Data-driven background estimation in the search for Higgs boson pair production in the HH $\rightarrow$ bbbb channel with the ATLAS experiment

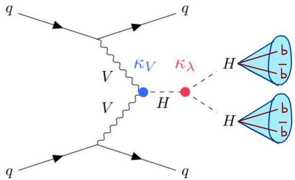
Master's Thesis Defense

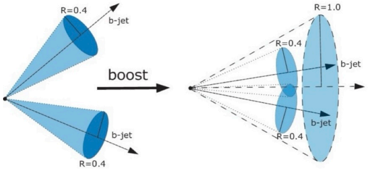
Emilio Apicella
First reviewer: R.C. Camacho Toro
Second reviewer: M. Franchini

29/09/2025

VBF boosted HH $\rightarrow$ 4b

Studying this channel helps to **test the electroweak symmetry breaking mechanism in the Standard Model** of particle physics and explore possible **new physics**.





Channel	Probability (%)
$HH \rightarrow b\bar{b}b\bar{b}$	$\sim$ 34
$HH \rightarrow b\bar{b}\tau^+\tau^-$	$\sim$ 7.3
$HH \rightarrow b\bar{b}W^+W^-$	$\sim$ 12.5
$HH \rightarrow b\bar{b}\gamma\gamma$	$\sim$ 0.13
$HH \rightarrow W^+W^-W^+W^-$	$\sim$ 4.6
$HH \rightarrow \gamma\gamma\gamma\gamma$	$\sim$ 0.0005

These events are hidden inside a large amount of QCD background composed of:

- Non-resonant multijet production with heavy quarks (b/c);
- $t\bar{t}$ events (approx. 10% of multijet bkg);
- light jets misidentified as b-jets;

Current Background Estimation in $HH \to b\bar{b}b\bar{b}$

arXiv:2404.17193v2 [hep-ex]

- Define a **Control Region (CR)**: a region where the **signal contamination is low (max 8%)**.
- **Event Selection:**
  - **1Pass:** only one boosted jet **is identified as a *b*-jet**.
  - **2Pass:** both boosted jets **are identified as *b*-jets**.
- In the CR, compute a **normalization factor**:

$$w = \frac{N_{\text{CR, 2Pass}}}{N_{\text{CR, 1Pass}}} = 0.0039 \pm 0.0002$$

- Apply this weight w to **1Pass events in the Signal Region (SR)** to estimate the background in **2Pass SR**.

Systematics: estimated from the difference of w in Validation Region (VR).

Problems: poor statistics in CR, high uncertainties

Solution

- **Approach:** Data-driven combined with Machine Learning techniques.
- **Data used:** Run 2 (2015–2018) + partial Run 3 (2022–2023).
- **Selection:**
  - VBF selection:
    - p_T of the two VBF jets > 20 GeV;
    - Invariant mass of the di-jet system: $m_{jj} > 1$ TeV;
    - $\left|\eta^{\text{vbfj1}} - \eta^{\text{vbfj2}}\right| > 3$;
  - Boosted topology selection:
    - p_T of the leading Higgs candidate > 450 GeV;
    - p_T of the subleading Higgs candidate > 250 GeV.

Analysis Regions Definitions

SR:

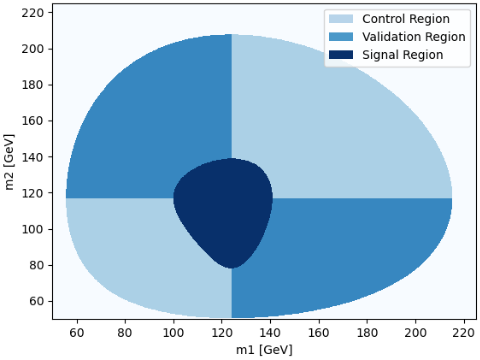
$$\sqrt{\left(\frac{m_{H_1} - 124\ \text{GeV}}{1500\ \text{GeV}/m_{H_1}}\right)^2 + \left(\frac{m_{H_2} - 117\ \text{GeV}}{1900\ \text{GeV}/m_{H_2}}\right)^2} < 1.6\ \text{GeV}$$

VR and CR:

$$\sqrt{\left(\frac{10(m_{H_1} - 124\ \text{GeV})}{\log m_{H_1}}\right)^2 + \left(\frac{10(m_{H_2} - 117\ \text{GeV})}{\log m_{H_2}}\right)^2} < 170\ \text{GeV}$$

&

$$(m_{H_1} > 124\ \text{GeV} \wedge m_{H_2} > 117\ \text{GeV}) \quad \text{or} \quad (m_{H_1} < 124\ \text{GeV} \wedge m_{H_2} < 117\ \text{GeV})$$



Tagging Strategies

Tag

Both large-R jets are bb-tagged

Pros:

- Kinematics of interest

Cons:

- Poor statistics

No Tag

Neither of the two large-R jets is bb-tagged

Pros:

- Very high statistics ($\sim$ 20000x events)

Cons:

- Opposit kinematics

$\rightarrow$ We will use the "No Tag" dataset and then we will apply a "reweighting".

Overview of the Analysis Strategy

1

Fit 2D mass distribution

$f(m_{h1}, m_{h2})$

Models tested: Gaussian Process, Polynomial

2

Train conditional Neural Network Flow

$f(x \mid m_{h1}, m_{h2}, \text{year})$

`zuko.NSF`

3

Compute Tag reweighting factor

$w(x) = \frac{f(x|\text{Tag}=1)}{f(x|\text{Tag}=0)}$

`XGBoosting`

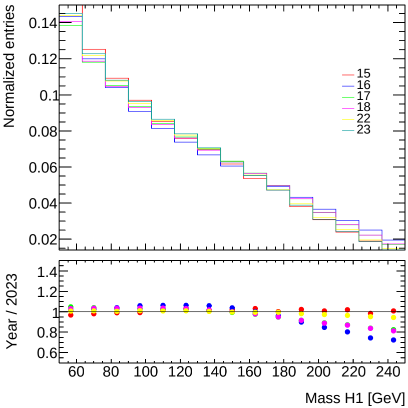
4

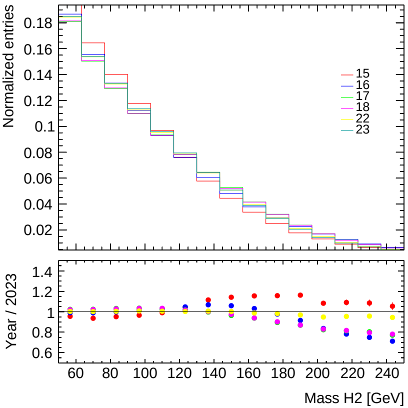
Uncertainty estimation

Background estimation in Signal Region

Mass Distributions

NoTag - CR+VR+SR



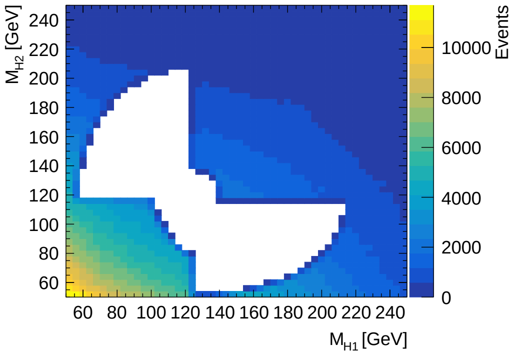


Data from different years show distinct distributions → treat them separately to ensure more accurate modeling.

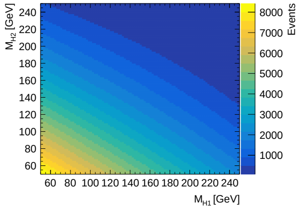
Fit Strategies

NoTag – 1 \ (VR ∪ SR)

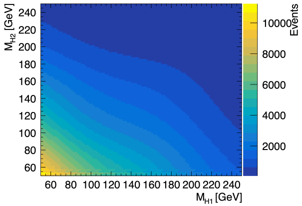
Data



Polynomial

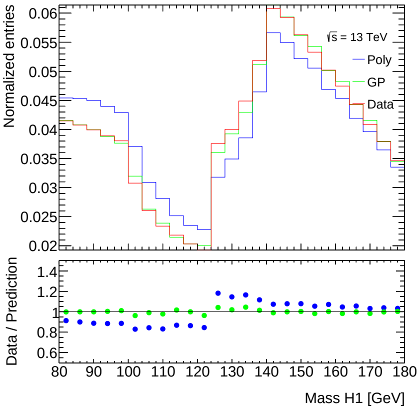


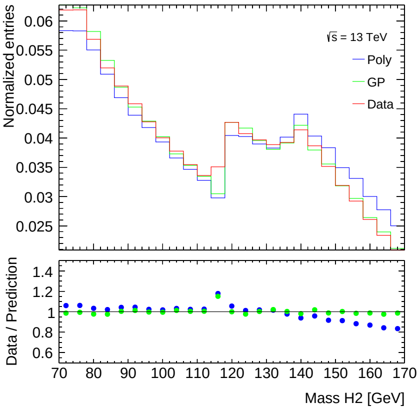
Gaussian Process



Fit Results

NoTag - VR





Compared to polynomial fits, the Gaussian Process regressor offers a more flexible and accurate description of the 2D mass distribution.

Neural Network Model

Features:

- Leading Higgs Candidate (H1): p_T^{h1}, ϕ^{h1}, η^{h1}.
- Subleading Higgs Candidate (H2): p_T^{h2}, ϕ^{h2}, η^{h2}.
- Leading VBF Jet: E^{vbfj_1}, $p_T^{vbfj_1}$, η^{vbfj_1}
- Subleading VBF Jet: E^{vbfj_2}, $p_T^{vbfj_2}$, η^{vbfj_2}
- Di-Jet system: m_{jj}

Conditions: m^{h1}, m^{h2}, `year`

```
flow=NSF(transforms=48,hidden_features=[256,256,256],bins=128)
```

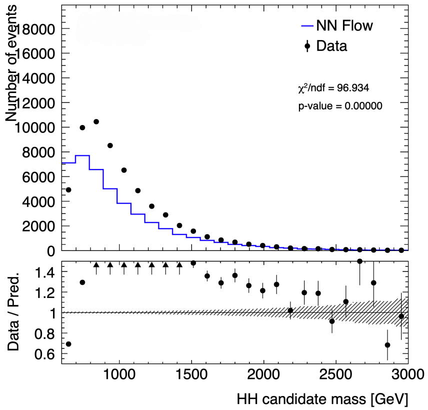
arXiv:1906.04032

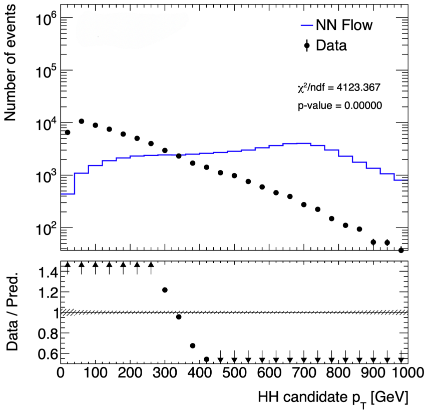
$$\mathrm{m}_{hh} = \sqrt{2\, p_T^{h_1}\, p_T^{h_2}\, (\cosh(\eta^{h_1} - \eta^{h_2}) - \cos(\phi^{h_1} - \phi^{h_2}))}$$

$$\mathrm{p}_T^{hh} = \sqrt{(p_T^{h_1})^2 + (p_T^{h_2})^2 + 2\, p_T^{h_1}\, p_T^{h_2} \cos(\phi^{h_1} - \phi^{h_2})}$$

Learning Correlation

NoTag - VR

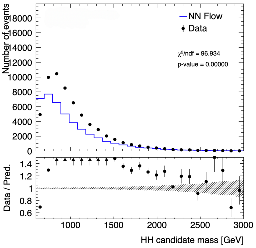


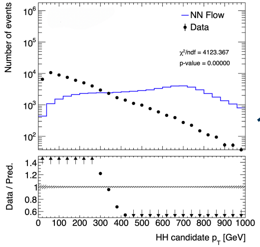


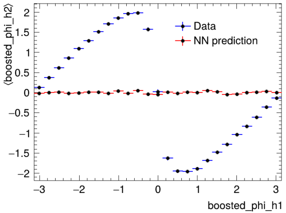
Even though the *primitive* variables are well modeled by the NN, derived/calculated observables are not consistent with data indicating a missing effect or assumption we did not account for.

Learning Correlation

NoTag - VR







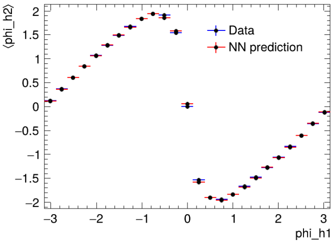
Correlation between ϕ^{h_1} and ϕ^{h_2}

The Neural Network doesn't learn the correlation among some variables.

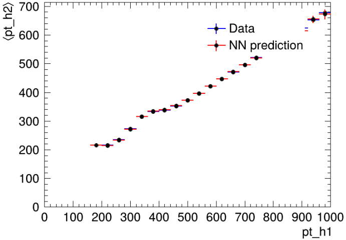
Learning Correlation

NoTag – VR

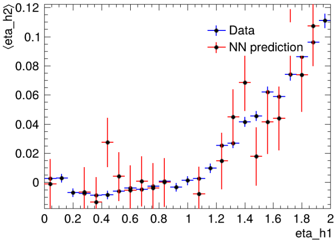
- $\phi^{h1}, \phi^{h2} \rightarrow \phi^{h1}, \Delta\phi$ where $\Delta\phi = (\phi^{h1} - \phi^{h2})$, $\phi^{h2} = \phi^{h1} - \Delta\phi$
- $\eta^{h1}, \eta^{h2} \rightarrow \eta^{h1}, \Delta\eta$ where $\Delta\eta = (\eta^{h1} - \eta^{h2})$, $\eta^{h2} = \eta^{h1} - \Delta\eta$



ϕ^{h1} vs. ϕ^{h2}



p_T^{h1} vs. p_T^{h2}

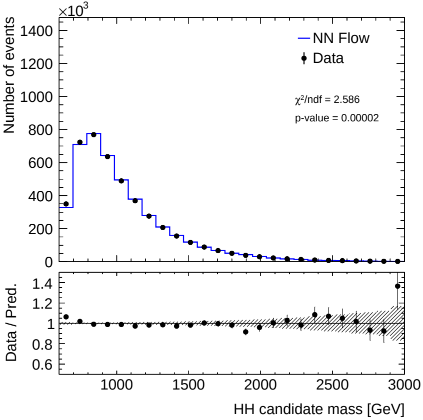


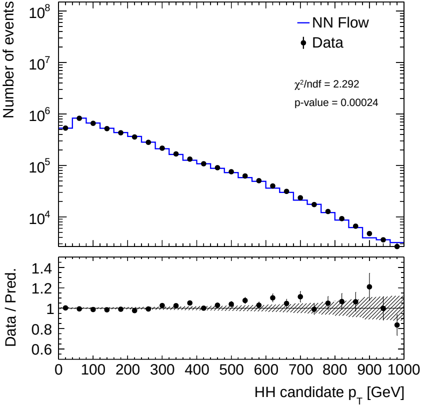
η^{h1} vs. η^{h2}

Now the correlation among all the variables is learned by the Neural Network.

Discriminant Variables Modeling Results

NoTag – VR





Optimal modeling in the VR obtained for the datasets with no boosted-Tag requirements; now we need to apply reweighting.

BDT Tagging Correction

XGBoosting [https://xgboosting.com]

Inputs:

- Leading Higgs Candidate (H1): m^{h1}, p_T^{h1}, ϕ^{h1}, η^{h1}.
- Subleading Higgs Candidate (H2): m^{h2}, p_T^{h2}, ϕ^{h2}, η^{h2}.
- Leading VBF Jet: E^{vbfj_1}, $p_T^{vbfj_1}$, η^{vbfj_1}
- Subleading VBF Jet: E^{vbfj_2}, $p_T^{vbfj_2}$, η^{vbfj_2}
- Di-Jet system: m_{jj}

Outputs:

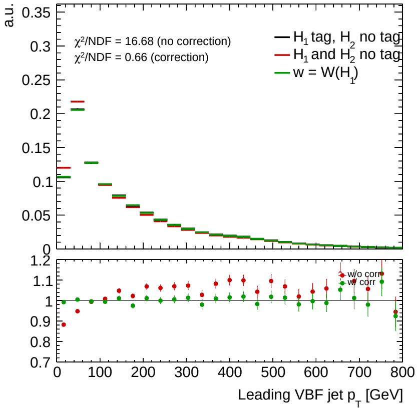
- `probability_tag_h1` = $\mathbb{P}(T = 1 \mid x)_{H1}$
- `probability_tag_h2` = $\mathbb{P}(T = 1 \mid x)_{H2}$

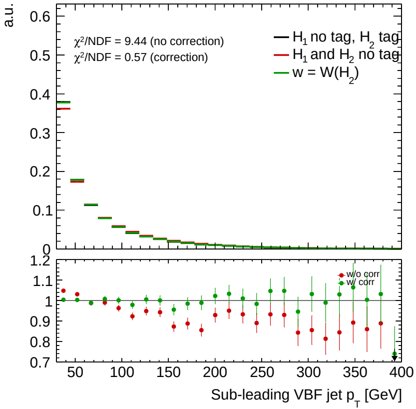
Correction:

$$w = \frac{p}{1-p}$$

Independent H1 and H2 Correction

CR

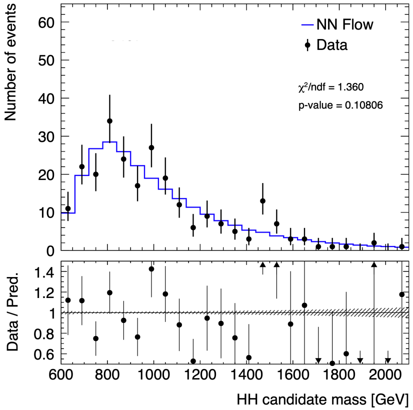


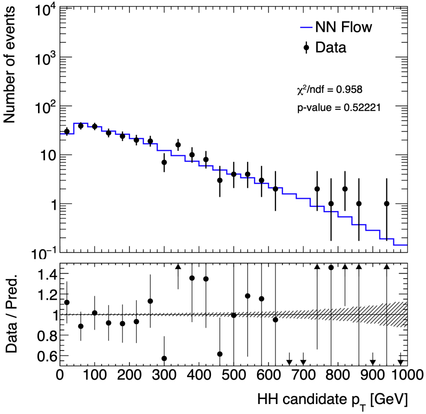


H1 and H2 correction work well, $prob_{h1}$ and $prob_{h2}$ are independent $\rightarrow$

$$\text{weight}_{h1h2} = \frac{prob_{h1} \cdot prob_{h2}}{1 - prob_{h1} \cdot prob_{h2}}$$

NN Flow Tagging Correction

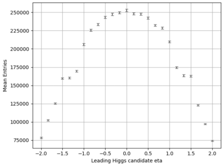


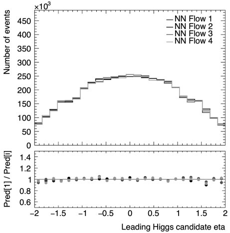


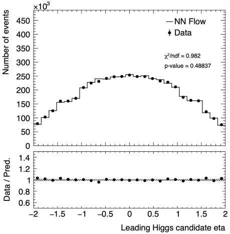
Apart from the poor statistics, the modeling is very good.

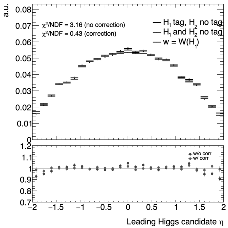
Uncertainty estimation

- Intrinsic statistics: distributions learned by the model.
- Neural Network training uncertainty: If we train the NN again, we would obtain different minimum $\rightarrow$ different parameters.
- Deviation from no-tagged data.
- BDT correction UNC: deviations from data in tagged events.



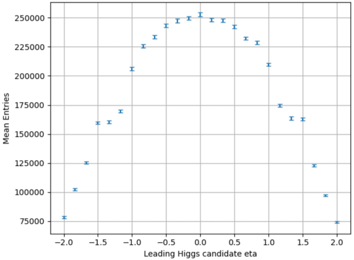


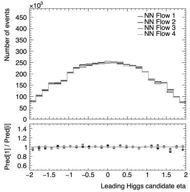


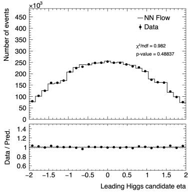


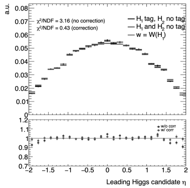
Uncertainty estimation

- **Intrinsic statistics: distributions learned by the model.**
- Neural Network training uncertainty: If we train the NN again, we would obtain different minimum $\rightarrow$ different parameters.
- Deviation from no-tagged data.
- BDT correction UNC: deviations from data in tagged events.



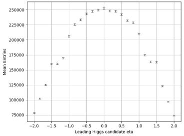


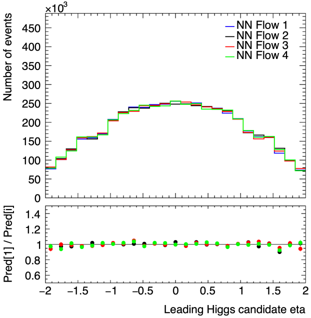


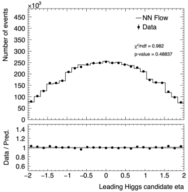


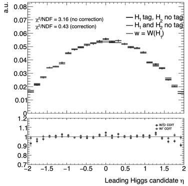
Uncertainty estimation

- Intrinsic statistics: distributions learned by the model.
- **Neural Network training uncertainty: If we train the NN again, we would obtain different minimum $\rightarrow$ different parameters.**
- Deviation from no-tagged data.
- BDT correction UNC: deviations from data in tagged events.



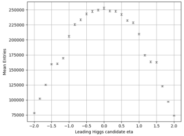


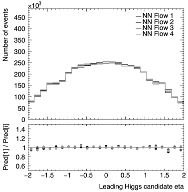


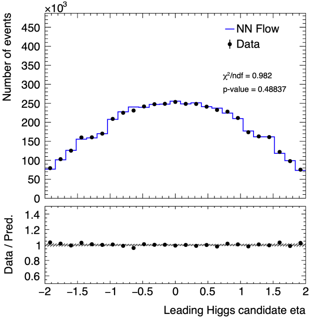


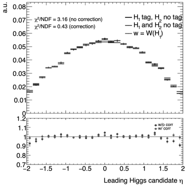
Uncertainty estimation

- Intrinsic statistics: distributions learned by the model.
- Neural Network training uncertainty: If we train the NN again, we would obtain different minimum $\rightarrow$ different parameters.
- **Deviation from no-tagged data.**
- BDT correction UNC: deviations from data in tagged events.



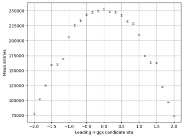


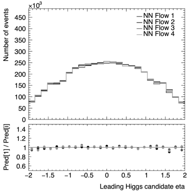


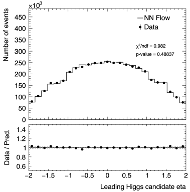


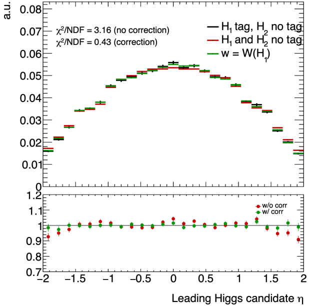
Uncertainty estimation

- Intrinsic statistics: distributions learned by the model.
- Neural Network training uncertainty: If we train the NN again, we would obtain different minimum $\rightarrow$ different parameters.
- Deviation from no-tagged data.
- **BDT correction UNC: deviations from data in tagged events.**



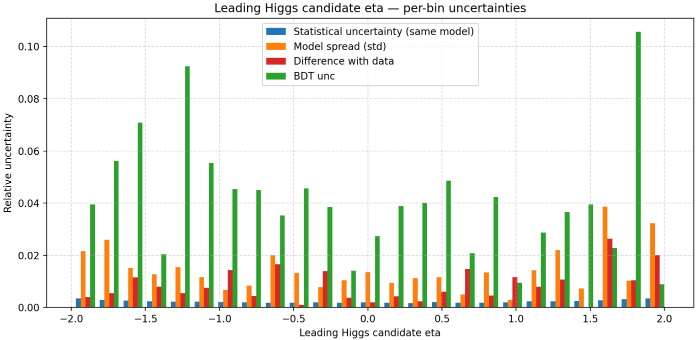




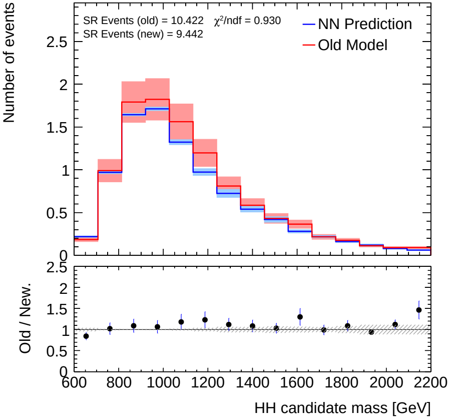


Uncertainty estimation

- **Different sources:**
  - Intrinsic statistics: distributions learned by the model.
  - Neural Network training uncertainty: If we train the NN again, we would obtain different minimum $\rightarrow$ different parameters.
  - Deviation from no-tagged data.
  - BDT correction UNC: deviations from data in tagged events.



Comparison with old model



- The new model reproduces the distributions in **agreement** with the old method.
- It allows us to generate **arbitrary statistics**.

→ This improves the precision of the background estimate: up to $\sim$ 90% **reduction** of the uncertainties.

→ It can be used to train ML classifiers for **signal/background discrimination**.

`generated_background.root`

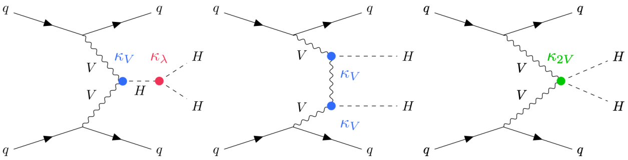
Ready to be shared for background analyses

THANK YOU !

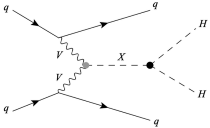
Backup

VBF Production

arXiv:2404.17193v2 [hep-ex]



(c) (d) (e)

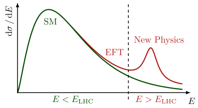


(f)

$$g_{hVV}^{\mathrm{SM}} = \frac{2m_V^2}{v} \qquad g_{hhVV}^{\mathrm{SM}} = \frac{2m_V^2}{v^2} \qquad \lambda_{hhh}^{\mathrm{SM}} = \frac{3m_h^2}{v}$$

(with $V = W, Z, \quad v \simeq 246\,\mathrm{GeV}$)

Boosted vs Resolved

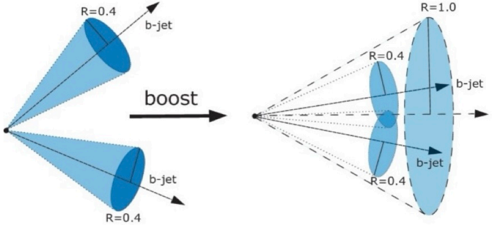


Resolved topology:

- Higgs bosons decay into 4 well-separated *b*-jets.
- Simpler reconstruction (small-R jets).
- **Large QCD background** makes signal extraction harder.

Boosted topology:

- Each Higgs is highly energetic ($p_T \gg m_H$).
- The two *b*-quarks merge into a single large-R jet.
- **Better background rejection** and mass resolution.



Note: *b*-tagging in boosted jets relies on advanced deep learning models, including the transformer-based **GN2X** architecture.

CR/VR yields

Region	noTag	Tag	noTag/Tag
CR	5166178	278	18583.4
VR	4657884	262	17778.2

Gaussian Process Regressor (GPR)

In a GPR, the function values follow a Gaussian distribution:

$$\mathbf{f} \sim \mathcal{N}\big(m(\mathbf{x}),\ K(\mathbf{x}, \mathbf{x}')\big),$$

where $m(\mathbf{x})$ is the mean function and K is the **kernel** encoding correlations.

Kernels used in this work:

- **Constant:** scales the overall variance.

$$k(\mathbf{x}, \mathbf{x}') = \sigma_c^2$$

- **RBF (Radial Basis Function):** smooth variations, different ℓ capture multiple scales.

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\Big(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\ell^2}\Big)$$

- **Dot Product:** adds a global linear trend.

$$k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 + \mathbf{x} \cdot \mathbf{x}'$$

- **White Noise:** models uncorrelated statistical noise.

$$k(\mathbf{x}, \mathbf{x}') = \sigma_n^2\, \delta_{\mathbf{x},\mathbf{x}'}$$

Computing Correction

$$\mathbb{P}(T=1 \mid x) = \frac{\mathbb{P}(x \mid T=1)\cdot\mathbb{P}(T=1)}{\mathbb{P}(x)} \qquad \mathbb{P}(T=0 \mid x) = \frac{\mathbb{P}(x \mid T=0)\cdot\mathbb{P}(T=0)}{\mathbb{P}(x)}$$

$$\frac{\mathbb{P}(T=1 \mid x)}{\mathbb{P}(T=0 \mid x)} = \frac{\mathbb{P}(x \mid T=1)\cdot\mathbb{P}(T=1)}{\mathbb{P}(x \mid T=0)\cdot\mathbb{P}(T=0)}$$

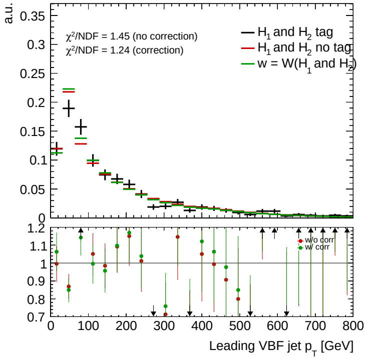
$$\mathbb{P}(T=1 \mid x) = p(x) \qquad \mathbb{P}(T=0 \mid x) = 1 - p(x)$$

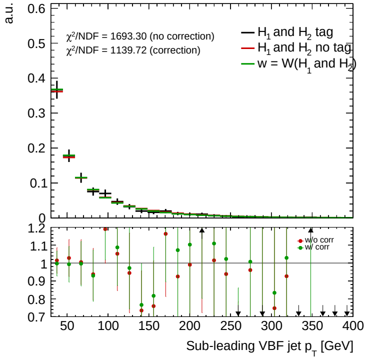
$$\Rightarrow \quad w = \frac{p(x)}{1-p(x)}$$

H1 and H2 simultaneous correction

H1 and H2 correction work well, $prob_{h1}$ and $prob_{h2}$ are independent $\rightarrow$

$$\text{weight}_{h1\,h2} = \frac{prob_{h1} \cdot prob_{h2}}{1 - prob_{h1} \cdot prob_{h2}}$$





Uncertainty estimation

Intrinsic statistics of the model

1. Generate 100 mass samples.

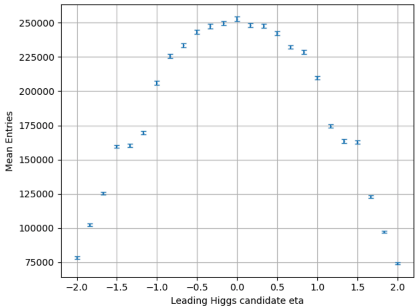
2. Apply the NN to each of them.

—> End up with 100 different distributions: compute **mean** and **standard deviation**.

1. Apply the NN 100 times on the same mass sample.

—> End up with 100 different distributions: compute **mean** and **standard deviation**.

Sum under $\sqrt{}$ the two standard deviations.



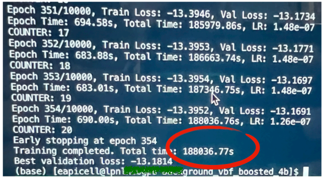
Uncertainty estimation

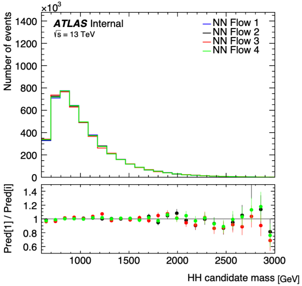
Model Spread

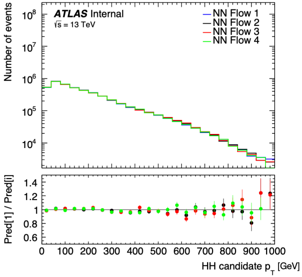
- Train the model as many times as we can.
  (4 for now)

- End up with different models and so different predictions

—> Compute **means** and **standard deviation**



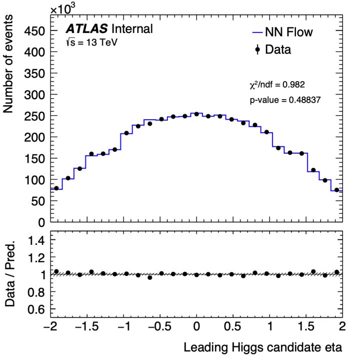


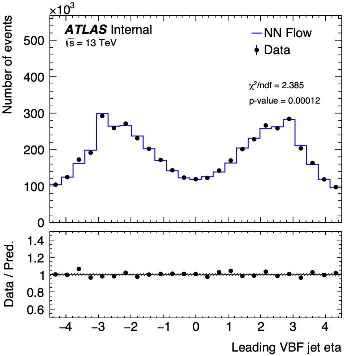


Uncertainty estimation

Deviation from data

Look at the relative (per bin) deviation between prediction and data.





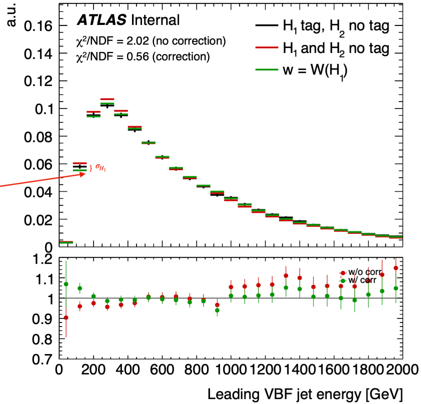
BDT Uncertainty

Add uncertainty due to Tagging (BDT) correction:

1. Apply re-weighting separately on H1 and H2

2. Compare with data -> $\sigma_{H_1}, \sigma_{H_2}$

3. $\sigma_{BDT} = \sqrt{\sigma_{H_1}^2 + \sigma_{H_2}^2}$



Comparison with old model

Estimation of the number of events with the old model:

- The background estimate in 2Pass SR is obtained **automatically** via:

$$N_{\text{SR}}^{\text{old}} = w \cdot N_{\text{SR, 1Pass}}$$

Estimation of the number of events with the new model:

1. Generate events in all regions using a 2D Gaussian Process regressor;
2. Evaluate the ratio
$$f = \frac{\text{SR events}}{\text{CR events}}$$
for 2Pass with the specified cuts;
3. Determine the number of events in the control region from real data, $N_{\text{CR}}^{\text{data}}$;
4. Compute the estimated number of events in the signal region as
$$N_{\text{SR}}^{\text{gen}} = f \cdot N_{\text{CR}}^{\text{data}}\ .$$

Neural Spline Flows

Durkan et al. 2019

Flexible Transformations for Normalizing Flows

- **Normalizing Flow**: invertible map from noise $z \sim p_z$ to data x:

$$p(x) = p_z(f^{-1}(x)) \left| \det \frac{\partial f^{-1}}{\partial x} \right|$$

- Standard flows use **affine transformations** limited flexibility.
- **Neural Spline Flows (NSF)**:
  - Replace affine maps with **monotonic rational-quadratic splines**.
  - Preserve **analytic invertibility + tractable Jacobian**.
- Applications: density estimation, VAEs, image generation.