VBF scalar to bb analysis

Fundamental Interactions & IAP meeting

P. Azzurri, D. Caiulo, S. Donato, G. Rauco, C. Vernieri
INFN and Scuola Normale, Pisa

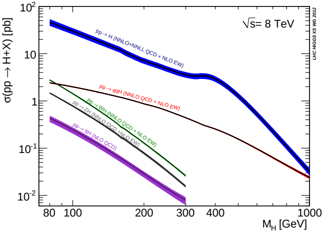
S. Alderweireldt, S. Bansal, T.Cornelis, X. Janssen,
J. Lauwers, N. van Remortel
University of Antwerp

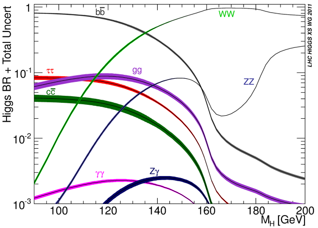
S. de Visscher, K. Kousouris
CERN

June 19th, 2015 – ULB, Brussels, Belgium

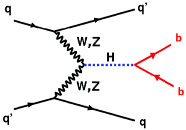
1. Introduction

Signal properties & search strategy



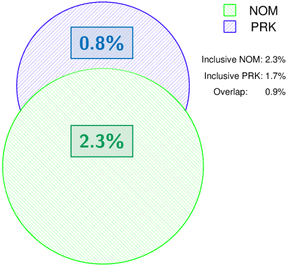


- **Properties of the VBF $H \to b\bar{b}$ channel**:
  - cross section significantly larger than for VH or ttH production
  - very large QCD background
  - trigger challenges
  - no such analysis from Atlas yet
- **4-jet signal topology**:
  - **2 central** b-jets
  - **2 more forward** jets with large $\Delta\eta$ and m_{jj}
  - suppressed colour-flow between VBF jets



- **Search strategy**:
  1. **topological trigger** on the signal main properties *(jets with large $\Delta\eta$, one b-jet, etc.)*
  2. use **multivariate methods** to exploit maximally the (significant) differences between signal and QCD *(don't use b-jet kinematics: stay orthogonal to the m_{bb})*
  3. perform a **fit of the $m_{b\bar{b}}$** spectrum

Data & background samples
Two phase spaces: NOM (set A) and PRK (set B)



- Primary data sample: 2012 nominal dataset → 19.8 fb^{-1}
  - **dedicated triggers in place** (VBF + b-jet criteria)
  - **treated as inclusive:**
    set A selection = set A trigger + set A preselection
  - VBF signal acceptance after trigger & preselection: **2.3%**
- Additional data sample: 2012 parked dataset → 18.2 fb^{-1}
  - **using general purpose triggers** (VBF criteria only)
  - **treated as exclusive:**
    set B selection = *veto set A selection* + set B trigger + set B preselection
  - VBF signal acceptance after trigger & preselection: **0.8%** → **35% more data**
- Background samples (in order of importance)
  1. QCD multijets
  2. Z+jets
  3. TT+jets
  4. single top
  5. W+jets

Event interpretation

When considering the 4-jet final state, an algorithm is required to identify the b-jet and VBF-tag jet candidates:

- **the 4 leading jets in p_T are the signal jets**
- **b-jet identification**
  - ✗ **straight-forward method**: take the two jets with the highest CSV b-tag value
  - ✓ **optimized**:
    - consider b-tag value, b-tag ordering, jet η and η-ordering of all jets and evaluate with **multivariate discriminant**.
    - jets originating from the Higgs boson: highest CSV b-tags and central rapidities $\rightarrow$ combination of variables should provide improvements over CSV only.
  - ✓ **improved signal acceptance**: $\sim 10\%$

2. Trigger and Preselection

Trigger paths

Optimized preselections

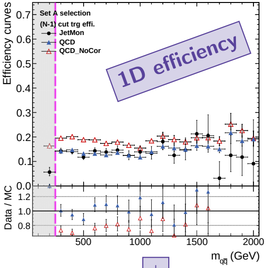
Trigger efficiency

Two sets of triggers: NOM (set A) and PRK (set B) & optimized, matching (pre)selections

	Set A (dedicated)	Set B (general)
L1	L1_TripleJet	L1_DiJet
	cuts on p_T + additional kinematic requirements	
HLT	HLT_QuadJet (10 Hz)	HLT_DiJet (100 Hz)
	4 p_T cuts (CALO or PF)	2 p_T cuts (CALO)
	b-tagging	no b-tagging (suboptimal)
	cuts ∼ VBF kinematics	cuts ∼ VBF kinematics (too high)
preselection	PF jetPt$_{0,1,2,3}$ > 80,70,50,40 GeV	PF jetPt$_{0,1,2,3}$ > 30,30,30,30 GeV
		jetPtAve(0,1) > 80 GeV
	m_{qq} > 250 GeV & $\Delta\eta_{qq}$ > 2.5	m_{qq} > 700 GeV & $\Delta\eta_{qq}$ > 3.5
		m_{qq}^{trig} > 700 GeV & $\Delta\eta_{qq}^{trig}$ > 3.5
	2x loose b-tag	1x medium + 1x loose b-tag
selection	$\Delta\phi_{bb} < 2.0$	
	loose PU ID	
	lepton veto	

Trigger efficiencies

1. **1D efficiencies are determined using an unbiased reference trigger, with N-1 preselection cuts applied.**
   - from simulation △
   - from data ●



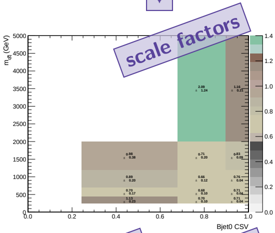
Efficiencies are

- small due to low heavy-quark content in preselected samples
- statistics limited (prescaled ref. trigger)

Preselection cuts optimized to

- reproduce trigger requirements
- sit sufficiently close to the turn-on plateau
- without reducing the signal acceptance

↓



2. **Scale factors are calculated**
   - between data / MC
   - to correct the mismodeling in trigger simulation wrt. data
   - binned in two variables → 2D correction map (choice following trigger logic and largest effects per phase space)
   - observe: uncorrected MC △ → corrected MC ▲
3. **Scale factors are convolved with 2D signal distributions to obtain trigger uncertainties**

x signal distr. = trigger uncertainties

3. Properties to exploit

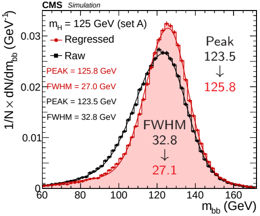
Jet p_T regression

Quark gluon likelihood

Additional hadronic activity

A. Jet transverse momentum regression

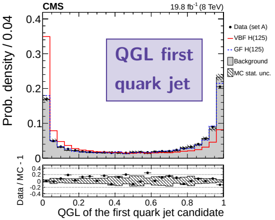
- jet properties are used to finetune CMS average jet calibration
- the re-calibration is done using a BDT
- an improved dijet invariant mass reconstruction is achieved



B. Quark gluon discrimination

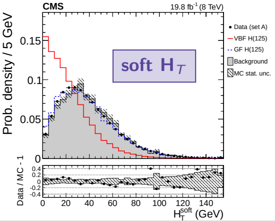
Tool developed by S. Bansal, T. Cornelis, P. Van Mechelen (UA)
Public document: CMS-PAS-JME-13-002

- **QCD background**:
  more jets originate from gluons than from quarks
- **VBF signal**: jets originate from quarks only
- jet composition and structure differences are exploited to make a discriminant



C. Additional hadronic activity

- VBF signal: electroweak production of jets
  → no QCD colour exchanged
- rapidity gap between VBF-tag jets
  → VBF signal is softer than QCD
- construct extra variables to exploit this property



4. Signal vs. Background discrimination

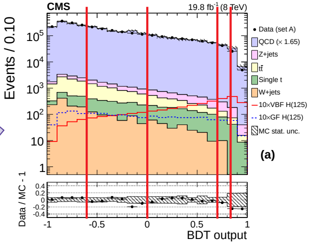
Defining the multivariate discriminant

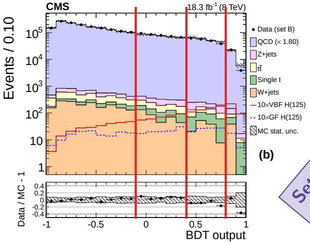
as orthogonal to m_{bb} as possible

The 12 input variables can be devided into five groups, and each group adds to the discriminating power:

- **VBF-topology**
  $\mathbf{m_{qq}}, |\Delta\eta_{qq}|, |\Delta\phi_{qq}|$
- **b-tagging**
  $\mathbf{CSV_0}, \mathbf{CSV_1}$
- **quark-gluon separation**
  $\mathbf{QGL_0}, \mathbf{QGL_1}, \mathbf{QGL_2}, \mathbf{QGL_3}$
- **soft activity**
  $\mathbf{H_T^{soft}}, \mathbf{N_2^{soft}}$
- **angles in CM frame**
  cos θ
  (4 jets CM frame: $\theta(qq, bb)$**)**

Set A
1+4 categories





Set B
1+3 categories

Events are divided in categories based on the value of the multivariate discriminant

- **-1 = background dominated ↔ +1 = signal dominated**
- category boundaries are optimized minimizing the expected limit and keeping in mind the need for sufficient MC statistics (required for templates later)
- the lowest category for both set A (BDT<-0.6) and set B (BDT<-0.1) is not used
- fits of the m_{bb} spectrum will be performed **simultaneously** in all categories

5. Fit for the H boson

Fit for the H boson: Strategy

Fit the $b\bar{b}$ invariant mass spectrum, separately for set A and B:

1. Derive the **QCD background template** from data in the signal depleted categories CAT1 and CAT5 (reference)
2. Take into account **residual correlation** between m_{bb} & the discriminant
   $\rightarrow$ derive data driven **transfer functions** from each category to the reference one
3. Extract **templates for other backgrounds**:
   - top (single top + ttjets)
   - Z+jets (including Z + W)
   - signal (including VBF + GF)
4. **Fit full model simultaneously in all categories (per set)**

$$f(m_{bb}) = N_{i,qcd} \cdot R_i(m_{bb}) \cdot Q(m_{bb}; \bar{p}) + N_{i,top} \cdot T_i(m_{bb}) + N_{i,Z} \cdot Z_i(m_{bb}) + \mu_{i,H} \cdot N_{i,H} \cdot H_i(m_{bb}; k_{JES}, k_{JER})$$

   - transfer functions R_i
   - floating normalizations N for top and Z (within 20% expected)
   - nuissance parameters k_{JES} and k_{JER}
   - signal strength $\mu_{i,H}$ for H

Systematic uncertainties

Source	VBF signal	GF signal
QCD shape parameters	determined by the fit	
QCD normalization	determined by the fit	
Top normalization	20%	
Z/W+jets normalization	20%	
JES (signal shape)	4%	
JER (signal shape)	10%	
Luminosity	2.6%	
JES (acceptance)	6 — 10%	4 — 12%
JER (acceptance)	1 — 4%	1 — 9%
B-jet tagging	3 — 9%	3 — 10%
Quark/gluon-jet tagging	1 — 3%	1 — 3%
Trigger	1 — 6%	5 — 20%
UE & PS	2 — 7%	10 — 45%
Process cross section	2.8%	10%
Scale variation	1 — 5%	xxxx
PDF	1.5 — 3%	3.5-5%

1a) **Leading uncertainty is the QCD background:**
$\rightarrow$ shape parameters + overall normalization: floating and determined by the fit

1b) **Other backgrounds:**
$\rightarrow$ Z/W+jets & top shapes taken from simulation + floating normalization within 20% of expectation

Systematic uncertainties

Source	VBF signal	GF signal
QCD shape parameters	determined by the fit	
QCD normalization	determined by the fit	
Top normalization	20%	
Z/W+jets normalization	20%	
JES (signal shape)	4%	
JER (signal shape)	10%	
Luminosity	2.6%	
JES (acceptance)	6 — 10%	4 — 12%
JER (acceptance)	1 — 4%	1 — 9%
B-jet tagging	3 — 9%	3 — 10%
Quark/gluon-jet tagging	1 — 3%	1 — 3%
Trigger	1 — 6%	5 — 20%
UE & PS	2 — 7%	10 — 45%
Process cross section	2.8%	10%
Scale variation	1 — 5%	xxxx
PDF	1.5 — 3%	3.5-5%

1a) **Leading uncertainty is the QCD background:**
$\rightarrow$ shape parameters + overall normalization: floating and determined by the fit

1b) **Other backgrounds:**
$\rightarrow$ Z/W+jets & top shapes taken from simulation + floating normalization within 20% of expectation

2) **JES and JER uncertainties affect the signal acceptance, and the m_{bb} & mvd shape**

Systematic uncertainties

Source	VBF signal	GF signal
QCD shape parameters	determined by the fit	
QCD normalization	determined by the fit	
Top normalization	20%	
Z/W+jets normalization	20%	
JES (signal shape)	4%	
JER (signal shape)	10%	
Luminosity	2.6%	
JES (acceptance)	6 — 10%	4 — 12%
JER (acceptance)	1 — 4%	1 — 9%
B-jet tagging	3 — 9%	3 — 10%
Quark/gluon-jet tagging	1 — 3%	1 — 3%
Trigger	1 — 6%	5 — 20%
UE & PS	2 — 7%	10 — 45%
Process cross section	2.8%	10%
Scale variation	1 — 5%	xxxx
PDF	1.5 — 3%	3.5-5%

1a) **Leading uncertainty is the QCD background:**
→ shape parameters + overall normalization:
floating and determined by the fit

1b) **Other backgrounds:**
→ Z/W+jets & top shapes taken from simulation
+ floating normalization within 20% of expectation

2) **JES and JER uncertainties affect the signal acceptance, and the m_{bb} & mvd shape**

3) **Flavour tagging uncertainties:**

- **CSV discriminant: affects event selection & mvd**
- **QGL discriminant: only affects mvd shape**

Systematic uncertainties

Source	VBF signal	GF signal
QCD shape parameters	determined by the fit	
QCD normalization	determined by the fit	
Top normalization	20%	
Z/W+jets normalization	20%	
JES (signal shape)	4%	
JER (signal shape)	10%	
Luminosity	2.6%	
JES (acceptance)	6 — 10%	4 — 12%
JER (acceptance)	1 — 4%	1 — 9%
B-jet tagging	3 — 9%	3 — 10%
Quark/gluon-jet tagging	1 — 3%	1 — 3%
Trigger	1 — 6%	5 — 20%
UE & PS	2 — 7%	10 — 45%
Process cross section	2.8%	10%
Scale variation	1 — 5%	xxxx
PDF	1.5 — 3%	3.5-5%

1a) **Leading uncertainty is the QCD background:**
→ shape parameters + overall normalization: floating and determined by the fit

1b) **Other backgrounds:**
→ Z/W+jets & top shapes taken from simulation + floating normalization within 20% of expectation

2) **JES and JER uncertainties affect the signal acceptance, and the m_{bb} & mvd shape**

3) **Flavour tagging uncertainties:**
- **CSV discriminant: affects event selection & mvd**
- **QGL discriminant: only affects mvd shape**

4) **The trigger uncertainty is propagated from the data/MC 2D scale factor:**
→ signal distribution × scale factor map convolution

Systematic uncertainties

Source	VBF signal	GF signal
QCD shape parameters	determined by the fit	
QCD normalization	determined by the fit	
Top normalization	20%	
Z/W+jets normalization	20%	
JES (signal shape)	4%	
JER (signal shape)	10%	
Luminosity	2.6%	
JES (acceptance)	6 — 10%	4 — 12%
JER (acceptance)	1 — 4%	1 — 9%
B-jet tagging	3 — 9%	3 — 10%
Quark/gluon-jet tagging	1 — 3%	1 — 3%
Trigger	1 — 6%	5 — 20%
UE & PS	2 — 7%	10 — 45%
Process cross section	2.8%	10%
Scale variation	1 — 5%	xxxx
PDF	1.5 — 3%	3.5-5%

1a) **Leading uncertainty is the QCD background:**
→ shape parameters + overall normalization: floating and determined by the fit

1b) **Other backgrounds:**
→ Z/W+jets & top shapes taken from simulation
+ floating normalization within 20% of expectation

2) **JES and JER uncertainties affect the signal acceptance, and the m_{bb} & mvd shape**

3) **Flavour tagging uncertainties:**
- **CSV discriminant: affects event selection & mvd**
- **QGL discriminant: only affects mvd shape**

4) **The trigger uncertainty is propagated from the data/MC 2D scale factor:**
→ signal distribution × scale factor map convolution

5) **Theoretical uncertainties:**
- **The uncertainty on the integrated luminosity is 2.6%**
- **Cross section uncertainties are taken from the Yellow Report**
- **PDF uncertainties are calculated following the PDF4LHC prescription**
- **Renormalization and factorization scale variations are studied with Powheg**
- **UE & PS variations are studied with Pythia 8**

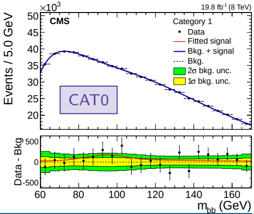
Fit for the H boson: Validate on the Z

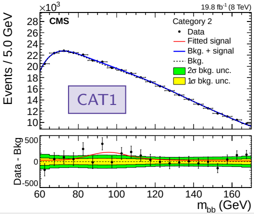
Validate the fit strategy by fitting the known Z resonance:

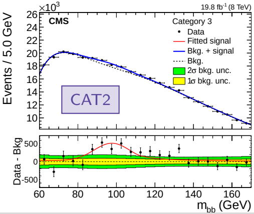
- The Z search uses the same techniques as the H search:
  - The event selection follows the set A selection + 1x medium CSV b-tag
  - A Fisher discriminant is used, again orthogonal to m_{bb}, and 3 categories are defined
  - The QCD background is derived from the signal depleted category, and linear transfer functions are constructed for the other categories
  - Top and Z templates are derived
  - The matching version of the fit model is fit simultaneously to all categories

$$f(m_{bb}) = N_{i,qcd} \cdot R_i(m_{bb}) \cdot Q(m_{bb}; \bar{p}) + N_{i,top} \cdot T_i(m_{bb}) + \mu_{i,Z} \cdot N_{i,Z} \cdot Z_i(m_{bb}; k_{JES}, k_{JER})$$

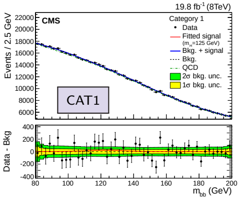
- Results:
  - Fitted signal stength: $\mu_\mathbf{Z} = \sigma/\sigma_\mathbf{SM} = \mathbf{1.28^{+0.50}_{-0.34}}$
  - Observed significance: $\mathbf{3.8}\sigma$ ($\mathbf{3.2}\sigma$ expected)

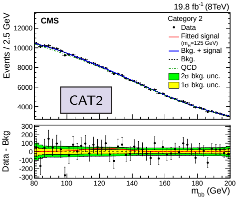


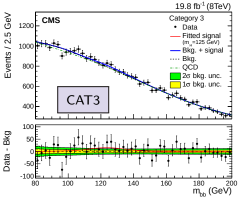


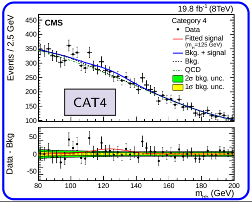


Fit for the H boson: Results







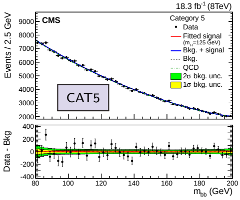


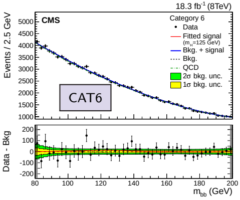
A binned maximum-likelihood fit is made per set, simultaneous in all respective categories:

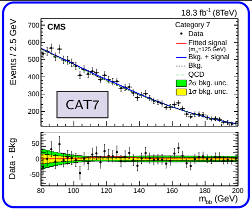
- CAT1-4 for set A
- CAT5-7 for set B
- highest CAT per set = most sensitive

Sensitivity:

- is highest in CAT3 and CAT4
- is worse in set B, due to:
  - worse m_{bb} resolution
  - larger transfer function uncertainties

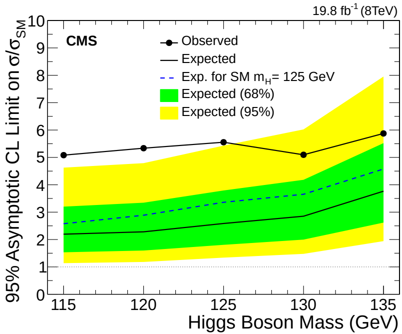






6. Limit, Combination & Summary

Expected limit: VBF H $\to b\bar{b}$



		Expected	Observed
Upper limits (95% CL)	range [115,135]	2.2–3.7	5.0–5.8
	@ 125 GeV	2.5	5.5
Signal significance	@ 125 GeV	0.8	2.2

$\to$ **the observed signal strength is $\mu = 2.8^{+1.6}_{-1.4}$, which is compatible to $\mu_{SM} = 1$ at the 8% level**

Combination: $H \to b\bar{b}$

Three CMS $H \to b\bar{b}$ channels were combined:

- VH $H \to b\bar{b}$ — doi:PhysRevD.89.012003[link]
- $t\bar{t}H$ $H \to b\bar{b}$ — doi:JHEP05(2013)145[link], doi:JHEP09(2014)087[link]
- VBF $H \to b\bar{b}$ — arXiv:1506.01010[link]

The combined signal strength is $\mu = 1.03^{+0.44}_{-0.42}$ with a significance of 2.6σ

$H \to b\bar{b}$ Channel	Best fit (68% CL) Observed	Upper limits (95% CL) Observed	Upper limits (95% CL) Expected	Signal significance Observed	Signal significance Expected
VH	0.89 ± 0.43	1.68	0.85	2.08	2.52
$t\bar{t}H$	0.7 ± 1.8	4.1	3.5	0.37	0.58
VBF	$2.8^{+1.6}_{-1.4}$	5.5	2.5	2.2	0.83
Combined	$1.03^{+0.44}_{-0.42}$	1.77	0.78	2.56	2.7

Summary

- **We have completed the VBF $Hb\bar{b}$ analysis:**
  - event selection and interpretation in two phase spaces and with improved b-jet identification
  - multivariate event discrimination orthogonal to m_{bb} and event categorization by mvd output
  - search for the SM Z boson as a cross check:
    $\rightarrow$ signal strength $\mu = 1.28^{+0.50}_{-0.34}$ and significance 3.8σ
  - search for the SM H boson following the same procedure, fitting the m_{bb} spectrum simultaneously in all categories
    $\rightarrow$ signal strength $\mu = 2.8^{+1.6}_{-1.4}$ and significance 2.2σ
- **The three CMS $H \rightarrow b\bar{b}$ results were combined**
  $\rightarrow$ signal strength $\mu = 1.03^{+0.44}_{-0.42}$ and a significance of 2.6σ
- **Our paper has been submitted to PRD (arXiv:1506.01010[link])**