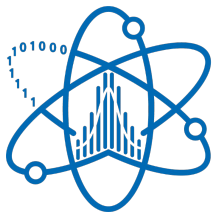




National Research
**Tomsk
State
University**



**Лаборатория
анализа данных
физики высоких энергий**

Томского
государственного
университета

Физический анализ данных

Томский Государственный Университет

19.01.2026

Направления работы

Повторение анализа ATLAS Open Data

Воспроизведение измерения сечений $t\bar{t}$ и Z на данных 2015 года ($\sqrt{s} = 13$ ТэВ)

Цель: валидация цепочки анализа и инфраструктуры.

✓ Верификация методов анализа

Пример: WVZ в pp -столкновениях

Цель: отработка методологии повышения точности.

✓ Алгоритмы машинного обучения

Пример: tW при 13 ТэВ, ATLAS

Цель: сравнение BDT, NN, Transformers, GNN и др. по метрикам качества классификации.

BBCI метод

BBCI (*Bias-Corrected Bootstrap for Confidence Intervals*) — модифицированный бутстрэп-метод, который улучшает оценку доверительных интервалов параметров за счёт коррекции смещения и нестабильности, возникающих при малых выборках.

Применение к результатам анализа **WVZ** позволяет проверить, насколько устойчивы оценки силы сигнала и систематических параметров по сравнению с классическим подходом.

Цель:

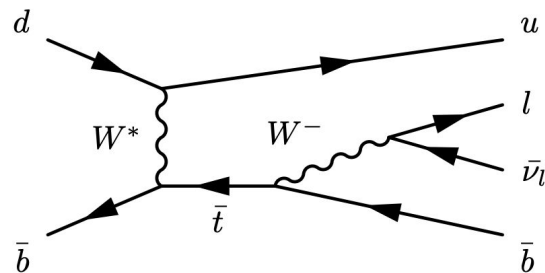
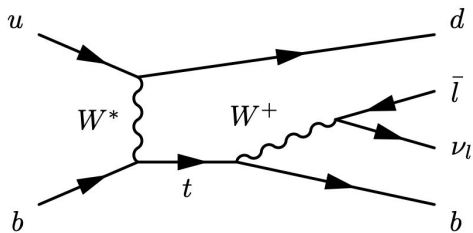
Проверить статистическую устойчивость и надёжность измерения силы сигнала (μ) в условиях ограниченной статистики и исследование влияния систематик на силу сигнала.

Идея:

Сгенерировать множество бутстрэп-реplik данных (случайная повторная выборка с заменой) и провести независимую аппроксимацию для каждой из них, чтобы получить распределение оценок μ и его дисперсию.

BCCI: t-channel production of single top quarks and antiquarks

- **Анализ WVZ:** первичная валидация bootstrap → корректное распределение μ
- **Метод на открытых данных HEPData** - “Measurement of t-channel production of single top quarks and antiquarks”
- В данном анализе используется 140 fb^{-1} данных эксперимента ATLAS при энергии столкновений 13 ТэВ для измерения сечений рождения процессов tW
 - Разделение на два канала обеспечивают чувствительность к PDF u- и d-кварков, поскольку доминирующие начальные состояния различаются для tW^+ (переход $u \rightarrow d$) и tW^- (переход $d \rightarrow u$).
- Для разделения сигнала tW и фона обучается нейронная сеть (NN), использующая кинематические переменные на уровне событий.
- **NN output** затем используется в качестве дискриминанта в **profile likelihood fit**.



Region strategy

Две комплементарные сигнальные области, оптимизированные под угловую корреляцию между лептоном и b-джетом:

- **SR-plus (SRp)**: события, в которых лептон и b-джет преимущественно сонаправлены (чувствительно к рождению tW^*).
- **SR-minus (SRn)**: события с противоположной угловой корреляцией (чувствительно к рождению tW^-).

Такое разделение повышает чувствительность к PDF и улучшает ограничения на модель сигнала.

Основные фоновые процессы

- **tt̄ (топ–анти топ)**: доминирующий фон в одно-лептонных конечных состояниях с b-джетами.
- **Одиночный топ (tW, t-канал)**: важный вклад, требующий точного моделирования, особенно в категориях с одним b-тегом.
- **W+jets**: критичен для областей с одним лептоном и MET.
- **Z+jets / дибозонные процессы**
- **Фейк лептоны / ошибочная идентификация заряда**: включаются там, где это необходимо.

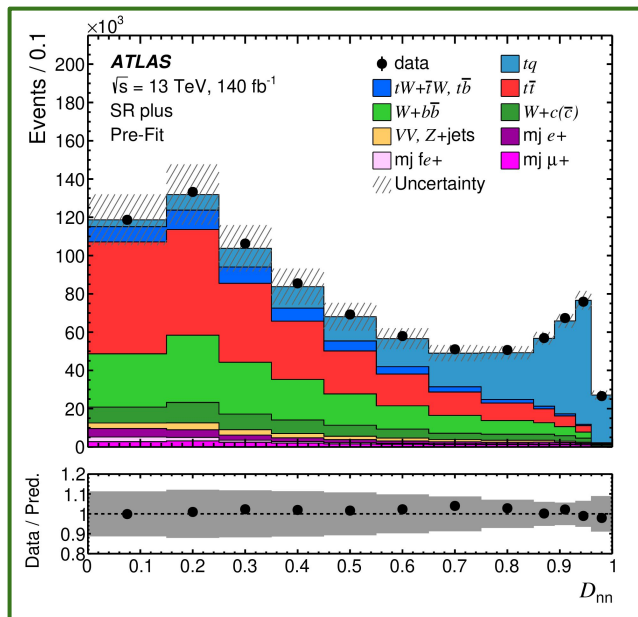
CR name	Requirement
B-e-plus	$q_e/e = +1, \eta(e) < 1.37, E_T^{\text{miss}} < 30 \text{ GeV}$
B-e-minus	$q_e/e = -1, \eta(e) < 1.37, E_T^{\text{miss}} < 30 \text{ GeV}$
EC-e-plus	$q_e/e = +1, \eta(e) > 1.52, E_T^{\text{miss}} < 30 \text{ GeV}$
EC-e-minus	$q_e/e = -1, \eta(e) > 1.52, E_T^{\text{miss}} < 30 \text{ GeV}$
CR μ -plus	$q_\mu/e = +1, 28 \text{ GeV} < p_T(\mu) < 40 \text{ GeV} \cdot \frac{ \Delta\phi(j_1, \ell) }{\pi}$
CR μ -minus	$q_\mu/e = -1, 28 \text{ GeV} < p_T(\mu) < 40 \text{ GeV} \cdot \frac{ \Delta\phi(j_1, \ell) }{\pi}$

Table 1: Summary of the definition of the CRs.

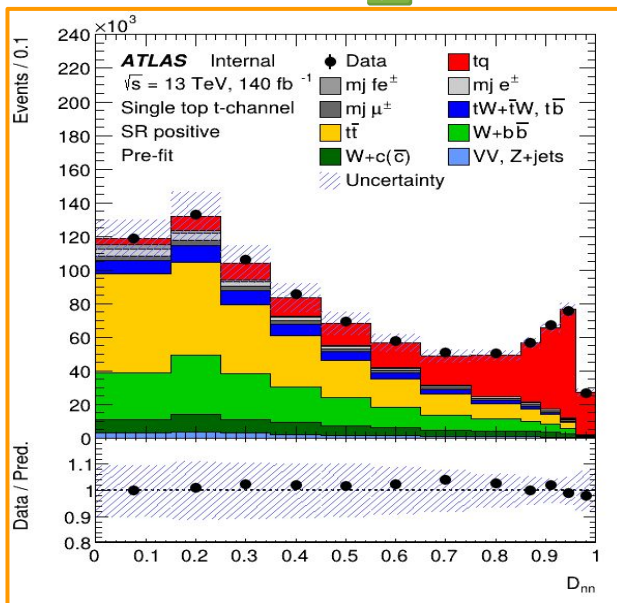
6 контрольных региона

- **Метод на открытых данных HEPData** - “Measurement of t -channel production of single top quarks and antiquarks”
 - **Реализованы два подхода:** прямой фит через Trefitter (воспроизведение результатов статьи) и bootstrap-реализация в PyHF

Reference Article



Results ✓



HEPData Record: [ins2764820](#)

[workspace.json](#) is a JSON specification of the statistical model.

8 channels: signal and control regions:
SR_p, **SR_n**, SR_{lelep}, SR_{lelepforw},
 SR_{muonp}, SR_{lelep}, SR_{lelepforw}, SR_{muonn}

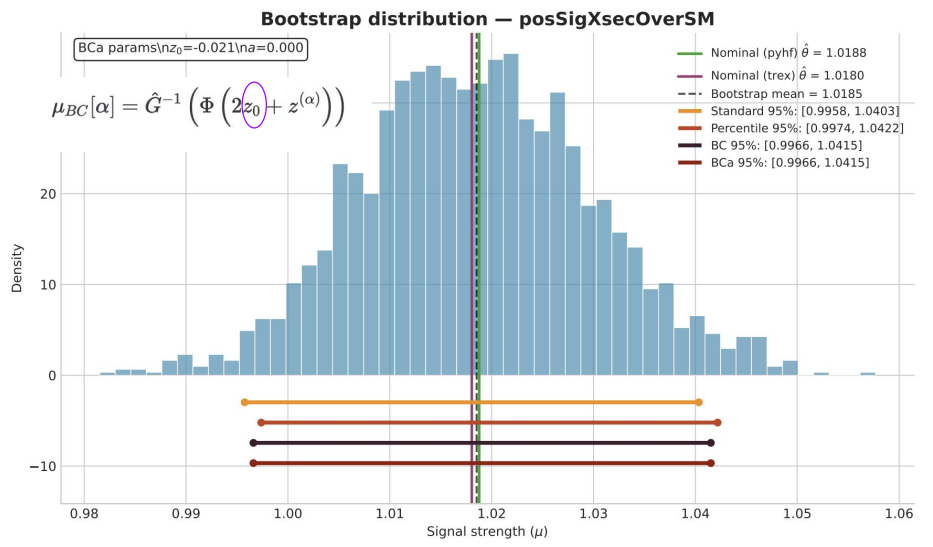
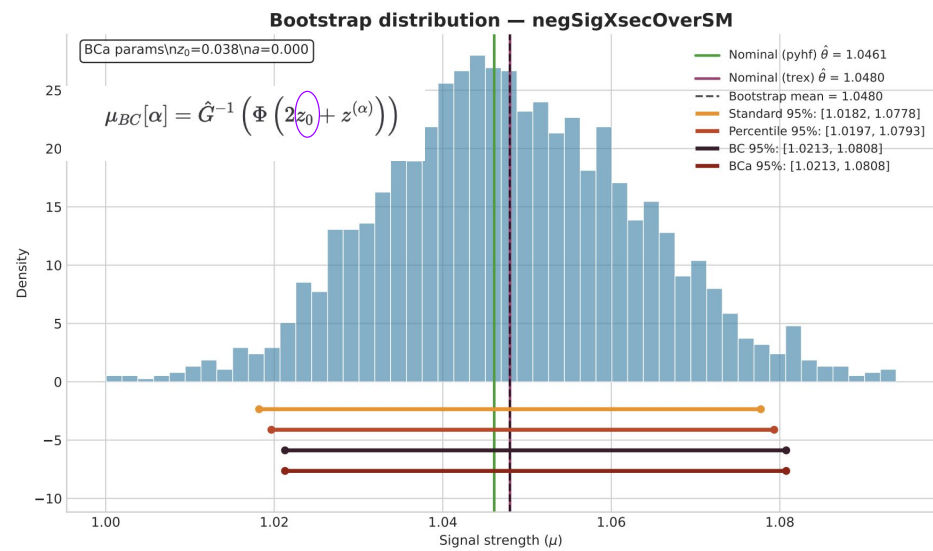
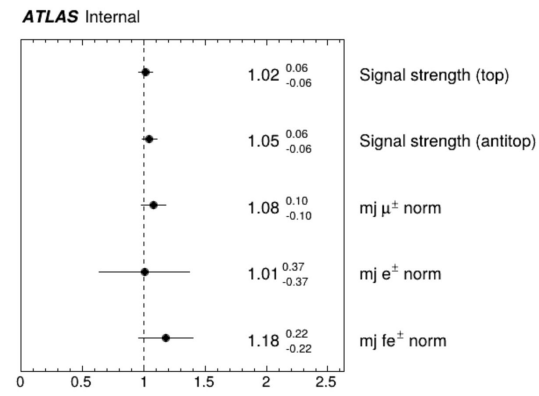
44 bins distributed across the 8 channels

Observed data provided per bin

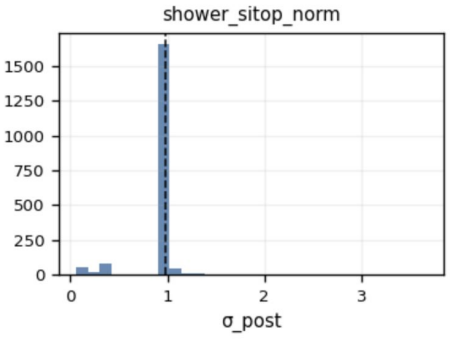
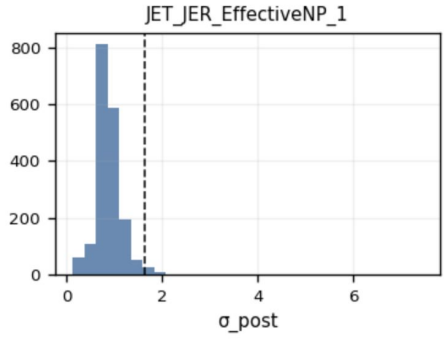
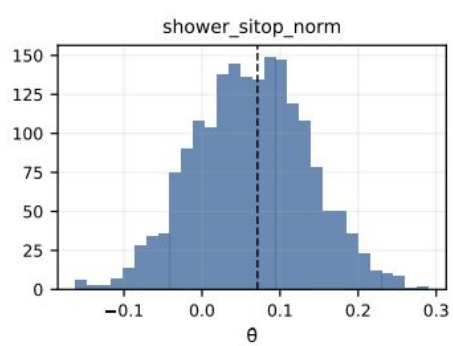
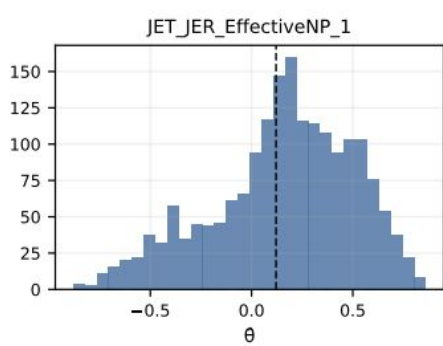
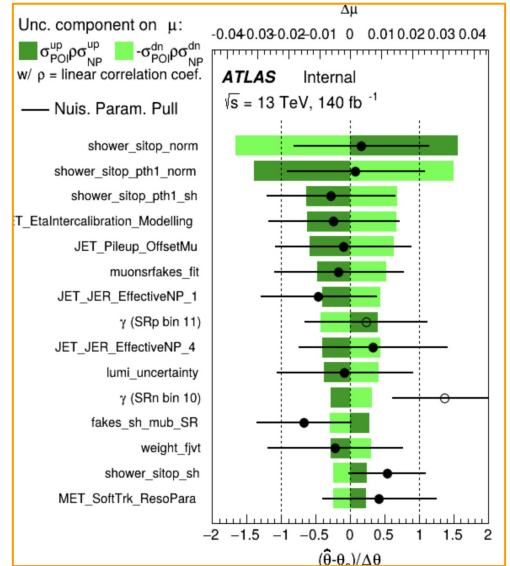
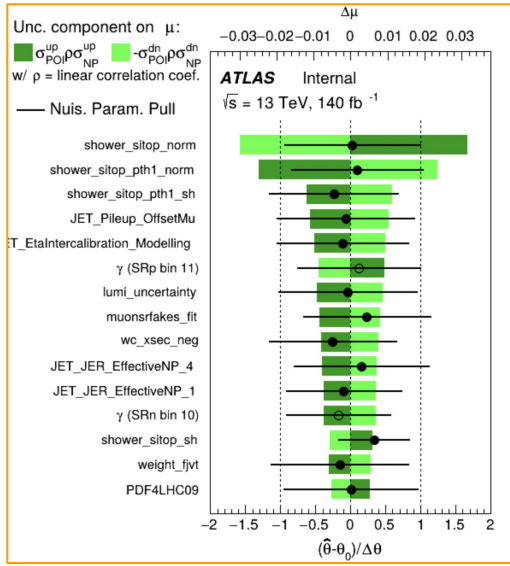
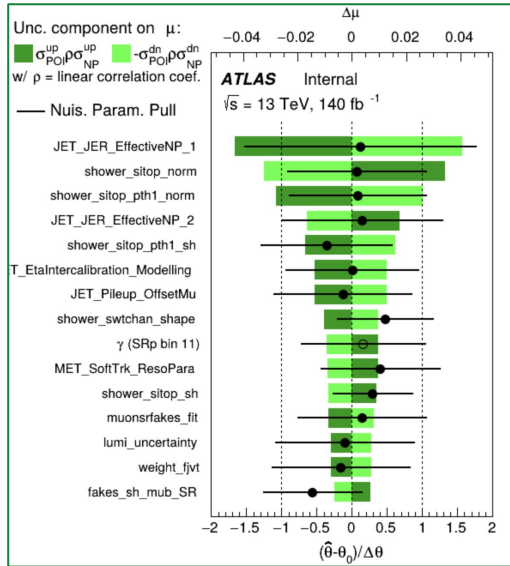
Expected model: signal + background + systematics (400+ nuisance parameters)

Parameter of Interest (POI):
negSigXsecOverSM

- 2000 bootstrap-реплик (Poisson), 100% сходимость
- Доверительные интервалы (μ): построены BC/BCa интервалы→ устойчивые CI при отсутствии асимптотических ошибок



Изучение систематик: preliminary results



Сравнение методов ML на HEP-данных

Цель:

Оценить эффективность современных алгоритмов машинного обучения для задач классификации в физике высоких энергий (HEP). Особое внимание уделено сравнению традиционных методов (BDT, Random Forest) и нейросетевых моделей.

Данные:

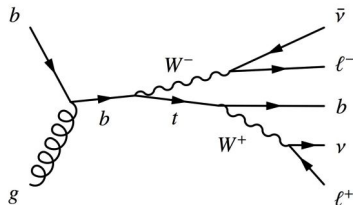
- [ATLAS Open Data \(2015–2016\)](#) — открытые данные эксперимента ATLAS.
- [HEPData Repository](#) — вспомогательные материалы и таблицы с реконструированными событиями.

Физический процесс:

Производство **топ-кварка** в ассоциации с **W-бозоном** при $\sqrt{s} = 13$ ТэВ (анализ ATLAS).

Требуется:

- Ровно **один b-tag джет**,
- И **два лептона**



Measurement of differential cross-sections of a single top quark produced in association with a W boson at $\sqrt{s} = 13$ TeV with ATLAS

The differential cross-section for the production of a W boson in association with a top quark is measured for several particle-level observables. The measurements are performed using 36.1 fb^{-1} of pp collision data collected with the ATLAS detector at the LHC in 2015 and 2016. Differential cross-sections are measured in a fiducial phase space defined by the presence of two charged leptons and exactly one jet matched to a b -hadron, and are normalised with the fiducial cross-section. Results are found to be in good agreement with predictions from several Monte Carlo event generators.

5 December 2017

Contact: [Top conveners](#)

internal

[Eur. Phys. J. C 78 \(2018\) 186](#)

[e-print arXiv:1712.01602](#) - [pdf from arXiv](#) - [Physics Briefing](#)

[Inspire record](#)

[Data points](#)

[Figures](#) | [Tables](#) | [Auxiliary Material](#)

Результаты обучения

- Получены результаты и оптимизированы результаты на увеличенной статистике (увеличена в 3 раза)
- Для каждого метода было определено ограниченное пространство гиперпараметров (например, число деревьев, глубина, скорость обучения для BDTG).
- Обучение и оценка нескольких конфигураций проводились на фиксированном наборе данных, с фиксированным отбором событий и согласованными весами событий.
- Каждая попытка оценивалась по метрике AUC (с дополнительной проверкой по подавлению фона при фиксированной эффективности сигнала).
- После чего лучшая конфигурация переобучалась для получения финальной ROC-кривой и диагностики переобучения.

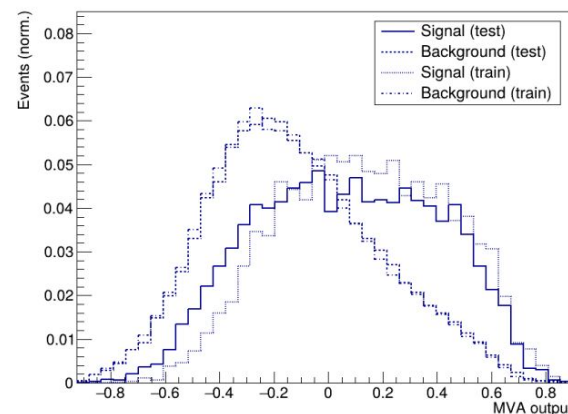
Результаты обучения

- Протестированы 10+ классических supervised ML алгоритмов
- Выявлена проблема с overtraining, начата работа по более аккуратной тренировке
- Идет подготовка сравнения разных реализаций: Python, Root/TMVA, R -> цель сравнить производительность одного и того же класса алгоритмов на разных платформах

Method	Auc_test	Best_balanced_acc	Best_fpr	Best_tpr	Rejb@effs=0.5	Ks_sig(test-vs-train)	Ks_bkg(test-vs-train)
MLP	0.655274	0.623852	0.277297	0.525	0.743824	0.074855	0.616657
Likelihood	0.643988	0.61763	0.35974	0.595	0.72806	1	0.99803
BDTG	0.639764	0.619807	0.295386	0.535	0.736573	3.30E-20	0.000252
HMatrix	0.637081	0.617358	0.360285	0.595	0.72197	0.232496	0.814639
BDT	0.635302	0.613723	0.277555	0.505	0.725977	2.53E-20	0.001526
KNN	0.63272	0.610293	0.334414	0.555	0.717576	8.93E-08	0.046508
Fisher	0.631464	0.605605	0.34379	0.555	0.703661	0.698448	0.992567
PDERS	0.629422	0.608107	0.378786	0.595	0.712034	0.041206	0.09604
LD	0.627842	0.602462	0.360076	0.565	0.695053	0.608545	0.931392
CutsGA	0.614133	0.606092	0.342816	0.555	0.620033		
RuleFit	0.531323	0.53141	0.43218	0.495	0.562337	0.96558	0.999997

Response Test on BDTG from TMVA

Overtraining check - BDTG



.....

Спасибо за внимание

Детали тренировки BDT

.....

Цель:

Разделить процессы **tw** (top+anti-top) и **tf** в области **1j1b** с помощью градиентного BDT.

Используемый инструмент:

Анализ выполнялся в пакете **TMVA (ROOT)** с использованием метода **BDTG** — Gradient Boosted Decision Trees.

Основные настройки и особенности:

- **Feature scaling:** применяется стандартная нормализация **z-score**, чтобы обучение шло на стандартизированных входных данных.
- **Разделение и нормировка:** используется случайное разбиение событий TMVA и нормировка по числу событий.
- **Оценка модели:** для применения обученной модели и построения распределений использовались **TMVA Reader + TTreeReader**.

Результаты обучения BDTG

- Основные переменные ($pT(\ell\ell E_T^{miss})$, $\Delta pT(\ell\ell bE_T^{miss})$) демонстрируют близкие значения значимости S.
- Незначительные различия (<10 %) возможно связаны с вариациями весов и параметров обучения.
- Проверка переобучения (overtraining check) показывает совпадение распределений обучающей и тестовой выборок, что подтверждает корректность тренировки модели.

Variable	Reference Article S ($\times 10^{-2}$)	Produced BDT S ($\times 10^{-2}$)
$pT(\ell\ell E_T^{miss})$	4.1	5.38
$\Delta pT(\ell\ell bE_T^{miss})$	2.5	4.19
ΣE_T	2.3	0.24
$\eta(\ell\ell E_T^{miss} b)$	1.3	2.85
$\Delta pT(\ell\ell E_T^{miss})$	1.1	0.93
$pT(\ell\ell b)$	1.0	0.62
$C(\ell\ell)$ (centrality)	0.9	0.46
$m(\ell_2, b)$	0.2	0.24
$m(\ell_1, b)$	0.1	0.08

Сравнение результатов обучения BDT с данными из референсной статьи показывает согласие в пределах статистических флуктуаций.

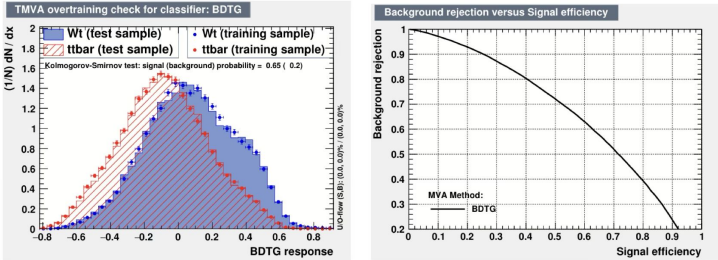
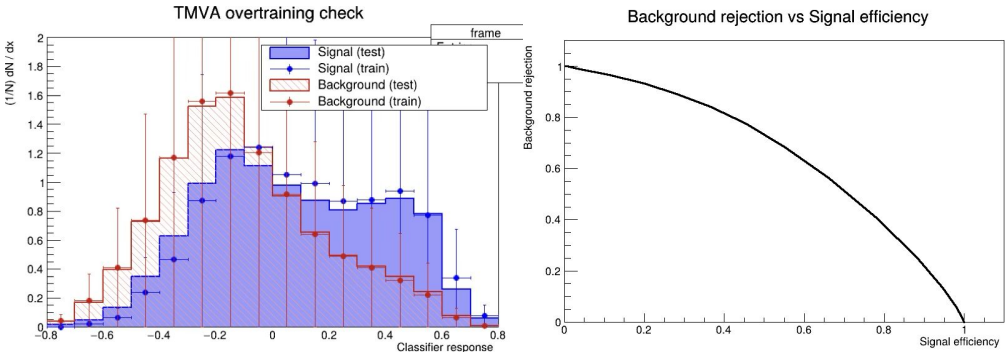
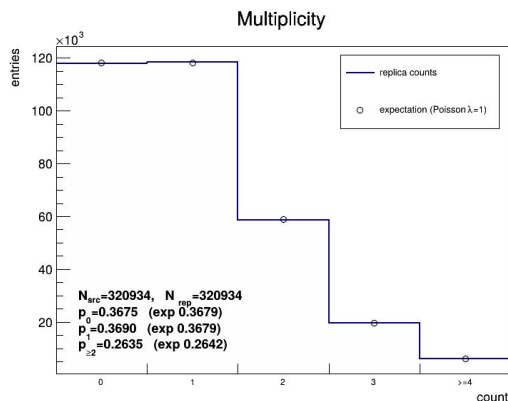


Figure 9: Comparison of test/training sample distributions and background rejection factor versus signal efficiency.

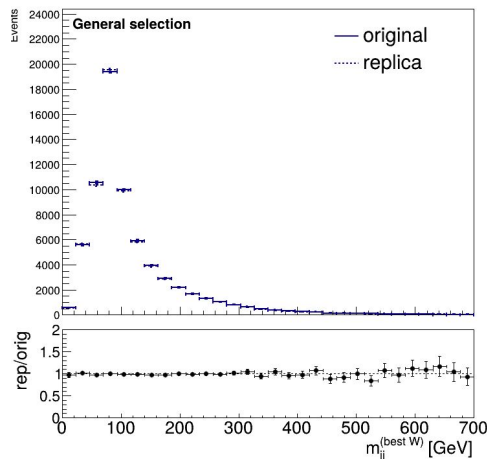


Подготовка реплик

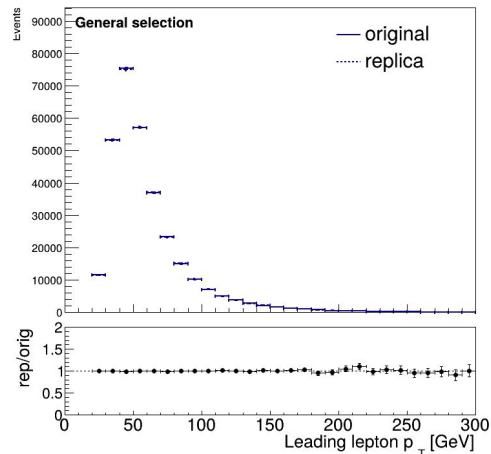
- Сгенерированы **1000 bootstrap-реплик** методом случайной повторной выборки с заменой:
 - для каждой реплики **фиксируется seed**, что обеспечивает **воспроизводимость результатов**.
- Каждая реплика сохраняет **общее количество событий** и **все корреляции между переменными**:
 - реплики формируются до разбиения на SR/CR/VR.



Распределение числа вхождений событий в репликах **согласуется** с теоретическим распределением Пуассона ($\lambda = 1$). Это подтверждает **корректность бутстрэп-процедуры**.

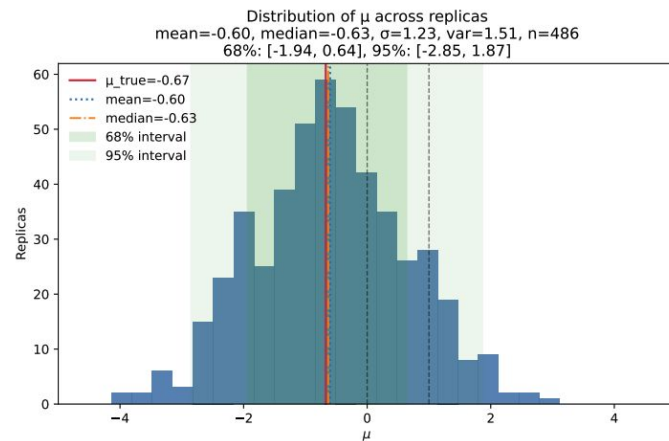
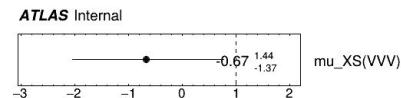
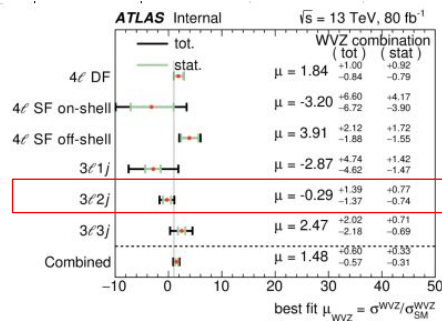
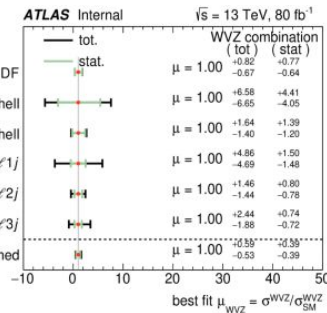
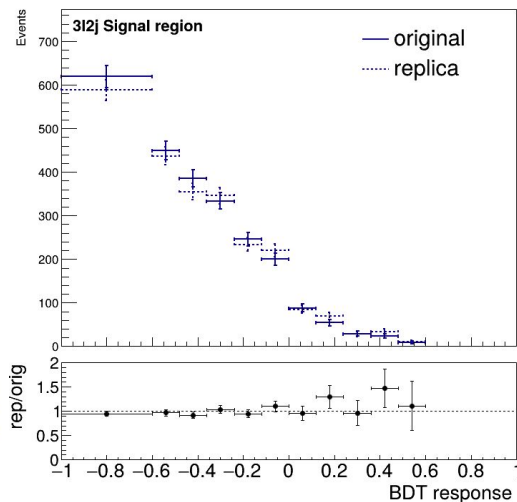
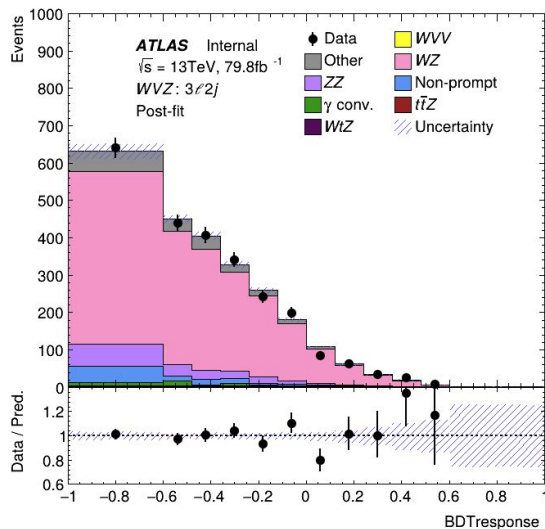


Сравнение форм распределений между оригинальными данными и репликами показывает **согласие** в пределах статистических флуктуаций.



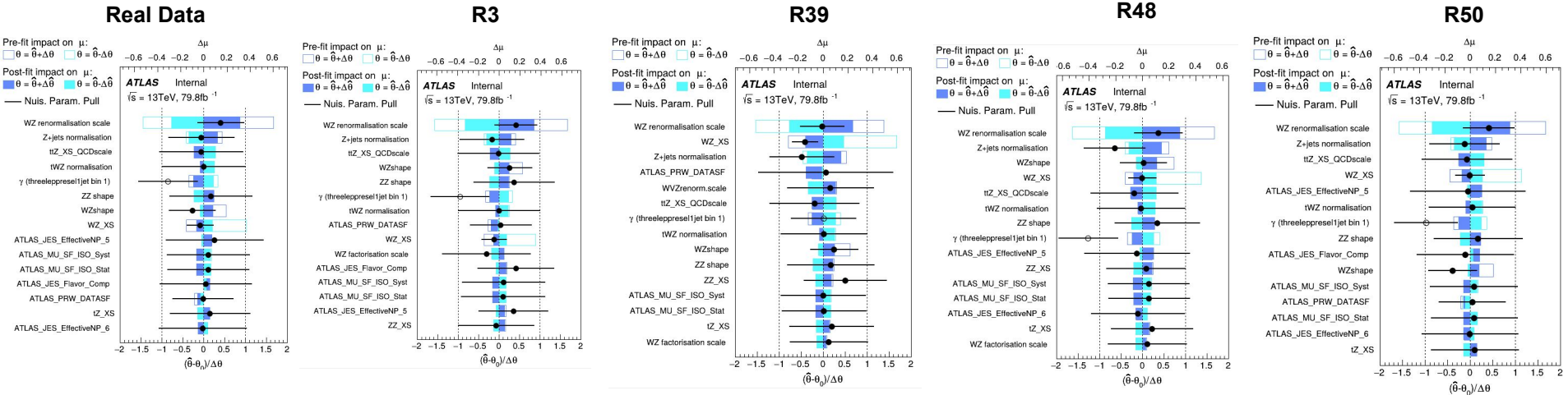
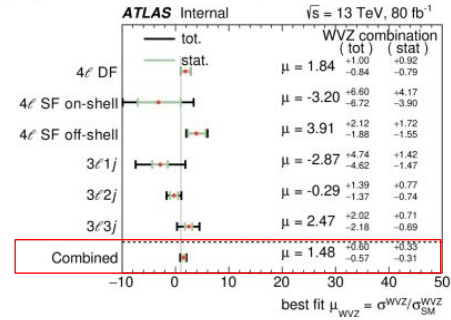
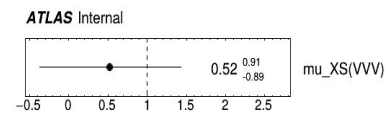
Trexfitter: 3l2j SR

- На данном этапе расширение анализа на наиболее чувствительный регион 3l2j
- Потеря статистики составляет около 1 %, что находится в пределах ожидаемых флуктуаций метода bootstrap.
 - 3l2j: orig = 2438, repl = 2407
- Метод подтверждает стабильность оценки μ и отсутствие смещения в пределах ожидаемых статистических флуктуаций



Combined regions (replicas)

Impact-параметры и pull-распределения для реальных данных и нескольких бутстрэп-реплик показывают согласие в пределах статистических флуктуаций.



Подготовка ntuples

Cutflow:

Все события нормализованы к одинаковой интегральной светимости.

- **tW**: 7042 события
- **t $\bar{t}$** : 46342 события

Отмечается различие в распределениях переменных, использованных для обучения BDT, в первую очередь в крайних бинах (первом и последнем).

- может означать проблемы с отбором.

События в области сигнала 1j1b

Process	Events
tW	$8\,300 \pm 1\,400$
$t\bar{t}$	$38\,400 \pm 6\,600$
$Z + \text{jets}$	620 ± 310
Diboson	230 ± 58
Fakes	220 ± 220
Predicted	$47\,800 \pm 7\,300$
Observed	45 273

