Search for non-resonant HH in the four b quark decay channel at CMS

CMS-PAS-HIG-20-005

VIRTUAL

11th Higgs Hunting Conference
Young Scientist Forum
September 22th, 2021

Daniel Guerrero on behalf of the CMS Collaboration
University of Florida
daniel.guerrero@cern.ch

Non-resonant HH production at the LHC

Direct access to measure the Higgs boson self-coupling (λ)

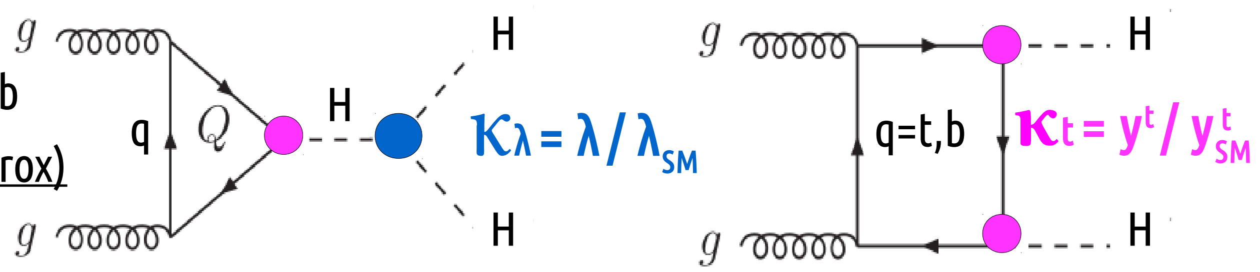
λ is connected to the Higgs potential shape → Crucial test of the SM (λ_{SM}~0.13)

Main production mechanisms: Gluon fusion (ggF) and Vector boson fusion (VBF)

ggF

σ_{ggF}~ 31.05 fb

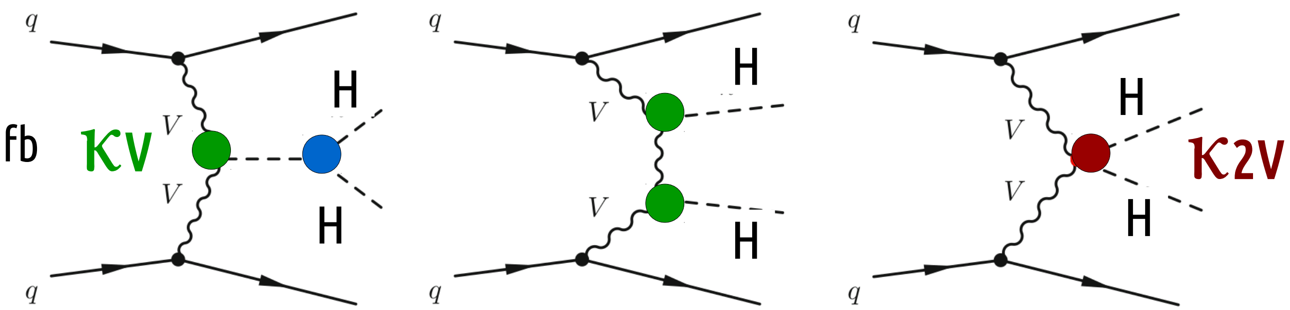
(NNLOFTapprox)

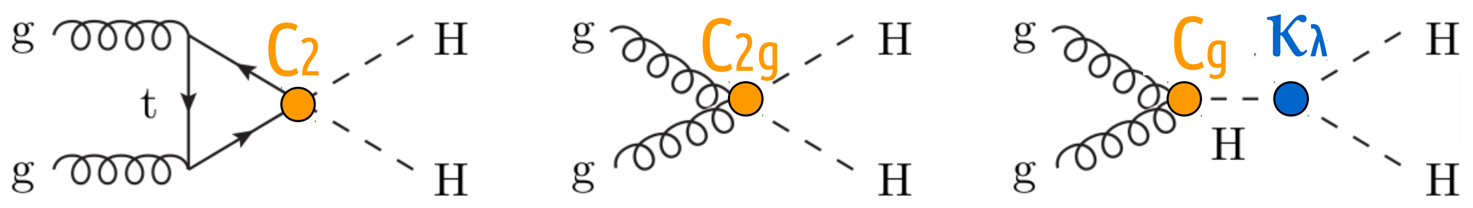


VBF

σ_{VBF} ~ 1.73 fb

(N3LO)





SM couplings

- Very small cross section (σ)
- Elusive process at current data luminosities

Anomalous Higgs couplings

- Parameterized with coupling modifiers (κ)
- Large change on σ and HH mass
- Sensitivity for new physics at the LHC

EFT approach for new physics

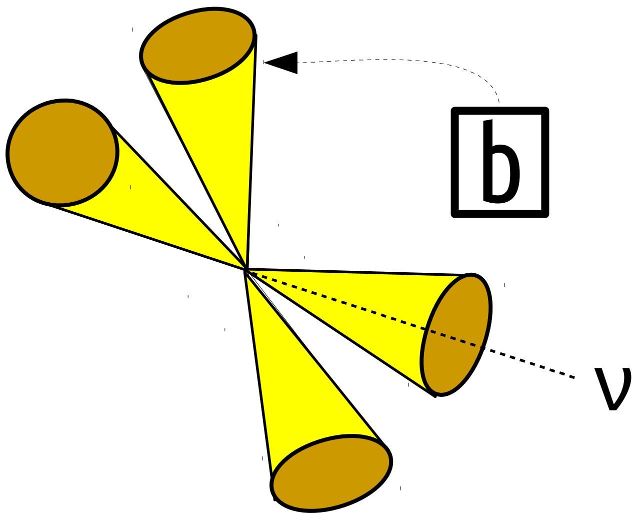
- 5 coupling modifiers : κ_λ, κ_t, C_2, C_{2g}, C_g
- 12 Benchmarks from clustering method

HH→ bbbb decay channel at CMS Run-2

It has the largest HH branching fraction (~33%) → ~1500 events produced during Run 2 (L=138 fb^{-1})
Signal reconstruction is ferociously challenged by the overwhelming production of multi-jet events

Expected signal

Four jets from b quark hadronization

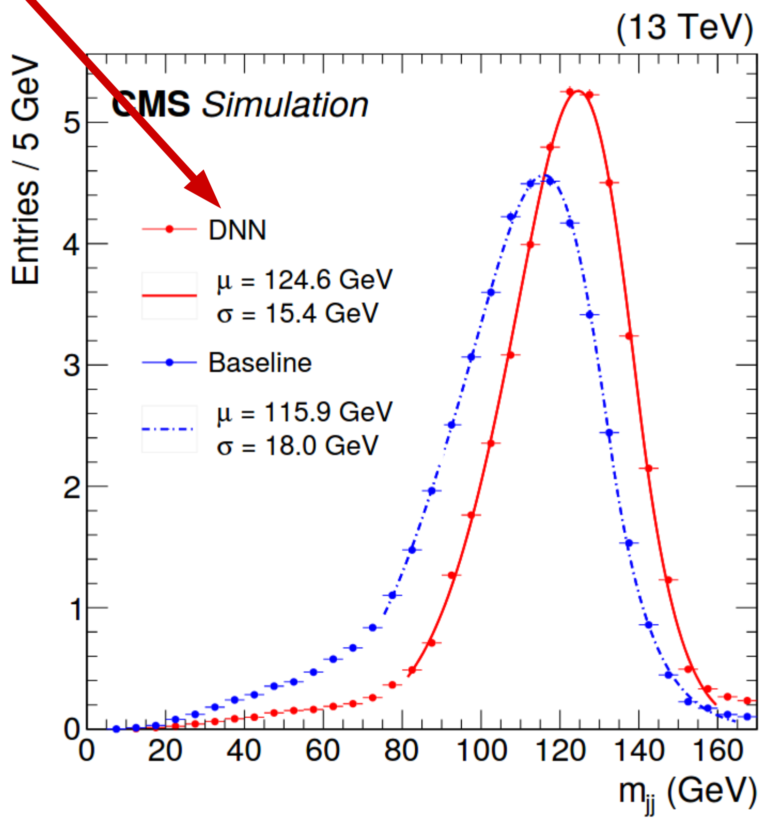


Reconstruction challenges:

- Jet identification: Large udsg/c/g jet background
- Higgs candidate reconstruction:
  - Jet combinatorics
  - Missing energy from neutrinos in B hadrons decays

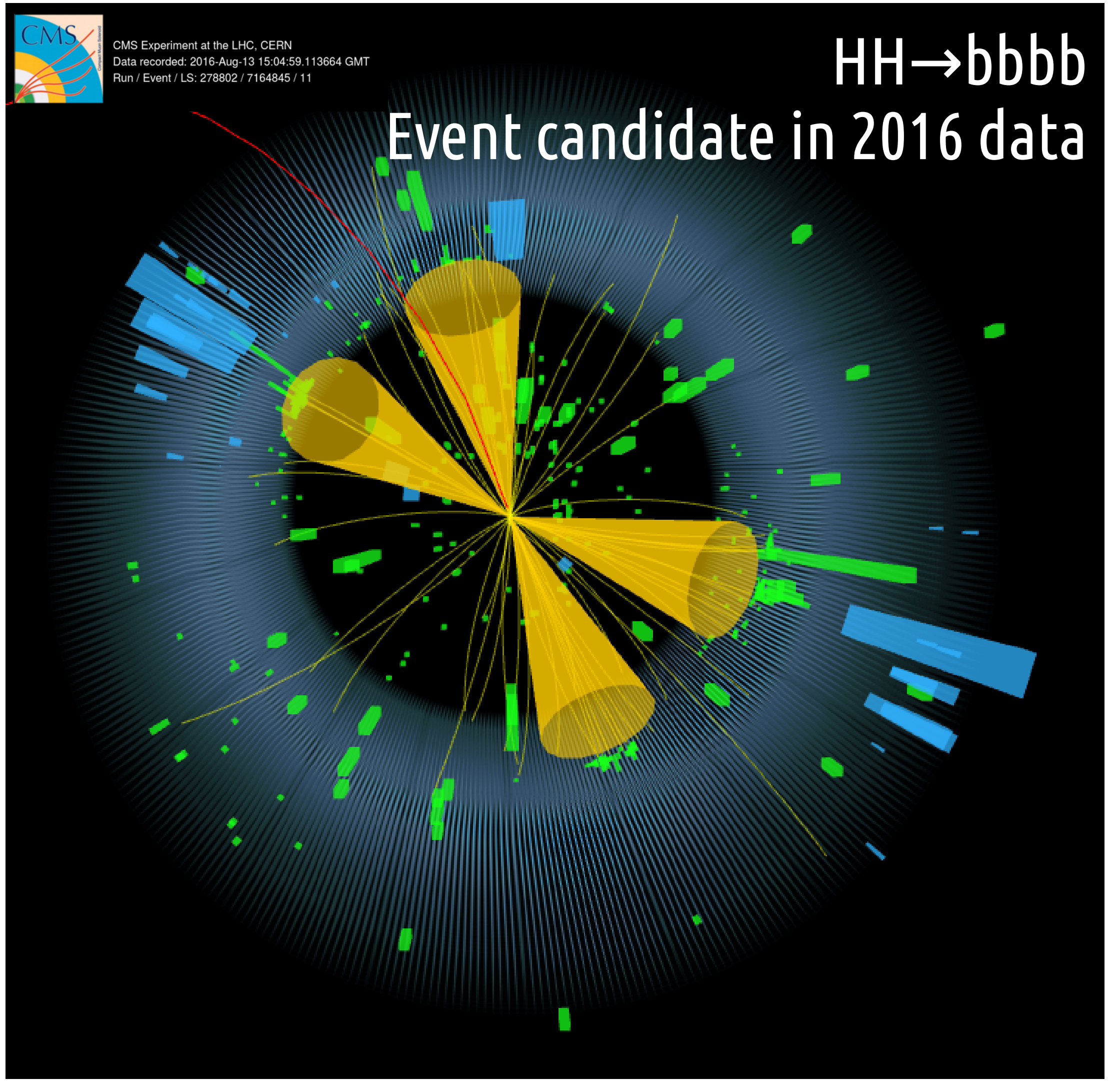
Advanced identification and reconstruction methods

- Jet flavor tagging using DeepJet (DNN)
- b-jet energy regression (DNN)



HH→ bbbb decay channel at CMS Run-2

It has the largest HH branching fraction (~33%) → ~1500 events produced during Run 2 (L=138 fb^{-1})
Signal reconstruction is ferociously challenged by the overwhelming production of multi-jet events

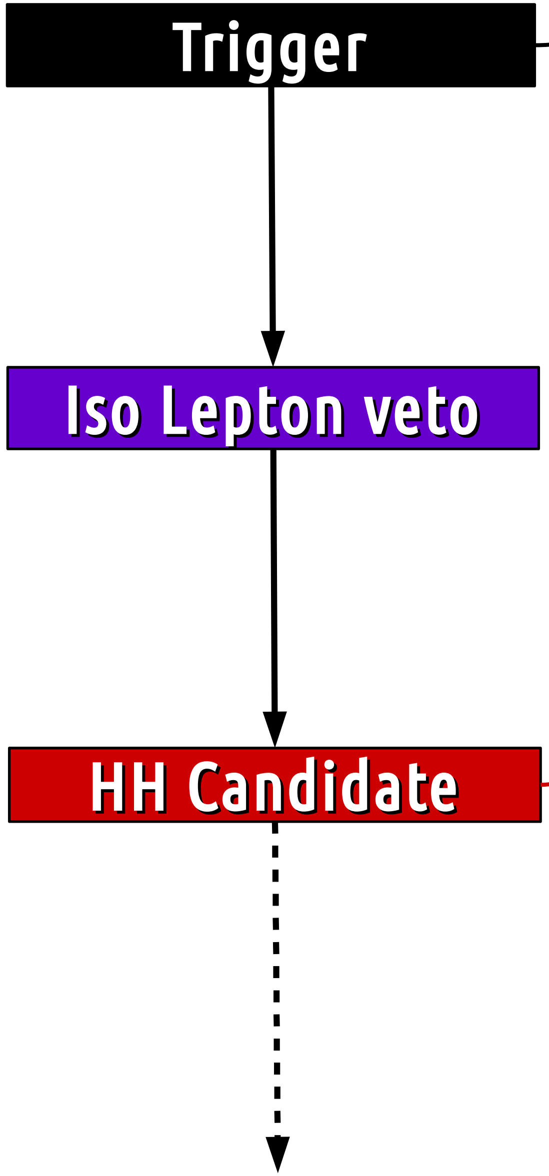


Built using innovative techniques :

- Novel jet pairing for identification of Higgs candidates
- Advanced ggF and VBF categorization
- Powerful background modeling using machine learning

VBF HH→ bbbb (highly boosted H's)
CMS-PAS-B2G-20-001, see Alessandra's talk!

Event preselection



Multijet triggers with 4 central jets, 3 b-tagged jets

Four b jet candidates:

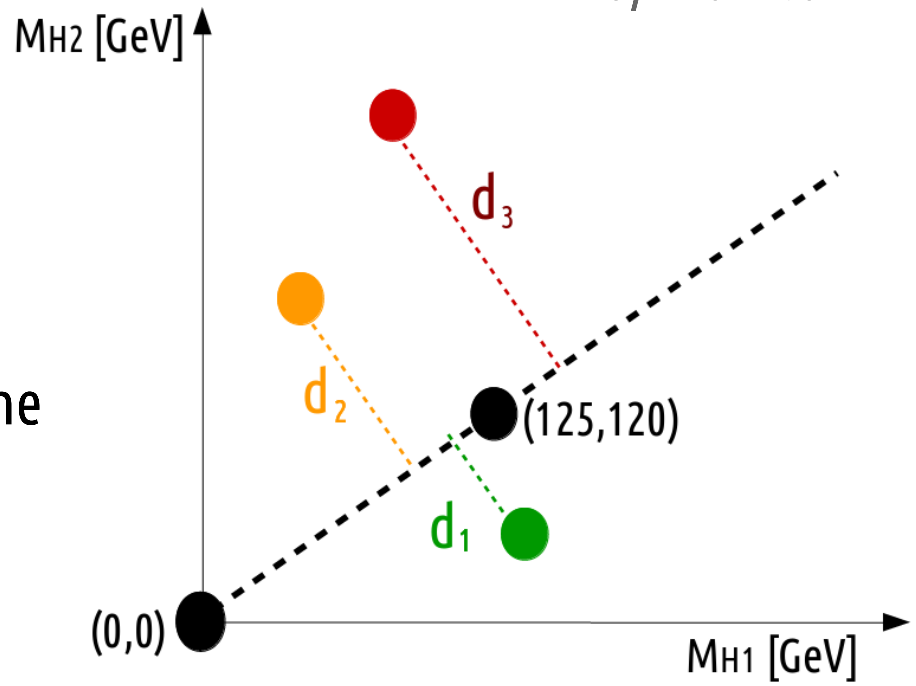
- Minimum P_T, central, quality requirements
- DeepJet cut (ε~75%, mistag-rate~1%)
- 4 most b-tagged jets

P_T (H1) > P_T (H2)
d1 < d2 < d3
k=125/120=1.04

H1, H2 identified by pairing algorithm:

- If |d1-d2| ≥ 30 GeV:
  d1 (a.k.a. closest to diagonal)
- Otherwise:
  d1 or d2 based on P_T(H) in 4-jet CM frame

$$d = \frac{|M_{H1} - k\ M_{H2}|}{\sqrt{1+k^2}}$$



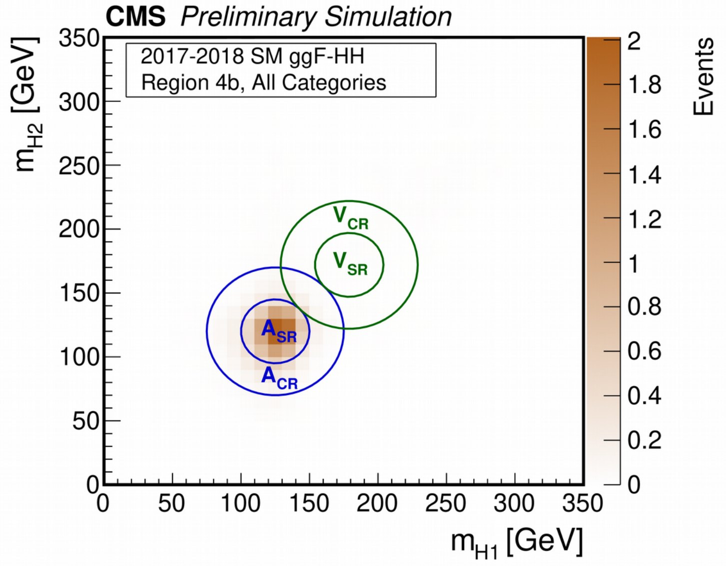
Very good performance w/o bkg. sculpting near the Higgs mass

Event regions

Several regions are defined to perform the analysis

Two b-tagging regions: '4b' (nominal selection) and '3b' (4^{th} most b-tagged fails DeepJet cut)

M_{H1}-M_{H2} plane regions:

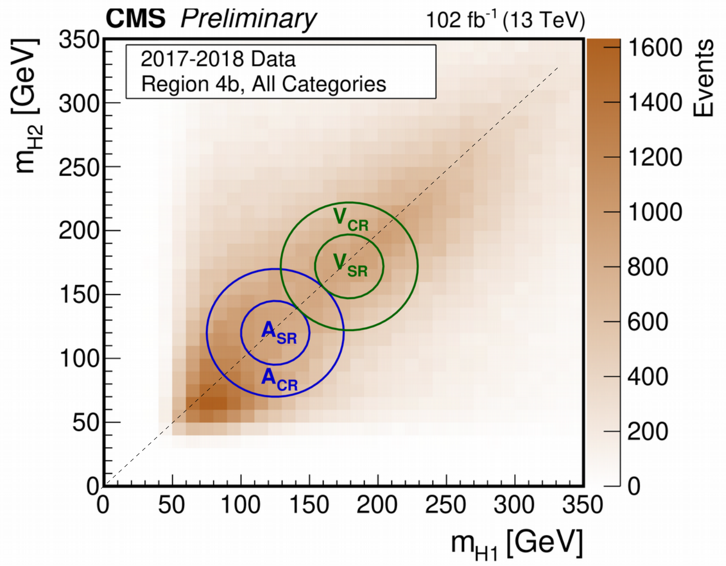


Analysis region (A)

- HH signal-enriched to perform search
- Centered at (C_1,C_2) = (125 GeV,120 GeV)
- Divided in signal (A_{SR}) and control regions (A_{CR})

Validation region (V)

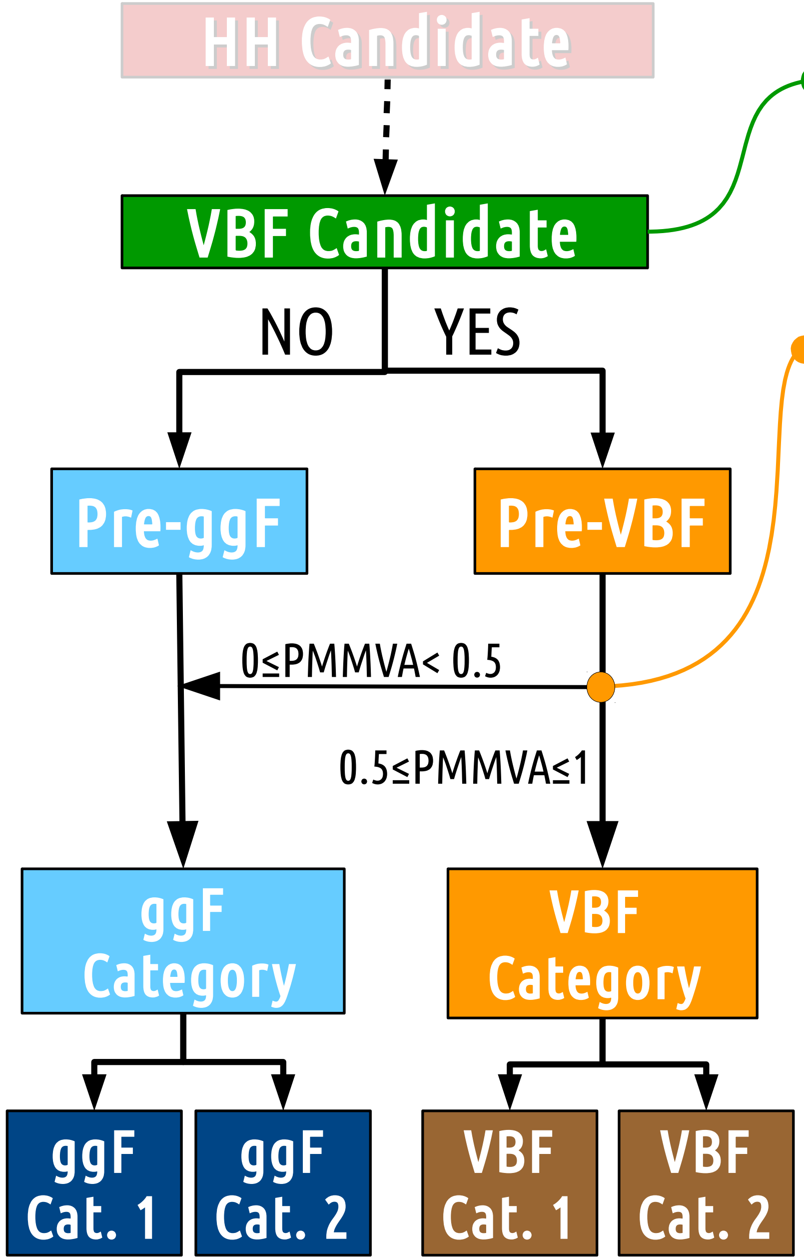
- Signal-free region used to verify full background model closure
- Aligned with pairing diagonal symmetry at (C_1,C_2) = (179 GeV, 172 GeV)
- Divided in validation signal (V_{SR}) and control regions (V_{CR})



SR: $$\sqrt{(\mathrm{m}_{\mathrm{H}_1}-\mathrm{C}_1)^2+(\mathrm{m}_{\mathrm{H}_2}-\mathrm{C}_2)^2} < 25\ \mathrm{GeV}$$

CR: $$25\ \mathrm{GeV} \leq \sqrt{(\mathrm{m}_{\mathrm{H}_1}-\mathrm{C}_1)^2+(\mathrm{m}_{\mathrm{H}_2}-\mathrm{C}_2)^2} < 50\ \mathrm{GeV}$$

Event categorization and subcategorization



VBF-jet candidates (excluding b-jets): Forward, quality requirements
VBF-jet pair selection: Two highest P_T jets with opposite-η hemispheres (η(j1) • η(j2) < 0)

A Production Mode MVA (PMMVA) is used to improve purity of the categories
- BDT trained with 13 variables to capture VBF vs ggF topologies in Pre-VBF events
- Samples: Signal (VBF κ2V=2) vs background (SM ggF HH)

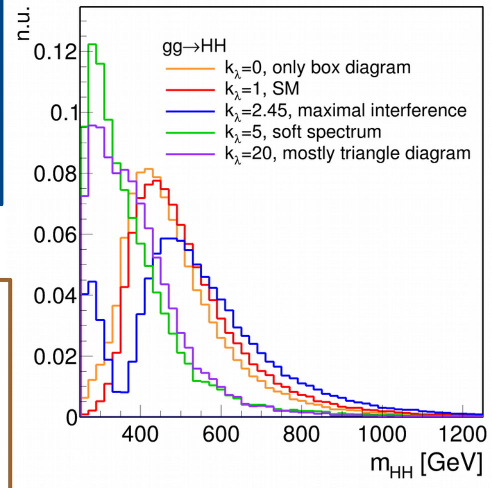
ggF subcategories

To address different m_{HH} kinematics
- Low m_{HH} (Cat.1): m_{HH} <450 GeV
- High m_{HH} (Cat.2): m_{HH} ≥450 GeV

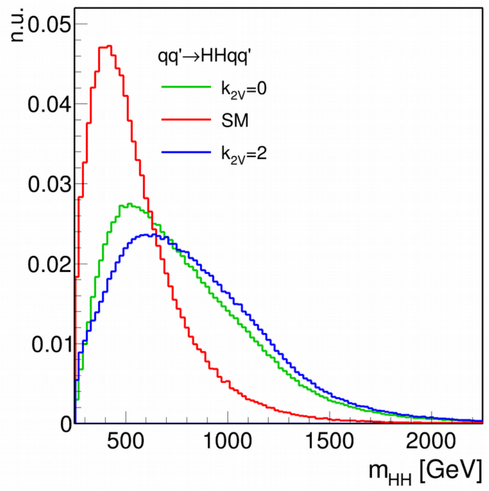
VBF subcategories

To address SM and BSM-κ_{2V} kinematics
- SM-like (Cat. 1): 0.5≤PMMVA<0.97
- BSM-κ_{2V} (Cat. 2): 0.97≤PMMVA≤1.0

ggF



VBF



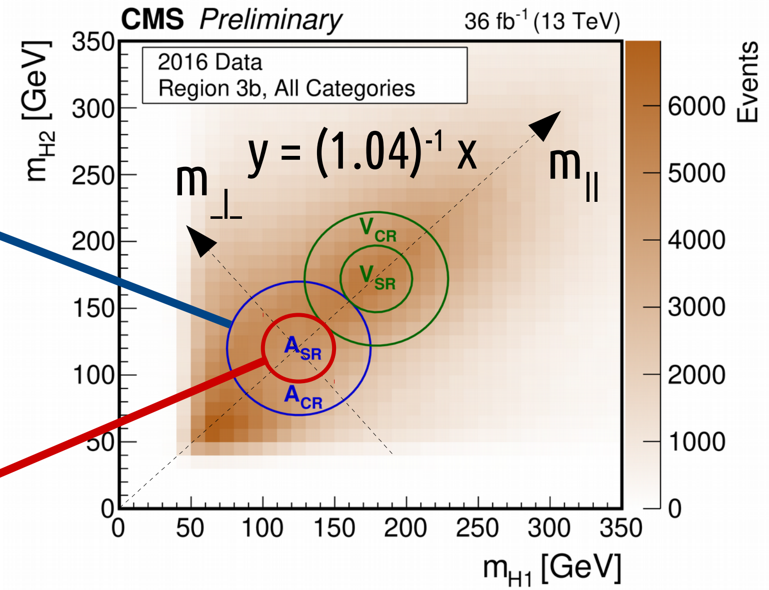
Background model overview

Data-driven multijet background model using ‘3b’ data to derive ‘4b’ background model

3b-to-4b shape differences are corrected with <u>BDT re-weighting</u>

The background model is built using ACR(3b) & ACR(4b) data

1. Normalization scaled by transfer factor α = NCR(4b) / NCR(3b)
2. Residual mismodeling on key variables are addressed via weights using a trained multidimensional BDT reweighter



Use ACR info + ASR(3b) data → ASR(4b) bkg model

Normalization: Transfer factor

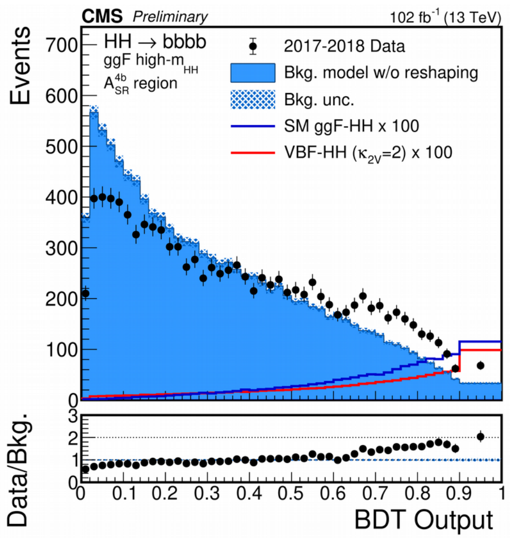
For ggF, it considers ‘parallel’ mass ($m_{||}$) dependency

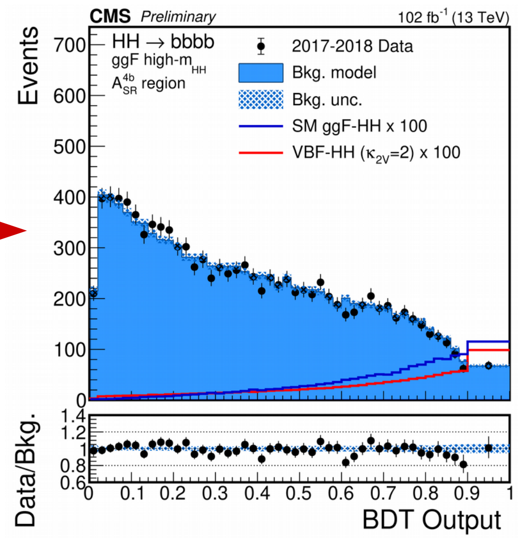
For VBF, it is constant

Shape : ASR(3b) distributions are re-shaped by reweighter

Full data/model closure is first verified in validation region

Performance in ASR(4b) region

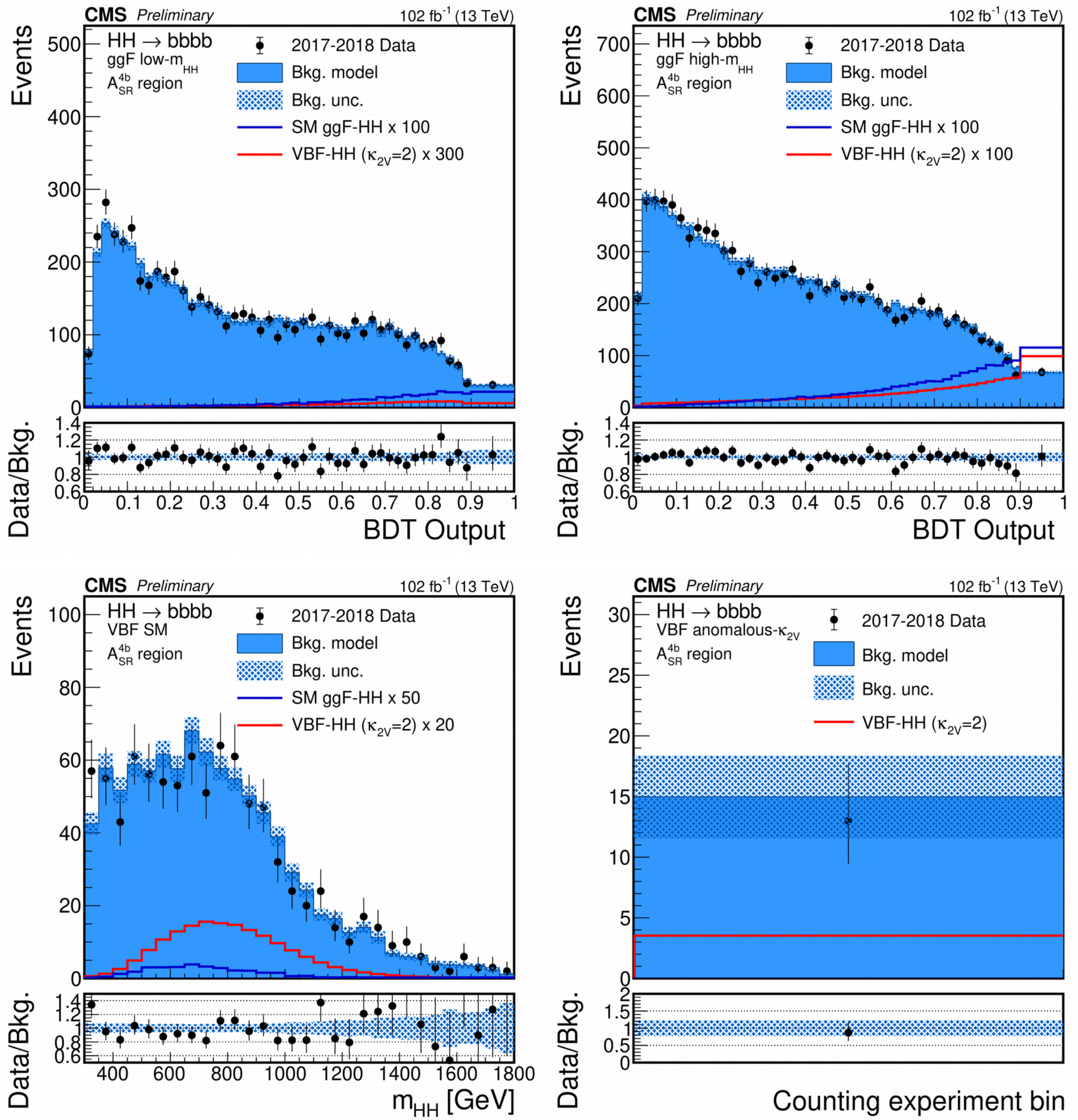




additiona

Systematics and signal extraction

2017-2018 observables



Systematic uncertainties:

Background model:

- Statistical uncertainty ASR(3b)
- 3b-to-4b transfer factor
- Validation: residual closure & limited precision
- Shape variation (e.g. ggF alternative training with CR variations)

Signal model: experimental, generator, and theoretical

Optimal observables for signal extraction:

ggF categories 1,2: Enhance HH signal with BDT distribution

- Trained by category using 16 variables (SM ggF vs bkg model)
- Bkg split in two:
  - Each half is used to train a classifier
  - Train classifier is applied to the other half

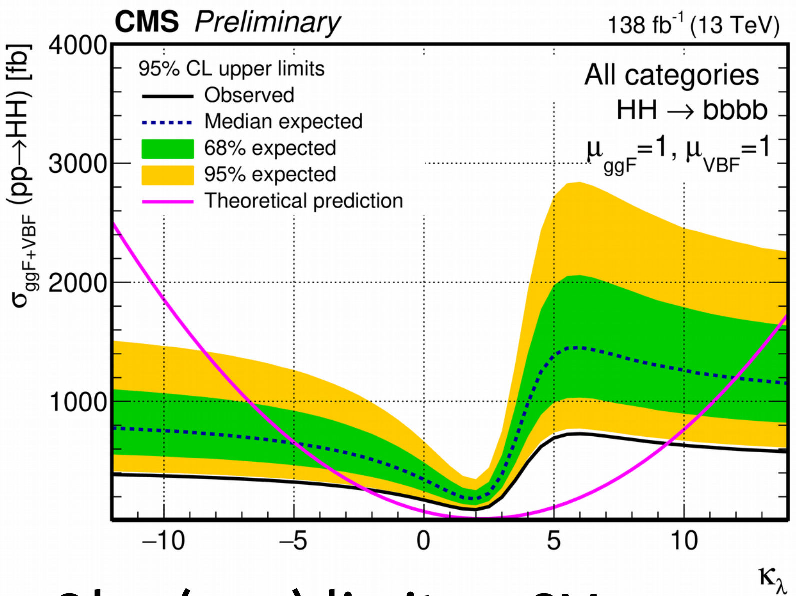
VBF category 1,2: MHH distribution, Counting experiment

Results: Upper limits

No excess of data events is observed relative to the background-only hypothesis

95% CL upper limits are set using the asymptotic CLs method

ggF + VBF



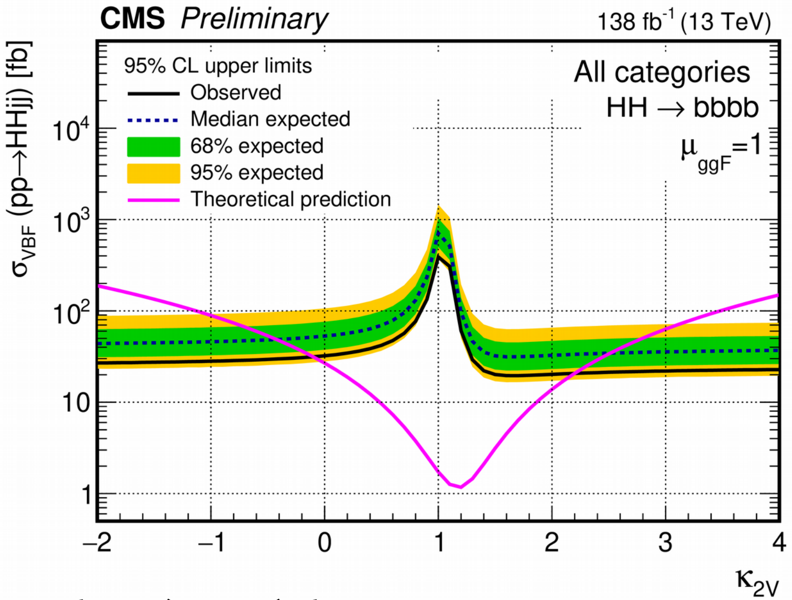
Obs. (exp.) limit on SM xs:

3.6 (7.3) x SM prediction

Obs. (exp.) allowed κ_λ interval:

$\kappa_\lambda \in$ [-2.3,9.4] ([-5.0, 12.0])

VBF-only



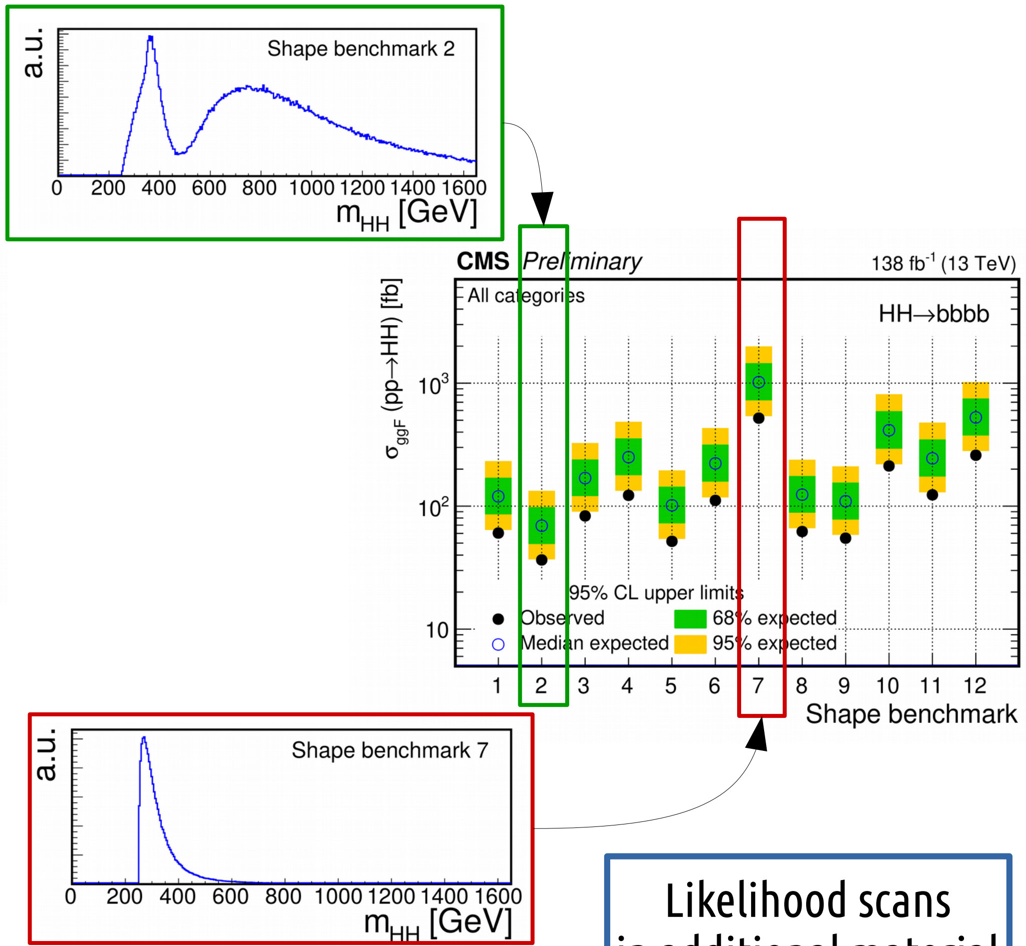
Obs. (exp.) limit on SM xs:

226 (413) x SM prediction

Obs. (exp.) allowed κ_{2V} interval:

$\kappa_{2V} \in$ [-0.1,2.2] ([-0.4,2.5])

EFT Benchmarks



Likelihood scans in additional material

Conclusions

CMS Experiment at the LHC, CERN
Data recorded: 2016-Aug-13 15:04:59.113664 GMT
Run / Event / LS: 278802 / 7164845 / 11

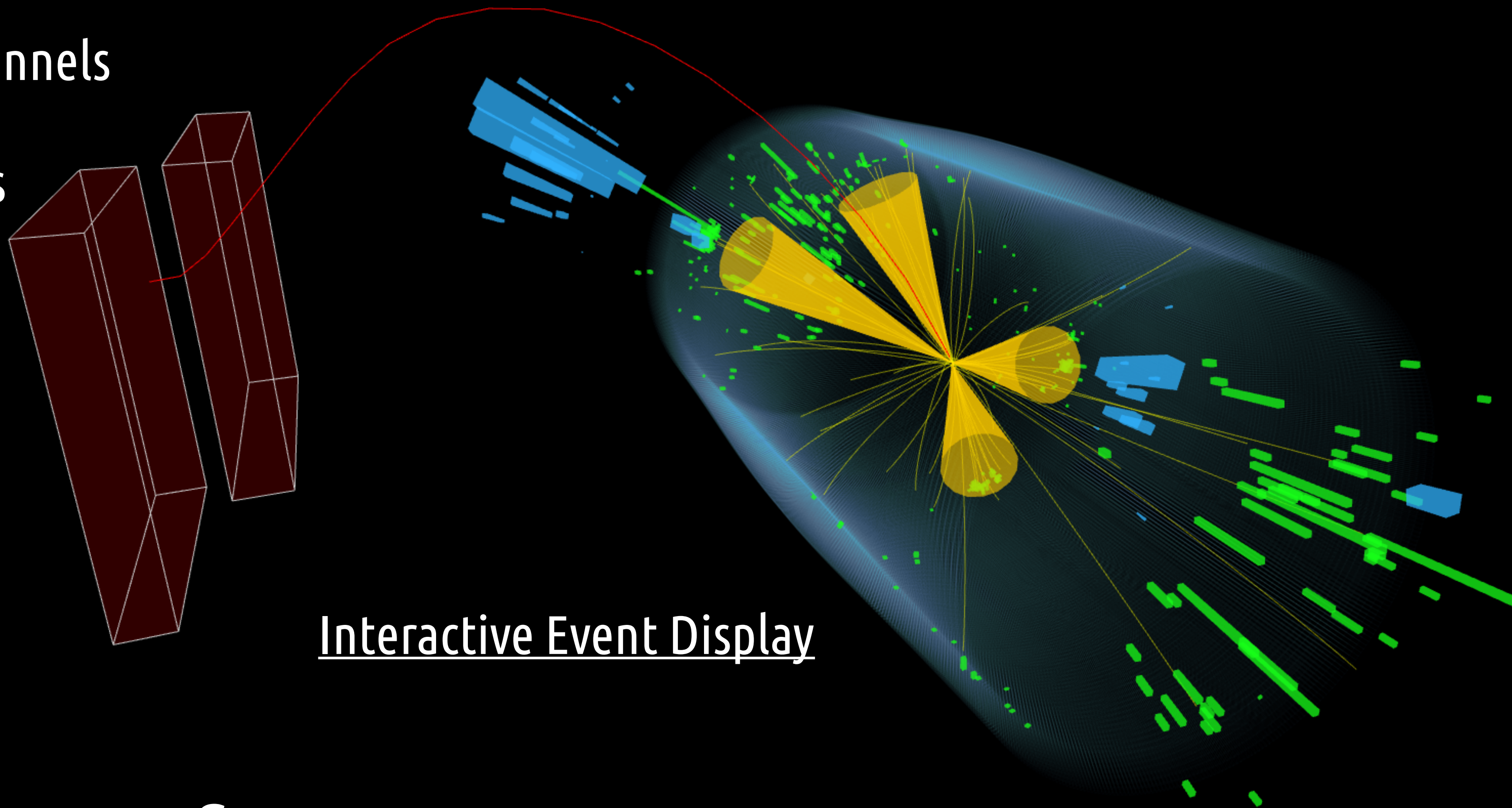
HH process can shed light on the structure of the Higgs potential

HH→bbbb is one of the most sensitive channels

- Leverage on innovative analysis methods
- 5 x better sensitivity than 2016 result

Best LHC constraints on SM production

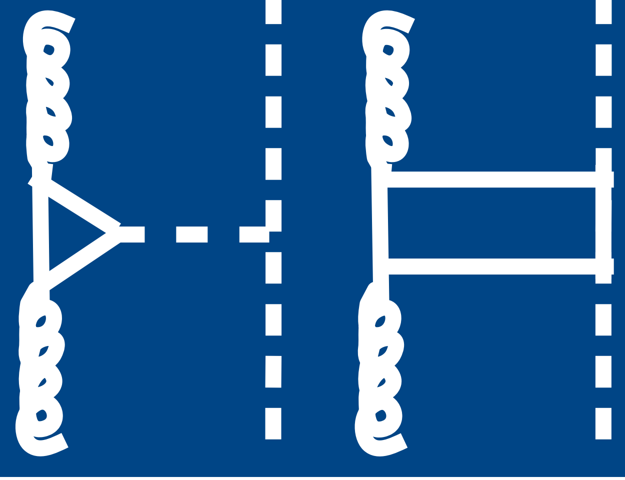
- Limit on HH xs : 3.6 x SM prediction
- Limit on VBF xs : 226 x SM prediction



Interactive Event Display

Tight constraints on anomalous couplings

Thank you for your attention!

Additional  Material

Data and MC samples

13 TeV pp collision data:

Dataset	Integrated Luminosity [fb^{-1}]
2016	36
2017-2018	102

Signal MC simulation:

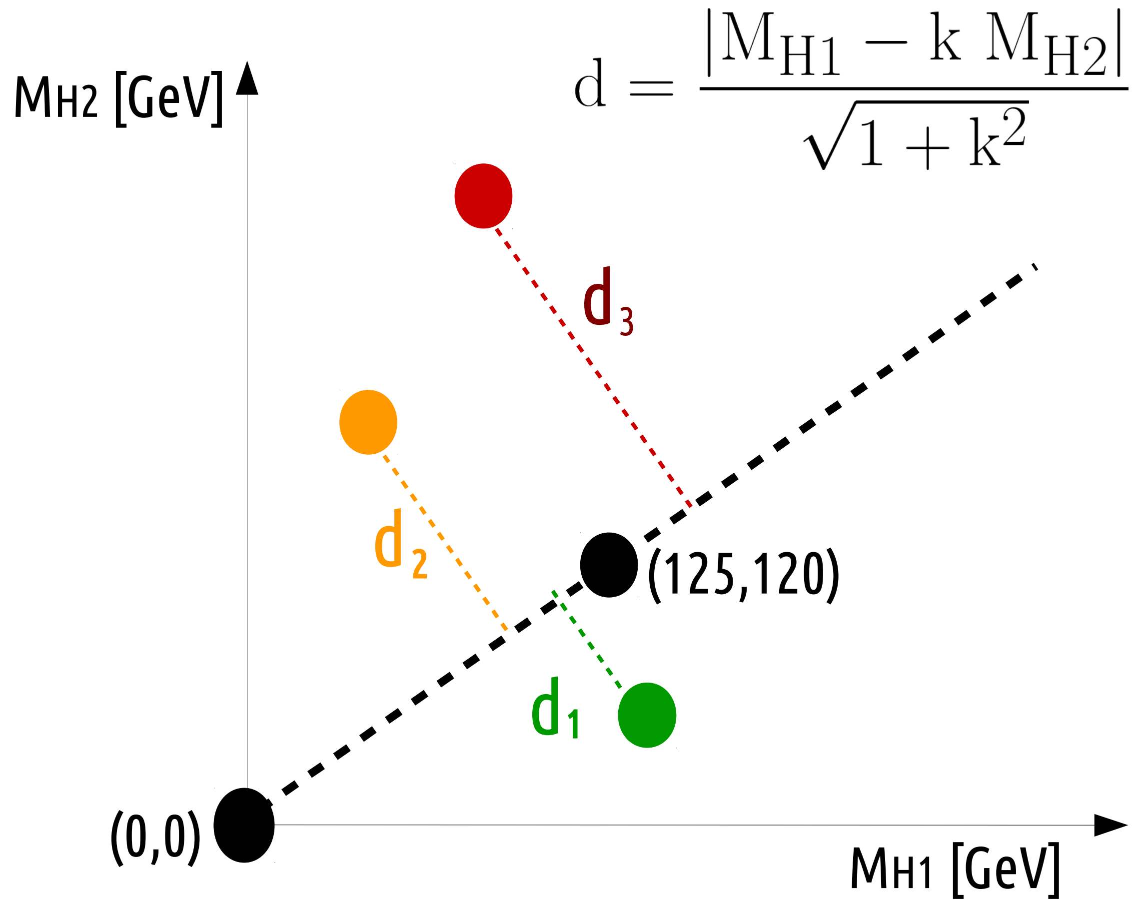
- 4 ggF samples (κ_λ=0,1,2.45,5) at NLO precision using PowHeg (3 for ggF modeling + 1 for cross-check)
- 7 VBF samples with κ_V, κ_{2V}, κ_λ combinations at LO using MadGraph (6 for VBF modeling + 1 cross-check)
  - 2 samples with alternative dipoleRecoil ON option (pythia dipole shower) for systematic uncertainties
- 12 EFT benchmarks = EFT LO samples re-weighted to NLO

Background MC simulation:

- QCD (HT-binned), ttbar, single Higgs are used for cross check studies

Jet Pairing for Higgs candidate reconstruction

Challenge: 4 preselected b-jets→ 3 possible pairings



Jet pairing method:

Step 1. Compute distance to the diagonal line (d) in plane

- Object ordering: pT(H1) > pT(H2)
- Diagonal defined with k = 125/120 = 1.04
- Pairs ordered by distance: $d_1 < d_2 < d_3$

Step 2. Select the pairing

If $\Delta d = |d_1 - d_2| \geq 30$ GeV:

- Choose d_1 pairing (closest to diagonal)

Otherwise:

- Choose d_1 or d_2 based on the highest pT(H) in 4-jet C.M. frame

Performance:

- Maximizes pairing performance w/o biasing bkg events near the Higgs mass (See slide 6)
- Correct pairing in 96% of SM ggF events. Ranging 82-96% (91-98%) depending on ggF (VBF) hypotheses

PMMVA

A classifier is trained to have categories with higher purity

Targets Pre-VBF events

Signal (S): VBF-HH (k2v=2)
- Signature with strongest contribution from longitudinal scattering amplitude V(L)V(L)→HH
- VBF-HH (k2v=0) has similar response

Background (B): NLO SM ggF-HH

Variable	Meaning
$p_T(H_1)$ ($p_T(H_2)$)	Tranverse momentum of the H_1 (H_2) candidate
$p_T(j_1)$ ($p_T(j_2)$)	Tranverse momentum of the j_1 (j_2) candidate
$\|\eta(jj)\|$	VBF-jet pair pseudorapity
M(jj)	VBF-jet pair invariant mass
$\Delta R(H_1,H_2)$	ΔR distance between two Higgs bosons
$\Delta R(H_1,j_1)$	ΔR distance between H_1 and j_1
$\Delta R(H_1,j_2)$	ΔR distance between H_1 and j_2
$\Delta R(H_2,j_1)$	ΔR distance between H_2 and j_1
$\Delta R(H_2,j_2)$	ΔR distance between H_2 and j_2
$\|\cos(\theta)^*(j1)\|$	$\|\cos(\theta)\|$ of j_1 in the six-jet center of mass frame
$\|\cos(\theta)^*(j2)\|$	$\|\cos(\theta)\|$ of j_2 in the six-jet center of mass frame
H1-centrality · H2-centrality	Product of the Higgs boson centralities

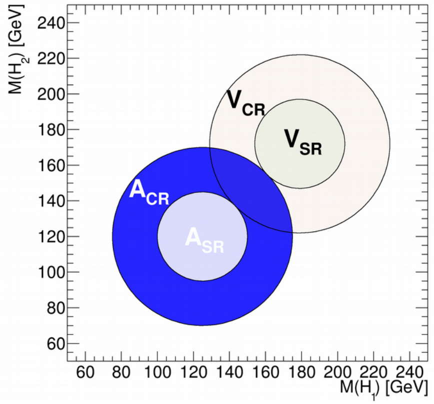
where:

H1-centrality · H2-centrality: $\exp[-(\frac{\eta(H_1)-\eta_{avg}}{\Delta\eta})^2 - (\frac{\eta(H_2)-\eta_{avg}}{\Delta\eta})^2]$.

$$\Delta\eta = \eta(j_1) - \eta(j_2) \qquad \eta_{avg} = \frac{\eta(j_1)+\eta(j_2)}{2}$$

Background model optimization and tests

Training in analysis control region:



- '3b' vs '4b' training variables used in BDT-reweighter
- ggF: b-jet PTs, input variables for BDT output
- VBF: b-jet PTs, M(HH) & correlated variables

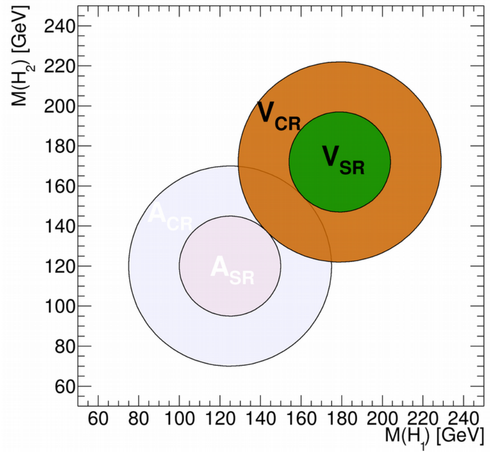
BDT-hyperparameters optimization

- K-S test improvement in individual variables
- A classifier is trained to separate 'target' from 'model'
  - if no separation is possible (AUC=0.5)
  - Then, the model is good

All variables are well-modeled in A_{CR} (4b)

Closure tests on the validation region:

Step 1: A background model is trained using validation control region data (V_{CR})



Step 2: Data/model distributions are compared in the validation region V_{SR}(4b) to verify the method closure

All variables are well-modeled in V_{SR}(4b)

Self-bias test to check signal contamination:

- Bias is negligible at our level of sensitivity

List of BDT-reweighting variables

GGF categories 1,2

BDT Reweighter Input variables
Regressed p_T of the leading-p_T b jet of the H_1 candidate
Regressed p_T of the trailing-p_T b jet of the H_1 candidate
Regressed p_T of the leading-p_T b jet of the H_2 candidate
Regressed p_T of the trailing-p_T b jet of the H_2 candidate
Mass of the H_1 candidate, $M(H_1)$
Mass of the H_2 candidate, $M(H_2)$
Mass of the Higgs pair system, m_{HH}
Transverse momentum of the H_1 candidate, $P_T(H_1)$
Transverse momentum of the H_2 candidate, $P_T(H_2)$
Pseudorapidity separation between the two Higgs candidates, $\Delta\eta(H_1, H_2)$
ΔR distance between two b jets of the H_1 candidate, $\Delta R(H_1(bb))$
ΔR distance between two b jets of the H_2 candidate, $\Delta R(H_2(bb))$
$\lvert\cos(\theta)^*\ (H)\rvert$ in HH frame
$\lvert\cos(\theta)^*\ (b)\rvert$ in H_1 frame
Sum of four b jets' regressed p_T
Transverse momentum of the HH system, $p_T(HH)$
Number of tight b-tags in 3 hightest b-tags
Sum of 3b's resolution scores
Minimal ΔR distance between two b jets, $\mathrm{Min}\lvert\Delta R(bb)\rvert$
Maximum pseudorapidity separation between two b jets, $\mathrm{Max}\lvert\Delta\eta(bb)\rvert$

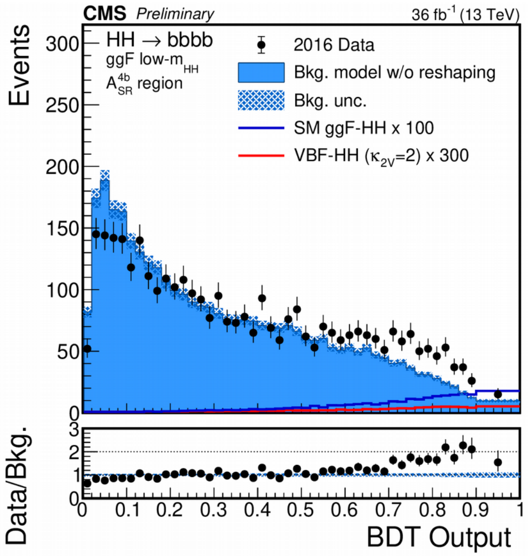
Input features
BDT output

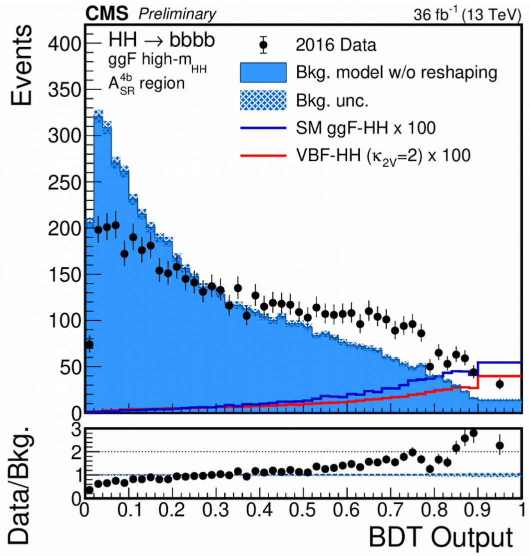
VBF category 1

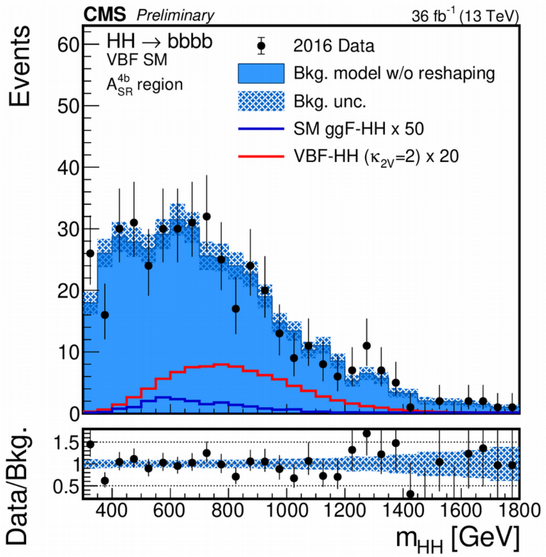
BDT Reweighter Input variables
Regressed p_T of the leading-p_T b jet of the H_1 candidate
Regressed p_T of the trailing-p_T b jet of the H_1 candidate
Regressed p_T of the leading-p_T b jet of the H_2 candidate
Regressed p_T of the trailing-p_T b jet of the H_2 candidate
Mass of the H_1 candidate, $M(H_1)$
Mass of the H_2 candidate, $M(H_2)$
Mass of the Higgs pair system, m_{HH}
Transverse momentum of the H_1 candidate, $P_T(H_1)$
Transverse momentum of the H_2 candidate, $P_T(H_2)$
Pseudorapidity separation between the two Higgs candidates, $\Delta\eta(H_1, H_2)$
Azimuthal angle separation between the two Higgs candidates, $\Delta\phi(H_1, H_2)$
Mass of the VBF-jet pair system, M(jj)
Pseudorapidity separation between the two VBF jets, $\Delta\eta(j1, j2)$
PMMVA score

Signal observables – Pre-fit w/o reshaping

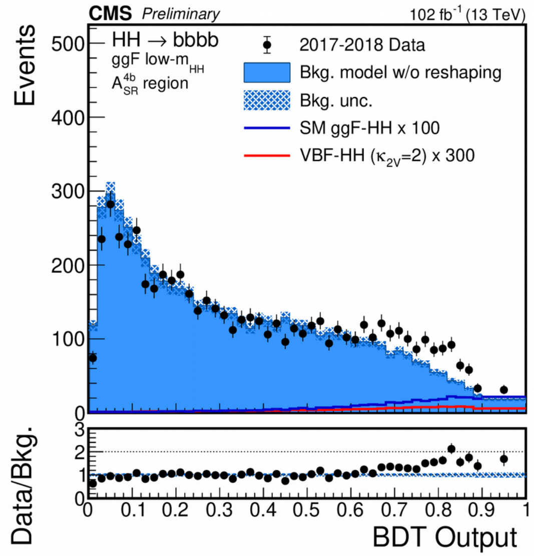
2016 dataset

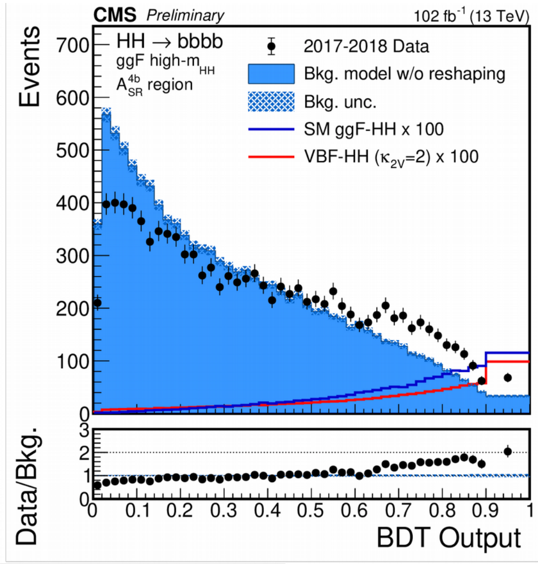


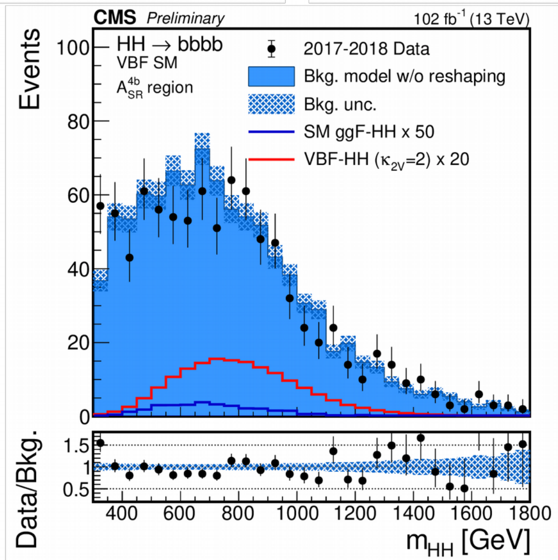




2017-2018 dataset

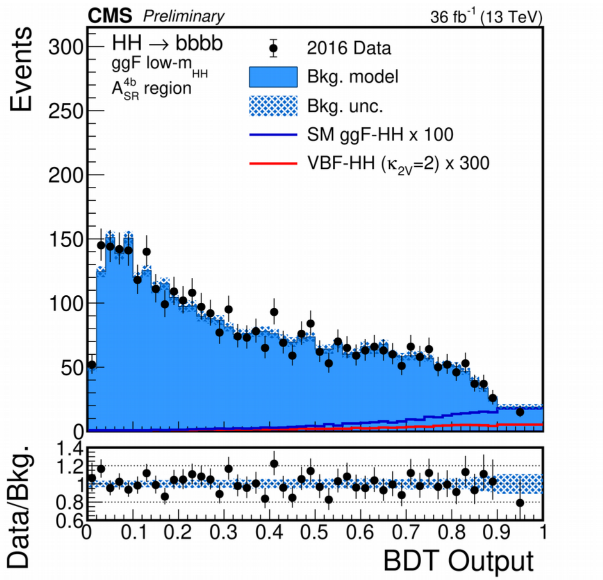


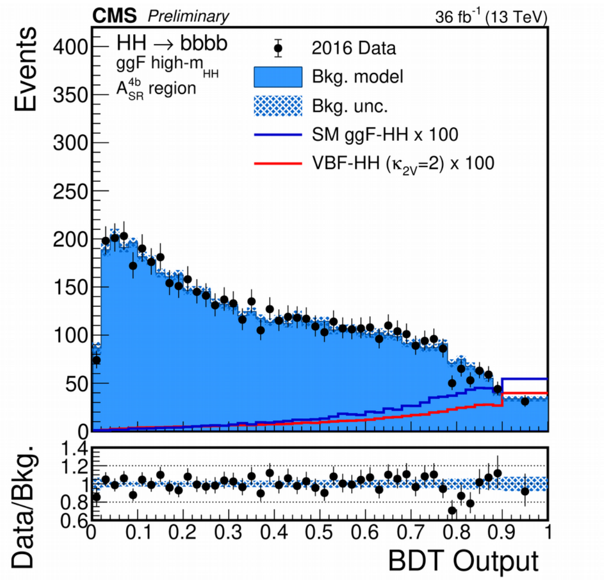


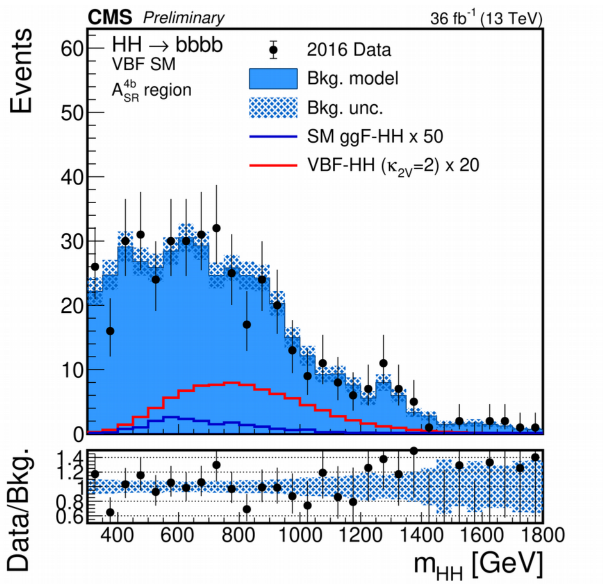


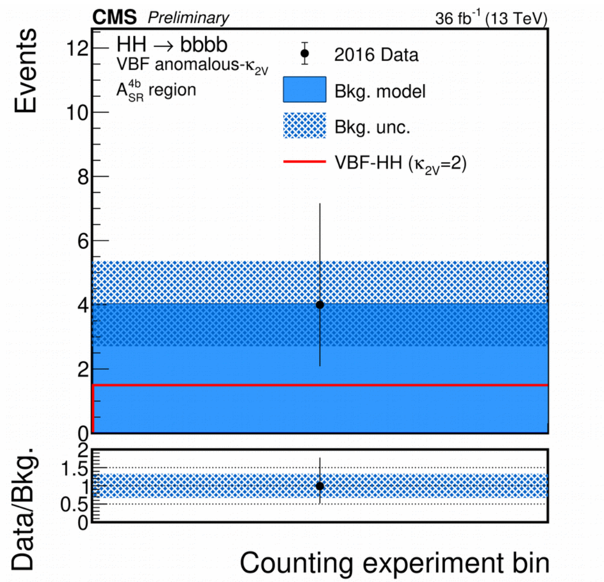
All signal extraction observables - post-fit

2016 dataset

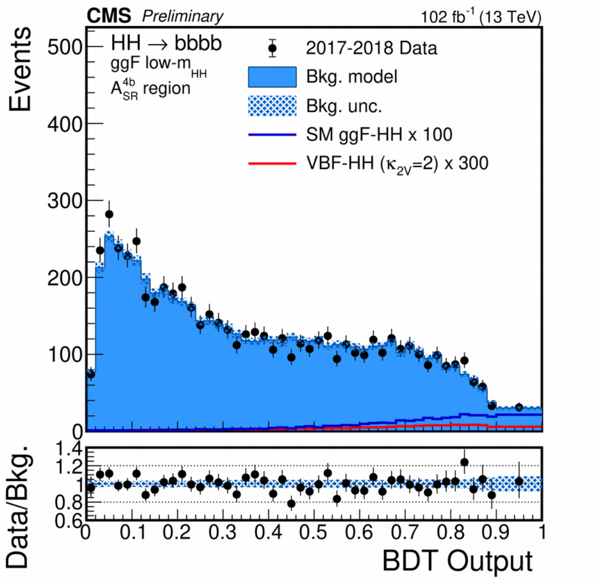


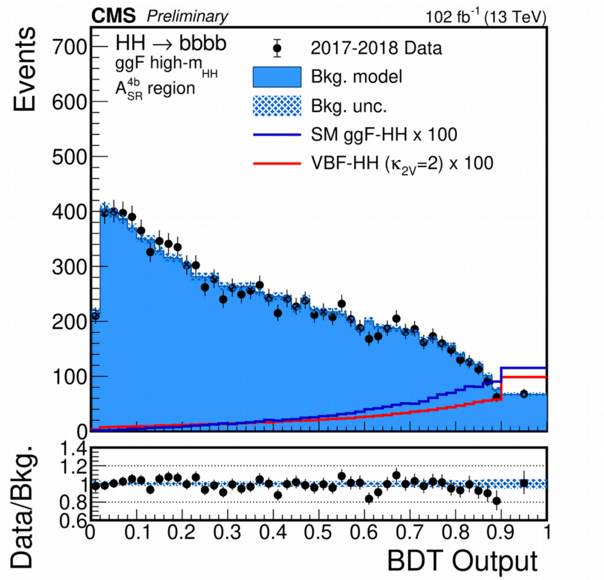


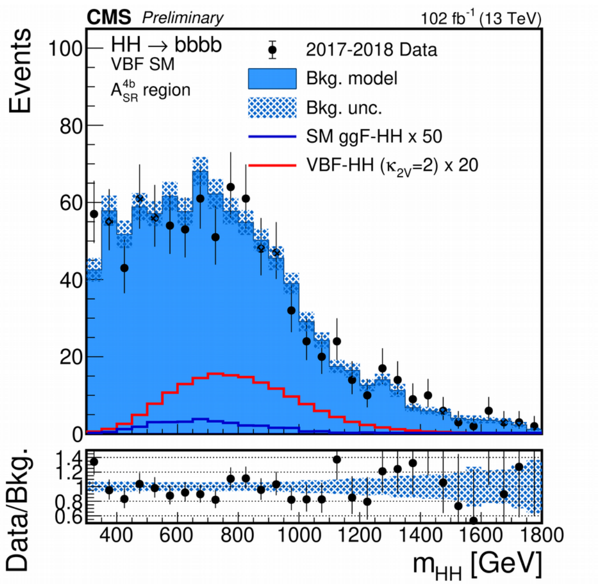


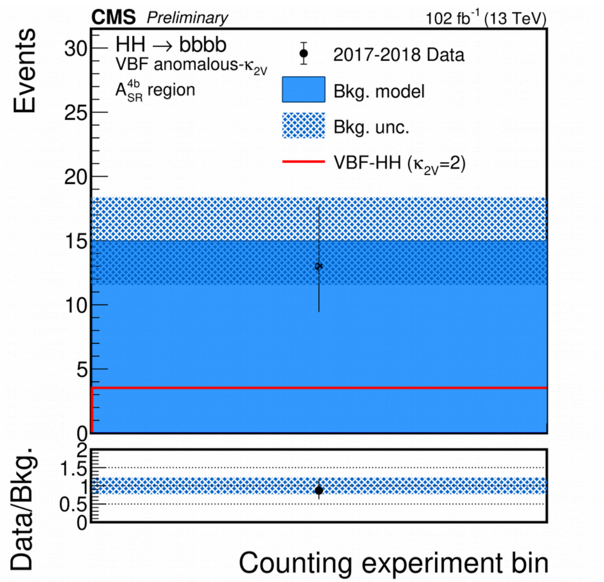


2017-2018 dataset





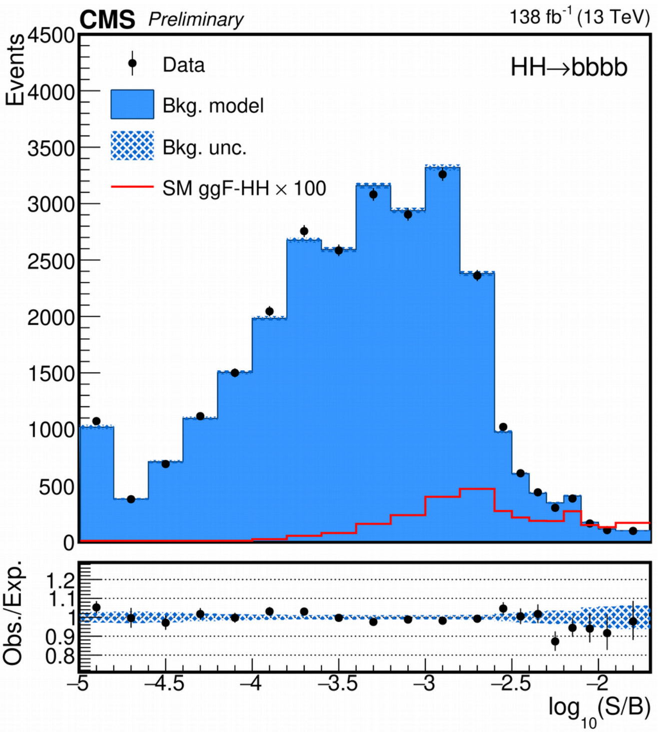


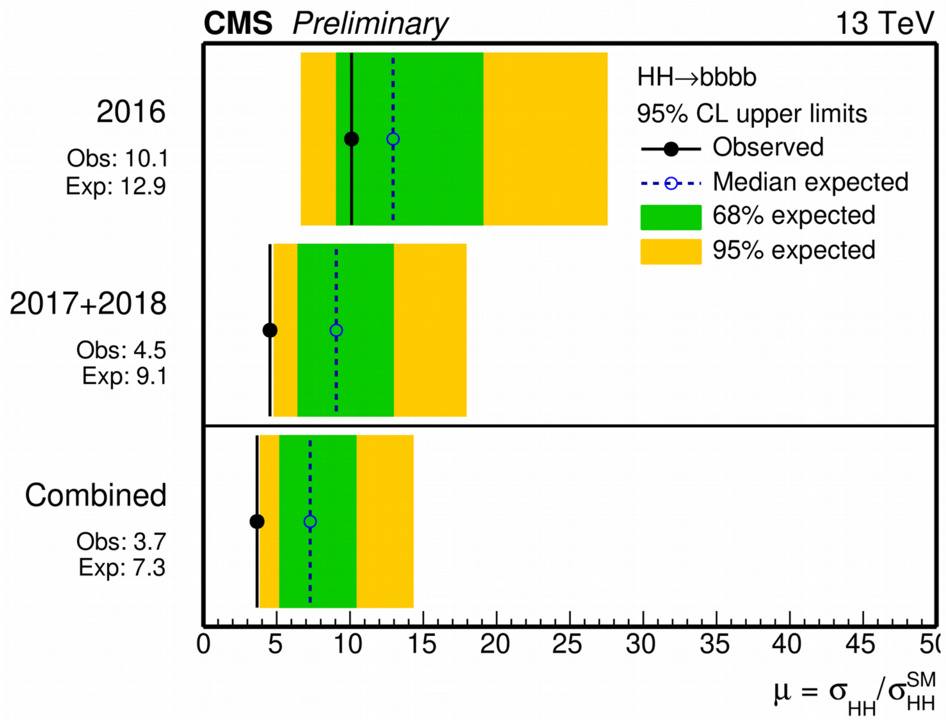


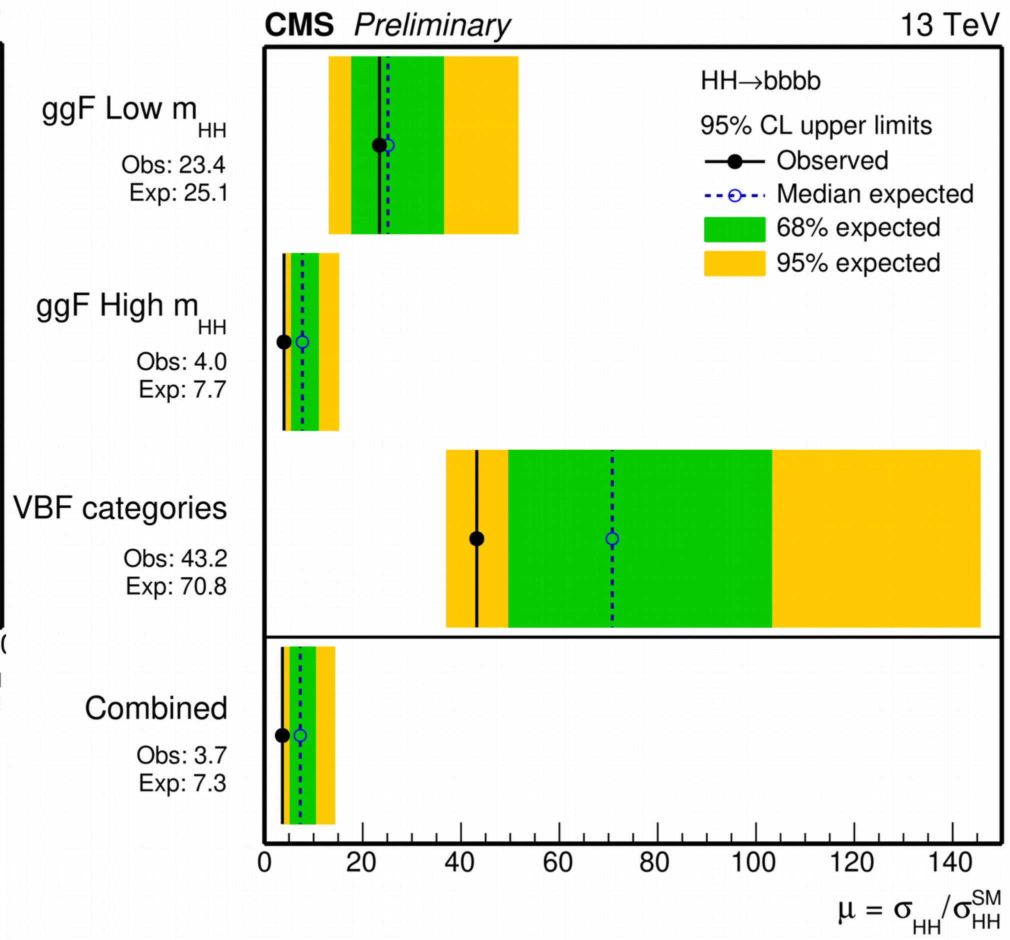
Upper limit on signal strength by year and categ

No excess data events is observed relative to the background-only expectation

95% CL upper limits are set using the asymptotic CLs method



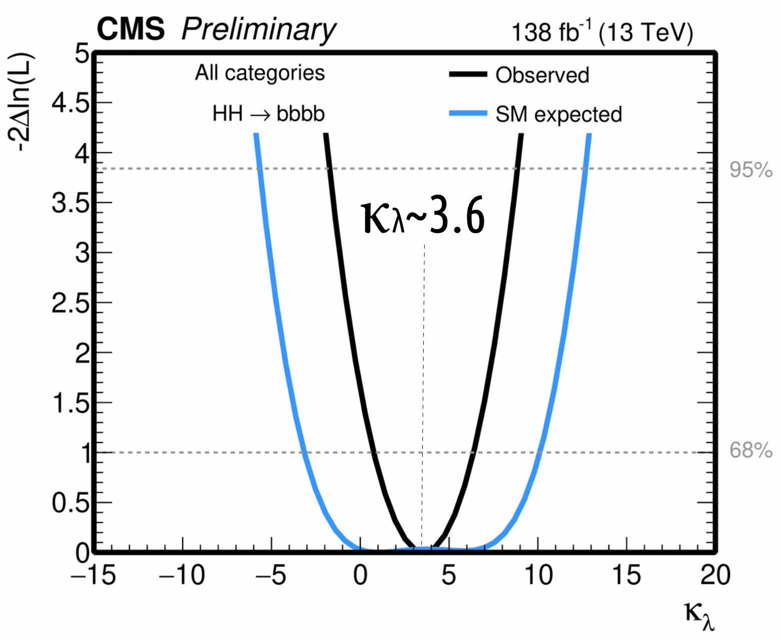


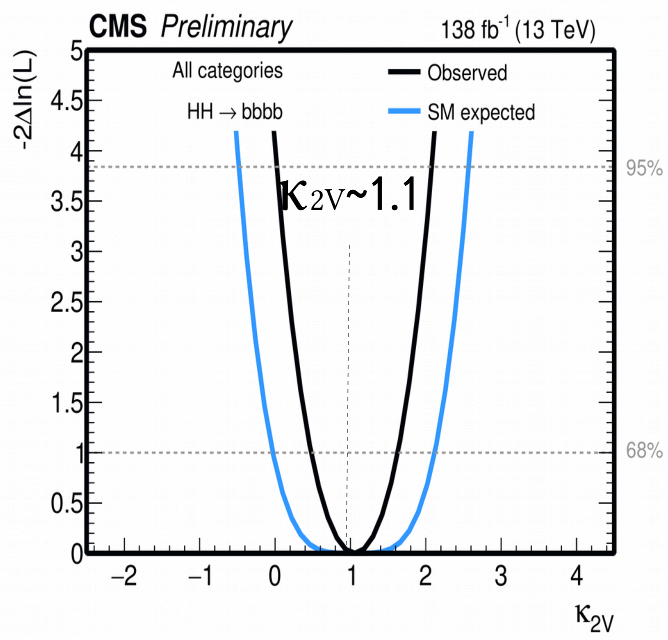


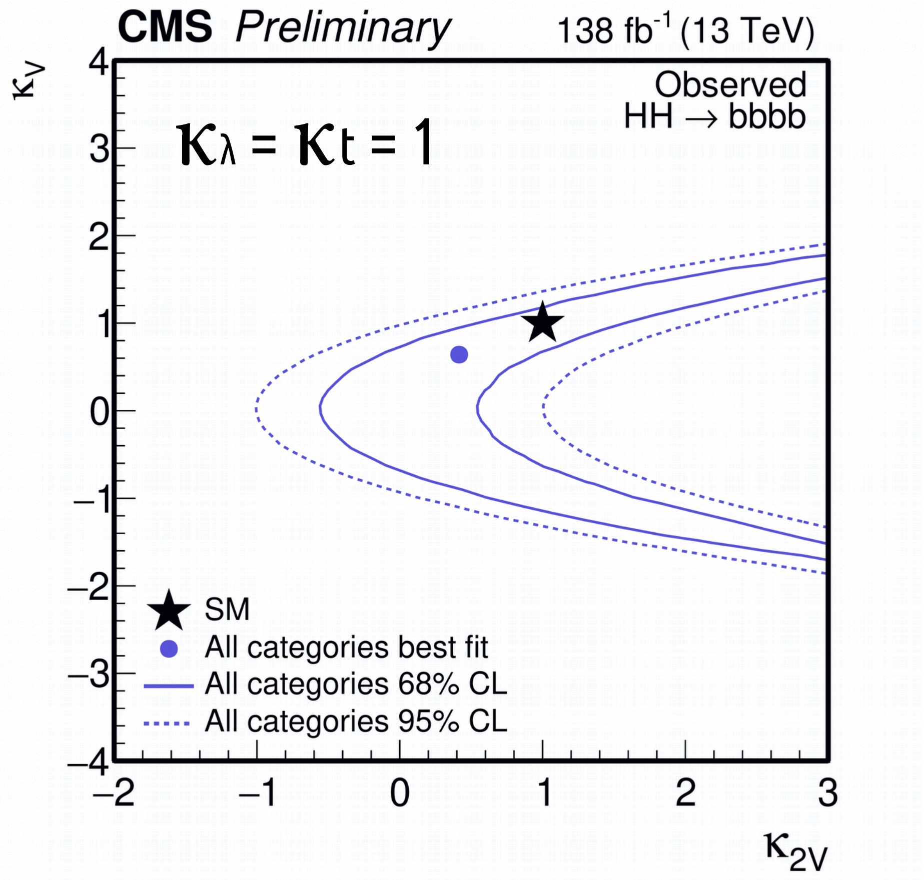
Additional results: Likelihoods scans

Assuming that a HH signal exist, one can measure couplings using via the negative log-likelihood scan

- Simultaneous fit of the ggF and VBF signal contributions as function of the couplings
- One-dimensional likelihoods → Besf-fit + 68% and 95%CL intervals
  - Scan for κ_λ, assuming $\kappa_{2V}=\kappa_V = \kappa_t = 1$
  - Scan for κ_{2V}, assuming $\kappa_\lambda= \kappa_V = \kappa_t = 1$
- Two-dimensional likelihoods→ Best-fit + 68% and 95%CL contours
- Observed best-fit values compatible with the SM at 95% CL

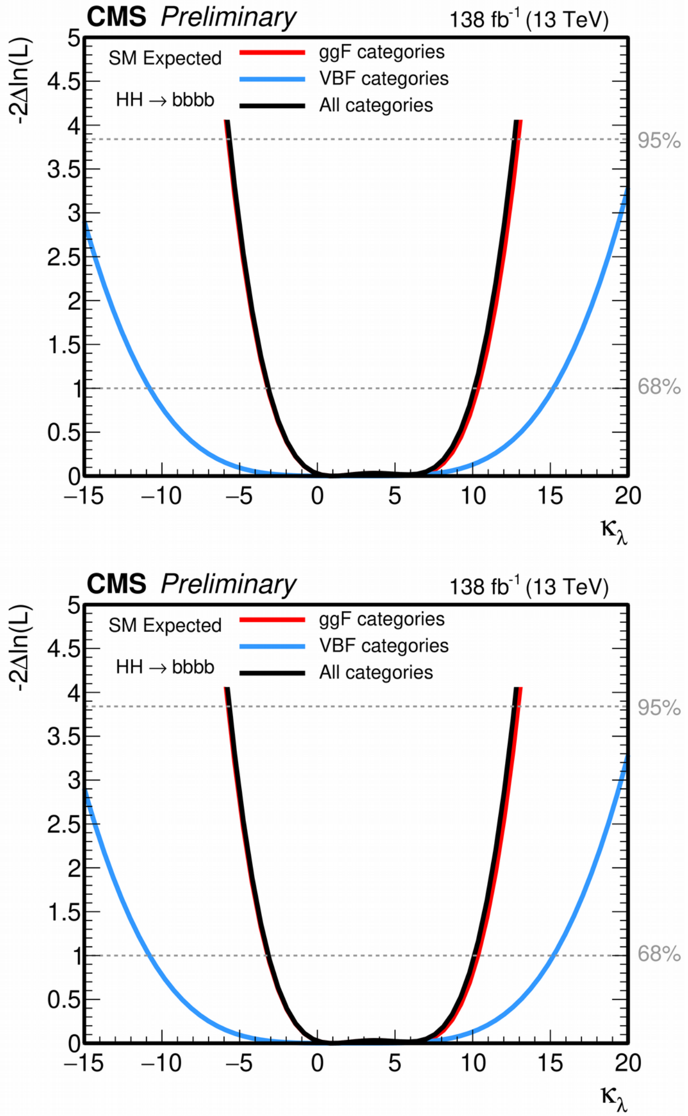




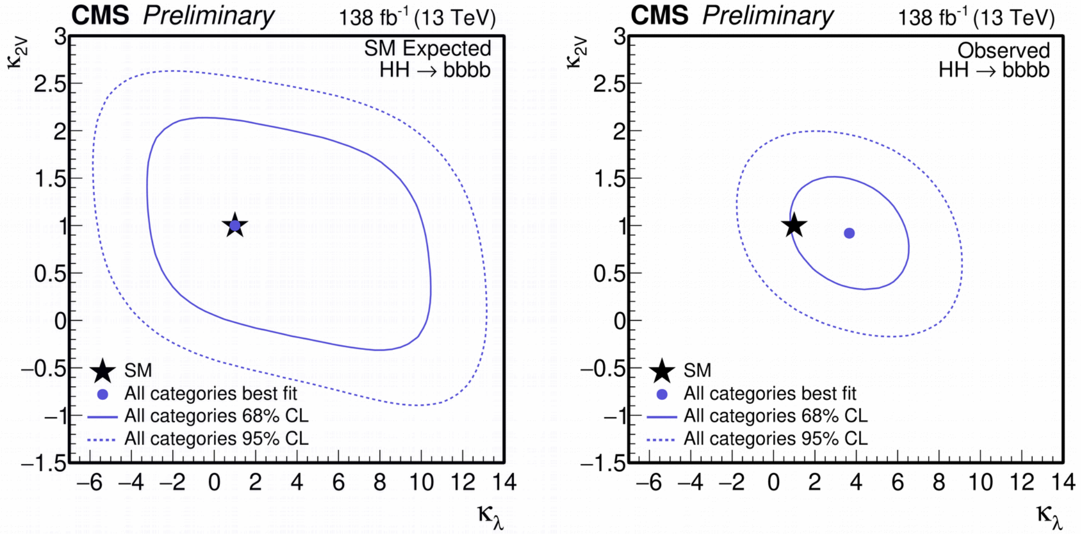


Likelihoods for various couplings

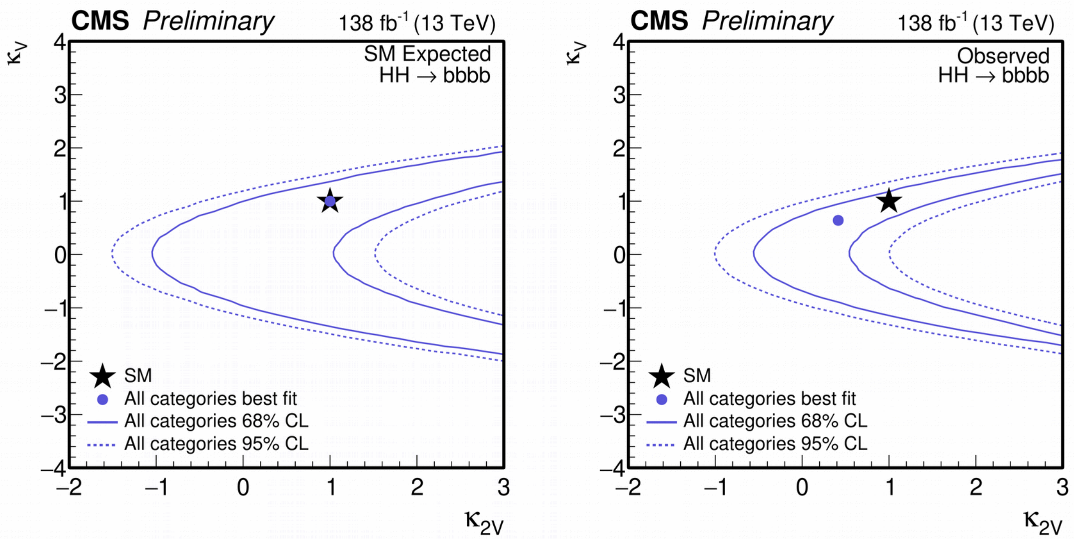
Scan versus κ_λ, assuming $\kappa_{2V}=\kappa_V=\kappa_t=1$



Contours in (κ_{2V},κ_V) plane



Contours in (κ_λ,κ_{2V}) plane



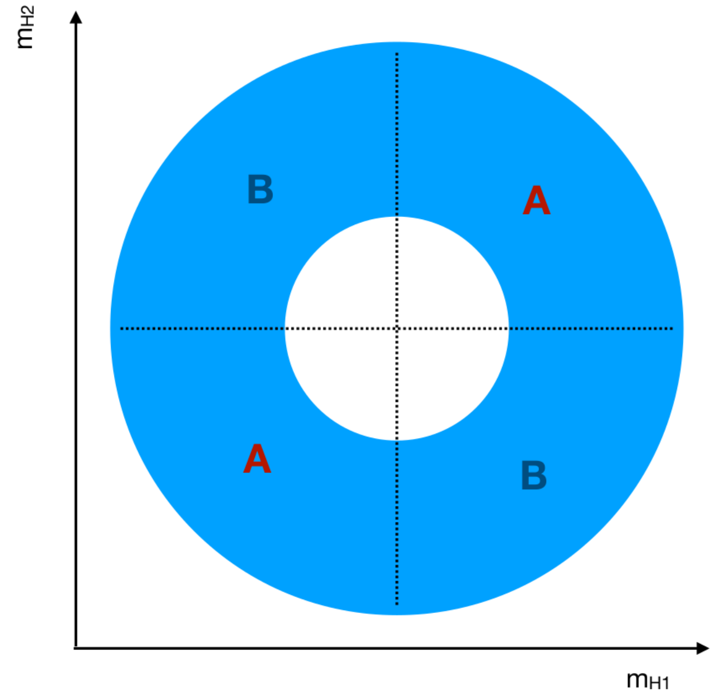
Systematic uncertainties

Signal experimental:

- Luminosity uncertainties in 2016 (1.2%), 2017 (2.3%) and 2018 (2.5%)
- Pile-up reweighting, L1 Pre-firing (2016, 2017)
- b-tagging and trigger efficiency
- Jet energy scale, jet energy resolution

Signal generation and theory:

- Factorization scales, Parton-Shower (PS), and PDF
- Event migration due to PS ISR recoil scheme (only for VBF signals)
- Cross section and final state branching fraction

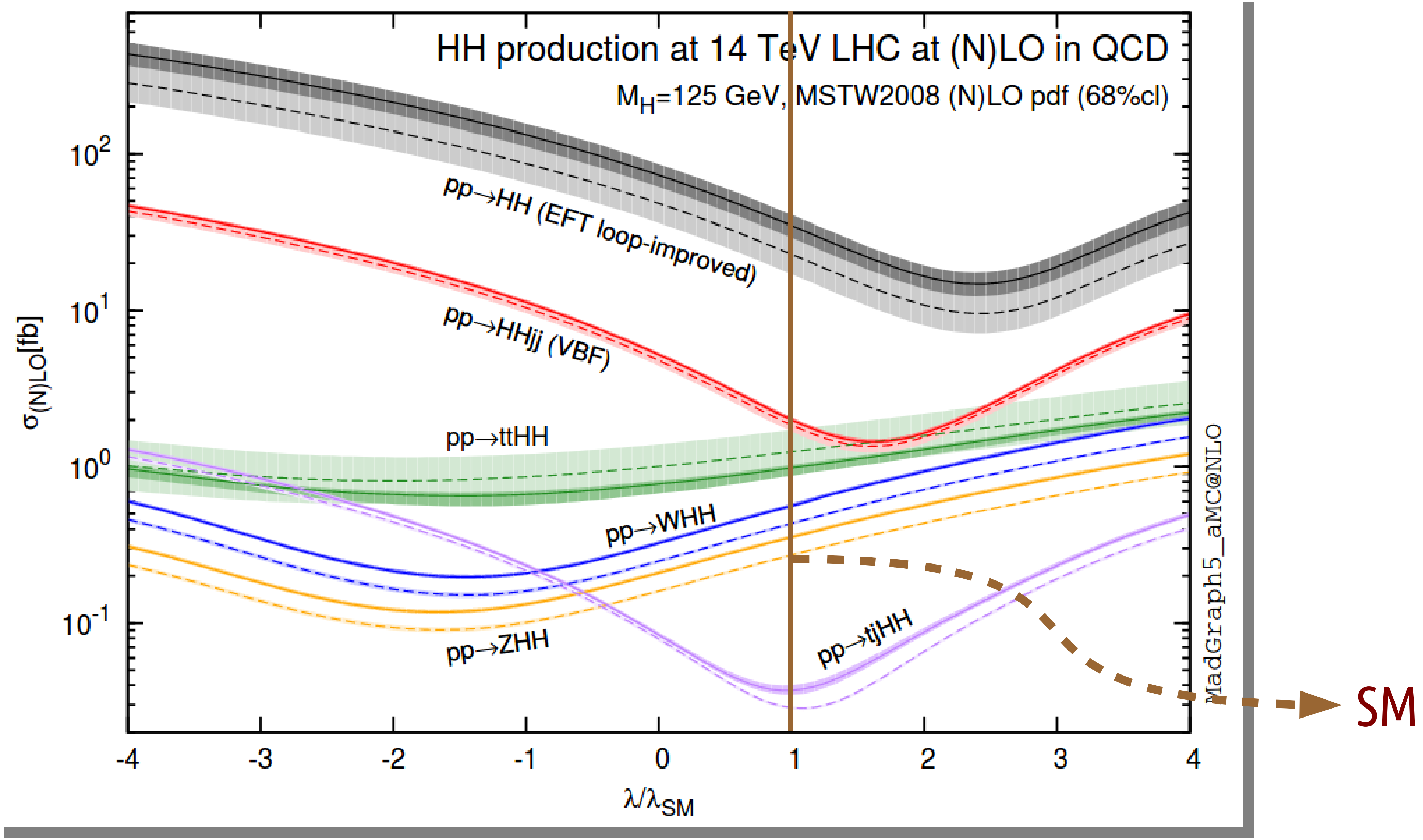


Background modeling:

- Bin-by-bin uncertainty: to account for Poisson fluctuations of the A_{SR}(3b) data
- 3b→4b Transfer factor statistical uncertainty: from limited CR statistics
- Shape uncertainty:
  - ggF cat. 1,2: Alternative shape derived using alternative CR definition
  - VBF cat. 1: Linear fit to M(HH) data/bkg ratio in validation region
- Uncertainty due to V_{SR}(4b) statistical power with respect to A_{SR}(4b)
- Uncertainty on normalization closure in V_{SR}(4b): 1.5 - 4.7% depending on the category and year

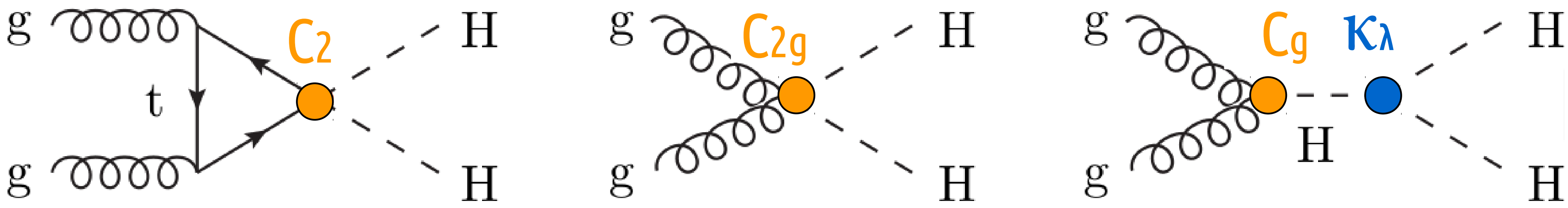
Higgs Pair Production Cross Section

Phys.Lett.B 732 (2014) 142-149



EFT approach for new physics

If the BSM physics scale is beyond the direct reach of the LHC,
its effects on the ggF HH production can be studied through a EFT model
with three contact interactions (coupling strenghts): ttHH (C2), ggHH (C2g) and ggH (Cg)



12 EFT Benchmarks are defined for LHC searches

They represent topologies of large regions of the 5-dimensional parameter space

	1	2	3	4	5	6	7	8	9	10	11	12
κ_λ	7.5	1.0	1.0	-3.5	1.0	2.4	5.0	15.0	1.0	10.0	2.4	15.0
κ_t	1.0	1.0	1.0	1.5	1.0	1.0	1.0	1.0	1.0	1.5	1.0	1.0
C_2	-1.0	0.5	-1.5	-3.0	0	0	0	0	1.0	-1.0	0	1.0
C_g	0	-0.8	0	0	0.8	0.2	0.2	-1.0	-0.6	0	1.0	0
C_{2g}	0	0.6	-0.8	0	-1.0	-0.2	-0.2	1.0	0.6	0	-1.0	0

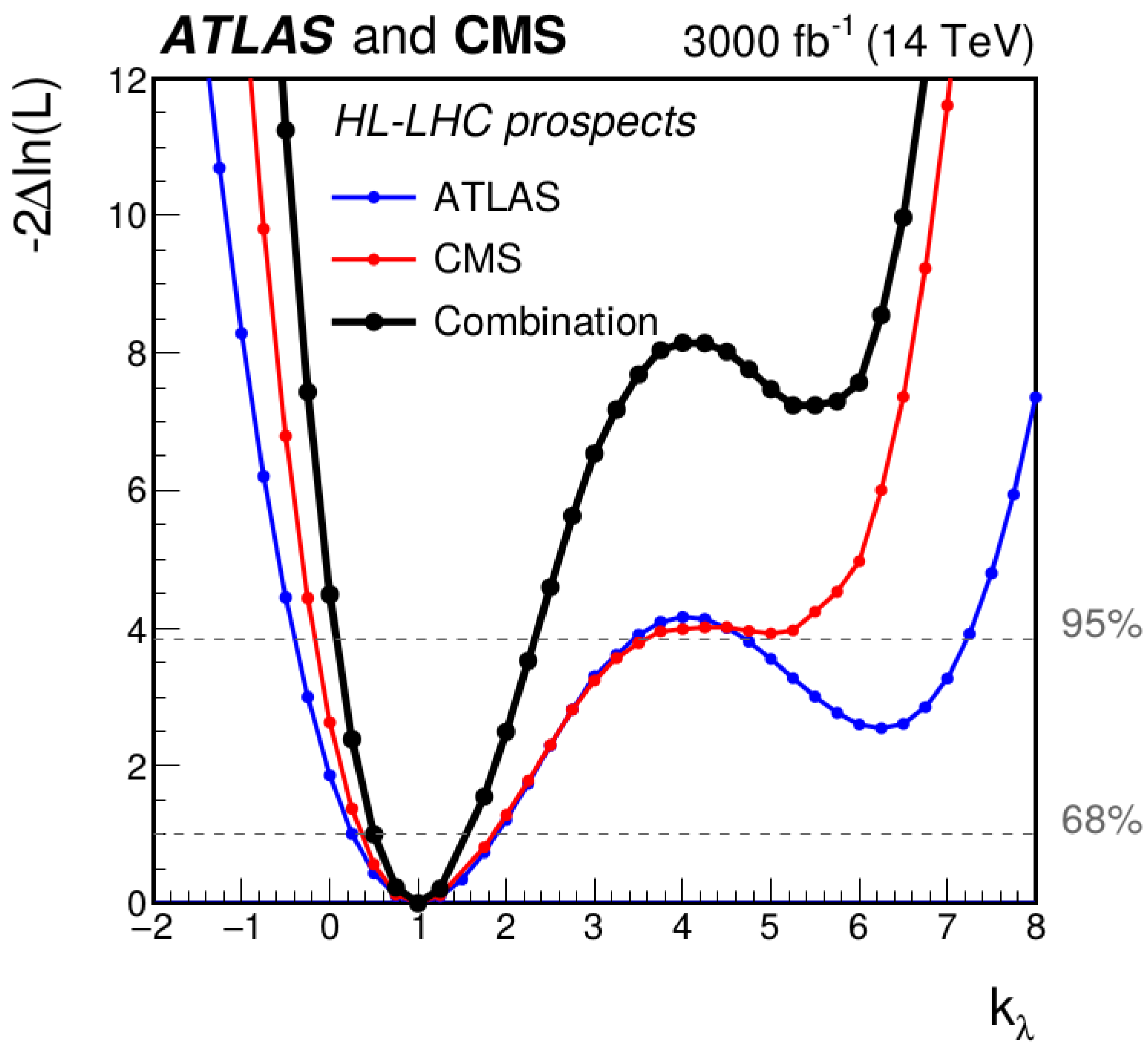
LO model arXiv:1806.05162
NLO corrections arXiv:1806.05162

Prospects at the High Luminosity-LHC

arXiv:1902.00134

	Statistical-only		Statistical + Systematic	
	ATLAS	CMS	ATLAS	CMS
bbbb	1.4	1.2	0.61	0.95
bbττ	2.5	1.6	2.1	1.4
bbγγ	2.1	1.8	2.0	1.8
bbVV→bblνlν	-	0.59	-	0.56
bbZZ→bb4l	-	0.37	-	0.37
Combination	3.5	2.8	3.0	2.6
	Combined		Combined	
	4.5		4.0	

Expected significance for SM HH production in standard deviations



Expected combined κ_λ interval

$0.57 \leq \kappa_\lambda \leq 1.5$ at 68% C.L.