Status of b-tagging performance

Alessandro Calandri
CPPM-Aix Marseille Université

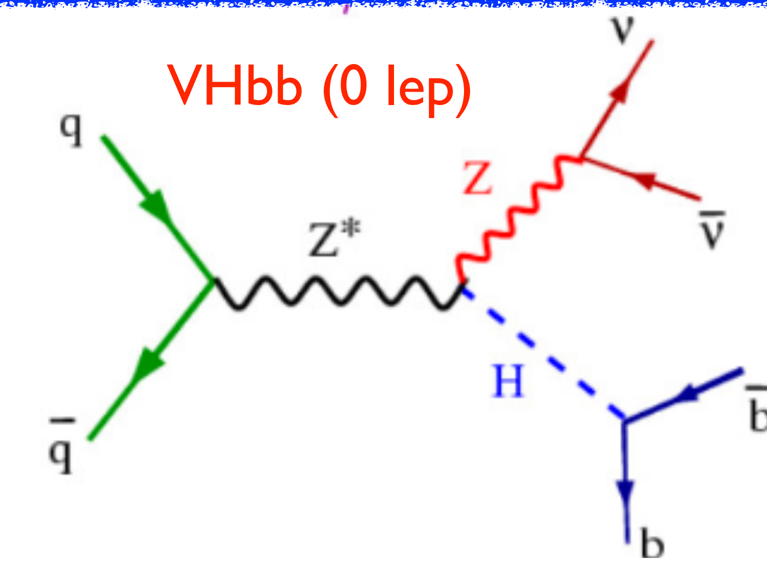
on behalf of the ATLAS and CMS Collaborations

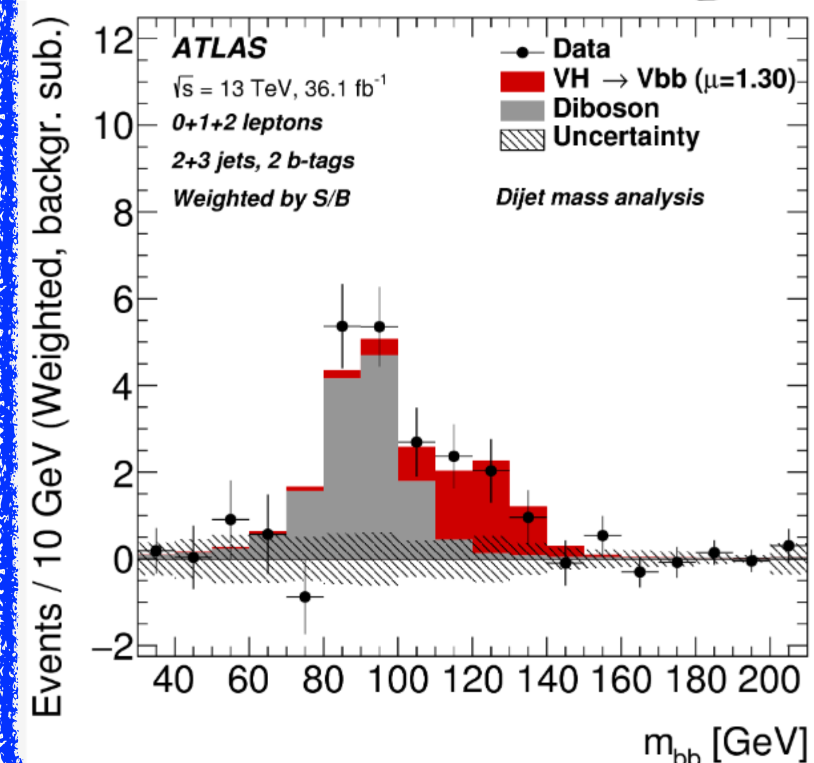
Top LHC France 2018, May 24th 2018 @ Paris

Motivation of b-tagging and outline of the talk

✔ b-jet identification (b-tagging) is crucial for Standard Model, Higgs and BSM physics at the LHC

- b-quarks present in the top quark decay
  V(tb)~1 → BR(t → Wb)~100%
- largest Higgs decay branching ratio (57%) is H → bb





✔ General remarks on b-tagging

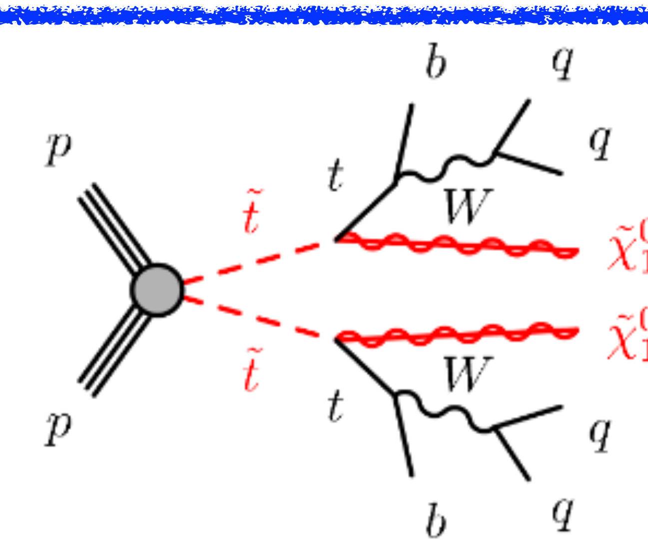
✔ Algorithm performance and results in ATLAS and CMS

✔ Looking at data...calibration in ATLAS and CMS

✔ A few words on upgrade studies for ATLAS/CMS Technical Design Reports

Documentation on CMS and ATLAS b-tagging available here:

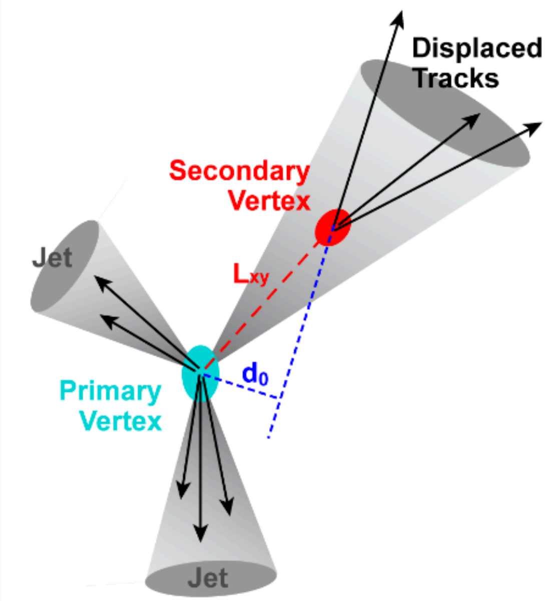
- https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsBTV
- https://twiki.cern.ch/twiki/bin/view/AtlasPublic/FlavourTaggingPublicResultsCollisionData



b-tagging

✓ b-jets stem from the process of hadronization of b-quark - 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 (~70%) jet momentum fraction carried by B-hadrons
- ✓ high charged tracks per decay (~ 5 tracks)
- ✓ presence of direct and indirect semileptonic decays, $b\to\mu\nu X$ (BR~12%), $b\to c\to\mu\nu X$ (BR~10%)
  - ✓ presence of soft muons or electrons in jets



✓ jet (exclusive) labelling in ATLAS

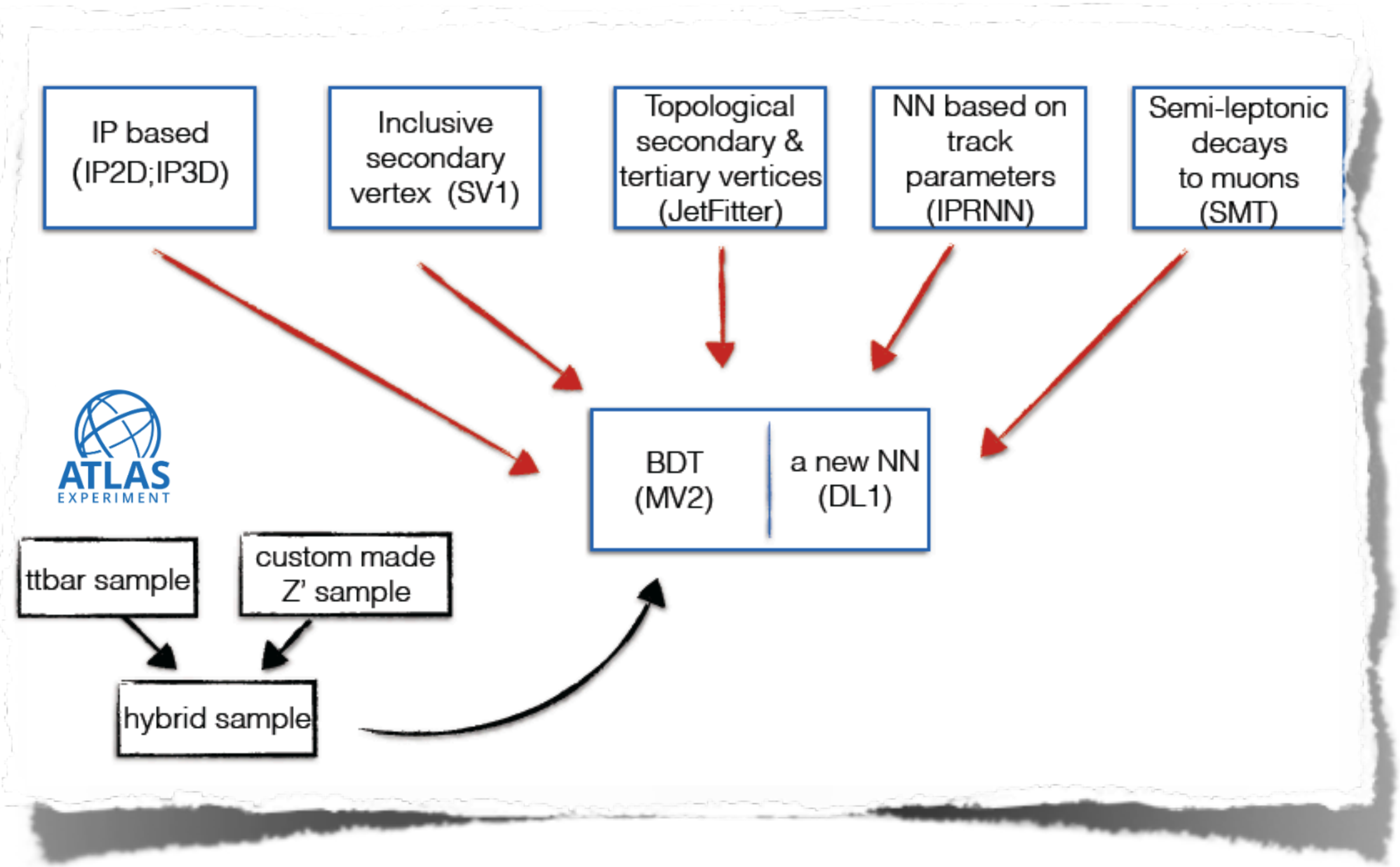
- ✓ b-jets: jets including one b-hadron in ΔR (b-hadron,jet)=0.3 with pt>5 GeV
- ✓ c-jets: jets including at least one c-hadron in ΔR=0.3 with pt>5 GeV
- ✓ τ-jets: no b/c-hadrons but at least one τ in ΔR=0.3 cone
- ✓ light-flavour jets: all the rest

✓ b-jet labelling in CMS

- ✓ generated heavy hadrons used in jet clustering with momentum set to negligible value (ghost-association)
- ✓ flavour assignment based on presence of ghost b/c
- ✓ pile-up jets are defined as jets not matched with generator level jets

b-tagging algorithms and performance

b-tagging chain @ ATLAS and CMS



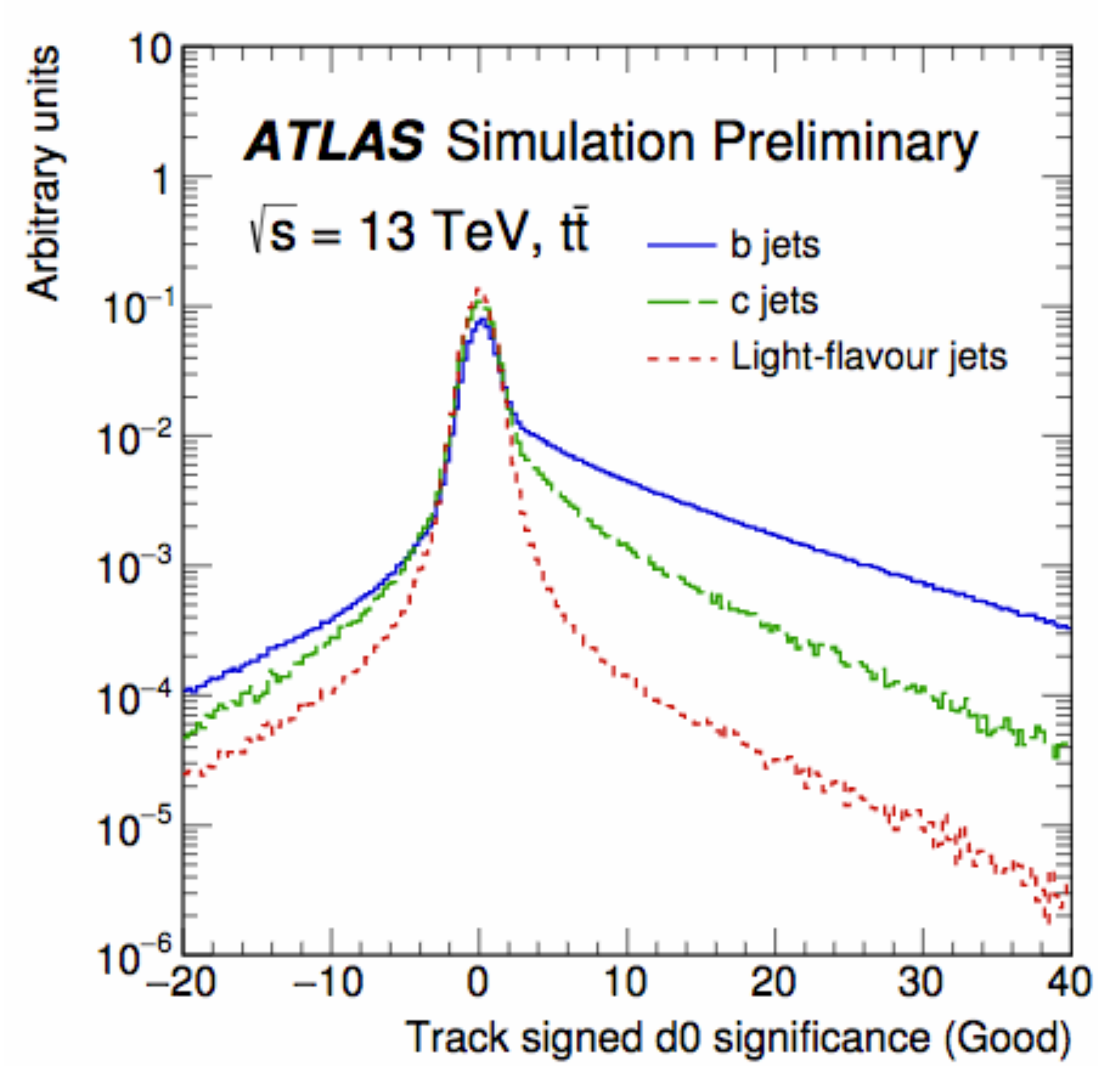
- ✓ Inputs from low-level taggers exploiting the features of the b-jet decay topology
  - impact-parameter tracks associated to jets, presence of secondary vertices, soft muons from semileptonic b-decays
- ✓ Combination of inputs using high-level tagger algorithms exploiting Boosted Decision Trees or Machine Learning techniques

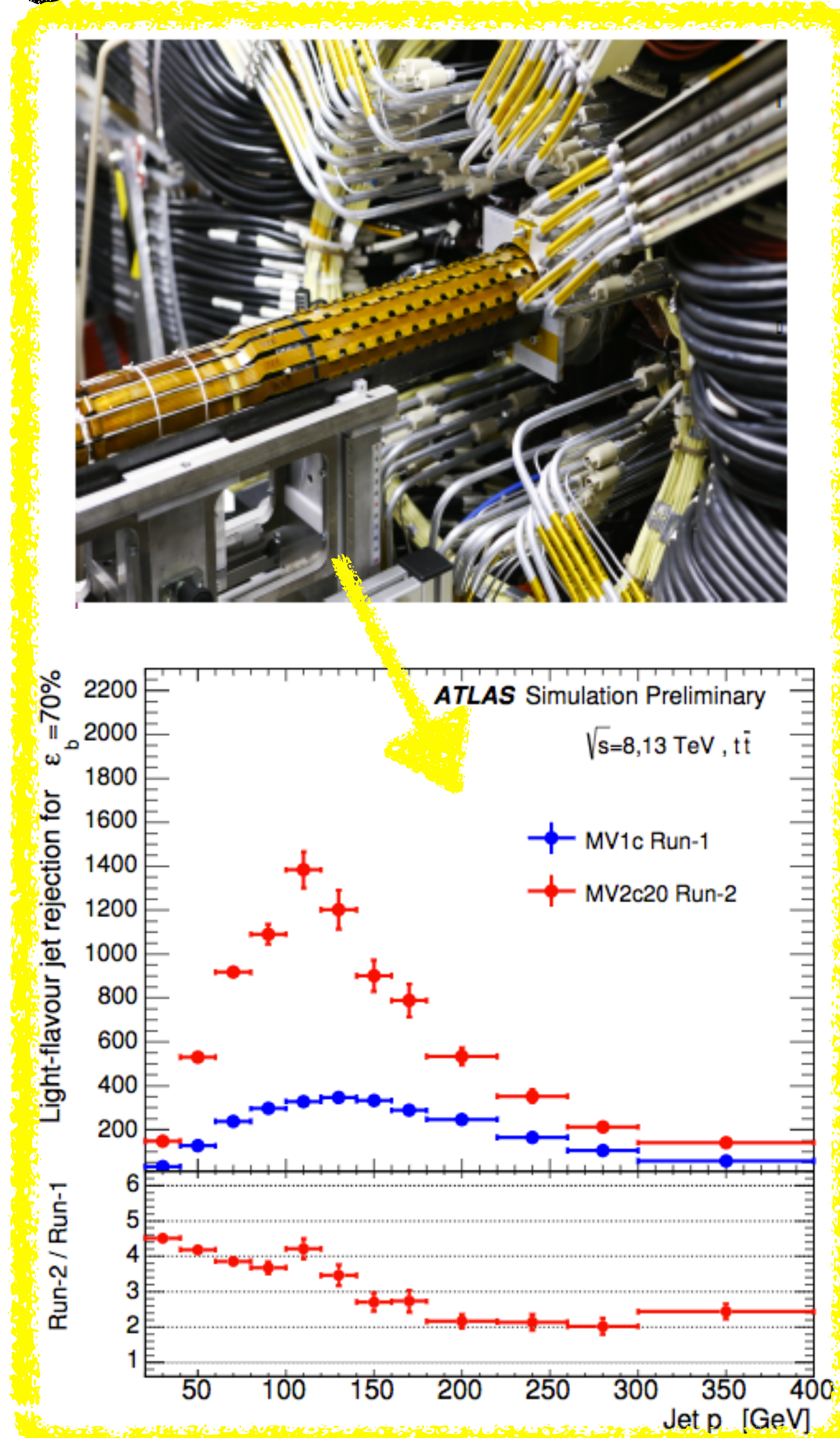
	ATLAS	CMS
Large impact parameter-tracks	IP2D, IP3D, RNNIP	TCHP, TCHE, JP, JBP
Secondary-vertex reconstruction	SV1, JetFitter	SSVHP, SSVHE
Soft leptons stemming from semileptonic b-decays	SMT	Soft Lepton Taggers
Combinations	MV2c10/DL1	CSVv2, cMVAv2, DeepCSV, DeepFlavour

Low-level tagger algorithms

✓ Account for features of the b-system (impact parameter of tracks associated to jets, SV reconstruction, soft leptons) and exploit discrimination wrt background jets (light-flavour and c-jets)

- used as inputs of high-level tagging discriminants
- detector response crucial to achieve good discrimination (e.g IBL in ATLAS ensures better track resolution and robust pattern recognition for IP2D/ IP3D algorithm - important for IP and SV algorithms)



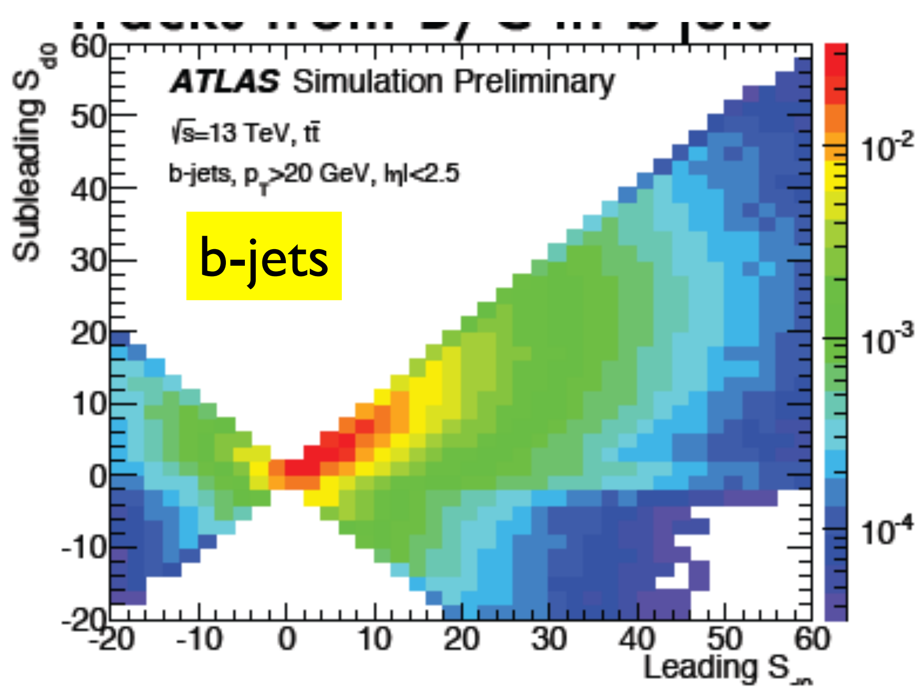


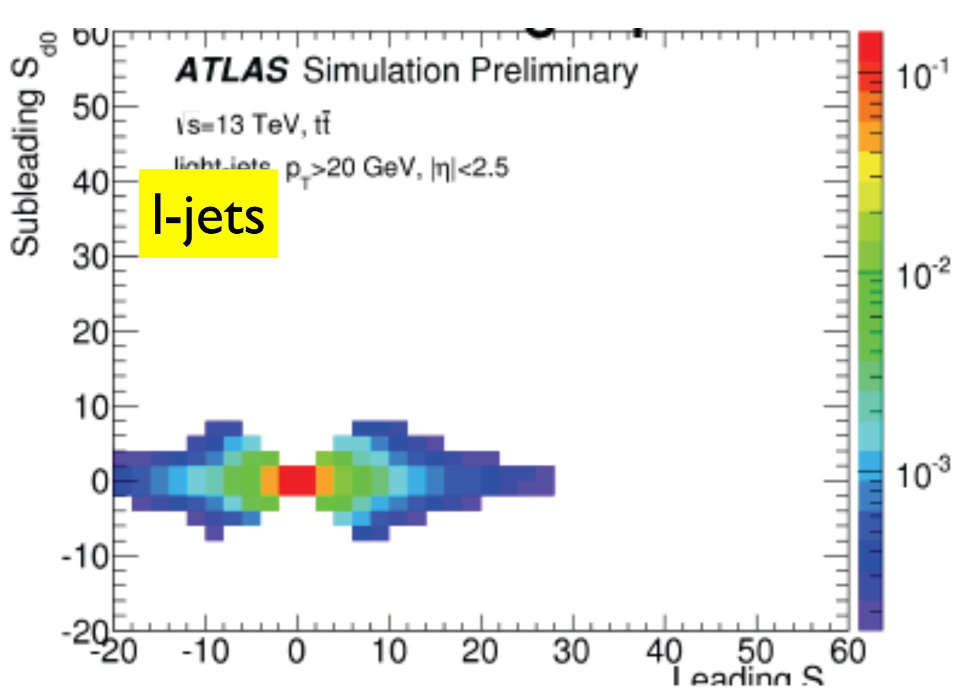
IPRNN

- Correlation between tracks associated to jets exploited with neural network techniques (Recurrent Neural Network tagger)

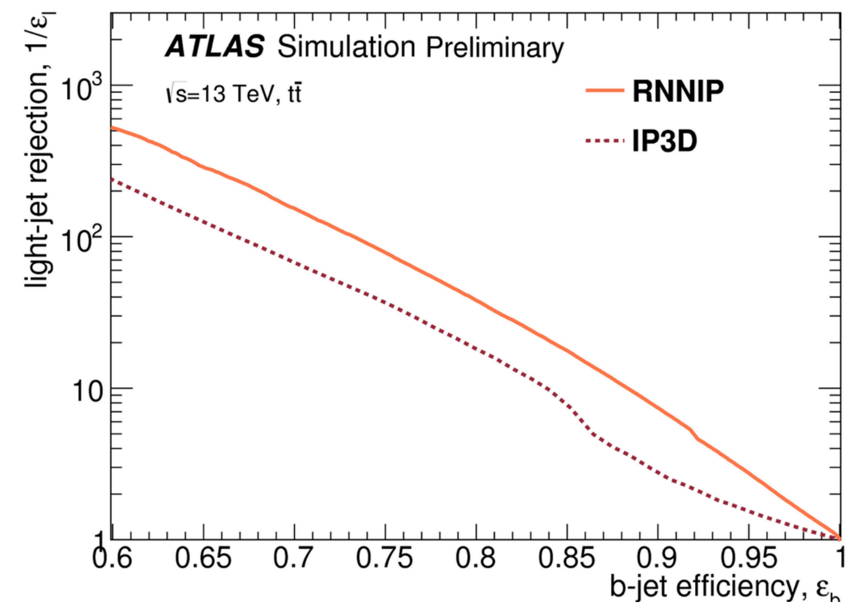
- IP2D/IP3D → properties of tracks are treated as independent and the template PDF's in hit categories are built neglecting track-to-track correlations

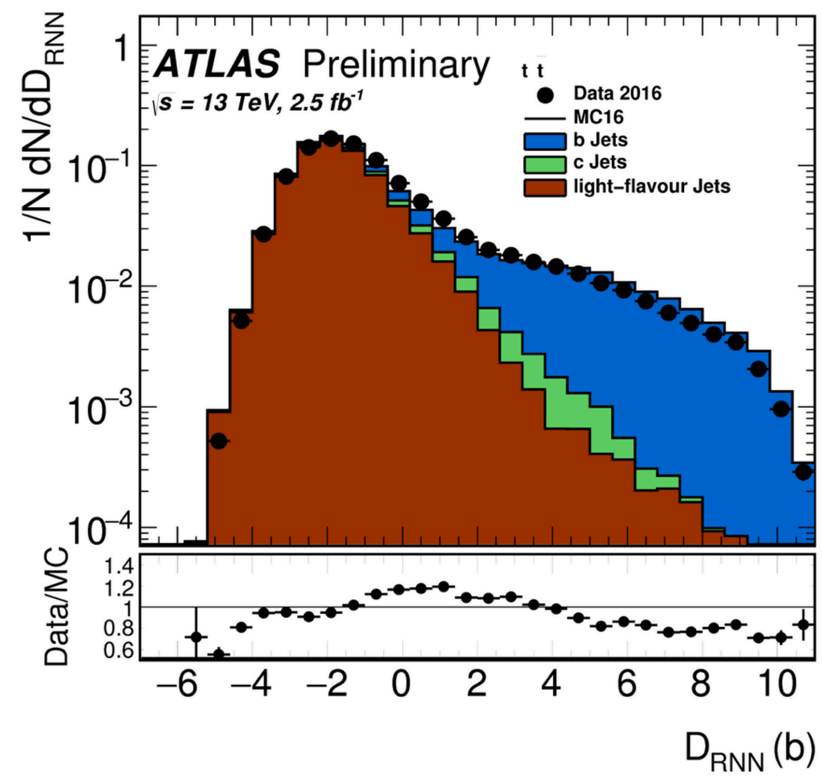
- Sequential dependencies between discriminating variables used for full characterization of properties of b-jets





Track Variable
S_{d_0}
S_{z_0}
Category
p_T^{frac}
ΔR(track, jet)



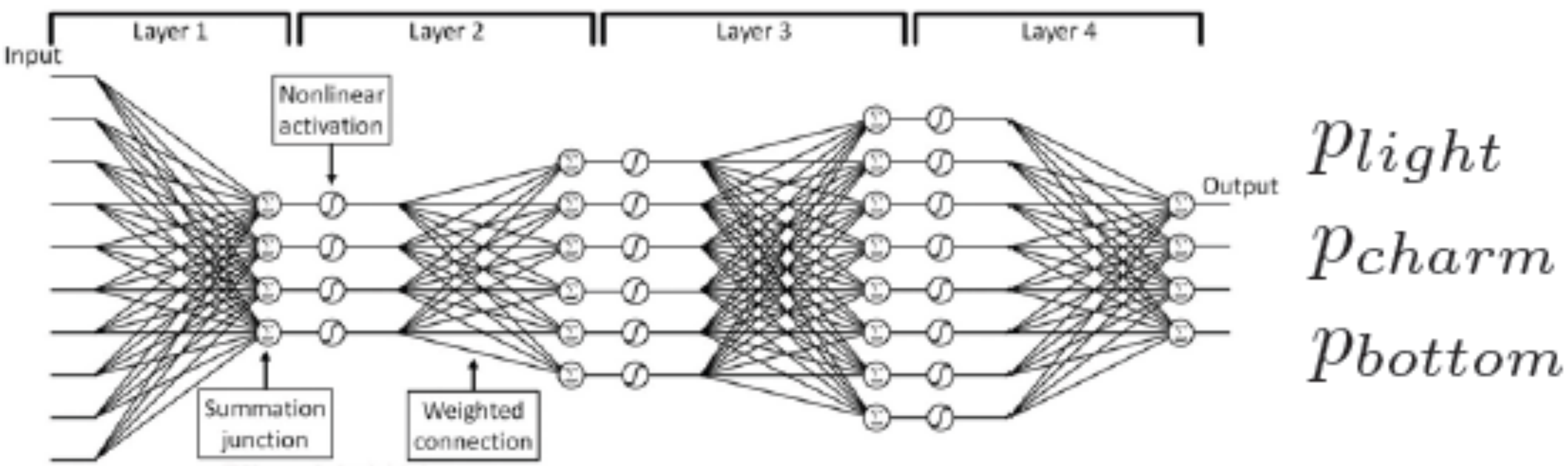


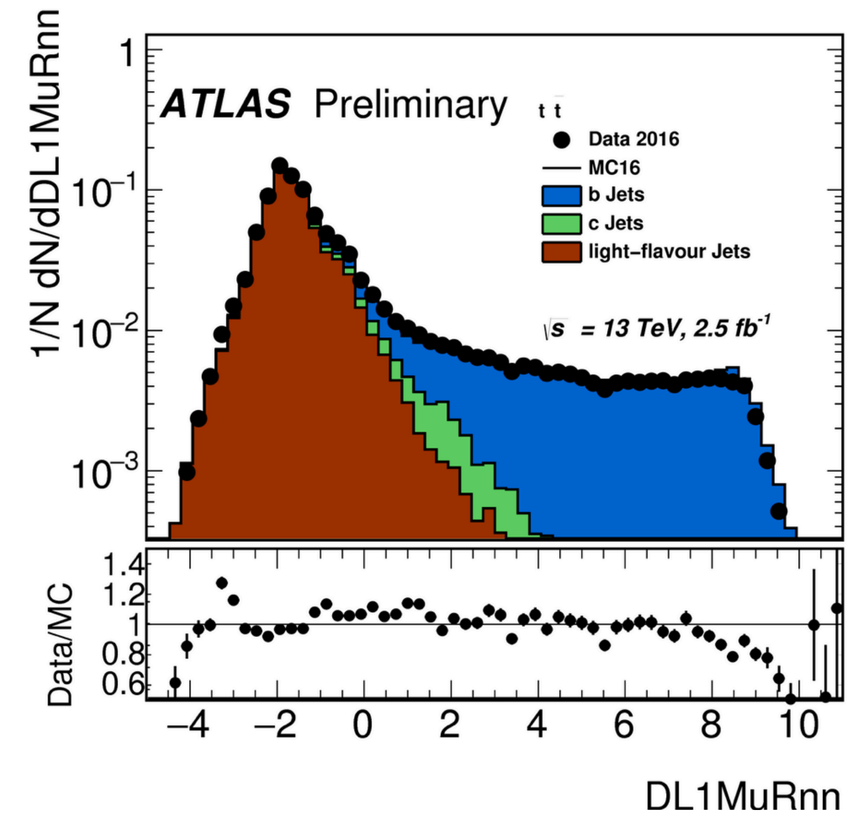
Deep Learning

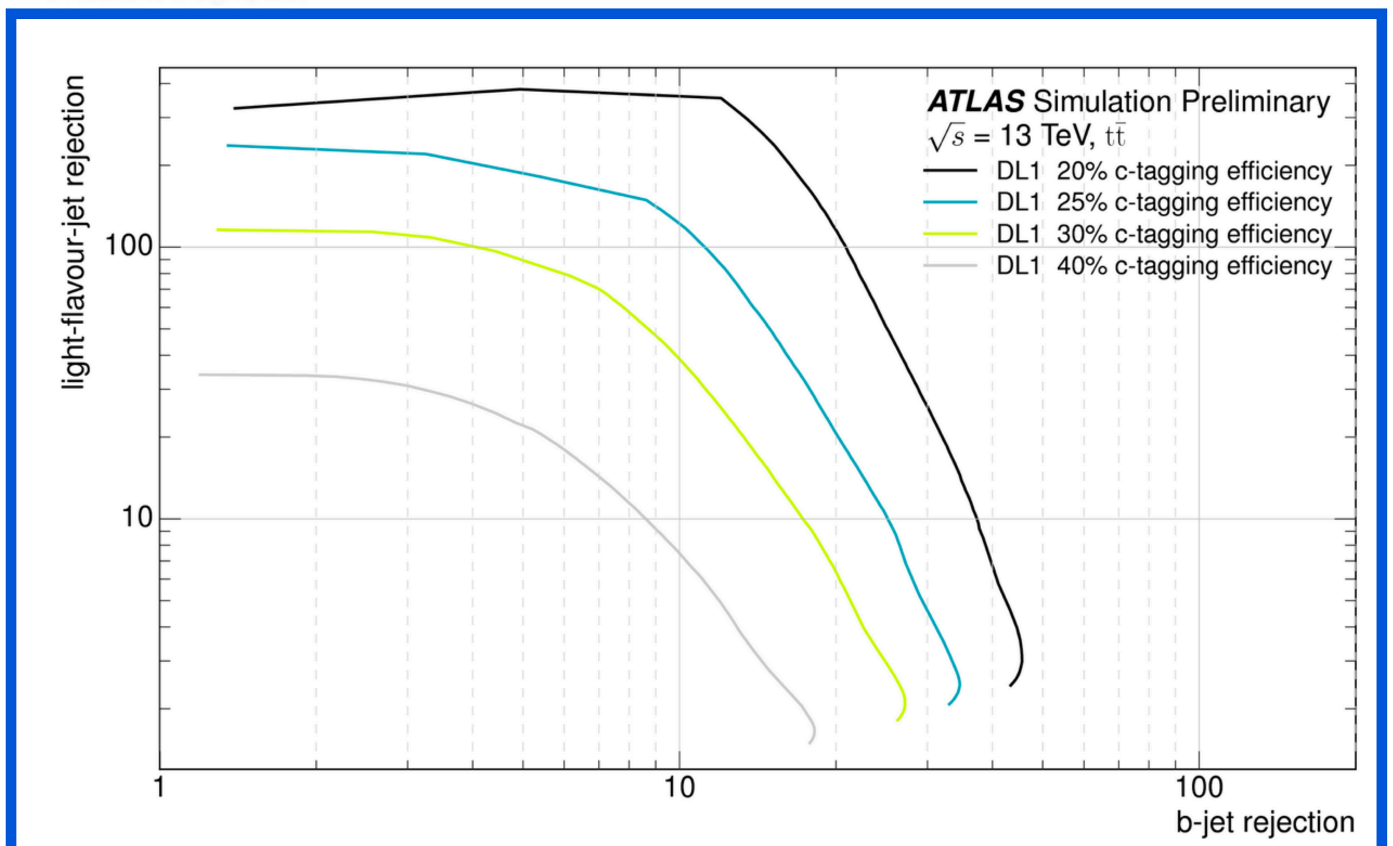
✓ Exploits the advantage of multivariate techniques with multiple output nodes - exploited for b-c tagging

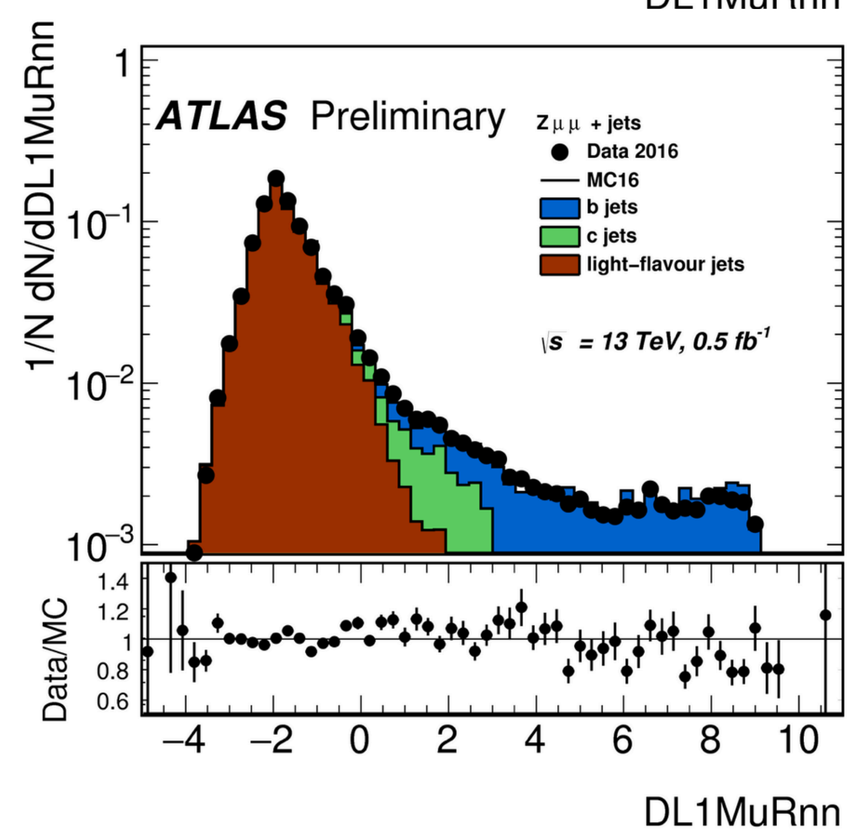
- fed with same input information as in MV2 → results in comparable performance

➡ Tunable c-fraction in the network → no need to retrain for different background components





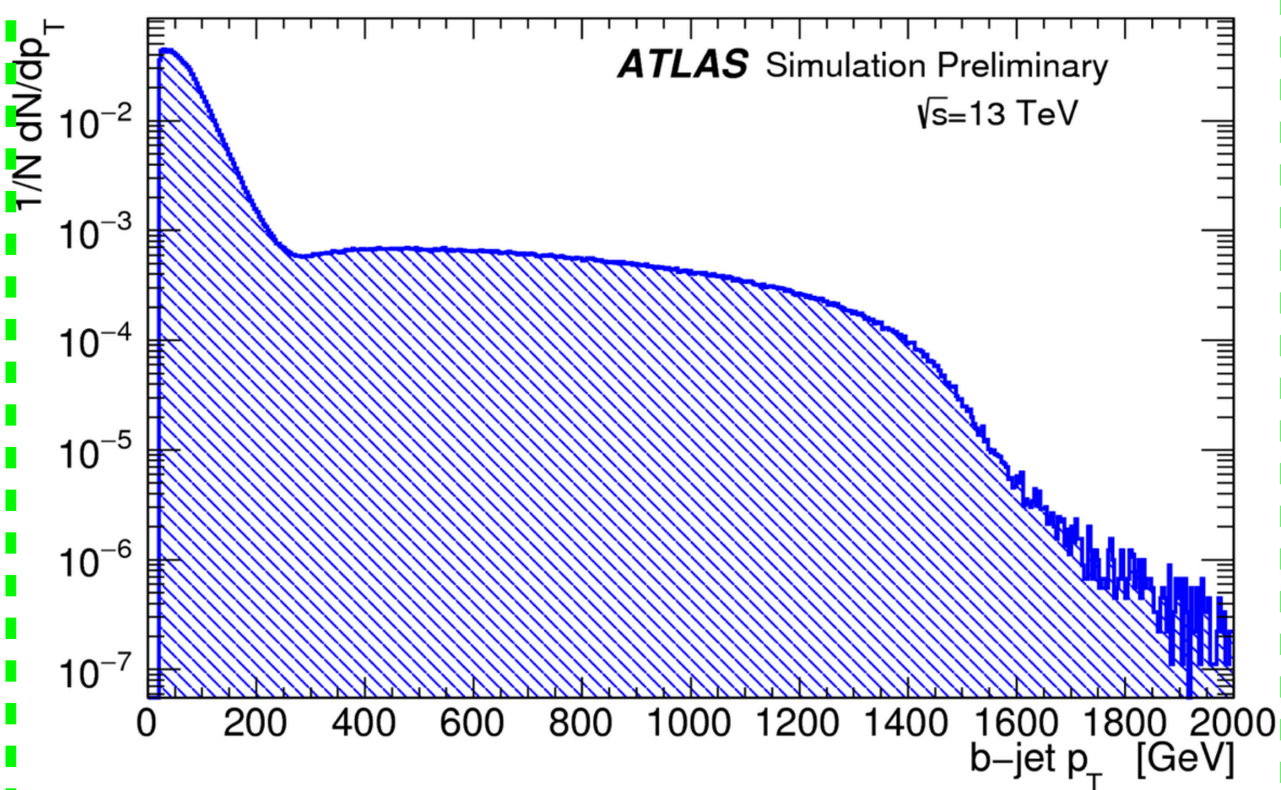


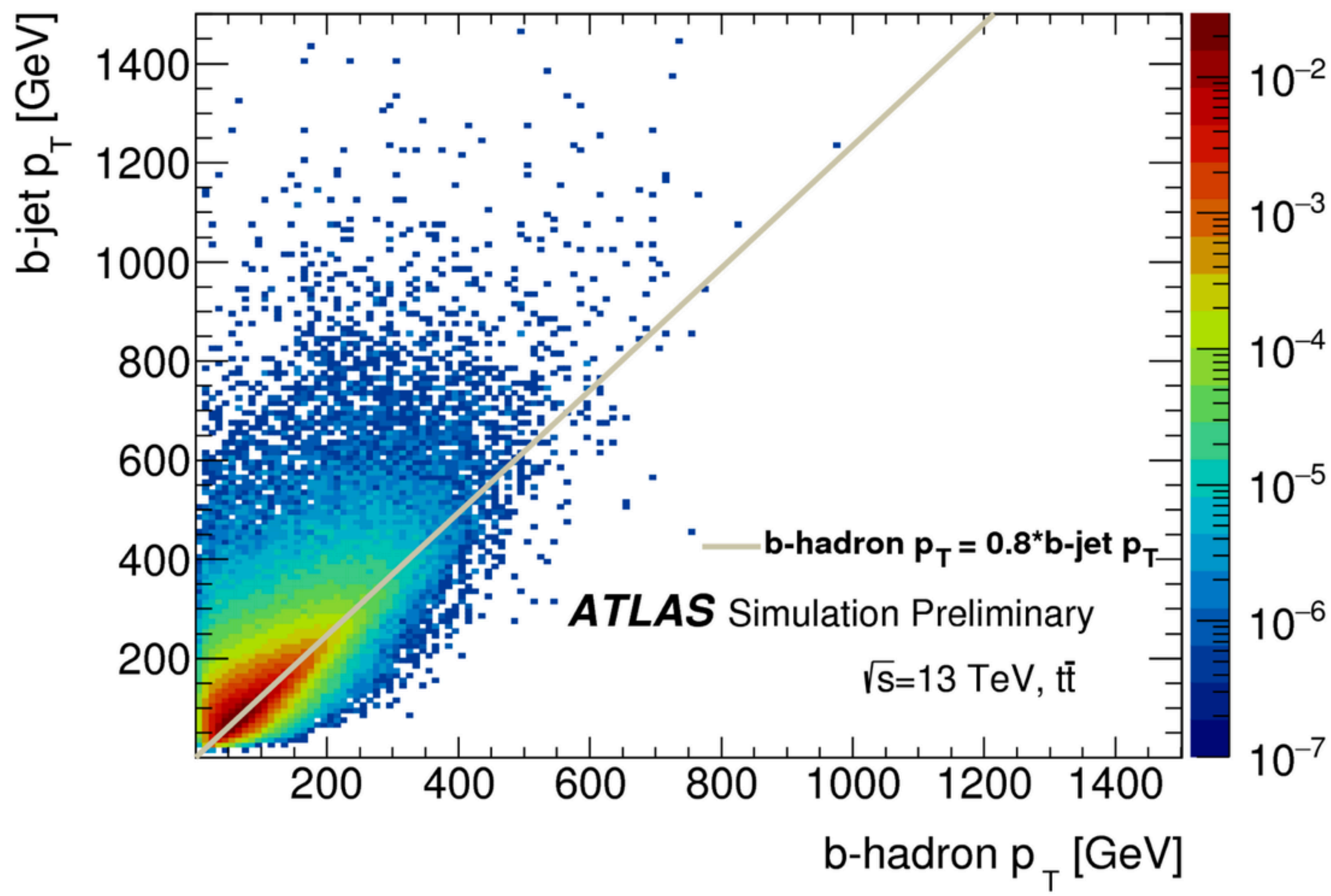


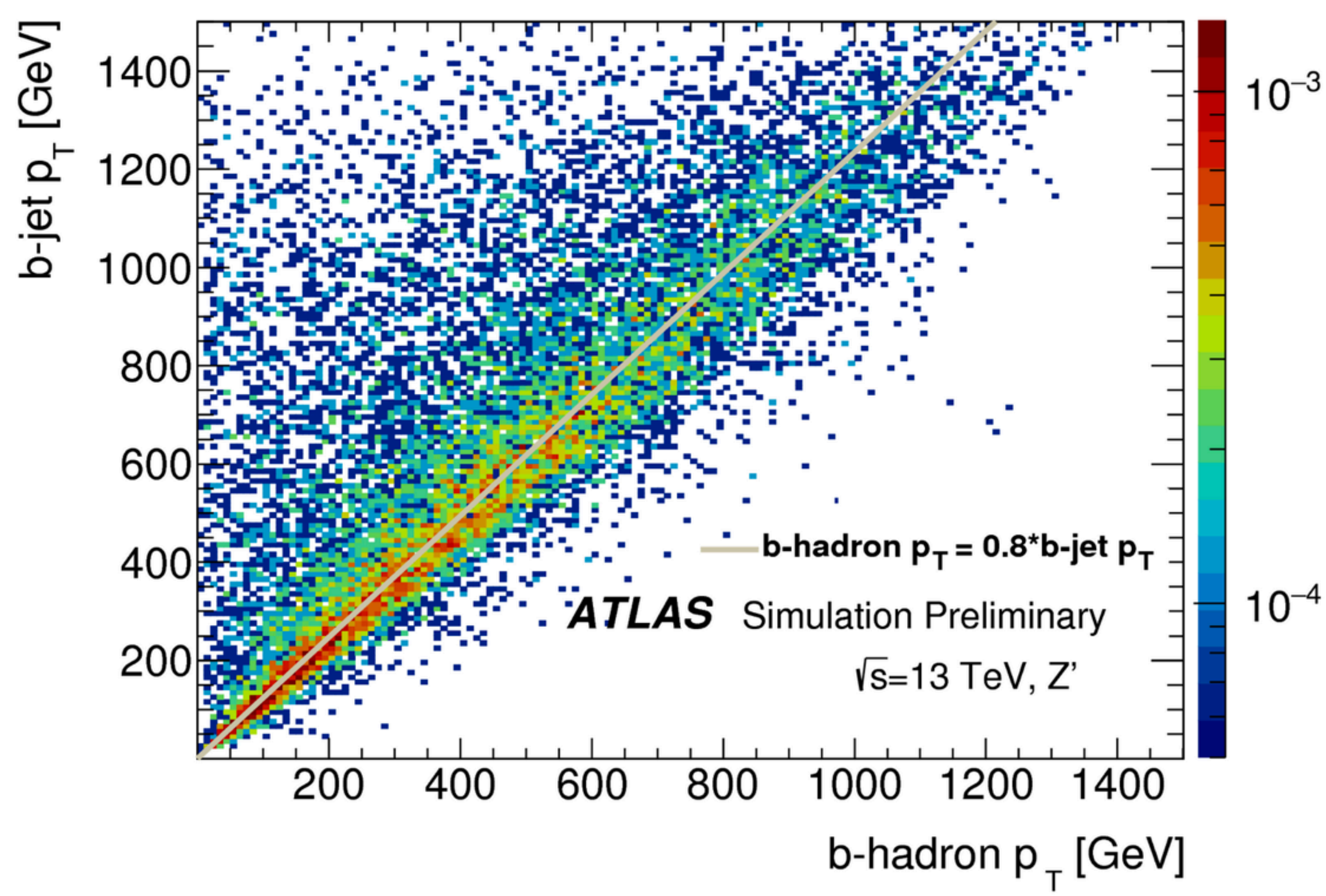
ATLAS EXPERIMENT

Algorithm training samples

- ✓ Studied b-hadron pt vs b-jet pt correlation in ttbar and broad Z' sample
  - ttbar sample looses correlation above mT (merging of jets), while Z' fully characterizes the high pt phase space
- ✓ New hybrid sample used for training of high level tagger algorithms
  - low pt ttbar + high pt Z' sample composition
  - similar performance at low pt but significantly larger rejections at high pt

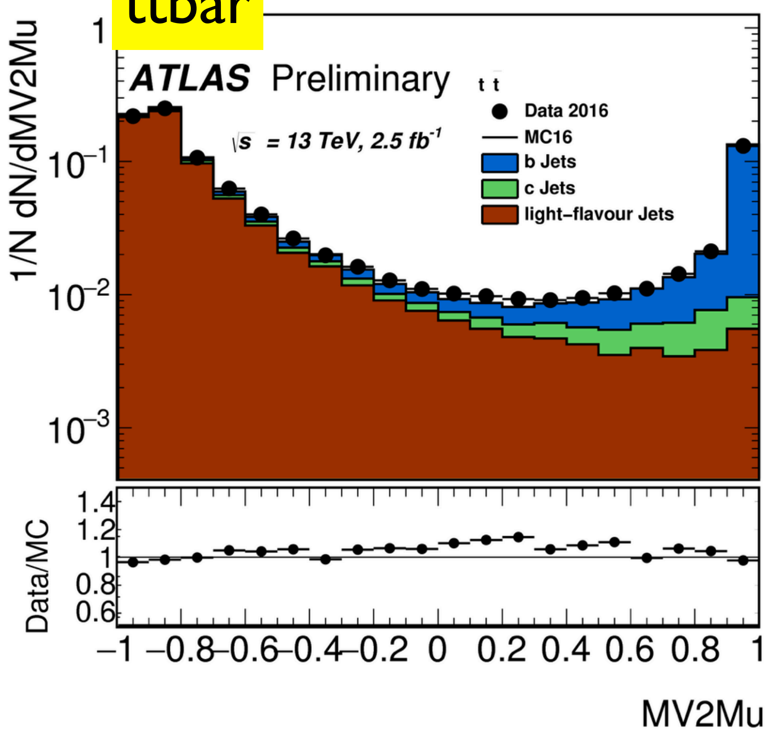




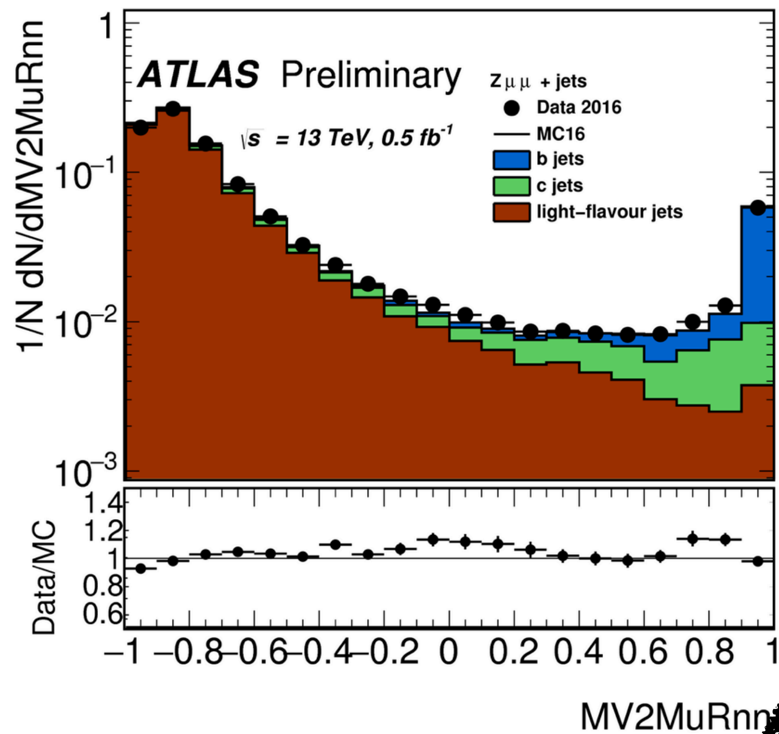


Algorithm performance

ttbar



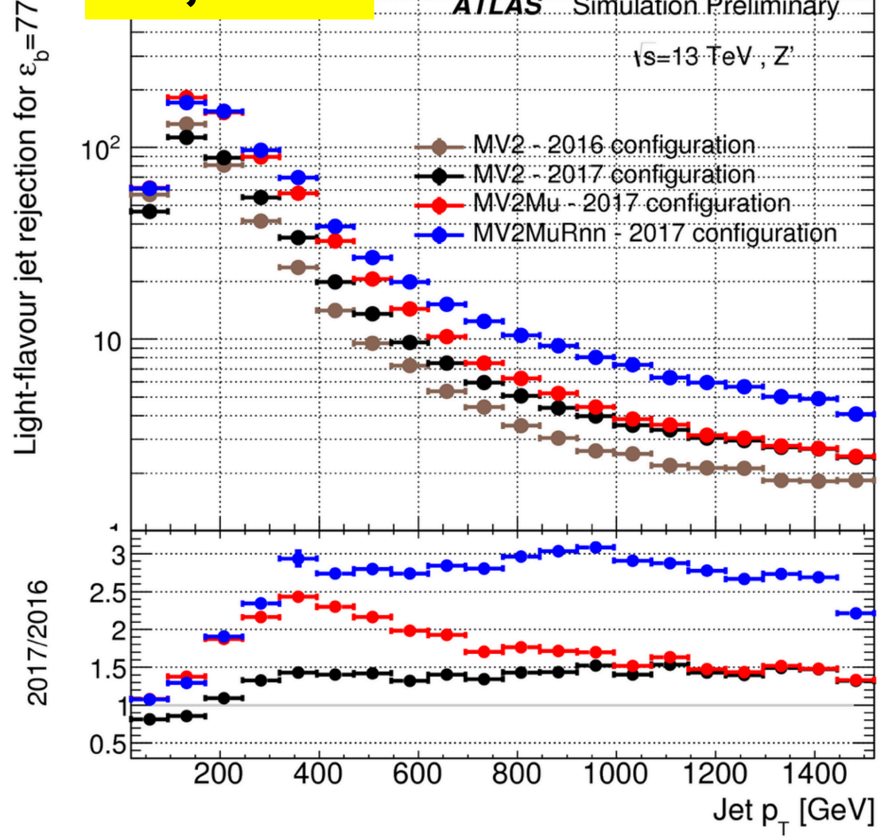
Zμμ+jets



✓ Three variants of the MV2 (and DL1) algorithms have been deployed

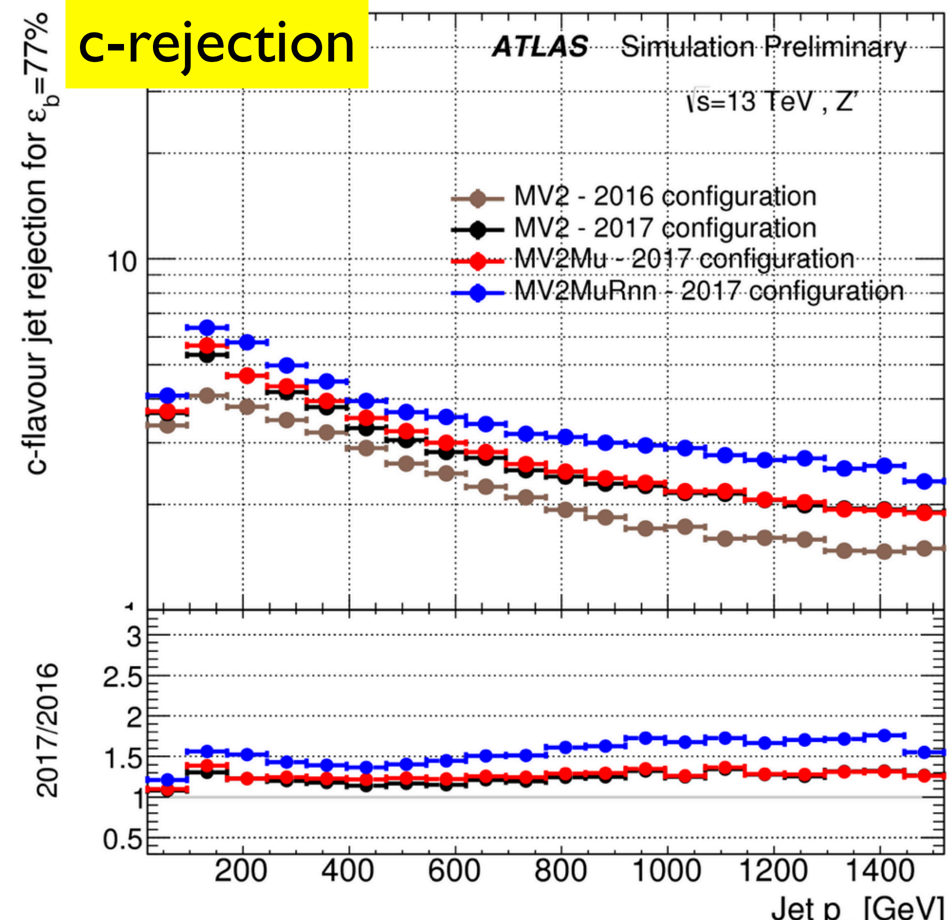
- standard impact parameter and secondary vertex-based inputs + kinematics of the jet (MV2/DL1)
- standard inputs + soft muon tagger (MV2Mu/DL1Mu)
- standard inputs + soft muon tagger+ RNNIP (MV2MuRnn/ DL1MuRnn)

l-rejection



Improvements on the full pt spectrum from various low-level tagger contributions (SMT at low-medium pt, RNN at high pt)

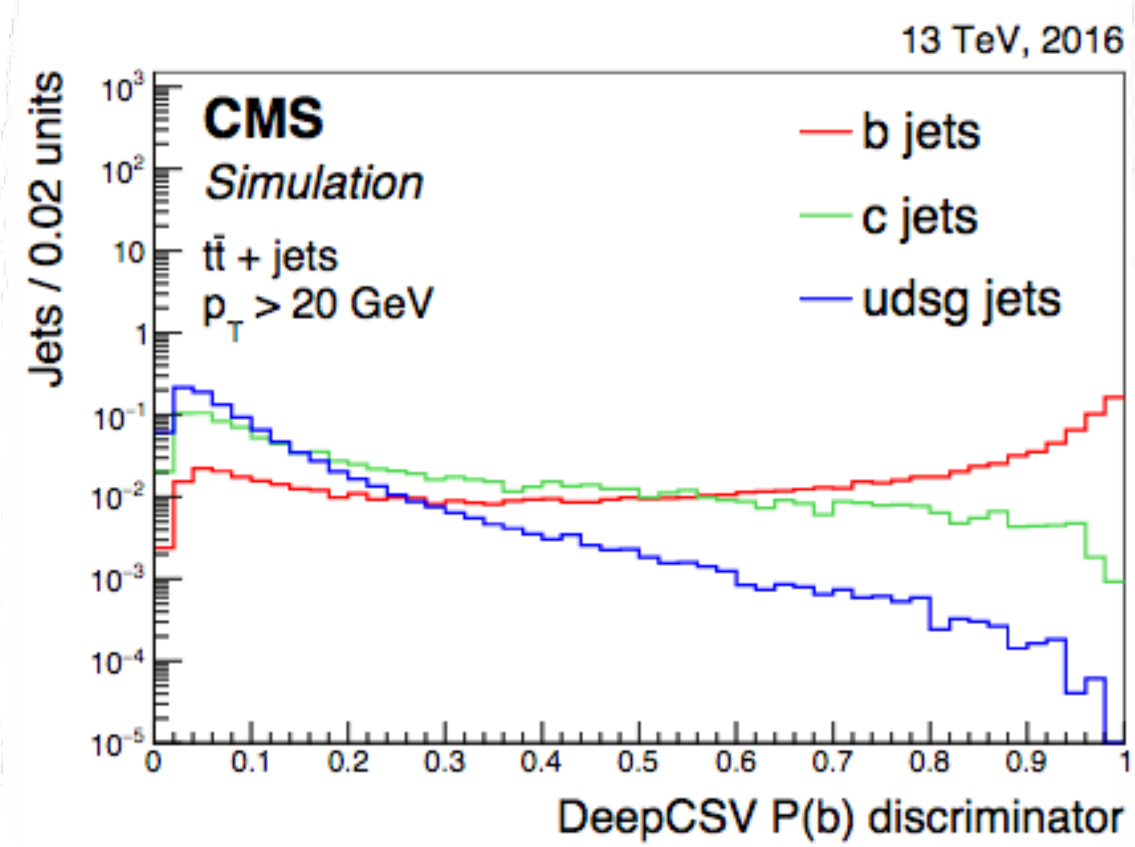
c-rejection



CSVv2 and DeepCSV

- ✓ Evolution of Run I Combined Secondary Vertex (CSV) algorithm for b- and c-tagging (CSVv2)
  - exploits kinematics of b/c-hadron decays and full decay chain
- ✓ Deep neural network techniques used to enhance the discrimination and to better exploit the phase-space
  - useful for multi-class training and classification and for the understanding of low-level information (hit pattern, tracks,...)
  - DeepCSV makes use of deep neural network (DNN) technology and is found to over-perform CSVv2
    - DNN architecture based on hidden layers and output nodes
    - additional tracks (first 6 most displaced tracks instead of 4 as for CSVv2) used in the algorithm with same track selection

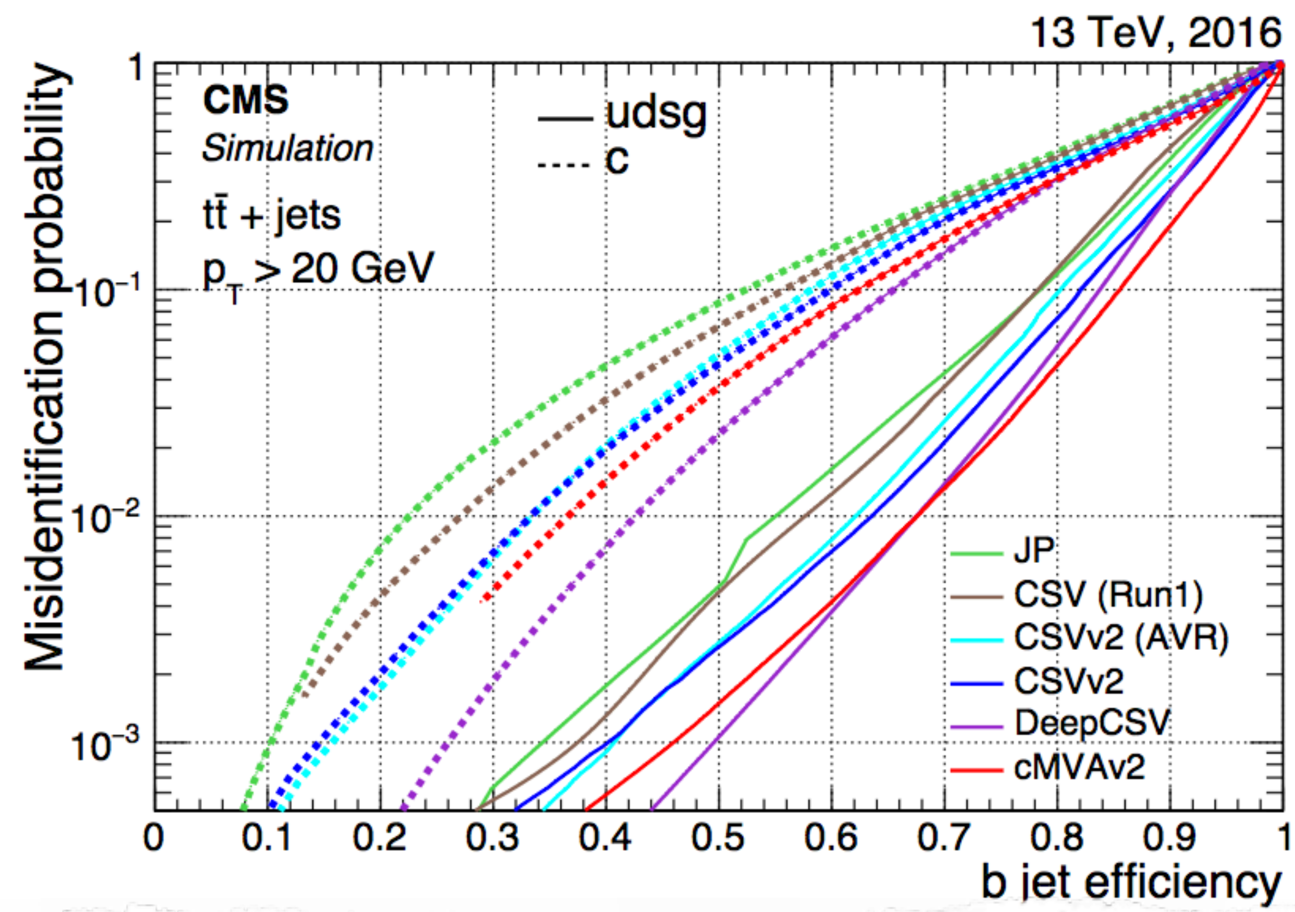
- ✓ Algorithm returns an output probability for the classes of jets employed in the training
  - jet with one or at least two b-quarks
  - jet with one or at least two c-quarks
  - all the rest

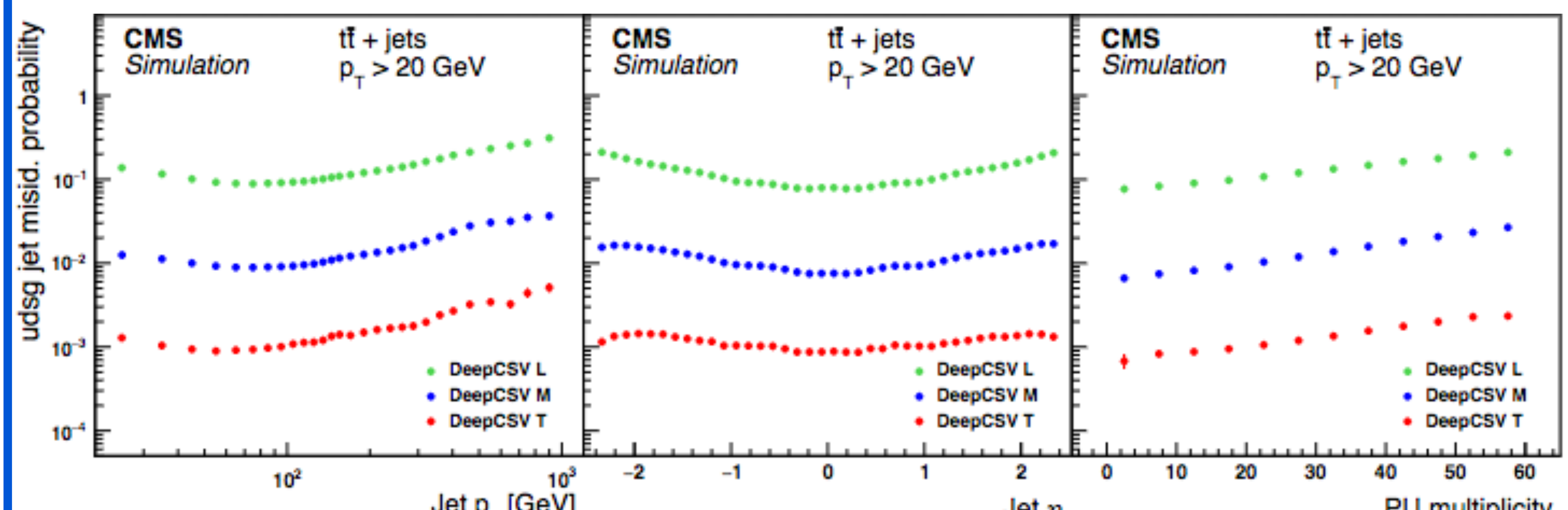


CSVv2 and DeepCSV (2)

✓ Large improvement in performance wrt CSVv2 algorithm

- same data/Monte Carlo agreement exhibited by CSVv2
- working points defined in order to fix the mis-identification rate of light jets to 10, 1 and 0.1%

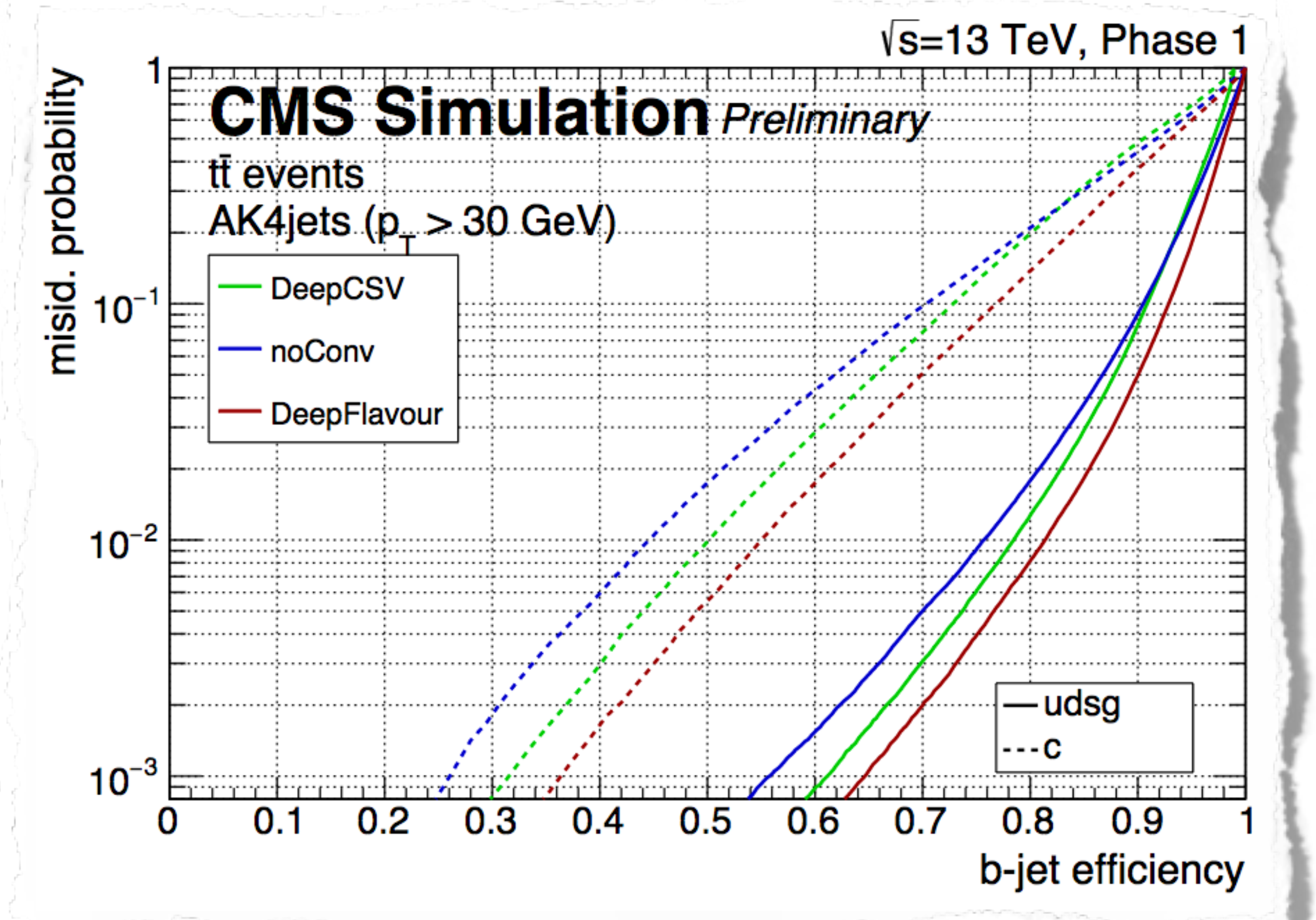


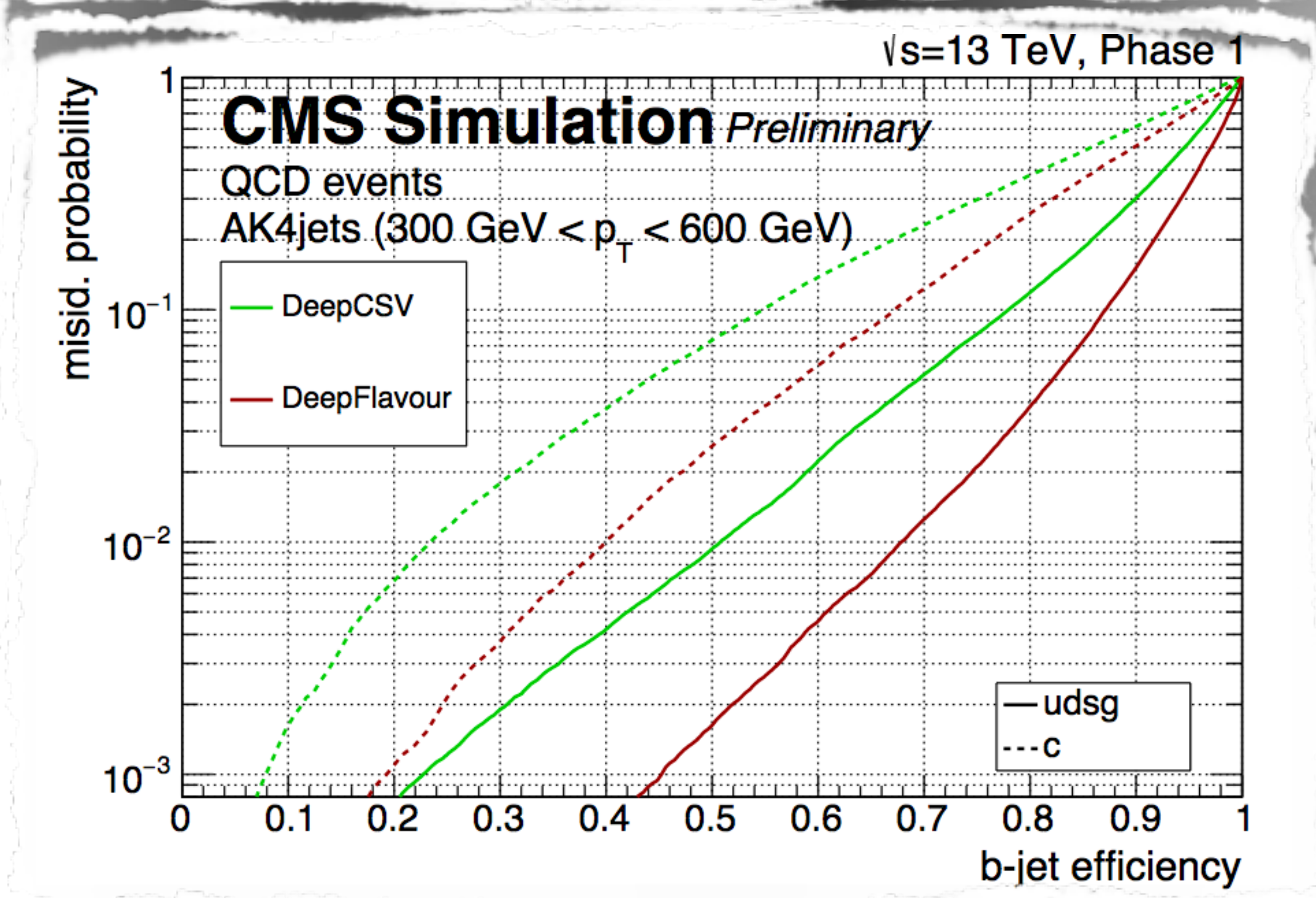


DeepFlavour Tagger

✓ Going further in exploiting DNN techniques → DeepFlavour Tagger

- very inclusive set of input tracks (no quality requirements applied)
- using properties of jet constituents and topological features of the reconstructed SV
- added a convolutional layer
- four output nodes for b, bb, c and light
- overall non-negligible improvement wrt DeepCSV (5% at 0.1% mistag rate)
- DeepFlavour expected to recover the performance loss at high jet momentum

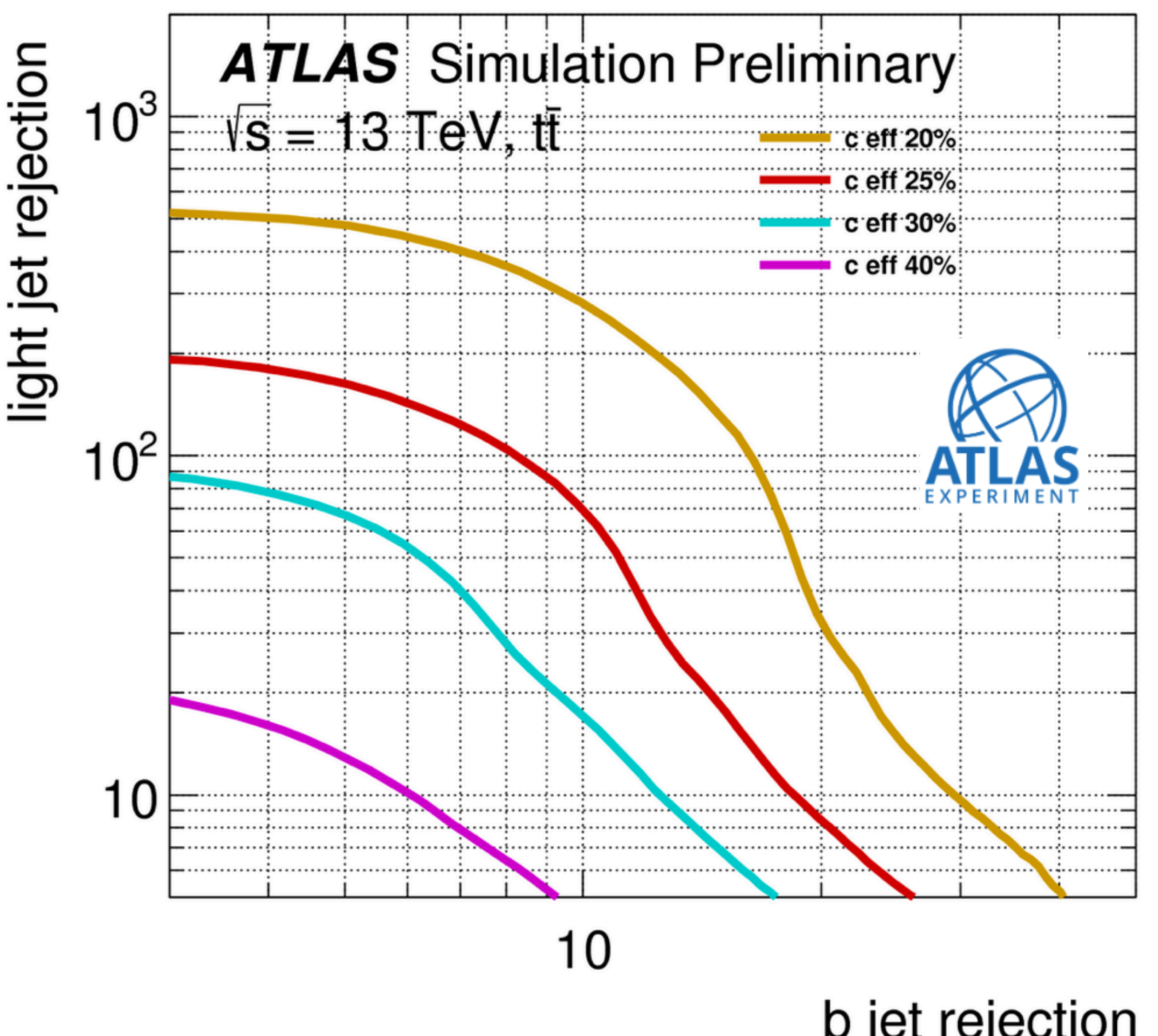




Identification of c-jets

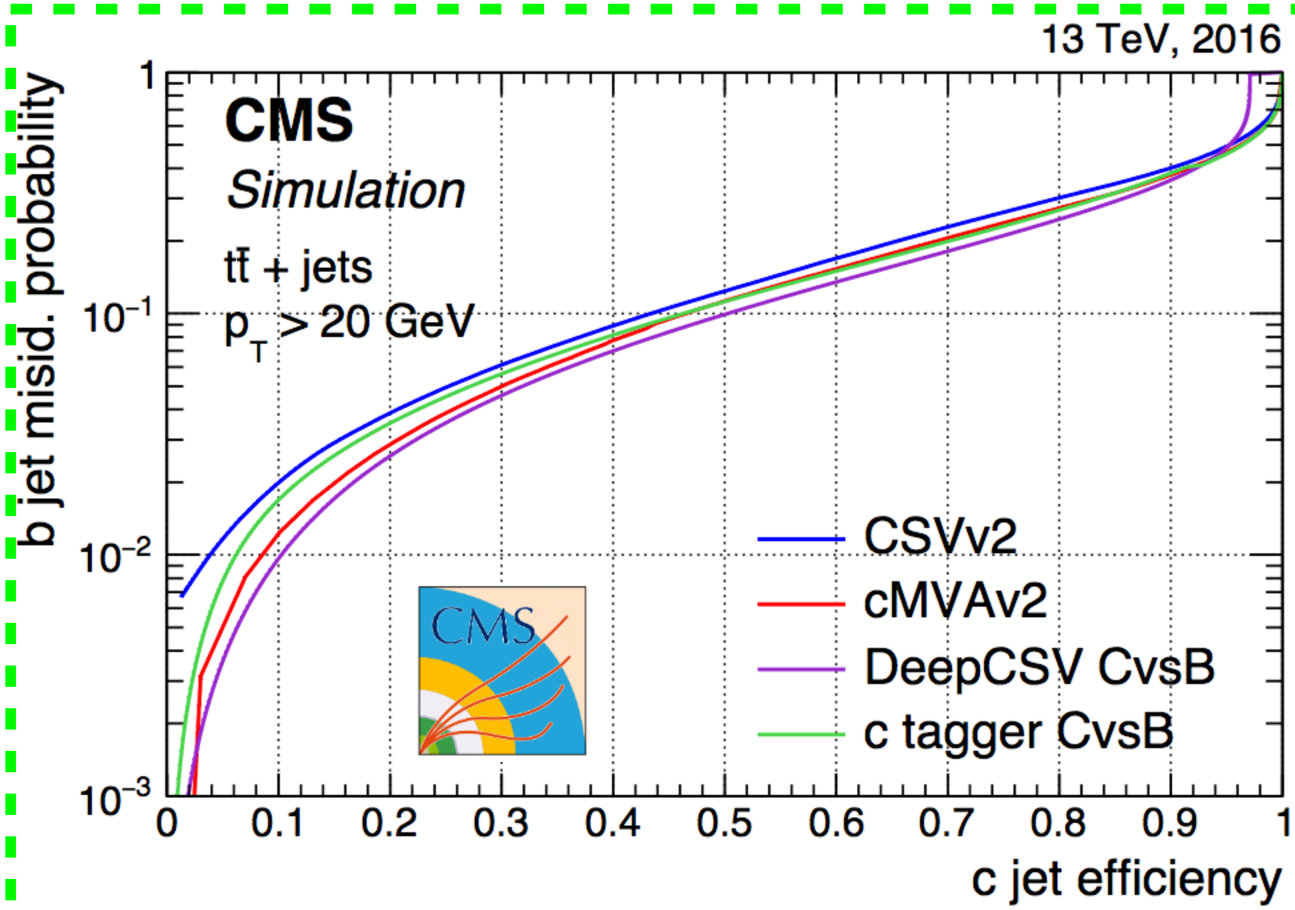
✓ Topology of the displaced vertex reconstructed by the JetFitter algorithm in addition to b-tagging inputs used in a dedicated BDT for c-tagging

- MV2c100 (b discrimination), MV2cl100 (light-flavour discrimination)
- developments for DL-based c-taggers also ongoing



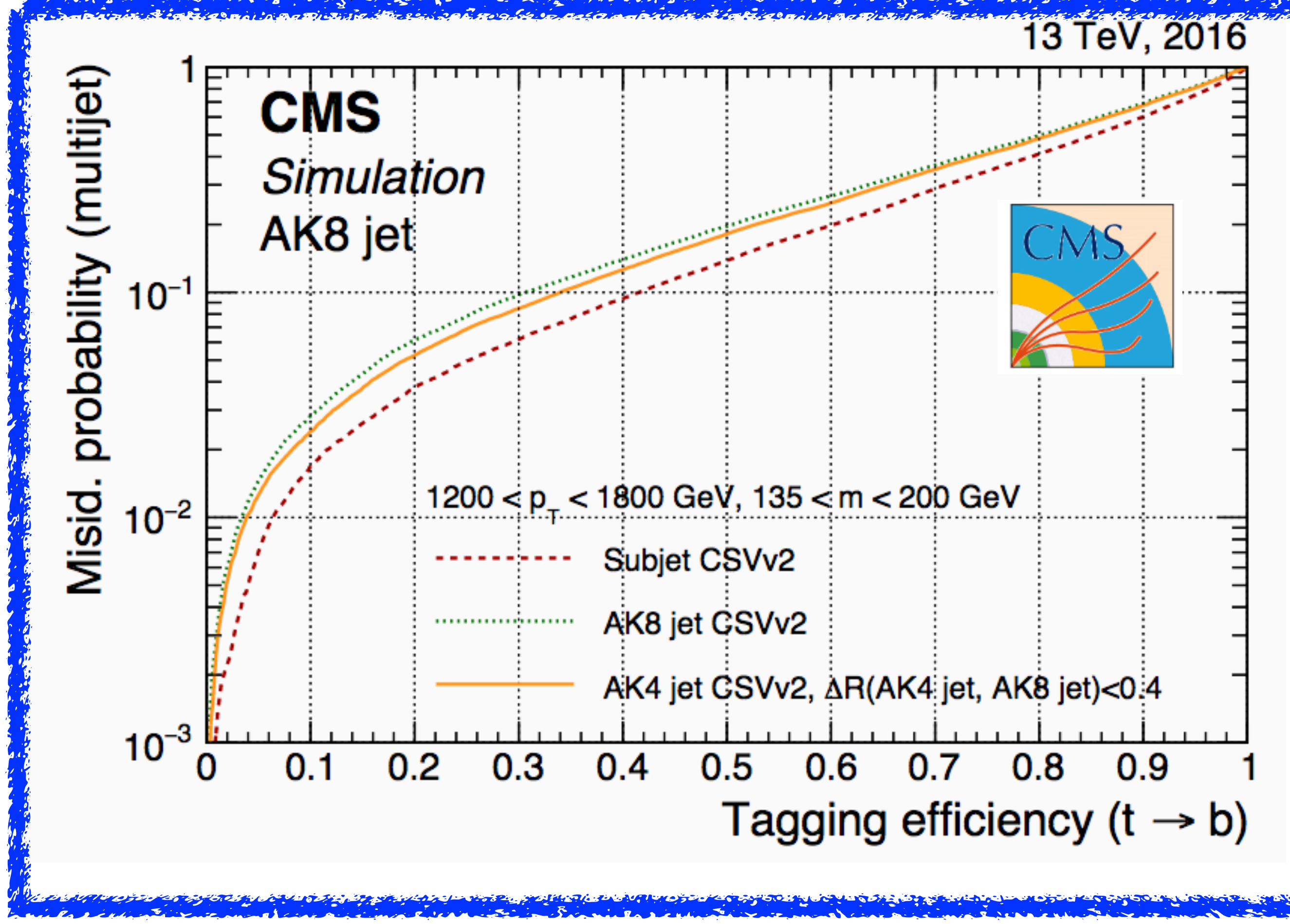
✓ c-tagging identification based on CSVv2

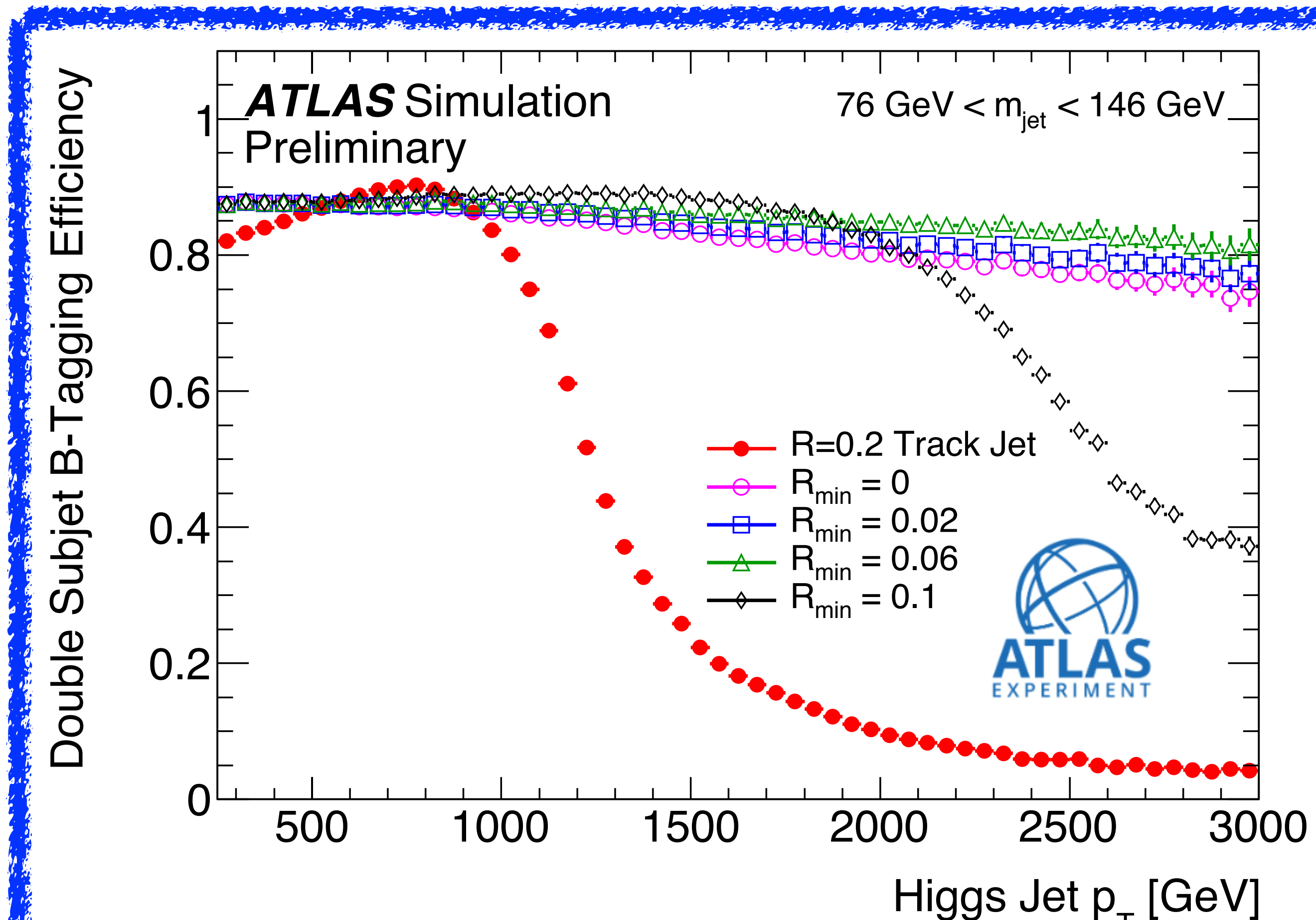
- similar inputs as in b-tagging + additional kinematics of the soft-lepton taggers
- discrimination exploited for c- vs light-flavour and c- vs b-jets (using Gradient Boosting Classifier)
- focus on DNN-based c-tagging response, i.e. DeepCSV - outperforming dedicated CSVv2 algorithm



b-tagging for boosted topologies

- ✓ b-quarks could be present in decays of boosted particles (relevant for BSM scenarios)
  - decay products clustered in a single fat (large-R, R=1.0/0.8) jet
  - usage of substructure techniques to reconstruct sub-jets and apply b-tagging
    - boosted H→bb , g→bb
- ✓ Dedicated effort for X→bb tagging in ATLAS and CMS
  - improves discrimination of boosted H→bb against boosted SM g→bb
  - uses variable-radius track-jets (ATLAS) to account for the boost of the parent particle (clustering radius as a function of pt)
  - other approaches also being investigated in ATLAS (exclusive-k_T-tagging and center-of-mass subjet reconstruction)
  - exploits the presence of the two b's in the jets and the correlation between their flight directions (CMS)



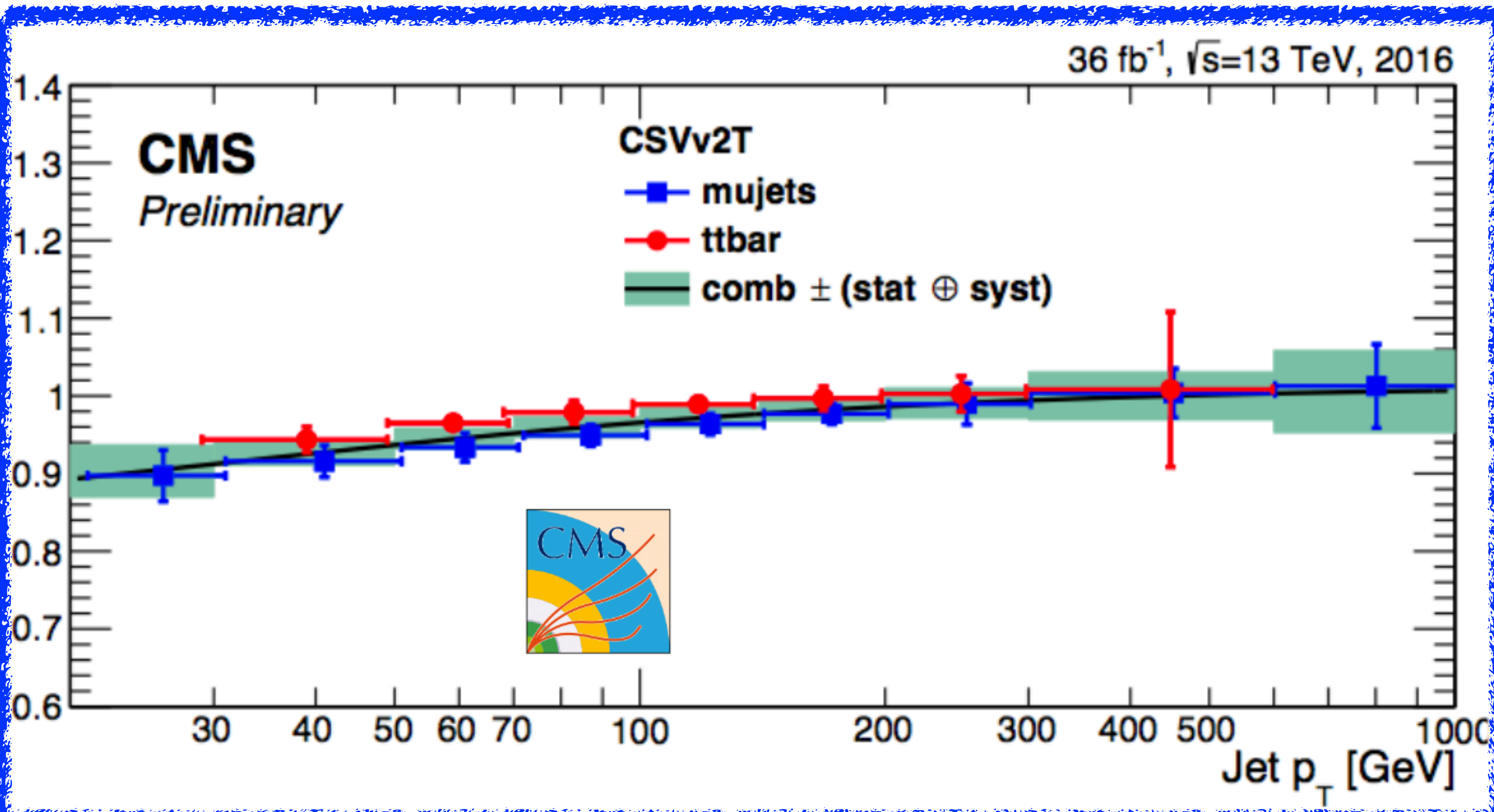


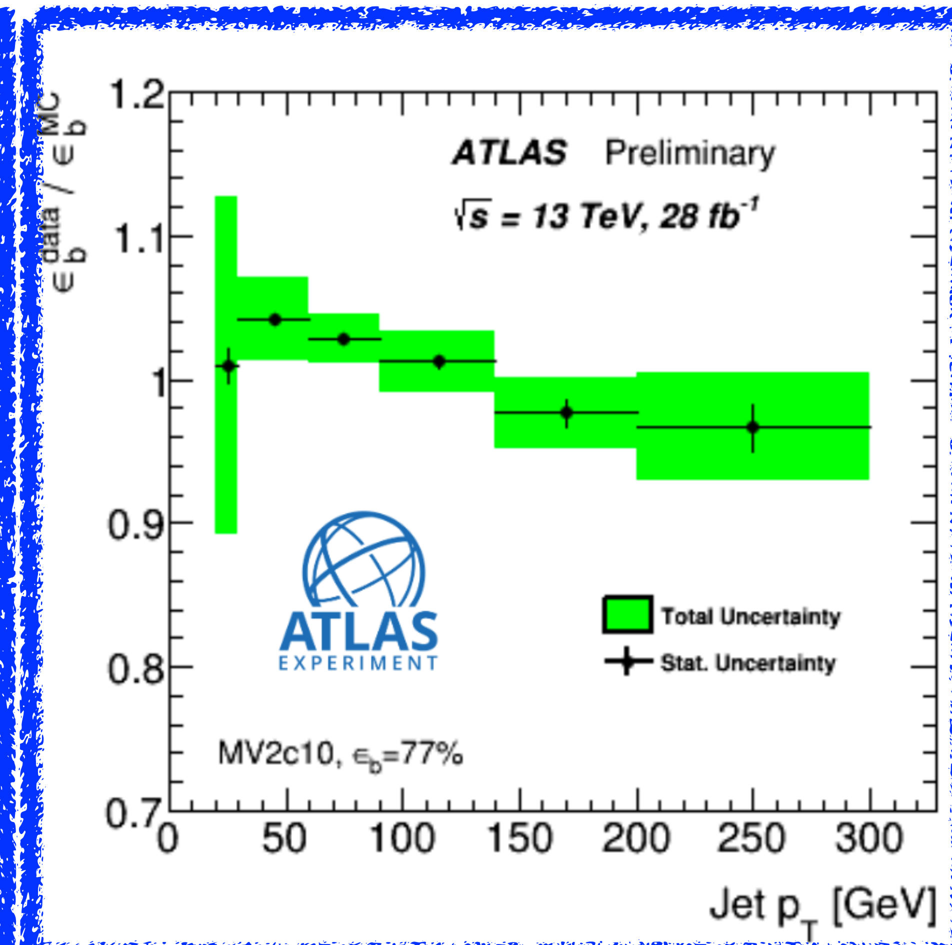
b-tagging performance in data (calibrations)

Efficiency measurements for b-jets

✓ Sample of true b-jets to extract scale factors as efficiencies in data and simulation

- ttbar dileptonic channel with opposite-sign eμ+jets to reduce Z→ll + jets background
- kinematic fits in data (using likelihood fit), additional method using tag and probe (ATLAS) with ttbar semileptonic and dilepton analysis and information on muon in jets (CMS)
- combination of various calibration methods ensures best precision
- MC-based high-pt extrapolation adopted in ATLAS while dedicated sample in CMS
- systematics uncertainties are small and mostly dominated by tt and HF modeling

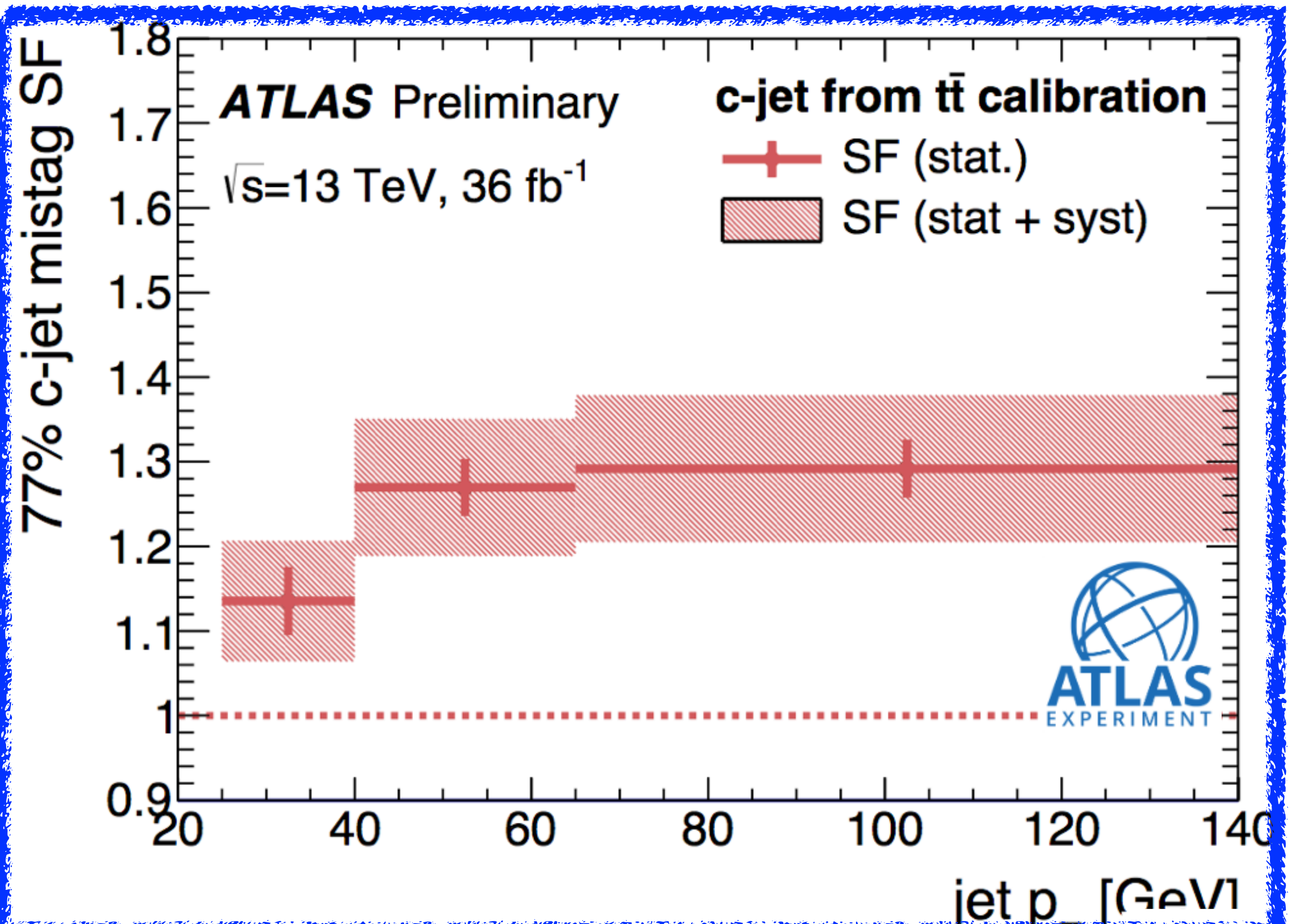


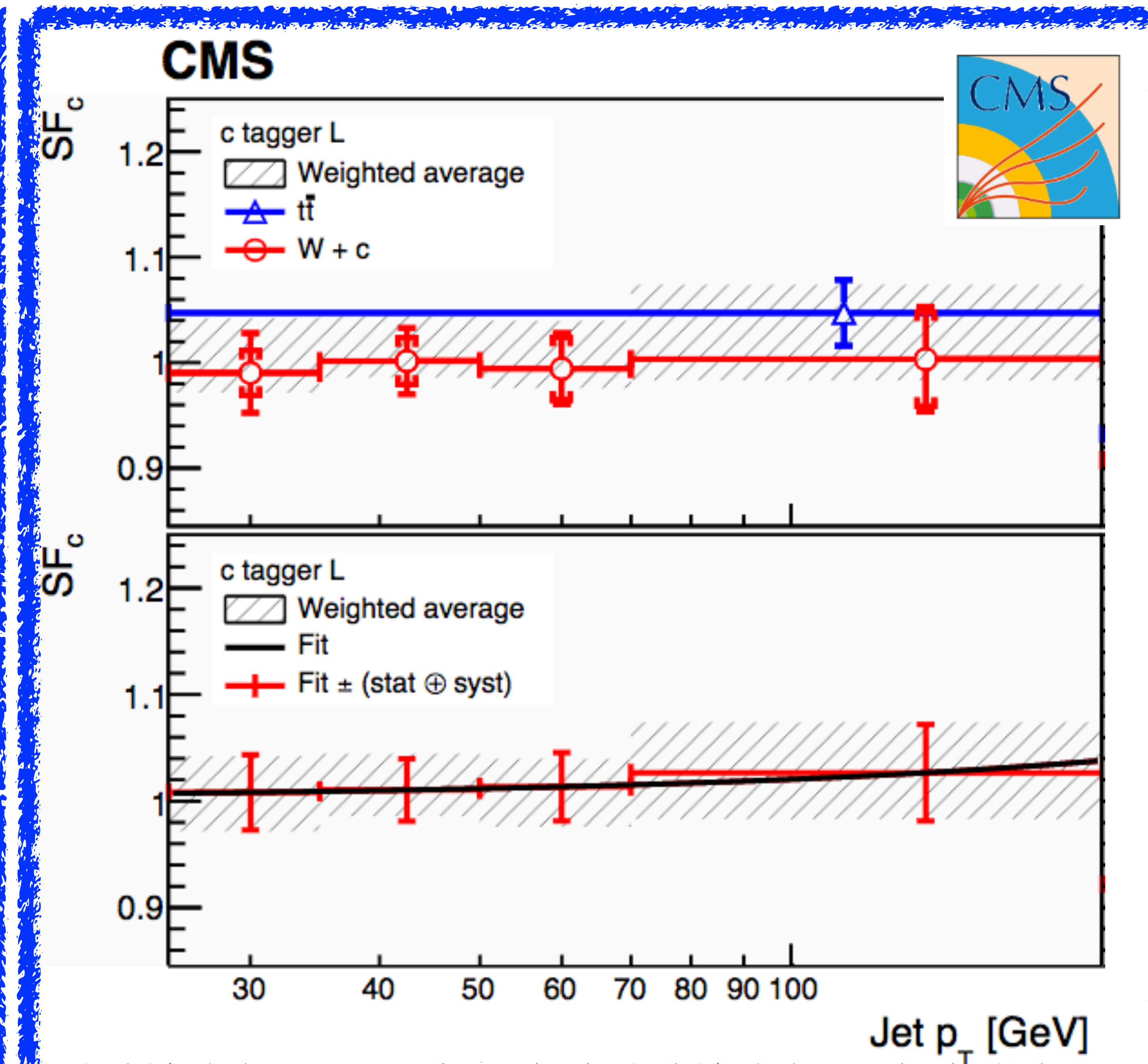


Efficiency measurements for c-jets (fake-rate)

✓ Sample of true c-jets before and after tagging to extract scale factors

- ATLAS and CMS use two c-enriched topologies, ttbar events in single lepton final state (W→lν, W→cs) and W+c selected by searching for a soft muon (W→μν) in the c-jets
- cut-and-count analysis in W+c (ATLAS) and fit to discriminant to extract c-tagging efficiency in ttbar (ATLAS and CMS)
- uncertainties dominated by ttbar modeling (10-20% depending on pt)

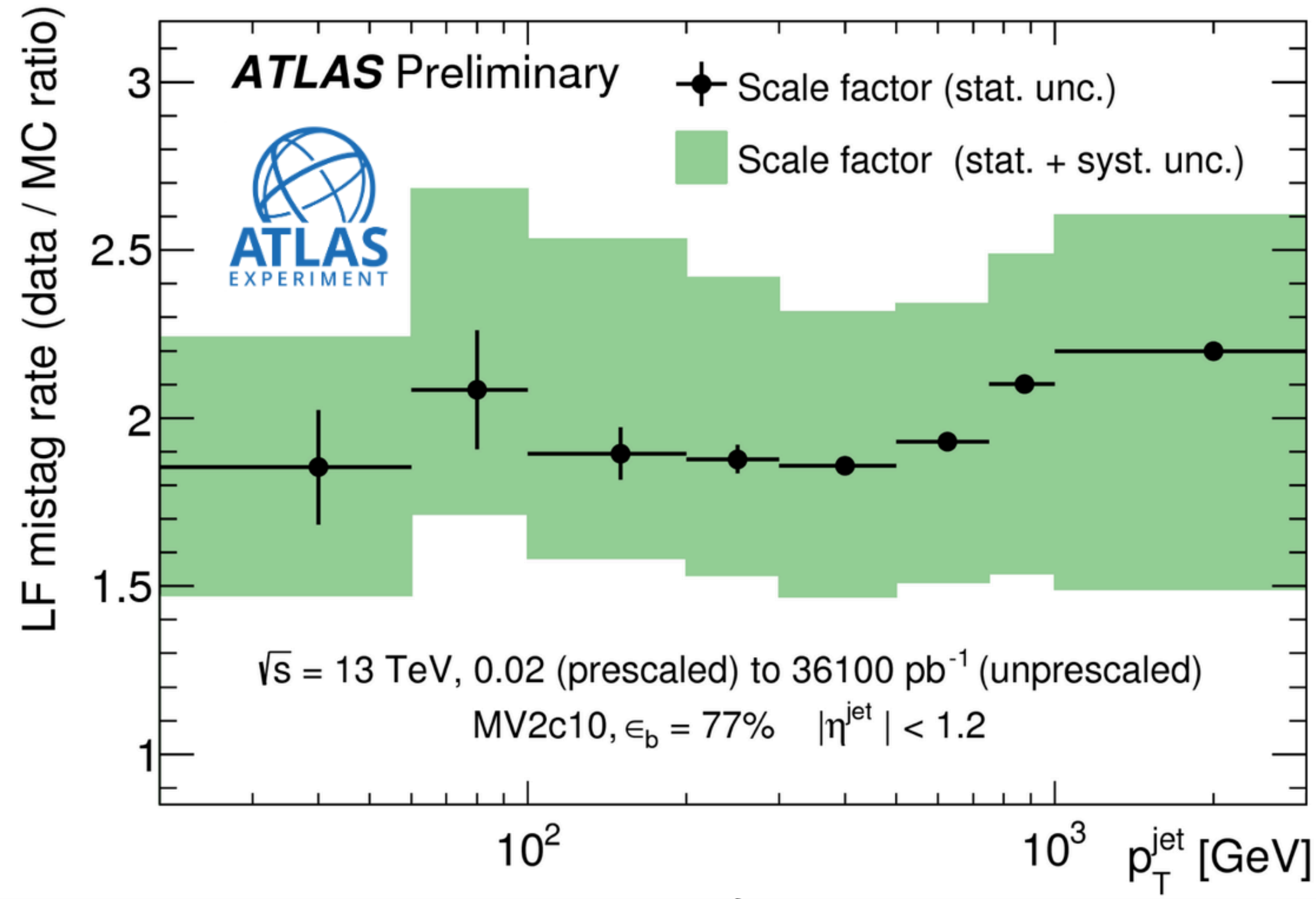


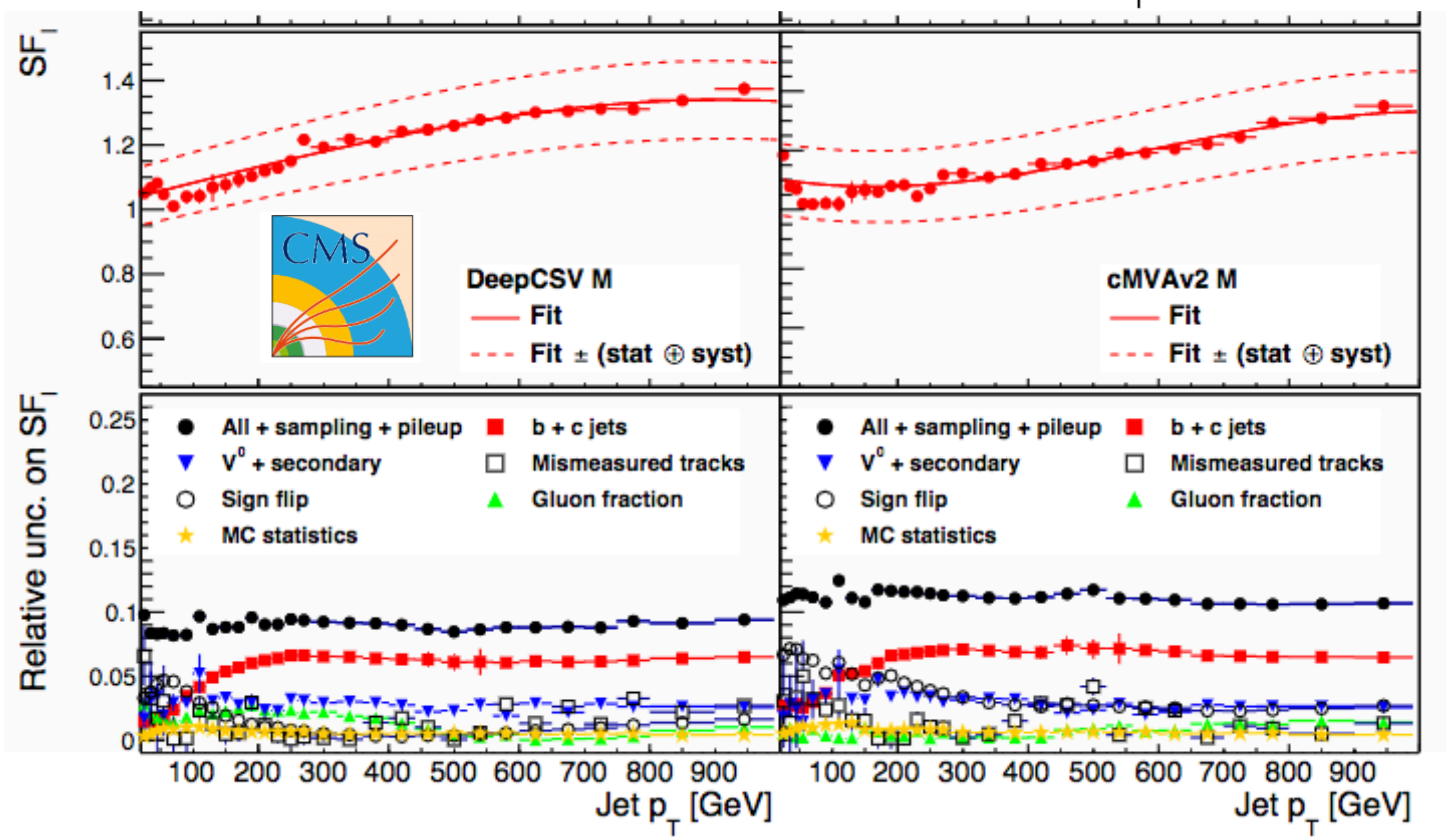


Efficiency measurements for light-jets (fake-rate)

✓ Sample of true light-flavour jets needed for this calibration

- flipped-taggers to calibrate fake-jets generated from track-resolution effects
- flipped taggers exhibit similar mistag rate for light-jets and much smaller discrimination power for b/c-tagging → light-enriched sample posttag
- large uncertainties (20-40%) concerning flipped tagging performance
- additional adjusted-Monte Carlo method (ATLAS) where data-driven tracking performance are propagated to the extraction of the mistag rate





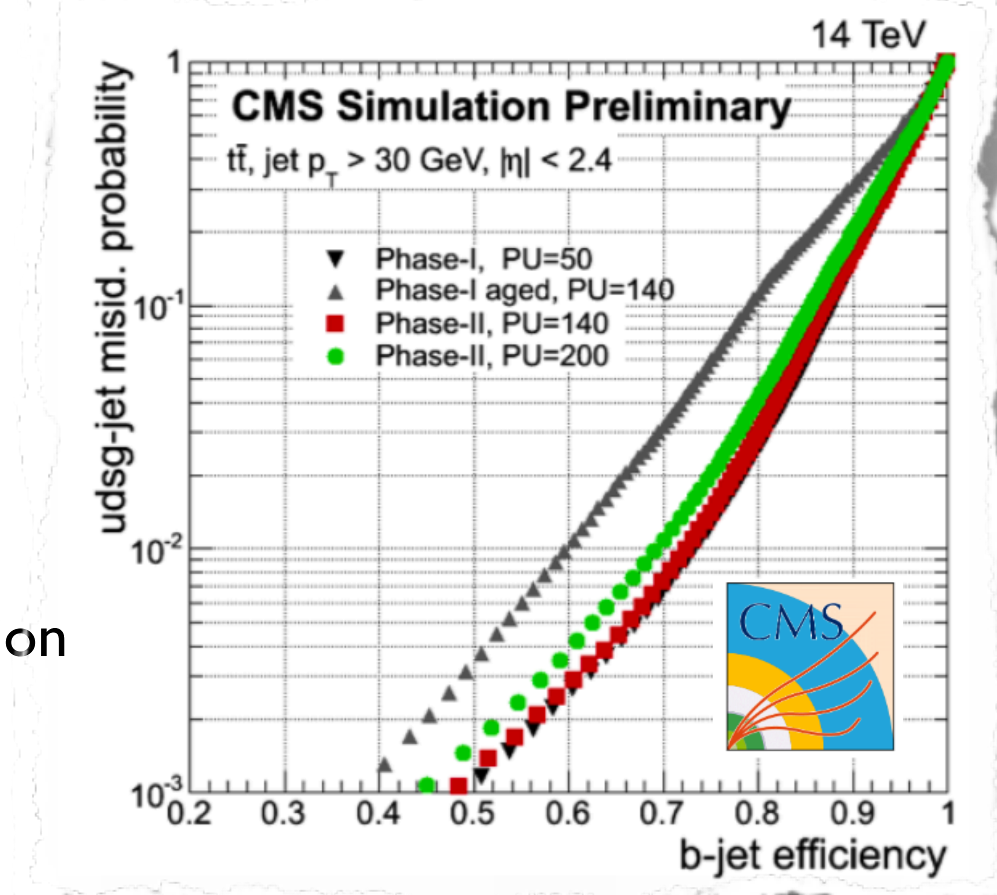
b-tagging upgrade studies

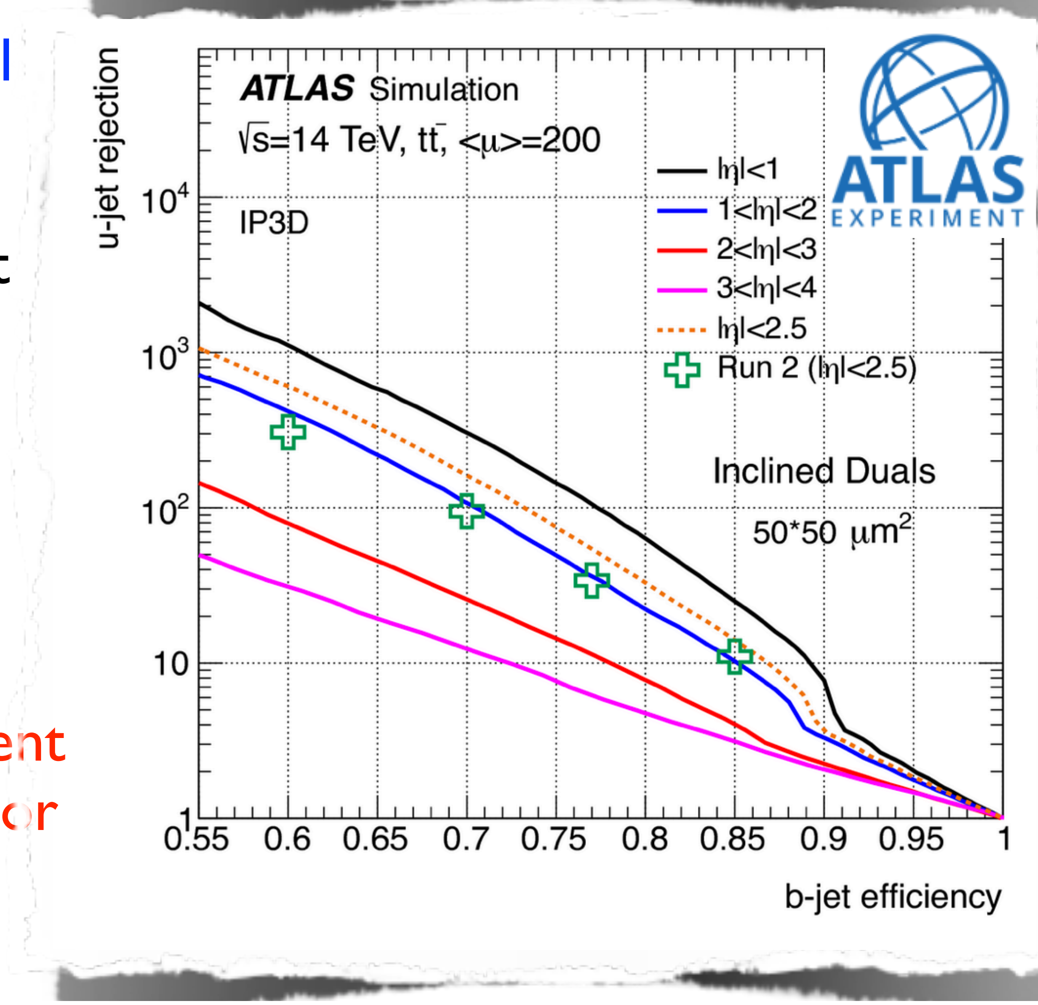
b-tagging for High-Luminosity Upgrades @ ATLAS&CMS

✓ Major upgrades of the ATLAS and CMS inner detectors to operate in the harsh High-Luminosity LHC environment

- extended tracking coverage to |η|=4, replacement of new detector (pixel/strip) with higher granularity
- high level of pile-up (<μ>=200) is a challenge for robust particle reconstruction/identification and pattern recognition
- b-tagging needs to account for updated inner detector geometry layout
- optimized b-tagging documented in ATLAS and CMS Pixel and Strip Technical Design Reports
- Re-definition of hit-motivated track categories for Impact Parameter-based taggers (ATLAS) enables to fully characterize the forward η region
- MV2c10 tuning re-optimized to account for geometry modifications in the ID

➡ b-tagging algorithms can cope with harsh HL-LHC environment and provide excellent background rejection in various detector regions





Conclusions

✓ b-tagging is a crucial tool for measurement and searches at the LHC

➡ Overview of algorithm developments in ATLAS and CMS for b-/c-jet identification, boosted topologies and upgrade studies presented together with a quick look on calibration techniques for b-jet efficiency, c- and light-flavour jet fake-rate

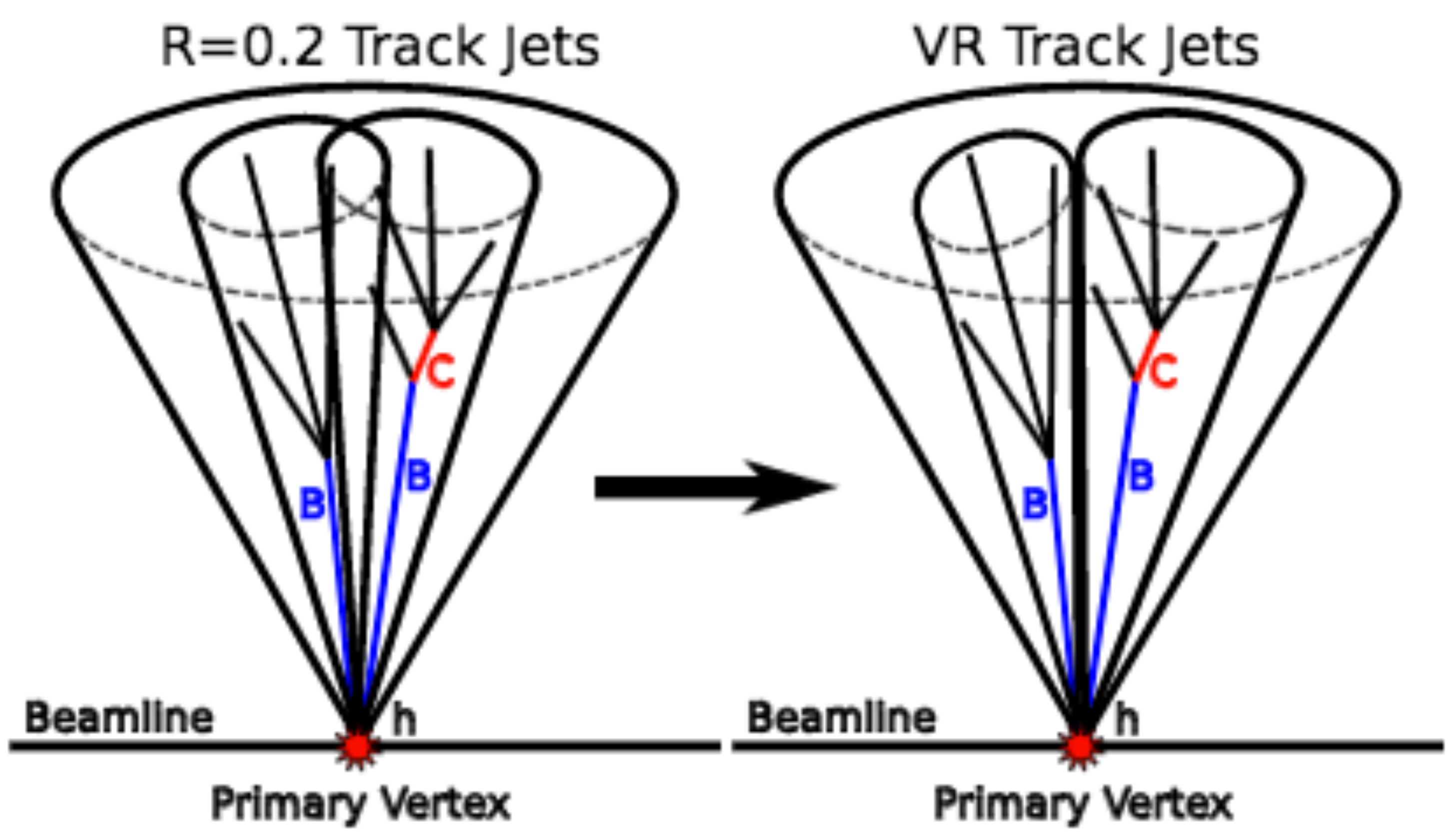
✓ High-level taggers strongly rely on inputs from low-level algorithms exploiting the kinematics of the b-jets to ensure separation against c-/light-flavour-jet backgrounds

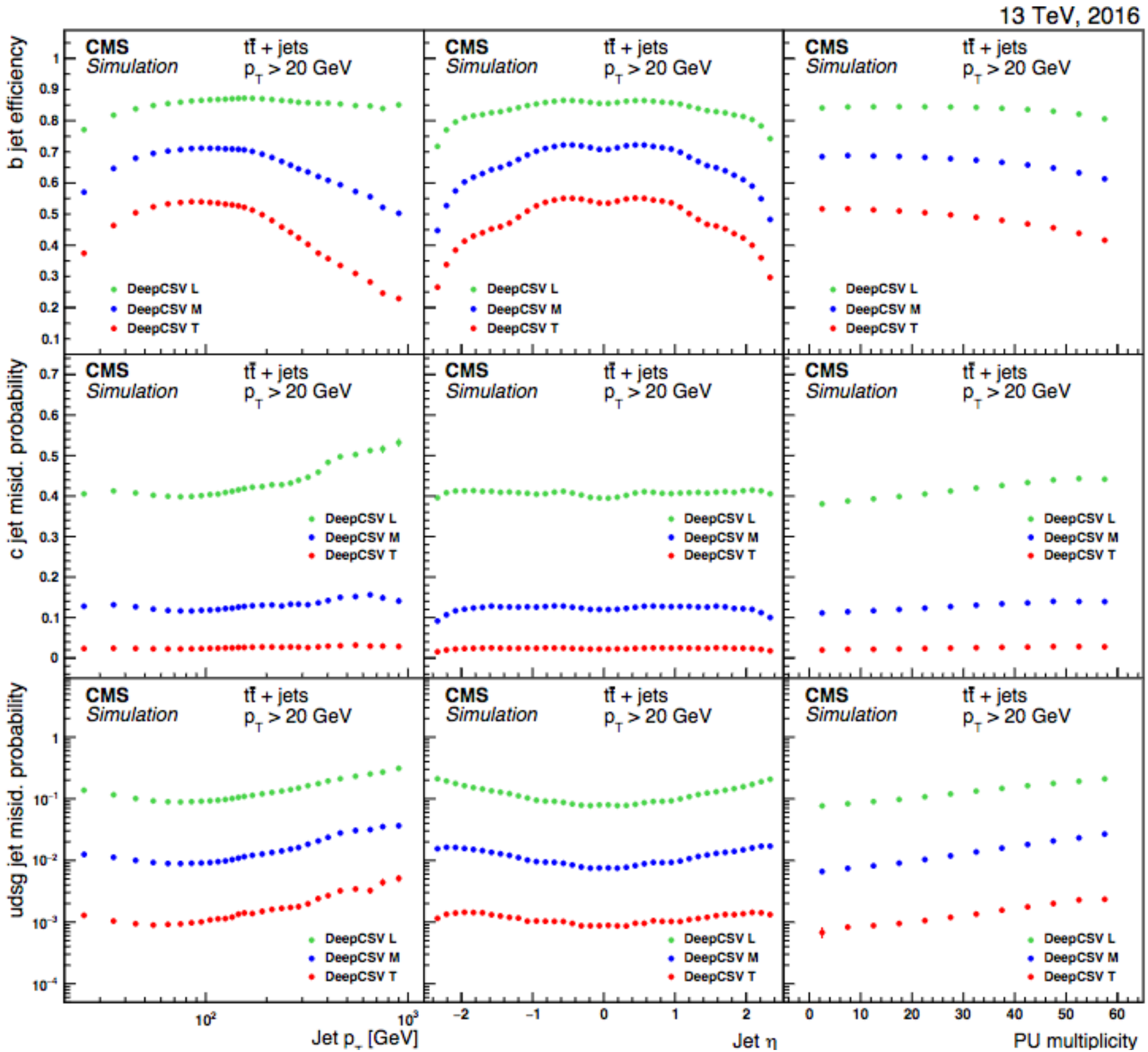
- deep understanding of jet topology and push for machine learning techniques has significantly improved the overall performance
- excellent level of background discrimination has definitely paid off → useful for physics measurements featuring b-/c-jets in the final state
- significant work on the upgrade side has allowed to achieve excellent discrimination power also in the harsh HL-LHC running conditions (average pile-up of ~200, extended tracking coverage, new geometry layout)

✓ Calibration of b-tagging algorithms is also an essential ingredient of the b-tagging chain

- similar approaches in ATLAS and CMS on how to tackle the extraction of the data/MC scale factors for b-, c- and light-flavour jets
- still some challenges ahead, i.e. nature of light-flavour fake-rates (resolution effects), high-pt extrapolation for b-jets,...

Additional slides

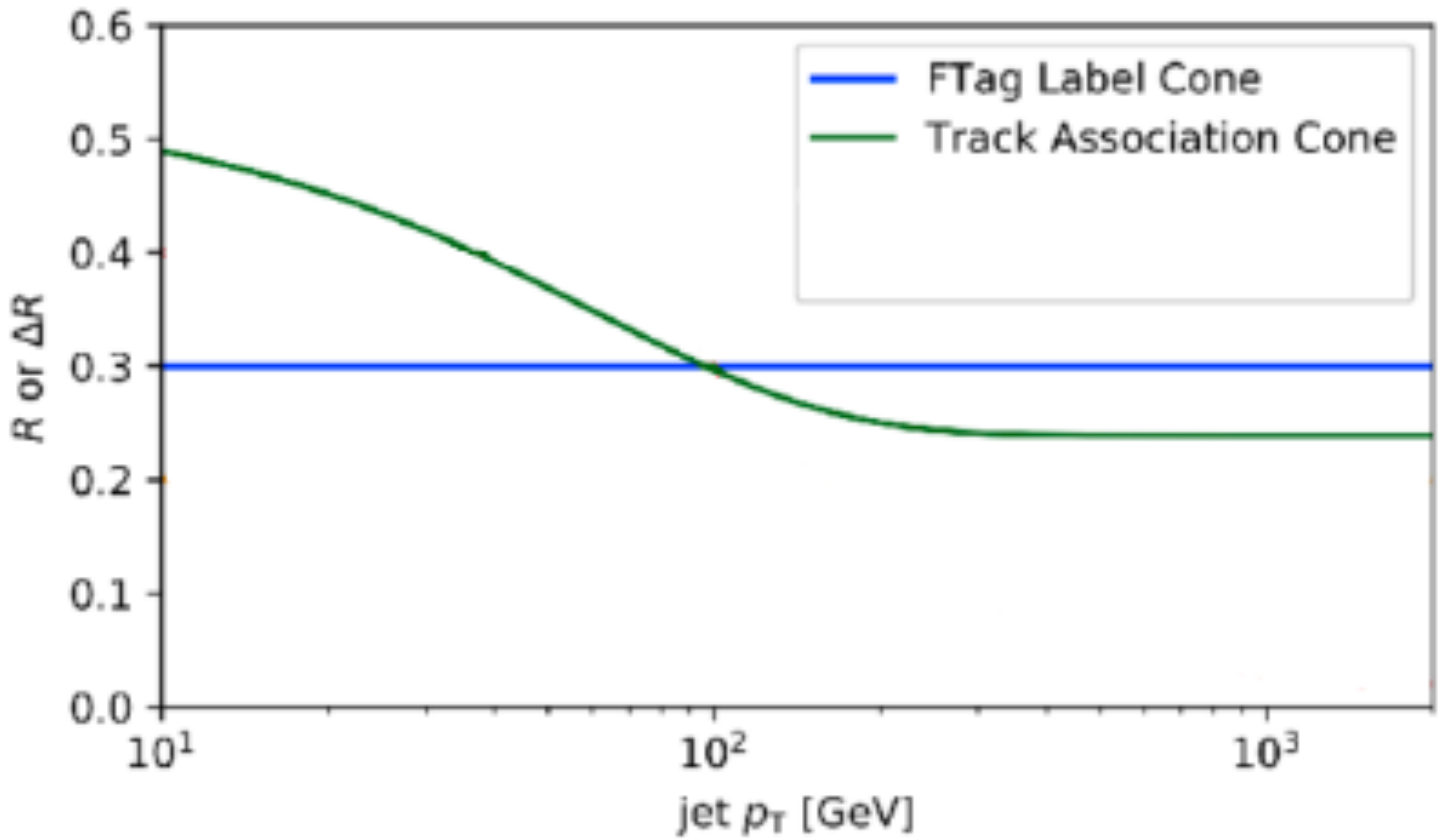

R=0.2 Track Jets
VR Track Jets
B
B
C
h
Beamline
Primary Vertex
B
B
C
h
Beamline
Primary Vertex



13 TeV, 2016
CMS
Simulation
$t\bar{t}$ + jets
p_T > 20 GeV
DeepCSV L
DeepCSV M
DeepCSV T
b jet efficiency
c jet misid. probability
udsg jet misid. probability
Jet p_T [GeV]
Jet η
PU multiplicity

Jet-to-track association in ATLAS

✓ Shrinking cone around jets to account for kinematics and group all tracks associated to the jet



ATLAS Pixel Technical Design Report

➡ Pixel Technical Design report completed and submitted to LHCC on December 21st

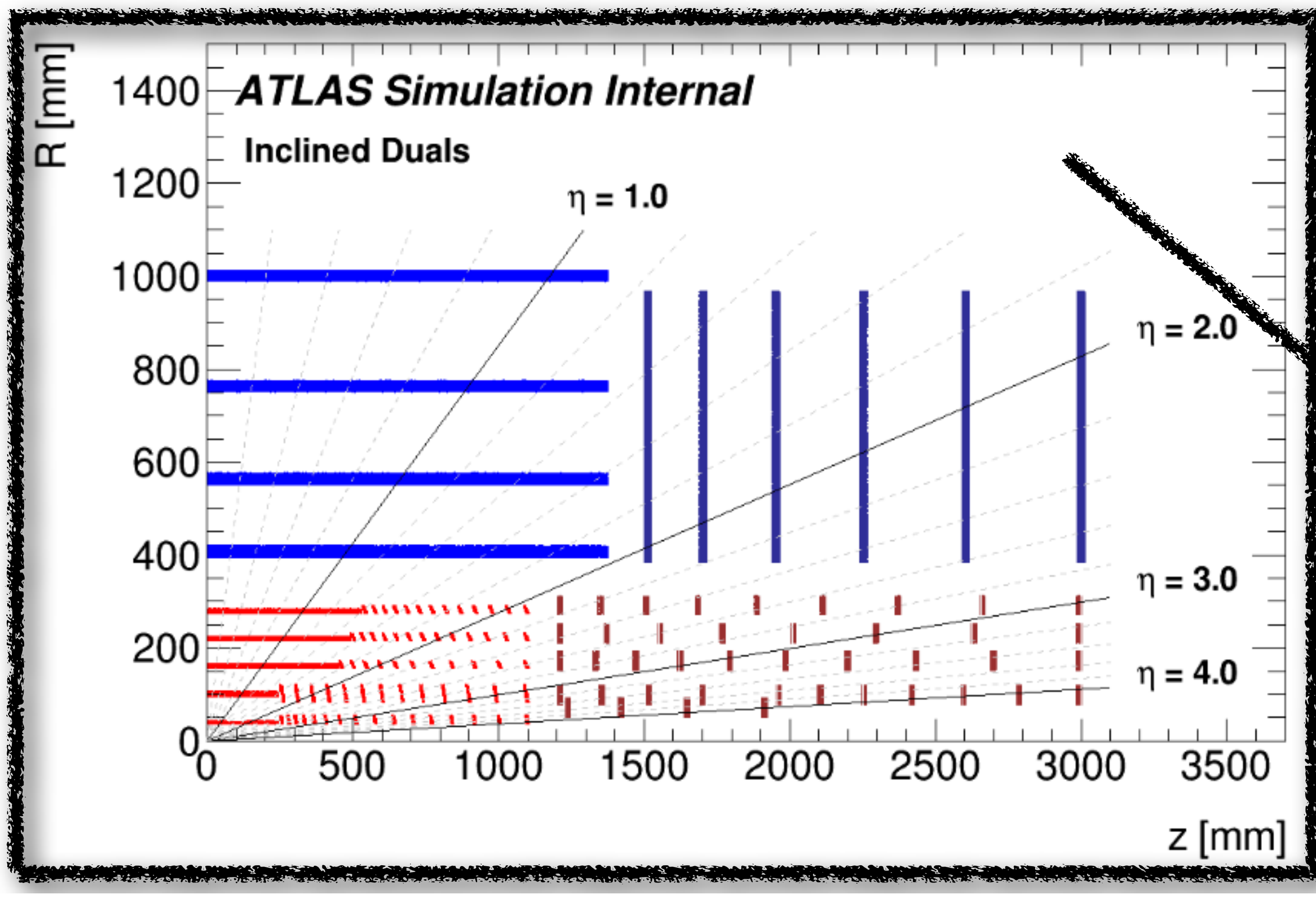
✔ performance studies on b-tagging at HL-LHC + prospects for physics using b-tagging in Chapter 3 (Tracking and Physics Performance)

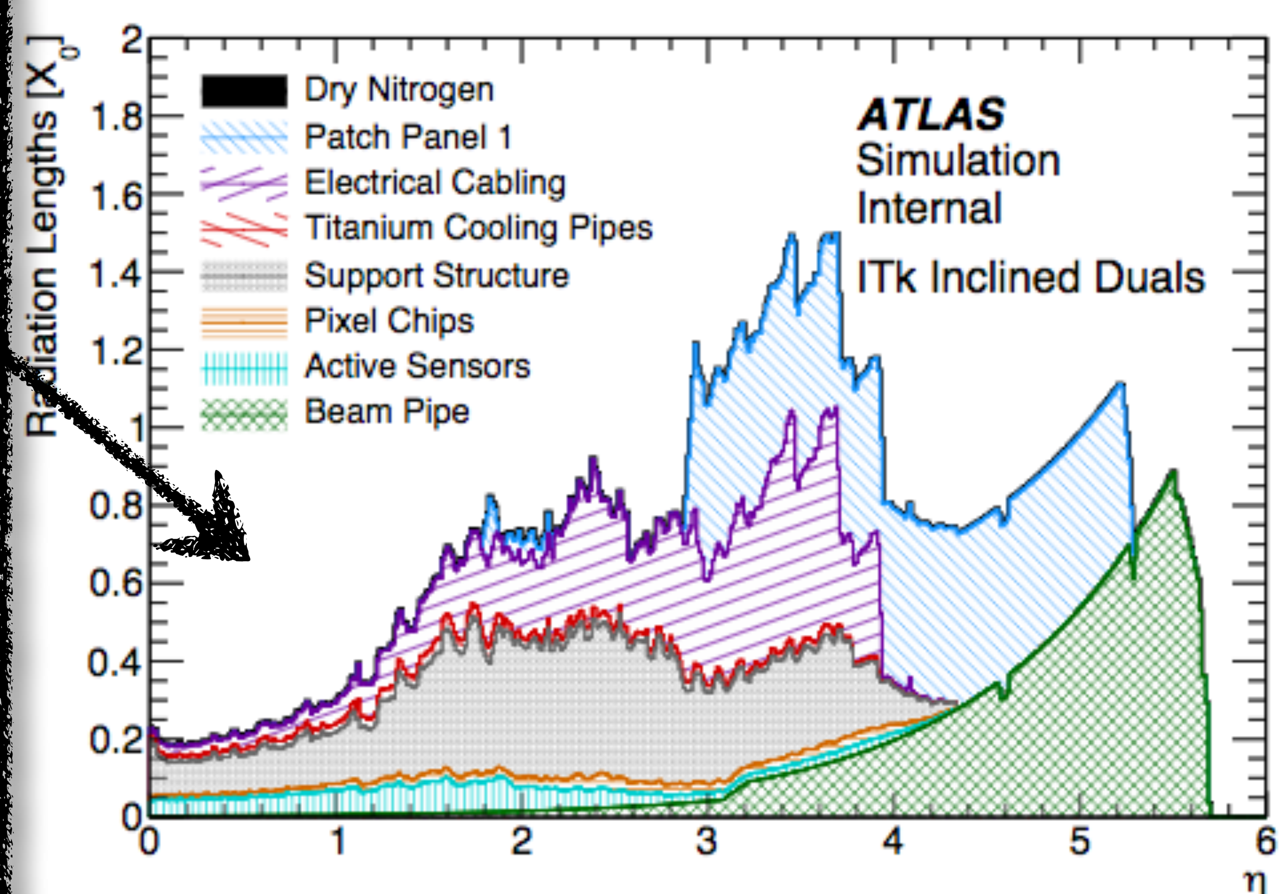
✔ Results using InclinedDuals (Step 2.2 layout) with extended tracking to |η|=4 at μ=200 (digital clustering for 50x50 μm^2 pitch)

ATLAS EXPERIMENT

▶ https://cds.cern.ch/record/2296611/

Technical Design Report
for the ATLAS Inner Tracker Pixel Detector





c-tagging

- Discrimination of c from b/light is very important for several physics studies
- Discrimination exploited by the topology and the kinematics of the displaced vertex reconstructed JetFitter - two taggers provided, MV2c100 (b/c discrimination), MV2cl100 (b/l discrimination)

Variable Name	Description
L_{xyz}	Three-dimensional displacement of secondary vertex from the primary vertex
L_{xy}	Transverse displacement of the secondary vertex
Y_{trk}^{min},Y_{trk}^{max},Y_{trk}^{avg}	Min, Max and Avg. track rapidity of tracks in jet
Y_{trk}^{min},Y_{trk}^{max},Y_{trk}^{avg} (2^{nd} vtx)	Min, Max and Avg. track rapidity of tracks at secondary vertex
m	Invariant mass of tracks associated to secondary vertex
E	Energy of charged tracks associated to secondary vertex
f_E	Energy fraction of charged tracks (from all tracks in the jet) associated to secondary vertex
N_{trk}	Number of tracks associated to the secondary vertex