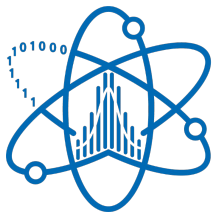




National Research
**Tomsk
State
University**



**Лаборатория
анализа данных
физики высоких энергий**

Томского
государственного
университета

Физический анализ данных

Томский Государственный Университет

08.12.2025

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:

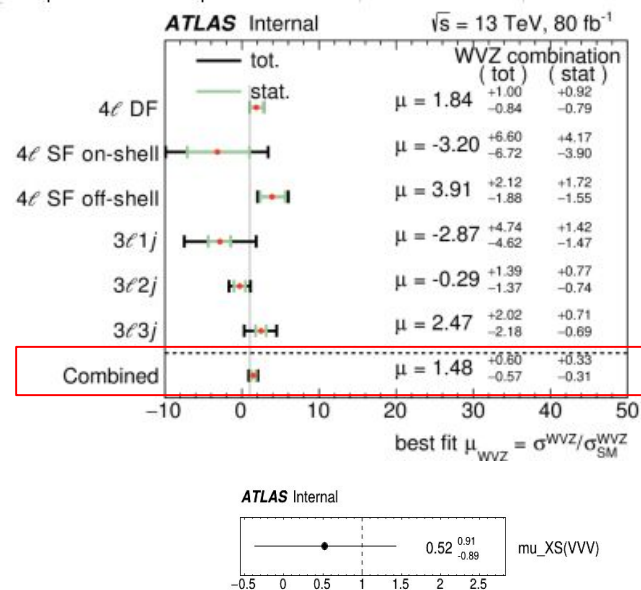
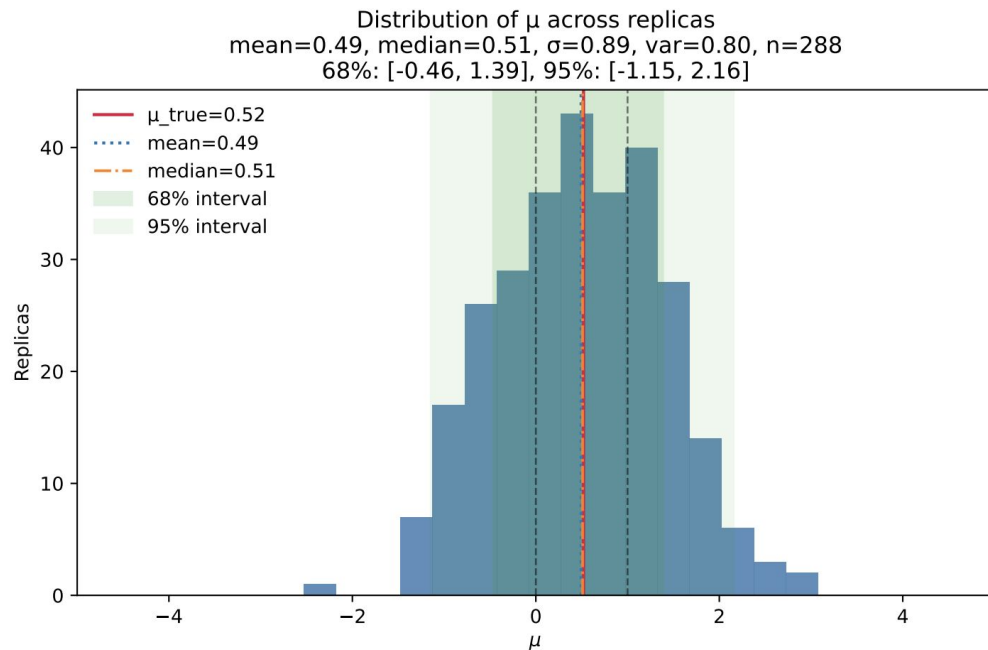
Проверить статистическую устойчивость и надёжность измерения силы сигнала (μ) в условиях ограниченной статистики.

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.

Trefitter: combined SRs

- Среднее значение силы сигнала (μ) совпадает с номинальным в пределах статистических неопределенностей, что подтверждает стабильность и корректность метода при комбинировании регионов.
- Следующий шаг:** валидация bootstrap подхода будет проведена на открытых данных HEPData



- Measurement of the total and differential cross-sections of $t\bar{t}W$ production: **хорошая стартовая точка**
 - Готовые bootstrap реплики уже в workspace ✓
 - Рабочий скрипты фитирования ✓
 - Отключение регуляризации, запуск фитирования (X RooFit) ✓
 - оценить чистую статистическую вариацию данных
 - Прогон 1000 реплик: в процессе ▶▶
- Measurement of t-channel production of single top quarks and antiquarks
 - JSON - файл с данными - нужно понимание структуры (или конвертация JSON → ROOT workspace)
 - Несколько вариантов фитирования: PyHF или TRExFitter

Результаты фитирования (без регуляризации):

```
Fit status: 0 (good)
Covariance Quality: 3 (good)
NLL: 454.056

Norm_ttW = 1.1393 ± 0.0932
```

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:

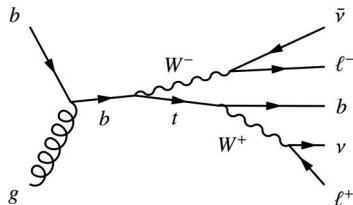
- [ATLAS Open Data \(2015–2016\)](#) — открытые данные эксперимента ATLAS.
- [HEPData Repository](#) — вспомогательные материалы и таблицы с реконструированными событиями.

Физический процесс:

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

Требуется:

- Ровно **один b-таг 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](#)

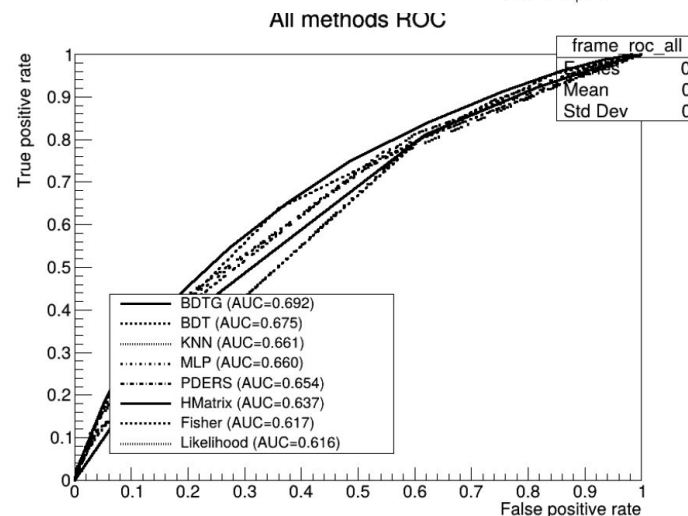
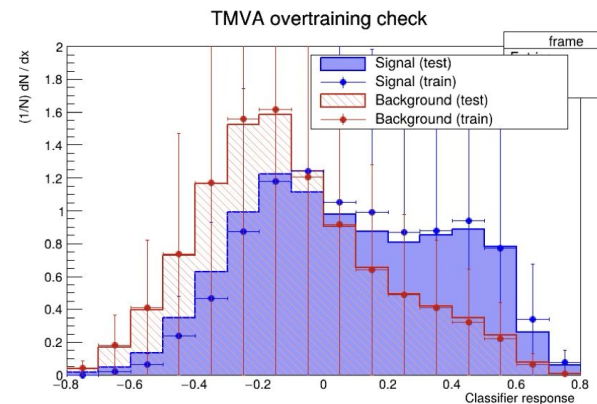
[Data points](#)

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

Результаты обучения

ROC-кривая демонстрирует, что BDT и BDTG показывают наилучшую производительность среди протестированных методов ($AUC \approx 0.69$).

- Проведена перекрёстная проверка выбора переменных и стратегии обучения.
- Тестировались альтернативные классификаторы (MLP, SVM, PCA в TMVA) для оценки относительной эффективности BDT.
- Начата работа по оптимизации гиперпараметров и порогов классификации с помощью **R** для увеличения эффективности сигнала при фиксированном уровне фона.
- Готовы ntuple-файлы с увеличенной статистикой, что позволит улучшить качество обучения и стабилизировать результаты в следующих итерациях → **увеличение статистики в 3 раза**



.....

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.

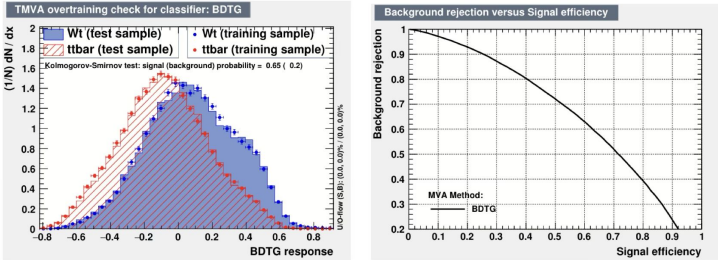
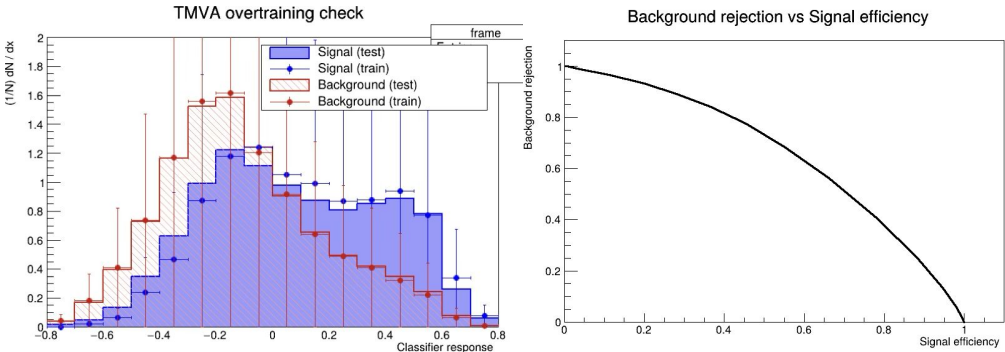
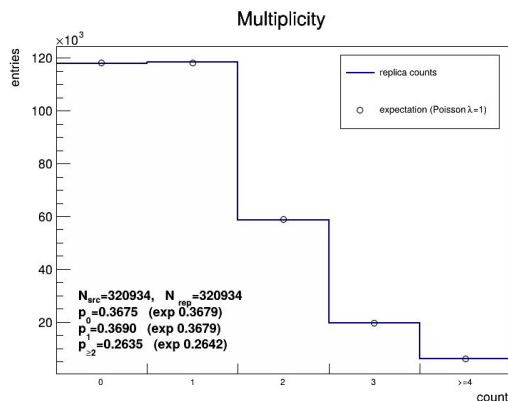


Figure 9: Comparison of test/training sample distributions and background rejection factor versus signal efficiency.

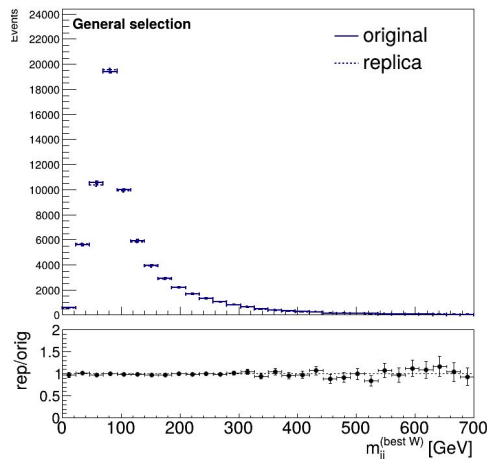


Подготовка реплик

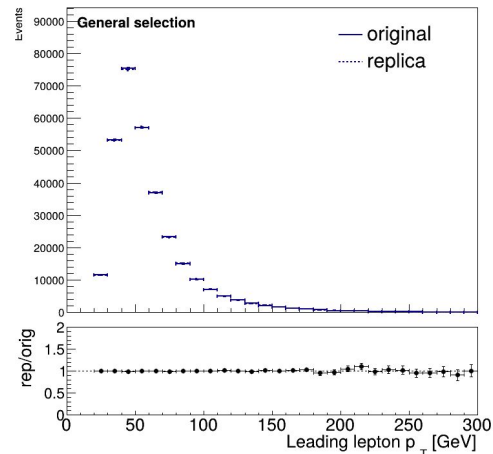
- Сгенерированы **1000 bootstrap-реплик** методом случайной повторной выборки с заменой:
 - для каждой реплики **фиксируется seed**, что обеспечивает **воспроизводимость результатов**.
- Каждая реплика сохраняет **общее количество событий** и **все корреляции между переменными**:
 - реплики формируются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



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

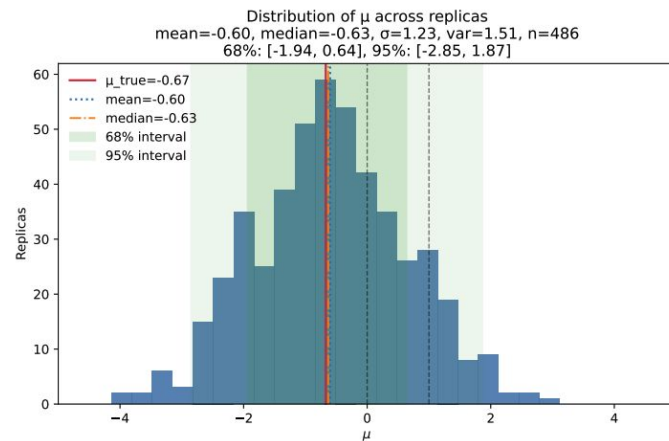
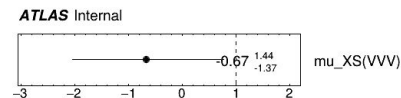
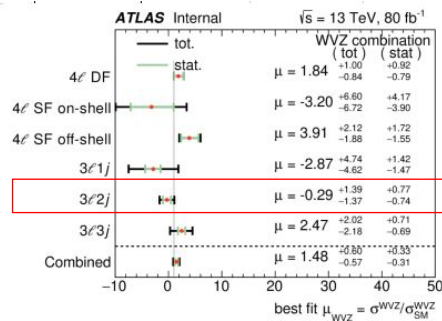
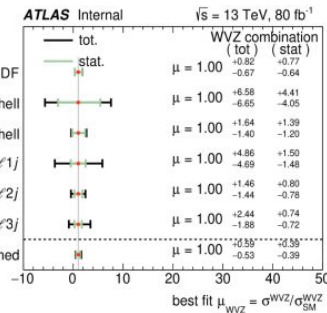
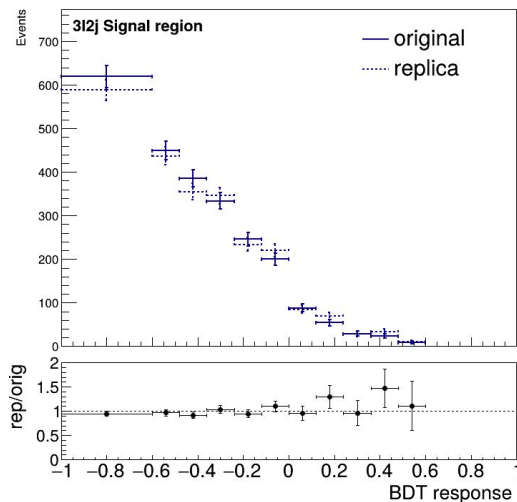
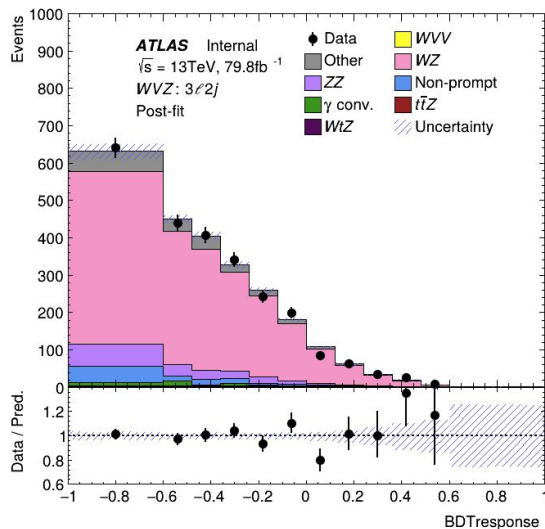


Сравнение форм распределений между оригинальными данными и репликами показывает **согласие** в пределах статистических флуктуаций.



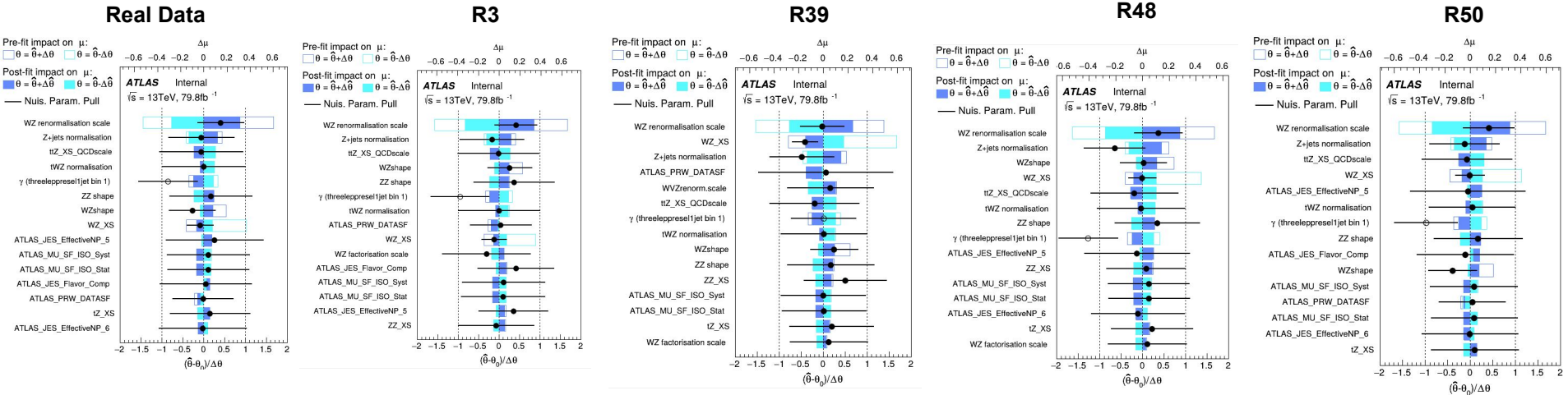
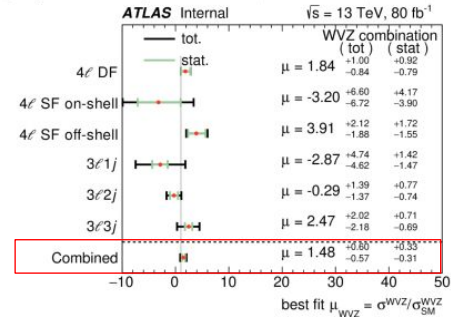
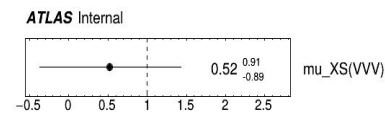
Trexfitter: 3l2j SR

- 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

