Signal extraction at inclusive $pp \to ZZ \to \ell\ell\nu\nu$ full Run2 analysis.

Zubov Dmitriy, Aleksandr Petukhov, Evgeny Soldatov

National Research Nuclear University «MEPhI»

MEPhI@ATLAS meeting

19.05.2023

Motivation and goals

Motivation:

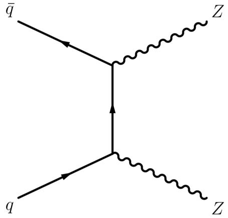
- No ZZZ or ZZγ couplings at tree level of SM $\rightarrow$ indirect search of the effects predicted by the theories beyond SM (BSM).
- $pp \rightarrow ZZ$ is an important background for on-shell (off-shell) Higgs boson production.

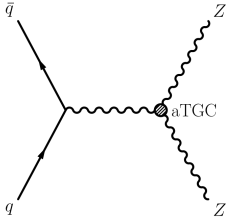
Goals of analysis:

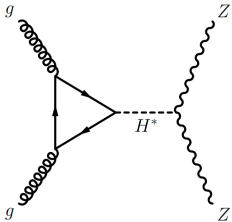
- Calculate integral and differential cross-sections in $p_T^{\ell\ell} = p_T^Z$, $\Delta\phi(\ell\ell)$, m_T^{ZZ}, N_{jets}, $m(j_1, j_2)$. The results are to be compared to the theoretical prediction with NNLO corrections in α_s ([1], [2]) and NLO corrections in α_{EWK} ([3]).
- Limits on the anomalous gauge-boson couplings (aTGCs) using vertex functions and EFT formalisms.

My goal:

- Measurement of the Z-boson pair production cross section using machine learning classifiers to improve estimation accuracy.







Integral cross section estimation approaches in the analysis.

1. Approach with «strict» cut-based selection of events:
   - Selection of the signal region with the best background suppression and with the best signal retention.
   - Development of a methodology for evaluating backgrounds in fit.
   - Evaluation of the integral cross section.

2. The approach with «loose» cut-based selection of events and application of a machine learning classifier:
   - Selecting a signal area with relatively good background suppression and great signal retention.
   - Training a machine learning classifier.
   - Development of a methodology for evaluating backgrounds in fit.
   - Evaluation of the integral cross section.

Inclusive $ZZ \to ll\nu\nu$

- Vertex with 2 tracks with $p_{\mathrm{T}} > 1$ GeV
- Two same flavour opposite-sign leptons (e+e- OR mu+mu-), leading p_{T}>30 GeV, subleading p_{T}>20 GeV
- Veto on any additional lepton with Loose ID and $p_{\mathrm{T}} > 7$ GeV
- $76 < M_{\ell\ell} < 106$ GeV
- $E_T^{miss} > 70$ GeV.

Signal	
ZZ ($\sim 0.7\%$)	production of two Z bosons and decay in $\ell\ell\nu\nu$
Background	
Zj ($\sim 85.6\%$)	Z boson and jet production, with the decay of the Z boson into a pair of charged leptons and with mismeasured $E_{\mathrm{T}}^{\mathrm{miss}}$
tt ($\sim 11.0\%$)	top-quark pair production and subsequent decay involving the final state $\ell\ell\nu\nu$ (nonresonant $\ell\ell\nu\nu$ production)
WZ ($\sim 1.0\%$)	the production of a pair of Z and W bosons, with the decay of the Z boson into a pair of charged leptons and the lepton decay of W when one missing ℓ mimics the signal topology
WW ($\sim 0.5\%$)	the production of the W pair with decay in $\ell\ell\nu\nu$ (nonresonant $\ell\ell\nu\nu$ production)
Wt ($\sim 0.9\%$)	W and top-quark production and decay to a final state containing $\ell\ell\nu\nu$ (nonresonant $\ell\ell\nu\nu$ production)
Other (4ℓ, $\ell\ell qq$, VVV, $Z(\tau\tau)$, $W +$ jets)	Background processes that contribute little to the total number of events and are evaluated via MC

Event selection optimization

• The optimization process looked for thresholds on the variables at which the maximum signal significance is achieved:

$$Z = \sqrt{2 \times [(S + B) \times \ln(1 + (S/B)) - S]}$$

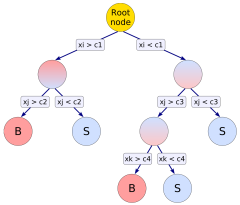
• The signal significance was considered as a function of several variables and the search for the optimal vector of optimized selections was carried out by enumerating all possible variants of the phase space constraint.

E_T^{miss}, ГэВ	—	>70
ΔR_{ll}	—	<1.8
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$	—	>2.3
N_{b-jets}	—	<1
E_T^{miss} significance	—	>10
Signal		
QCD ZZ	7600±30	1946±15
EWK ZZ	262±2	13.0±0.04
Total signal	**7860±30**	**1959±15**
Background		
Zj	963000±4000	180±20
WZ	11340±30	945±8
tt	123340±80	131±2
WW	5093±13	64.0±1.5
Wt	10250±40	41±3
VVV	41.8±0.3	7.88±0.10
Other	282±2	0.79±0.11
Total background	**1123000±4000**	**1368±20**
Z	**5.43±0.02**	**44.7±0.4**

Decision trees with gradient boosting (BDTG)

Decision tree

- A decision tree is a binary tree : a sequence of cuts paving the phase-space of the input variables
- Repeated yes/no decisions on each variables are taken for an event until a stop criterion is fulfilled
- Trained to maximize the purity of signal nodes (or the impurity of background nodes)



Advantages:

- Decision trees are independent of monotonous variable transformations
- Weak variables are ignored and do not deteriorate performance

Disadvantages:

- Decision trees are extremely sensitive to the training samples, therefore to overtraining
- Slightly different training samples can lead to radically different DT

Boosting

- Sequentially apply the DT algorithm to reweighted (boosted) versions of the training data
- Each model in the series trains upon its predecessor's mistakes, trying to correct them
- Works very well on non-optimal decision tree (small number of nodes)
- There are different boosting algorithms and in our work we use the gradient descent

Classifier training parameters

Hyperparameters:

- Number of trees;
- Max depth of the decision tree allowed
- Minimum percentage of training events required in a leaf node
- Number of grid points in variable range used in finding optimal cut in node splitting
- Shrinkage (Learning rate)

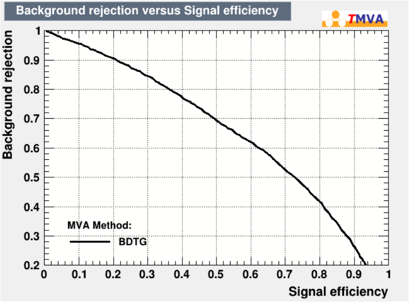
Separation into training and test sample:

- Random division in equal proportion

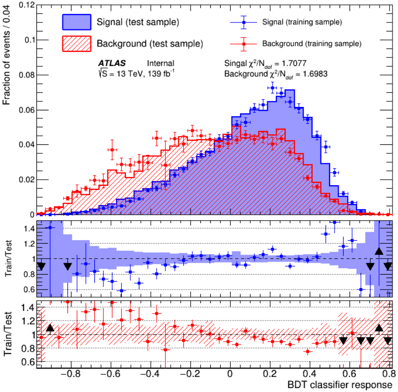
1. $E_{\mathrm{T}}^{\mathrm{miss}}$;
2. Object-based $E_{\mathrm{T}}^{\mathrm{miss}}$-significance;
3. $m_{\mathrm{T}}(ZZ)$;
4. $\Delta R(\ell\ell)$;
5. $p_{\mathrm{T}}^{\ell 1}$;
6. $p_{\mathrm{T}}^{\ell 2}$;
7. $p_{\mathrm{T}}^{Z} = p_{\mathrm{T}}^{\ell\ell}$
8. $\dfrac{p_{\mathrm{T}}^{Z}}{m_{\mathrm{T}}(ZZ)}$;
9. $H_{\mathrm{T}} = p_{\mathrm{T}}^{\ell 1} + p_{\mathrm{T}}^{\ell 2} + \sum_i p_{\mathrm{T}}(j_i)$;
10. $E_{\mathrm{T}}^{\mathrm{miss}}/H_{\mathrm{T}}$;
11. $\rho_Z = \dfrac{p_{\mathrm{T}}^{Z}}{p_{\mathrm{T}}^{\ell 1} + p_{\mathrm{T}}^{\ell 2}}$;
12. $\Delta\varphi(\vec{p}_{\mathrm{T}}^{\,\ell 1}, \vec{p}_{\mathrm{T}}^{\,\ell 2})$;
13. $\Delta\varphi(\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}, \vec{p}_{\mathrm{T}}^{\,\ell\ell})$;
14. $\dfrac{p_{\mathrm{T}}^{\ell 1}}{p_{\mathrm{T}}^{\ell 2}}$;
15. $\Delta\eta(\ell\ell)$;
16. $m(\ell\ell)$;
17. y_Z;
18. N_{jets};
19. $\dfrac{E_{\mathrm{T}}^{\mathrm{miss}} + \sum_i p_{\mathrm{T}}(j_i) - p_{\mathrm{T}}^{Z}}{p_{\mathrm{T}}^{Z}}$;
20. V_{T}

Training results of the classifier with «strict» event preselection

Variable	«Strict» cut
E_T^{miss}, GeV	>70
E_T^{miss} significance	>10
ΔR_{ll}	<1.8
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$	>2.3
N_{b-jets}	<1



- No good separation of signal and background
- The classifier is overtrained



Selecting the signal region «Loose»

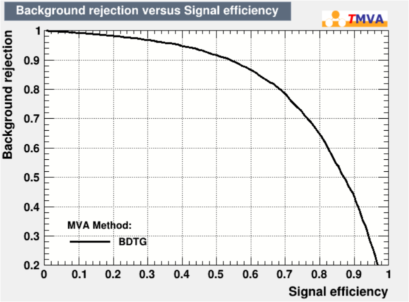
- To isolate the signal region with relaxed thresholds, optimization of the thresholds on the variables was performed.

- This optimization looked for a vector of thresholds on the variables corresponding to the maximum signal significance provided that the number of signal events >4500.

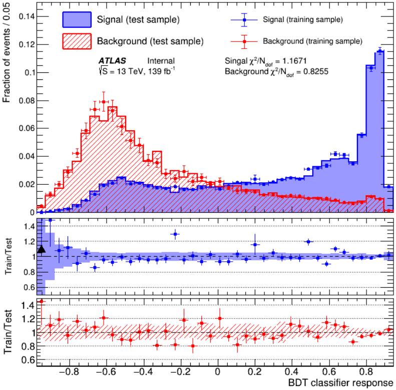
Variable	Loose cut	Strict cut
E_T^{miss}, GeV	>70	>70
E_T^{miss} significance	>7	>10
ΔR_{ll}	<2.2	<1.8
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$	>1.3	>2.3
Число b-струй	<1	<1
Signal		
QCD ZZ	4410±20	1946 ±15
EWK ZZ	57.8 ±0.9	13.0 ±0.4
Total signal	**4470±20**	**1959±15**
Background		
Zj	12200±300	180±20
WZ	3116 ±15	945 ±8
tt	2829 ±11	131 ±2
WW	1352 ±7	64.0 ±1.5
Wt	729 ±10	41 ±3
VVV	1771 ±0.17	7.88±0.10
Other	4.46 ±0.26	0.79±0.11
Total background	**20400±300**	**137±20**

Classifier training result with loose event preselection

Variable	Cut
E_T^{miss}, GeV	>70
E_T^{miss} significance	>7
ΔR_{ll}	<2.2
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$	>1.3
N_{b-jets}	<1



- Signal and background are well separated
- Maximum Signal significance 46.8 ± 0.4
- In the classifier the internal settings (hyperparameters) and the set of variables were optimized.



Fit description

The integral cross-section and the backgrounds are estimated in the **maximum likelihood fit**, by maximizing the following function in terms of μ and η:

$$\mathcal{L}(\mu, \theta) = \prod_{r}^{\text{regions}} \left[\prod_{i}^{\text{bins} \in r} \text{Pois}(N_i^{\text{data}} | \mu \nu_i^s \eta^s(\theta) + \nu_i^b \eta^b(\theta)) \right] \cdot \prod_{i}^{\text{nuis. par.}} \mathcal{L}(\theta_i),$$

N (ν) — observed (predicted) event yields
μ — signal normalization coefficient (signal strength), $\mu = \nu^s / N^s$.
θ — background normalization coefficients and systematic uncertainties nuisance parameters.
η — parameterize effect of θ on the predicted yields.

The fit has 4 regions (including signal region) and 4 normalization coefficients (including signal strength).

Right now fit to the observed data is performed **only in control regions** and Asimov dataset is used instead of the observed data in signal region.

The following statistic is used to compute the discovery significance and the uncertainties of $\hat{\mu}$:

$$q(\mu, \hat{\mu}, \hat{\theta}) = -2 \ln \lambda(\mu, \hat{\mu}, \hat{\theta}) = -2 \ln \frac{\mathcal{L}(\mu, \hat{\hat{\theta}}(\mu))}{\mathcal{L}(\hat{\mu}, \hat{\theta})}, \; Z_{\text{disc}}^{\text{exp.}} = \sqrt{q(\mu = 1)_A}. \quad (1)$$

Definition of control and signal regions.

Loose variant of phase space:

Variable	SR	WZ (3ℓ)	NR ($e\mu$)	Zj
$E_T^{miss.}$, GeV	>70		>70	
ΔR_{ll}	<2.2		<2.2	
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$, rad	>1.3		>1.3	
$E_T^{miss.}$ significance	>7			[4;7]
m_T^W, GeV		>60		

Fit in the signal region was performed using the variable *BDTscore*, in the control region using the variable $p_T^{\mathbf{Z}}$

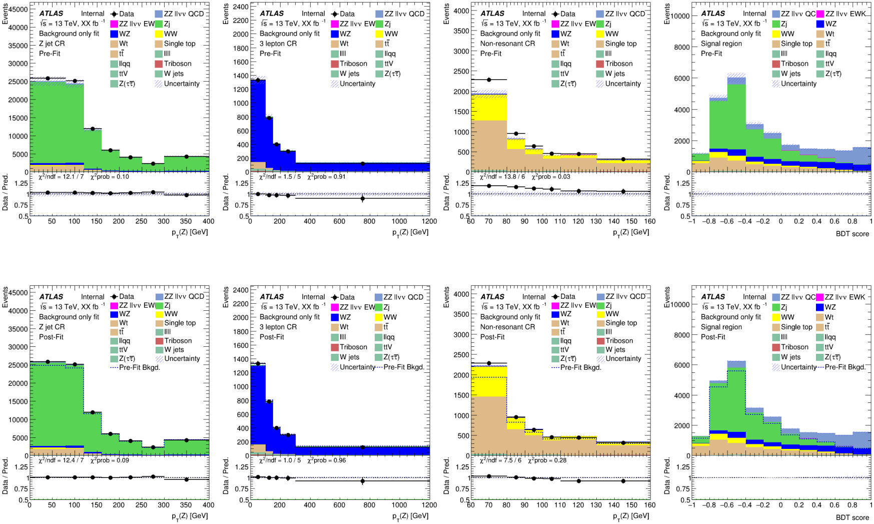
Strict variant of phase space::

Variable	SR	WZ (3ℓ)	NR ($e\mu$)	Zj
$E_T^{miss.}$, ГэВ	>70		>70	
ΔR_{ll}	<1.8		<1.8	
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$, rad	>2.3		>2.3	
$E_T^{miss.}$ significance	>10			[4;9]
m_T^W, GeV		>60		

Fit in the signal and control regions was performed using the variable $p_T^{\mathbf{Z}}$

- **SR** - region of the phase space in which the fraction of signal events is maximal.
- **WZ(3l)** - region of the phase space in which the fraction of events of the WZ process is maximal.
- **Non-resonant** - region of the phase space in which the fraction of events of the $\ell^+\ell^-$ nonresonant production processes is maximal.
- **Zj** - region of the phase space in which the fraction of events of the process Zj is maximal.

Fit. Before and after distributions for the loose version of the phase space.



Fit. Results.

The resulting signal strength value is applied in calculation of the observed cross section: $\sigma_{meas.} = \mu_{ZZ} \cdot \sigma_{SM}$

	«Strict» fit with $p^T(Z)$	«Loose» fit with $p^T(Z)$	«Loose» fit with BDT score
μ_{ZZ}	$1.00^{+0.04}_{-0.04}$(stat)$^{+0.06}_{-0.05}$(syst)	$1.00^{+0.04}_{-0.04}$(stat)$^{+0.06}_{-0.05}$(syst)	$1.00^{+0.03}_{-0.03}$(stat)$^{+0.06}_{-0.05}$(syst)
μ_{Zj}	$1.31^{+0.03}_{-0.03}$(stat)$^{+0.07}_{-0.07}$(syst)	$1.13^{+0.01}_{-0.01}$(stat)$^{+0.06}_{-0.06}$(syst)	$1.13^{+0.01}_{-0.01}$(stat)$^{+0.06}_{-0.06}$(syst)
μ_{NR}	$1.11^{+0.08}_{-0.07}$(stat)$^{+0.05}_{-0.05}$(syst)	$1.15^{+0.02}_{-0.02}$(stat)$^{+0.05}_{-0.05}$(syst)	$1.15^{+0.02}_{-0.02}$(stat)$^{+0.05}_{-0.05}$(syst)
μ_{WZ}	$1.01^{+0.05}_{-0.05}$(stat)$^{+0.06}_{-0.05}$(syst)	$0.97^{+0.02}_{-0.02}$(stat)$^{+0.06}_{-0.05}$(syst)	$0.97^{+0.02}_{-0.02}$(stat)$^{+0.06}_{-0.05}$(syst)
Expected significance	16.8	16.2	26.1

The BDT classifier response fit in relaxed phase space shows greater expected significance and reduced statistical error.

Changes in the signal region

- Not to use selection on the variable $E_{\mathrm{T}}^{\mathrm{miss}}$-significance because of the difficulty of using it at the truth level
- Adding a selection to the $E_{\mathrm{T}}^{\mathrm{miss}}/H_{\mathrm{T}}$ variable
- Optimization of selection was also performed through maximization of signal significance Z using a multivariate approach.
- Additional $m_{\ell\ell}$ was added for the **strict** selection as it increased the estimate of the statistical significance.

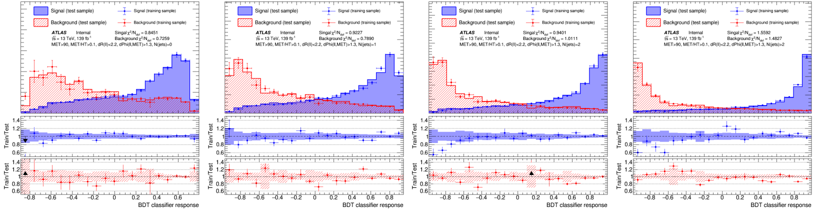
	Strict	**Loose**
$m_{\ell\ell}$, GeV	$\in [80; 100]$	$\in [76; 106]$
$E_{\mathrm{T}}^{\mathrm{miss.}}$, GeV	> 110	> 90
$\Delta R(\ell\ell)$	< 1.8	< 2.2
$\Delta\varphi(\mathrm{miss}, \ell\ell)$	> 2.7	> 1.3
$E_{\mathrm{T}}^{\mathrm{miss.}}/H_{\mathrm{T}}$	> 0.65	> 0.1
$N_{\text{b-jets}}$	$= 0$	$= 0$

	Strict	**Loose**
Signal	1562 ± 15	3810 ± 20
Background	1007 ± 17	25000 ± 300
Z	41.1	23.5

Changes in the classifier

- Training of classifiers in 4 regions with number of jets 0, 1, 2 and more than 2
- Adding variables related to jets
- Hyperparameter optimization

$N_{jets}=0$ $N_{jets}=1$ $N_{jets}=2$ $N_{jets}>2$



$N_{jets}=0$	$N_{jets}=1$	$N_{jets}=2$	$N_{jets}>2$
p-Value(Signal) = 0.65	p-Value(Signal) = 0.55	p-Value(Signal) = 0.53	p-Value(Signal) = 0.057
p-Value(Bckg) = 0.80	p-Value(Bckg) = 0.72	p-Value(Bckg) = 0.44	p-Value(Bckg) = 0.080

Optimal hyperparameters:

- NTrees: 400
- Shrinkage: 0.2
- MinNodeSize: 10%

	Signal	Bkg	Signif.
$N_{jets}=0$	1311.96	690.149	40.51
$N_{jets}=1$	639.818	611.829	22.63
$N_{jets}=2$	234.449	266.918	12.78
$N_{jets}>2$	139.62	278.742	7.78
Total	2325.85	1847.91	46.36

Conclusion

- Statistical error reductions were shown when using the classifier response in signal extraction.
- In the near future, the plan is to do a fit in a *loose* signal region and compare it to a fit in a *strict* signal region.

backup

Object selection

Electrons

- Likehood medium
- lead > 30 ГэВ
- sublead > 20 ГэВ
- $|\eta|$ calo cluster $<$ 2.47
- $|\Delta(z_0) \cdot \sin(\theta)| < 0.5$ мм
- $|d_0$-significance$| < 5$
- Isolation WP FixedCutLoose
- Crack region veto
- Исключение пересечений с мюонами и струями

Muons

- Medium
- $|\eta| < 2.5$
- lead > 30 ГэВ
- sublead > 20 ГэВ
- Combined muons
- $|\Delta(z_0) \cdot \sin(\theta)| < 0.5$ мм
- $|d_0$-significance$| < 3$
- Isolation WP PflowLoose_FixedRad
- Исключение пересечений со струями

Jets

- AntiKt4EMPFlow
- > 30 ГэВ
- $|\eta| < 4.5$
- JVT > 0.5
- Event-level cleaning for LooseBad jets

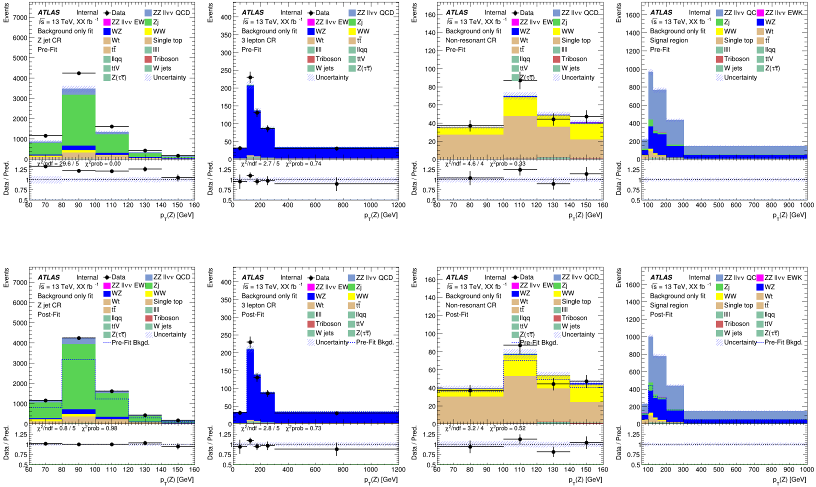
E_T^{miss}

- Tight WP, rebuilt with METMaker using selected leptons and calibrated jets

Selection optimization details

- $E_T^{miss.}$, $[50;1500]$ GeV, a step of 10 GeV;
- $\Delta R(\ell\ell)$, $[0;4]$, a step of 0.1;
- $\Delta\varphi(\text{miss},\ell\ell)$, $[0;3.15]$, a step of 0.1;
- $E_T^{miss.}/H_T$, $[0;2]$, a step of 0.05;
- $N_{\text{b-jets}}$, events with $\{0,1,2,3,\geq 4\}$ b-jets.

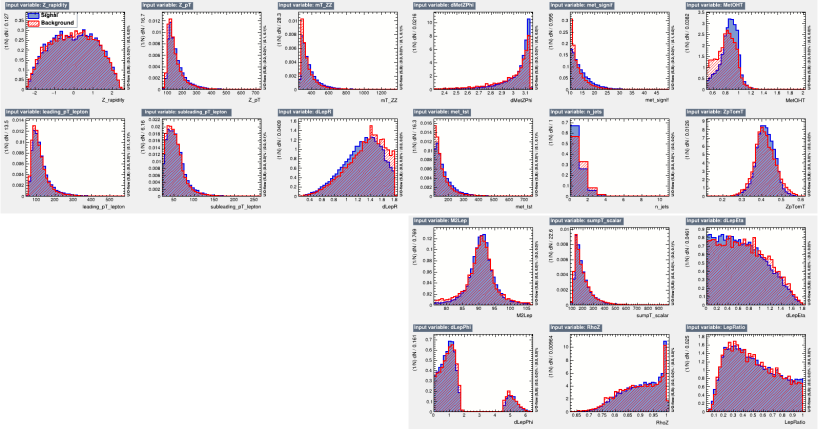
Fit. Before and after distributions for the strict version of the phase space



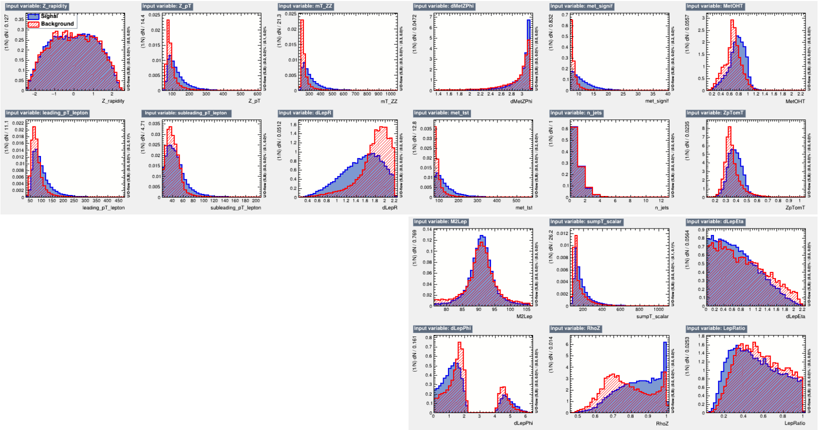
Theoretical		Experimental	
PDF	3.5%	Lepton.	2.0%
Scale	2.0%	Jet.	2.0%
UEPS	2.0%	$E_T^{miss.}$	1.1%

Таблица: Main sources of Theoretical и Experimental errors

Variables. Strict preselection.

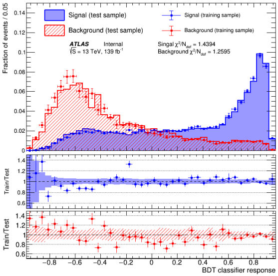


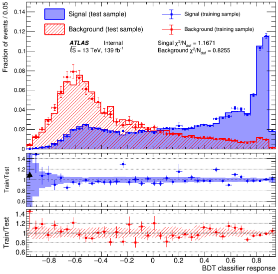
Variables. Loose preselection



Отбор переменных

Variable	auROC
M2Lep	0.79267
dMetZPhi	0.80029
dLepR	0.80434
MetOHT	0.80567
n_jets	0.80690
leading_pT_lepton	0.80715
mT_ZZ	0.80740
subleading_pT_lepton	0.80748
frac_pT	0.80818
sumpT_scalar	0.80809
met_tst	0.80813
LepRatio	0.80799
dLepPhi	0.80787
dLepEta	0.80693
Z_rapidity	0.80623
Z_pT	0.80393
sumpT_vector	0.80371
ZpTomT	0.79994
RhoZ	0.79746





- Идея в том, чтобы измерить важность переменной, глядя на сколько увеличивается *auROC*, когда переменная добавляется.
- Отбор начинается с одной переменной с наибольшим *auROC* и последовательно добавляет переменную из оставшихся $N - n$ с самым высоким *auROC*.
- Это предполагает обучение BDT для каждого из $N - n$ комбинации для определения *auROC* и нахождения лучшей комбинации.

Увеличение значимости с 46.1 ± 0.4 до 46.8 ± 0.4

Training setup

Variable	$N_{jets}=0$	$N_{jets}=1$	$N_{jets}=2$	$N_{jets}>2$
mT_ZZ	✓	✓	✓	✓
$leading\_pT\_lepton$	✓	✓	✓	✓
$subleading\_pT\_lepton$	✓	✓	✓	✓
$dLepR$	✓	✓	✓	✓
$dMetZPhi$	✓	✓	✓	✓
$E_T^{miss}signif.$	✓	✓	✓	✓
$frac\_pT$	✓	✓	✓	✓
$MetOHT$	✓	✓	✓	✓
$M2Lep$	✓	✓	✓	✓
$leading\_jet\_pt$	—	✓	✓	✓
$leading\_jet\_rapidity$	—	✓	✓	✓
$second\_jet\_pt$	—	—	✓	✓
$second\_jet\_rapidity$	—	—	✓	✓
mjj	—	—	✓	✓
$dYjj$	—	—	✓	✓
jet_vsum_pt	—	—	✓	✓
jet_vsum_eta	—	—	✓	✓
jet_vsum_phi	—	—	✓	✓

Preselection	
E_T^{miss}, GeV	>90
ΔR_{ll}	<2.2
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$	>1.3
N_{b-jets}	<1
E_T^{miss}/H_T	>0.1

Hyperparameter optimization

- Hyperparameter optimisation (HPO) has been used to improve separation power while maintaining stability.
- Optimised metrics — Z, with condition p-valuse>0.05%
- Training and test samples were randomly allocated each time, avoiding bias (SplitMode=random:SplitSeed=0)
- Hyperparameters (HP) under consideration:
  - NTrees: 100, 200, 300, 400, 600, 800, 1000
  - Shrinkage: 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.3, 0.4, 0.5, 1.0
  - MinNodeSize: 0.2%, 0.5%, 1%, 2%, 5%, 10%, 20%

$$Z = \sqrt{2 \times [(S + B) \times \ln(1 + (S/B)) - S]}$$

A detailed comparison of all classifiers is shown in the table.

Selection of classifier hyperparameters

- The performance of the classifier varies from training to training. In other words, a classifier that previously showed excellent performance may perform worse the next time with the same settings.
- During HPO, specific HPs were not selected, but patterns in setting, stability, and separation power were observed:
  0. The larger Z is, the more unstable the classifier is. Therefore, it is essential to find a compromise set of HP.
  1. Classifiers with the highest Z usually have a relatively large number of trees.
  2. Classifiers with the highest Z usually have a relatively large number of trees (400-1000) and shrinkage (0.1-0.5).
  3. At large values of minnodesize (20%), the classifiers had the lowest separating power, but at the value of this parameter of 10%, stable classifiers with large Z were observed.
- Thus, one set of hyperparameters was defined for all categories, at which the classifier has good stability and separability:
  - NTrees: 400
  - Shrinkage: 0.2
  - MinNodeSize: 10%

Inclusive $ZZ \to ll\nu\nu$

- Vertex with 2 tracks with $p_{\mathrm{T}} > 1$ GeV
- Two same flavour opposite-sign leptons (e+e- OR mu+mu-), leading p_{T}>30 GeV, subleading p_{T}>20 GeV
- Veto on any additional lepton with Loose ID and $p_{\mathrm{T}} > 7$ GeV
- $76 < M_{\ell\ell} < 106$ GeV
- $E_T^{miss} > 70$ GeV.

Process	**% of background**	
	Strict	**Loose**
$WZ \to \ell\nu\ell\ell$ — one missing ℓ mimics the signal topology	68%	12%
$Z(\to \ell\ell) + \text{jets}$ — lepton pair with mismeasured $E_{\mathrm{T}}^{\mathrm{miss}}$	15%	69%
$WW \to \ell\nu\ell'\nu'$ — non resonant production of a lepton pair	3%	3%
Wt, t, $t\bar{t}$, ttV — non resonant production of a lepton pair via t-quark	9%	15%
Other backgrounds: 4ℓ, $\ell\ell qq$, VVV, $Z(\tau\tau)$, $W+$ jets	5%	1%