Scalar Search and Study in Belgium

VBF $H \rightarrow b\bar{b}$

P. Azzurri, J. Bernardini, D. Caiulo, P. Spagnolo, C. Vernieri
INFN and Scuola Normale, Pisa

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

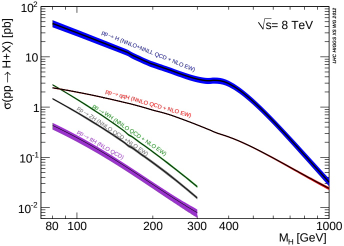
S. de Visscher, K. Kousouris
CERN

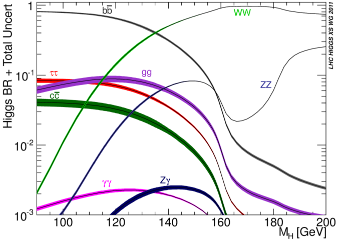
January 23-24th, 2014 - IIHE, Brussels, Belgium

Outline

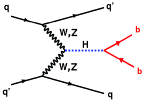
1. Introduction
2. Nominal analysis
   - overview
   - ARC/CWR comments
3. Round two analysis
   - parked data
   - new elements
4. Summary and Outlook

Introduction





- Properties of the VBF H $\rightarrow b\bar{b}$ channel:
  - cross section significantly larger than for VH or ttH production
  - very large QCD background
  - trigger challenges
- 4-jet signal topology:

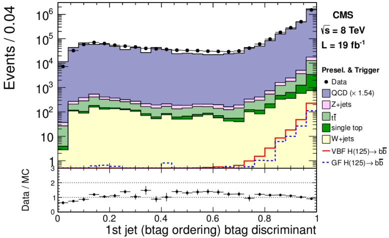


- **Search strategy**:
  - topological trigger on the signal main properties *(jets with large $\Delta\eta$, two b-jets, etc.)*
  - use multivariate methods to exploit maximally the (significant) differences between signal and QCD *(maintain the orthogonality to the mbb)*
  - perform a fit of the $m_{b\bar{b}}$ spectrum

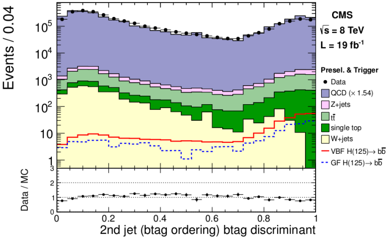
Nominal analysis

Nominal analysis HIG-13-011

- **Nominal analysis performed on full 2012 dataset (1)**
  - Several **dedicated triggers** (L1 + HLT, different jet-p_T thresholds) were used
  - **Event interpretation** is based on requiring 4-jet events with a good primary vertex and additional pile-up and jet IDs
  - **Event reconstruction** uses particle-flow algorithms and R=0.5 anti-k_T jet clustering, and identification criteria are applied to the jets (against fake jets and pile-up contamination)
  - **B-jet identification** is done using the CSV b-tagger, both at trigger level and offline



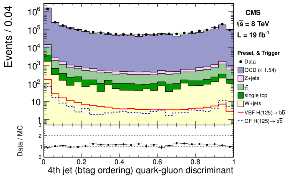
(1) highest CSV-tag jet: b1



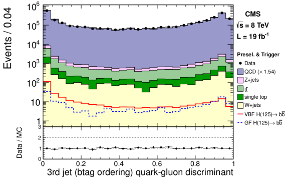
(2) 2nd highest CSV-tag jet: b2

Nominal analysis HIG-13-011

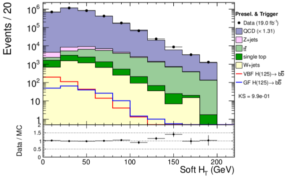
- **Nominal analysis performed on full 2012 dataset (2)**
  - Several **dedicated triggers** (L1 + HLT, different jet-p_T thresholds) were used
  - **Event interpretation** is based on requiring 4-jet events with a good primary vertex and additional pile-up and jet IDs
  - **Event reconstruction** uses particle-flow algorithms and R=0.5 anti-k_T jet clustering, and identification criteria are applied to the jets (against fake jets and pile-up contamination)
  - **B-jet identification** is done using the CSV b-tagger, both at trigger level and offline
  - The event selection is improved using **quark-gluon discrimination** (to determine if the final state light quarks originate from light quark hadronization (signal) or gluons (bkg)) and looking at **additional hadronic activity** between the VBF tagging jets ($q\bar{q}$) (other than that of the central H decay products ($b\bar{b}$))



(1) lowest CSV-tag jet: q1



(2) 2nd lowest CSV-tag jet: q2

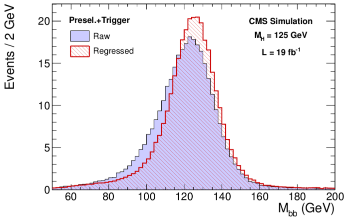


(3) soft Ht [GeV]

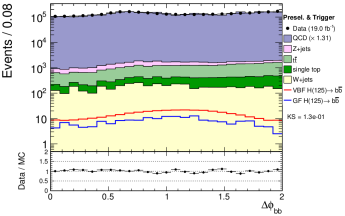
Nominal analysis HIG-13-011

Nominal analysis performed on full 2012 dataset (3)

- **B-jet identification** is done using the CSV b-tagger, both at trigger level and offline
- The event selection is improved using **quark-gluon discrimination**
  (to determine if the final state light quarks originate from light quark hadronization (signal) or gluons (bkg))
  and looking at **additional hadronic activity** between the VBF tagging jets ($q\bar{q}$)
  (other than that of the central H decay products ($b\bar{b}$))
- The $b\bar{b}$ mass resolution is improved by applying **jet energy corrections** on top of the CMS standard ones (determined using regression techniques similar to those used in the VH $H \to b\bar{b}$ analysis)
- The **offline event selection** uses cuts based on the trigger logic, with an extra $\Delta\phi_{b\bar{b}} < 2$ cut to exclude QCD events with back-to-back $b\bar{b}$ pairs



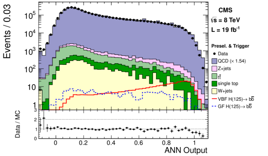
(1) raw & regressed $m_{b\bar{b}}$ invariant mass



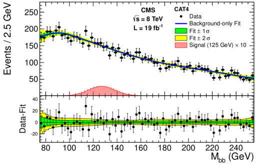
(2) $b\bar{b}$-pair $\Delta\phi$

Nominal analysis HIG-13-011

- **Nominal analysis performed on full 2012 dataset (2)**
  - **Multivariate techniques (ANN)** are employed to maximally separate signal and background
  - Since the final search uses a data-driven fit of the $m_{b\bar{b}}$ spectrum, only variables **orthogonal to $m_{b\bar{b}}$** are used in the construction of the multivariate discriminant
  - The events are split up into five **categories, based on the ANN reponse (1)**; the search is then conducted in the highest four
  - A **fit of the $m_{b\bar{b}}$ spectrum (2)** is performed in each category, using a 3 and 4-part background model (QCD: Bernstein, Z/W,T: crystal ball, from simulation, (signal))
  - **Systematic uncertainties** are attributed to trigger efficiencies, elements affecting the signal acceptance, elements affecting the Z-template and uncertainties on the integrated luminosity and the process cross sections
  - **Limits** are computed with the asymptotic CLs method



(1) ANN distribution after offline preselection



(2) Fit to the $m_{b\bar{b}}$ distribution in CAT4

Nominal analysis: ARC/CWR comments (1)

$$\mathcal{N}_{\rm QCD} \cdot \mathcal{B}_{\rm QCD}\left(m_{b\bar{b}}\right) + \mathcal{N}_{\rm Z} \cdot \mathcal{Z}\left(m_{b\bar{b}}\right) + \mathcal{N}_{\rm top} \cdot \mathcal{T}\left(m_{b\bar{b}}\right)$$

$$\mathcal{N}_{\rm sig} \cdot \mathcal{CB}\left(m_{b\bar{b}}\right) + \left(1 - \mathcal{N}_{\rm sig}\right) \cdot \mathcal{B}_3\left(m_{b\bar{b}}\right)$$

- **Extension of the $m_{b\bar{b}}$ fit bias studies**
  - Bias studied in CAT4 only
  - Fit function: 5th order Bernstein polynomial
  - Alternative models: exp. power law, tanh and modified Gaussian
  - Fit range 70-250 GeV

  → **bias $\sim$ 10% and $<$ 30% in CAT4**

Nominal analysis: ARC/CWR comments (1)

$$\mathcal{N}_{\text{QCD}} \cdot \mathcal{B}_{\text{QCD}}\left(m_{b\bar{b}}\right) + \mathcal{N}_{Z} \cdot \mathcal{Z}\left(m_{b\bar{b}}\right) + \mathcal{N}_{\text{top}} \cdot \mathcal{T}\left(m_{b\bar{b}}\right)$$

$$\mathcal{N}_{\text{sig}} \cdot \mathcal{CB}\left(m_{b\bar{b}}\right) + \left(1 - \mathcal{N}_{\text{sig}}\right) \cdot \mathcal{B}_3\left(m_{b\bar{b}}\right)$$

- **Extension of the $m_{b\bar{b}}$ fit bias studies**
  - Bias studied in ~~CAT4 only~~
    → *extended to all categories*
  - Fit function: 5th order Bernstein polynomial
  - Alternative models: exp. power law, tanh and modified Gaussian
  - Fit range 70-250 GeV

Nominal analysis: ARC/CWR comments (1)

$\mathcal{N}_{QCD} \cdot \mathcal{B}_{QCD}\left(m_{b\bar{b}}\right) + \mathcal{N}_Z \cdot \mathcal{Z}\left(m_{b\bar{b}}\right) + \mathcal{N}_{top} \cdot \mathcal{T}\left(m_{b\bar{b}}\right)$

$\mathcal{N}_{sig} \cdot \mathcal{CB}\left(m_{b\bar{b}}\right) + \left(1 - \mathcal{N}_{sig}\right) \cdot \mathcal{B}_3\left(m_{b\bar{b}}\right)$

- **Extension of the $m_{b\bar{b}}$ fit bias studies**
  - Bias studied in ~~CAT4 only~~
    $\rightarrow$ *extended to all categories*
  - Fit function: ~~5th order~~ Bernstein polynomial
    $\rightarrow$ *bias is problematic especially in CAT1*
    $\rightarrow$ *increasing the order to 6 yields acceptable results*
  - Alternative models: exp. power law, tanh and modified Gaussian
  - Fit range 70-250 GeV

Nominal analysis: ARC/CWR comments (1)

$$\mathcal{N}_{\mathrm{QCD}} \cdot \mathcal{B}_{\mathrm{QCD}}\left(\mathrm{m}_{\mathrm{b\bar{b}}}\right) + \mathcal{N}_{\mathrm{Z}} \cdot \mathcal{Z}\left(\mathrm{m}_{\mathrm{b\bar{b}}}\right) + \mathcal{N}_{\mathrm{top}} \cdot \mathcal{T}\left(\mathrm{m}_{\mathrm{b\bar{b}}}\right)$$

$$\mathcal{N}_{\mathrm{sig}} \cdot \mathcal{CB}\left(\mathrm{m}_{\mathrm{b\bar{b}}}\right) + \left(1 - \mathcal{N}_{\mathrm{sig}}\right) \cdot \mathcal{B}_3\left(\mathrm{m}_{\mathrm{b\bar{b}}}\right)$$

- **Extension of the $m_{b\bar{b}}$ fit bias studies**
  - Bias studied in ~~CAT4 only~~
    → *extended to all categories*
  - Fit function: ~~5th order~~ Bernstein polynomial
    → *bias is problematic especially in CAT1*
    → *increasing the order to 6 yields acceptable results*
  - Alternative models: exp. power law, tanh and modified Gaussian
    → *functions extended with additional parameter*
  - Fit range 70-250 GeV

Nominal analysis: ARC/CWR comments (1)

$\mathcal{N}_{QCD} \cdot \mathcal{B}_{QCD}\left(m_{b\bar{b}}\right) + \mathcal{N}_{Z} \cdot \mathcal{Z}\left(m_{b\bar{b}}\right) + \mathcal{N}_{top} \cdot \mathcal{T}\left(m_{b\bar{b}}\right)$

$\mathcal{N}_{sig} \cdot \mathcal{CB}\left(m_{b\bar{b}}\right) + \left(1 - \mathcal{N}_{sig}\right) \cdot \mathcal{B}_3\left(m_{b\bar{b}}\right)$

- **Extension of the $m_{b\bar{b}}$ fit bias studies**
  - Bias studied in ~~CAT4 only~~
    → *extended to all categories*
  - Fit function: ~~5th order~~ Bernstein polynomial
    → *bias is problematic especially in CAT1*
    → *increasing the order to 6 yields acceptable results*
  - Alternative models: exp. power law, tanh and modified Gaussian
    → *functions extended with additional parameter*
  - Fit range ~~70-250 GeV~~
    → *range optimized to 90-255 GeV*

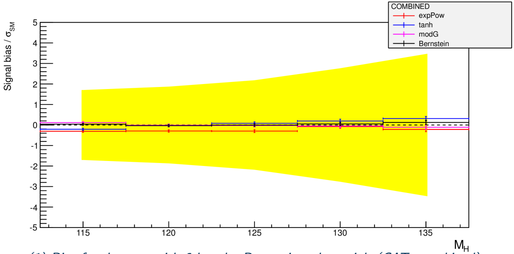
Nominal analysis: ARC/CWR comments (1)

$$\mathcal{N}_{QCD} \cdot \mathcal{B}_{QCD}\left(m_{b\bar{b}}\right) + \mathcal{N}_{Z} \cdot \mathcal{Z}\left(m_{b\bar{b}}\right) + \mathcal{N}_{top} \cdot \mathcal{T}\left(m_{b\bar{b}}\right)$$

$$\mathcal{N}_{sig} \cdot \mathcal{CB}\left(m_{b\bar{b}}\right) + \left(1 - \mathcal{N}_{sig}\right) \cdot \mathcal{B}_3\left(m_{b\bar{b}}\right)$$

- **Extension of the $m_{b\bar{b}}$ fit bias studies**
  - Bias studied in ~~CAT4 only~~
    $\rightarrow$ *extended to all categories*
  - Fit function: ~~5th order~~ Bernstein polynomial
    $\rightarrow$ *bias is problematic especially in CAT1*
    $\rightarrow$ *increasing the order to 6 yields acceptable results*
  - Alternative models: exp. power law, tanh and modified Gaussian
    $\rightarrow$ *functions extended with additional parameter*
  - Fit range ~~70-250 GeV~~
    $\rightarrow$ *range optimized to 90-255 GeV*

$\rightarrow$ **bias $<$ 20% in all CATs and new study with revised pT cuts and lower turnons ongoing**



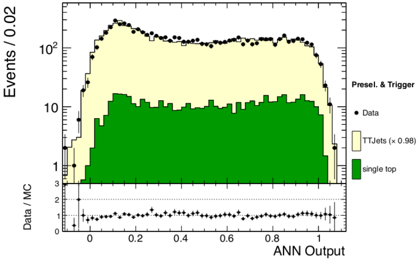
(1) Bias for the case with 6th order Bernstein polynomials (CATs combined)

Nominal analysis: ARC/CWR comments (2)

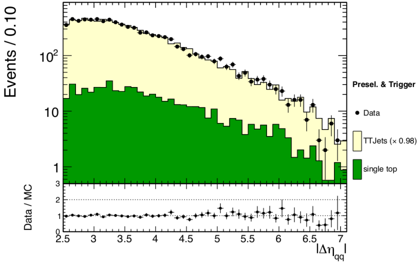
- **Bug fix for setup of background model in *combine*** (ref: H working meeting 29 Nov '13)
  - $\rightarrow$ **insert 3 separate bkg sources instead of 1, and leave combine to build the extended PDF**
- **Data/MC discrepancies in the ANN tail** (ref: H working meeting 29 Nov '13)

Nominal analysis: ARC/CWR comments (2)

- **Bug fix for setup of background model in *combine*** (ref: H working meeting 29 Nov '13)
  - → **insert 3 separate bkg sources instead of 1, and leave combine to build the extended PDF**
- **Data/MC discrepancies in the ANN tail** (ref: H working meeting 29 Nov '13)
  - Not attributable to low stats in 100-250 QCD slice
  - Not attributable to variable correlations
  - Studied extra ttbar control region (QCD free): ANN proves reliable
  - → **final treatment: systematic uncertainty anti-correlating the signal yields in CAT3 & 4**



(1) ANN output in ttbar control region

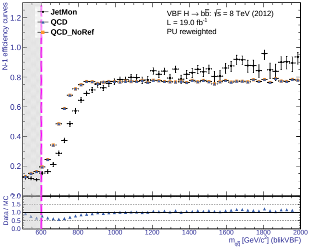


(2) $q\bar{q}$-pair $\Delta\eta$ in ttbar control region

Round two analysis

Round two analysis: parked data

- Data streams
  - 2012A: No VBFParked
  - 2012B: /VBF1Parked/Run2012B-22Jan2013-v1/AOD
  - 2012C: /VBF1Parked/Run2012C-22Jan2013-v1/AOD
  - 2012D: /VBF1Parked/Run2012D-22Jan2013-v1/AOD
- Included triggers:
  - HLT_DiJet35_MJJ650_AllJets_DEta3p5_VBF (L1: L1_HTT150, L1_HTT175, L1_HTT200, L1_ETM40)
  - HLT_DiJet35_MJJ700_AllJets_DEta3p5_VBF (L1: L1_HTT175, L1_HTT200, L1_ETM40)
  - HLT_DiJet35_MJJ750_AllJets_DEta3p5_VBF (L1: L1_HTT175, L1_HTT200, L1_ETM40)
- Offline preselection studied following the trigger logic:
  - $\Delta\eta_{q\bar{q}} > 3.5$
  - $m_{q\bar{q}} > 600$
  - jetPt[1] > 35 GeV
  - $\Delta\phi_{b\bar{b}} < 2.0$
- Parked data amounts to 18.2 fb^{-1}
- Possible reference triggers for efficiency study
  - HLT_DiPFJetAve40
  - HLT_DiPFJetAve80

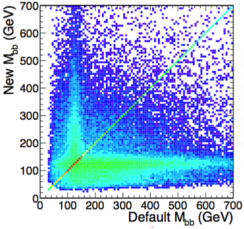


(1) trigger efficiency curve for $q\bar{q}$ -pair invariant mass

Round two analysis: new elements (1)

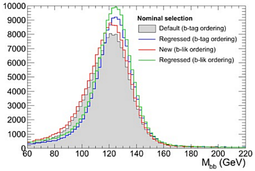
B-jet discrimination ∼ b-tag likelihood (1)

- Nominal analysis event interpretation:
  - select 4 leading jets in p_T
  - order in CSV b-tag:
    - 2 leading ones (b-jets)
    - 2 trailing ones (q-jets)
  - *problem:* sometimes the b-tag ordering is incorrect



(1) mbb scatter plot

- Round two analysis event interpretation:
  - select 4 leading jets in p_T
  - order in CSV b-tag **and** η
  - build a **b-likelihood BDT** based on btagIdx, etaIdx, btag and eta
  - order in b-likelihood: again 2 leading (=b) and 2 trailing (=q)



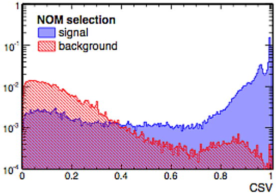
(2) mbb peak

- **Improvement**:
  - coherently use b-tag and eta ordering of the candidates
  - regain some in-peak contribution to $m_{b\bar{b}}$, other than from the regression
  - increased significance

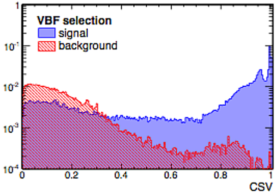
Round two analysis: new elements (1)

B-jet discrimination ∼ b-tag likelihood (2)

- BDT training on VBF@125 sample, separately in nominal and parked data phase space



(1) result with CSV for nominal phase space

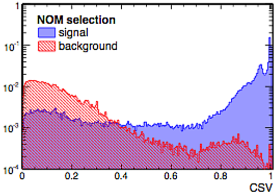


(2) result with CSV for parked data phase space

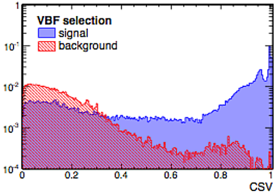
Round two analysis: new elements (1)

B-jet discrimination ∼ b-tag likelihood (2)

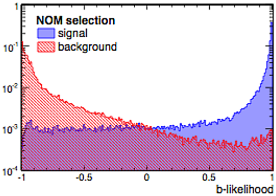
- BDT training on VBF@125 sample, separately in nominal and parked data phase space



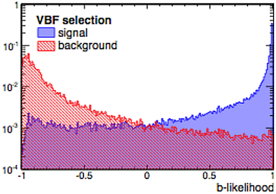
(1) result with CSV for nominal phase space



(2) result with CSV for parked data phase space



(3) result with b-lik for nominal analysis

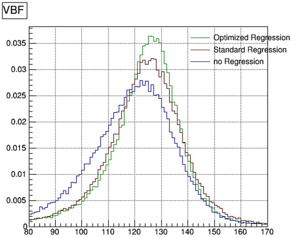


(4) result with b-lik for parked data phase space

Round two analysis: new elements (2)

Optimization of the b-jet energy regression

- Use case:
  - b-jet resolution is suboptimal when compared to light-quark/gluon-jet resolution
  - use regression techniques to derive a **correction factor per b-jet** and in one go improve the $b\bar{b}$ invariant mass resolution
- Recent steps taken:
  - **Comparison** with the regression from the **VH analysis** (ref: AN-13-069)
    → *VH uses also SoftLepton information*
  - Addition of input variables
  - Training optimized according to the two **different phase space regions** covered by the VBF analysis
- **Result**: a sizeable improvement in $b\bar{b}$ invariant mass resolution is achieved
  - around 5-15% when compared to the result using the standard regression
- **To do**: evaluate the effect on the sensitivity



(1) optimization of the regression in the VBF phase space

Summary & Outlook

- Some extra issues concerning the first full iteration of the analysis as presented in the summer of 2013 were addressed
- The parked data, amounting to $\sim 18\ \mathrm{fb}^{-1}$ is now being studied
- New idea added to the analysis: b-likelihood
- Elements currently being looked at or to follow soon include:
  (all to be considered in both the nominal and the parked data vbf phase space)
  - optimization of the preselection cuts, following the trigger logic
  - evaluation of the trigger efficiency
  - retraining the categorization
  - redoing the fit of the $b\bar{b}$ invariant mass spectrum
  - repeating the bias studies