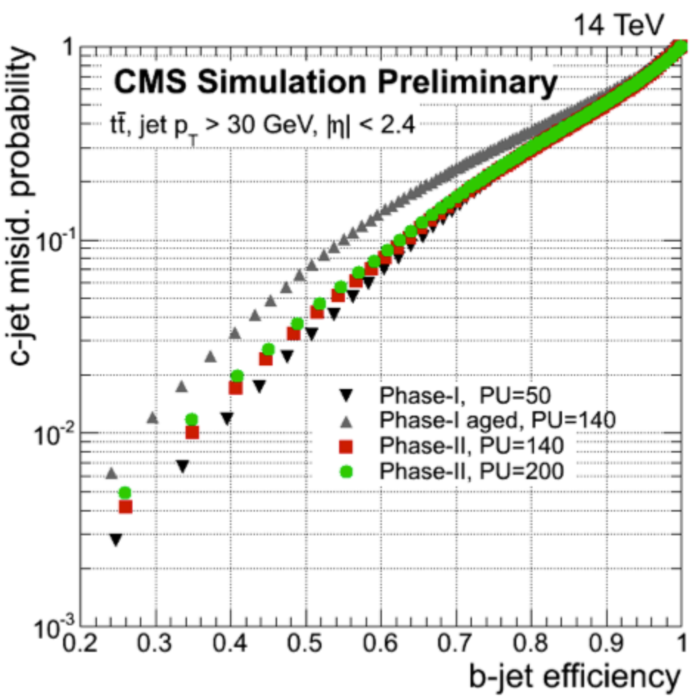


High Luminosity Upgrades

CMS DP-2016/065

- Major upgrades of the CMS detector planned to operate during the High Lumiosity (HL) LHC phase
  - Trackers will be replaced with new detector with higher granularity, radiation robustness and extended coverage
- First studies show that the b-tagging algorithms can operate in the complex high PU environment expected during HL-LHC





Conclusions

- b-Tagging is a fundamental tools in most physics analyses
- CMS reached a significant improvement on their algorithms in RunII, and new promising ideas for further developments are being explored
- Not only algorithms, but also the measurements of their performance on data had benefited from new ideas (and of increased sample statistics) in 2016
- Techniques are being extended to cover more specific topologies becoming ever more important with the increase of the LHC collisions center-of-mass energy
- More challenge ahead: already working to maintain b-tagging a successful tool in the next decade of data taking

Backup Material

Track Selection

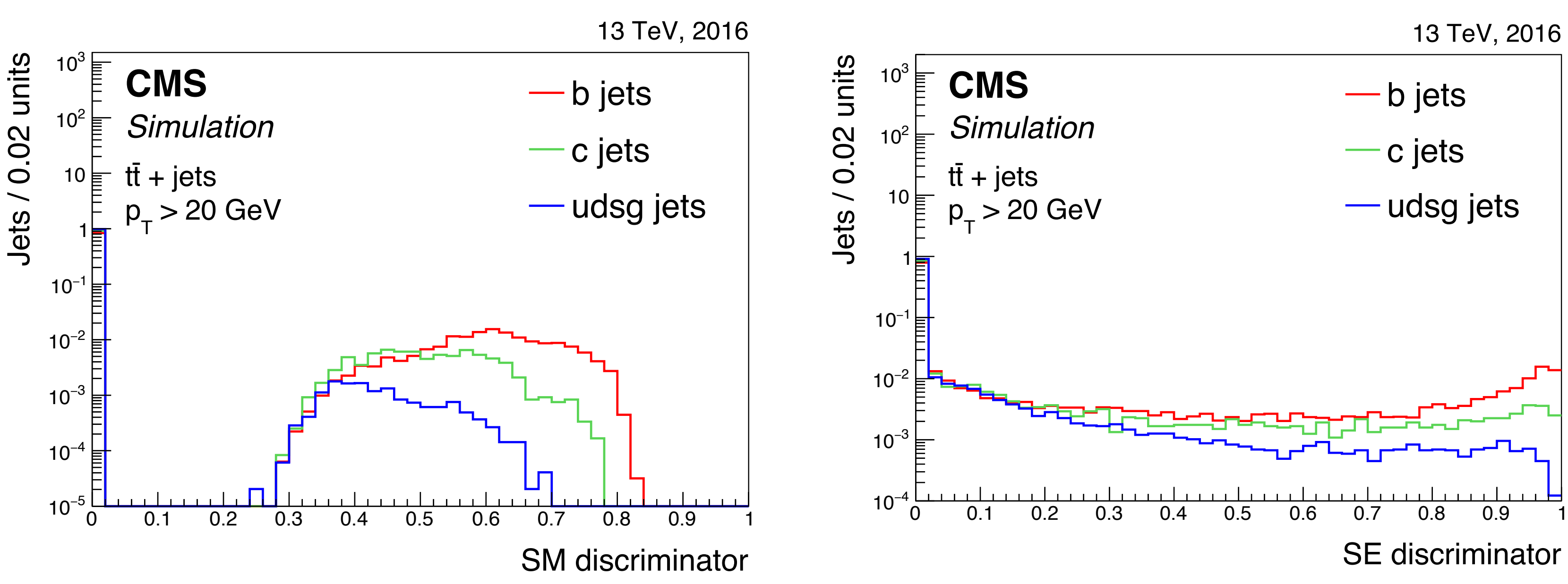
- The tracks used in the algorithms of b-jet identification must satisfy the following quality criteria:
  - Transverse momentum p_T>1 GeV
  - Normalized χ^2<5
  - At least one hit in the pixel layers of the tracker
    - This requirement has been significantly loosened with respect to Run1 to cope with the reduced hit efficiency at high luminosity
  - Transverse impact parameter IP_{xy}<0.2 cm
  - Longitudinal impact parameter IP_z<17 cm
  - Distance between track and jet axis at their point of closest approach D<0.07 cm
  - Decay length L<5 cm

Secondary Vertices (SV)

- Adaptive vertex reconstruction (AVR) algorithm:
  - It is the algorithm used for b-tagging during LHC Run1
  - Track associated to the jet are fitted through the adaptive vertex fitter
  - Several selection criteria applied to remove secondary vertices less likely to originate from a B hadron decay
- Inclusive vertex finder (IVF):
  - Using inclusively the tracks in the event, without prior associations with the jets
  - Cluster of tracks are identified and fitted around displaced "seed" tracks with IP>50 μm and IP significance > 1.2
  - Tracks in common with the event primary vertex are arbitrated, and the secondary vertex is refitted if at least two tracks remain

Soft Lepton Taggers

- Soft lepton variables are used to build a soft lepton tagger
- A Boosted Decision Tree (BDT) is used to combine:
  - 2D and 3D impact parameter significance of the lepton
  - ΔR(jet,lep), p_T^{lep}/p_T^{jet}, lepton p_T^{rel}
  - For soft electron: MVA-based electron identification



DeepJet

- Probability for gluon jets to be misidentified as a light quark (uds) jet, as a function of the efficiency to correctly identify light quark jets

CMS DP-2017/027