An Unexpected Application of Fairness to Higgs Boson Detection

Karel de Vries

Nikhef

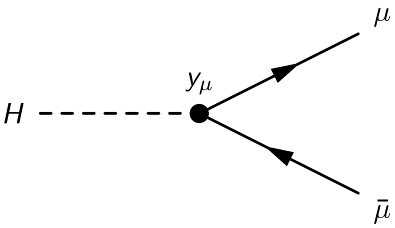
November 8, 2024

Table of Contents

Higgs decay to Muons

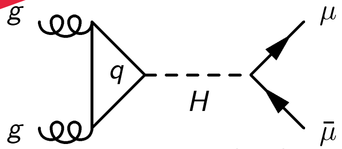
$H \to \mu\mu$

- Yukawa coupling $\propto$ fermion mass
- Fermion masses are free parameters of SM and have to be determined experimentally
- Coupling to muon (μ) not observed
- ATLAS and CMS found evidence
- Simulated ATLAS Run 2 data
- $\sqrt{s} = 13\,\text{TeV}$, $\mathcal{L} = 139\,\text{fb}^{-1}$

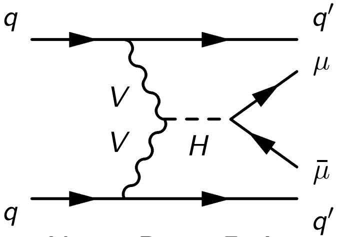


$$M_\mu = \frac{y_\mu \cdot v}{\sqrt{2}}$$

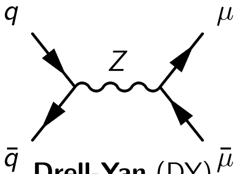
$\mu\mu$ Production



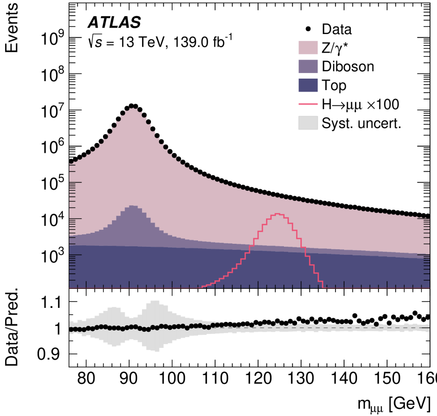
Gluon Fusion (ggF)



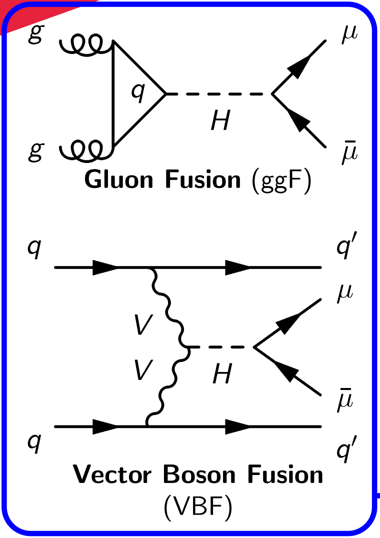
Vector Boson Fusion (VBF)

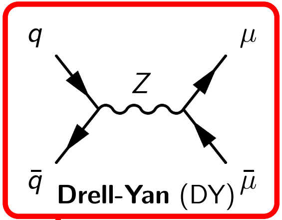


Drell-Yan (DY)



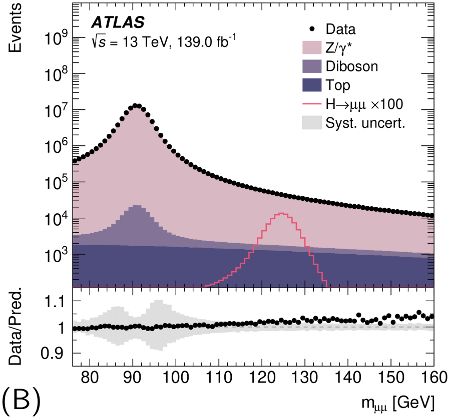
$\mu\mu$ Production





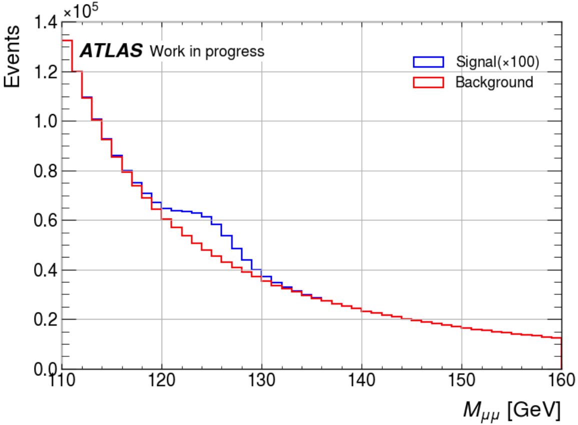
Background (B)

Signal (S)



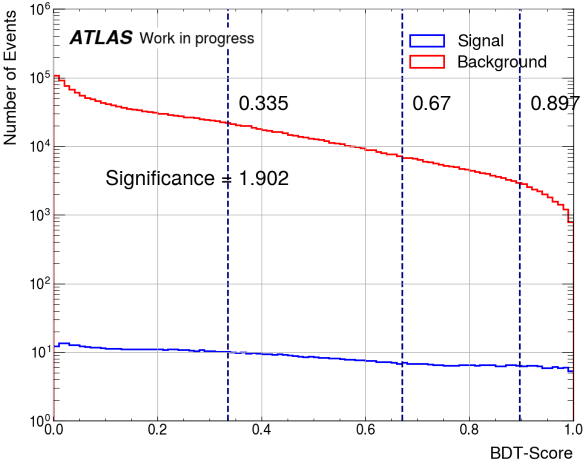
Outline Analysis Strategy

- Fit S+B-model to dimuon mass $M_{\mu\mu}$ spectrum
- Significance $= S/\Delta$
- Uncertainties:
  - Statistical $\Delta_{stat} = \sqrt{B}$
  - Systematic Δ_{syst}
- Machine Learning (ML)



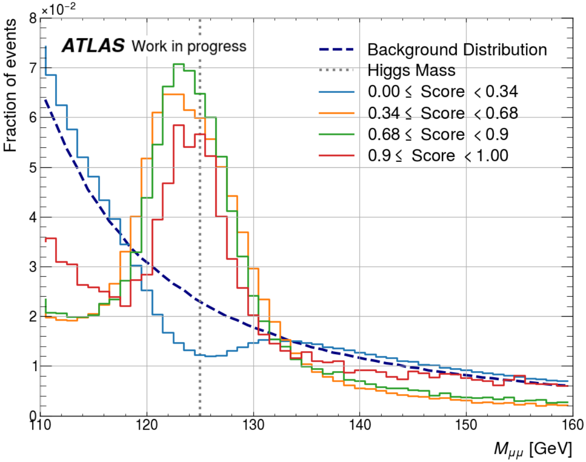
Enrich S/B with Machine Learning

- Train ML model using detector observables
- Boosted Decision Tree (BDT)
- Categorise by BDT score
- Extract S and B in each category
- Maximise the total significance



Mass Sculpting

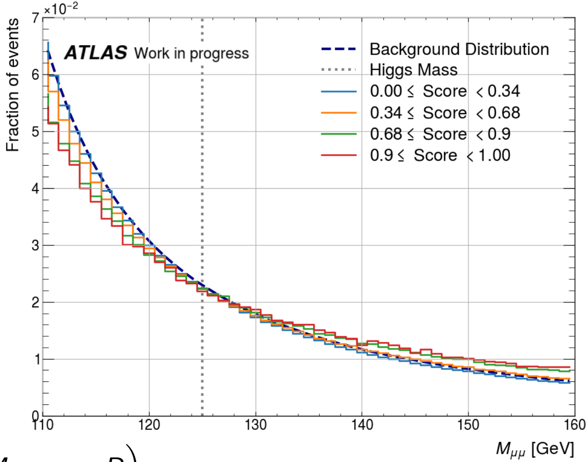
- Perform fit to $M_{\mu\mu}$ spectrum of each category
- Classifier can change $M_{\mu\mu}$ spectrum
- Fit too much S
- Mass sculpting can cause Δ_{syst}
- Run 2 Legacy (R2L): trained on events with $M_{\mu\mu} \in [120, 130]$ GeV



Fairness

Fairness in Particle Physics

- Use fairness to reduce Δ_{syst}
- Same shape $M_{\mu\mu}$ distribution of B events for each category
- Equal Opportunity for B (EOP_{B})
- Fairness



$$P\Big(R(x) \in [r_1, r_2] \big| M_{\mu\mu}, y = B\Big) = P\Big(M_{\mu\mu}, y = B\Big)$$

Example of Discrimination

- Example from: *Hardt, Price, Srebro, 2016*
  `https://arxiv.org/pdf/1610.02413.pdf`
- ML and bank loans
- Black people got rejected the **most** given they never defaulted on a loan
  Asian people got rejected the **least** given they never defaulted on a loan
- They did not have EOP of getting the loan

Equal Odds

- Equal Odds (EOD) is when EOP is satisfied for both classes:

$$P\Big(R(x) \in [r_1, r_2] | M_{\mu\mu}, y\Big) = P\Big(M_{\mu\mu}, y\Big)$$

- Stronger than EOP
- It turns out that in the case of $H \to \mu\mu$: $\mathrm{EOP_S}$ always satisfied
- In the case of $H \to \mu\mu$: $\mathrm{EOD} = \mathrm{EOP_B}$

Forcing EOP: Post Integration

- Strategy from the literature: Post Integration (PI)
- Train classifier R with $M_{\mu\mu}$ as input
- Integrate out $M_{\mu\mu}$:

$$R_{\text{PI}}(x) = \int_{110}^{160} R(M_{\mu\mu}, x) P(M_{\mu\mu}) dM_{\mu\mu}$$

- Effective, but can decrease performance a lot. Therefore used in combination with R2L (R2L+PI)
- It is applied after training, therefore the actual ML trained classifier is not fair

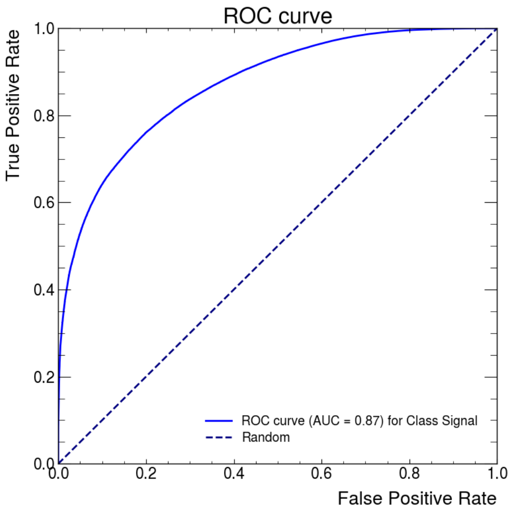
ROC-Split

ROC-curve

- Given a threshold t: $R(x) \geq t$ is classified as S and $R(x) < t$ as B
- True positive rate (tpr) is the chance of correctly classifying S
- False positive rate (fpr) is the chance of falsely classifying B as S
- Receiver Operator Characteristic (ROC):

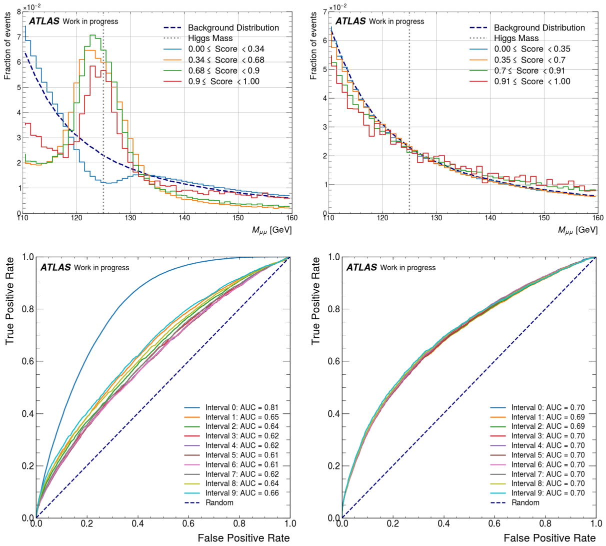
$$\mathrm{ROC}(t) = \big(\mathrm{fpr}(t), \mathrm{tpr}(t)\big)$$

- Area Under the Curve (AUC)



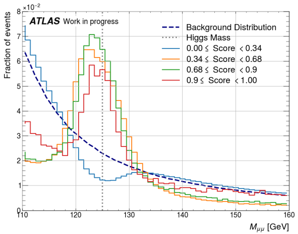
EOD and ROC-curves

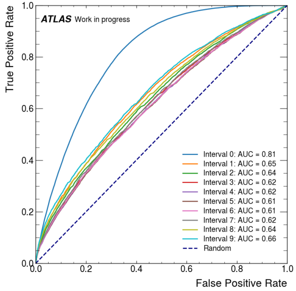
- EOD is satisfied when the ROC-curve is independent of $M_{\mu\mu}$
- When EOP_S is satisfied: $EOD = EOP_B$
- Consequence: EOP_B is satisfied when EOP_S is satisfied and the path of the ROC-curve is independent of $M_{\mu\mu}$



ROC-Split

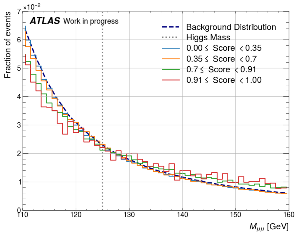
- Algorithm to train classifiers satisfying EOP:
  1. Divide $M_{\mu\mu}$ up in bins and determine $\{\mathrm{AUC}_i\}$
  2. Sample from a bin with $p_i = 2(1 - \mathrm{AUC}_i)$
  3. Train model on this new set and repeat
- Can be applied to ML architectures using epochs
- Flexibility: choice between fairness and performance

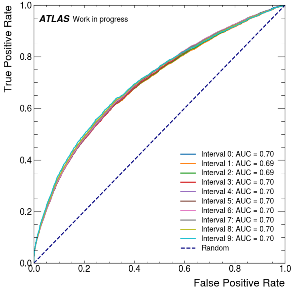




ROC-Split

- Algorithm to train classifiers satisfying EOP:
  1. Divide $M_{\mu\mu}$ up in bins and determine $\{\mathrm{AUC}_i\}$
  2. Sample from a bin with $p_i = 2(1 - \mathrm{AUC}_i)$
  3. Train model on this new set and repeat
- Can be applied to ML architectures using epochs
- Flexibility: choice between fairness and performance





Results

Results

- Similar significance for the three methods
- $\Delta_{\text{stat}} >> \Delta_{syst}$
- Impact of fairness limited for this analysis with the current available data

	Significance
R2L	1.42
ROC-Split	1.43
R2L+PI	1.43

Conclusion & Outlook

Conclusion & Outlook

- Two new methods for reducing ML bias for $H \to \mu\mu$:
  1. ROC-Split
  2. R2L+PI
- Both similar significance as R2L
- $\Delta_{\text{stat}} >> \Delta_{syst}$
- Reduction of Δ_{syst} becomes more important as more data becomes available
- Create a measure to quantify ML biases
- Construct a general decorrelation strategy with fairness

Thank you!

Event Selection

	Event selection
Muons	At least one $\mu^+\mu^-$ pair $\lvert\eta\rvert < 2.7$ $p_T^{\mu_1} > 27$ GeV $p_T^{\mu_2} > 15$ GeV
ggF/VBF	No extra leptons No b-jet

Event selection

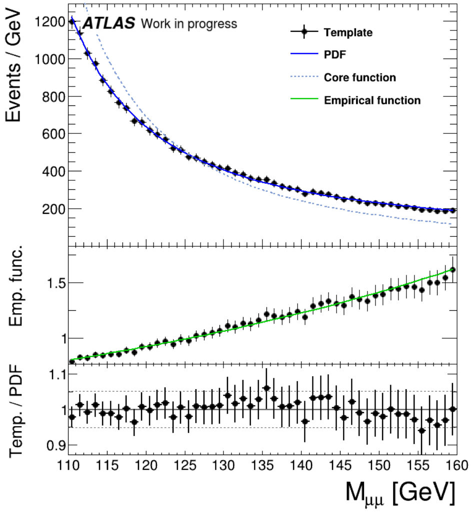
Channel Name	Event Selection
0Jet	$N_j = 0$
1Jet	$N_j = 1$
2Jet	$N_j \geq 2$ $m_{jj} < 400 \quad \text{and} \quad \|\eta_{j^l} - \eta_{j^s}\| < 2.5$
VBF	$N_j \geq 2$ $m_{jj} > 400 \quad \text{or} \quad \|\eta_{j^l} - \eta_{j^s}\| > 2.5$
ggFAll	$N_j < 2 \quad \text{or} \quad (m_{jj} < 400 \quad \text{and} \quad \|\eta_{j^l} - \eta_{j^s}\| < 2.5)$
AllJet	No selections

Input observables

Selections	Variable	Description
All selections	$p_T^{\mu\mu}$	Transverse momentum of the dimuon system
	$y_{\mu\mu}$	Rapidity of the dimuon system
	$\cos\theta^\star$	Cosine of the muon decay angle
Events with 1+ jets	$p_T^{j^l}$	Transverse momentum of the leading jet
	η_{j^l}	Pseudo rapidity of the leading jet
	$\Delta\phi_{\mu\mu,j^l}$	$\lvert\phi_{\mu\mu} - \phi_{j^l}\rvert$
	$N_{\text{tracks}}^{j^l}$	Number of ID tracks of the leading jet
Events with 2+ jets	$p_T^{j^s}$	Transverse momentum of the subleading jet
	η_{j^s}	Pseudo rapidity of the subleading jet
	$\Delta\phi_{\mu\mu,j^s}$	$\lvert\phi_{\mu\mu} - \phi_{j^s}\rvert$
	$N_{\text{tracks}}^{j^s}$	Number of ID tracks of the subleading jet
	p_T^{jj}	Transverse momentum of the dijet system
	m_{jj}	Mass of the leading jet
	y_{jj}	Rapidity of the dijet system
	$\Delta\phi_{\mu\mu,j^l}$	$\lvert\phi_{\mu\mu} - \phi_{j^s}\rvert$
	H_T	Scalar sum of jet transverse momenta
	$\not{p}_T$	Missing transverse momentum
No jet selections	N_j	Number of jets

Model

- Fit S+B-model to $M_{\mu\mu}$ spectrum
- S: Gaussian-like
- Theoretical core function: Breit-Wigner(BW) or Drell-Yan (DY)
- Empirical function $\mathcal{F}_{\mathcal{E}}$
- B-function: core function $\times\mathcal{F}_{\mathcal{E}}$



Signal Model

- S: double-sided Cristal Ball (CB)
- Fit on simulated data
- Each category separately
- Shape of S fixed in S+B-model

$$CB = \begin{cases} e^{-\frac{1}{2}t^2} & \text{for } -\alpha_{\text{left}} \leq t \leq \alpha_{\text{right}} \\ e^{-\frac{1}{2}\alpha_{\text{left}}^2}\left[\frac{\alpha_{\text{left}}}{n_{\text{left}}}\left(\frac{n_{\text{left}}}{\alpha_{\text{left}}} - \alpha_{\text{left}} - t\right)\right]^{-n_{\text{left}}} & \text{for } t < -\alpha_{\text{left}} \\ e^{-\frac{1}{2}\alpha_{\text{right}}^2}\left[\frac{\alpha_{\text{right}}}{n_{\text{right}}}\left(\frac{n_{\text{right}}}{\alpha_{\text{right}}} - \alpha_{\text{right}} + t\right)\right]^{-n_{\text{right}}} & \text{for } t > \alpha_{\text{right}}, \end{cases}$$

Background functions

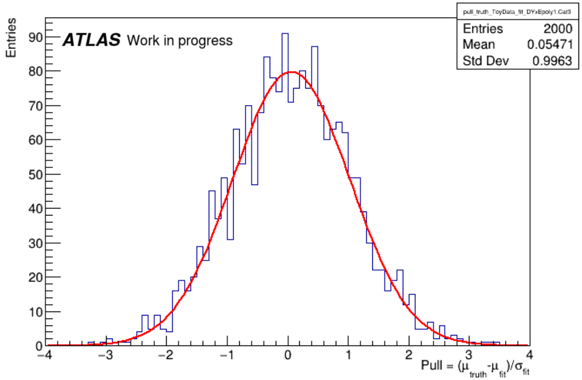
$$\text{BW} = \frac{1}{(M_{\mu\mu} - m_Z)^2 + \frac{\Gamma_Z^2}{4}}$$

$$\text{DY} = \frac{k}{(M_{\mu\mu}^2 - m_Z^2)^2 + m_Z^2\Gamma_Z^2}$$

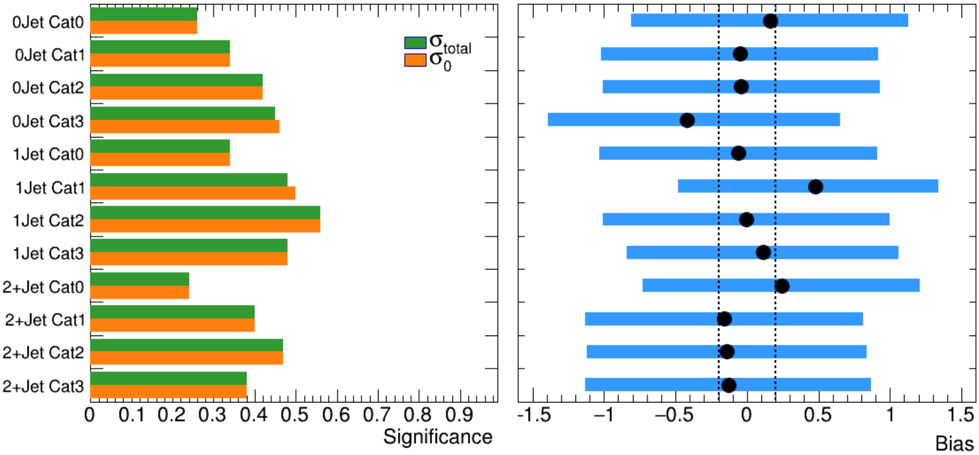
$$\mathcal{F}_{\mathcal{E}} = \begin{cases} \text{PowerN} & = M_{\mu\mu}^{a_0 + \cdots + a_{N-1} M_{\mu\mu}^{N-1}} \\ \text{EpolyN} & = e^{a_1 M_{\mu\mu} + \cdots + a_N M_{\mu\mu}^N} \\ \text{PolyN} & = a_1 M_{\mu\mu} + \cdots + a_N M_{\mu\mu}^N \end{cases}$$

Bias Studies

- Signal strength: $\mu = \frac{S}{S_{SM}}$
- Fit S+B-model on 2000 toy sets
- Pull $= \frac{\mu_{truth} - \mu_{fit}}{\sigma_{fit}}$
- Mean pull is spurious signal uncertainty Δ_{ss}



Significance



More Data

- High Luminosity LHC
- Extrapolated dataset to $\mathcal{L} = 3000\,\text{fb}^{-1}$
- $\Delta_{\text{stat}} \leq \Delta_{ss}$