Flavour Tagging at ATLAS

Hannah Arnold

Albert-Ludwigs-Universität Freiburg

Heavy Flavour Production at the LHC, IPPP Durham

20-22 April 2106

Introduction

- *b*-tagging = the identification of jets containing *b*-hadrons
  → important tool for background suppression in a wide range of analyses
  (e.g. top quark and Higgs sector, in the search for new phenomena)

- several dedicated algorithms with varying levels of complexity and performance
- exploiting the long lifetime, high mass and decay multiplicity of *b*-hadrons and the hard *b*-quark fragmentation
  → reconstructable secondary vertices and distinct track properties
- performance characterized by power to separate *b*-, *c*- and light jets in simulated events (requires jet flavour labelling based on simulation's truth record)
  - **challenges**: dense and boosted environments, double *b*-hadron jets,...

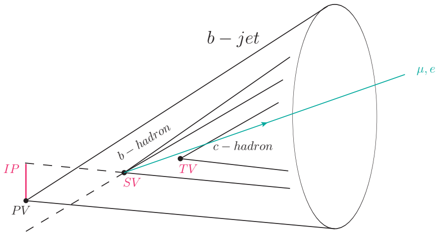
- for use in physics analyses measurements of *b*-jet tagging efficiency, *c*-jet tagging efficiency and mistag rate required → data-to-MC scale factors (SF)
  → eliminating / reducing dependence on simulations
- several (complementary) calibration methods developed → combinations
  - **challenges**: selecting samples with strong predominance of a single flavour, SF applicability to different simulations / processes,...

Outline

b-Tagging algorithms

b-Tagging algorithms

1. lifetime-based
   impact parameter-based: (JetProb), IP3(2)D
   vertex-based: SV(0)1, JetFitter
   → combinations: IP3D+SV1, IP3D+JetFitter, **MV1**, **MV2** (Run 2)
2. muon-based
3. *b*-jet trigger



Key objects

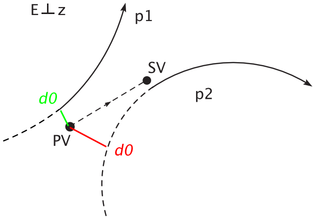
1. (small-R) calorimeter jets
   → jet axis used to **define *b*-hadron flight path**
2. tracks reconstructed in the Inner Detector
3. signal primary vertex (PV) of HS collision (with ≥ 2 tracks)

tracks-to-calo. jets association:

- based on angular separation ΔR(track, jet)
- ΔR cut varies as a function of the jet p_{T}
  (decreasing with increasing jet p_{T})
- exclusive

Impact parameter(IP)-based algorithms I

transverse IP d_0: distance of closest approach of track to PV in the r-ϕ projection
longitudinal IP z_0: distance between z coordinates of the PV and the track at closest approach in r-ϕ

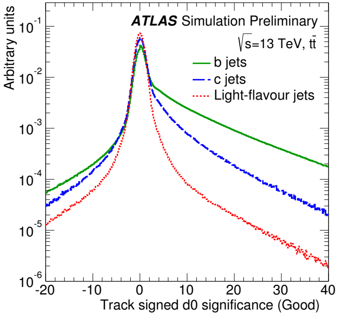


improving separation power by:

- introducing sign: "+" ("-") - track intersects the jet axis in front of (behind) the PV
- using significance, e.g. d_0/σ_{d_0}

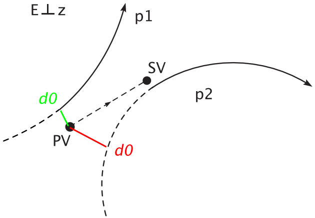
→ tracks from *b*-/*c*-hadron decays: large d_0 and z_0, "+" sign
→ exp. resolution generates a random sign,
tails at "+" from long-lived particles, conversions,...

→ **JetProb**: relies on d_0 significance; simple, robust, no simulation dependence



Impact parameter(IP)-based algorithms II

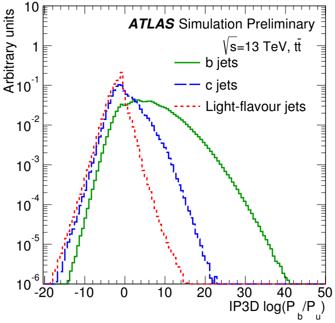
transverse IP d_0: distance of closest approach of track to PV in the r-ϕ projection
longitudinal IP z_0: distance between z coordinates of the PV and the track at closest approach in r-ϕ



improving separation power by:

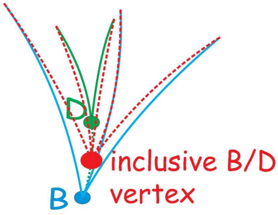
- introducing sign: "+" ("-") - track intersects the jet axis in front of (behind) the PV
- using significance, e.g. d_0/σ_{d_0}

$\rightarrow$ tracks from b-/c-hadron decays: large d_0 and z_0, "+" sign
$\rightarrow$ exp. resolution generates a random sign, tails at "+" from long-lived particles, conversions,...



IP3D: LLR-based, relies on d_0 and z_0 significances as well as correlations
$\rightarrow$ more powerful, 2D-PDFs from simulation

Vertex-based algorithms: SV algorithm II

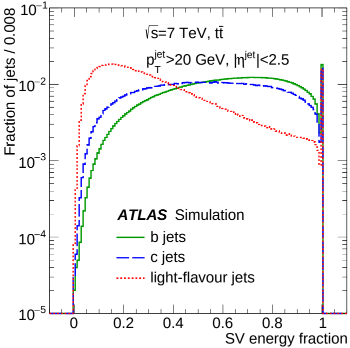


Explicit reconstruction of a *single, inclusive* secondary vertex (SV).

- using all associated, displaced tracks to form vertex candidates from track pairs (χ^2 based)
- vertices compatible with long-lived particles and material interactions are rejected
- iterative procedure to combine all tracks from 2-track-vertices into single inclusive vertex

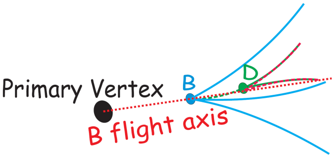
→ small mistag rate
→ SV finding efficiency: $\sim 70\%$

→ **SV1**: LLR-based, exploiting vertex mass, energy fraction, number of 2-track vertices, ΔR(jet,PV-SV)



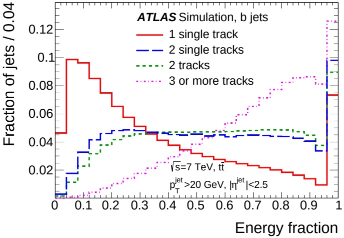
Vertex-based algorithms: JetFitter algorithm

Explicit reconstruction of the complete *b*-hadron decay chain.



- exploiting topological structure of weak *b*- and *c*-hadron decays inside jet
- uses Kalman filter to find common line between and position of PV, SV and TV (tertiary, *c*-hadron decay vertex)
  → approximating *b*-hadron flight path
- one track sufficient to built vertex!

→ six variables: decay topology + vertex information



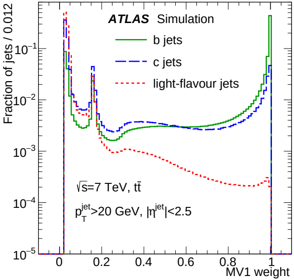
→ input for artificial neural network

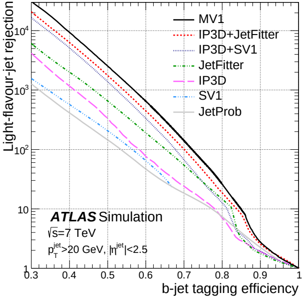
→ three output nodes corresponding to the *b*-,*c*- and light jet hypotheses $P_{b/c/l}$

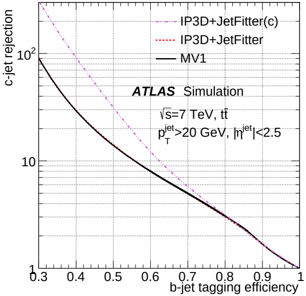
→ **final discriminating variables**: $\ln(P_b/P_l)$ (JetFitter), $\ln(P_b/P_c)$ (JetFitter(c))

Performance I

- figure of merit: *b*-jet tagging efficiency vs. light (c-) jet rejection
- defined by placing cuts on *b*-/*c*/-light jet distributions of the disciminating variables in simulated $t\bar{t}$ events
- clear hierarchy between standalone and combined algorithms
  - MV1 = IP3D+SV1+(IP3D+JetFitter) (NN)



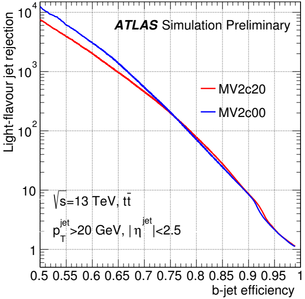


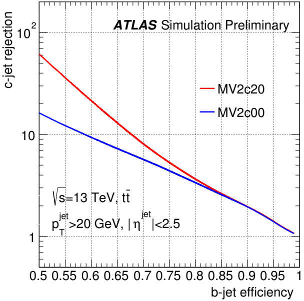


Performance II: Run 2 improvements

Improved performance in Run 2 due to

- upgrade of the Inner Detector: added fourth pixel layer, Insertable B-Layer (IBL)
- improved track reconstruction (shared hits at high jet p_{T})
- revisited *b*-tagging algorithms: **MV2cxx**
  - BDT combining 24 input variables from SV, IP(2)3D, JetFitter algorithms
  - "xx" = *c*-jet fraction in simulated $t\bar{t}$ events used for training



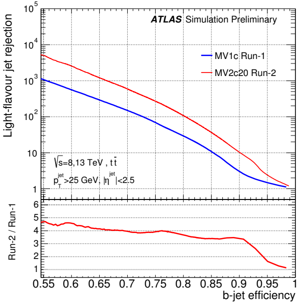


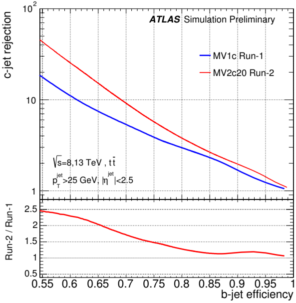
→ small admixture of *c*-jets only slightly degrades light-jet reject, significantly improves *c*-jet rejection

Performance III: Run 2 improvements

Improved performance in Run 2 due to

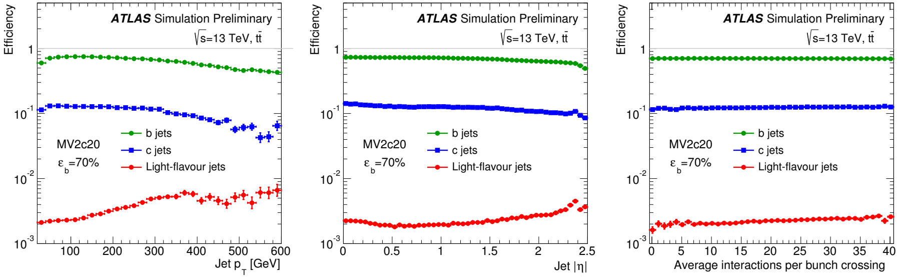
- upgrade of the Inner Detector: added fourth pixel layer, Insertable B-Layer (IBL)
- improved track reconstruction (shared hits at high jet p_{T})
- revisited *b*-tagging algorithms: **MV2cxx**
  - BDT combining 24 input variables from SV, IP(2)3D, JetFitter algorithms
  - "xx" = *c*-jet fraction in simulated $t\bar{t}$ events used for training





→ @70% *b*-tagging eff.: factor 4 (1.5-2) improved light-(c-)jet rejection!

Performance IV: differential



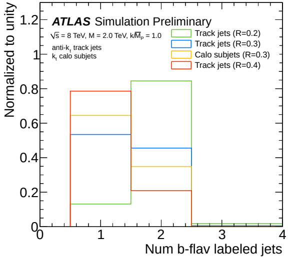
- *b*-tagging depends on jet p_{T} and η, as well as average number of pile-up interactions, different impact on *b*-/*c*- and light-jet efficiencies
- degradation of *b*-jet tagging efficiency at
  - very low jet p_{T}: the *b*-hadron flight path is (too) short, tracks are (too) soft and not very displaced
  - high jet p_{T}: merging of tracks, detector inefficient for very late decays, increased track multiplicity due to fragmentation

Performance V: dense environment and boosted topologies

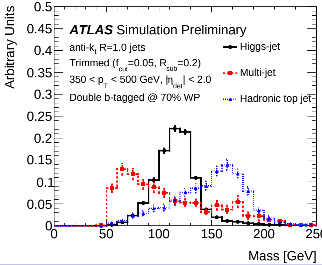
Scenario: searches for new physics at high masses

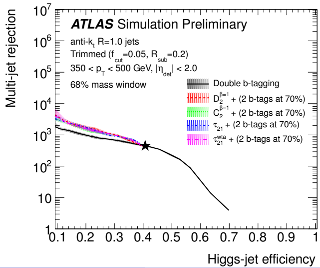
→ highly boosted and collimated decay products, e.g. $h \to b\bar{b}$

- R=0.4 calo. jets might not be able to resolve the products from the two *b*-hadron decays
  → large-R (e.g. 1.0) calo. jets: "Higgs jet"
- apply *b*-tagging to ghost-associated (R=0.2) track jets
  - small *R* parameters possible
  - low PU sensitivity
  - good angular resolution also in dense environment



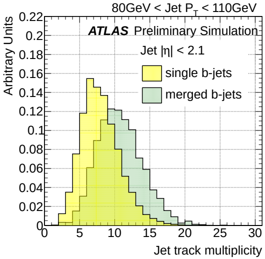
- exploit large-R jet substructur to improve bkg. rejection?

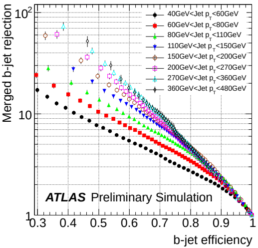




Performance VI: double *b*-hadron jets

- *b*-tagging algorithms currently optimized on $t\bar{t}$ samples, i.e. for jets containing a single *b* hadron $\rightarrow$ no information on the number of *b*-hadrons
- competing effects if two *b*-hadron decays inside one jet ($\geq$ 2 SVs, higher multiplicity of displaced tracks,...) likely improve (decrease) performance of IP-based and SV taggers (JetFitter)
- possible strategies to separate single and double *b*-hadron jets:
  - excplicitely reconstruct two *b*-hadron decay vertices / chains
  - exploit substructure, kinematic differences





Enriched *b*-jet sample using MV1@60%. Other MVA inputs: track-jet width, distances between tracks / subjets

Jet truth flavour labelling

Run 1

- select partons with $p_{\mathrm{T}} > 5$ GeV
- cone-based matching to reco. jet: $\Delta R < 0.3$

Run 2

- select weakly decaying hadrons with $p_{\mathrm{T}} > 5$ GeV
- *exclusive* cone-based matching to reco. jet: $\Delta R < 0.3$
- alternative: ghost-association

Order of labelling

- if b parton(hadron) is found $\rightarrow$ b jet
- else if c parton(hadron) is found $\rightarrow$ c jet
- else if tau is found $\rightarrow$ tau jet
- else light jet

Comments

- several analyses already used Run 2 baseline in Run 1
- cone-based matching and ghost-association agree at the 1% level in case of b-jets (in $t\bar{t}$)
- difference has non-negligible impact on light-jet rejection
- cone-based labelling coherent with track-jet association $\rightarrow$ baseline!
- $g \rightarrow b\bar{b}/c\bar{c}$ **jets are labelled as b-/c-jets** $\leftarrow$ agrees with b-tagging point of view

Calibration of b-tagging algorithms

- the performance of b-tagging algorithms is optimized and evaluated using simulations, usually of inclusive $t\bar{t}$ events
- cannot expect simulations to describe all effects (modelling, detector) that impact the performance of b-tagging algorithms accurately
  $\rightarrow$ measurements of the b-/c- and light-jet tagging efficiencies needed;
  as function of jet p_{T} and η
- requires extracting samples of jets dominated by a single jet flavour
- results are presented in terms of data-to-simulation scale factors

$$\mathrm{SF} = \varepsilon_x^{\mathrm{data}} / \varepsilon_x^{\mathrm{MC}}, \quad x = b, c, \mathrm{light}$$

- **assumption**: SFs are process independent
  i.e. SF measured in a $t\bar{t}$ dilepton sample, applicable to a W+jets sample
- SFs are MC generator dependent
  $\rightarrow$ current approach: MC-to-MC SFs
  better: derive SFs for different MC generators

Calibration methods: an overview

b-jet tagging efficiency calibration

- muon-based methods (dijet samples)
  - p_T^{rel}
  - system8
- $t\bar{t}$-based methods (single-/dilepton $t\bar{t}$) ← inclusive!
  - tag counting
  - kinematic selection
  - kinematic fit
  - combinatorial likelihood ← most precise!

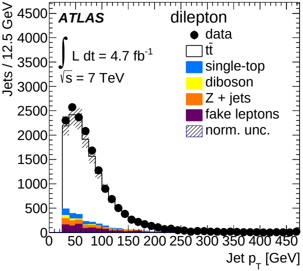
    → **combined** with muon-based methods

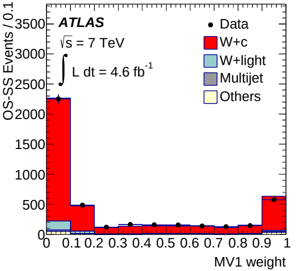
c-jet tagging efficiency calibration

- W+c method (soft-muon tagged)
- D^* method ($D^* \to D^0(\to K\pi)\pi$)

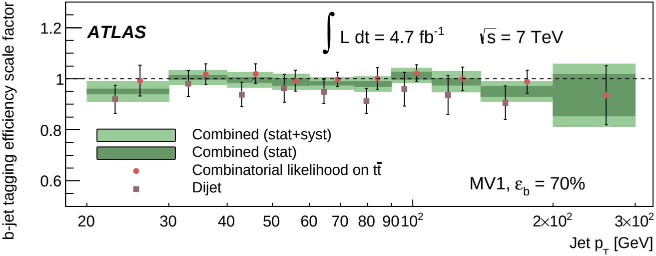
mistag rate calibration: negative tag method

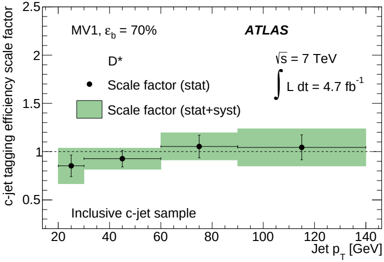
Differences: purity, inclusiveness, simulation dependence, precision

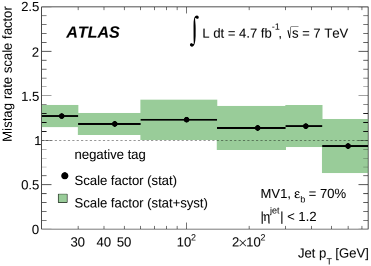




Calibration results

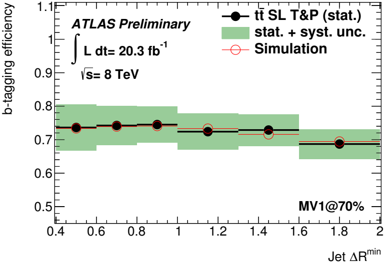


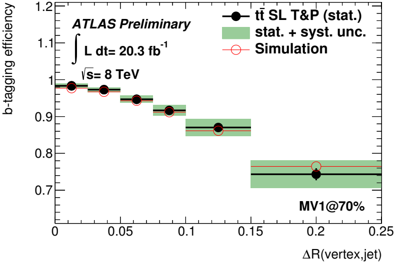




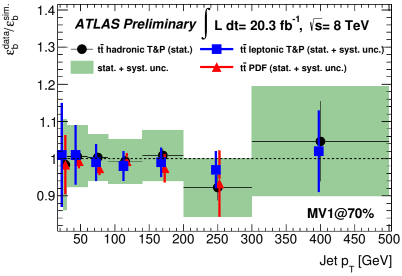
Total uncertainties for MV1@70%: *b*-jet calib. 2-8% (comb. like.), *c*-jet calib. 8-15% (D^*), light-jet calib. 15-40 %

Calibration results: dense invironments I

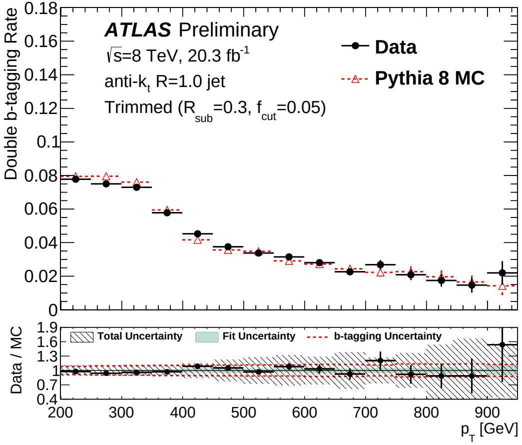




- **top**: b-jet tagging efficiency for jets from the hadronic top decay in single-lepton $t\bar{t}$ events as function of
  left: the minimal distance to a close-by jet
  right: the distance between the jet axis and the PV-SV axis
  $\rightarrow$ SFs: $\sim$ 1 and flat across jet p_{T}
- **bottom**: comparison of SFs with jets from the leptonic top decay $\rightarrow$ compatible



Calibration results: dense invironments II



Efficiency for two b-tagged (MV1@70%) track jets ghost-associated to large-R calo. jet.

Impact of (heavy flavour) modelling on b-tagging efficiencies

- **calibration's goal**: resolve (minimize) simulation dependence (including detector response, reconstruction,...)
- calibration methods exploiting *inclusive* jet samples are doing a good job
  - still rely on simulation for e.g. (ratios of) flavour fractions, distributions of discriminant variables,... $\rightarrow$ accounted by applying dedicated uncertainties
  - modelling uncertainties dominant for certain methods and phase space regions $\rightarrow$ can be resolved by refined, more data-driven methods, e.g. combinatorial likelihood method
- more difficult in case of methods relying on *non-inclusive* jet samples:
  - SFs applicable to *inclusive* jet samples if the differences of jet properties (affecting b-tagging) between the *non-inclusive* and the *inclusive* sample are well described

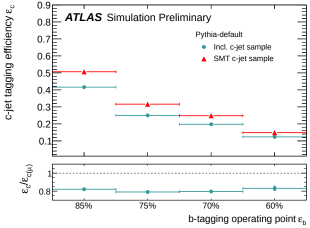
$$\varepsilon = \alpha \cdot \varepsilon_{\text{excl}}$$

$$\text{SF} = \frac{\alpha_{\text{data}}}{\alpha_{\text{MC}}} \cdot \text{SF}_{\text{excl}}$$

$$\text{if } \alpha_{\text{data}} = \alpha_{\text{MC}} \rightarrow \text{SF} = \text{SF}_{\text{excl}}$$

$$\alpha_{\text{data}} = \alpha_{\text{MC}}?$$

If not $\rightarrow$ need to extrapolate, i.e. estimate $\frac{\alpha_{\text{data}}}{\alpha_{\text{MC}}}$.



Extrapolation to inclusive jet samples

- **goal**: estimate α_{data} from simulation where heavy quark fragmentation and heavy hadron decay properties are corrected to best knowledge:

$$\alpha_{\text{data}} \approx \alpha_{\text{MC}}^{\text{corr}}$$

$$\rightarrow \ \text{SF} \approx \frac{\alpha_{\text{MC}}^{\text{corr}}}{\alpha_{\text{MC}}} \cdot \text{SF}_{\text{excl}}$$

- ratio $\rightarrow$ SF extrapolation relatively independent from reference simulation

Example: corrections applied in W+c calibration (reference MC: Pythia 6)

- the fragmentation fractions (e^+e^- and $e^\pm p$ data)
- the *total* branching ratio of the semileptonic decay of c-hadrons (PDG)
- the branching ratios of some *exclusive* semileptonic decays (PDG)
- the topological branching ratios of hadronic n-prong decays (PDG / EvtGen)
- the momentum of the decay muon in the rest frame of the c-hadron p^* (EvtGen)

(the impact of the c-quark fragmentation function was evaluated, but not corrected)

Extrapolation to inclusive jet samples

- **goal**: estimate α_{data} from simulation where heavy quark fragmentation and heavy hadron decay properties are corrected to best knowledge:

$$\alpha_{\text{data}} \approx \alpha_{\text{MC}}^{\text{corr}}$$

$$\rightarrow \text{SF} \approx \frac{\alpha_{\text{MC}}^{\text{corr}}}{\alpha_{\text{MC}}} \cdot \text{SF}_{\text{excl}}$$

- ratio → SF extrapolation relatively independent from reference simulation

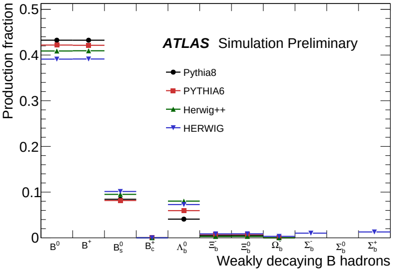
Example: corrections applied in W+c calibration (reference MC: Pythia 6)

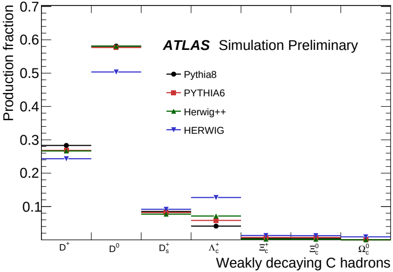
- the production fractions (e^+e^- and $e^\pm p$ data)
- the *total* branching ratio of the semileptonic decay of c-hadrons (PDG) ← today: EVTGEN*!
- the branching ratios of some *exclusive* semileptonic decays (PDG) ← today: EVTGEN*!
- the topological branching ratios of hadronic n-prong decays (PDG / EVTGEN) ← does not cancel!
- the momentum of the decay muon in the rest frame of the c-hadron (EVTGEN)

*with ATLAS decay table

(the impact of the c-quark fragmentation function was evaluated, but not corrected)

Heavy hadron production fractions





- large differences in production fractions between generators
- existing measurements
  - for *b*-hadrons: HFAG (Tevatron), HFAG (Tevatron + LEP + LHCb)
  - for *c*-hadrons: several measurements in e^+e^- and $e^\pm p$ data $\rightarrow$ combination

  differ among each other and with simulation (Pythia 6 is rather close!)
- impact on *b*-tagging efficiencies is rather small; more pronounced for *c*-hadrons whose tagging efficiencies are rather different

  for MV1@70%: $\varepsilon(D^+)/\varepsilon(D^0) \sim 1.8$

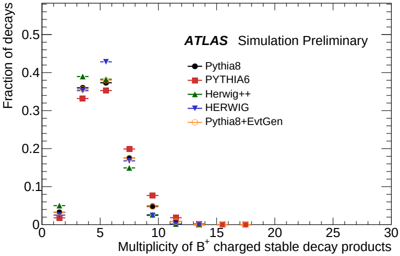
  $\varepsilon(B^+)/\varepsilon(B^0) \sim 1.03$

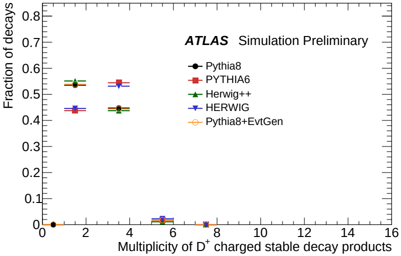
Heavy hadron branching fractions

- inclusive and prominent exclusive semi-leptonic decays of dominant, weakly decaying heavy hadrons measured to high precision
  $\rightarrow$ implemented in EVTGEN via ATLAS dec. table!
  (other generators show large differences)
- many exlcusive hadronic decays are measured, but still make up only a relatively small fraction of all hadronic decays
  e.g. D^+: 60%, b-baryons: much less
- large differences between the number of implemented exclusive hadronic decays in different generators (EvtGen » Pythia8 » Pythia6 » Herwig) and their branching fractions
- BUT: mainly the number of charged decay products has an impact on the tagging efficiency - unfortunately here the situation is not much better

Inclusive charged particle multiplicities

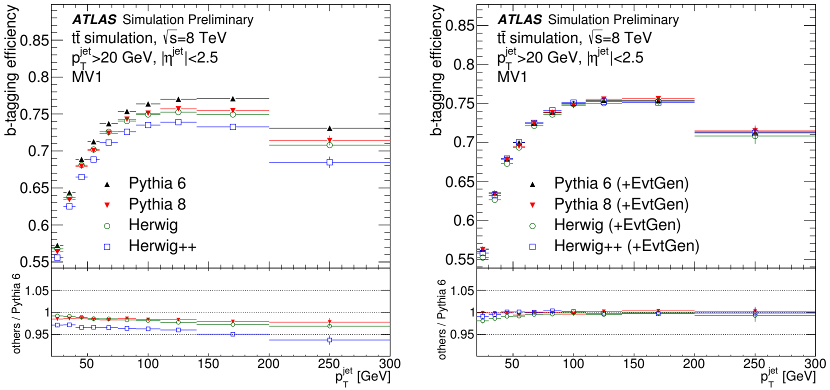
- only for the D^0 the topological (inclusive) n-prong fractions are measured from which one can infer the hadronic ones...
- the mean of inclusive charged particle multiplicities is measured for the admixture of B^0, B^+, B_s, Λ_b: EvtGen (4.76) « 4.97 $\pm$ 0.07 « Pythia 6 (5.20)
- BUT: b-tagging efficiency depends on spectrum - large differences between generators





$\rightarrow$ **dominant (extrapolation) uncertainty in $p_{\mathrm{T}}^{\mathrm{rel}}$, W+c and D^* calibrations!**

Using EvtGen ...



- harmonizes heavy hadron decays
- reduces the differences in the *b*-jet tagging efficiency predicted by generators significantly; only small differences at the % level, mainly at low jet p_{T} remain
- harmonization less pronounced in case of the *c*-jet tagging efficiency - fragmentation fractions, tracks from fragmentation, etc. have a stronger impact

Summary on HF modelling

- using EVTGEN harmonizes the *b*-jet tagging efficiencies significantly, but
- cannot resolve the simulation dependency of SFs completely (especially for *c*-jets), since only affects heavy flavour decay modelling
- other corrections are still needed when extrapolating SFs to inclusive samples
- only comparisons between different generators allow to assess modelling uncertainties other than semi-leptonic branching fractions
- theoretically varying those within their measured uncertainties would suffice, but more complicated to do
- the estimated SF extrapolation factors (rel. unc.) are for MV1:
  $p_{\mathrm{T}}^{\mathrm{rel}}$: ~ 1 (3%); W+c: $0.86 - 0.95$ (3-7%); D^*: $0.82 - 0.92$
  (the corrected eff. extrapolation factors are W+c: $0.7 - 0.75$ « D^*: $0.5 - 0.6$)

Summary

- several dedicated algorithms to identify jets containing *b*-hadrons
- high performing combined algorithm further improved for Run 2
- several methods to calibrate *b*-jet, *c*-jet tagging efficiencies and light-jet mistag rate
- using EVTGEN consistently to decay heavy hadrons in Run 2 simulations reduced generator dependency of the *b*-jet tagging efficiency signficantly
- *b*-tagging of small-R track jets ghost-associated to large-R calo. jets improves *b*-tagging in dense and boosted topologies
- work ongoing to better exploit jet substructure and to improve single and double *b*-hadron jet separation

BACKUP

References

link	title
1512.01094.pdf	“Performance of b-Jet Identification in the ATLAS Experiment“
ATLAS-CONF-2012-100.pdf	“Identification and Tagging of Double b-hadron jets with the ATLAS Detector”
ATLAS-CONF-2013-109.pdf	“Calibration of the b-tagging efficiency for c jets with the ATLAS detector using events with a W boson produced in association with a single c quark “
ATLAS-CONF-2016-001.pdf	“Calibration of ATLAS b-tagging algorithms in dense jet environments”
ATLAS-CONF-2016-002.pdf	“Studies of b-tagging performance and jet substructure in a high p_T $g \rightarrow b\bar{b}$ rich sample of large-R jets from pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector”
ATL-PHYS-PUB-2014-013.pdf	“Flavor Tagging with Track Jets in Boosted Topologies with the ATLAS Detector”
ATL-PHYS-PUB-2014-014.pdf	“b-tagging in dense environments“
ATL-PHYS-PUB-2015-022.pdf	“Expected performance of the ATLAS b-tagging algorithms in Run-2”
ATL-PHYS-PUB-2015-035.pdf	“Expected Performance of Boosted Higgs ($\rightarrow b\bar{b}$) Boson Identification with the ATLAS Detector at $\sqrt{s} = 13$ TeV”
ATL-PHYS-PUB-2016-004.pdf	“Simulation of top quark production for the ATLAS experiment at $\sqrt{(s)} = 13$ TeV “
ATL-PHYS-PUB-2014-008	“Comparison of Monte Carlo generator predictions for bottom and charm hadrons in the decays of top quarks and the fragmentation of high p_T jets “