A data-driven background estimation method in $H \to Z\gamma$

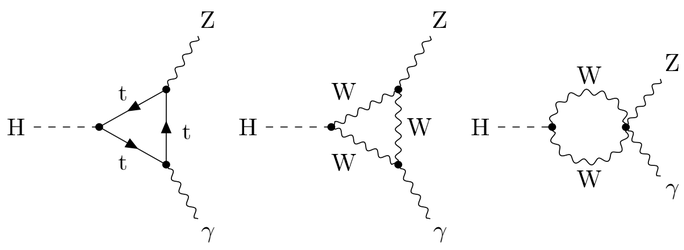
Zhanxiang Gao, Ligang Xia, Shan Jin

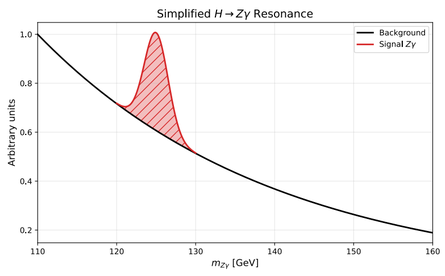
NanJing University

zhanxiang.gao@cern.ch

Introduction

- Higgs decays into $Z + \gamma$ due to loop diagram is rare with branch fraction 0.15%
- The $Z + \gamma$ signal features an evident peek on a smooth background.





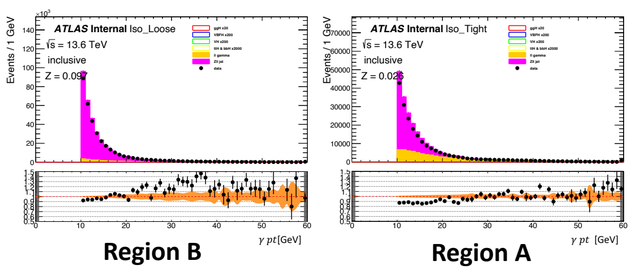
Motivation

- **The signal is searched with the $l^+l^-\gamma$ final state.**
- **Two main backgrounds:**
  - $Z(l^+l^-) + jets$: A jet misidentified as a photon.
  - $Z(l^+l^-) + \gamma$: Irreducible background.
- **Great challenge:**
  - Find a good empiric function to model the background in each signal category.
- **Proposal:**
  - A data-driven method is proposed to estimate these two main backgrounds.

$Z + jets$ (Data-Driven Fake Factor)

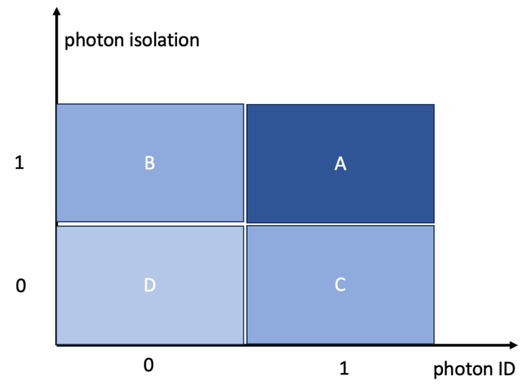
$Z(l^+l^-) + jets$ background estimation (jet →γ)

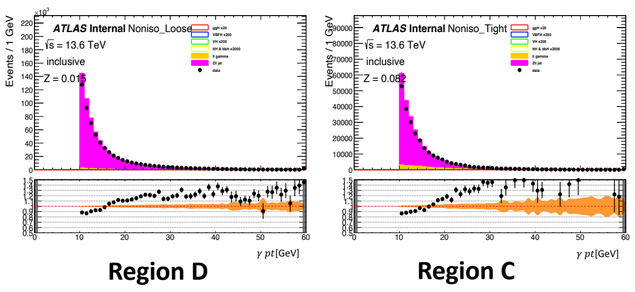
- **ABCD method**:
  - Four categories built according to the photon ID and isolation states.
  - The data in the ”C” region (non-Iso but tight-ID) is used to estimate the background in the “A” region (signal region, isolated and tight-ID).
  - The Scale Factors from C to A (Fake Factors) are obtained from the “B” and “D” regions (loose-ID regions).
  - The Fake Factor (**FF**) is expressed as a function of $\gamma\ pt$



Region B

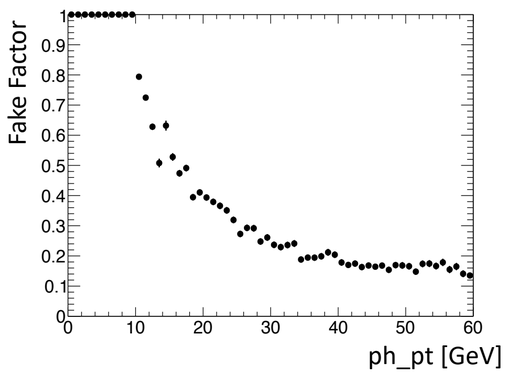
Region A





Region D

Region C



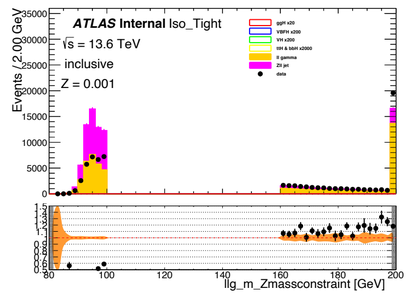
$H \to Z\gamma$(BDT Scale Factor)

$Z(l^+l^-) + \gamma$ background estimation

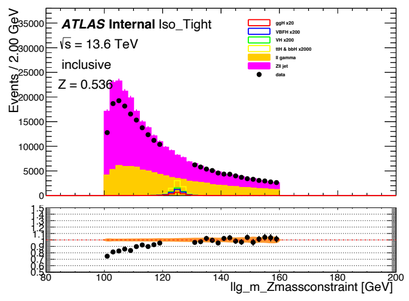
- **MC modified by Scale Factors (SF):**
  - **Control Region (far-sideband):** $M_{Z\gamma} < 100\ GeV$ or $\mathrm{M}_{Z\gamma} > 160$ GeV.
    - Used to extract Scale Factor
    - Dominated by the $Z(l^+l^-) + \gamma$ background.
  - **Scale Factors:** Obtained from the “far-sideband” control region.
    - It is based on a BDT model.
  - **Application to Signal Region**: $100 < M_{Z\gamma} < 160$ GeV

How to do that ?

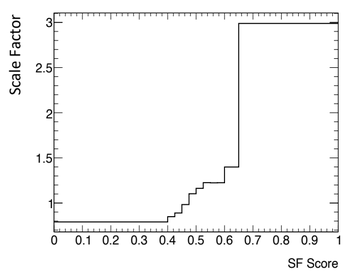
- BDT uses multiple variables as inputs including lepton and photon information since we may have mismodeling on both lepton and photon.
- The BDT-based SF is obtained from the far-sideband CR and applied to the SR.
- The $Z + \gamma$ SF is expressed as a function of a BDT score.



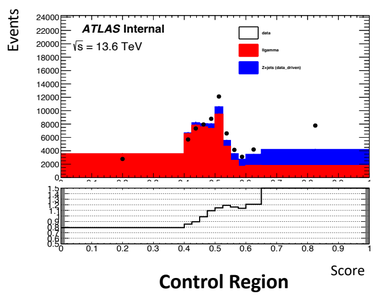
Control Region



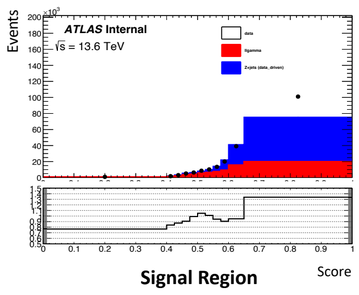
Signal Region



Scale Factor

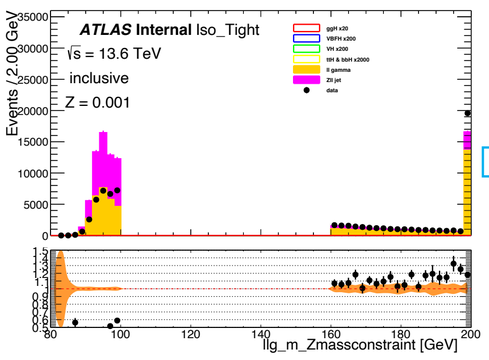


Control Region



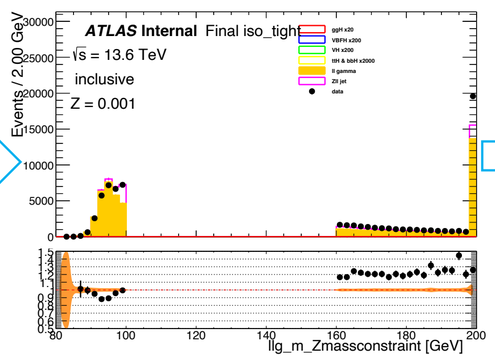
Signal Region

Results



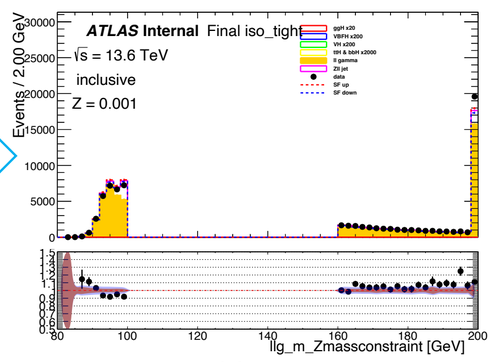
(a)CR

FF

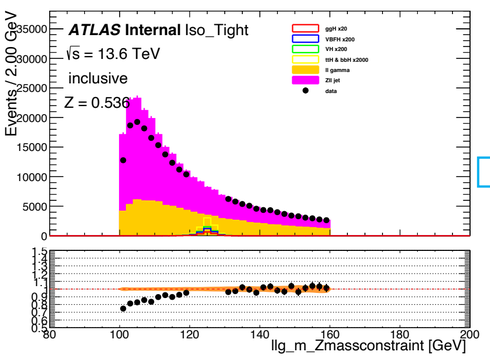


(b)CR

SF

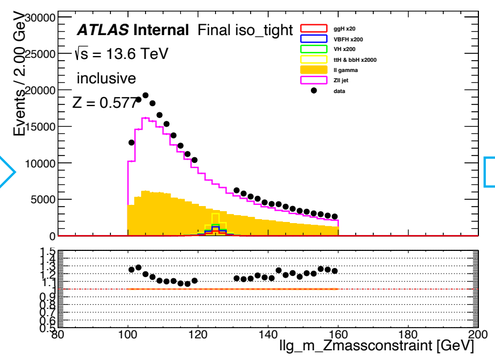


(c)CR



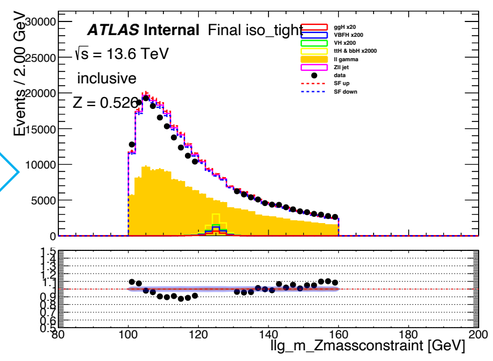
(a)SR

FF



(b)SR

SF



(c)SR

Figures (a) : The nominal distributions from Data and MC.
(b) : The $Z + jets$ background is estimated using data in the control region with the Fake Factors.
(c) : The $l^+l^-\gamma$ background MC is reweighted with the BDT-based Scale Factors.

Conclusion & Outlook

- **Conclusion:**
  - $Z + \gamma$ background MC is modified by data-driven BDT-based Scale Factors.
  - $Z + jets$ background is estimated using data in Control Region with Fake Factors.
  - Data and MC are in agreement within 15%.
- **Outlook:**
  - More detailed study of photon ID and isolation variable to validate the ABCD method.
  - Spurious signal test based on the background template from our method.
  - Try to tune the hyperparameters in the training model.