GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Fake Estimation in the $t\bar{t}H(H \to b\bar{b})$ Analysis in the Single Lepton Channel with the ATLAS Experiment

Johannes Mellenthin

Supervised by Arnulf Quadt

II. Physikalisches Institut, Georg-August-Universität Göttingen

11th Annual Meeting of the Helmholtz Alliance “Physics at the Terascale”

28.11.2017

SPONSORED BY THE

Federal Ministry of Education and Research

BMBF-Forschungsschwerpunkt ATLAS-EXPERIMENT

Physik bei höchsten Energien mit dem ATLAS-Experiment am LHC

FSP 103 ATLAS

$t\bar{t}H$ Analysis

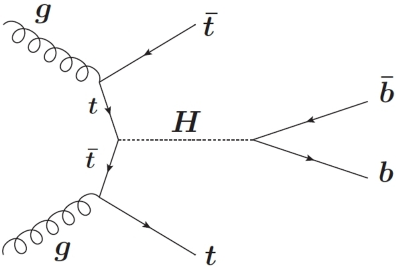
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

$t\bar{t}H(H \to b\bar{b})$

- Direct sensitivity to top and *b*-quark Yukawa coupling
- Study Higgs fermion couplings
- $H \to b\bar{b}$ has the highest BR (58%)

Results

- Full $t\bar{t}H(H \to b\bar{b})$ analysis with 36.1 fb^{-1} data gives $\mu > 2.0$ excluded at 95% C.L.
- Significance: 1.4 σ (1.6 σ exp.)
- *ATLAS-CONF-2017-076*



Event Topology

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

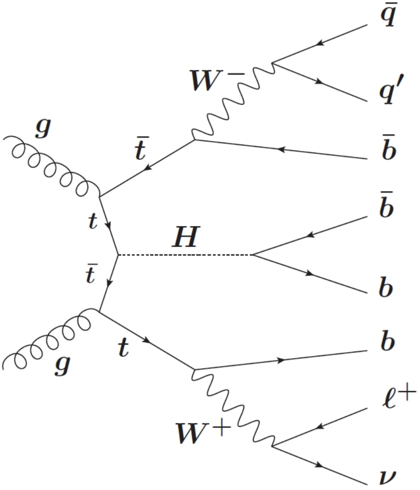
Single Lepton Channel

- 1 leptonic W decay
- 1 electron or 1 muon
- 6 jets with 4 b-jets

Very challenging analysis

- 4 b-jets in the signal region
- Large background from $t\bar{t}$ + jets

$\rightarrow$ Strategy: Divide into different regions



Selection

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Event Selection

- Events triggered by single lepton triggers
- At least 5 jets
- At least 2 *b*-tags at 60% working point (WP) or at least 3 *b*-tags at 77% WP

Categorisation

- Split in *N* jets
  - 5 jets
  - at least 6 jets
- Split in *N* *b*-tags
  - Use 4 working points for 5 ranges

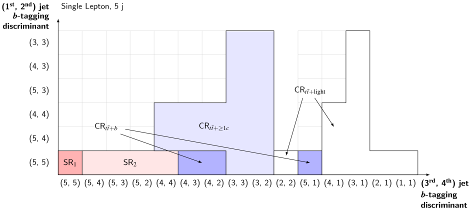
	none (1)	loose (2)	medium (3)	tight (4)	very tight (5)
WP	[100%, 85%]	[85%, 77%]	[77%, 70%]	[70%, 60%]	[60%, 0%]

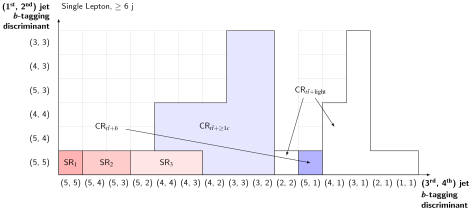
- Define 11 regions enriched in $t\bar{t}H$, $t\bar{t}+b$, $t\bar{t}+c$, and $t\bar{t}+$ light jets

Categorisation

Analysis Regions

- 5 (ultra) pure signal regions (SR); train BDT in each SR
- 6 control regions (CR) which are enriched either in $t\bar{t}+b$, $t\bar{t}+c$, and $t\bar{t}+$ light jets to constrain systematic uncertainties on the background





$\rightarrow$ Combined profile likelihood fit to all regions is performed

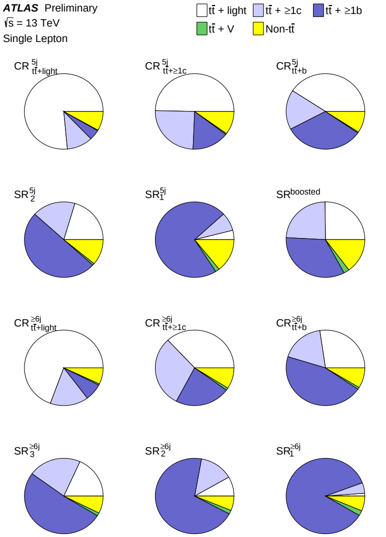
Signal & Background

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Composition

- Signal process $t\bar{t}H$
  - Cross-section: $\sim 0.5\,\text{pb}$
  - MadGraph5_aMC@NLO+Pythia8
- Dominating background
  - $t\bar{t}$ + jets
  - Cross-section: $\sim 832\,\text{pb}$
  - Powheg+Pythia8
- Other backgrounds
  - $t\bar{t}V$, single top, W/Z + jets, diboson
  - Multi-jet (Fakes and non-prompt) ($\sim 4.5\,\%$ in CRs and $\sim 1.3\,\%$ in SRs)

→ Good estimate of multi-jet background improves data/MC agreement and has impact on $t\bar{t}$+HF normalisation

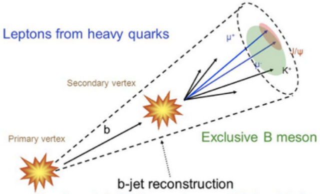


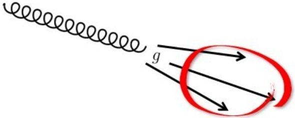
Multi-jet Backgrounds

Fake and non-prompt leptons

- Non-prompt leptons
  - Semi-leptonic decays from *c*- and *b*-quarks
  - Photon conversion
  - Kaon decay (usually small)
- Fake leptons
  - Jets can be misidentified as reconstructed lepton (mostly electrons)
- Consider electrons and muons separately
- Difficult to model this from MC, huge statistics needed

→ Very important to measure this type of background from data





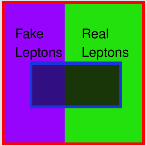
Matrix Method

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

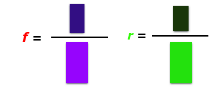
Selection

- Fully data-driven estimate using matrix method
- Introducing a loose lepton selection alongside the tight
- Measure behaviour of real and fake leptons in CRs
  - CRs contain events that are enriched in fakes and non prompt leptons
  - Use an extrapolation weight to relate these events to the background in the SR

$$N^{loose} = N^{loose}_{real} + N^{loose}_{fake}$$
$$N^{tight} = N^{tight}_{real} + \underline{N^{tight}_{fake}}$$



- Define fake efficiency f and real efficiency r as



$$N^{loose} = N^{loose}_{real} + N^{loose}_{fake}$$
$$N^{tight} = r \cdot N^{loose}_{real} + f \cdot N^{loose}_{fake}$$

Efficiencies

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Lepton Selection

- Introducing a loose lepton selection alongside the tight

	Loose selection	Tight selection
Electron identification level	MediumLH	TightLH
Muon identification level	Medium	Medium
Lepton isolation requirement	None	Gradient

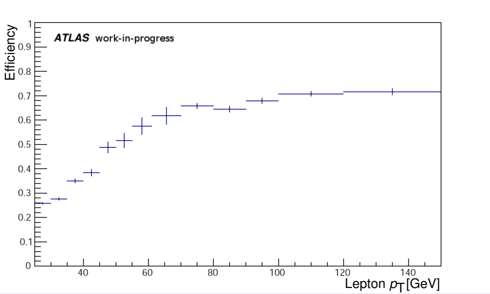
- Measure behavior of fakes in fake enriched regions
  - Electrons: single charged lepton + neutrino and low E_T^{miss}
  - Muons: decay from semileptonic *b*-quarks

Efficiencies

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Fake Efficiency

$$\frac{N\,(\text{Data} - \text{MC}_{\text{real}})_{\text{tight}}}{N\,(\text{Data} - \text{MC}_{\text{real}})_{\text{loose}}}$$



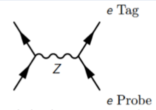
- Efficiencies vary as function of kinematic quantities such as lepton p_T, lepton η, leading jet p_T, E_T^{miss}

Efficiencies

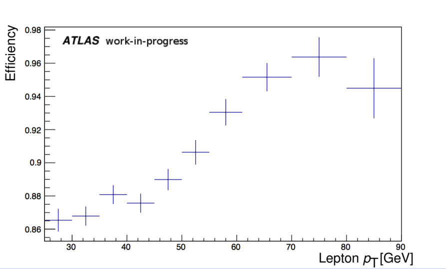
Real Efficiency

- Tag-and-Probe on $Z \rightarrow \ell\ell$ events

$$\frac{\text{Number of probes that pass tight}}{\text{Number of all probes}}$$



- Events with a pair of SFOS loose or tight leptons and at least one jet



Matrix Method

Weight

- Calculate a per-event weight and apply it to loose selection data

$$w = \frac{f}{r - f}(r - P)$$

  where P is 1 if the loose event passes the tight selection and 0 otherwise
- Negative bin entries can occur if most of the loose events pass the tight selection
  - Need to go to a looser lepton definition
  - High fake rates

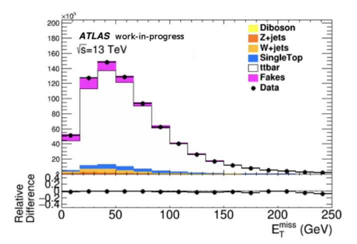
  → Use pre-scaled (PS) triggers (only every n^{th} event passing a (loose) trigger is saved)

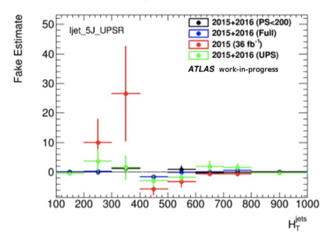
Matrix Method

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Limitations

- Good modelling in both electron and muon channels, but:
  - Only in low *b*-tagging multiplicity regions
  - "Spikes" in larger *b*-tagging multiplicity regions
    - Events from the electron channel with high trigger pre-scales
    - Bins with low statistics containing only tight (loose) events → negative (very large) weights





→ Use a tag-rate-function (TRF) to increase statistics

Tag Rate Function (TRF)

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

What is TRF?

- Usually used to avoid having fluctuations in MC in regions with low statistics due to large *b*-tagging multiplicities (we have ≥ 4 *b*-tag jets in SR)
- Idea: Instead of removing events if they do not pass the number of required *b*-tags, consider all events, reweighted based on the **probability of the event to contain *n b*-tags**

→ Keep inclusive statistics

- Given $\epsilon(f, \eta, p_{\mathrm{T}})$ being the efficiency for a jet with η, , p_{T}, and flavour f to be *b*-tagged, the probability for an event with N jets to contain 1 *b*-tag is

$$P_{=1} = \sum_{i=1}^{N} \left(\epsilon_i \prod_{i \neq j} (1 - \epsilon_j) \right)$$

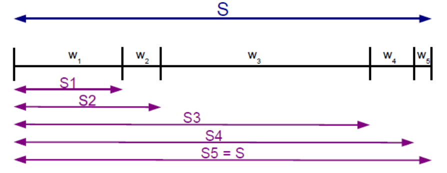
- Inclusive b-tagging regions are computed with $P_{\mathrm{incl}} = 1 - P_{=0}$

Tag Rate Function (TRF)

How to compute it

- To compute discriminating variables, *b*-tagged jets probability needs to be known
- Compute sum of TRF weights S of all permutations corresponding to the desired number of *b*-tags (e.g. 1 *b*-tag among 5 jets)
- Some of the weights w_i can be much larger than others
- Throw a random number uniformly distributed and see in which w_i segment it falls
- *b*-tag the jets based on the configuration of the i^{th} permutation

→ Probability to pick a permutation i is proportional to its TRF weight



Tag Rate Function (TRF)

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

TRF with Matrix Method

- Typically applied to MC samples where the truth flavour of a jet is known
- This information is not available for the matrix method (data-driven)
- But *b*-tagging can be used to label the jets as "*b*" or "not-*b*"
- Efficiencies are flavour-blind

$$\epsilon\left(x|N_{\text{jets}}\right) = \frac{x_{\text{b-tagged}}}{x_{\text{all}}}$$

- **Hybrid TRF approach**: Reduce possible modeling discrepancies by extrapolating from inclusive sample with *m* *b*-tags to desired *n* b-tags ($n > m$)
  - → Smaller statistics gain compared to 0 inclusive *b*-tags

→ Use matrix method to estimate fakes in a *b*-tag inclusive region

→ Apply TRF weight to estimate on desired high *b*-tagging multiplicity region

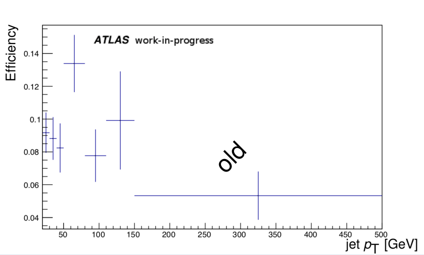
TRF Efficiency

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

$t\bar{t}H$ Analysis

- Extrapolate from at least 1 *b*-tagged jet with a WP of 85%
- Lepton η, leading jet p_T and Delta R as variables for the parametrisation of the fakes (MM)
- Without PS trigger for the low p_T region in 2016 data

- Computation of TRF efficiencies
  - For WPs of 60%, 70%, 77%, 85%
  - Hybrid approach: ≥ 1 *b*-jet
  - 4 jet inclusive
  - Rebinning
  - $\rightarrow$ Decrease uncertainty
- Different *b*-tagging efficiency parametrisations for TRF, but focus on jet p_T (most significant)
  - 00001: jet p_T
  - 00011: jet p_T and jet η
  - 01001: Delta R(jet, ℓ) and jet p_T
  - 10001: Delta ϕ(jet, E_T^{miss}) and jet p_T



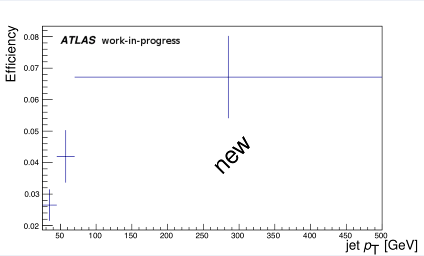
TRF Efficiency

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

$t\bar{t}H$ Analysis

- Extrapolate from at least 1 *b*-tagged jet with a WP of 85%
- Lepton η, leading jet p_T and Delta R as variables for the parametrisation of the fakes (MM)
- Without PS trigger for the low p_T region in 2016 data

- Computation of TRF efficiencies
  - For WPs of 60%, 70%, 77%, 85%
  - Hybrid approach: $\geq$ 1 *b*-jet
  - 4 jet inclusive
  - Rebinning
  - $\rightarrow$ Decrease uncertainty
- Different *b*-tagging efficiency parametrisations for TRF, but focus on jet p_T (most significant)
  - 00001: jet p_T
  - 00011: jet p_T and jet η
  - 01001: Delta R(jet, ℓ) and jet p_T
  - 10001: Delta ϕ(jet, E_T^{miss}) and jet p_T

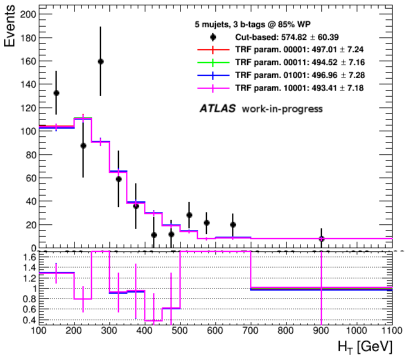


Fakes

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Results

- H_T^{had}: scalar sum of the jets' p_T
- Comparison of cut-based with TRF
- Agreement of both methods for electrons as well as muons
- Reduction of the statistical uncertainties with TRF
- Very good agreement between the different parametrisations $\rightarrow$ use jet p_T
- Removal of spikes
- Smoothens distribution



Conclusions & Outlook

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Fakes in the $t\bar{t}H$ Analysis

- Estimating fakes and non-prompt leptons is important for a good $t\bar{t}H$ measurement
- Uncertainties for TRF efficiencies are small
- Different TRF parameterisations very similar $\rightarrow$ use jet p_{T}
- TRF smoothens distributions

$\rightarrow$ Implementation for continuous *b*-tagging

Thank you for your attention!