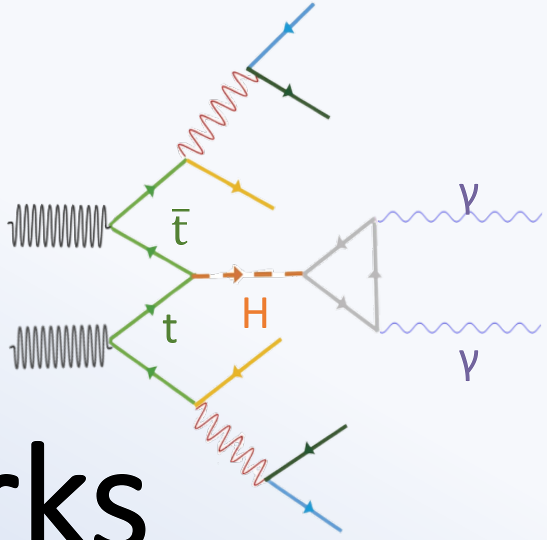


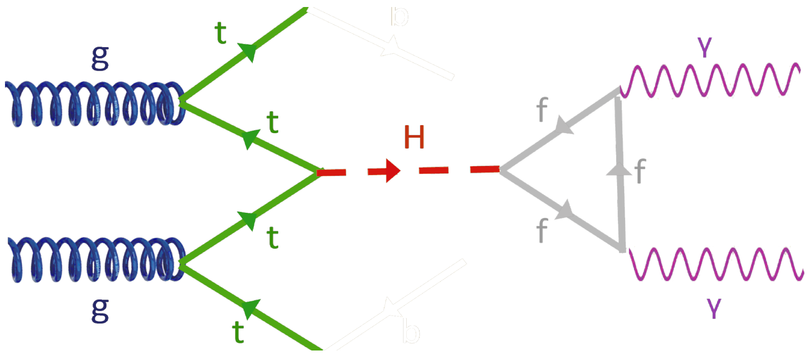
Adversarial Neural Networks for ttH ($H\rightarrow\gamma\gamma$)

Prof. Philip Clark, Emily Takeva

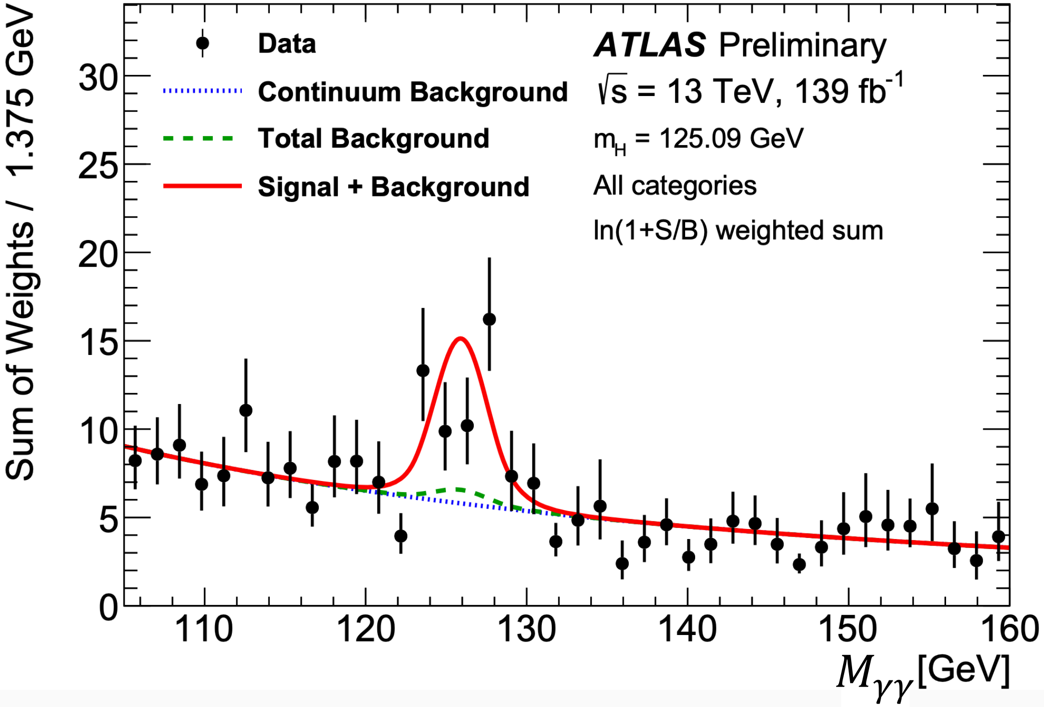
12/04/2021

Physics Motivation for ttH(γγ)

We need to measure it as accurately as possible, in order to unravel the mysteries of the new fundamental Top quark – Higgs boson interaction discovered in 2018.



- SM ttH cross section: **σ = 0.507 pb**
- SM H→γγ branching ration: $\boldsymbol{B_{\gamma\gamma} = 2.27\times10^{-3}}$
- Very high signal purity and fully reconstructable invariant mass
- High photon reconstruction and isolation efficiency due to the high resolution of the ATLAS electromagnetic calorimeter
- Backgrounds determined in fit of $M_{\gamma\gamma}$ side-bands with 1-parameter function

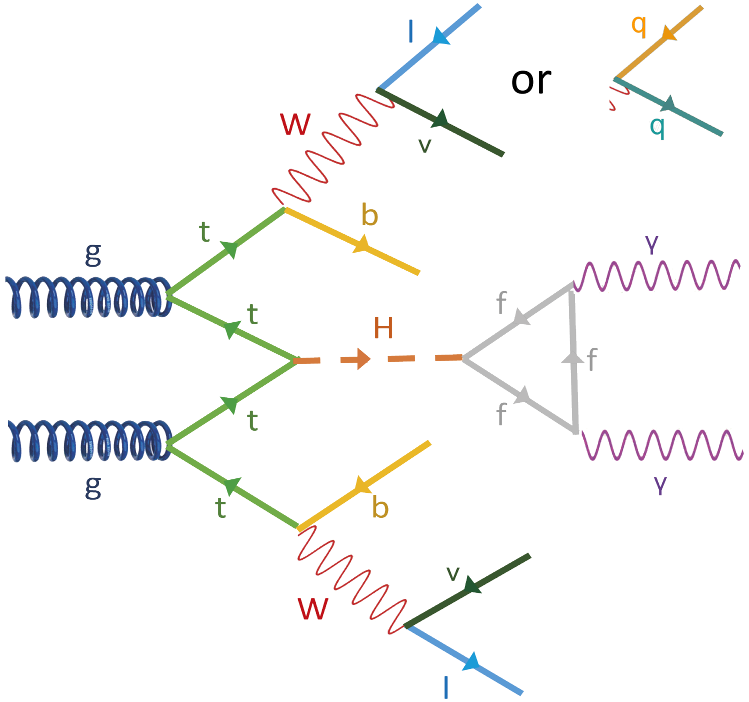


http://cds.cern.ch/record/2668103/files/ATLAS-CONF-2019-004.pdf

- Photons: reconstructed from calorimeter clusters formed using a dynamical, topological cell clustering based algorithm, selection requires ≥ 2, where the photons with highest p_T are selected as candidates for the diphoton system

- Jets: reconstructed using anti-K_T algorithm

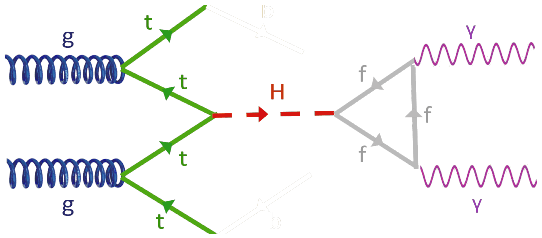
- BDT used to reconstruct top decays and define event categories



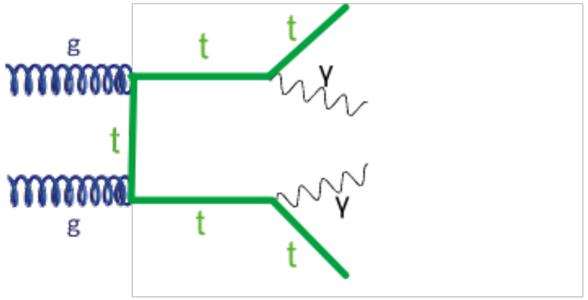
Data and Simulated Samples

Simulated MC data with

- 105 – 160 GeV mass range for $M_{\gamma\gamma}$
- Signal ttH :



Background ttyy:

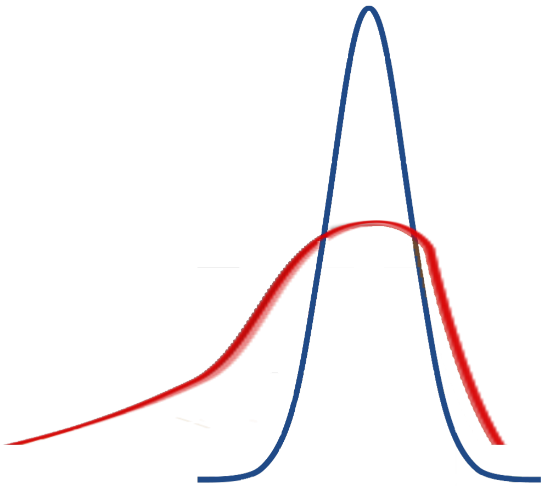


NTNI (non tight, non isolated photons) data as an approximation of background with Full Run2 dataset, all analysis cuts + # of jets $\gtrapprox$ 3

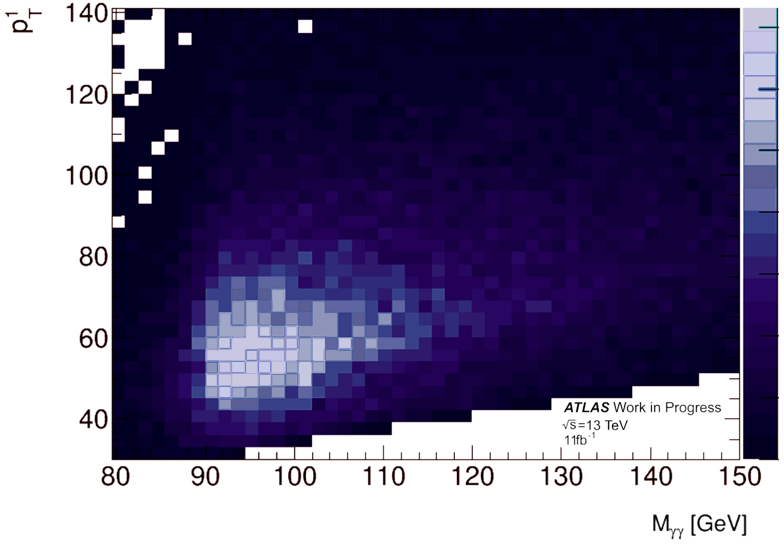
- Tight refers to identification requirement, which accounts for photon shape in the calorimeter. Tight is used for when calorimeter assigns higher degree of confidence that this is a prompt photon, loose for smaller confidence
- Isolated refers to hadronic activity (tracks, calorimeter signals) around a photon. It is used to separate QCD jet from photons , QCD jets have a lot of activity, prompt photons have little
- TI (tight, isolated photons) is used for extracting final result (best approximation to signal)

Problem

Rejecting background using photon kinematics could sculpt the background in case of correlations of the photon kinematics with $M_{\gamma\gamma}$



Example of sculpting, which would prevent the 1-parameter function fit to the side-band where in our case blue integrated area distribution could be the signal, and red the background after background rejection .



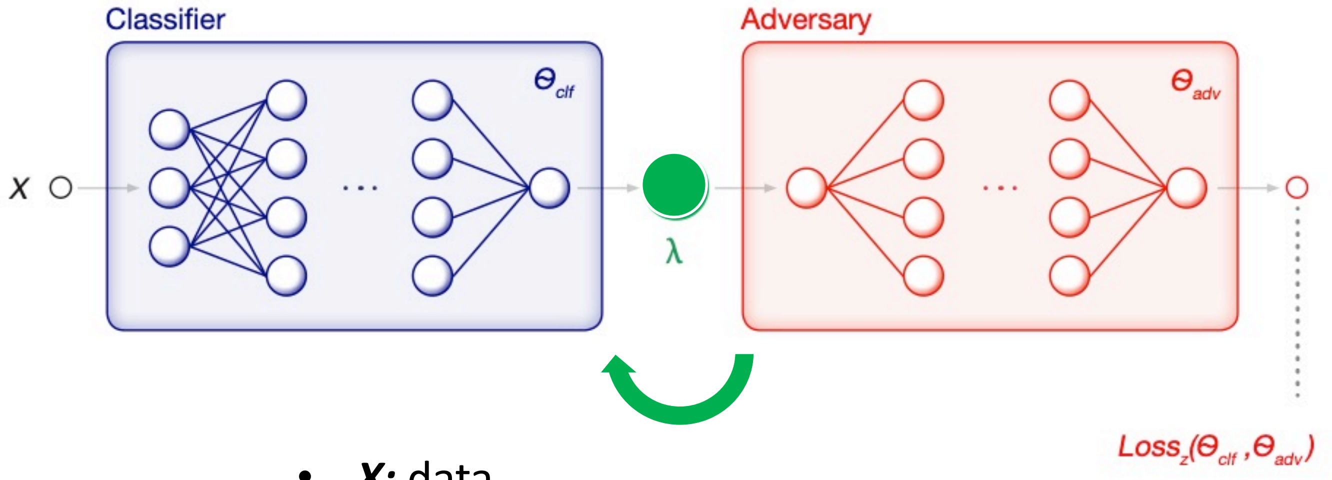
The reason for the sculpting is the strong correlations between some of the photon kinematic variables with the $M_{\gamma\gamma}$ distribution. Example above is the leading photon's distribution in NTNI data.

Idea

- Solution: Adversarial Neural Networks
- Binary classifier function is trained using two neural networks with the idea to find the balance between minimizing loss function J_{cls} and maximizing J_{adv}:

$$min_{\theta_{cls.}} max_{\theta_{adv.}} J_{FinalClassifier} = J_{cls}(\theta_{cls.}) - \lambda J_{adv}(\theta_{cls.}\theta_{adv.})$$

Classifier: trained to use the photon kinematic variables to reject the background

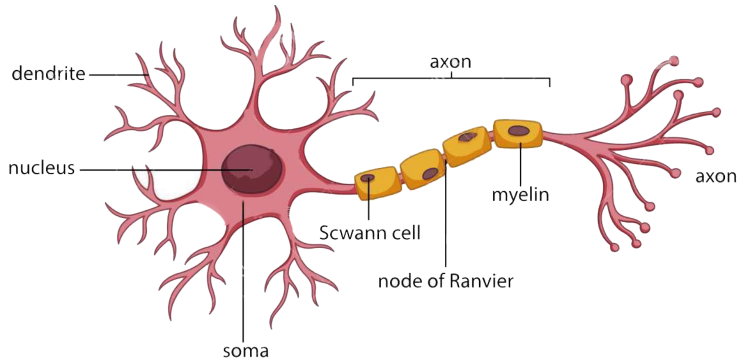


Adversary: trained to decorrelate the variables from $M_{\gamma\gamma}$

- ***X:*** data
- $\boldsymbol{\theta_{cls.}}$ and $\boldsymbol{\theta_{adv.}}$: weights parametrizing classifier and adversary
- ***λ > 0:*** controls the performance of J

Neural Networks

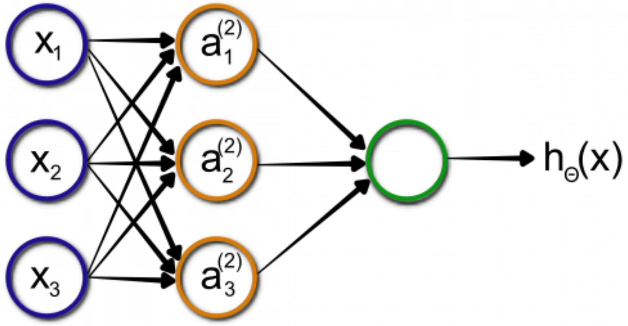
How the idea was born



- Neurons in the brain carry information by transmitting electrical impulses (signals) and have three basic parts: a cell body, an axon and dendrites

- The dendrites receive information (input), the nucleus processes the received information and the axon sends the processed information to other neurons (output).

Implementation in ML



- A neural network in ML is a collection of units (neurons), which transmit and process information
- Hypothesis function h(x) :

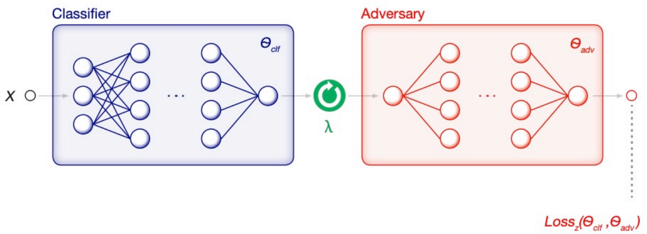
$$h_\Theta(x) = g(\Theta_{10}^{(2)} a_0^{(2)} + \Theta_{11}^{(2)} a_1^{(2)} + \Theta_{12}^{(2)} a_2^{(2)} + \Theta_{13}^{(2)} a_3^{(2)})$$

where Θ are the weights of the cost function

- Cost function:

$$J(\Theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} h_\Theta(x^{(i)}) + (1-y_k^i)\log(1-h_\Theta(x^{(i)})k)] + \frac{r}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(\Theta_{ij}^{(l)})^2$$

The complex dynamic of two neural networks



- Adversarial Neural Networks
- Find best balance between **using the photon kinematic variables for further background rejection** and **fixing the problem which comes from that, by de-correlating those variables from $M_{\gamma\gamma}$**

Jenson-Shannon Divergence (JSD)

How do we quantify sculpting?

- Idea is to construct a metric of background rejection ($\varepsilon_{bkg.}$ = $\frac{N_{bkg}^{accept}}{N_{bkg}^{total}}$) vs. background sculpting (JSD factor)
- JSD is a generalization of the Kullback-Leibler divergence:

$$KL(A \,||\, B) = -\underbrace{\textstyle\sum_i A_i \log_n B_i}_{1} + \textstyle\sum_i A_i \log_n A_i$$

A = $M_{\gamma\gamma}$
B = $M_{\gamma\gamma}^{ANN}$

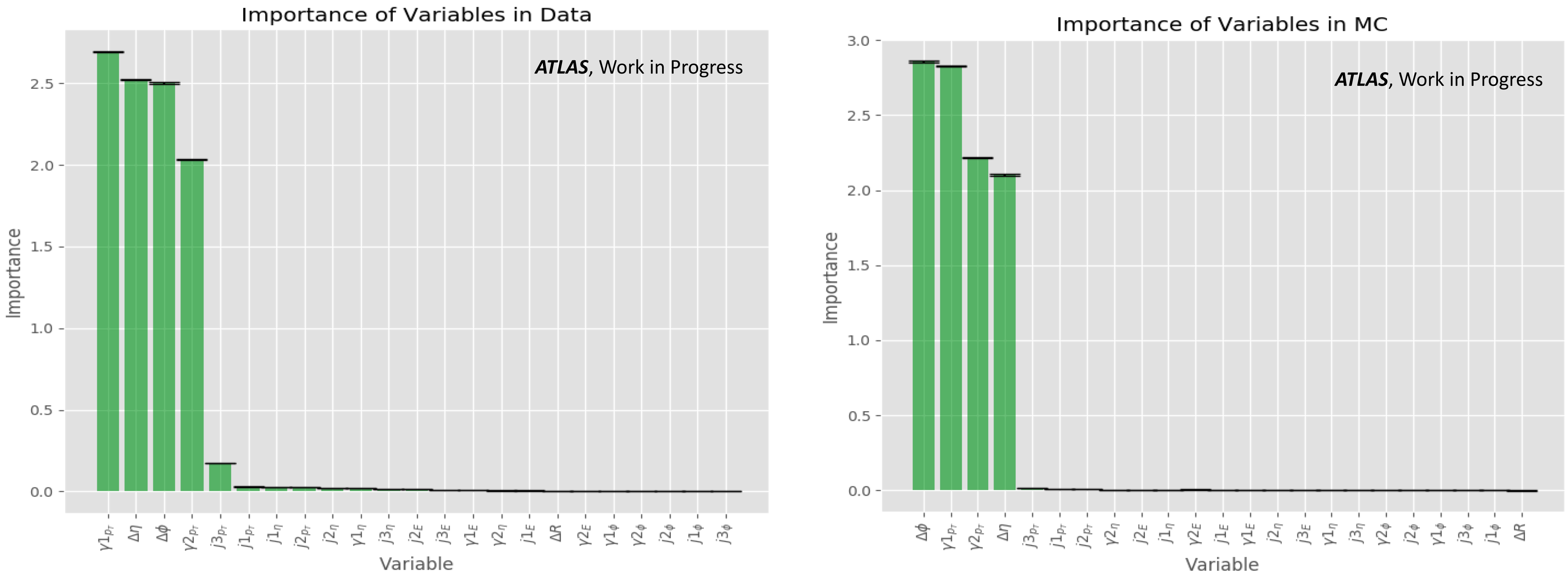
where A and B are the two distributions we are comparing, i are the discrete bins

- For identical A and B, KL = 0, for completely different A and B, KL = 1
- **JSD** avoids the instabilities in KL (ex. For every bin i where A_i > 0 but B_i=0, 1→∞)

$$JSD(A \,||\, B) = \frac{1}{2}(KL(A \,||\, M) + KL(B \,||\, M))$$

M = $\frac{A+B}{2}$

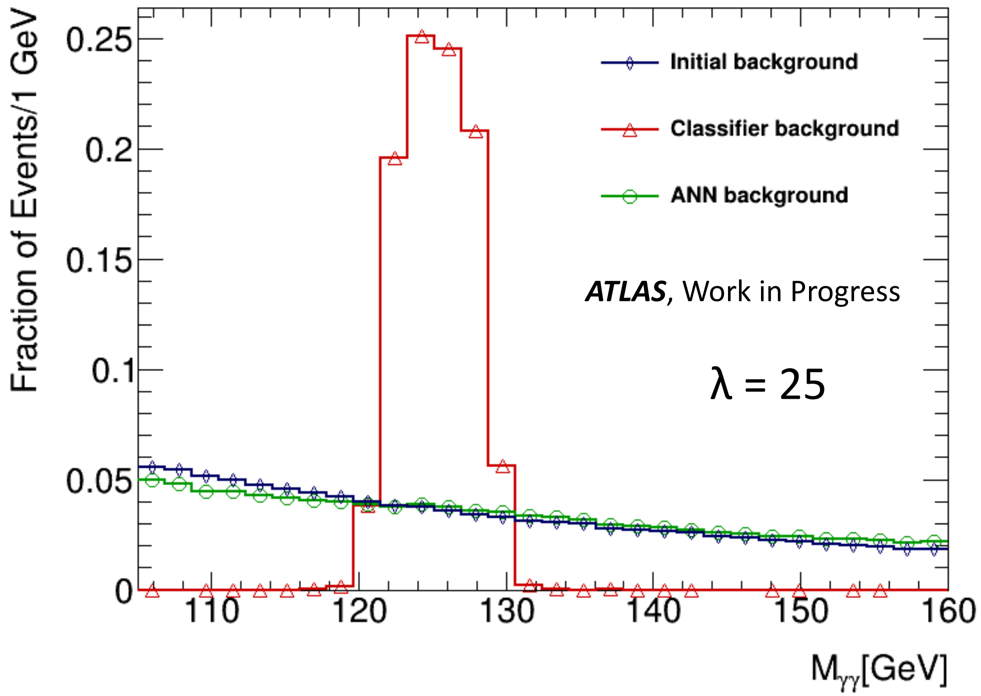
Variable Ranking



Variables used for training:

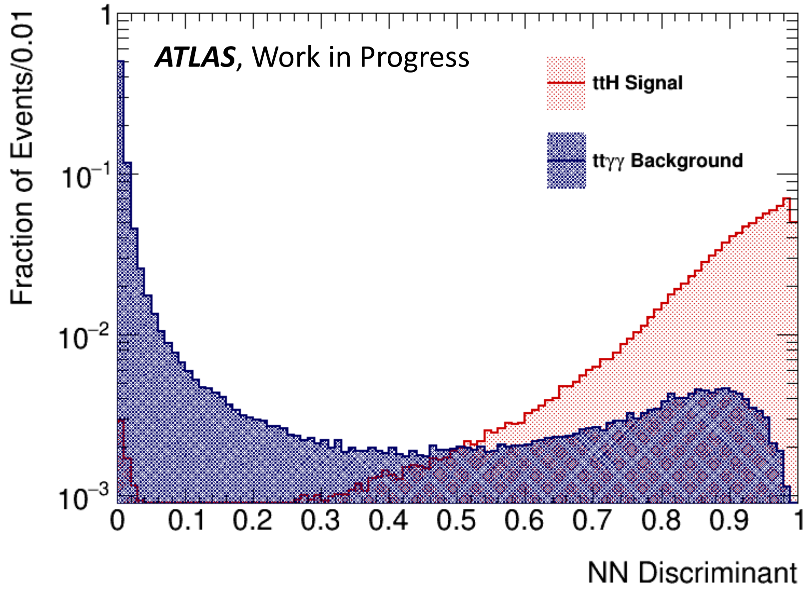
- p_T, E, η, ϕ, Δϕ, Δη and ΔR of the leading and sub-leading photons
- p_T, E, η, ϕ of leading, sub-leading and third jets

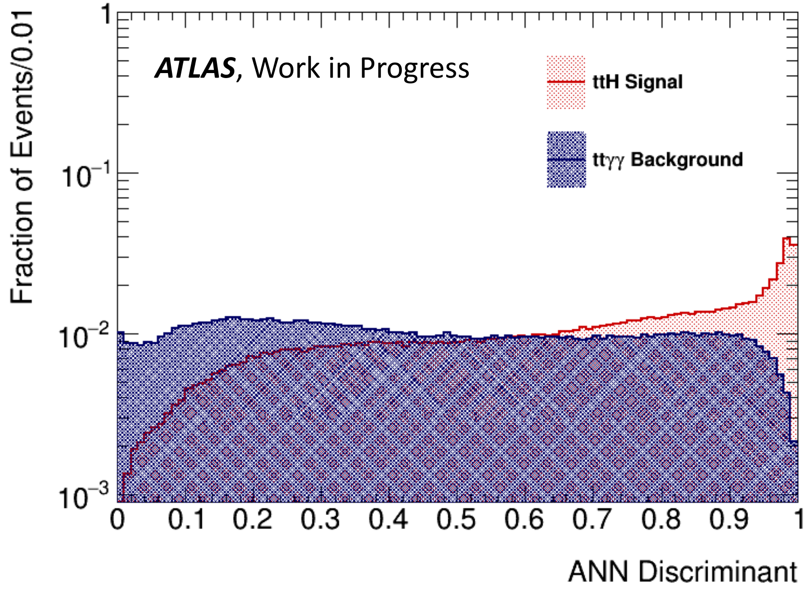
Simulated Samples, Results



1) classifier accuracy: 90.1 %
2) classifier accuracy in signal: 94.5 %
3) classifier accuracy in background: 85.6 %
----
4) ANN accuracy: **60.5%**
5) ANN accuracy in signal: **66.8%**
6) ANN accuracy in background: **54.3%**
ROC = 67.0 %
$JSD_{(105-160)\ GeV}$ **= (0.04 ± 0.01) %**
$JSD_{(120-130)\ GeV}$ **= (0.04 ± 0.04) %**

Sculpting minimized!





Fitting the $M_{\gamma\gamma}$ distribution after ANN training shows a good agreement with an exponential of first order.

Initial with first order exponential:

$\frac{\chi^2}{ndf}$ = 0.76

prob = 82%

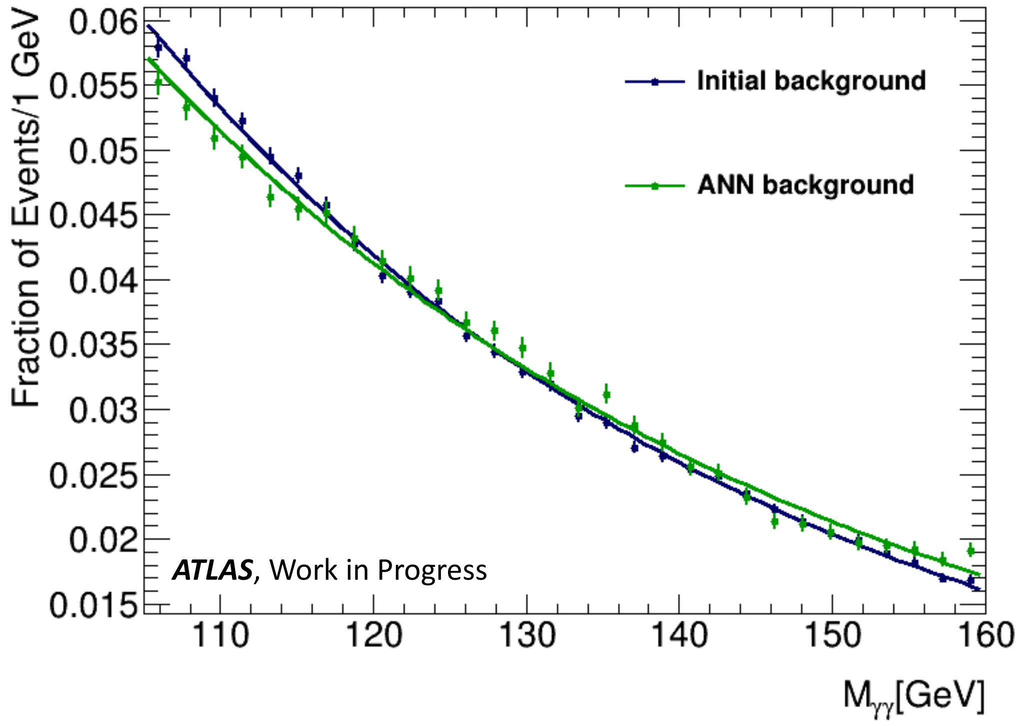
ANN with first order exponential:

$\frac{\chi^2}{ndf}$ = **1.95**

prob = 2%

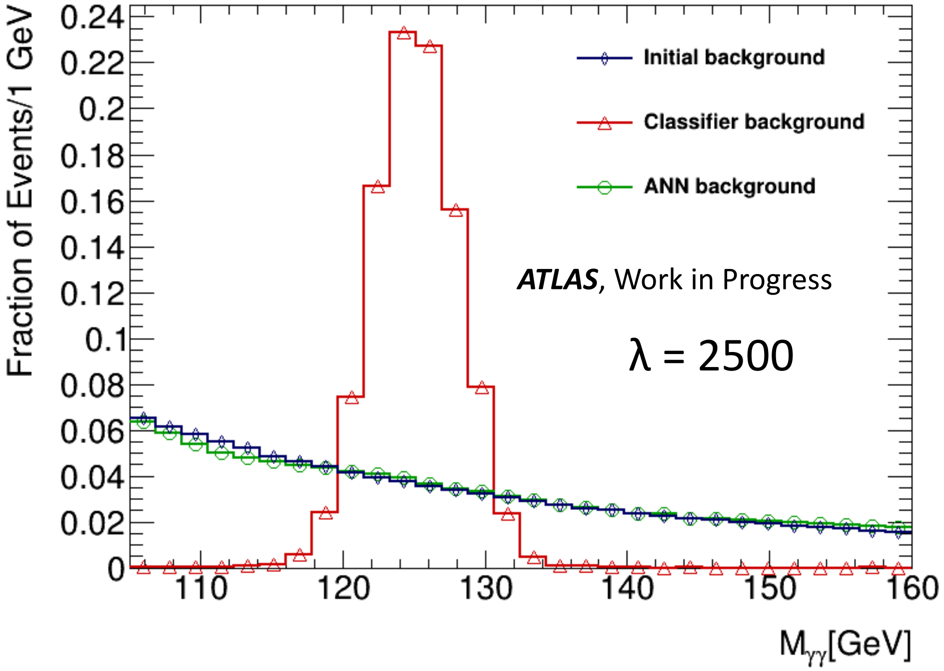
$\frac{\chi^2}{ndf}$ -> goodness of fit

prob -> probability that the values are independent, or significance of $\frac{\chi^2}{ndf}$



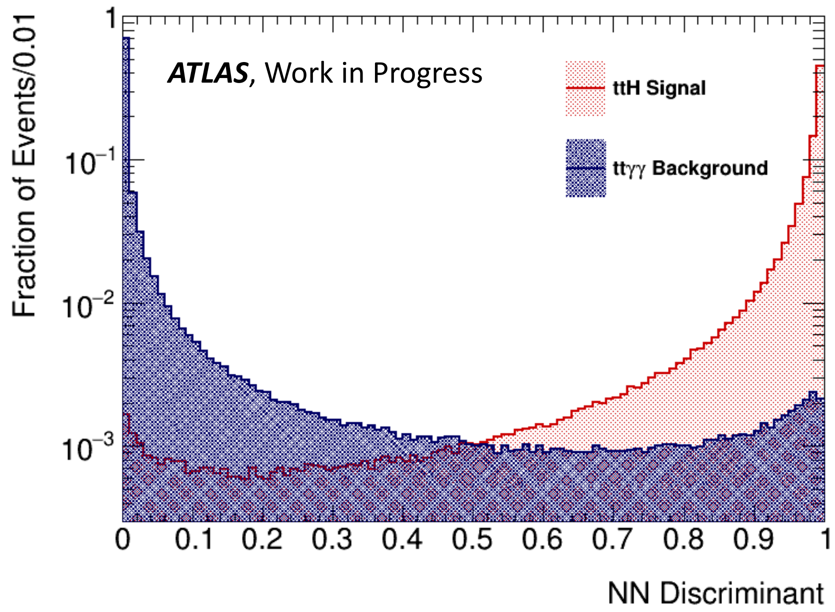
Overall, the sculpting in simulated data was removed, while keeping efficiency optimal and modeling simple.

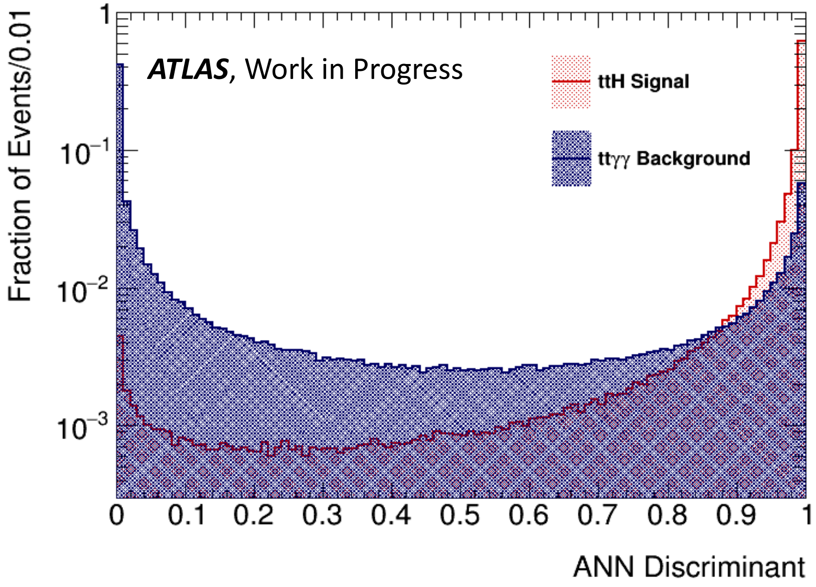
NTNI Data, Results



1) classifier accuracy: 95.2 %
2) classifier accuracy in signal: 96.1 %
3) classifier accuracy in background: 94.3 %
----
4) ANN accuracy: 83.4 %
5) ANN accuracy in signal: 95.7 %
6) ANN accuracy in background: 71.1 %
ROC = 0.94
$JSD_{(105-160)\ GeV}$ = **(1.14 ± 0.01) %**
$JSD_{(120-130)\ GeV}$ = **(0.05 ± 0.01) %**

Sculpting minimized!





NTNI Data Results, Fitting

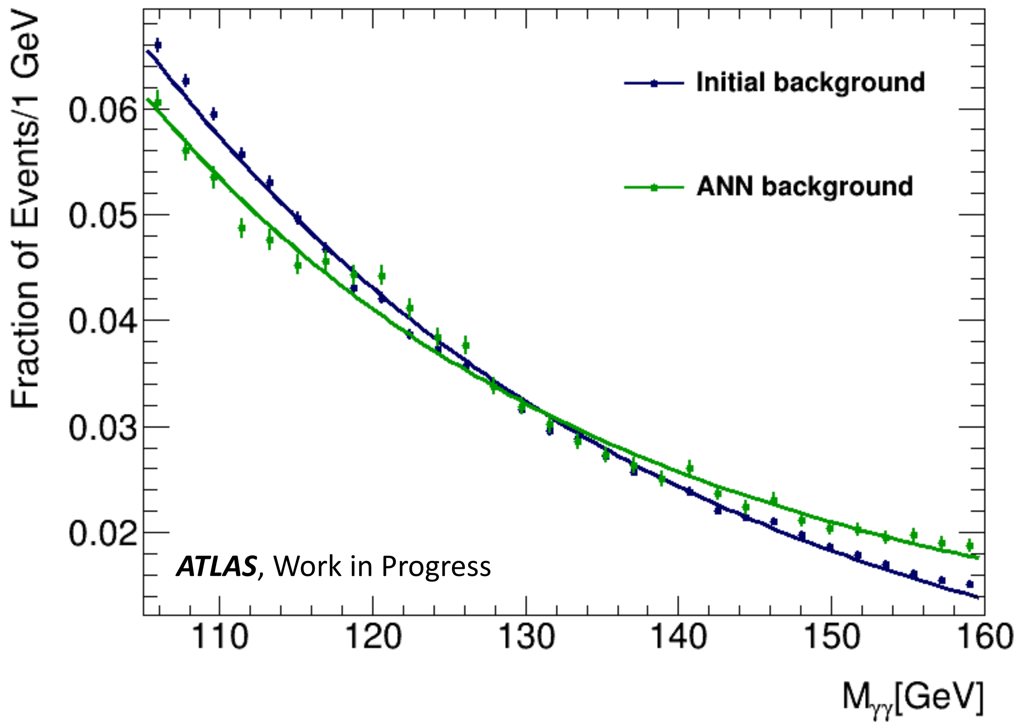
Fitting the $M_{\gamma\gamma}$ distribution after ANN training shows a similar agreement to the initial distribution with an exponential of second order.

Initial with first order exponential:

$\frac{\chi^2}{ndf}$ = 3.32

ANN with first order exponential:

$\frac{\chi^2}{ndf}$ **= 2.84**



Overall, the sculpting in real NTNI data was also removed, while keeping efficiency excellent and modeling simple.

Conclusions and Next Steps

Next Steps

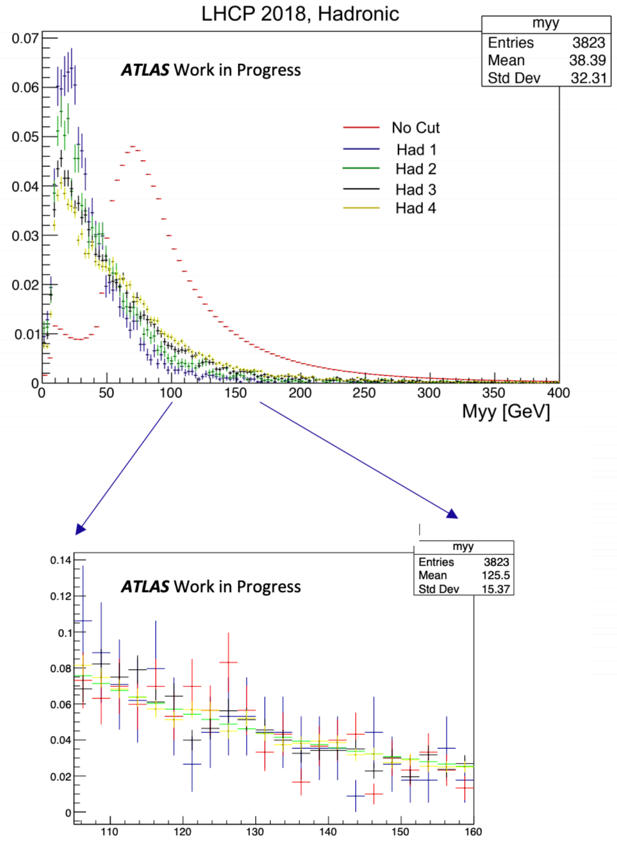
- Improve the background sculpting tests
- Improve the sensitivity with feature engineering
- Determine the optimal event categorization, which yields the optimal sensitivity

Conclusions

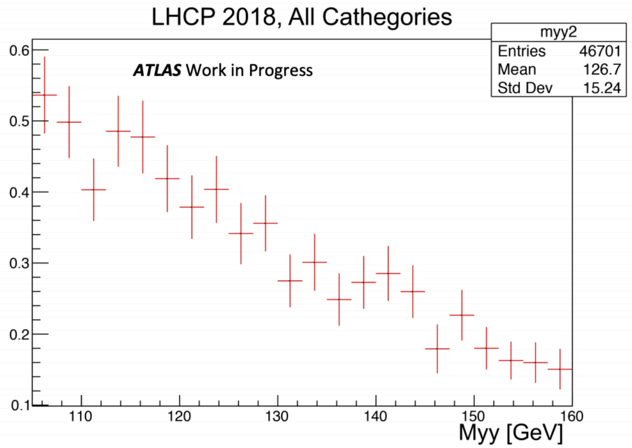
- Rejecting background using photon kinematics sculpts the background due to correlations of the photon kinematics with $M_{\gamma\gamma}$
- An adversarial neural network platform was proposed and adapted for the purpose of rejecting background events with maximum efficiency in the ttH(γγ) channel while dealing with the problem of sculpting.
- Significant reduction of the sculpting observed in MC and Data, while efficiencies are kept optimal.
- We are a step closer to better constrains on the top-Higgs Yukawa coupling, whose precise measurement could be a doorway towards exciting new physics.

Backup

Scaled JSD Check in Run2 NTNI Data



Sculpting seen is due to the change of slope.



JSD factors in %:
(Range 105 – 160 GeV)

Had 1: **8.1 ± 0.2**
Had 2: **6.0 ± 0.1**
Had 3: **3.6 ± 0.0**
Had 4: **2.4 ± 0.0**
Lep 1: **2.8 ± 0.6**
Lep 2: **1.2 ± 0.6**
Lep 3: **0.5 ± 0.7**

Backup

Scaled JSD Factors in MC / %

ttyy, TI, Hadronic

Had 1: **0.9 ± 0.2**
Had 2: **0.3 ± 0.1**
Had 3: **0.1 ± 0.1**
Had 4: **0.1 ± 0.1**

ttyy, Relax tight and isolated criteria, All Hadronic

Had 1: **1.4 ± 0.1**
Had 2: **0.4 ± 0.1**
Had 3: **0.2 ± 0.0**
Had 4: **0.1 ± 0.0**

ttyy, TI, Leptonic

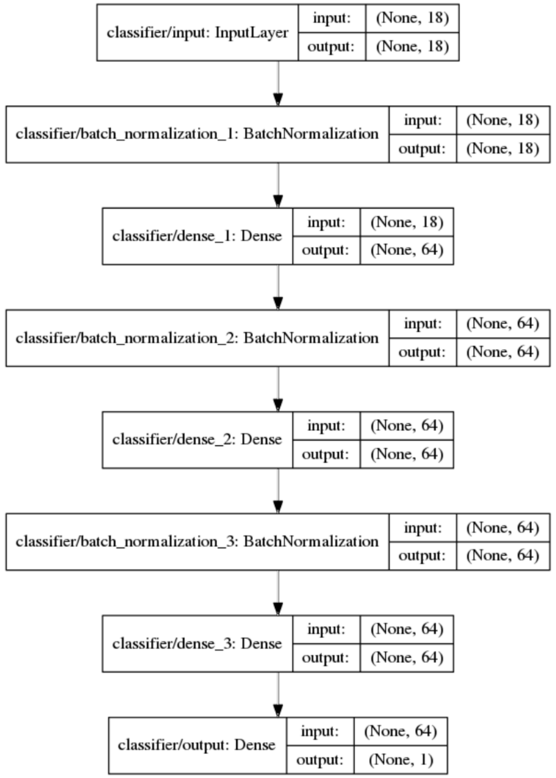
Lep 1: **0.1 ± 0.1**
Lep 2: **0.1 ± 0.1**
Lep 3: **0.1 ± 0.1**

ttyy, Relax tight and isolated criteria, Leptonic

Lep 1: **0.1 ± 0.1**
Lep 2: **0.1 ± 0.1**
Lep 3: **0.1 ± 0.1**

No significant sculpting observed in ttyy MC.

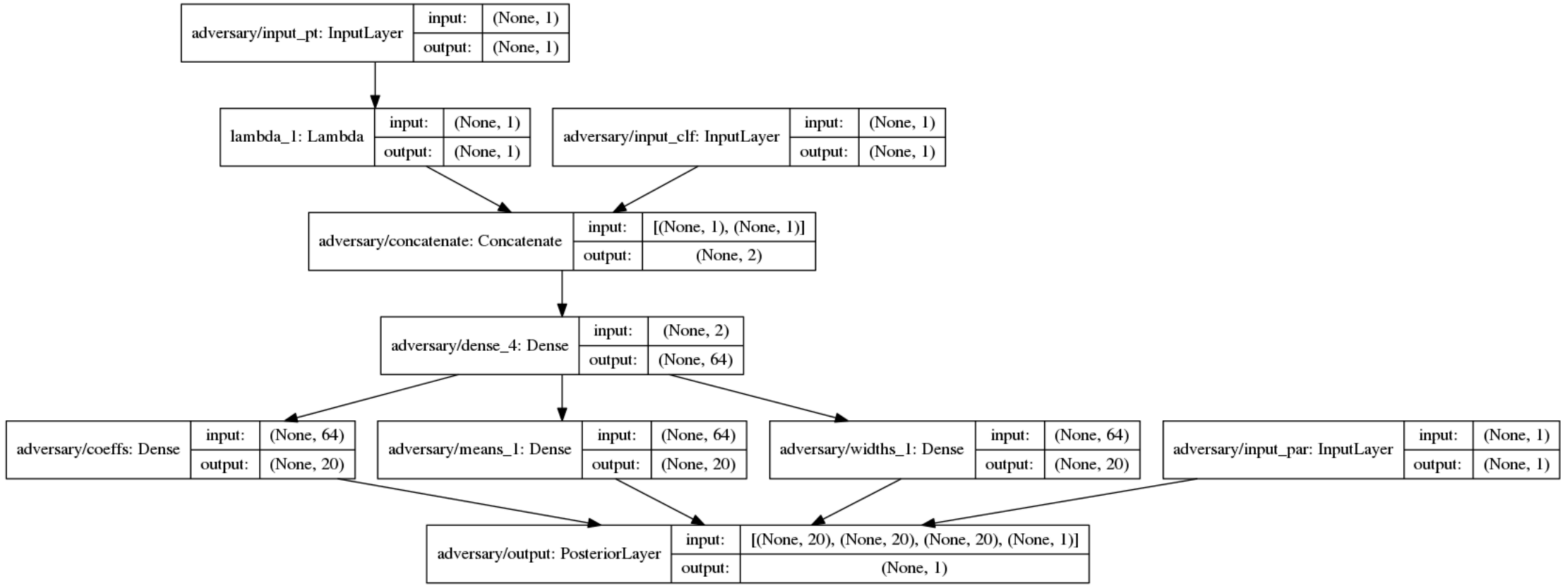
Backup, Classifier Model



- BatchNormalization layer: standardise the variables from the preceding layer (scales them so they have a mean of 0 and SD = 1.

- Dense layer: there exists a connection between every node in the previous layer and every node in the current layer.
- If the previous layer has M nodes, and the current layer has N nodes, the weight matrix has dimensions M x N, and every entry is trainable.
- If any node = 0 (no existing connection) -> a sparsely connected layer.

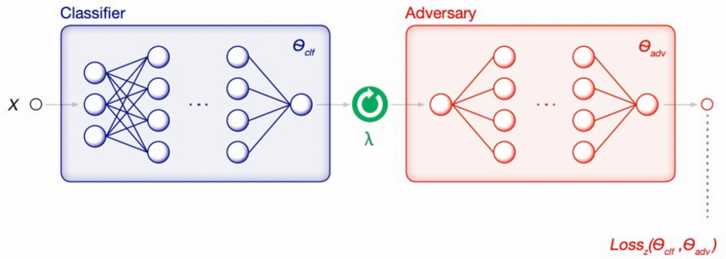
Backup, Adversary Model



Backup

Adversarial Neural Networks

Mathematical prospective



Cost function of classifier

$$J_{cls}(\vec{\theta}_{cls}) = \frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

Cost function of adversary

$$J_{adv}(\vec{\theta}_{adv}) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log p_{adv}(M_{\gamma\gamma}|\theta_{adv}, J_{cls}(\vec{\theta}_{cls}), a)]$$

where *m* is the number of iterations for finding the minima of the cost function, *Θ* the weights/parameters, which are updated after each iteration, *h(Θ)* the hypothesis function, y^i the current calculating of the function at iteration *i* and *a* represents any auxiliary inputs to the adversary

Balance between them to be achieved: $min_{\theta_{cls}} max_{\theta_{adv}} J_{ANN} = J_{cls}(\theta_{cls}) - \lambda J_{adv}(\theta_{adv}\theta_{cls})$

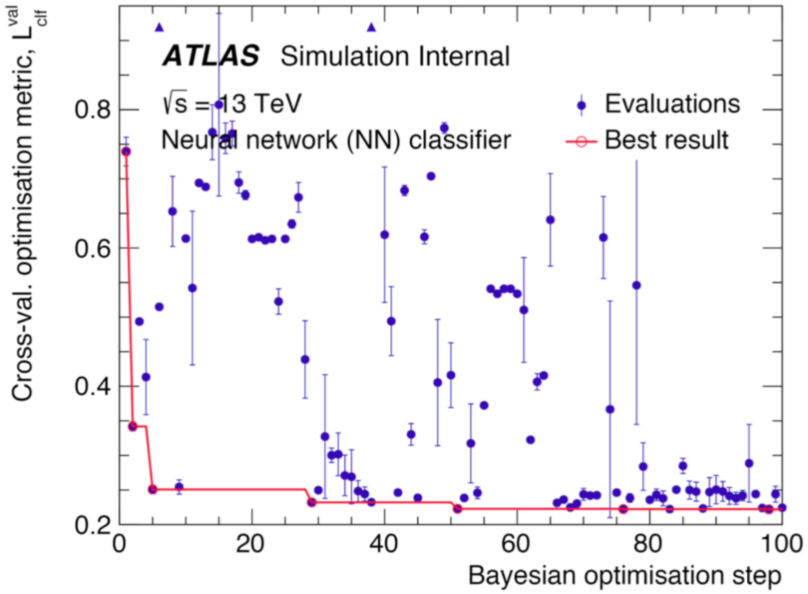
Backup

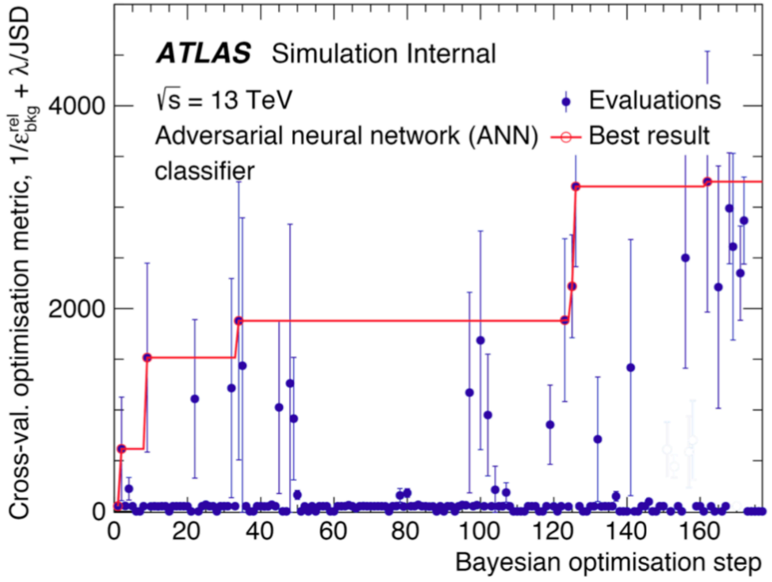
Spearmint hyperparameter optimization

Parameter	Range	Scale	Chosen value CN
Learning rate	$[10^{-5}, 10^{-1}]$	log	10^{-3}
Learning rate decay	$[10^{-6}, 10^{-2}]$	log	10^{-6}
Hidden layers	[1 , 6]	linear	4
Nodes per hidden layer	[2, 512]	log 2	512
Dropout regularization	[0, 0.5]	linear	0.3
Hidden layer activation function	{RELU, tanh}	choice	RELU

Backup

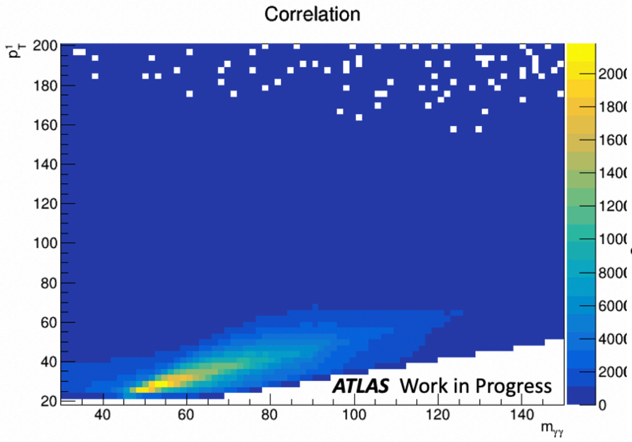
Spearmint hyperparameter optimisation

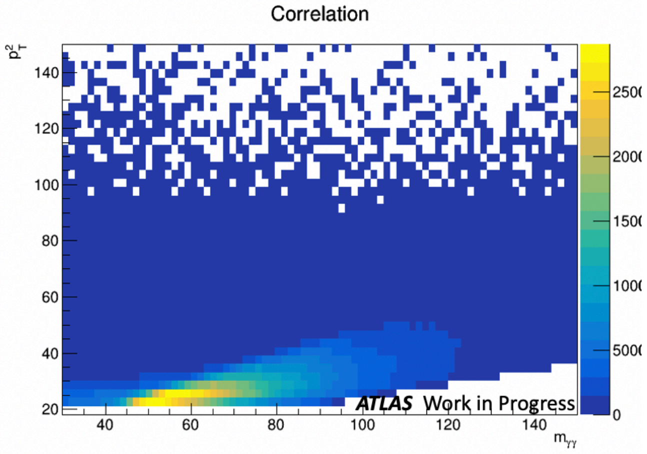




Backup

data





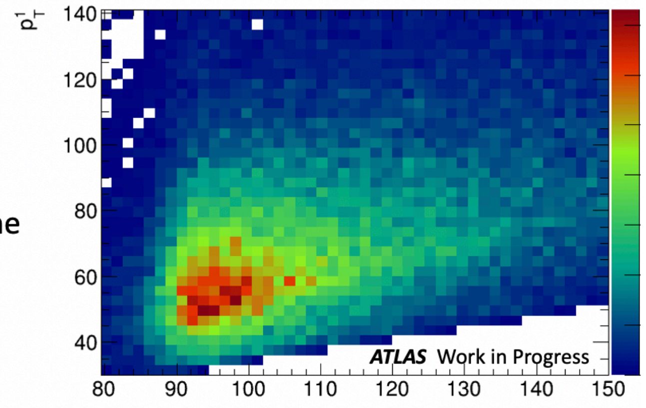
So far: correlations

- Definite positive correlation of the transverse momentum of both photons with $m_{\gamma\gamma}$

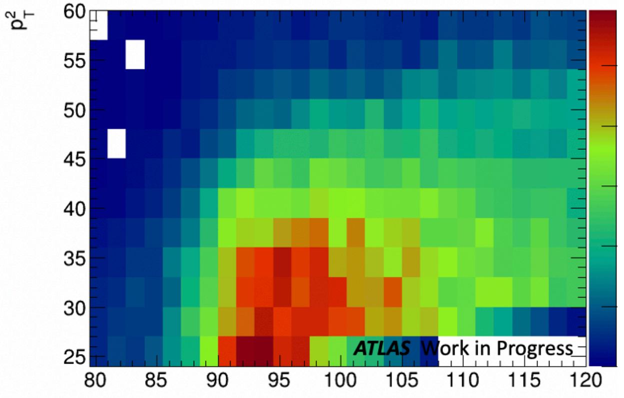
Reminder:

- Correlations of classifier input variables with $m_{\gamma\gamma}$ lead to background sculpting and we'll train the second NN to avoid tis.

ttγγ



Correlation factor: 0.271



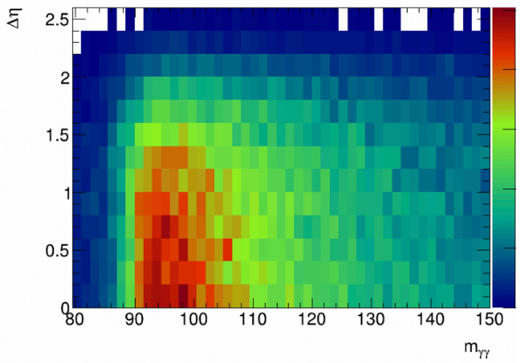
Correlation factor: 0.358

Backup

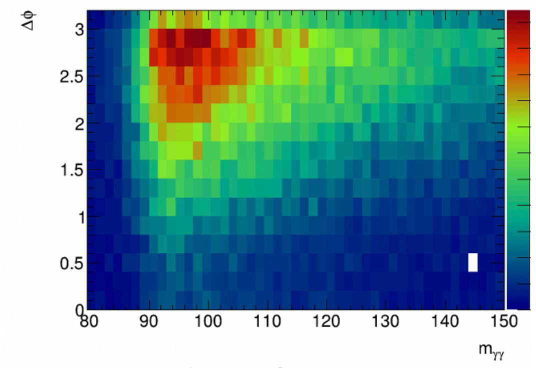
ttγγ

So far: correlations

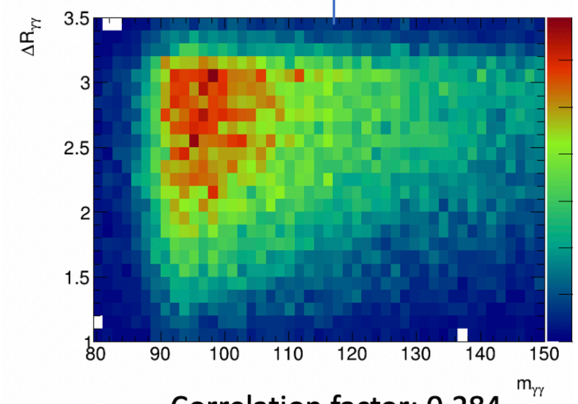
$M_{\gamma\gamma}^2 = 2\, E_{\gamma 1}\, E_{\gamma 2}(1 - \cos\vartheta)$



Correlation factor: 0.097



Correlation factor: 0.221



Correlation factor: 0.284

η 1	η 2	ϕ_1	ϕ_2	p_1^T	p_2^T	ΔR	Δη	Δϕ
-0.006	-0.001	0.005	0.002	0.271	0.358	0.284	0.097	0.221

Positive correlation: relationship between two variables in which both of them change in the same way.