bbbbackground estimation

Challenges and techniques in HH / HY → 4b analyses

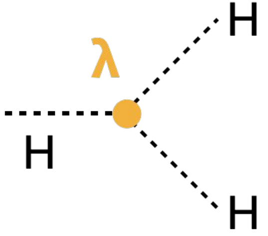
Nicole Hartman (nicole.hartman@tum.de)
Technical University of Munich
on behalf of the ATLAS collaboration

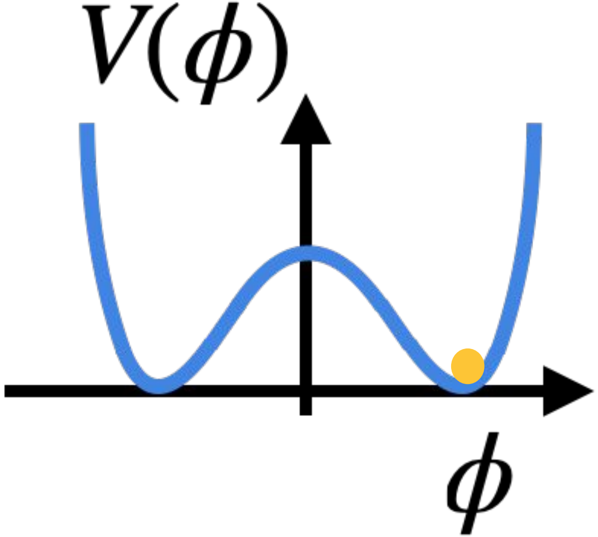
Matej Roguljić (matej.roguljic@cern.ch)
Johns Hopkins University
on behalf of the CMS Collaboration

Higgs 2023
28.11.2023

Why is 4b awesome?

$$V(x) = \mu^2 h(x)^2 + \lambda v h(x)^3 + \frac{1}{4}\lambda h(x)^4$$





4b: most signal

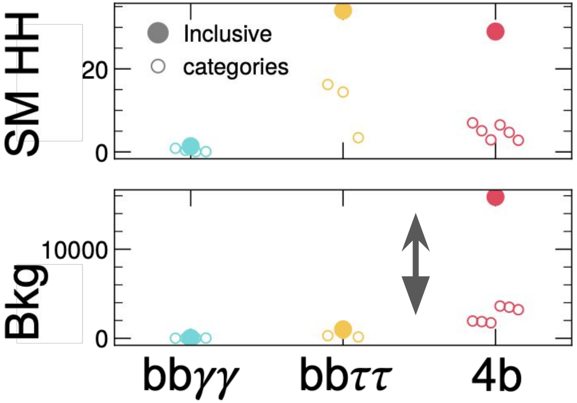
Higgs 1 decay (columns) / Higgs 2 decay (rows)

	bb	WW	ττ	ZZ	γγ
bb	34%				
WW	25%	4.6%			
ττ	7.3%	2.7%	0.39%		
ZZ	3.1%	1.1%	0.33%	0.069%	
γγ	0.26%	0.10%	0.028%	0.012%	0.0005%

Why is 4b HHard?

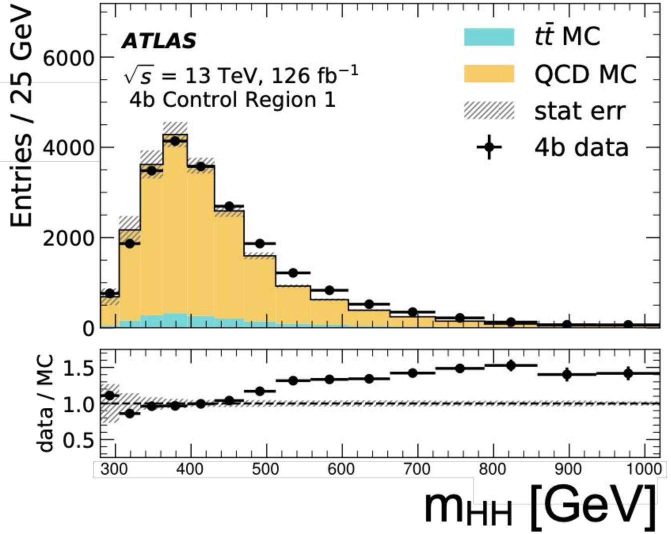
1 Backgrounds are BIG

Large multi-jet (QCD) backgrounds in hadronic final states



Fom ATLAS bb$\gamma\gamma$ HDBS-2018-34, bb$\tau\tau$ HDBS-2018-40, and 4b HDBS-2019-29 analyses.

Orders of magnitude higher bkg for 4b



2 Hard to simulate

- Multijet events modelled at leading order
- Often lack of statistics

Needs a data-driven background prediction!

The 4b analysis landscape

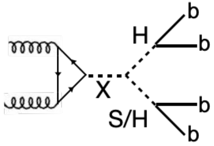
Publications

✔ ATLAS

✔ CMS

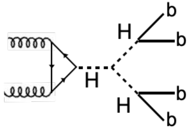
Resonant

X➞HH / X➞SH

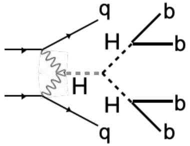


Non-Resonant

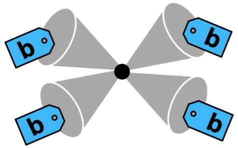
ggF



VBF



Resolved



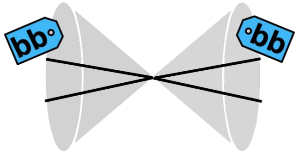
✔ HH: Phys. Rev. D 105 (2022) 092002

✔ HH: JHEP08(2018)152 (36 fb^{-1})

✔ ggF/VBF: Phys. Rev. D 108 (2023) 052003

✔ ggF/VBF: PhysRevLett.129.081802

Boosted



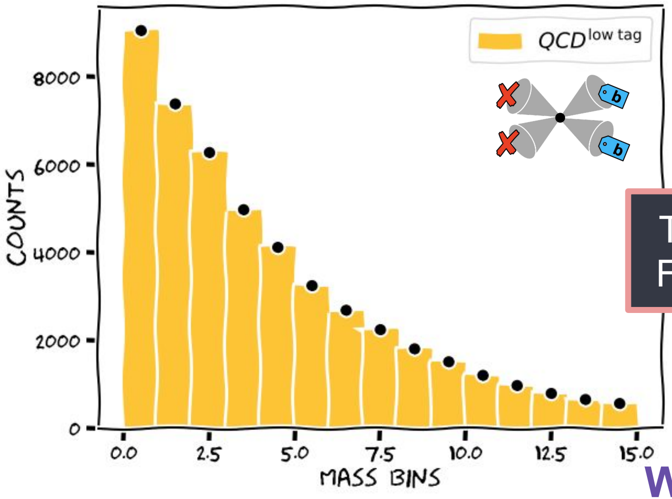
✔ HH: Phys. Rev. D 105 (2022) 092002

✔ HH: CMS-PAS-B2G-20-004

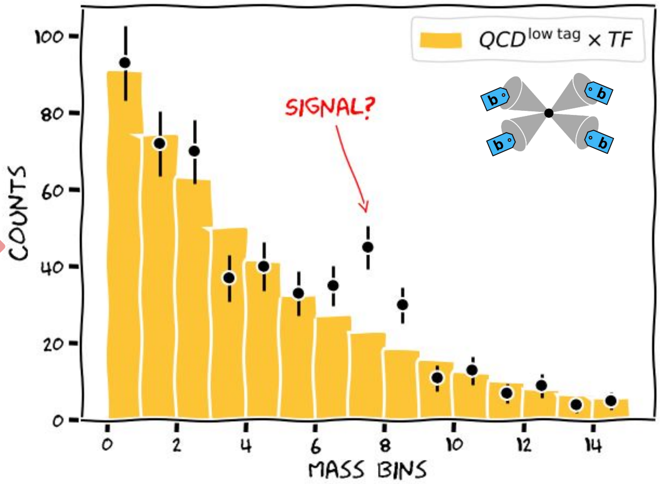
✔ HY: PhysLetB.2022.137392

✔ ggF/VBF: PhysRevLett.131.041803

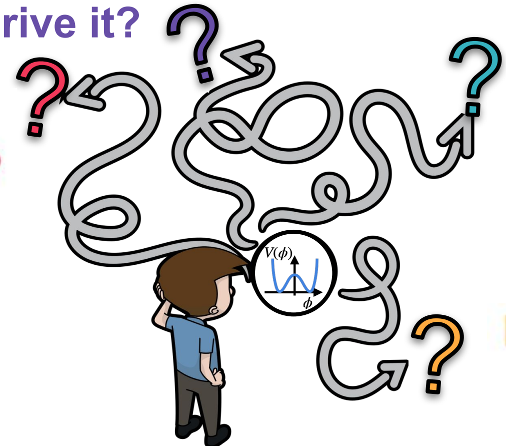
Transfer function method



Transfer Function



Where to derive it?



How to estimate the uncertainty

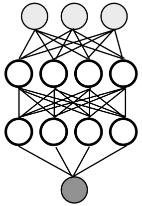
What function?

constant

polynomials

BDT

NN

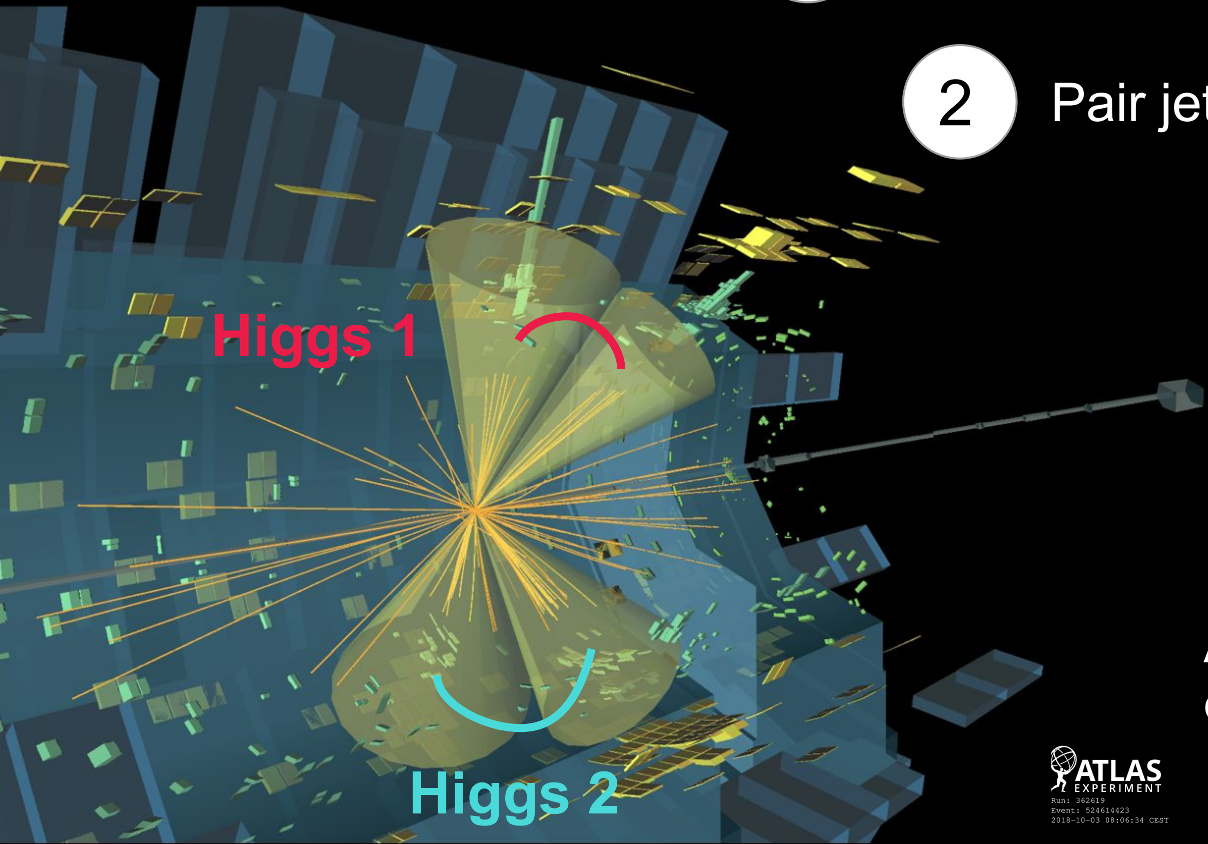


Method validation?

Resolved Analyses

1 Small-R jet b-tagging

2 Pair jets into Higgs Candidates



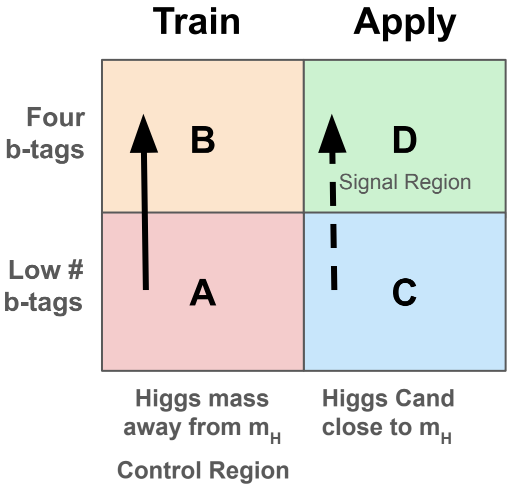
Also the two main handles for estimating the backgrounds

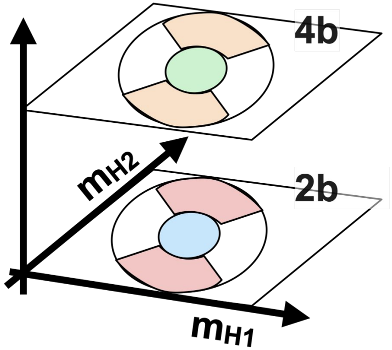
Resolved analyses

	ATLAS HH resonant	ATLAS HH non-res	CMS HH non-res
Fit variable(s):	m_{HH}	$m_{HH}, \Delta\eta_{HH}, X_{HH}$	BDT

Need a multi-dim background!

Generalized ABCD method





Trick: Classifiers are likelihood ratios, e.g, $w(x) = p^{CR}_{4b}(x) / p^{CR}_{2b}(x)$

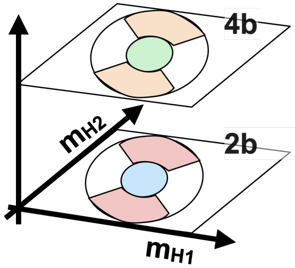
$$p_{4b}(x) = w(x) \cdot p_{2b}(x), x \in \mathbb{R}^n$$

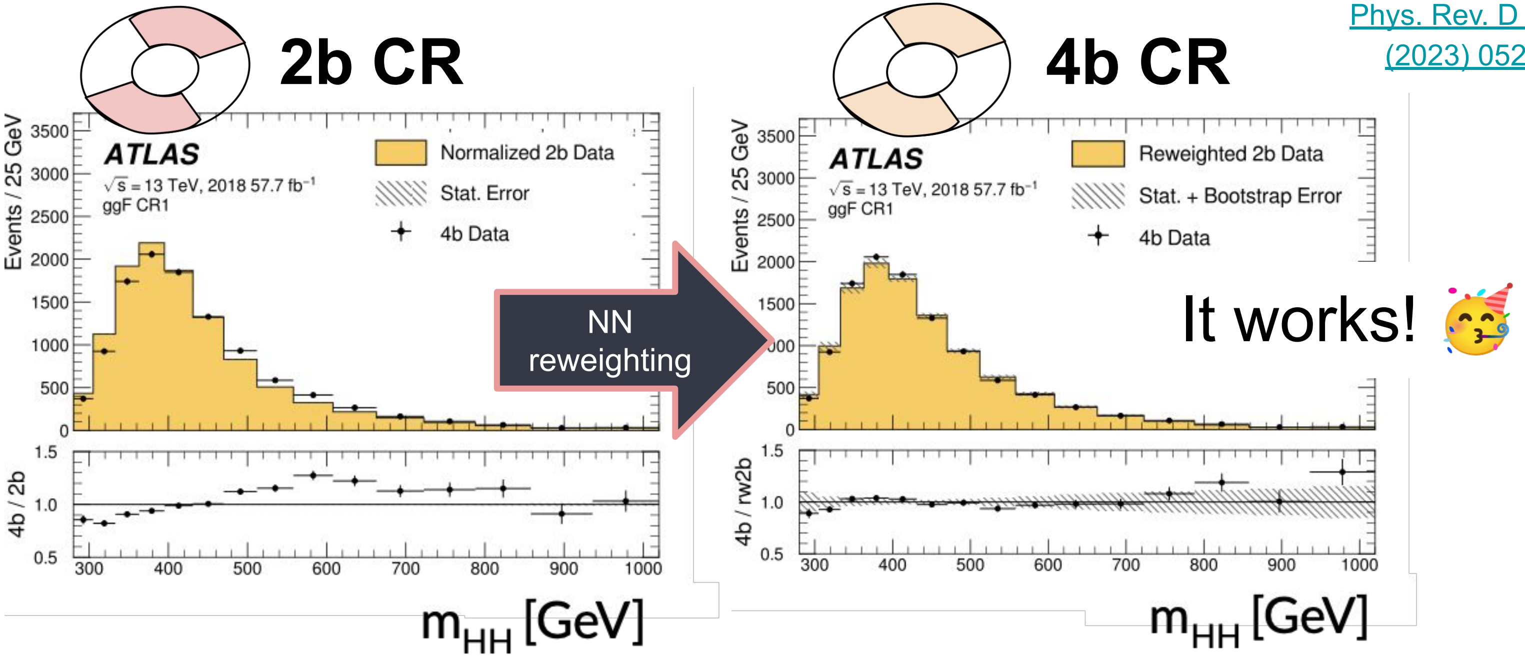
w(x) can be NN (ATLAS) or BDT (CMS)

Graphic with ATLAS HH4b non-res ABCD regions, same idea applies for other resolved analyses.

Train NN* in **Control Region**

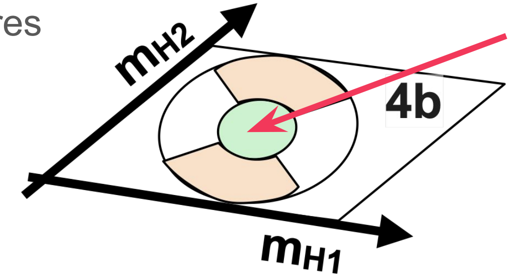
[Phys. Rev. D 108 (2023) 052003]





Note: here m_{HH} was *not* used to in the reweighting features

X (above plots), $x \in \mathbb{R}^{12}$: including some jet pTs, angular variables, and jet multiplicity



How to quantify an error on this bkg pred?

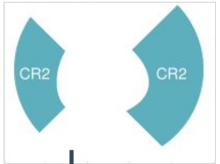
* Or BDT for the CMS model

Background uncertainties

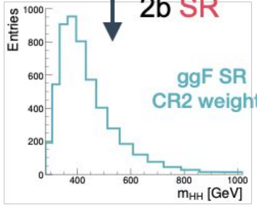
1. Choice of CR

Alternative estimate

Train NNs in CR2



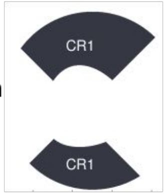
Apply to 2b SR



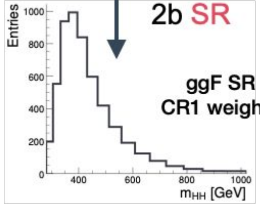
/

Nominal estimate

Train NNs in CR1

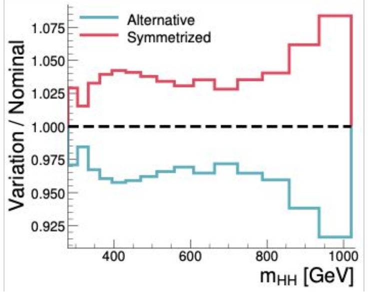


Apply to 2b SR



||

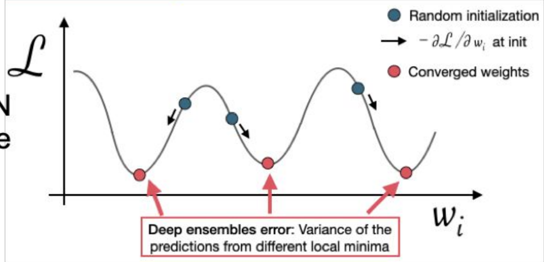
Take the ratio as an error bar

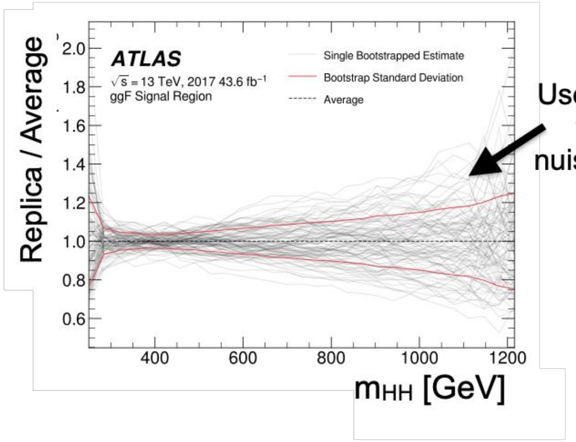


2. NN initialization

[1612.01474](1612.01474) + fig modified from [1912.02757](1912.02757)

Retraining the NN 100x captures the uncertainty from multiple minima





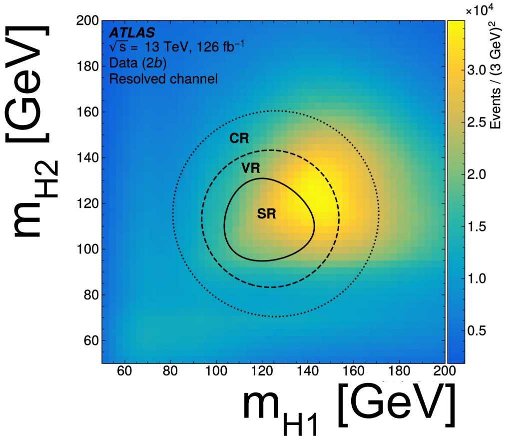
Use the variation of trainings as a nuisance parameter

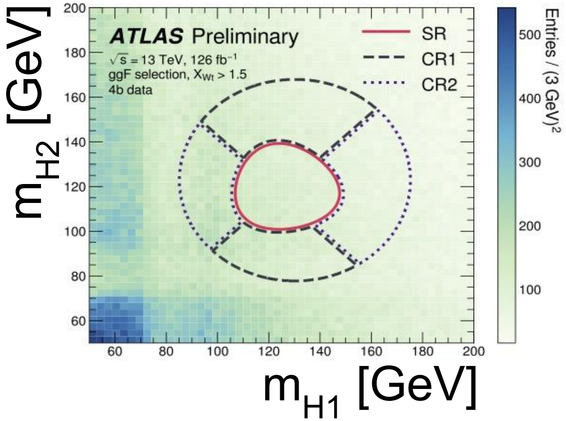
Phys. Rev. D 108 (2023) 052003

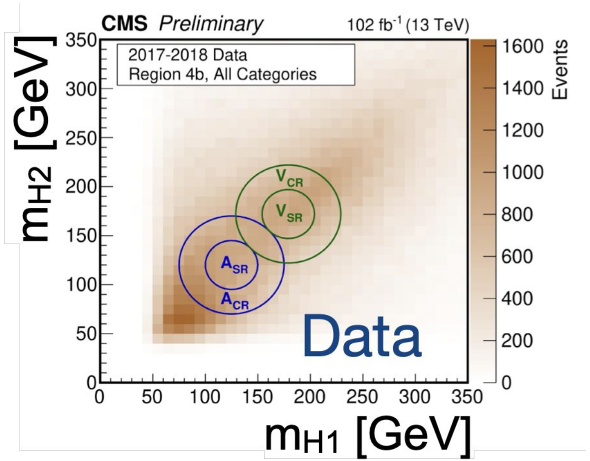
Background unc: comparisons

	ATLAS HH resonant	ATLAS HH non-res	CMS HH non-res
Source region	**2b → 4b**	**2b → 4b**	**3b → 4b**

Training regions







Uncertainties

ATLAS HH resonant:
- → **CR shape**
- → **NN retraining**
- → Stat unc (2b)

ATLAS HH non-res:
- → **CR shape**
- → **NN retraining**
- → Stat unc (2b)
- → 3b1f non-closure unc

CMS HH non-res:
- → **Stat unc (3b) (dominant)** ☆
- → Norm uncertainty 4b / 3b
- → Validation non-closure
- → Validation stats
- → CR shape

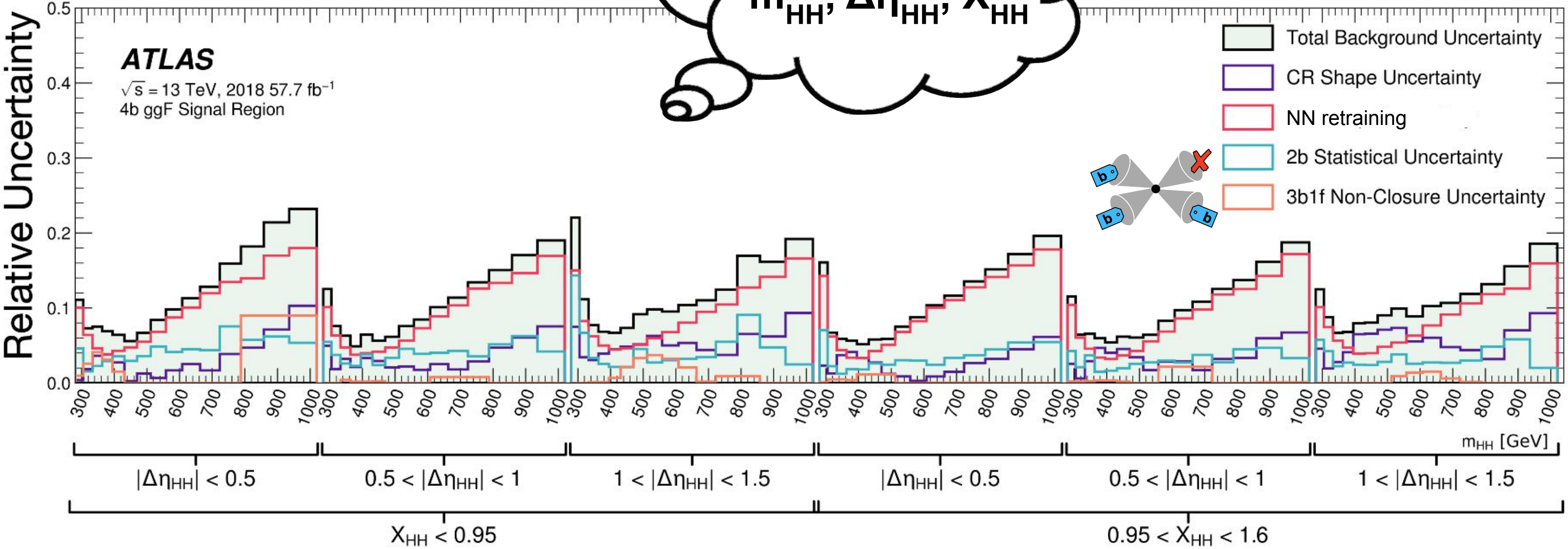
rinse repeat & **Validation is key**

Systematics

HH non-res

m_{HH}, $\Delta\eta_{HH}$, X_{HH}

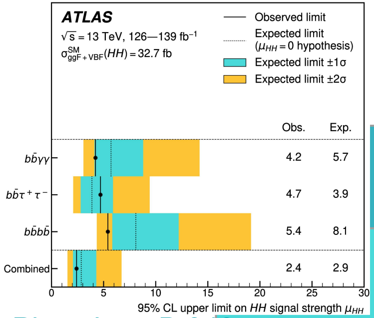
Phys. Rev. D 108 (2023) 052003



$$X_{HH} = \sqrt{\left(\frac{m_{H1} - 124\ \text{GeV}}{0.1 m_{H1}}\right)^2 + \left(\frac{m_{H2} - 117\ \text{GeV}}{0.1 m_{H2}}\right)^2}$$

Different NN trained for each of the three years

Impact of systematics: ATLAS HH NR



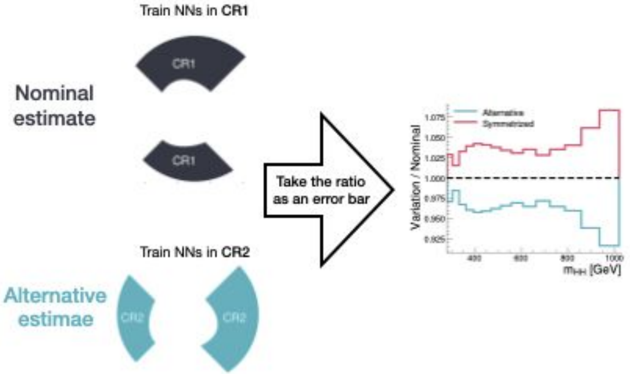
Phys. Lett. B 843 (2023) 137745

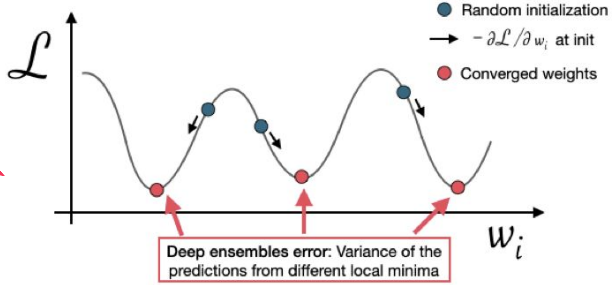
Recall from Marco Valente's talk : ATLAS **4b upper limit on HH signal strength: 5.4 (8.1) obs (exp).***

Phys. Rev. D 108 (2023) 052003

Uncertainty		$\Delta \mu_{ggF} / \mu_{ggF}$
Theoretical	Uncertainty on signal rate	9.0%
	All other theory uncertainties	1.4%
Background modelling	Control Region Interpolation	7.5%
	NN retraining (100x) + bootstrapped dataset	7.1%
	3b non-closure	2.0%

All other experimental systematic uncertainties < %-level impact



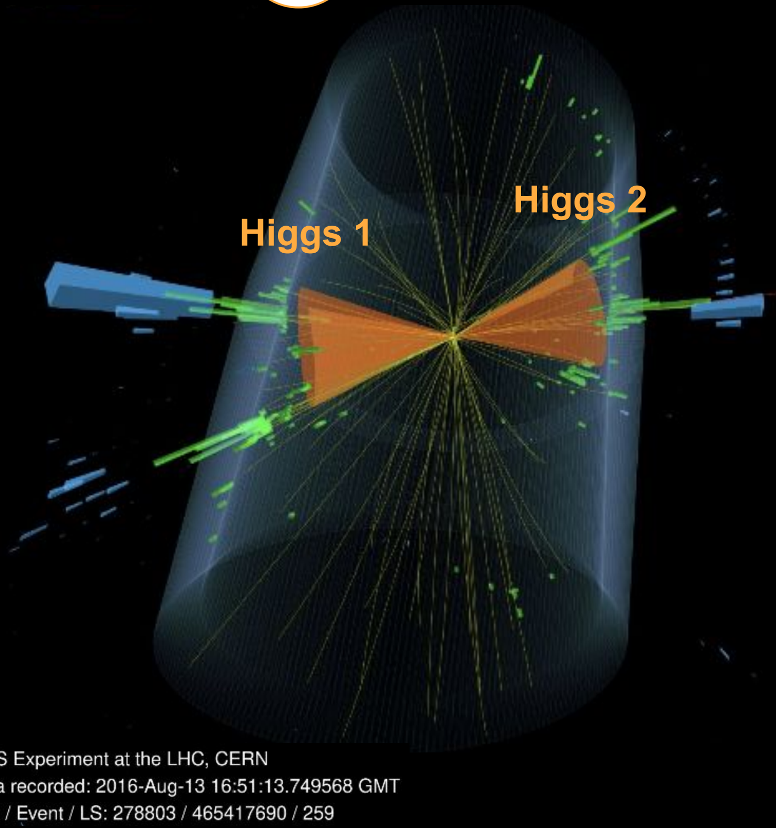


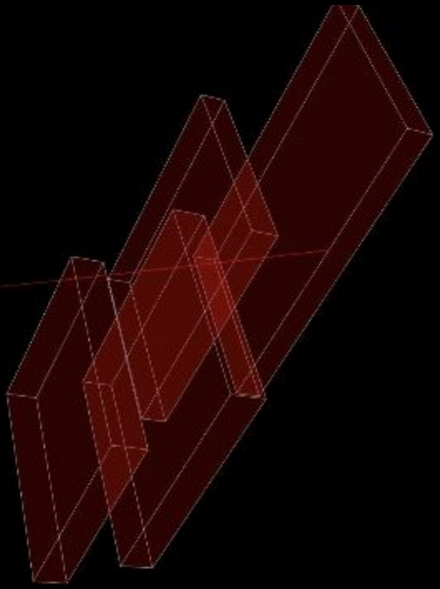
Background model drives the sensitivity for 4b analyses!

* CMS 4b resolved: 3.9 (7.8) for obs (exp) upper limit HH signal strength

Boosted analyses

1 Less multijet background at high energy





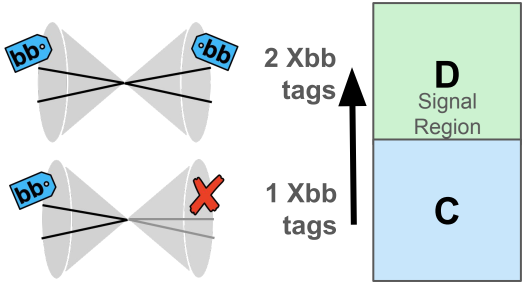
2 $t\bar{t}$ becomes more prominent

Boosted analyses

	HH/HY (CMS)	HH (ATLAS)	HH NR (CMS)
Fit variable(s):	$m_{HH(HY)}, m_{H(Y)}$	m_{HH}	m_{HH} (VBF), m_H (ggF)

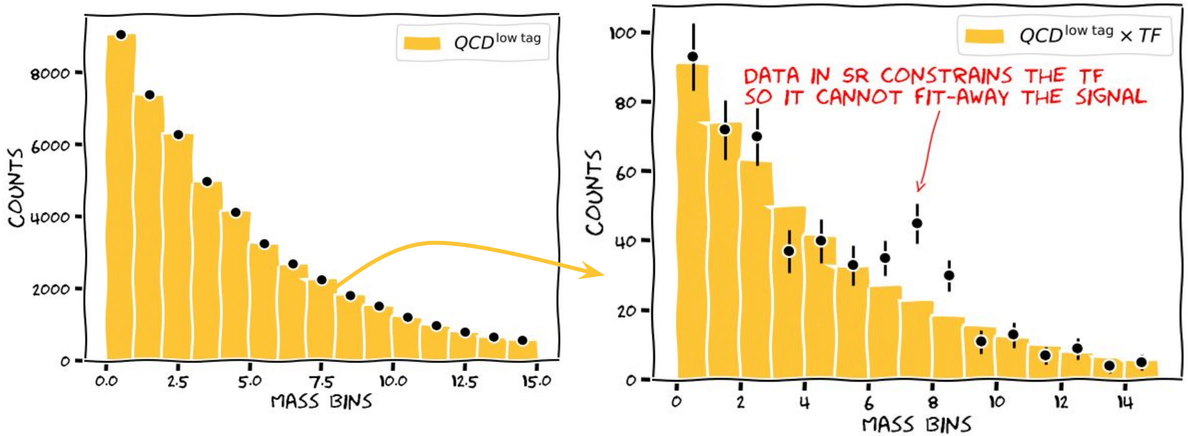
Similar methods as resolved analyses

Measure & apply transfer function



TF in this case is called "pass-to-fail" ratio: $R_{P/F}$

In-situ transfer function measurement



In-situ Transfer Function measurement

If transfer function (or $R_{P/F}$) is difficult to model…

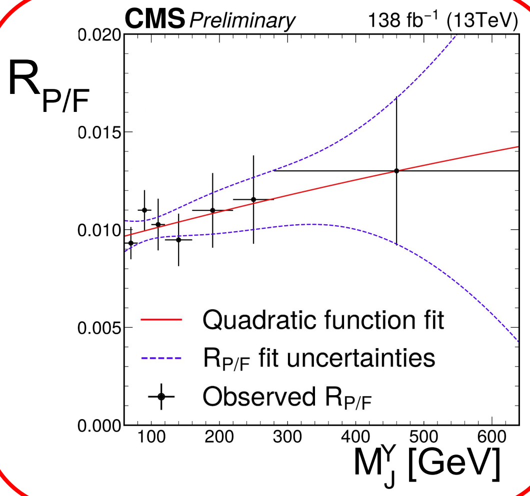
(1) Do an initial estimate

$$\frac{R^{true}_{P/F}}{R^{init}_{P/F}} \longrightarrow R_{Ratio}$$

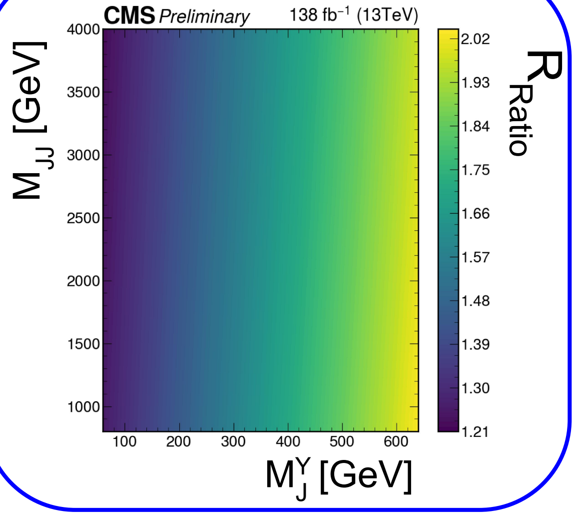
- In simulation (CMS-PAS-B2G-20-004)
- In CR (PhysLetB.2022.137392)

(2) Fit the residual difference

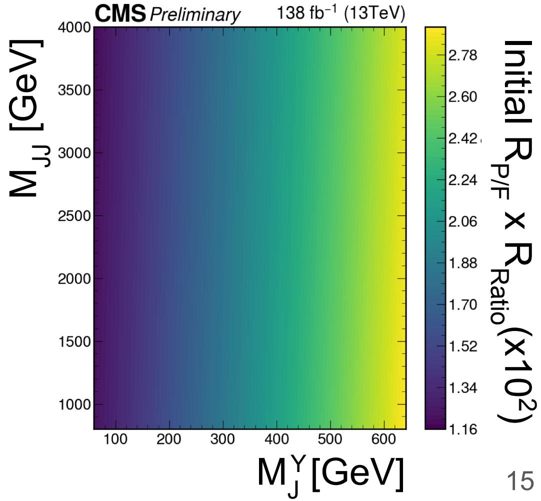
CMS resonant HY4b search (PhysLetB.2022.137392)



X

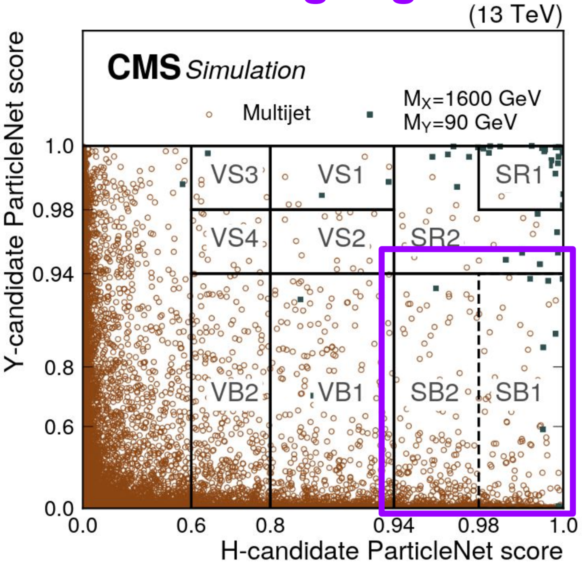


=

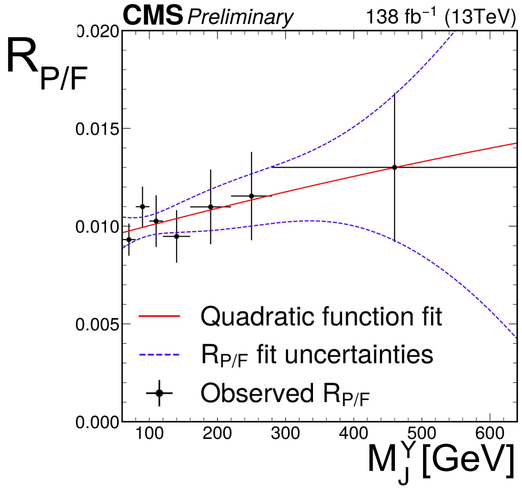


Background uncertainties

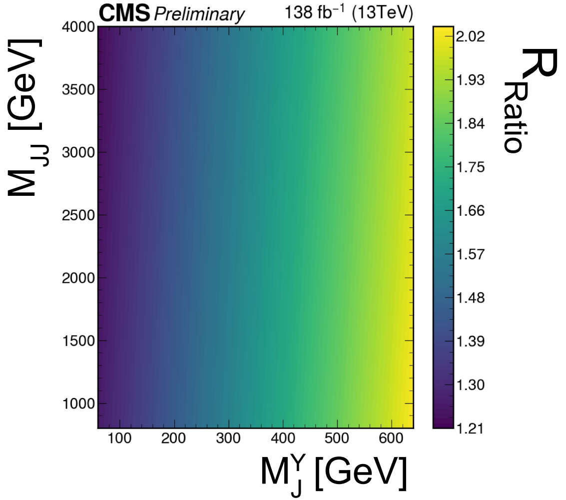
Data stat. uncertainties
in low-tag regions



Uncertainty on initial $R_{P/F}$



Uncertainty on R_{Ratio} parameters



Figures from
CMS resonant HY4b search
(PhysLetB.2022.137392)

Impact of background uncertainties

HH resonant

Phys. Rev. D 105 (2022) 092002

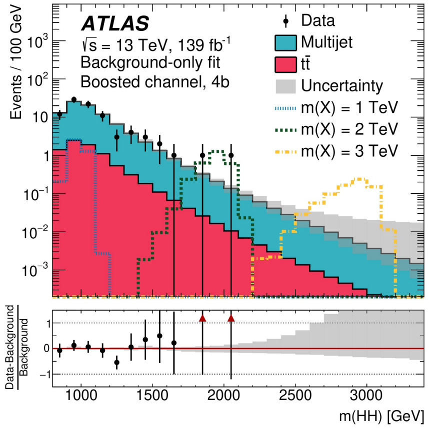
HY resonant

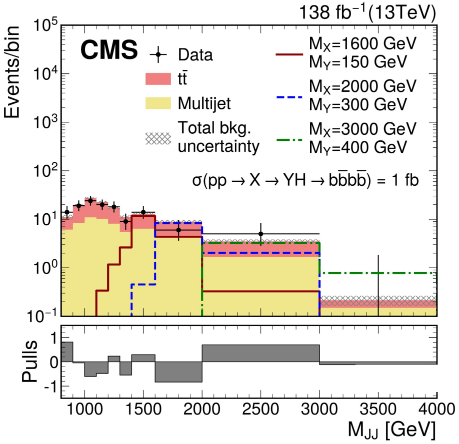
PhysLetB.2022.137392

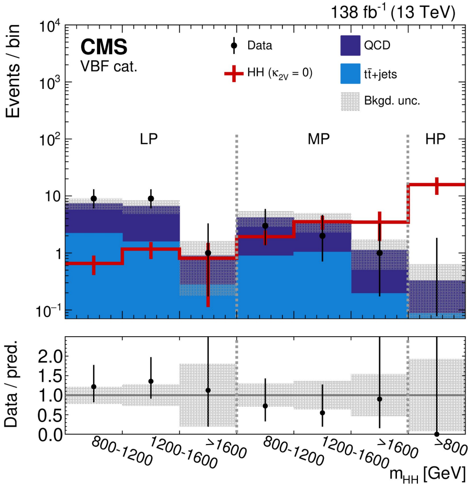
HH NR

PhysRevLett.131.041803

Postfit signal regions (one of)





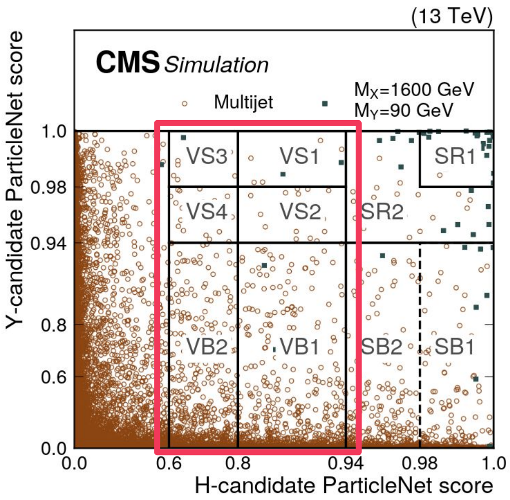


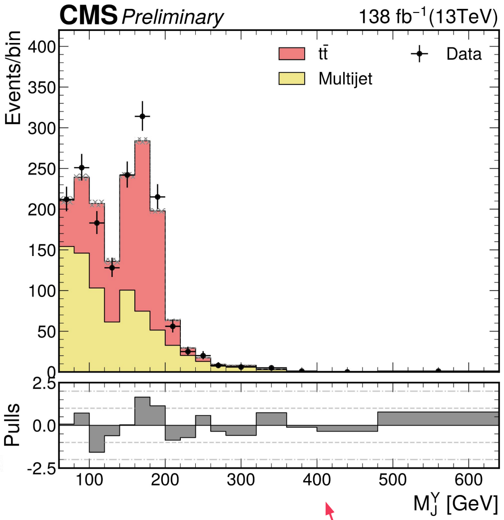
- ➔ **Choice of CR**
- ➔ **CR stat. uncertainty**
- ➔ **Transfer factor unc.**

Statistical uncertainty dominates

Does it work?

Confirm the method by fitting in the
validation regions

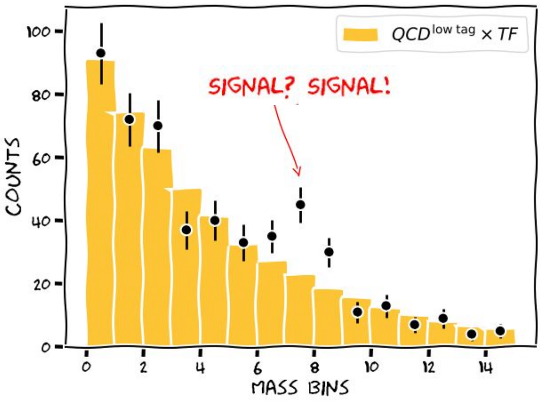




CMS resonant HY4b search
(PhysLetB.2022.137392)

Goodness-of-fit p-value > 0.05

Generate toy datasets and run
bias and signal injection tests

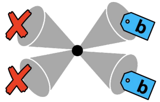


Transfer function method

How tos:

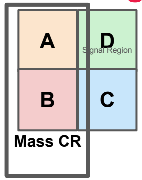
1. Background-rich source region

- Invert a cut that doesn't distort the shape of the fitted distribution
- Usually b-tagging requirement(s)

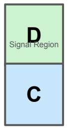


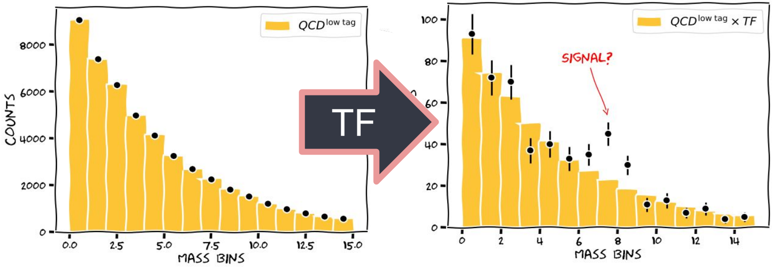
2. Transfer function: fit in

control region



direct fit to the data





3. Apply TF to source region for a background estimate in the signal region

4. Determine the estimation uncertainty?

TF uncertainty

Source shape uncertainty

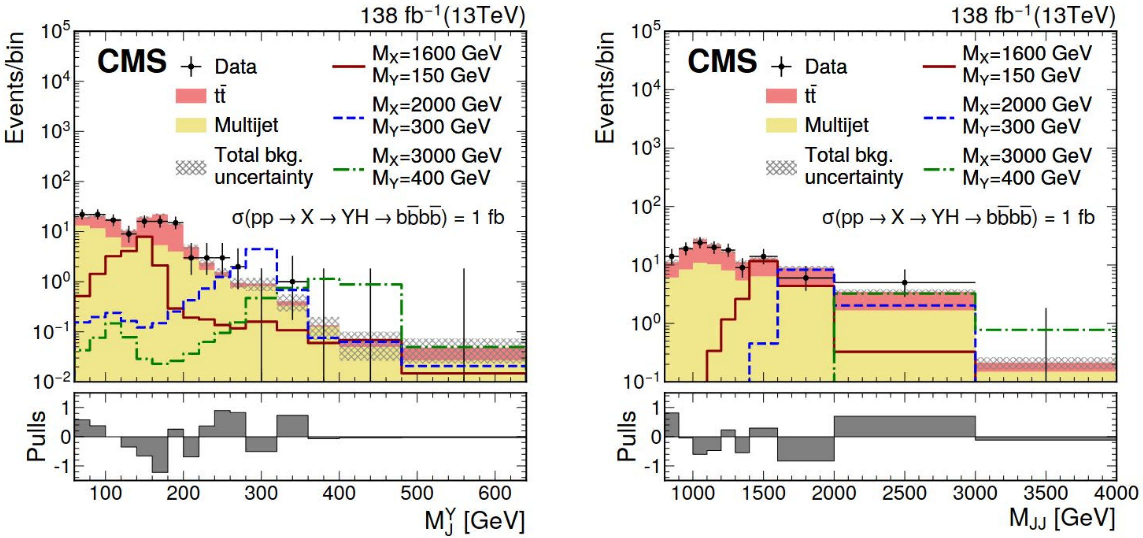
CR selection

non-closure

5. Validate the method in a signal-depleted region in data

What about $t\bar{t}$ background?

High p_T regime suppresses QCD so $t\bar{t}$ becomes significant in the signal regions



CMS resonant HY4b search (PhysLetB.2022.137392)

Similar shape as QCD

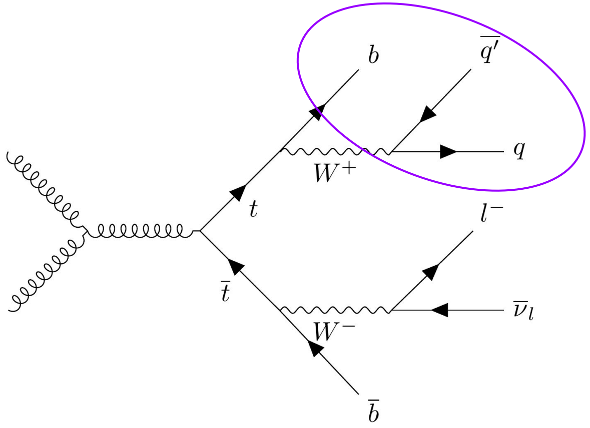
➔ Jointly fit using TF method

Different shape from QCD

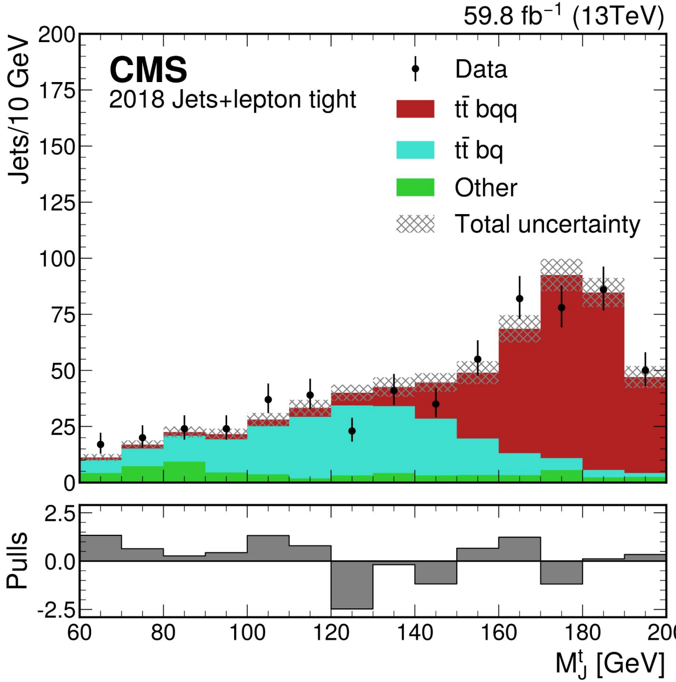
➔ Model separately using simulation

Correcting $t\bar{t}$ simulation

1 b-tagged jet and a lepton allow us to select a clean set of hadronically decaying top jets



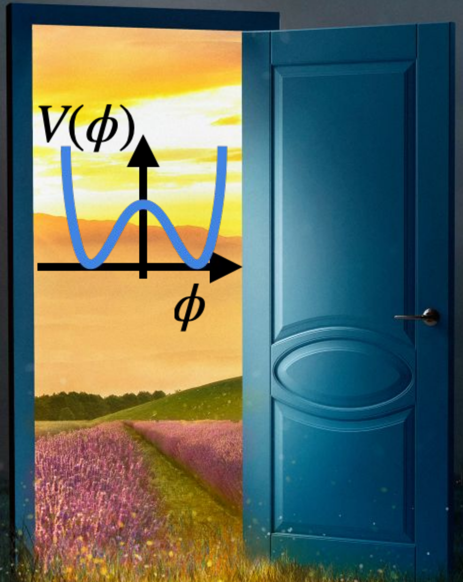
2 Use them to extract data-to-simulation correction factors



3 Apply correction factors to simulation in the SR

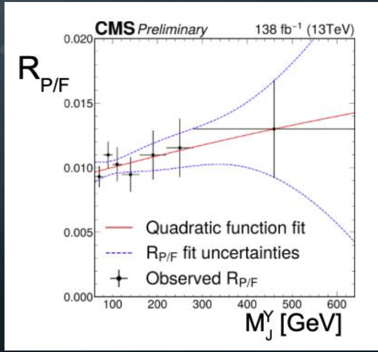
CMS resonant HY4b search
(PhysLetB.2022.137392)

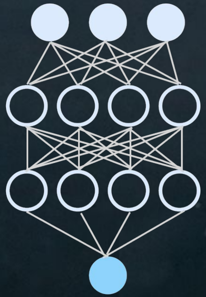
In summary



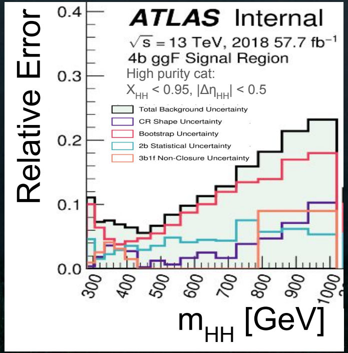
Data-driven methods crucial for HH4b analyses

Transfer functions (low → high tag) mostly used

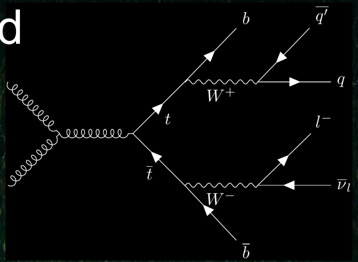




Error estimation and validation is the name-of-the-game



$t\bar{t}$ relevant in boosted



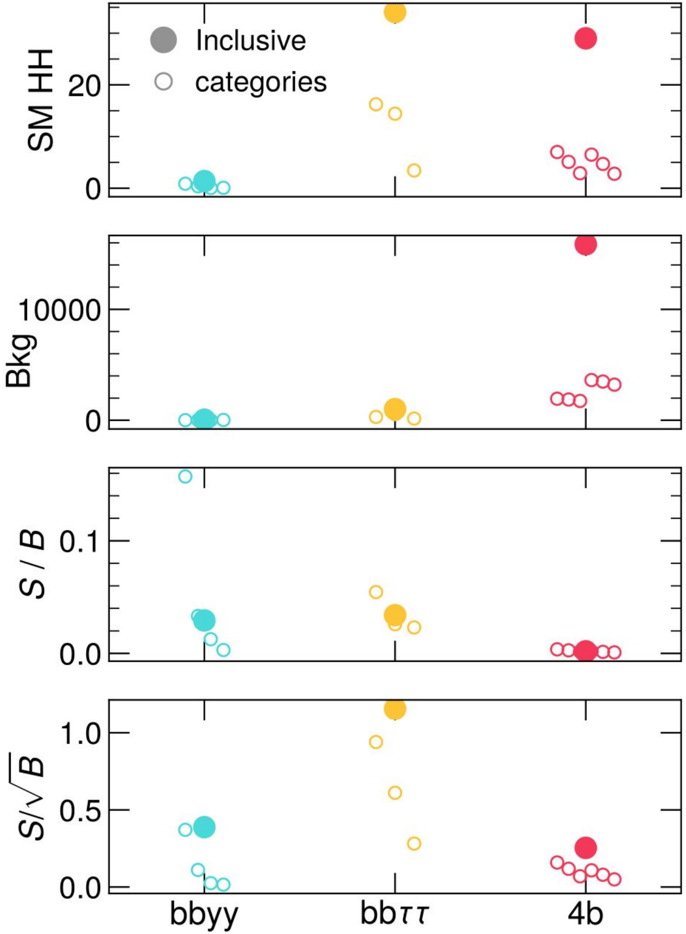
Exciting for Run 3 and bbbbeyond

ATLAS : Comparison of the NR channels

Comparison between the signal and backgrounds for ATLAS's three most sensitive HH channels:

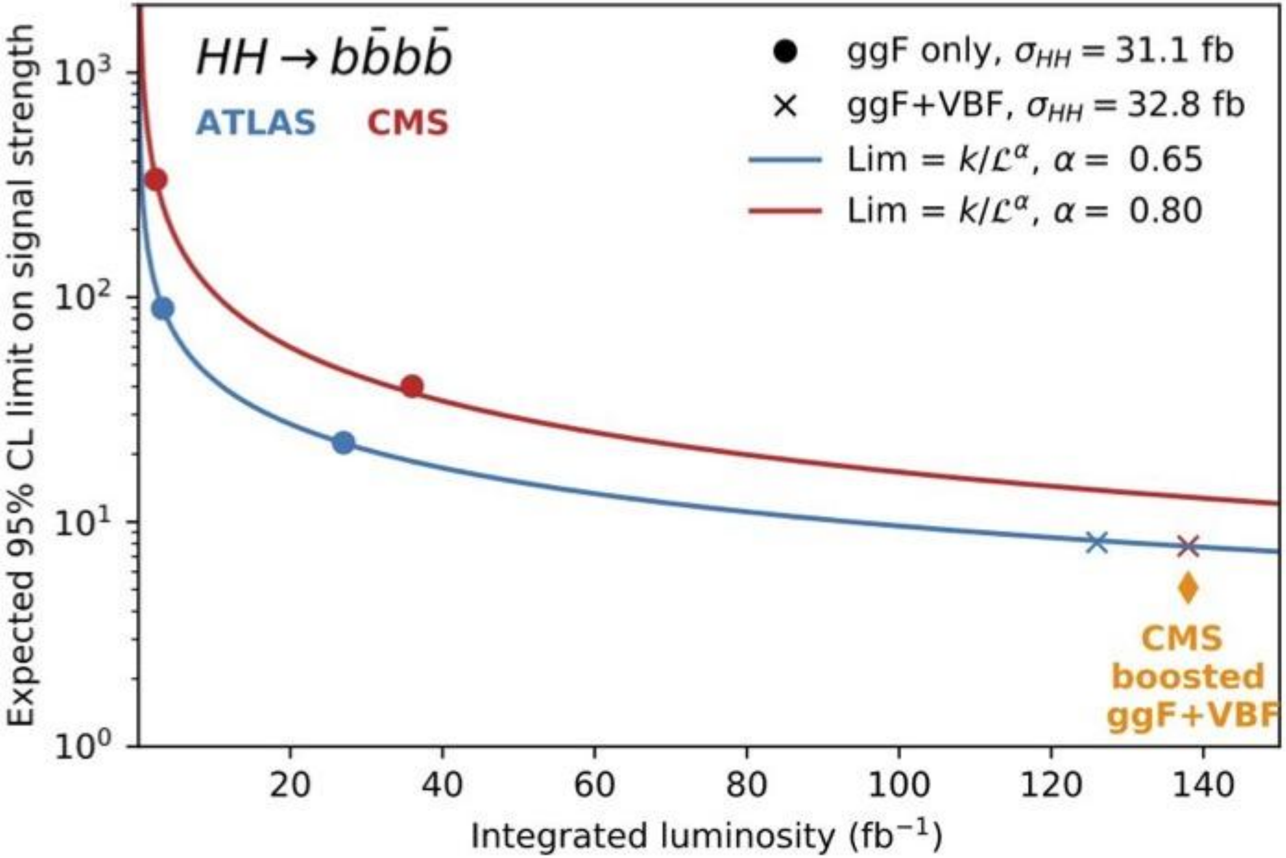
- bb$\gamma\gamma$ (HDBS-2018-34)
- bb$\tau\tau$ (HDBS-2018-40)
- 4b (HDBS-2019-29)

For this heuristic comparison, a loose cut on the bb$\tau\tau$ MVA discriminants was used to mimic the bb$\gamma\gamma$ BDT categories.



ATLAS / CMS comparison

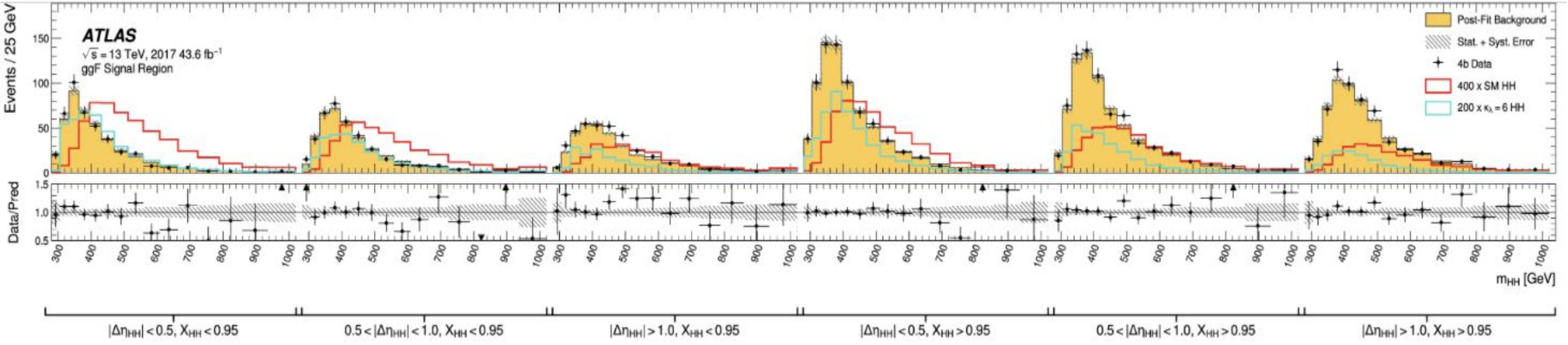
Slide from Rafael's



It works

Phys. Rev. D 108 (2023) 052003

Post fit plot for ATLAS HH4b non-res in analysis categories



Input variables used for the resolved analyses reweighting

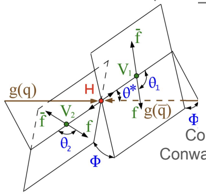
ATLAS: NN 2b ➜ 4b

Variable description	ggF	VBF
$\log(\Delta R_1)$: between the closest two HC jets	✓	
$\log(\Delta R_2)$ between the other two HC jets	✓	
$\log(p_T)$ of the 4th leading HC jet	✓	
$\log(p_T)$ of the 2nd leading HC jet	✓	
$\langle\lvert\eta\rvert\rangle$: average absolute value of the HC jets η	✓	✓
Number of jets in the event	✓	
$\log(p_{T,HH})$	✓	
ΔR_{HH}	✓	
$\Delta\phi$ between the jets in the leading HC	✓	
$\Delta\phi$ between the jets in the subleading HC	✓	
$\log(X_{Wt})$	✓	✓
Trigger bucket index	✓	✓
Year index		✓
Second smallest ΔR between the jets in the leading HC (out of the three possible pairings)		✓
Maximum di-jet mass out of the possible pairings of HC jets		✓
Minimum di-jet mass out of the possible pairings of HC jets		✓
Energy of the leading HC		✓
Energy of the subleading HC		✓

Same variables for non-res ggF and resonant resolved analyses.

CMS: BDT 3b ➜ 4b

Variable description	ggF	VBF
$p_{T,1}, p_{T,2}, p_{T,3}, p_{T,4}$: p_T of the four chosen b-jets	✓	✓
$m_H H$ 4-jet invariant mass	✓	✓
m_{H1}, m_{H2}: invariant mass of the Higgs Candidates	✓	✓
$p_{T,H1}$, $p_{T,H2}$: transverse momentum of the Higgs Candidates	✓	✓
$\lvert\Delta\eta(H1, H2)\rvert$	✓	✓
Scalar sum p_T of b-jets	✓	
Vector sum p_T of b-jets	✓	
$\Delta R^{H1}(bb)$, $\Delta R^{H2}(bb)$: opening angle between jets in HCs	✓	
ΔR_{min} out of the three possible pairings between b-jets	✓	
$\lvert\Delta\eta_{max}\rvert$ out of the three possible pairings between b-jets	✓	
$\lvert\cos\theta^*\rvert$: Abs value of the angle of one of the HCs with respect to the beam line in the center-of-mass frame of the four jets	✓	
$\lvert\cos\theta_b^{H1}\rvert$: Angle of one of the b-jets in the leading Higgs in the Higgs reference frame	✓	
$\sum R_e$: Sum of the resolution estimators of the three tightest WP b-tagged jets (based on DeepJet score)	✓	
N_B^T: Number of the above three jets passing the tight DeepJet WP	✓	
$\lvert\Delta\phi(H_1, H_2)\rvert$		✓
$m_j j$ between the VBF jets		✓
$\lvert\Delta\eta_{jj}\rvert$ between the VBF jets		✓
MVA score of ggF vs VBF BDT		✓



Colored graphic from J. Conway's 2018 SSI lecture

Likelihood ratio trick

Suppose that we train a classifier, $D(x) \in (0,1)$, to classify between 2 classes,

- 0: 2b class
- 1: 4b class

Train with **binary cross entropy** (maximize the probability of the correct class, or minimize negative log likelihood = loss)

If D(x) is p_{4b}, then 1-D(x) is p_{2b}, because there are only 2 classes and the prob need to sum to 1

$$\mathcal{L} = - \mathbb{E}_{x\sim p4b} [\log D(x)] - \mathbb{E}_{x\sim p2b} [\log(1 - D(x))]$$

The minimum of this loss is: $D^*(x) = \dfrac{1}{1 + p_{2b}(x) / p_{4b}(x)}$

Sanity check:

- If p_{4b} is high and p_{2b} is ~ 0, D*(x) = 1
- As p_{2b} is high and $p_{4b} \rightarrow 0$ D*(x) = 0 as expected

Rearrange for the **weight:** $w(x) = p_{4b}(x) / p_{2b}(x) = \dfrac{1}{D^*(x)} - 1$

This classifier D*(x) gives us the likelihood ratio $p_{4b}(x) / p_{2b}(x)$ 🙂

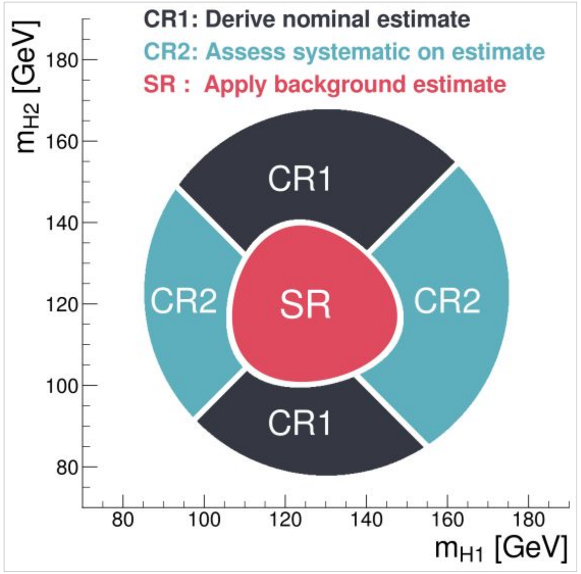
Background reweighting

Loss function ATLAS 4b resolved actually uses… iteration on a ~~classical~~ classifier theme.

- Multi-dimensional reweighting 2b ➜ 4b.

$$p_{4b} = w(x) \cdot p_{2b}(x)$$

Want to learn this



- Let Q(x) be a NN mapping from 2b ➜ 4b.

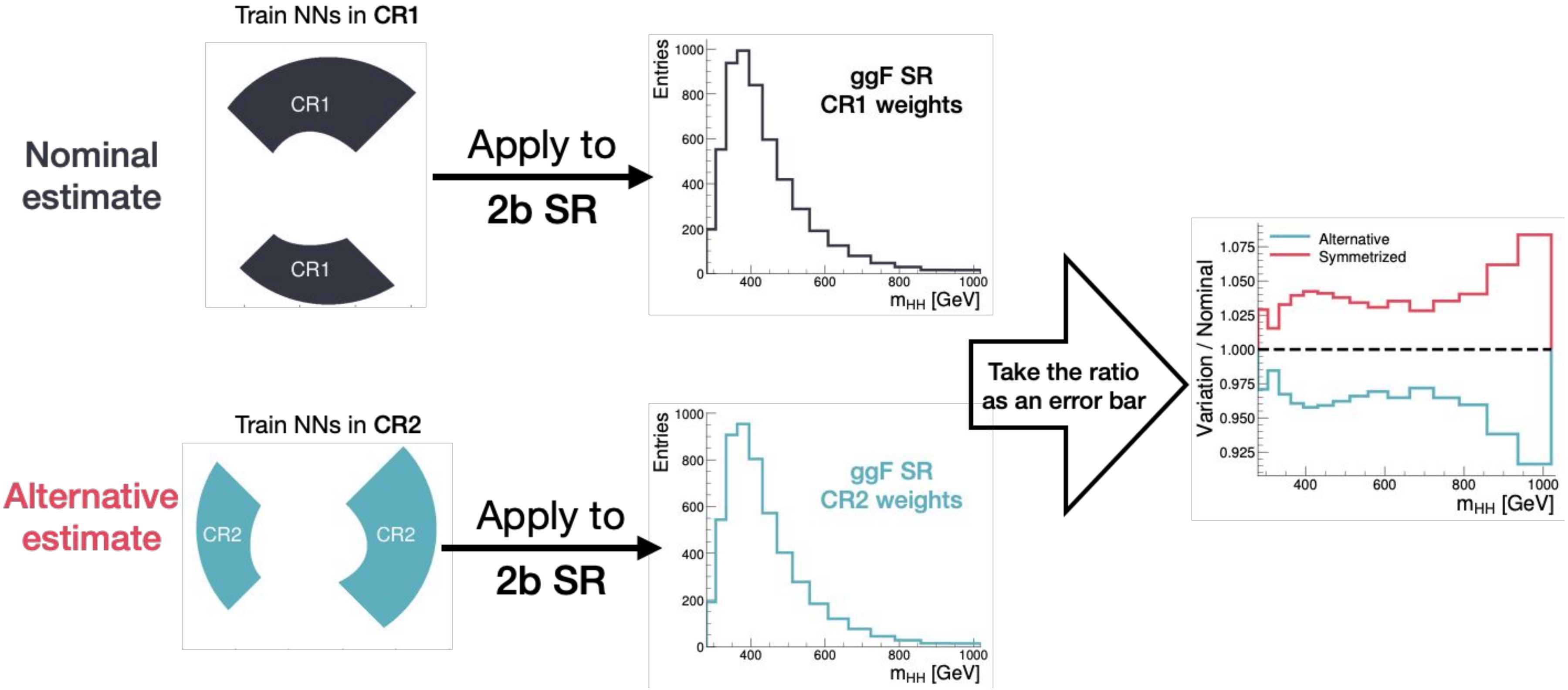
$$\mathscr{L}[Q(x)] = \mathbb{E}_{x \sim p_{2b}}\left[\exp\left(\frac{Q(x)}{2}\right)\right] + \mathbb{E}_{x \sim p_{4b}}\left[\exp\left(-\frac{Q(x)}{2}\right)\right]$$

Minimize loss in Control Region

$$Q^*(x) = \arg\min_{Q} \mathscr{L}[Q(x)] = \log \frac{p_{4b}(x)}{p_{2b}(x)} \implies w(x) = e^{Q^*(x)}$$

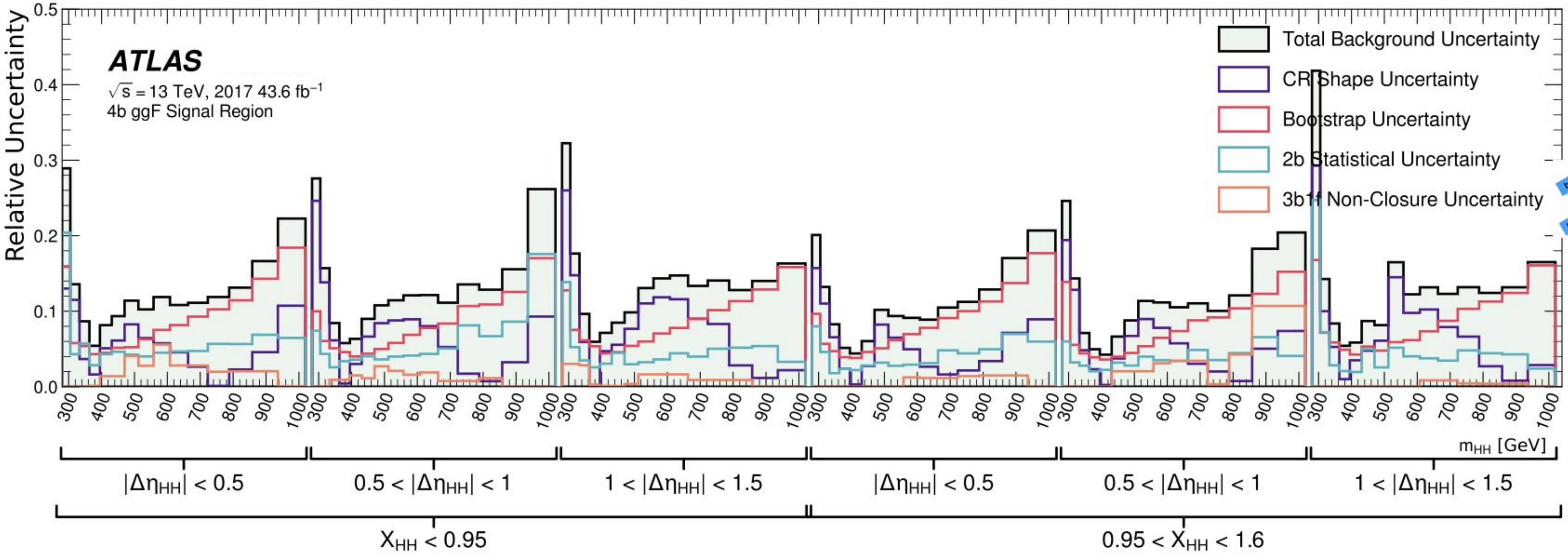
Apply w(x) to 2b Signal Region to get 4b prediction

Choice of Control Region



Systematics

[Phys. Rev. D 108 (2023) 052003]



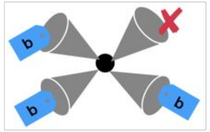
Background validation

Q: Does this proposal work?

→ Invert *every cut!!*

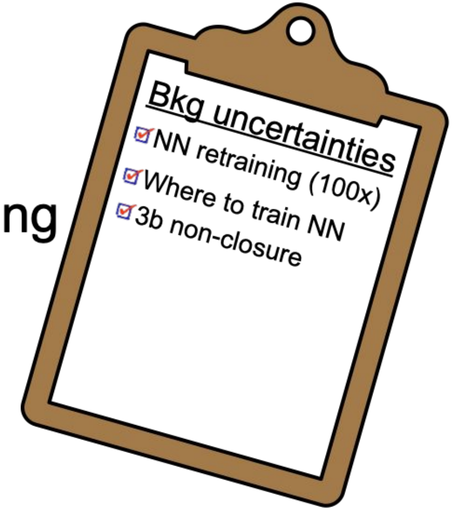
Trigger → ~~4~~ b-jets → ~~$\Delta\eta_{HH}$ < 1.5~~ → SR

3 b-jets



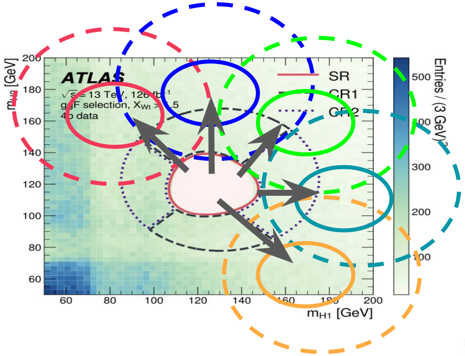
❌ Some mismodeling

→ Add uncertainty



$\Delta\eta_{HH}$ > 1.5

Shift the center (5x)



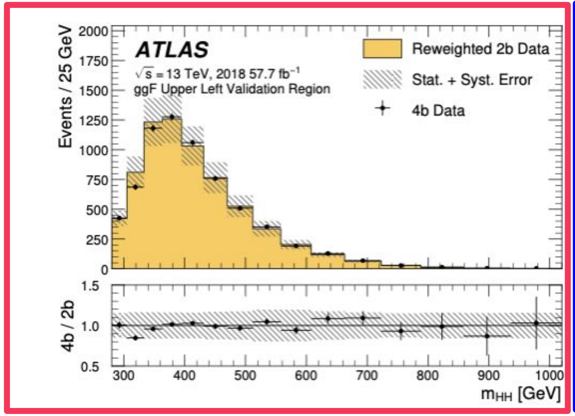
$$X_{HH} = \sqrt{\left(\frac{m_{H1} - 124\ \text{GeV}}{0.1 m_{H1}}\right)^2 + \left(\frac{m_{H2} - 117\ \text{GeV}}{0.1 m_{H2}}\right)^2}$$

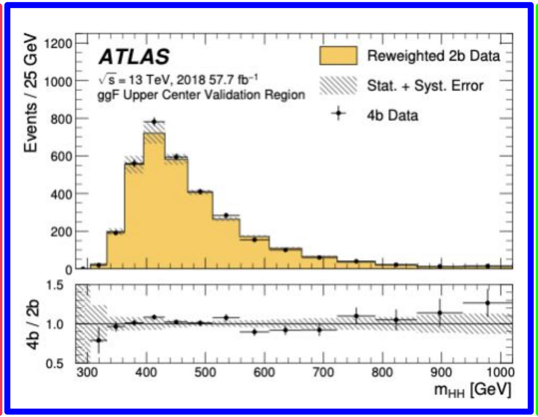
SR: $X_{HH} < 1.6$

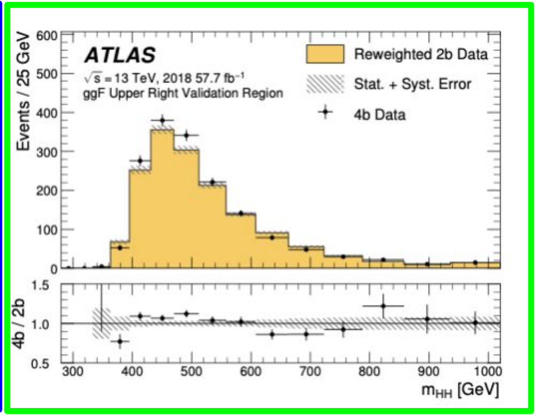
✅ Modeled by existing uncertainties

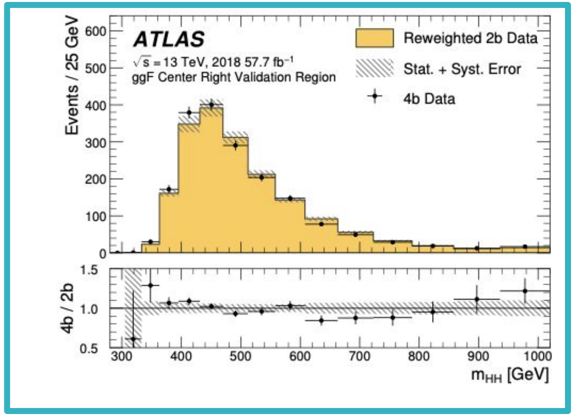
Phys. Rev. D 108 (2023) 052003

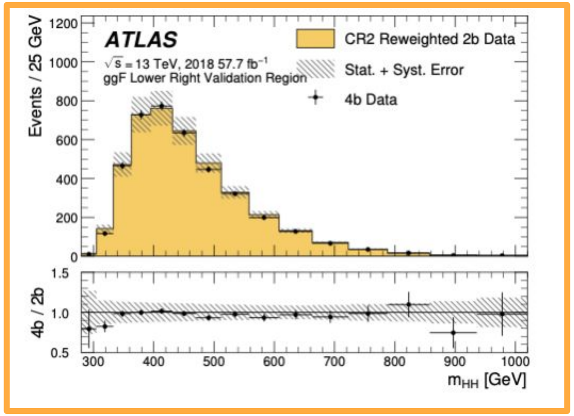
Background Validation







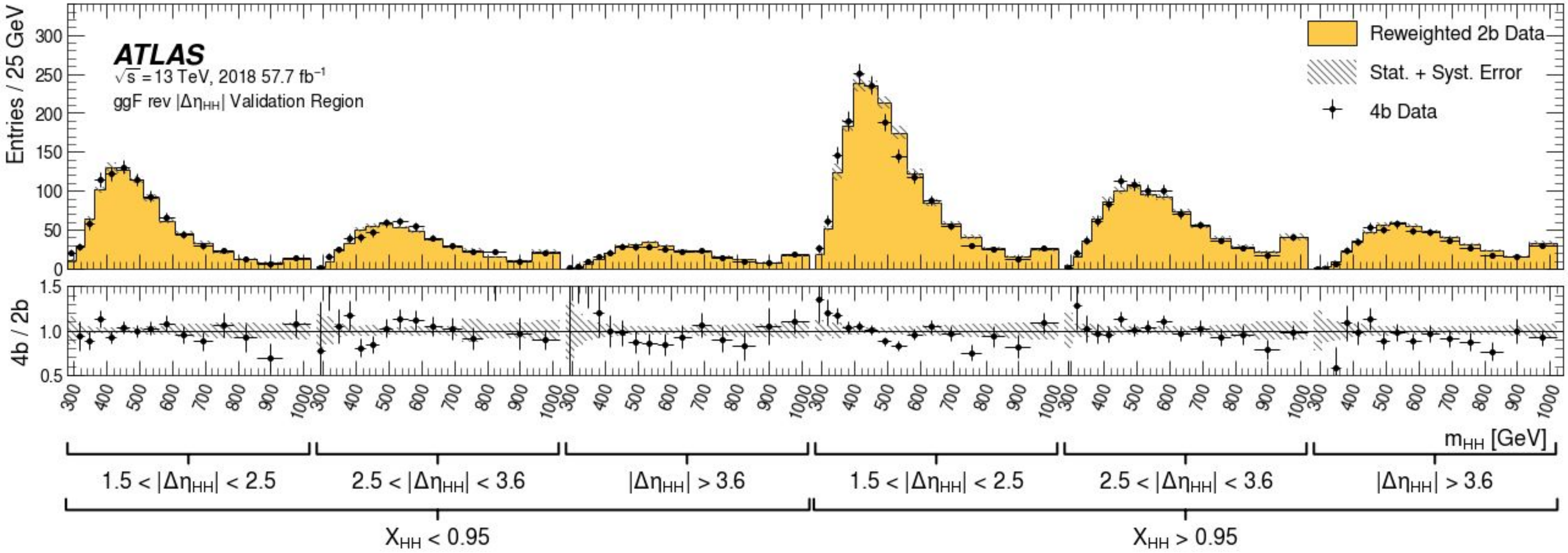




Phys. Rev. D 108 (2023) 052003

Background Validation: rev $\Delta\eta_{HH}$

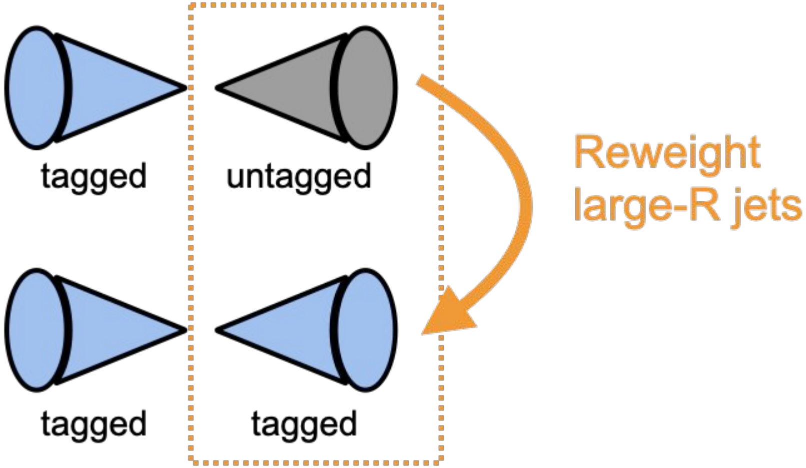
[Phys. Rev. D 108 (2023) 052003]



4b boosted ATLAS (resonant)

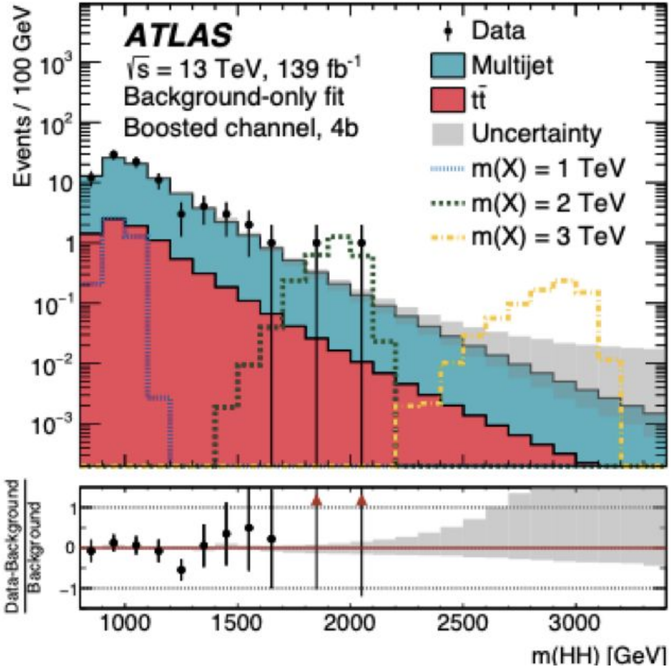
4b resonant

Still a multidimensional reweighting (Low tag -> high tag)



QCD fit in CR with

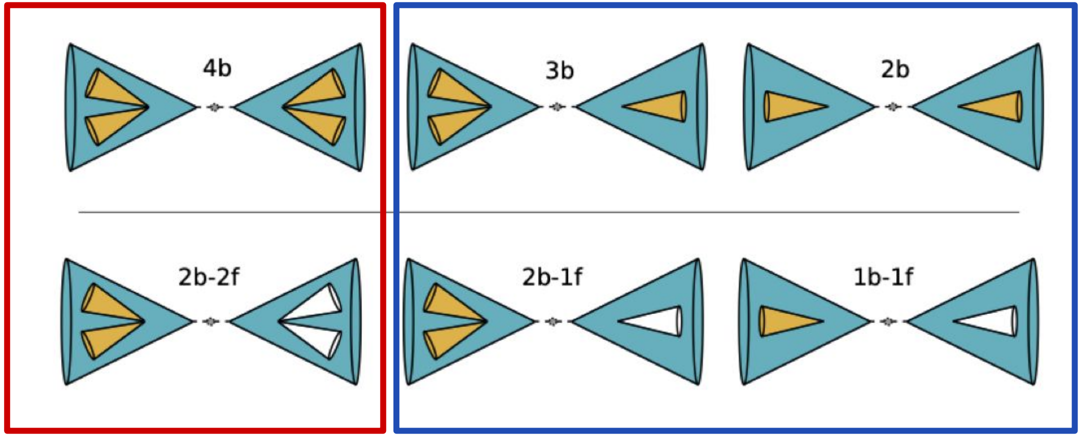
- Shape: iterative spline reweighting
- Norm: combined fit with **$t\bar{t}$** to m_{H1} shape



Dedicated categories for number of b-tagged track jets: 4b, 3b and 2b-split.

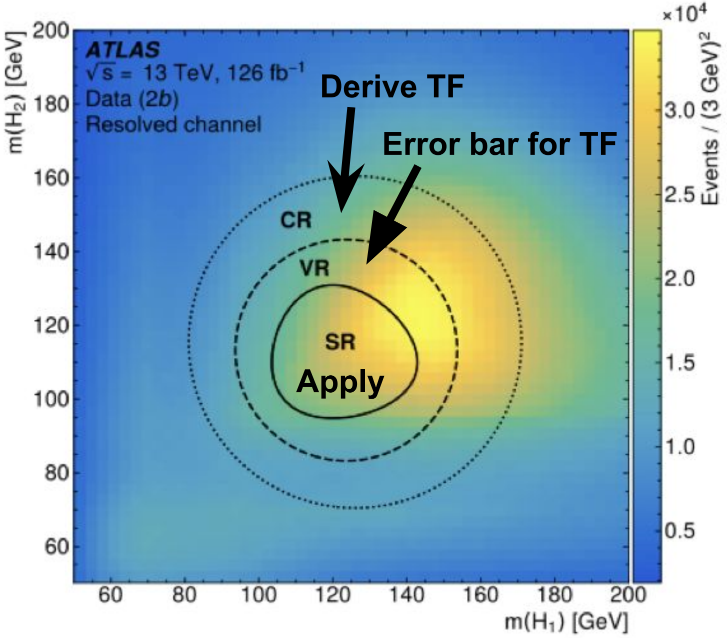
Details of the background estimate

Still a multidimensional reweighting (Low tagged jet -> high tagged jet)



In this region, the stat unc is so high, that the 2b-2f shape is taken directly as is (no transfer function).

Iterative spline reweighting
(transfer function)



Iterative spline reweighting

For each "iteration", sequentially correct a set of reweighting variables

Spline fit to ratio of tagged / untagged 1d histograms

$$W_i = W_{i-1} \times \left[\lambda_i \prod_j \left(f_{ij}(x_j) - 1 \right) + 1 \right]$$

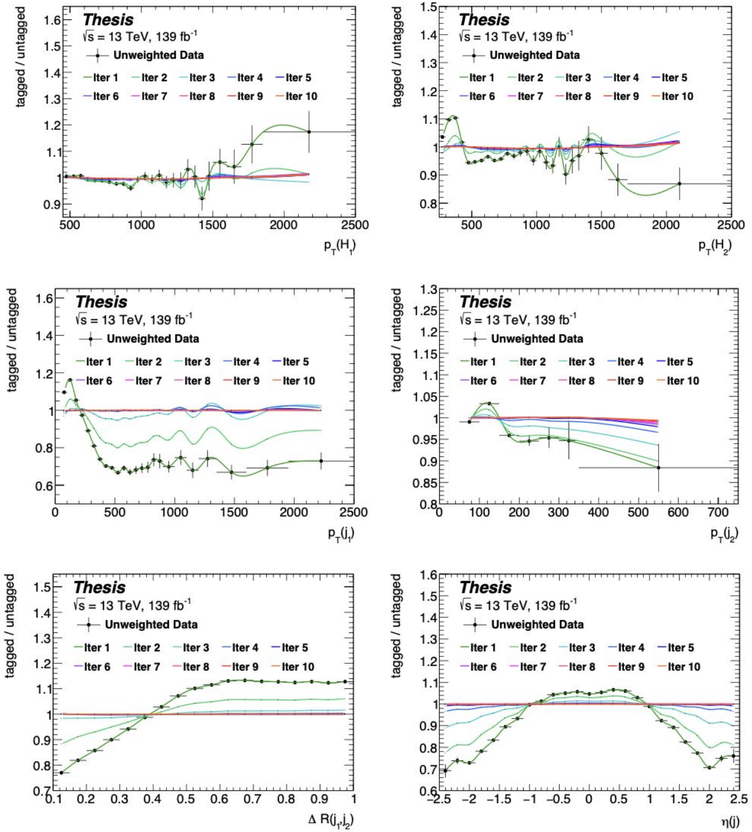
"j" are the kinematic reweighting variables:

- p_T of the "tagged" large-R jet
- p_T and η of the b-tagged VR track jet
- ΔR(lead trk jet, subl trk jet) [when applicable]

After 10 iterations, the fit has converged.

This shape reweighting happens **directly on data** inclusively for QCD and tt̄.

Alex Emerman's thesis



Normalizations

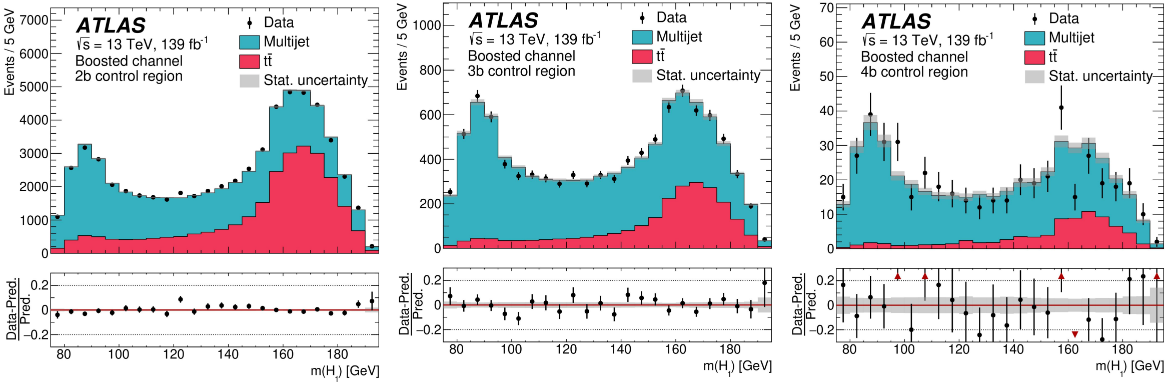
[Phys. Rev. D 105 (2022) 092002]

Combined fit for the QCD and $t\bar{t}$ normalizations with the m_{H1} shape.

- QCD yield from the low-tagged region
- $t\bar{t}$ taken from MC

$$N^{tag}_{i,\mathrm{data}} = \mu_{QCD}\left(N^{untag}_{i,\mathrm{data}} - N^{untag}_{i,t\bar{t}}\right) + \alpha_{t\bar{t}} - N^{tag}_{i,t\bar{t}}$$

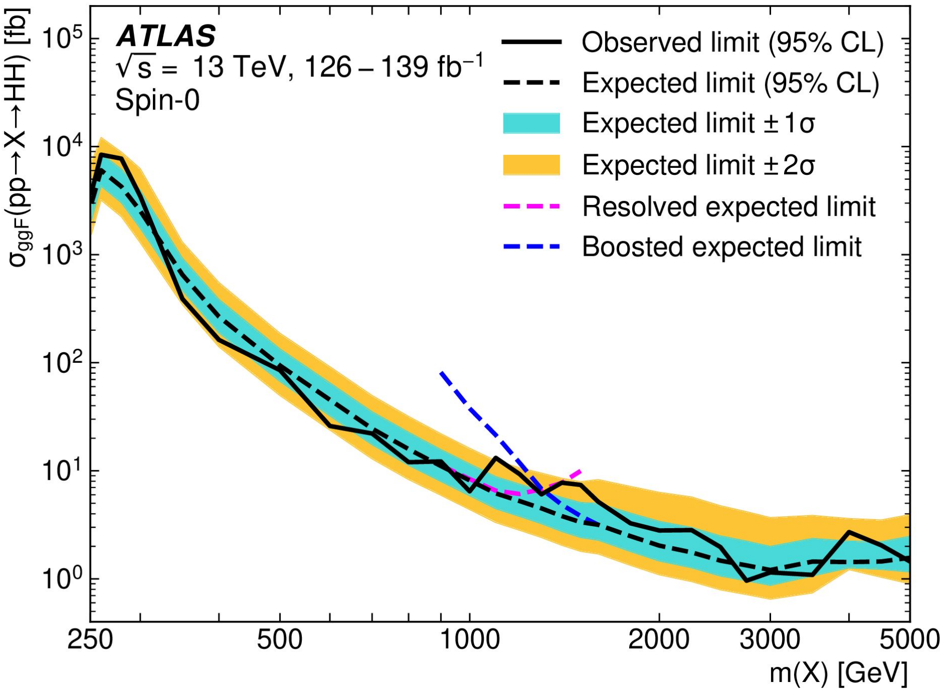
Fit separately for each of the three SRs (4b, 3b and 2b split), **6 norm factors**



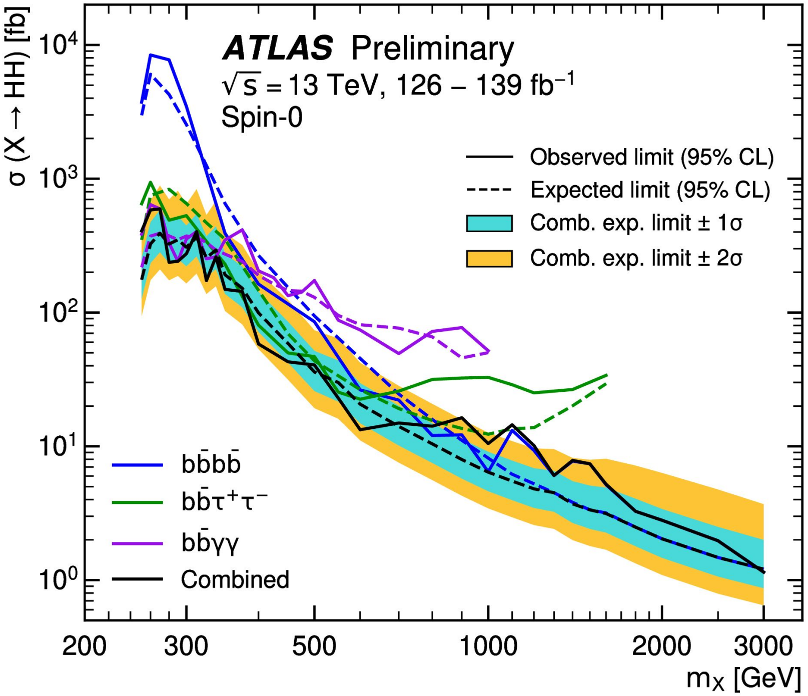
Region	$2b$	$3b$	$4b$
μ_{MJ}	0.05435 ± 0.00056	0.1204 ± 0.0023	0.0272 ± 0.0015
$\alpha_{t\bar{t}}$	0.863 ± 0.011	0.786 ± 0.042	1
Correlation	−0.74	−0.74	0

ATLAS resonant combination

[Phys. Rev. D 105 (2022) 092002]



[ATLAS-CONF-2021-052]

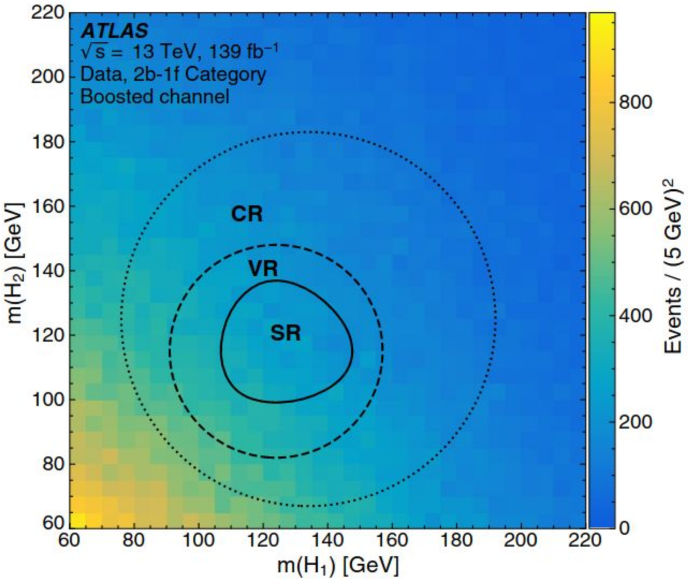


More Backup slides

Definition of boosted control regions (1)

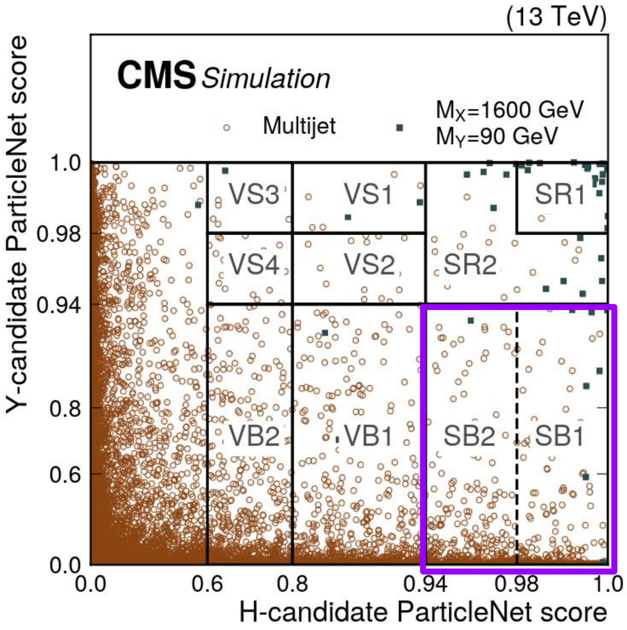
HH resonant

Phys. Rev. D 105 (2022) 092002



HY resonant

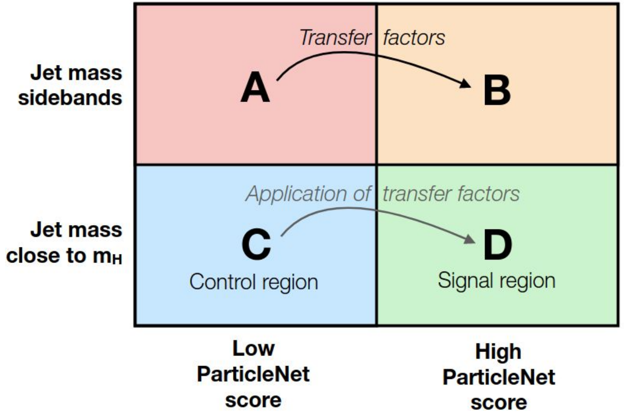
PhysLetB.2022.137392



Definition of boosted control regions (2)

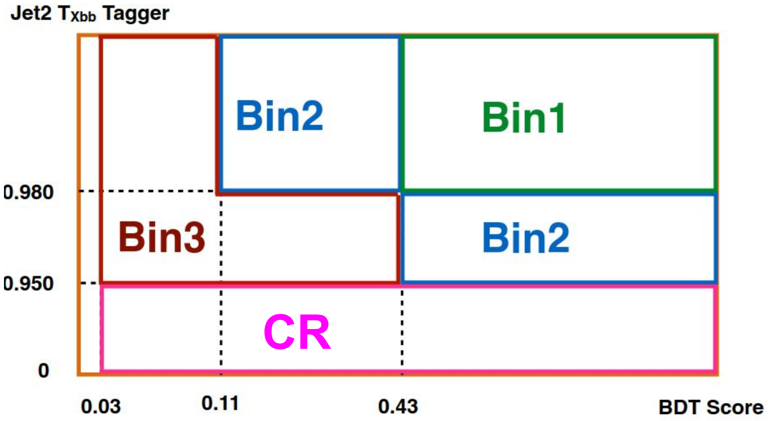
HH NR VBF

PhysRevLett.131.041803



HH NR ggF

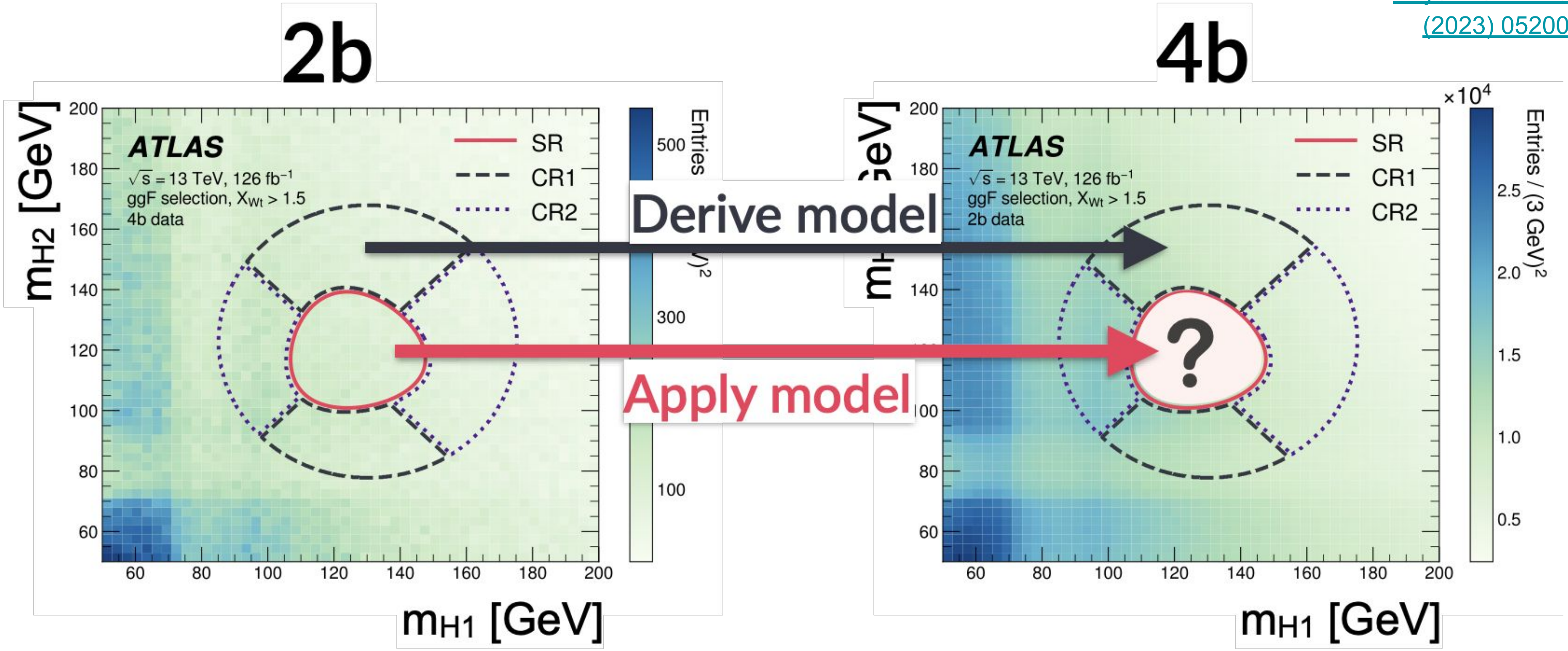
PhysRevLett.131.041803



AK8 m_{reg} region	$50 < m_{reg}^{lead} < 110\,GeV$	$110 < m_{reg}^{lead} < 150\,GeV$	$150 < m_{reg}^{lead} < 200\,GeV$
$50 < m_{reg}^{subl} < 90\,GeV$	Transfer factor regions (A & B)		
$90 < m_{reg}^{subl} < 145\,GeV$	Validation region (D)	search region (D)	Validation region (D)
$145 < m_{reg}^{subl} < 200\,GeV$	Transfer factor regions (A & B)		

The background model

Phys. Rev. D 108 (2023) 052003



CMS HH(4b) NR resolved

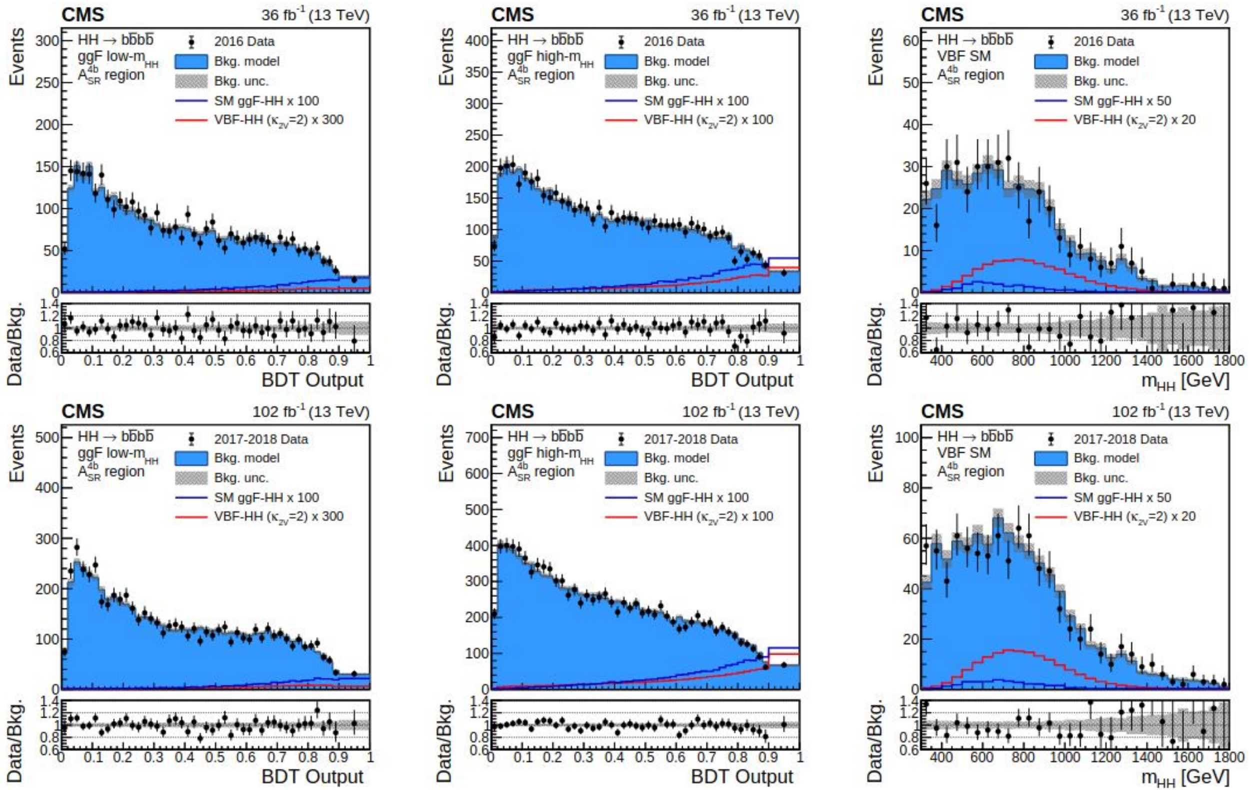


Figure 1: Distributions of the events observed in the A^{4b}_{SR} signal region for 2016 (top) and 2017–2018 (bottom) data. The two leftmost columns show the BDT output in the low- and high-mass categories, and the rightmost column shows the m_{HH} distribution in the VBF SM-like category.

CMS HH(4b) NR boosted: VBF

[PhysRevLett.131.041803]

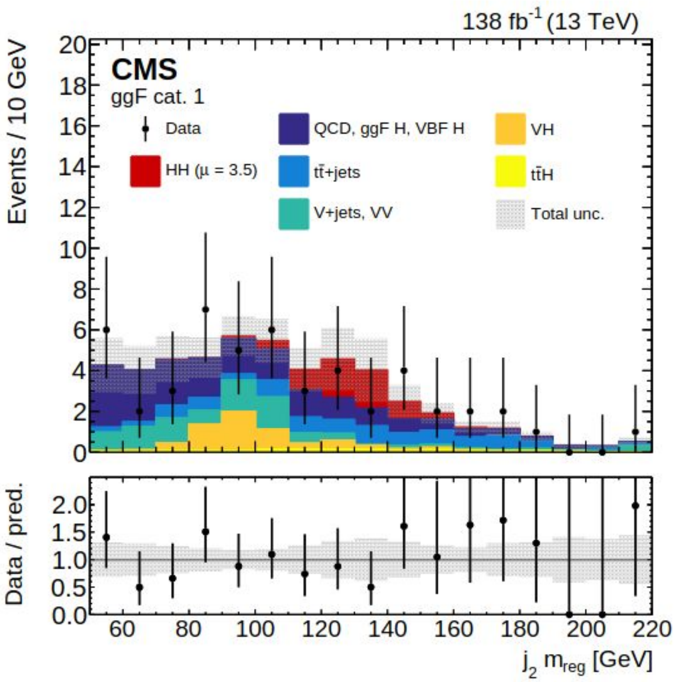
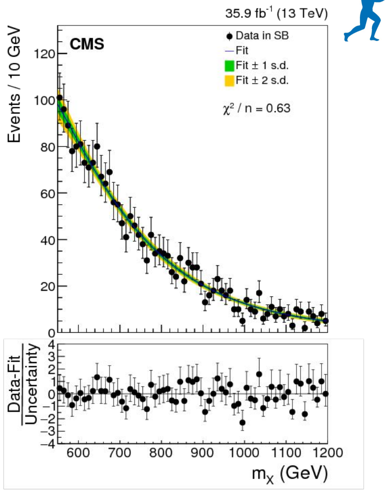


Figure 1: The data and fitted signal and background distributions for the $D_{b\bar{b}}$-subleading jet regressed mass are shown for the ggF BDT event category 1, the category accounting for most of the sensitivity to the ggF HH signal. The SM HH ($\kappa_{2V} = \kappa_V = \kappa_\lambda = 1$) signal is shown scaled to the best fit signal strength $\mu = 3.5$. The lower panel shows the ratio of the data and the total prediction, with its uncertainty represented by the shaded band. The error bars on the data points represent the statistical uncertainties.

Parametric function bkg. modelling

- Directly model the shape with a function
- Can be used when searching for a resonance on a smoothly falling background
  - Turn-on effects may be problematic



- Resonant HH→4b search, 2016 data (JHEP08(2018)152)
  - Functional forms chosen in studies performed before unblinding, using **control regions**
    - Signal-free regions with kinematic properties similar to events in signal regions