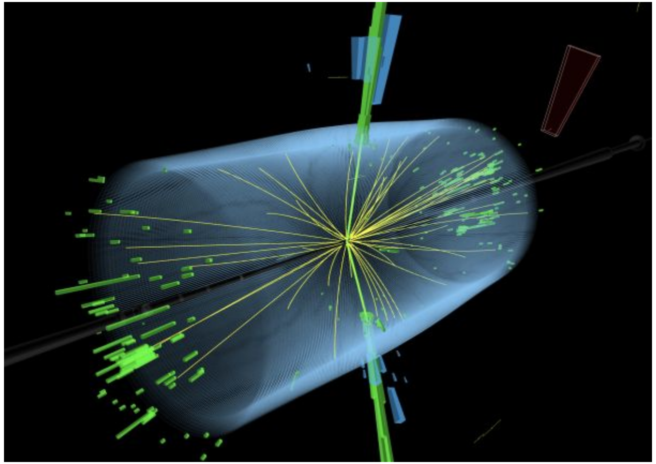
A search for light Higgs in CMS

Anirban Bala[1], Rajdeep Mohan Chatterjee[1], Monoranjan Guchait[1], Gobinda Majumder[1], Soumya Mukherjee[2]

[1] Department of High Energy Physics, Tata Institute of Fundamental Research
[2] University of San Diego

On behalf of CMS collaboration



DAE-BRNS HEP symposium, 19-23 Dec 2024

Banaras Hindu University, Varanasi

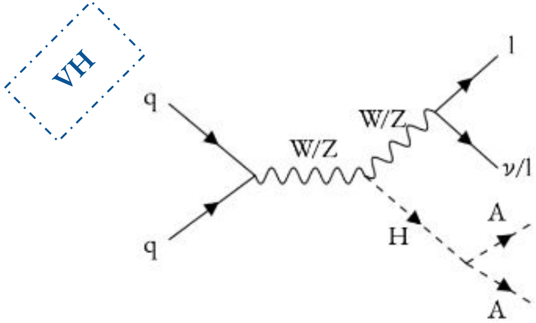
Motivation

- SM Higgs discovery ➢ precision measurement.
- But H → invisible not ruled out; constrained to be ≤ 8%. [link]
- Two Higgs doublet model (2HDM) with an additional Higgs singlet contains total 7 Higgses.
    - 3 are CP even (H_1, H_2, H_3)
    - 2 are CP odd (A_1, A_2)
    - 2 are charged Higgs ($H^{\pm}$)

Among the CP even Higgs, either H_1 or H_2 can be SM like.

- In supersymmetry, next-to-minimal supersymmetric standard model (NMSSM) is one such model.

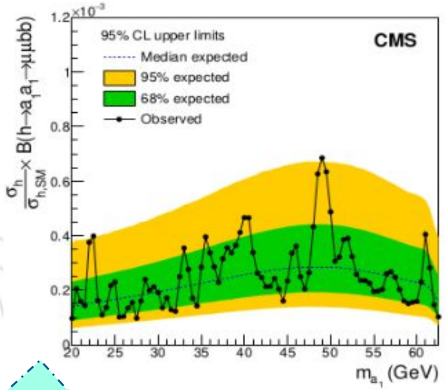
- The lightest H_1/A_1 can be also lighter than H_2 in a certain region of model parameter space. If H_1 / A_1 be dominantly singlet like, then H_1 / $A_1 \rightarrow \gamma\gamma$ branching ratio can be very high (upto ~80%) in certain region of parameter spaces.
[Monoranjan Guchait, Jacky Kumar, arXiv: 1608.05693]
- For such a scenario light Higgs bosons can be produced through SM like Higgs production.



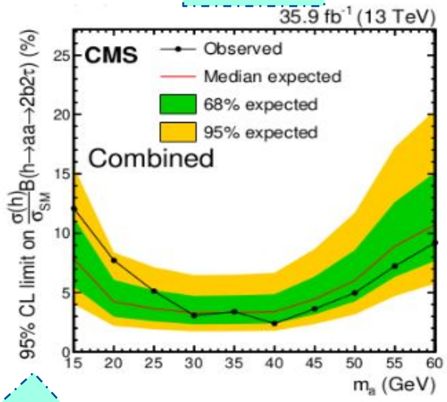
Some recent searches for light Higgs in CMS

“ggH” production mode

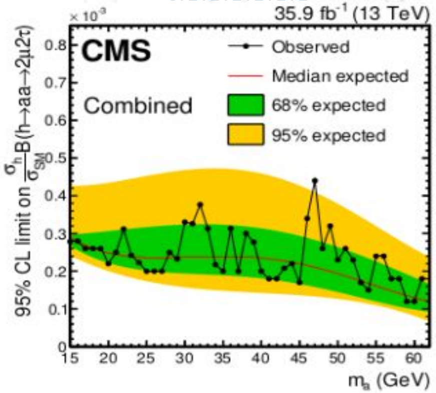
$2\mu 2b$



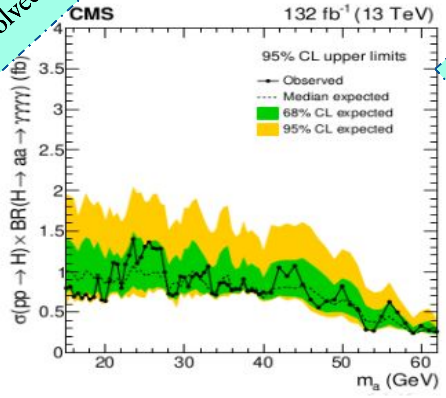
$2\tau 2b$



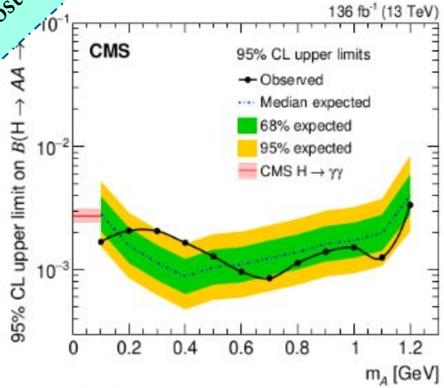
$2\mu 2\tau$



4γ (resolved)



4γ (boosted)

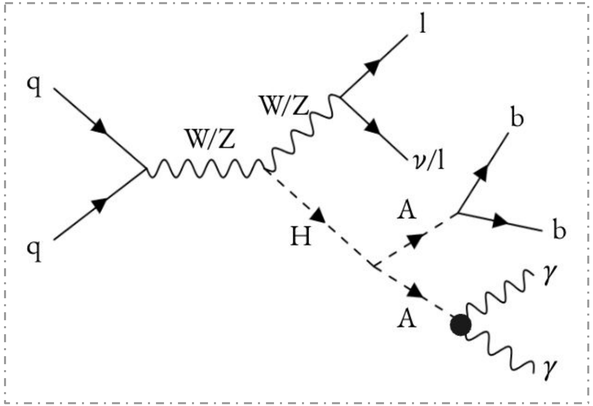


	$\gamma\gamma$	$\mu\mu$	$\tau\tau$	bb
$\gamma\gamma$	☑	×	×	×*
$\mu\mu$	×	×	☑	☑
$\tau\tau$	×	☑	×	☑
bb	×*	☑	☑	☑

Analyzing 'VH' production mode

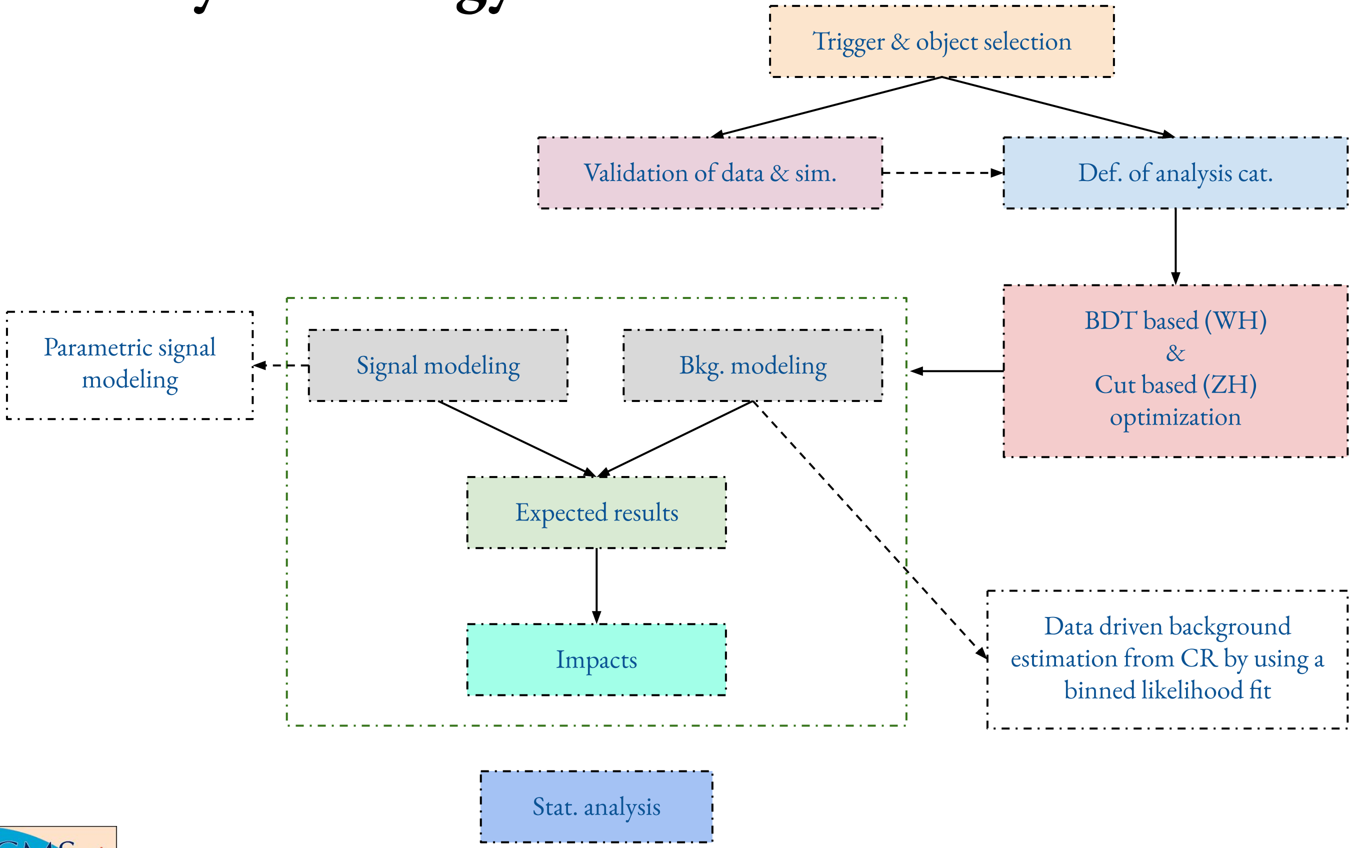
- **Full Run 2 (137 fb^{-1}) analysis is based on the VH (V=W/Z) production mode using leptonic trigger.**
- Targeting $H \rightarrow AA \rightarrow bb\gamma\gamma$ (20 GeV < M_A < 60 GeV) in the VH production mode
- The analysis is optimized assuming Br ($H \rightarrow AA$) = 10%; Br ($A \rightarrow bb$) = Br ($A \rightarrow \gamma\gamma$) = 50%

- **Final states are: one (two) triggered lepton (s) coming from W (Z) along with two b-jets and two photons.**



Signal MC	X-sec (pb)
WH/ZH	0.01/0.003
Bkg MC	**X-sec (pb)**
tt (semi lepton)	365.34
tt (dilepton)	88.29
tt + γ + jets	4.078
DY + jets	5343.25

Analysis strategy



Analysis is currently blinded and only control region has been looked at for collision data

Trigger & object selection

WH

Trigger
- Single electron trigger with p_T threshold 32 GeV
- Single muon trigger with p_T threshold 24 GeV

- Only 1 identified & isolated electron (e channel)
  - $p_T >= 34$ GeV, $|\eta| <= 2.5$
- Only 1 identified & isolated muon (μ channel)
  - $p_T >= 27$ GeV, $|\eta| <= 2.5$
- At least 2 identified b-jets
  - $p_T >= 20$ GeV, $|\eta| <= 2.4$
- At least 2 photons
  - $p_T >= 10$ GeV, $|\eta| <= 2.5$

ZH

Trigger
- Double electron trigger with p_T threshold 23, 12 GeV
- Double muon trigger with p_T threshold 17, 8 GeV

- At least 2 identified & isolated electrons (e channel)
  - $p_T >= (24, 13)$ GeV, $|\eta| <= 2.5$
- At least 2 identified & isolated muons (μ channel)
  - $p_T >= (18, 9)$ GeV, $|\eta| <= 2.5$
- At least 2 identified b-jets
  - $p_T >= 20$ GeV, $|\eta| <= 2.4$
- At least 2 photons
  - $p_T >= 10$ GeV, $|\eta| <= 2.5$

→ Additionally, the angular separation between any two final state objects, ΔR (obj1,obj2) > 0.4

→ Events are vetoed with additional leptons with loose-id, to make the WH channel orthogonal with the ZH channel.

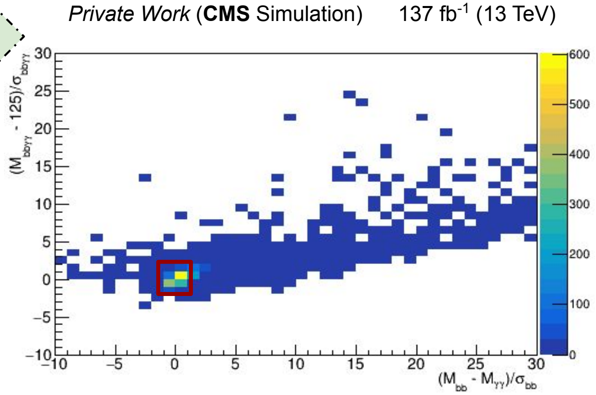
Introducing a discriminating variable 'χ^2'

We are defining a variable "χ^2" following one similar analysis strategy which is :

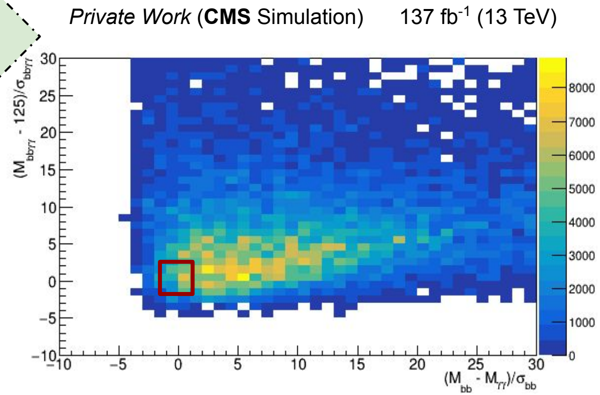
$$\chi^2 = (M_{bb} - M_{\gamma\gamma})^2 / \sigma_{bb}^{\ 2} + (M_{bb\gamma\gamma} - 125)^2 / \sigma_{bb\gamma\gamma}^{\ 2}$$

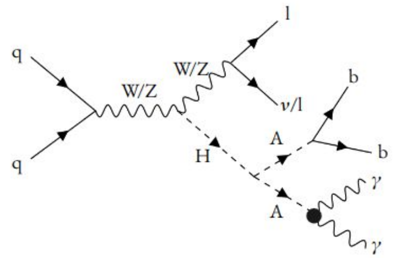
σ_{bb} ~ 5 GeV
$\sigma_{bb\gamma\gamma}$ ~ 12 GeV

Sig



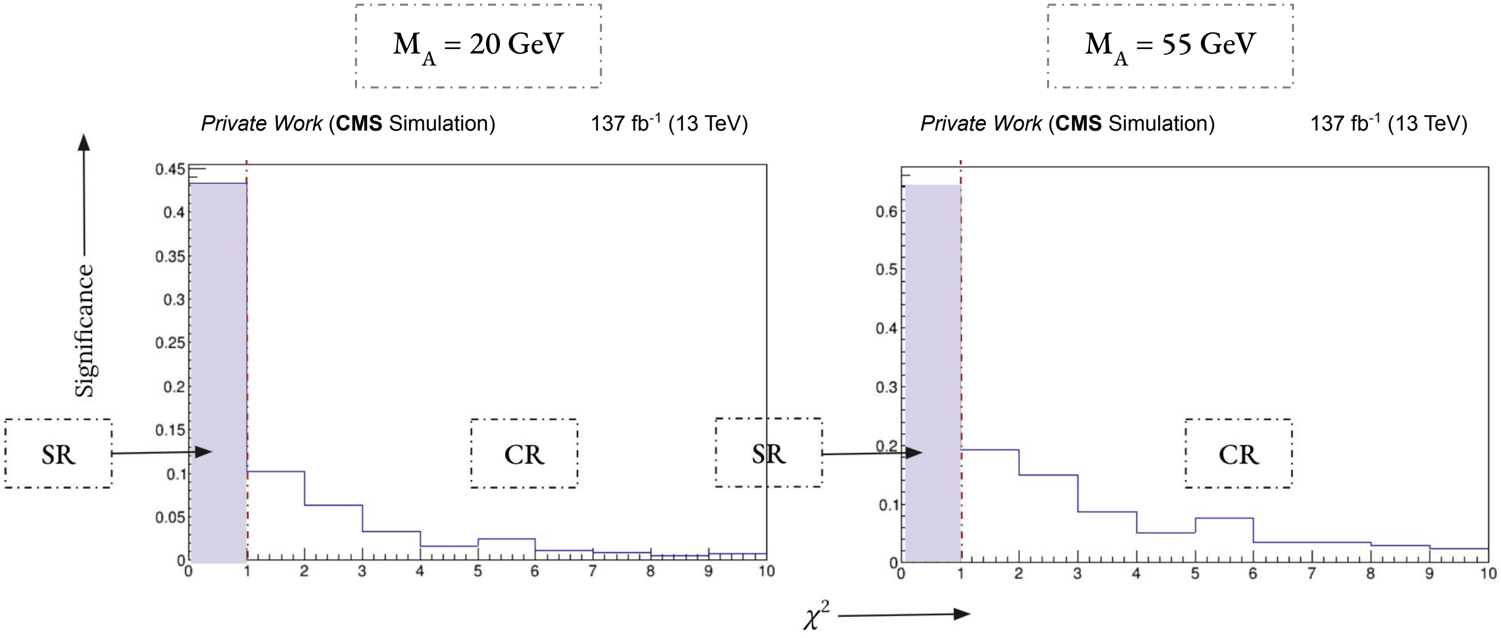
Bkg





- χ^2 distribution is concentrated near zero for the signal points whereas for bkg it is quite wide.
- We can use this variable to efficiently select our SR.
- This "χ^2" variable will be used later for background estimation.

Signal region optimization



- **We have defined our signal region (SR) : $\chi^2 \leq 1$ and control region (CR) : $\chi^2 > 3$** (Signal contamination ~0.12%).

BDT for WH channel

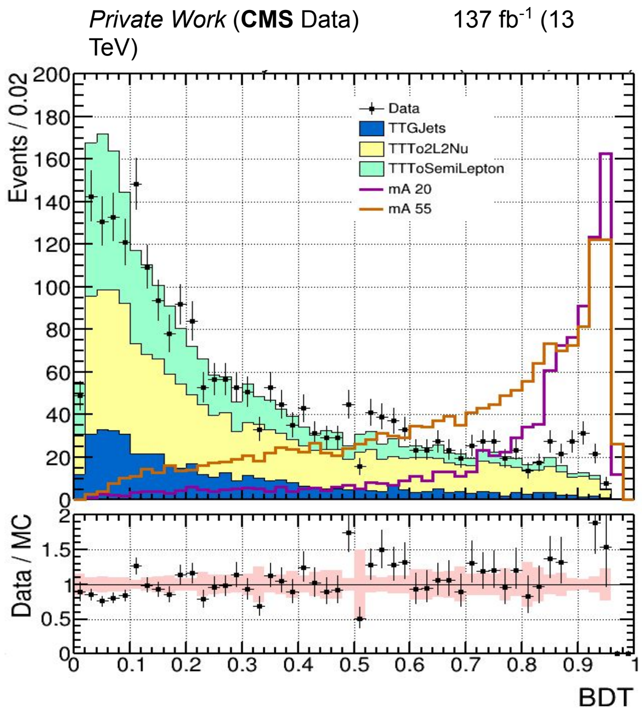
- Signal mass points: ‘A’ mass 20-60 GeV in the interval of 5 GeV.
- Different types of BDT training are possible:
    - With all the signal points together (20 + 25 + 30 + … + 60)
    - With two consecutive signal points together (20 + 25, 25 + 30, … , 55 + 60)
    - With each signal point separately (20, 25, 30, … , 60)
    - “Parametric” BDT (where “mass” is treated as an input parameter)
        - The samples corresponding to all signal mass points are merged. For the signal, an extra input feature is added which is the “true mass” that the signal event belongs to.
        - For the MC background events, an extra input feature is added which can take any random discrete value from the list of available signal mass points.
        - Boundary optimization is done independently for each signal mass point.
        - **Interpolation strategy:** For those mass points where MC signal samples are not available, the efficiency x acceptance & the boundaries are taken from the nearest mass point where samples are available.
- Any input feature for which the correlation with respect to M_A is > 10% is not used in the BDT training
- **Input features:**
    - Total 19 variables considered for training

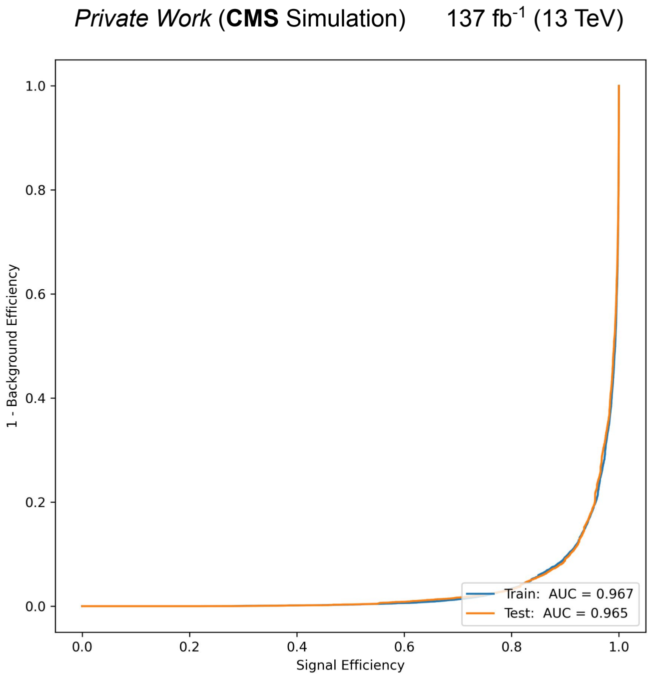
BDT input features

- Lepton (e/μ)
  - p_T, η
- b-jets
  - η of leading and sub-leading b-jet
  - $p_T^{\text{leading b}}/M_{bb}$, $p_T^{\text{sub-leading b}}/M_{bb}$
  - Identification score of leading and sub-leading b-jet
- No. of jets
- Photons
  - η of leading and sub-leading γ
  - $p_T^{\text{leading }\gamma}/M_{\gamma\gamma}$, $p_T^{\text{sub-leading }\gamma}/M_{\gamma\gamma}$
  - Identification score of leading and sub-leading γ
- Multi-object features
  - $\Delta\varphi_{\gamma\gamma}$, $\Delta\varphi_{bb}$, $\Delta\varphi_{(bb)(\gamma\gamma)}$, $\Delta\varphi_{(\gamma\gamma)(MET)}$

- We are dividing the p_T of leading and sub-leading b-jets (photons) by M_{bb} ($M_{\gamma\gamma}$) just to decorrelate the variables with M_A.

Signal-background separation and ROC

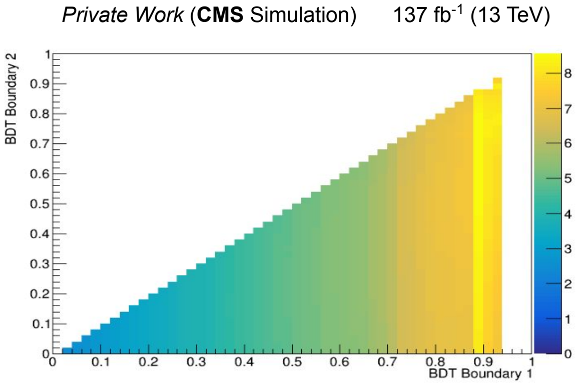




No evidence of over-training

Event categorization

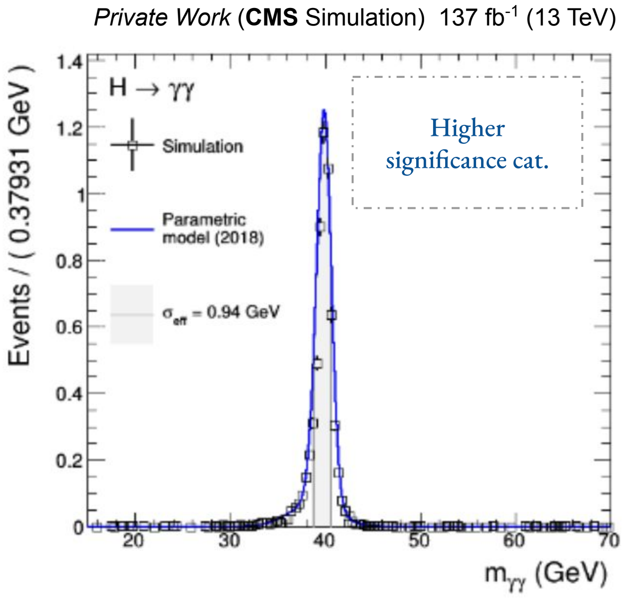
- ➢ Event categorization based on BDT.
- ➢ Categorized into two dimensional phase spaces.



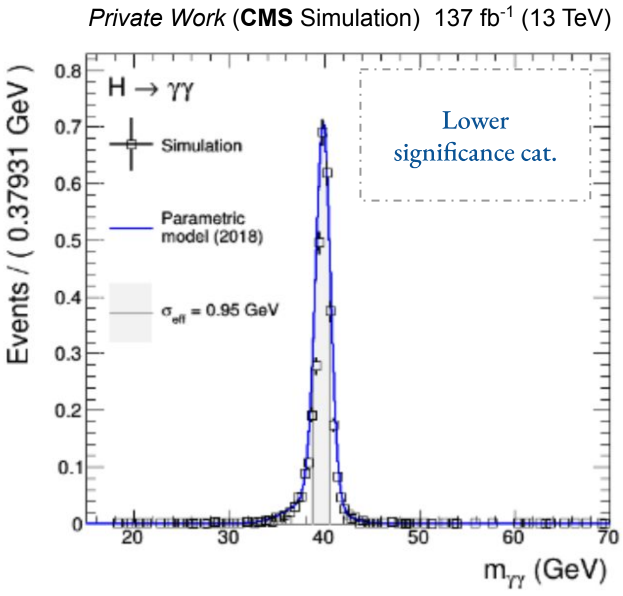
- ➢ **2D phase space**
  - Two boundaries
  - Here the constraints are: the second boundary should be always left to the first one & at least seven background events in each category
  - SR1, SR2 and CR defined by χ^2 and BDT
  - **SR1 ➢ $\chi^2 \leq 1$; BDT ≥ 0.89**
  - **SR2 ➢ $\chi^2 \leq 1$; $0.77 \leq$ BDT < 0.89**
  - **CR ➢ $\chi^2 > 3$; BDT < 0.77**
- ➢ Formal boundary optimization has been done by a 2D "grid search" method.

Signal modeling

- Parametric signal modeling constructed by fitting all the signal mass points (20-60 GeV) simultaneously.
- Modeling is done using sum of Gaussians (upto 5).
- Here sum of 3 gaussians is used for WH channel and sum of 2 gaussians is used for ZH channel. (using χ^2/ndf gof figure of merit)



WH channel; Tag 0



WH channel; Tag 1

Overview of background estimation

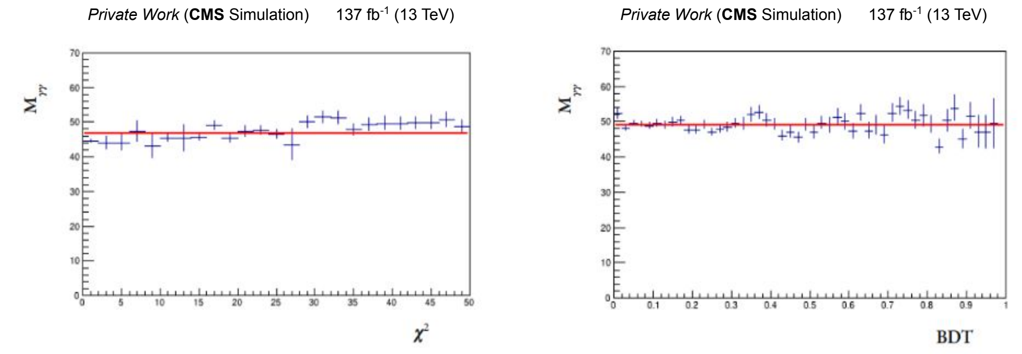
The background estimation is done using the "envelope" method developed for the [$H \to \gamma\gamma$ analysis].

In the standard $H \to \gamma\gamma$ analysis, the bkg shape is a falling distribution between 100 - 180 GeV. SR is between 115 - 135 and the remaining part is the sideband region. But for our case, we have a peaking bkg distribution between 15-70 GeV, also we do not have any lower sideband region. Currently our analysis is blinded.

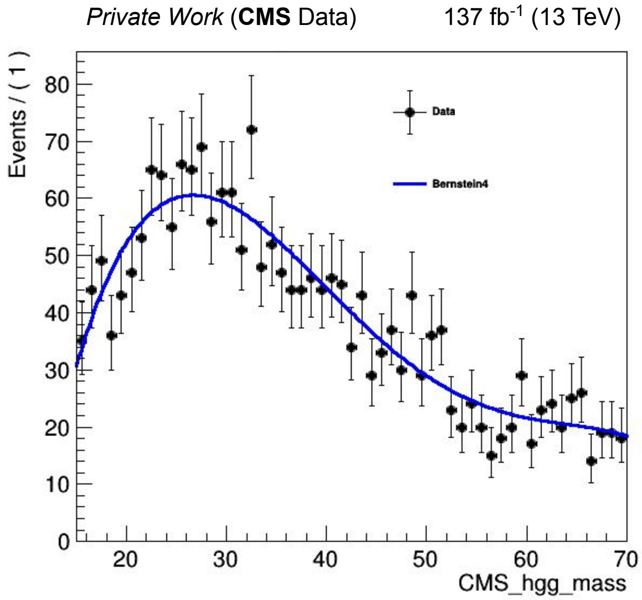
(1) Define a control region using χ^2 and BDT.

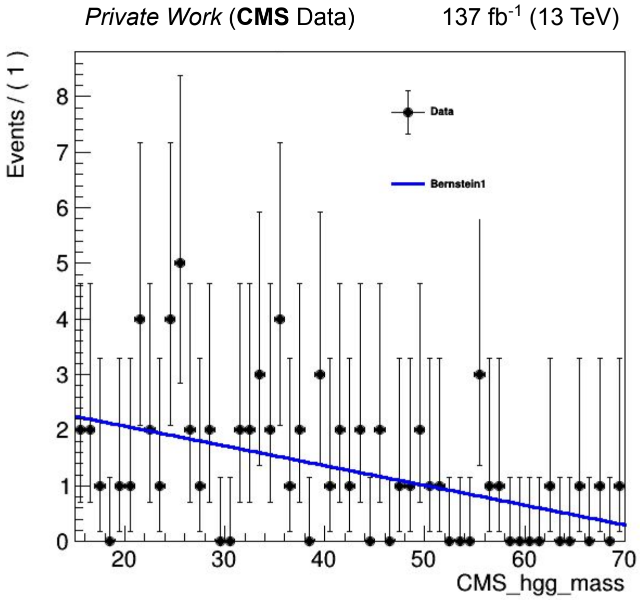
(2) Propagate shape of $M_{\gamma\gamma}$ from CR to SR.

(3) Correlation of $M_{\gamma\gamma}$ with χ^2 and BDT was checked.



Background modeling (data driven)



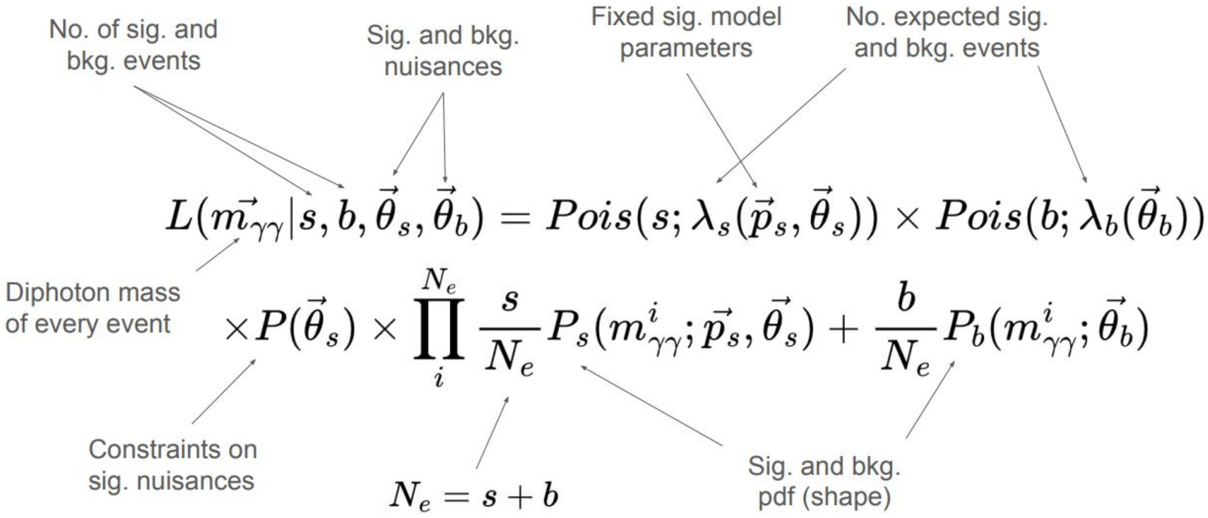


WH channel

Modeling is done on CR data

ZH channel

Likelihood analysis



$$L(\vec{m_{\gamma\gamma}}|s, b, \vec{\theta}_s, \vec{\theta}_b) = Pois(s; \lambda_s(\vec{p}_s, \vec{\theta}_s)) \times Pois(b; \lambda_b(\vec{\theta}_b))$$

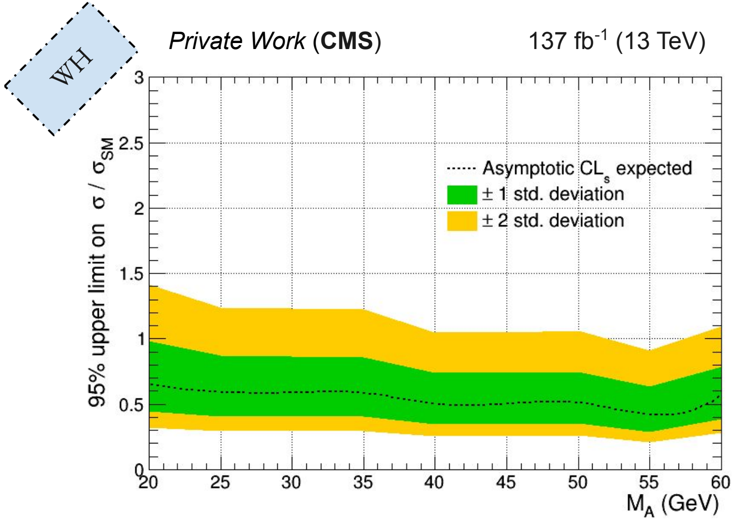
$$\times P(\vec{\theta}_s) \times \prod_{i}^{N_e} \frac{s}{N_e} P_s(m^i_{\gamma\gamma}; \vec{p}_s, \vec{\theta}_s) + \frac{b}{N_e} P_b(m^i_{\gamma\gamma}; \vec{\theta}_b)$$

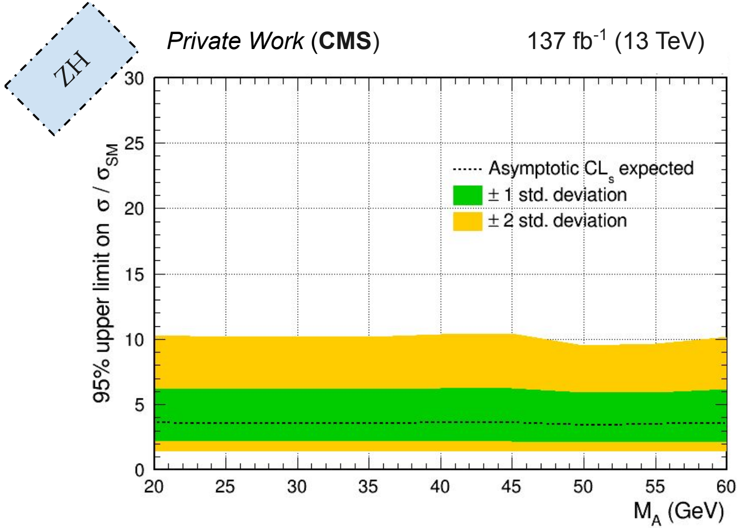
$$N_e = s + b$$

- Main role of this statistical analysis is to create a likelihood that can be used further to extract the exclusion limit.
- Here, we can reparametrize the likelihood in terms of 'μ' which is $\mu = s/\lambda_s$ as we want the signal strength (μ) as our parameter of interest (POI).
- More is here.

Expected limit

- Our analysis is stat. dominated



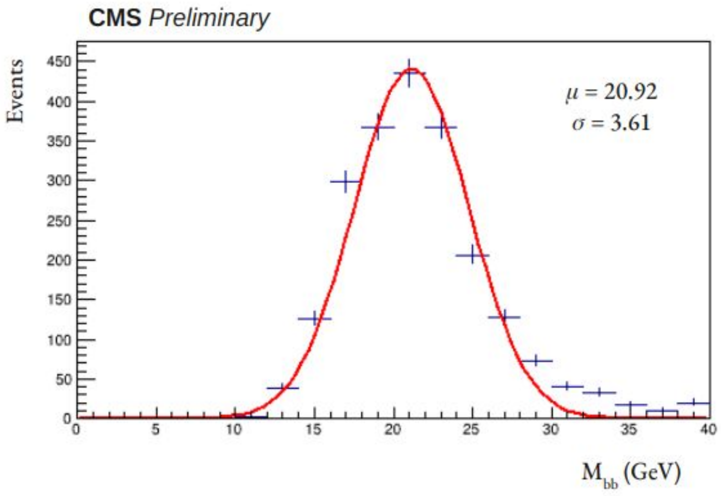


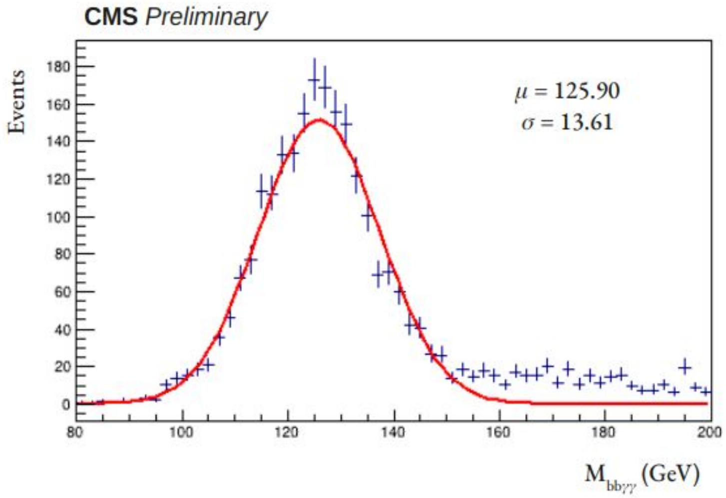
Summary

- This analysis is being explored for the first time in CMS.
- This search channel is motivated by NMSSM. We are targeting a phase space where $A \rightarrow \gamma\gamma$ branching ratio is very high (due to singlet nature) along with the significant $A \rightarrow bb$ decay mode to probe this channel.
- The main challenges of this analysis are two folds.
    - For Run-2, non availability of the diphoton trigger to target the 'ggH' production mode for this analysis.
    - The background structure is peaking instead of falling.
- Future stages:
    - For Run-3 this kind of analysis can be extended as for those triggers the diphoton invariant mass cut is not there. Also for 2018 era (Run-2), the diphoton invariant mass cut is absent.
    - Similar searches can be possible in Run-3, the strategy may be almost similar to this.

Backup

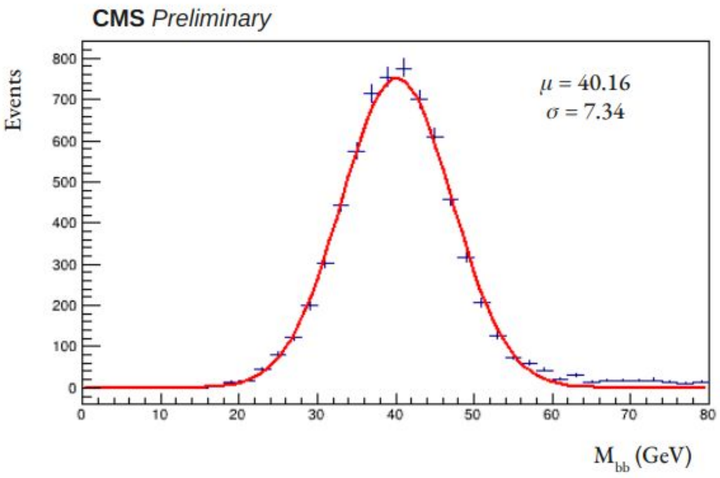
M_A = 20 GeV

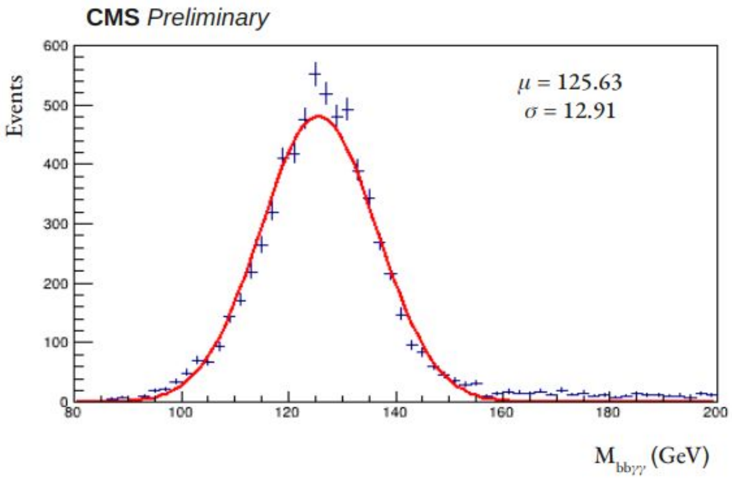




Fitting with Gaussian function

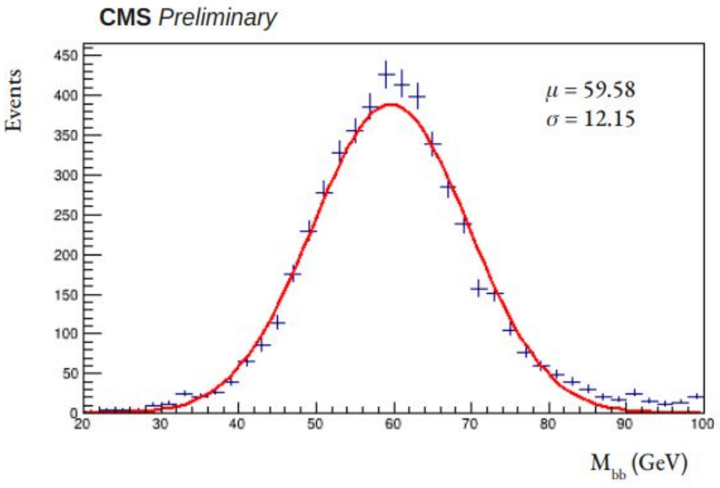
$M_A = 40$ GeV

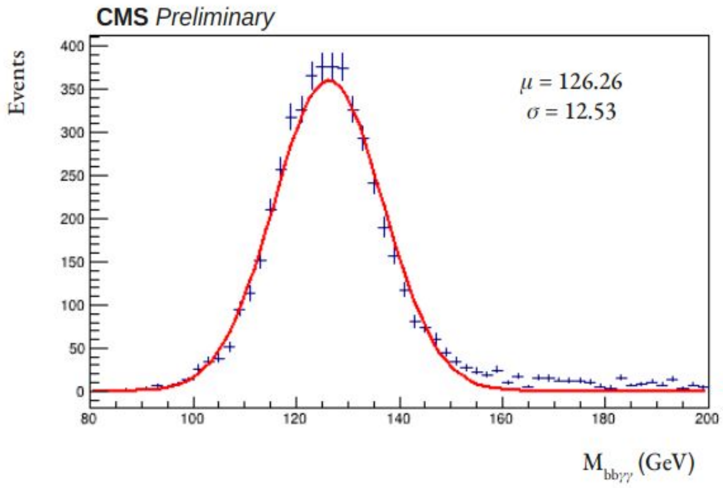




Fitting with Gaussian function

M_A = 60 GeV





Fitting with Gaussian function

σ_{bb} and $\sigma_{bb\gamma\gamma}$ as a function of M_A

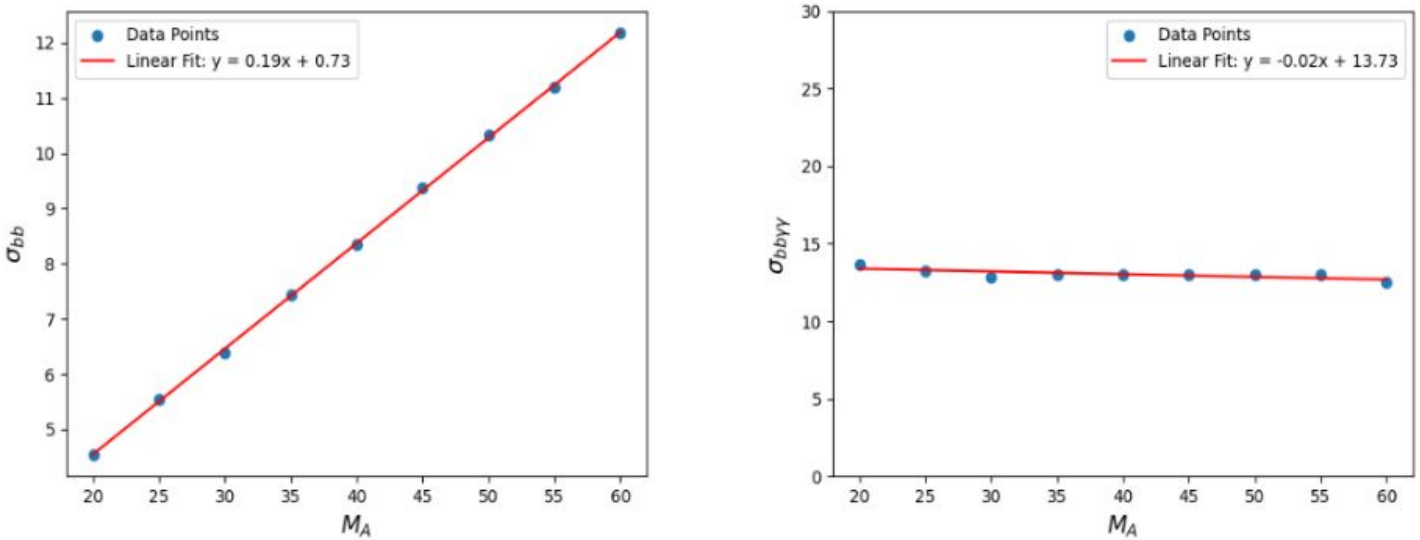
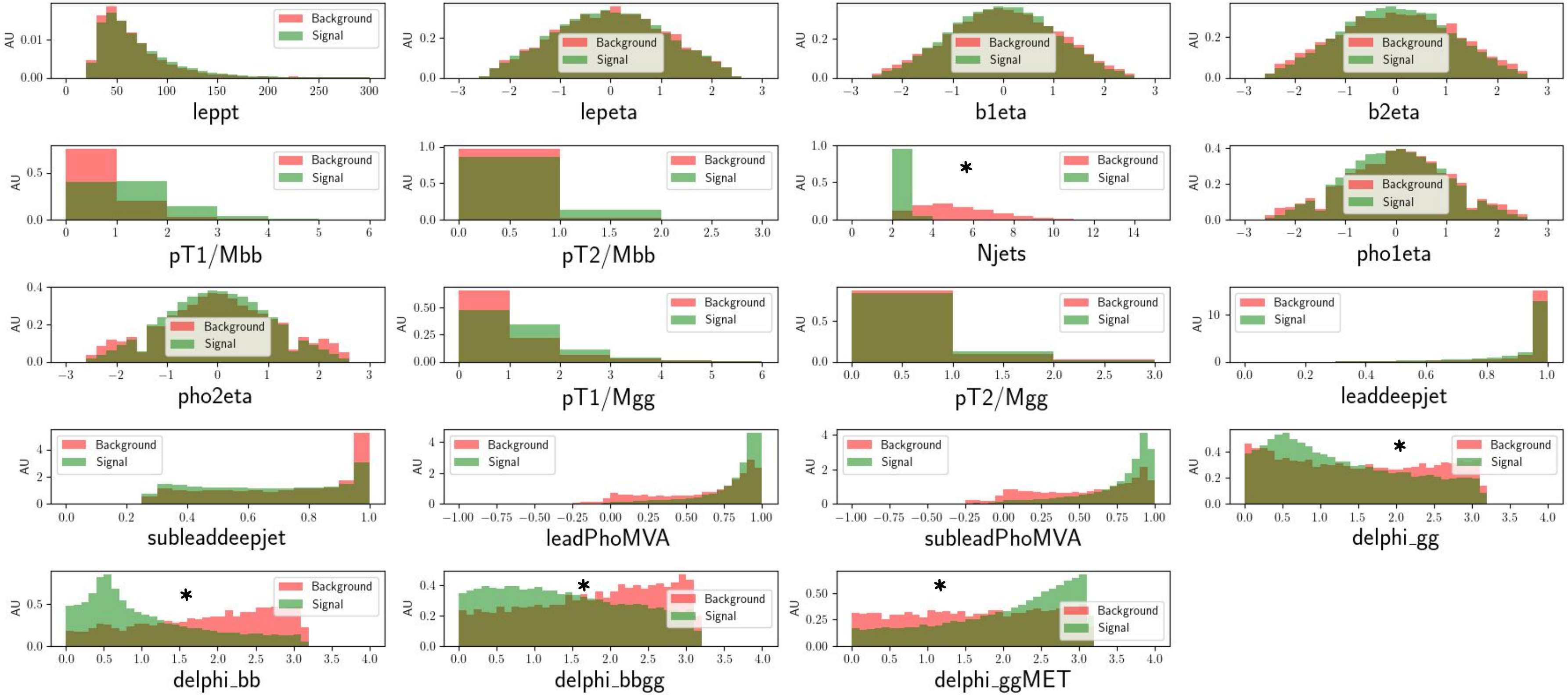


Figure 6: In the upper plot we are showing that σ_{bb} gradually increases with M_A whereas $\sigma_{bb\gamma\gamma}$ is almost independent of M_A.

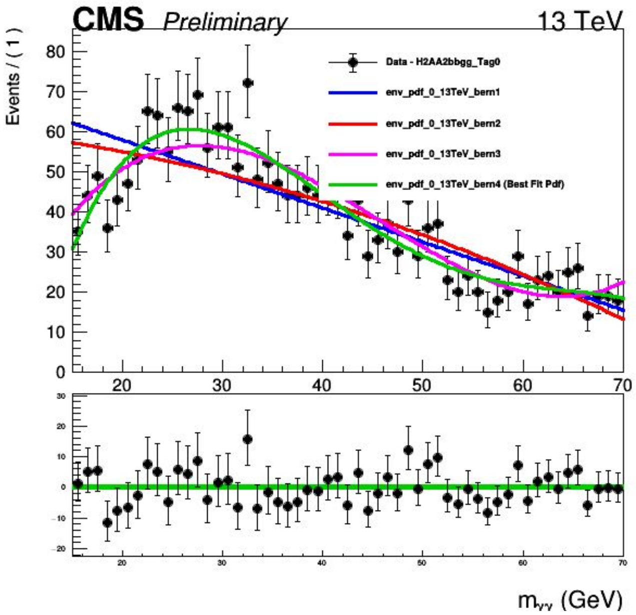
Boundary optimization by 2D grid search (WH)

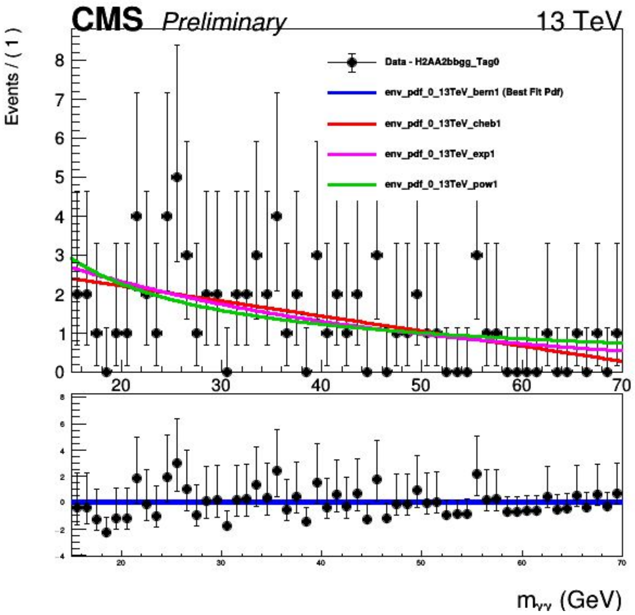
M_A	Left boundary	Sig. events	Bkg. events	Significance	Right boundary	Sig. events	Bkg. events	Significance	Total significance
20	0.81	1.91	20.25	0.41	0.91	3.86	11.86	1.06	1.14
25	0.81	2.59	20.25	0.56	0.91	5.27	11.86	1.43	1.54
30	0.79	3.37	24.17	0.67	0.91	5.59	11.86	1.51	1.65
35	0.77	2.83	28.70	0.52	0.91	8.33	11.86	2.19	2.25
40	0.77	3.21	28.70	0.59	0.91	7.59	11.86	2.01	2.10
45	0.77	3.79	28.70	0.69	0.91	5.91	11.86	1.59	1.74
50	0.79	2.79	24.17	0.55	0.91	8.57	11.86	2.25	2.32
55	0.79	2.99	24.17	0.59	0.91	7.92	11.86	2.09	2.18
60	0.77	4.01	25.44	0.77	0.89	4.70	15.12	1.15	1.39

Separation of input variables for signal & background



Background modeling (data driven)





WH channel

ZH channel

Modeling is done on CR data

Systematic uncertainties

Table 18: The shape and yield uncertainties.

Type of uncertainties	Source of uncertainties
Shape	Photon energy scale, smearing
Yield	Luminosity, lepton efficiency, b-tagging, photon efficiency, pileup modeling, JEC/JER (exp.); PDF, QCD, ISR/FSR (theo.)

Exp. unc.

- Event level (which can change event yield / category migration) :
  - Trigger SF [1]
  - b-tag (shape) [2]
  - Pileup reweighting [3]
  - Lepton ID / Iso [4]
  - Photon ID [5]
- Object level (which can change object property) :
  - JEC [6]
  - JER [7]
  - Scale (change shape of $M_{\gamma\gamma}$)
  - Smearing (change shape of $M_{\gamma\gamma}$)
  - Rochester correction
- Luminosity [8]

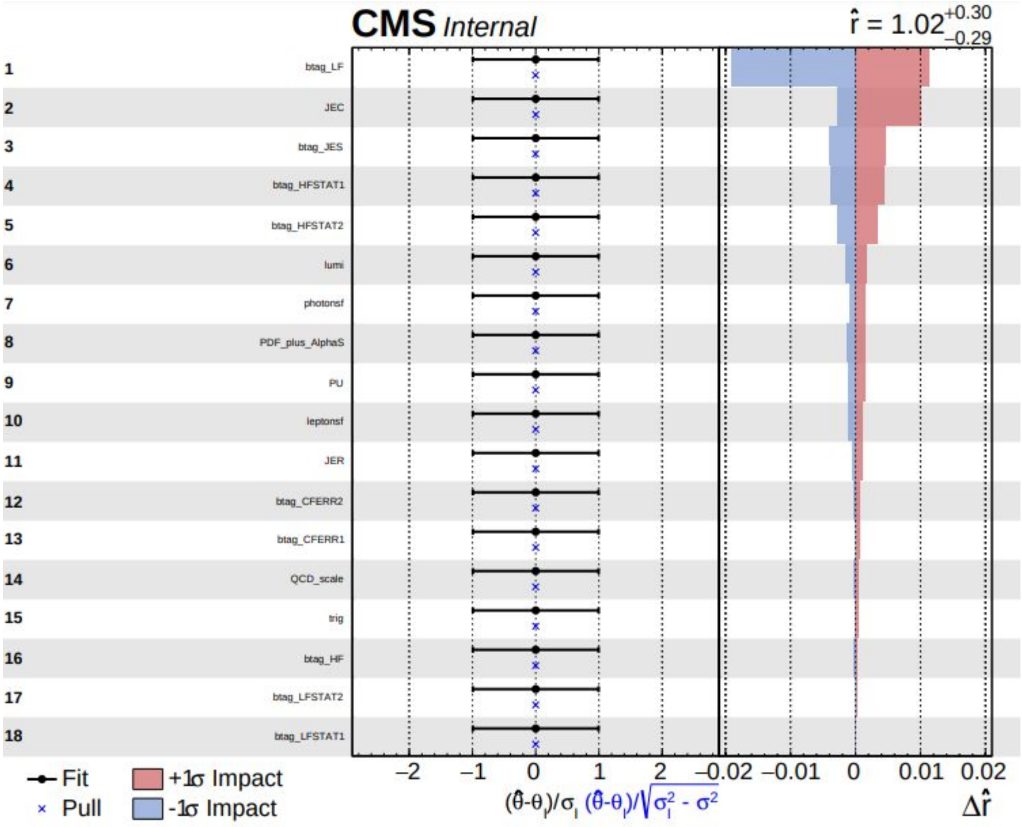
Theory unc.

- PDF [9]
- QCD scale [10]
- ISR & FSR [11]

- Our analysis is stat. dominated
- Impact plot kept in backup

Impact plot

$M_A = 40$ GeV



CMS Internal
r̂ = 1.02 +0.30 −0.29
1 btag_LF
2 JEC
3 btag_JES
4 btag_HFSTAT1
5 btag_HFSTAT2
6 lumi
7 photonsf
8 PDF_plus_AlphaS
9 PU
10 leptonsf
11 JER
12 btag_CFERR2
13 btag_CFERR1
14 QCD_scale
15 trig
16 btag_HF
17 btag_LFSTAT2
18 btag_LFSTAT1
Fit
Pull
+1σ Impact
-1σ Impact
−2 −1 0 1 2
−0.02 −0.01 0 0.01 0.02
Δr̂

Outline

- Motivation
- Some recent searches for “light” Higgs in CMS
- Analysis strategy
  - BDT based for WH channel
  - Cut based for ZH channel
- Statistical analysis
  - Parametric signal modeling
  - Data driven background modeling (by “envelope” method)
  - Systematic uncertainties
  - Expected results
- Summary

Statistical analysis

- Parametric signal modeling
- Data driven background estimation using "envelope" method
  - Unlike in the standard model $H \rightarrow \gamma\gamma$ analysis we have a CR in place of a sideband region
- Systematic uncertainties
- Expected result
- Impact plot