

Background and Signal Shapes

Nicolas **Morange**, *IJCLab*

Workshop on Systematic Effects and Nuisance Parameters in
Particle Physics Data Analyses, 25/04/2023



Background and Signal shapes ?

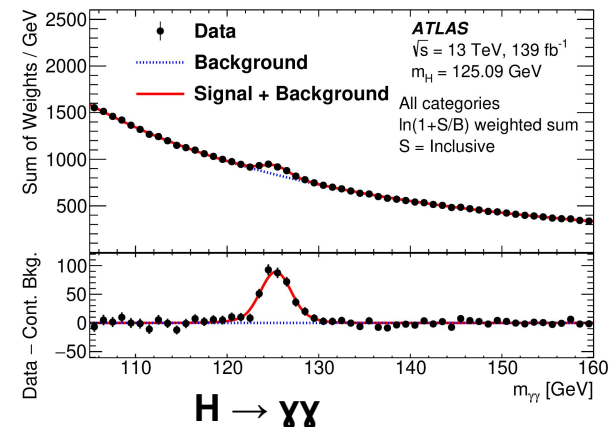
Every analysis is a (multidimensional) shape analysis

- Traditional split between “cut-and-count” and shape analyses
 - Cut-and-count: evaluate number of events in signal region after selections
 - Shape fits: fit full distribution to extract signal
 - Usually more sensitive
- Hidden shapes
 - No analysis is just 1 signal region
 - Multiple signal regions, control regions
 - Extrapolating from one region to another is a shape effect
- We need accurate signal and background shapes in all cases !

High- E_T selection	A	B	C	D
Observed data	22	7	233	131
<i>a priori</i>				
Estimated background	12.4 ± 4.7	7 ± 2.6	233 ± 15	131 ± 11
<i>a posteriori (background-only fit)</i>				
Fitted background	18.8 ± 3.5	10.2 ± 3.2	236 ± 15	128 ± 11
<i>a posteriori (signal-plus-background fit)</i>				
Fitted background	10.0 ± 6.0	5.7 ± 2.4	230 ± 15	131 ± 11
Fitted signal ($(m_\phi, m_\pi) = (600, 150) \text{ GeV}$)	12.2 ± 8.7	1.4 ± 1.0	3.4 ± 2.5	< 1

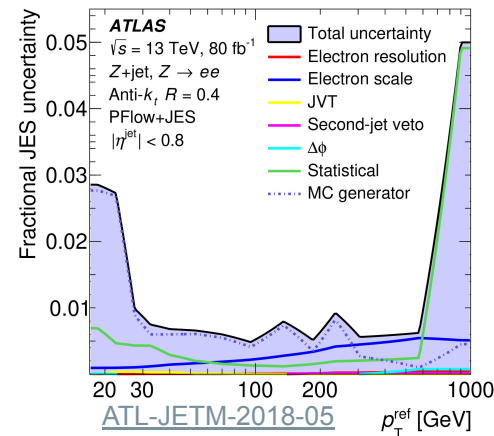
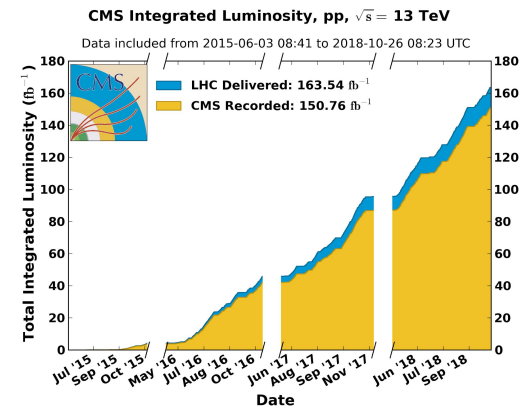
Low- E_T selection	A	B	C	D
Observed data	23	3	220	61
<i>a priori</i>				
Estimated background	10.8 ± 6.6	3 ± 1.7	220 ± 15	61 ± 7.8
<i>a posteriori (background-only fit)</i>				
Fitted background	20.6 ± 4.0	5.4 ± 2.3	222 ± 15	59 ± 7.7
<i>a posteriori (signal-plus-background fit)</i>				
Fitted background	8.4 ± 7.7	2.4 ± 1.5	217 ± 15	61 ± 7.8
Fitted signal ($(m_\phi, m_\pi) = (125, 55) \text{ GeV}$)	14.6 ± 9.9	< 1	3.2 ± 2.2	< 1

LLP “CaRatio” search



With shapes come modelling uncertainties

- Large datasets
 - $\sim 140 \text{ fb}^{-1}$ collected by ATLAS and CMS in Run 2
 - Already 40 fb^{-1} of Run 3 data
 - Statistical uncertainties smaller and smaller
- Large datasets: precision calibrations
 - Electron and muon uncertainties at per-mille level
 - Jet energy scales at sub-percent precision
 - B-tagging efficiency uncertainty at $<1\%$
 - => Large reduction in experimental uncertainties
- Therefore signal and background shapes need to be known with adequate precision
 - Meaning small modelling uncertainties

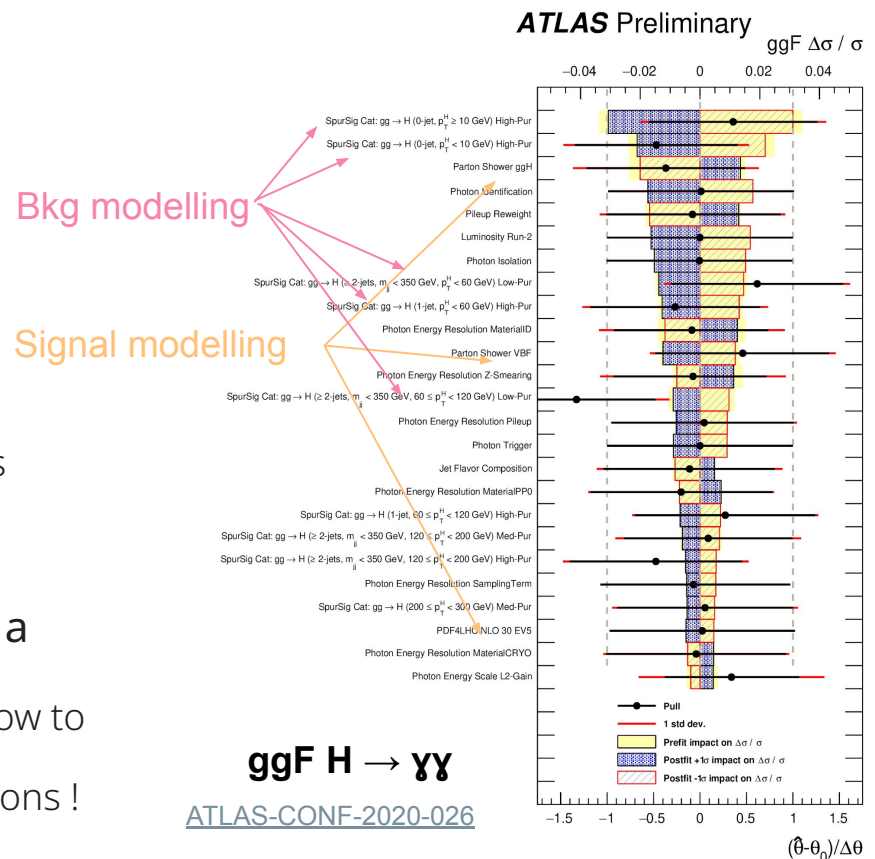


Modelling: leading concern in many analyses

- **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/ML 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**
 - Fine enough differential measurements allow to get rid of signal modelling uncertainties
 - But uncertainties come back in interpretations !



The best Monte-Carlo is the data

Analyses make use of the data as much as possible

Theory / Monte-Carlo driven

- Signal uncertainties
 - Bkgs without good CRs
- ⇒ Uncertainties from MC variations or comparisons
⇒ Apply on full phase space
⇒ See presentations by **G. Jones** and **F. Tackmann**

- Bkgs with good CRs
- ⇒ Uncertainties from MC variations or comparisons
⇒ Constrained by profiling
⇒ Apply on extrapolation from CR to SR
⇒ See e.g presentations on Optimal Transport by **T. Manole** and **P. Windischhoffer**

Data driven

- Embedding techniques
 - Smooth background descriptions (e.g analytical)
- ⇒ Dedicated uncertainty evaluation

Slides heavily based on a presentation given at Higgs 2021 jointly with **Adinda De Wit** (LLR)
Credits to her !!

Full spectrum of techniques to get shapes and uncertainties

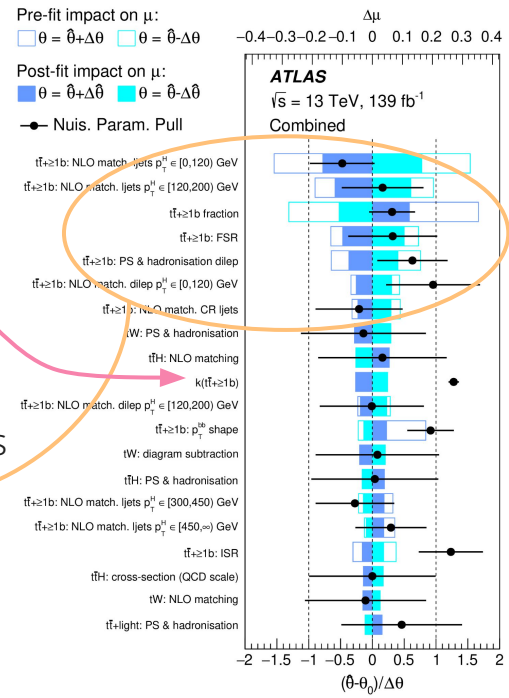
Background shapes

MC-based textbook example: $t\bar{t}b\bar{b}$, for $t\bar{t}Hb\bar{b}$

- $t\bar{t}b\bar{b}$ dominant bkg and low S/B
 - Complex process to model by MC
 - Control Regions not enough

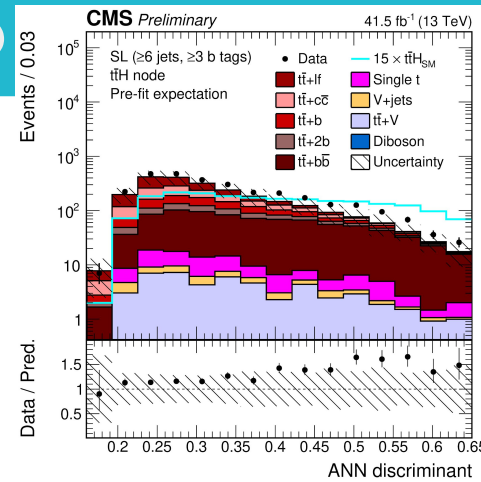
- Very large theory uncertainty
 - Cross-section well constrained by profiling, measured $\sim 1.3x$ expectation
 - Modelling systematics == collection of 2-point systematics
 - ME matching and PS uncertainties esp. give large shape/extrapolation effect

- Different setup by ATLAS/CMS but similar modelling impact:
 - ATLAS: $\Delta\mu = 0.25$
 - CMS: $\Delta\mu = 0.15$



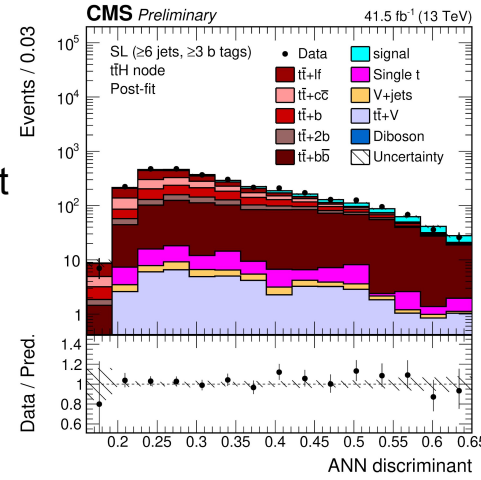
JHEP 06 (2022) 97

Prefit



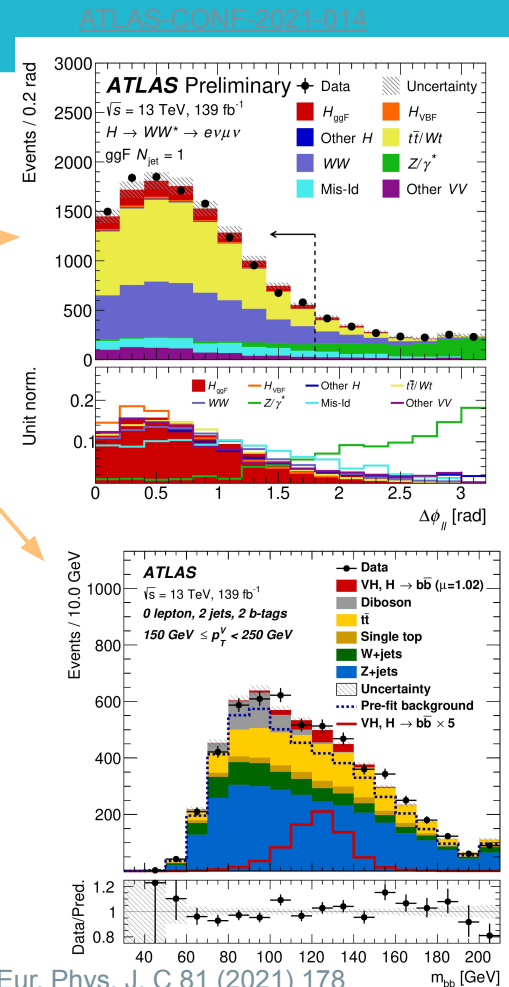
CMS-PAS-HIG-18-030

Postfit



Good modelling everywhere is hard: $t\bar{t}$

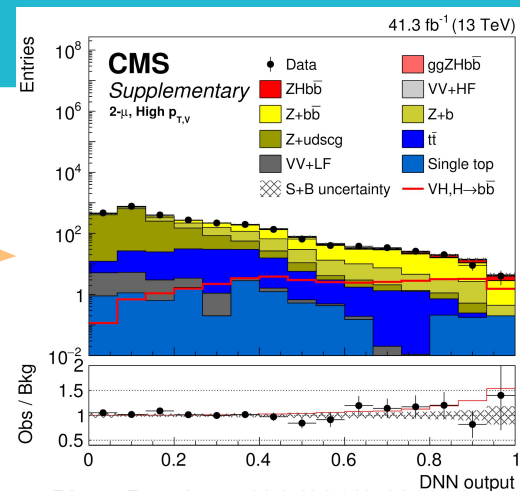
- The LHC is a top factory
 - $t\bar{t}$ is a bkg to almost any final state
 - Limited experimental efficiencies (b-veto)
 - Weird corners of the phase space (acceptance)
- $t\bar{t}$ modelling
 - Good modelling of bulk of phase space by the NLO generators after tuning
 - Though sizable discrepancies 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



VHbb: W/Z+hf backgrounds

Good MC modelling: costly but worth it

- 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
 - Large uncertainty associated
- ATLAS uses Sherpa NLO samples
 - Countless CPU hours required for MC generation
 - Filters (in)efficiency, spread of MC weights



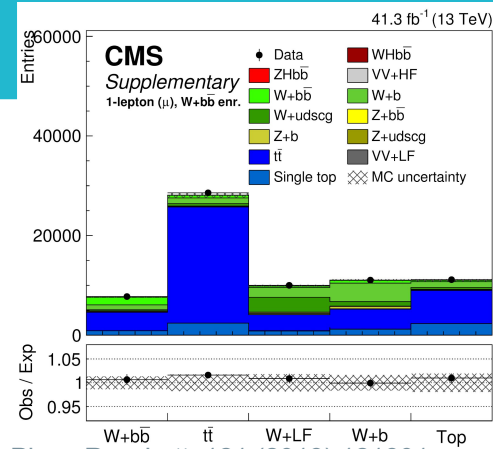
[Phys. Rev. Lett. 121 \(2018\) 121801](#)

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

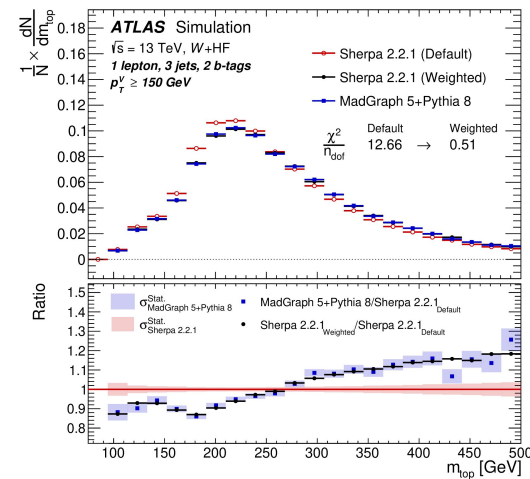
VHbb: W/Z+hf backgrounds estimation

Controlled use of systematics profiling

- Taking advantage of good control regions
 - Control regions “pretty close” to signal regions
 - Use of ΔR_{bb} / m_{bb} sidebands + multiclass BDT
 - Purity to specific backgrounds from “good” to “excellent”
- Profiling at work
 - CRs allow to constrain background cross-sections
 - And some background shapes
 - What remain are smaller extrapolation uncertainties
- Caveats
 - Choice of the 2-point systematics, e.g 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](#)



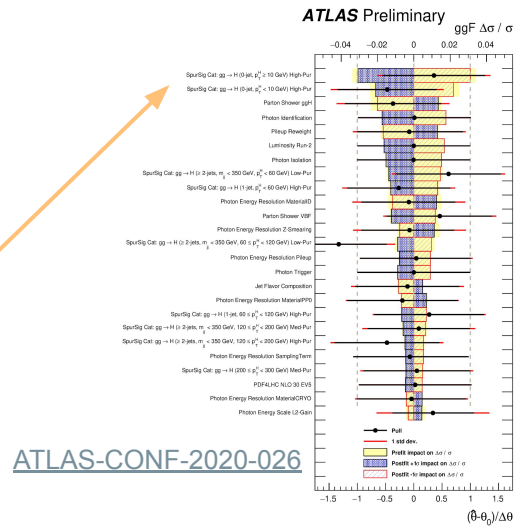
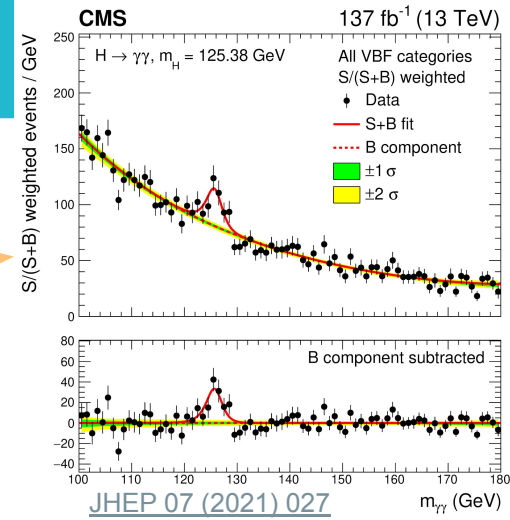
[Eur. Phys. J. C 81 \(2021\) 178](#)

Modelling smooth backgrounds

See Model selection talk by C. Schafer

- Textbook $H \rightarrow \gamma\gamma$ example
 - Narrow resonance on top of smoothly falling bkg
 - Use of semiparametric models
 - Fit of analytical functions more accurate than $\gamma\gamma$ / γ -jet MC samples
 - Also applies to $H \rightarrow \mu\mu$, $H \rightarrow Z\gamma$...

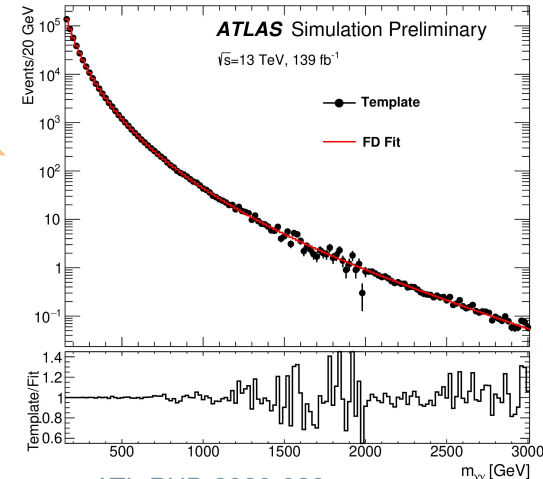
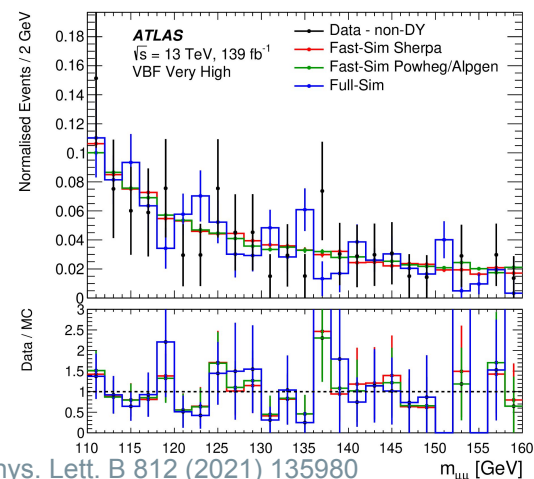
- Procedures well established since Run-1
 - ATLAS-CMS disagreement also when established
 - **CMS**: Discrete profiling. Choice of function embedded in a nuisance parameter
 - Residual uncertainty very small
 - **ATLAS**: Select function, and estimate maximum bias 'spurious signal'
 - Requires vast amounts of MC events
 - Limitation for high luminosity



Smooth backgrounds: new techniques

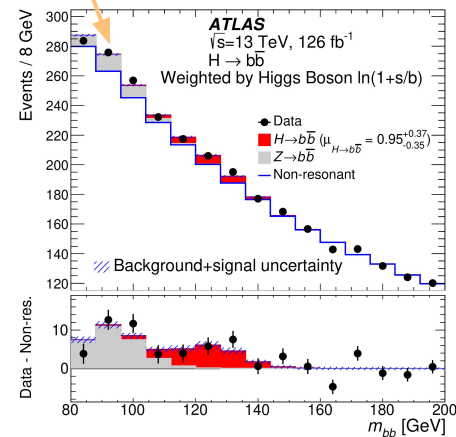
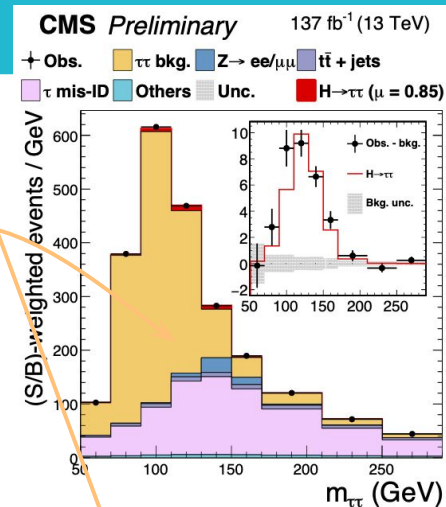
New techniques to overcome limitations of spurious signal evaluation

- Use of very fast sim ($H \rightarrow \mu\mu$):
 - 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



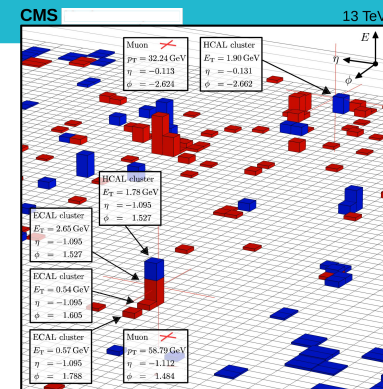
Resonant backgrounds - embedding

- E.g. Z boson decays in fermionic channels
- Same signature as the signal, except for mass
 \Rightarrow 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

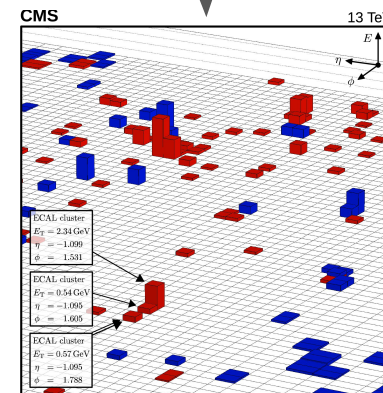


Embedding - principle

- Principle in a nutshell:
 - Select a well-understood process in data, in our case $Z \rightarrow \mu\mu$
 - 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% \rightarrow cannot merge at level of hits/deposits
 - Cannot obtain perfect closure \rightarrow residual corrections
 - Spin correlations for simulated taus ignored
- Less complex procedure (re-scaling, not replacing) also in use in ATLAS ($\tau\tau$)
 - Trade complexity for accuracy



Remove muon deposits



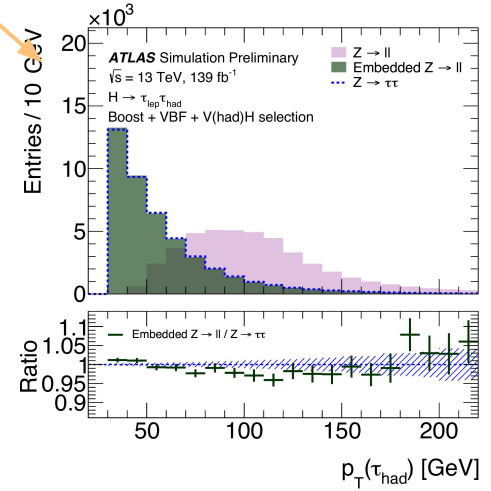
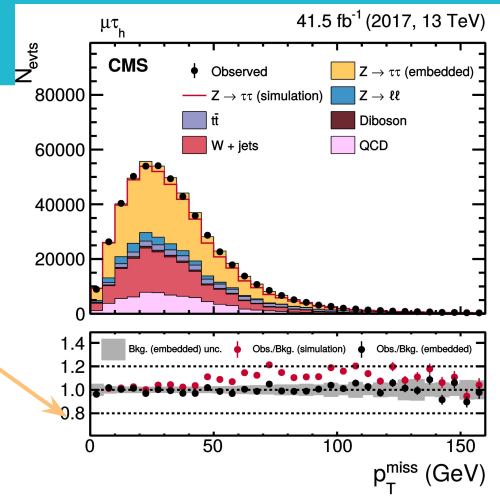
Calorimeter deposits
before and after
removing muon deposits

Embedding - achievements

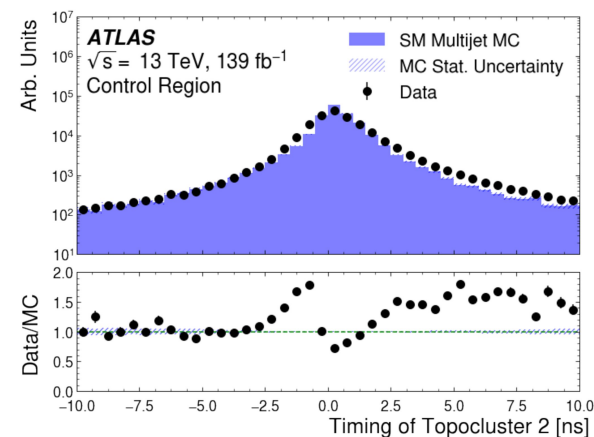
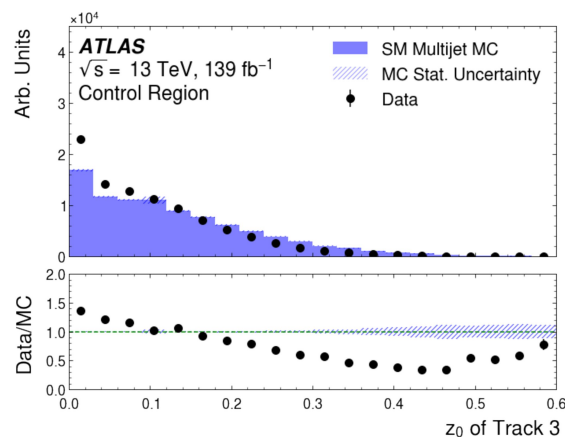
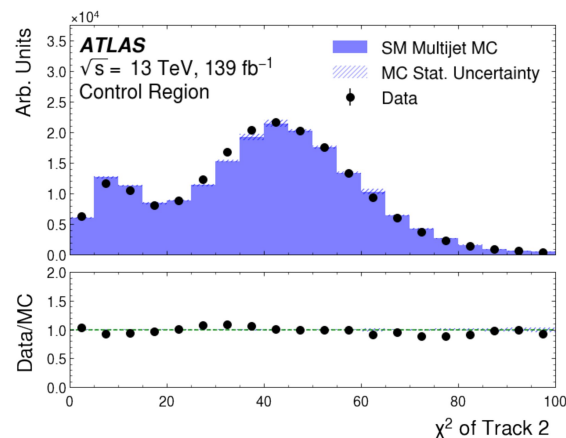
- Better modelling of kinematic distributions with embedded samples than simulation
- Helps reduce some uncertainties
- Simplified procedure provides a control region in data
- Even better modelling (smaller uncertainties?)
→ more work needed!

Uncertainty	$\sigma(\mu_H)$	$\sigma(\mu_{\text{VBF}})$
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)

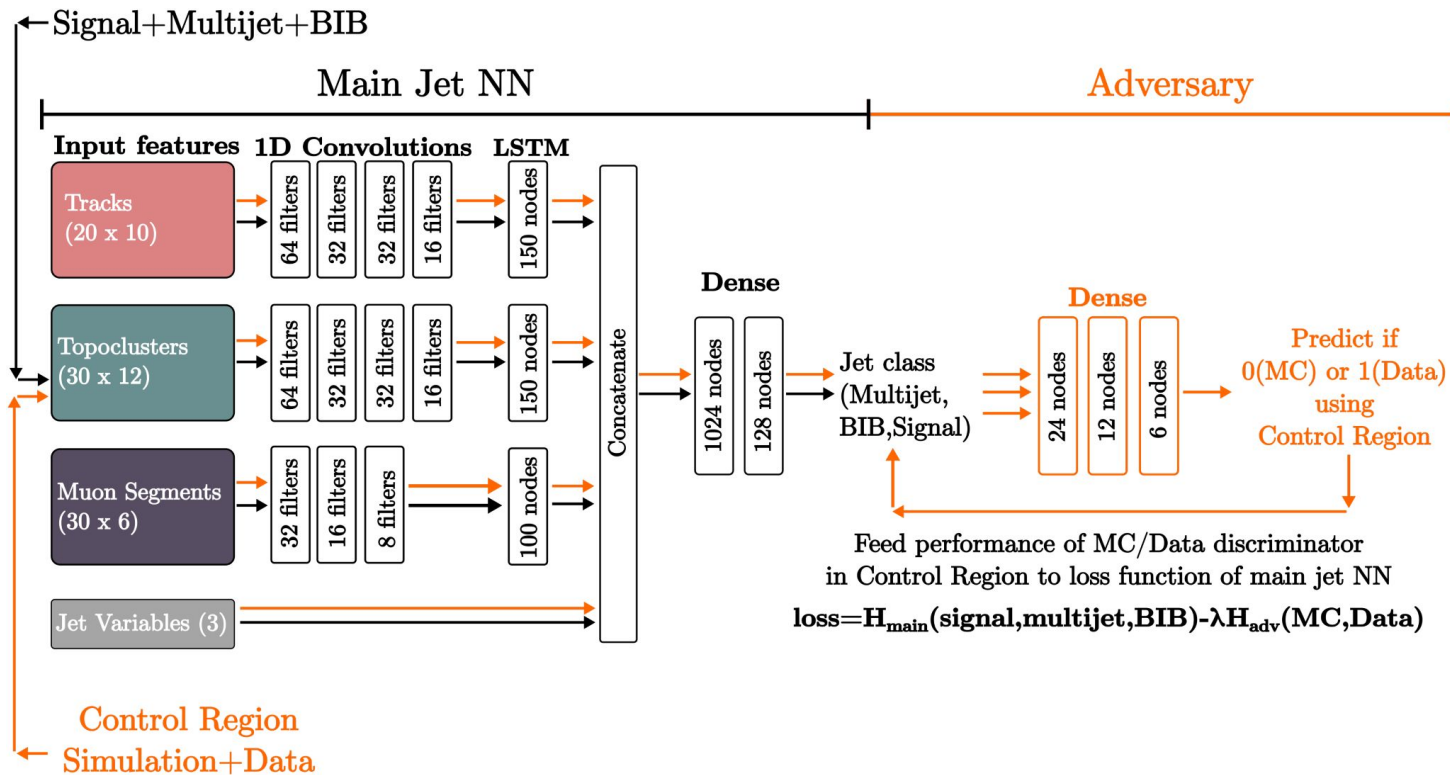


- Search for LLP using strange “CalRatio” jets
 - Build multiclass NN to separate signal CalRatio jets (**MC**), QCD (**MC**), Beam-Induced-Background (from **data CR** defined at trigger level)
 - But BIB-data sample is known to have significant fraction of QCD-data contamination
 - And certain input variables, such as jet timing, are important discriminators, but are not perfectly modelled
- **NN learns to separate data/MC because of QCD events in BIB sample...**



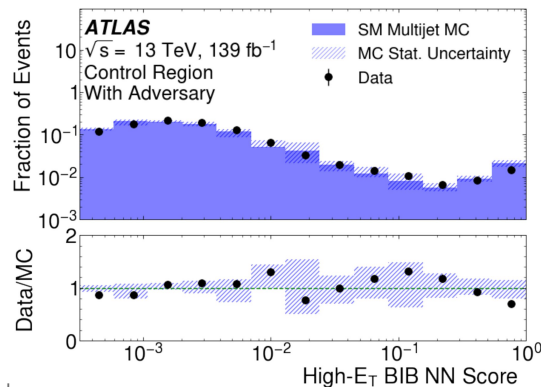
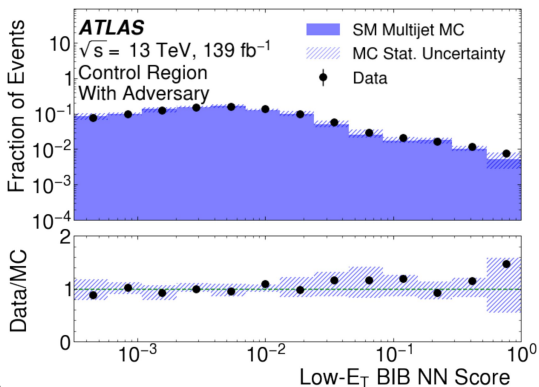
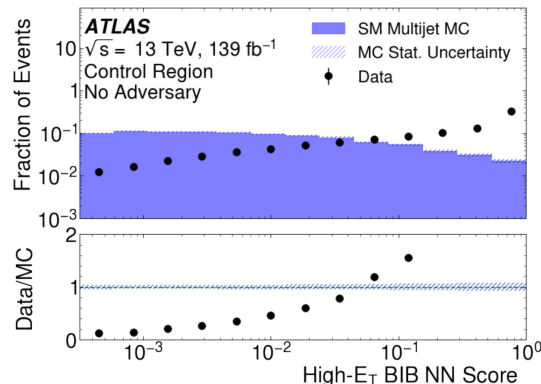
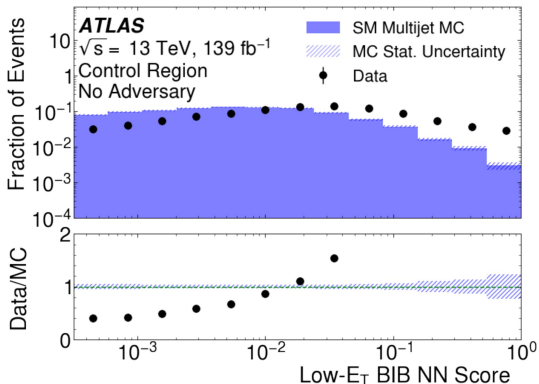
Adversarial NN to the rescue

- Adversary trained to distinguish data from MC in dijet control region
- Feeds into main NN as penalty in loss function



Adversarial NN results

- Huge improvement
- Residual discrepancies covered by systematic uncertainty



No adversary

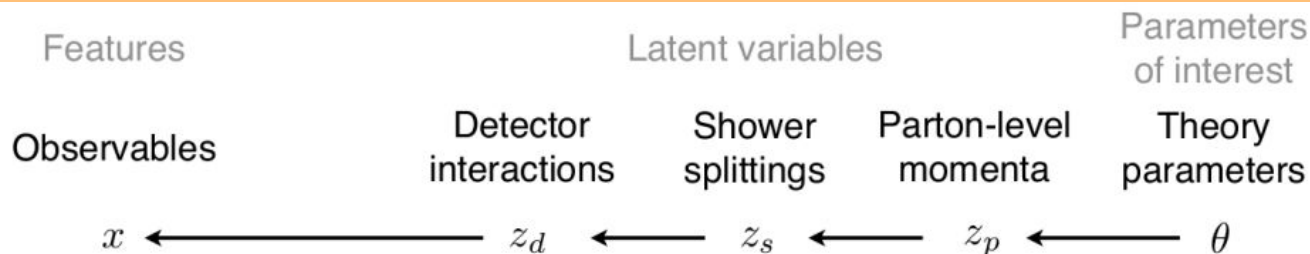


Adversary

Signal shapes

Signal shapes ?

Signal shapes are the convolution of theory predictions in the form of MC samples, and of experimental (detector) effects



$$p(x|\theta) = \underbrace{\iiint p(z_p|\theta)p(z_s|z_p)p(z_d|z_s)p(x|z_d)dz_pdz_sdz_d}_{\text{intractable}}$$

From Gilles Louppe

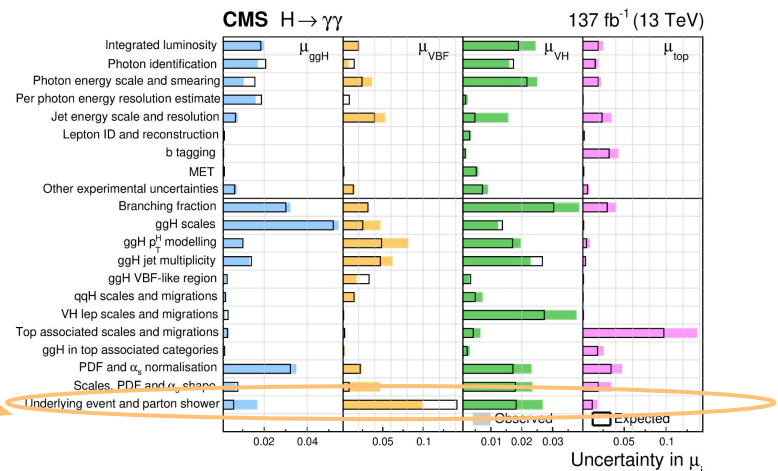
- Uncertainties affect all terms in the convolution
- For background shapes, control regions and data-driven techniques allow to short-circuit some of the uncertainties
- For signal shapes we need to have them all

Examples in Higgs: Underlying event & parton shower

ATLAS-CONF-2020-026

Uncertainty source	ggF+ bbH $\Delta\sigma$ (%)	VBF $\Delta\sigma$ (%)	WH $\Delta\sigma$ (%)	ZH $\Delta\sigma$ (%)	ttH+ tH $\Delta\sigma$ (%)
Underlying Event and Parton Shower (UEPS)	± 2.3	± 10	$< \pm 1$	± 9.6	± 3.5
Modeling of Heavy Flavor Jets in non-ttH Processes	$< \pm 1$	$< \pm 1$	$< \pm 1$	$< \pm 1$	± 1.3
Higher-Order QCD Terms (QCD)	± 1.6	$< \pm 1$	$< \pm 1$	± 1.9	$< \pm 1$
Parton Distribution Function and α_S Scale (PDF+ α_S)	$< \pm 1$	± 1.1	$< \pm 1$	± 1.9	$< \pm 1$
Photon Energy Resolution (PER)	± 2.9	± 2.4	± 2.0	± 1.3	± 4.9
Photon Energy Scale (PES)	$< \pm 1$	$< \pm 1$	$< \pm 1$	± 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	$< \pm 1$	$< \pm 1$	$< \pm 1$	$< \pm 1$	$< \pm 1$
Leptons	$< \pm 1$	$< \pm 1$	$< \pm 1$	$< \pm 1$	$< \pm 1$
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	$< \pm 1$	$< \pm 1$	$< \pm 1$	± 3.7	± 1.9

- Significant component of the theoretical uncertainty in several measurements, e.g. $H \rightarrow \gamma\gamma$
 - Particularly in VBF phase space
- 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



Going for differential measurements: Higgs STXS

Differential measurements: instead of measuring 1 signal cross-section, measure simultaneously Higgs cross-section in well-defined parts of phase space based on production kinematics

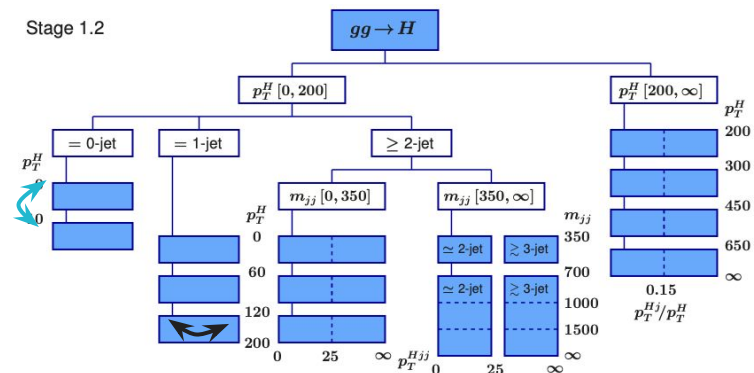
- Higgs Simplified Template Cross-sections

- Agreement between ATLAS CMS and theorists on “good” partition of phase space
- Selected so that relevant theory uncertainties can be provided
- Good sensitivity to new physics at high momentum

- Requires a much more refined set of theory 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

- Overall net reduction of signal uncertainties

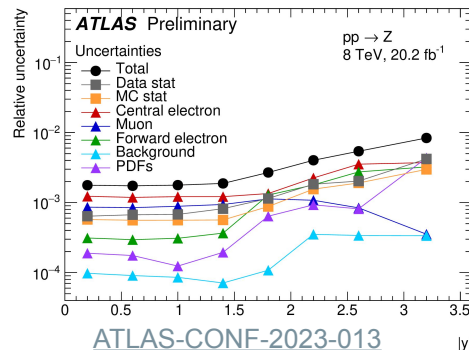


<https://twiki.cern.ch/twiki/bin/view/LHCPHysics/LHCHWG/FiducialAndSTXS>

Uncertainties in interpretations of measurements

Differential measurements allow to factorize, but do not make uncertainties magically disappear

- Measurement of transverse momentum and rapidity of Z boson using Run 1 data
 - Joint measurement of **1584** parameters (cross-sections + polarization coefficients) !
 - Extremely precise data
 - Negligible modelling uncertainties
- Interpretation of these measurements: determination of α_s
 - Relate all these measurements to common underlying theory parameters
 - Modelling uncertainties dominate
 - Missing higher order corrections
 - Parton density functions



Experimental uncertainty	+0.00044	-0.00044
PDF uncertainty	+0.00051	-0.00051
Scale variations uncertainties	+0.00042	-0.00042
Matching to fixed order	0	-0.00008
Non-perturbative model	+0.00012	-0.00020
Flavour model	+0.00021	-0.00029
QED ISR	+0.00014	-0.00014
N4LL approximation	+0.00004	-0.00004
Total	+0.00084	-0.00088

ATLAS-CONF-2023-015

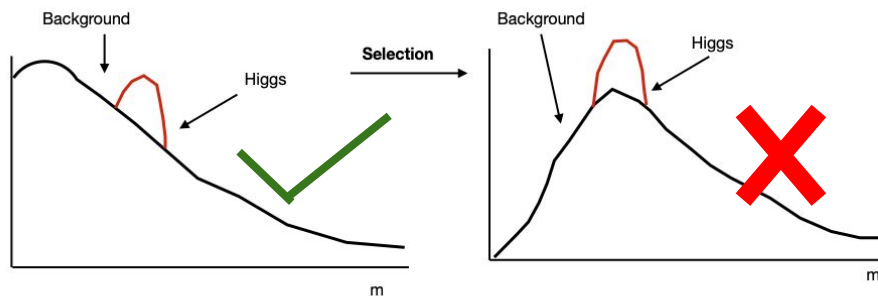
Summary

- Getting the right signal and background shapes (i.e with small associated uncertainties) is a major topic when going for precision measurements or measurements of low processes with low S/B
- Large field of analysis techniques to use data more and rely less on MC predictions
 - Very active field esp. using techniques from the ML world
- Progress requires close collaboration experimentalists / theorists / statisticians
 - Simulations of complex final states ($t\bar{t}b\bar{b}$, $W/Z+hf\dots$)
 - Simulations of difficult phase space (Higgs VBF, high p_T)
 - Agreement on “adequate” uncertainties in the shapes

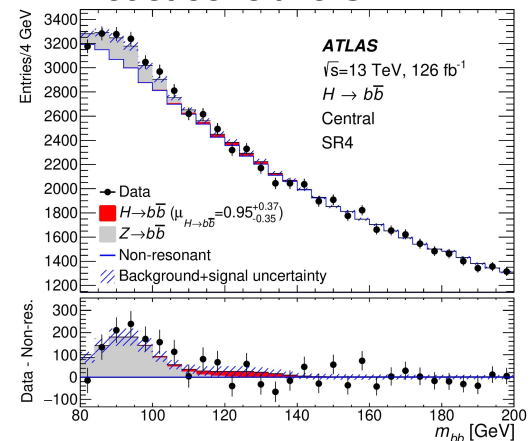
Additional Material

Smooth backgrounds: sculpting

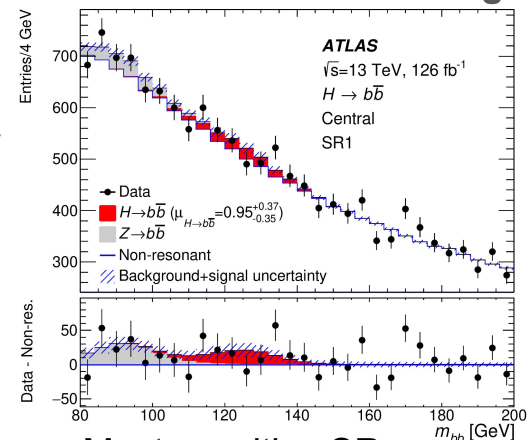
- Analysis selection should avoid sculpting background
 - Loss of sensitivity, difficulty modelling data-driven background
- Mitigation strategies in $H \rightarrow b\bar{b}$ analyses
 - "Basic" selection: mass-decorrelated double-b taggers for boosted $H \rightarrow b\bar{b}$
 - Event classification: mass-decorrelated ANN for VBF $H \rightarrow b\bar{b}$



Least sensitive SR



Similar non-resonant bkg shapes!

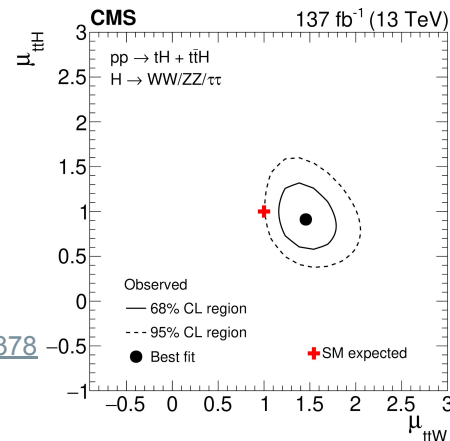
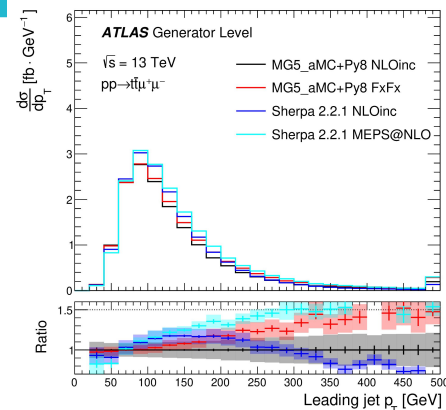
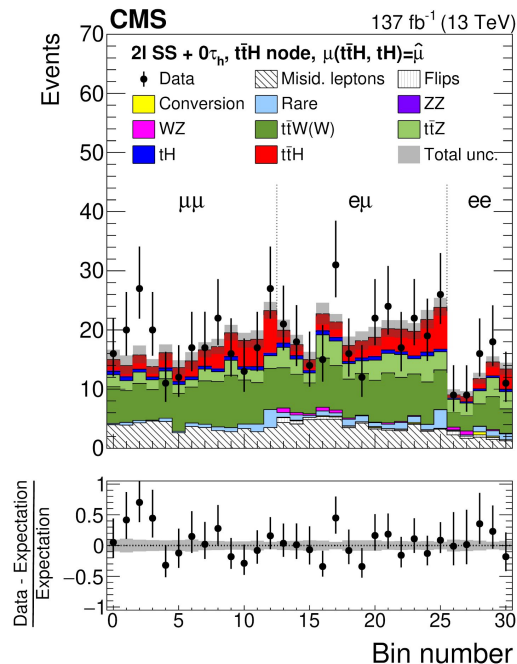


Most sensitive SR

ttH in multilepton final states: ttW/ttZ

ATLAS-PHYS-PUB-2020-024

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



ATLAS-CONF-2019-045

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

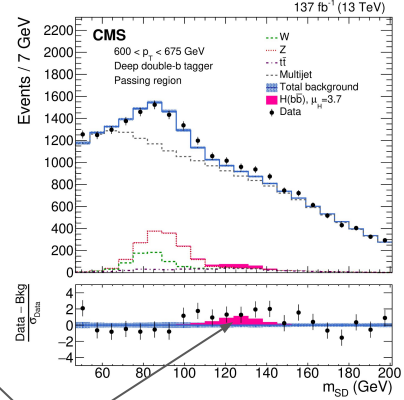
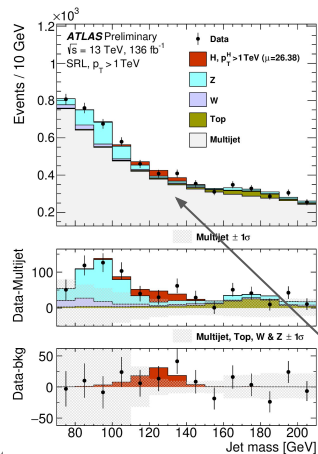
Phase space modelling - Higgs pT

- 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	$p_T^H > 450 \text{ GeV}$	$p_T^H > 1 \text{ TeV}$
Total	3.3	31
Statistical	2.8	30
Jet Systematics	1.2	7
Modeling and Theory Sys.	1.0	1
Flavor Tagging Sys.	0.5	3
Total Systematics	1.7	8

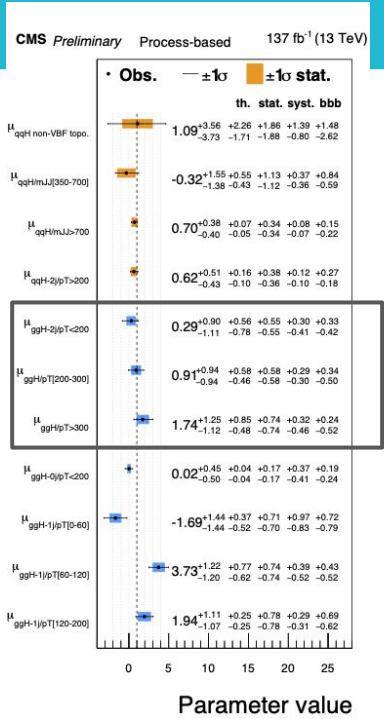
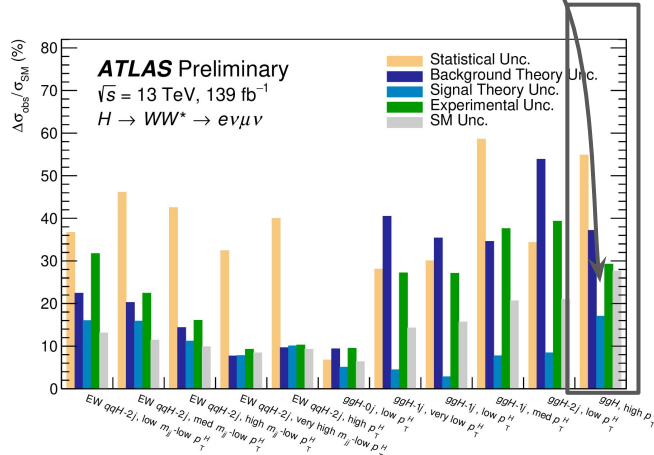
	2016	2017	2018	Combined
Expected μ_Z	$1.00^{+0.38}_{-0.28}$	$1.00^{+0.42}_{-0.29}$	$1.00^{+0.43}_{-0.29}$	$1.00^{+0.23}_{-0.19}$
Observed μ_Z	$0.86^{+0.32}_{-0.24}$	$1.11^{+0.48}_{-0.33}$	$0.91^{+0.37}_{-0.26}$	$1.01^{+0.24}_{-0.20}$
HJ-MiNLO				
Expected μ_H	$1.0^{+3.3}_{-3.5}$	1.0 ± 2.5	$1.0^{+2.3}_{-2.4}$	1.0 ± 1.4
Observed μ_H	$7.9^{+3.4}_{-2}$	$4.8^{+2.6}_{-2.5}$	1.7 ± 2.3	$3.7^{+1.6}_{-1.5}$
Expected H significance ($\mu_H = 1$)	0.3σ	0.4σ	0.4σ	0.7σ
Observed H significance	2.4σ	1.9σ	0.7σ	2.5σ
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 pT spectrum				
Expected μ_H	1.0 ± 1.5	$1.0^{+1.1}_{-1.0}$	$1.0^{+1.1}_{-1.0}$	$1.0^{+0.7}_{-0.7}$
Observed μ_H	$4.0^{+1.9}_{-1.0}$	$2.2^{+1.4}_{-1.2}$	1.1 ± 1.1	$1.9^{+0.9}_{-0.7}$
Expected H significance ($\mu_H = 1$)	0.7σ	0.9σ	1.0σ	1.7σ
Observed H significance	2.6σ	1.8σ	1.1σ	2.9σ
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
POWHEG 1J, pT reweight



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

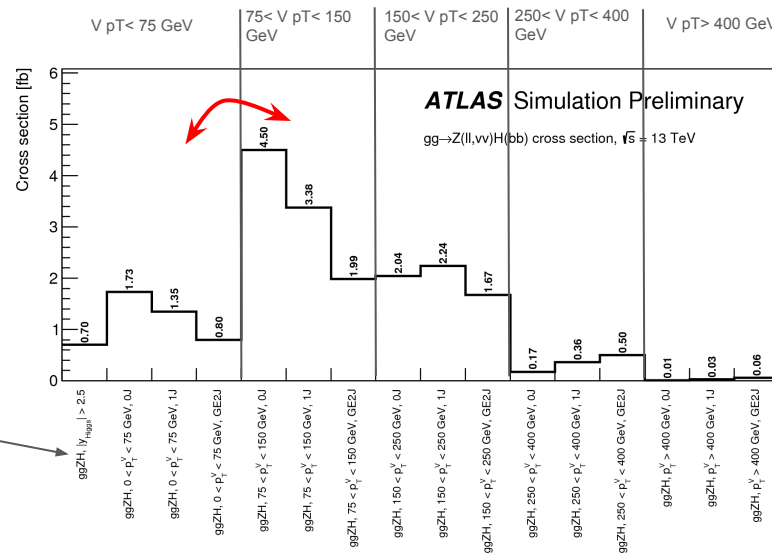


CMS-PAS-HIG-19-010

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 $0 < p_T < 75 \text{ GeV}$; 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”!

