A Boosted Decision Tree approach to the search for dark photons from Higgs boson decays in events with a photon and missing transverse momentum with the ATLAS detector

Tesi di Laurea Triennale in Fisica di: Giulia Maineri

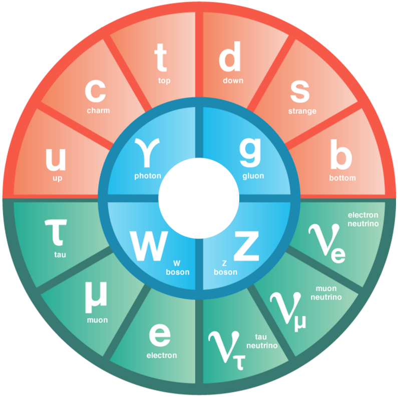
Relatori: Prof. Marcello Fanti, Dott.ssa Silvia Resconi, Dott.ssa Federica Piazza

UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

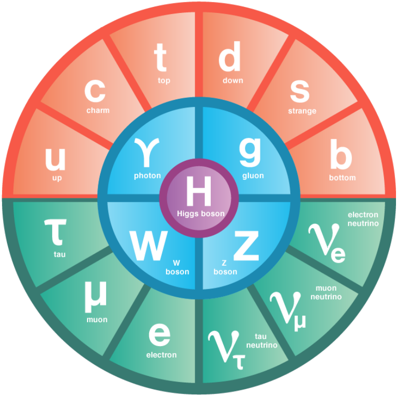
Dark Photon

Higgs boson discovery at LHC in 2012 completed the Standard Model (SM).



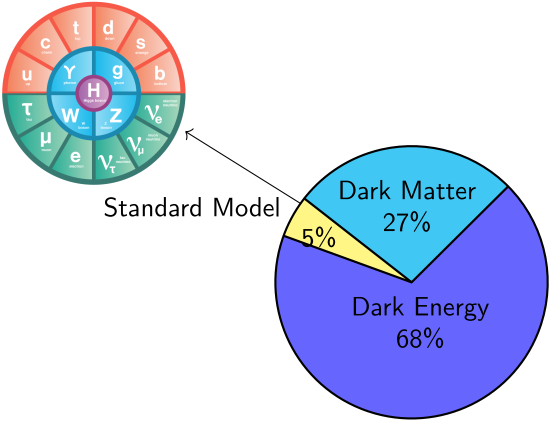
Dark Photon

Higgs boson discovery at LHC in 2012 completed the Standard Model (SM).



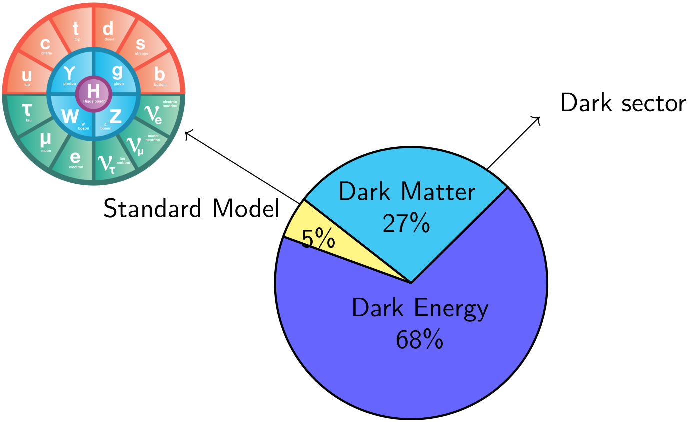
Dark Photon

But Standard Model accounts only for the $\simeq 5\%$ of the mass and energy content of the universe.



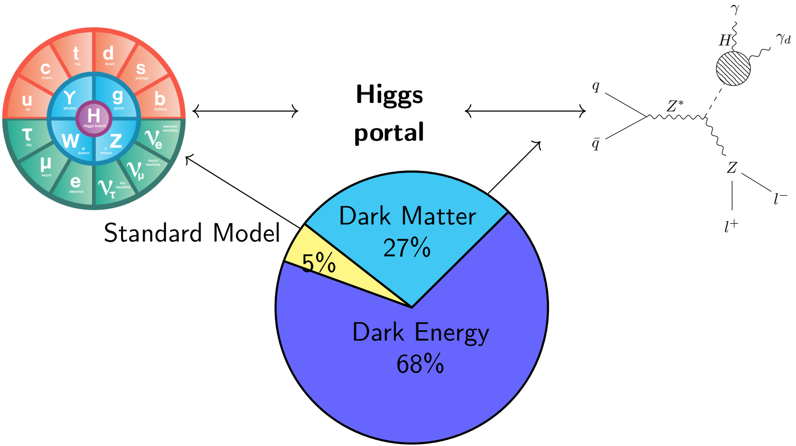
Dark Photon

But Standard Model accounts only for the $\simeq 5\%$ of the mass and energy content of the universe.

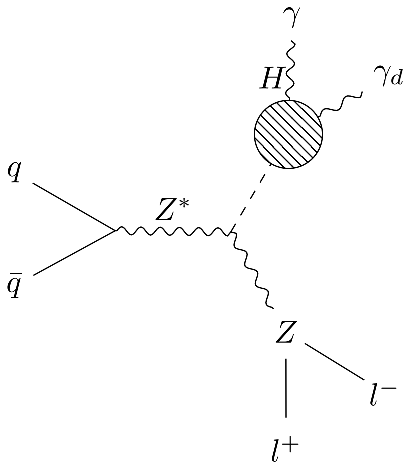


Dark Photon

But Standard Model accounts only for the $\simeq 5\%$ of the mass and energy content of the universe. $\Longrightarrow$ search for Dark Photons from Higgs boson decays



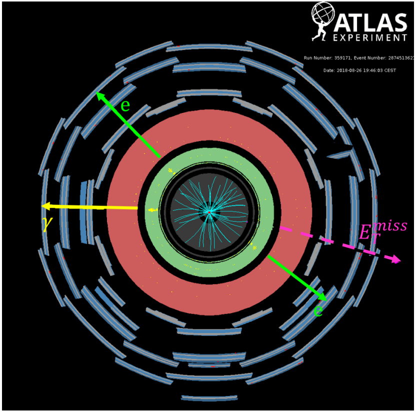
Signal Region

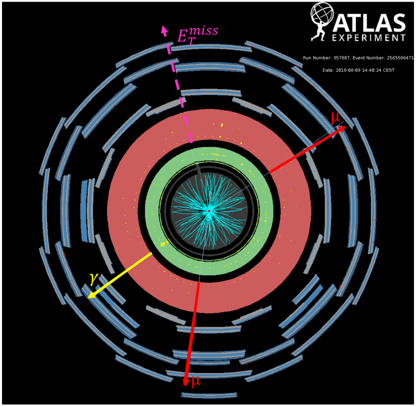


Signal: $Z(\to l^+l^-)H(\to \gamma\gamma_d), l \in \{e, \mu\}$

- 1 photon $N_\gamma = 1$
- 2 leptons, $N_e = 2$ or $N_\mu = 2$
- 60 GeV $\leq m_{ll} \leq$ 116 GeV
- missing transverse momentum $E_T^{miss} > 60$ GeV

Signal Region



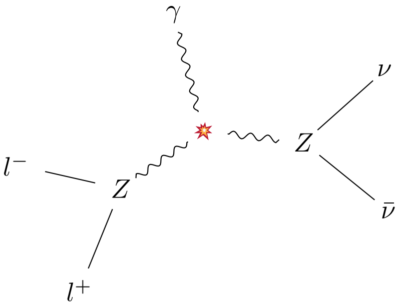


$$\vec{E}_T^{miss} = -\left[\sum_e \vec{p}_T^{(e)} + \sum_\mu \vec{p}_T^{(\mu)} + \sum_\gamma \vec{p}_T^{(\gamma)} + \sum_\tau \vec{p}_T^{(\tau)} + \sum_{jet} \vec{p}_T^{(jet)} + \sum_x \vec{p}_T^{(x)}\right]$$

Backgrounds

Backgrounds:

- irreducible: $VV\gamma$, $V \in \{Z, W\}$

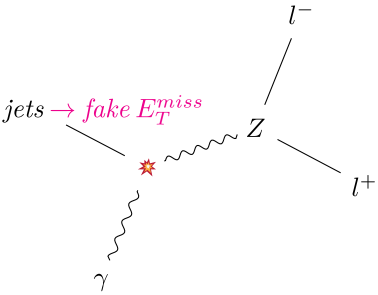


💥 represents the collision point or primary vertex

Backgrounds

Backgrounds:

- irreducible: $VV\gamma$, $V \in \{Z, W\}$
- reducible: fake E_T^{miss}($Z\gamma$+jets, Z+jets, etc.), $e \to \gamma$ (VV, VVV, $Vtll$, $t\bar{t}VV$), top backgrounds ($Wt\gamma$, t, $t\bar{t}$, ttV), Higgs ($ttH(\to Z\gamma)$, $VH(\to Z\gamma)$) and $W\gamma$

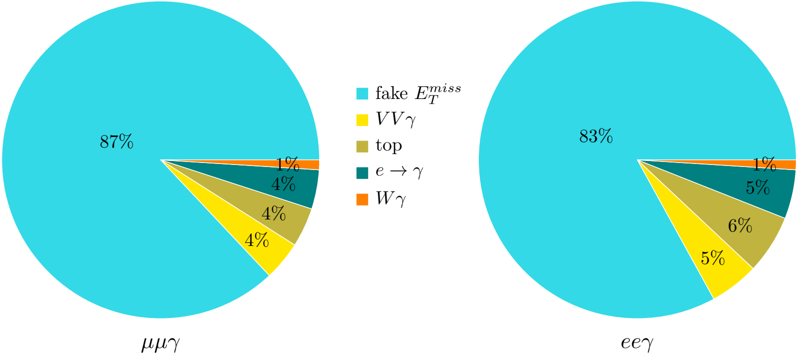


✸ represents the collision point or primary vertex

Backgrounds

Backgrounds:

- irreducible: $VV\gamma$, $V \in \{Z, W\}$
- reducible: fake E_T^{miss}($Z\gamma$+jets, Z+jets, etc.), $e \to \gamma$ (VV, VVV, $Vtll$, $t\bar{t}VV$), top backgrounds ($Wt\gamma$, t, $t\bar{t}$, ttV), Higgs ($ttH(\to Z\gamma)$, $VH(\to Z\gamma)$) and $W\gamma$



MultiVariate Analysis

Different variables distributions can be exploited to separate signal from backgrounds

$\Longrightarrow$ MultiVariate Analysis

$\Longrightarrow$ Boosted Decision Trees

truly is: \ accepted as:	Sig	Bkg
Sig	☺	Type-2 error
Bkg	Type-1 error	☺

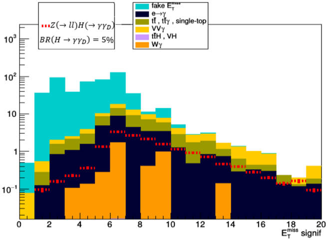
MultiVariate Analysis

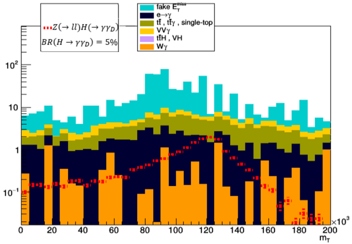
E_T^{miss} **significance** $S = \frac{|\vec{E}_T^{miss}|}{\sigma_{E_T^{miss}}}$

Transverse mass

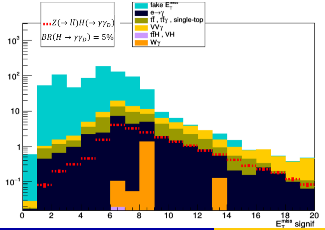
$$m_T = \sqrt{2 p_T^{\gamma} E_T^{miss} (1 - \cos(\Phi^{\gamma} - \Phi^{E_T^{miss}}))}$$

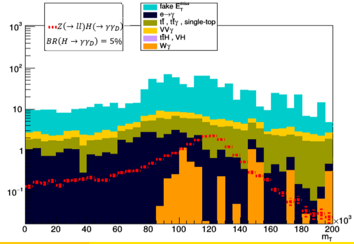
$ee\gamma$ channel





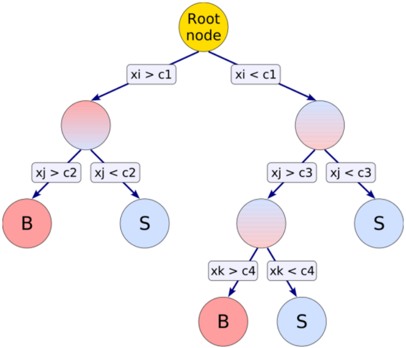
$\mu\mu\gamma$ channel





Decision Tree

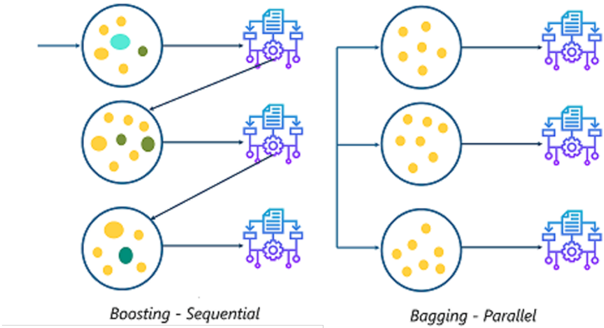
- is a binary tree structured **classifier**
- can distinguish data of two or more **different types**
- uses **one** discriminating variable at each node
- ends with **leaves** when a **stopping criterion** is fulfilled
- needs to be **trained** on a known dataset
- needs a known dataset to **test** the performance
- suffers from **instability** due to statistical fluctuations in the training sample



Boosted Decision Tree

The sum of weak learners results in a stronger and more stable learner $\Longrightarrow$ Boosting procedure

- generates a **forest** from one single tree
- subsequently modifies the events **weights** in the sample
- can be done with different **algorithms** (AdaBoost, Gradient Boost + Bagging)



BDTs have been implemented in TMVA environment

BDT optimization

Optimized parameters:

- BoostType
- Bagging
- Learning rate
- Number of folds
- Number of trees
- Separation Index
- Max Depth
- Min Node Size
- Number of cut values

Input features:

9 kinematic variables

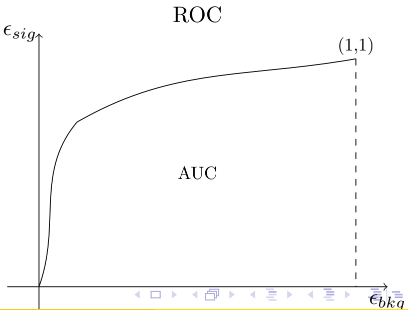
Adopted Figures of Merit:

- Approximate Median Significance

$$AMS = \sqrt{2\left[(s+b)\log\left(1+\frac{s}{b}\right)-s\right]}$$

s=signal yield, b=background yield.

- Receiver Operating Characteristic



BDT optimization

Optimized parameters:

- BoostType
- Bagging
- Learning rate
- Number of folds
- Number of trees
- Separation Index
- Max Depth
- Min Node Size
- Number of cut values

Input features:

9 kinematic variables

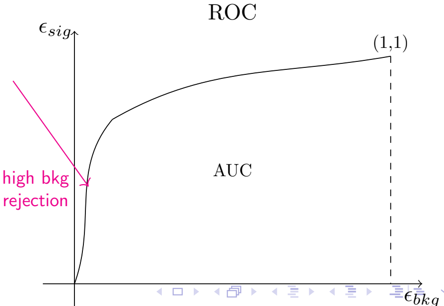
Adopted Figures of Merit:

- Approximate Median Significance

$$AMS = \sqrt{2\left[(s+b)\log\left(1+\frac{s}{b}\right)-s\right]}$$

s=signal yield, b=background yield.

- Receiver Operating Characteristic

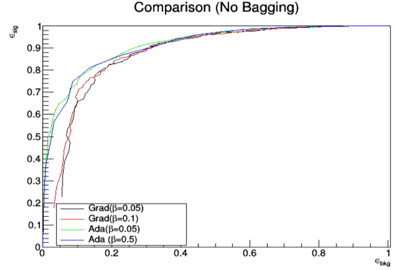


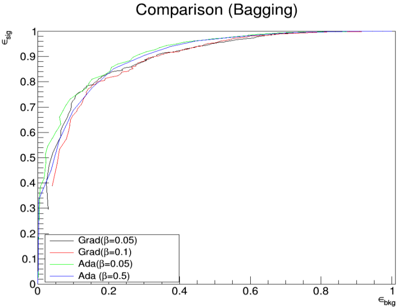
BDT optimization

Optimized parameters:

- BoostType : AdaBoost
- Bagging : No
- Learning rate
- Number of folds
- Number of trees
- Separation Index
- Max Depth
- Min Node Size
- Number of cut values

⚠ These plots concern test sample



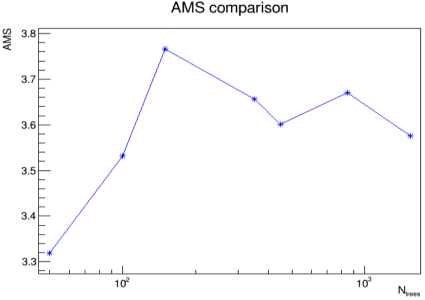


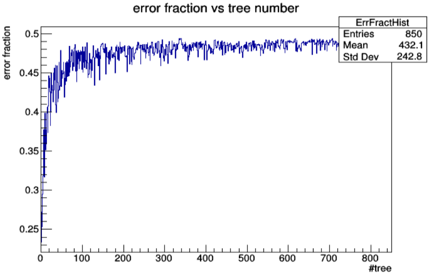
BDT optimization

Optimized parameters:

- BoostType : AdaBoost
- Bagging : No
- Learning rate : $\beta = 0.2$
- Number of folds : $Nfolds = 2$
- Number of trees : $Ntrees = 150$
- Separation Index
- Max Depth
- Min Node Size
- Number of cut values

⚠ These plots concern test sample





BDT optimization

Optimized parameters:

- BoostType : AdaBoost
- Bagging : No
- Learning rate : $\beta = 0.2$
- Number of folds : $Nfolds = 2$
- Number of trees : $Ntrees = 150$
- Separation Index : Gini Index
- Max Depth : $MaxDepht = 3$
- Min Node Size : $MinNodeSize = 5\%$
- Number of cut values : $NCuts = 20$

$$G_{Gini} = p(1-p)$$

$$\textit{where purity:} \quad p = \frac{s}{s+b}$$

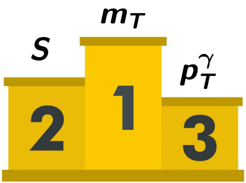
Gain definition	*AMS*
Gini Index	3.76654
Cross Entropy	3.63069
Misclassification Error	3.61978

Variable Ranking

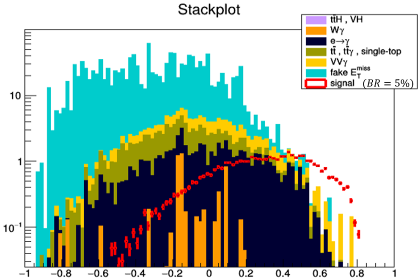
Variables have been deleted one at a time to see their contribution to AMS; the least discriminant has been removed at each step.

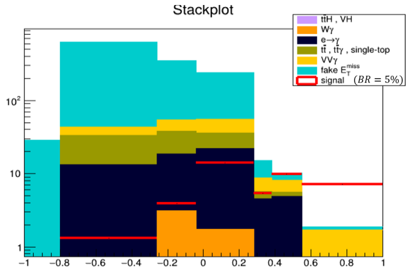
1. Transverse mass $m_T \equiv \sqrt{2p_\gamma^T E_T^{miss}(1 - \cos\Delta\Phi(\vec{p}_T^{\gamma}, \vec{E}_T^{miss}))}$
2. E_T^{miss} significance $S \equiv \frac{E_T^{miss}}{\sigma_{E_T^{miss}}}$
3. Photon transverse momentum p_T^{γ}
4. $p_T^{balance} \equiv \frac{p_T^{\gamma + E_T^{miss}}}{p_T^{ll}}$

variable	*AMS*	Δ*AMS*
all	3.36141	
m_T	2.26116	-1.10025
S	2.26925	-1.09216
p_T^{γ}	2.69211	-0.66930
$p_T^{balance}$	3.15523	-0.20618



BDT score distributions - Monte Carlo

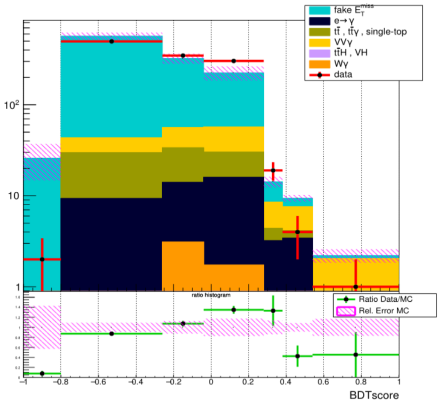




The MC simulations provide weighted events samples and suffer by large statistical fluctuations $\Longrightarrow$ rebinning to 7 bins to optimize the signal sensitivity

BDT score distributions - Data

Log scale

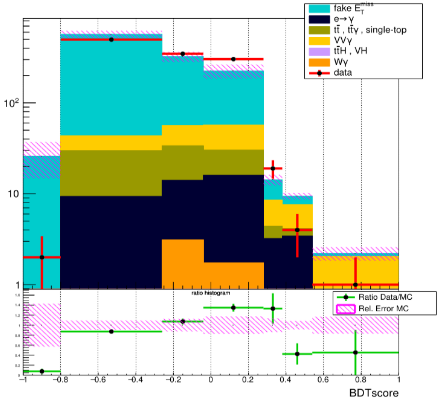


only statistical uncertainties included

- data from Run 2 at ATLAS ($\mathcal{L} = 139\,\mathrm{fb}^{-1}$)

BDT score distributions - Data

Log scale

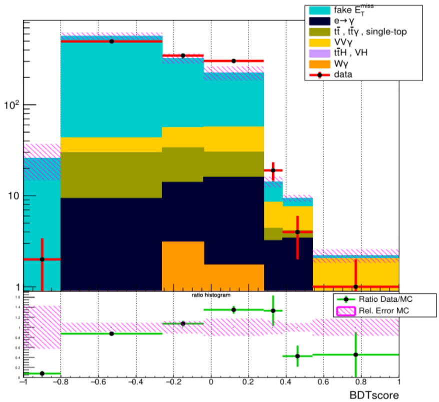


only statistical uncertainties included

- data from Run 2 at ATLAS ($\mathcal{L} = 139\,\text{fb}^{-1}$)
- template fit
  $\Longrightarrow$ upper limit for $BR(H \to \gamma\gamma_d)$

BDT score distributions - Data

Log scale

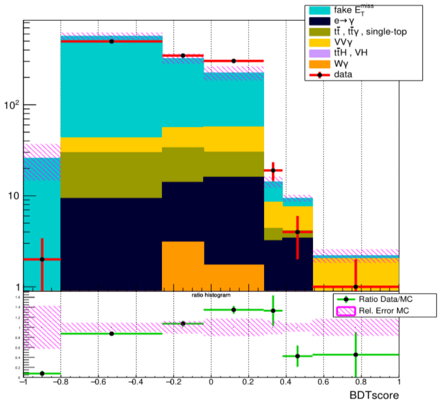


only statistical uncertainties included

- data from Run 2 at ATLAS ($\mathcal{L} = 139\,\mathrm{fb}^{-1}$)
- template fit $\implies$ upper limit for $BR(H \to \gamma\gamma_d)$
- HistFitter, a tool for statistical analysis used at ATLAS

BDT score distributions - Data

Log scale



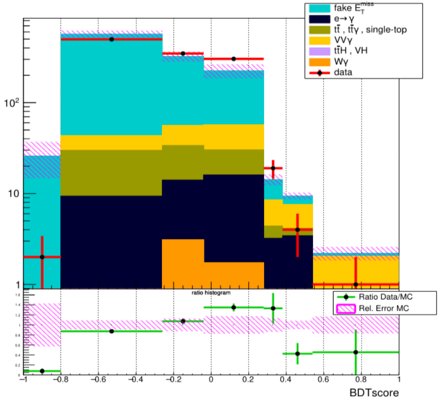
only statistical uncertainties included

- data from Run 2 at ATLAS ($\mathcal{L} = 139\,\text{fb}^{-1}$)
- template fit
  $\Longrightarrow$ upper limit for $BR(H \to \gamma\gamma_d)$
- HistFitter, a tool for statistical analysis used at ATLAS

	$BR(H \to \gamma\gamma_d)$ this analysis
expected	2.25%
observed	1.79%

BDT score distributions - Data

Log scale



only statistical uncertainties included

- data from Run 2 at ATLAS ($\mathcal{L} = 139\,\text{fb}^{-1}$)
- template fit $\Longrightarrow$ upper limit for $BR(H \to \gamma\gamma_d)$
- HistFitter, a tool for statistical analysis used at ATLAS

	$BR(H \to \gamma\gamma_d)$ this analysis	official[1]
expected	2.25%	2.41%
observed	1.79%	1.87%

[1] https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2022-064/

Conclusions

- An **exotic signal** with final state $ll\gamma + E_T^{miss}$ has been studied.

Conclusions

- An **exotic signal** with final state $ll\gamma + E_T^{miss}$ has been studied.
- The main issue is to increase the sensitivity in signal and **background** distinction.

Conclusions

- An **exotic signal** with final state $ll\gamma + E_T^{miss}$ has been studied.
- The main issue is to increase the sensitivity in signal and **background** distinction.
- A MultiVariate Analysis has been performed: **Boosted Decision Trees**.

Conclusions

- An **exotic signal** with final state $ll\gamma + E_T^{miss}$ has been studied.
- The main issue is to increase the sensitivity in signal and **background** distinction.
- A MultiVariate Analysis has been performed: **Boosted Decision Trees**.
- Settings and parameters of the BDT algorithms have been properly **optimized**; the most discriminant input variables have been found.

Conclusions

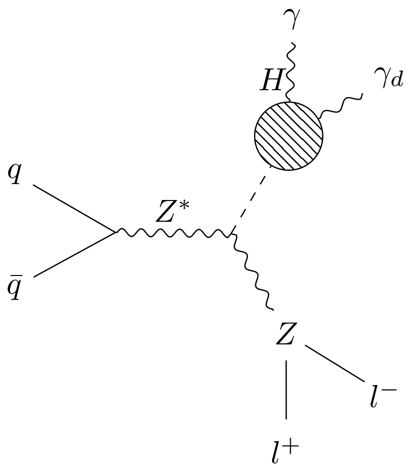
- An **exotic signal** with final state $ll\gamma + E_T^{miss}$ has been studied.
- The main issue is to increase the sensitivity in signal and **background** distinction.
- A MultiVariate Analysis has been performed: **Boosted Decision Trees**.
- Settings and parameters of the BDT algorithms have been properly **optimized**; the most discriminant input variables have been found.
- A fit to the data in the Signal Region has been performed in order to get the **exclusion limit** on the branching ratio of the decay $H \to \gamma\gamma_d$.

Conclusions

- An **exotic signal** with final state $ll\gamma + E_T^{miss}$ has been studied.
- The main issue is to increase the sensitivity in signal and **background** distinction.
- A MultiVariate Analysis has been performed: **Boosted Decision Trees**.
- Settings and parameters of the BDT algorithms have been properly **optimized**; the most discriminant input variables have been found.
- A fit to the data in the Signal Region has been performed in order to get the **exclusion limit** on the branching ratio of the decay $H \rightarrow \gamma\gamma_d$.
- Results are consistent with the ones obtained in the official analysis.

Backup

Signal Region

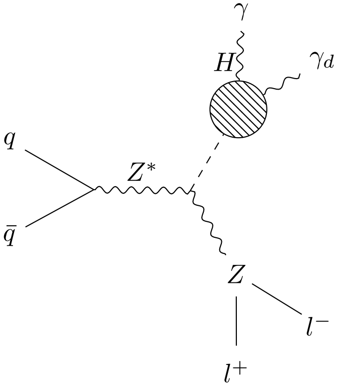


Signal: $Z(\to l^+l^-)H(\to \gamma\gamma_d), l \in \{e, \mu\}$

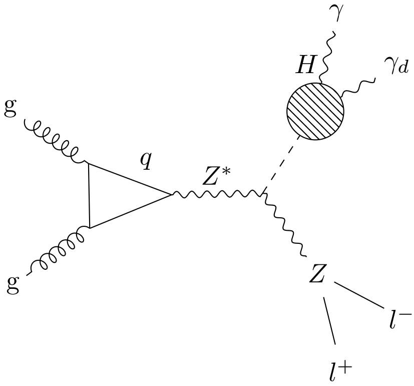
- 1 photon $N_\gamma = 1$ with $p_T^\gamma > 25$ GeV
- 2 leptons, $N_e = 2$ or $N_\mu = 2$
  - one leading lepton with $p_T^{l_1} > 27$ GeV
  - one subleading lepton with $p_T^{l_2} > 20$ GeV
- 60 GeV $\leq m_{ll} \leq$ 116 GeV
- $m_{ll\gamma} > 100$ GeV
- missing transverse momentum $E_T^{miss} > 60$ GeV
- $\Delta\Phi(\vec{E}_T^{miss}, \vec{p}_T^{\gamma ll}) > 2.4$
- $N_{jets} \leq 2$
- $N_{bjets} = 0$

Signal production

ZH production via quark

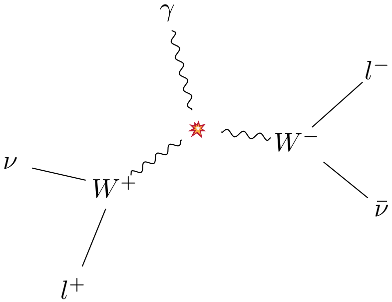


ZH production via gluons fusion



Backgrounds

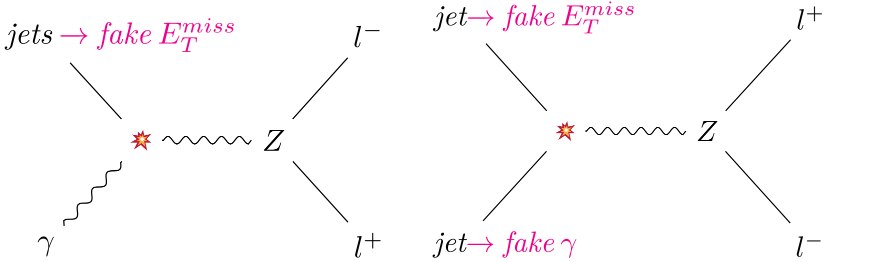
- irreducible: $VV\gamma$, $V \in \{Z, W\}$



✸ represents the collision point or primary vertex

Backgrounds

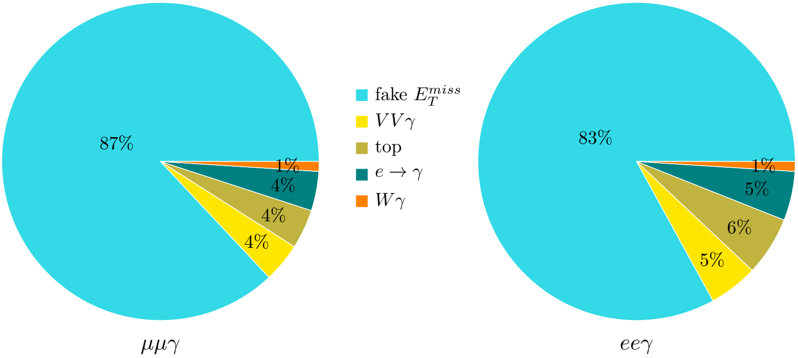
- irreducible: $VV\gamma$, $V \in \{Z, W\}$
- reducible: fake E_T^{miss}($Z\gamma$+jets, Z+jets, etc.), $e \to \gamma$ (VV, VVV, $Vtll$, $t\bar{t}VV$), top backgrounds ($Wt\gamma$, t, $t\bar{t}$, ttV), Higgs ($ttH(\to Z\gamma)$, $VH(\to Z\gamma)$) and $W\gamma$



💥 represents the collision point or primary vertex

Backgrounds

- irreducible: $VV\gamma$, $V \in \{Z, W\}$
- reducible: fake E_T^{miss}($Z\gamma$+jets, Z+jets, etc.), $e \to \gamma$ (VV, VVV, $Vtll$, $t\bar{t}VV$), top backgrounds ($Wt\gamma$, t, $t\bar{t}$, ttV), Higgs ($ttH(\to Z\gamma)$, $VH(\to Z\gamma)$) and $W\gamma$

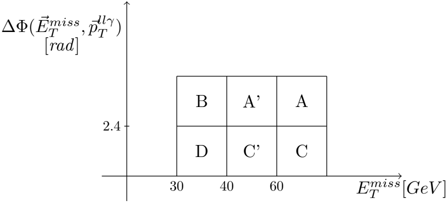


Backgrounds

- $VV\gamma$, $V \in \{Z, W\}$: shape from MC and normalization data-driven estimated
- fake E_T^{miss}: shape from MC and normalization data-driven estimated $\Longrightarrow$ ABCD method
- $e \to \gamma$: pure data-driven estimates $\Longrightarrow$ $f_{e \to \gamma}$ applied to $eee/\mu\mu e$ CRs
- top backgrounds: MC + 20% uncertainty
- Higgs, $W\gamma$: pure MC

ABCD method

1. $\vec{E}_T^{miss}$ and $\Delta\Phi(\vec{E}_T^{miss}, \vec{p}_T^{ll\gamma})$ used to define 6 regions;
2. Signal will be mostly located in A; fake $\vec{E}_T^{miss}$ be located between A and B as between C and D;
3. Fake $\vec{E}_T^{miss}$ events can be estimated and used to rescale events in Signal Region



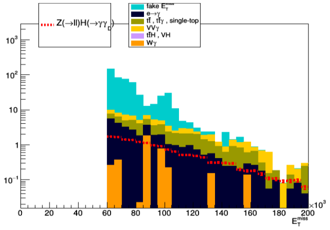
$$\frac{N_A}{N_B} = \frac{N_C}{N_D}$$

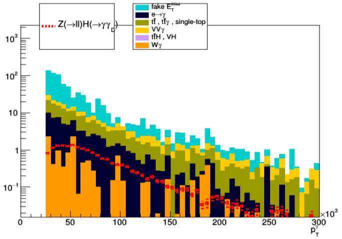
Variables distributions

Missing transverse momentum E_T^{miss}

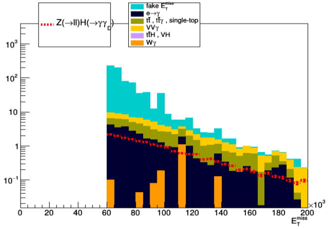
Photon transverse momentum p_T^{γ}

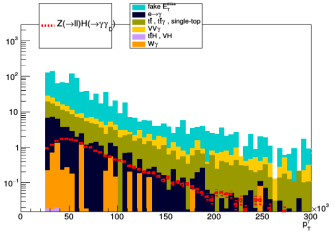
$ee\gamma$ channel





$\mu\mu\gamma$ channel

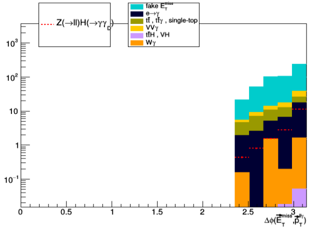


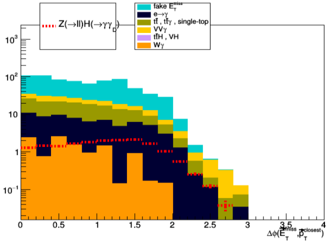


Variables distributions

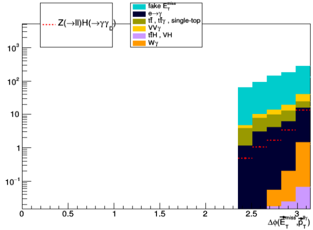
$\Delta\Phi(\vec{E}_T^{miss}, \vec{p}_T^{\gamma ll})$ $\Delta\Phi(\vec{E}_T^{miss}, \vec{p}_T^{closest})$

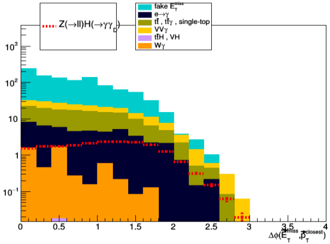
$ee\gamma$ channel





$\mu\mu\gamma$ channel





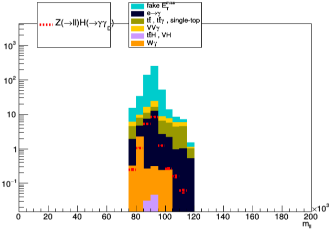
Variables distributions

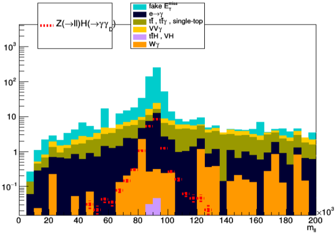
Invariant mass

$$m_{ll} = \sqrt{2p_T^{l_1} p_T^{l_2}[\cosh \Delta\eta - \cos \Delta\Phi]}$$

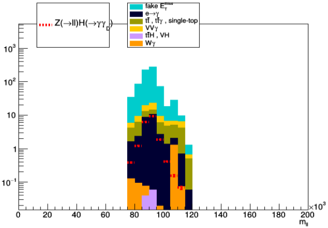
(No cuts)

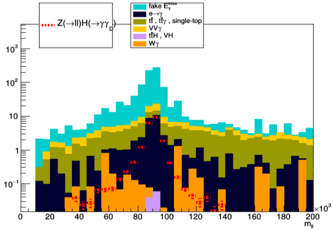
$ee\gamma$ channel





$\mu\mu\gamma$ channel



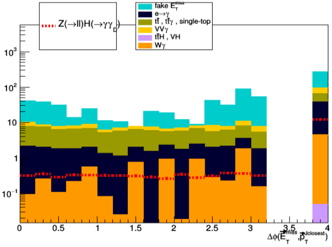


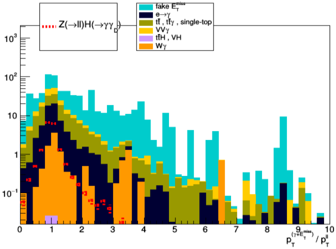
Variables distributions

$\Delta\Phi(\vec{E}_T^{miss}, \vec{p}_T^{closest\ jet})$

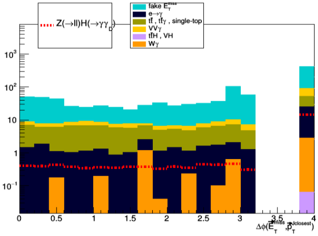
$p_T^{balance} = p_T^{\gamma + E_T^{miss}} / p_T^{ll}$

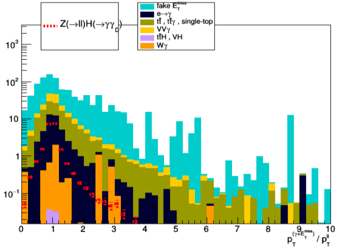
$ee\gamma$ channel





$\mu\mu\gamma$ channel





TMVA Analysis

BDTs have been implemented in TMVA (a Toolkit for MultiVariate Analysis) environment

1. Pre-Analysis
2. Training
    1. root node
    2. one single variable and cut
    3. stopping criterion
    4. leaf nodes
    5. classification according to purity $p = \frac{s}{s+b}$
3. Applying

⚠ Overtraining

truly is: \ accepted as:	Sig	Bkg
Sig	☺	Type-2 error
Bkg	Type-1 error	☺

AdaBoost and Gradient Boost

AdaBoost

- A boost weight α is assigned to each new tree:

$$\alpha = \frac{1-E}{E}$$

- The new added tree will focus on events mis-classified by the previous tree
- The boost weight can be given a power β, the learning rate
- The output of the classifier is a weighted sum of DTs votes:

$$y(\vec{x}) = \frac{1}{\sum_i^{N_{trees}} \ln(\alpha_i)} \sum_i^{N_{trees}} \ln(\alpha_i) h_i(\vec{x})$$

Gradient Boost

- The idea is to approximate the final output as a expansion series of DTs output:

$$F(\vec{x}, P) = \sum_{m=0}^{M} \gamma_m f(x; \alpha_m)$$

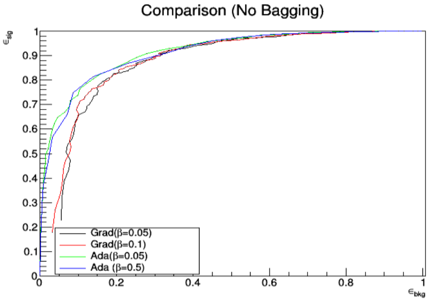
$$P \in \{\gamma_m; \alpha_m\}_0^M$$

- The weights P are chosen such that $F(\vec{x}, P)$ minimizes the loss function:

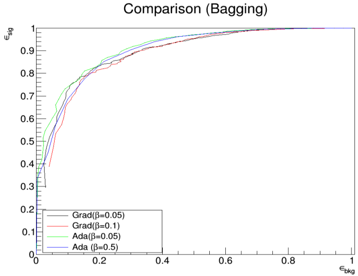
$$L(F, y) = \log\left(1 + e^{-2yF(\vec{x})}\right)$$

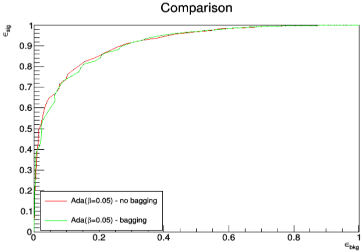
BDT optimization

BoostType



Bagging





BDT optimization

Number of folds

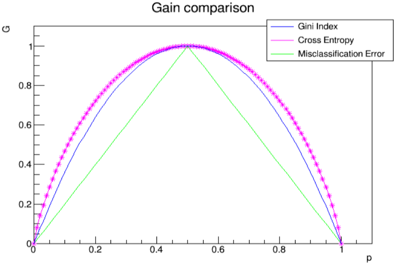
	AMS	
N_{trees}	$N_{folds} = 2$	$N_{folds} = 5$
150	3.76654	3.69253
850	3.66992	3.64944

Number of trees

N_{trees}	AMS
50	3.31854
100	3.53190
150	3.76654
350	3.65642
450	3.60072
850	3.66992
1550	3.57516

Separation Index

Gain definition	*AMS*
Gini Index	3.76654
Cross Entropy	3.63069
Misclassification Error	3.61978



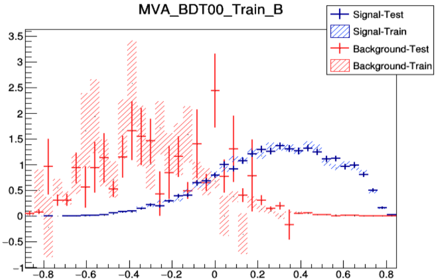
$$G_{Gini} = p(1-p)$$

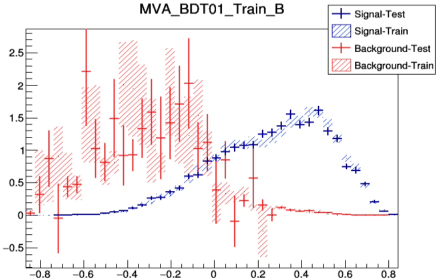
$$G_{entropy} = -[p \log p + (1-p)\log(1-p)]$$

$$G_{mis} = 1 - \max\{p, (1-p)\}$$

Check overtraining

BDT score distributions for the final configuration (AdaBoost, No Bagging, $\beta = 0.2$, 2 folds, 150 trees, Gini Index)

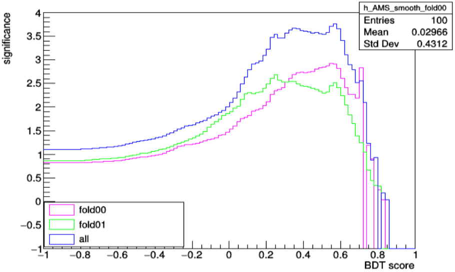




	χ^2	Ndf	$p-value$
sig	39.4	37	0.36
bkg	50.5	39	0.10

	χ^2	Ndf	$p-value$
sig	36.6	38	0.53
bkg	30.7	38	0.79

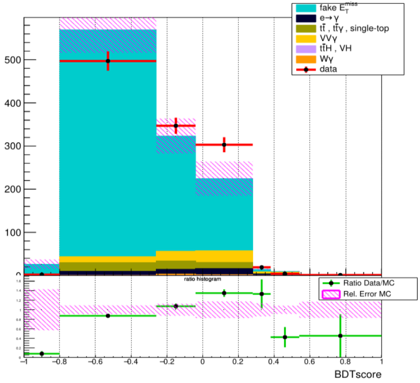
AMS distributions



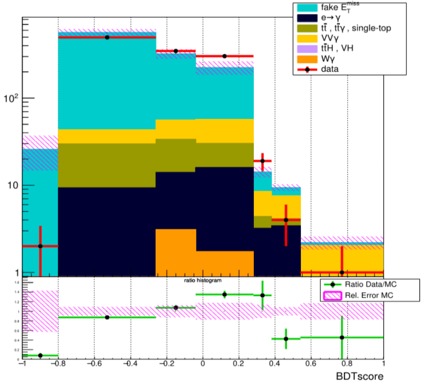
The curve is built moving from left to right, cutting step by step the BDT score and taking only the integrals of signal and background distributions at the right of the cut.

It starts from 1 at the left (all signal and all background included) and ends at 0 at the right (no signal included).

Linear scale



Log scale



only statistical uncertainties included