$h \to aa \to qqqq$ Analysis

Lessons from $h \to aa \to gggg$ Analysis

Sanha Cheong

Weekly Analysis Team Progress Meeting
Nov. 29th, 2022

SLAC NATIONAL ACCELERATOR LABORATORY
ATLAS EXPERIMENT
HDBS
Stanford University
U.S. DEPARTMENT OF ENERGY

$h \rightarrow aa \rightarrow gggg$ Analysis

- **Internal Note: ATL-COM-PHYS-2022-017**
- Analysis by Murtaza Safdari
- Very mature, near complete
- Analogous final state
- **Many lessons for us to benefit from**
  - Overall analysis strategy
  - Background estimation
  - Control / Signal definitions
  - Expansions, improvements, etc.

Changes / Improvements / Discussion Points highlighted in Green

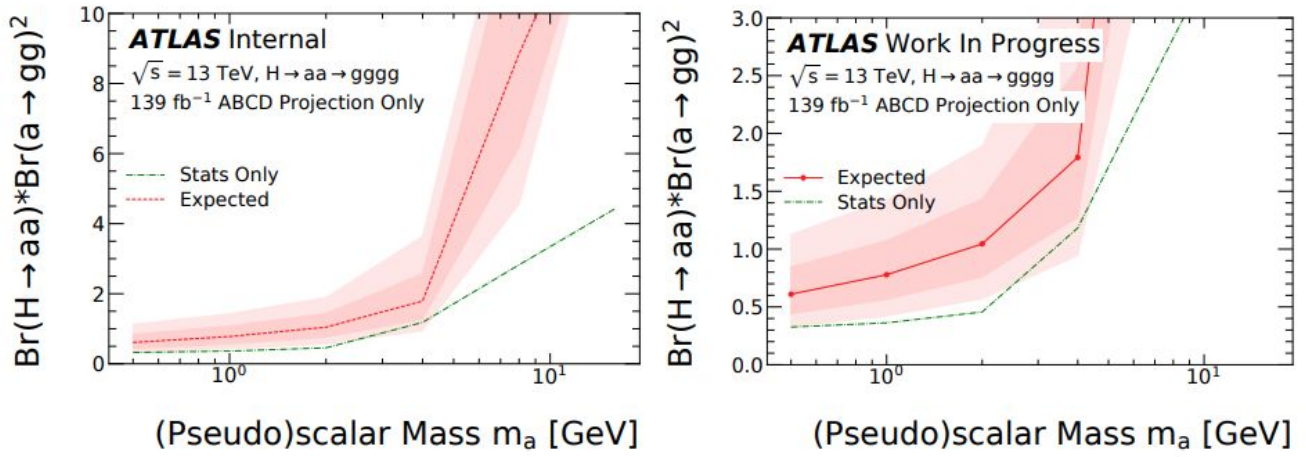


Figure 18: Expected upper limits on $Br(H \rightarrow aa) \times Br(a \rightarrow gg)^2$ across the range of signal a masses tested. Note here that full Run 2 data is being used to estimate the background counts in the signal region using the ABCD estimation method and the MC derived closure.

SLAC

Overview

- **Full Run-II data**
- $Z + h$ production with $Z \rightarrow \mu\mu$
  - **We could probably extend to $Z \rightarrow ee$**
  - Cut-flow baselines from other Z+jets and Vh analyses
- Background
  - **Dominated by Z+jets**
  - ttbar contribution negligible (<10% impact on closure)
  - **This "could" be different for us**

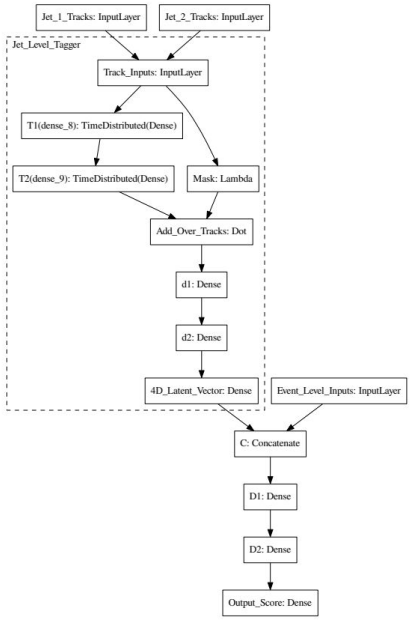
    (HF might remove more Z+jets while leaving ttbar somewhat significant)

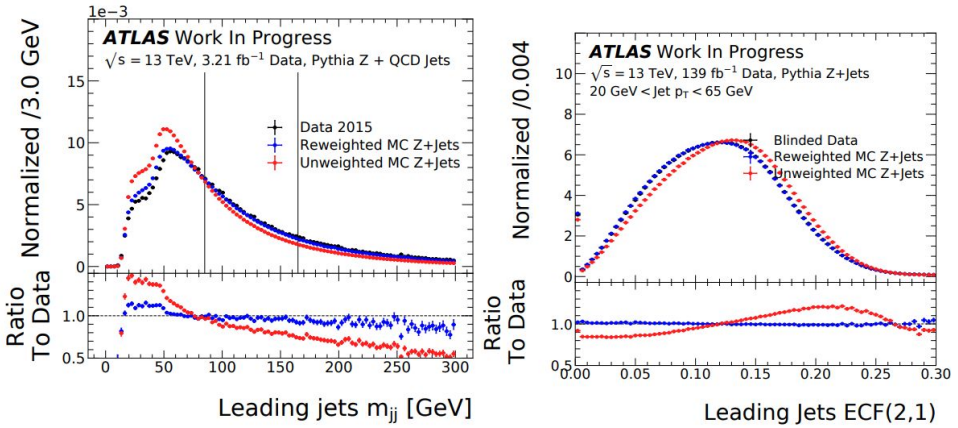
SLAC

Signal v.s. Background

- NN to distinguish *S* from *B*
  - **Event-level tagger** for analysis
  - Two leading jets encoded by PFN
  - **We (I) would like to use GNN**
  - Combined with other info, e.g. m_{jj}, $Z \rightarrow l^+ l^-$ kinematics
- Calibration or Data / MC issues
  - **No explicit calibration at the jet-level**

    (no known source of color singlet → gg jets in SM)
  - Uses NN to reweight Z+jets MC to data
    - Trained on m_{jj} side-band, excluding [85, 165] GeV
  - **Reweighted to match m_{jj} and ECF(2, 1) data profile**
- **Could/Should we do jet-level tagger instead?**
  - **"Event-level" information is not very complex**



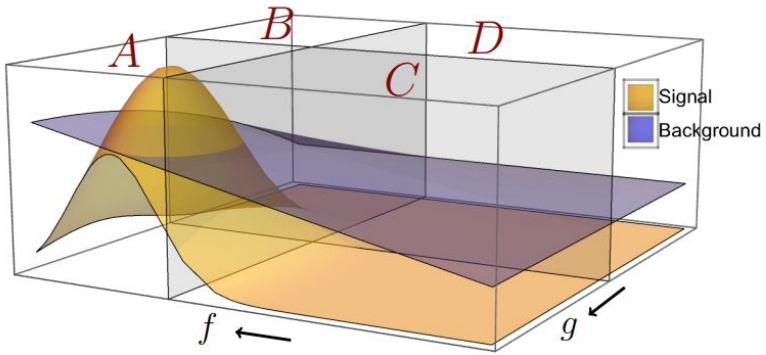


(a) Invariant mass of the two leading jets, m_{jj}.

(b) Two point energy correlation ECF(2,1).

Background Estimation: ABCD Method

- **Data-driven bkg. estimation using ABCD method**
  - Two axes should be independent for background
  - Signal should be well-contained within one region
- Two axes: m_{jj} and event-level *NN* score
- **Main challenge: ensure m_{jj} and *NN* independence**
  - DisCo loss (Distance Correlation)
  - Regularization term that penalizes correlation
  - This is challenging to train
- **Could we use other methods?**
  - **Adversarial or other method for decorrelation**
  - **Not using ABCD-method**
- Both Pythia and Sherpa bkg. samples used



$$N_A = N_C N_B / N_D$$

$$\mathrm{dCov}^2(X,Y) = \langle |X-X'||Y-Y'| \rangle + \langle |X-X'| \rangle \langle |Y-Y'| \rangle - 2\langle |X-X'||Y-Y''| \rangle$$

$$\mathrm{dCorr}^2(X,Y) = \frac{\mathrm{dCov}^2(X,Y)}{\mathrm{dCov}(X,X)\mathrm{dCov}(Y,Y)}$$

$$L = L_{\text{classifier}}(\vec{y}, \vec{y}_{\text{true}}) + \lambda\, \mathrm{dCorr}^2_{y_{\text{true}}=0}(\vec{m}_{jj}, \vec{y}),$$

SLAC

More-or-less agreed in Orange
(feel free to dispute if you have other thoughts)

Summary

- Any questions?
- **Main points that I would like to discuss**
  - Can we include Z → *ee* channel?
    - Less trivial than Z → $\mu\mu$, but known standard procedures
    - **We should do it, but this will take "some" time**

    (VBF???)
  - **We should re-explore ttbar background**
  - Shall we use GNN for signal jets?
    - **Yes** (could be non-trivial, but most likely worth it)
- **Discussion points (continued)**
  - Jet-level v.s. event-level tagger
    - **No clear preference**
    - **Jet-level tagger fine for bump-hunt**, might be inappropriate for ABCD
  - Background estimation method (ABCD or not)
    - **Some group preference towards bump-hunt in m_{jj}**, if we can keep the signal peak tight (expect low stats. issue at the higher m_{jj} side for bkg.)
    - If bump-hunt, **dedicated mjj regression** will probably help significantly
  - If ABCD, decorrelation method?

SLAC