Jet flavour identification at CMS

Petra Van Mulders
Vrije Universiteit Brussel

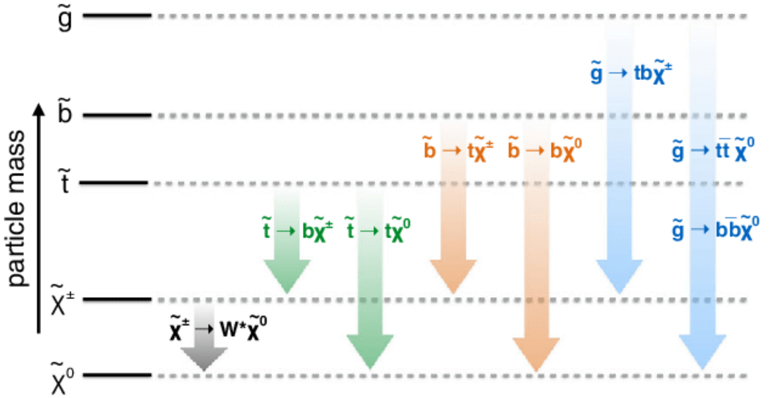
for the CMS Collaboration

Game of Flavours
CMS Heavy flavour tagging workshop 2019
30th of April - 3rd of May

CMS-PAS-HIG-13-011

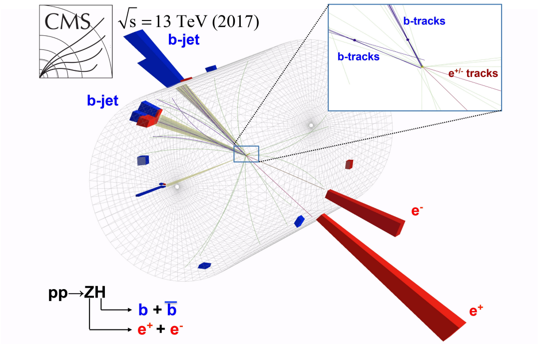
Jet flavour identification – motivation

- **Jet flavour identification is crucial for Standard Model studies and searches, e.g.:**
  - Higgs sector: BR(H → bb) ~ 60%
  - Top quark sector: BR(t → bW) ~ 100%
  - Sensitivity for H → cc
  - New particles decaying to t, H, b or c quarks
  - …



Potential decay modes in simplified natural SUSY model [Phys. Rev. D 91 (2015) 052018]

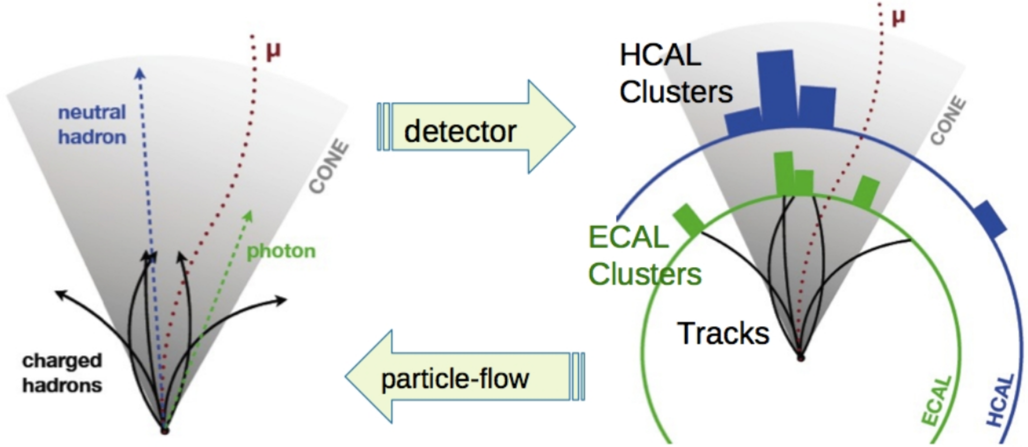
- **Highlights of analyses for which jet flavour identification/tagging is vital:**
  - VH (H → bb) → Chris Palmer and Valerio Dao
  - ttH → Joshuha Thomas-Wilsker
  - Gluon splitting → Ben Nachman



Candidate ZH → eebb event [Phys. Rev. Lett. 121 (2018) 121801]

Jet reconstruction and jet flavour in simulation

- **Quarks will hadronize/fragment into colorless hadrons forming a jet of particles**
- **Particles in CMS are reconstructed with the particle flow algorithm**



Particle flow reconstruction
[JINST 12 (2017) P10003]

- **The reconstructed particles are clustered into jets, which have a momentum close to that of the parent quark**
- **The jet flavour in simulation is obtained by clustering the generated heavy hadrons in the jet, rescaling their momentum to a negligible value (ghost hadrons)**
  - The presence of a clustered b or c ghost hadron in the jet determines the jet flavour

Jet flavour identification – the basics

- **Compared to light quarks/hadrons, heavy-flavour (b and c) quarks/hadrons have:**
  - Larger mass and harder fragmentation (fraction of initial quark momentum carried by the corresponding hadron)
  - Longer lifetime → displaced decays for b/c hadron
  - For b (c) hadrons: 20 (10) % of the decays is to leptons
- **Algorithms for heavy-flavour jet identification exploit these properties**
  - Information from the reconstructed particles is combined using multivariate analysis / deep learning tools
  - Accurate + efficient reconstruction of charged particle trajectories (tracks) in the detector is essential → Mia Tosi

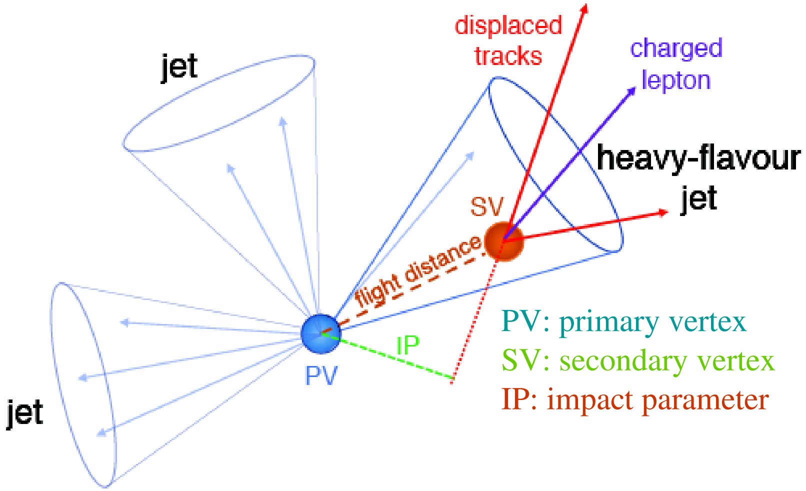


Illustration of some of the heavy-flavour jet properties [JINST 13 (2018) P05011]

- **Accurate modelling/simulation of heavy-flavour jet production at the LHC is vital to build realistic jet flavour identification algorithms and to design calibration strategies**
  - Event generation in CMS and ATLAS → Qiang Li, Chris Pollard
  - Heavy-flavour production/jet modelling in Herwig/Sherpa → Simon Plätzer, Gurpreet Singh Chahal

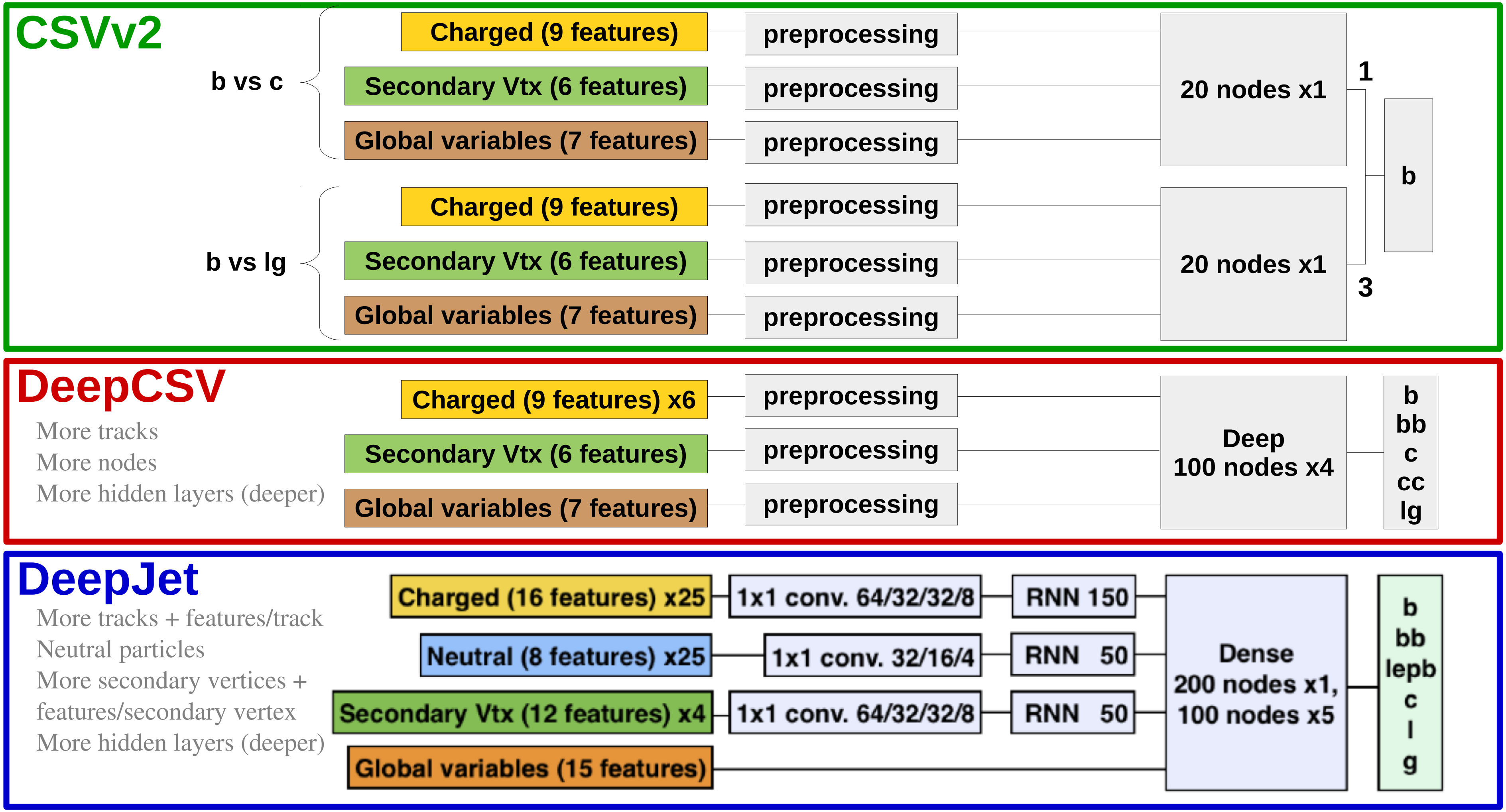
Outline

- **Heavy-flavour identification in standard topologies**
  - Evolution of the algorithms and their performance
  - Algorithm calibration (aka scale factors) – standard methods
  - Discriminator distribution shape calibration using adversarial neural network

- **Heavy-flavour identification in boosted topologies**
  - Tagging jets with two heavy-flavour quarks from a boosted H or Z boson decay
  - Recent developments for double-b/c tagging algorithms
  - Boosted top quark identification

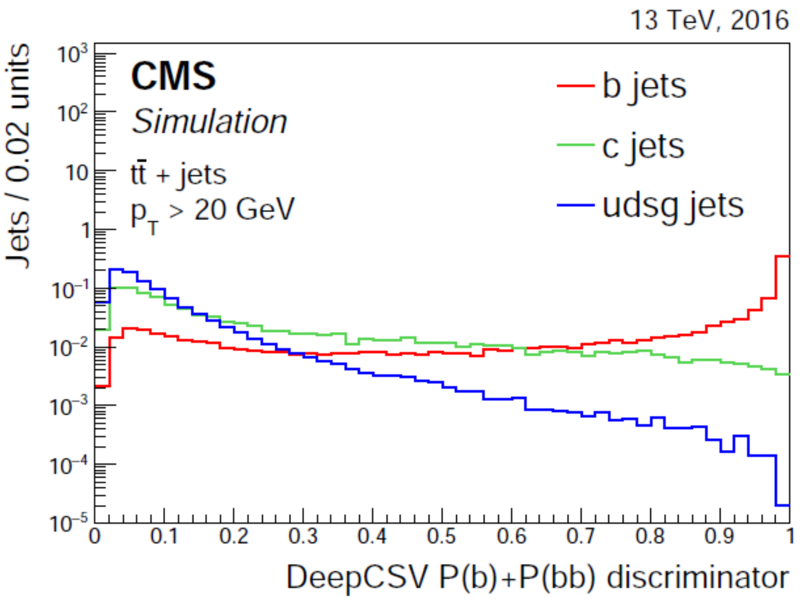
- **Outlook and conclusion**

Algorithm evolution: more info and deeper



Quantifying the performance of algorithms

- **The performance is evaluated by evaluating the efficiency for b jets (ε_b) and the misidentification probability ($\varepsilon_{non\text{-}b}$) for different thresholds on the discriminator**



[JINST 13 (2018) P05011]

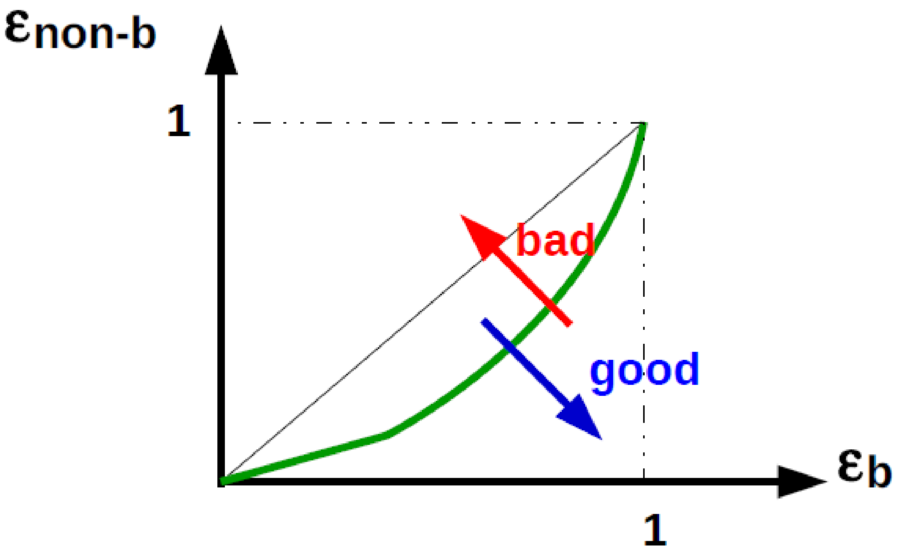
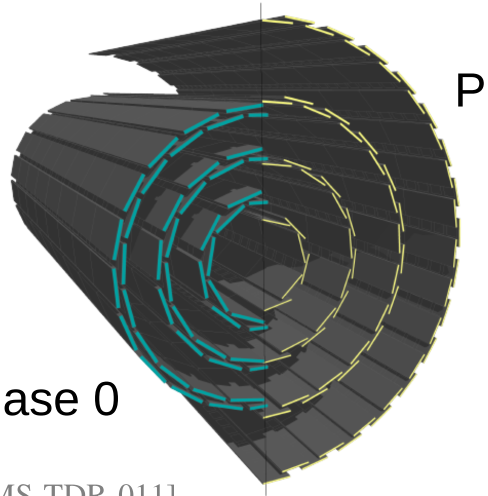


Illustration performance (ROC) curve

Performance of the flagship algorithms has greatly improved over the last years

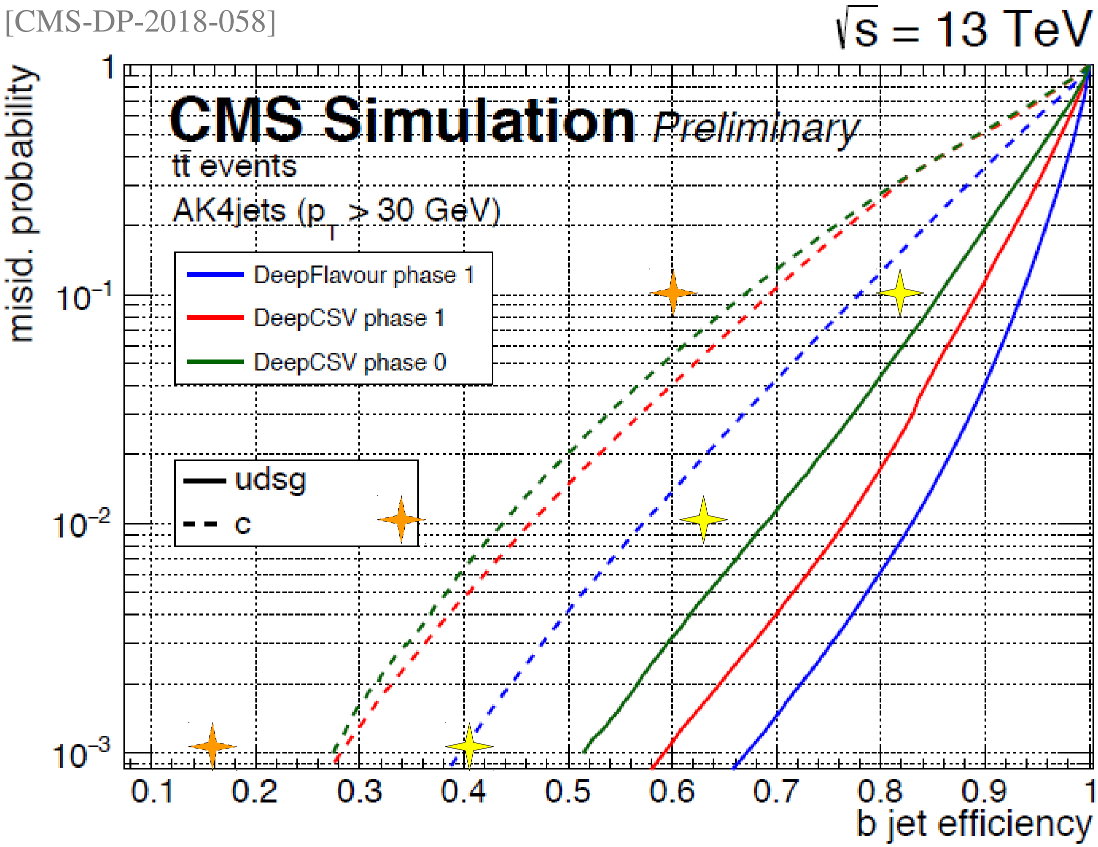
Between 2016 and now:
an improvement of 10-15% in *absolute* b jet identification efficiency!

- Algorithms were improved
- Pixel tracker has now 4 layers and first layer closer to the beam pipe



[CMS-TDR-011]

[CMS-DP-2018-058]



CSVv2 phase 0 – udsg misid. probability

CSVv2 phase 0 – c misid. probability

Approximation based on [JINST 13 (2018) P05011]

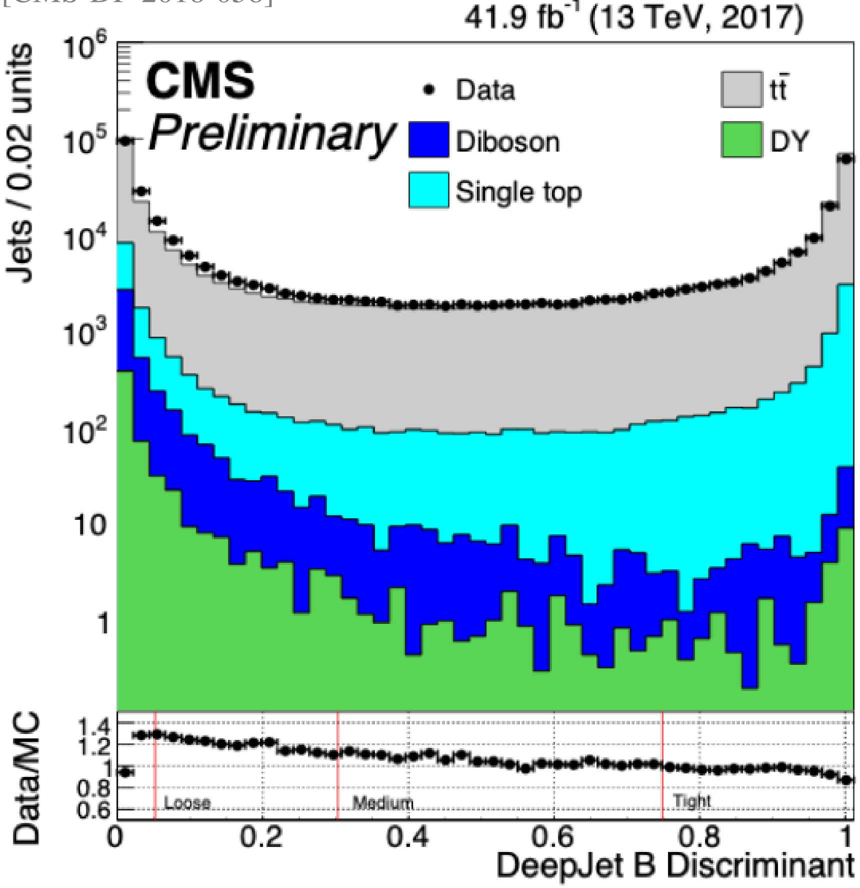
Reweighting the simulation to the data

- **Three 'working points' are defined:**
  - Loose (L), Medium (M), Tight (T)
  - Choice depending on analysis needs
  - Note: could also be used to veto b jets
- **Calibration (scale factors) for each working point is needed since discriminator shape in data and simulation may differ**
  - ε_b and $\varepsilon_{non\text{-}b}$ will be different in simulation and data
  - Scale factors (SF) depend on the jet flavour *f* and kinematics (p_T/η)

$$SF_f = \varepsilon_f^{\mathrm{data}}(p_{\mathrm{T}},\eta)/\varepsilon_f^{\mathrm{MC}}(p_{\mathrm{T}},\eta)$$

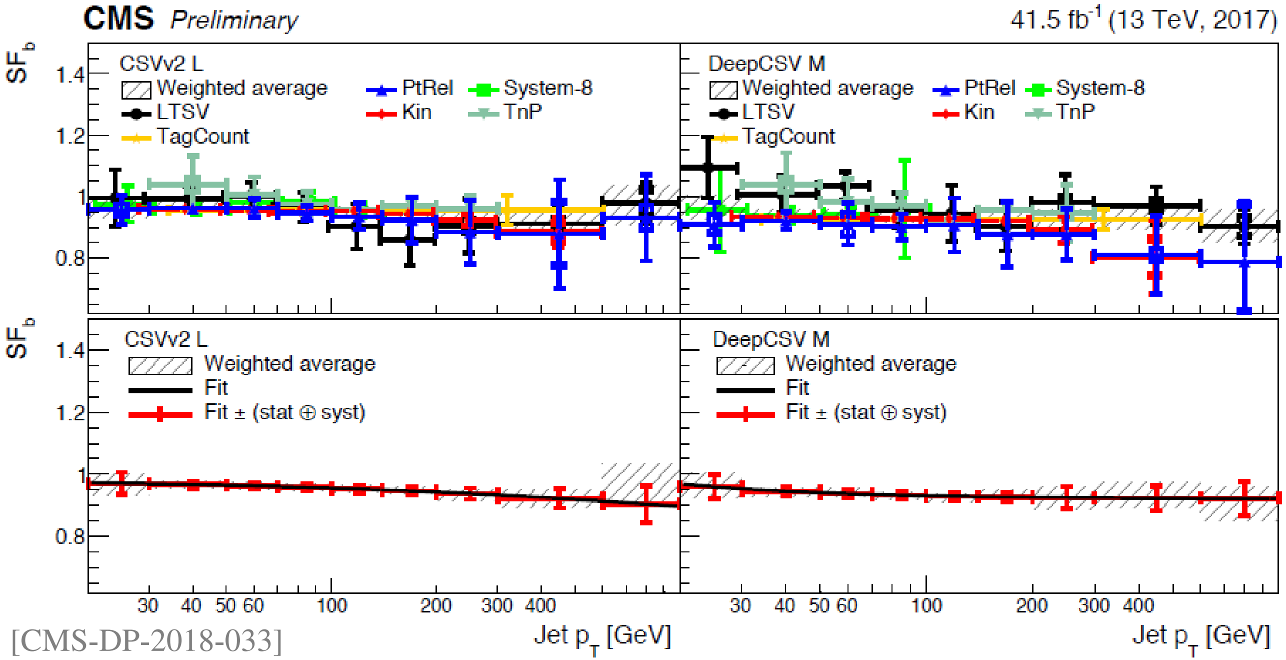
- **The scale factors are then used to reweight the simulation to data depending on the number of jets of each flavour**

[CMS-DP-2018-058]



Scale factors measurements

- **The efficiency in data is obtained by selecting a sample of jets of a certain flavour**
  - For b jets: $t\bar{t} \rightarrow$ dilepton, $t\bar{t} \rightarrow$ lepton+jets and muon-enriched QCD multijet events $\rightarrow$ 6 measurements
  - For c jets: W+c and $t\bar{t} \rightarrow$ lepton+jets events $\rightarrow$ 2 measurements
  - For light jets: QCD multijet events $\rightarrow$ 1 measurement
- **Techniques are described in JINST 13 (2018) P05011**



- **Measurements are combined**
- **Statistical uncertainty typically a factor 10 smaller than the systematic uncertainty (except for SF_c)**
- **Dominant systematic uncertainties for most measurements are related to the flavour purity of the jet sample**

[CMS-DP-2018-033]

Performance of the flagship taggers in data

- **The scale factor for light jets is typically larger than 1**

  → larger misidentification probability in data compared to simulation

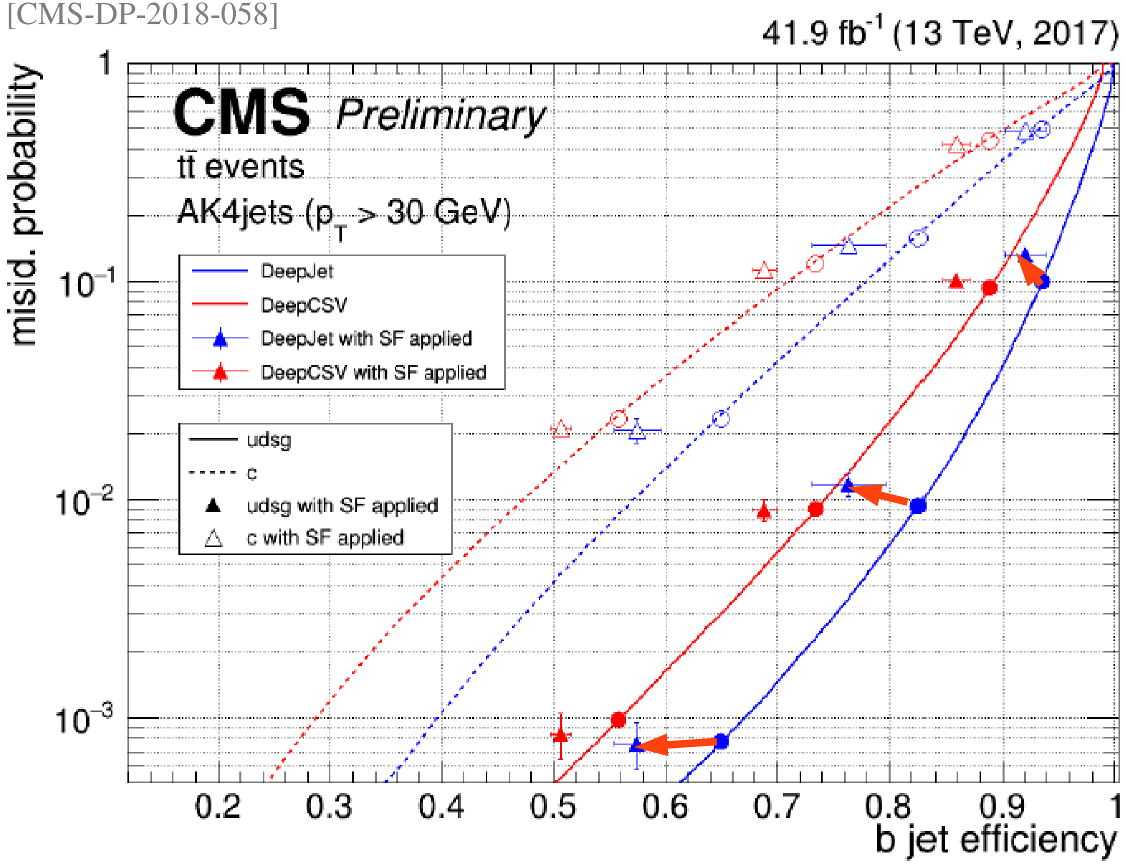
- **The scale factor for b jets is typically smaller than 1**

  → smaller identification efficiency in data compared to simulation

- **Similar performance loss for the various algorithms**

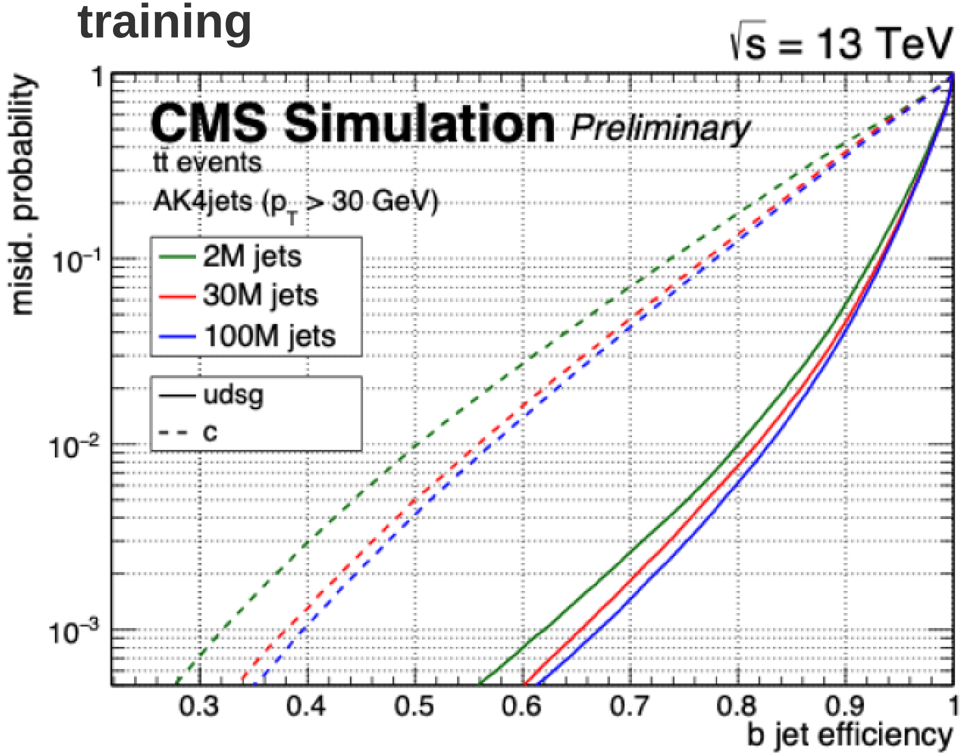
  → important to keep in mind when optimizing/developing algorithms for heavy-flavour jet identification!

[CMS-DP-2018-058]

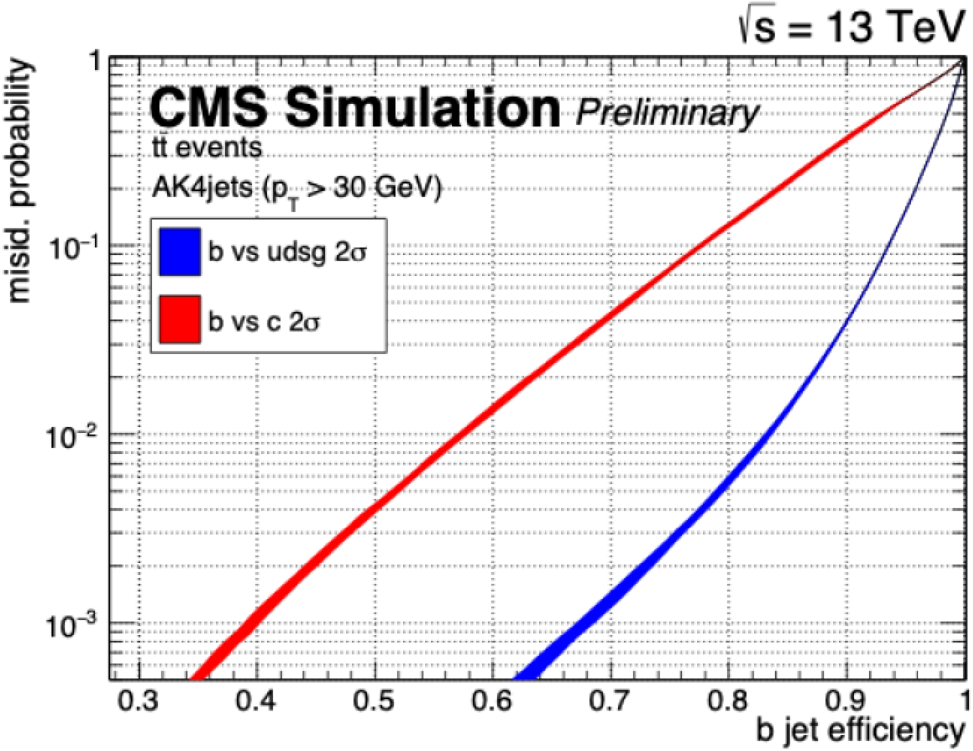


DeepJet training

- **The performance depends on the size of the training sample**
- **Typically ~100M jets are used for training**



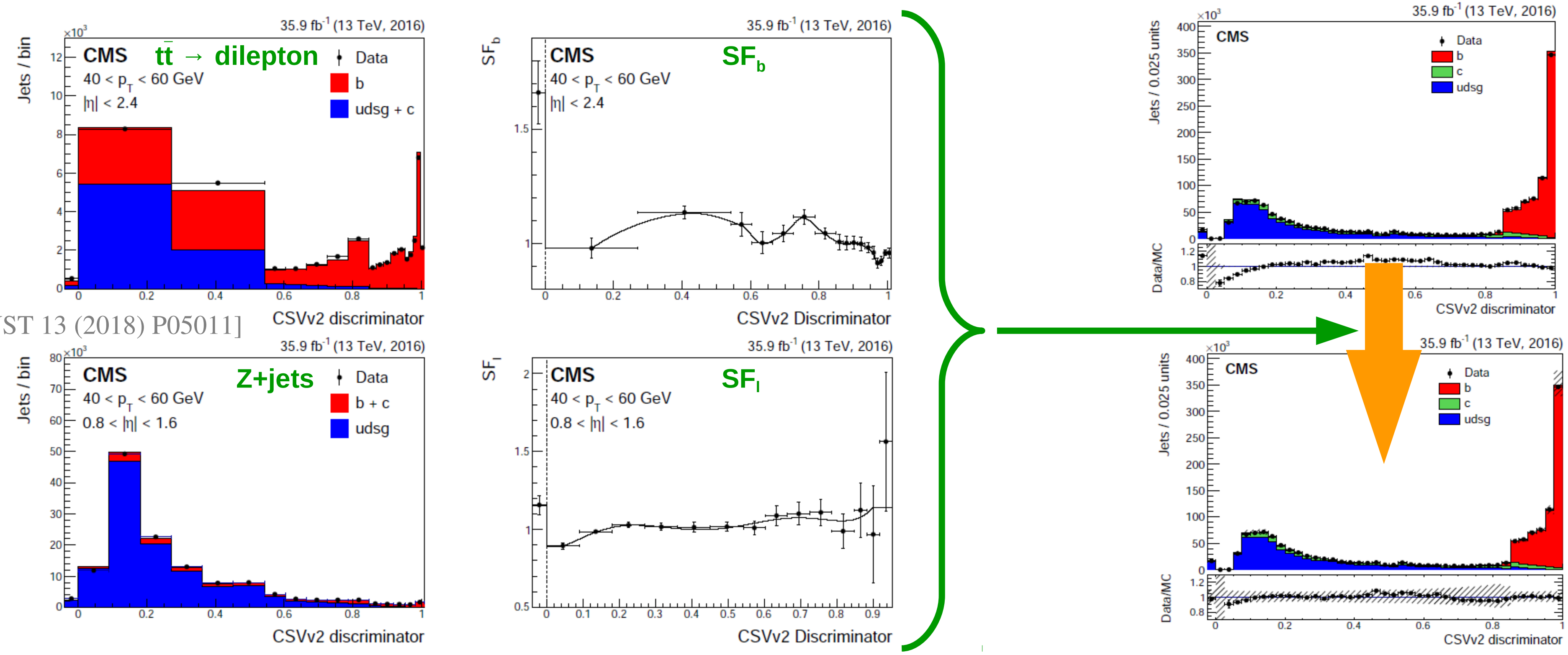
- **The performance does not really depend on the initialization of the neural network weights in the training**



[CMS-DP-2018-058]

Discriminator shape calibration – Iterative Fit

- **Some physics analyses use the shape of the algorithm discriminant**
  → shape calibration is required (scale factors depending on discriminant value)
- **Scale factors for b and light jets are simultaneously determined by an iterative procedure in two samples: $t\bar{t}$ → dilepton and Z+jets events**

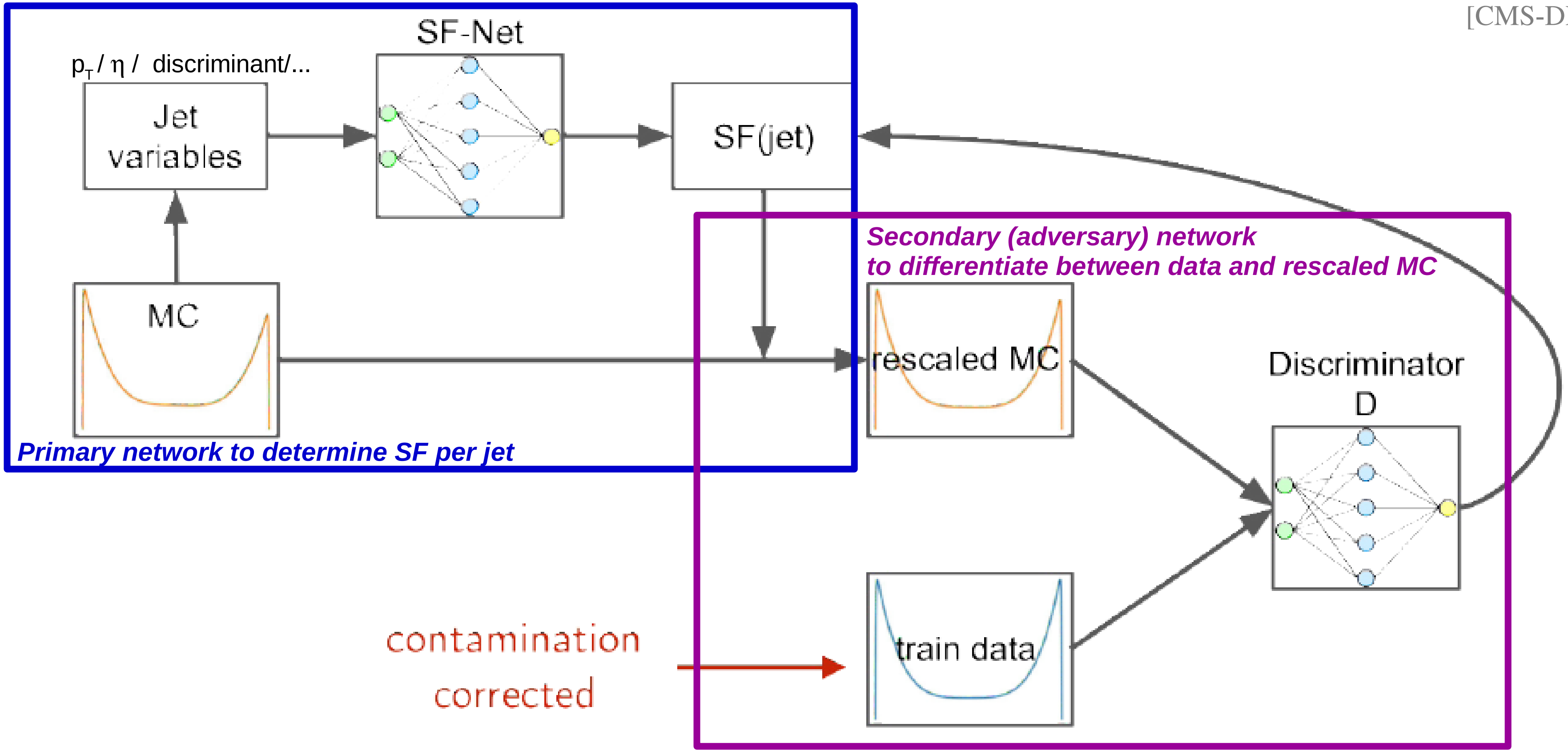


[JINST 13 (2018) P05011]

Calibration with an adversarial neural network

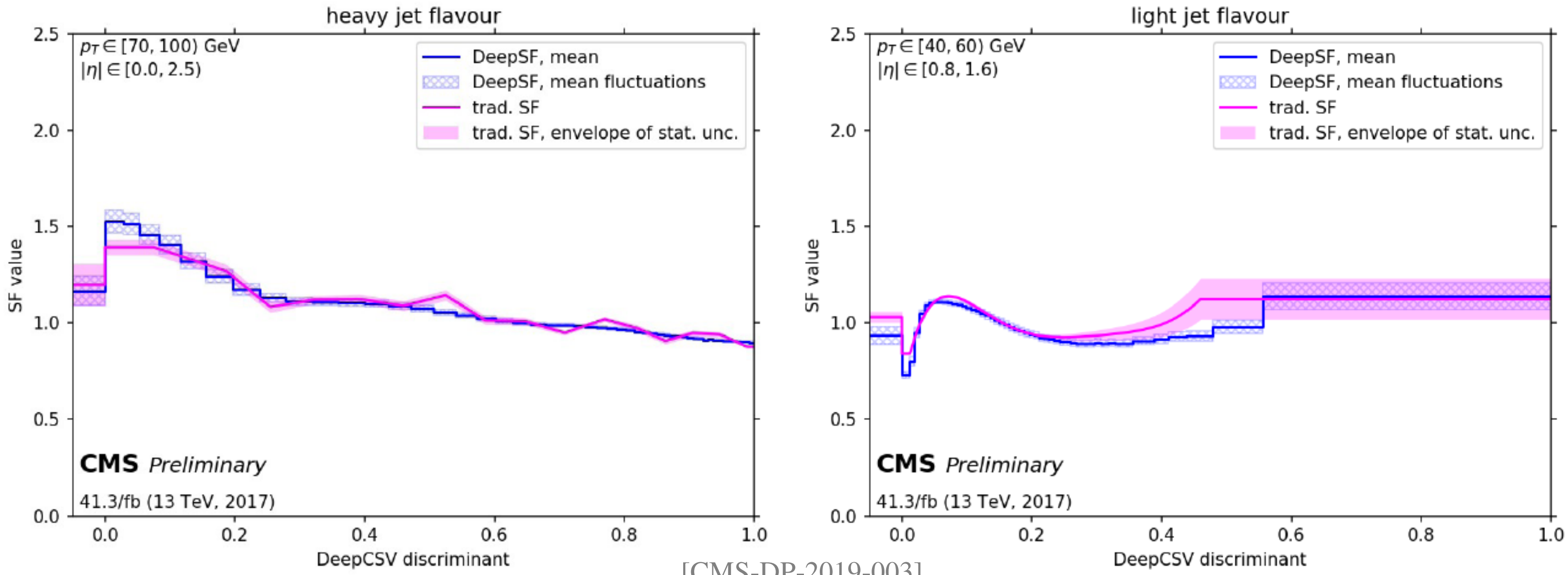
- **Previous procedure: coarse p_T/η bins and polynomial functions or splines**
- **New procedure: determine the scale factors with an adversarial neural network**

[CMS-DP-2019-003]



Adversarial neural network gives smoother SF

- **Comparison of scale factors obtained with the iterative fit procedure and the adversarial neural network (new procedure)**
  - The scale factor in each bin is given by the mean of 25 trainings (different seeds)
  - Uncertainty on the mean is an indication for the stability of the procedure



[CMS-DP-2019-003]

- **Smoother dependence with the new procedure + scale factors are compatible!**

Outline

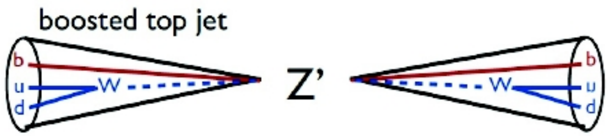
- **Heavy-flavour identification in standard topologies**
  - Evolution of the algorithms and their performance
  - Algorithm calibration (aka scale factors) – standard methods
  - Discriminator distribution shape calibration using adversarial neural network

- **Heavy-flavour identification in boosted topologies**
  - Tagging jets with two heavy-flavour quarks from a boosted H or Z boson decay
  - Recent developments for double-b/c tagging algorithms
  - Boosted top quark identification

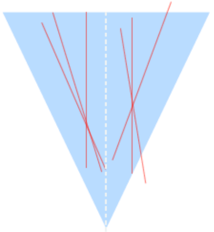
- **Outlook and conclusion**

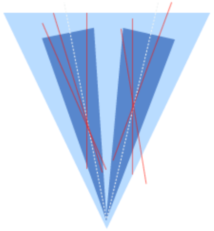
Boosted topologies with b jets

- **Boosted particles decaying to b quarks, e.g.:**
  - t → bW
  - H or Z → bb

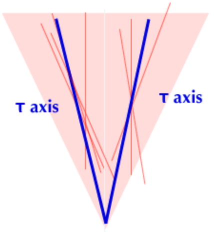


- **General approaches for b jet identification in boosted topologies:**
  - **Fat jet (AK8) b tagging:** retraining algorithms with relaxed criteria for the association of tracks or secondary vertices to the jet
  - **Subjet b tagging:** resolve jet substructure with soft drop jet declustering and apply b jet identification algorithm on subjets





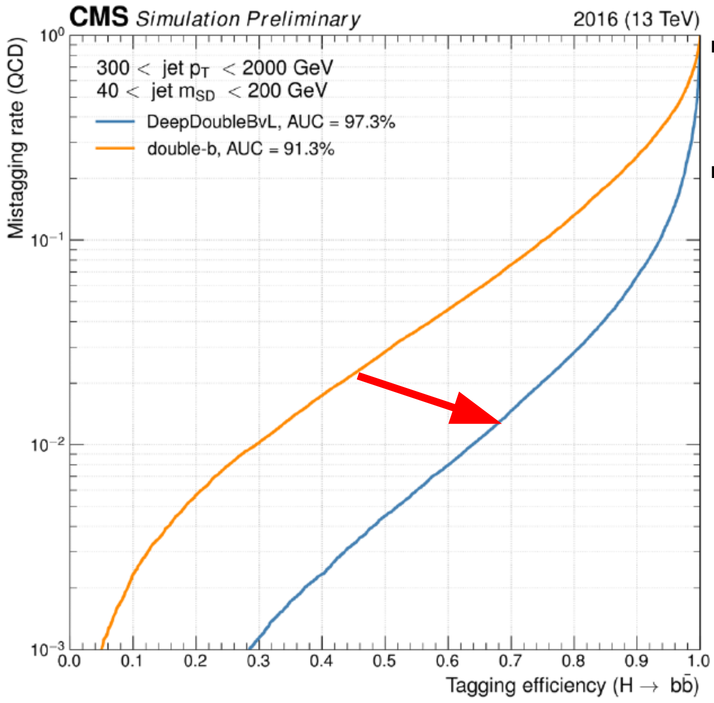
- **Dedicated approaches for b jet identification in boosted topologies:**
  - **t→ bW → bqq' tagging:** "the top tagger" is a deep neural network combining >150 features from all the jet constituents → see CMS-DP-2017-049
  - **H/Z→ bb tagging:** "the double-b tagger" combines 27 jet properties exploiting the correlations between the flight directions of the b quarks with a boosted decision tree (BDT) → see JINST 13 (2018) P05011



New algorithms for boosted H/Z→bb and H/Z→cc decays have been developed: DeepDoubleBvL, DeepDoubleCvL and DeepDoubleCvB

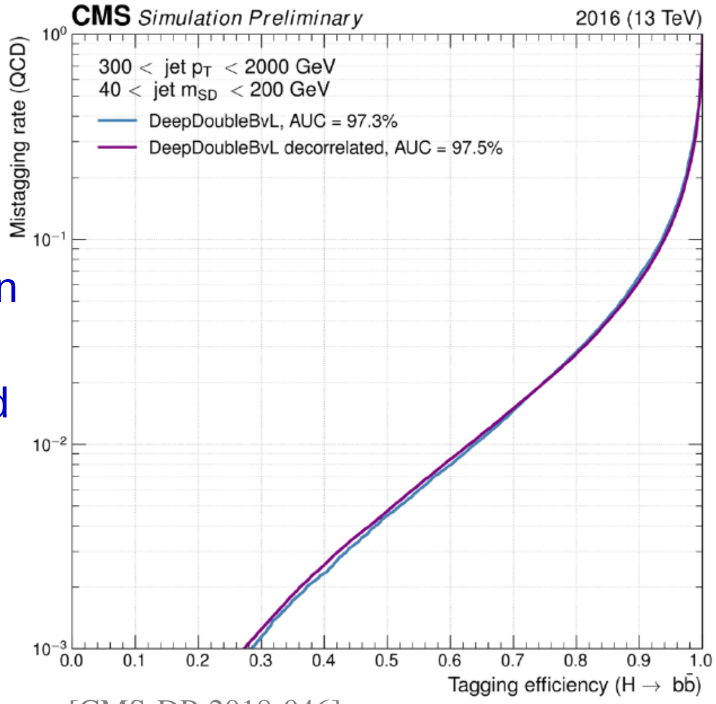
The DeepDoubleBvL algorithm

- **The DeepDoubleBvL algorithm is based on a deep neural network with a similar architecture as DeepJet algorithm**
  - Same properties as for double-b tagger
  - Additionally, 8 features of up to 50 tracks and 2 features of up to 5 secondary vertices



[CMS-DP-2018-046]

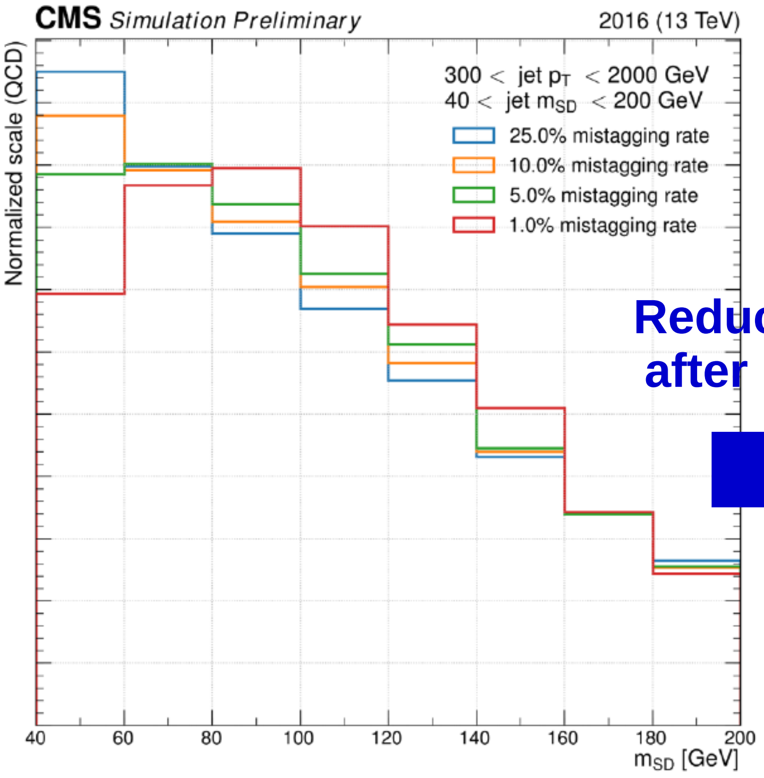
- **Left: large performance gain for new approach**
- **Right: same performance after mass decorrelation**
  - Tagger should not depend on mass of the jet
  - Mass decorrelation achieved by adding a penalty term to the loss function in the neural network training → *penalizes the difference between the mass distribution for tagged and untagged jets*



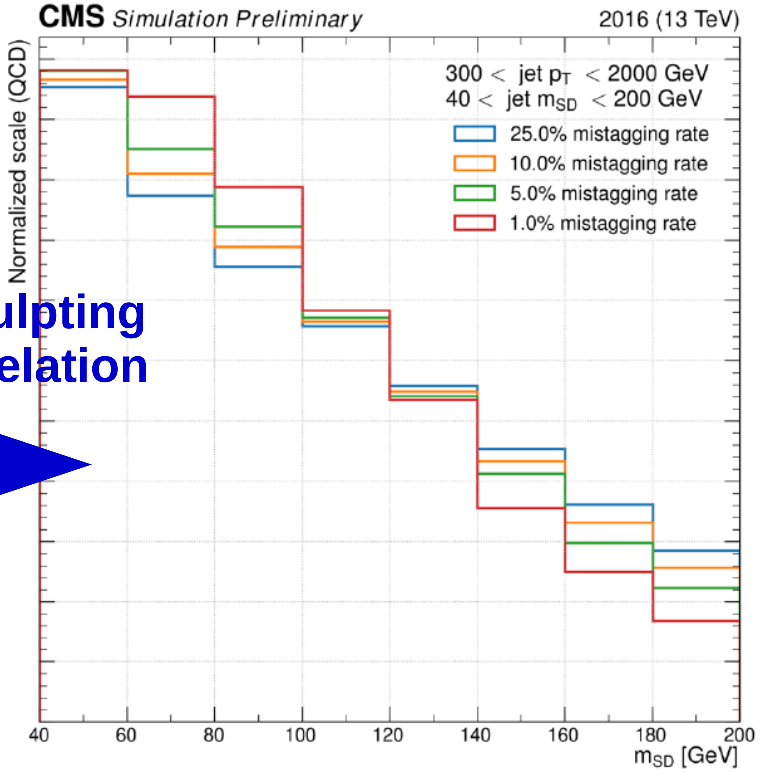
[CMS-DP-2018-046]

Mass decorrelation for DeepDoubleBvL

- **The jet soft-drop mass distribution is shown for misidentified jets in QCD multijet events for four fixed misidentification probabilities of the DeepDoubleBvL tagger**
- **Mass sculpting reduced after mass decorrelation**



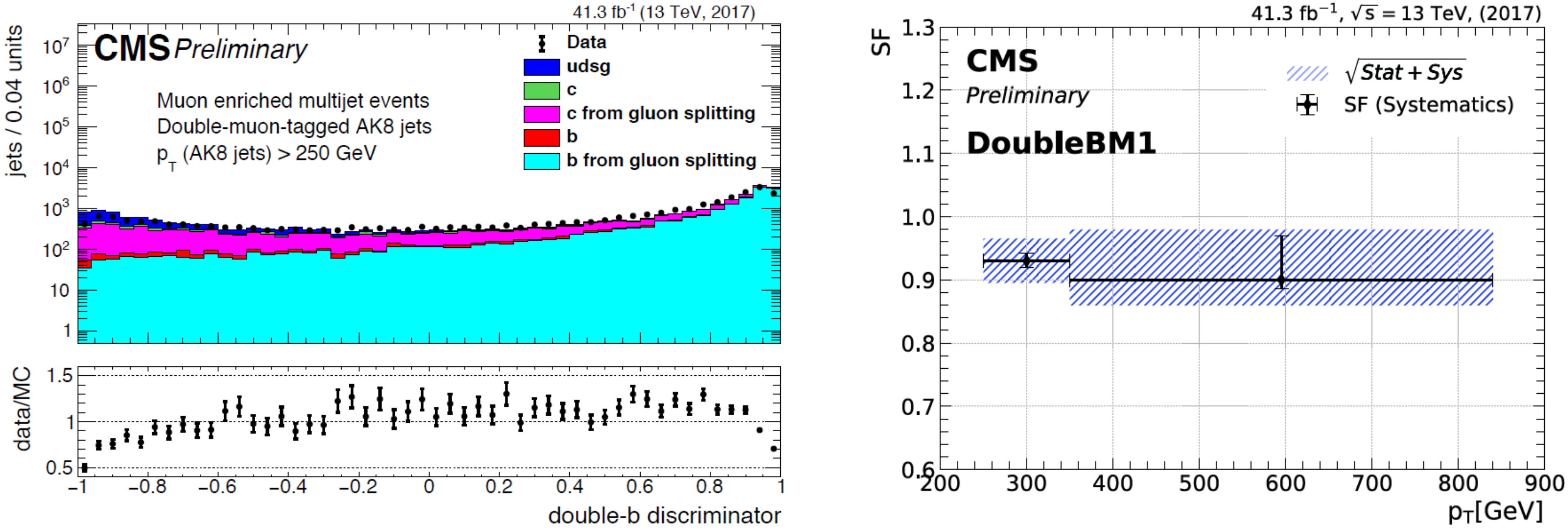
Reduced mass sculpting after mass decorrelation



[CMS-DP-2018-046]

Calibration of the double-b tagger

- **Calibration of the double-b tagger is achieved by selecting muon-enriched QCD multijet events: AK8 jets with 2 muon-tagged subjets**



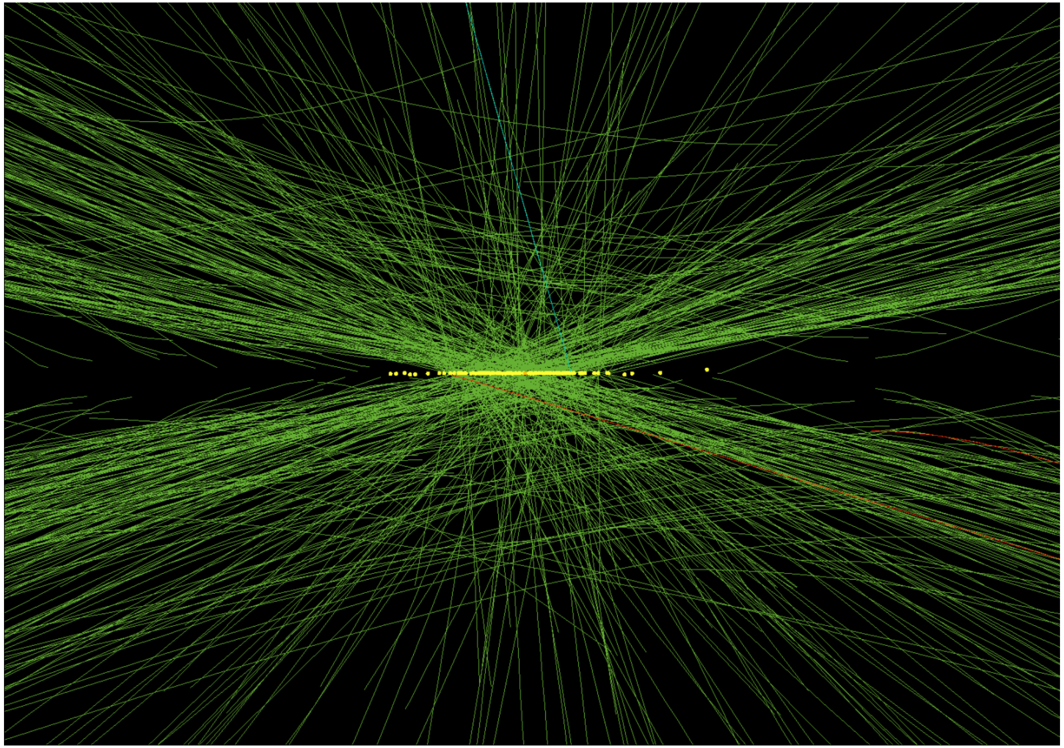
[CMS-DP-2018-033]

- **Misidentification probability is determined in control region for most analyses**

A flavour of the future → tomorrow

- **Machine learning and potential**
  - What is happening in ATLAS and CMS
    → Jean-Roch Vlimant and Tobias Golling
  - Fast inference on FPGA (flavour identification at lowest level trigger)
    → Sioni Paris Summers

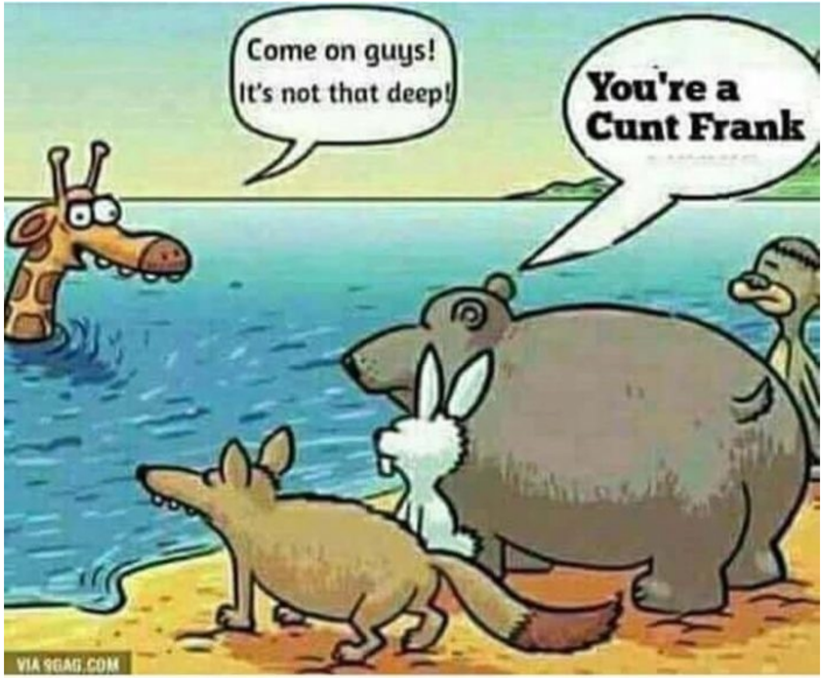
- **Preparation of CMS for the HL-LHC with 140-200 pileup collisions**
  - Flavour tagging algorithms
    → Daniel Bloch
  - Status of the MIP Timing Detector
    → Paolo Meridiani
  - Physics potential with flavour tagging
    → Jyothsna Rani Komaragiri



An event display with 78 reconstructed collisions [Andre Holzner/CERN]

Conclusion

- Over the last year(s) many developments happened:
  deep → deeper → deepest



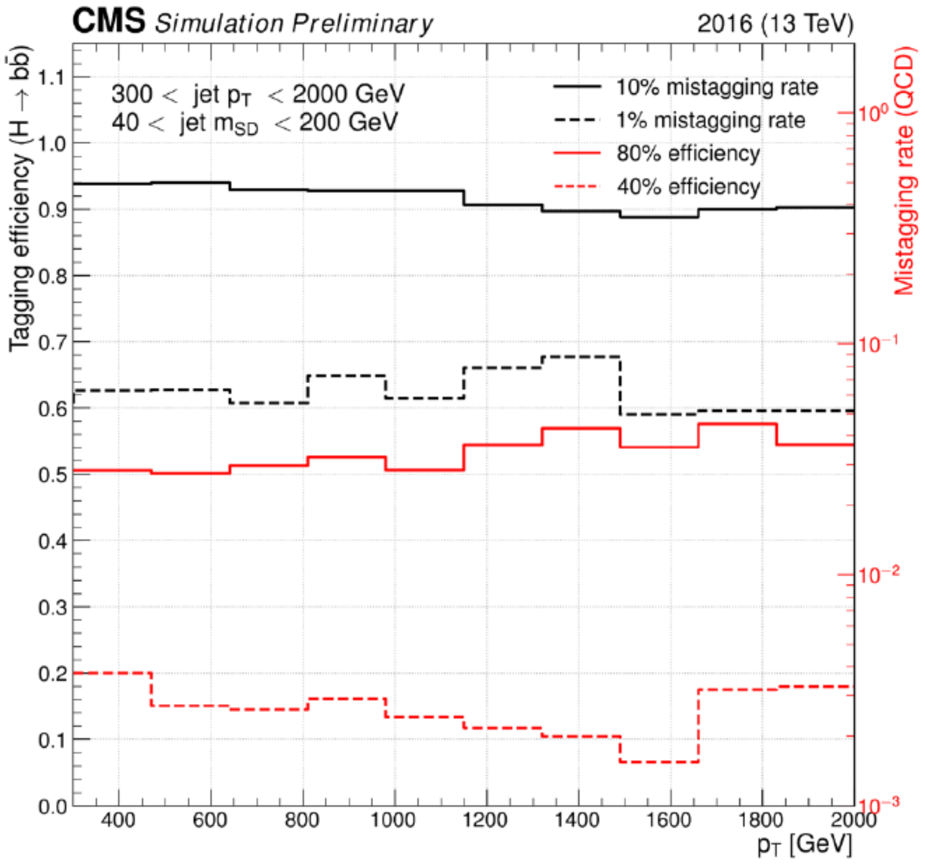
- New paper in preparation!
- Further improvements are happening, but new ideas are always welcome!

Additional material

- **DeepDoubleBvL performance dependence on jet pT**
- **DeepDoubleCvL and DeepDoubleCvB algorithms**

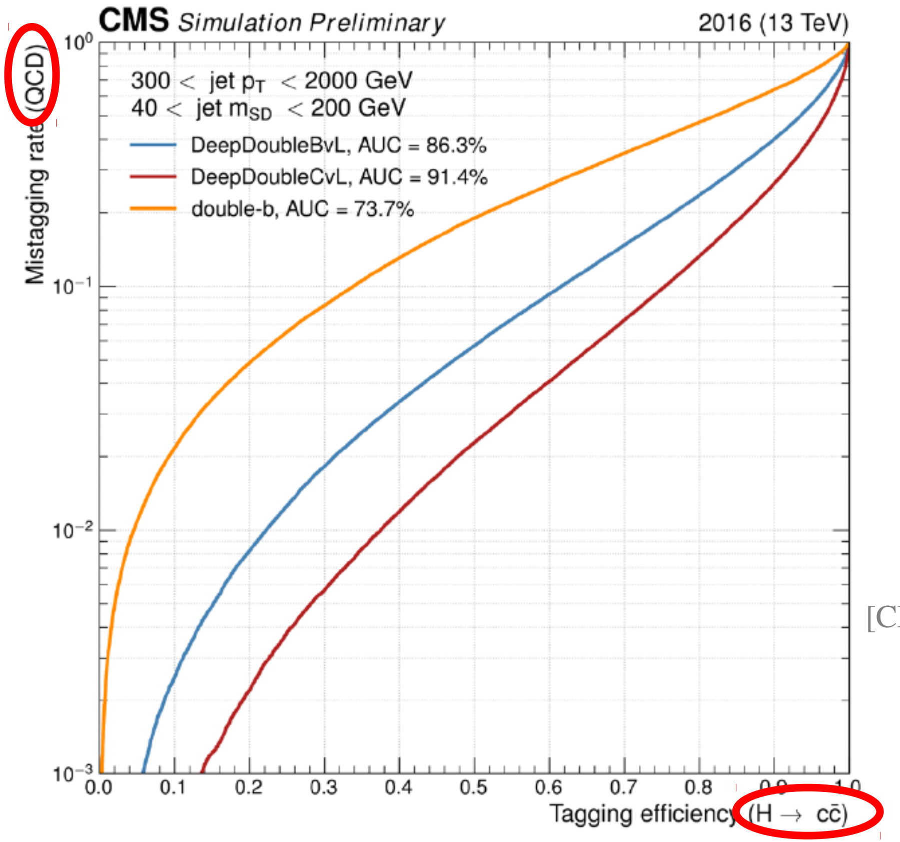
Performance dependence on jet p_T

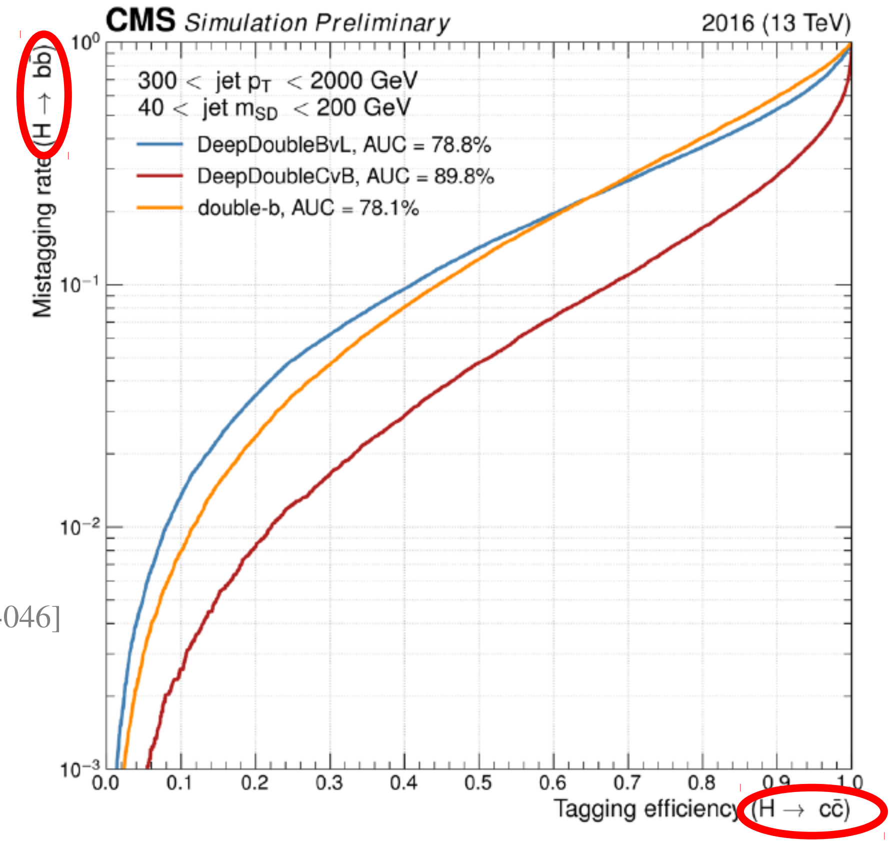
- **The DeepDoubleBvL performance is relatively stable with the jet p_T**



The DeepDoubleCvL / CvB algorithm

- **Identical as DeepDoubleBvL, but with the aim to identify boosted H→cc decays**
- **Left (right): the DeepDoubleCvL (DeepDoubleCvB) outperforms the other algorithms**





[CMS-DP-2018-046]