Search for associated $t\bar{t}H$ production in the $H \rightarrow b\bar{b}$ decay channel at CMS using the Matrix Element Method

Daniel Salerno

Reference

CMS PAS: HIG-14-010

2014 PhD seminar

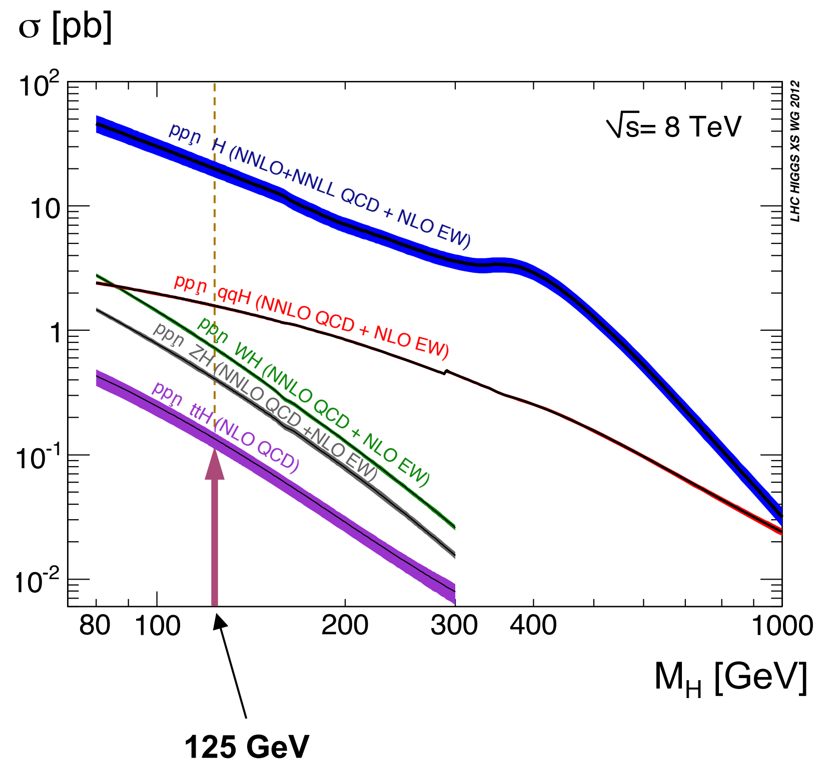
11 September 2014, Universität Zürich

Standard model ttH production

Motivation

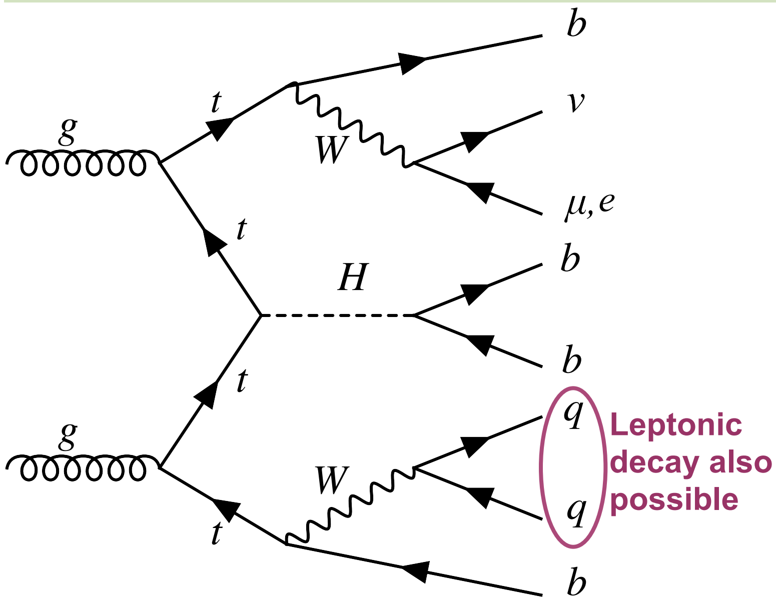
- Higgs boson with 125 GeV mass discovered by CMS and ATLAS
  - Focus now on studying its properties
- ttH provides a direct probe of the Higgs/top Yukawa coupling y_t
  - Most important fermion coupling
  - Only one with $y_t \sim 1$
  - Provides insight to possible new physics
- This search is at CMS
  - Multipurpose detector at the LHC

Production cross section at LHC



ttH (H→bb) channel

Feynman diagram



- 4 b jets (2 from H, 1 from each top)
- 2 (0) light flavour jets (from W)
- 1 (2) leptons – μ or *e* (from W)
- Missing energy (from ν)

Characteristics

- H→bb has largest BR (≈58%)
  - Fully reconstructed final state
- Leptonic final state
  - Greatly reduced background

Background processes

tt + jets
Dominant background

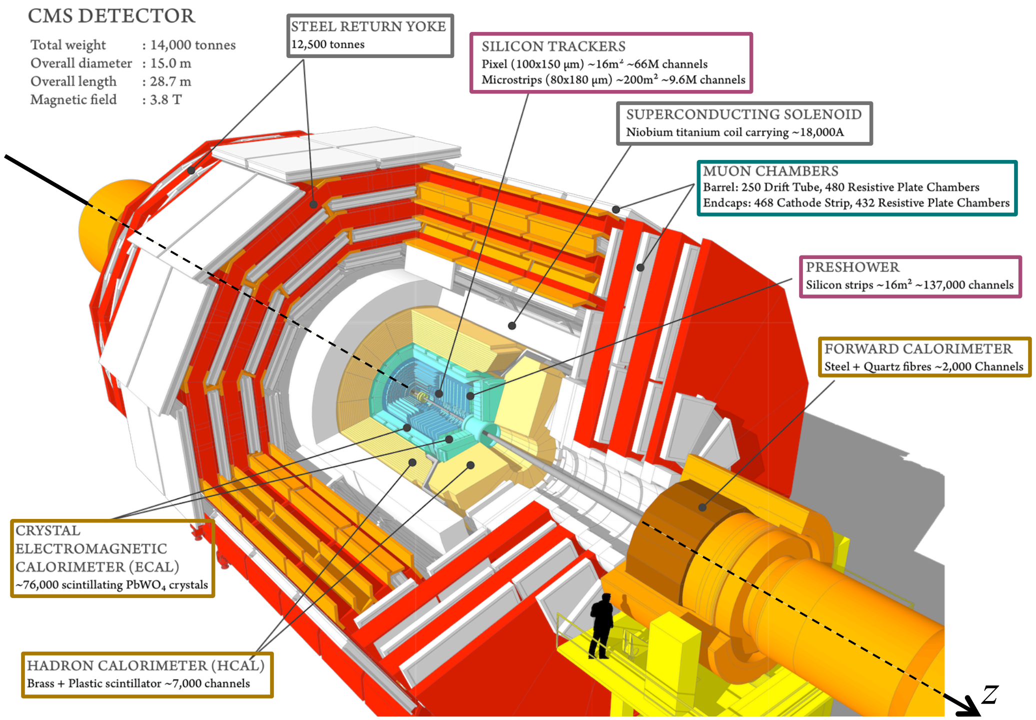
- $t\bar{t}$ + lf
- $t\bar{t}$ + cc
- $t\bar{t}$ + b
- $t\bar{t}$ + bb ← **Irreducible**

Minor backgrounds

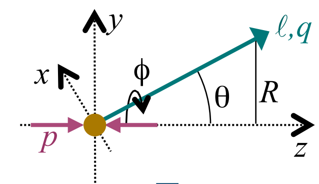
- EWK
- Single t
- $t\bar{t}$ + W,Z
- VV ← **Eliminated by final selection**

The CMS detector

- Located at the LHC – a proton-proton collider
  - Centre-of-mass energy of 8 TeV in 2012



Coordinate system

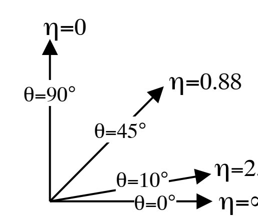


Useful variables

E_T and p_T defined in the x-y plane

Pseudorapidity:

$\eta = -\ln(\tan(\theta/2))$



$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$$

Inner detector (ID) | Calorimeters | Solenoid magnet | Muon detectors

Data and preselection

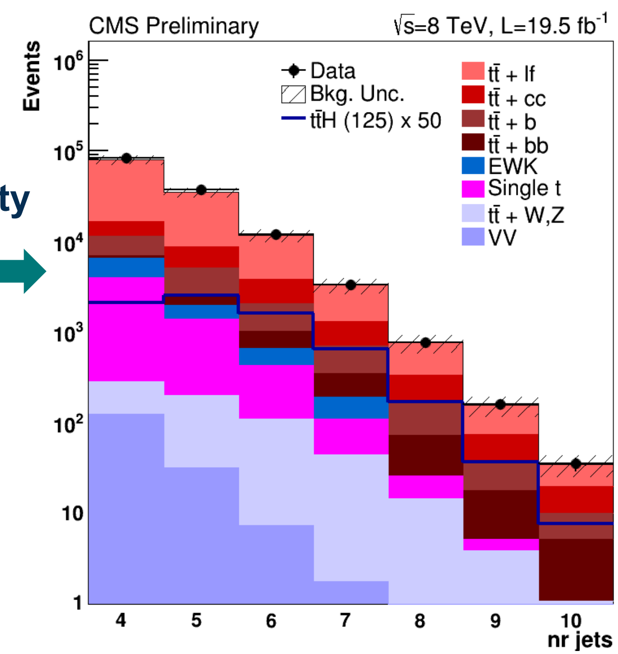
Data

- **19.5 fb^{-1}:** 8 TeV 2012 data sample
- Single-electron trigger: isolated, $p_T > 27$ GeV (e)
- Single-muon trigger: isolated, $p_T > 24$ GeV (μ, $\mu\mu$, μe)
- Double-electron trigger: isolated, $p_T > 17, 8$ GeV (ee)

Preselection

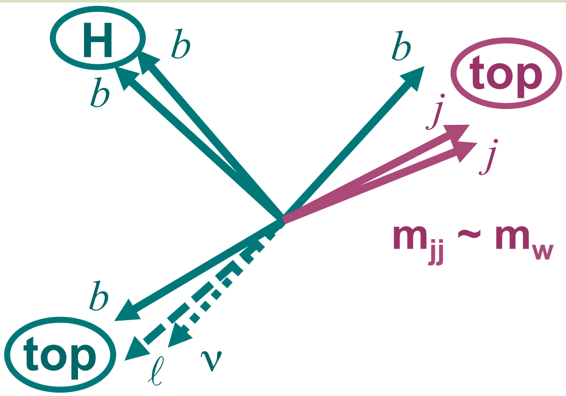
- Jets
  - $p_T > 30$ GeV, $|\eta| < 2.5$
  - 2 b-tagged jets
- Single lepton (SL)
  - $p_T > 30$ GeV, $|\eta| < 2.5$ (e), $|\eta| < 2.1$ (μ)
- Double lepton (DL)
  - $p_T > 20$ GeV, $|\eta| < 2.5$ (e), $|\eta| < 2.4$ (μ)

Jet multiplicity
SL events

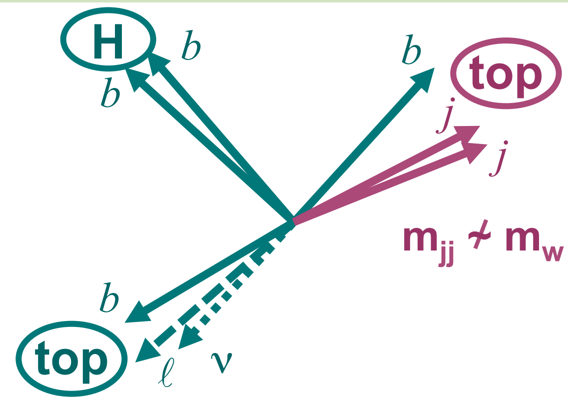


4 event categories

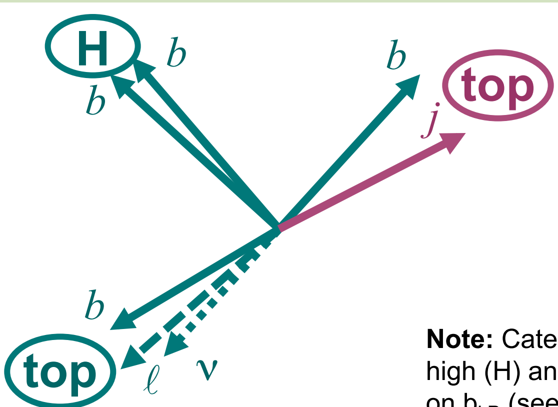
SL – Category 1: ≥6j, 4b, 1ℓ



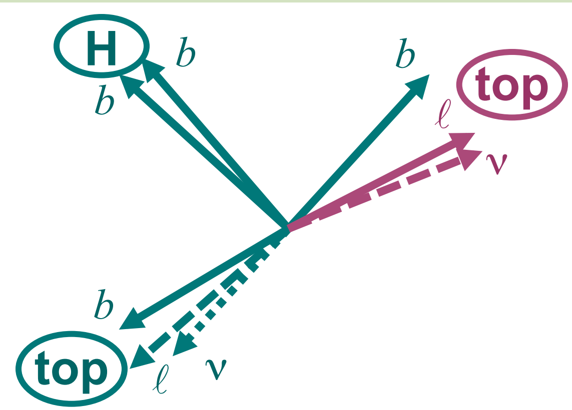
SL – Category 2: ≥6j, 4b, 1ℓ



SL – Category 3: 5j, 4b, 1ℓ



Double Lepton: ≥4j, 4b, 2ℓ

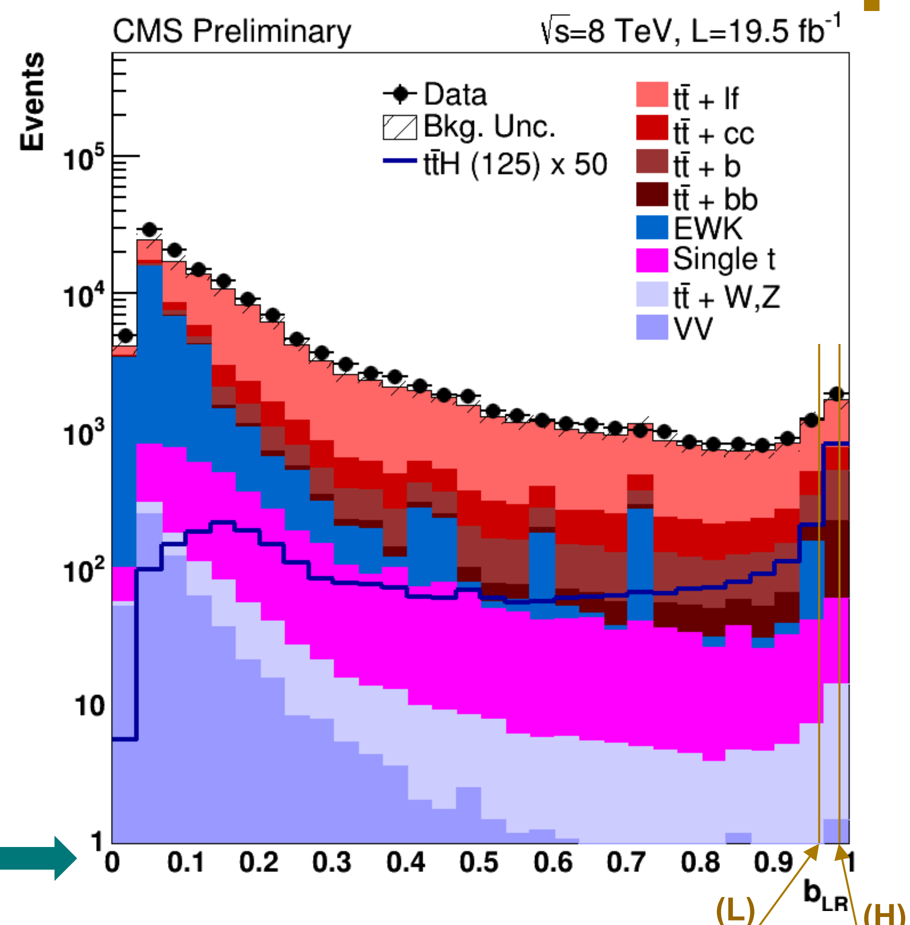


Note: Categories further split into high (H) and low (L) purity based on b_{LR} (see slide 6)

b-tag likelihood ratio

- Events further selected based on a b-tag likelihood ratio discriminant b_{LR}
  - For each jet, b-tagging algorithm combines information from track IP and secondary vertex: CSV parameter (ζ)
  - $\zeta_1, \ldots, \zeta_{njets}$ used in a likelihood function for 4 b- and 2 b-quark hypotheses
  - $$b_{\mathrm{LR}} = \frac{\mathcal{L}_{bbbb}(\zeta_1, \ldots, \zeta_n)}{\mathcal{L}_{bbbb}(\zeta_1, \ldots, \zeta_n) + \mathcal{L}_{bbqq}(\zeta_1, \ldots, \zeta_n)}$$
  - A cut on b_{LR} is made in each category to define high (H) and low (L) purity subcategories

b_{LR} discriminant
- SL events
- 5 jets

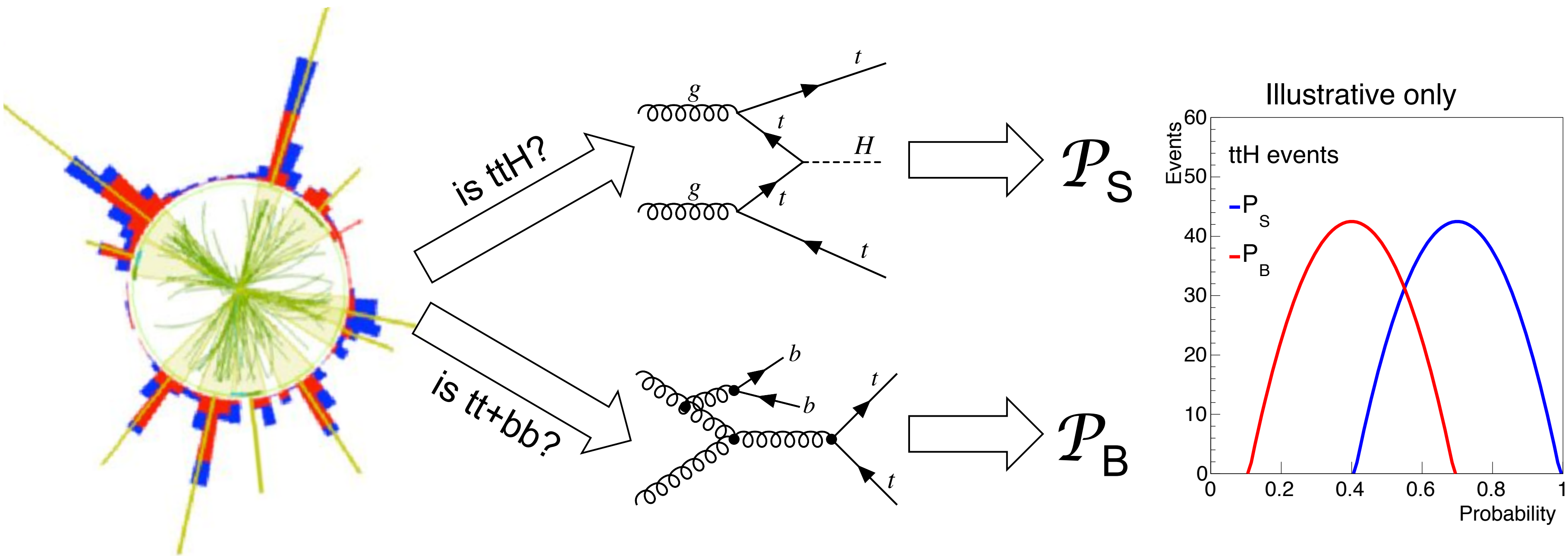


Cuts for this category: 0.96 0.995

The Matrix Element Method

Overview

- Provides optimal separation of signal and background
- Reduces combinatorial self-background
- Calculates the probability of an event being signal/background



The MEM event probabilities

The Variables

- Measured kinematical variables (**y**) used as input
  - Lepton energy and direction is assumed to be perfectly measured
  - Jet direction is assumed to be perfectly measured
  - Integration over poorly measured variables (E_{jet}, p_ν)

The Formula

- Sum over all possible permutations of jet–quark matching

$$w_i(\mathbf{y}) = \frac{1}{\sigma_i}\sum_{\text{perm}}\int_\Omega d\mathbf{x}\int dx_a dx_b \Phi(x_a, x_b)\delta^4\{(x_a P_a + x_b P_b) - \sum p(\mathbf{x})\}|\mathcal{M}_i(\mathbf{x})|^2 W(\mathbf{y}|\mathbf{x})$$

  - Ω = phase space volume of final particles **x**, $x_{a,b}$ = parton momentum fraction
  - Φ = parton flux factor, $\mathcal{M}_i$ = scattering amplitude of process i (i = ttH, tt+bb)
  - W = transfer function: probability of measuring **y** given **x**

The Prob-abilities

- 3 different probabilities are determined
  - $\mathcal{P}_S(\mathbf{y}) = w_S(\mathbf{y})\mathcal{L}_{bbbb}(\zeta)$
  - $\mathcal{P}_{B1}(\mathbf{y}) = w_B(\mathbf{y})\mathcal{L}_{bbbb}(\zeta)$
  - $\mathcal{P}_{B2}(\mathbf{y}) = w_B(\mathbf{y})\mathcal{L}_{bbqq}(\zeta)$
    - Where $\mathcal{L}_{bbqq}(\vec{\zeta}) = \sum_i P(\zeta_1, ..., \zeta_6|\{bbqqqq\}_i)$ is the b-tag likelihood

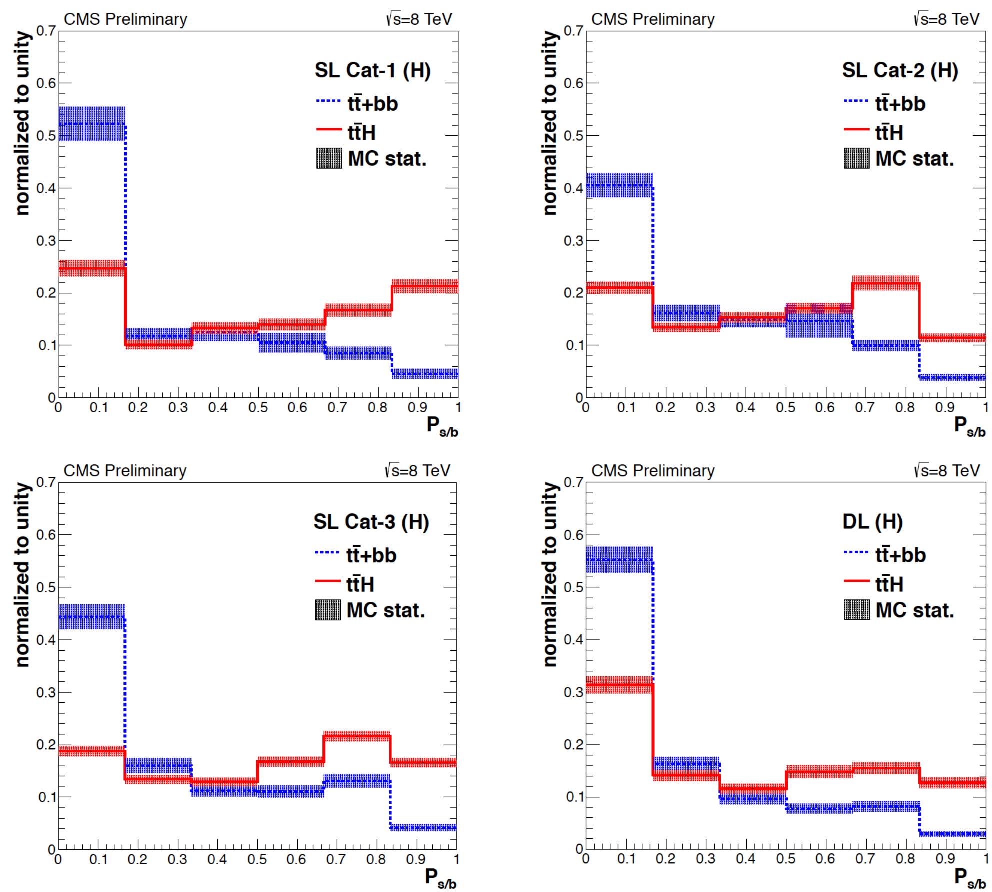
The final discriminant

Calculation

- For each event $\mathcal{P}_S$ and $\mathcal{P}_{B1}$ and $\mathcal{P}_{B2}$ are calculated
- Final discriminant is built

$$P_{s/b} = \frac{\mathcal{P}_S}{\mathcal{P}_S + \mathcal{P}_{B1} + \mathcal{P}_{B2}}$$

Expected distribution

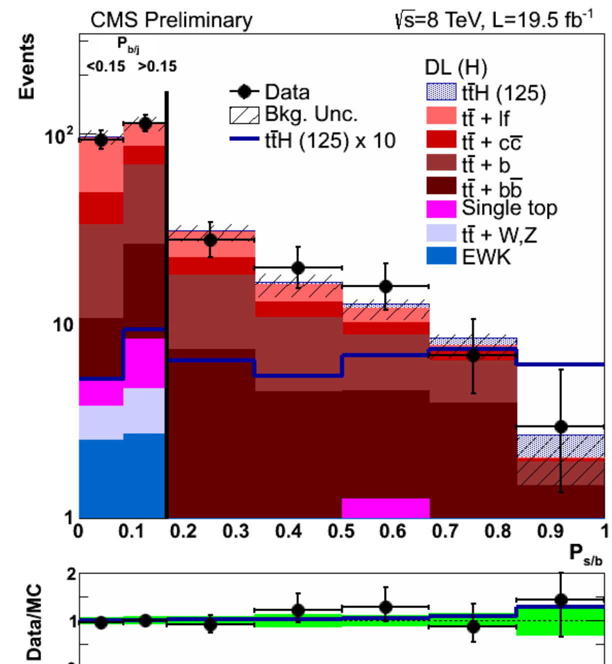


The final picture

Systematic uncertainties

- Signal and background predictions affected by experimental and theoretical uncertainties
- Dominant systematics are
  - Jet energy resolution
  - CSV uncertainty
  - tt+bb uncertainty
- Systematic uncertainties constrained by fitting the MC to the observed distributions
- Ultimately the uncertainty is dominated by the limited data

Post-fit distribution of $P_{s/b}$ (DL)

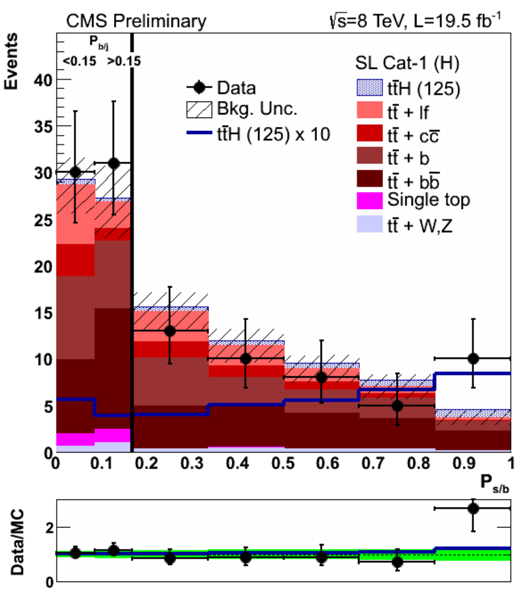


- Events in the first bin are split into 2 bins based on $P_{b/j}$:
  - Separates tt+bb and tt+lf

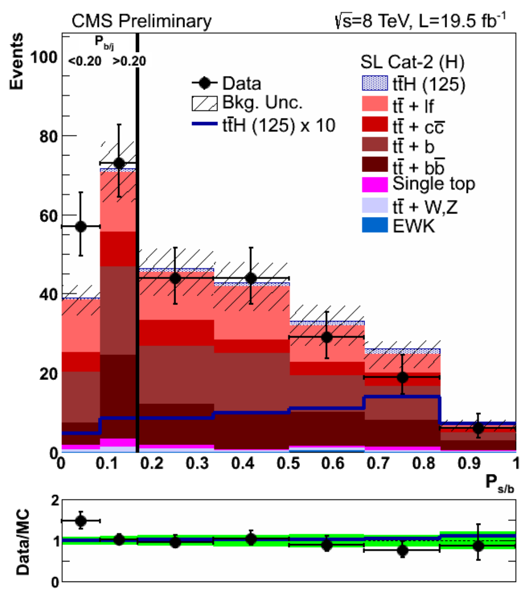
Post-fit discriminant distribution

First presented in July!

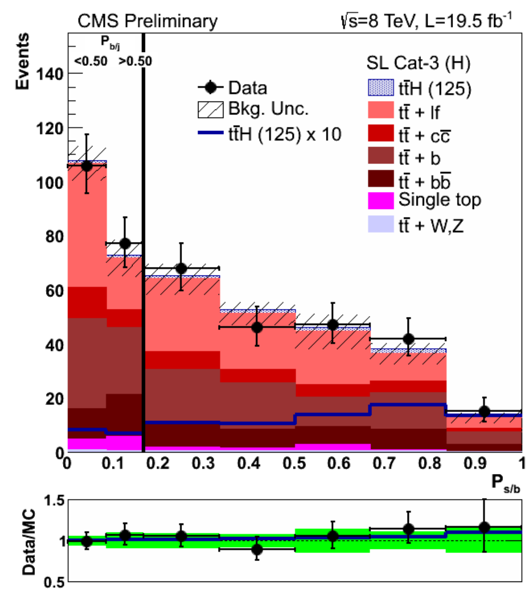
SL category 1



SL category 2



SL category 3



Signal expected to peak towards the right

2 rightmost bins provide the best signal/background discrimination

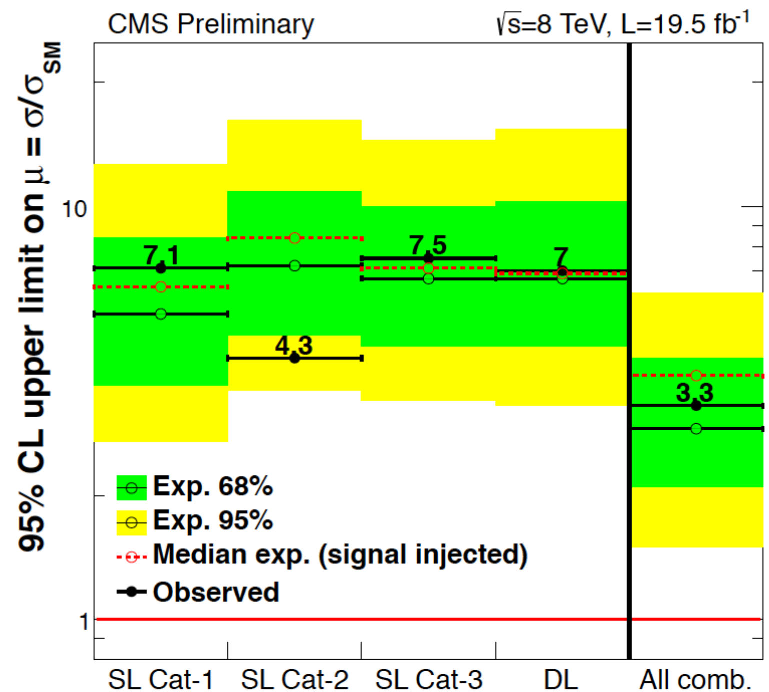
Exclusion limits

First presented in July!

Statistical interpretation

- Insufficient data for discovery
  - Analysis limited by statistics
- An upper limit can be placed on the ttH cross section
  - Signal strength modifier: $\mu = \sigma_{ttH}/\sigma_{SM}$
- Best fit value of μ after combining all categories is $\mu = 0.7 \pm 1.4$
  - Large uncertainty due to limited statistics

95% CL Upper limits on $\mu=\sigma/\sigma_{SM}$



Expected (observed) limit is μ < 2.9 (3.3)

Conclusion

Summary	Defined a signal/background discriminant based on the MEM Set an upper limit on the ttH cross section ($\mu = \sigma_{ttH}/\sigma_{SM}$) Expected upper limit is $\mu < 2.9$, observed limit is $\mu < 3.3$
Comparison	This analysis represents ~30% improvement over the previous CMS MVA analysis (HIG-13-019) ▸ Expected upper limit of $\mu < 4.1$, observed limit of $\mu < 5.2$ Improvement mostly due to better discrimination against tt+bb
Next steps	Expansion of current analysis ▸ Include all hadronic and boosted final states, and $H \rightarrow \tau\tau$ Looking forward to run at 13 TeV ▸ More data will provided a stronger result

Backup

Samples used in analysis

Data	■ **19.5 fb^{-1}:** 8 TeV 2012 data sample ▸ 7 TeV 2011 sample not considered in this analysis ■ Single-electron trigger: isolated, p_T > 27 GeV (e) ■ Single-muon trigger: isolated, p_T > 24 GeV (μ, $\mu\mu$, μe) ■ Double-electron trigger: isolated, p_T > 17, 8 GeV (ee)

Monte Carlo	■ **Signal:** $gg \to t\bar{t}H \to t\bar{t}b\bar{b}$ with M_H = 125 GeV	(PYTHIA)
	■ **tt+jets:** $gg \to t\bar{t}q\bar{q}\ ,\quad q = b, c, s, u, d$	(MadGraph)
	■ **ttV:** $t\bar{t} + W, Z$	(MadGraph)
	■ **Single top:** $t,\ tW,\ \bar{t},\ \bar{t}W$	(POWHEG)
	■ **EWK:** $q\bar{q} \to Z/\gamma^* \to \ell^+\ell^-$ and $W \to \ell\nu$	(MadGraph)
	■ **VV:** $WW,\ WZ,\ ZZ$	(PYTHIA)

b-tag likelihood ratio

b-tag likelihood ratio

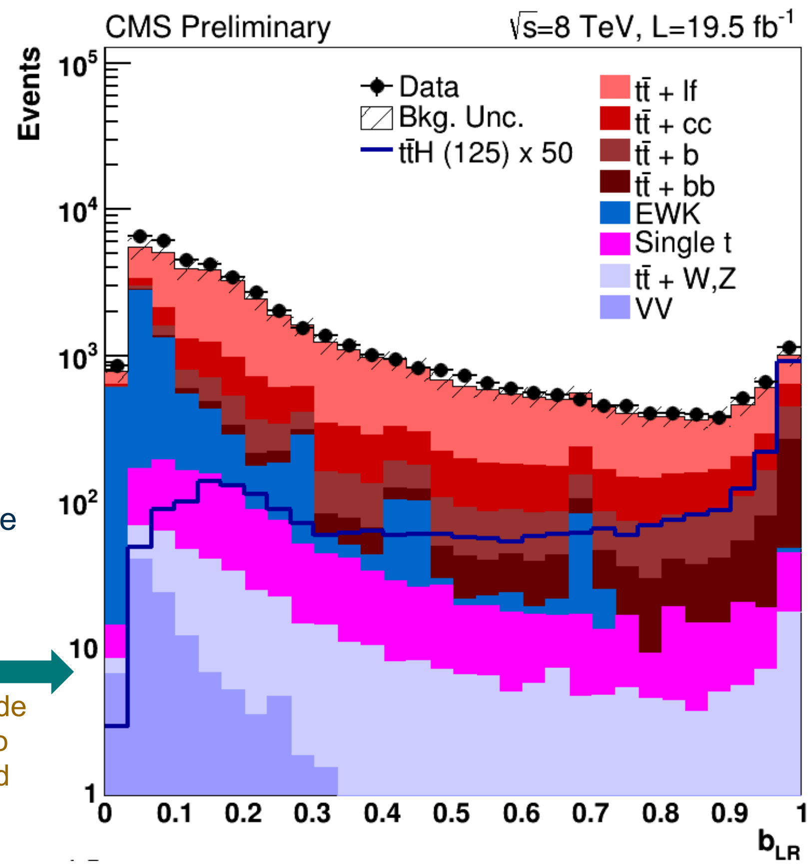
- Events selected based on the b-tag likelihood ratio discriminant
  - Jets sorted by CSV value (ζ)
    - A variable used to identify b jets
  - Top 4 to 6 jets used to calculate b_{LR}:

$$b_{LR} = \frac{\sum_i P(\zeta_1, ..., \zeta_6 | \{bbbbqq\}_i)}{\sum_i P(\zeta_1, ..., \zeta_6 | \{bbbbqq\}_i) + \sum_i P(\zeta_1, ..., \zeta_6 | \{bbqqqq\}_i)}$$

Note: Sum is over all possible permutations of jet–quark matching

Distribution of the b_{LR} discriminant
- SL events
- 6 or more jets

A cut on b_{LR} is made in each category to define high (H) and low (L) purity subcategories



The final discriminant

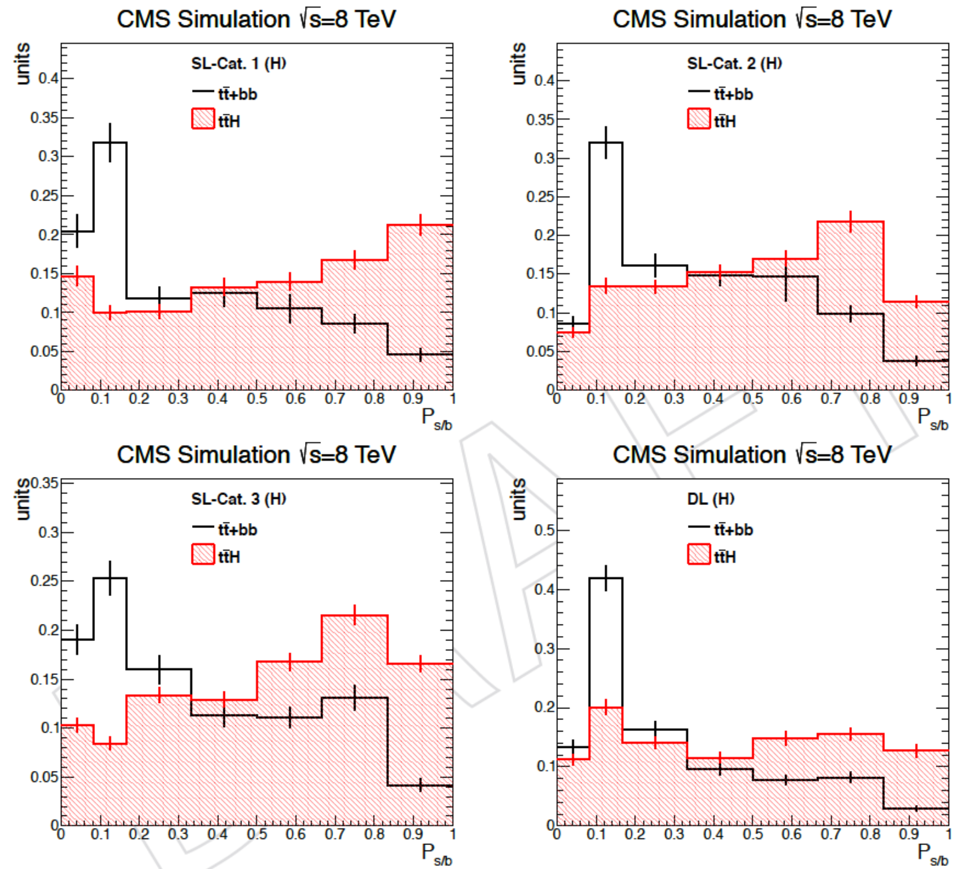
Calculation

- 3 different probabilities are determined
  - $\mathcal{P}_S(\mathbf{y}) = w_S(\mathbf{y})\mathcal{L}_{bbbb}(\zeta)$
  - $\mathcal{P}_{B1}(\mathbf{y}) = w_B(\mathbf{y})\mathcal{L}_{bbbb}(\zeta)$
  - $\mathcal{P}_{B2}(\mathbf{y}) = w_B(\mathbf{y})\mathcal{L}_{bbqq}(\zeta)$
    - Where $\mathcal{L}_{bbqq} = \sum_i P(\zeta_1, ..., \zeta_6 | \{bbqqq\}_i)$ is the b-tag likelihood
- Final discriminant is built

$$P_{s/b} = \frac{\mathcal{P}_S}{\mathcal{P}_S + \lambda_{b/j}\mathcal{P}_{B1} + (1-\lambda_{b/j})\mathcal{P}_{B2}}$$

  - $\lambda_{b/j}$ sets the relative ratio between tt+bb and tt+jj backgrounds

Expected distribution



Systematics and the fit

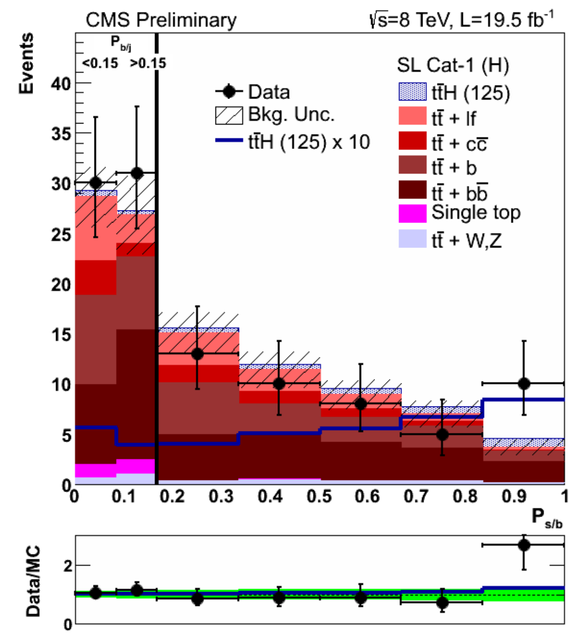
Systematic uncertainties

- Signal and background predictions affected by experimental and theoretical uncertainties

Luminosity	2.6%
Pile-up	omitted
Trigger and ID efficiency	2.0%
Jet energy scale and resolution	shape
b-tagging	shape
tt+jets modelling	shape
tt+ heavy flavour	50%
Parton density function	3-9%
QCD scale	1-20%
Limited MC statistics	bin-by-bin

- MC simulations are fitted to data allowing the systematics to float
  - Background shape and normalisations change depending on data
  - Constrains systematics, improves the power of the analysis

Post-fit distribution of $P_{s/b}$



- Events in the first bin are split into 2 bins based on: $$P_{b/j} = \frac{\mathcal{P}_{B1}}{\mathcal{P}_{B1} + \mathcal{P}_{B2}}$$
  - Value chosen to get ~50% tt+lf in each