Data-driven estimates for ATLAS ttH(bb)

John Keller
(Carleton University)

Carleton UNIVERSITY

ATLAS

Motivation

- Both ATLAS and CMS ttH(bb) analyses limited by background modelling, particularly tt+hf which is taken from MC.
- What can be done?
  - Reduce the background, e.g. with improved MVAs.
  - Collect more data, constrain the nuisance parameters further.
  - Improve uncertainty on theory inputs.
  - Reduce dependence on theory inputs via data-driven approach.

ATLAS

Uncertainty source	$\Delta\mu$	
$t\bar{t}+\geq 1b$ modeling	+0.46	−0.46
Background-model stat. unc.	+0.29	−0.31
b-tagging efficiency and mis-tag rates	+0.16	−0.16
Jet energy scale and resolution	+0.14	−0.14
$t\bar{t}H$ modeling	+0.22	−0.05
$t\bar{t}+\geq 1c$ modeling	+0.09	−0.11
JVT, pileup modeling	+0.03	−0.05
Other background modeling	+0.08	−0.08
$t\bar{t}+$ light modeling	+0.06	−0.03
Luminosity	+0.03	−0.02
Light lepton (e,μ) id., isolation, trigger	+0.03	−0.04
Total systematic uncertainty	+0.57	−0.54
$t\bar{t}+\geq 1b$ normalization	+0.09	−0.10
$t\bar{t}+\geq 1c$ normalization	+0.02	−0.03
Intrinsic statistical uncertainty	+0.21	−0.20
Total statistical uncertainty	+0.29	−0.29
Total uncertainty	+0.64	−0.61

CMS

Uncertainty source	$\pm\Delta\mu$ (observed)	$\pm\Delta\mu$ (expected)
Total experimental	+0.15/−0.16	+0.19/−0.17
b tagging	+0.11/−0.14	+0.12/−0.11
jet energy scale and resolution	+0.06/−0.07	+0.13/−0.11
Total theory	+0.28/−0.29	+0.32/−0.29
tt̄+hf cross section and parton shower	+0.24/−0.28	+0.28/−0.28
Size of the simulated samples	+0.14/−0.15	+0.16/−0.16
Total systematic	+0.38/−0.38	+0.45/−0.42
Statistical	+0.24/−0.24	+0.27/−0.27
Total	+0.45/−0.45	+0.53/−0.49

Current data-driven estimates

- Several backgrounds are already estimated with data-driven methods to a certain extent.
- Z+hf: MC with scale factor derived inside the Z mass peak.
- Fake/non-prompt leptons in dilepton channel: MC with scale factor derived from same-sign data.
- Fake/non-prompt leptons in single-lepton channel: Matrix Method.
- Multijet in all-hadronic channel: TRF_{MJ}.

The Matrix Method

- Define a Loose sample by removing some lepton quality requirements.
- Obtain the following 2 equations:

$$N^{\mathrm{l}} = N^{\mathrm{l}}_{\mathrm{r}} + N^{\mathrm{l}}_{\mathrm{f}},$$
$$N^{\mathrm{t}} = \varepsilon_{\mathrm{r}} N^{\mathrm{l}}_{\mathrm{r}} + \varepsilon_{\mathrm{f}} N^{\mathrm{l}}_{\mathrm{f}}$$

known

what we want

measured separately

- Do a little bit of algebra:

$$N^{\mathrm{t}}_{\mathrm{f}} = \frac{\varepsilon_{\mathrm{f}}}{\varepsilon_{\mathrm{r}} - \varepsilon_{\mathrm{f}}} (\varepsilon_{\mathrm{r}} N^{\mathrm{l}} - N^{\mathrm{t}})$$

- This is equivalent to giving each event a weight:

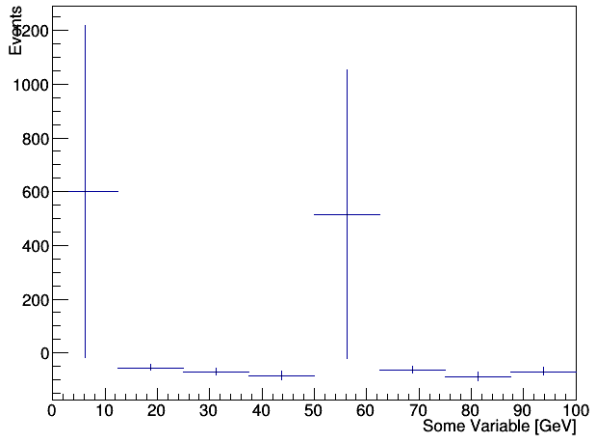
$$w_i = \frac{\varepsilon_{\mathrm{f}}}{\varepsilon_{\mathrm{r}} - \varepsilon_{\mathrm{f}}} (\varepsilon_{\mathrm{r}} - \delta_{\mathrm{i}})$$

1 if loose lepton is also tight
0 otherwise

which allows us to parametrize the efficiencies in event-level variables.

Matrix Method challenges

- There is one very large challenge with the Matrix Method: the lepton quality requirements imposed at the trigger level are not much looser than our tight selection.
- This means most loose events pass tight, and we're left with a bunch of events with small negative weights, and a handful with large positive weights, and if we're not careful plots that look like this:



- After careful study and validation, the estimate in the tightest signal regions was deemed small and consistent with 0, so the background was not considered in these regions.

The Matrix Method Reloaded

- One way to overcome this challenge is the Likelihood Matrix Method.
- Treat $\varepsilon_{r,f}$ and $N_T^{r,f}$ as parameters, and find values which maximize:

$$L(Data) = Prob(\epsilon)Prob(f)Prob(N_T)Prob(N_L)$$

where

$$Prob(N_T) = Pois(N_T|(\overline{N_T^R} + \overline{N_T^F}))$$
$$Prob(N_L) = Pois(N_L|(\overline{N_L^R} + \overline{N_L^F}))$$

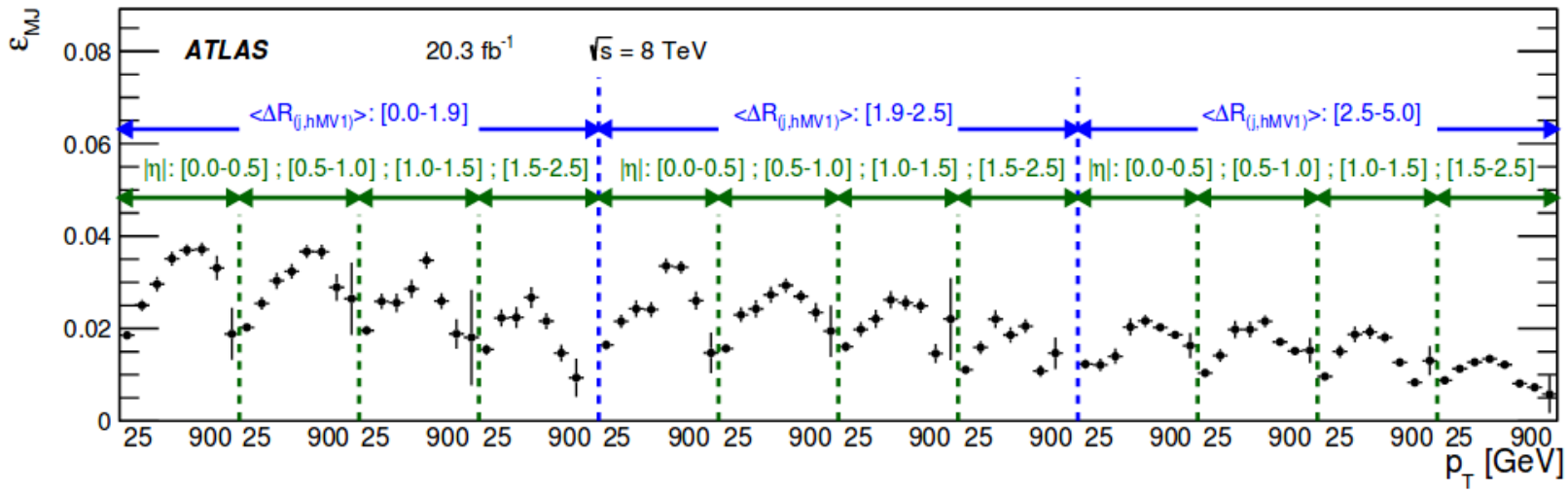
- Gives physical results and deals better with fluctuations, at the cost of additional computation needed.
- Could also consider TRF-based method presented in next slide.

- However fake/non-prompt leptons form a very small background for us: maybe it is not worth bending over backwards for them.

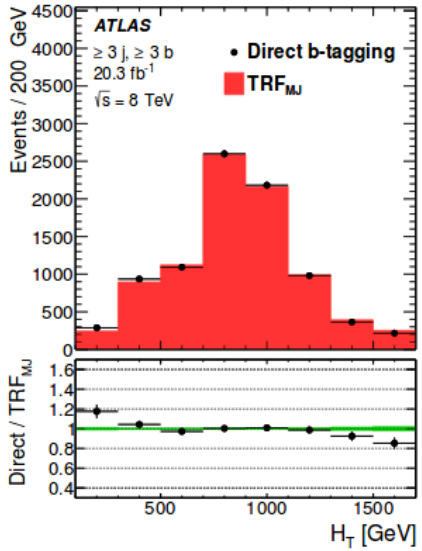
1604.03812

Multijet in all-had channel

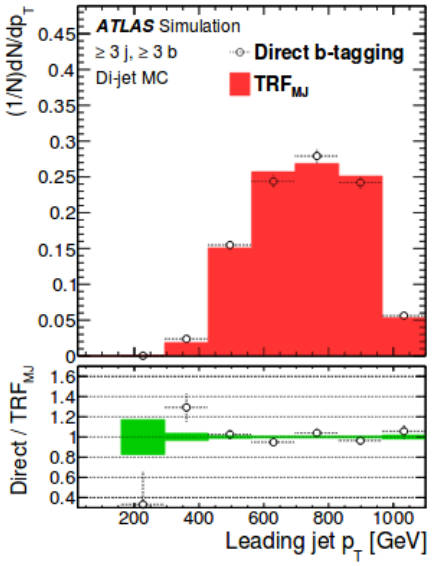
- Dominant multijet background to ttH(bb) all-had estimated using Tag Rate Function for MultiJet (TRF_{MJ}) method.
- Use multijet-enriched region with 3+ jets, 2+ b-tags, to measure the probability ε_{MJ} that a third jet is b-tagged, parametrized by η, p_T, and average distance to first two b-tagged jets.
- Apply these weights to 2-tag events at higher jet multiplicity, to get predictions for regions with 3 or more tags.



Multijet in all-had channel



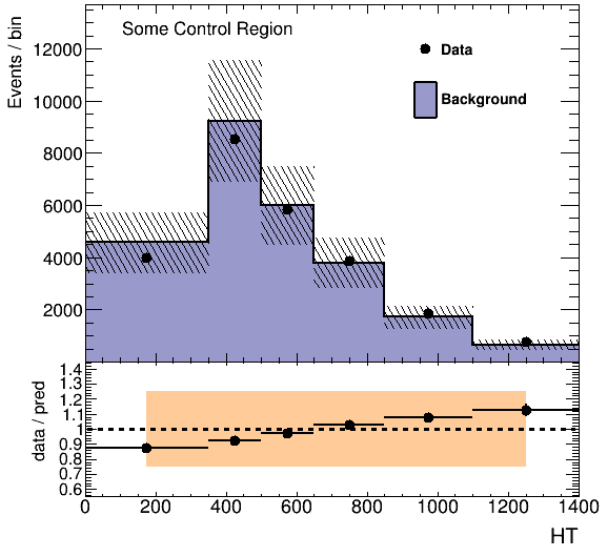
- Method is validated by comparing to direct tagging in lower jet multiplicity regions, in both data and MC.
- Systematics assessed based on alternative parametrizations, changing the set of jets used to calculate ε_{MJ}, and residual mis-modelling in the ε_{MJ} extraction region.



TRF_{MJ} predictions	Parameterisation variables in the TRF_{MJ} method
Nominal set	p_T, $\lvert\eta\rvert$, $\langle\Delta R_{(j,\mathrm{hMV1})}\rangle$
Multijet set 1	p_T, ΔR_{MV1}, $\Delta R^{\mathrm{min}}_{(j,\mathrm{hMV1})}$
Multijet set 2	p_T, ΔR_{MV1}, $\Delta R^{\mathrm{min}}_{(j,j)}$
Multijet set 3	p_T, $\lvert\eta\rvert$, $\Delta R^{\mathrm{min}}_{(j,\mathrm{hMV1})}$
Multijet set 4	p_T, $\lvert\eta\rvert$, ΔR_{MV1}, $\Delta R^{\mathrm{min}}_{(j,\mathrm{hMV1})}$
Multijet set 5	p_T, ΔR_{MV1}, $\langle\Delta R_{(j,\mathrm{hMV1})}\rangle$
Multijet lowest MV1	Nominal set removing the two lowest MV1 jets from computation
Multijet random MV1	Nominal set removing randomly two MV1 jets from computation
Multijet HT RW	Nominal set with H_T reweighting
Multijet ST RW	Nominal set with S_T reweighting

Data-driven $t\bar{t}$?

- In a sense, current tt+b/c/light estimates *are* data-driven: MC templates are given large uncertainties, and fit to data together with signal-depleted control regions.
- However this frequently leads to the Profiler's Dilemma:



- See this in a control region, what do you do?
  - Fit: and risk biasing the signal region by pulling wrong parameters.
  - Don't fit: and risk leaving a mis-modelling uncorrected.

What we talk about when we talk about data-driven

- The Profiler's Dilemma comes from a simple fact: by using our uncertainties to extrapolate from the CRs to SRs, we cannot put any uncertainty on the extrapolation.
- A data-driven approach will take as a background template in the SR(s) the data in some control region(s), with some adjustments applied for the extrapolation.
- This extrapolation may itself be derived from data or it may be from MC or a combination. The key point is that it will have its own uncertainties.

Example from a related search

- SUSY/4-top single-lepton search: counting experiment with high jet multiplicity (up to 12) and either 0 or 3+ b-tags.
- $t\bar{t}$ estimate starts from $n_{b\text{-tag}}$ distribution in 5-jet region in data.
- Assume a parametrization to extrapolate $n_{b\text{-tag}}$ distribution from j to j+1 jets:

$$N_{j,b}^{t\bar{t}+\text{jets}} = N_{j}^{t\bar{t}+\text{jets}} \cdot f_{j,b},$$

$$f_{(j+1),b} = f_{j,b} \cdot x_0 + f_{j,(b-1)} \cdot x_1 + f_{j,(b-2)} \cdot x_2$$

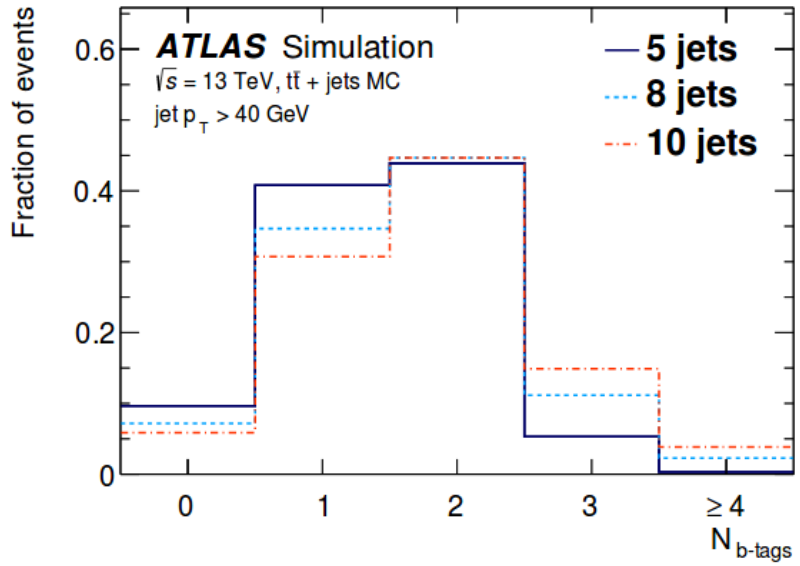
- Here $f_{j,b}$ is the fraction of *j*-jet $t\bar{t}$ events, which have *b* b-tags.
- The *x* parameters represent:
  - x_0: Probability additional jet is not tagged.
  - x_1: Probability additional jet is tagged.
  - x_2: Probability additional jet is tagged, and moves another b-jet into the acceptance.

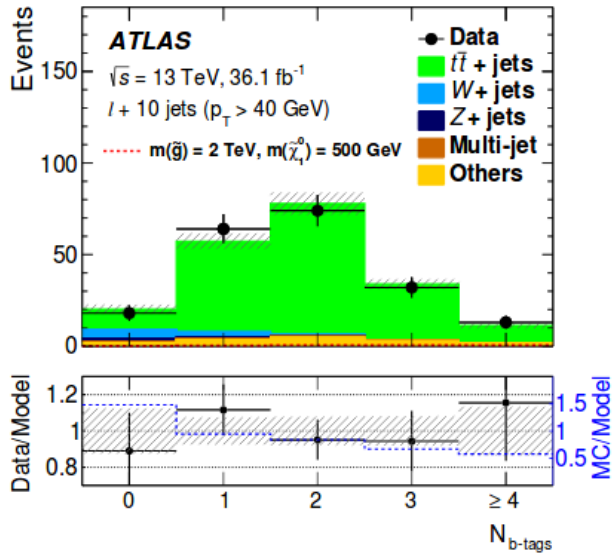
Example from a related search

- Overall normalization in each jet slice is assumed to scale with the following parametrization:

$$N_{j+1}^{t\bar{t}+\text{jets}} / N_{j}^{t\bar{t}+\text{jets}} \equiv r^{t\bar{t}+\text{jets}}(j) = c_0^{t\bar{t}+\text{jets}} + c_1^{t\bar{t}+\text{jets}} / (j + c_2^{t\bar{t}+\text{jets}})$$

- *x*, *c* parameters and the starting 5-jet $n_{\text{b-tag}}$ distribution are determined via likelihood fit to data.





Application to ttH

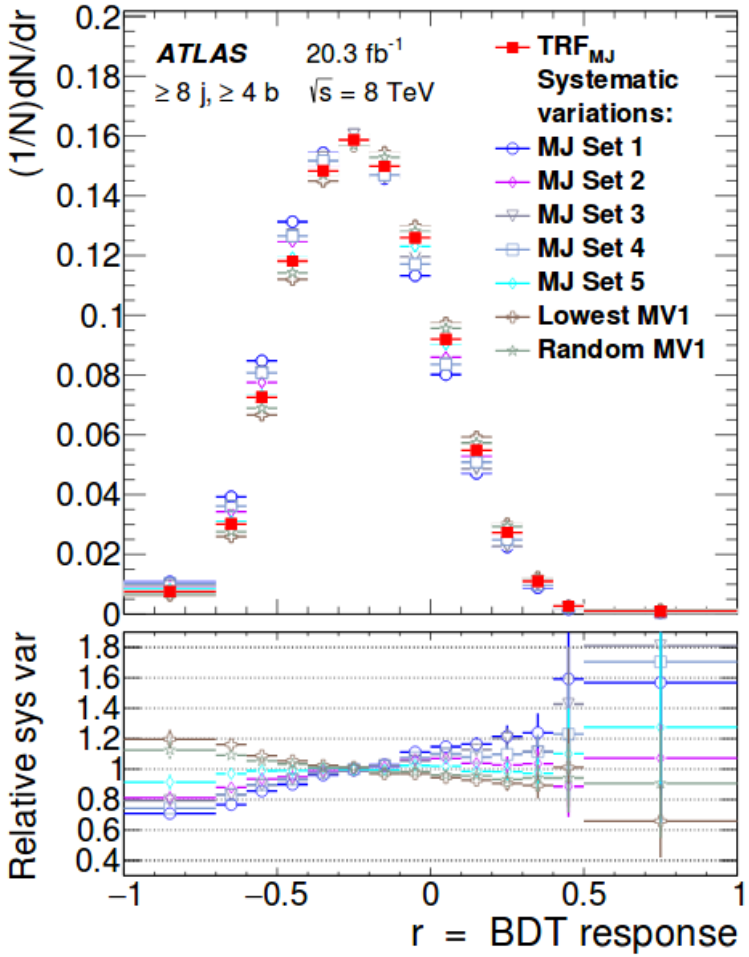
- Previous method works for counting the number of b-tags in different jet multiplicities, but not immediately applicable to kinematic variables within b-tag bins.
  - How to parametrize scaling of H_T say with additional jet?
- Could a method similar to TRF_{MJ} work for $t\bar{t}$? i.e. “promoting” un-tagged additional jets to b-jets using efficiencies derived at lower jet multiplicity.
- Alternatively use $t\bar{t}$+b/c/light ratios in MC to extrapolate from low to high b-tag regions?

Summary

- Reducing theory uncertainties: not only a job for theorists!
- Using a data-driven approach is a promising way to improve both the sensitivity and the robustness of the analysis, but non-trivial to apply to our dominant $t\bar{t}$+hf background.
- Any method is likely to rely on simulation to a certain extent, therefore these efforts are complementary to MC improvements in $t\bar{t}$+hf modelling.

Backup

TRF_{MJ} systematics



Njet scaling in tt