GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

b-tagging calibration using the $t\bar{t}$ PDF method with the ATLAS experiment in Run II

Jannik Geisen
Supervised by: Arnulf Quadt

II. Physikalisches Institut, Georg-August-Universität Göttingen

Special thanks to the whole $t\bar{t}$ PDF b-calibration team

Helmholtz Alliance Terascale Meeting 2017
DESY, 28.11.2017

Bundesministerium für Bildung und Forschung

BMBF-Forschungsschwerpunkt ATLAS-EXPERIMENT
Physik bei höchsten Energien mit dem ATLAS-Experiment am LHC
FSP 103 ATLAS

Outline

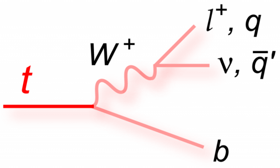
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

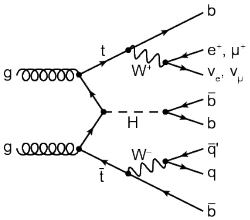
1. *b*-tagging calibration in ATLAS
    - Motivation
    - The MV2c10 algorithm in ATLAS
2. The PDF likelihood method in dileptonic $t\bar{t}$ events
    - Idea
    - Setup
    - Studies & Results
3. Outlook
    - Next steps
    - Further improvements

b-tagging calibration in ATLAS

Why is b/flavour-tagging so important?

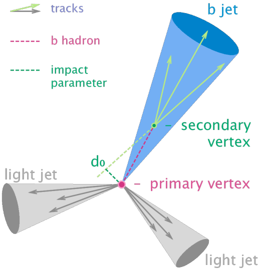
- LHC collides protons $\Rightarrow$ always have many jets (also from pileup)
- LHC is a "top factory" $\Rightarrow$ BR($t \to Wb$) $\approx$ 100%
- b-tagging one of leading uncertainties in many ATLAS top analyses
- BR($H \to b\bar{b}$) $\approx$ 58%
  $\Rightarrow$ b-tagging also important for $t\bar{t}H$, VH & BSM analyses
- Some analyses must reject b-jets! (e.g. $H \to \tau\tau$)

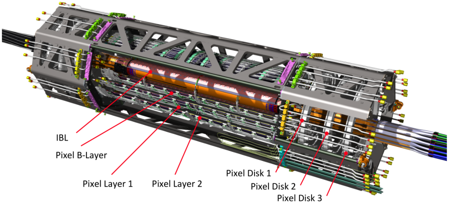




Multivariate analysis approach to identify *b*-jets

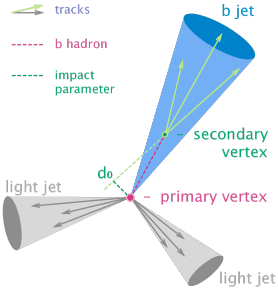
- Most vital input: charged particle tracks reconstructed in inner detector
- Information from tracks associated to jets used in different algorithms
- Run II: Insertable B-Layer (IBL)! → positive impact on *b*-tagging
  (*b*-tagging performance plots with/without IBL)
- Tracking improved in Run II (+), but jets have higher boost on average (-)





ATLAS uses 3 distinct basic *b*-tagging algorithms

1. Impact parameter based
2. Inclusive secondary vertex reconstruction
3. Decay chain multi-vertex reconstruction



- Complementary algorithms, output combined in multivariate discriminant
- Trained on $\approx$ 5M $t\bar{t}$ events:
  - Signal: *b*-jets
  - Background: mixture of 10% c-jets + 90% light-jets

$\Rightarrow$ Best separation for different jet flavours: MV2c10

Improvements from Run I to Run II: ATL-PHYS-PUB-2015-022
MV2 optimisation studies: ATL-PHYS-PUB-2016-012

What does 'calibration' mean in *b*-tagging?

- Choose working point of your algorithm, e.g. $\epsilon_b = 70\%$
  $\Rightarrow$ 70% of *b*-jets are *b*-tagged in MC
  $\Rightarrow$ What about data?
- Perform calibration: Estimate ϵ_b^{data} and corresponding uncertainties
- Analysers apply SF $\kappa = \epsilon_b^{\text{data}}/\epsilon_b^{\text{MC}}$ to correct data/MC difference
- κ should be close to 1 $\rightarrow$ shows good performance of algorithm

Examples of different strategies & environments

- Run II: *b*-jets in dileptonic $t\bar{t}$ with combinatorial likelihood approach (Run I: ATLAS-CONF-2014-004)
- *b*-jets in dileptonic $t\bar{t}$ with tag-and-probe method
- *b*-jets in l+jets $t\bar{t}$ with tag-and-probe method
- light-flavour jets in di-jet events with negative tag method
- *c*-jets with "JetFitterCharm" in l+jets $t\bar{t}$ (ATL-PHYS-PUB-2015-001)

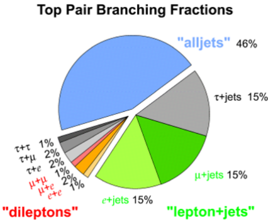
The **P**robability **D**ensity **F**unction likelihood method in dileptonic $t\bar{t}$ events

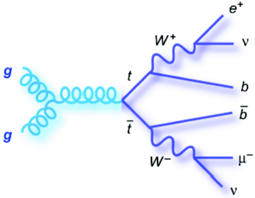
PDF likelihood method

Motivation

b-tagging calibration in dileptonic $t\bar{t}$ using PDF likelihood method

- Select as many events rich in b-jets as possible
- Dileptonic $t\bar{t}$ events $\rightarrow$ pure final state $\left(l^+ l^- \nu \bar{\nu} b \bar{b}\right)$, low c-jet contribution
  $\Rightarrow$ Main background: light-jets from initial/final state radiation
- Split into 6 channels for control: $(ee, \mu\mu, e\mu)$ X (2j, 3j)
- Perform fit to likelihood function in all data events
  $\Rightarrow$ Extract ϵ_b as function of jet p_T
- Derive data/MC scale factors and systematic uncertainties





PDF likelihood method

Likelihood method

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Per-event likelihood function for two jet case:

$$\begin{aligned}\mathcal{L}\left(p_{T,1}, p_{T,2}, w_1, w_2\right) = [\ & f_{bb}\mathcal{P}_{bb}\left(p_{T,1} p_{T,2}\right)\mathcal{P}_b\left(w_1|p_{T,1}\right)\mathcal{P}_b\left(w_2|p_{T,2}\right) \\ &+ f_{bl}\mathcal{P}_{bl}\left(p_{T,1} p_{T,2}\right)\mathcal{P}_b\left(w_1|p_{T,1}\right)\mathcal{P}_l\left(w_2|p_{T,2}\right) \\ &+ f_{ll}\mathcal{P}_{ll}\left(p_{T,1} p_{T,2}\right)\mathcal{P}_l\left(w_1|p_{T,1}\right)\mathcal{P}_l\left(w_2|p_{T,2}\right) \\ &+ 1 \leftrightarrow 2]/2\end{aligned}$$

- $f_{bb}, f_{bl}, f_{ll} = 1 - f_{bb} - f_{bl}$: two jet flavour fractions
- $\mathcal{P}_f\left(w|p_T\right)$: PDF for b-tag weight for jet flavour f, depending on jet p_{T}
- $\mathcal{P}_{f_1 f_2}\left(p_{T,1}, p_{T,2}\right)$: two-dim PDF for combination of $[p_{T,1}, p_{T,2}]$ & $[f_1, f_2]$
- PDFs implemented as binned histograms

Define total likelihood function, to be maximised on all data events:

$$\mathcal{L}_{\text{tot}} = \exp\left(-N\right)\prod_{i=1}^{N_{\text{events}}}\mathcal{L}\left(p_{T,1,i}, p_{T,2,i}, w_{1,i}, w_{2,i}\right)$$

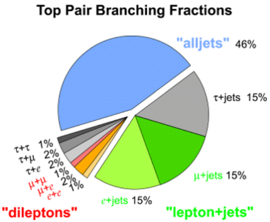
$$\text{Extract } \mathcal{P}_b\left(w|p_T\right) \text{ from fit } \Rightarrow \underline{\epsilon_b(p_T)} = \int_{w_{\text{cut}}}^{\infty} \mathrm{d}w' \mathcal{P}_b\left(w'|p_T\right)$$

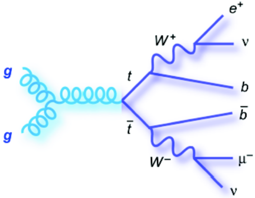
PDF Likelihood method

Event selection

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

- Exactly 2 oppositely charged (OS) leptons (e, μ):
  - Trigger-matched, tight isolation, no overlap with jets
  - $p_{\mathrm{T}} > 28$ GeV, $|\eta_{\mu}| < 2.5$, $|\eta_{\mathrm{el}}| < 2.47$ (exclude $1.37 < |\eta_{\mathrm{el}}| < 1.52$),
- Exactly two or three jets:
  - Anti-k_t with $\Delta R = 0.4$ from the calorimeter
  - jet $p_{\mathrm{T}} \geq 20$ GeV, jet $|\eta| < 2.5$
- End up with 6 channels:
  - $ee, \mu\mu$ dominated by Z+jets (smaller background: single top, Diboson) $\Rightarrow$ Allow to constrain Z+jets and improve stat
  - $e\mu$ highest purity; main bkgd: single top $\Rightarrow$ Consider $e\mu$+2j for calibration





PDF Likelihood method

Analysis setup

- Analyse ATLAS 2015+2016 data at $\sqrt{s} = 13$ TeV with $\int L \, \mathrm{d}t = 36.1 \text{ fb}^{-1}$
- Compare to MC simulation for dileptonic $t\bar{t}$ signal & background:

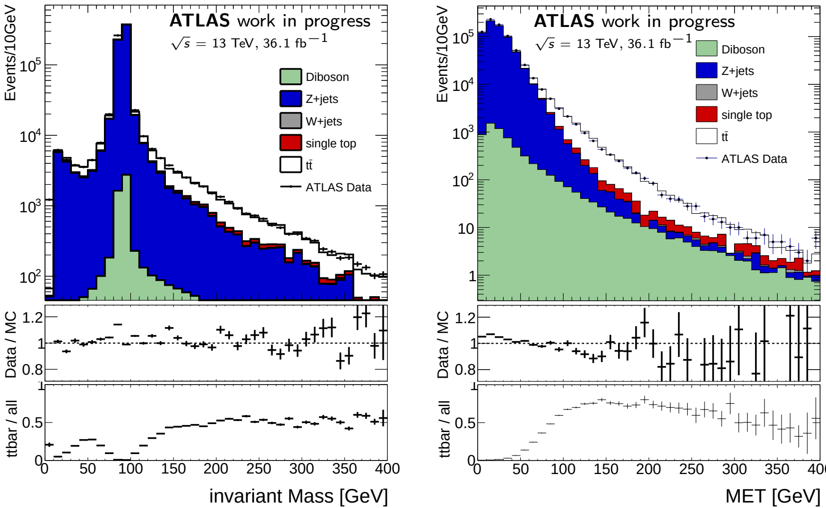
Process	Nominal MC	Alternative MC
Signal: $t\bar{t}$	Powheg+Pythia8	Sherpa 2.2.1
Background:		
Z+jets	Sherpa 2.2.1	Powheg+Pythia8
Single top	Powheg+Pythia6	-
Diboson	Powheg+Pythia8	Sherpa 2.2.1
Lepton fakes	From data	-
W+jets	Sherpa 2.2.1	-

This is work in progress, more alternative samples are being produced!

PDF Likelihood method

Initial control plots in ee+2j

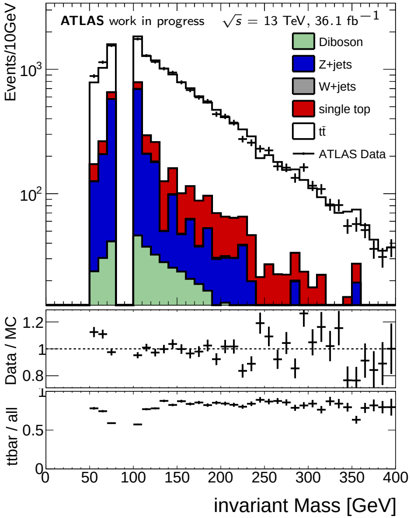
Z+jets dominant background in $ee, \mu\mu$

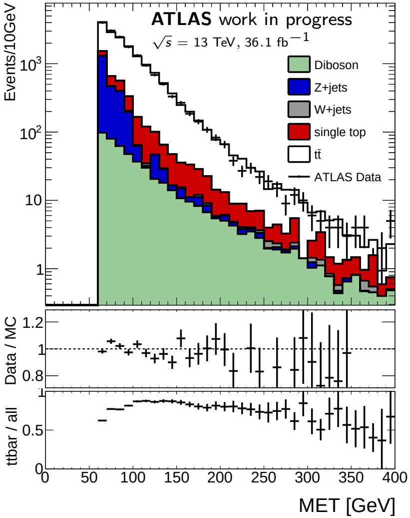


$\Rightarrow$ Require $E_T^{miss} > 60$ GeV and 50 GeV $< m_{ll} <$ 80 GeV or $m_{ll} > 100$ GeV

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

After cut on $E_T^{\text{miss}} > 60$ GeV and 50 GeV $< m_{ll} <$ 80 GeV or $m_{ll} >$ 100 GeV:





PDF Likelihood method

Additional cut in $e\mu$

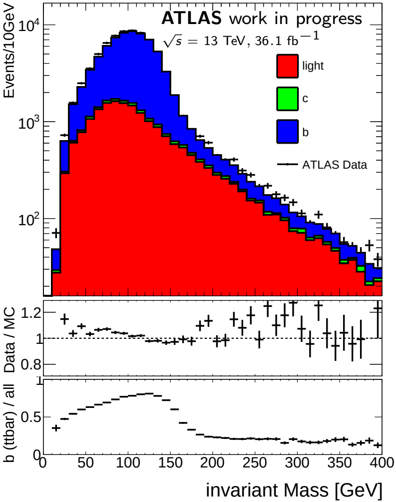
- $e\mu$ channel has highest purity ($\approx 7 - 14\%$ larger than $ee, \mu\mu$)
- Low Z+jets background $\Rightarrow$ no MET or m_{ll} cuts applied

<u>Think of further cut:</u>

- Build all possible pairs of leptons and jets $\Rightarrow$ look at $m_{l,j}$
- If lepton and jet from same top, $m_{l,j}$ has end point at $\approx m_{\text{top}}$
- Combine leptons with jets such that $\left(m^2_{l_1,j_1} + m^2_{l_2,j_2}\right)$ minimal

$\Rightarrow$ Require $m_{l,j} < 175$ GeV for one or both pairs

Check $m_{l,j}$ split by flavours:

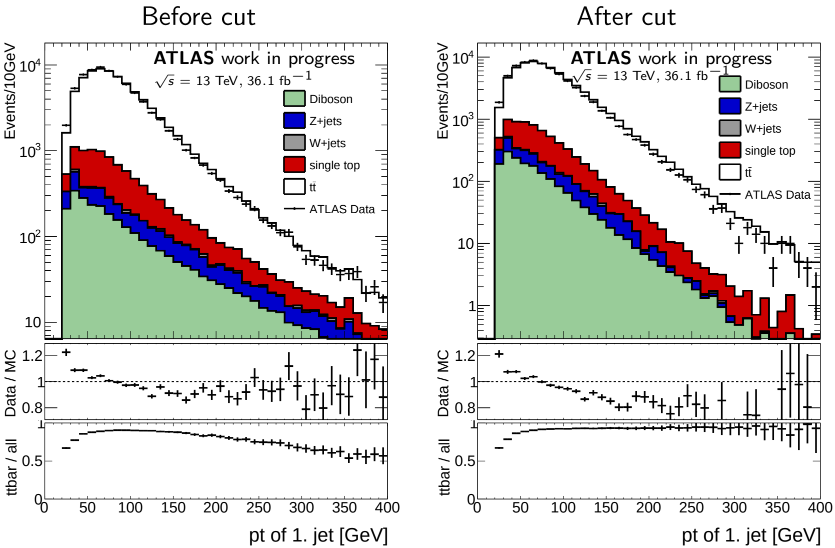


PDF Likelihood method

Apply 2 lepton cut in $e\mu$+2j

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Before cut | After cut



⇒ Signal purity increases especially in high jet p_{T} regions

PDF Likelihood method

Renormalisation of MC

Select regions to extract data/MC renormalisation scale factors for $t\bar{t}$ & Z+jets
$\Rightarrow$ Improve Data/MC agreement before fit
Use same-sign (SS) lepton events to estimate non-prompt and fake leptons

Process	$e\mu$	$ee, \mu\mu$
$t\bar{t}$	2 OS leptons Different lep flavour	2 OS leptons Same lep flavour 50 GeV $< m_{ll} <$ 80 GeV or $m_{ll} >$ 100 GeV; $E_T^{\text{miss}} >$ 60 GeV
Z+jets	2 OS leptons Same lep flavour 80 GeV $< m_{ll} <$ 100 GeV	2 OS leptons Same lep flavour 80 GeV $< m_{ll} <$ 100 GeV; $E_T^{\text{miss}} >$ 60 GeV
Non-prompt/ fake leptons	2 SS leptons Different lep flavour	2 SS leptons Same lep flavour 50 GeV $< m_{ll} <$ 80 GeV or $m_{ll} >$ 100 GeV; $E_T^{\text{miss}} >$ 60 GeV

- Z+jets SF between 0.96 - 1.05
- $t\bar{t}$ signal SF between 0.98 - 1.03
- Contribution from fakes found to be negligible

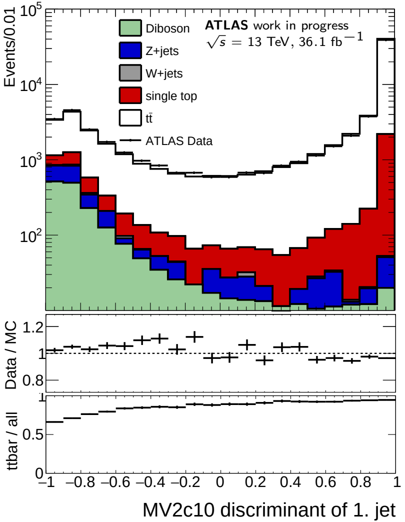
PDF Likelihood method
Systematics

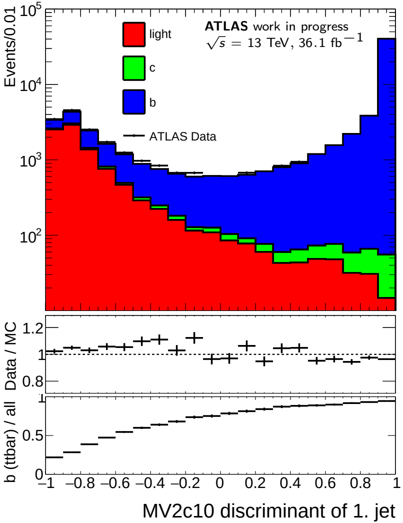
Main systematics:

- Monte Carlo Generator
  ⇒ Currently: only Powheg+Pythia8 vs. Sherpa available
  ⇒ Future: Powheg vs. MadGraph5_aMC@NLO; Pythia8 vs. Herwig7
- Pileup modelling
- Jets
  - Jet energy scale
  - Jet energy resolution
  - Jet vertex tagger
- Leptons
  - Trigger
  - Identification
  - Isolation

Fit input

MV2c10 for first jet in $e\mu$+2j



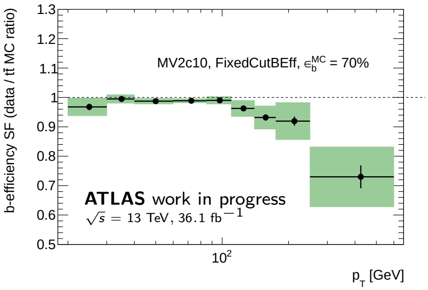


Fit result

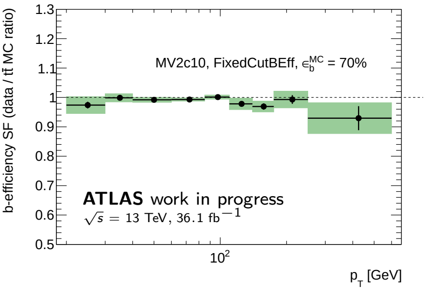
Data/MC SF for MV2c10 (Fixed cut)

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

$\epsilon_b = 70\%$: Before cut



After 2 lepton cut



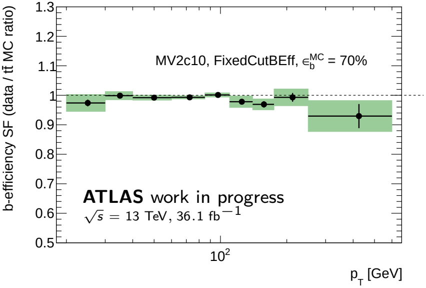
- Scale factor decreases after jet $p_T \approx 150$ GeV
- Data/MC agreement on 2-jet flavour fraction improved after 2 lepton cut
- Systematics in the order of few % $\rightarrow$ main syst. uncertainty: MC modelling

Summary

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

- b-tagging is an indispensable tool in LHC analyses
- Calibration of b-tagging algorithms necessary to correct MC eff. to data
- Performed b-tagging calibration in $e\mu$+2j channel of $t\bar{t}$ events
- PDF likelihood fit results with MV2c10 (fixed cut) shown
  - Good data/MC agreement observed
  - SF divergence from 1 at high jet p_T reduced by 2 lepton cut

Keep in mind: very preliminary results!



Outlook

Next steps

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

- Include 2017 data asap
  $\Rightarrow$ Need appropriate MC simulation including pileup modelling
- Studying new tagging algorithms in ATLAS
  - Deep learning tagging algorithm "DL1" (based on neural network)
  - Hybrid cut instead of fixed cut
    (equal to fixed cut for jet $p_{\mathrm{T}} < 250$ GeV; afterwards p_{T}-dependent efficiency)
- Implement BDT to further discriminate signal from background

Thank you for your attention!

Backup

MC Normalisation SF