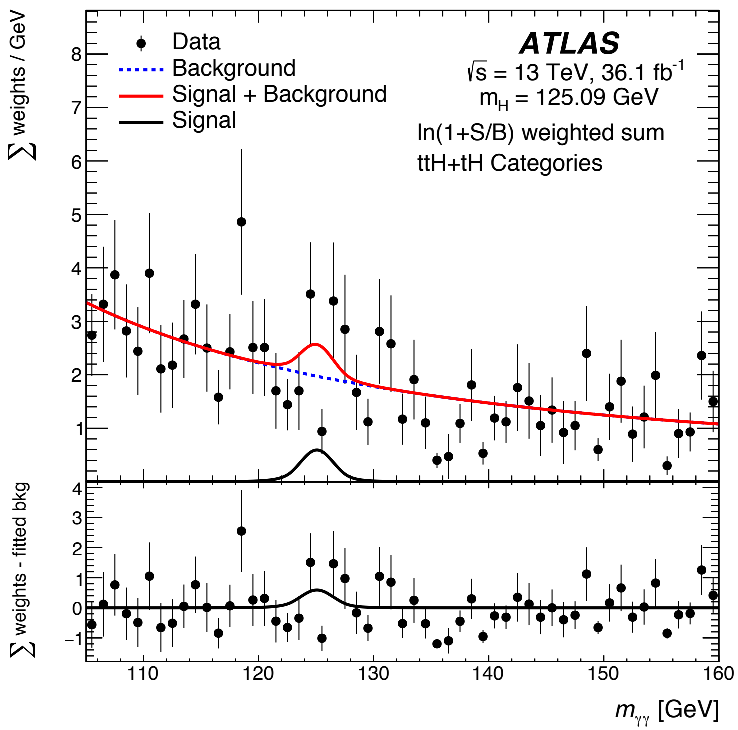
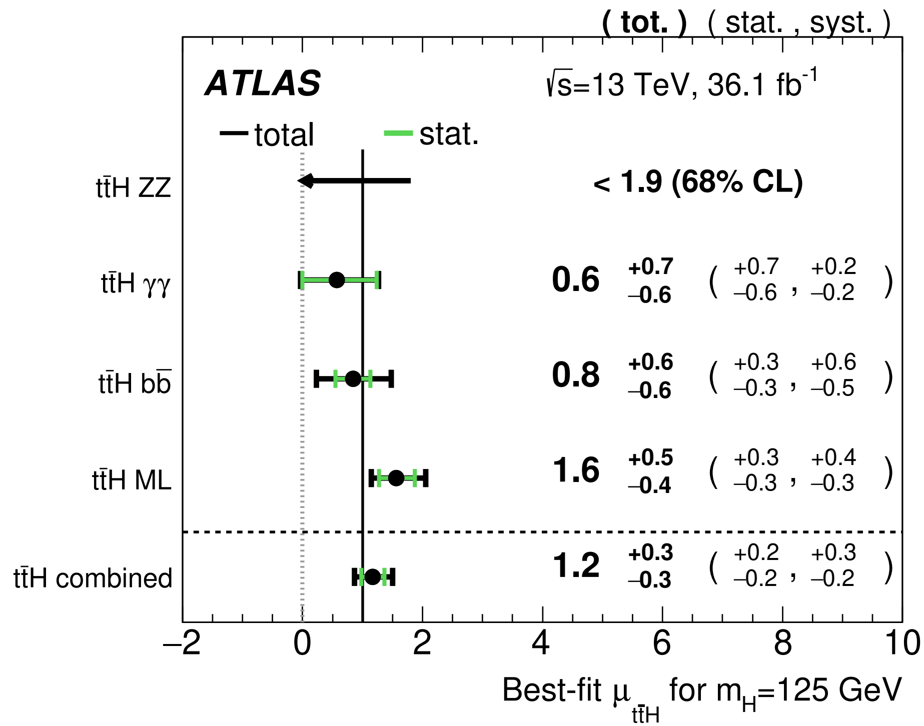
Data driven background estimation for ttH → γγ measurements

Haichen Wang
Lawrence Berkeley National Laboratory

Higgs Toppings Workshop
Probing Top-Higgs Interaction at the LHC
Benasque, Spain
28-May, 2018

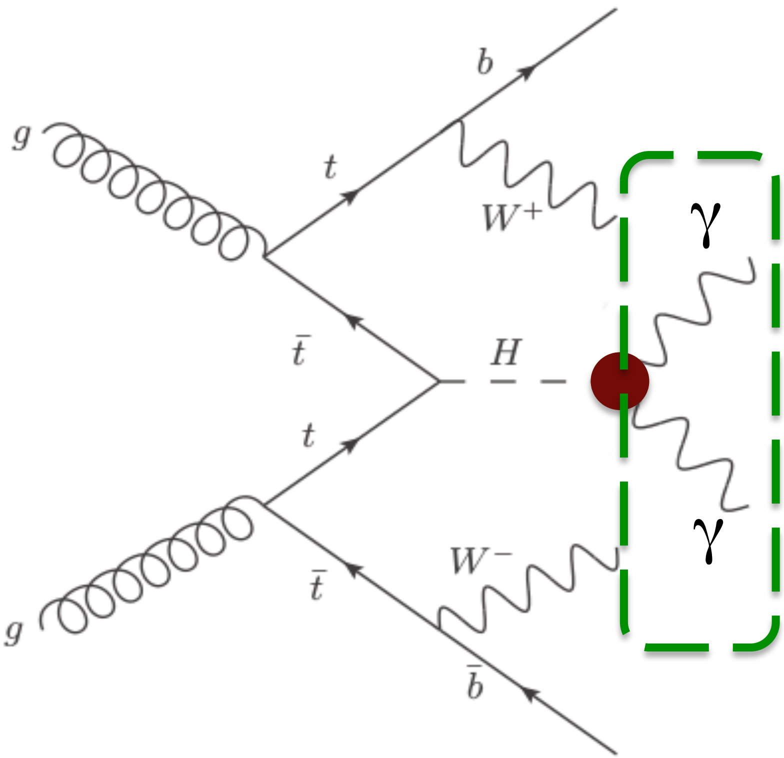
Understanding of the background in the ttH γγ analysis becomes increasingly important

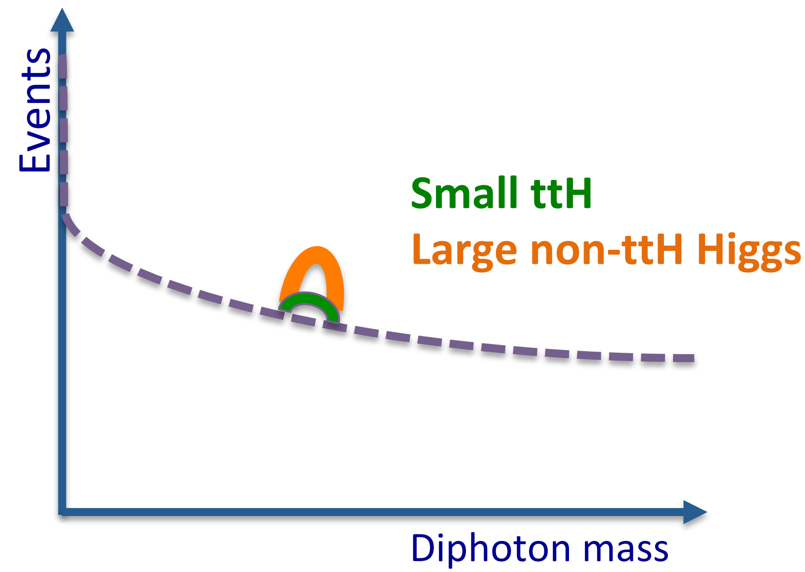




- Analysis recap
- ***Continuum*** background composition
- ***Resonant*** background (non-ttH Higgs) uncertainty

We start with events with two good photons

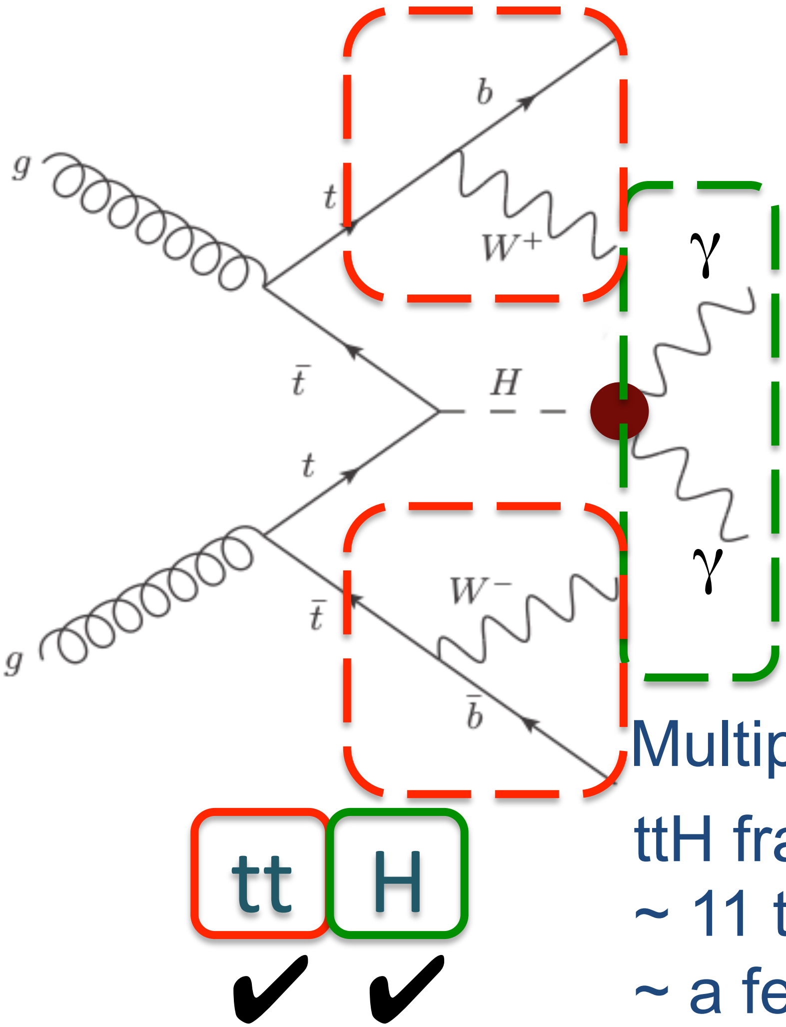




tt H ✔

~ 1% signal from ttH production,
~16 ttH signals, ~ 1700 SM Higgs,
~ 2 million continuum background
@36 fb^{-1}

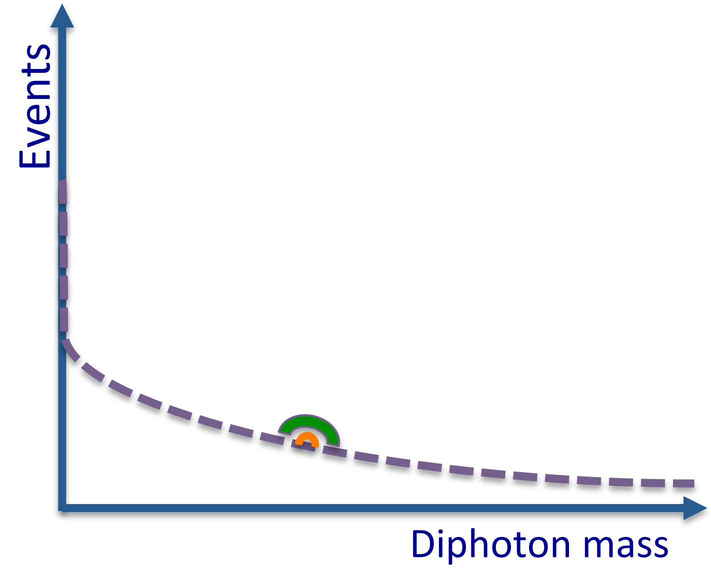
After requiring the event very likely to have a ttbar system -



SM continuum bkg reduced

Non-ttH Higgs now small

ttH remains



Multiple signal regions

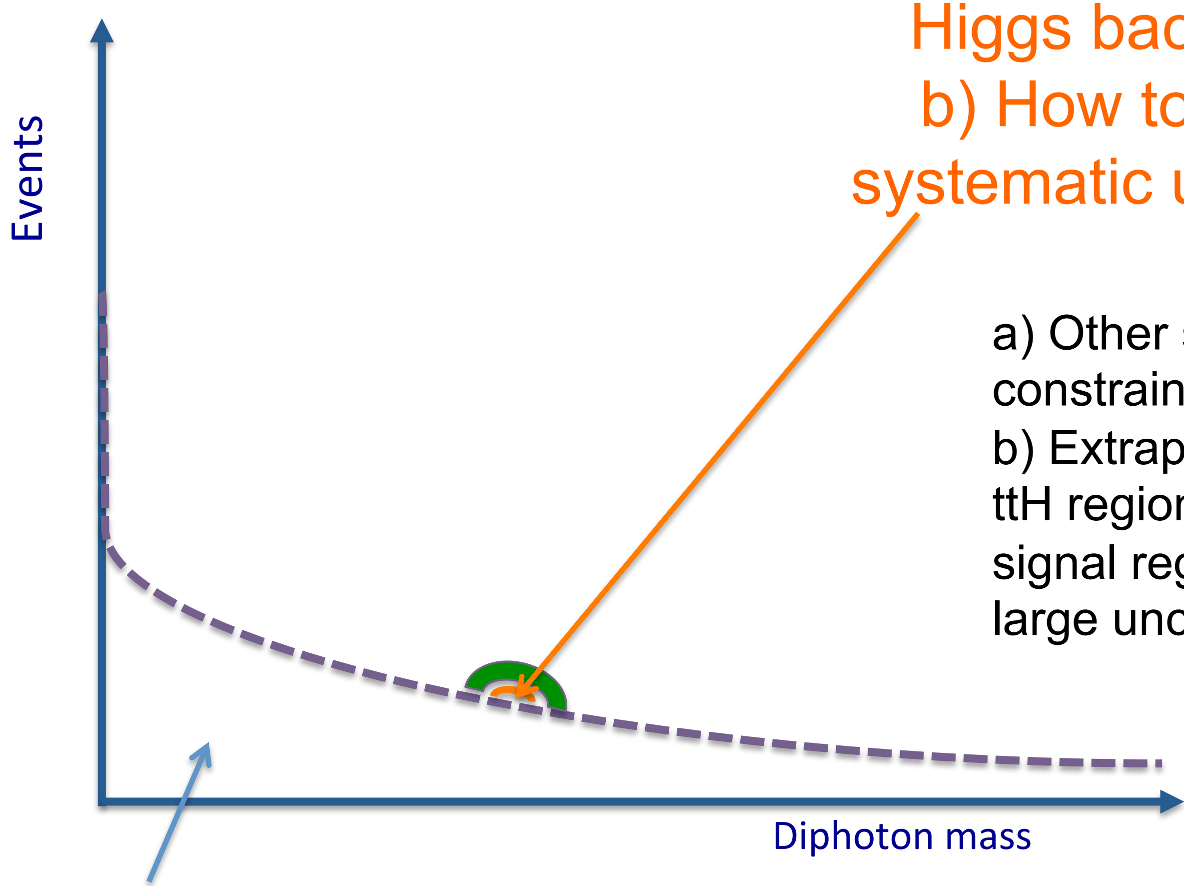
ttH fraction can be as high as 90%

~ 11 ttH signal divided into 7 categories

~ a few hundred continuum bkg evts @ 36fb^{-1}

What do we care about?

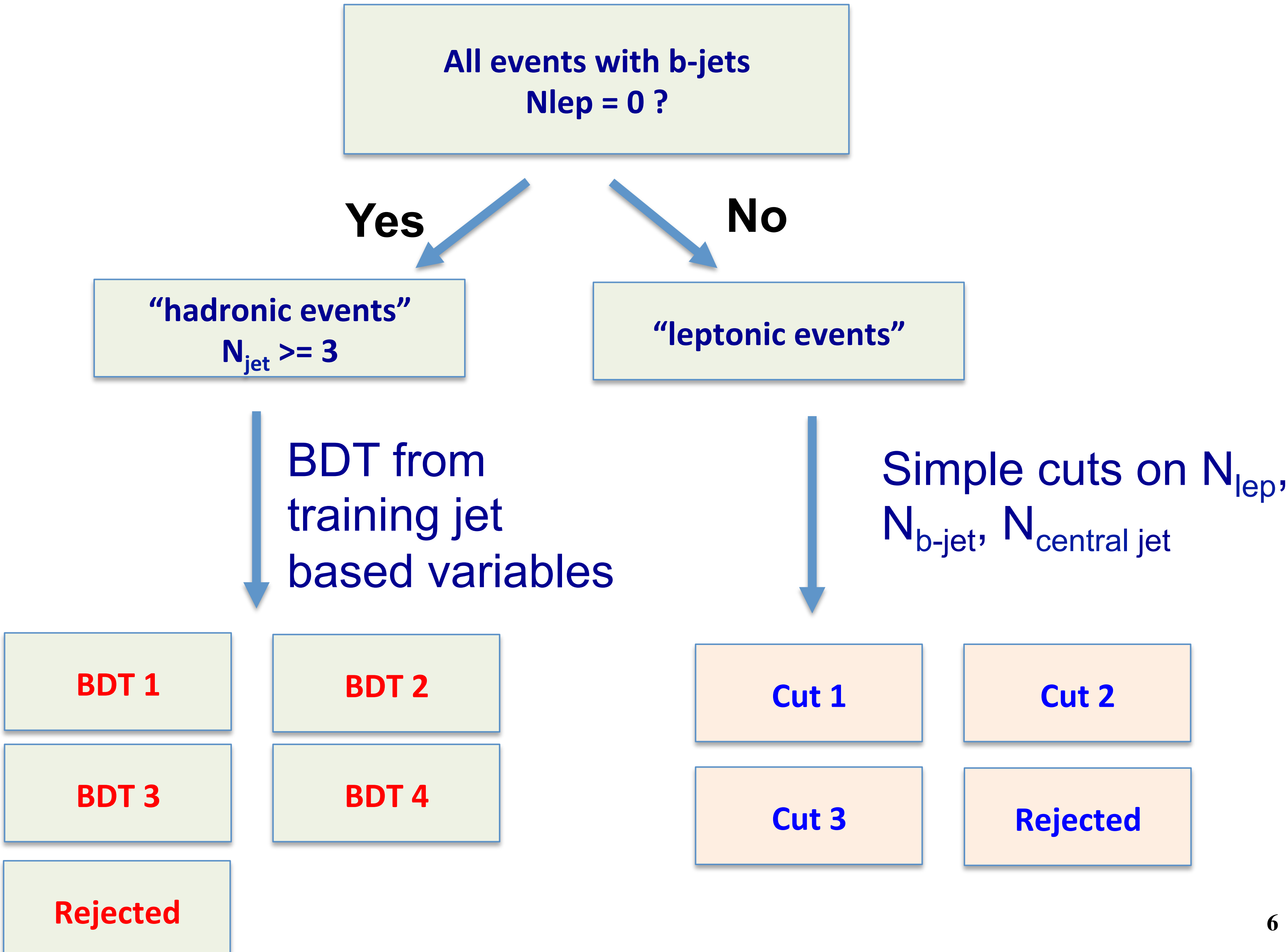
a) How large is the non-ttH Higgs background?
b) How to assign a systematic uncertainty?



a) Other signal regions constrain ggH, VBF, VH
b) Extrapolation from non-ttH regions to ttH-like signal regions subject to large uncertainties

What is the composition of the continuum background?
Is tt + γγ a large fraction? If so, future optimization will target the difference between ttH and tt+ISR/FSR photons

Categorization



Continuum background

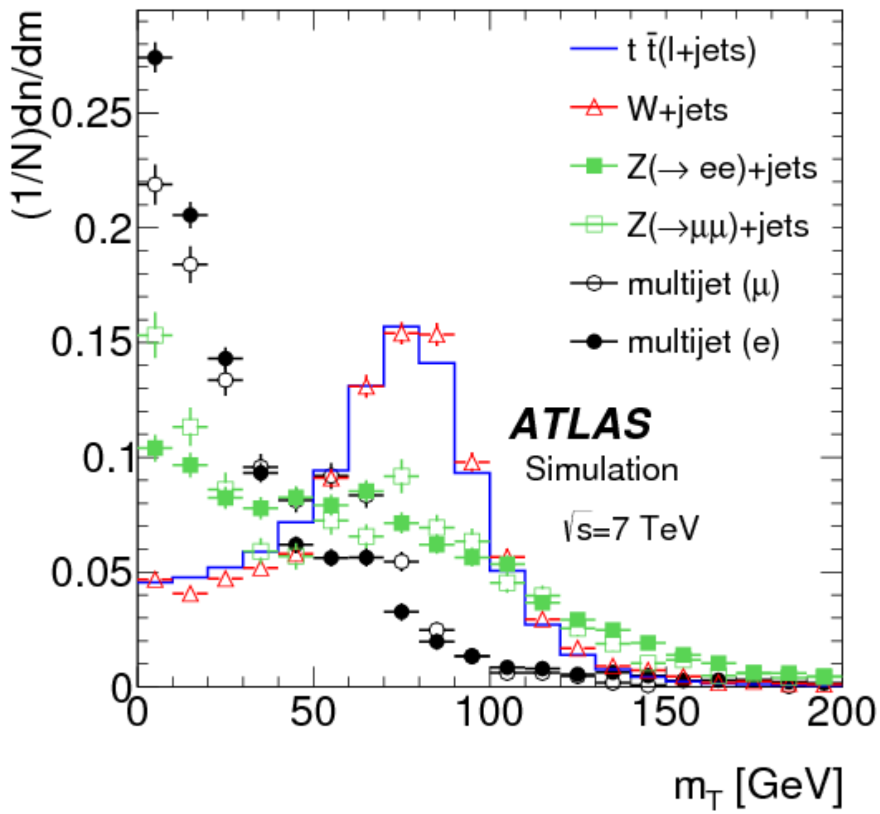
- Background normalization under the Higgs peak → well established fit procedure

- Background composition → an accurate understanding allows targeted optimization
  - Future analysis improvement may focus on the separation between ttH and tt + FSR / ISR photons

- No public statement on the composition was made by the LHC experiments
  - Some ideas on using templates to decompose background for events passing preselection
  - One can imagine the tt + γγ background is large, if we are getting close to tt H

Background composition – leptonic channel

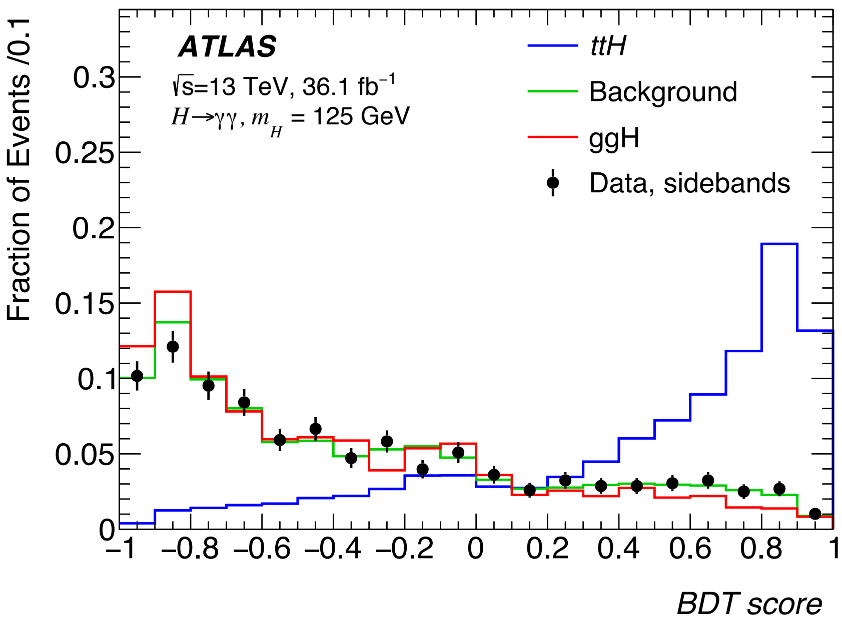
- Leptonic events
  - Nbjet > 0 , Nlep > 0; Additional requirements for categorization
  - tt + γγ should not be trivial
  - Not clear about V + γγ + b
- m_T template decomposition
  - m_T shape different between top, jet fakes
  - tt + γγ is the majority of the background here
- What can be useful for us
  - A good modeling of ttgamma gamma



Background composition

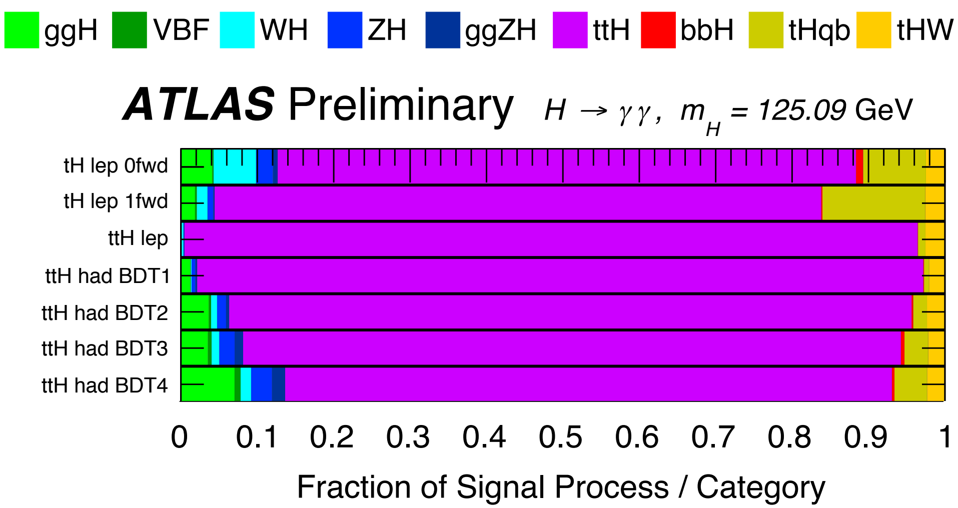
- Hadronic events
  - $N_{jet} \geq 3$, $N_{bjet} \geq 1$, $N_{lep} = 0$, additional BDT requirements to categorize events
  - Overwhelming majority γγ+ jets
  - A small fraction can be tt + γγ

- No obvious candidate observable for template fit
  1) Multivariate discriminant
  (based on jet variables) can be used (see T. Schwarz and R. Hyneman's talk)
  2) Template shapes show separation between top and jets
- Best categories have very large top contributions



Resonant Higgs processes contribution

- Non-ttH Higgs events is a non-trivial fraction of the Higgs events in the ttH categories
  - ~ 10-20% of the total Higgs events



- Central values of these background processes are constrained by other Higgs signal regions
- Question is about how well we can extrapolate from other regions to the ttH regions?

  - Currently, 100% is assigned, which we hope we can reduce in the next round

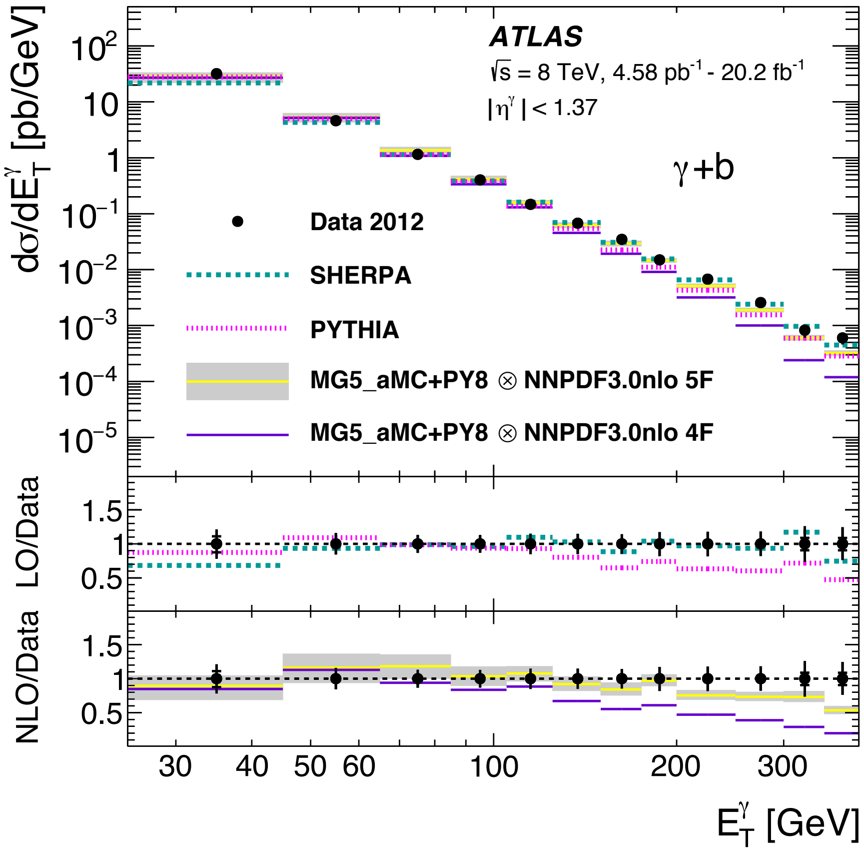
Resonant Higgs processes contribution

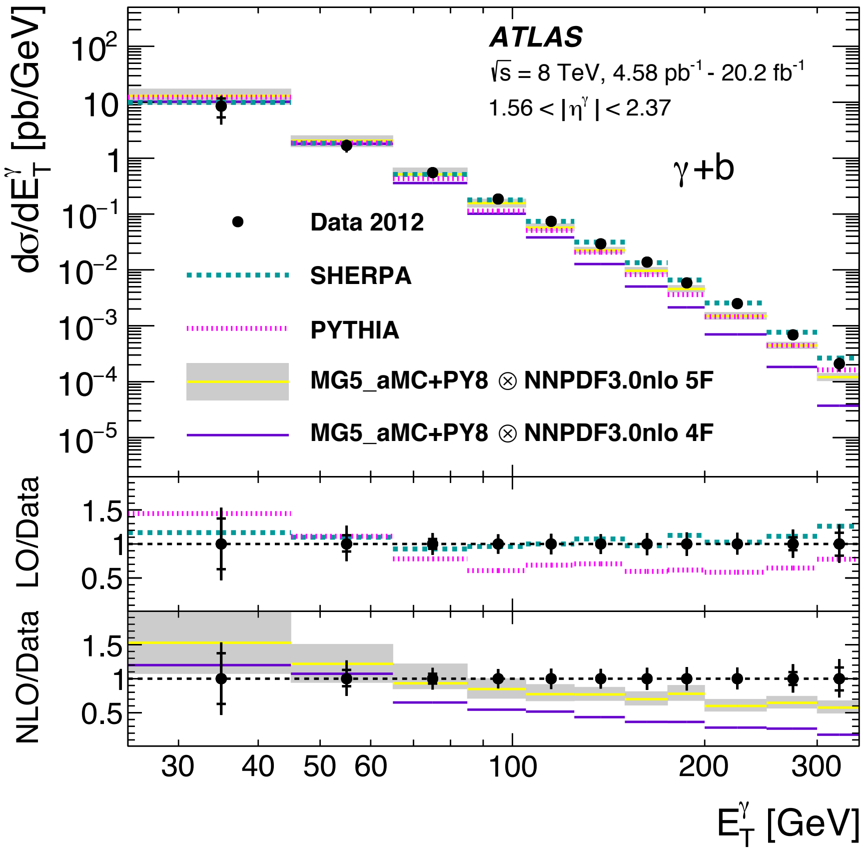
- What do we know about processes with b-jets coming from sources similar to b-jets in the non-ttH Higgs events, from data?

  - LHC vector boson + jets measurements
  - H ($\rightarrow$ 4l) + b-jet measurement
  - Data/MC comparison in diphoton sidebands

- The central value is not a concern – determined from combined fits of categories optimizing for different production modes
- The question is how well we can extrapolate from more inclusive phase space to the ttH phase space, specifically, H + b-jets

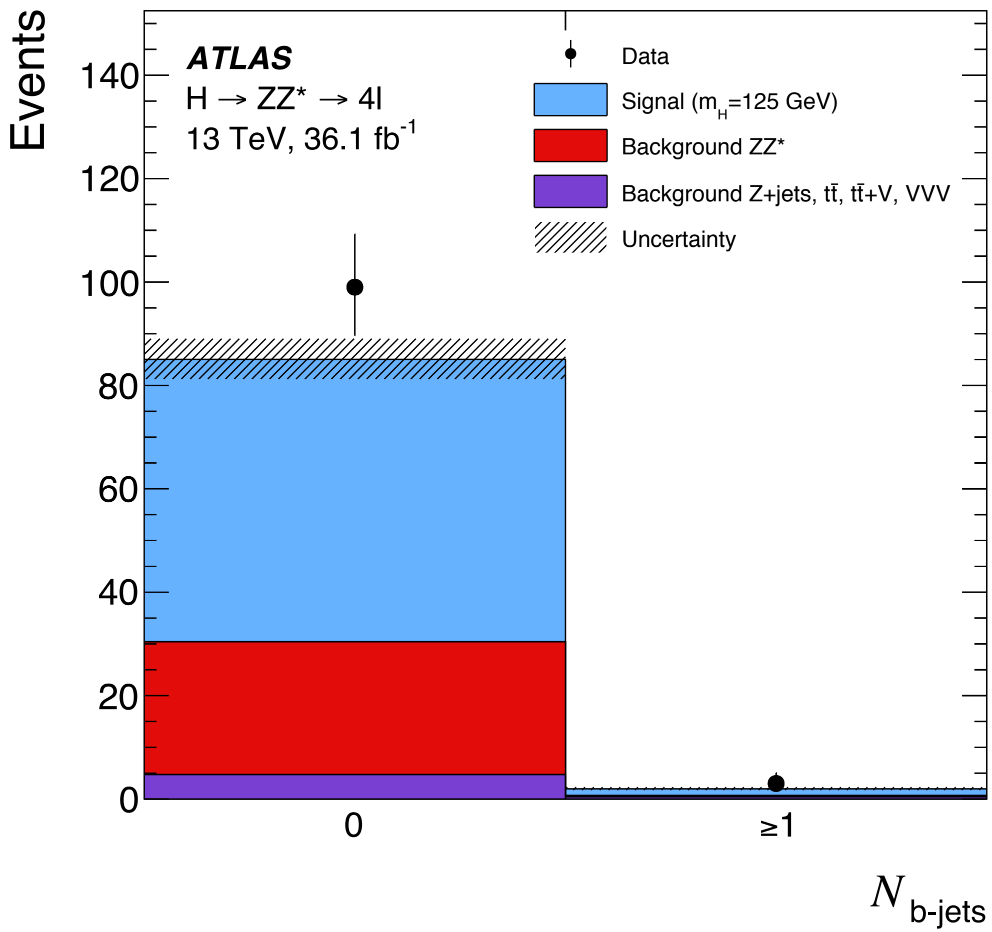
- **Single photon + bjets measurement**
  - Larger discrepancy for high E_T photon in forward region





Resonant background estimate

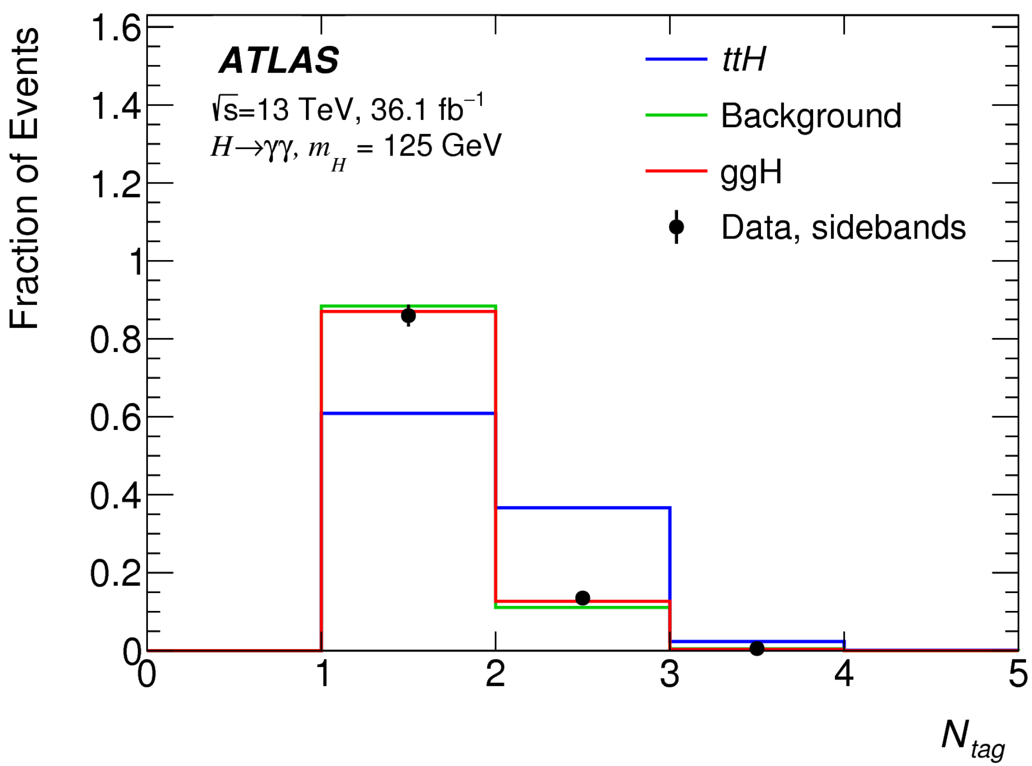
- **Use H → 4l measurement – orthogonal to diphoton data set – full data driven information**



Statistics limited – can give an upper bound on the σ/σ_{SM} of < ~ a couple hundreds percent for N-bjet > 0

Resonant background estimate

- **Compare data and MC (Sherpa) sample in the diphoton mass sidebands (| $m_{\gamma\gamma}$ -125 | > 5 GeV)**
- Use the difference to provide an estimate



Left plot has everything normalized to 1 with N_{bjet} > 0

Different generators used for signal and background

Currently, a 100% uncertainty is assigned to the contamination of non-ttH production mode

Summary

- **Continuum background**
  - How large is the fraction of tt + γγ events?
  - Data-driven methods indicate that in the signal regions that tt + γγ is majority
  - Better modeling of tt + γγ, tt + γ will be helpful

- **Resonant background** – non-ttH + b-jets
  - The current 100% uncertainty is probably too conservative
  - Data/MC comparison indicates smaller discrepancy
  - Input from theoretical community useful

  - Uncovered – tH background; perhaps hard to estimate from data?

References

- Reference :
  - ATLAS HIGG-2016-21 (https://arxiv.org/abs/1802.04146) , submitted to PRD
  - CMS HIG-16-040 (https://arxiv.org/abs/1804.02716), submitted to JHEP

ATLAS MC Generators

Process	Generator	Showering	PDF set	σ [pb] $\sqrt{s} = 13$ TeV	Order of calculation of σ
ggH	Powheg NNLOPS	Pythia8	PDF4LHC15	48.52	$N^3LO(QCD)+NLO(EW)$
VBF	Powheg-Box	Pythia8	PDF4LHC15	3.78	NNLO(QCD)+NLO(EW)
WH	Powheg-Box	Pythia8	PDF4LHC15	1.37	NNLO(QCD)+NLO(EW)
$q\bar{q}' \to ZH$	Powheg-Box	Pythia8	PDF4LHC15	0.76	NNLO(QCD)+NLO(EW)
$gg \to ZH$	Powheg-Box	Pythia8	PDF4LHC15	0.12	NLO+NLL(QCD)
$t\bar{t}H$	MG5_aMC@NLO	Pythia8	NNPDF3.0	0.51	NLO(QCD)+NLO(EW)
$b\bar{b}H$	MG5_aMC@NLO	Pythia8	CT10	0.49	5FS(NNLO)+4FS(NLO)
t-channel tH	MG5_aMC@NLO	Pythia8	CT10	0.07	4FS(LO)
W-associated tH	MG5_aMC@NLO	Herwig++	CT10	0.02	5FS(NLO)
$\gamma\gamma$	Sherpa	Sherpa	CT10		
$V\gamma\gamma$	Sherpa	Sherpa	CT10		

CMS MC Generators

Simulated signal events are generated using MADGRAPH5_aMC@NLO v2.2.2 at next-to-leading order (NLO) [26] in perturbative quantum chromodynamics (QCD) with FxFx merging [27], the parton level samples being interfaced to PYTHIA8.205 [28] for parton showering and hadronization. The CUETP8M1 PYTHIA underlying event tune parameter set is used [29]. Events produced via the gluon fusion mechanism are weighted as a function of the Higgs boson p_T and the number of jets in the event, to match the prediction from the NNLOPS program [30]. Parton distribution functions (PDFs) are taken from the NNPDF3.0 [31] set. The signal cross sections and branching fraction recommended by the LHC Higgs cross section working group are used [32].

The dominant background to $H \rightarrow \gamma\gamma$ consists of the irreducible prompt diphoton production, and the reducible backgrounds from γ + jet and dijet events where the jets are misidentified as isolated photons. Background events, used for the trainings of multivariate discriminants and

for category optimization, have been simulated using various event generators. The diphoton background is modeled with the SHERPA v.2.2.1 [33] generator. It includes the Born processes with up to 3 additional jets as well as the box processes at leading order. Multijet and γ + jet backgrounds are modeled with PYTHIA, with a filter applied to enhance the production of jets with a large fraction of electromagnetic energy. The $W\gamma$ and $Z\gamma$ samples are generated with MADGRAPH5_aMC@NLO at leading order, while Drell–Yan events are simulated with the same generator at NLO precision.