# $HH \to 4b$ analysis at LHC-ATLAS

Shota Hayashida (Nagoya Univ.)

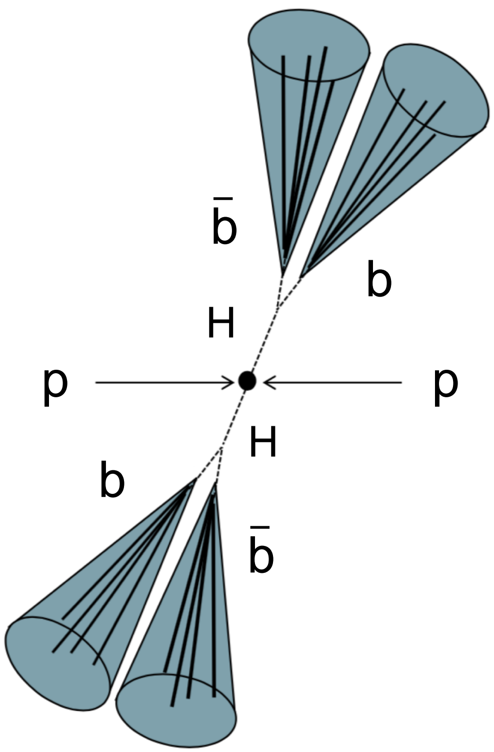
22. Feb. 2023

ILC Physics Meeting

# ➢ Introduction



- ✓ The $b\bar{b}b\bar{b}$ final state on $HH$ search:
  - One of the most sensitive channels thanks to the largest branching ratio (~34%).
- ✓ This presentation will focus on $HH \rightarrow b\bar{b}b\bar{b}$ analysis at the LHC-ATLAS experiment [arXiv:2301.03212].
  - analysis selection, background estimation etc
- ✓ However, it may not be helpful for the analysis at the ILC, because the background components will be different between the LHC and the ILC ☹
- ✓ Even so, I hope it will be helpful for other analyses.

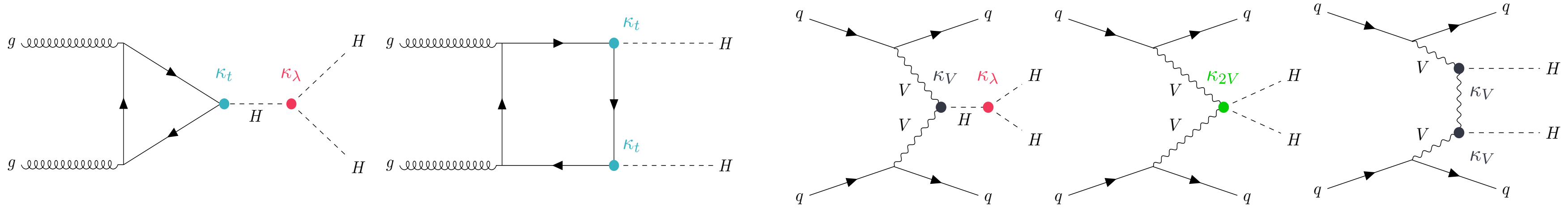


# ➢ Overview of the analysis strategy



✓ The analysis targets $HH \rightarrow b\bar{b}b\bar{b}$ in the ggF and VBF production modes.

- SM ggF $\sigma_{NNLO} \approx$ 31.05 fb @$\sqrt{s} = 13$ TeV
- SM VBF $\sigma_{N3LO} \approx$ 1.726 fb @$\sqrt{s} = 13$ TeV

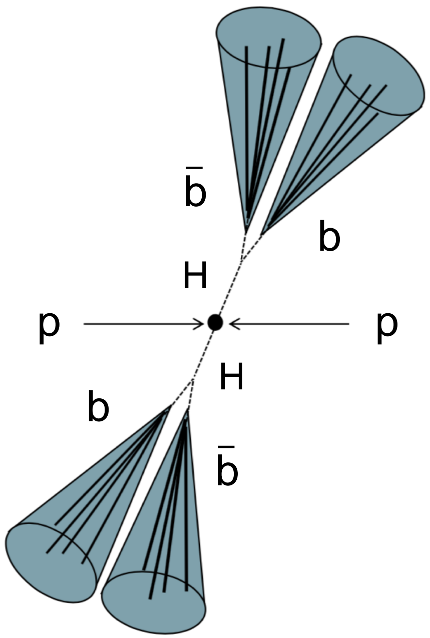



✓ The analysis will select events that have

- at least four $b$-jets
- invariant masses of $b$-jet pairs around Higgs boson mass

✓ Experimental challenges:

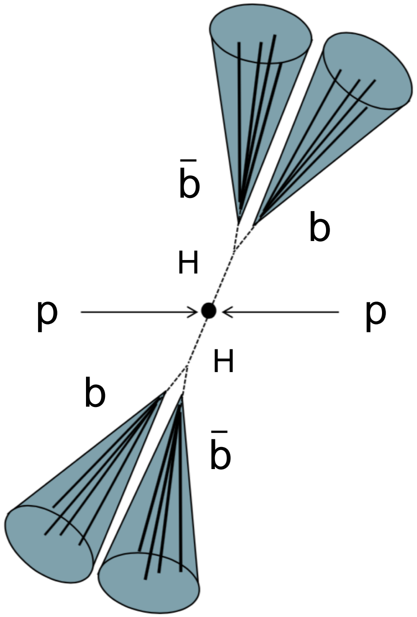
- Higgs boson reconstruction
- Rejection and modelling of QCD multijet background

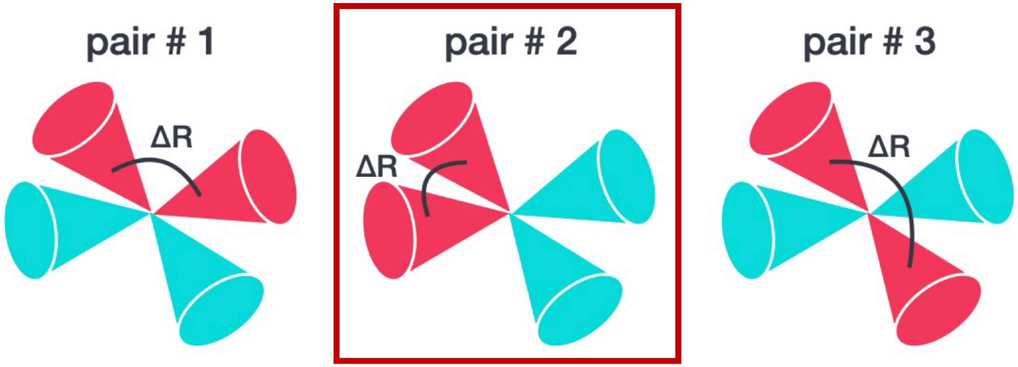


# ➢ $HH \to b\bar{b}b\bar{b}$ reconstruction



1. Select events with at least four $b$-jets (4b events)
   - $p_{\mathrm{T}} > 40$ GeV & $|\eta| < 2.5$ & $b$-tagging @77% WP
2. Pair the four $b$-jets to reconstruct two Higgs candidates (HC)
   - Exploit a principle that the decay products of the Higgs boson are usually collimated due to the Higgs boson's initial momentum.
   - Select a pairing with the smallest $\Delta R_{jj}$ between two jets forming the leading HC.




90% (80%) accuracy for
SM ggF (VBF) events where
the four $b$-jets are four $b$-quarks
from Higgs decays

Leading Higgs Candidate
Subleading Higgs Candidate

☝ the smallest $\Delta R_{jj}$

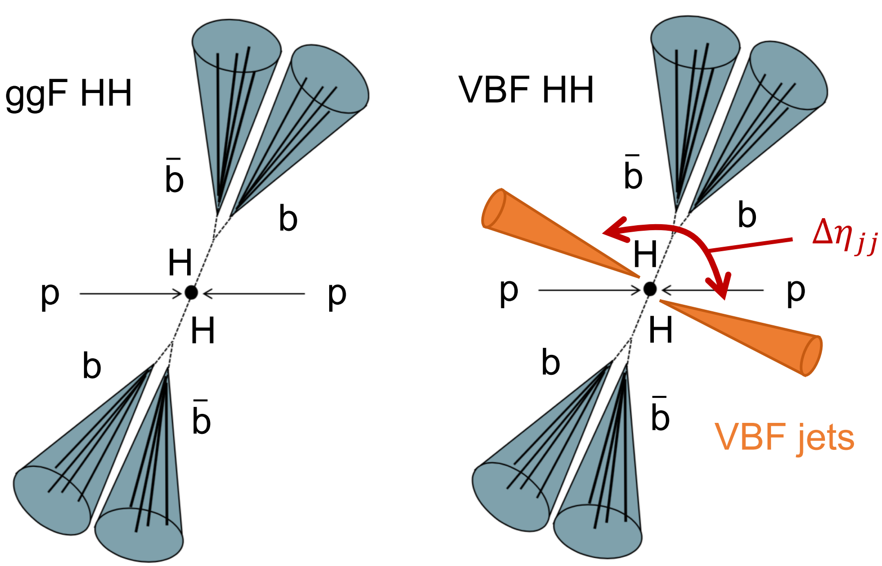
# ➢ ggF and VBF categorization

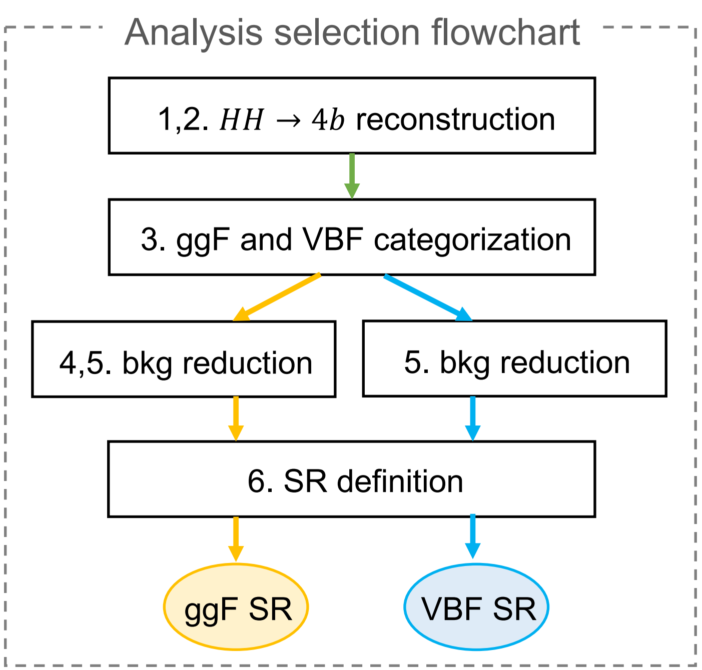


3. Events are categorized into the ggF selection or the VBF selection based on if a pair of VBF jets exist.

- # of jets ≥ 6
- $|\Delta\eta_{jj}| > 3$ & $m_{jj} > 1$ TeV
- $(\sum_i p_i)_{\mathrm{T}} < 65$ GeV ($i$: four $b$-jets + VBF jets)

If any of three selections failed
→ Pass to the ggF selection
If all selections passed
→ Pass to the VBF selection

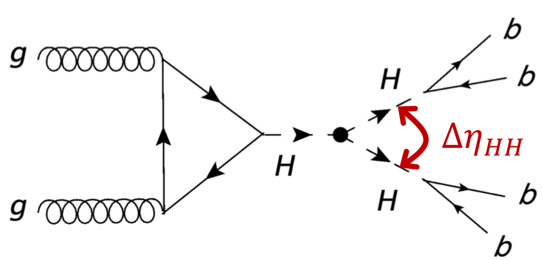


# ➢ Background reduction

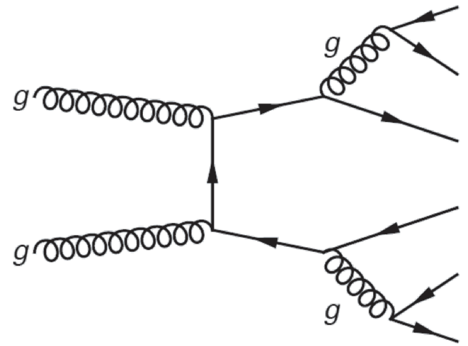
**Signal**

$HH \rightarrow b\bar{b}b\bar{b}$ signal

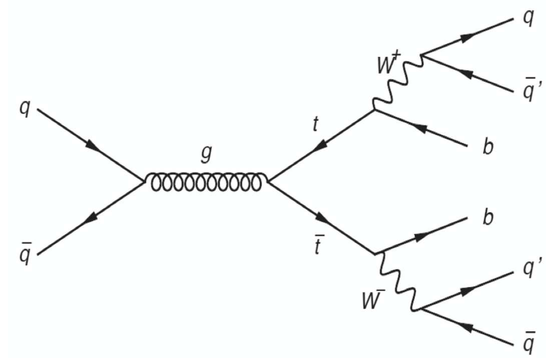



**Main background components**

QCD multijet background




$t\bar{t}$ background




4. QCD multijet background can be reduced by $|\Delta\eta_{HH}| < 1.5$.
   - This cut is not used for the VBF selection due to high sensitivity to SM VBF signal in the region.
5. $t\bar{t}$ background also can be reduced by suppressing events coming from top decays $\left(t \rightarrow bW(\rightarrow q\bar{q})\right)$.

Plot from my doctoral dissertation

# ➢ Signal Region (SR) definition



✓ To define signal region, a discriminator $X_{HH}$ is defined as

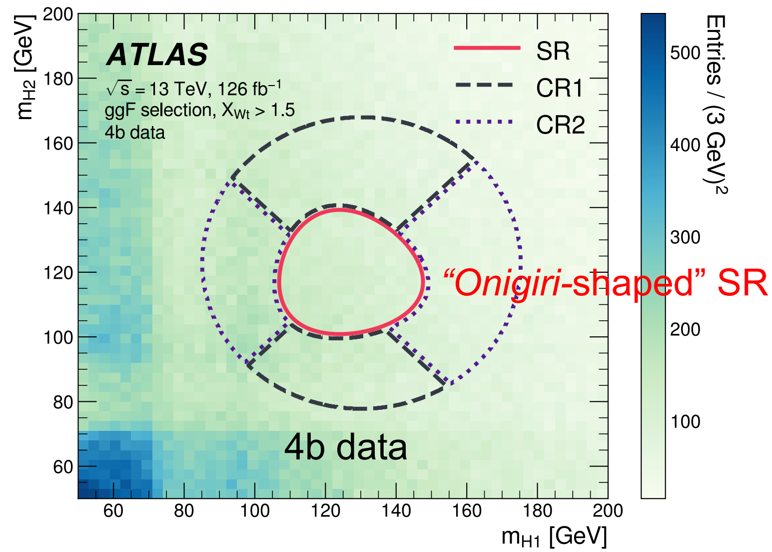
$$X_{HH} = \sqrt{\left(\frac{m_{H1} - 124\,\text{GeV}}{0.1\, m_{H1}}\right)^2 + \left(\frac{m_{H2} - 117\,\text{GeV}}{0.1\, m_{H2}}\right)^2},$$

$m_{H1}$: leading Higgs boson candidate,
$m_{H2}$: sub-leading Higgs boson candidate

which indicates an agreement of HC masses with the expected masses.

6. Events are required to pass $X_{HH} < 1.6$ in both the ggF and VBF selections.

Plot from my doctoral dissertation




# ➢ Data and signal cutflow table

| | Data | ggF Signal SM | ggF Signal $\kappa_\lambda = 10$ | VBF Signal SM | VBF Signal $\kappa_{2V} = 0$ |
|---|---|---|---|---|---|
| **Common preselection** | | | | | |
| Preselection (backup) | $5.70 \times 10^8$ | 530 | 7300 | 22 | 630 |
| Trigger class (backup) | $2.49 \times 10^8$ | 380 | 5300 | 16 | 410 |
| **ggF selection** | | | | | |
| Fail VBF selection | $2.46 \times 10^8$ | 380 | 5200 | 14 | 330 |
| At least 4 $b$-tagged central jets | $1.89 \times 10^6$ | 86 | 1000 | 1.9 | 65 |
| $\lvert\Delta\eta_{HH}\rvert < 1.5$ | $1.03 \times 10^6$ | 72 | 850 | 0.94 | 46 |
| $X_{Wt} > 1.5$ | $7.51 \times 10^5$ | 60 | 570 | 0.74 | 43 |
| $X_{HH} < 1.6$ (ggF signal region) | $1.62 \times 10^4$ | 29 | 180 | 0.24 | 23 |
| **VBF selection** | | | | | |
| Pass VBF selection | $3.30 \times 10^6$ | 5.2 | 81 | 2.2 | 71 |
| At least 4 $b$-tagged central jets | $2.71 \times 10^4$ | 1.1 | 15 | 0.74 | 28 |
| $X_{Wt} > 1.5$ | $2.18 \times 10^4$ | 1.0 | 11 | 0.67 | 26 |
| $X_{HH} < 1.6$ | $5.02 \times 10^2$ | 0.48 | 3.1 | 0.33 | 17 |

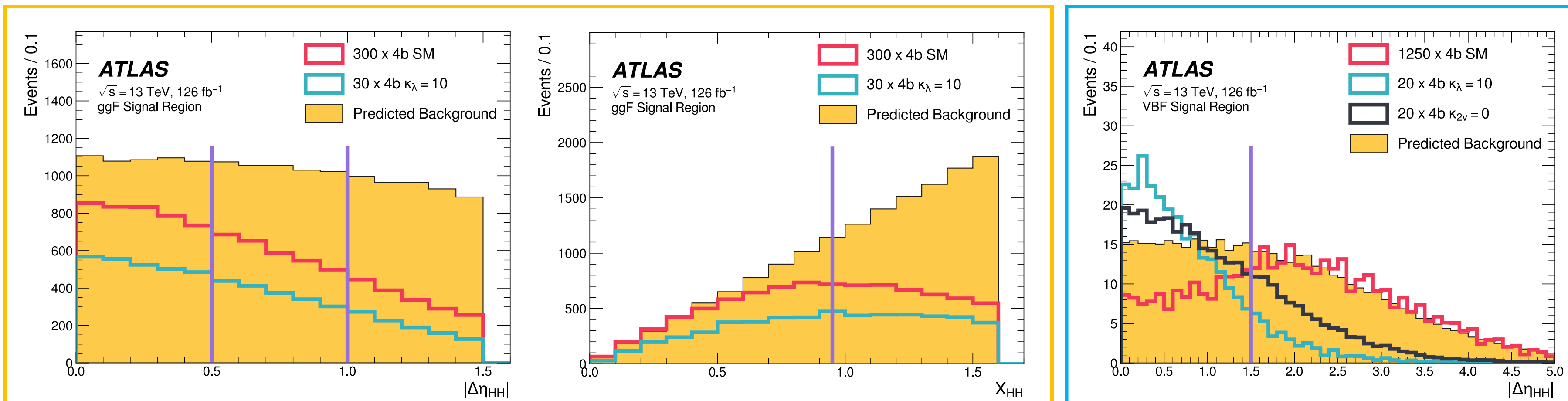
✓ Data (considered as backgrounds) are reduced to $10^{-4}$ or $10^{-6}$ (0.01 or 0.0001%), while HH signals are kept to $10^{-1}$ or $10^{-2}$ (10 or 1%) in the ggF and VBF selection.

# ➢ Analysis categorization in SRs



✓ Analysis categorization is adopted to improve sensitivity.

- $|\Delta\eta_{HH}|$ and $X_{HH}$ are used in the ggF selection.
  - Totally 6 categories are provided per years.
  - Low $|\Delta\eta_{HH}|$ and low $X_{HH}$ category derives the highest sensitivity
- $|\Delta\eta_{HH}|$ boundary at 1.5 is used in the VBF selection.
  - Low $|\Delta\eta_{HH}|$ is more sensitive to BSM while high $|\Delta\eta_{HH}|$ is more sensitive to SM




✓ 30-40% improvements on the SM ggF and VBF cross-section expected limits with respect to ones without any analysis categorization .

# ➢ Strategy of background estimation



- ✓ Main background: QCD multijet background (~90%) and $t\bar{t}$ background (~10%)
- ✓ QCD multijet background is hard to model in simulation ☹
- ✓ This analysis uses a fully data-driven approach using 2b data (events in data with exactly two $b$-jets) to estimate 4b background (2b data →4b bkg).

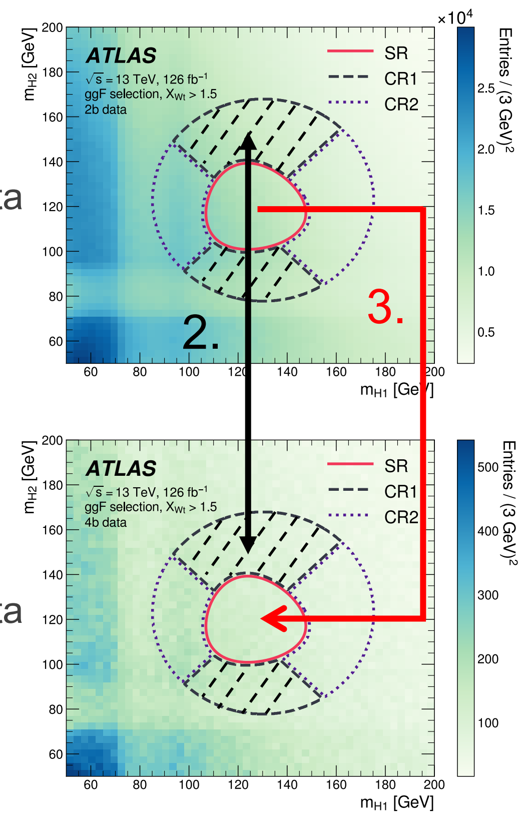
  ← 2b data is preferred in this analysis because of having more statistics and similar kinematics to 4b.

- ✓ The background estimation strategy:
  1. Define Control Region 1 and 2 (CR1 and CR2)
  2. Derive weights ($w = 2b/4b$) to transfer 2b data to 4b bkg using 2b and 4b data in CR1
  3. Apply the weights to 2b data in SR

     (reweighted 2b data = background prediction)
  4. Take a difference between CR1 and CR2 derived weights as CR12 shape systematic uncertainty to account for kinematic differences between CR1 and SR.

2b data

4b data



# ➢ Neural Network reweighting



✓ The relation between 2b data and 4b data is not simple.
e.g. physics process, analysis selection etc

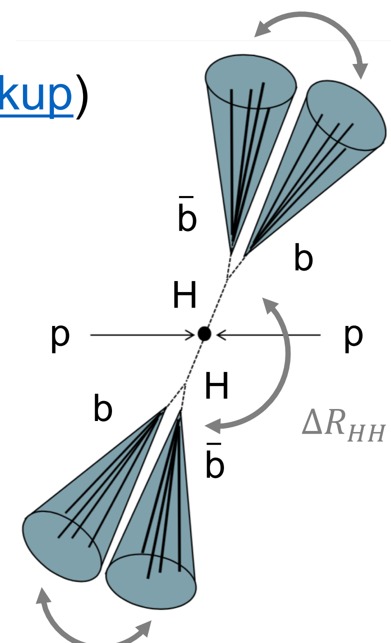
➔ Need to see differences and correlations in many variables.

✓ This analysis utilizes a machine learning approach using a neural network (NN) to derive weights [1,2].

(The NN structure developed by keras can be found at backup)

✓ Totally 12 (9) training variables are used for the ggF (VBF) channel (backup).

- $\Delta R_{jj}$ between two jets forming Higgs candidate
- $\Delta R_{HH}$ between two Higgs candidates
- # of jets in the event etc



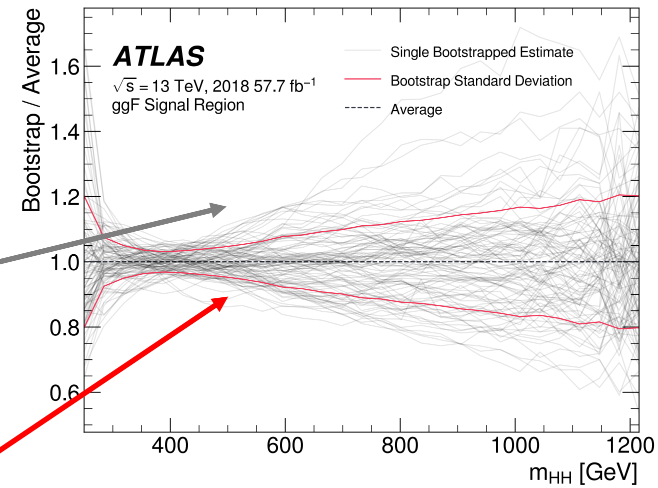

$$\Delta R_{jj} = \sqrt{(\Delta\eta_{jj})^2 + (\Delta\phi_{jj})^2}$$

[1] J. Mach. Learn. Res. 10 (2009) 1391–1445
[2] arXiv:1911.00405 [eess.SP]

# ➢ 2b poisson + bootstrap uncertainties



- ✓ NN training is dependent on CR statistics, NN initialization and NN training stochasticity.
- ✓ To account for the variations,
  1. Generate 100 different training samples (CR2b + 4b) with the bootstrap resampling technique [1].
  2. Train an ensemble of 100 NNs with each different training sample.
     → a set of 100 background estimations is provided.
  3. **Take a mean of the 100 background estimation as the nominal background estimation.**
  4. Take the standard deviation as bootstrap uncertainty.

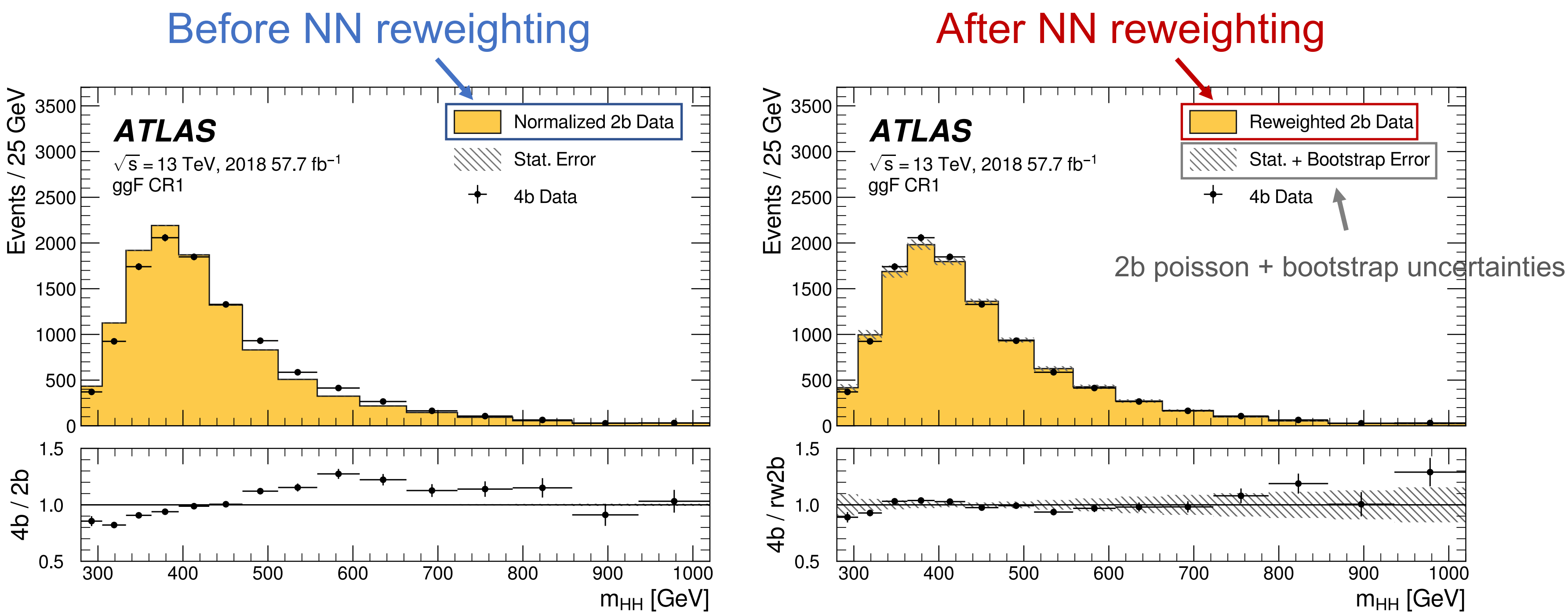



- ✓ 2b poisson uncertainty (of reweighted 2b data) is also included as a statistical uncertainty via sum of square weights.

[1] B. Efron, *Bootstrap Methods: Another Look at the Jackknife*, The Annals of Statistics 7 (1979) 1–26

# ➢ NN reweighting performance






- ✓ Good agreements between reweighted 2b data and 4b data in CR1.
- ✓ NN reweighting is validated to have good performance to learn the relation between 2b and 4b data ☺

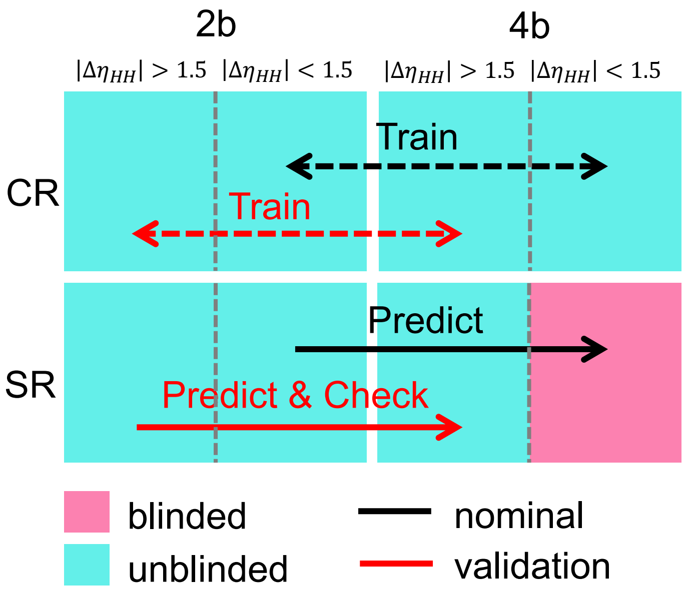
# ➢ NN reweighting validation



✓ To validate the weights derived in CR1 work in SR, some validation studies were performed.

✓ e.g. reversed-$\Delta\eta_{HH}$ ($|\Delta\eta_{HH}| > 1.5$) data

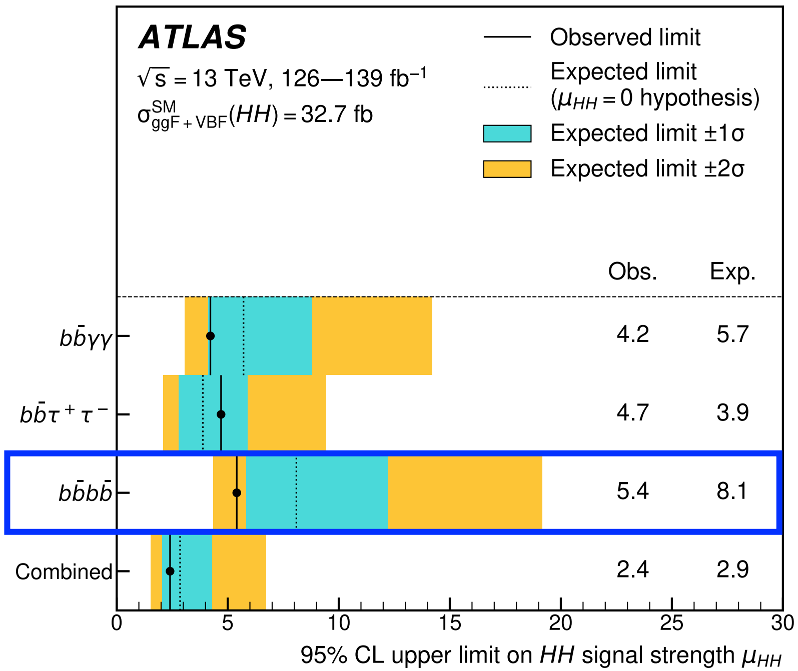
- Train NNs using reversed-$\Delta\eta_{HH}$ data instead of the nominal data ($|\Delta\eta_{HH}| < 1.5$) with the same setup as the nominal background estimation
- Check the NN reweighting performance in the reversed-$\Delta\eta_{HH}$ SR.
- Good agreement between the background estimation and the target 4b data is observed ☺

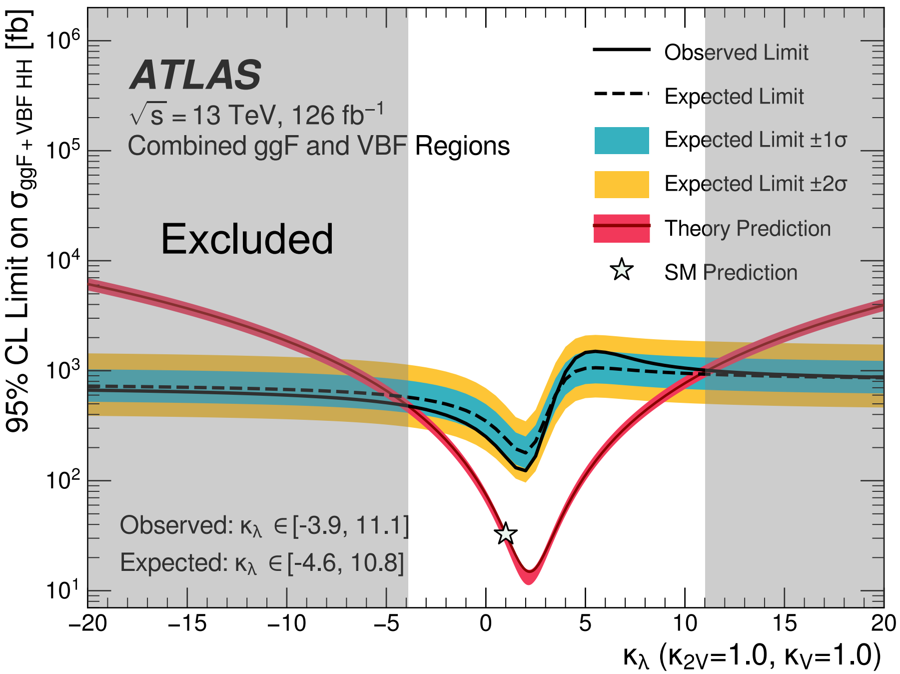



✓ The other validation studies (using shifted regions, MC samples, 3b data) also show good agreements.

✓ NN reweighting is validated to work in the SR ☺

# ➢ Results of $HH \rightarrow b\bar{b}b\bar{b}$ analysis




- ✓ 95% CL upper limit of 5.4 (8.1) on the signal strength is achieved.
- ✓ Trilinear Higgs self-coupling constraint is $\kappa_\lambda \in [-3.9, 11.1]([-4.6, 10.8])$.

# ➢ Systematic uncertainty



Impact of the dominant uncertainties on the expected upper limit on the signal strength of SM ggF+VBF HH cross-section ($\mu$) when fixing the nuisance parameter(s) with respect to $\mu$ with all free nuisance parameters

| Source of Uncertainty | $\Delta\mu/\mu$ |
|---|---|
| **Theory uncertainties** | |
| Theory uncertainty in signal cross-section | $-9.0\%$ |
| All other theory uncertainties | $-1.4\%$ |
| **Background modeling uncertainties** | |
| Bootstrap uncertainty | $-7.1\%$ |
| CR to SR extrapolation uncertainty | $-7.5\%$ |

✓ Background modeling uncertainties are the dominant uncertainties.

✓ Bootstrap uncertainty will be improved by increasing statistics (this analysis used 126 $fb^{-1}$).

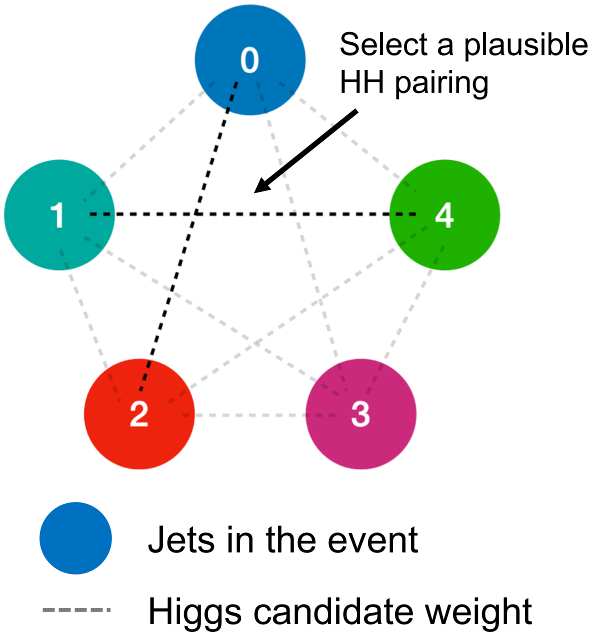
# ➢ Ideas for future experiments (1)



*(My interest)*

✓ <u>Higgs boson reconstruction from four $b$-jets</u>

- Motivation:
  - 25% of $HH \to b\bar{b}b\bar{b}$ signals miss a part of $b$-jets from Higgs decays when selecting four $b$-jets ($p_\mathrm{T} > 40$ GeV & $b$-tagging @77% WP) ☹
  - Low pairing accuracy for BSM $HH \to b\bar{b}b\bar{b}$ signals in the smallest $\Delta R$ pairing due to targeting higher $p_\mathrm{T}$ regimes, e.g. 50% for $\kappa_\lambda = 10$ ☹
- Idea:
  - Utilize a machine learning approach using a graph neural network (GNN) [1].
  - Select a plausible $HH \to b\bar{b}b\bar{b}$ system from five or more jets in the event by learning kinematical correlations of the jets.
  - Maybe improve the pairing accuracy at the ILC too ☺




[1] Peter W. Battaglia et., *Relational inductive biases, deep learning, and graph networks*, arXiv:1806.01261

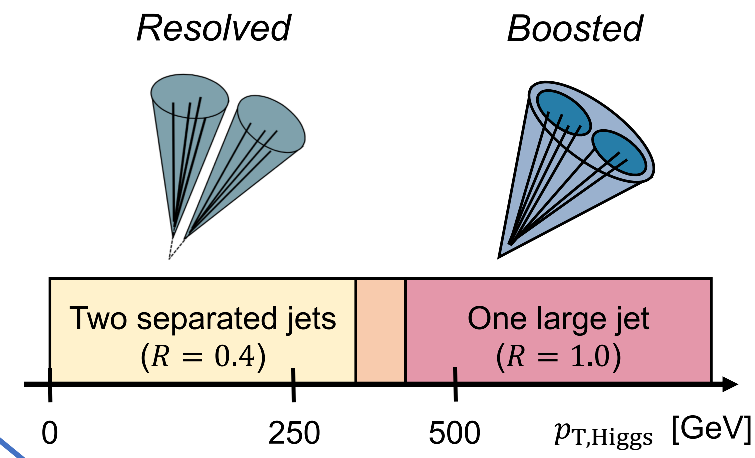
# ➢ Ideas for future experiments (2)

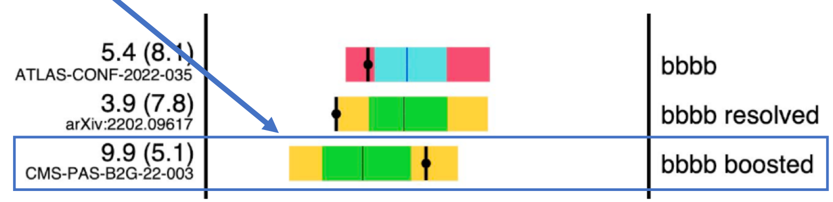


*(My interest)*

✓ Boosted topology of $HH \to b\bar{b}b\bar{b}$

- Motivation:
  - $H \to b\bar{b}$ experimental fingerprint often reconstructed as one large jet (boosted).
  - CMS experiment achieved the high sensitivity in the boosted analysis [1].
- Key:
  - Strong boosted $H \to b\bar{b}$ tagger is a key to suppress QCD multijet background.
  - NN-based boosted $H \to b\bar{b}$ tagger is developed in the ATLAS [2].
    → almost the same performance as CMS




- Maybe hard to utilize the boosted topology at the ILC due to lack of events in HHZ cross-section, but the technique will be interesting to improve sensitivity.

[1] CMS, Search for nonresonant pair production of highly energetic Higgs bosons decaying to bottom quarks, arXiv:2205.06667
[2] ATLAS, Identification of boosted Higgs bosons decaying Into $b\bar{b}$ with neural networks and variable radius subjets in ATLAS, ATL-PHYS-PUB-2020-019

# ➢ Conclusion



- ✓ This presentation focused on the analysis selection and the background estimation for $HH \rightarrow b\bar{b}b\bar{b}$ analysis at the LHC-ATLAS experiment.
- ✓ The analysis at the LHC is challenged by a huge amount of QCD multijet background.
- ✓ The analysis selection is optimized to suppress QCD multijet background, and the background estimation is developed for the precise modeling.
- ✓ Not sure that the techniques will be helpful for $HH \rightarrow b\bar{b}b\bar{b}$ analysis at the ILC, but hope it will be helpful for other analyses.

# ➢ backup

# ➢ Preselection



| | 0 lepton |
|---|---|
| Lepton | No lepton: no tight electron and good muon with $p_T > 25$ GeV |
| Jet | Two loosely $b$-tagged jets (77% eff. WP) with $p_T > 25$ GeV, and two extra jets with $p_T > 25$ GeV |

| | 1 lepton |
|---|---|
| Lepton | One lepton: one tight electron or good muon with $p_T > 25$ GeV |
| Jet | Two loosely $b$-tagged jets (77% eff. WP) with $p_T > 25$ GeV |

Table 6.1 Summary of pre-selection used in this analysis.

# ➢ Triggers



| Year | Requirement | ATLAS Terminology | Type |
|---|---|---|---|
| 2016 | Two *b*-jet (60% eff. WP) with $E_T > 55$ GeV and one extra jet with $E_T > 100$ GeV | HLT_j100_2j55_bmv2c2060_split | 2b1j |
| | Two *b*-jet (60% eff. WP) with $E_T > 35$ GeV and two extra jet with $E_T > 35$ GeV | HLT_2j35_bmv2c2060_split_2j35_ L14J15.0ETA25 | 2b2j |
| 2017 | Two *b*-jet (70% eff. WP) with $E_T > 55$ GeV and one extra jet with $E_T > 150$ GeV | HLT_j110_gsc150_boffperf_split_ 2j35_gsc55_bmv2c1070_split_L1J85_3J30 | 2b1j |
| | Two *b*-jet (40% eff. WP) with $E_T > 35$ GeV and two extra jet with $E_T > 35$ GeV | HLT_2j15_gsc35_bmv2c1040_split_ 2j15_gsc35_boffperf_split_L14J15.0ETA25 | 2b2j |
| 2018 | Two *b*-jet (70% eff. WP) with $E_T > 55$ GeV and one extra jet with $E_T > 150$ GeV | HLT_j110_gsc150_boffperf_split_ 2j45_gsc55_bmv2c1070_split_L1J85_3J30 | 2b1j |
| | Two *b*-jet (60% eff. WP) with $E_T > 35$ GeV and two extra jet with $E_T > 35$ GeV | HLT_2j35_bmv2c1060_split_2j35_ L14J15.0ETA25 | 2b2j |

Table 5.1 List of multi *b*-jet triggers used for this analysis.

# ➢ Data and signal cutflow table



| | Data | ggF Signal SM | ggF Signal $\kappa_\lambda = 10$ | VBF Signal SM | VBF Signal $\kappa_{2V} = 0$ |
|---|---|---|---|---|---|
| Common preselection | | | | | |
| Preselection | $5.70 \times 10^8$ | 530 | 7300 | 22 | 630 |
| Trigger class | $2.49 \times 10^8$ | 380 | 5300 | 16 | 410 |
| ggF selection | | | | | |
| Fail VBF selection | $2.46 \times 10^8$ | 380 | 5200 | 14 | 330 |
| At least 4 $b$-tagged central jets | $1.89 \times 10^6$ | 86 | 1000 | 1.9 | 65 |
| $\lvert\Delta\eta_{HH}\rvert < 1.5$ | $1.03 \times 10^6$ | 72 | 850 | 0.94 | 46 |
| $X_{Wt} > 1.5$ | $7.51 \times 10^5$ | 60 | 570 | 0.74 | 43 |
| $X_{HH} < 1.6$ (ggF signal region) | $1.62 \times 10^4$ | 29 | 180 | 0.24 | 23 |
| VBF selection | | | | | |
| Pass VBF selection | $3.30 \times 10^6$ | 5.2 | 81 | 2.2 | 71 |
| At least 4 $b$-tagged central jets | $2.71 \times 10^4$ | 1.1 | 15 | 0.74 | 28 |
| $X_{Wt} > 1.5$ | $2.18 \times 10^4$ | 1.0 | 11 | 0.67 | 26 |
| $X_{HH} < 1.6$ | $5.02 \times 10^2$ | 0.48 | 3.1 | 0.33 | 17 |
| $m_{HH} > 400$ GeV (VBF signal region) | $3.57 \times 10^2$ | 0.43 | 1.8 | 0.30 | 16 |

This presentation omits to mention a cut of $m_{HH} > 400$ GeV for the VBF selection, but we required the cut due to the poor background modelling in the region.

# ➢ NN structure



✓ The NN used for the ggF (VBF) channel consists of

- one input layer
- three densely connected hidden layers of 50 (20) nodes each with ReLU activation functions
- a single node linear output

✓ The loss function is based on Ref. [1,2]

[1] J. Mach. Learn. Res. 10 (2009) 1391–1445
[2] arXiv:1911.00405 [eess.SP]

$$\mathcal{L}(Q(x)) = \mathbb{E}_{x\sim p_{2b}}\left[\sqrt{e^{Q(x)}}\right] + \mathbb{E}_{x\sim p_{4b}}\left[\frac{1}{\sqrt{e^{Q(x)}}}\right]$$

$Q(x)$: an estimator dependent on input variables $x$,
$\mathbb{E}_{x\sim p2b(4b)}$: expectation value with respect to the 2b (4b) probability density.

The reweighting function $w = p_{4b}/p_{2b}$ can be derived once NN is trained.

$$Q_{\min}(x) \equiv \arg\min_{Q} \mathcal{L}(Q(x)) = \log \frac{p_{4b}(x)}{p_{2b}(x)}$$

# ➢ NN training variables



| ggF | VBF |
|---|---|
| 1. $\log(p_T)$ of the $2^{nd}$ leading Higgs boson candidate jet<br>2. $\log(p_T)$ of the $4^{th}$ leading Higgs boson candidate jet<br>3. $\log(\Delta R)$ between the closest two Higgs boson candidate jets<br>4. $\log(\Delta R)$ between the other two Higgs boson candidate jets<br>5. Average absolute $\eta$ value of the Higgs boson candidate jets<br>6. $\log(p_T)$ of the di-Higgs system<br>7. $\Delta R$ between the two Higgs boson candidates<br>8. $\Delta\phi$ between jets in the leading Higgs boson candidate<br>9. $\Delta\phi$ between jets in the subleading Higgs boson candidate<br>10. $\log(X_{Wt})$<br>11. Number of jets in the event<br>12. Trigger class index as one-hot encoder | 1. Maximum dijet mass from the possible pairings of the four Higgs boson candidate jets<br>2. Minimum dijet mass from the possible pairings of the four Higgs boson candidate jets<br>3. Energy of the leading Higgs boson candidate<br>4. Energy of the subleading Higgs boson candidate<br>5. Second-smallest $\Delta R$ between the jets in the leading Higgs boson candidate (from the three possible pairings for the leading Higgs candidate)<br>6. Average absolute $\eta$ value of the four Higgs boson candidate jets<br>7. $\log(X_{Wt})$<br>8. Trigger class index as one-hot encoder<br>9. Year index as one-hot encoder (for years inclusive training) |