Выделение инклюзивного процесса $ZZ \to \ell\ell\nu\nu$ из фоновых процессов в данных pp столкновений с энергией 13 ТэВ в эксперименте ATLAS

Зубов Д.В.

НИЯУ МИФИ

Научный руководитель: Солдатов Е.Ю.

Консультанты: Петухов А.М., Пятиизбянцева Д.Н.

Москва, 2022

Введение

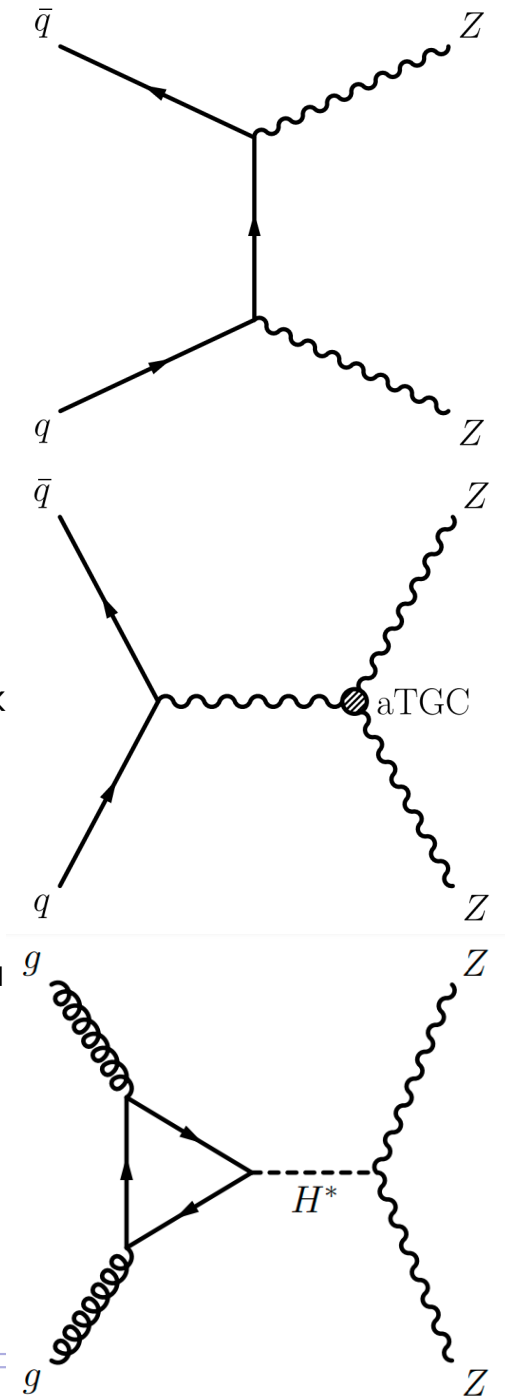
Актуальность и мотивация:

- Прецизионное измерение двухбозонных процессов один из способов проверки Стандартной модели и пертурбативной КХД на масштабе энергий несколько ТэВ
- Измерение aTGC и aQGC является косвенным поиском новой физики
- Многие расширения СМ предсказывают новые скалярные, векторные или тензорные резонансы, которые могут распадаться на пары электрослабых бозонов.

Цель работы:

- Выделение процесса $ZZ \to \ell\ell\nu\nu$ из фоновых процессов с большей точностью за счет улучшения методов отбора и выделения сигнальных событий.

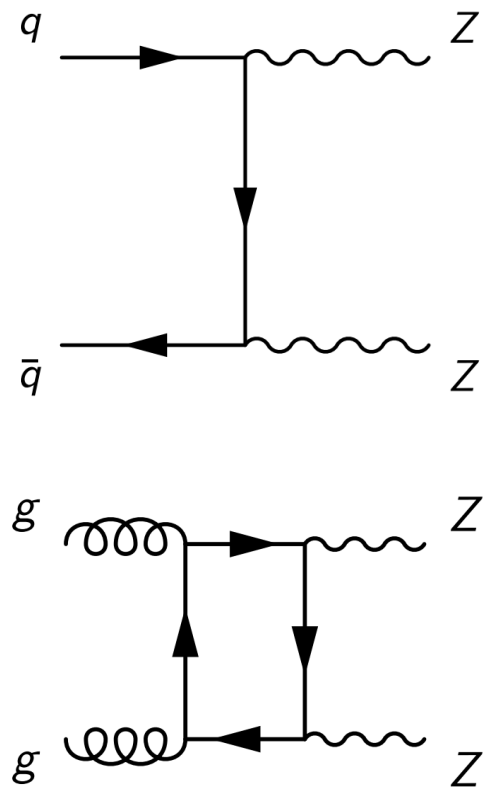
Сечение инклюзивного процесса $ZZ \to \ell\ell\nu\nu$ уже было измерено (arXiv:1905.07163) коллаборацией ATLAS на данных набранных в течение 2015-2016 годов со стат. и сист. ошибками 5.5% и 4.3% соответственно.



Инклюзивный процесс $ZZ \to ll\nu\nu$

- В событии два разноименно-заряженных лептона одного аромата (e+e- или mu+mu-), при этом, поперечный импульс первого больше 30 ГэВ, второго больше 20 ГэВ;
- Вето на третий заряженный лептон;
- 76 ГэВ $< M_{ll} <$ 106 ГэВ;
- $E_T^{miss} > 70$ ГэВ.

Сигнал	
ZZ ($\sim 0.7\%$)	КХД рождение двух Z-бозонов и последующий распад в $ll\nu\nu$
Фон	
Zj ($\sim 85.6\%$)	рождение Z-бозона и струи, с распадом Z-бозона в пару заряженных лептонов и большим ложным потерянным поперечным импульсом
tt ($\sim 11.0\%$)	рождение пары топ-кварков и последующим распадом включающим конечное состояние $ll\nu\nu$ (не резонансное рождение $ll\nu\nu$)
WZ ($\sim 1.0\%$)	рождение пары бозонов Z и W, с распадом Z-бозона в пару заряженных лептонов и лептонным распадом W
WW ($\sim 0.5\%$)	рождение пары W с распадом в $ll\nu\nu$ (не резонансное рождение $ll\nu\nu$)
Wt ($\sim 0.9\%$)	рождение W и топ-кварка и распадом в конечное состояние, содержащее $ll\nu\nu$ (не резонансное рождение $ll\nu\nu$)
VVV	рождение трех векторных бозонов (V = W или Z)
Other (ttV, ttVV)	рождение пары топ-кварков и одного или двух векторных бозонов

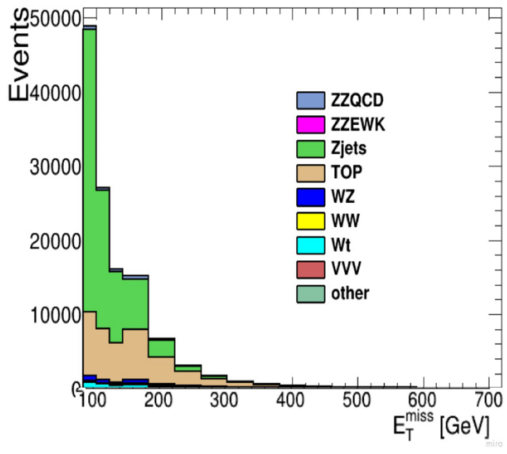


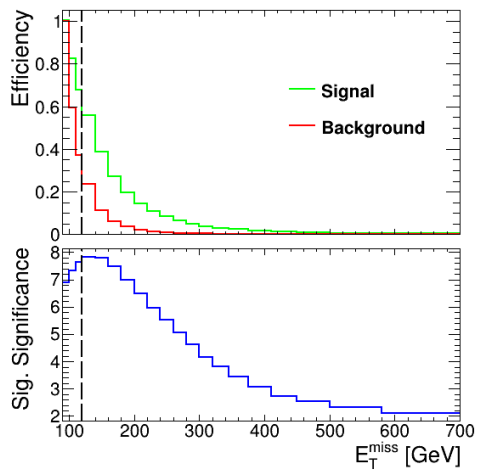
Оптимизация отбора событий «жадным» алгоритмом

- В «жадном» алгоритме оптимизации пороги на переменные определяются последовательно.
- Оптимальный порог соответствует максимуму сигнальной значимости.
- Распределение каждой следующей переменной строится с учетом оптимального порога полученного на предыдущем шаге.
- Используемые переменные в порядке рассмотрения:
  - E_T^{miss}
  - ΔR_{ll}
  - $\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$
  - N_{b-jet}
  - E_T^{miss}-значимость

Предотборы:

- Начальные предотборы

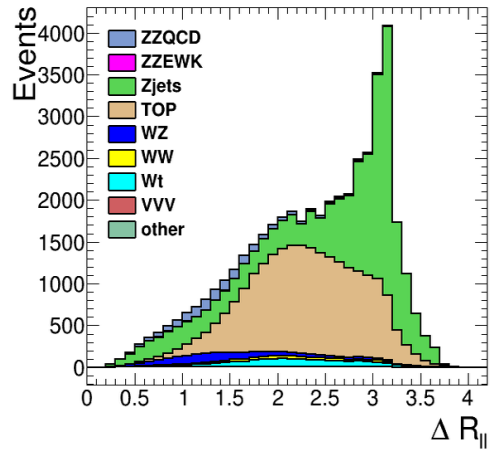


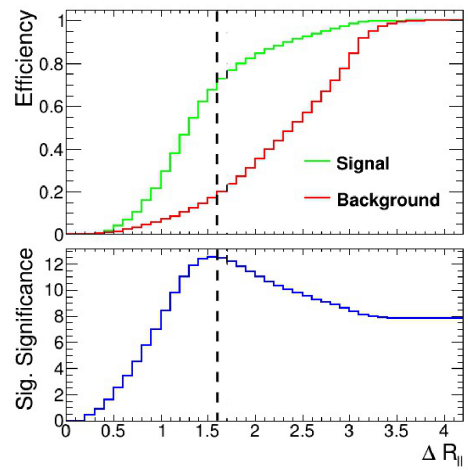


Порог E_T^{miss}>120 ГэВ
Значимость 5.43 → 10.87

Предотборы:

- Начальные предотборы + E_T^{miss}>120 ГэВ





Порог $\Delta R_{ll} < 1.6$
Значимость 10.87 → 17.88

Оптимизация отбора событий «жадным» алгоритмом

В процессе оптимизации было достигнуто значительное подавление фона и увеличение сигнальной значимости.

Переменная	До	После
E_T^{miss}, ГэВ	—	>120
ΔR_{ll}	—	<1.6
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$	—	>2.6
N_{b-jets}	—	<1
E_T^{miss} значимость	—	>11

Недостатки «жадного» алгоритма:

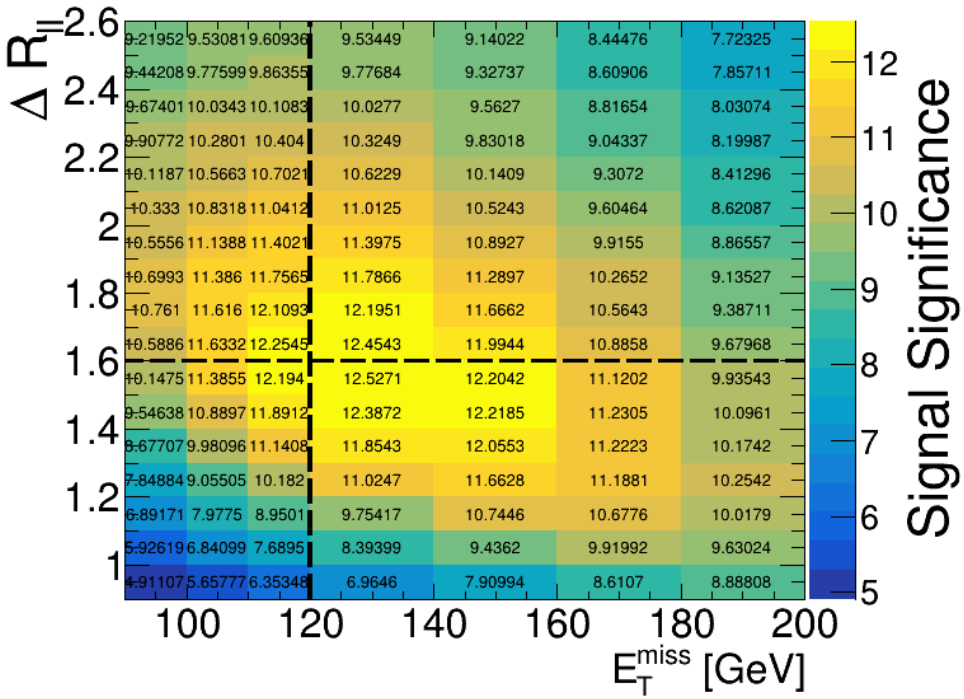
- Результат зависит от последовательности, в которой оптимизируются переменные
- Результат зависит от корреляций между переменными
- Метод не охватывает весь спектр возможных решений

Суммарное число сигнальных событий	7860±30	1322±13
Суммарное число фоновых событий	1123000±4000	790 ± 8
Сигнальная значимость	5.43 ± 0.02	38.9±0.4

Оптимизация отбора событий многомерным алгоритмом

- Многомерный метод лишен недостатков «жадного» алгоритма
- В нем реализован одновременный поиск порогов на все рассматриваемые переменные

- Значимость сигнала рассчитывается для каждого бина многомерной гистограммы, содержащей сигнальные и фоновые события.
- Бининг определяет точность и область фазового пространства, в которой будет искаться решение.



- Многомерный трудно визуализировать, когда число переменных больше двух. На рисунке показана метода в двумерном случае.

Оптимизация отбора событий многомерным алгоритмом

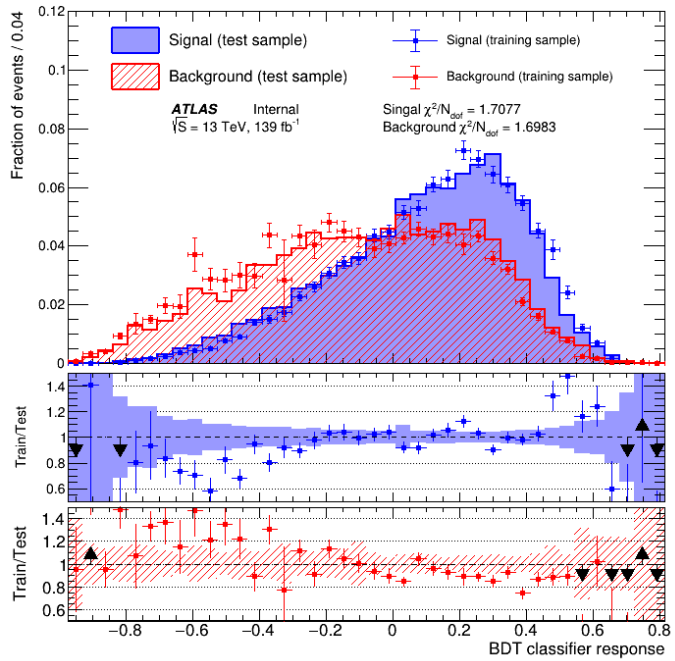
Результаты оптимизации «жадным» и многомерным алгоритмами.

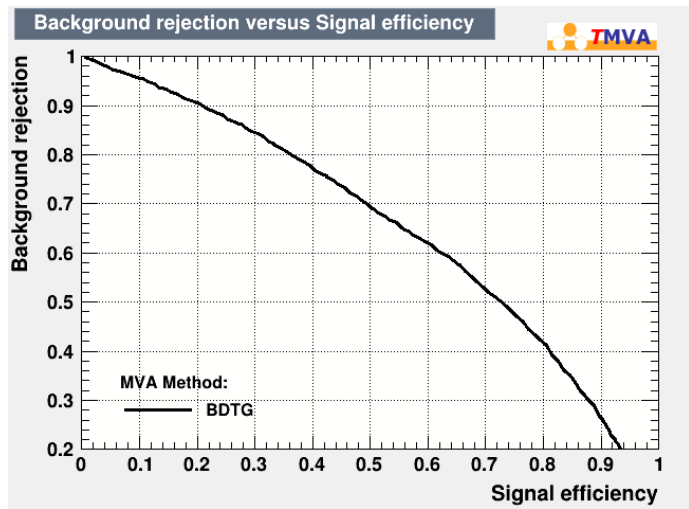
Переменная	До	«Жадный» алгоритм	Многомерный алгоритм
E_T^{miss}, ГэВ	—	> 120	> 70
ΔR_{ll}	—	< 1.6	< 1.8
$\Delta\phi(\mathbf{E}_T^{miss}, \mathbf{p}_T^{ll})$	—	> 2.6	> 2.3
N_{b-jets}	—	< 1	< 1
$\frac{E_T^{miss}}{\text{значимость}}$	—	> 11	> 10
Сигнал	7860 ± 30	1322 ± 13	1926 ± 15
Фон	1123000 ± 4000	790 ± 8	1368 ± 20
Сигнальная значимость	5.43 ± 0.02	38.9 ± 0.4	44.0 ± 0.4

- Жадный алгоритм оптимизации приводит к большему подавлению фона и более высокому отношению сигнал/фон
- Многомерный метод посредством более расслабленных ограничений приводит к более высокому значению значимости сигнала и большему числу сигнальных событий
- Метод многомерной оптимизации можно считать наилучшим, поскольку основной целью исследования является поиск максимальной значимости сигнала

Результаты тренировки классификатора с «жестким» предотбором событий

Переменная	Жесткий отбор
E_T^{miss}, ГэВ	>70
E_T^{miss} значимость	>10
ΔR_{ll}	<1.8
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$	>2.3
N_{b-jets}	<1

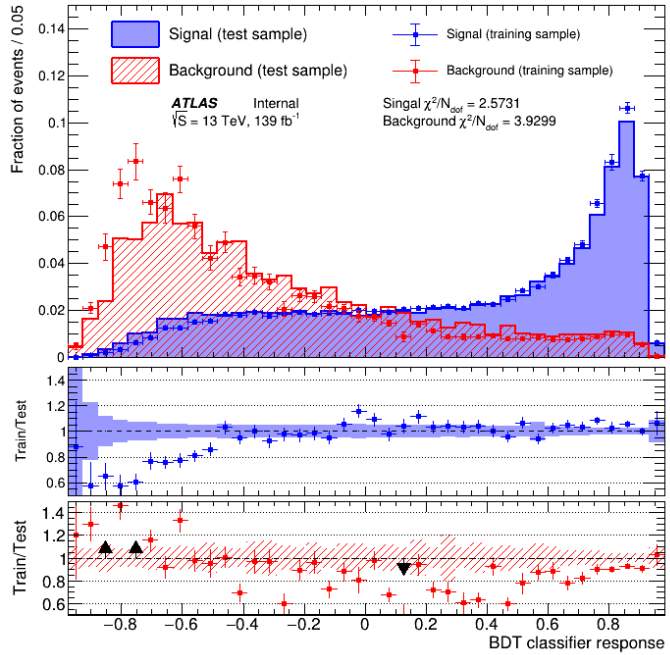


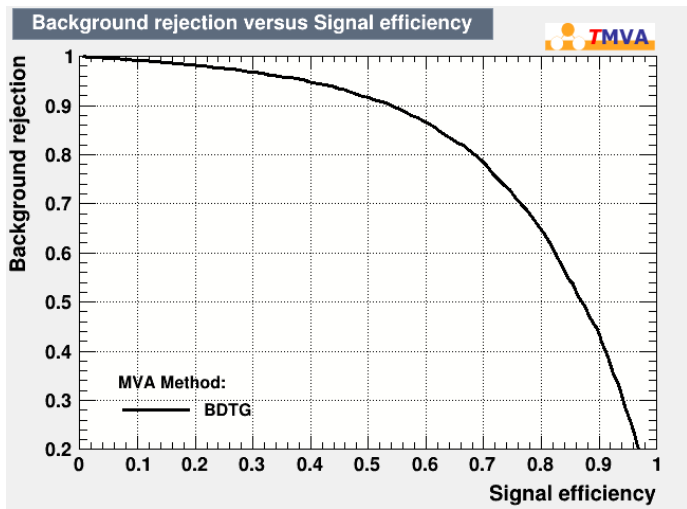


- Не удалось достичь хорошего разделения сигнала и фона
- Классификатор переобучен
- Максимальная сигнальная значимость 44.0±0.4

Результаты тренировки классификатора с расслабленным предотбором событий

Переменная	Расслабленный отбор
E_T^{miss}, ГэВ	>70
E_T^{miss} значимость	>7
ΔR_{ll}	<2.2
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$	>1.3
N_{b-jets}	<1





- Сигнал и фон хорошо разделяются
- Максимальная сигнальная значимость 44.3±0.4
- Есть пространство для улучшения классификатора

Оптимизация и настройка классификатора

6 различных критериев разделения:

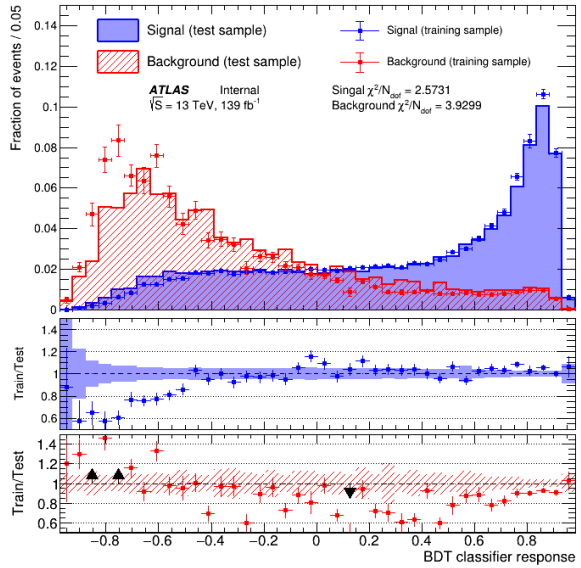
- CrossEntropy
- GiniIndex (Default)
- GiniIndexWithLaplace
- MisClassificationError
- SDivSqrtSPlusB ($S/\sqrt{S+B}$)
- RegressionVariance

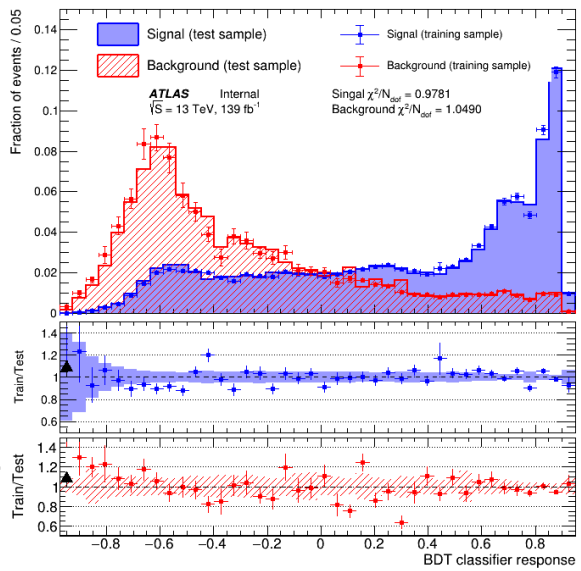
Оптимизация гиперпараметров:

Option	Default	Best
NTrees	400	200
Shrinkage	0.1	0.3
MaxDepth	3	2
MinNodeSize	5%	0.2%

Увеличение значимости с 44.3 ± 0.4 до 46.1 ± 0.4

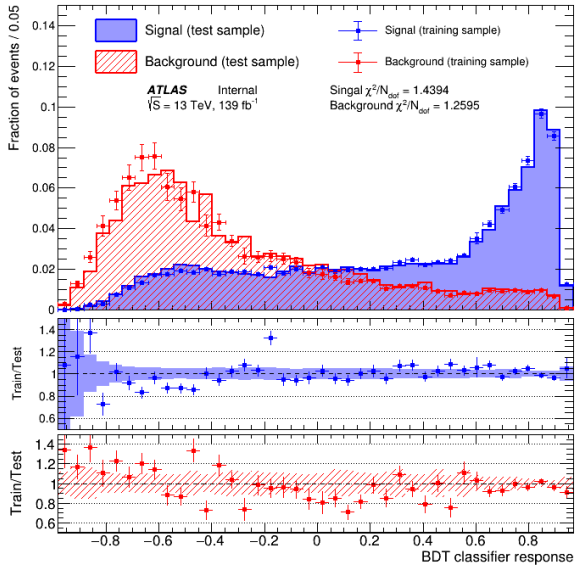
Результаты оптимизации гиперпараметров:

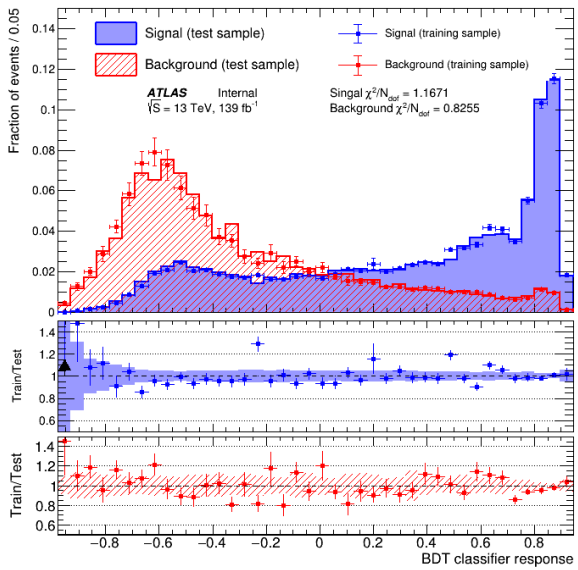




Отбор переменных

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

Одновременный фит

Оценка ожидаемого числа событий выполнялась путем оценки силы сигнала:

$$\mu = \frac{N_{meas.}}{N_{SM}}$$

Фит производился для оценки ожидаемого числа сигнальных событий в два этапа:

1. Фитирование данных предсказаниями о фонах в контрольных регионах для получения оценки нормировочных коэффициентов фона (μ_{Zj}, μ_{WZ}, μ_{NR});
2. Рассматриваются как контрольные, так и сигнальный регионы, а μ_{ZZ} используется в качестве свободного параметра. Поскольку для сигнального региона информация о данных недоступна на текущем этапе анализа, то вместо них используются данные Азимова.

Для описания статистической модели физического эксперимента вводилась функция правдоподобия:

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

Определение контрольных и сигнального регионов.

Расслабленный вариант ФП:

Переменная	SR	WZ(3ℓ)	NR ($e\mu$)	Zj
$E_T^{miss\cdot}$, ГэВ	>70		>70	
ΔR_{ll}	<2.2		<2.2	
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$, рад	>1.3		>1.3	
$E_T^{miss\cdot}$ значимость	>7			[4;7]
m_T^W, ГэВ		>60		

Фит в сигнальном регионе происходил по переменной $BDTscore$, в контрольных по переменной p_T^Z

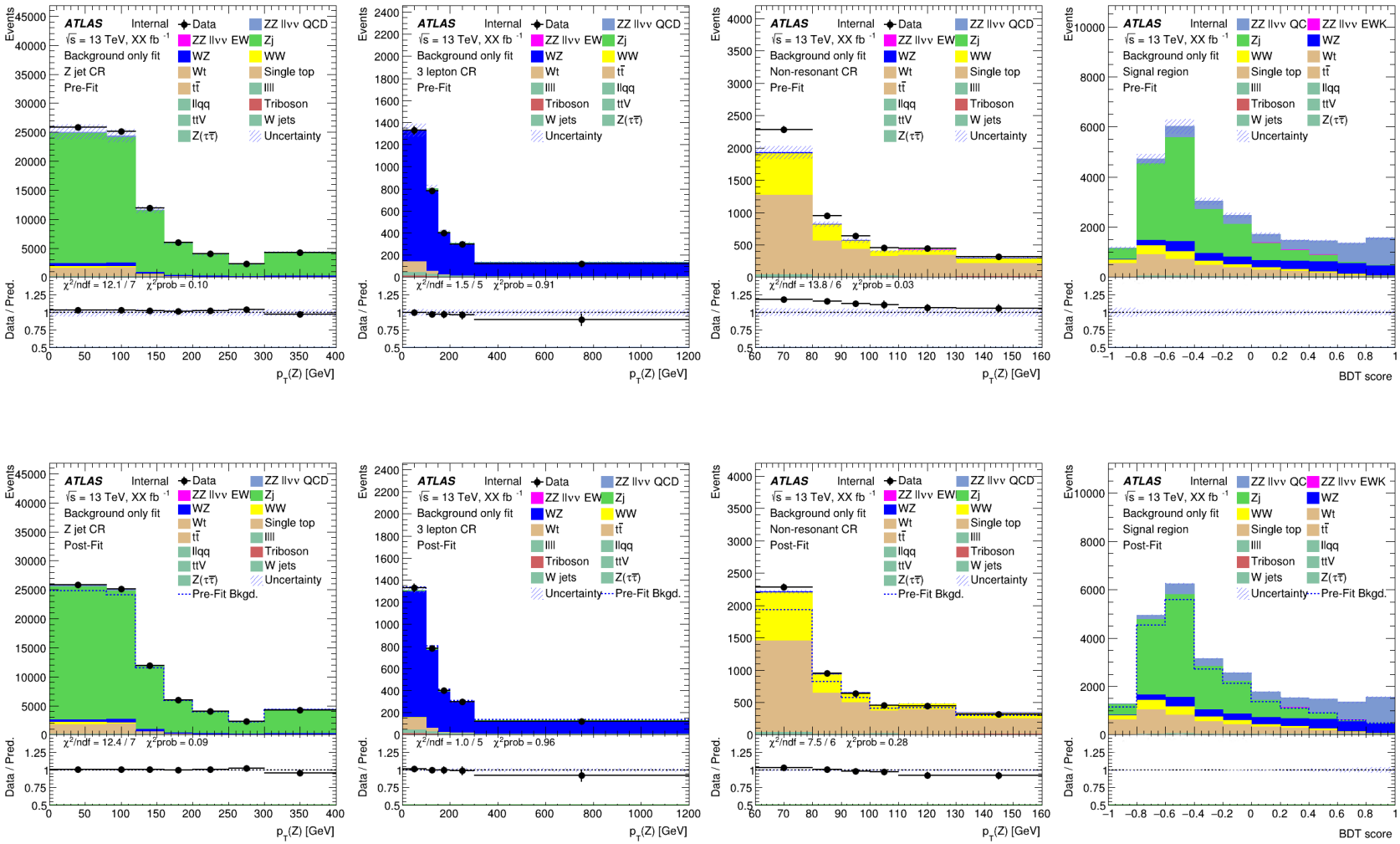
Жесткий вариант ФП:

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

Фит в сигнальном и контрольных регионах происходил по переменной p_T^Z

- **SR** - регион фазового пространства, в котором доля сигнальных событий максимальна.
- **WZ(3l)** - регион фазового пространства, в котором доля событий процесса WZ максимальна.
- **Non-resonant** - регион фазового пространства, в котором доля событий процессов нерезонансного рождения l^+l^- максимальна.
- **Zj** - регион фазового пространства, в котором доля событий процесса Zj максимальна.

Фит. Распределения до и после для расслабленного варианта ФП.



- В ходе фита получены значения нормфакторов μ_{Zj}, μ_{WZ}, μ_{NR}, а также ожидаемые погрешности для нормфактора μ_{ZZ}.

Фит. Результаты.

	«Жесткий» фит	«Расслабленный» фит
μ_{ZZ}	$1.00^{+0.04}_{-0.04}(\text{stat})^{+0.06}_{-0.05}(\text{syst})$	$1.00^{+0.03}_{-0.03}(\text{stat})^{+0.06}_{-0.05}(\text{syst})$
μ_{Zj}	$1.31^{+0.03}_{-0.03}(\text{stat})^{+0.07}_{-0.07}(\text{syst})$	$1.13^{+0.01}_{-0.01}(\text{stat})^{+0.06}_{-0.06}(\text{syst})$
μ_{NR}	$1.11^{+0.08}_{-0.07}(\text{stat})^{+0.05}_{-0.05}(\text{syst})$	$1.15^{+0.02}_{-0.02}(\text{stat})^{+0.05}_{-0.05}(\text{syst})$
μ_{WZ}	$1.01^{+0.05}_{-0.05}(\text{stat})^{+0.06}_{-0.05}(\text{syst})$	$0.97^{+0.02}_{-0.02}(\text{stat})^{+0.06}_{-0.05}(\text{syst})$
Ожидаемая значимость, σ	16.8	26.1

Полученное значение силы сигнала μ_{ZZ} применяется при вычисления наблюдаемого сечения: $\sigma_{\text{meas.}} = \mu_{ZZ} \cdot \sigma_{\text{SM}}$

Заключение

Целью данного исследования является изучение инклюзивного процесса рождения пары Z-бозонов с последующим распадом в конечное состояние $\ell\ell\nu\nu$ и выделение его из фоновых процессов.
В соответствии с поставленной целью в результате данной работы

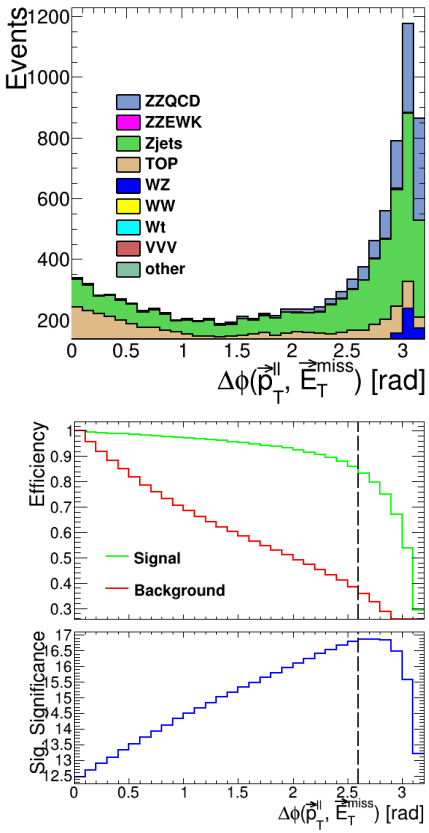
1. Предложен и реализован многомерный метод оптимизации отбора событий, который демонстрирует свою эффективность по сравнению с «жадным» методом поиска оптимальных отборов ($38.9 \pm 0.4\sigma \rightarrow 44.0 \pm 0.4\sigma$)и успешно применяется в других задачах оптимизации отбора событий.
2. Впервые применены алгоритмы машинного обучения в контексте изучения инклюзивного процесса $ZZ \rightarrow \ell\ell\nu\nu$.
3. Произведена настройка и оптимизация классификатора BDTG, в ходе которой получен эффективный алгоритм, несклонный к переобучению и имеющий стабильную разделяющую способность, которая выше чем при выделении сигнала простыми порогами на перемнные ($44.7 \pm 0.4\sigma \rightarrow 46.8 \pm 0.4\sigma$).
4. Произведен фит в сигнальном регионе по распределению отклика оптимизированного классификатора. Результат фита по отклику классификатора показывает значительно большую ожидаемую значимость ($26.1\ \sigma$) и меньшие погрешности определения нормировочных коэффициентов μ по сравнению с фитом в сигнальном регионе по распределению переменной p_T^Z ($16.8\ \sigma$).

backup

Оптимизация отбора событий «жадным» алгоритмом

Предотборы:

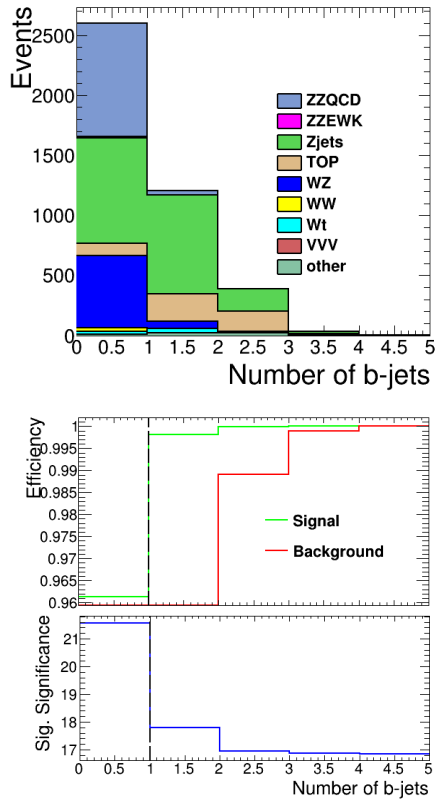
- Начальные предотборы + E_T^{miss}>120 ГэВ + ΔR_{ll} < 1.6



Порог $\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll}) > 2.6$
Значимость 17.88 → 23.6

Предотборы:

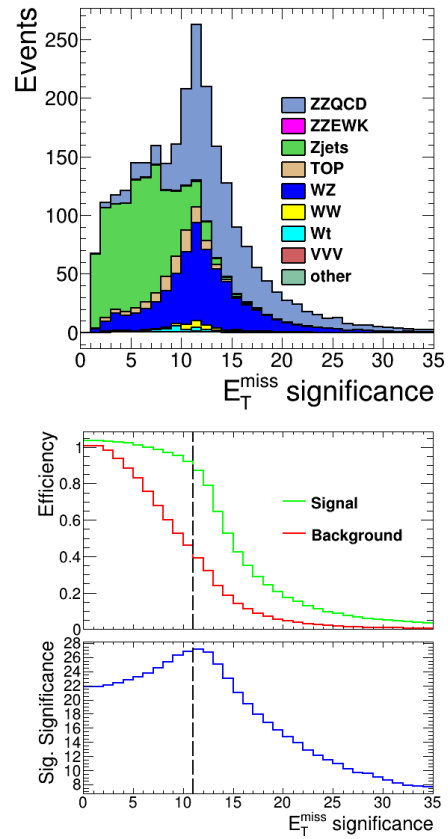
- Начальные предотборы + E_T^{miss}>80 ГэВ + ΔR_{ll} < 1.6 + $\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$ >2.6



Порог $N_{b-jets} = 0$
Значимость 23.6 → 29.4

Предотборы:

- Начальные предотборы + E_T^{miss}>120 ГэВ + ΔR_{ll} < 1.6 + $\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$ >2.6 + $N_{b-jets} = 0$



Порог E_T^{miss}-значимость > 11
Значимость 29.4 → 38.9

Переменная	Отбор	Число сигнальных событий	Число фоновых событий	Сигнальная значимость
E_T^{miss}, ГэВ	> 120	2736 ± 17	$(625 \pm 6) \cdot 10^2$	10.87 ± 0.08
ΔR_{ll}	< 1.6	1864 ± 14	$(103 \pm 5) \cdot 10^1$	17.88 ± 0.14
$\Delta\phi(\mathbf{E}_T^{miss}, \mathbf{p}_T^{ll})$, рад	> 2.6	1594 ± 14	$(405 \pm 2) \cdot 10^1$	23.6 ± 0.2
Число b-струй	< 1	1548 ± 14	2299 ± 21	29.4 ± 0.3
E_T^{miss} значимость	> 11	1322 ± 13	790 ± 8	38.9 ± 0.4

Таблица: Полученные пороги по переменным, количество сигнальных и фоновых события и значение сигнальной значимости на каждом этапе оптимизации «жадным» методом.

	До	После
	Сигнал	
QCD ZZ	$(760 \pm 3) \cdot 10^1$	1317 ± 13
EWK ZZ	262 ± 2	4.35 ± 0.03
Суммарное число сигнальных событий	$(786 \pm 3) \cdot 10^1$	1322 ± 13
	Фон	
Zj	$(963 \pm 4) \cdot 10^3$	64 ± 5
WZ	$(1134 \pm 3) \cdot 10^1$	632 ± 6
tt	$(12334 \pm 8) \cdot 10^1$	19.7 ± 0.9
WW	5093 ± 13	24.2 ± 0.9
Wt	$(1025 \pm 4) \cdot 10^1$	4.3 ± 0.8
VVV	41.8 ± 0.3	11.23 ± 0.15
Other	282 ± 2	0.42 ± 0.07
Суммарное число фоновых событий	$(1123 \pm 4) \cdot 10^3$	790 ± 8
Сигнальная значимость	5.43 ± 0.02	38.9 ± 0.4

Таблица: Число событий и значимость сигнала до и после оптимизации «жадным» методом.

	«Жадный» алгоритм	Многомерный алгоритм
	Сигнал	
QCD ZZ	1317 ±13	1946 ±15
EWK ZZ	4.35 ±0.03	13.0 ±0.4
Суммарное число сигнальных событий	1322 ±13	1959 ±15
	Фон	
Zj	64 ±5	$(18 \pm 2) \cdot 10^1$
WZ	632 ±6	945 ±8
tt	19.7 ±0.9	131 ±2
WW	24.2 ±0.9	64.0 ±1.5
Wt	4.3 ±0.8	41 ±3
VVV	11.23 ±0.15	7.88 ±0.10
Other	0.42 ±0.07	0.79 ±0.11
Суммарное число фоновых событий	790 ± 8	$(137 \pm 2) \cdot 10^1$

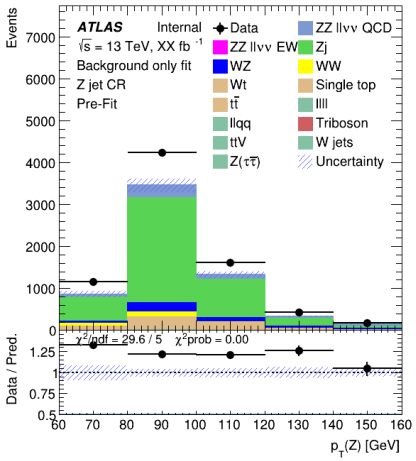
Таблица: Количество событий от разных процессов в фазовом пространстве ограниченном отборами полученными «жадным» и многомерным методами.

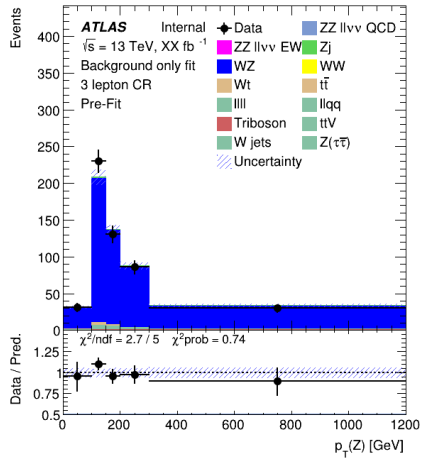
Разделение сигнальных и фоновых событий с помощью алгоритмов машинного обучения

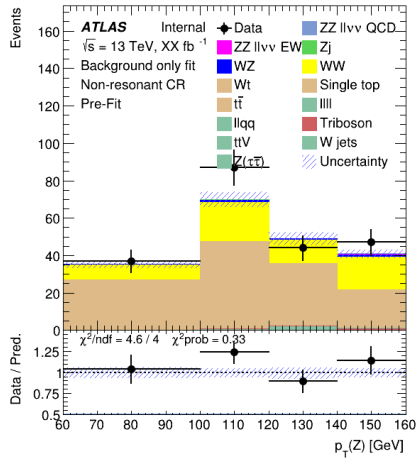
- Машинное обучение применялось в целях максимального разделения сигнальных и фоновых событий и более точной оценки числа сигнальных событий
- Более точная оценка сигнальных событий позволит измерить сечение с меньшей погрешностью
- В качестве алгоритма был выбран BDTG - алгоритм на основе леса деревьев решений использующий градиентный спуск.

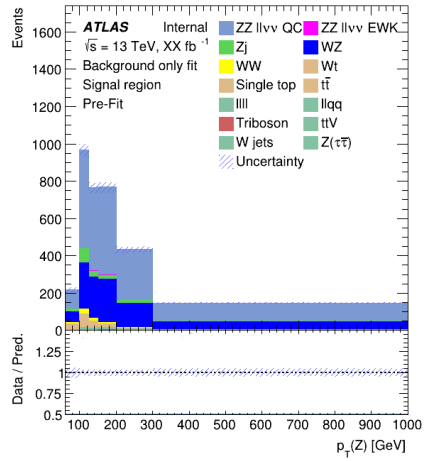
Фит. Распределения до и после для строгого варианта ФП.

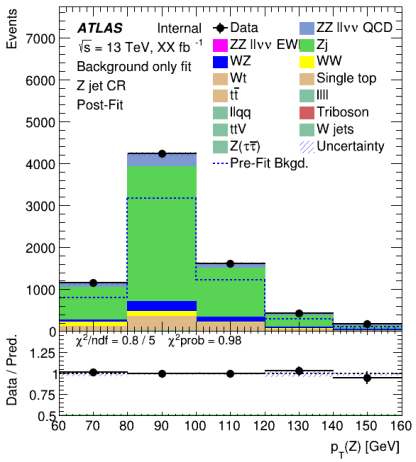
- Сначала фит производился в контрольных регионах на данных для определения нормфакторов основных фоновых процессов.
- Затем с полученными нормфакторами производился фит во всех четырех регионах для определения нормфактора для сигнала

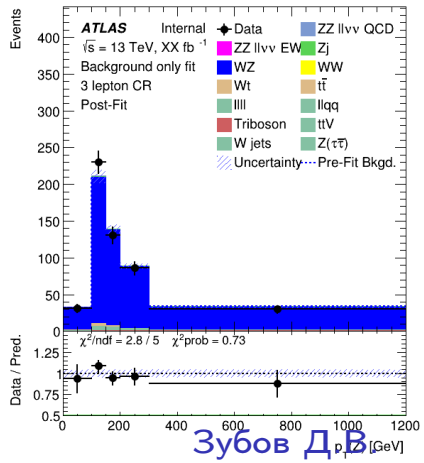


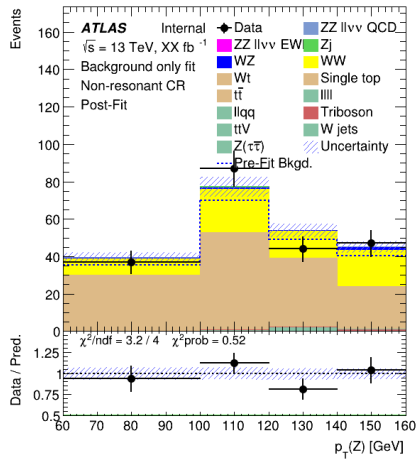


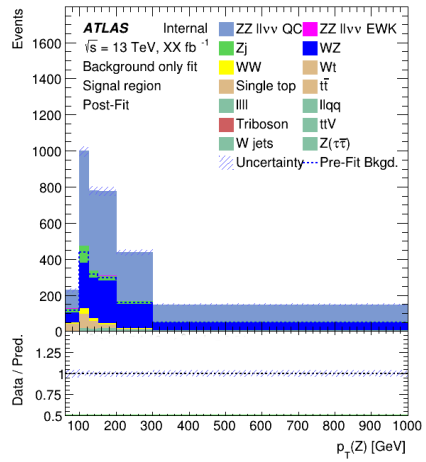












Теоретические		Экспериментальные	
PDF	3.5%	Лептон.	2.0%
Scale	2.0%	Струй.	2.0%
UEPS	2.0%	$E_T^{miss.}$	1.1%

Таблица: Основные источники экспериментальных и теоретических погрешностей

В качестве критерия оценки, по которому определяются значения отборов на переменные, рассматривалась:

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

где S.S. - сигнальная значимость, S - число сигнальных событий, B - число фоновых событий.
Для вычисления значимости открытия и неопределенности оценок $\hat{\mu}$ и $\hat{\theta}$ используется следующая статистика:

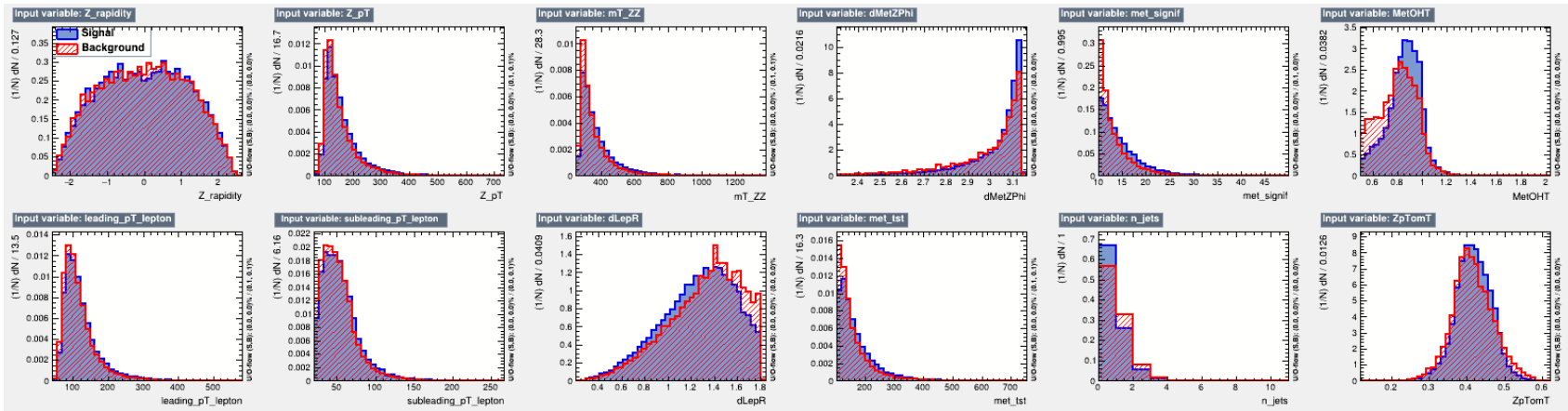
$$q(\mu, \hat{\mu}, \hat{\theta}) = -2 \ln \lambda(\mu, \hat{\mu}, \hat{\theta}) = -2 \ln \frac{\mathcal{L}(\mu, \hat{\hat{\theta}}(\mu))}{\mathcal{L}(\hat{\mu}, \hat{\theta})},$$

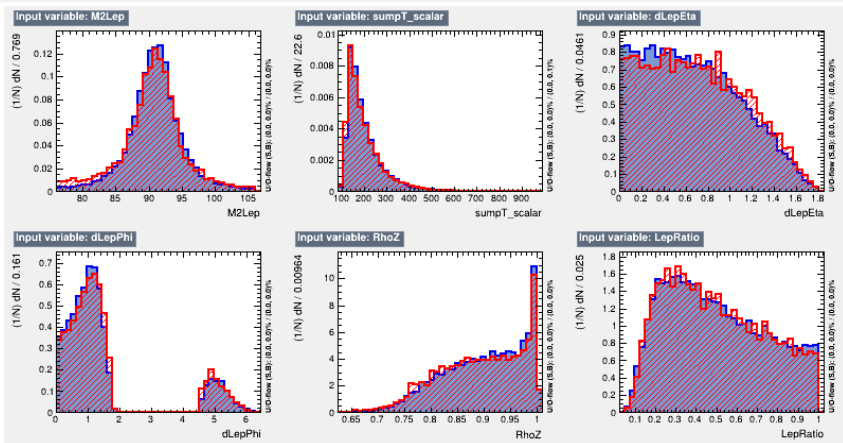
где $\lambda(\mu, \hat{\mu}, \hat{\theta})$ — отношение правдоподобия, а $\hat{\hat{\theta}}(\mu)$ - множество значений θ, минимизирующих $-\ln \mathcal{L}(\mu, \theta)$ для любого заданного μ.
В соответствии с, ожидаемая медианная значимость обнаружения может быть рассчитана следующим образом:

$$Z_{\text{disc}}^{\text{exp.}} = \sqrt{q(\mu=1)_A},$$

где $q(\mu=1)_A$ рассчитывается с использованием набора данных Азимова.

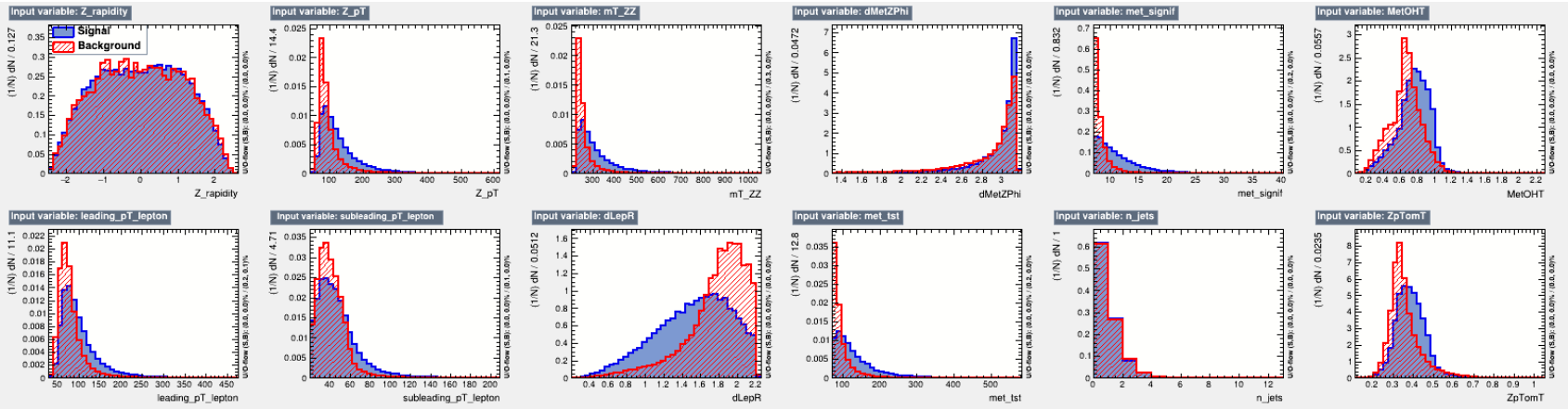
Используемые переменные. Жесткий предотбор





- Распределения рассматриваемых переменных для сигнала и фона сильно пересекаются
- Переменные не имеют разделяющей силы

Используемые переменные. Расслабленный предотбор



- Распределения для сигнала и фона с расслабленной преселекцией различаются в большей степени.