

Поиск распадов Бозона Хиггса на частицы темной материи в эксперименте ATLAS

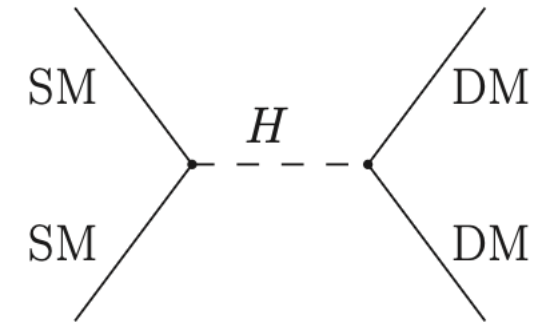
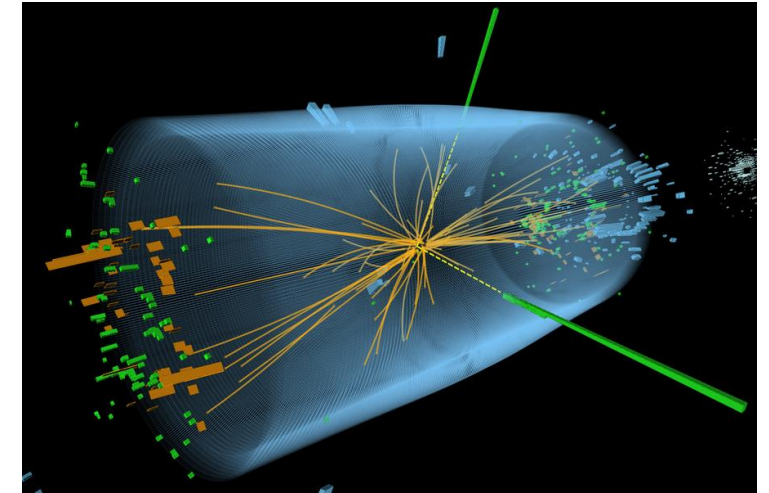
Марина Покидова,
ЛАФ ОФВЭ ПИЯФ



Научный руководитель:
с.н.с., кандидат физ.-мат.наук
Нарышкин Юрий Германович

Введение

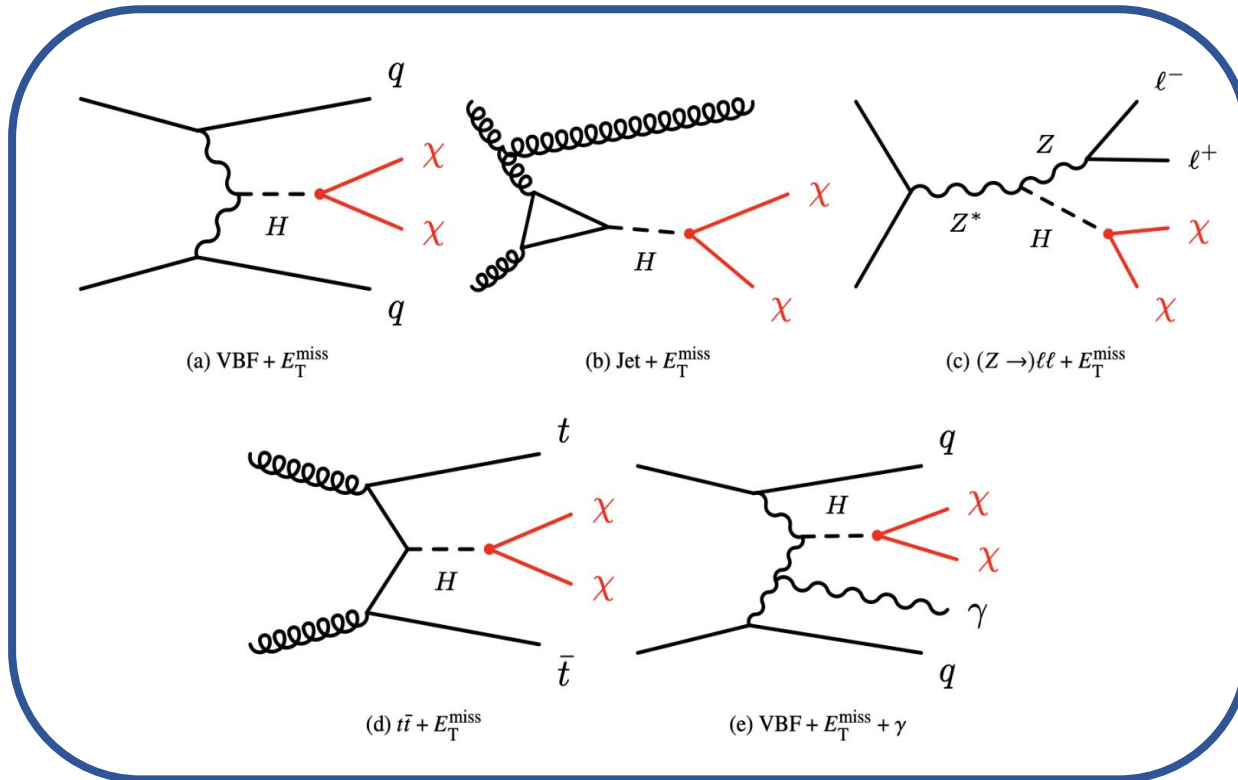
- Поиски новой физики являются одной из актуальных задач на БАК, например, поиски частиц Темной Материи (ТМ)
- Бозон Хиггса обнаружен ATLAS и CMS в 2012 году.
- В некоторых расширениях СМ бозон Хиггса действует как частица-посредник между частицами ТМ и частицами СМ. Такие модели называются “порталом Хиггса”
- В рамках СМ существует «невидимый распад» $H \rightarrow ZZ^* \rightarrow 4\nu$
 $BR(H \rightarrow ZZ^* \rightarrow 4\nu) \sim 0.1\%$
- За пределами СМ существует распад $H \rightarrow \text{WIMP}$



Модели «портала Хиггса» в основе исследования по поиску $H \rightarrow \text{inv}$ в эксперименте ATLAS на БАК

Поиск $H \rightarrow \text{inv}$ в эксперименте ATLAS

В последних исследованиях в эксперименте ATLAS поиск $H \rightarrow \text{inv}$ происходит в основных каналах рождения бозона Хиггса с использованием данных за Run 2 БАК



- $VBF + MET \sim 3,8 \text{ pb}$
- $Jet + MET \sim 49 \text{ pb}$
- $MET + Z(\ell\ell) \sim 0,88 \text{ pb}$
- $t\bar{t} + MET \sim 0,5 \text{ pb}$
- $VBF + MET + \gamma \sim 0,015 \text{ pb}$

Подробнее: [Phys. Lett. B 842 \(2023\) 137963](#)

Участие в исследованиях ATLAS, которые войдут в диссертационную работу

- Mono-Z анализ Run2 [Phys. Lett. B 829 \(2022\) 137066](#)
 - Оценка вклада фонового процесса Z+jets с помощью методов ABCD и One sideband
 - Исследования систематики Z+jets
- Комбинация результатов по поиску невидимых распадов бозона Хиггса Run2 полученных в различных каналах [Phys. Lett. B 842 \(2023\) 137963](#)
 - Исследования перекрытия событий между каналами (overlap check)

$\text{BR}(\text{H} \rightarrow \text{inv}) < 10.7 (7.7) \% \text{ при } 95\% \text{ CL}$ – полученные ATLAS ограничения являются наиболее строгими на данный момент

Представление результатов исследований на конференциях

- IX Всероссийский с международным участием Молодежный научный форум «Open Science 2022», Гатчина, 16–18 ноября 2022, «Комбинация результатов по поиску невидимых распадов Бозона Хиггса на частицы темной материи в эксперименте АТЛАС», Покидова М.В., Бердников Я.А. Нарышкин Ю.Г.
- Неделя науки ФизМех, Санкт-Петербург, 3–7 апреля 2023 г. Статистическая комбинация поисков невидимых распадов бозона Хиггса в эксперименте АТЛАС, Покидова М.В., Бердников Я.А. Нарышкин Ю.Г.
- X Всероссийский с международным участием Молодежный научный форум «Open Science 2023», Гатчина, 15–17 ноября 2023, Результаты поисков распадов бозона хиггса на частицы темной материи в эксперименте атлас Покидова М.В., Бердников Я.А., Нарышкин Ю.Г.
- Научная сессия секции ядерной физики ОФН РАН, 1-5 апреля 2024, Дубна, "Searches for Higgs Boson decays into Dark Matter particles in the ATLAS experiment", Покидова М.В., Бердников Я.А. Нарышкин Ю.Г.
- Седьмая Международная конференция по физике частиц и астрофизике (The 7th International Conference on Particle Physics and Astrophysics (ICPPA-2024)), Москва, Россия, 22-25 октября 2024, «Latest results of searches of Higgs Boson decays into Dark Matter particles in the ATLAS experiment», Покидова М.В., Бердников Я.А. Нарышкин Ю.Г.
- Сессия-конференция ОФН РАН «Физика фундаментальных взаимодействий», г. Москва 17- 21 февраля 2025 «Последние результаты поисков распада бозона Хиггса на частицы темной материи в эксперименте АТЛАС», г. Москва 17- 21 февраля 2025
- LXXV Международная конференция «ЯДРО-2025», Санкт-Петербург, 1–6 июля 2025 г., « Latest results of searches for invisible Higgs Boson decays in events with vector-boson fusion signatures in the ATLAS experiment», Покидова М.В., Бердников Я.А. Нарышкин Ю.Г.
- 26-й Международный Балдинский Семинар по Проблемам Физики Высоких Энергий «Релятивистская Ядерная Физика и Квантовая Хромодинамика», Россия, г. Дубна, 15-20 сентября 2025 г. « Latest results of searches for invisible Higgs Boson decays in in the ATLAS experiment», Покидова М.В., Бердников Я.А. Нарышкин Ю.Г.

Представление результатов исследований

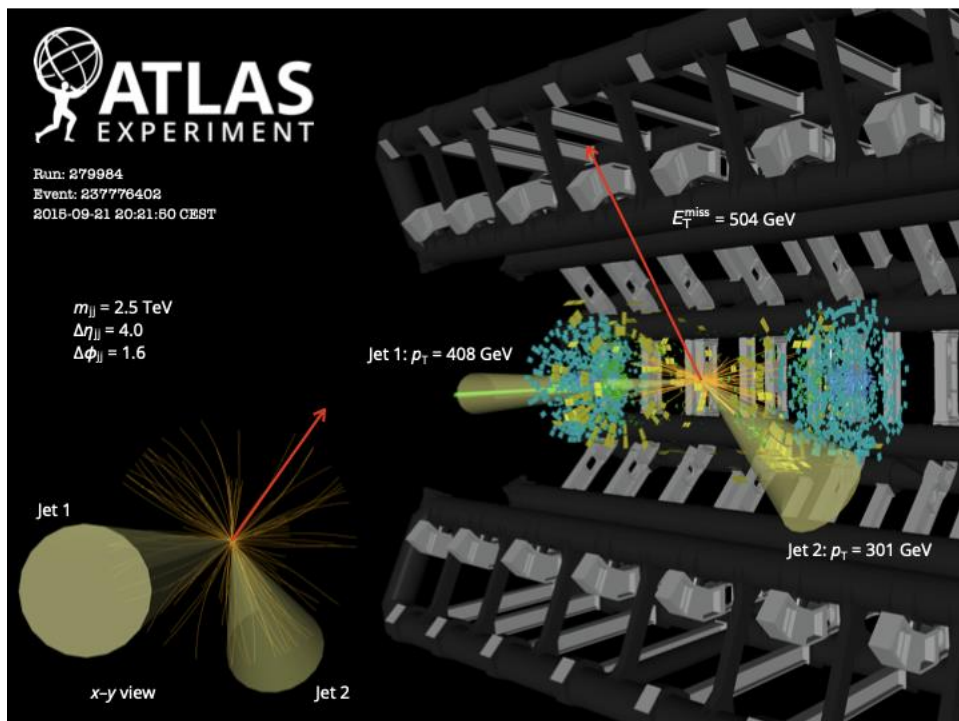
Просидинги конференций:

- «Latest results of searches of Higgs Boson decays into Dark Matter particles in the ATLAS experiment», 7 стр, «Physics of atomic Nuclei», Покидова М.В., Бердников Я . А . Нарышкин Ю.Г.
- «Searches for Higgs Boson decays into Dark Matter particles in the ATLAS experiment» (« Поиски распадов Бозона Хиггса на частицы Темной Материи в эксперименте АТЛАС»), 10 стр, Physics of Elementary Particles and Atomic Nuclei, Покидова М.В., Бердников Я . А . Нарышкин Ю.Г.

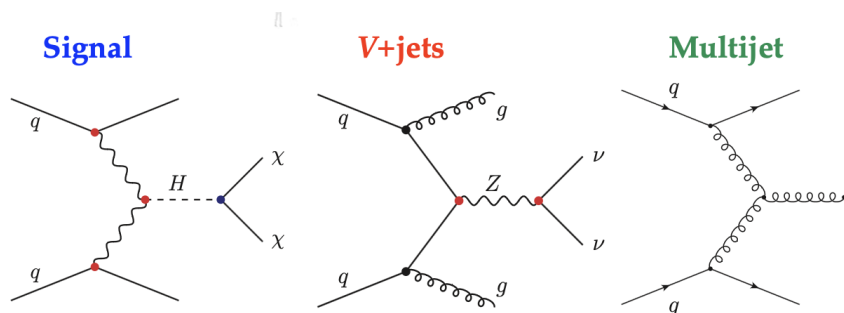
Статьи:

- Статья « Поиск частиц Темной Материи в распадах бозона Хиггса в экспериментах ATLAS и CMS Большого Адронного Коллайдера» принята к опубликованию в журнал "Научно-технические ведомости СПбГПУ. Физико-математические науки", №1-2027

Задачи $H \rightarrow \text{inv}$ поиска VBF+MET в эксперименте АТЛАС



- В Run3 энергия центра масс достигла 13.6 ТэВ → увеличение статистики
- Исследования в процессах, которые статистически ограничены могут достигнуть большую чувствительность
- Анализ VBF+MET для комбинации данных Run2 + частично Run3
 - Оптимизация отбора фоновых процессов и оценка вклада фоновых процессов с помощью методов машинного обучения (MVA/ML подходов)
- Определение верхних пределов на ширину распада $\text{BR}(H \rightarrow \text{inv})$ в VBF+MET канале (pyhf)



$H \rightarrow inv$ при рождении в слиянии векторных бозонов (VBF+MET)

Самым «чувствительным» процессом рождения бозона Хиггса для $H \rightarrow inv$ является VBF+MET канал

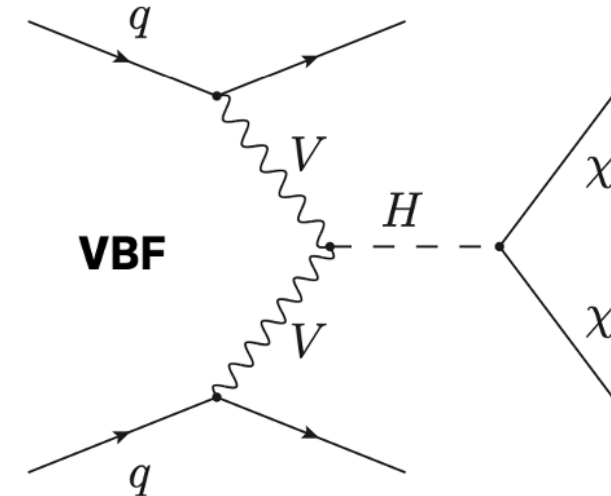
В обоих экспериментах нет статистически значимого отклонения от прогнозов СМ

ATLAS Run2:

- $BR(H \rightarrow inv) < 0.145$ (0.103) при 95% CL

CMS Run2:

- $BR(H \rightarrow inv) < 0.17$ (0.11) при 95% CL



$H \rightarrow inv$ при рождении в слиянии векторных бозонов (VBF+MET)

В VBF+MET анализе исследуемое конечное состояние $H \rightarrow inv$ характеризуется двумя струями с большим разделением по псевдобыстроте и большой инвариантной массой двух струй и E_T^{miss} .

VBF topology

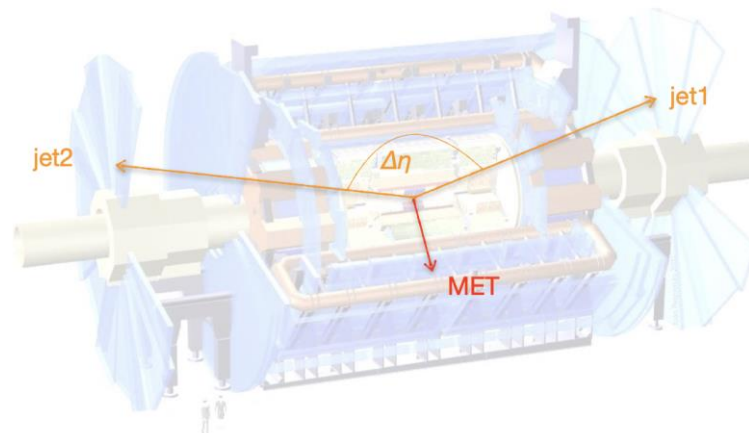
- Two jets with $p_T(j1/j2) > 80/50$ GeV
- Small add. jet activity: $p_T(j3) < 25$ GeV ...
3 or 4 jets, if compatible with FSR
- Jets in opposite hemispheres
- $\Delta\eta(jj) > 3.8$
- $m(jj) > 0.8$ TeV

EWK suppression

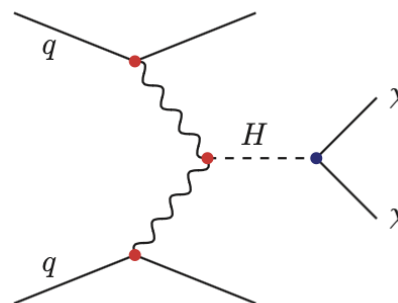
- Lepton (electron and muon) veto

Multijet suppression

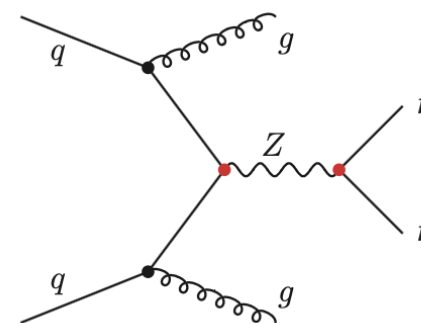
- MET > 160 GeV
- $\Delta\Phi(jj) < 2.0$



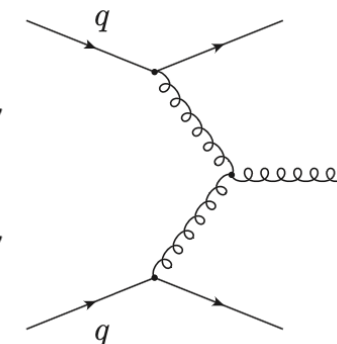
Signal



V+jets

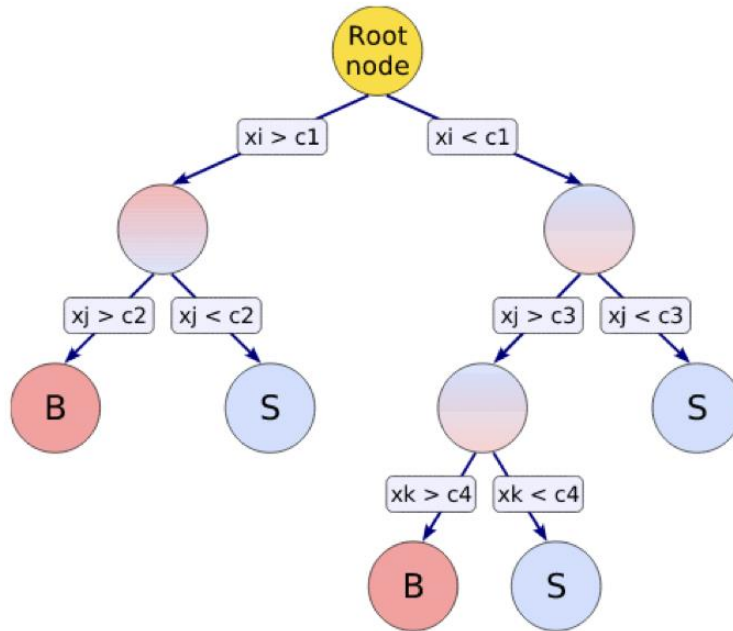


Multijet



Ускоренное дерево решений Boosted decision tree (BDT)

Классификатор на основе ML-фреймворка TMVA root для оптимизации критериев отбора сигнальных и фоновых событий



Для задач классификации сигнал/фон обычно используют BDT с градиентным бустингом – классификатор на основе деревьев решений, где каждое следующее дерево строится на основе минимизации ошибки уже построенного ансамбля.

Сигнал: $VBF\ H \rightarrow inv$,

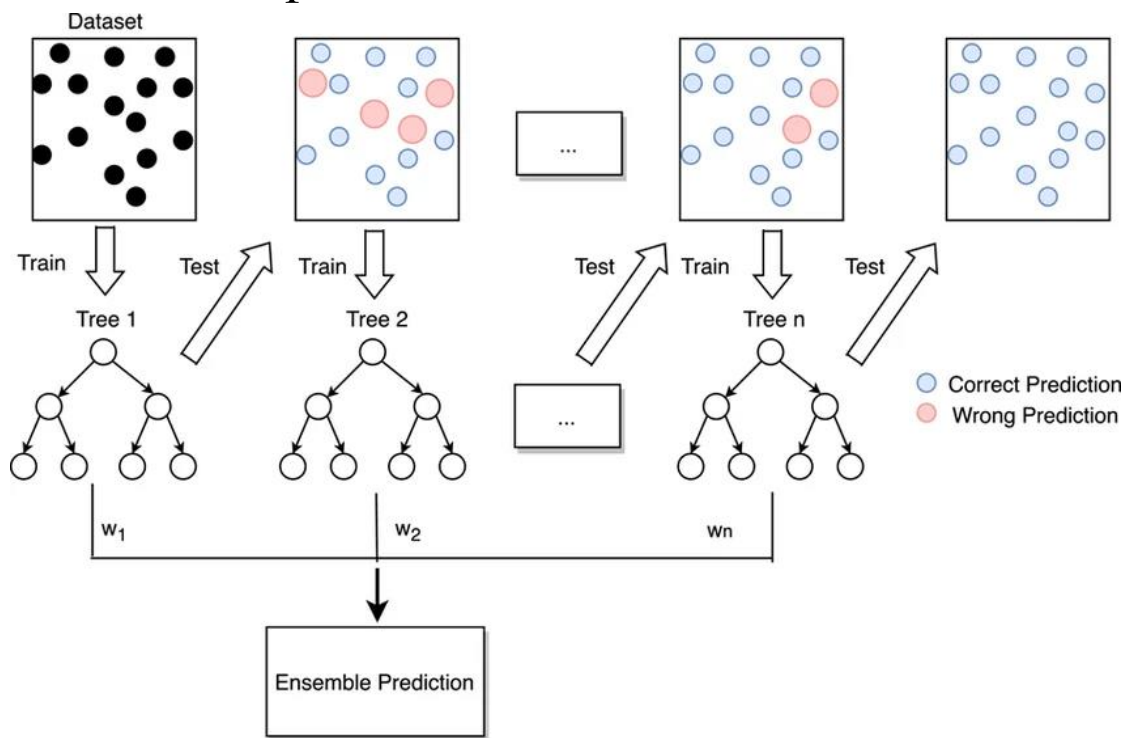
Фоновые процессы:

- $W\ strong$
- $Z\ strong$
- $W\ EWK$
- $Z\ EWK$
- $Multijet$
- $\partial pyque$
(*singletop*,
ttbar, etc)

Алгоритм обучается различать распределения сигнальных и фоновых событий и присваивает каждому событию некоторое значение: $-1 \lesssim \text{BDT score} \lesssim 1$, где 1 соответствуют большей «сигналоподобности» события, -1 – событие похоже на фон.

Ускоренное дерево решений Boosted decision tree (BDT)

Классификатор на основе ML-фреймворка TMVA root для оптимизации критериев отбора сигнальных и фоновых событий



Для задач классификации сигнал/фон обычно используют BDT с градиентным бустингом – классификатор на основе деревьев решений, где каждое следующее дерево строится на основе минимизации ошибки уже построенного ансамбля.

Сигнал: $VBF H \rightarrow inv$,

Фоновые процессы:

- W strong
- Z strong
- W EWK
- Z EWK
- Multijet
- $\partial\pi\pi\pi$
(singletop, $t\bar{t}$ bar, etc)

Алгоритм обучается различать распределения сигнальных и фоновых событий и присваивает каждому событию некоторое значение: $-1 \lesssim \text{BDT score} \lesssim 1$, где 1 соответствуют большей «сигналоподобности» события, -1 – событие похоже на фон.

Оптимизация ML модели

Эффективность классификации зависит от репрезентативности тренировочной/тестовой выборки событий, входных переменных и параметров модели классификатора (для BDTG – это число деревьев, глубина и тд)

- Тренировка/тест модели обычно происходит на четной/нечетной части событий.
- Эффективные входные переменные для разделения событий должны быть не сильно коррелированы между собой и иметь высокую оценку variable importance для классификации
- Оптимальные параметры модели выбираются на основе ROC-int оценки с учетом перетренировки модели (K-S тест)

Входные переменные, которые использовались в Run2:

- $m_{jj}, \Delta\eta_{jj}, \Delta\phi_{jj}$
- $met_tst, met_tst_et, met_tst_nolep_et$
- $jet1\ p_T, jet2\ p_T$
- $\Delta\phi(jet1, MET), \Delta\phi(jet2, MET)$

Какие переменные можно рассмотреть в Run3:

Для $i = 3, 4$

- $f_{i1}^\eta = \frac{(\eta_1 - \eta_i) - (\eta_i - \eta_2)}{\eta_1 - \eta_2}$
- $f_{i1}^\phi = \frac{\Delta\phi_{1i} - \Delta\phi_{i2}}{\Delta\phi_{12}}$
- $f_{i1}^p = \frac{2p_{T,i}}{p_{T,1} + p_{T,2}}$
- $cent_{i1} = C_i = \exp\left(-\frac{4}{(\eta^{j1} - \eta^{j2})^2} \left(\eta^i - \frac{\eta^{j1} + \eta^{j2}}{2}\right)^2\right)$
- $mi_rel = m_i^{rel} = \frac{\min\{m_{j1,i}, m_{j2,i}\}}{m_{jj}}$

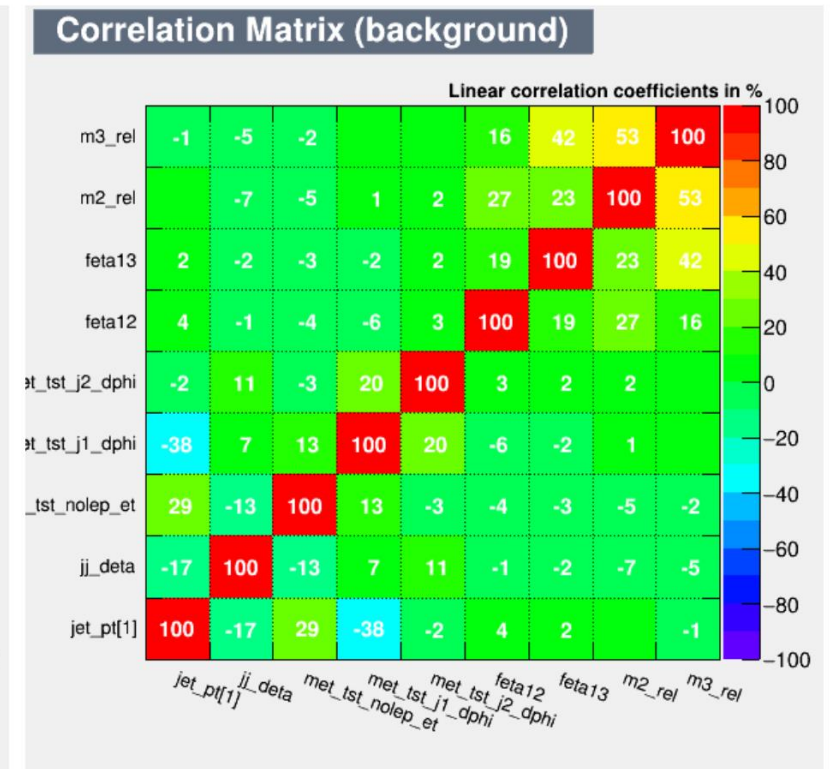
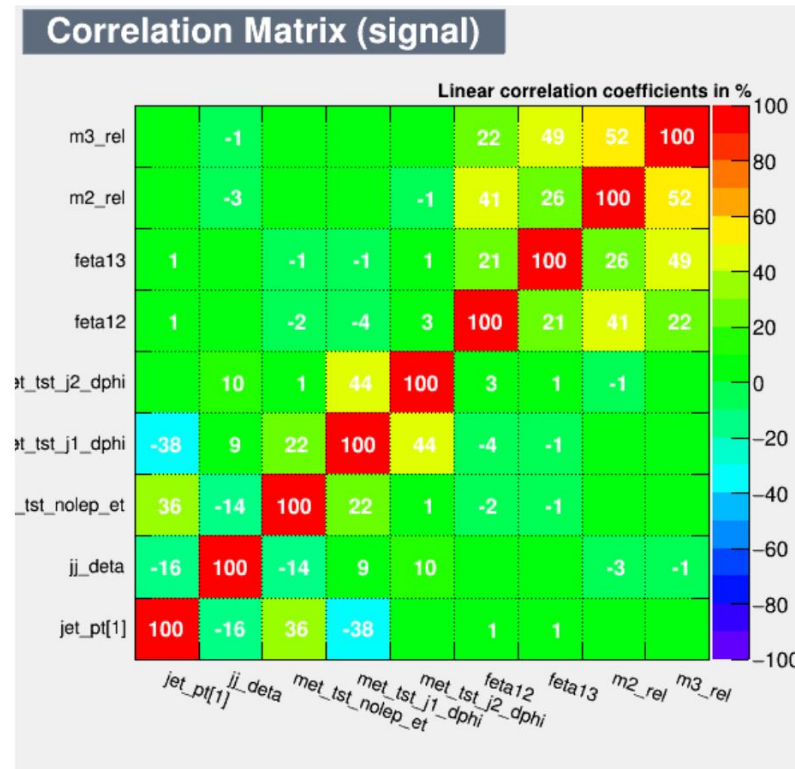
in progress

Оптимизация отбора событий с помощью MVA/ML подходов

- Были выбраны входные переменные, которые дают вклад в классификацию событий по рангу переменных и корреляционной матрице переменных и оптимальные параметры модели (по значениям Roc-int и K-S теста).

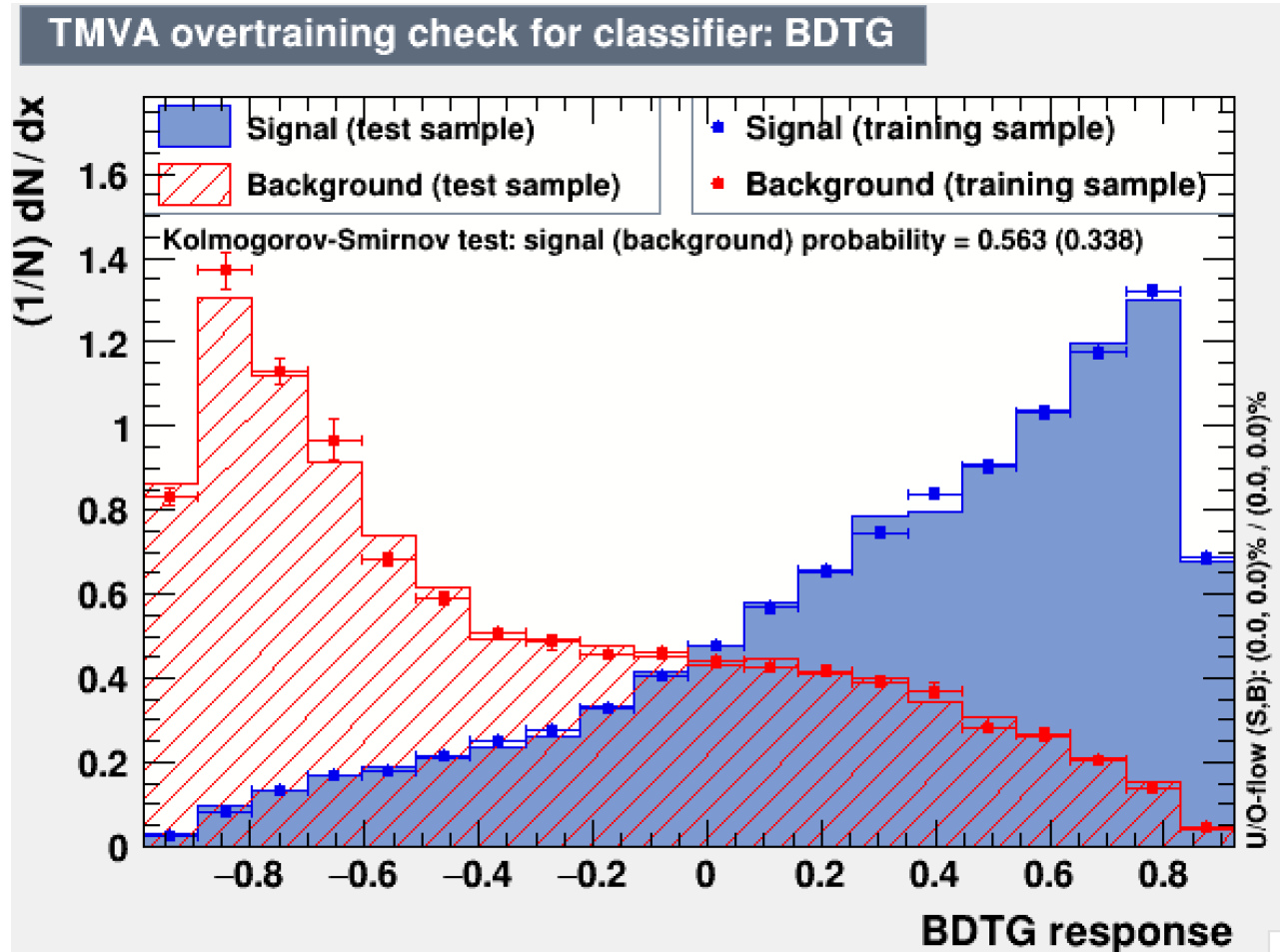
- Оптимальные параметры модели:

NTrees = 100,
MinNodeSize = 2.5%,
nCuts = 20,
MaxDepth = 4,
Shrinkage=0.10



in progress

Результаты BDTG (четные/нечетные события тренировка/тест)



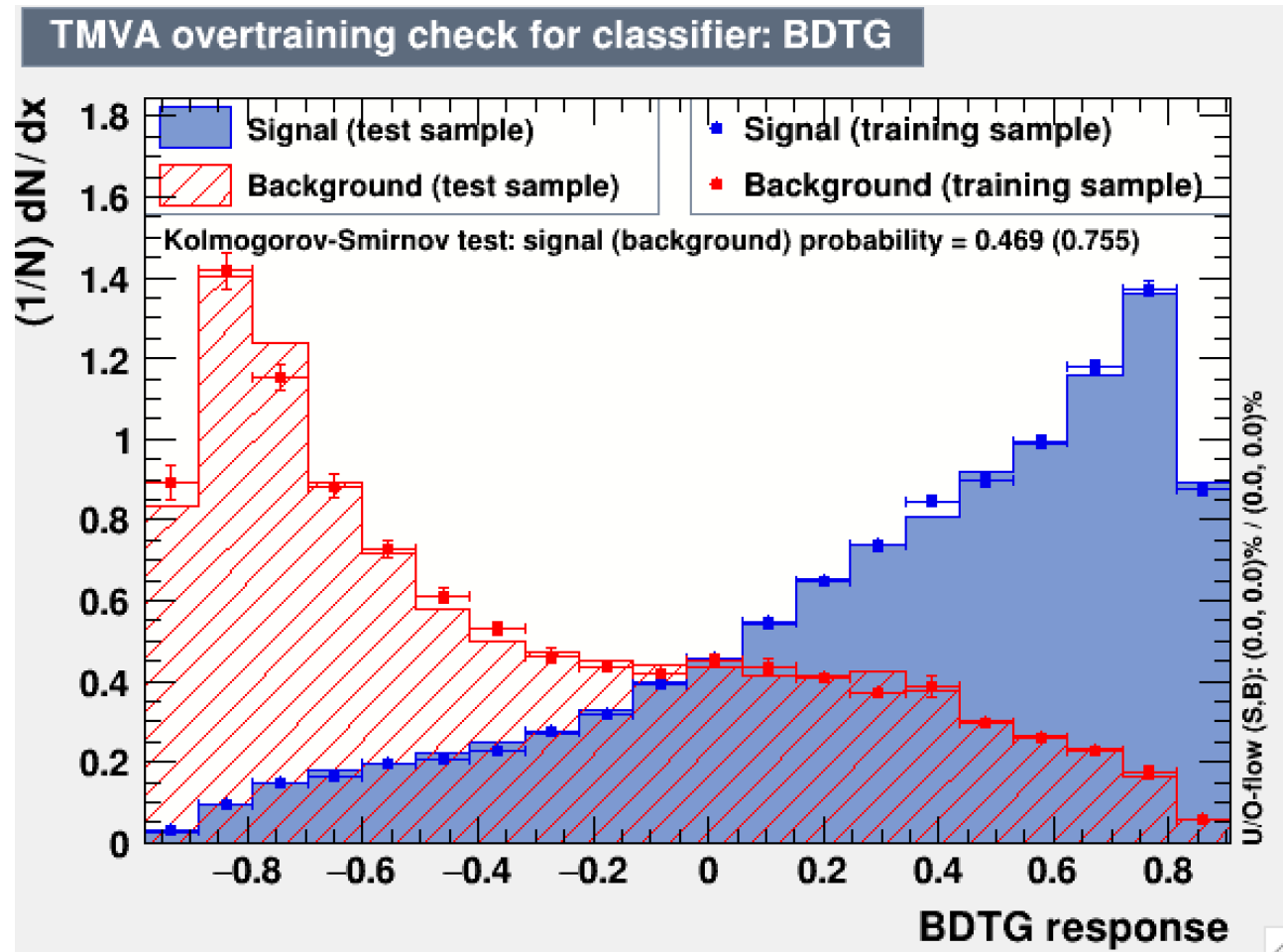
Signal	-- training events	: 47330
Signal	-- testing events	: 47191
Signal	-- training and testing events:	94521
Background	-- training events	: 1830376
Background	-- testing events	: 1833694
Background	-- training and testing events:	3664070

Rank	Variable	Variable Importance
1	feta12	1.311e-01
2	m2_rel	1.283e-01
3	met_tst_j2_dphi	1.135e-01
4	met_tst_j1_dphi	1.128e-01
5	feta13	1.108e-01
6	m3_rel	1.093e-01
7	jj_deta	1.080e-01
8	met_tst_et	9.929e-02
9	jet_pt[1]	8.693e-02

Goodness of signal (background) consistency: 0.563 (0.338)

ROC-int : 0.836

Результаты BDTG (нечетные/четные события тренировка/тест)



```
Signal      -- training events      : 47191
Signal      -- testing events       : 47330
Signal      -- training and testing events: 94521
Background  -- training events      : 1833694
Background  -- testing events       : 1830376
Background  -- training and testing events: 3664070
```

Rank	Variable	Variable Importance
1	feta12	1.341e-01
2	m2_rel	1.291e-01
3	met_tst_j1_dphi	1.283e-01
4	jj_deta	1.179e-01
5	feta13	1.105e-01
6	met_tst_j2_dphi	1.028e-01
7	m3_rel	1.018e-01
8	met_tst_et	8.891e-02
9	jet_pt[1]	8.649e-02

Goodness of signal (background) consistency: 0.469 (0.755)

ROC-int : 0.835

Применение оценки классификатора для верхних пределов на вероятность распада

- Оценка классификатора BDT позволяет получить распределение, обеспечивающее высокую чувствительность в смысле разделения сигнальных и фоновых событий (соотношение S/B для каждого бина гистограммы).

Лучшее разделение $\rightarrow$ Большая чувствительность $\rightarrow$ Более строгий верхний предел

- Впоследствии распределение BDT используется в специальном исследовании по определению верхних пределов на $BR(H \rightarrow \text{inv})$. Для проведения этого исследования применяется библиотека `pyhf` — реализация фреймворка HistFactory на языке Python.

in progress

- В данный момент происходит работа над исследованиями $H \rightarrow \text{inv}$ в канале VBF+MET с использованием данных полученных за Run2 и частично Run3
 - Результаты буду включены в диссертационную работу, ноту и планируемую статью
“Search for invisibly decaying Higgs bosons with the Vector-Boson Fusion plus missing transverse momentum signature with Run 2 and Run 3 data from the ATLAS detector”

Спасибо за внимание

Backup slides

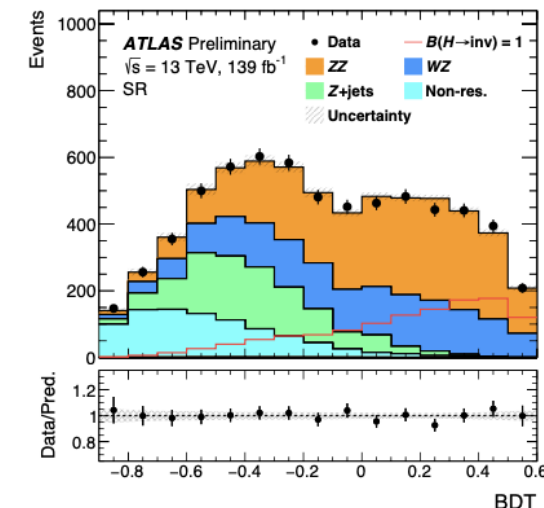
Результаты $H \rightarrow \text{inv}$ поиска в моно-Z канале

Boosted Decision Tree (BDT) – многопараметрический метод анализа для оптимизации сигнальной области

Кинематические переменные, которые входят в состав BDT как параметры:

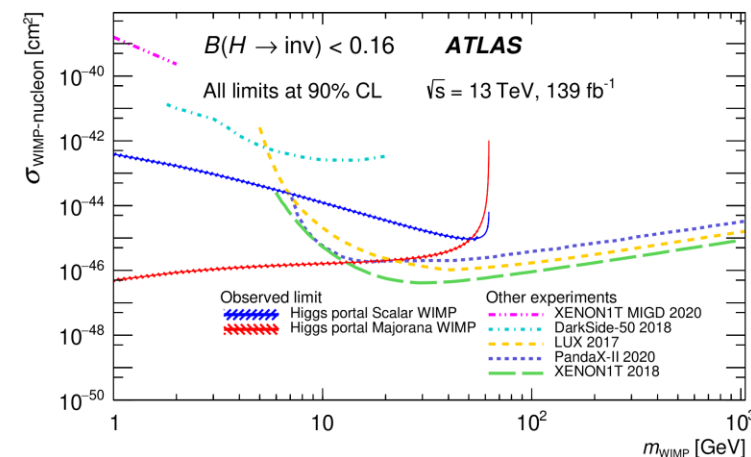
$\Delta R(\ell\ell), m_{\ell\ell}, MET / HT, MET \text{ significance}, \Delta\Phi(E_T^{\text{miss}}, Z), \eta_Z, \text{frac}_{p_T}$
 скалярная сумма p_T

	SR	$e\mu$ CR	3ℓ CR	4ℓ CR
Observed events	6382	891	11415	314
Expected yields after fit	6385 ± 80	894 ± 29	11410 ± 110	295 ± 11
$ZH \rightarrow \ell\ell + \text{inv}$	-4 ± 110	-	-	-
$ZZ \rightarrow \ell\ell\nu\nu$	2669 ± 110	-	443.4 ± 7.5	-
WZ	1624 ± 28	11.59 ± 0.23	10646 ± 110	-
$Z + \text{jets}$	1110 ± 100	0.802 ± 0.018	237.6 ± 4.0	-
Non-resonant	876 ± 39	878 ± 29	-	-
$ZZ \rightarrow \ell\ell\ell\ell$	85.2 ± 5.5	-	-	295 ± 11
$t\bar{t}V$	12.5 ± 1.1	1.769 ± 0.036	48.98 ± 0.82	-
Triboson	12.2 ± 1.4	2.886 ± 0.076	35.65 ± 0.60	-



Вклад фоновых процессов для сигнальной и контрольных областей.

Подробнее: [Phys. Lett. B 829 \(2022\) 137066](#)



Результаты $H \rightarrow \text{inv}$ поиска в эксперименте ATLAS для Run2

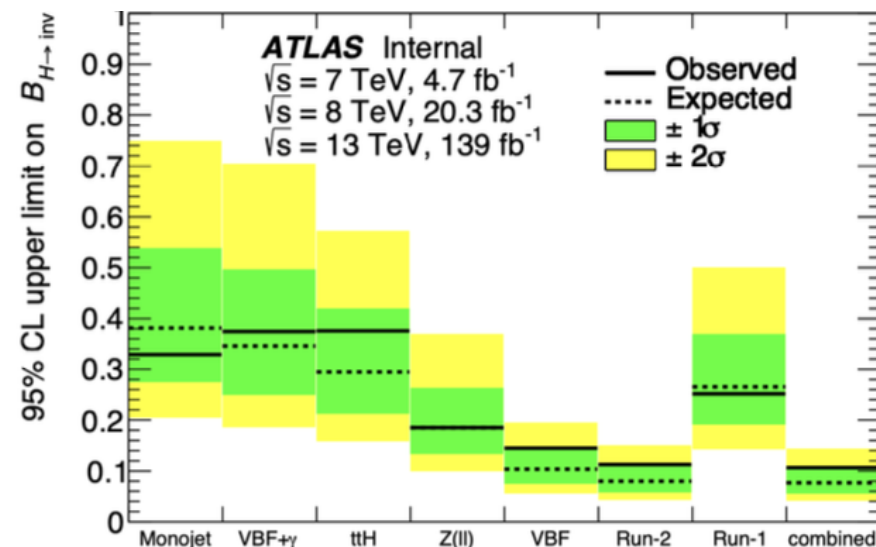
- Значения верхнего предела на вероятность $H \rightarrow \text{inv}$ представлено для всех каналов, Run 1, Run 2 и их комбинации

$B(H \rightarrow \text{inv}) < 10.7 \text{ (7.7) \%}$ при 95% CL

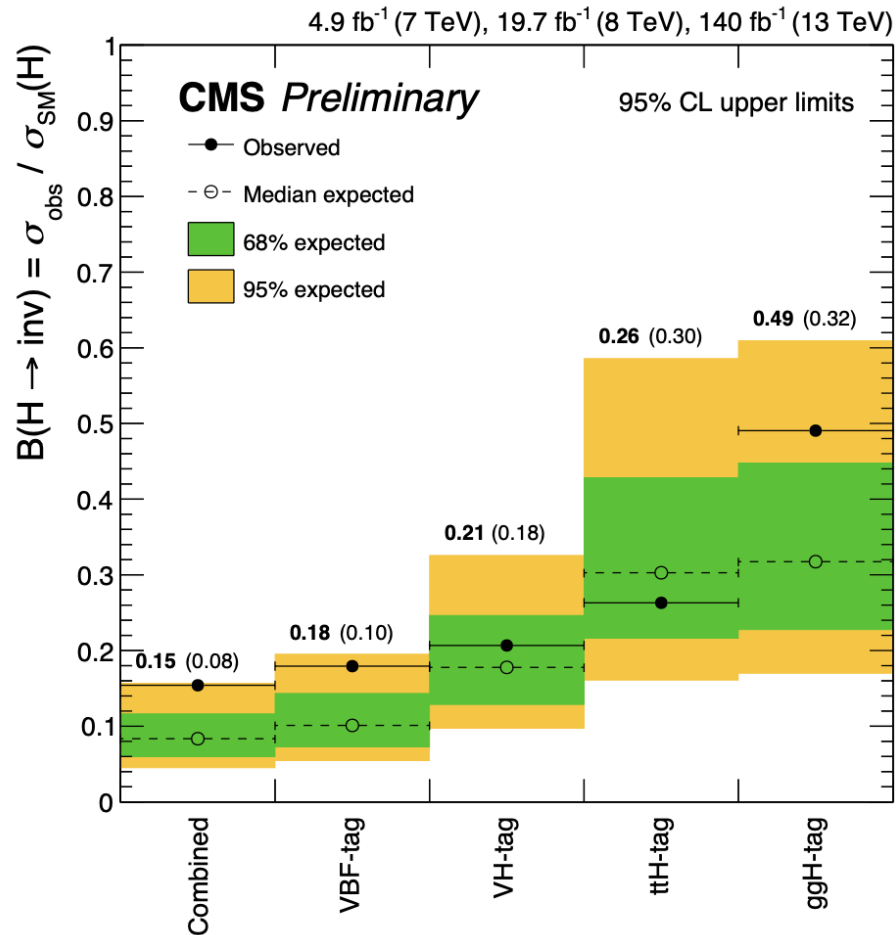
- Полученные ограничения являются наиболее жесткими на данный момент

Подробнее: [Phys. Lett. B 842 \(2023\) 137963](#)

Results	Expected	+1 σ	-1 σ	+2 σ	-2 σ	Observed
Jet + E_T^{miss}	0.381	0.538	0.275	0.749	0.205	0.329
VBF + $E_T^{\text{miss}} + \gamma$	0.346	0.497	0.249	0.704	0.186	0.375
$t\bar{t}$	0.295	0.420	0.212	0.573	0.158	0.376
$(Z \rightarrow) \ell\ell + E_T^{\text{miss}}$	0.185	0.263	0.133	0.370	0.099	0.185
VBF + E_T^{miss}	0.104	0.144	0.075	0.195	0.056	0.145
Run-1	0.265	0.370	0.191	0.501	0.142	0.252
comb. Run2	0.080	0.111	0.058	0.151	0.043	0.113
combined	0.077	0.107	0.055	0.144	0.041	0.107

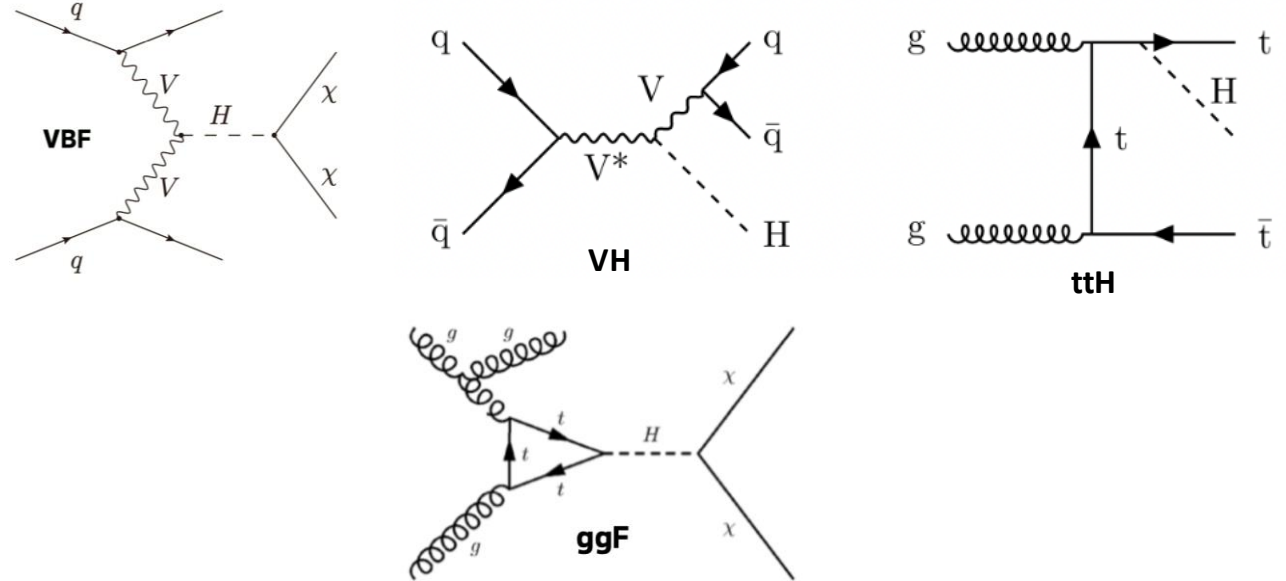


Сравнение с результатами CMS



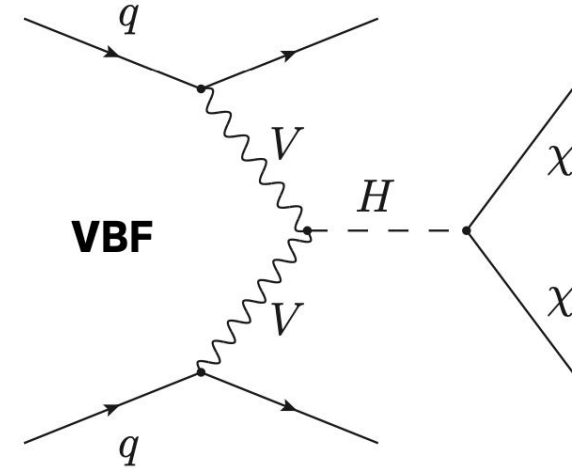
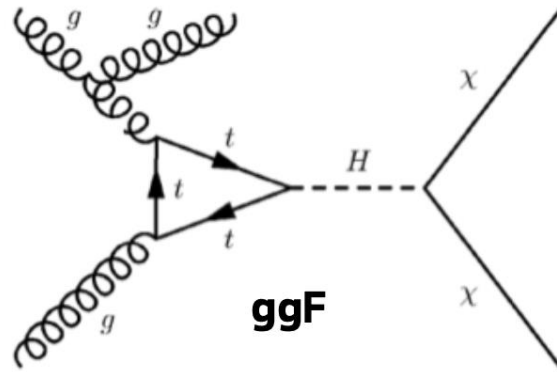
- Полученные в эксперименте CMS значения верхнего предела на вероятность $H \rightarrow \text{inv}$ для всех каналов, Run 2 и их комбинации

$B(H \rightarrow \text{inv}) < 15.0$ (8.0) % при 95% CL



Подробнее: [CMS-PAS-HIG-21-007](https://cds.cern.ch/record/2788112/files/CMS-PAS-HIG-21-007.pdf)

почему VBF+MET наиболее чувствительный канал ?

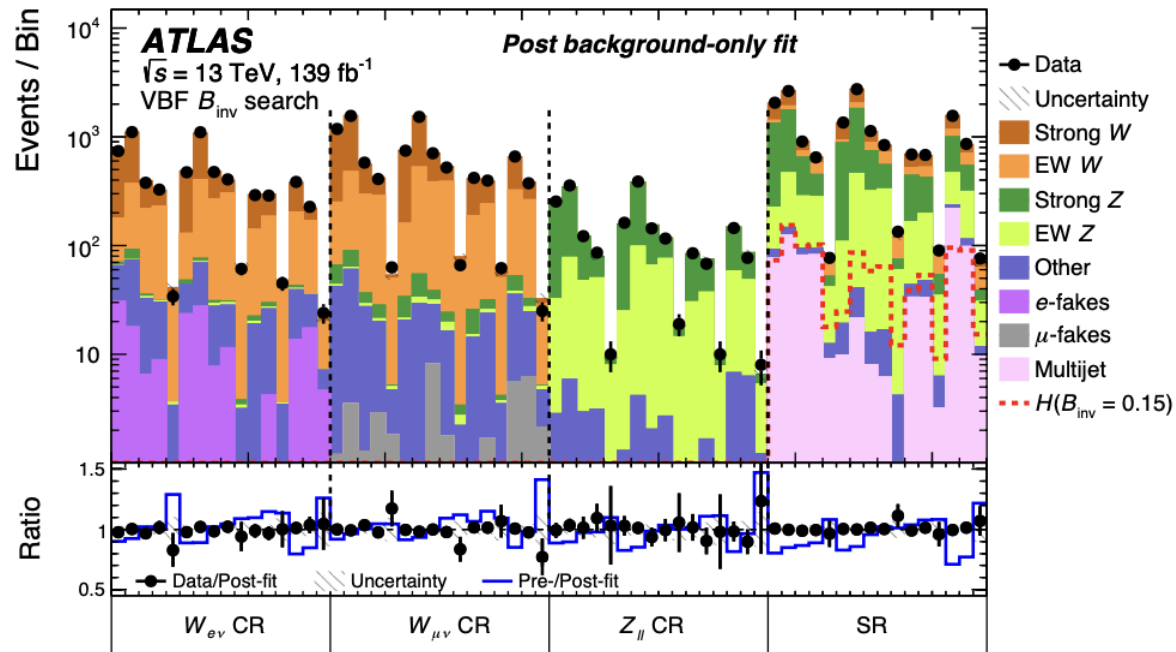


- $Z \rightarrow \nu\nu$ (strong production) is large compared to ggF:
$$\frac{\sigma_{\text{ggF}}(H + \text{jet})}{\sigma(Z + \text{jet}) \cdot \text{Br}(Z \rightarrow \nu\nu)} \sim \frac{1}{300}$$

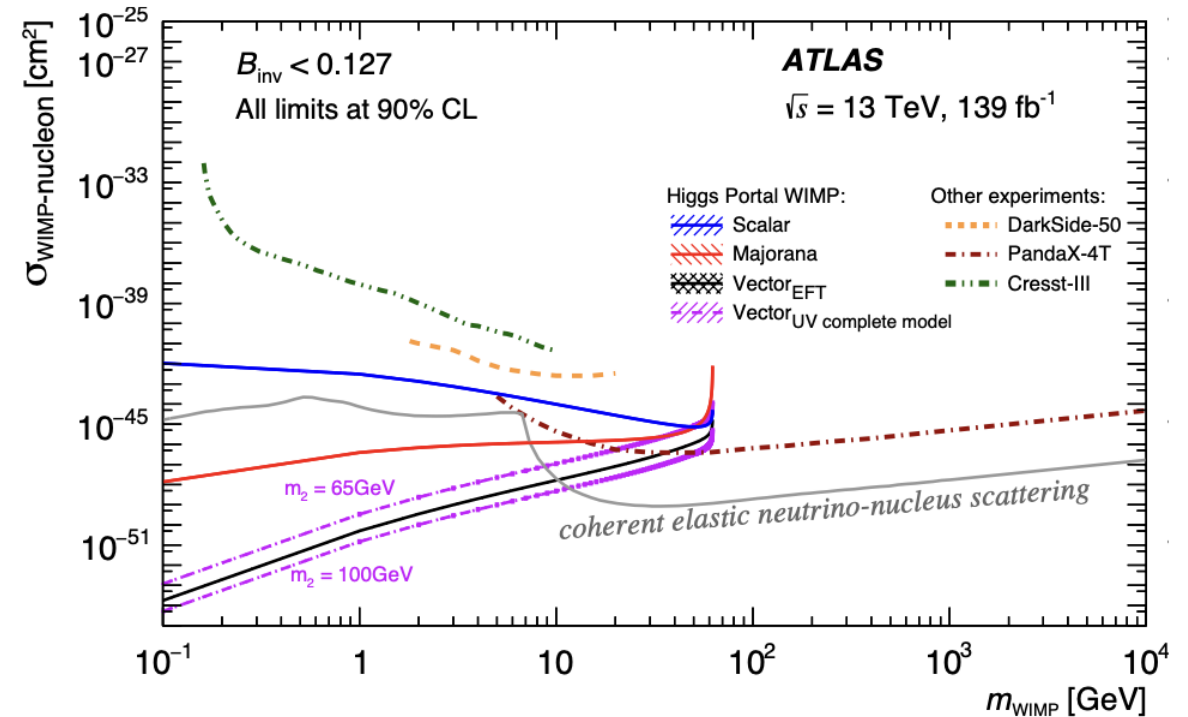
- VBF: signal and bg are comparable
$$\frac{\sigma_{\text{VBF}}(H + qq)}{\sigma_{\text{VBF}}(Z + qq) \cdot \text{Br}(Z \rightarrow \nu\nu)} \sim \frac{4}{3}$$

$H \rightarrow inv$ при рождении в слиянии векторных бозонов (VBF+MET)

- Нет значительного отклонения от прогнозов СМ.
- $BR(H \rightarrow inv)$: 14,5% (наблюдаемый) при 10,3% (ожидаемый)



- SR и все CR области после фитирования и неопределенности

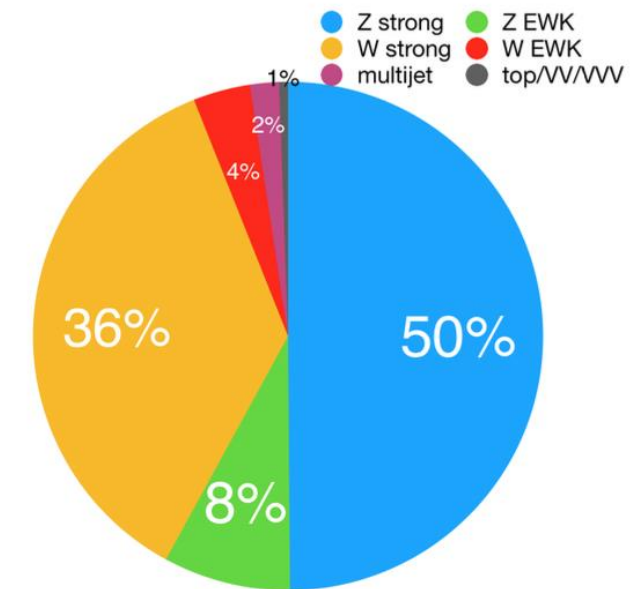


- Верхние пределы спин-независимого поперечного сечения WIMP–нуклона с использованием интерпретаций портала Хиггса B_{inv} при 90% CL в зависимости от m_{WIMP}

Оценка фоновых процессов

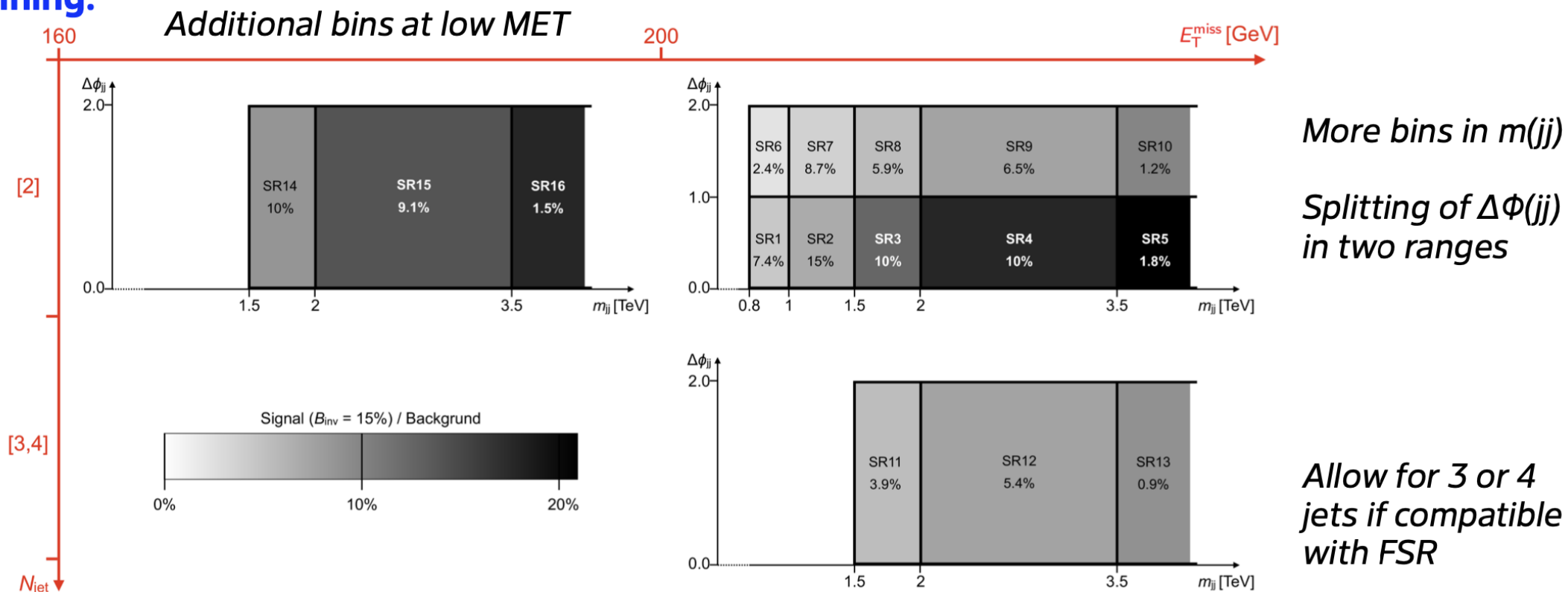
- Хорошее разделение фоновых и сигнальных событий – разделение SR на 16 категорий. CR заданы отдельно для методов оценки фоновых событий.
- V +jets фоновые процессы составляют около 95% фоновых событий в SR, их вклад оценивается с помощью выделенных контрольных областей.
- Меньший вклад обусловлен многоструйными (*Multijet*) процессами и оценивается data-driven методами.
- Незначительный вклад от tt , VV , VVV оценивается с помощью MC моделирования.

N _{jet} =2 & MET>200 GeV				
$\Delta\phi_{jj}<1$ $0.8<m_{jj}<1$ bin1	$\Delta\phi_{jj}<1$ $1<m_{jj}<1.5$ bin2	$\Delta\phi_{jj}<1$ $1.5<m_{jj}<2$ bin3	$\Delta\phi_{jj}<1$ $2<m_{jj}<3.5$ bin4	$\Delta\phi_{jj}<1$ $3.5<m_{jj}$ bin5
$1\leq\Delta\phi_{jj}<2$ $0.8<m_{jj}<1$ bin6	$1\leq\Delta\phi_{jj}<2$ $1<m_{jj}<1.5$ bin7	$1\leq\Delta\phi_{jj}<2$ $1.5<m_{jj}<2$ bin8	$1\leq\Delta\phi_{jj}<2$ $2<m_{jj}<3.5$ bin9	$1\leq\Delta\phi_{jj}<2$ $3.5<m_{jj}$ bin10
N _{jet} >2 MET>200 GeV		N _{jet} >2 $\Delta\phi_{jj}<2$ $1.5<m_{jj}<2$ bin11	N _{jet} >2 $\Delta\phi_{jj}<2$ $2<m_{jj}<3.5$ bin12	N _{jet} >2 $\Delta\phi_{jj}<2$ $3.5<m_{jj}$ bin13
N _{jet} =2 160<MET<200 GeV		$\Delta\phi_{jj}<2$ $1.5<m_{jj}<2$ bin14	$\Delta\phi_{jj}<2$ $2<m_{jj}<3.5$ bin15	$\Delta\phi_{jj}<2$ $3.5<m_{jj}$ bin16



36 fb⁻¹ result: only three $m(jj)$ bins (1–1.5, 1.5–2, >2 TeV); MET > 180 GeV (no bins with 3 or 4 jet)

New binning:

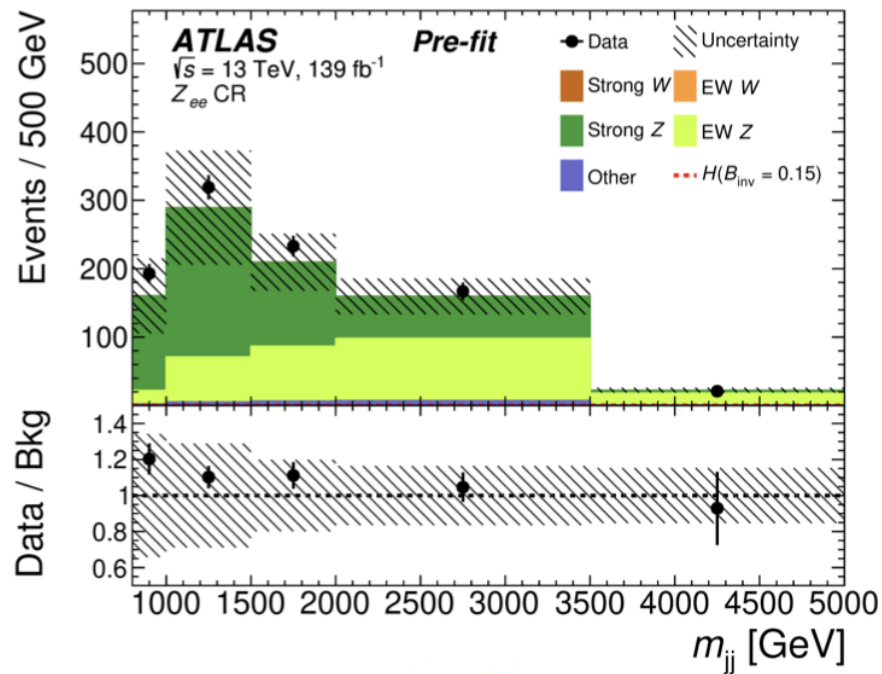


Shading indicates S/B , assuming BR of Higgs boson to invisible particles of 0.15

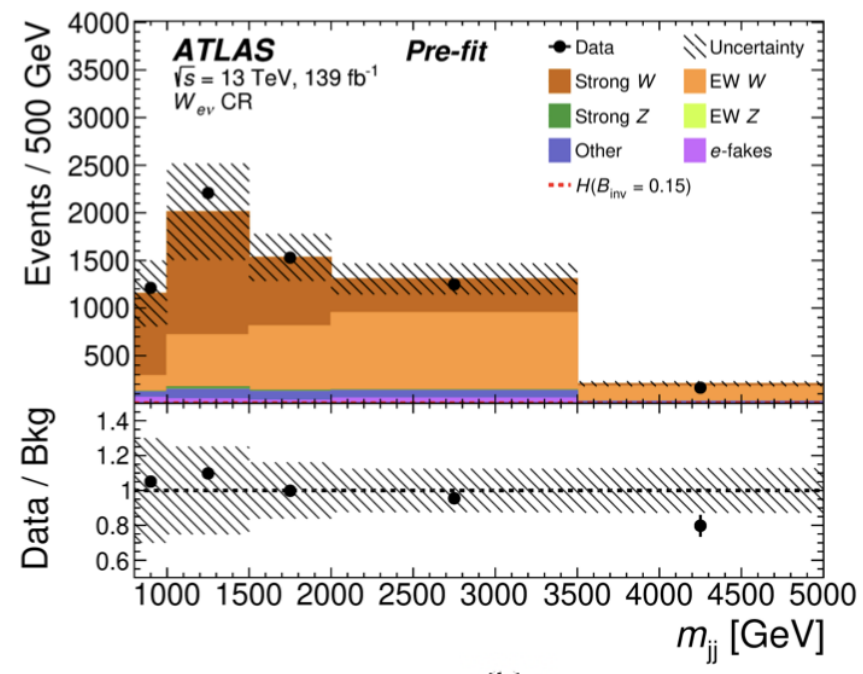
Percentage gives approximate contribution of signal from invisibly decaying Higgs bosons relative to total signal yield in inclusive SR

$V + jets$ фоновый процессы

- Вклад $V+jets$ фоновых событий оценивается с помощью метода «контрольных областей» содержащих выбранные $W (\rightarrow lv) + jets$ (1L) и $Z (\rightarrow ll) + jets$ (2L) события



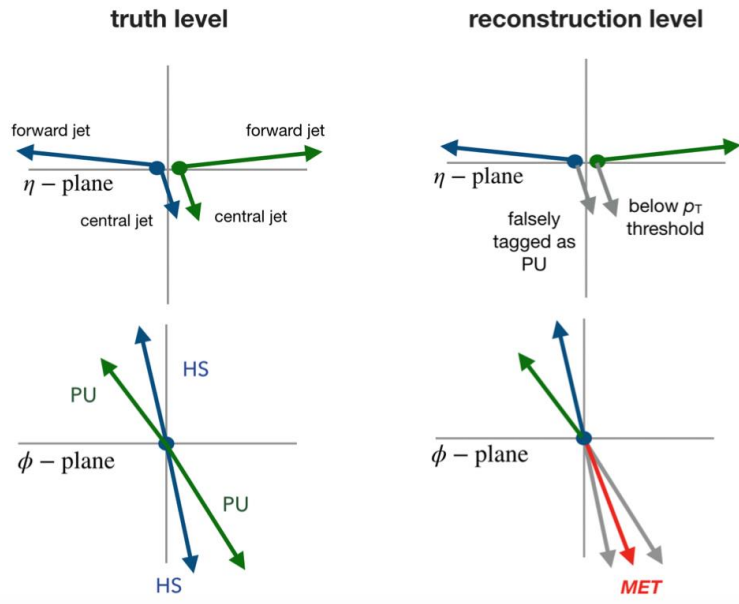
(a)



(b)

Полученные в Run2 распределения по m_{jj} для (a) $W (\rightarrow lv)$ и (б) $Z (\rightarrow ll)$ контрольных областей

Multijet фоновый процесс



При наличии большого количества перекрывающихся событий (Pile-up – PU) неправильное определение струй может привести к симуляции сигнального события фоновым.

- Вклад КХД многоструйного фонового процесса в данном анализе ожидается небольшой за счет критериев отбора событий.
- Но оценку данного фонового процесса получить не просто, так как помимо недостаточной статистики, и из-за ошибок при реконструкции струй и неточной идентификации сигнальных событий, MC samples имеют большие неопределенности измерения.

В Run2 многоструйный фон оценивался с помощью комбинации двух независимых data-driven методов:

- «**Rebalance and Smear**» (R+S)
- «**Pileup/FJVT CR**» метода

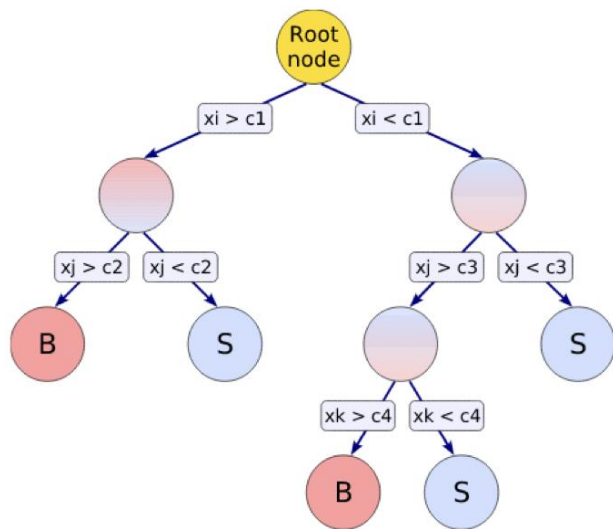
Планы Run3 $H \rightarrow \text{inv}$ поиска в эксперименте АТЛАС

- В Run3 энергия центра масс достигла 13.6 ТэВ $\rightarrow$ увеличение статистики
- Исследования в процессах, которые статистически ограничены (такие как VBF+MET, monoZ, tt+MET и VBF+MET+gamma), могут достигнуть большую чувствительность
- В некоторых процессах, рассмотренных в Run2 продолжают исследования, с использованием частичных данных Run3

Что такое BDT?

Boosted Decision Tree (BDT) – многопараметрический метод анализа для оптимизации сигнальной области в $H \rightarrow inv$

- Оптимизация и определение выполняются с помощью МК с использованием наилучших доступных оценок.
- Исключены сильно коррелированные переменные.



Input variable	Definition
Z rapidity	
$\Delta R(\ell\ell)$	
$\Delta\phi(\vec{E}_T^{\text{miss}}, \vec{p}_T^Z)$	
E_T^{miss} -significance	
fractional p_T	$ \vec{E}_T^{\text{miss}} + \sum \vec{p}_T^{\text{jets}} - p_T^Z /p_T^Z$
H_T	scalar sum of hard objects p_T
$\frac{E_T^{\text{miss}}}{H_T}$	
$m_{\ell\ell}$	

переменные входящие в состав BDT
для H(inv) mono-Z анализа

- Дерево решений в этом случае работает как двоичный классификатор, в котором набор данных последовательно разбивается в попытке максимально разделить полученные данные на один из двух классов, в нашем случае 'Сигнал' или "Фон".
- В каждой точке ветвления ("узле") в дереве классификатор определяет, путем минимизации заранее определенной функции потерь/критерия Джини, оптимальные переменные для отборов. Ветви дерева создаются в соответствии с тем, лежат ли события выше или ниже этого значения отбора. Это выполняется итеративно до достижения заданной глубины дерева или до тех пор, пока набор данных не будет полностью отнесен к одному из классов.

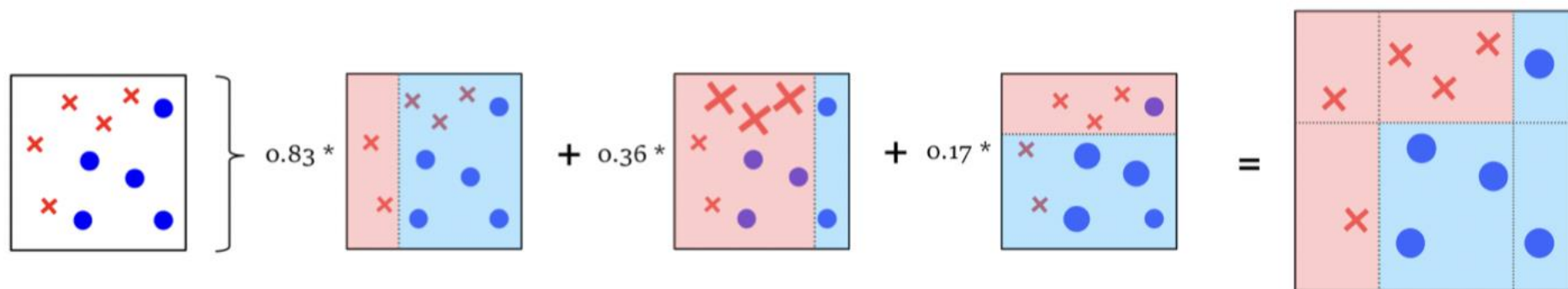
- В TMVA для BDT variable importance основана на том, насколько использование данной переменной способствует улучшению разделения в узлах деревьев. Затем вклад суммируется по деревьям и нормируется. В результате получаются относительные значения importance.

Rank	: Variable	: Variable Importance

1	: feta12	: 1.311e-01
2	: m2_rel	: 1.283e-01
3	: met_tst_j2_dphi	: 1.135e-01
4	: met_tst_j1_dphi	: 1.128e-01
5	: feta13	: 1.108e-01
6	: m3_rel	: 1.093e-01
7	: jj_deta	: 1.080e-01
8	: met_tst_et	: 9.929e-02
9	: jet_pt[1]	: 8.693e-02

Boosting включает в себя объединение нескольких деревьев решений (порядка сотен) с помощью эффективного средневзвешенного значения результатов классификации.

Наиболее распространенные алгоритмы boosting генерируют деревья решений последовательно, учитывая ошибки, допущенные предыдущим классификатором, как показано на рисунке. Комбинируя множество слабых классификаторов, таким образом, можно получить более сильный классификатор, а именно Усиленное дерево решений (BDT).



Набор данных классифицируется в фазовом пространстве с двумя переменными с помощью нескольких последовательных деревьев решений, каждое из которых учитывает ошибки предыдущего классификатора при выполнении своей собственной классификации. Более крупные точки данных представляют больший вес, придаваемый соответствующим событиям при классификации.

Объединив все слабые классификаторы, мы получаем более сильный классификатор.

Adaptive Boosting aka Adaboost	Gradient Boosting
Both Adaboost (the first boosting algorithm) and Gradient Boosting are boosting algorithms, which combines predictions from multiple weak learners, usually decision stumps to form a strong learner	
In AdaBoost, the weights of the samples are adjusted at each iteration.	No reweighting of the samples take place in GBM
Algorithm: Training process starts with a decision stump (usually). At every step, the weights of the training samples which are misclassified are increased for the next iteration. The next tree is built sequentially on the same training data but using the newly weighted training samples. This process is repeated until a desired performance is achieved.	Algorithm: GBM uses gradient descent to iteratively fit new weak learners to the residuals of the previous ones, minimizing a loss function. There are several loss functions to choose from, Mean Squared Error being most common for Regression and Cross Entropy for Classification. GBM uses Decision Trees as the weak learners.
Both Adaboost and GBM are stage-wise additive models (greedy algorithm), meaning new trees in the model are built without changing the previous existing trees	
The final model is formed by combining the predictions from individual trees through a weighted sum. For a classification problem, prediction is given by: $Prediction = sign(\sum_{m=1}^M \alpha_m * F_m(x))$ where, $F_m(x)$ is the output of each model and α_m are the weights computed by the boosting algorithm, m is the number of iterations	The final model is an equal-weighted sum of all of the individual trees. Prediction for a regression problem in GBM is given by: $Prediction = \hat{y} + \eta * \sum_{m=2}^M \hat{r}_{m-1}$ where, $\hat{y}$ is the prediction from the first tree, η is the learning rate, $\hat{r}_i$ is the prediction of residuals, m is the no. of iterations
Both Adaboost and Gradient boosting can be used for both Classification and Regression problem	

© AIML.com Research

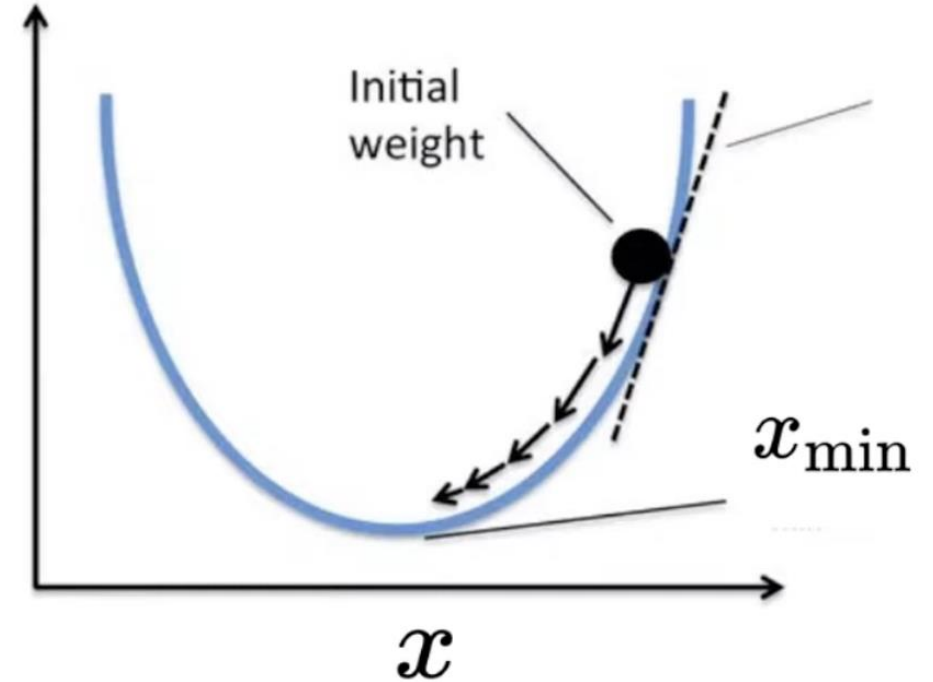
Напоминание: градиентный спуск

- Минимизируем функцию $f(x)$
- Выбираем x_0 — начальное приближение
- Пусть x_n — текущая найденная точка
- Вычисляем градиент (производную) $f'(x_n)$
- Вычисляем следующее приближение:

$$x_{n+1} = x_n - \alpha f'(x_n)$$

α — скорость сходимости к точке минимума (*learning rate*)

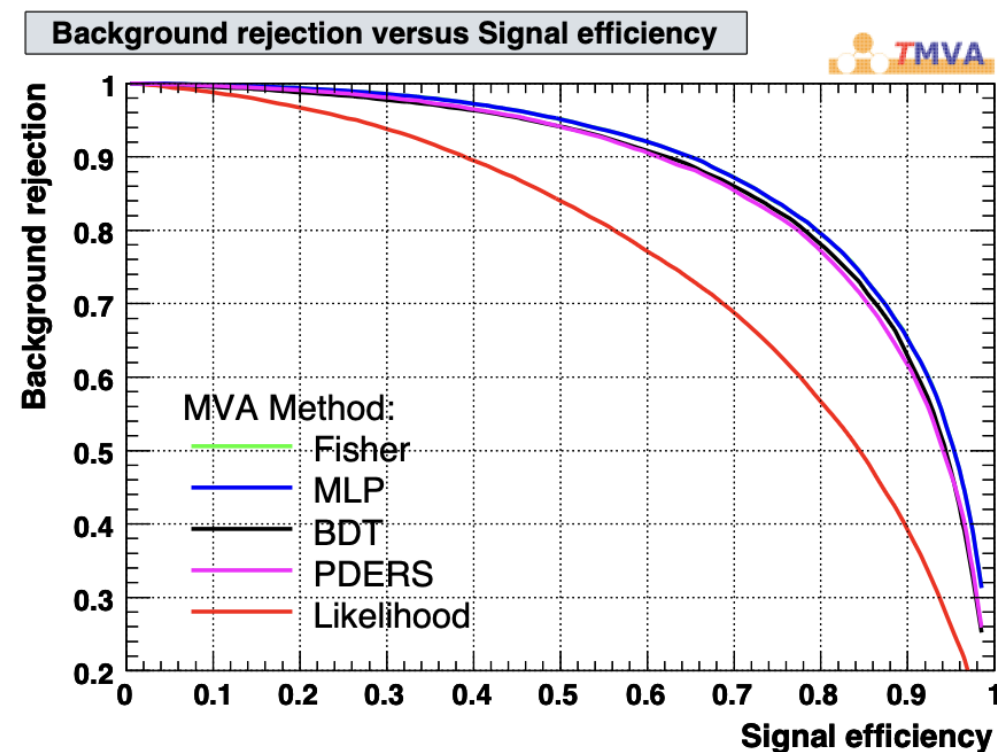
При должном выборе шага α алгоритм быстро сходится к точке локального минимума



- **ROC-кривая (Receiver Operating Characteristic)** — это графический инструмент для оценки качества моделей бинарной классификации, который показывает соотношение между долей истинно положительных срабатываний (TPR) и долей ложноположительных срабатываний (FPR) при изменении порога классификации
- Здесь происходит сравнение background rejection и signal efficiency при разных значениях BDT score

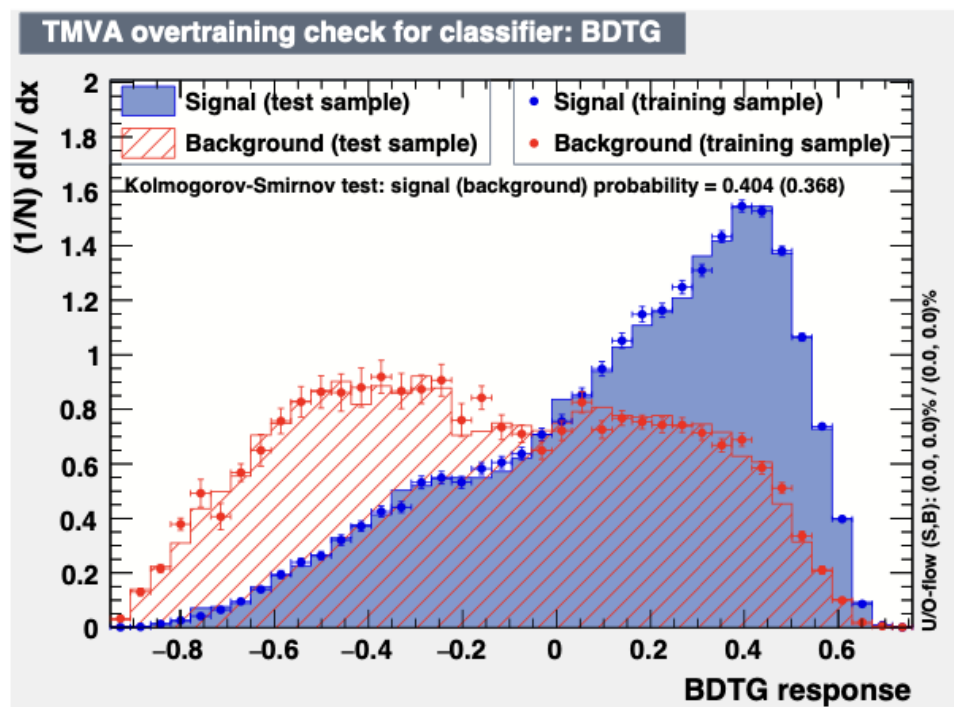
Для числовой оценки качества модели используют метрику **AUC** — площадь под ROC-кривой (это тоже самое что Roc-int).

Она показывает способность модели разделять классы и чем это значение ближе к 1, тем модель лучше.



Полученная в результате величина представляет собой Дискриминационную переменную BDT, которая также называется «BDT score».

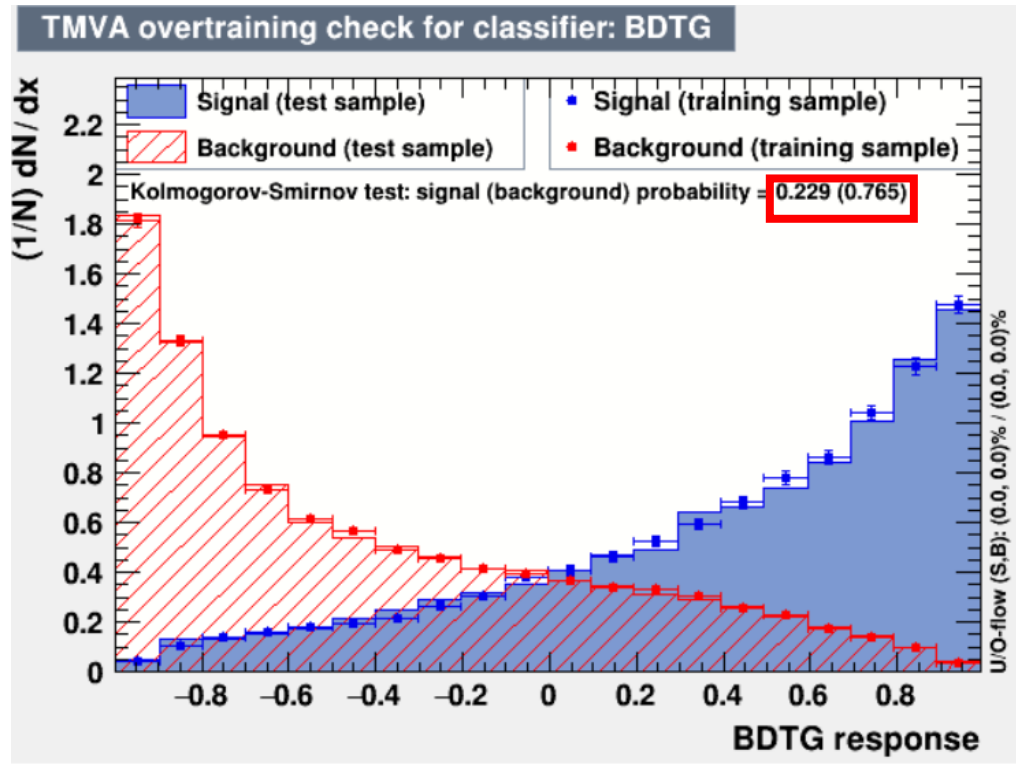
Этот показатель, измеряемый от -1,0 до 1,0, является вероятностной мерой того, что событие является сигнальным (выход = 1,0) или фоновым (выход = -1,0).



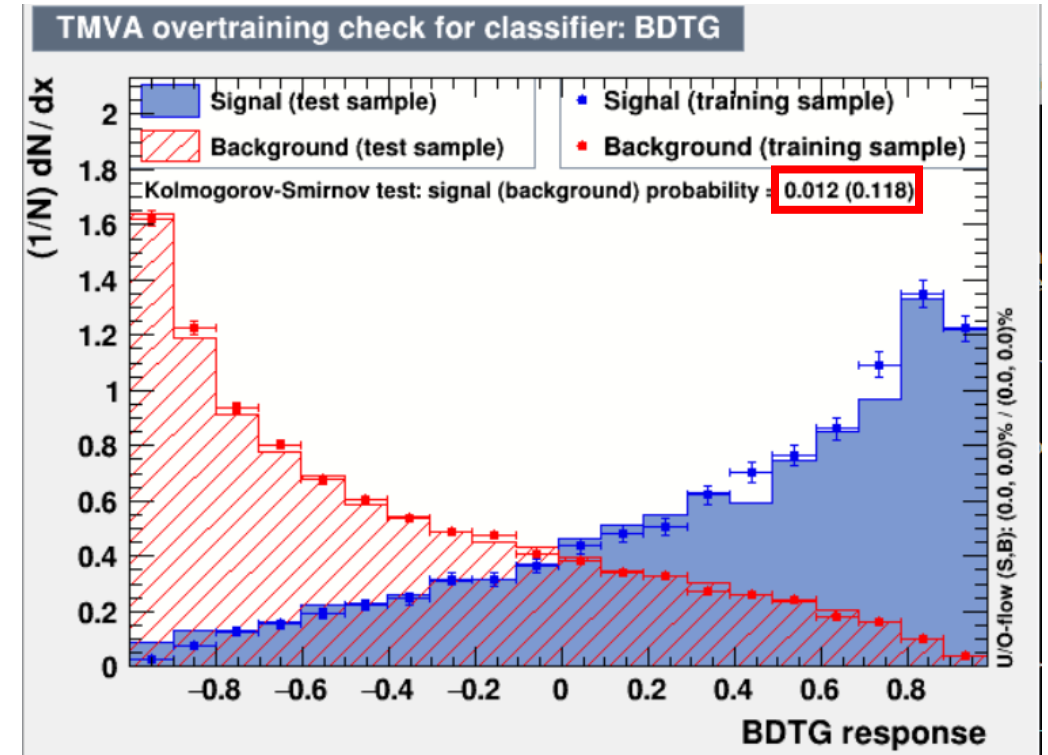
Таким образом в результате BDT , идеальный классификатор классифицировал бы все сигнальные события как смещенные вправо, а все фоновые события (показаны смещенным влево).

распределения BDT для сигнальных и фоновых событий в моно-Z анализе H(inv).

Kolmogorov-Smirnov test



Var 1 : with old variables

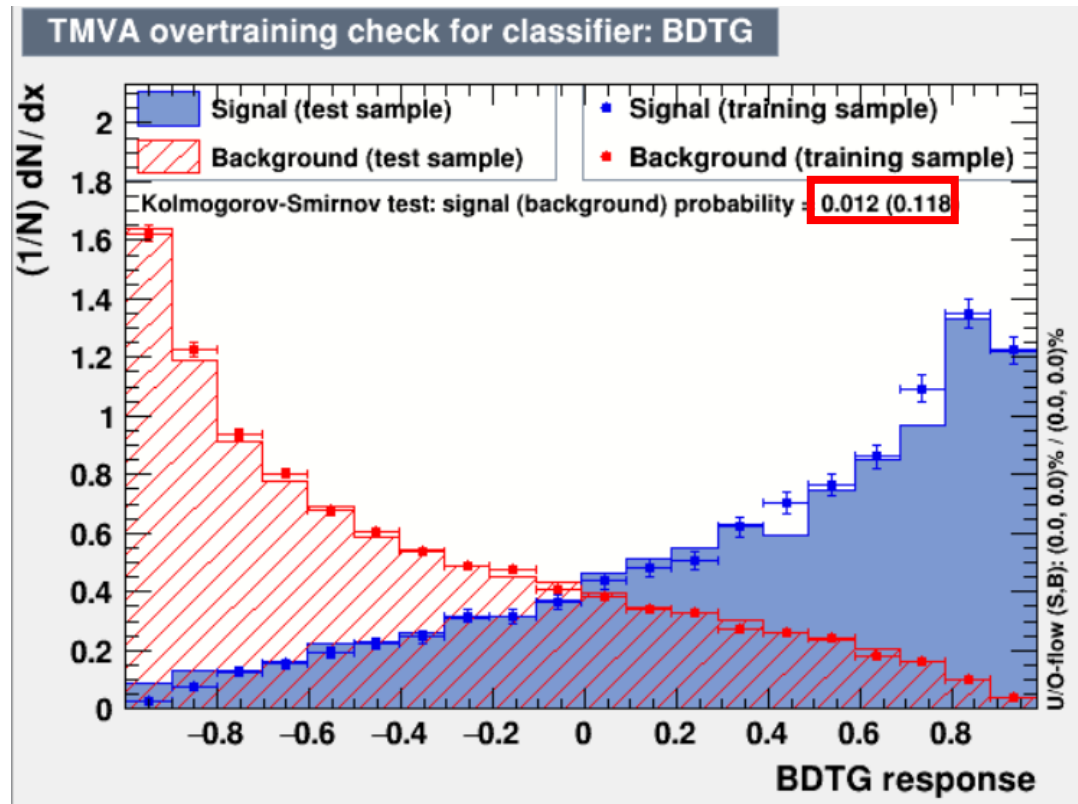


Var 2: old+ new variables

- A Kolmogorov Smirnov test is applied between the training and the test output distribution.
- In case the two distributions are compatible a random value between 0 and 1 should be obtained, if they are not you will observe a difference and this might be an indication of overtraining.
- So a very small value is not good.
- A value which is not small is good, while a value too close to 1 is also not so good, because indicates that the statistical fluctuations are too small, and this can be also an indication of some potential problems.

Kolmogorov-Smirnov test

- Тест Колмогорова-Смирнова (KS-тест) сравнивает распределение полученное для тестовой и тренировочной выборки и проверяет, одинаково ли BDTG ведёт себя на данных, на которых обучался, и на независимых данных.



KS probability	Интерпретация
$p_{KS} \sim 0.5$	очень хорошо
$p_{KS} \sim 0.1$	обычно нормально
$p_{KS} \sim 0.05$	стоит внимательно проверить
$p_{KS} \ll 0.05$	перетренировка

Применение оценки классификатора для верхних пределов на вероятность распада

Фитирование - это процесс подгонки теоретической модели под реальные экспериментальные данные методом максимального правдоподобия (Maximum Likelihood Fit).

- Строится глобальная математическая функция правдоподобия, которая представляет собой произведение распределений Пуассона по всем бином (интервалам распределения) во всех SR и CR областях

Main likelihood (LL) or fitting model:

$$\mathcal{L}(\mu, \vec{\beta}_V, \vec{n}_{\text{PU-MJ/fake-}e/\text{fake-}\mu}, \vec{\theta}) = \mathcal{L}^{\text{SR}} \cdot \mathcal{L}^{\text{PU-CR}} \cdot \mathcal{L}^{\text{V+jets-CR}} \cdot \mathcal{L}^{\text{fake-CR}} \cdot \mathcal{L}^{\text{NP}},$$

Likelihood for Signal region (SR), for other fit regions can be found in [run 2 paper](#):

$$\mathcal{L}^{\text{SR}} = \prod_i \mathcal{P} \left(N_i^{\text{SR}} \mid \beta_i \cdot B_{Z,i}^{\text{SR}} + \beta_i \cdot B_{W,i}^{\text{SR}} + B_{\text{MJ},i}^{\text{SR}} + B_{\text{other},i}^{\text{SR}} + \mu \cdot S_i^{\text{SR}} \right),$$

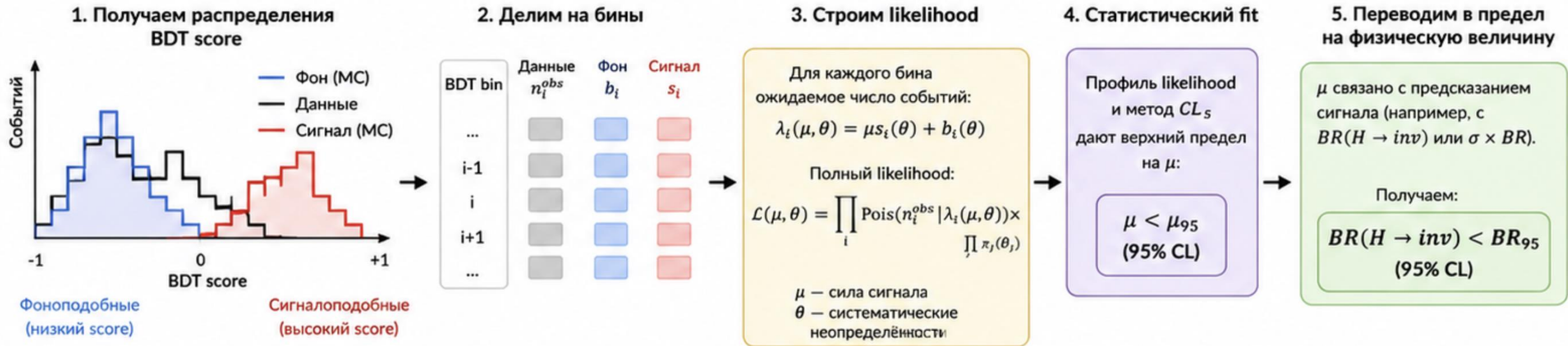
Likelihood for nuisance parameter (modeling systematic sources):

$$\mathcal{L}^{\text{NP}} = \prod_j \mathcal{G} (0 \mid \theta_j),$$

пакет HistFactory начинает варьировать μ и все параметры θ одновременно, чтобы максимизировать функцию правдоподобия и найти $\hat{\mu}$, $\hat{\theta}$

Если после фитирования значение $\hat{\mu}$ оказалось близким к нулю (то есть явного избытка событий сигнала над фоном в SR не обнаружено), мы устанавливаем верхний предел (upper limit) на BR

Применение оценки классификатора для верхних пределов на вероятность распада



in progress