Search for the Standard Model Higgs boson in $H \to WW \to \ell \nu q\bar{q}$

Nural Akchurin[1], Jake Anderson[2], Chayanit Asawatangtrakuldee[12], Andrea Benaglia[3], Andrew Beretvas[2], Jeffrey Berryhill[2], Pushpa Bhat[2], Sarah Boutle[4], Chris Clarke[5], Fabio Colombo[3], Analu Custodio[10], Jordan Damgov[1], Leonardo Di Matteo[3], Phil Dudero[1], Ricardo Eusebi[6], Pietro Govoni[13], Dan Green[2], Joey Goodell[4], Robert Harr[5], Pratima Jindal[8], Ajay Kumar[7], Gordon Kaussen[11], Kristina Krylova[5], Kevin Lannon[9], Sung-Won Lee[1], Qiang Li[12], Shuai Liu[12], Wuming Luo[9], Yajun Mao[12], Kellen McGee[5], Kalanand Mishra[2], Md. Naimuddin[7], Chris Neu[4], Ilya Osipenkov[6], Alexx Perloff[6], Kirti Ranjan[7], Sasha Sakharov[5], Ram K Shivpuri[7], Kevin Siehl[5], Andre Sznajder[10], Nhan V. Tran[2], Zijun Xu[12], Weimin Wu[2], John Wood[4], Fan Yang[2], Francisco Yumiceva[2], and Wei Zou[12]

[1] Texas Tech University, Lubbock, Texas, USA
[2] Fermi National Accelerator Laboratory, Batavia, Illinois, USA
[3] Milano-Bicocca University and INFN, Milan, Italy
[4] University of Virginia, Charlottesville, Virginia, USA
[5] Wayne State University, Detroit, Michigan, USA
[6] Texas A&M University, College Station, Texas, USA
[7] Delhi University, Delhi, India
[8] Princeton University, New Jersey, USA
[9] University of Notre Dame, Notre Dame, Indiana, USA
[10] Universidade do Estado do Rio de Janeiro (UERJ), Brazil
[11] University of Hamburg, Hamburg, Germany
[12] University of Peking, China
[13] CERN

HIG-12-021
CMS AN AN-12-193

2011 analysis: HIG-12-003

Public note:
http://cdsweb.cern.ch/record/1449158?ln=en

Public twiki:
https://twiki.cern.ch/twiki/bin/view/CMSPublic/Hig12003TWiki

Outline

HIG-12-021[*] (ICHEP 2012 analysis) is a development of the recently CMS approved HIG-12-003

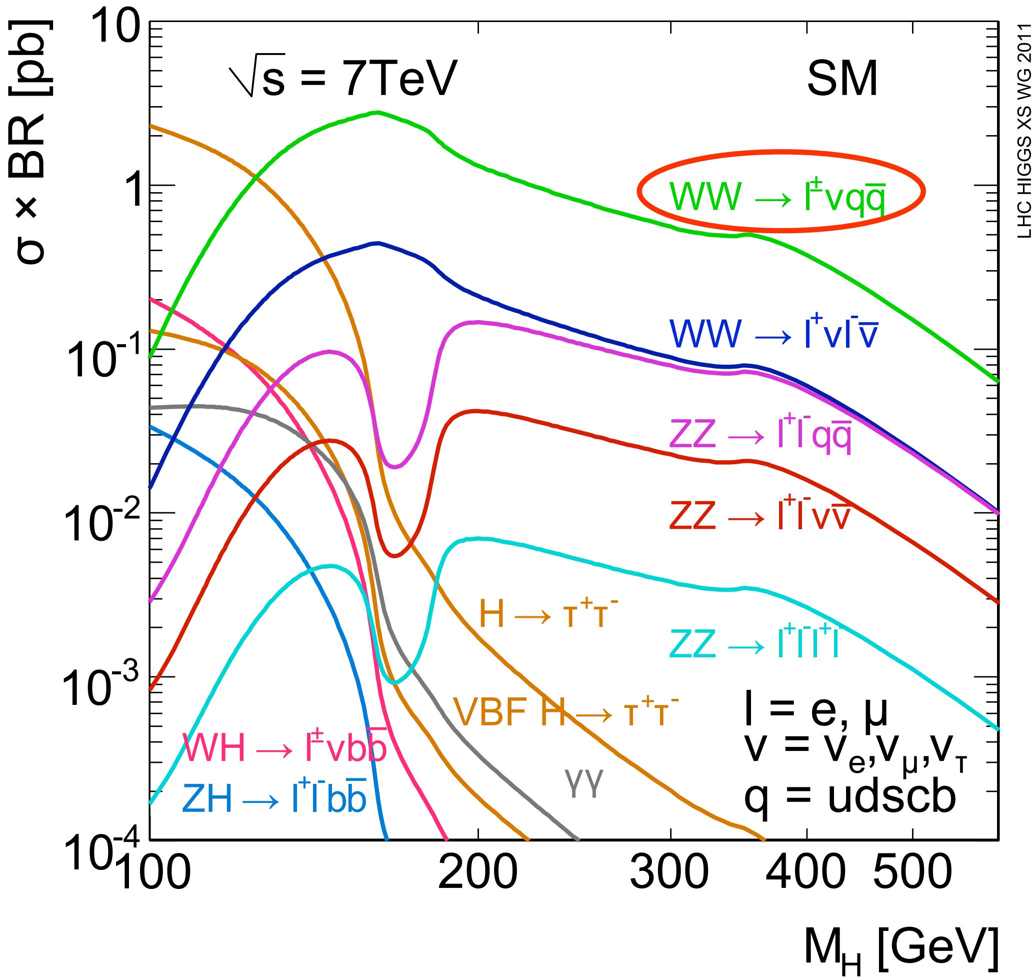
- **In this talk:**
  - Analysis overview and event selection
  - Technical details
  - Data/MC comparisons with 1.6 fb^{-1}
  - Signal modeling
  - First limit extraction (fit-analysis) with 1.6 fb^{-1}
  - Advanced limit extraction (mva-analysis) with 1.6 fb^{-1}

Today updated to 3.5 /fb

[*] Supported by AN-12-193

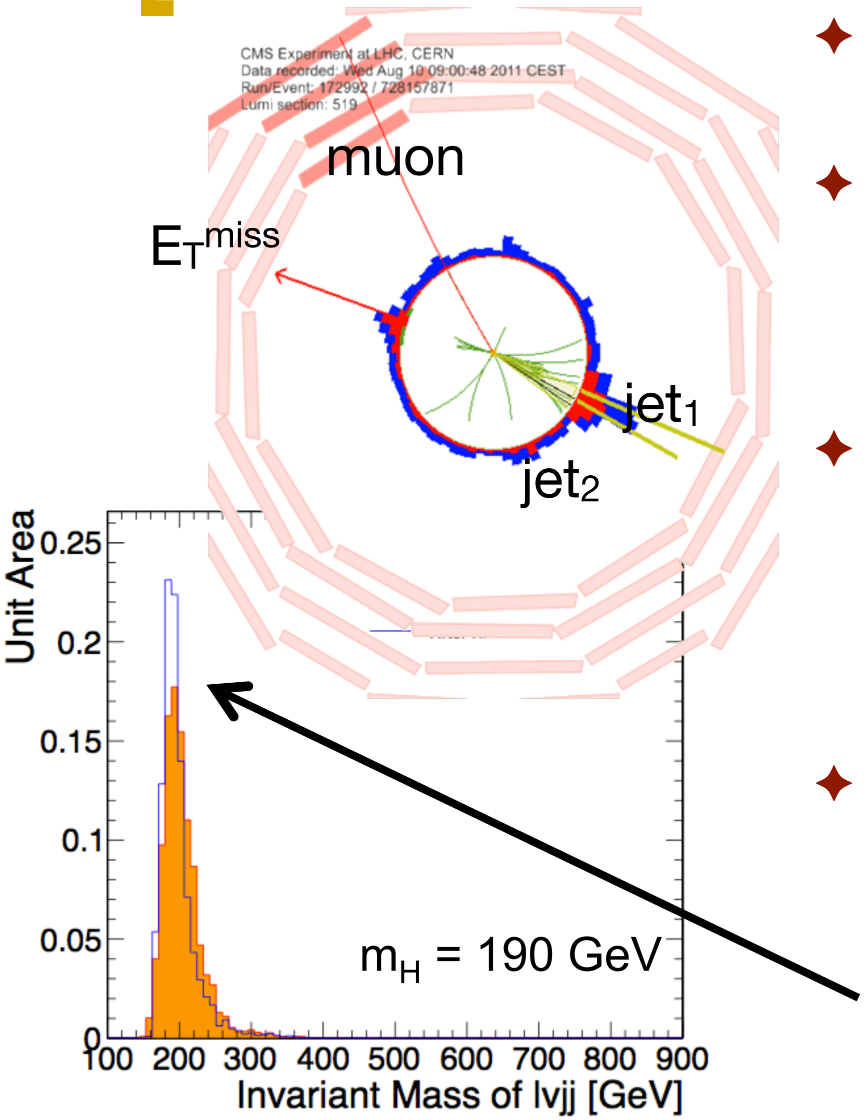
Review Q &A twiki:
https://twiki.cern.ch/twiki/bin/viewauth/CMS/HIG-12-021-ARC

Analysis overview



- H→WW→ℓνjj does a lot of heavy lifting.
  - largest BR × σ over most of the mass range.
  - Using a W mass constraint, the decay is sufficiently reconstructed to produce a mass peak.
- Principal drawback is the large W+jet background
  - We employ data-driven techniques to understand and control this process.

Event (pre-) selection



- One isolated, high-p_T lepton
  - p_T > 30 (25) GeV for electrons (muons)
- High E_T^{miss} from 1 neutrino
  - E_T^{miss} > 30 (25) GeV for electrons (muons)
  - $m_T(\text{lepton}+E_T^{miss})$ > 30 GeV
- Two high pT jets with m_{jj} ~ 80 GeV
  - Anti-Kt 0.5 particle flow jets
  - p_T > 30 GeV, $|\eta|$ < 2.4
  - ΔR(jet-lepton) > 0.3
  - $N_{extra\text{-}jets}$ = 0,1
- WW inv. mass ➔ reconstruct Higgs signal
  - neutrino p_Z from m_W constraint
  - We do a kinematic fit on lepton,E_T^{miss}, hadronic W to improve Higgs mass resolution and to remove the correlation between m_{WW} from m_{jj}.

Technical details

2012 physics objects

- Detailed summary in Sarah Boutle talk [*]:

[*]https://indico.cern.ch/getFile.py/access?contribId=45&sessionId=1&resId=0&materialId=slides&confId=188820

Short summary

- Use of PFBReco objects
  - PF isolation for electrons and muons
  - MVA Id for electrons (WP70) and cut based for muons (tight WP)
- AK5PFJets CHS Jets with JEC
  - Cut based JetId - loose WP
- PFMET with type 1 correction

Data and MC samples: Summer11/ CMSSW 5.2.X

dataset name Data	run range	int. lumi[*]
/SingleElectron/Run2012A-PromptReco-v1/AOD /SingleElectron/Run2012B-PromptReco-v1/AOD	190456-194479	1.599 fb^{-1}
/SingleMu/Run2012A-PromptReco-v1/AOD /SingleMu/Run2012B-PromptReco-v1/AOD	190456-194479	1.606 fb^{-1}

Now processed **3.5 fb^{-1}** using JSON file of June 12. Will use for unblinding.

dataset name MC	x-sec (pb)	eq. lumi (fb^{-1})
/WJetsToLNu_TuneZ2Star_8TeV-madgraph-tarball/Summer12-PU_S7_START52_V9-v1/AODSIM	36257	0.5
/DYJetsToLL_M-50_TuneZ2Star_8TeV-madgraph-tarball/Summer12-PU_S7_START52_V9-v1/AODSIM	3503	4.1
/TTJets_TuneZ2star_8TeV-madgraph-tauola/Summer12-PU_S7_START52_V9-v1/AODSIM	225	30
/WW_TuneZ2star_8TeV_pythia6_tauola/Summer12-PU_S7_START52_V9-v1/AODSIM	57	170
/WZ_TuneZ2star_8TeV_pythia6_tauola/Summer12-PU_S7_START52_V9-v1/AODSIM	32	300
/ZZ_TuneZ2star_8TeV_pythia6_tauola/Summer12-PU_S7_START52_V9-v1/AODSIM	8	1000
/T_TuneZ2_*_TuneZ2star_8TeV-powheg-tauola/Summer12-PU_S7_START52_V9-v1/AODSIM	10-50	~100
/Tbar_TuneZ2_*_TuneZ2star_8TeV-powheg-tauola/Summer12-PU_S7_START52_V9-v1/AODSIM	10-30	~100

So far generated 7 signal mass points at 8 TeV: **180, 200, 300, 400, 450, 500, 600 GeV**
Enough to plot limit curve with good granularity. Will have all mass points next week.

Signal cross section x BR from "Yellow Report"

Sample	Xsec (8 TeV) pb	Xsec (8 TeV) with VBF pb
GluGluToHToWWToLNuQQ_M-200_8TeV	7.127	7.7641
GluGluToHToWWToLNuQQ_M-250_8TeV	4.802	5.2324
GluGluToHToWWToLNuQQ_M-300_8TeV	3.606	3.9071
GluGluToHToWWToLNuQQ_M-350_8TeV	3.406	3.6192
GluGluToHToWWToLNuQQ_M-400_8TeV	2.924	3.086
GluGluToHToWWToLNuQQ_M-450_8TeV	2.003	2.1268
GluGluToHToWWToLNuQQ_M-500_8TeV	1.283	1.37797
GluGluToHToWWToLNuQQ_M-550_8TeV	0.8141	0.88788
GluGluToHToWWToLNuQQ_M-600_8TeV	0.523	0.58071

BR H > WW > lvjj	Charge conjugation factor	Number of lepton factor	Xsec X BR (8 TeV)
0.108	2	1.5	2.5155684
0.103	2	1.5	1.6168116
0.101	2	1.5	1.1838513
0.099	2	1.5	1.0749024
0.0852	2	1.5	0.7887816
0.0808	2	1.5	0.51553632
0.0799	2	1.5	0.330299409
0.0806	2	1.5	0.214689384
0.0818	2	1.5	0.142506234

Trigger requirements

- 2012 triggering is much simpler wrt 2011
  - Single lepton trigger for both electrons and muon
    - no MT cut applied online → gaining signal acceptance

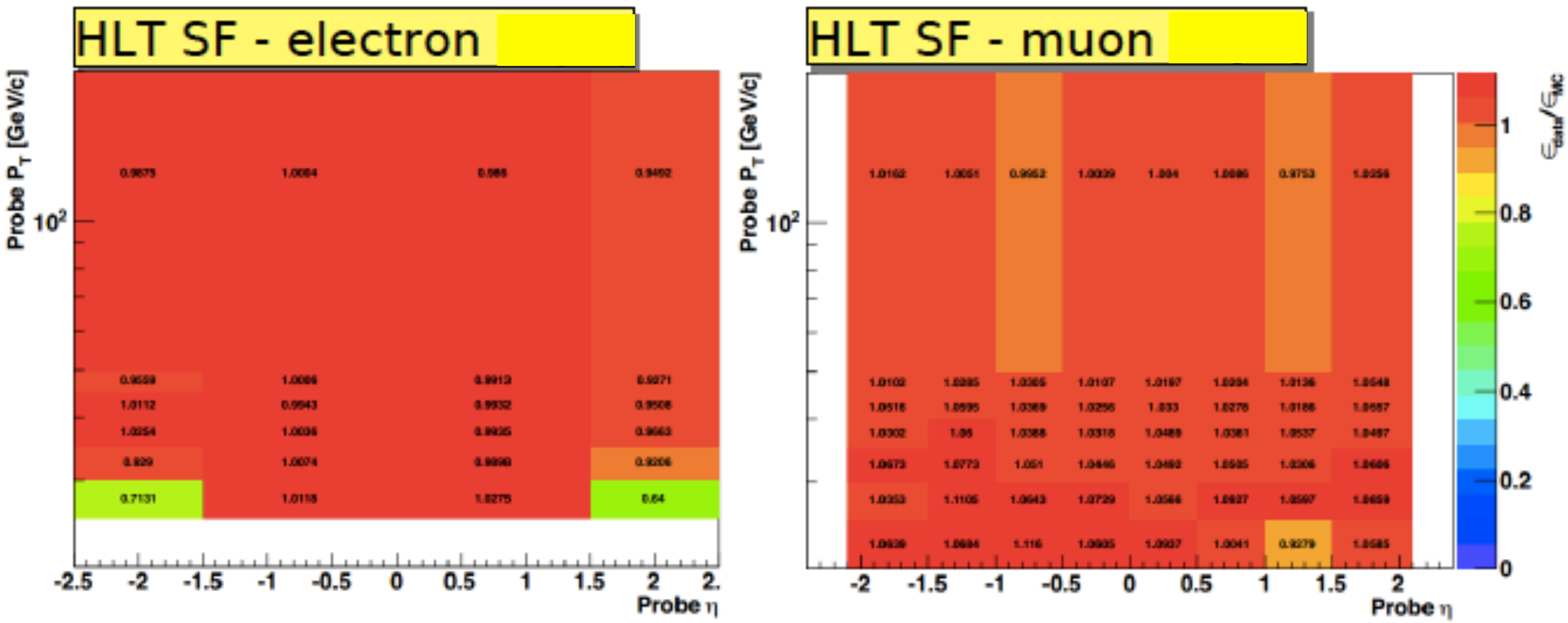
electron trigger paths	run range
HLT_Ele27_WP80_v*	190456- June 12 cutoff

muon trigger paths	run range
HLT_IsoMu24_v*, HLT_IsoMu24_eta2p1_v*	190456- June 12 cutoff

Efficiency corrections

- MC lepton efficiencies corrected with **“Tag and Probe”** on Z events
  - **HLT efficiency** and **reconstruction/identification** scale factors
  - Detailed description in Gordon Kaussen talk [*]

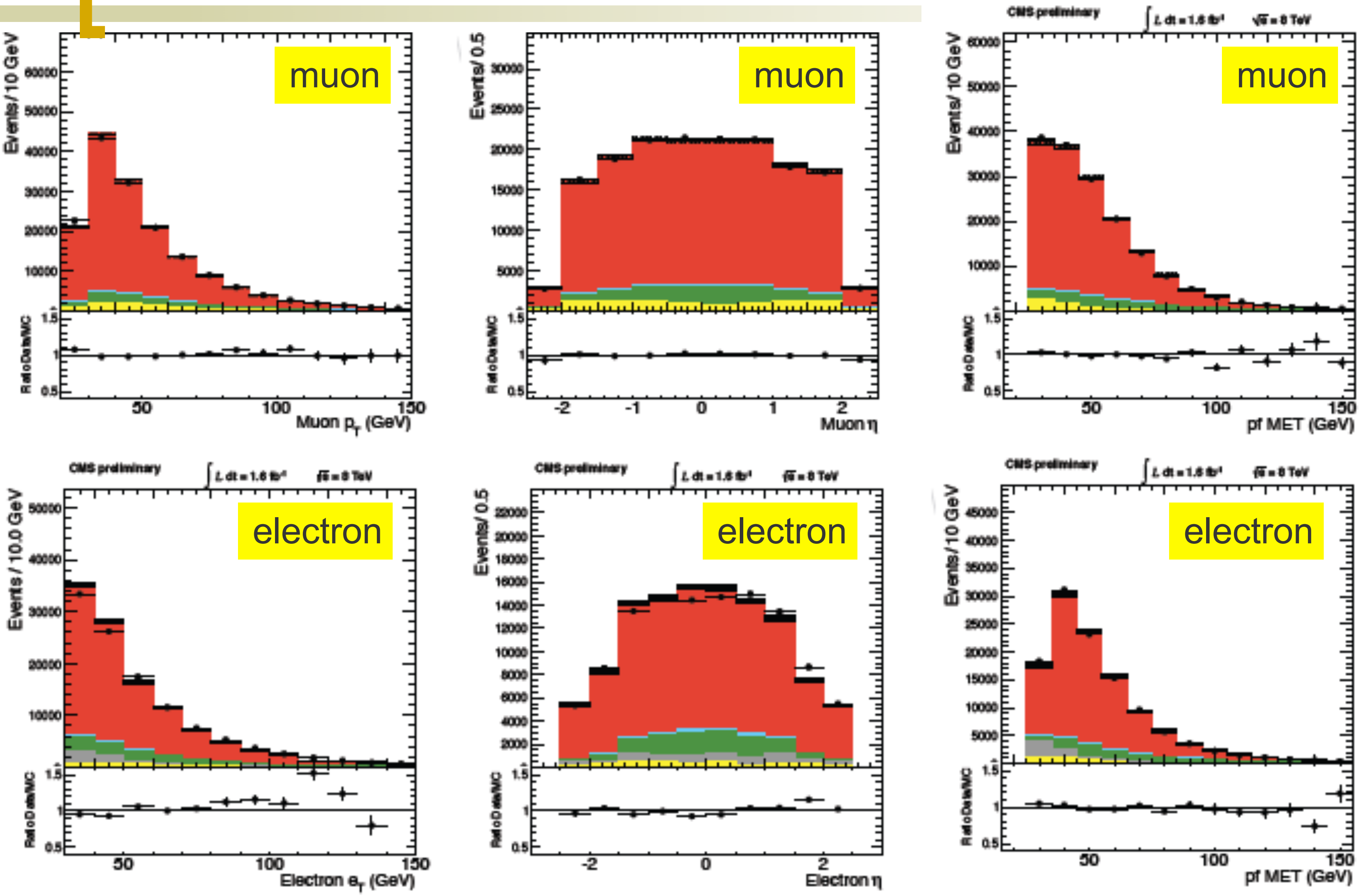
[*]https://indico.cern.ch/getFile.py/access?contribId=23&sessionId=10&resId=0&materialId=slides&confId=193619



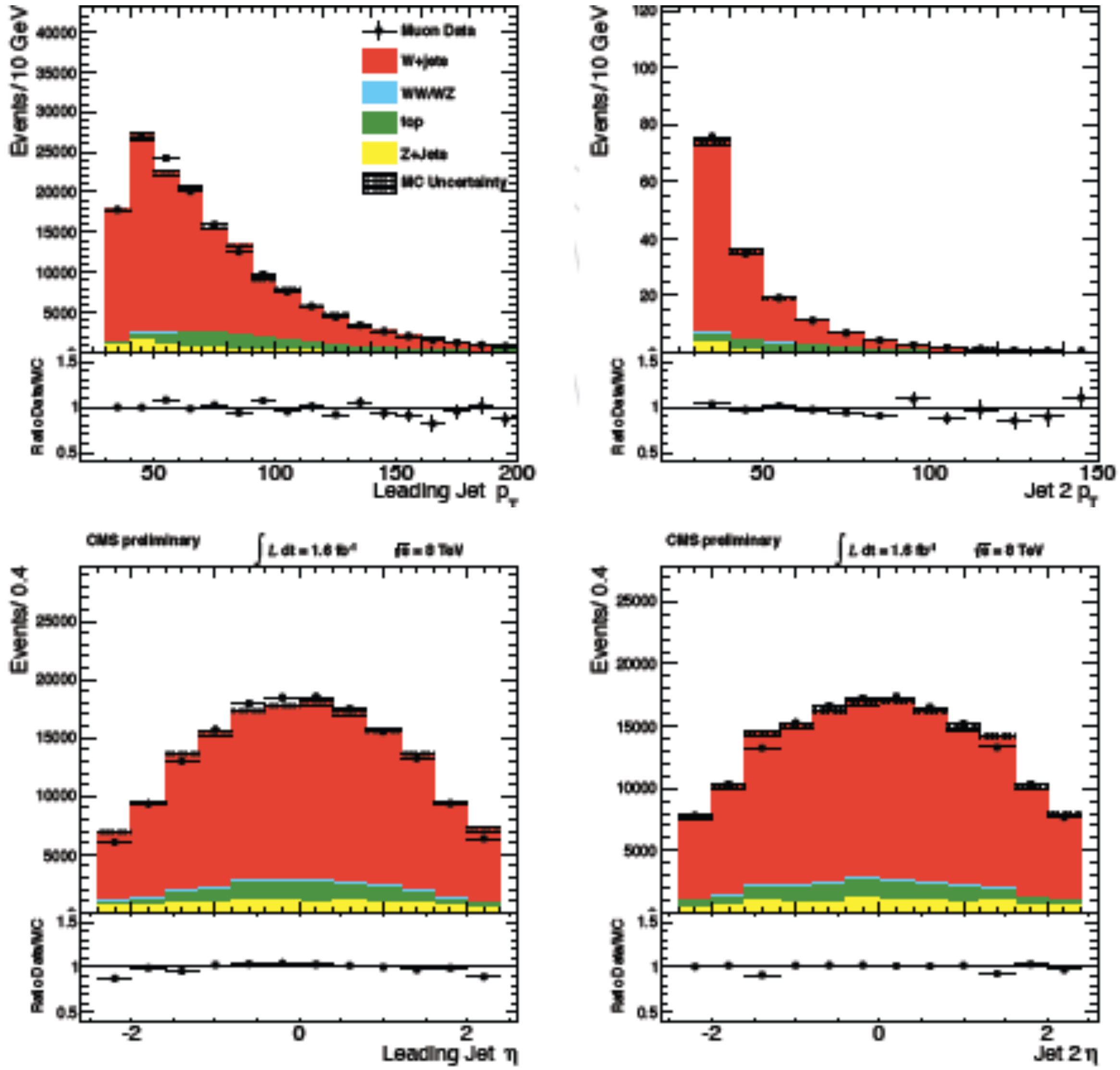
Apply similar scale factor for lepton reconstruction and ID efficiencies.

Data/MC comparison at pre-selection level

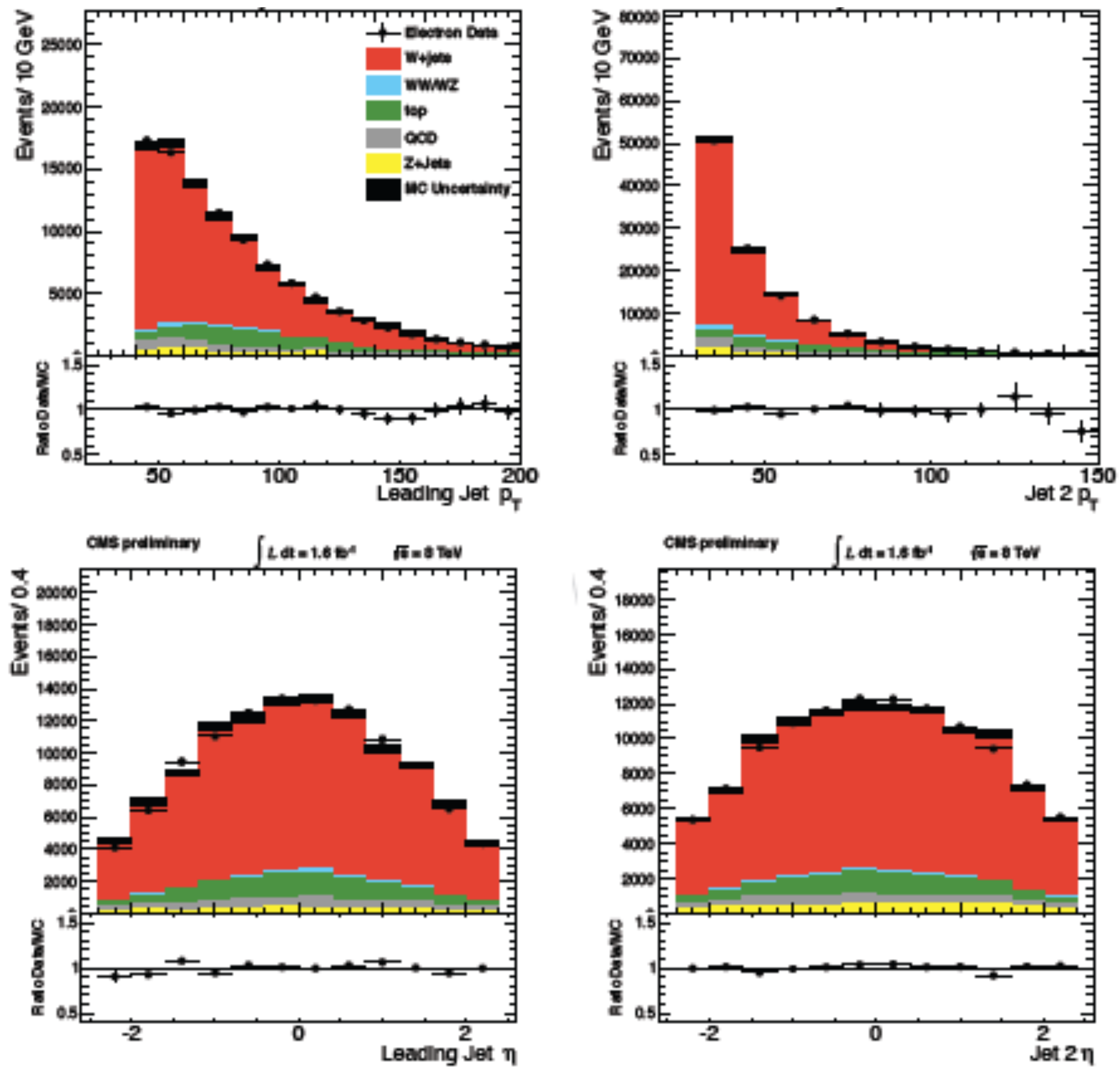
Comparison plots: leptons



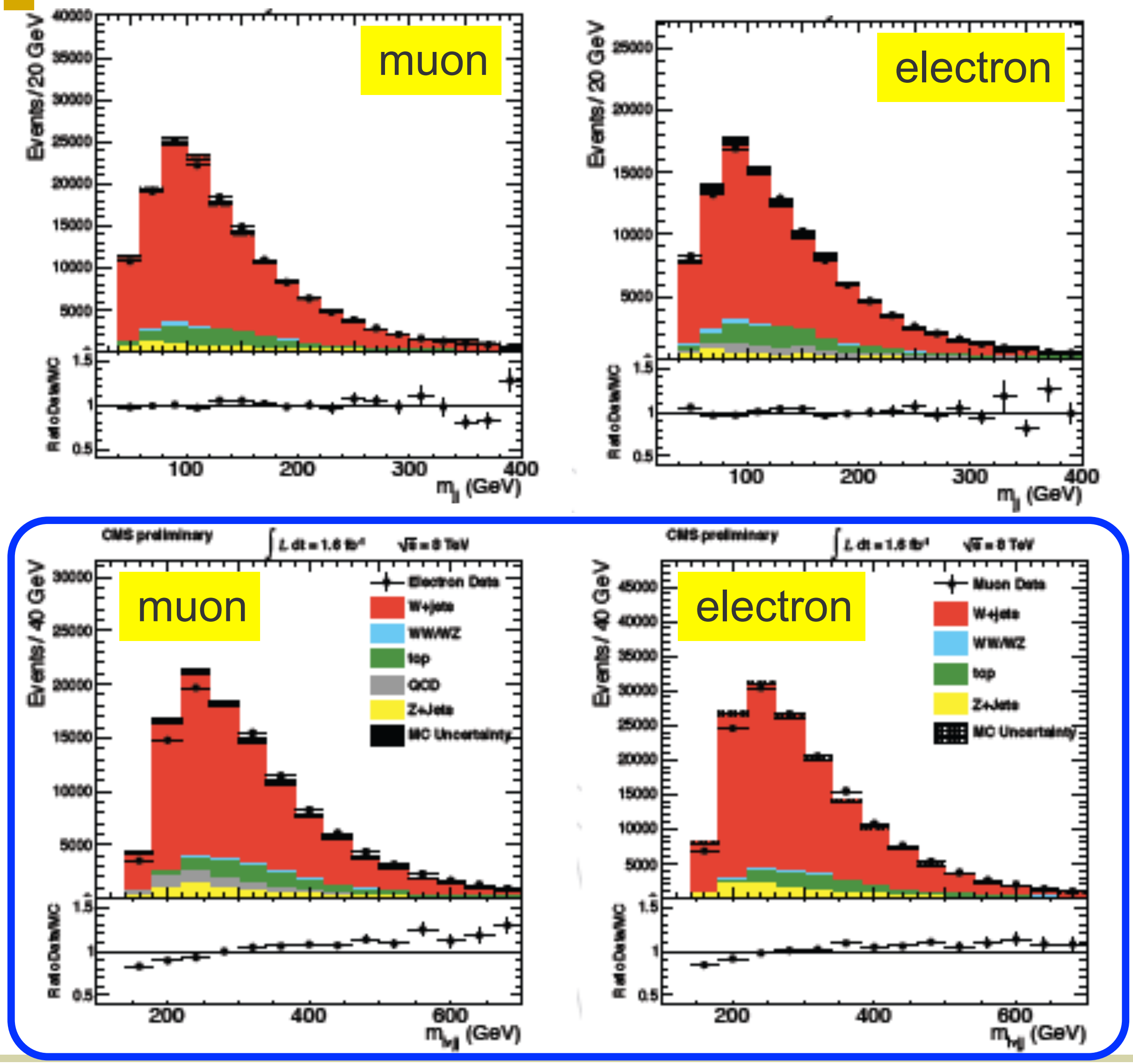
Comparison plots: jets (muon data)



Comparison plots: jets (electron data)



Comparison plots: m_{jj} , m_{lvjj}



The disagreement seen here is understood. Basically, the LO MadGraph simulation doesn't get the m_{lvjj} spectrum right. Hence, we use data-driven technique for W+jets shape.

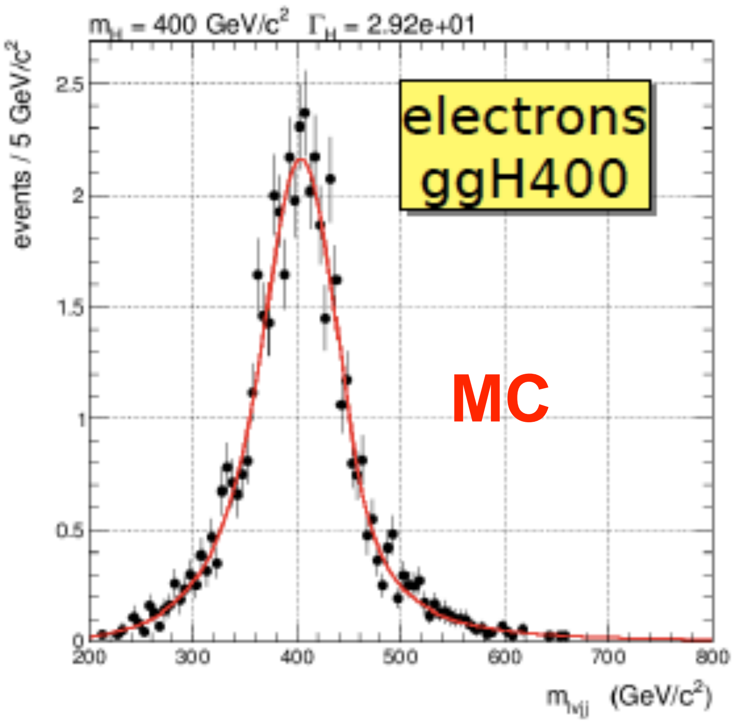
First limit extraction: fit to m_{WW} spectrum

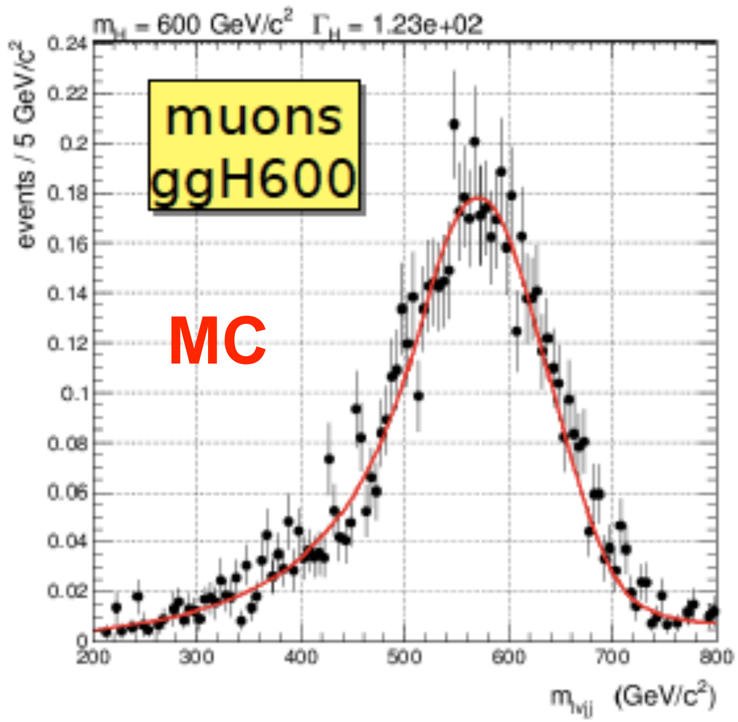
Rationale and methodology

- The idea is to try and **extract a limit** on SM Higgs x-section **just after common pre-selections**
  - **lots of events** survive: ~12k (~14k) events in the electron (muon) channel
  - the signal shape is modeled by **fitting the MC shapes** (double crystal ball)
  - the background shape is modeled by a suitable **analytic function** (attenuated power law)
- Limit extracted performing **shape analysis** on **m_{lvjj} spectrum**
  - data are fitted within the **combination tool** with a B+μS hypothesis to calculate the test statistics
  - only Higgs mass points within 250 and 600 GeV/c^2 are considered
- Only **few additional cuts** applied on top of pre-selections to enhance S/B:
  - centrality cuts: $|\eta_{lepton}| < 1.5$, $|\eta_{j}| < 3.$, $|\Delta\eta_{jj}| < 1.5$
  - cut on had. W resonance: $65 < m_{jj} < 95$ GeV/c^2

Signal shape

- **Signal shape** in the limit extraction is modeled with a **smooth function**
  - gaussian core with two power laws to describe the leading/trailing edge (**double crystal-ball**)
  - fitted to the signal MC samples for each mass





Background modeling

- A **Fermi turn-on ⊗ power law** is used to model the background
  - 4 parameters fitted to the data

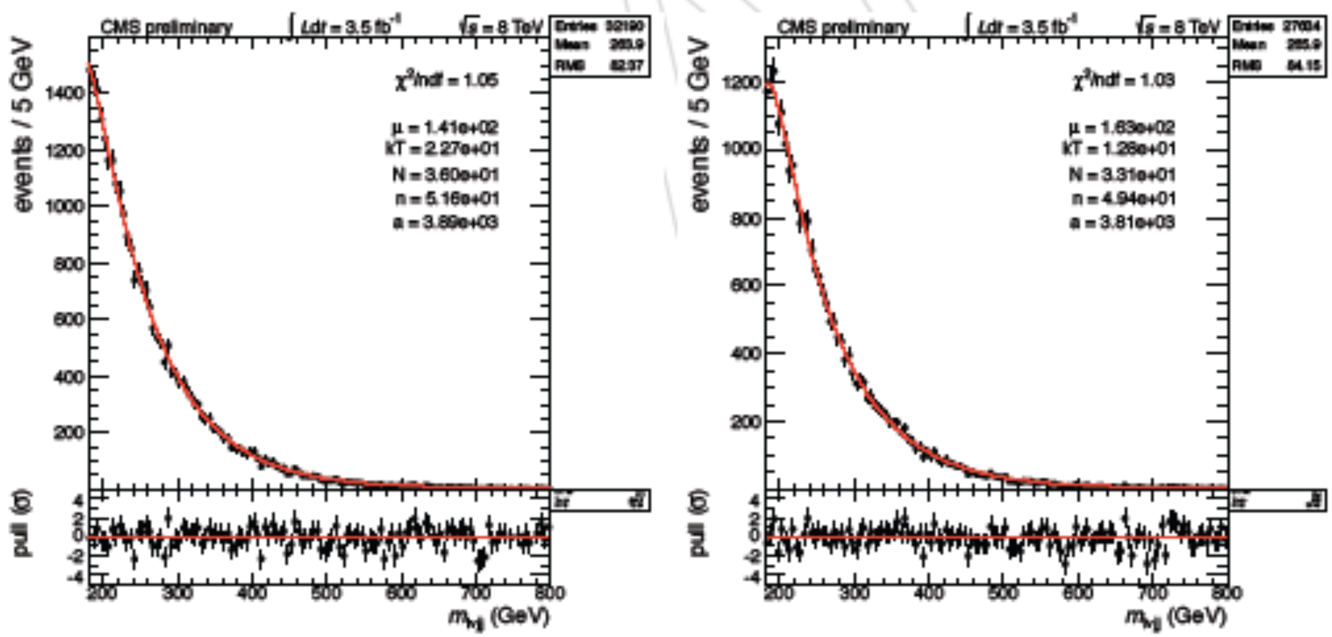
$$pdf_B = 1/(1+e^{-(x-\mu)/kT}) \cdot [(500+a)/(x+a)]^n$$

  - T.O. needed to describe the function attenuation at low mass
- Two fits obtained outside the comb. tool for illustrative purposes
  - χ^2/ndf close to 1

fit range: 180-800 GeV/c²

bin size: 5 GeV/c²

m_{jj} cut relaxed



Systematic uncertainties

background fit error (*)	~2 %	background systematics
background model choice (**)	see backup	
inclusive cross-section	15-20%	signal systematics
gluon-fusion higgs line shape	10-30%	
acceptance uncertainty due to pdfs	1-2%	
luminosity	2.2%	
jet energy scale and E_T^{miss}	shape	
lepton efficiencies	2%	
lepton trigger efficiencies	1%	
pile-up	< 1%	
b-tagging	< 1%	

- (*) this is automatically **taken into account** by the **combination tool**
- (**) this is **translated** into an **equivalent error** on the **signal** - more in the next slides

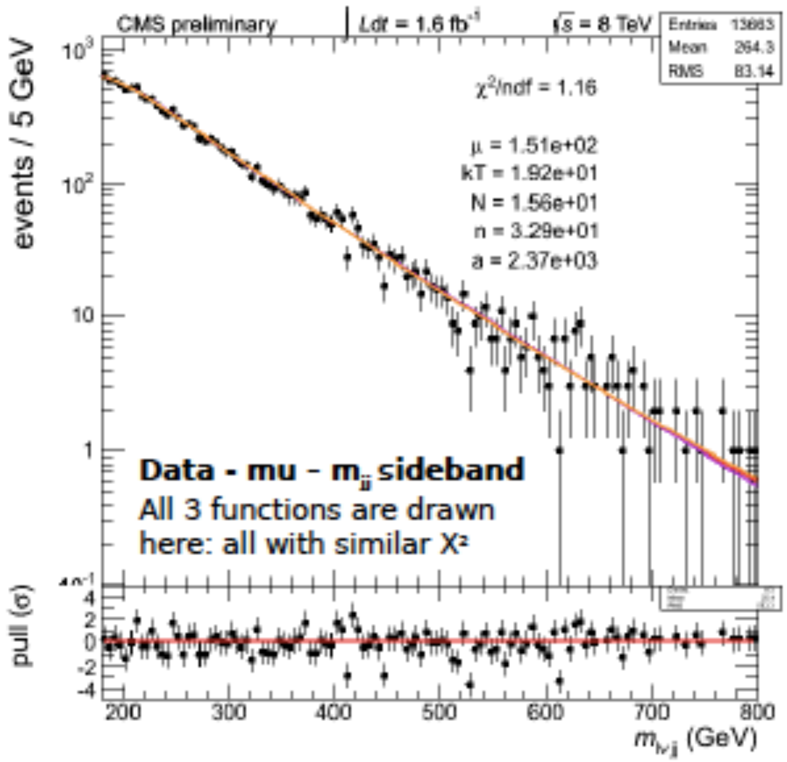
Background model systematics

- We want to determine the **systematic uncertainty** related to the **choice** of the **background shape**

$$\mathcal{F}_{\rm B} = \frac{1}{1+e^{-(x-\mu)/\tau}} \cdot \left(\frac{500+a}{x+a}\right)^n$$

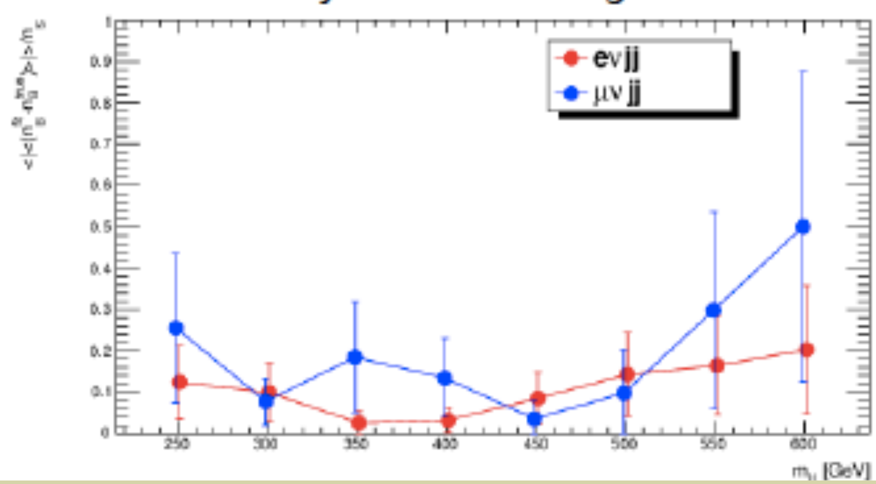
$$\mathcal{F}_{\rm B,\ alt} = \frac{1}{1+e^{-(x-\mu)/\tau}} \cdot \left(e^{ax^3+bx^2+cx}\right)$$

$$\mathcal{F}_{\rm B,\ alt} = \frac{1}{1+e^{-(x-\mu)/\tau}} \cdot \left(e^{-\lambda_1 x} + N_2 e^{-\lambda_2 x}\right)$$

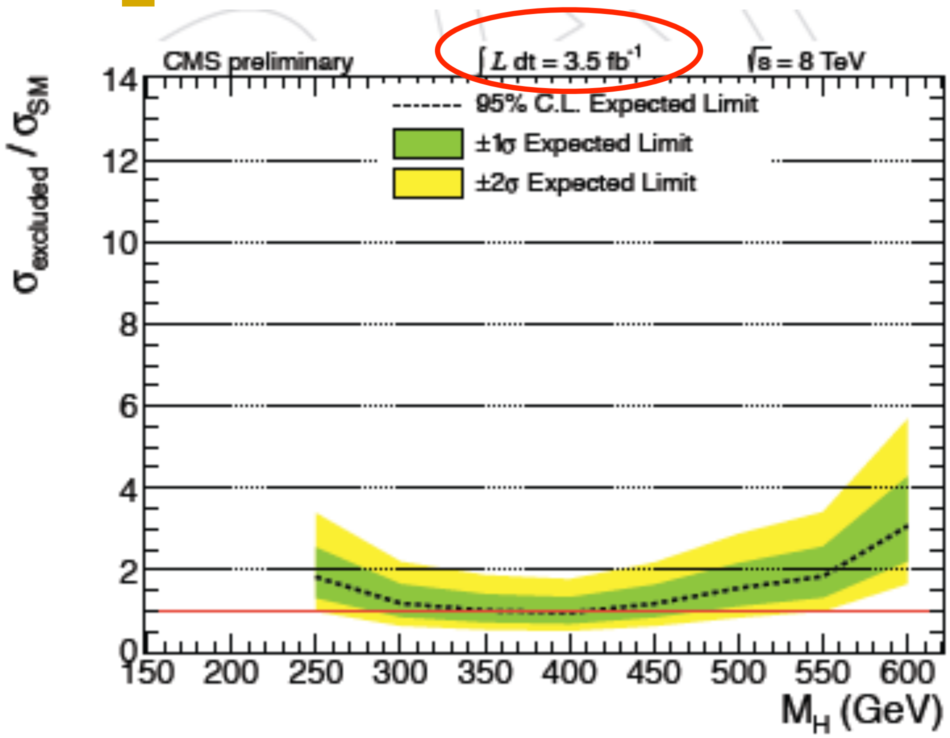


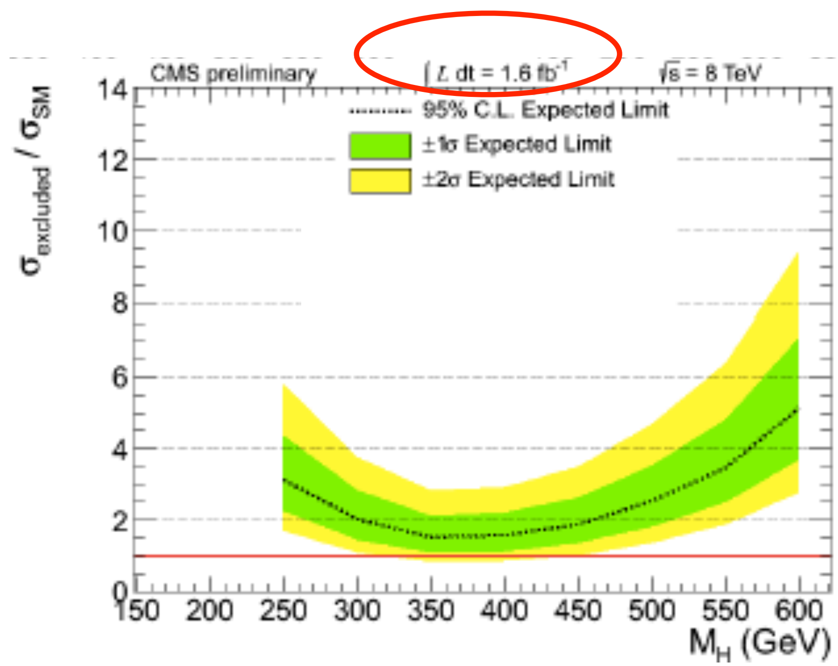
- we find a set of suitable functions and we initialize them on data
- we consider the difference for **each mass hypothesis** with respect to the nominal choice as the uncertainty on our choice
- we convert the error to a signal uncertainty

→ 2-50% syst. error on signal



Limits on SM Higgs x-section: 3.5 fb^{-1}





Advanced limit extraction: MVA (LH) approach

Use optimized selection to improve sensitivity

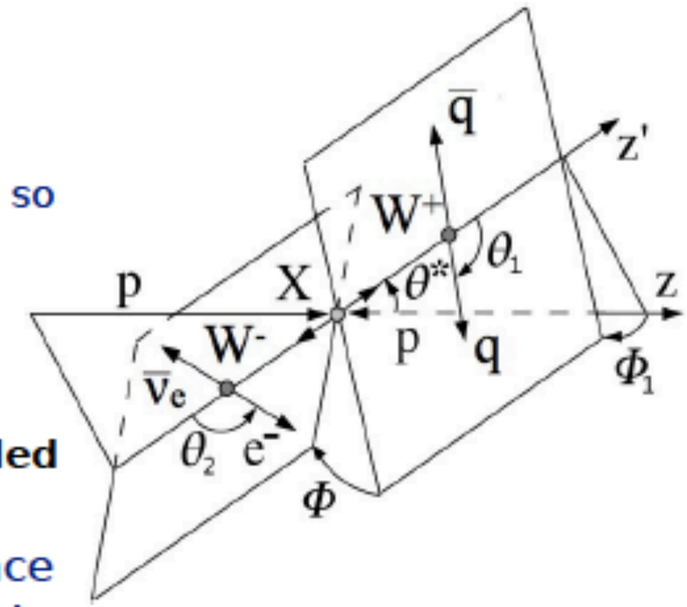
- To **improve the limit** on SM Higgs cross-section and **reduce the systematics** on the background evaluation:
  - drastically reduce the number of background events
  - preserve a reasonable efficiency on the signal

  ⇒ **MVA approach**

- Use ROOT's TMVA package to build a simple **likelihood** discriminator:
  - a different likelihood is built for each different final state
    - 12 mass points × 2 flavours (e/μ) × 2 jet bins (2j/3j) = **48 different likelihoods**

- To extend the analysis to low mass (down to 170 GeV/c^2), a **signal extraction technique** other than the fit method is needed

Likelihood input variables

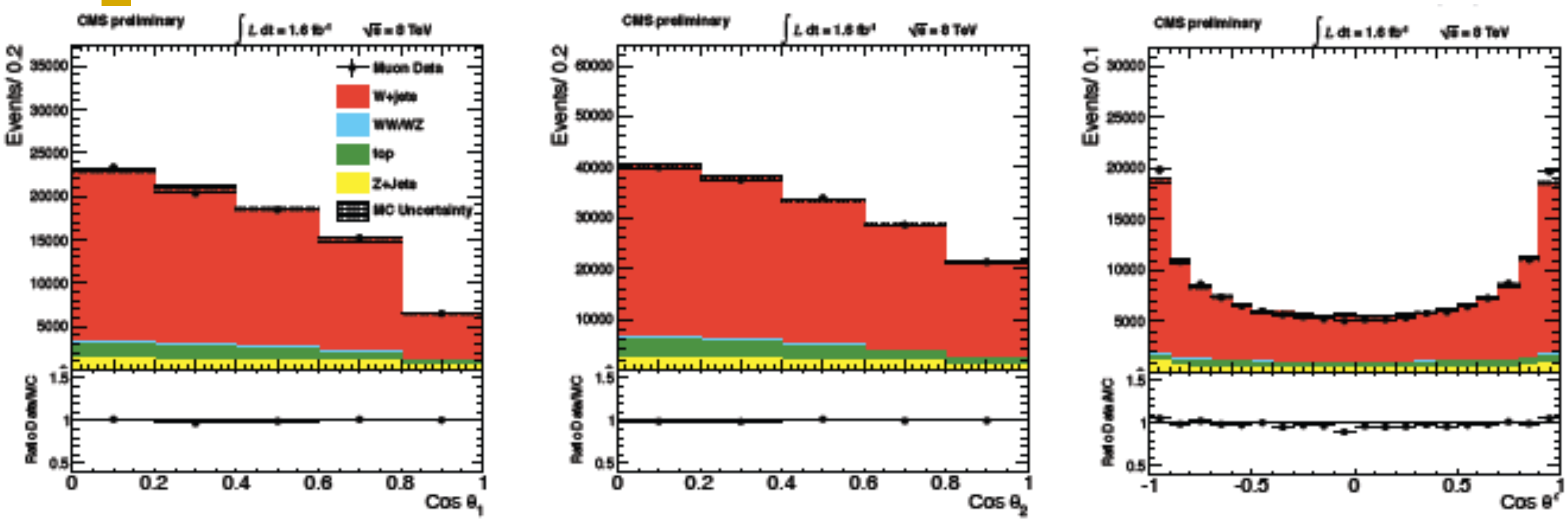


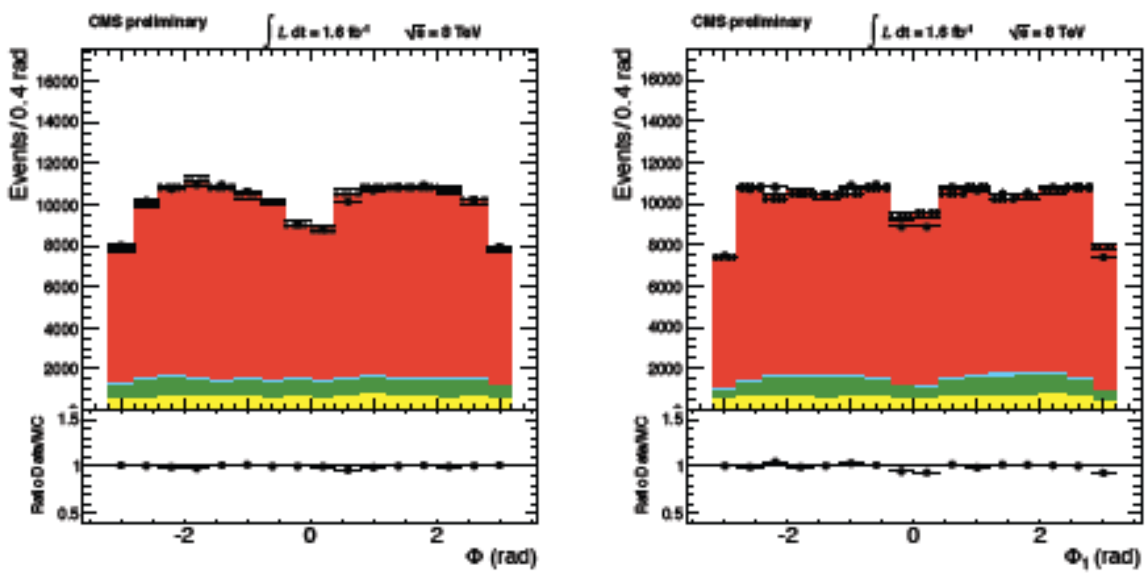
- The higgs boson decay kinematics is fully described by → $\{m_{WW}, m_{jj}, \theta_1, \theta_2, \theta', \phi, \phi_1\}$
  - **m_{WW}** is the variable we use to extract the limit, so it is not included
  - **m_{jj}** used to estimate the huge W+jets background, so it is not included
  - the other **five angular variables are included**
- The **lepton charge** is a good variable since signals are charge-symmetric, while W+jets production is not
- Higgs kinematic variables such as **$(p_T)_{WW}$** and **y_{WW}** help as well

minimal full set of variables describing the Higgs decay:

$\{\theta_1, \theta_2, \theta^*, \phi, \phi_1, (p_T)_{WW}, y_{WW}$, lepton charge$\}$

Data/MC comparison for inputs (angular vars)

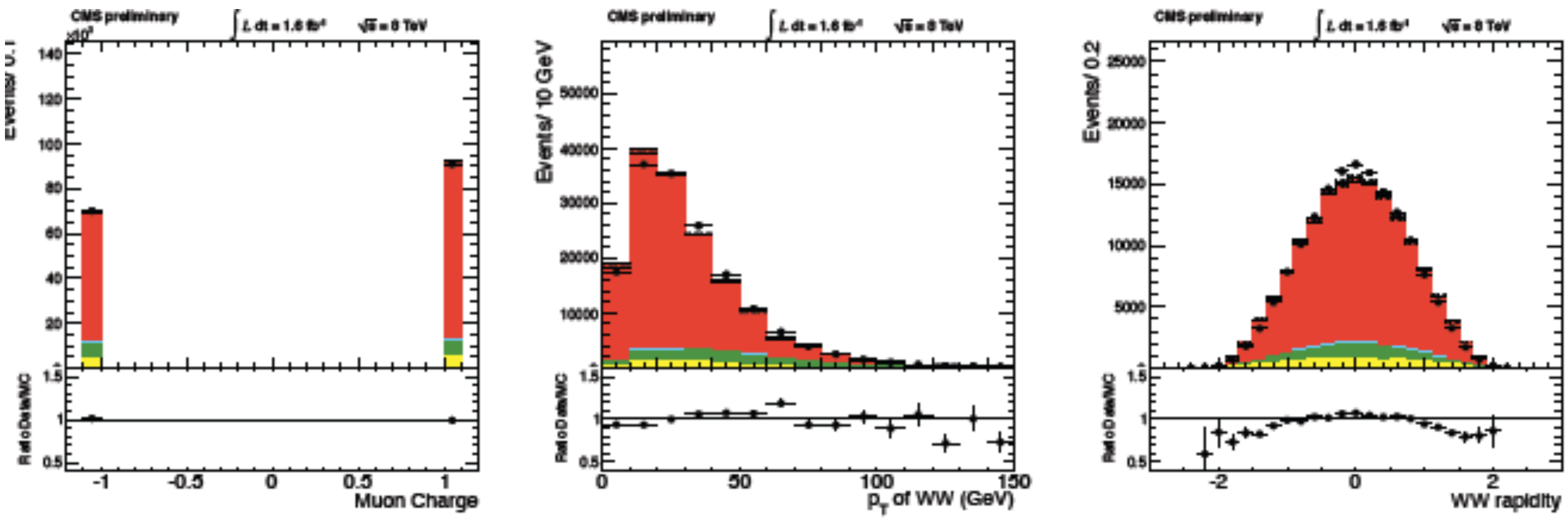




muon data

Reasonable modeling of kinematics in simulation.

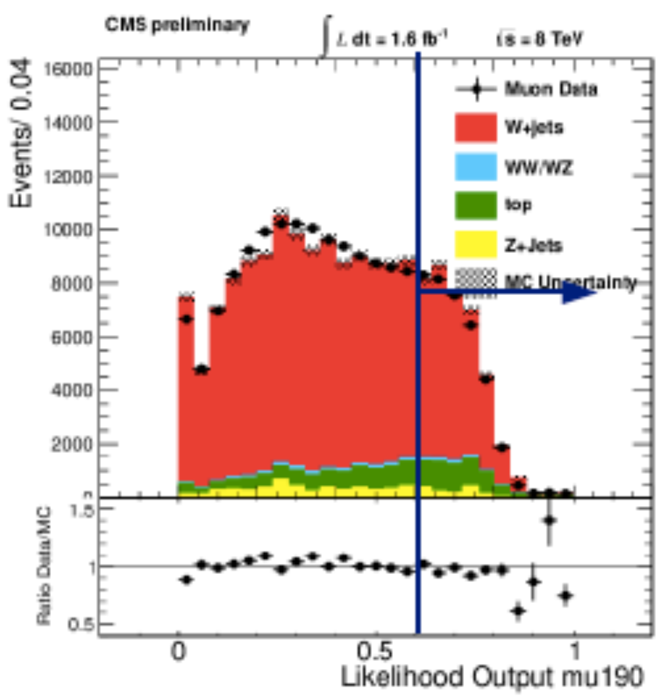
Data/MC comparison for inputs (other 3 vars)



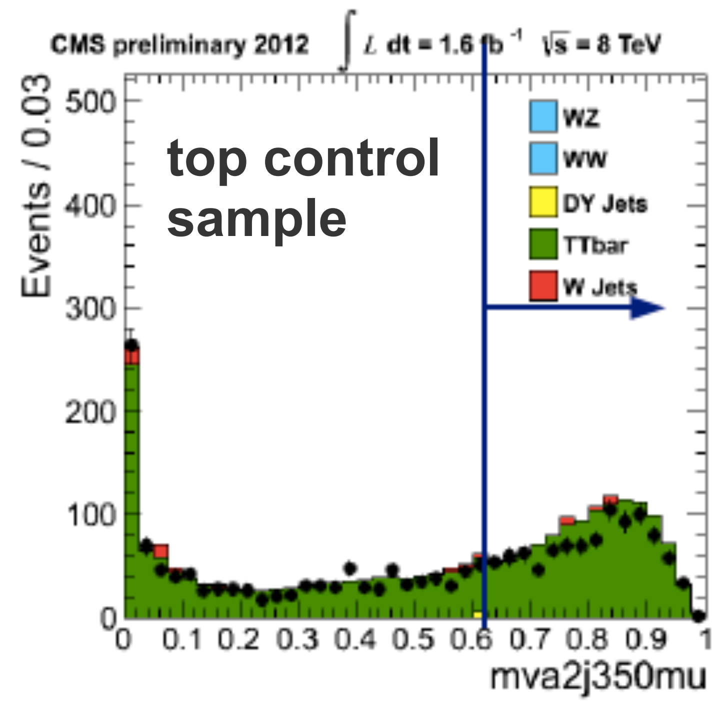
Agreement at the same level as in last year's data.

Example of likelihood output

- We look at data/MC comparison of the likelihood for the analysis final state, dominated by **W+jets** (left), and for a **tt-enriched** final state (right)
- Residual differences between tt-data and tt-MC taken as **systematic error on MVA-signal efficiency**

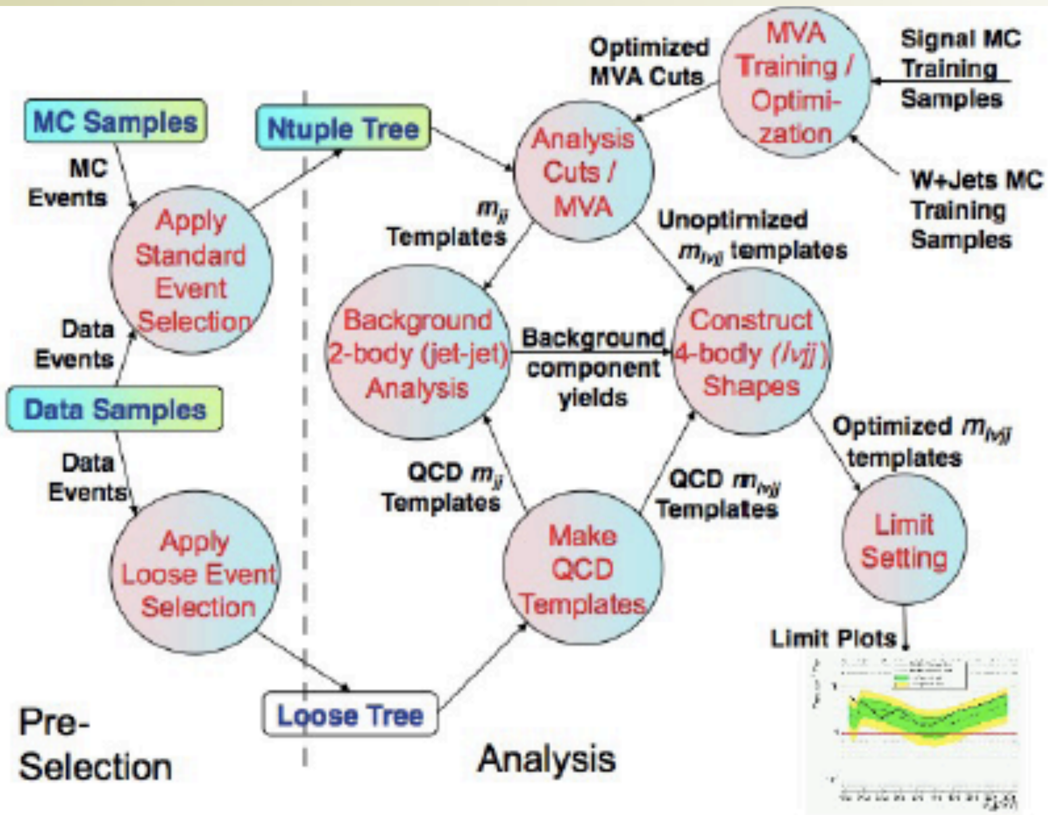


H190 - mu - 2j



H350 - mu - 2j

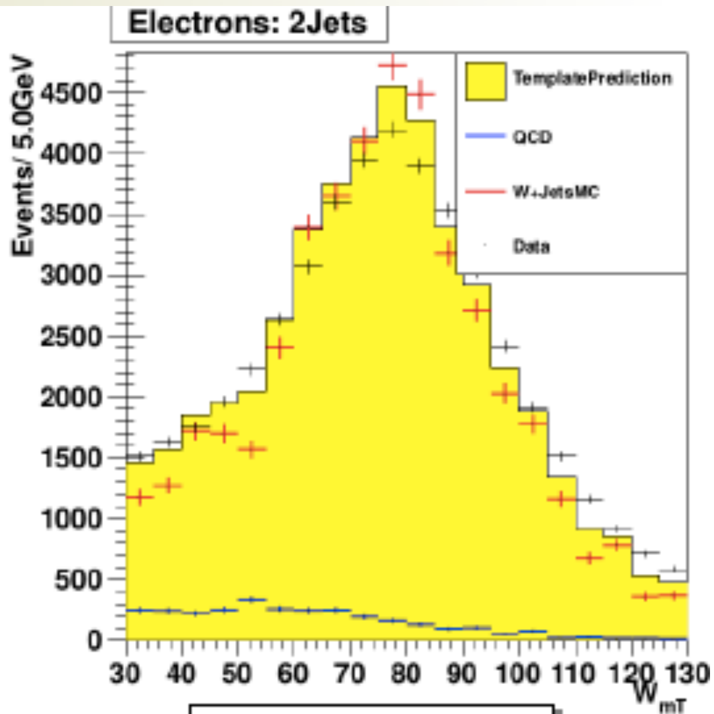
Signal extraction technique



- **Two** main conceptual **steps**:
  - The **two-body mass distribution** m_{jj} used to get the **background normalization**
  - The **four-body mass distribution** m_{lvjj} used to extract the **limit**

QCD data driven

- To get the **QCD shape from data**: define a ~ pure QCD sample applying all standard cuts but
  - invert isolation cut, i.e. $I_{comb}/p_T > 0.3$
  - for electrons, remove the ID requirement
  - relax E_T^{miss} cut to be > 20 GeV instead of 30 GeV
- Take the **W+jets** shape from **MC**
- To get the **QCD fraction** in **data**
  - fit the lepM_T distribution in data with the above described shapes
  - account for different acceptances between E_T^{miss} > 20 and E_T^{miss} > 30 GeV



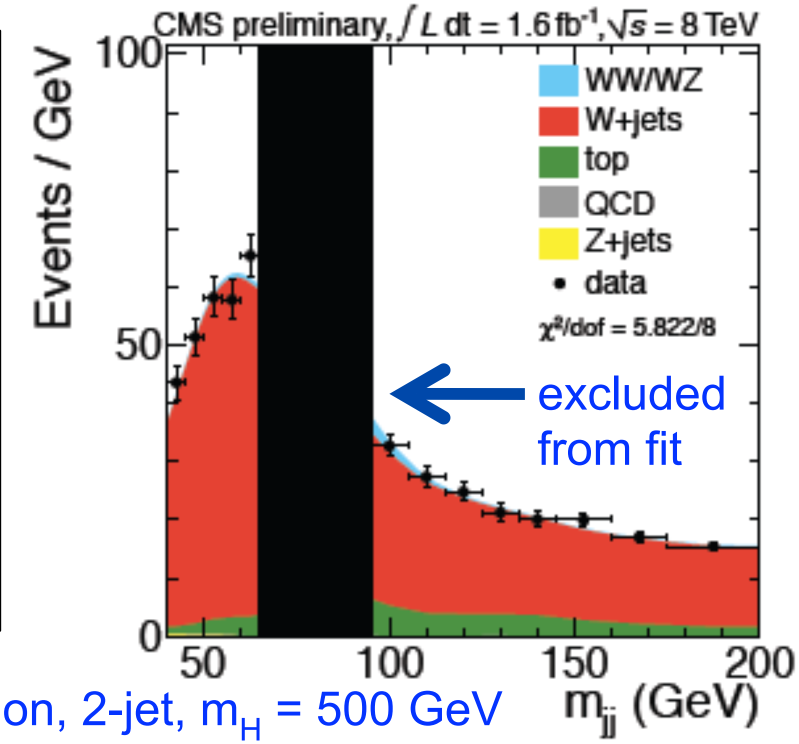
QCD fractions in data

Stat. Error Only	2 jets	3 jets
electron	$6.4 \pm 0.7\%$	$2.1 \pm 0.7\%$
muon	$0.2 \pm 0.4\%$	$0.0 \pm 0.4\%$

Syst. Error	2 jets	3 jets
electron	50%	50%
muon	100%	100%

Relative normalization from fit to m_{jj}

- Determine backgrnd composition in a 1D, unbinned, max likelihood fit to the di-jet invariant mass spectrum.
- Background shapes are taken from:
  - MC for all minor backgrounds (not for W+jets and QCD)
  - data-driven approach for QCD
  - analytic description/ MC for W+jets



muon, 2-jet, m_H = 500 GeV

Process	Shape	External constraint on normalization
W+jets	data/MC	Unconstrained
Diboson	MC	Constrained: (NLO) 89.4 pb ± 10% [48]
$t\bar{t}$	MC	Constrained: (approx. NNLO) 225 pb ± 7% [49]
single top	MC	Constrained: (approx. NNLO) 113 pb ± 5% [50–52]
Drell-Yan+jets	MC	Constrained: (NNLO, $m_{\ell\ell} > 50$ GeV) 3504 pb ± 4.3% [53]
Multi-jet	data	Constrained: $\not{E}_T$ fit in data ± 50% (100%) for electrons (muons)

Limit setting with the m_{WW} spectrum

- The fit to the m_{jj} spectrum determines the relative normalization of the backgrounds.
- The background components are stacked up and compared with the data with the additional selection ($65 < m_{jj} < 95$) GeV.
- Shapes of the minor backgrounds are taken from MC.
  - Again, QCD is taken from the data-driven sample.
  - The W+jets shape is constructed from the m_{jj} sidebands (see next two slides).

W+jets shape (1/2)

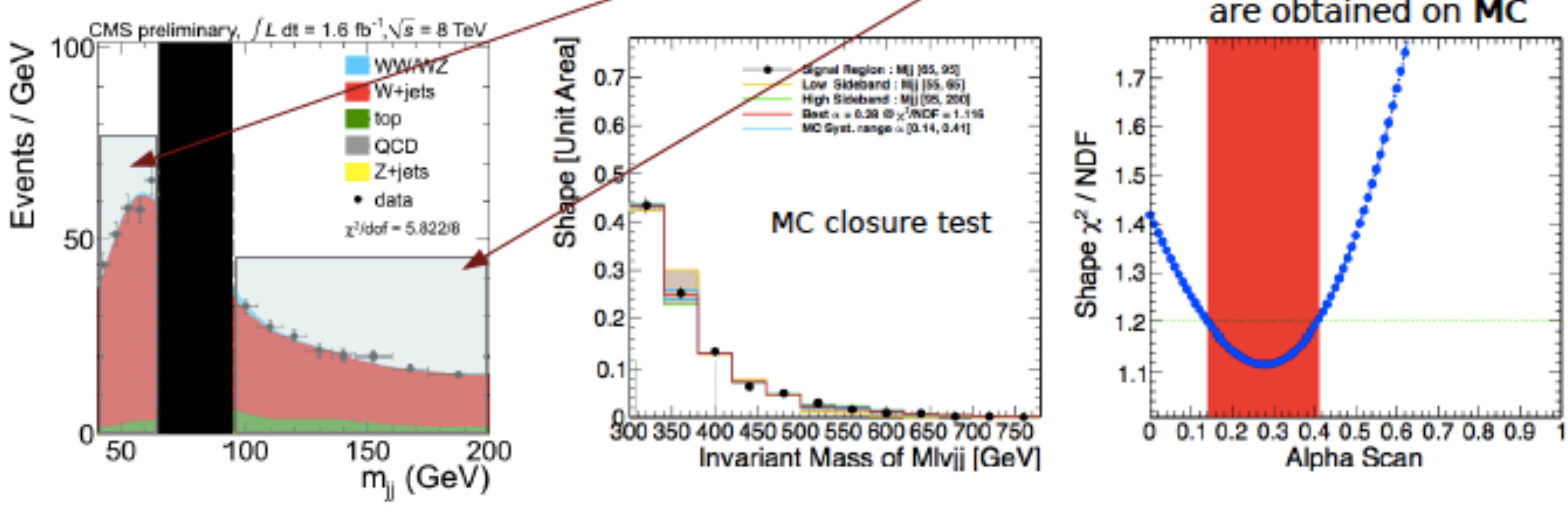
- Following the sideband method devised for **HIG-12-003**

- For **each** of the **48 working points**, 3 regions are defined
  - $m_{jj} \in [65,95]$ GeV/c² (signal region)
  - $m_{jj} \in [55,65]$ GeV/c² (lower sideband, SBL)
  - $m_{jj} \in [95,115]$ or $[95,200]$ GeV/c² for $m_H <$ or ≥ 250 GeV/c² (higher sideband, SBH)

- The shape is extracted, in each bin *i*, as

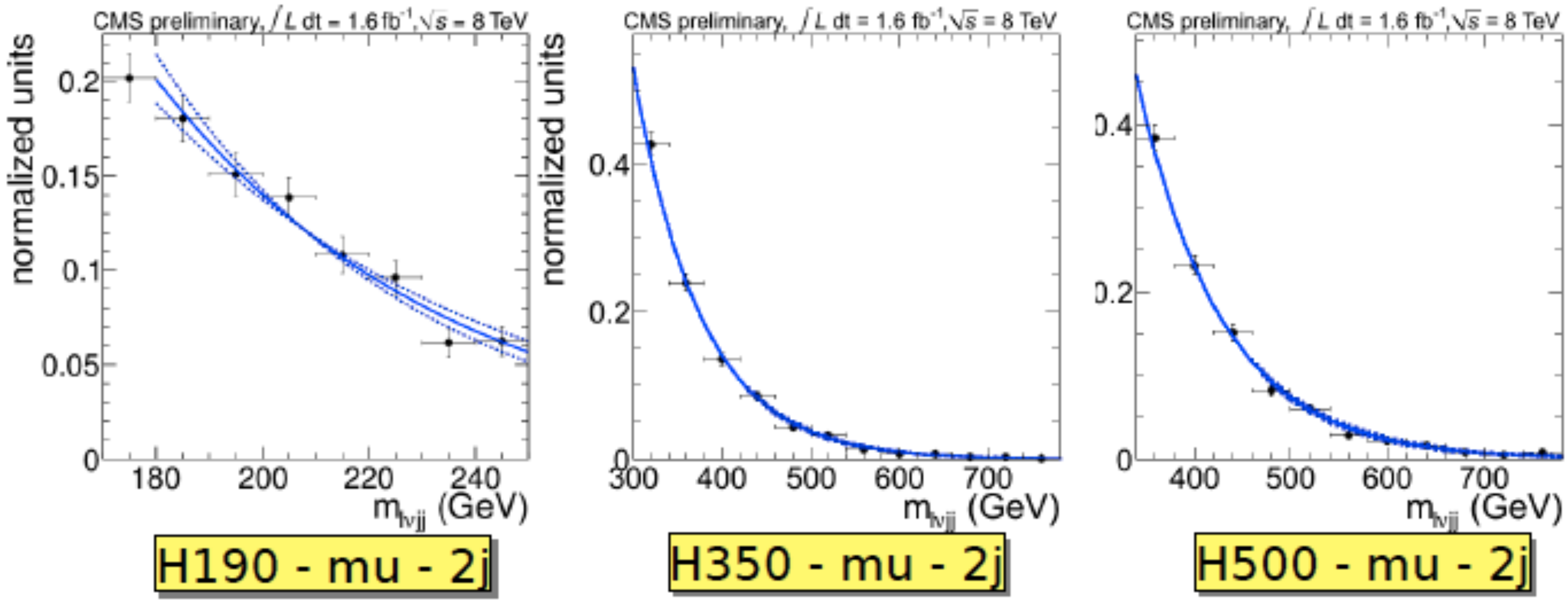
$$m_{l\nu jj}{}^{i} = \alpha \cdot m_{SBL}{}^{i} + (1-\alpha) \cdot m_{SBH}{}^{i}$$

the **best α value**, as well as its **uncertainty**, are obtained on **MC**

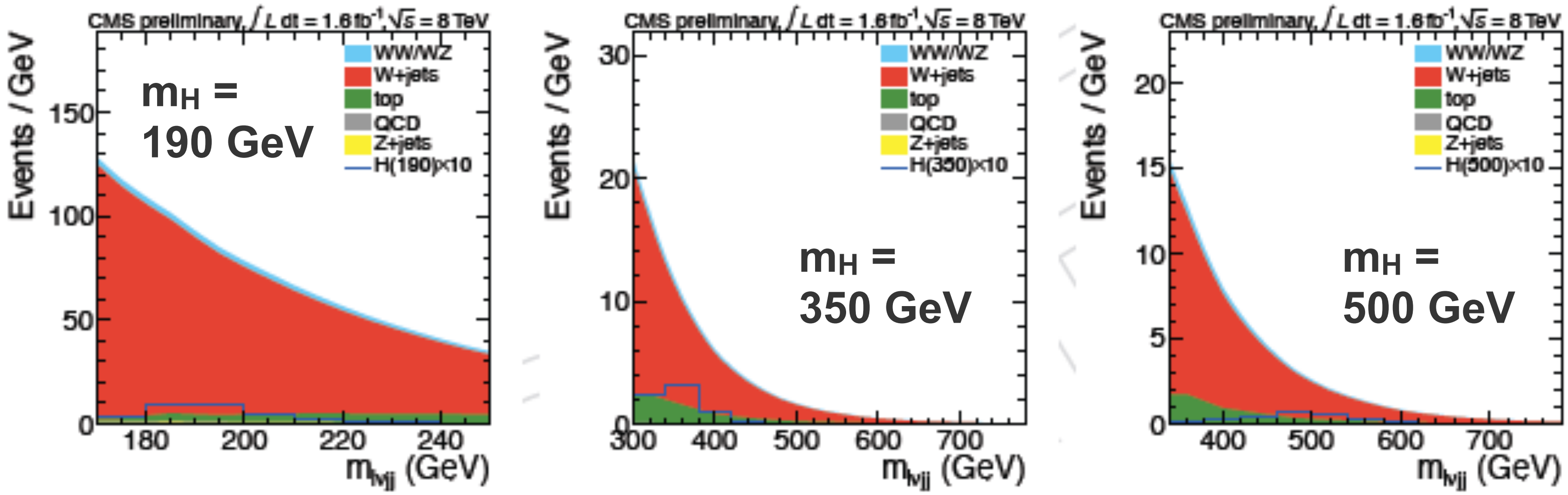


W+jets shape (2/2)

- The background-subtracted, alpha-combined sideband shape from data is smoothed using an exponential function.
  - The statistical uncertainty of the smoothing is **combined** with the uncertainty due to α and used as a **systematic error**
  - The dotted lines are the total shape systematic envelope.



The m_{WW} spectrum



Mass peak makes it straightforward to interpret any observed excess in data.

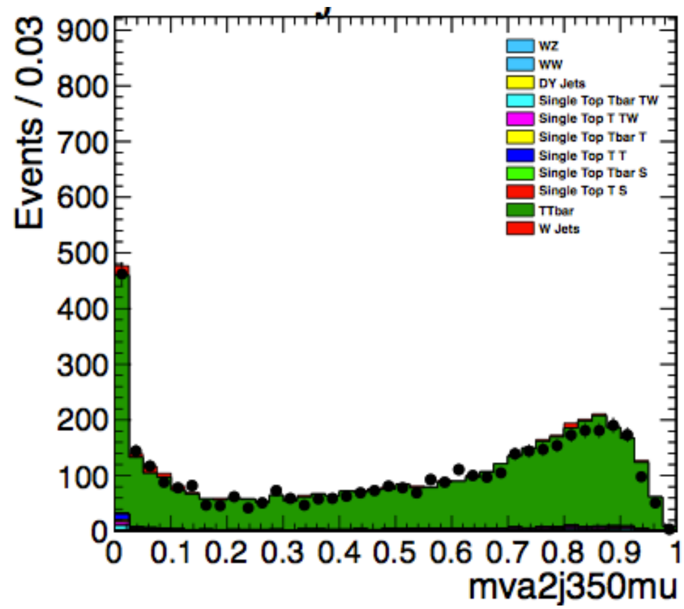
Systematic uncertainties

Source	**uncertainty**
Higgs line shape	0 – 30%
Signal cross-section	15 – 20%
Signal efficiency x acceptance	10%
Luminosity	5%
Jet energy scale, resolution and MET	< 1%
Theory (PDFs)	1 – 2%
Lepton trigger efficiency	1%
Lepton selection efficiency	1 – 2%
Pile-up	< 1%

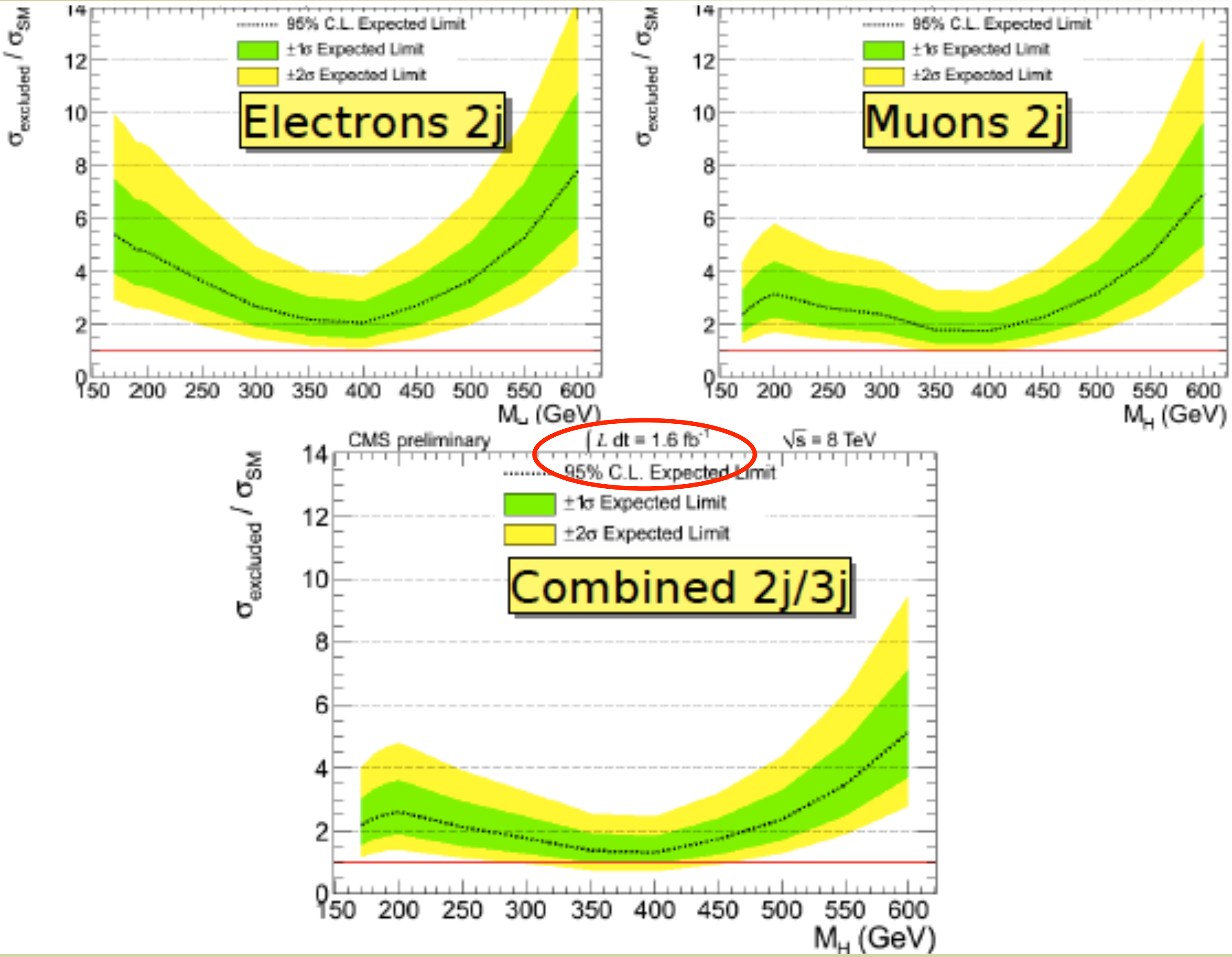
Background systematics

normalization uncertainty	<= 2 %
W+jets fit uncertainty	shape

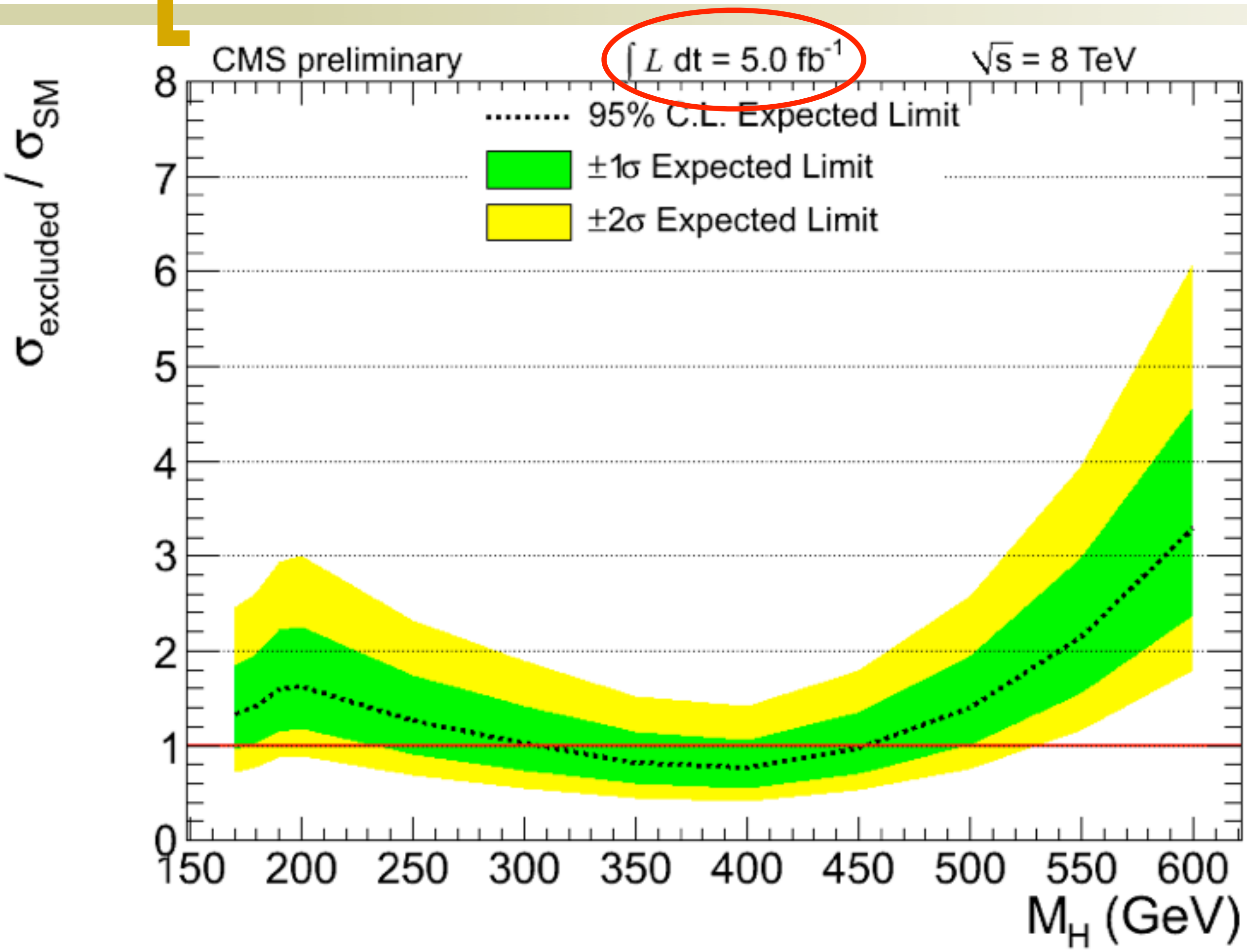
- Efficiency x acceptance syst. is evaluated, using a pure top sample, as the difference between data and MC.



Limit using 1.6 fb^{-1} 8 TeV data



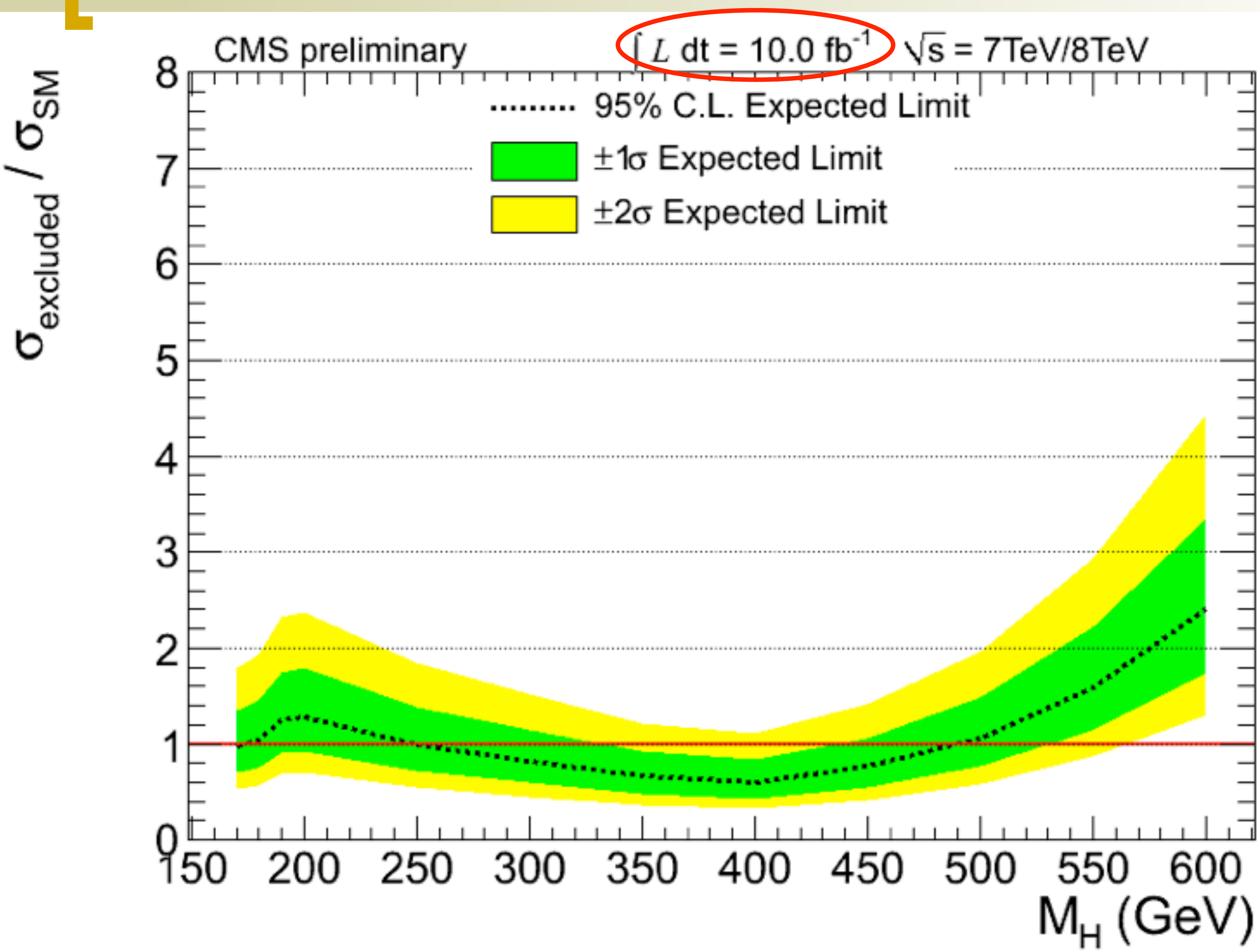
Projection to 5 fb^{-1}



✦ About 10−20% improvement over 2011 sensitivity

- higher Higgs xsection at 8 TeV
- improved trigger

Combination of 7 TeV and 8 TeV results



Summary

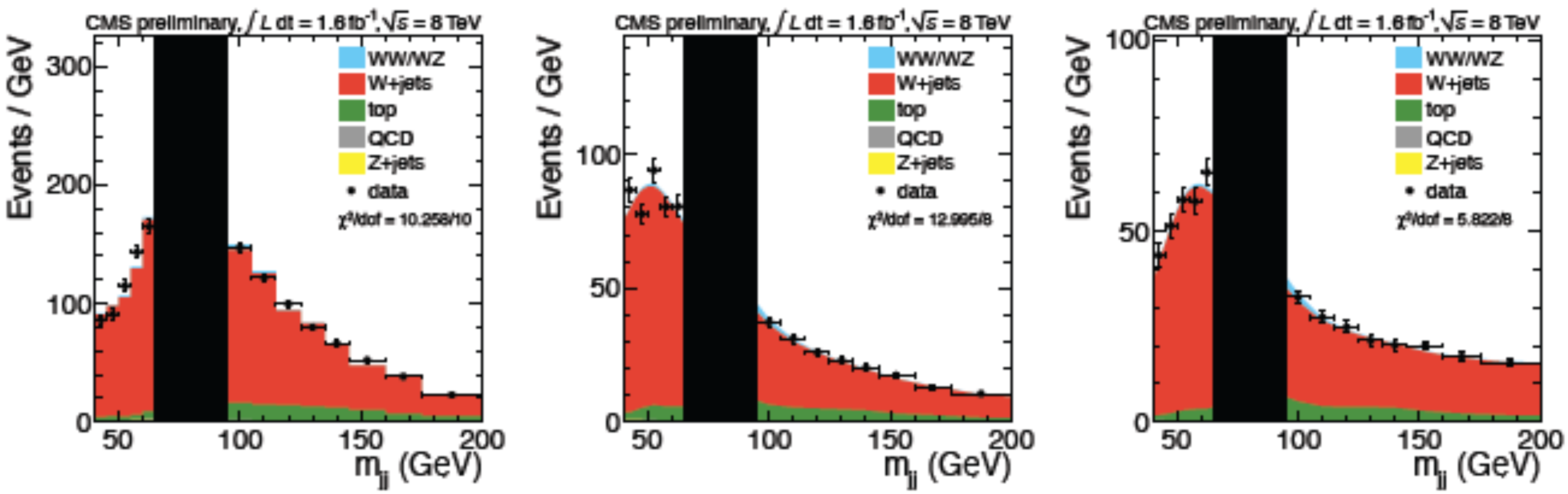
- The 2012 ICHEP analysis for Higgs searches in the **WW → lν jj** has been presented
  - New 2012 data well understood
  - MC simulation issues addressed and covered for both extraction techniques
  - Signal and background systematics re-evaluated with new samples

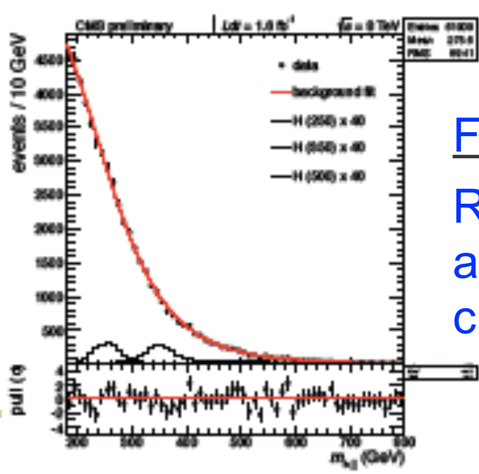
- **Two** subsequent **limit extractions** presented with 1.6 fb^{-1} of data:
  - a first one, after baseline mass-independent cuts
  - a second one, after optimized mass-dependent likelihood selections

- Ready to look into the signal region

Unblinding procedure

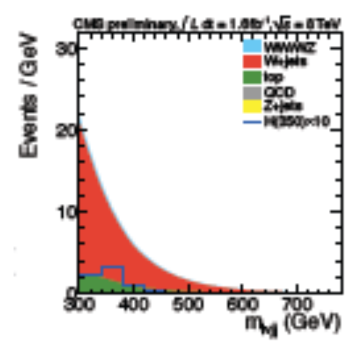
Open the box in the signal window (i.e., remove the cover) and derive observed limits





Fit-based analysis

Remake this plot after applying mjj cut. Then set limit.



MVA-based analysis

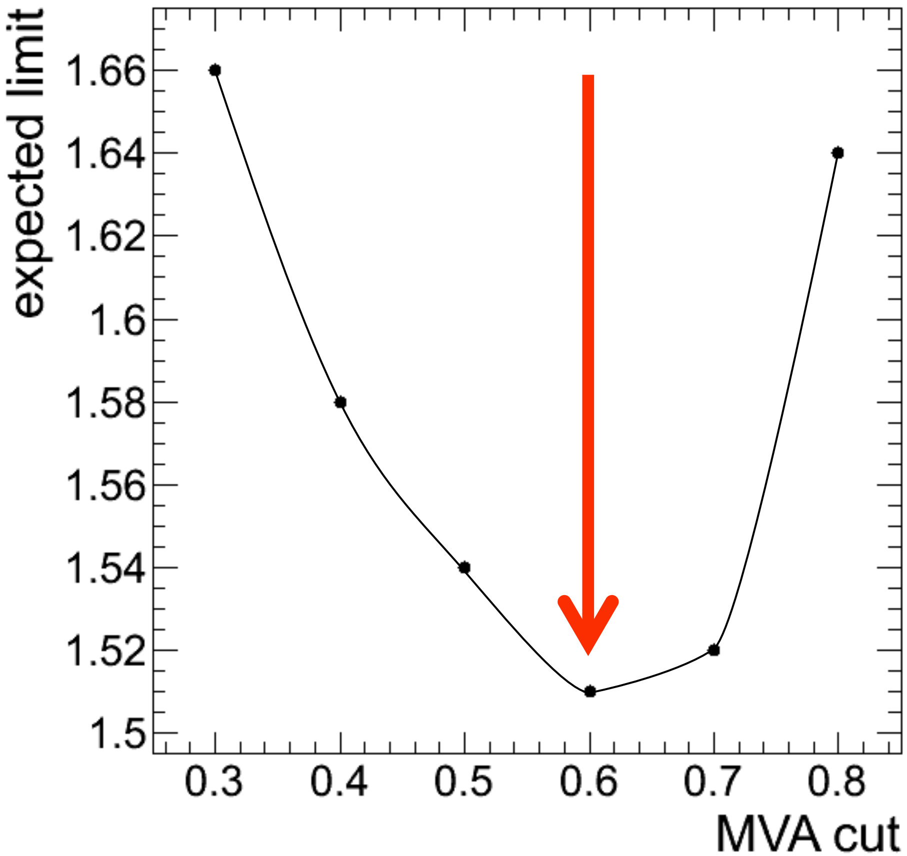
Plot the data points from the blackened signal region. Then set limit.

backup slides

MVA cut optimization

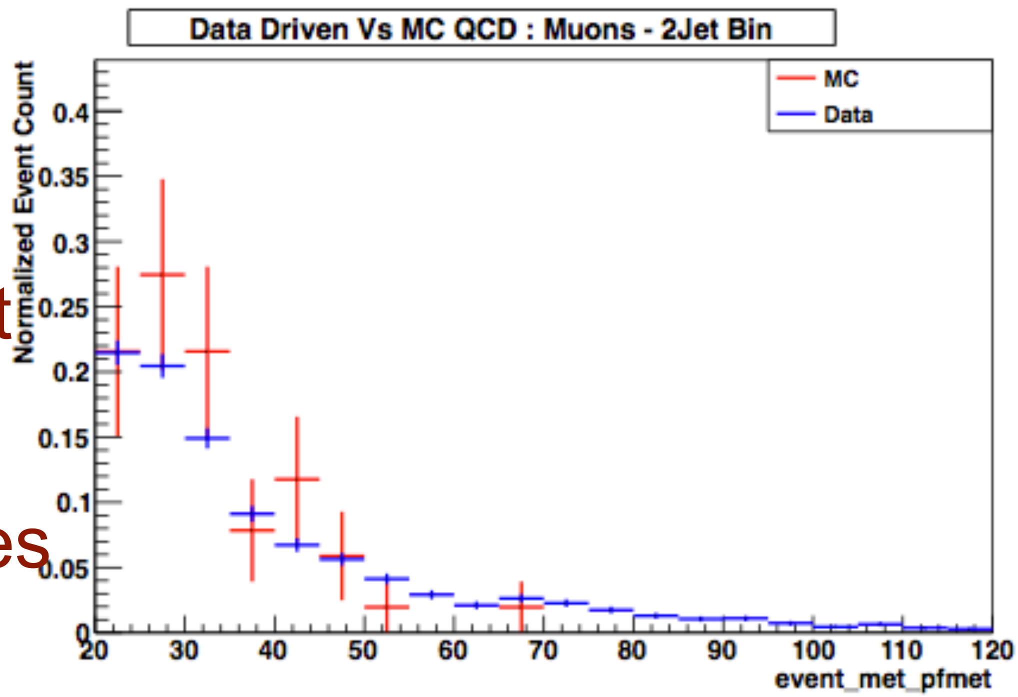
- We select the MVA cut value based on running the full asymptotic limit setting machinery and using the expected limit.
- Once the optimal cut is selected for each of the 48 analysis points they can be combined using the standard Higgs combination package.

muon, 2-jets, m_H = 350 GeV



Data driven QCD estimation

- We derive the QCD shape and normalization from the data.
  - invert the isolation requirements
  - relax ID requirements
  - relax the MET cut
- We can fit the MET distribution in data to get the normalization of the QCD contribution after accounting for differences due to the MET cut.
- The shape is also taken from this data as the MC is statistics starved.



W+jets background

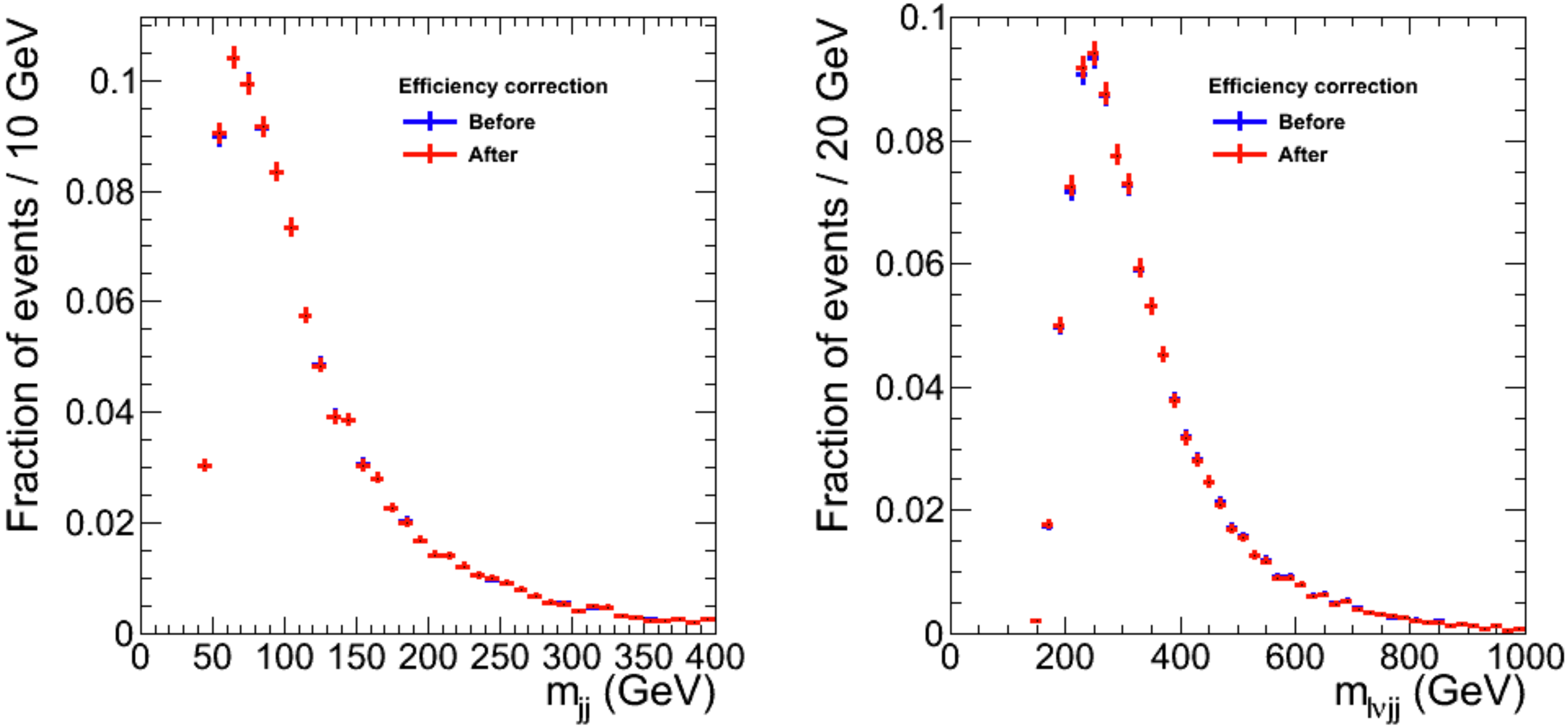
- The W+jets process is the dominant background.
- For $m_H \leq 180$ GeV, MC is used as the shape template, because statistics in this case are plentiful.
- For higher masses MC statistics are much lower so we take an analytic approach.
- The shapes are inspired by MC but ultimately the parameters of the functions are determined from the fit to the data.

$$\mathcal{F}_{\mathrm{W+jets}} = \mathrm{erf}(m_{jj}; m_0, \sigma) \times \left[(m_{jj})^{-\alpha-\beta \ln(m_{jj}/\sqrt{s})}\right]$$

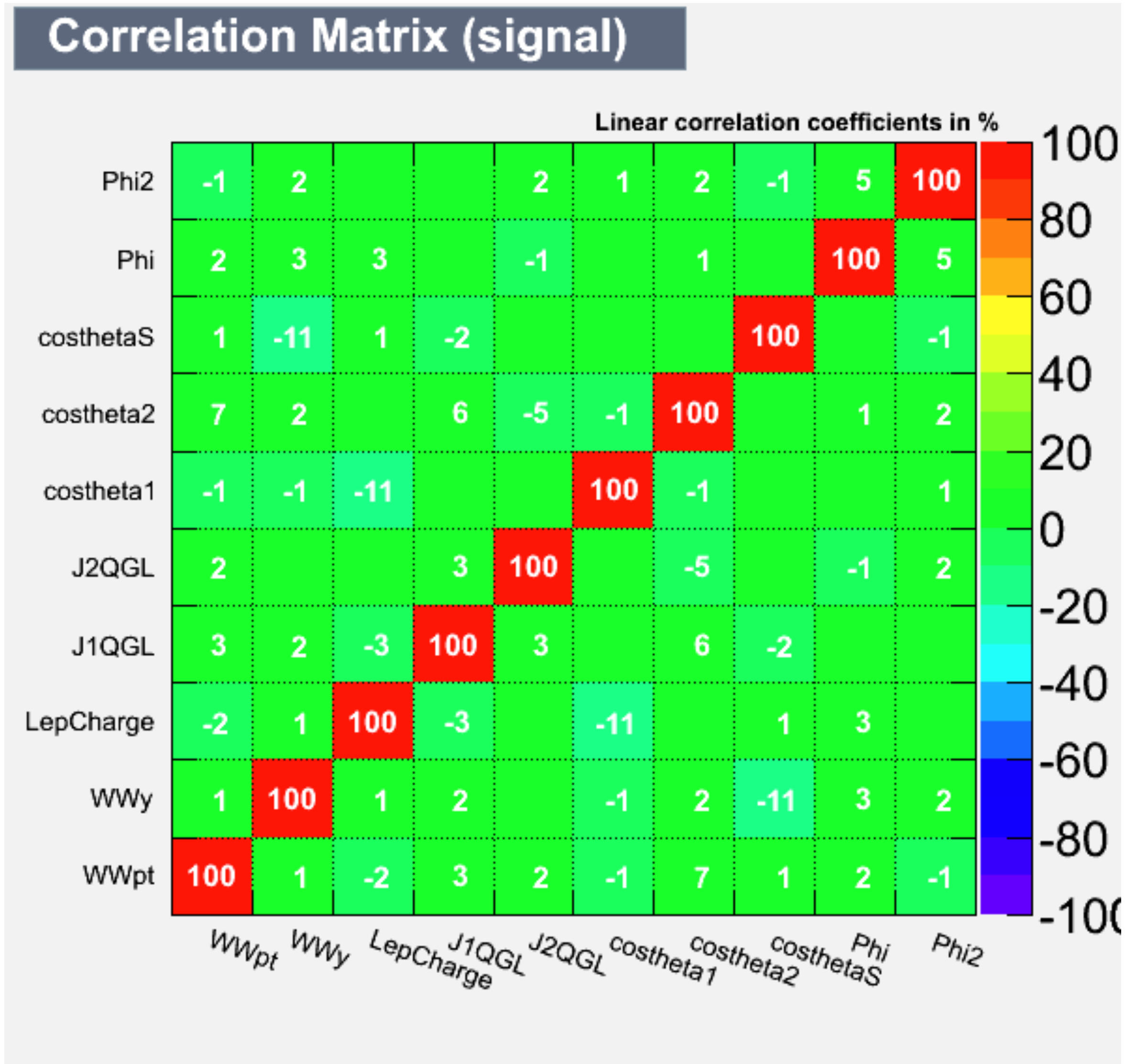
$$\mathcal{F}^{\text{low mass}}_{\text{W+jets, 2-jets}} = \mathrm{erf}(m_{jj}; m_0, \sigma) \times (m_{jj})^{-\alpha} \times e^{m_{jj}\tau}$$

$$\mathcal{F}^{\text{low mass}}_{\text{W+jets, 3-jets}} = (m_{jj})^{-\alpha-\beta \ln(m_{jj}/\sqrt{s})} \times e^{m_{jj}\tau}$$

Trigger effect on key distributions



MVA correlations



Likelihood selection efficiency