V+jets background modelling in ATLAS

Maria Mironova (LBNL)

The 19th Workshop of the LHC Higgs Working Group

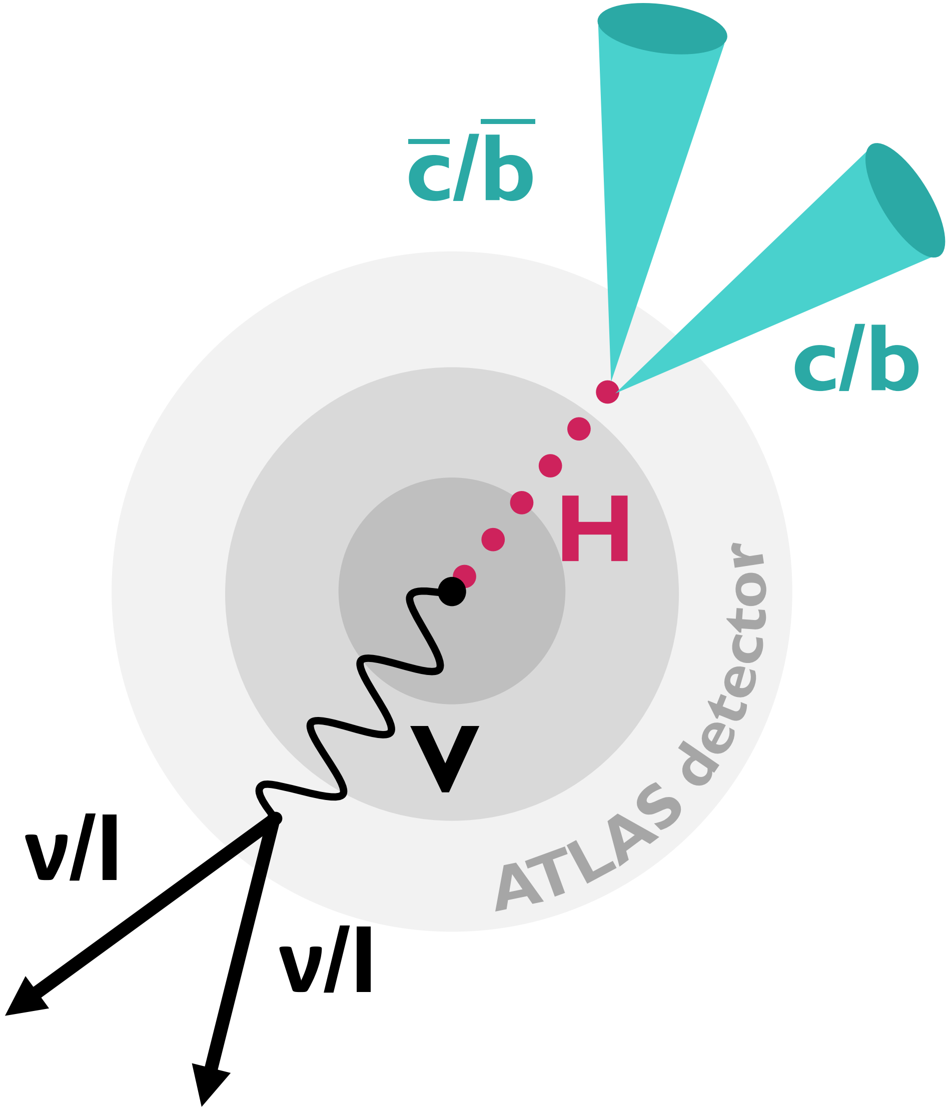
28/11/2022

BERKELEY LAB

ATLAS EXPERIMENT

Introduction

- Overview of the V+jets samples and modelling approaches commonly used by ATLAS analyses
- **V+jets is an important background in many ATLAS analysis and good modelling is crucial**
- For example: in the VH(→bb) and VH(→cc) analyses
- Brief reminder of **VH(→bb) and VH(→cc) strategy:**
  - Targeting H→bb and H→cc decays in the VH production mode
  - Categorisation into channels based on vector boson decay (Z→νν, W→lν, Z→ll)
  - Identification of b- and c-jets with the use of jet flavour tagging
  - Categorisation of events by p_T of of vector boson and jet multiplicity
  - Fit to di-jet invariant mass (in VH(cc)), or BDT distribution (in VH(bb)) to extract signal strengths or cross-sections in the STXS scheme



V+jets in VH(→bb)

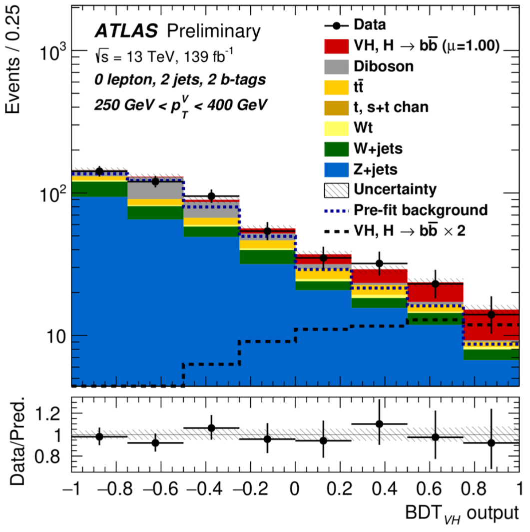
- **W+jets** and **Z+jets** backgrounds are a major background in the VH(bb) analysis, mainly W/Z+bb

→ contribution larger than 50% for most analysis regions

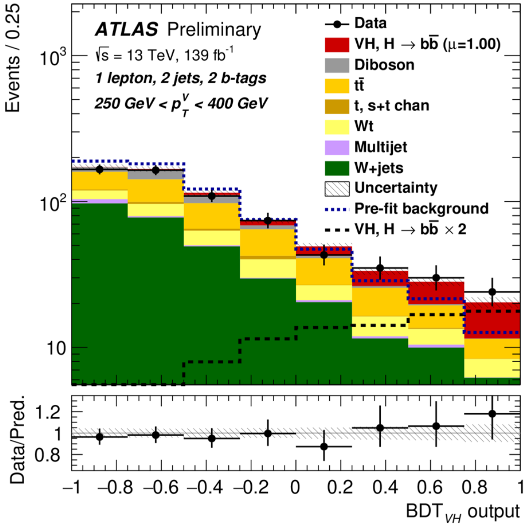
- V+jets modelling uncertainties have a sizeable contribution to the total uncertainty

→ especially W+jets is important for the WH measurement, and both W and Z+jets are important in the low p_T^V bins of the STXS measurement

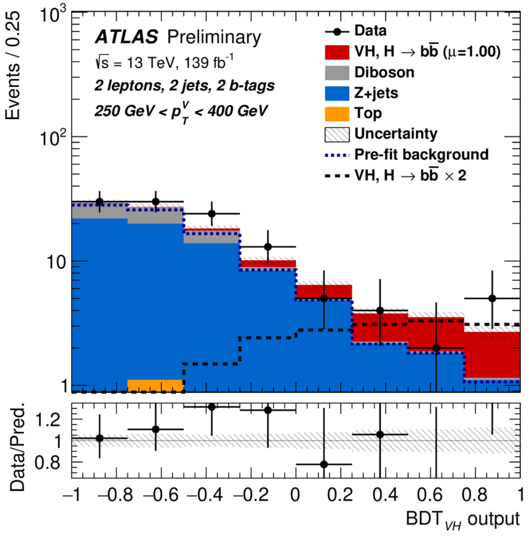
0-lepton



1-lepton



2-lepton



Breakdown of uncertainties for VH(bb) signal

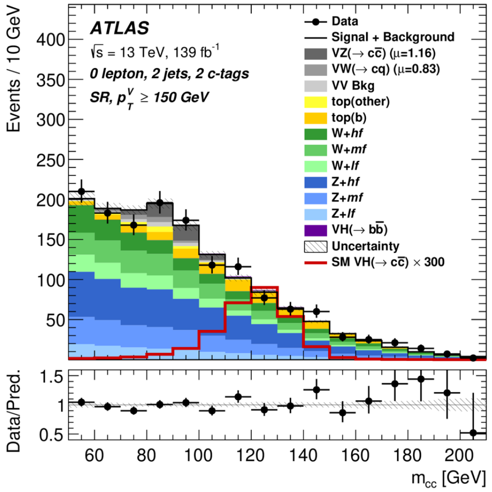
Source of uncertainty		VH (σ_μ)	WH (σ_μ)	ZH (σ_μ)
Total		0.177	0.260	0.240
Statistical		0.115	0.182	0.171
Systematic		0.134	0.186	0.168
Statistical uncertainties				
Data statistical		0.108	0.171	0.157
$t\bar{t}$ $e\mu$ control region		0.014	0.003	0.026
Floating normalisations		0.034	0.061	0.045
Experimental uncertainties				
Jets		0.043	0.050	0.057
E_T^{miss}		0.015	0.045	0.013
Leptons		0.004	0.015	0.005
b-tagging	b-jets	0.045	0.025	0.064
	c-jets	0.035	0.068	0.010
	light-flavour jets	0.009	0.004	0.014
Pile-up		0.003	0.002	0.007
Luminosity		0.016	0.016	0.016
Theoretical and modelling uncertainties				
Signal		0.072	0.060	0.107
Z + jets		0.032	0.013	0.059
W + jets		0.040	0.079	0.009
$t\bar{t}$		0.021	0.046	0.029
Single top quark		0.019	0.048	0.015
Diboson		0.033	0.033	0.039
Multi-jet		0.005	0.017	0.005
MC statistical		0.031	0.055	0.038

V+jets in VH(→cc)

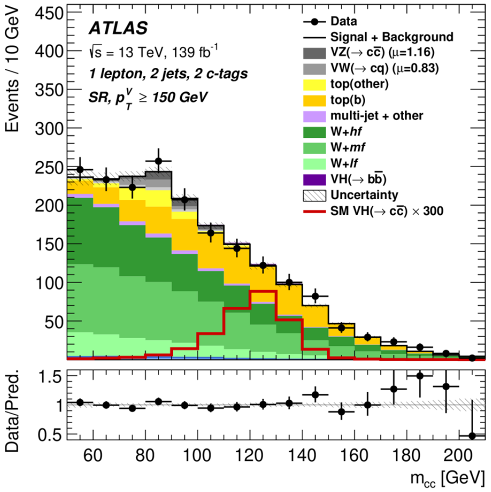
- Similar to VH(bb), V+jets also a major background in VH(cc), with a more diverse flavour composition → mainly enriched in W/Z+cc and W/Z+cl
- **Z+jets** modelling uncertainties are the leading systematic uncertainty, and **W+jets** uncertainties are also sizeable
- Additionally, due to low c-tagging efficiency, **simulation statistics** have a large impact and are mitigated through truth-tagging (details in backup)

→ As the main background, small statistical uncertainties in simulation are important for V+jets

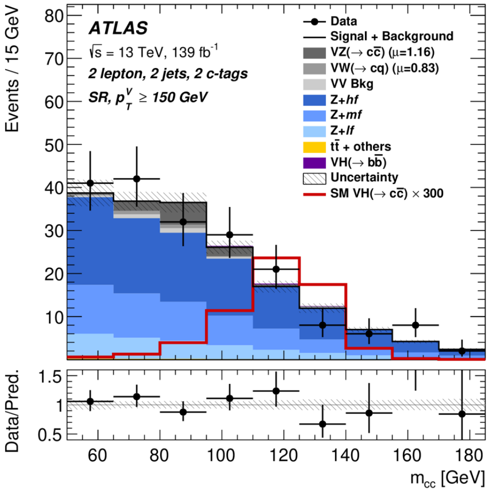
0-lepton



1-lepton



2-lepton



Breakdown of uncertainties for VH(cc) signal

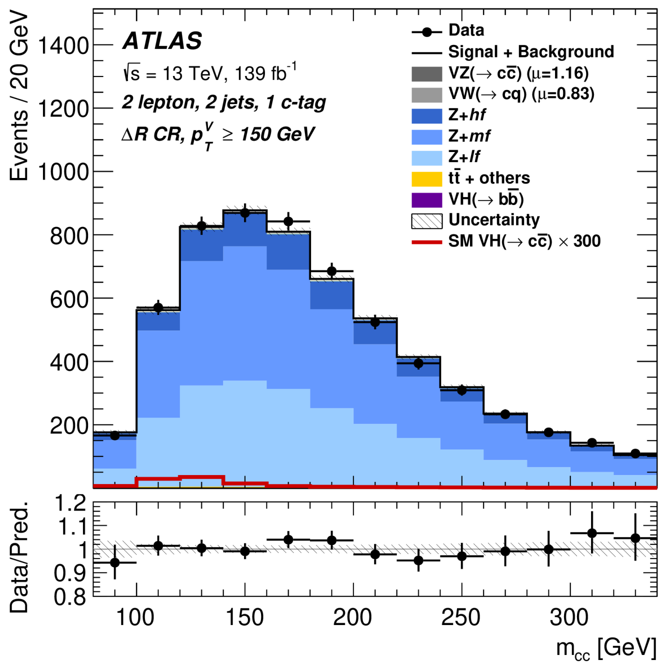
Source of uncertainty		$\mu_{VH(c\bar{c})}$
Total		15.3
Statistical		10.0
Systematic		11.5
Statistical uncertainties		
Signal normalisation		7.8
Other normalisations		5.1
Theoretical and modelling uncertainties		
$VH(\rightarrow c\bar{c})$		2.1
$Z+$jets		7.0
Top quark		3.9
$W+$jets		3.0
Diboson		1.0
$VH(\rightarrow b\bar{b})$		0.8
Multi-jet		1.0
Simulation samples size		4.2
Experimental uncertainties		
Jets		2.8
Leptons		0.5
E_T^{miss}		0.2
Pile-up and luminosity		0.3
Flavour tagging	c-jets	1.6
	b-jets	1.1
	light-jets	0.4
	τ-jets	0.3
Truth-flavour tagging	ΔR correction	3.3
	Residual non-closure	1.7

V+jets modelling approach

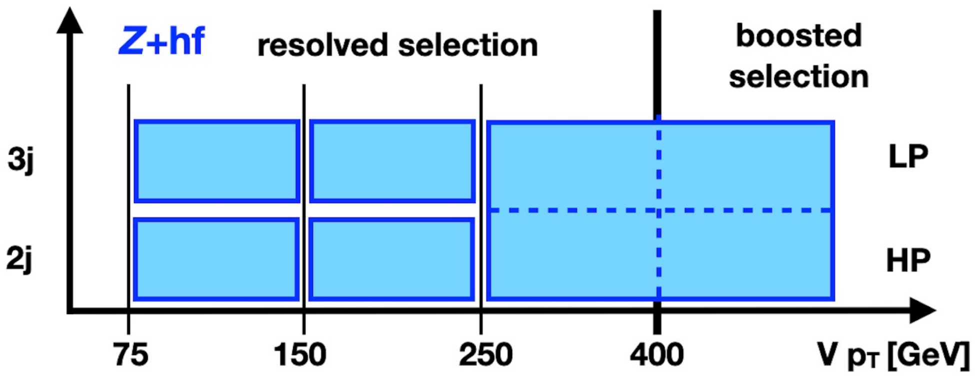
- Start from **nominal simulated samples**
  - Nominally simulated with **Sherpa 2.2.1 5F MEPS@NLO** (NLO-accurate ME for up to 2 jets, LO-accurate ME for up to four jets)
  - Samples produced in slices of max(H_T, p_T^V) to control phase space sampling
  - Filters are applied to select events with heavy flavour jets
  - More details on generator setup here
- Constrain **normalisations** (and m_{cc} shapes) of V+jets in dedicated control regions, e.g. through selecting events with high ΔR between jets
- Float normalisations based on di-jet flavour:
  - VH(bb): Float V+hf (bb,bc,bl,cc) separately and take remaining components as predicted by simulation + uncertainty
  - VH(cc): Float separately V+hf (bb,cc), V+mf (bc,bl,cl) and V+l

  → In both cases, with uncertainties applied on flavour composition
- Determine floating normalisations with as much granularity as data allows (in different bins of jet multiplicity, p_T of vector boson)

Example of V+jets control region in VH(cc)



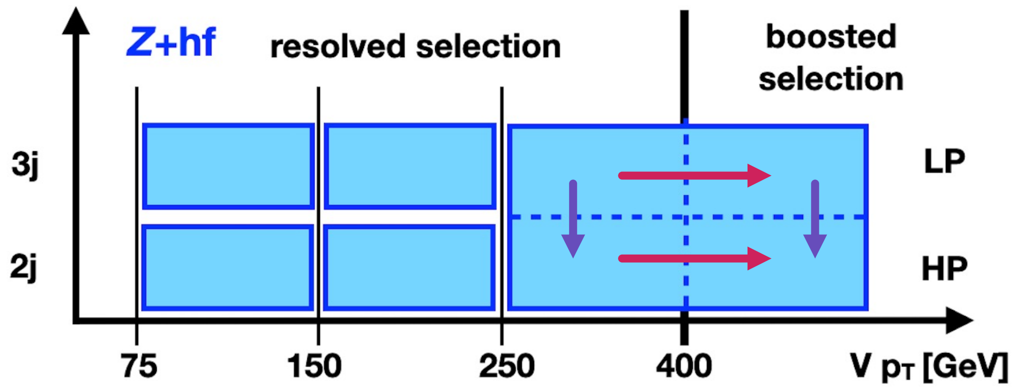
Example of floating normalisation scheme in VH(bb)



V+jets modelling approach

- Derive uncertainties by considering different variations
  - **MadGraph+Pythia8 5F MEPS@LO** (up to 4 partons) → dominant uncertainty
  - Renormalisation/factorisation scale (μ_R, μ_F) variations
  - CKKW and matching scale variation in Sherpa 2.2.1 sample → studied in VH(bb), small effect with limited statistics
- Calculate shape and normalisation effects of each alternative generator
- Group normalisation effects together, to calculate:
  - **Overall normalisation** uncertainties on smaller V+jets components
  - **Extrapolation uncertainties** between different analysis regions and on the **flavour composition** of backgrounds

*Example of floating normalisation scheme in VH(bb) → p_T^V and **jet multiplicity** acceptance uncertainties are highlighted*



Extrapolation uncertainties calculated from yields n_1 and n_2 from regions 1 and 2 (e.g. SR and CR):

$$Acc.ratio = \sqrt{\sum_i \left(\frac{\left(\frac{n_1}{n_2}\right)_i}{\left(\frac{n_1}{n_2}\right)_{nominal}} - 1 \right)^2}$$

Different sources added in quadrature

V+jets modelling approach

- **Shape uncertainties:** Consider also variations on the shapes of kinematic distributions based on the alternative samples, and include shape uncertainties in the analysis
- Different approaches possible, depending on fit discriminant and available statistics:
  - **VH(cc):** Fit uses **Higgs candidate invariant mass** as variable, so directly parametrise the ratio of nominal and alternative generators
  - **W+jets in VH(bb):** Use **BDT_R** technique → parametrise shape effect on multiple kinematic variables using BDT
  - **Z+jets in VH(bb):** Instead of using MadGraph as alternative samples, use **data-driven estimation** of shape systematic from sideband data

Illustration of shape systematics for mass-based fit

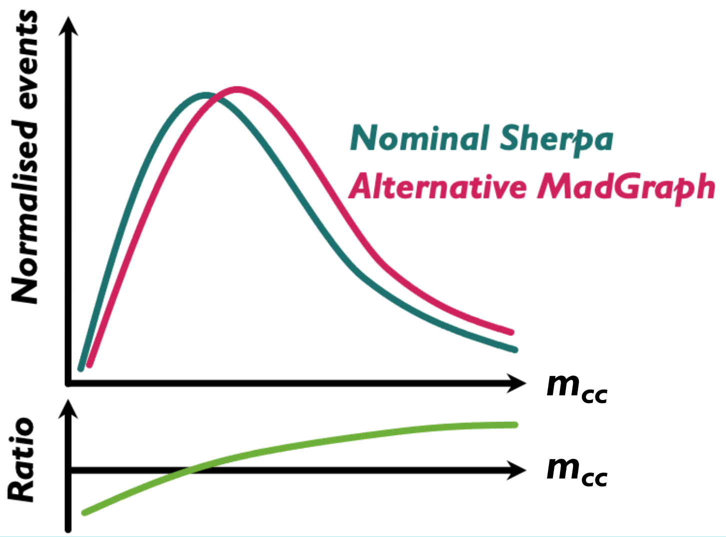
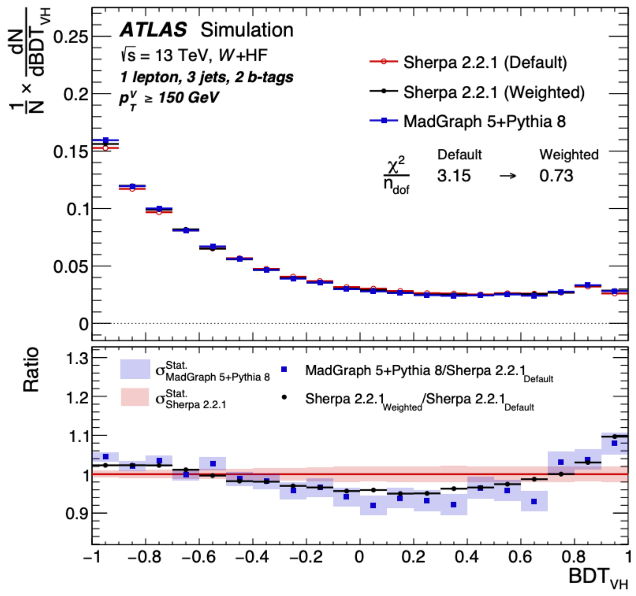


Illustration of shape systematics for BDT-based fit

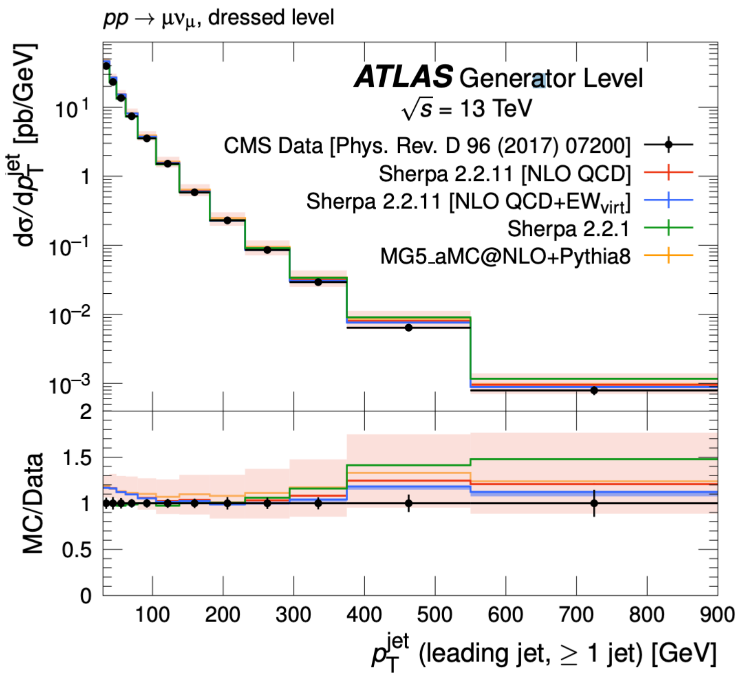


Recent developments

Several recent developments in V+jets event generation in ATLAS (details here)

- **Sherpa 2.2.11 setup** with several improvements:
  - Corrected heavy flavour hadron production fractions
  - Inclusion of higher-order QCD and EW corrections, updated EW input scheme, and additional specialised treatments
  - Additional computational improvements reduce CPU resources needed per event
- **MadGraph5_aMC@NLO+Pythia8** w/ up to 3 additional partons at NLO, using FxFx ME and PS merging prescription, is also available as an alternative generator

Comparison of different Sherpa and MadGraph setups in comparisons to W+jets data



Mean CPU time per event for Sherpa 2.2.1 and 2.2.11

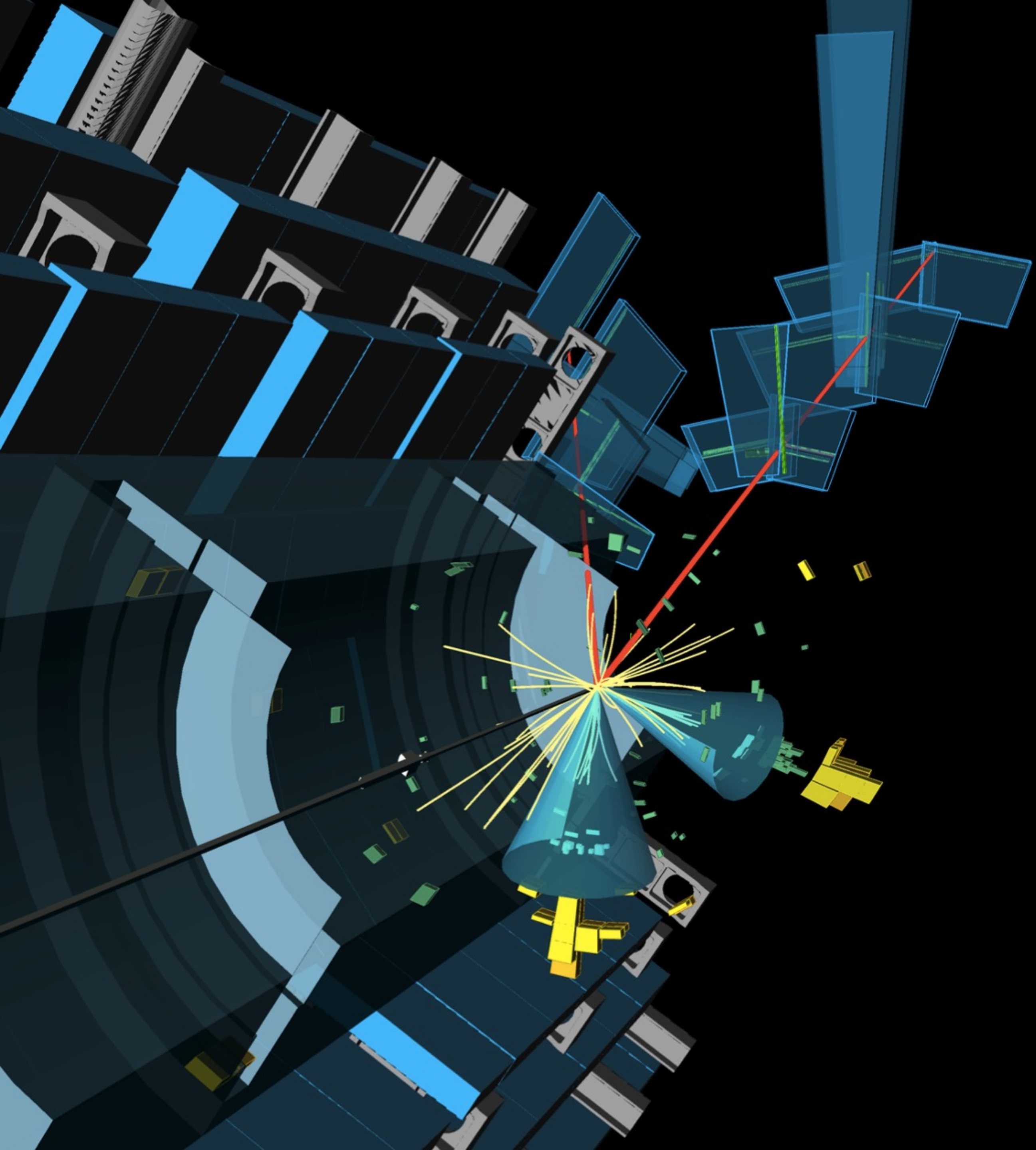
Phase-space strategy	Mean [s/event]	Mean [HS06 s/event]	Fraction of events [%]
Sherpa 2.2.11 configuration			
$\left(\frac{\max(H_T, p_T^V)}{20}\right)^2$ analytic enhancement	17.9 ± 0.2	375 ± 4	100
Sherpa 2.2.1 configuration			
$0 < \max(H_T, p_T^V) < 70$ GeV	4.7 ± 0.5	99 ± 11	31
$70 < \max(H_T, p_T^V) < 140$ GeV	34.6 ± 2.3	725 ± 48	27
$140 < \max(H_T, p_T^V) < 280$ GeV	36.8 ± 1.2	772 ± 25	19
$280 < \max(H_T, p_T^V) < 500$ GeV	53.7 ± 2.2	1126 ± 46	11
$500 < \max(H_T, p_T^V) < 1000$ GeV	67.6 ± 3.0	1418 ± 63	9
$\max(H_T, p_T^V) > 1000$ GeV	108.4 ± 5.7	2273 ± 120	3

Summary

- Accurate prediction of V+jets background is crucial for many ATLAS analysis, e.g. VH(→bb) and VH(→cc)

→ discussed the V+jets treatment in these analyses in detail

- Nominal V+jets samples are generated using **Sherpa 2.2.1 5F MEPS@NLO**
- Normalisation of main V+jets background components derived in control regions from data
- Modelling uncertainties assessed as two-point systematics using different alternative generators, e.g. **MadGraph+Pythia8 5F MEPS@LO** (dominant uncertainty), renormalisation/factorisation scale etc
- **Normalisation and acceptance effects** are considered separately from shape uncertainties and derived between analysis categories and flavour composition
- **Shape uncertainties** are derived for each source of uncertainty using different techniques (generator comparison in fitted distribution, BDTr, data-driven)
- V+jets **simulated statistics** can have a sizeable impact on analyses
- Recent work in ATLAS provides new options for V+jets generation: **Sherpa 2.2.11** and **MadGraph5_aMC@NLO+Pythia8** with theoretically motivated and computational improvements



Thank you!
Any questions?

ATLAS EXPERIMENT

Run: 303892
Event: 4866214607
2016-07-16 06:20:19 CEST

MC samples

- **V+jets:** (Details)
  - Nominally simulated with Sherpa 2.2.1
    - NLO-accurate matrix elements for up to 2 jets, LO-accurate ME for up to four jets in five-flavour scheme are calculated with Comix
    - b- and c-quarks are treated as massless
    - QCD corrections for ME @ NLO by OpenLoops
    - NNPDF3.0NNLO PDF
    - Max(H_T, p_T^V) slides with boundaries [0, 70, 140, 280, 500, 1000, 6500] GeV
  - Alternative samples simulated with MadGraph5_aMC@NLO 2.6.5
    - Showering and hadronisation with Pythia 8.240 with A14 tune and NNPDF2.3LO PDF set
    - Full 5-flavour scheme with massless quarks in ME calculation

Summary of Sherpa configurations

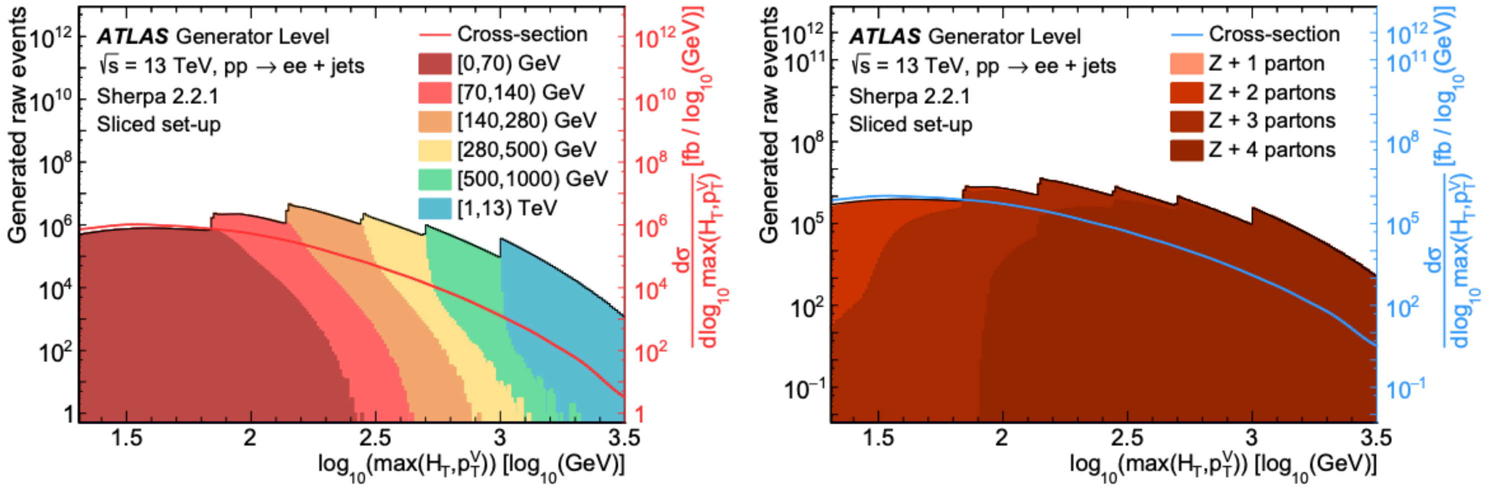
https://arxiv.org/pdf/2112.09588.pdf

Table 1: Summary of the Sherpa 2.2.1 and 2.2.11 configurations.

Configuration	Sherpa 2.2.1	Sherpa 2.2.11
Generator version	Sherpa 2.2.1	Sherpa 2.2.11
PDF set	NNPDF3.0NNLO	NNPDF3.0NNLO
EW input scheme	Effective	$\sin^2\theta_{\text{eff}}$
QCD accuracy	0–2j@NLO+3,4j@LO	0–2j@NLO+3,4,5j@LO
NLO EW_{virt} corrections	No	Yes
Subtraction scheme	Default	Modified Catani–Seymour
Special treatment for unordered histories	No	Yes
Scale for $\mathbb{H}$-events	STRICT_METS	H'_{T}
Gluon colour/spin exact matching	Yes	No
Core process for K-factor	$2 \to 4$	$2 \to 2$
Phase-space strategy	Sliced in $\max(H_{\text{T}}, p_{\text{T}}^{V})$	Analytic enhancement

Phase space sampling

https://arxiv.org/pdf/2112.09588.pdf



(a) Boundary Breakdown

(b) Jet Multiplicity Breakdown

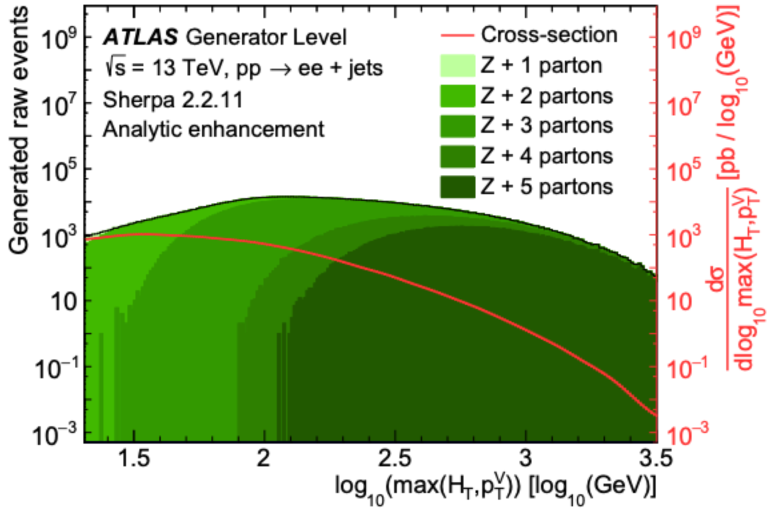


Figure 10: Distribution of unweighted $pp \to e^+e^- +$jets MC sampled events and differential cross-section for the sliced enhancement as a function of the slicing observable, $\log_{10}(\max(H_T, p_T^V))$. Events are sliced according to the $\max(H_T, p_T^V)$ variable. The unweighted distribution is split either (a) according to the phase-space sampling slices, or (b) according to the final-state jet multiplicity. In (b) the Z+1 parton contribution is small and only visible in the top left of the distribution.

MC samples

Process	ME generator	ME PDF	PS and hadronisation	Tune	Cross-section order
$qq \to VH$ $(H \to c\bar{c}/b\bar{b})$	POWHEG-BOX v2 + GOSAM + MINLO	NNPDF3.0NLO	PYTHIA 8.212	AZNLO	NNLO(QCD) +NLO(EW)
$gg \to ZH$ $(H \to c\bar{c}/b\bar{b})$	POWHEG-BOX v2	NNPDF3.0NLO	PYTHIA 8.212	AZNLO	NLO+NLL
$t\bar{t}$	POWHEG-BOX v2	NNPDF3.0NLO	PYTHIA 8.230	A14	NNLO +NNLL
t/s-channel single top	POWHEG-BOX v2	NNPDF3.0NLO	PYTHIA 8.230	A14	NLO
Wt-channel single top	POWHEG-BOX v2	NNPDF3.0NLO	PYTHIA 8.230	A14	Approx. NNLO
V+jets	SHERPA 2.2.1	NNPDF3.0NNLO	SHERPA 2.2.1	Default	NNLO
$qq \to VV$	SHERPA 2.2.1	NNPDF3.0NNLO	SHERPA 2.2.1	Default	NLO
$gg \to VV$	SHERPA 2.2.2	NNPDF3.0NNLO	SHERPA 2.2.2	Default	NLO

Event selection / modelling uncertainities

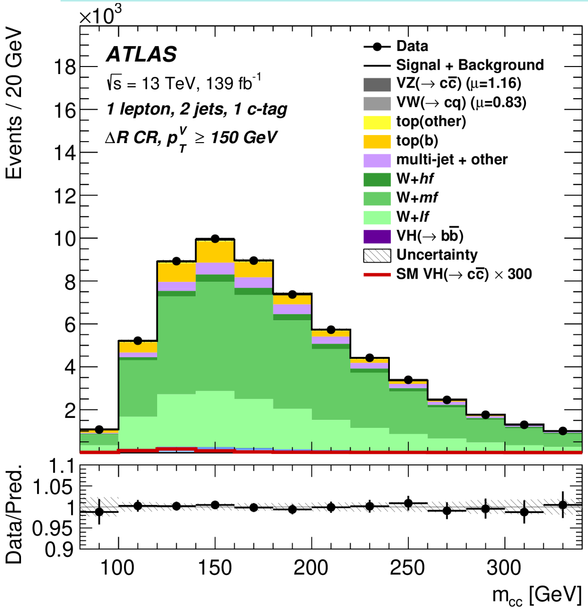
Common Selections	
Central jets	≥ 2
Signal jet p_T	≥ 1 signal jet with $p_T > 45$ GeV
c-jets	1 or 2 c-tagged signal jets
b-jets	No b-tagged non-signal jets
Jets	2, 3 (0- and 1-lepton), 2, ≥ 3 (2-lepton)
p_T^V regions	75–150 GeV (2-lepton) > 150 GeV
ΔR(jet 1, jet 2)	$75 < p_T^V < 150$ GeV: $\Delta R \leq 2.3$ $150 < p_T^V < 250$ GeV: $\Delta R \leq 1.6$ $p_T^V > 250$ GeV: $\Delta R \leq 1.2$
0 Lepton	
Trigger	E_T^{miss}
Leptons	0 *loose* leptons
E_T^{miss}	> 150 GeV
p_T^{miss}	> 30 GeV
H_T	> 120 GeV (2 jets), > 150 GeV (3 jets)
$\min \lvert\Delta\phi(E_T^{miss}, \text{jet})\rvert$	$> 20°$ (2 jets) , $> 30°$ (3 jets)
$\lvert\Delta\phi(E_T^{miss}, H)\rvert$	$> 120°$
$\lvert\Delta\phi(\text{jet1}, \text{jet2})\rvert$	$< 140°$
$\lvert\Delta\phi(E_T^{miss}, p_T^{miss})\rvert$	$< 90°$
1 Lepton	
Trigger	e sub-channel: single electron μ sub-channel: E_T^{miss}
Leptons	1 *tight* lepton and no additional *loose* leptons
E_T^{miss}	> 30 GeV (e sub-channel)
m_T^W	< 120 GeV
2 Lepton	
Trigger	single lepton
Leptons	2 *loose* leptons Same flavour, opposite-charge for $\mu\mu$
m_{ll}	$81 < m_{ll} < 101$ GeV

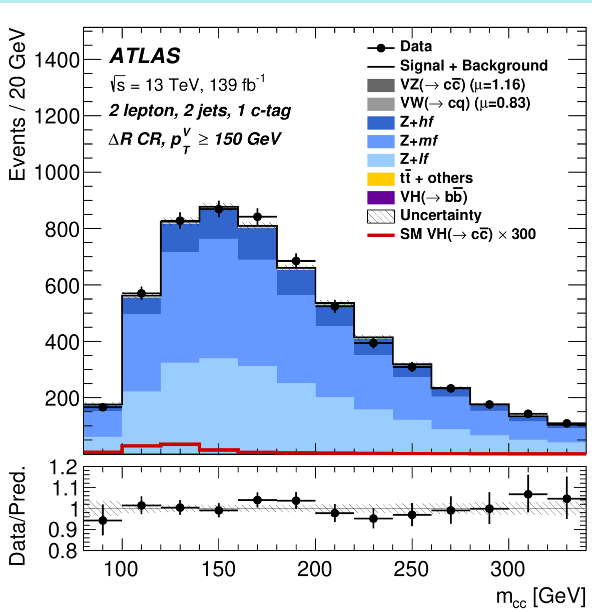
$VH(\to b\bar{b})$	
$WH(\to b\bar{b})$ normalisation	27%
$ZH(\to b\bar{b})$ normalisation	25%
Diboson	
$WW/ZZ/WZ$ acceptance	10/5/12%
p_T^V acceptance	4%
N_{jet} acceptance	7 – 11%
Z+jets	
Z+hf normalisation	Floating
Z+mf normalisation	Floating
Z+lf normalisation	Floating
$Z + bb$ to $Z + cc$ ratio	20%
$Z + bl$ to $Z + cl$ ratio	18%
$Z + bc$ to $Z + cl$ ratio	6%
p_T^V acceptance	1 – 8%
N_{jet} acceptance	10 – 37%
High ΔR CR to SR	12 – 37%
0- to 2-lepton ratio	4 – 5%
W+jets	
W+hf normalisation	Floating
W+mf normalisation	Floating
W+lf normalisation	Floating
$W + bb$ to $W + cc$ ratio	4 – 10 %
$W + bl$ to $W + cl$ ratio	31 – 32 %
$W + bc$ to $W + cl$ ratio	31 – 33 %
$W \to \tau\nu(+c)$ to $W + cl$ ratio	11%
$W \to \tau\nu(+b)$ to $W + cl$ ratio	27%
$W \to \tau\nu(+l)$ to $W + l$ ratio	8%
N_{jet} acceptance	8 – 14%
High ΔR CR to SR	15 – 29%
$W \to \tau\nu$ SR to high ΔR CR ratio	5 – 18%
0- to 1-lepton ratio	1 – 6 %
Top quark (0- and 1-lepton)	
top(b) normalisation	Floating
top(other) normalisation	Floating
N_{jet} acceptance	7 – 9%
0- to 1-lepton ratio	4%
SR/top CR acceptance ($t\bar{t}$)	9%
SR/top CR acceptance (Wt)	16%
Wt / $t\bar{t}$ ratio	10%
Top quark (2-lepton)	
Normalisation	Floating
Multi-jet (1-lepton)	
Normalisation	20 – 100%

V+jets background

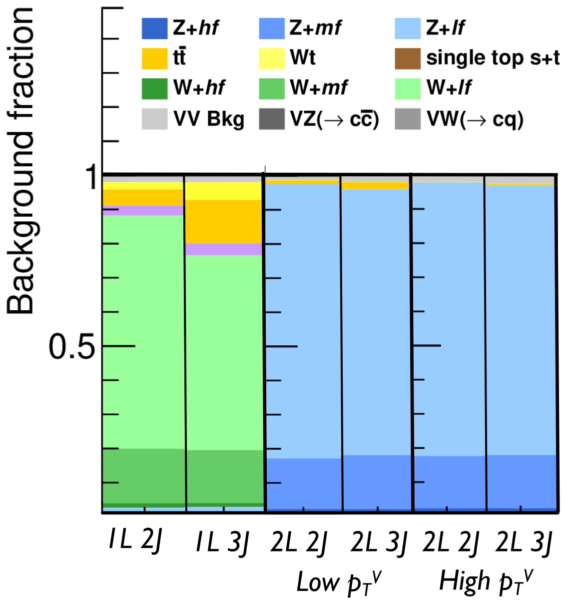
- V+jets (split as W and Z+jets) split into flavours:
  - **V+hf: V+cc,** V+bb
  - **V+mf: V+cl,** V+bc, V+bl [1]
  - **V+lf** [2]
- All V+jets normalisations floating in fit, separated as V+hf, V+mf and V+lf
- V+hf and V+mf floating normalisations determined with the help of a **high ΔR_{cc} control region**
- One ΔR_{cc} CR for each corresponding SR:
  - Low pTV: 2.3 < ΔR_{cc} < 2.5
  - Medium pTV: 1.6 < ΔR_{cc} < 2.5
  - High pTV: 1.2 < ΔR_{cc} < 2.5
- Upper cut added to stay close to SR phase space
- V+lf floating normalisations determined **in 0 c-tag CR** and 1 and 2 lepton → same kinematic selection as SR

High ΔR_{cc} control region (postfit)





0 c-tag CR

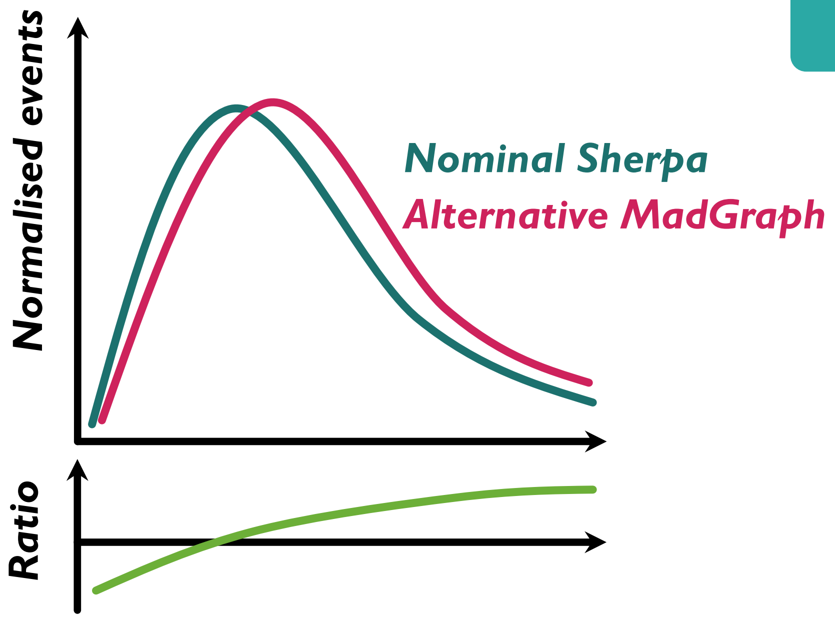


[1] including W(τν)+b, W(τν)+c in 0 lepton
[2] including W(τν)+l in 0 lepton

V+jets m_{cc} shape uncertainties

- Extrapolation from ΔR CR to SR is more complicated, as m_{cc} and ΔR_{cc} are correlated
- Two sets of shape uncertainties defined, from comparisons of Sherpa2.2.1 and MadGraph5

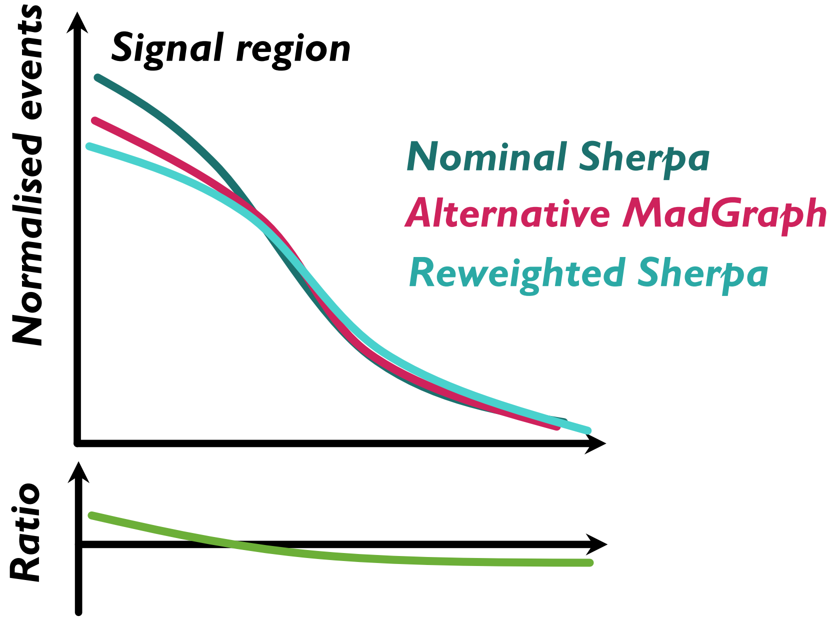
Systematic 1



- Derived in the ΔR CR, applied to SR and ΔR CR, correlated shape+normalisation effect

→ Provides constraints on m_{cc} shape in SR from ΔR CR, and takes care of acceptance effect

Systematic 2



- Reweight Sherpa MC in the SR by Syst 1 and calculate residual difference to MG
- Applied to SR only as shape-only

→ Provides additional freedom on m_{cc} shape in SR

V+jets normalisations

- All V+jets normalisations floating in fit and constrained from signal and control regions
- Common normalisations for all data-taking periods, as SRs are not split by years
- Decorrelations between n_{jet} and p_T^V regions as much as possible within the stat uncertainties
- Nominal MC generator is **Sherpa 2.2.1**
- **0- and 1-lepton:**
  - Common normalisations for all categories for W+hf and W+mf
  - Separate floating normalisations for W+lf in n_{Jet} due to 0 c-tag CR with high statistics
- **0- and 2-lepton:**
  - Floating normalisations split by p_T^V categories (low p_T^V only in 2-lepton)
  - Split normalisations in n_{Jet} for Z+lf

W+jets floating normalisations

Background	p_T^V	Jets	Value
W+hf			1.16 ± 0.35
W+mf			1.28 ± 0.35
W+lf		2	1.02 ± 0.04
		3	0.97 ± 0.05

Z+jets floating normalisations

Background	p_T^V	Jets	Value
Z+hf	>150 GeV		1.19 ± 0.22
	75-150 GeV		1.25 ± 0.25
Z+mf	>150 GeV		1.10 ± 0.15
	75-150 GeV		1.11 ± 0.15
Z+lf	>150 GeV	2	**1.07 ± 0.03**
		3	**1.08 ± 0.05**
	75-150 GeV	2	**1.12 ± 0.04**
		3	**1.07 ± 0.06**

*Most normalisations in agreement with 1 (**highlighted** otherwise)*
Similar normalisations also seen in VH(bb) with smaller uncertainties

V+jets uncertainties

- **Acceptance ratios** between channels, flavour components and jet multiplicity categories
- Comparison of Sherpa 2.2.1 and MadGraph5 and μ_R, μ_F scale variations
- **m_{cc} shape uncertainties** derived from the same sources
- Largest uncertainties from Sherpa/MadGraph comparisons, followed by μ_R scale variation

	Uncertainty	Prior
Z+jets	Z+bb to Z+cc ratio	20 %
	Z+bl to Z+cl ratio	18 %
	Z+bc to Z+cl ratio	6 %
	p_T^V acceptance	1-8 %
	n_{Jet} acceptance	10-37 %
	0-lepton/2-lepton ratio	4-5 %
W+jets	W+bb to W+cc ratio	4-10 %
	W+bl to W+cl ratio	31-32 %
	W+bc to W+cl ratio	31-33 %
	W(τν)+c to W+cl ratio	11 %
	W(τν)+b to W+cl ratio	27 %
	W(τν)+l to W+l ratio	8 %
	n_{Jet} acceptance	8-14 %
	W(τν) SR/ΔR CR ratio	5-18 %
	0-lepton/1-lepton ratio	1-6 %

Truth-tagging

- Due to moderate c-tagging efficiency (27%), the available MC statistics are significantly reduced in the VH(cc) analysis

→ Additional MC statistics, especially for V+jets, would mean a significant improvement

- Mitigation possible through the use of **truth-tagging**

→ Instead of using direct cut on flavour tagging requirements (direct tagging), weigh event based on probability of passing c-tagging

- Weights calculated from flavour tagging efficiency stored in 2D map as function of p_T and η
- Used in VH(cc) analysis to improve statistical uncertainty on simulated background events for V+jets and other backgrounds by ~ factor 3
- Also used in VH(bb) on non-b jets ("hybrid tagging")
- Closure with direct tagging not perfect → requires additional uncertainties
- Recent promising developments in truth-tagging using GNN (link paper)

c-tagging efficiency as a function of jet p_T

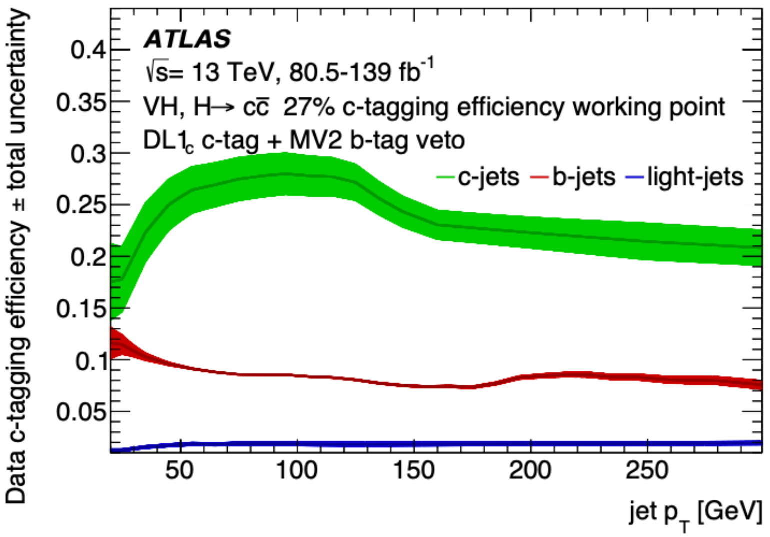
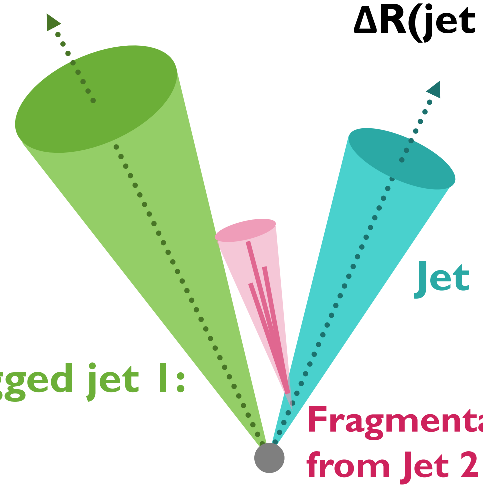


Illustration of close-by jets which can cause disagreement with direct and truth tagging



BDTr approach

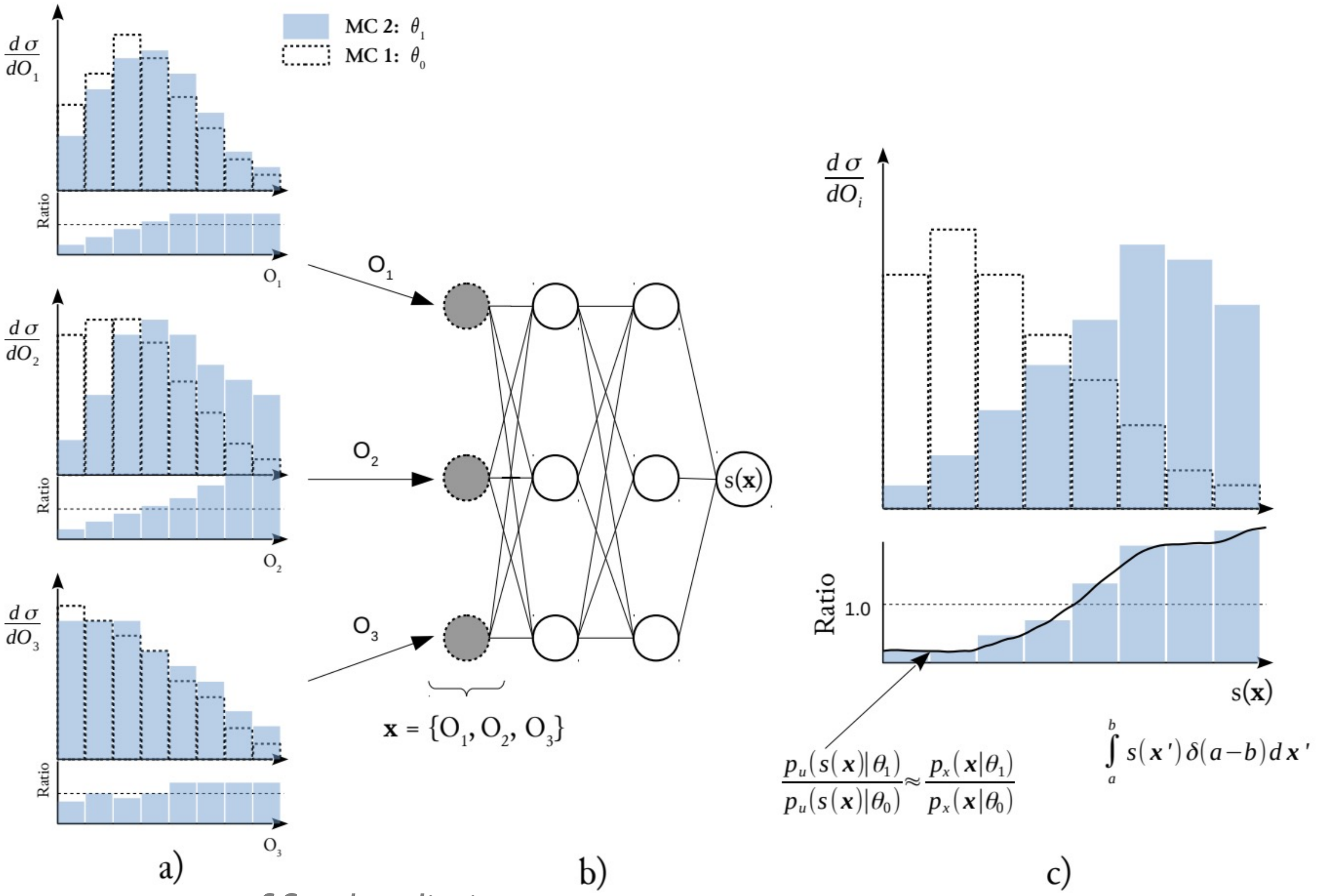


Diagram courtesy of Stephen Jiggins

BDTr approach:

1. Train BDT classifier to separate nominal and alternative MC model
2. Evaluate classifier response for both MC models
3. Parametrise ratio of classifier response for both models
4. Reweight nominal MC by parametrisation and use as systematic uncertainty

For W+jets in VH(bb), factorise p_T^V as independent shape variation due its importance in the categorisation

Data-driven approach for Z+jets

- Use data-driven approach for Z+jets modelling in VH(bb), due to high purity of 2-lepton channel
- Sum SR+CR and subtract data-driven ttbar estimate from templates and data
- Parametrise the data/MC ratio for the m_{bb} and p_T^V distributions, while excluding m_{bb} [80,140] GeV (to remove VH and Diboson)

→ Use parametrised ratio as the uncertainty