Heavy-Flavour tagging in CMS Run 2

Alessandro Calandri - ETH Zürich
on behalf of the CMS Collaboration

ICHEP 2020 - 40th International Conference on High Energy Physics

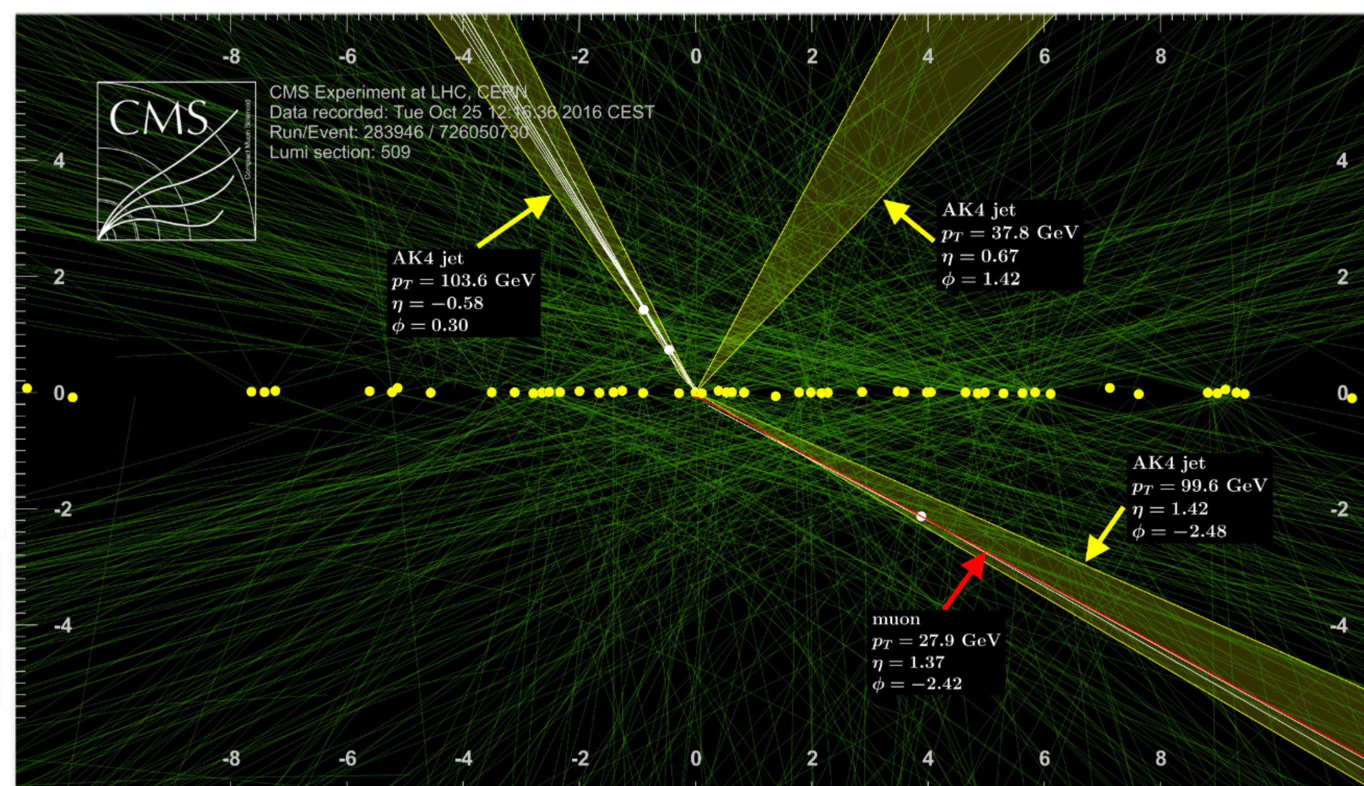
Outline of the talk

- ➡ Jet flavour identification and the corresponding calibration of the simulated response to data are a crucial feature for Higgs, SM and BSM physics at the LHC
  - b-quarks present in the top-quark decay
  - largest Higgs decay branching ratio in H→bb

- ➡ General introduction on CMS heavy flavour tagging
- ➡ Algorithm performance and results for resolved and boosted jets
- ➡ Highlights on flavour tagging applications to CMS physics measurements and for CMS upgrade projects
- ➡ Calibration of b-tagging response from simulation to data - discussion of calibration techniques for b-jets, c-jets and light-flavour mistag rate measurement
- ➡ Wrapping-up and conclusions

Unless otherwise specified, all results presented in this talk are published in CMS-BTV-16-002

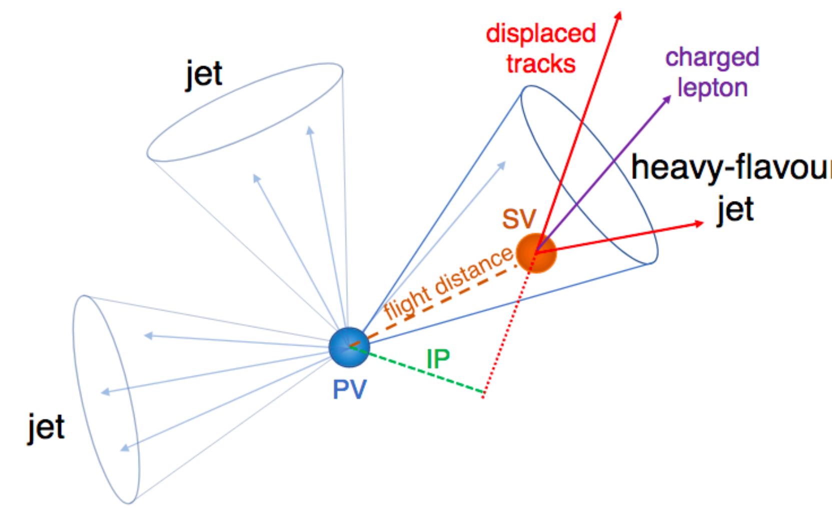
A CMS collision events with 3 jets, and a multitude of tracks, reconstructed primary and secondary vertices



Heavy flavour tagging

➡ b-jets stem from the process of hadronisation of b-quarks - B-hadron properties used to identify b-jets

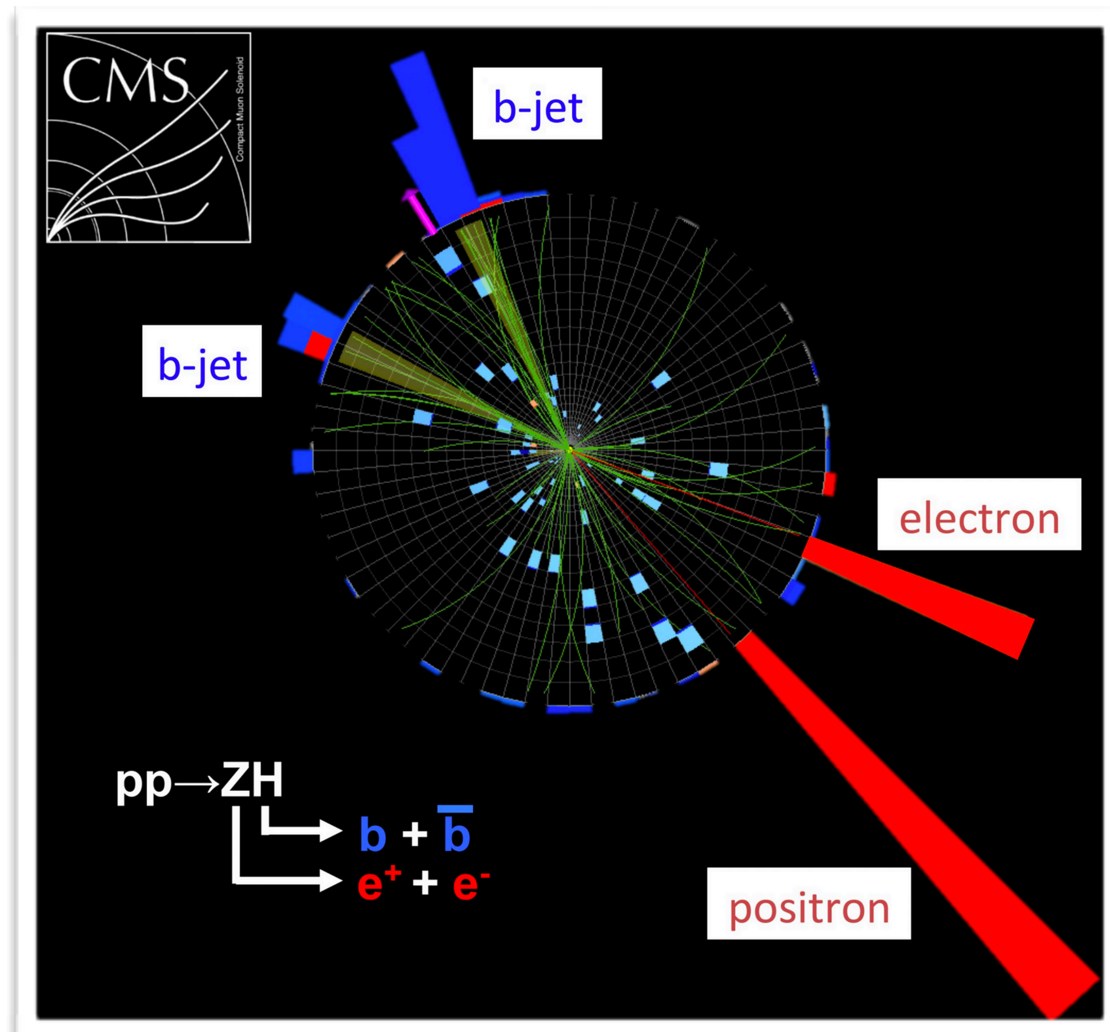
- large mass, few GeV
- long lifetime, $\beta\gamma c\tau$, order of mm
  - displaced tracks and secondary vertices
- large momentum (around 70%) carried by B-hadrons



- presence of direct and indirect semileptonic decays, i.e. $b \to \mu\nu X$ (BR $\sim$ 12%), $b \to c \to \mu\nu X$ (BR $\sim$ 10%)

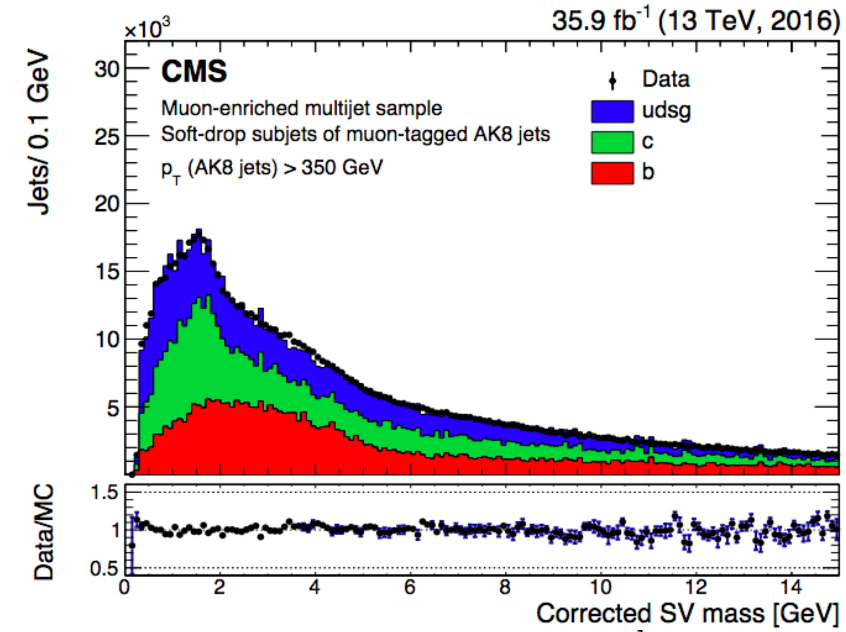
➡ Inputs from low-level taggers exploiting features of the b-jet decay topology

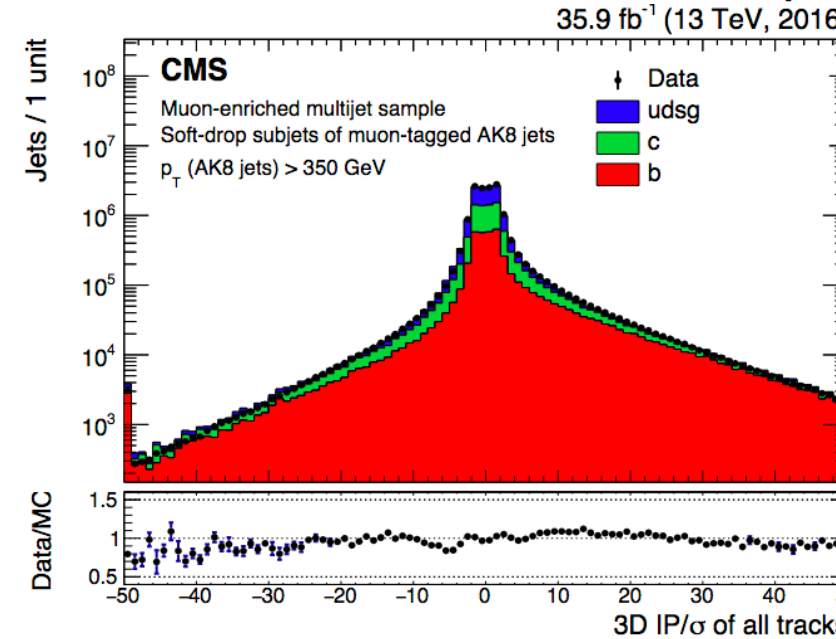
- impact-parameter of tracks associated to jets, presence of secondary vertices
- combination of high-level observables and low-level properties (from particle-flow candidates) included in DNN
- strong push for development of Deep Neural-Network based techniques for CMS b-tagging LHC Run 2 legacy results

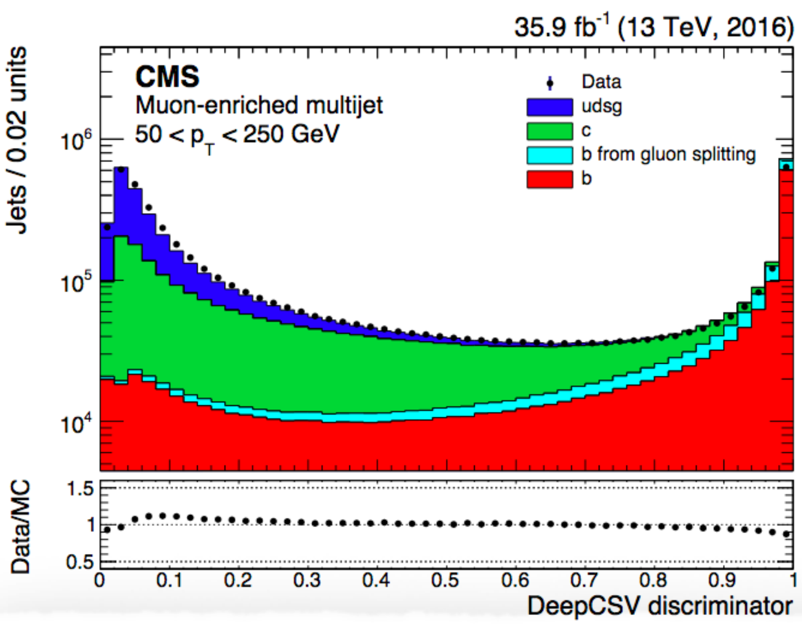


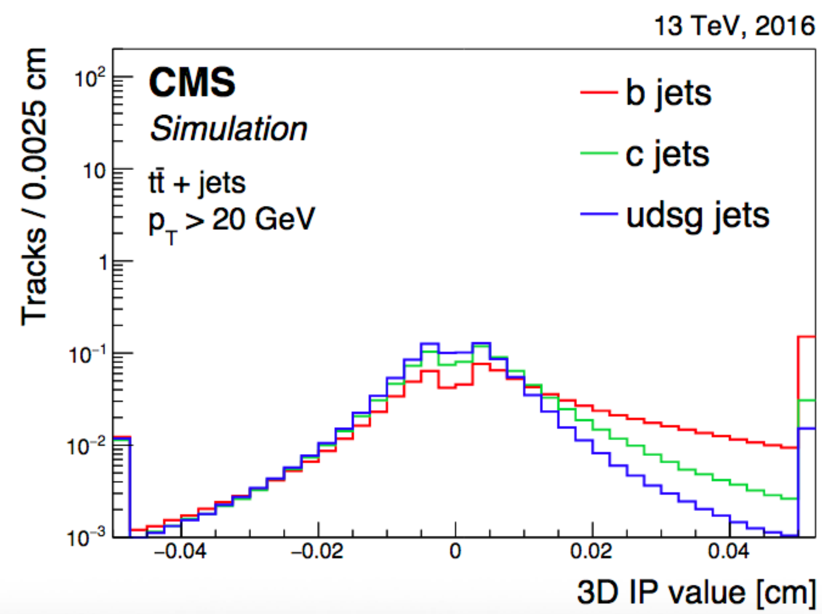
Tagger algorithms for resolved jets

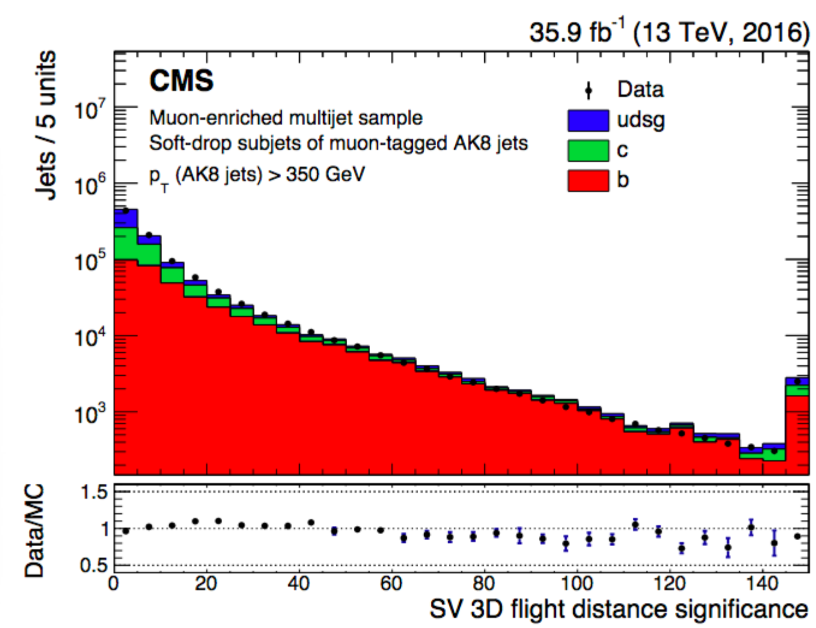
- Input variables discriminating between light-flavour-jets, c-jets and b-jets
  - track impact parameters, SV features (mass, flight-distance),…
- Deep Neural Network algorithm (DeepCSV) trained on multi jet and top quark pair events making use of high-level features and providing b vs light-flavour and b vs c-jet separation
- c-tagging is an additional challenge given the c-jet properties
  - c-discrimination made possible due to the multi class output structure of DeepCSV → CvsL and CvsB discrimination

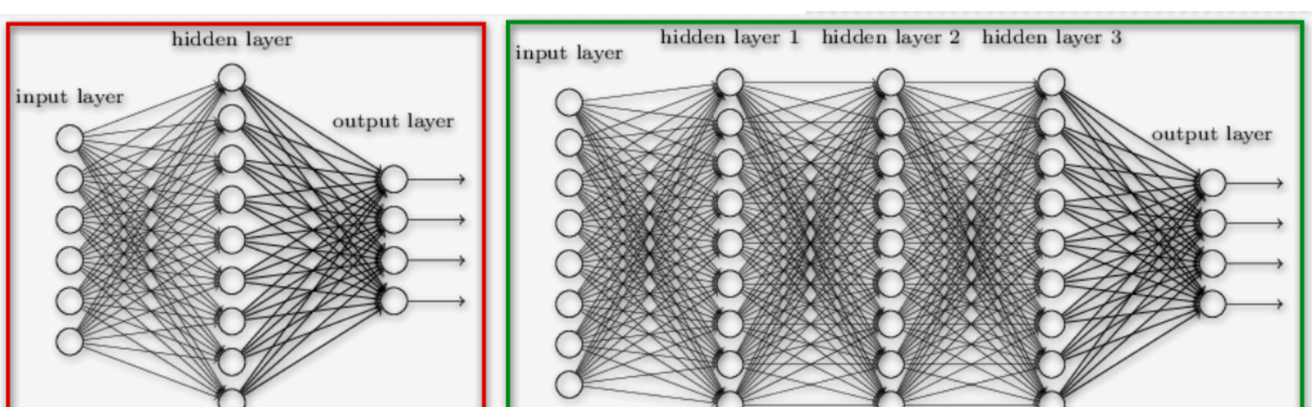












DNN architecture for DeepCSV and DeepJet

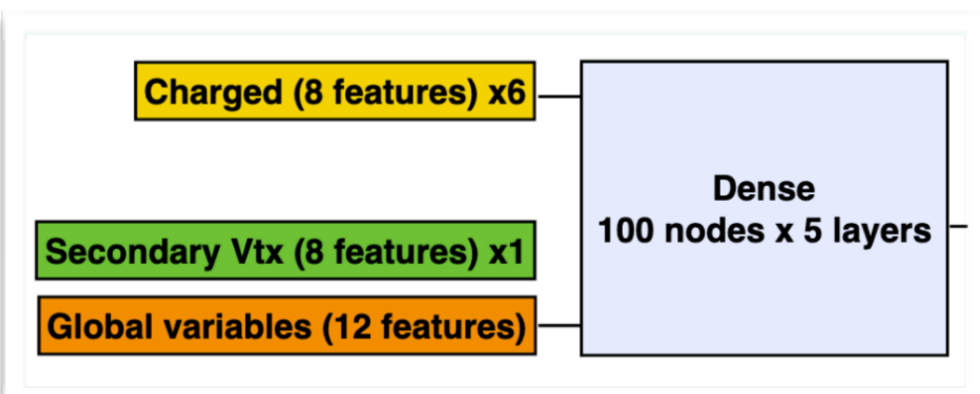
Tagger algorithms for resolved jets - improving architecture

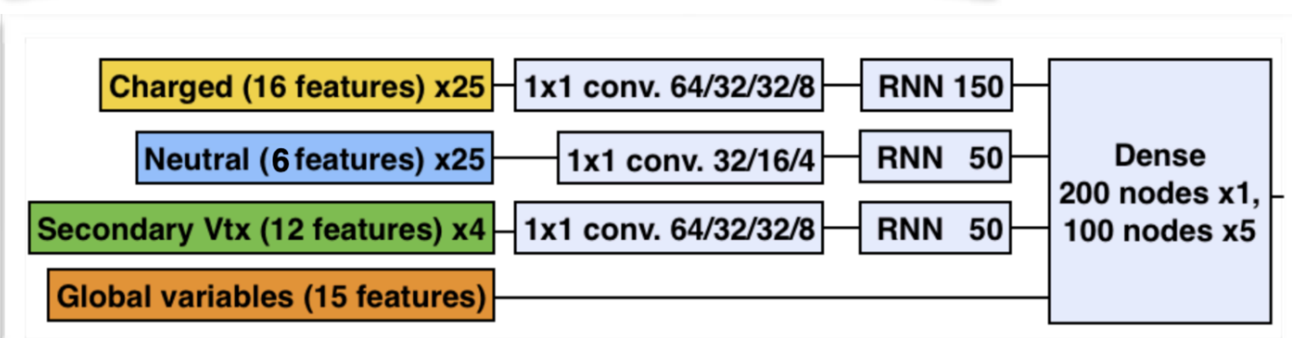
- Strong push to further improve b-tagging performance by exploiting more performant taggers for Run 2 legacy results

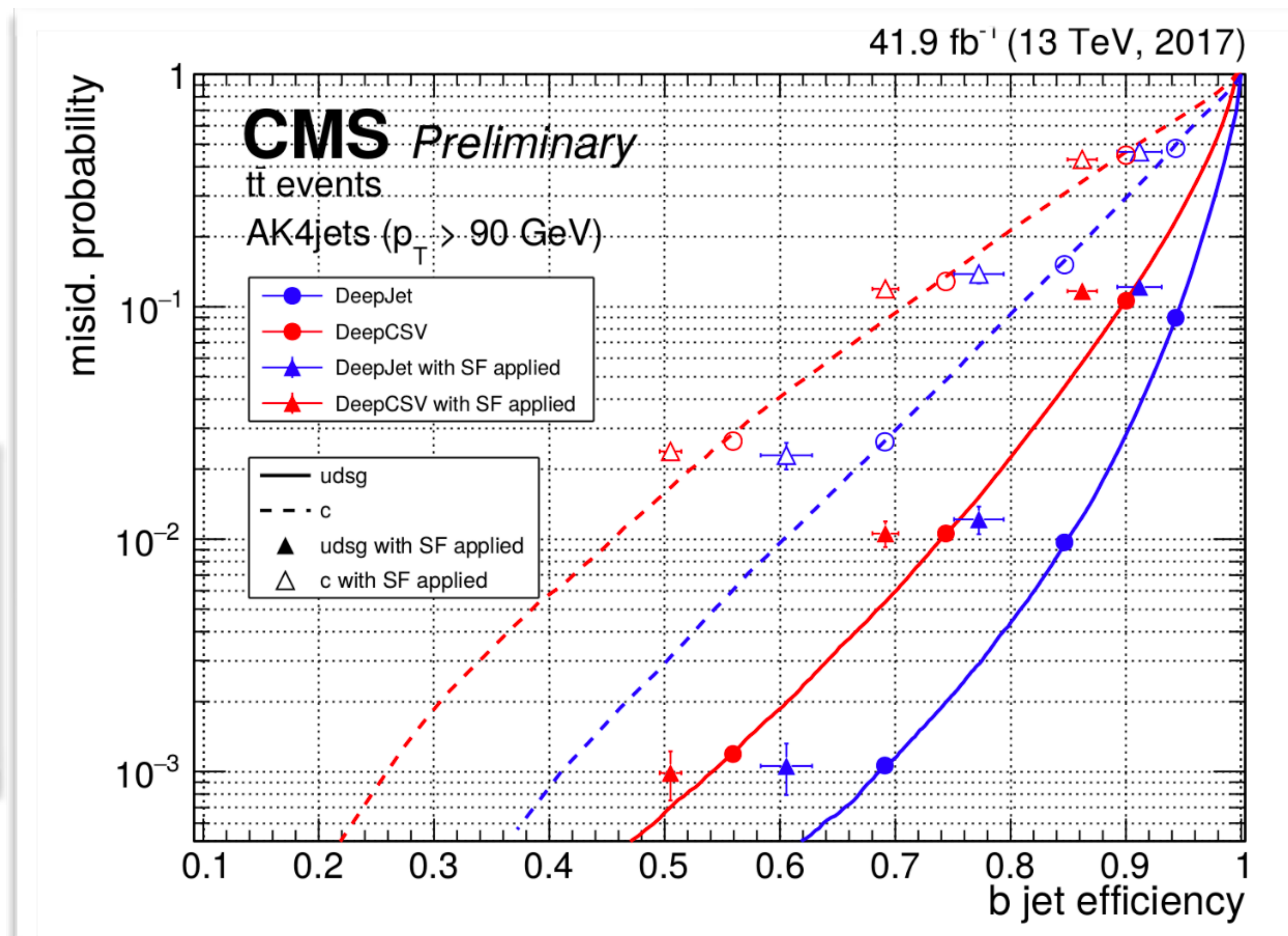
- DeepJet being developed as a successor of DeepCSV

  - inclusion of low-level input features from particle-flow candidates (properties of charged and neutral particles and secondary vertices)

  - significant performance gain over DeepCSV coming from usage of larger set of low-level inputs, less restrictive track selection and a suitable (more complex) NN architecture



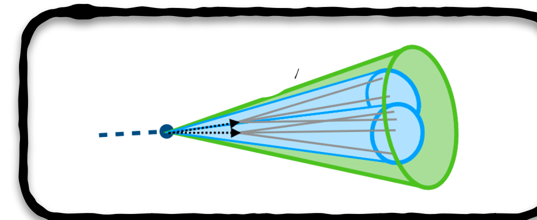




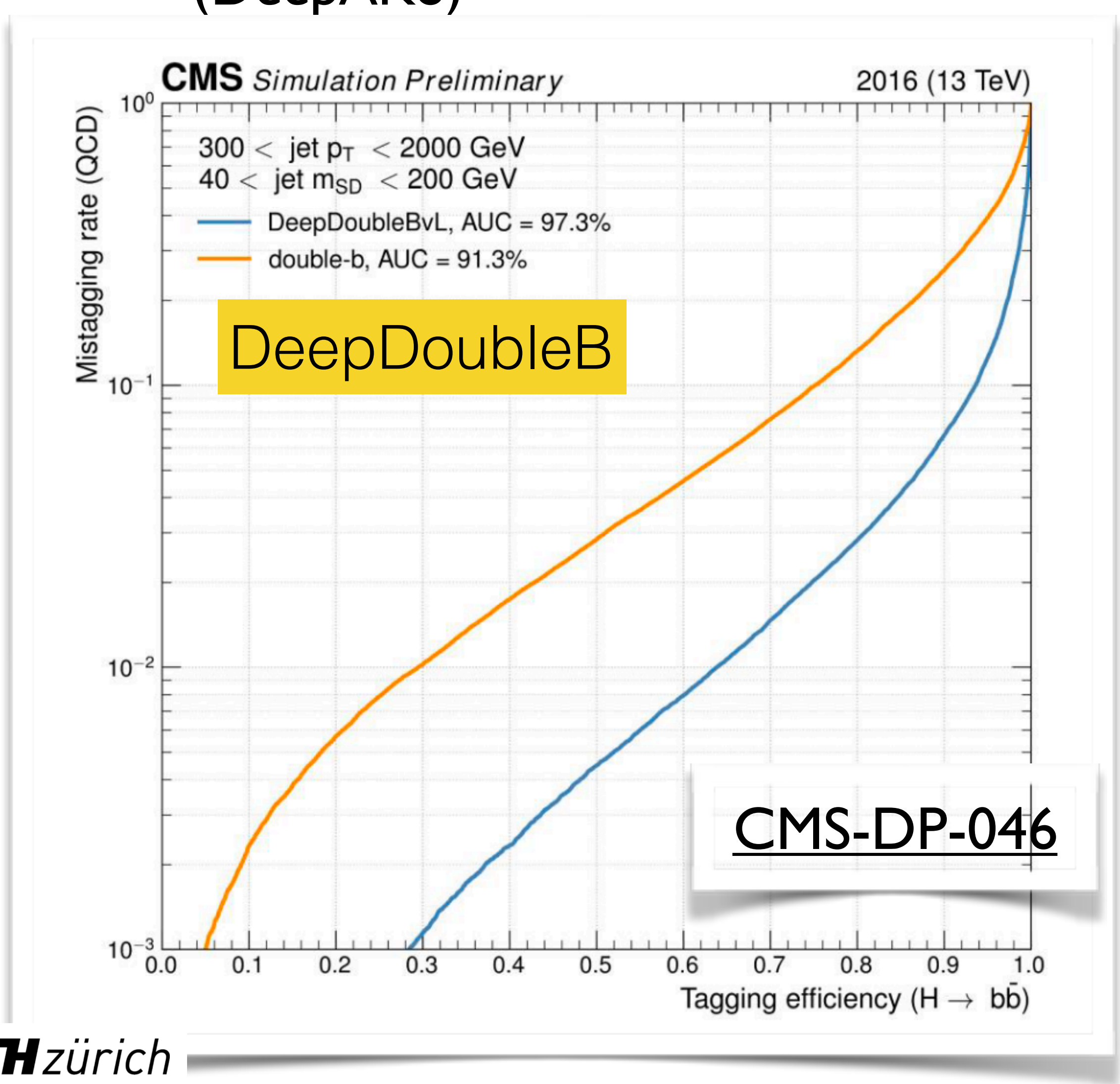
CMS DP-2018-058

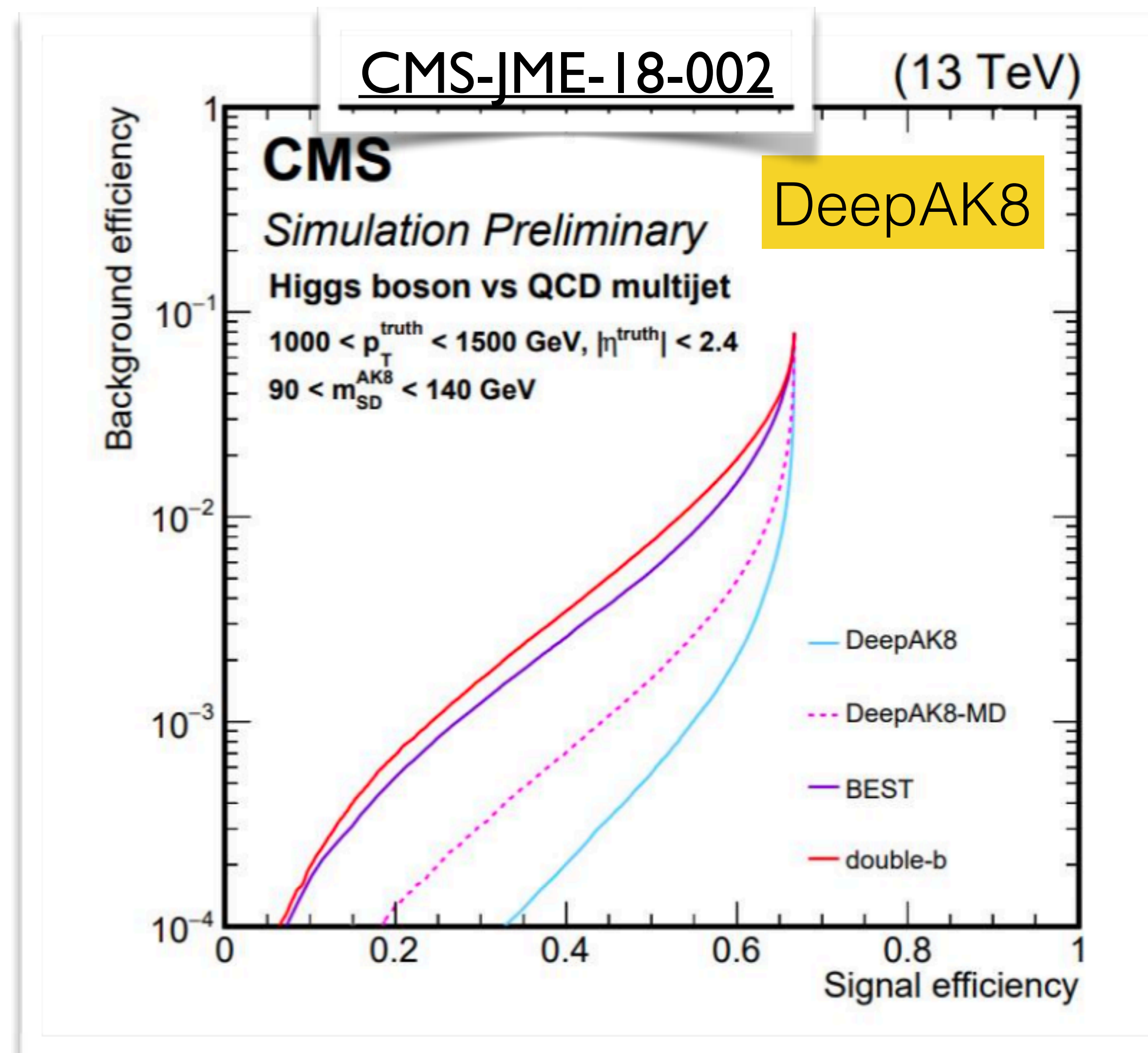
Tagger algorithms for boosted jets

- Several physics measurements explore high pt regime to mitigate overwhelming QCD background
  - collimated b-jets reconstructed as AK8 jets - significant improvement in sensitivity results due to boosted taggers - extensively used in several CMS measurements



- Two main algorithms developed by CMS for boosted regime:
  - DeepDoubleB and DeepAK8 - dedicated DNN trainings using boosted features providing distinct probabilities for bb Vs light-flavour, top, QCD
  - large gain wrt previous algorithms (DoubleB) based on BDT training - successful mass decorrelation to avoid background sculpting, either via the loss function (DeepDoubleB) or by using adversarial training (DeepAK8)



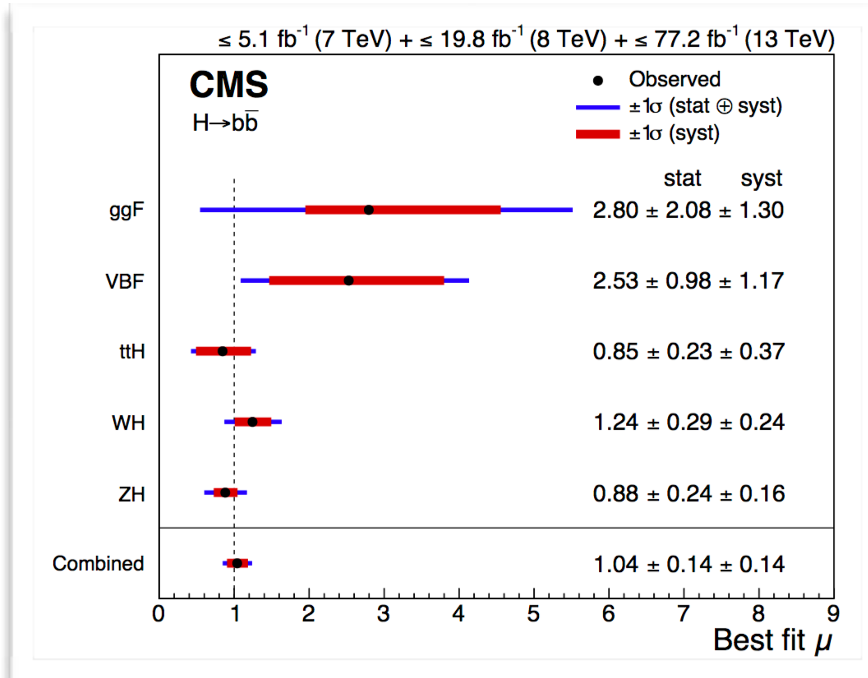


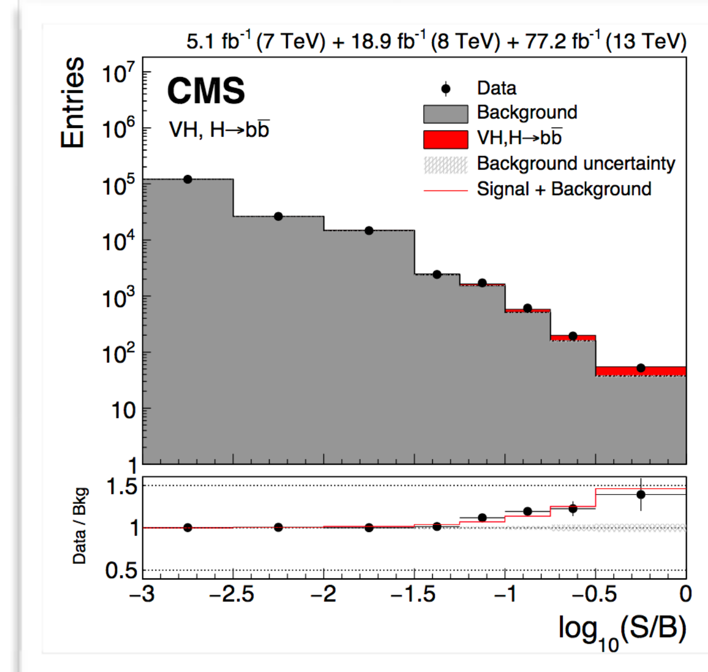
Highlights on b-tagging in physics and upgrades

★ Several CMS physics analyses and upgrade projects benefit from improved b-tagging performance

➡ Flavour tagging for resolved and boosted topologies plays a fundamental role in the observation of the Higgs boson decay to bottom quarks PRL 121, 121801 and in the inclusive high pt ggH→bb search CMS-HIG-19-003

CMS Run 2 results

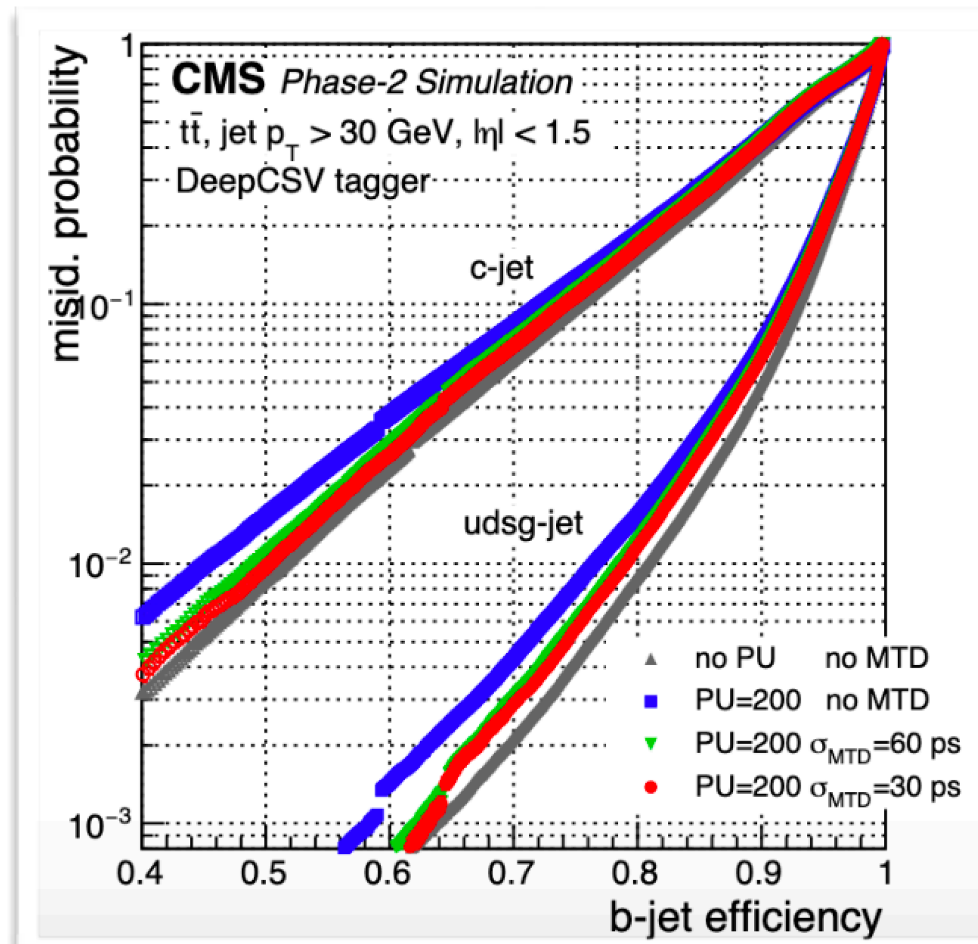




CMS Upgrade for HL-LHC

➡ Adding timing information for CMS MIP Timing Detector improves b-tagging characterisation and increases overall performance by allowing to recover performance in high pile-up regime

- timing information used for track selection
- additional results available CMS-TDR-020

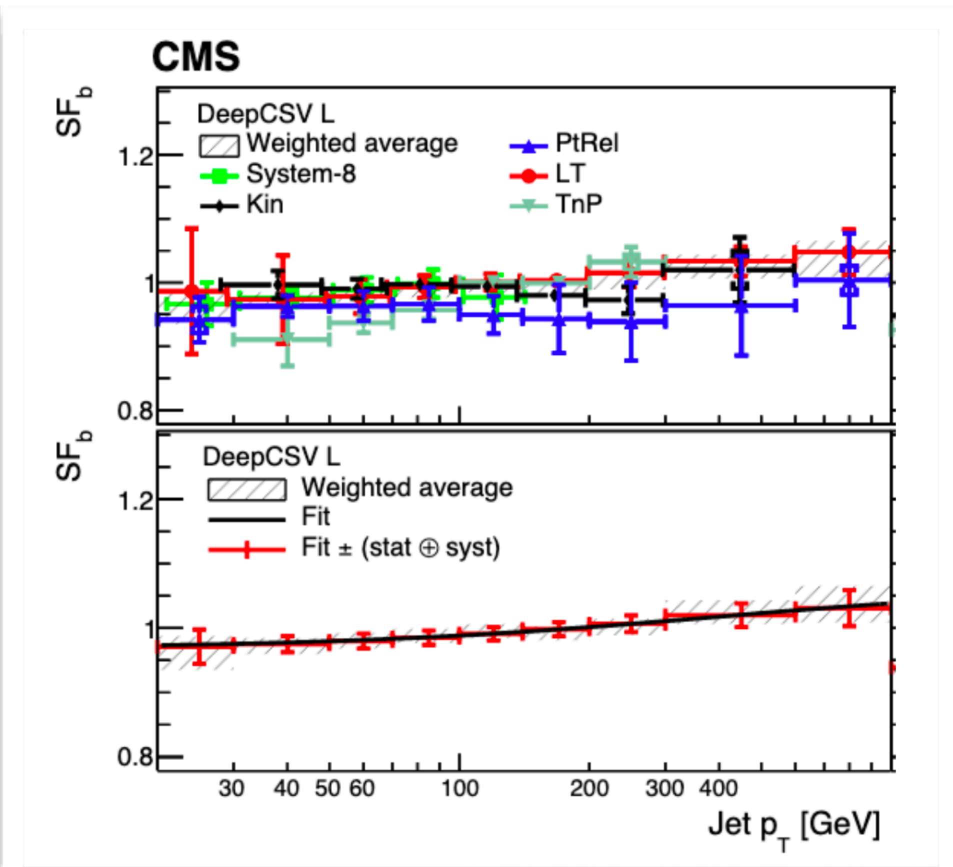


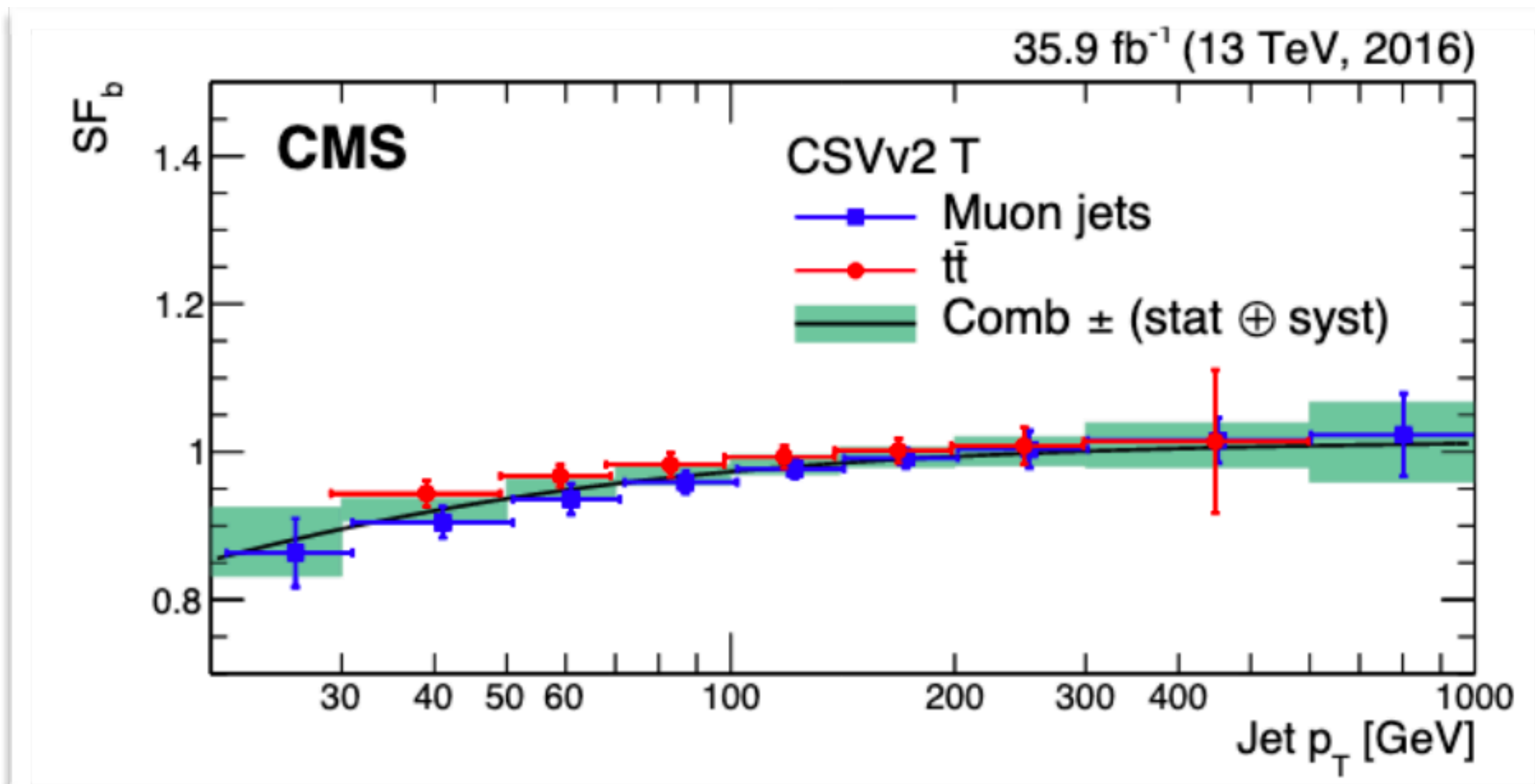
Calibration on data - overview of scale factor methods

- Calibration techniques derive correction factors to ensure the performance (efficiency) of heavy-flavour tagging in simulation is the same as in data

- Several topologies selected for extraction of scale factors enriched in b-jets, c-jets, light-flavour jets
  - b-jet calibration: ttbar, μ-enriched, c-jet calibration: W+c, light-flavour jet calibration: QCD multijet
  - majority of calibration methods calculate SF for a fixed cut on the discriminator values, full b-tagging discriminant shape correction also available for DeepCSV and DeepJet algorithms

- Full set of calibration methods to ensure redundancy and usage of orthogonal phase-spaces (μ-enriched, ttbar, ..) → exploit combination of results as input to physics measurements





Examples of combination for b-jet calibrations to probe consistency of SF results

AK4 - Calibration for b-jets

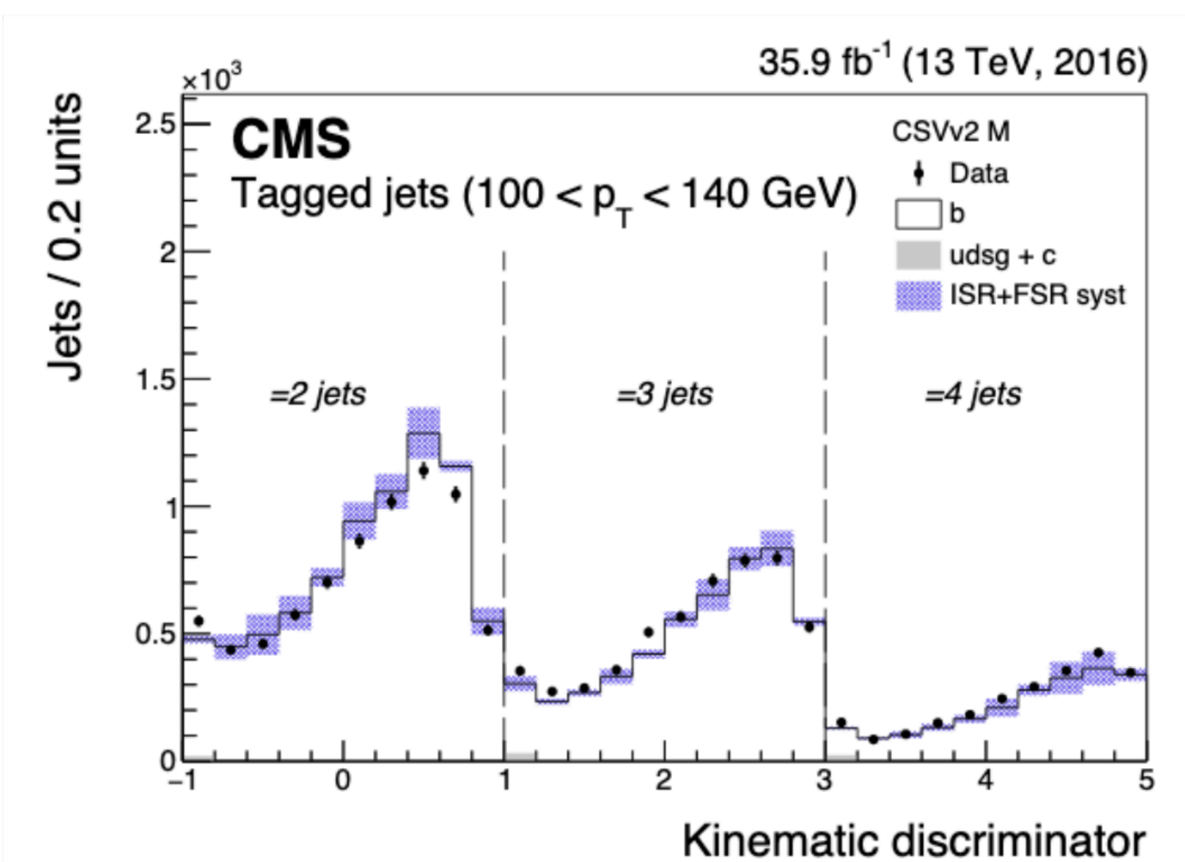
➡ Usage of discriminating observables based on event topology to mitigate presence of light-flavour jet polluting the phase space for the efficiency measurements on b-jets

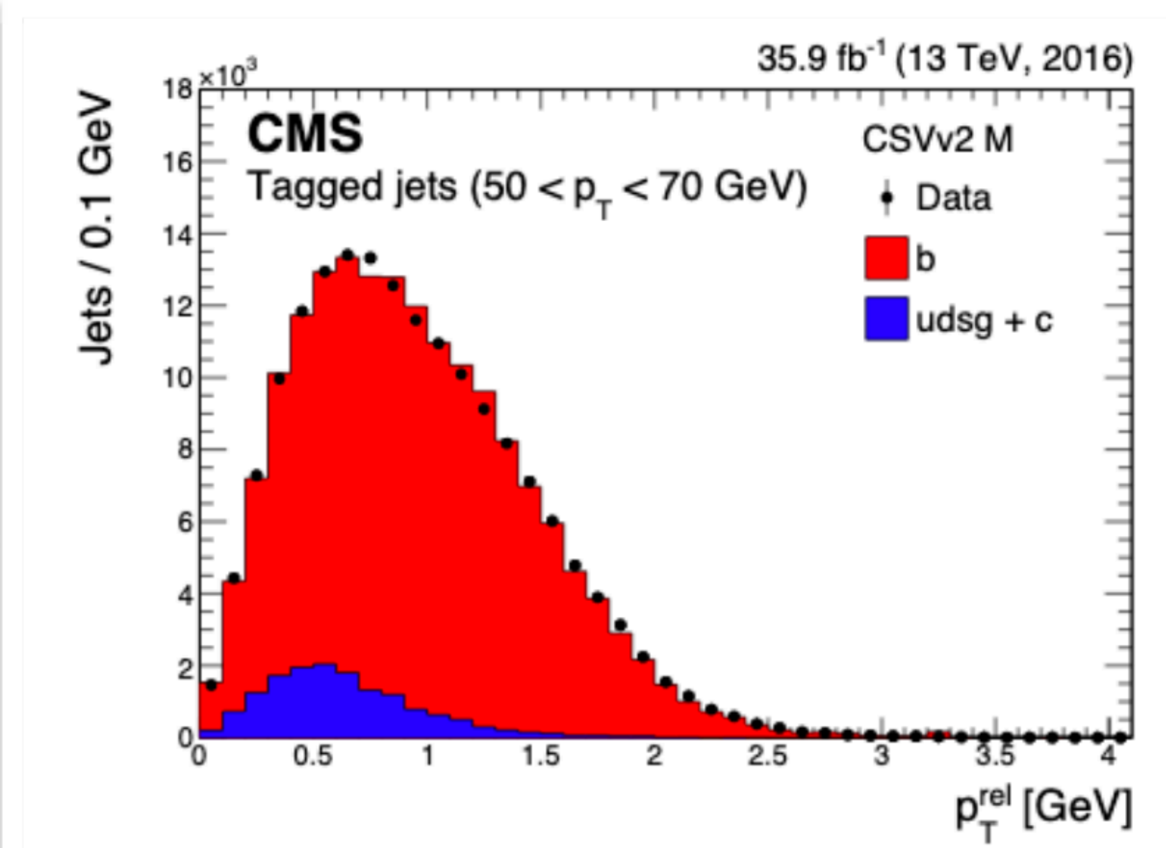
➡ tt-based calibrations:

- Tag&Probe (semileptonic tt): probe-jet to determine efficiency of b-jets in data and simulation
- Kinematic fit (dileptonic tt): usage of BDT with kinematic variables to discriminate jets in the tt system from ISR/FSR jets

➡ μ-enriched calibrations:

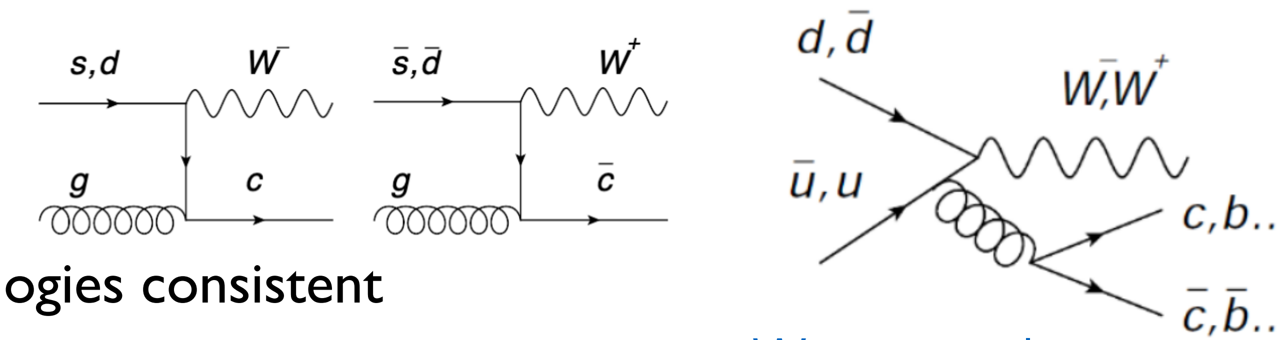
- ptRel, System8, LT methods using semileptonic b-decays - momentum of the muon wrt jet axis larger for b-jets and powerful for b-jet discrimination - usage of ptRel templates from simulation

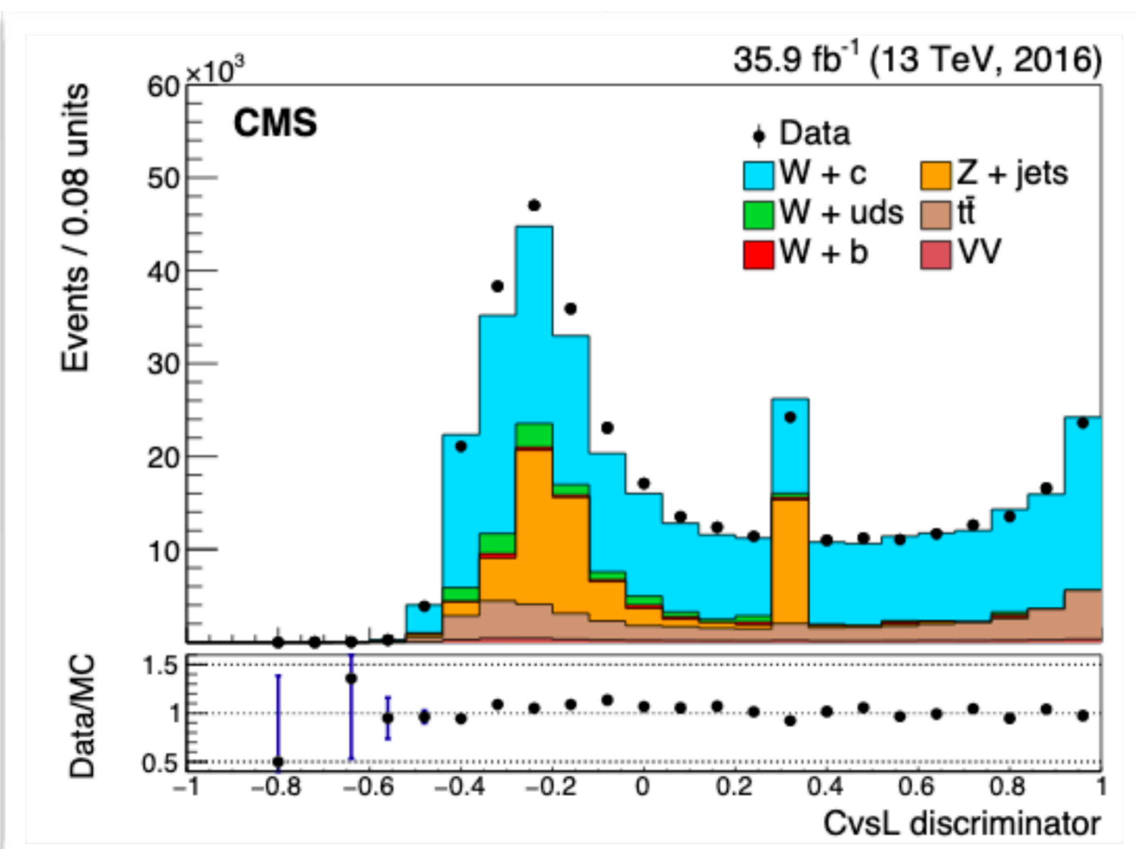




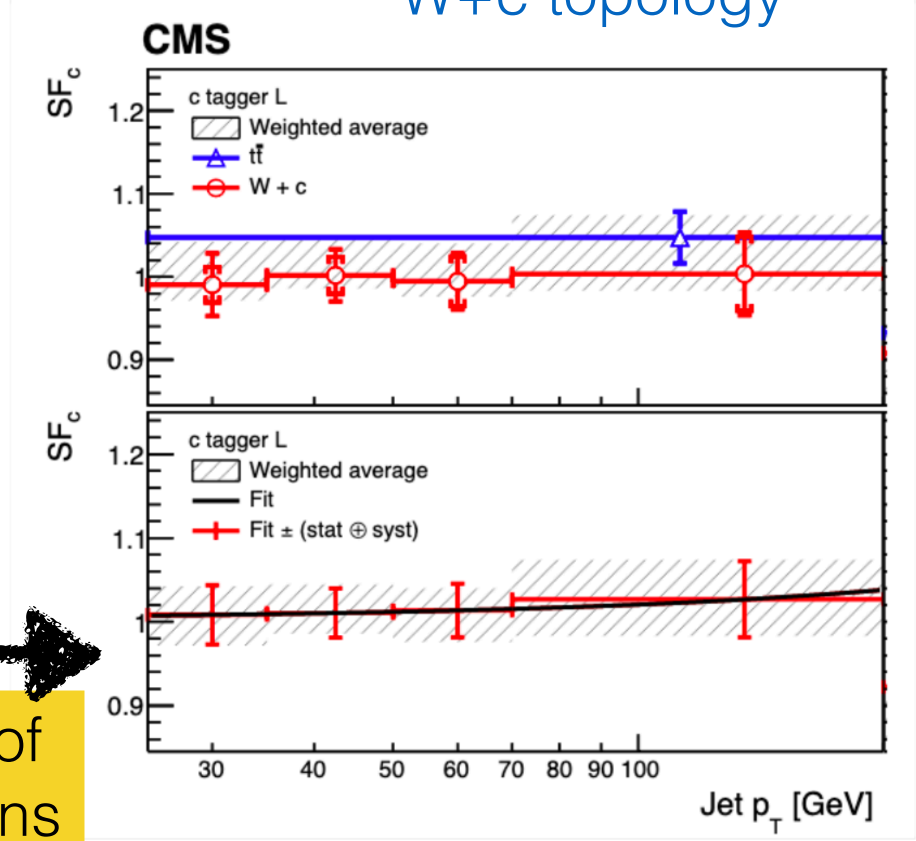
AK4 - Calibration for c-jets

- W+c topology - W+c production mainly due to processes in which W and c are opposite sign, dominant background (W+qq) balanced OS/SS rate
  - OS/SS subtraction provides large enrichment in W+c with a signal purity of around 60%(80%) for W→eν (μν)
  - efficiency measurements in data and in simulation as a function of jet pt used to extract SF's for different working points
- Additional method making use of c-jets from ttbar semileptonic events using a kinematic reconstruction to target c-jets stemming from the W decay
  - limited statistics for SF extraction
- Results from measurements using different topologies consistent





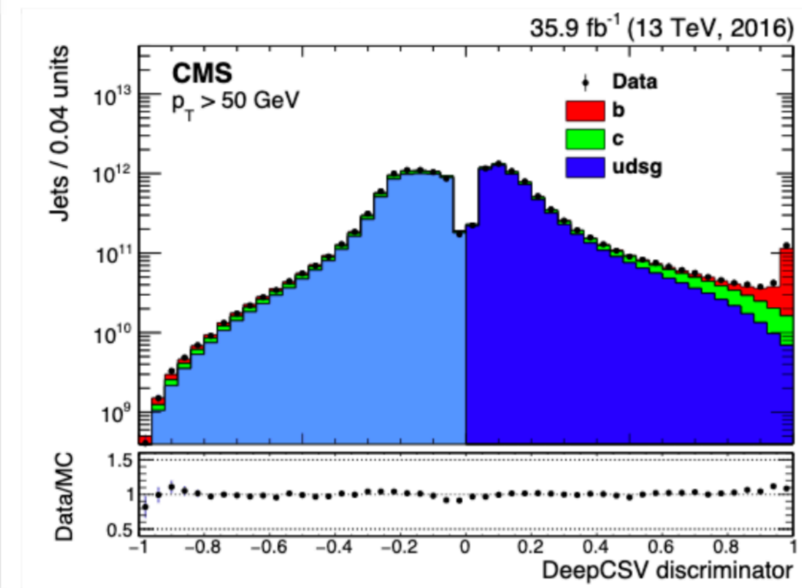
W+c topology



Combination of c-jet calibrations

AK4 - Calibration for light-flavour jets

- Inclusive multi-jet events - default tagger but use only negative impact parameter values and SV's with negative flight distance
- Distribution should be approximately symmetric for light-flavour jets - non-zero values of IP/SV flight distance coming from resolution effects



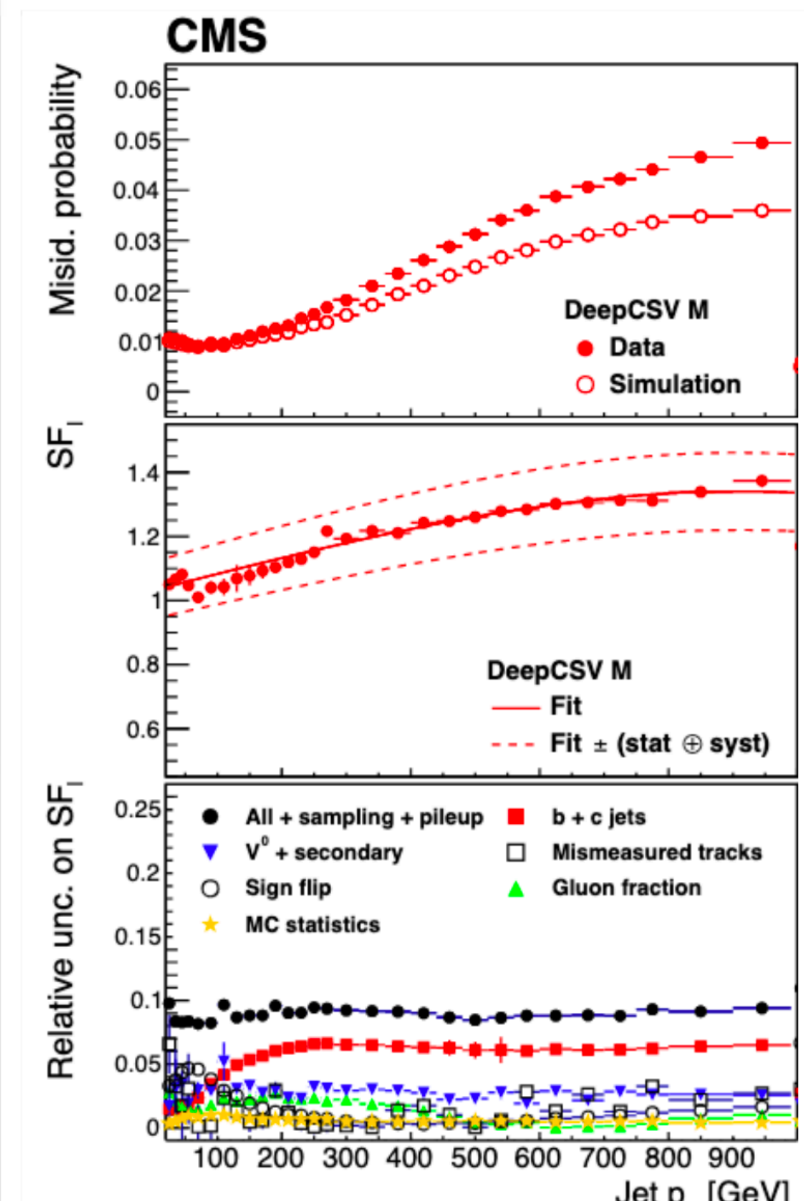
$$\epsilon_\ell = \epsilon^- R_{LF}$$

Fraction of -ve tagged jets passing WP in inclusive multi-jet sample

Ratio of mis-id probability for light-jets w.r.t. the negative tagging efficiency of all jets in MC

$$R_{LF} = \frac{\epsilon_\ell^{MC}}{\epsilon^{-,\,MC}}$$

Definition of light-flavour jet mistag-rate



Wrapping-up & conclusions

★ Heavy flavour tagging is a crucial ingredient for measurements and searches at the LHC

➡ Presented overview of algorithm developments in CMS for b/c-jet identification for resolved and boosted topologies

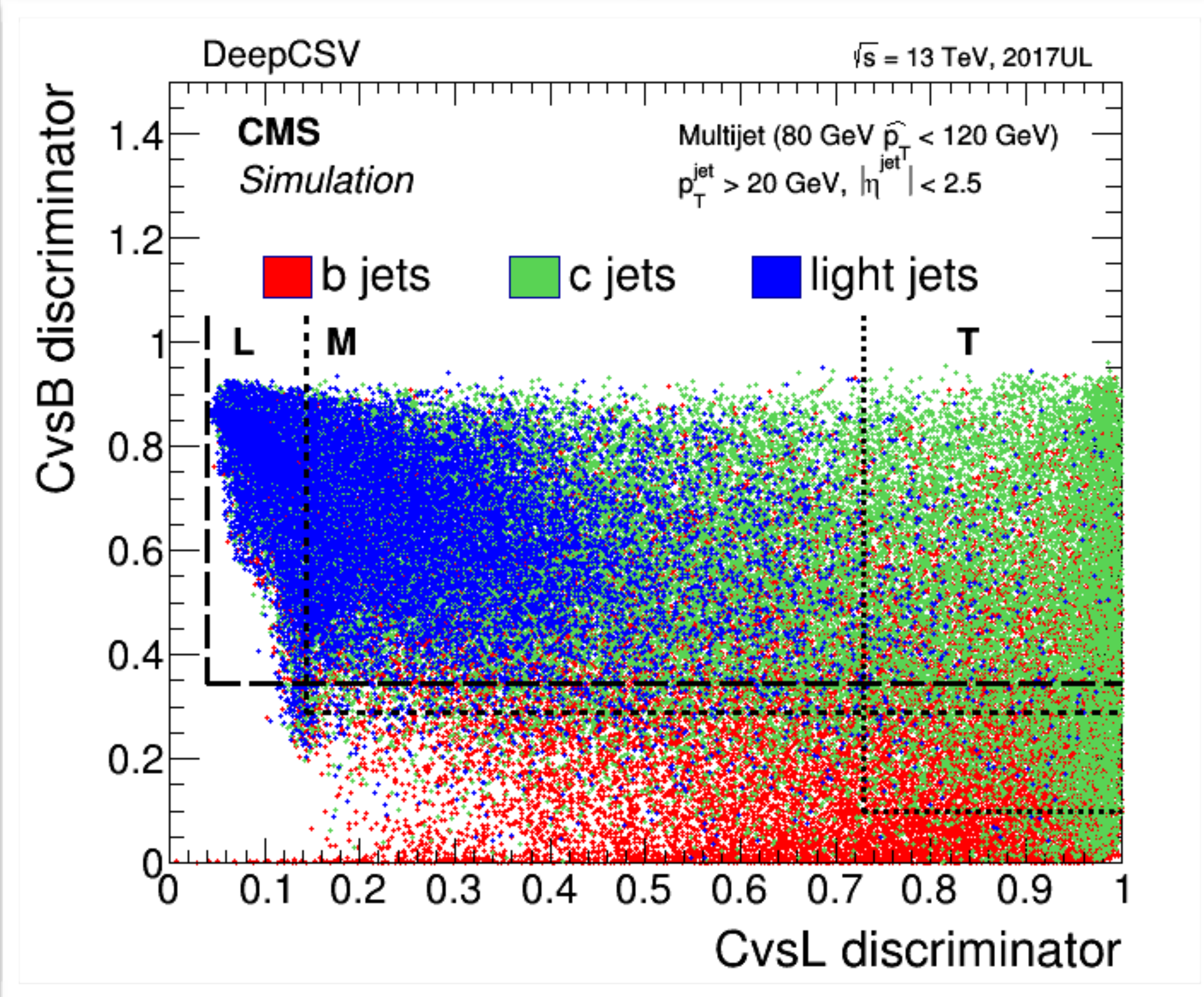
- several new algorithms developed for CMS Run 2 legacy measurements based on better understanding of deep learning techniques and jet topology features
- excellent level of background discrimination has definitely paid off - very useful for searches and measurements with b/c-jets in the final state and for CMS flagship analyses such as VH→bb observation with 2016+2017 data, first VH→cc search with 2016 data, inclusive boosted search in bb final state
- significant work on the upgrade side has also allowed to achieve excellent performance by making use of novel detector features embedded in the b-tagging chain (e.g. timing information in MTD)

➡ Calibration of b/c-tagging algorithms and light-flavour mistag rate measurements are also an essential features of heavy flavour tagging

- calibration analyses for b-, c-jet and light-flavour measurements are complex and have largely benefitted from combination of results expected from several methods targeting different phase-spaces

Additional slides

2D c-tagger working points & algorithm definitions



	BvsC/L	CvsB	CvsL
DeepCSV:	P(b)+P(bb)	$\frac{P(c)}{P(c)+P(b)+P(bb)}$	$\frac{P(c)}{P(c)+P(udsg)}$
DeepJet:	$P(b)+P(bb)+P(b_{lep})$	$\frac{P(c)}{P(c)+P(b)+P(bb)+P(b_{lep})}$	$\frac{P(c)}{P(c)+P(uds)+P(g)}$