Modelling of signal and background processes: where is a better understanding needed to make progress

Adinda de Wit, Nicolas Morange

Higgs 2021 conference, October 21st

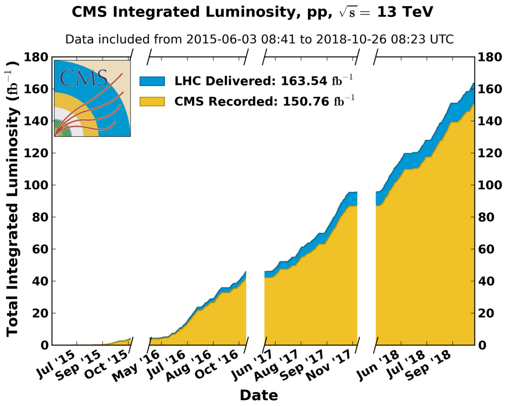
Universität Zürich UZH

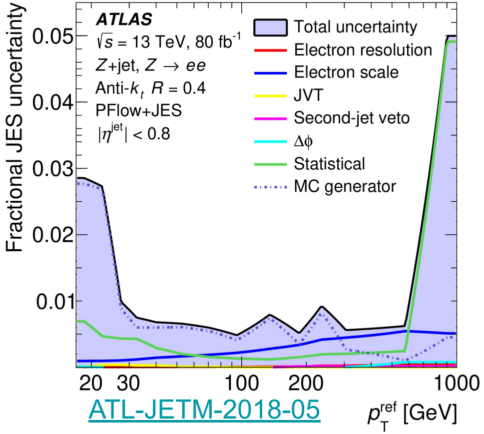
IJCLab Irène Joliot-Curie Laboratoire de Physique des 2 Infinis

cnrs

Introduction (or 'why do we care ?')

- Run 2: large datasets
  - ~140 fb^{-1} collected by ATLAS and CMS
  - Statistical uncertainties smaller and smaller
- Large datasets: precision calibrations
  - Electron and muon uncertainties at per-mille level
  - JES at sub-percent precision
  - B-tagging efficiency uncertainty at <1%
  - => Large reduction in experimental uncertainties
- Hence modelling more and more crucial topic

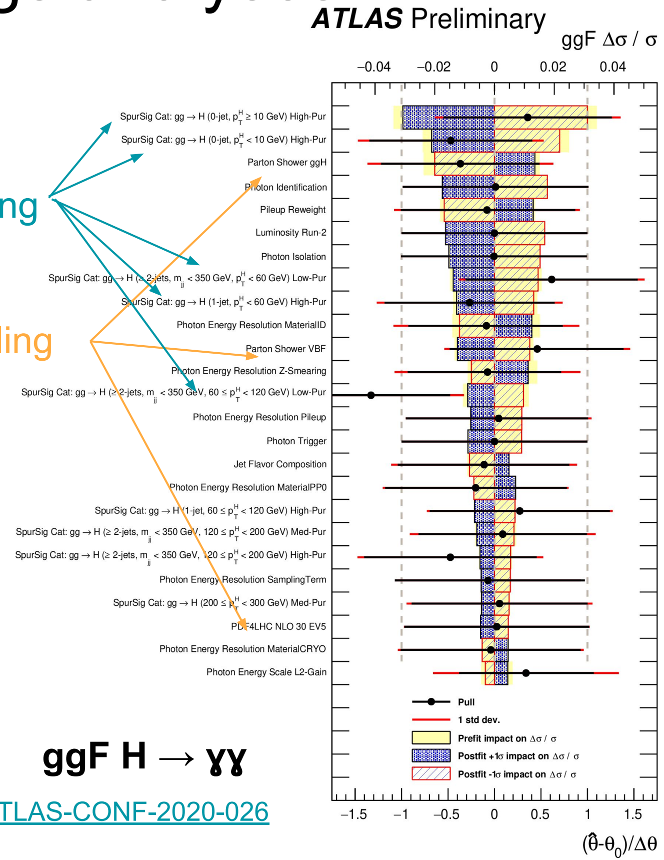




ATL-JETM-2018-05

Modelling: leading concern in all Higgs analyses

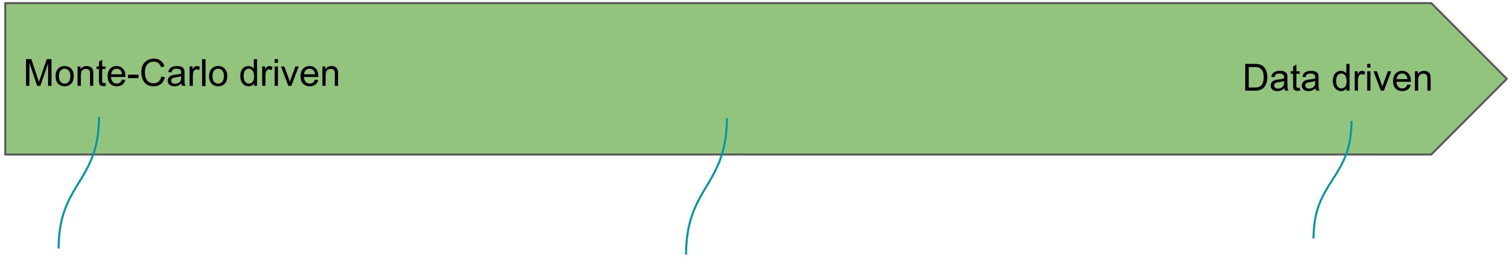
- Goal #1: good modelling out-of-the-box
  - NLO generators for ~ all processes:
    Huge success from past years
    Large effort on parameter tuning from the collaborations
  - MVA techniques require excellent modelling of correlations
- Goal #2: small modelling uncertainties
  - Easier to achieve when Goal #1 fulfilled
  - Keeping them small at the heart of analysis design
  - Lots of techniques involved
- Note: Differential measurements are not a miraculous solution
  - Stat. uncertainties dominate in STXS measurements
  - But modelling uncertainties are correlated
  - Thus important for interpretations



ATLAS-CONF-2020-026

The best Monte-Carlo is the data

Analyses make use of the data as much as possible



- Signal uncertainties
- Bkgs without good CRs

⇒ Uncertainties from MC variations or comparisons
⇒ Apply on full phase space

- Bkgs with CRs

⇒ Uncertainties from MC variations or comparisons
⇒ Constrained by profiling
⇒ Apply on extrapolation from CR to SR

- Embedding techniques
- Smooth background descriptions (e.g analytical)

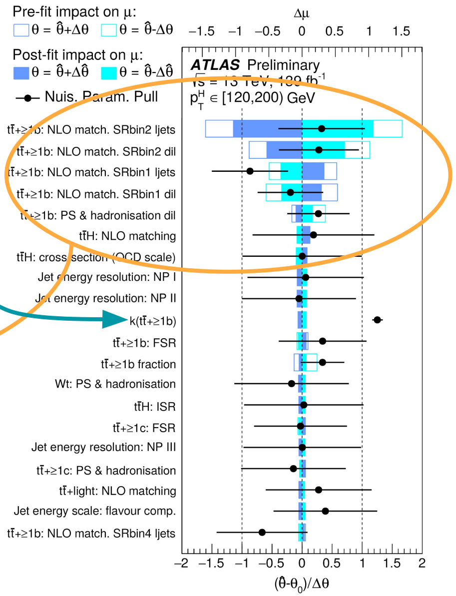
⇒ Dedicated uncertainty evaluation

Let’s explore those cases !

Background modelling

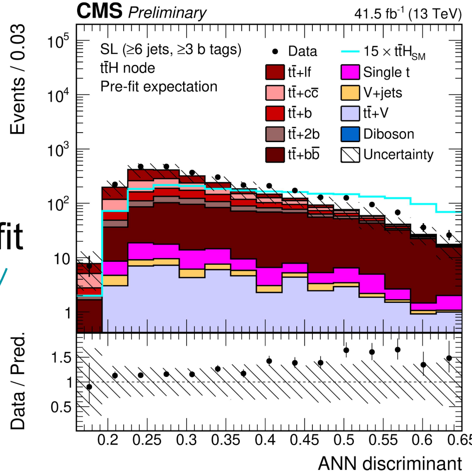
Textbook example: $t\bar{t}b\bar{b}$, for ttHbb

- $t\bar{t}b\bar{b}$ dominant bkg and low S/B
  - Complex process to model by MC
- Very large theory uncertainty
  - Cross-section well constrained by profiling, measured ~1.3x expectation
  - But ME matching and PS uncertainties give large shape/extrapolation effect
- Different setup by ATLAS/CMS but similar modelling impact:
  - ATLAS: $\Delta\mu$ = 0.25
  - CMS: $\Delta\mu$ = 0.15



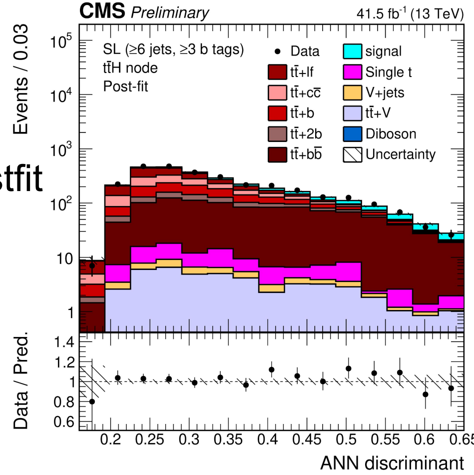
ATLAS-CONF-2020-058

Prefit



CMS-PAS-HIG-18-030

Postfit

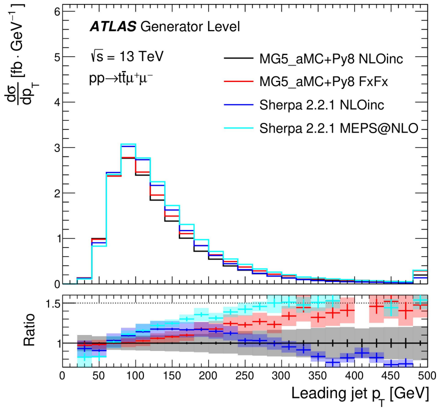


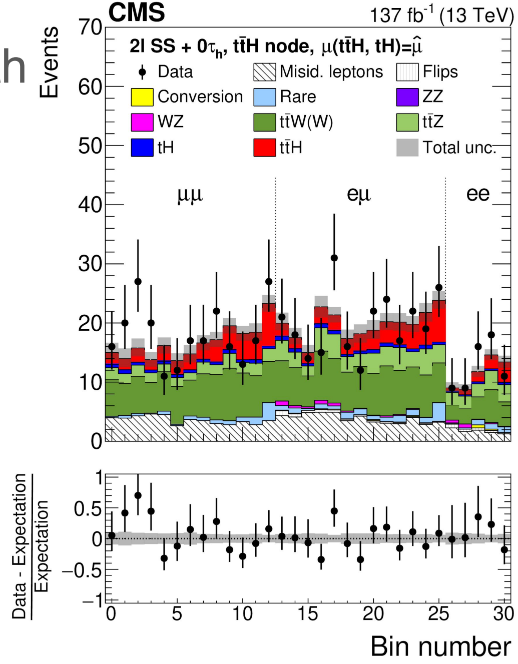
ttH in multilepton final states: ttW/ttZ

- ttH ML: complex final states with many bkgs
- ttW/ttZ leading ones
  - Description by MC complex
  - Significant differences between generators
- Extensive use of multiclass ML techniques to separate signal / bkgs and fit ttW/ttZ
  - Impact of bkg modelling contained
  - Large μ(ttW)~1.5 in ATLAS and CMS

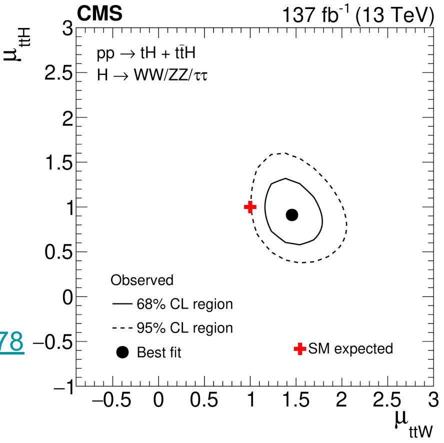
ATLAS-CONF-2019-045

ATL-PHYS-PUB-2020-024





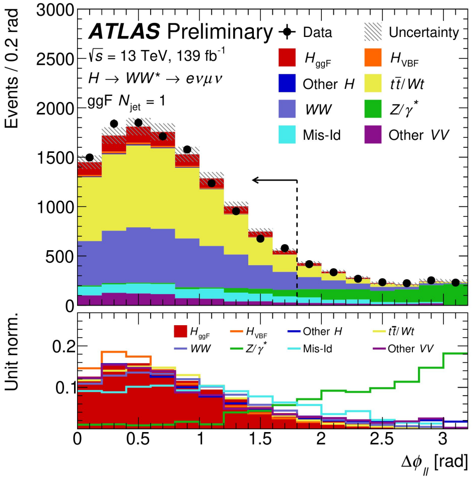
Eur. Phys. J. C 81 (2021) 378

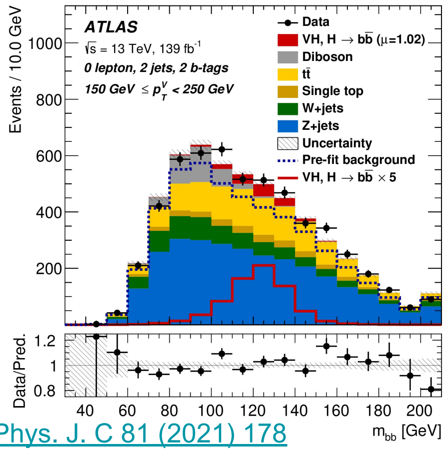


An ubiquitous background: $t\bar{t}$

- The LHC is a top factory
  - $t\bar{t}$ is a bkg to almost any final state
  - Even H→4ℓ
  - HWW: large bkg when Njet≥1, despite b-veto
  - VHbb: large bkg even in 0-lepton, 2 b-jets
- $t\bar{t}$ modelling
  - Good modelling of bulk of phase space by the NLO generators after tuning
    - Though sizable discrepancies remain in some cases
  - Difficulty: uncertainties in tails / corners of phase space
    - Not easy to get enough MC statistics:
      - filtering / slicing strategies
      - Future common ATLAS/CMS MC samples may help: ATL-PHYS-PUB-2021-016
    - Extrapolation from ‘bulk’ (CR) to ‘corner’ (SR) of phase space
    - Ambiguity between $t\bar{t}$ and Wt processes
  - Result in sizable $t\bar{t}$ modelling uncertainties in those analyses

ATLAS-CONF-2021-014

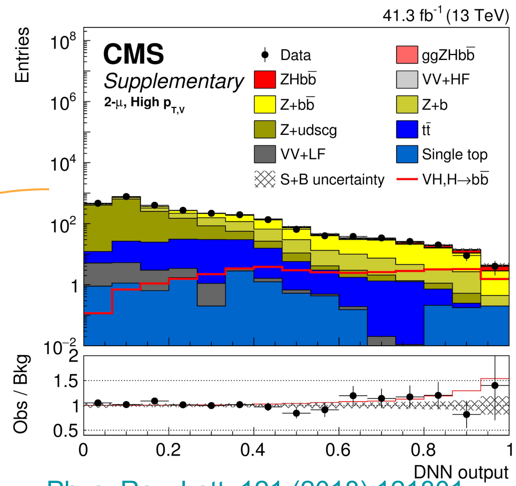




Eur. Phys. J. C 81 (2021) 178

VHbb: W/Z+hf backgrounds

- W/Z+$b\bar{b}$ largest bkgs in VHbb search
- Difficulty: generate enough MC events in relevant phase space (high pT(V)), filtered for W/Z+hf
- CMS analysis (2018) uses MadGraph LO samples
  - Reweighting in pT(V) used
  - Very large uncertainty associated
- ATLAS uses Sherpa NLO samples
  - Countless CPU hours required for MC generation
  - Filters (in)efficiency, spread of MC weights

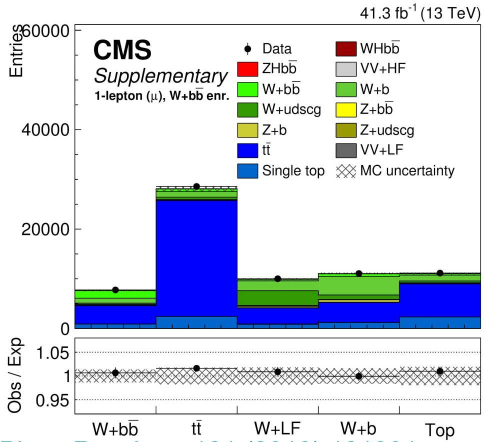


[Phys. Rev. Lett. 121 (2018) 121801](https://doi.org/10.1103/PhysRevLett.121.121801)

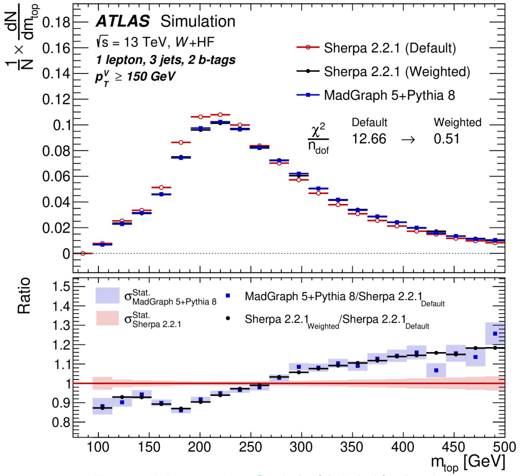
Uncertainty source	$\Delta\mu$	
Statistical	+0.26	−0.26
Normalization of backgrounds	+0.12	−0.12
Experimental	+0.16	−0.15
b-tagging efficiency and misid	+0.09	−0.08
V+jets modeling	+0.08	−0.07
Jet energy scale and resolution	+0.05	−0.05
Lepton identification	+0.02	−0.01
Luminosity	+0.03	−0.03
Other experimental uncertainties	+0.06	−0.05
MC sample size	+0.12	−0.12
Theory	+0.11	−0.09
Background modeling	+0.08	−0.08
Signal modeling	+0.07	−0.04
Total	+0.35	−0.33

VHbb: W/Z+hf backgrounds estimation

- Uncertainties constrained by profiling
  - Use of ΔRbb / mbb sidebands + multiclass BDT
  - 2-lepton: excellent control over Zbb (high purity)
  - 1-lepton: less so for Wbb ($t\bar{t}$ bkg)
- Still sizable impact from extrapolation uncertainties
  - Wbb dominant one
  - Sherpa/MadGraph difference much larger than Sherpa scale / matching variations
  - MC stat noise in uncertainty evaluation smoothed by use of ML techniques for n-dim reweighting



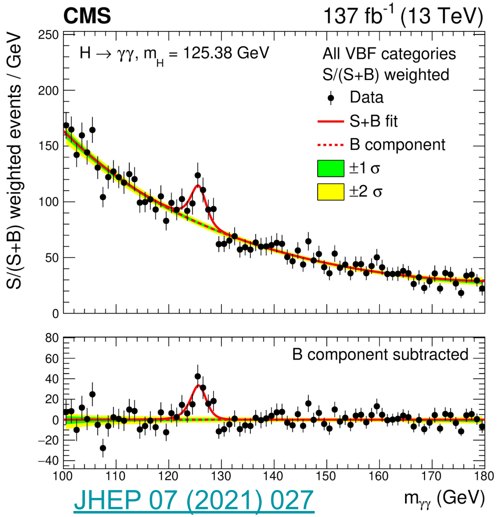
[Phys. Rev. Lett. 121 (2018) 121801](#)



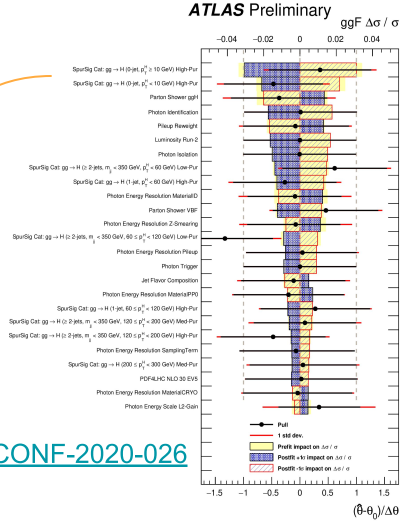
[Eur. Phys. J. C 81 (2021) 178](#)

Modelling smooth backgrounds

- Textbook H → ɣɣ example
  - Narrow resonance on top of smoothly falling bkg
  - Fit of analytical functions more accurate than ɣɣ / ɣ-jet MC samples
  - Also applies to H→μμ, H→Zɣ...
- Procedures well established since Run-1
  - CMS: Discrete profiling. Choice of function embedded in a nuisance parameter
    - Residual uncertainty very small
  - ATLAS: Select function, and estimate maximum bias ‘spurious signal’
    - Requires vast amounts of MC events
    - Limitation for high luminosity



JHEP 07 (2021) 027

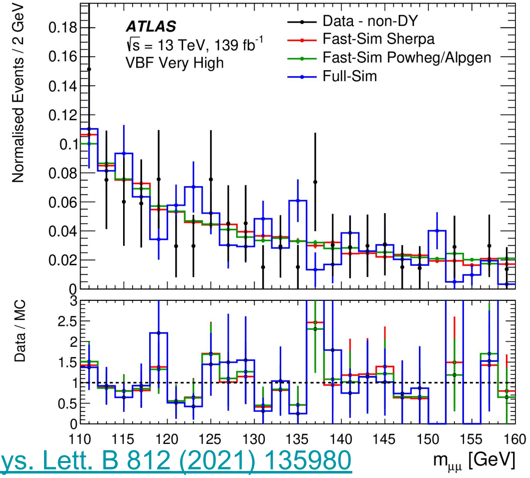


ATLAS-CONF-2020-026

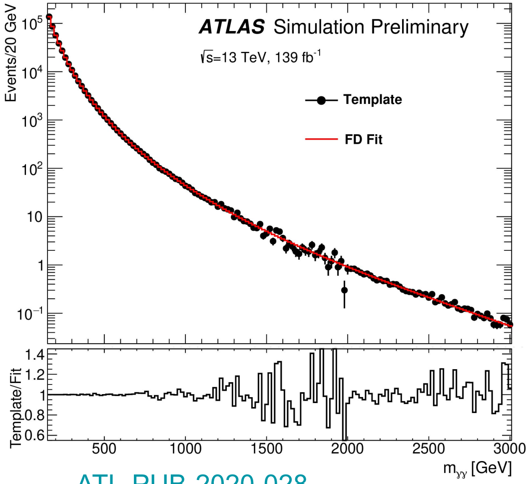
Smooth backgrounds: new techniques

ATLAS: new techniques to overcome limitations of spurious signal evaluation

- Use of very fast sim (H→μμ):
  - LO DY samples at parton-level, with parameterised detector effects
  - Spurious signal evaluated on these samples
- Functional Decomposition
  - Use series expansion to parameterize bkg shape
  - Either replacement of functional form, or use for spurious signal evaluation
- Gaussian Processes
  - Kernel encodes width of features
  - Either replacement of functional form, or use for spurious signal evaluation



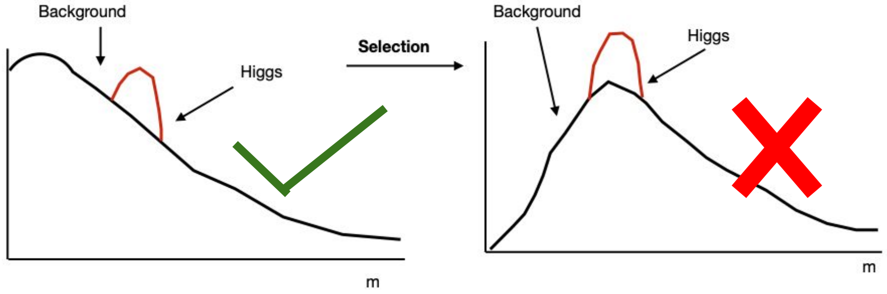
Phys. Lett. B 812 (2021) 135980



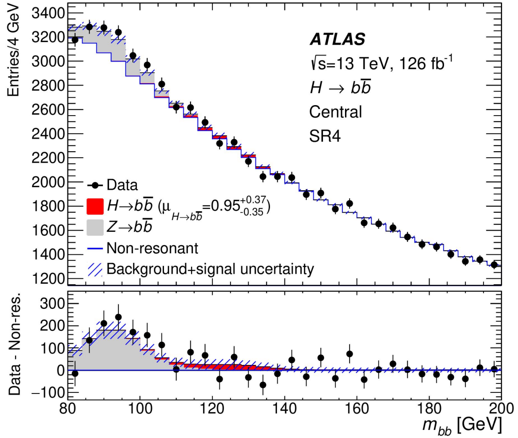
ATL-PUB-2020-028

Smooth backgrounds: sculpting

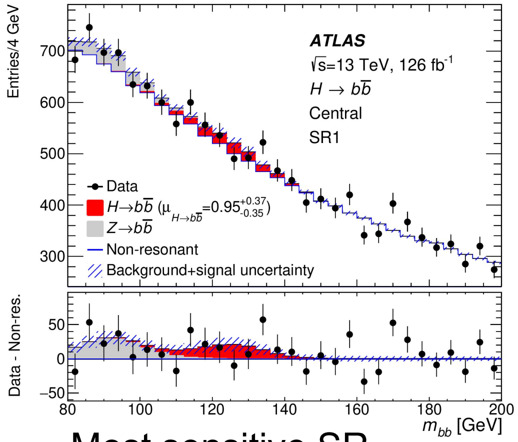
- Analysis selection should avoid sculpting background
  - Loss of sensitivity, difficulty modelling data-driven background
- Mitigation strategies in H→bb analyses
  - “Basic” selection: mass-decorrelated double-b taggers for boosted H→ bb
  - Event classification: mass-decorrelated ANN for VBF H→bb



Least sensitive SR



Similar non-resonant bkg shapes!



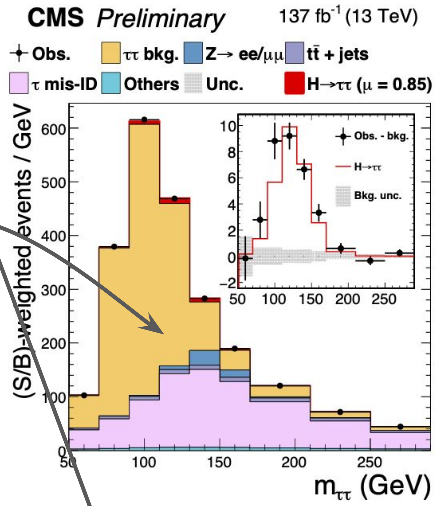
Most sensitive SR

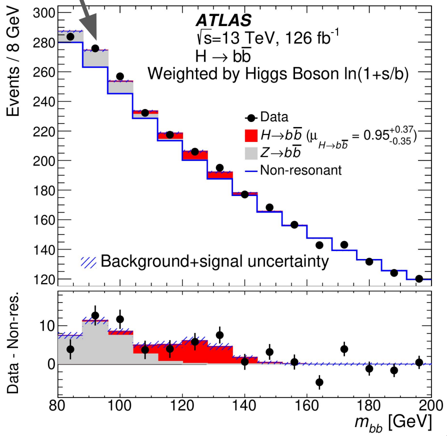
Eur. Phys. J. C.81(2021) 537

Resonant backgrounds - embedding

- E.g. Z boson decays in fermionic channels
- Same signature as the signal, except for mass = hard to model using data control regions
    - “Good” control for the background likely not signal-depleted
- MC simulation does not always adequately describe data
- Even if it does - would need very large samples to avoid large MC statistical uncertainties
- Hybrid solution: Embedding

CMS-PAS-HIG-19-010

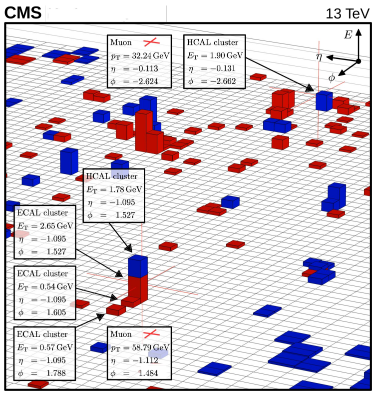




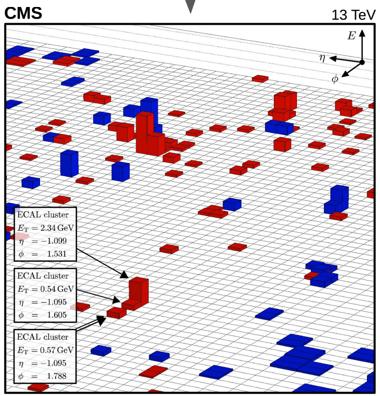
Eur. Phys. J. C.81(2021) 537

Embedding - principle

- Principle in a nutshell:
  - Select a well-understood process in data, in our case Z→μμ
  - Replace the muons by simulated particles of interest: τ's (ATLAS,CMS), b's (ATLAS)
- A simple idea?
  - Simulated/Real geometry don't match 100% → cannot merge at level of hits/deposits
    - Cannot obtain perfect closure → residual corrections
  - Spin correlations for simulated taus ignored
- Less complex procedure (re-scaling, not replacing) also in use in ATLAS (ττ)
  - Trade complexity for accuracy



Remove muon deposits

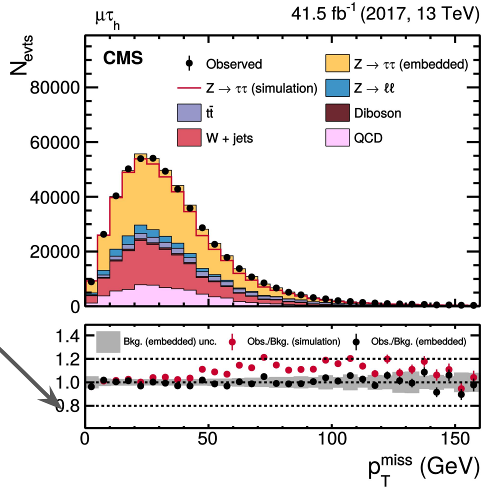


Calorimeter deposits before and after removing muon deposits

JINST 14 (2019) P06032

Embedding - achievements

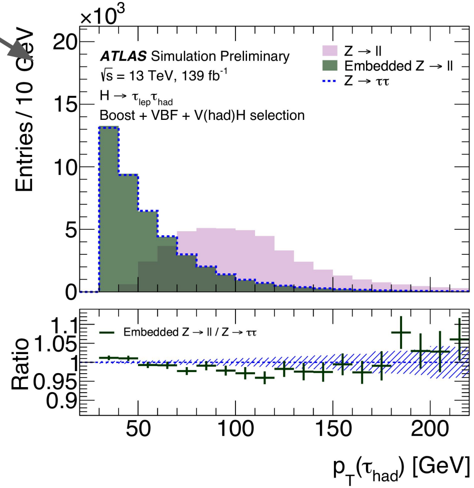
- Better modelling of kinematic distributions with embedded samples than simulation
- Helps reduce some uncertainties
- Simplified procedure provides a control region in data
- Even better modelling (smaller uncertainties?) → more work needed!



Uncertainty	$\sigma(\mu_H)$	$\sigma(\mu_{\mathrm{VBF}})$
Total statistical uncertainty	$+1.3 - 1.3$	$+1.6 - 1.5$
Data statistical uncertainty	$+0.6 - 0.6$	$+0.9 - 0.9$
Nonresonant background	$+1.0 - 1.0$	$+1.2 - 1.2$
$Z +$ jets normalization	$+0.5 - 0.5$	$+0.5 - 0.5$
Total systematic uncertainty	$+0.6 - 0.4$	$+0.6 - 0.5$
Higgs boson modeling	$+0.3 - 0.1$	$+0.2 - 0.1$
JES/JER	$+0.3 - 0.2$	$+0.4 - 0.2$
b-tagging (including trigger)	$+0.2 - 0.1$	$+0.2 - 0.1$
Other experimental uncertainty	$+0.4 - 0.3$	$+0.4 - 0.4$
Total	$+1.4 - 1.3$	$+1.7 - 1.6$

VBF H→bb analysis with 2016 data - Z+jets normalization uncertainty significant. Removed thanks to embedding (trade: 20% closure uncertainty)

Phys. Rev. D98 052003 (2018)



ATLAS-CONF-2021-044

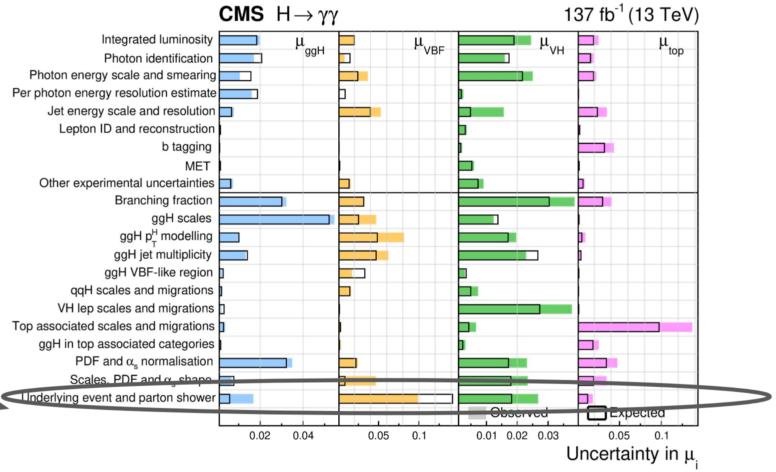
Signal modelling

Underlying event & parton shower

ATLAS-CONF-2020-026

Uncertainty source	ggF+ $b\bar{b}H$ $\Delta\sigma[\%]$	VBF $\Delta\sigma[\%]$	WH $\Delta\sigma[\%]$	ZH $\Delta\sigma[\%]$	$t\bar{t}H+tH$ $\Delta\sigma[\%]$
Underlying Event and Parton Shower (UEPS)	±2.3	±10	$<\pm1$	±9.6	±3.5
Modeling of Heavy Flavor Jets in non-$t\bar{t}H$ Processes	$<\pm1$	$<\pm1$	$<\pm1$	$<\pm1$	±1.3
Higher-Order QCD Terms (QCD)	±1.6	$<\pm1$	$<\pm1$	±1.9	$<\pm1$
Parton Distribution Function and α_S Scale (PDF+α_S)	$<\pm1$	±1.1	$<\pm1$	±1.9	$<\pm1$
Photon Energy Resolution (PER)	±2.9	±2.4	±2.0	±1.3	±4.9
Photon Energy Scale (PES)	$<\pm1$	$<\pm1$	$<\pm1$	±3.4	±2.2
Jet/E_T^{miss}	±1.6	±5.5	±1.2	±4.0	±3.0
Photon Efficiency	±2.5	±2.3	±2.4	±1.4	±2.4
Background Modeling	±4.1	±4.7	±2.8	±18	±2.4
Flavor Tagging	$<\pm1$	$<\pm1$	$<\pm1$	$<\pm1$	$<\pm1$
Leptons	$<\pm1$	$<\pm1$	$<\pm1$	$<\pm1$	$<\pm1$
Pileup	±1.8	±2.7	±2.1	±3.8	±1.1
Luminosity and Trigger	±2.1	±2.1	±2.3	±1.1	±2.3
Higgs Boson Mass	$<\pm1$	$<\pm1$	$<\pm1$	±3.7	±1.9

- Significant component of the theoretical uncertainty in several measurements, e.g. H→γγ
- Several ways in use to estimate these:
  - Difference between two showering/hadronization programs
  - Difference between a main tune and alternative tune, using the same showering/hadronization program
  - In this case: ATLAS: PY8 vs Herwig7, CMS: PY8 tune variation



JHEP 07 (2021) 027

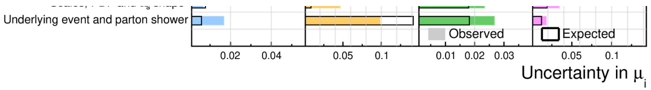
Underlying event & parton shower

H→γγ ATLAS-CONF-2020-026

Uncertainty source	ggF+ $b\bar{b}H$ $\Delta\sigma[\%]$	VBF $\Delta\sigma[\%]$	WH $\Delta\sigma[\%]$	ZH $\Delta\sigma[\%]$	$t\bar{t}H$+ tH $\Delta\sigma[\%]$
Underlying Event and Parton Shower (UEPS)	±2.3	±10	< ±1	±9.6	±3.5
Modeling of Heavy Flavor Jets in non-$t\bar{t}H$ Processes	< ±1	< ±1	< ±1	< ±1	±1.3

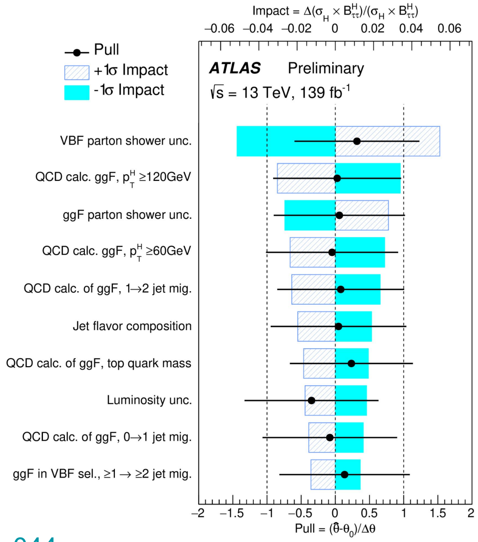
- This uncertainty is particularly large for VBF
- Leads to large theory uncertainties for VBF STXS measurements
  - For now, statistical uncertainty dominates
- Consolidating the estimation of these effects would be beneficial

H→γγ VBF JHEP 07 (2021) 027



STXS bin				SM prediction	Result	Stat. unc.	Syst. unc. [pb]		
Process	m_{jj} [GeV]	$p_T(H)$ [GeV]	N_{jets}	[pb]	[pb]	[pb]	Th. sig.	Th. bkg.	Exp.
ggF + $gg \to Z(\to qq)H$	[0, 350]♠	[60, 120]	≥ 1	0.39 ± 0.06	0.17 ± 0.39	±0.22	±0.06	±0.15	±0.29
		[120, 200]	= 1	0.047 ± 0.011	0.018 ± 0.030	±0.018	±0.004	±0.004	±0.019
	[0, 350]	[120, 200]	≥ 2	0.059 ± 0.020	0.036 ± 0.039	±0.027	±0.009	±0.009	±0.025
		[200, 300]	≥ 0	0.030 ± 0.009	0.031 ± 0.011	±0.009	±0.003	±0.001	±0.006
		[300, ∞[	≥ 0	0.008 ± 0.003	0.009 ± 0.004	±0.003	±0.001	±0.000	±0.001
	[350, ∞[	[0, 200]	≥ 2	0.055 ± 0.013	0.14 ± 0.11	±0.05	±0.06	±0.01	±0.07
EWK	[60, 120]		≥ 2	0.033 ± 0.001	0.031 ± 0.020	±0.017	±0.003	±0.001	±0.010
	[350, ∞[		≥ 2	0.090 ± 0.002	0.071 ± 0.017	±0.014	±0.010	±0.002	±0.006
$t\bar{t}H$				0.031 ± 0.003	0.047 ± 0.046	±0.032	±0.011	±0.027	±0.018

H→ττ



ATLAS-CONF-2021-044

STXS uncertainties

- Measuring STXS requires updated uncertainty model compared with inclusive measurements
- Two types of uncertainties
  - **Between** STXS bins
    - Not a measurement uncertainty when measuring cross sections
    - Enters when merging bins
    - Enters for interpretations (μ,κ, EFT)
  - **Within** STXS bins
    - Accounts for differences in acceptance

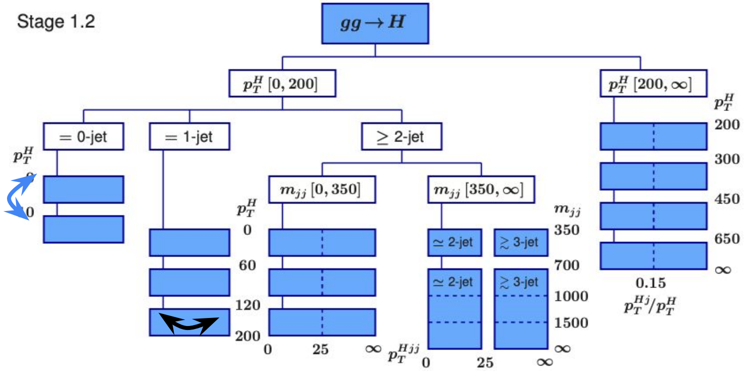
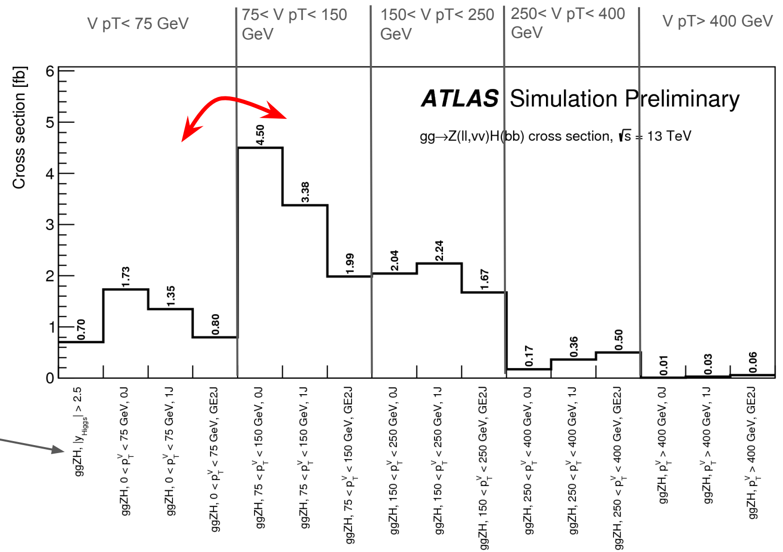


Figure from https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHWGFiducialAndSTXS

STXS uncertainties between bins

ATLAS-PHYS-PUB-2018-035

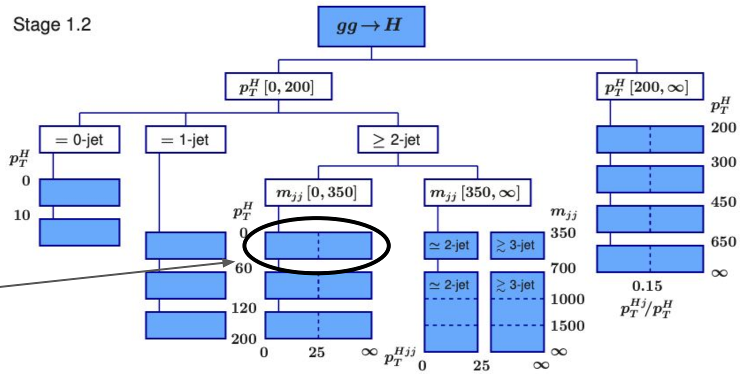
- Generally based on scale/pdf variations with uncertainties acting across bin boundary
  - E.g. change in cross section above the boundary when applying variations → uncertainty
  - Uncertainty acts across boundary (relative)
  - Difficulty in certain cases
- Important to agree on values of these → e.g. re-interpreting measurements/comparing interpretations
- Common scheme being completed in LHC Higgs WG



E.g. cross section 0-75 GeV < 75-150 GeV; migration across 75 GeV bin boundary can lead to a very large uncertainty in the first bin:
25% uncertainty above the 75 GeV boundary → 100% uncertainty below.

STXS uncertainties within bins

- Multiple possible approaches:
- Additional bin boundaries
  - Same approach as for between-bin uncertainties
  - Centralised calculation possible
  - Only captures acceptance effect across (conveniently placed) boundaries
- Within-STXS bin scale variations
  - Analysts ensure inclusive STXS bin cross section remains invariant
  - Does not necessarily encapsulate all relevant effects
- These uncertainties should be **small**
  - Does not mean “negligible”!



Phase space modelling - Higgs pT

- Modelling of Higgs boson pT spectrum particularly important for analyses looking at the boosted regime
    - Example of where recent progress has been incorporated in the analyses!
- However, large theory/modelling systematics in the ggH high pT spectrum remain → dwarfed by the statistical uncertainty in highly boosted analyses...

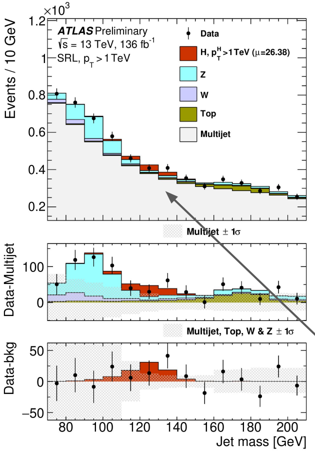
Uncertainty Contribution	$p_T^H > 450$ GeV	$p_T^H > 1$ TeV
Total	3.3	31
Statistical	2.8	30
Jet Systematics	1.2	7
Modeling and Theory Systs.	1.0	1
Flavor Tagging Systs.	0.5	3
Total Systematics	1.7	8

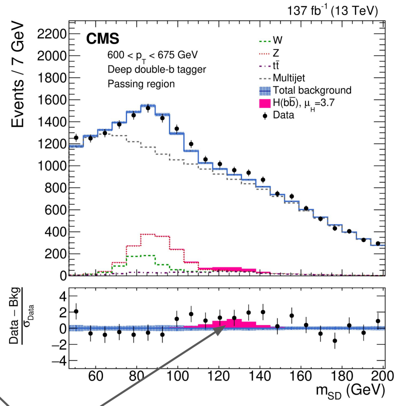
ATLAS-CONF-2021-010

HJ-MiNLO

POWHEG 1J, pT reweight

	2016	2017	2018	Combined
Expected μ_Z	$1.00^{+0.38}_{-0.28}$	$1.00^{+0.42}_{-0.29}$	$1.00^{+0.43}_{-0.29}$	$1.00^{+0.23}_{-0.19}$
Observed μ_Z	$0.86^{+0.32}_{-0.24}$	$1.11^{+0.48}_{-0.33}$	$0.91^{+0.37}_{-0.26}$	$1.01^{+0.24}_{-0.20}$
HJ-MiNLO				
Expected μ_H	$1.0^{+3.3}_{-3.5}$	1.0 ± 2.5	$1.0^{+2.3}_{-2.4}$	1.0 ± 1.4
Observed μ_H	$7.9^{+3.4}_{-3.2}$	$4.8^{+2.6}_{-2.5}$	1.7 ± 2.3	$3.7^{+1.6}_{-1.5}$
Expected H significance ($\mu_H = 1$)	$0.3\,\sigma$	$0.4\,\sigma$	$0.4\,\sigma$	$0.7\,\sigma$
Observed H significance	$2.4\,\sigma$	$1.9\,\sigma$	$0.7\,\sigma$	$2.5\,\sigma$
Expected UL μ_H ($\mu_H = 0$)	<6.8	<5.0	<4.7	<2.9
Observed UL μ_H	<8.0	<4.8	<1.7	<3.7
Ref.[23] H p_T spectrum				
Expected μ_H	1.0 ± 1.5	$1.0^{+1.1}_{-1.0}$	$1.0^{+1.1}_{-1.0}$	$1.0^{+0.7}_{-0.6}$
Observed μ_H	$4.0^{+1.9}_{-1.8}$	$2.2^{+1.4}_{-1.2}$	1.1 ± 1.1	$1.9^{+0.9}_{-0.7}$
Expected H significance ($\mu_H = 1$)	$0.7\,\sigma$	$0.9\,\sigma$	$1.0\,\sigma$	$1.7\,\sigma$
Observed H significance	$2.6\,\sigma$	$1.8\,\sigma$	$1.1\,\sigma$	$2.9\,\sigma$
Expected UL μ_H ($\mu_H = 0$)	<3.4	<2.4	<2.3	<1.4
Observed UL μ_H	<4.0	<2.2	<1.1	<1.9



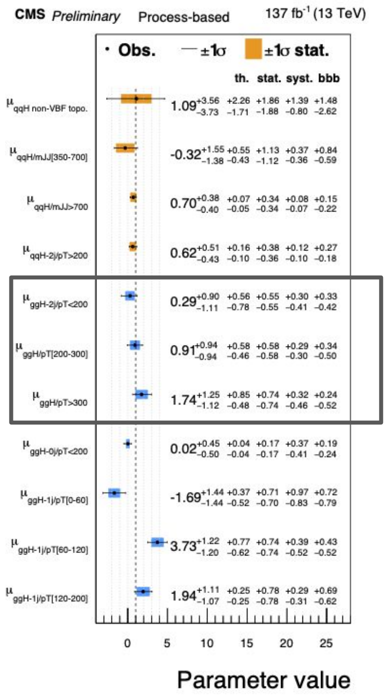


HJ-MiNLO

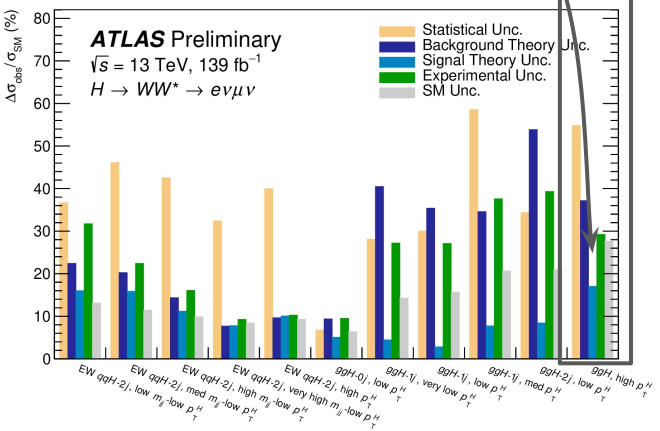
JHEP 12 (2020) 085

Phase space modelling - Higgs pT

- … but not necessarily in less boosted phase spaces - e.g. signal strength measurement ggH+2jet / high pT in H→ττ
- In H→WW STXS cross section measurements also a more important component at high pT than in other bins



[CMS-PAS-HIG-19-010]



[ATLAS-CONF-2021-014]

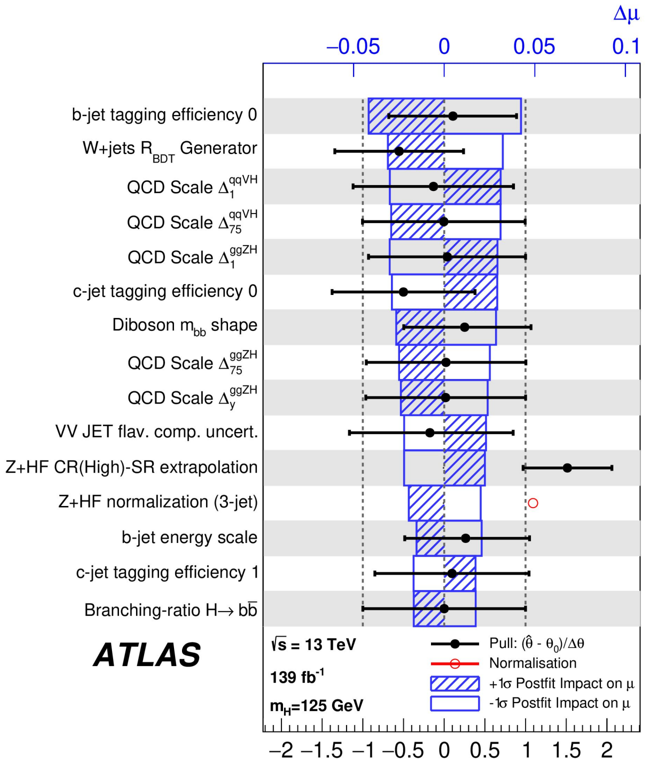
Summary

- Modelling and associated uncertainties are a major topic when going for precision measurements or measurements of low S/B processes
- Large field of analysis techniques to use data more and rely less on MC predictions
- Still, need a lot of help from our theory / MC generators colleagues
  - Simulations of complex final states ($t\bar{t}b\bar{b}$, W/Z+hf…)
  - Simulations of difficult phase space (Higgs VBF, high pT)
  - Parton shower uncertainties also a concern
  - ⇒ We want N3LO accuracy for all processes, at the speed of LO generators !

Backup

VHbb uncertainties

Source of uncertainty		σ_μ VH	WH	ZH
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
$E_{\mathrm{T}}^{\mathrm{miss}}$		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



ATLAS-HIGG-2018-51