VH(bb) Simplified template cross section measurements with the CMS experiment

Saswat Mishra
On behalf of the CMS collaboration

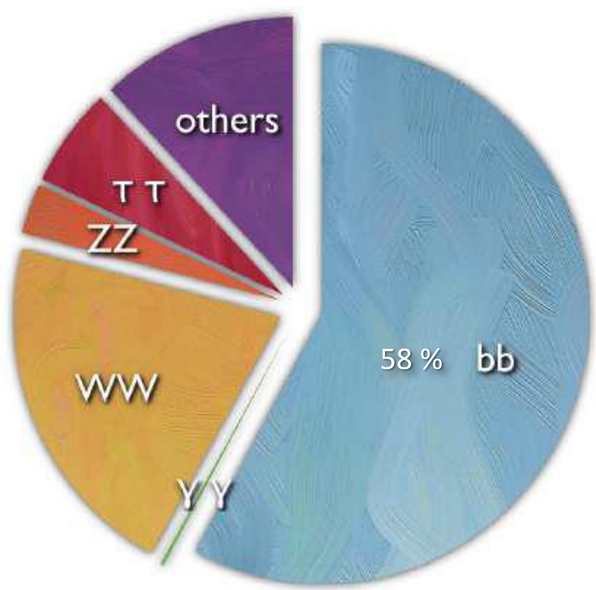
at LHC days in Split, 2022

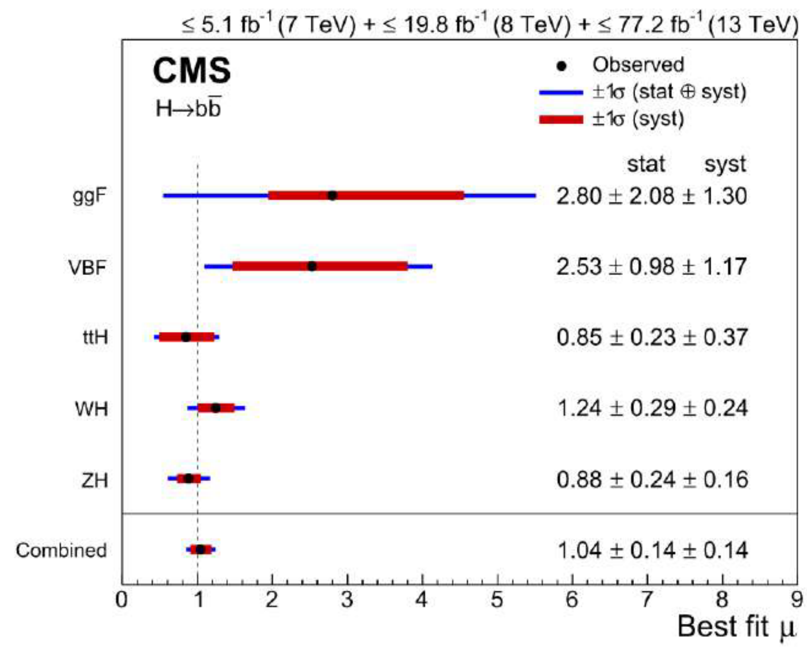
October 3rd, 2022

Supported by:
HRZZ
Croatian Science Foundation

Introduction

- ✓ $H \rightarrow b\bar{b}$ is the dominant decay mode with a branching fraction of 58% (Phys. Rev. Lett. 121 (2018) 12)
  - Observed by the CMS experiment in 2018 with observed (expected) significance of 5.6 (5.5)
- ✓ $VH \rightarrow b\bar{b}$ process provides highest sensitivity among all Higgs production modes
  - Effective in reducing multi-jet background events
  - Ability to use lepton and MET (Missing Transverse Energy) triggers from the Vector boson
  - W/Z produced generally back-to-back w.r.t Higgs

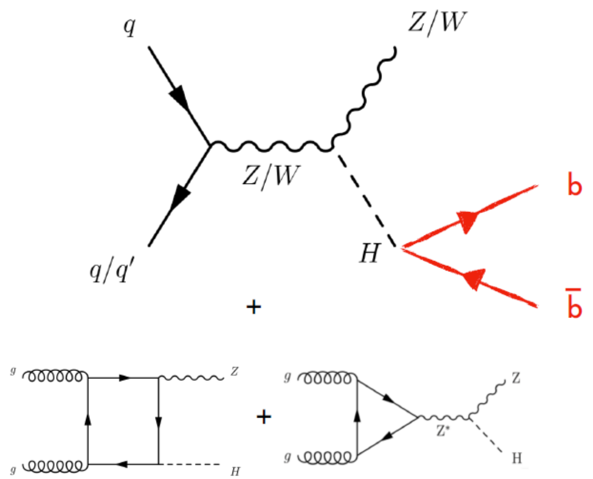




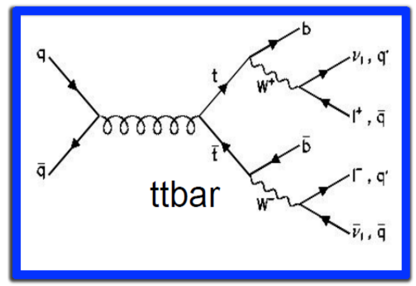
$H \rightarrow b\bar{b}$ observation via all higgs production modes

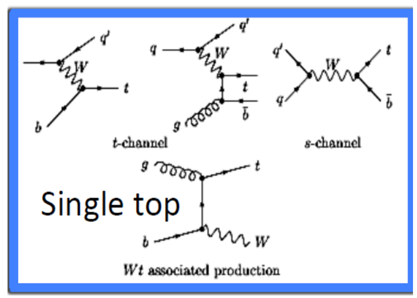
$VH \to b\bar{b}$ process

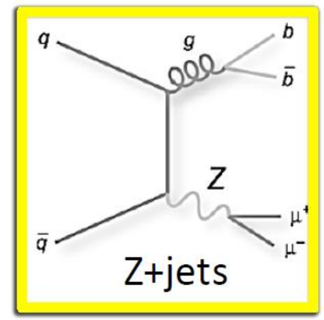
signal process

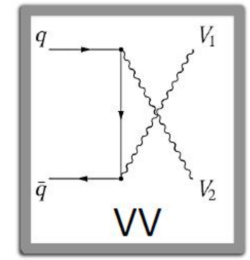


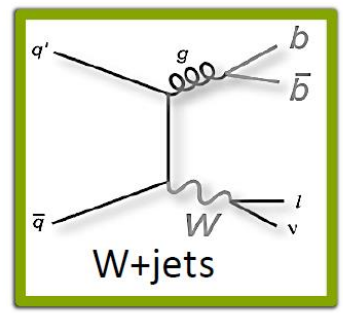
Major Backgrounds







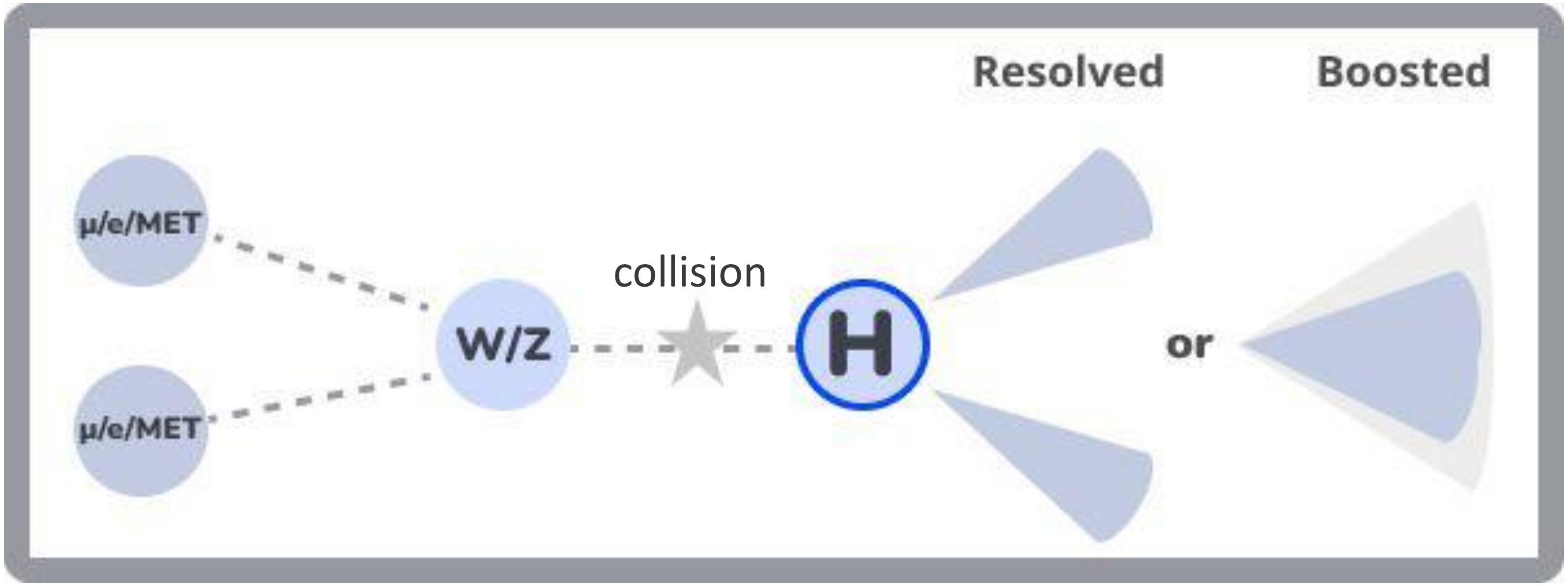




- ✓ 3 channels based on vector boson final states
  - 0-lepton ($Z \to \nu\nu$), 1-lepton ($W \to l\nu$), 2-lepton($Z \to l\bar{l}$)
- ✓ Different kinds of background dominate in different channels:
  - $t\bar{t}$, W+Jets and Z+Jets in 0-lepton
  - $t\bar{t}$ and W+jets in 1-lepton
  - DY+jets in 2-lepton

Higgs Topology

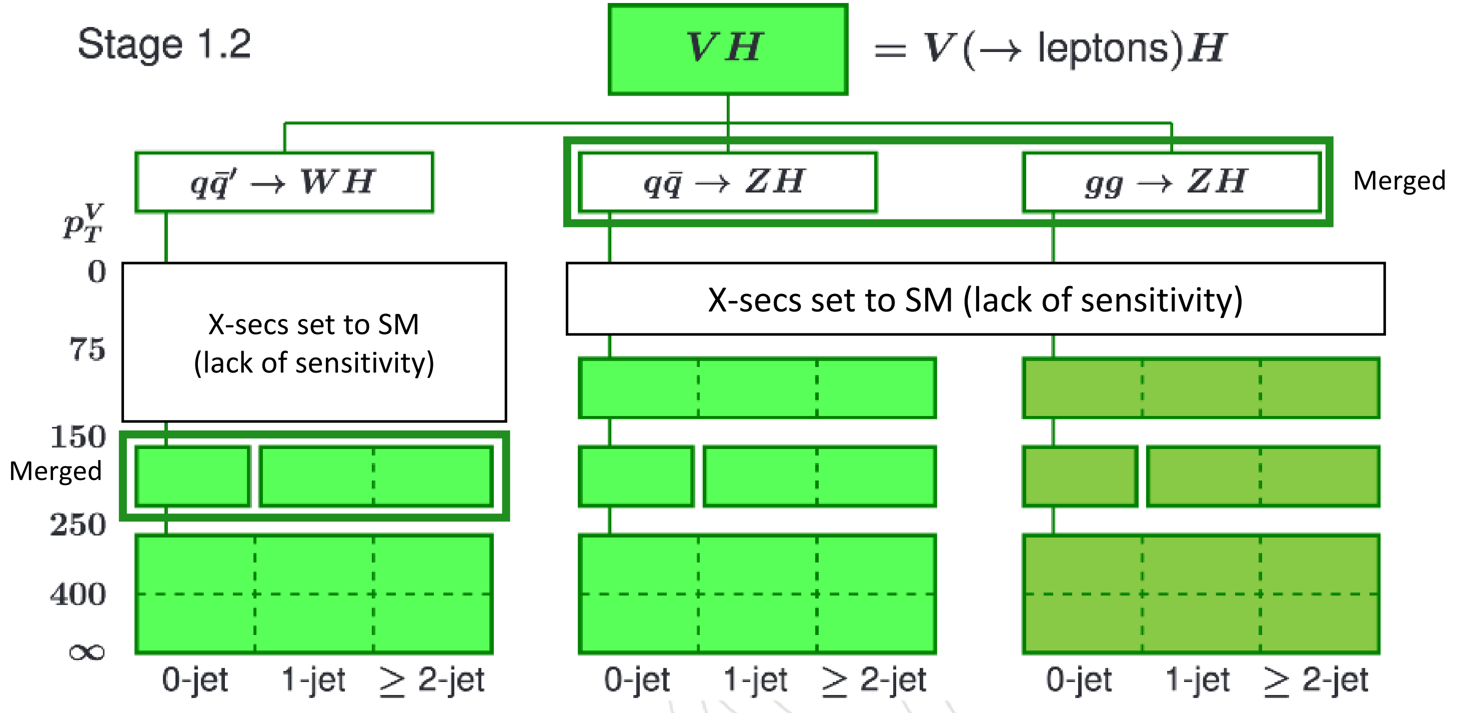
- ✓ With higher Lorentz boost ($p_T V > 250$ GeV), angular separation between pair of b-jets reduces to form a single jet cone of large R
- ✓ Dedicated boosted analysis for Higgs decays resulting in two merged b jets clustered into a single large cone jet
  - Both resolved and boosted analyses are combined to maximize sensitivity
- ✓ 2 different jet clustering algorithms are used:
  - Normal resolved jets are clustered using Anti-Kt algorithm with a radius of 0.4 (AK4 jets)
  - Large cone jets (Fat jets) are clustered with a radius of 0.8 with same algorithm (AK8 jets)
- ✓ b-tagging algorithms are used to distinguish b-jets from c/light flavored jets:
  - Resolved b-jets are tagged using the DeepCSV algorithm
  - Boosted fat jets are tagged using the DeepAK8 algorithm



STXS Framework

✓ Simplified Template Cross-Section (STXS) framework is used for fiducial cross-section measurements
  - Measure cross-section in mutually exclusive phase space regions
  - Reduction of theoretical uncertainties
  - Possibility to combine ATLAS and CMS measurements easily
  - Cross-section measured in bins of $p_T V$ adopted by STXS (stage 1.2) categorization (below)

arXiv:1610.07922, arXiv:1605.04692

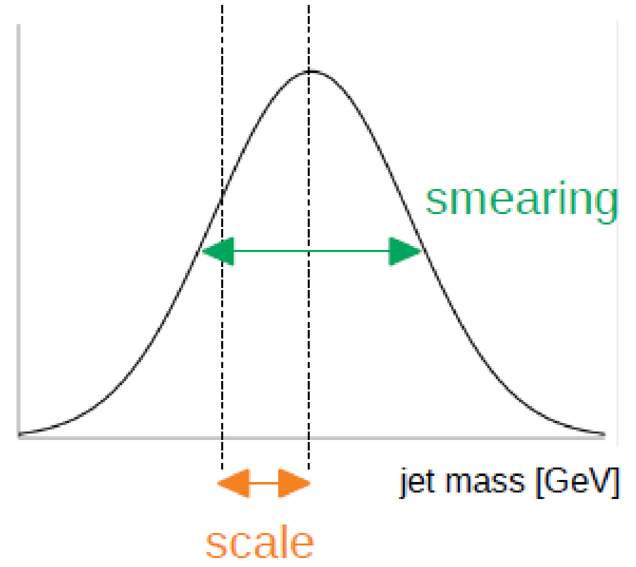


Improvement in resolution of jet parameters

✓ **Dedicated b-jet energy regression**

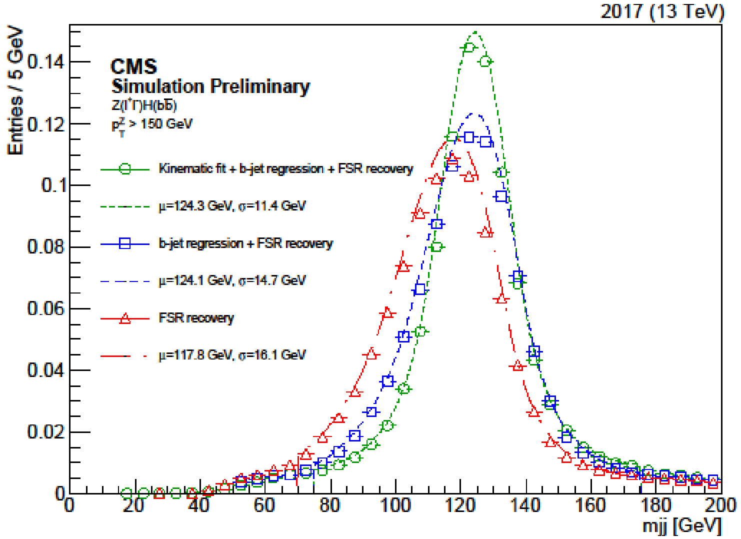
(Comput. and Soft. for Big Sci. 4, 10 (2020))

- DNN based regression
- Energy correction to account for escaping neutrino from semi-leptonic decay
- Improve the detector response



✓ **Dedicated Jet smearing and scaling correction**

- fit uses 2-jet event topology in which the jet resolution can be measured by the jet system balance against the Z in the transverse plane



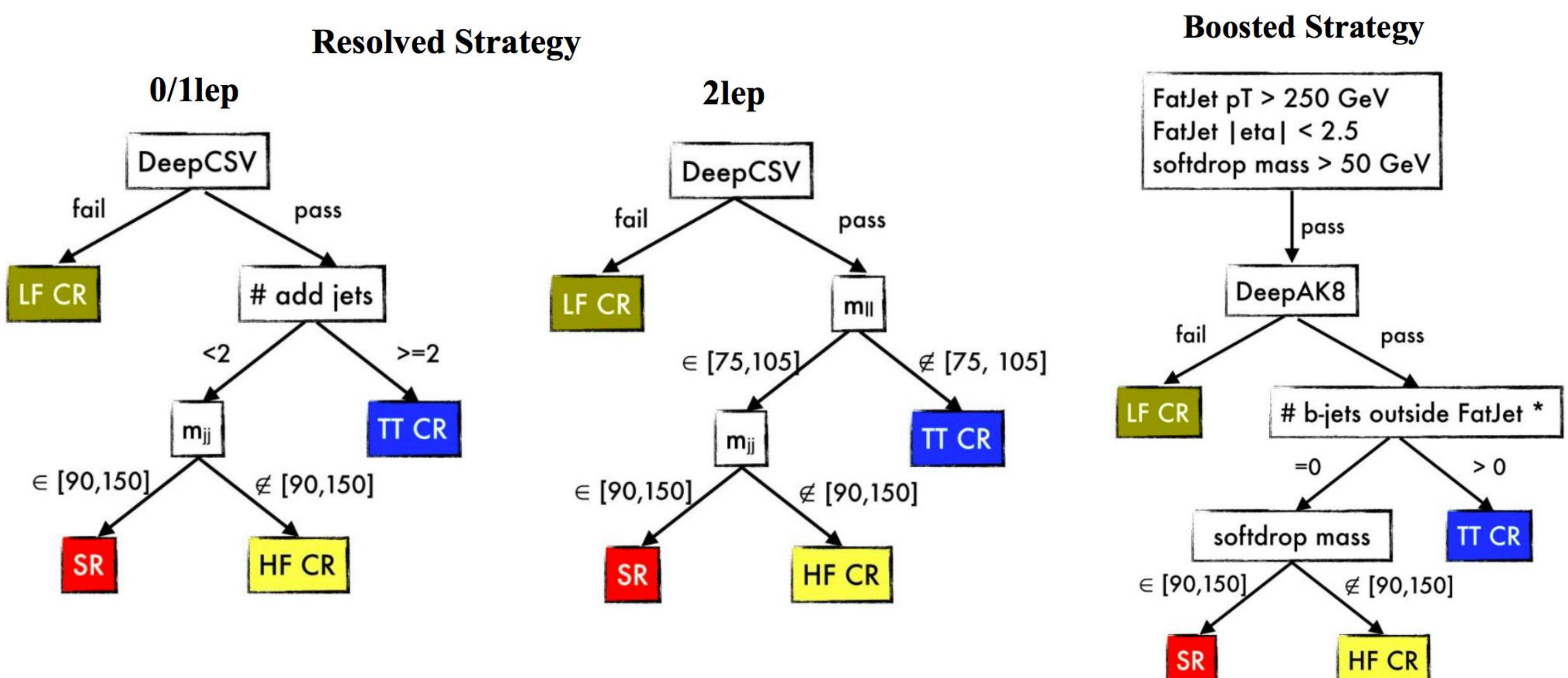
✓ **Kinematic fit in 2-lepton**

- Possible in 2-lepton due to better resolution of lepton momenta than jets and absence of intrinsic MET
- Fit leptons and jets within uncertainties using the constraints:

  $m(ll) = m(Z)$ & $p_T(total) = 0$

Analysis selection & strategy

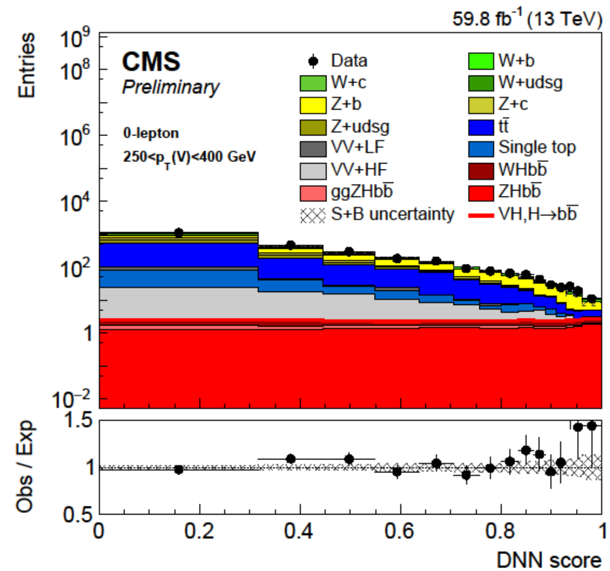
- ✓ 3 lepton channels based on vector boson decay, split in STXS bins
- ✓ 3 control regions (CRs) enriched in main backgrounds: V+LF, V+HF and TT
  - Split into low (2-lepton channel), medium and high $p_T V$ bins
- ✓ Overlap between resolved and boosted events optimized to ensure maximum sensitivity
  - Most relevant for high $p_T V$ bins
- ✓ Total: 51 control regions, 30 signal regions per year → 243 regions for Full Run 2 !!!
- ✓ **Simultaneous maximum likelihood fit of SR+CR using MVA output score in SR and V+HF regions**
  - Dedicated process scale factors to constrain major backgrounds in their respective control regions



Multivariate methods

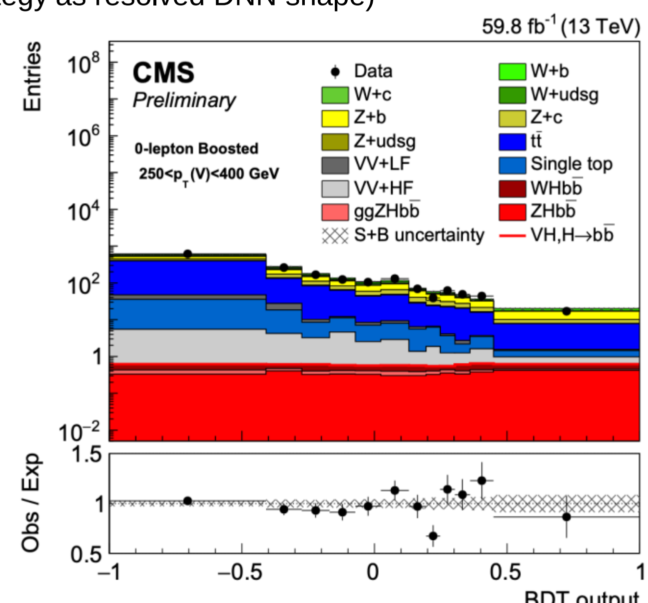
DNN based discriminant for resolved analysis

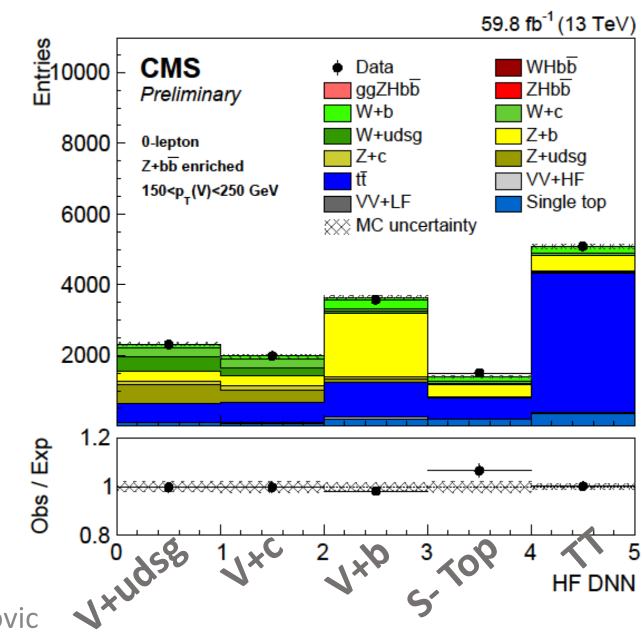
- Input features: channel-dependent high-level features(e.g. dijet properties, vector boson kinematics)
- SR: binary signal vs. background classification (binning in signal region optimized for sufficient and equal contribution of signal in each bin)
- HF CR (0/1 lepton): multi-classification into five classes (V+udsg, V+c, V+b, single top, tt)



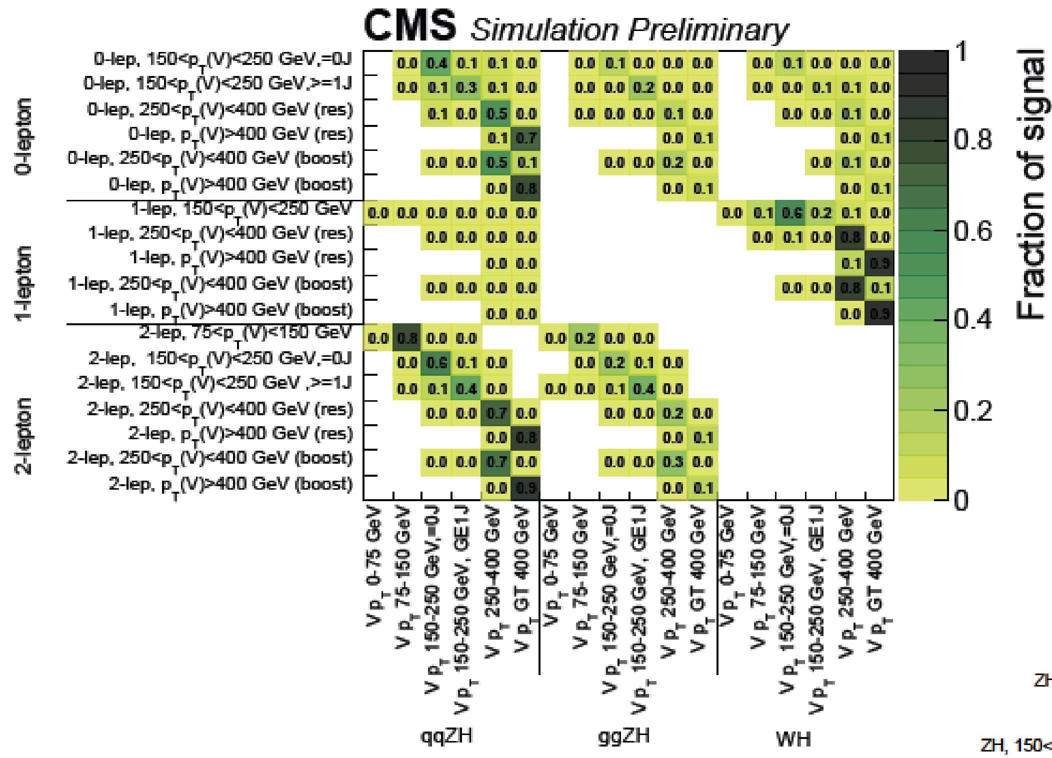
BDT-based discriminant for boosted analysis

- Input features: Fat-Jet information (softdrop mass, DeepAK8 bbVsLight output, pT), reconstructed vector boson properties, resolved features for overlap events
- SR: binary signal vs. background classification (same binning strategy as resolved DNN shape)

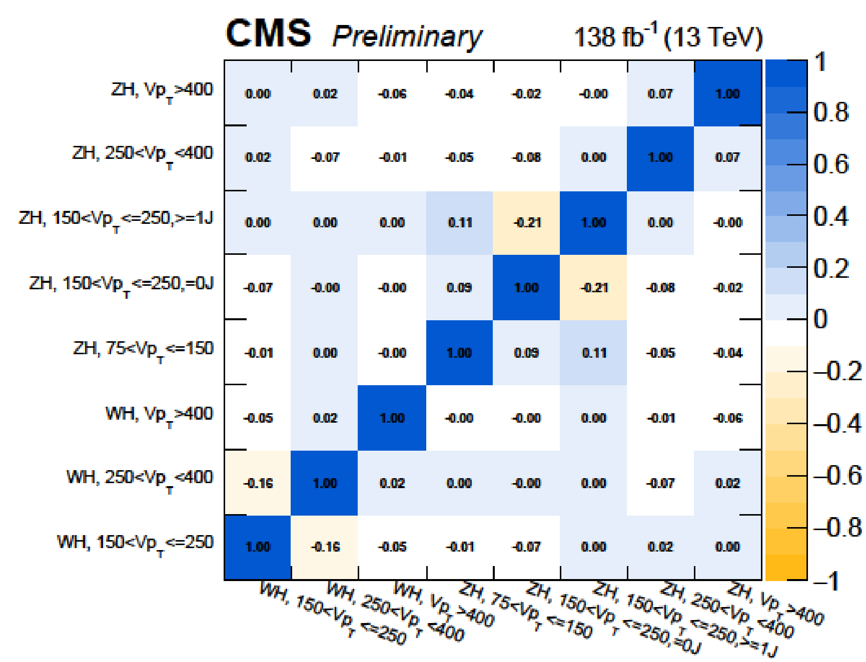




Signal Composition



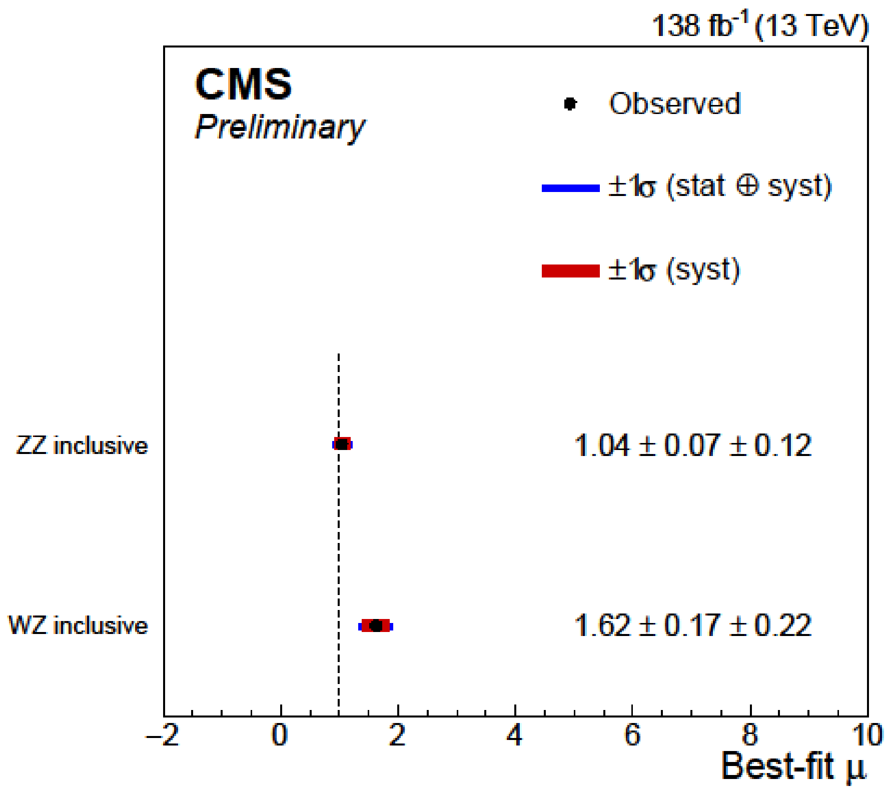
- ✓ qqZH and ggZH processes merged in corresponding STXS bins
- ✓ signal purity is higher in 2lep channel compared to 0lep/1lep across all the STXS bins
- ✓ Good agreement between the reconstruction categories and gen-level STXS bins



- ✓ correlation matrix roughly diagonal across all STXS bins:
  - Largest correlation (20%) in μ for medium $p_T V$ ZH region in 0-jet and >1-jet bins

$VZ \to b\bar{b}$ cross-check analysis

- ✓ Cross check analysis targeting $VZ \to b\bar{b}$ process as standard candle to validate VHbb analysis
  - Similar signal topology but much higher production cross section than $VH \to b\bar{b}$ (~110 times higher)
  - Mass window altered to match Z boson peak instead of Higgs peak
  - Dedicated MVAs trained in corresponding regions similar to VHbb analysis
- ✓ Inclusive μ=1.16±0.13 corresponding to 9.6σ (8.9σ) observed (expected) significance



Systematic Uncertainties

	$\Delta\mu$
Background (theory)	+0.067 −0.064
Signal (theory)	+0.082 −0.060
MC stats.	+0.092 −0.093
Sim. modelling	+0.070 −0.066
b tagging	+0.059 −0.041
Jet energy resolution	+0.045 −0.057
Luminosity	+0.041 −0.034
Jet energy scale	+0.029 −0.036
LeptonID	+0.016 −0.002
Trigger(MET)	+0.001 −0.001

significant theory corrections on background and signal

dominant source of uncertainty: MC statistics

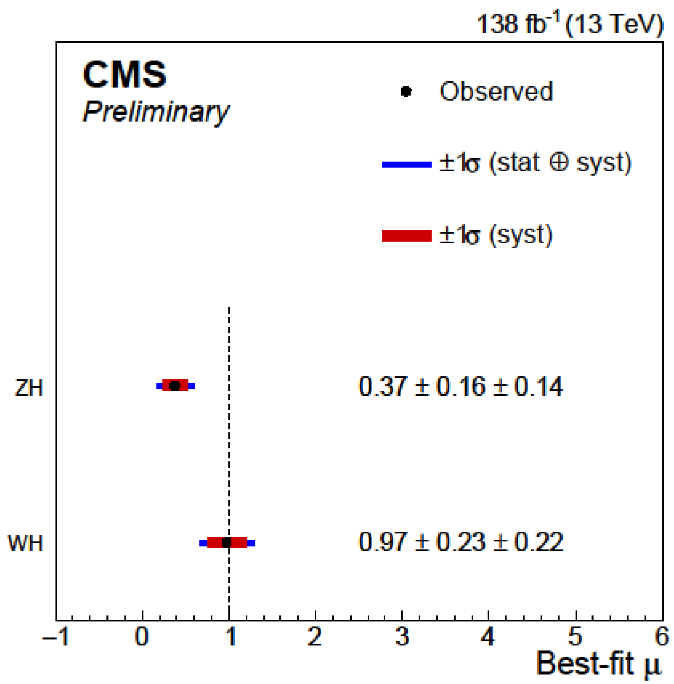
modeling of V+jets backgrounds:
- reweighting of LO samples in 2016
- $\Delta R(b\bar{b})$ reweighting of NLO samples 2017/2018

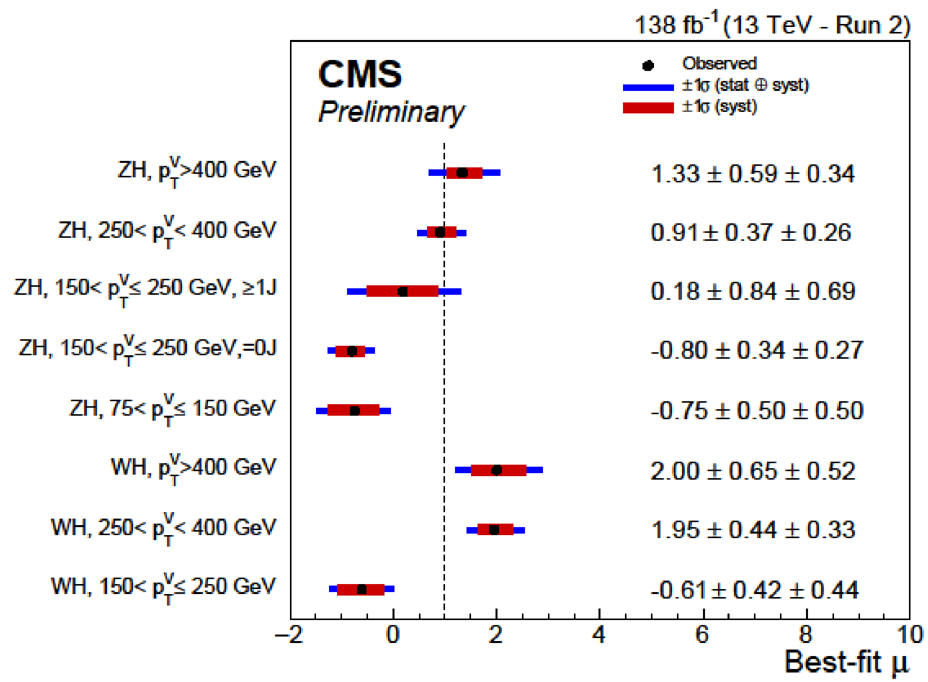
b tagging uncertainties for resolved and boosted tagger

- ratios of variation/nominal are **symmetrized and smoothed** for all JES, JER, b-regression uncertainties and the uncertainty assigned to the pile-up reweighting
- **correlated** uncertainties: theoretical uncertainties (PDF, scale, cross sections, branching ratio, …) and experimental uncertainties (JES, JER, pile-up reweighting)
- **uncorrelated** uncertainties: b tagging, regression scaling and smearing, category migration uncertainties, background normalization scale factors

Results

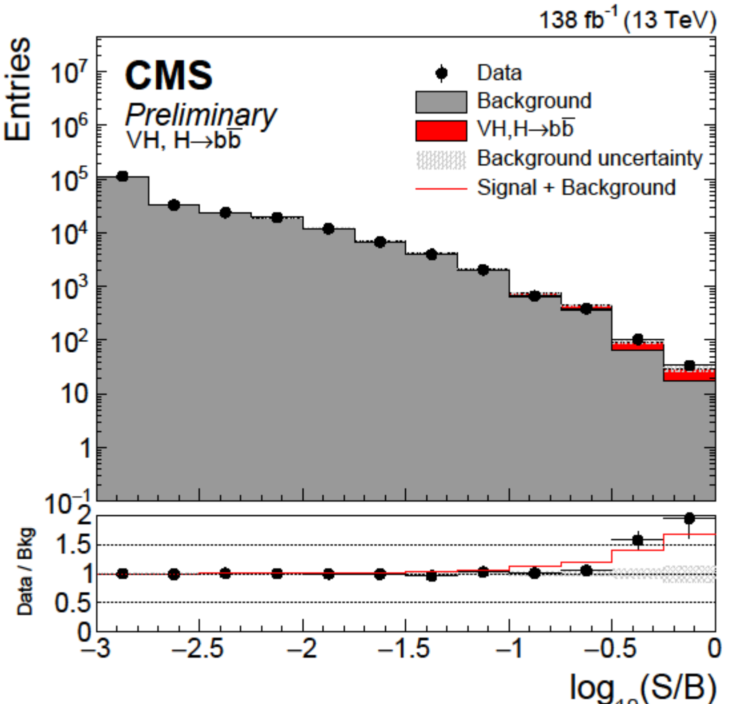
- ✓ μ extracted from binned maximum likelihood fit for each STXS bin
- ✓ Inclusive Run 2 signal strength $\mu = 0.58^{+0.19}_{-0.18}$ [± 0.15 (stat) ± 0.12(syst)]
  - Corresponding Observed (expected) significance: 3.3σ (5.2σ)
- ✓ Per-channel compatibility with SM is 0.4% (2.9σ)
- ✓ Compatibility between 0lep and 2lep against inclusive ZH measurement around 20%

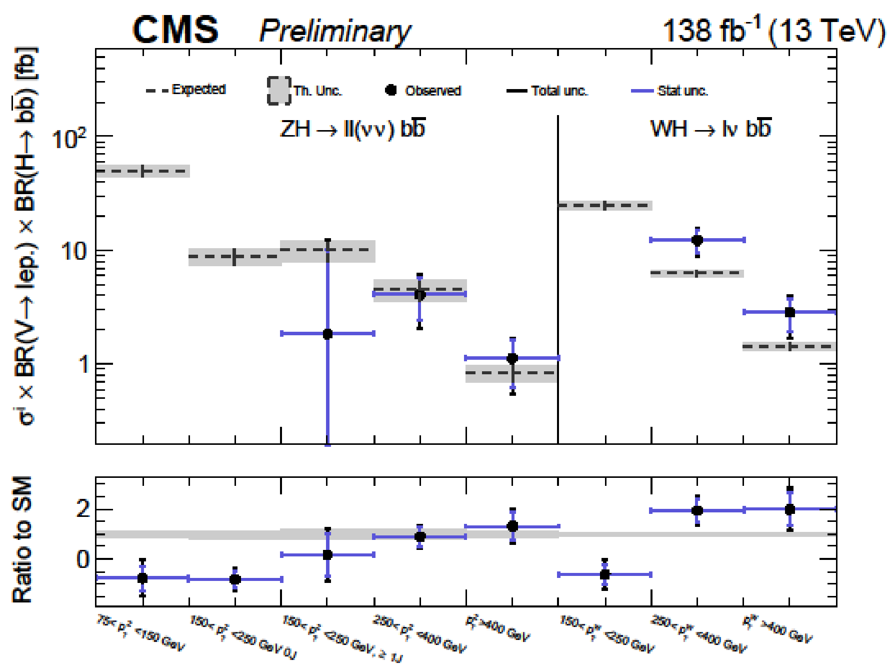




Summary

- ✓ After the discovery and subsequent observation of the Higgs boson decays into 5 different final states (γγ, ZZ, WW, ττ, bb) in Run 2 LHC has delivered a wealth of data that permits experiments to enter the era of Higgs precision measurements
- ✓ Presented an overview of CMS measurement of the VHbb cross section in the context of the STXS framework using the full Run 2 data (138 fb^{-1})
  - Currently in process of collaboration wide review [CDS]





Back Up

Physics Objects

Leptons

Electrons:

- MVA based ID
- 1-lepton: WP90 with rel. Iso. 0.15
- 2-lepton: WP80 with rel. Iso. 0.12

Muons:

- 1-lepton: tight cut based ID with rel. Iso. 0.06
- 2-lepton: loose ID with rel. Iso. 0.25

Missing transverse energy (MET):

- pf MET is used which is defined as, $\text{pf MET} = -\sum \vec{p_T}$
- Correction applied in x-y plane

Jets

Resolved:

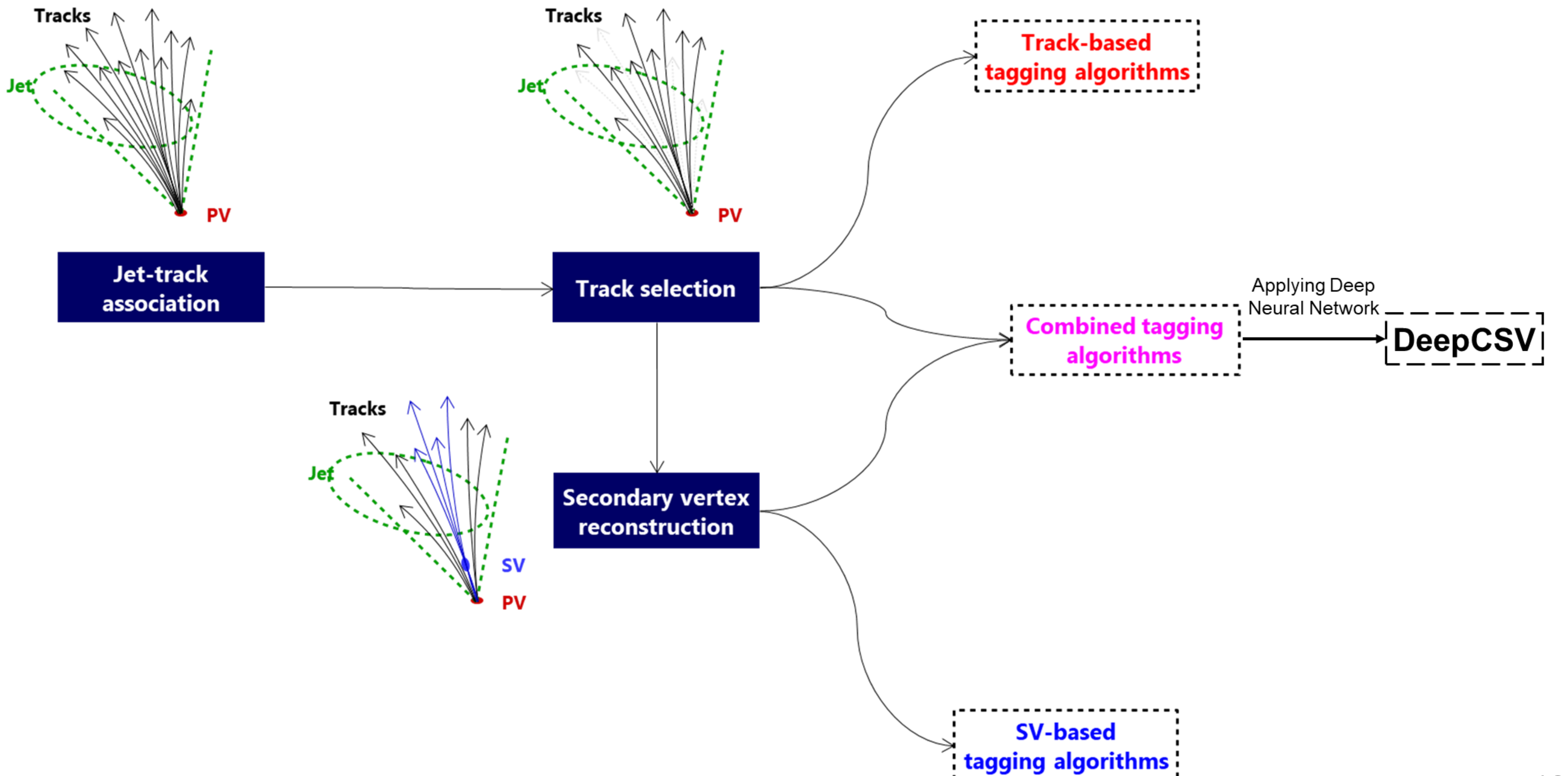
- AK4 CHS, JEC/JER applied
- b-jet energy regression
- calibrated and apply smearing to MC
- tight jet ID, tight PU jet ID
- DeepCSV b-tagger

Boosted:

- AK8 CHS, JEC/JER applied
- PUPPI soft-drop mass
- DeepAK8 b-tagger bbVsLight output
- No geometric overlap with selected leptons
- p_T >250 GeV

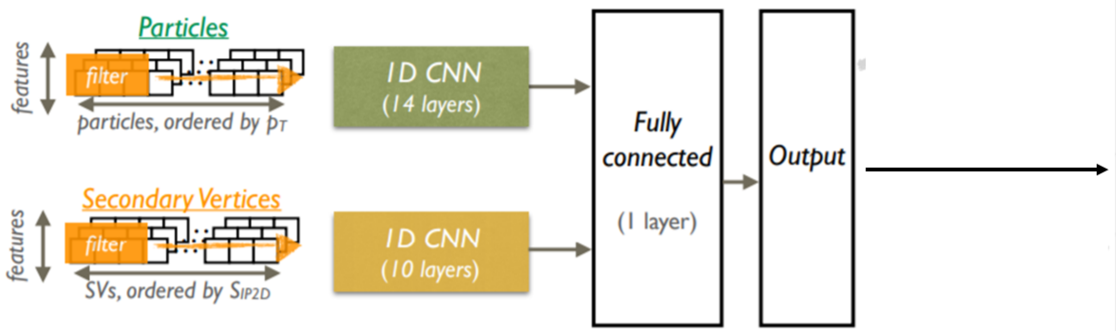
DeepCSV algorithm

- Combined Secondary Vertex version 2(CSVv2) algorithm exploits both track information and the secondary vertex information of jets to tag bjets.
- DeepCSV uses deep neural network on same algorithm for accuracy in tagging b jets (4% more efficient)



DeepAK8 algorithm

- Deep Neural Network based algorithm exploiting properties of fat jets (AK8 jets)
- Multi-class classifier for t/W/Z/H tagging
- Categories subdivided based on decay modes (e.g., Z→bb, Z→cc, Z→qq)
- Scores can be aggregated for many purposes, e.g., bb vs cc vs light tagging
- Directly uses jet constituents (secondary vertices/ particles information from jets)

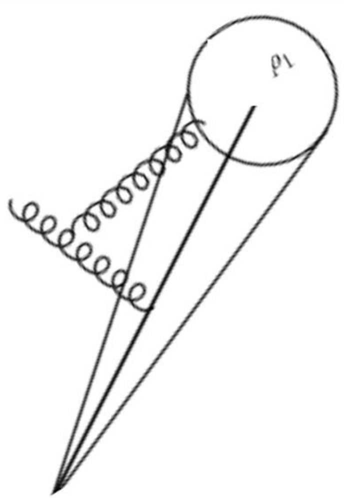


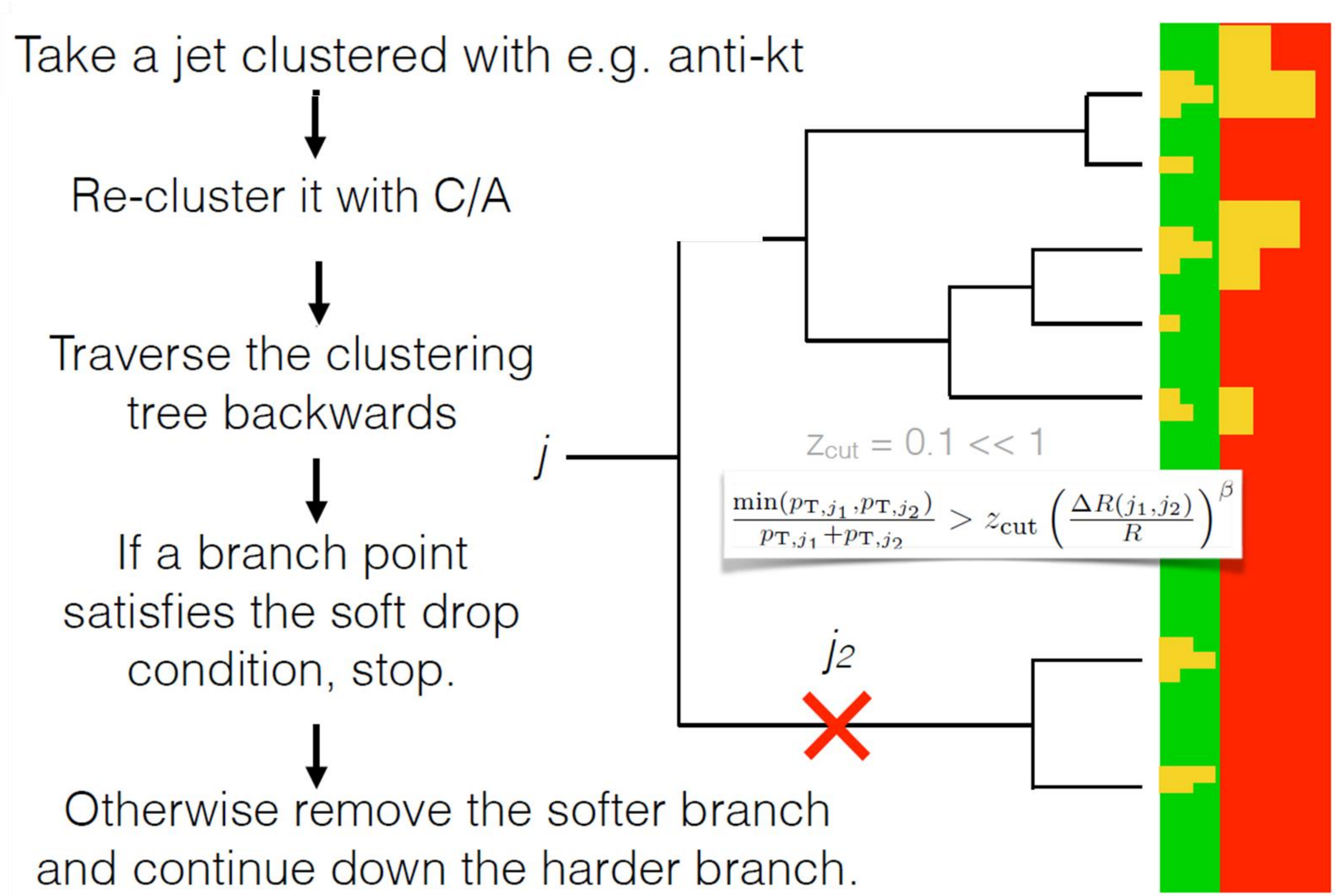
Category	Label
Higgs	H (bb)
	H (cc)
	H (VV*→qqqq)
Top	top (bcq)
	top (bqq)
	top (bc)
	top (bq)
W	W (cq)
	W (qq)
Z	Z (bb)
	Z (cc)
	Z (qq)
QCD	QCD (bb)
	QCD (cc)
	QCD (b)
	QCD (c)
	QCD (others)

Soft-Drop Mass

Soft drop is a **jet grooming** algorithm

- Removes soft and wide angle radiation from a jet





b-jet tagging

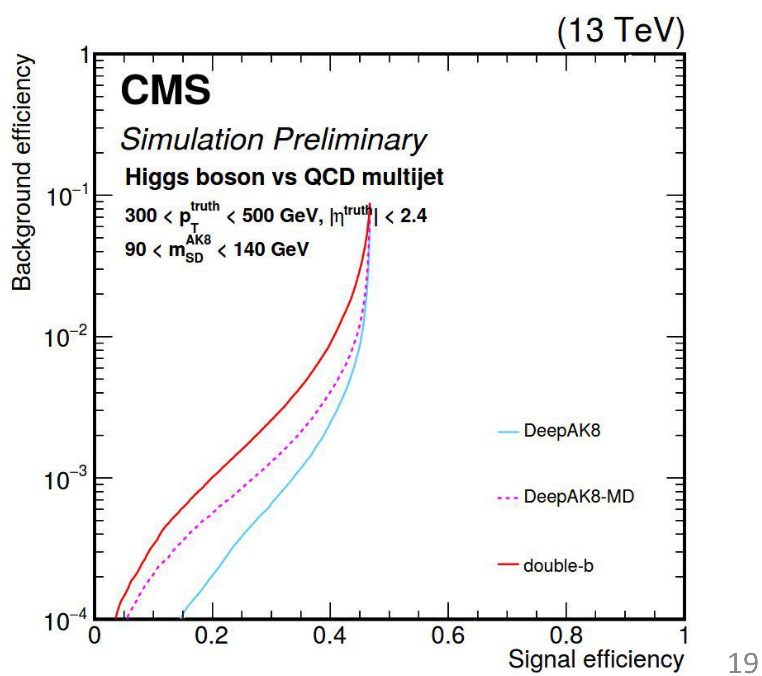
Resolved b-tagging

- DeepCSV (Deep Combined Secondary Vertex) is a DNN based b-tagging algorithm
- Continuous discriminator output values for b-jets vs jets of different flavors
- Scale-Factors derived for various working points based on mistag efficiency
- The DeepCSV score is very important variable for,
  - ✓ selecting b-jets in an event
  - ✓ Applying cuts for selecting signal events
  - ✓ Usage in multi-variate trainings for signal vs background classification

WP Name	Mistag efficiency	CSVv2	CMVA	DeepCSV
Loose (2018)	10%	n/a	n/a	0.1241
Medium(2018)	1%	n/a	n/a	0.4184
Tight (2018)	0.1%	n/a	n/a	0.7527
Loose (2017)	10%	0.5803	n/a	0.1522
Medium(2017)	1%	0.8838	n/a	0.4941
Tight (2017)	0.1%	0.9693	n/a	0.8001
Loose (2016)	10%	0.5426	-0.5884	0.2219
Medium(2016)	1%	0.8484	0.4432	0.6324
Tight (2016)	0.1%	0.9535	0.9432	0.8958

Boosted b-tagging

- DeepAK8 (Deep Anti-Kt R=0.8) algorithm is used for identifying fat jets of bigger radius of 0.8
- Multiple output classes (bb vs. light output node is
- used)
- Significant improvement with respect to double-B algorithm(usage of a DNN and low-level information)
- Output decorrelated from the soft-drop mass
- Efficiency measurement for gg->bb process used for signal



V+Jets Modeling

- LO corrected to NLO in bins of Δη(bb) used for 2016,
- Inconsistent generation of LO V+jets for 2017 and 2018
- Using NLO V+jets samples MG5_aMCatNLO w/ FXFX merging in 2017 and 2018 :
  - Mismodeling (ΔR(bb) < 1.0)
    - Reweight V+Jets to data in LF CRs with ΔR(bb)
    - Derived per lepton channel and pT(V) bin and extrapolated to HF & SR V+Jets
    - Added systematic uncertainty (low impact in fit)
    - Similar to VH(cc) analysis
  - Mismodeling (DeepCSV < loose WP):
    - 2D reweighting (leading and sub-leading b-tag) derived in LF CRs with ΔR(bb)>1.0
    - Extrapolated to ΔR(bb)<1.0 LF CRs
    - Other regions not affected due to b-tag selection
    - Applied first, before ΔR(bb) reweighting

Resolved Selection

0-lepton

Variable	SR	Z + b jets	Z + light jets	$t\bar{t}$
Common selection:				
min(pfMET,H_T^{miss})	> 100	-//-	-//-	-//-
p_T^{miss}	> 170	-//-	-//-	-//-
p_T^{j1}	> 60	-//-	-//-	-//-
p_T^{j2}	> 35	-//-	-//-	-//-
p_T(jj)	> 120	-//-	-//-	-//-
$\Delta\phi$(Z,H)	> 2.0	-//-	-//-	-//-
Different between SR and CRs:				
N_{aj}	≤ 1	≤ 1	≤ 1	≥ 2
M(jj)	$\in$[90-150]	$\notin$[90-150]	-	-
$btag_{max}$	> medium	> medium	< medium	> medium
$\Delta\phi$(pfMET,trkMET)	< 0.5	< 0.5	< 0.5	-
min $\Delta\phi$(pfMET,J)	-	-	-	$< \pi/2$

2-lepton

Variable	SR	Z + b jets	Z + light jets	$t\bar{t}$
$btag_{max}$	>medium	>medium	<loose	>tight
$btag_{min}$	>loose	>loose	<loose	>loose
$M(V)$	[75,105]	[85,97]	[75,105]	[10,75] and <120
M(jj)	[90,150]	$\notin$[90,150]	[90,150]	-
$\vec{p}_T^{miss}$	-	<60	-	-
$\Delta\phi(H,V)$	-	> 2.5	> 2.5	-

Variable	SR	W + b jets	W + light jets	$t\bar{t}$
Common selection:				
p_T(jj)	> 100	-//-	-//-	-//-
$p_T(V)$	> 150	-//-	-//-	-//-
N_{lep}	< 1	-//-	-//-	-//-
p_T^{j1}	> 25	-//-	-//-	-//-
p_T^{j2}	> 25	-//-	-//-	-//-
$\Delta\phi$(lep, pfMET)	< 2	-//-	-//-	-//-
Difference between SR and CRs:				
$btag_{max}$	>medium	>medium	[loose-medium]	>tight
$btag_{min}$	>loose	-	-	-
M(jj)	[90,150]	[150,250] and <90	<250	< 250
N_{aj}	< 2	< 2	-	>1
σ(pfMET)	-	> 2	> 2	-
$\Delta\phi(H,V)$	< 2.5	-	-	-

1-lepton

Boosted Selection

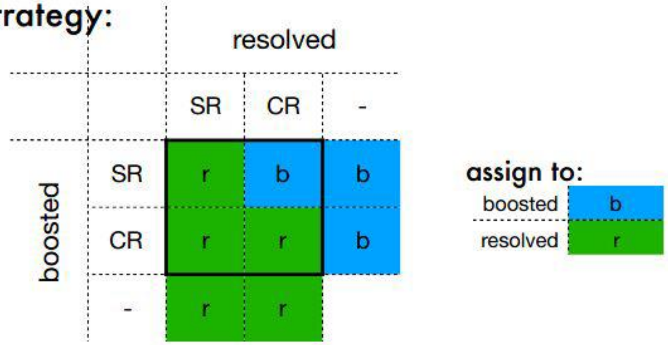
0-lepton Variable	SR	Z + b jets	Z + light jets	$t\bar{t}$
DeepAK8 (bbVsLight)	> 0.8	> 0.8	< 0.8	> 0.8
$M(\text{jj})$	$\in[90,150]$	$\notin[90,150]$	> 50	> 50
N_{al}	$= 0$	$= 0$	$= 0$	> 0
N_{aj}	$= 0$	$= 0$	$= 0$	> 1

1-lepton Variable	SR	W + b jets	W + light jets	$t\bar{t}$
DeepAK8 (bbVsLight)	> 0.8	> 0.8	< 0.8	> 0.8
$M(\text{jj})$	$\in[90,150]$	$\notin[90,150]$	> 50	> 50
N_{al}	$= 0$	$= 0$	$= 0$	> 0
N_{aj}	$= 0$	$= 0$	$= 0$	> 1

2-lepton Variable	SR	Z + b jets	Z + light jets	$t\bar{t}$
DeepAK8 (bbVsLight)	> 0.8	> 0.8	< 0.8	> 0.8
$M(\text{jj})$	$\in[90,150]$	$\notin[90,150]$	> 50	> 50
$M(V)$	$\in[75,105]$	$\in[75,105]$	$\in[75,105]$	$\notin[90,150]$

Overlap region

- some events pass selection for boosted as well as for resolved analysis (=overlap)
- have to assign to either of them for combination
- 4 different overlap strategies tested
  - compared expected sensitivity for WH>250, ZH>250 STXS signal strengths
- best strategy:



- event priorities for overlap events:
  1. Resolved SR
  2. Boosted SR
  3. Resolved CRs
  4. Boosted CRs

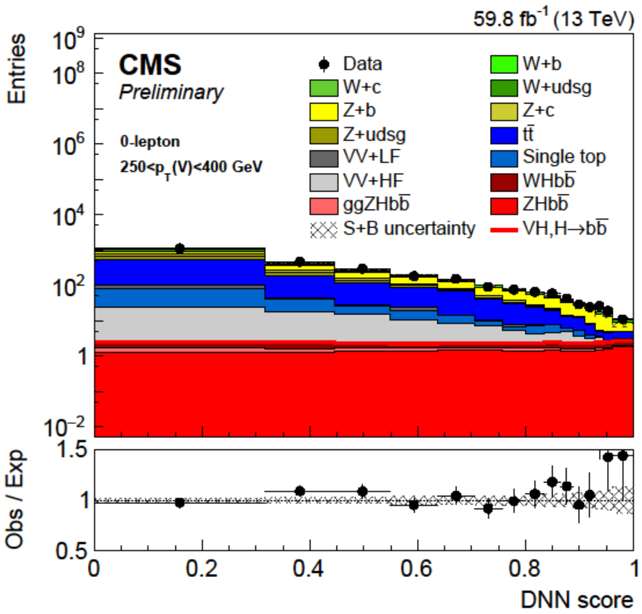
- adding the boosted analysis lowers uncertainties on WH>250, ZH>250 STXS signal strengths

Signal Extraction

- Signal extracted using binned maximum likelihood fit to the MVA score distributions

$$Expected\ events\ in\ a\ bin\ (r) = \mu . S(\theta) + B(\theta)$$

$$binned\ liklihood: L(n|\mu,\theta) = \prod_{i\ \epsilon\ bins} P_{possion}(n_i|\mu . S(\theta) + B(\theta)\ )$$



- Background processes are constrained by control region fits which are performed simultaneously with the signal region fit.

DNN variables in resolved SR

Variable	Description	0-lepton	1-lepton	2-lepton
m(jj)	Dijet invariant mass	✓	✓	✓
pT(jj)	Dijet transverse momentum	✓	✓	✓
pT(MET)	MET transverse momentum	✓	✓	✓
V(mt)	Transverse mass of vector boson		✓	
V(pt)	Transverse momentum of vector boson		✓	✓
pT(jj)/pT(V)	Ratio of momentum of vector boson and Higgs boson		✓	✓
$\Delta\phi(V,H)$	Azimuthal angle between vector boson and dijet directions	✓	✓	✓
btag_{max}	Working point b-tagging score of leading jet	✓	✓	✓
btag_{min}	Working point b-tagging score of sub-leading jet	✓	✓	✓
$\Delta\eta(jj)$	Pseudorapidity difference between leading and sub-leading jet	✓	✓	✓
$\Delta\phi(jj)$	Azimuthal angle between leading and sub-leading jet	✓	✓	
$pT_{max}(j_1, j_2)$	Maximum transverse momentum of jet between leading and sub-leading jet	✓	✓	
$\text{pT}(j_2)$	Maximum transverse momentum of jet between leading and sub-leading jet	✓	✓	
SA5	Number of soft-track jets with momentum greater than 5 GeV	✓	✓	✓
N_{aj}	Number of additional jets	✓	✓	
$\text{btag}_{\text{max}}(add)$	Maximum btagging discriminant score among additional jets	✓		
$\text{pT}_{\text{max}}(\text{add})$	Maximum transverse momentum among additional jets	✓		
$\Delta\phi(jet, MET)$	Azimuthal angle between additional jet and MET	✓		
$\Delta\phi(lep, MET)$	Azimuthal angle between lepton and MET		✓	
M_t	Reconstructed top quark mass		✓	
$pT(j_1)$	Transverse momentum of leading jet			✓
M_t	Transverse momentum of sub-leading jet			✓
$m(V)$	Reconstructed vector boson mass			✓
$\Delta R(V,H)$	Angular separation between vector boson and Higgs boson			✓
$\Delta R(V,H)$	Angular separation between leading and sub-leading jets			✓
$\sigma(m(jj))$	Resolution of dijet invariant mass			✓
N_{rec}	Number of recoil jets			✓

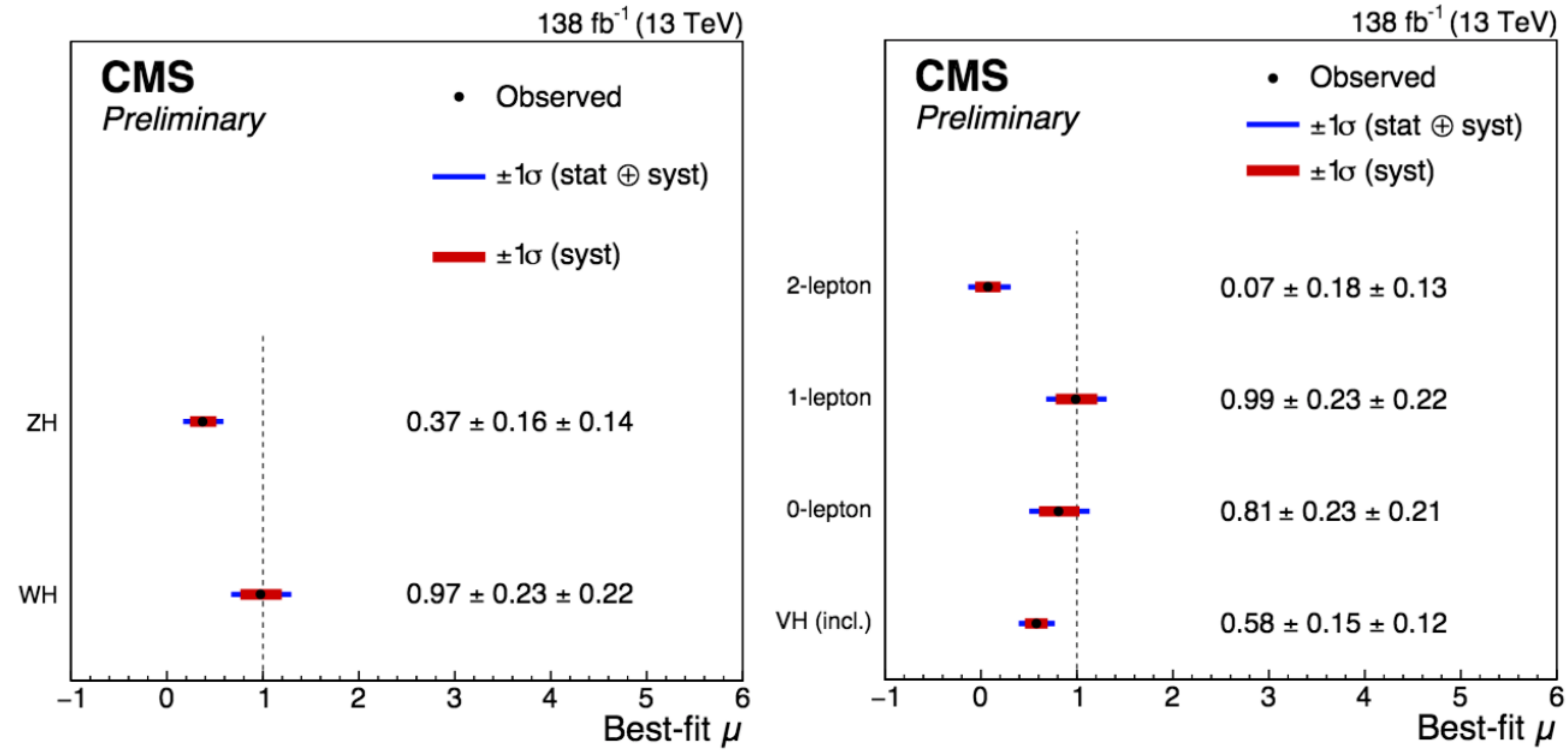
BDT variables in boosted SR

BDT training in boosted region is carried by using both boosted and resolved features of events:

- Kinematic information and deepAK8 score for exclusively boosted events
- Resolved kinematic information for events passing overlap category

Exclusive Boosted Variables	Resolved variables for Overlap events
Missing Transverse Momentum (MET)	Di jet transverse momentum
Deep AK8 binned WP	$\Delta\phi$(H, MET)
Fat jet transverse momenta	Di jet Mass
Fat jet Mass	AK4 Jet with highest momenta
Fat jet pseudo rapidity	$\Delta\phi$(leading AK4 Jet, MET)
$\Delta\phi$(V, FatJet)	# of additional AK4 Jets
	$\Delta\phi$(AK4 Jet 1, AK4 Jet 2)
	$\Delta\eta$(AK4 Jet 1, AK4 Jet 2)
	Leading AK4 Jet momenta
	Sub leading AK4 Jet momenta

Per channel scans



Jackknife Compatibility

- Jackknife resampling is a **non-parametric method of estimating uncertainty on a parameter by removing partitions from total event dataset**.
- we divide the **2017 dataset** combined from 2 analyses into g equal-sized orthogonal partitions
- for each partition *i*:
  - Remove that set of events from each analyses datacards
  - Redo both the fits to get μ_i
- jackknife estimate of the variance on $\Delta\mu$ is calculated from the variance of $\Delta\mu_i$:

$$var_I(\Delta\mu) = \frac{g-1}{g}\sum(\Delta\mu_{(i)} - \overline{\Delta\mu_d})^2 = \frac{(g-1)^2}{g} var(\Delta\mu_{(i)})$$

- Compute disagreement on $\Delta\mu$ ($\sigma_{\Delta\mu}$)

$$\sigma_{\Delta\mu} = \frac{\overline{\Delta\mu}}{\sqrt{var(\Delta\mu)}} = \frac{\overline{\Delta\mu}}{\frac{(g-1)}{\sqrt{g}} \times std.dev_{\Delta\mu}}$$

$\sigma_{\Delta\mu} = 2\sigma$
correlation $\rho \sim 0.5$

- for our purpose we are using 1000 equal sized partitions