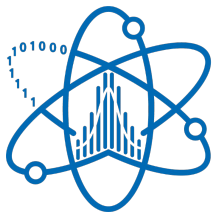




National Research
**Tomsk
State
University**



**Лаборатория
анализа данных
физики высоких энергий**

Томского
государственного
университета

Физический анализ данных

Томский Государственный Университет

16.02.2026

Направления работы

✓ **Повторение анализа ATLAS Open Data** (решили техническую проблему, пока приоритет на других задачах)

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 (Bootstrap)**

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

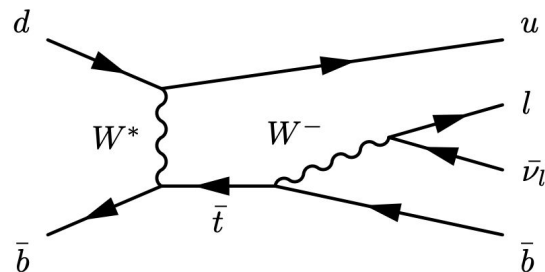
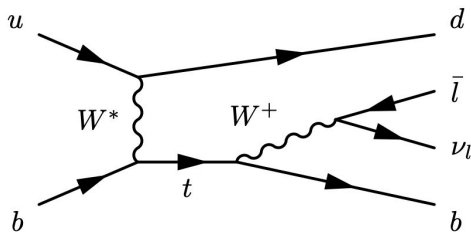
BBCI (*Bias-Corrected Bootstrap for Confidence Intervals*) — модифицированный бутстрэп-метод, который улучшает оценку доверительных интервалов параметров за счёт коррекции смещения и нестабильности, возникающих при малых выборках.

Цель:

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

BCL: 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**.

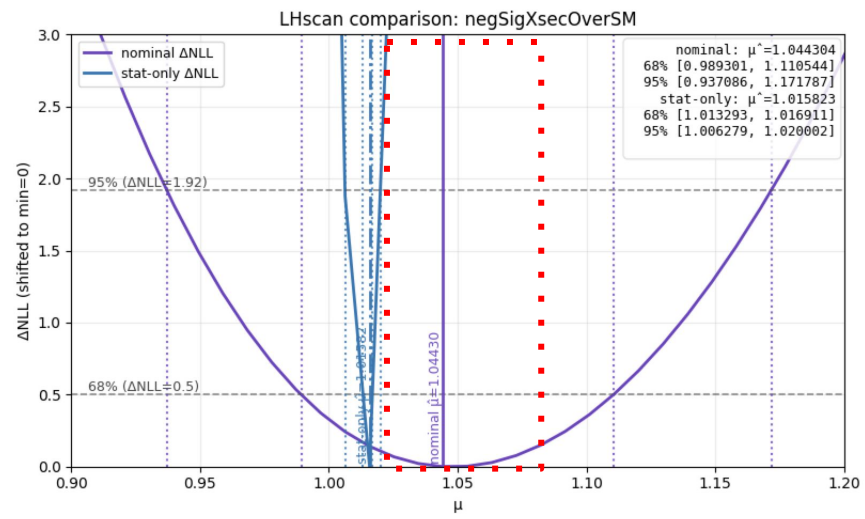
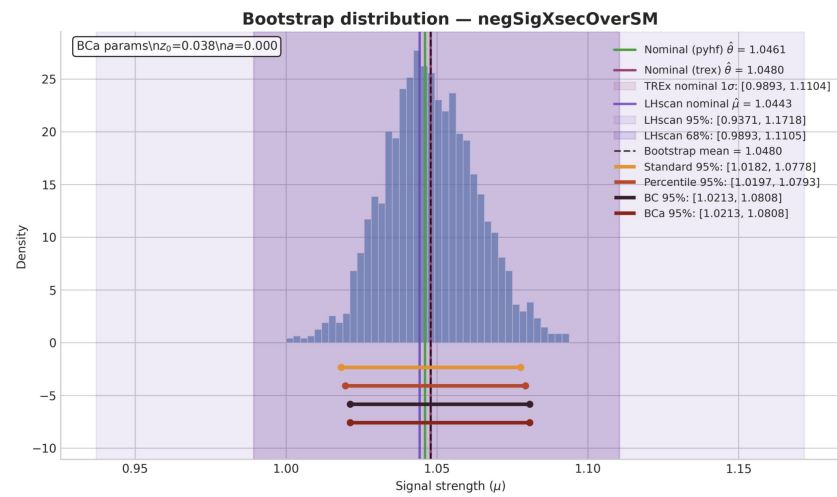


Сравнение интервалов (Trexfitter)

Для сравнения с Bootstrap интервалами добавлены номинальные 68% и 95% доверительные интервалы LHscan номинального фита.

- **Nominal LHscan** — полный фит с профилированием nuisance-параметров
- Был выполнен stat-only fit (LHscan): он дал значительно более узкие интервалы.

Бутстрапный доверительный интервал находится между статистическим и номинальным интервалами LHscan. Выполненный bootstrap по наблюдаемым данным (Poisson-fluctuation) воспроизводит преимущественно статистическую компоненту неопределённости.



Параметрический бутстрап (переход на руНФ)

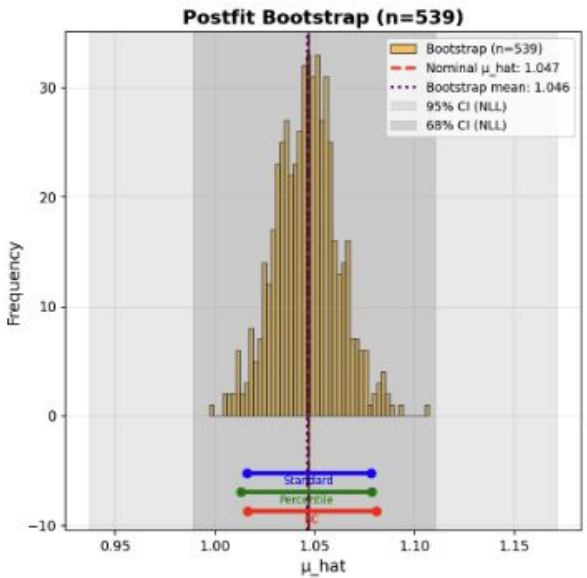
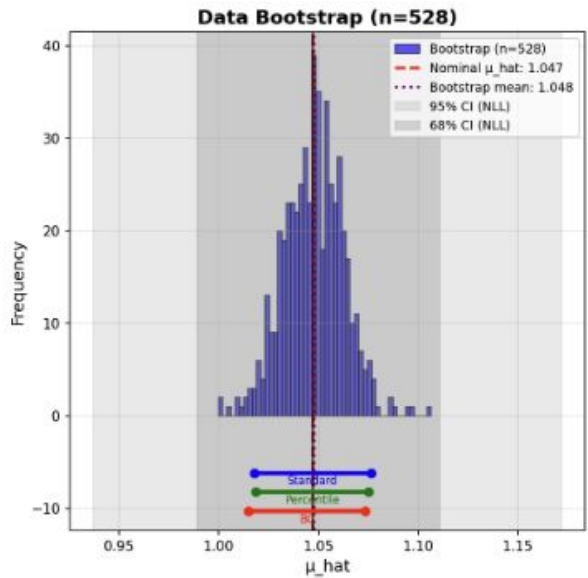
Profile likelihood включает также вклад систематических неопределённостей через профилирование nuisance-параметров θ .
Для корректной проверки полной профилированной ошибки необходимо:

- учитывать пост-фит значения ($\hat{\mu}$, θ),
- воспроизводить ожидаемые события при этих параметрах,
- флуктуировать их согласно статистической модели.

Выполнен параметрический бутстрап, используя псевдоэксперименты, сгенерированные на основе ожидаемого значения после фитирования в подобранной точке (μ, θ).

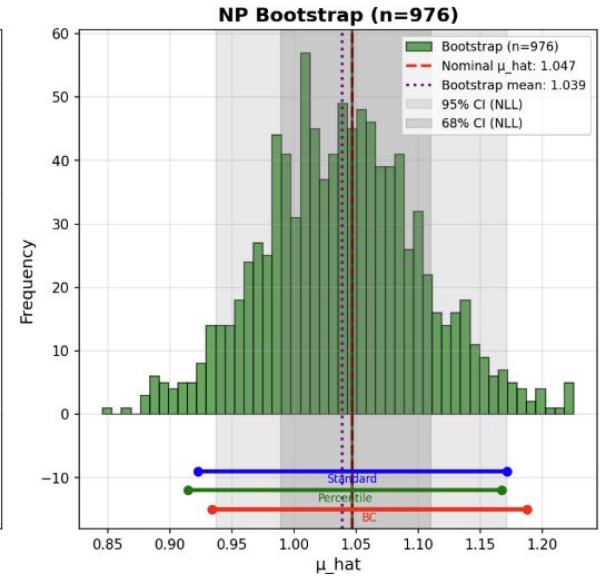
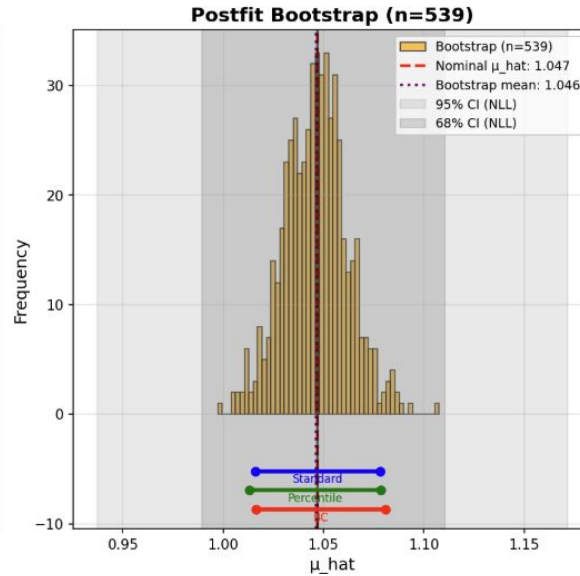
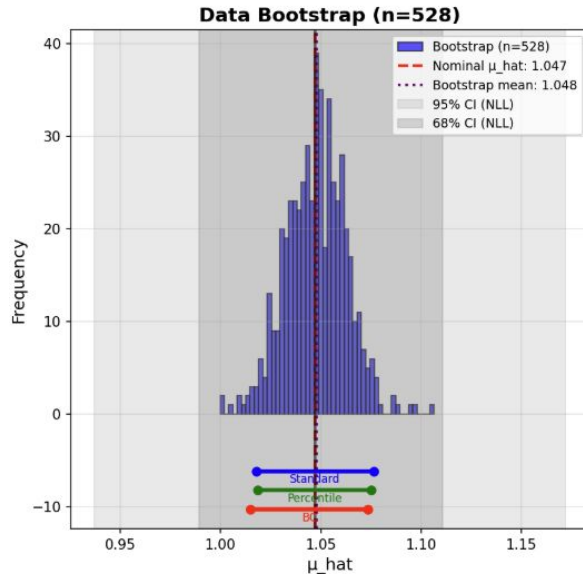
Это позволяет воспроизвести вклад как статистических, так и систематических неопределённостей и сравнить ширину распределения $\hat{\mu}$ с полной профилированной ошибкой

$\sigma_{\text{stat}} = 0.00510$
 $\sigma_{\text{syst}} = 0.06724$
 $\sigma_{\text{total}} = 0.06743$



Параметрический бутстреп: NP sampling (preliminary)

- Из номинального фита получены $\hat{\mu}$, $\hat{\theta}$ и ковариационная матрица $\text{Cov}(\theta)$.
- Для каждой bootstrap-реплики:
 - Сэмплируем nuisance-параметры: $\theta_{\text{sample}} \sim N(\hat{\theta}, \text{Cov}(\hat{\theta}))$
 - Вычисляем ожидаемые события: $\lambda(\hat{\mu}, \theta_{\text{sample}})$
- Генерируем псевдо-данные по распределению Пуассона
- Выполняем полный профилированный фит и извлекаем $\hat{\mu}$



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:

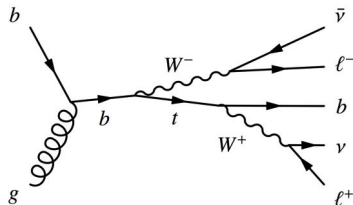
- [ATLAS Open Data \(2015–2016\)](#) — открытые данные эксперимента ATLAS.
- [HEPData Repository](#) — вспомогательные материалы и таблицы с реконструированными событиями.

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

Проведена систематическая оценка различных алгоритмов машинного обучения с использованием nested cross-validation — 5 внешних и 3 внутренних fold'ов для подбора гиперпараметров.

В таблице представлены значения mean_fold_auc, стандартное отклонение по fold'ам и итоговый out-of-fold AUC.

По результатам сравнения можно выделить три наиболее эффективных алгоритма:

- **Gradient Boosting** — AUC ≈ 0.670
- **SVC** — AUC ≈ 0.668
- **Random Forest** — AUC ≈ 0.667

Лучший результат демонстрирует Gradient Boosting, однако различия между тремя лидерами статистически невелики

Algorithm (Python)	outer_k	inner_k	n_iter	mean_fold_auc	std_fold_auc	overall_oof_auc	mean_fold_wacc	std_fold_wacc	overall_oof_wacc
AdaBoost	5	3	25	0.663612	0.006768	0.660667	0.62078	0.004724	0.62078
Bagging	5	3	25	0.647918	0.007531	0.647809	0.502429	0.000625	0.502429
ExtraTrees	5	3	25	0.667711	0.006485	0.667169	0.622879	0.005654	0.622879
GaussianNB	5	3	25	0.636746	0.008755	0.636637	0.59563	0.004879	0.59563
GradientBoosting	5	3	25	0.670242	0.007283	0.670033	0.625561	0.006146	0.62556
HistGradientBoosting	5	3	25	0.668588	0.006794	0.668074	0.624038	0.005583	0.624038
KNN	5	3	25	0.649829	0.006528	0.649802	0.500247	0.000234	0.500247
LDA	5	3	25	0.633161	0.009455	0.632984	0.500227	4.95E-05	0.500227
LinearSVC	5	3	25	0.638989	0.00895	0.638861	0.602301	0.008101	0.6023
LogisticRegression	5	3	25	0.641267	0.008797	0.641124	0.604398	0.007429	0.604398
MLP	5	3	25	0.66697	0.007911	0.666548	0.50295	0.00077	0.50295
QDA	5	3	25	0.645348	0.007435	0.645249	0.51138	0.002164	0.511379
RandomForest	5	3	25	0.667747	0.006265	0.667129	0.620657	0.006323	0.620657
SGDClassifier	5	3	25	0.652858	0.006399	0.652702	0.570652	0.004155	0.570652
SVC	5	3	12	0.668126	0.006504	0.667912	0.62335	0.00485	0.62335

Результаты обучения

- Лучшие алгоритмы (Gradient Boosting / BDTG) дают $AUC \approx 0.67$.
- Результаты согласуются между Python и TMVA.
- Существенного выигрыша от более сложных моделей нет.
- Основное ограничение связано с физической информативностью входных переменных.
- Методология nested CV обеспечивает корректную оценку качества.

Platform	Best algorithm	Best score (AUC)	Score type (as reported)
TMVA	BDTG	0.670094	best_mean_auc (k-fold)
Python	GradientBoosting	0.670033	overall_oof_auc (out-of-fold)
R	MLP	0.660152	auc_oof_weighted (out-of-fold, weighted)

.....

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 ($pT(\ell\ell E_T^{miss})$, $\Delta pT(\ell\ell b E_T^{miss})$) 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
$\Delta pT(\ell\ell b E_T^{miss})$	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.

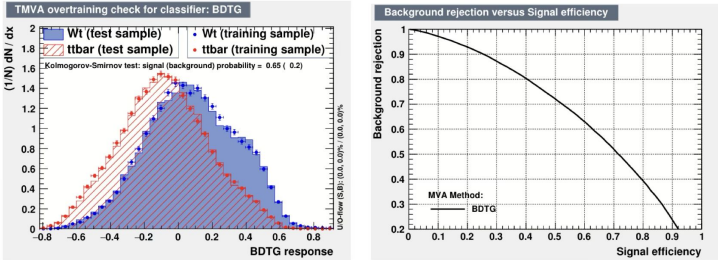
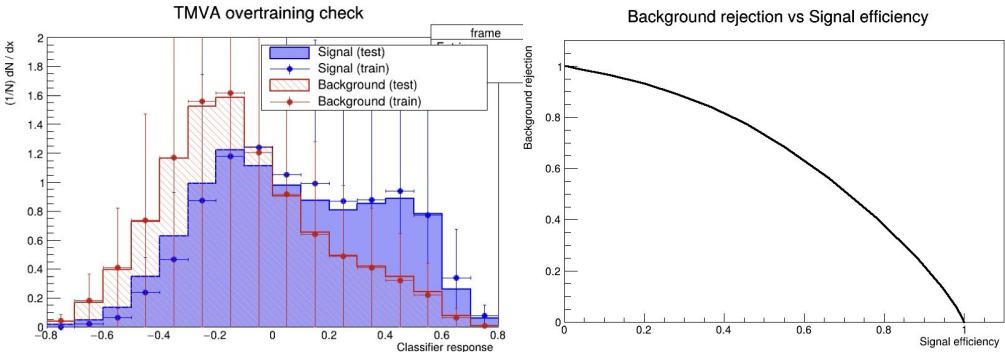
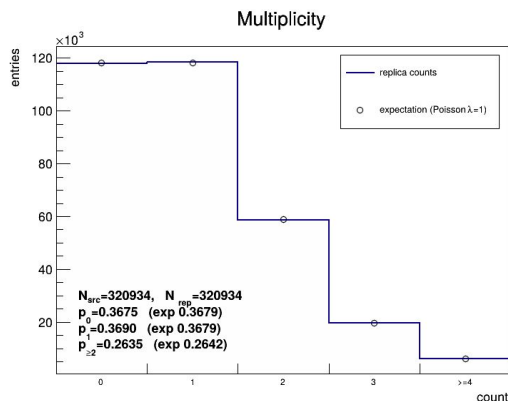


Figure 9: Comparison of test/training sample distributions and background rejection factor versus signal efficiency.

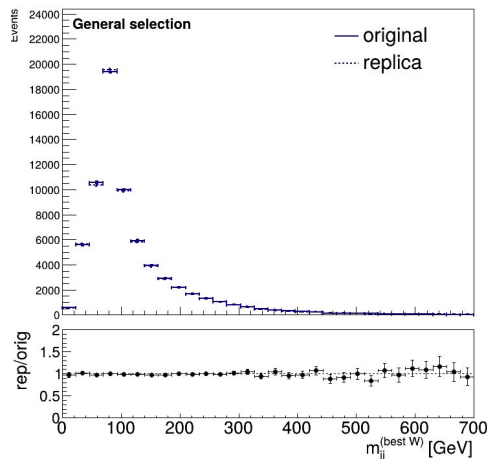


Подготовка реплик

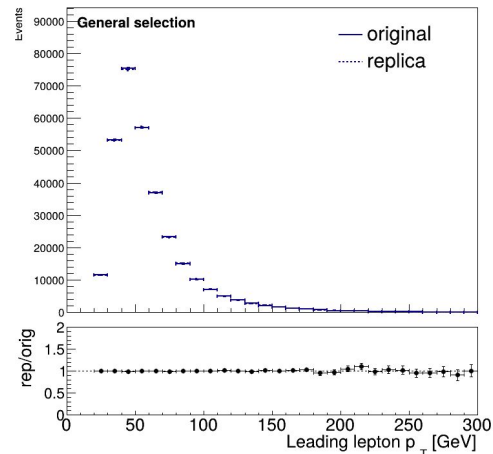
- Сгенерированы **1000 bootstrap-реплик** методом случайной повторной выборки с заменой:
 - для каждой реплики **фиксируется seed**, что обеспечивает **воспроизводимость результатов**.
- Каждая реплика сохраняет **общее количество событий** и **все корреляции между переменными**:
 - реплики формируются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



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

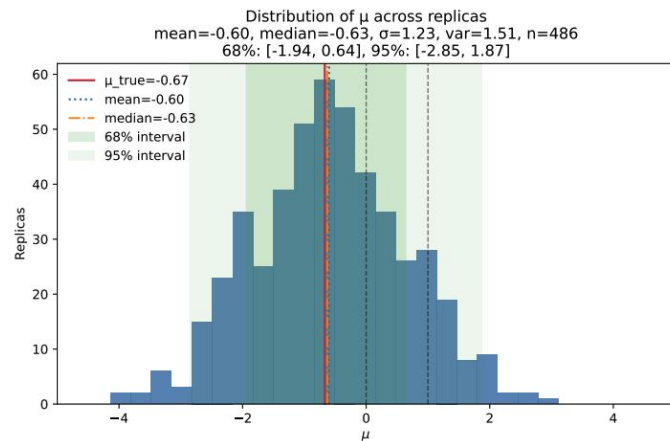
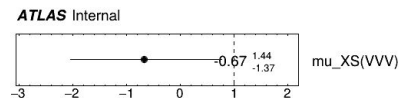
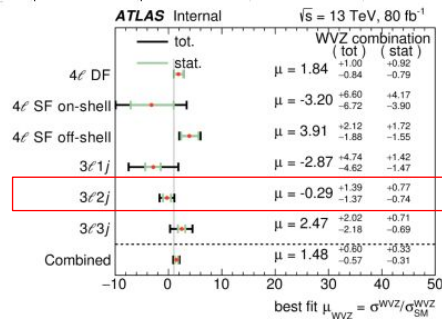
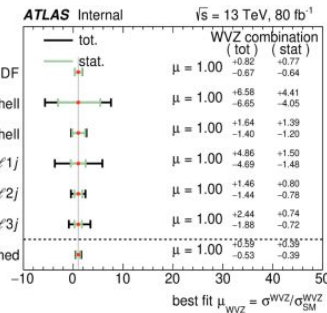
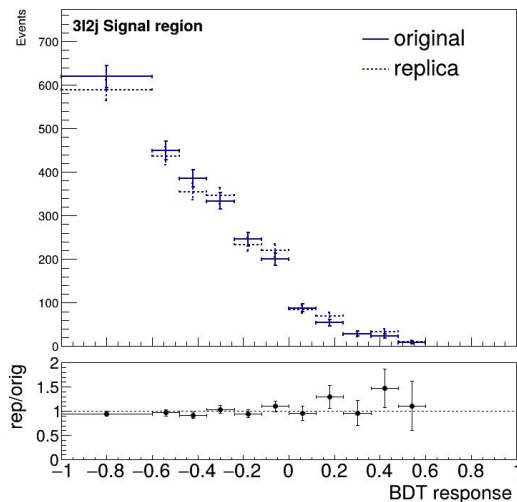
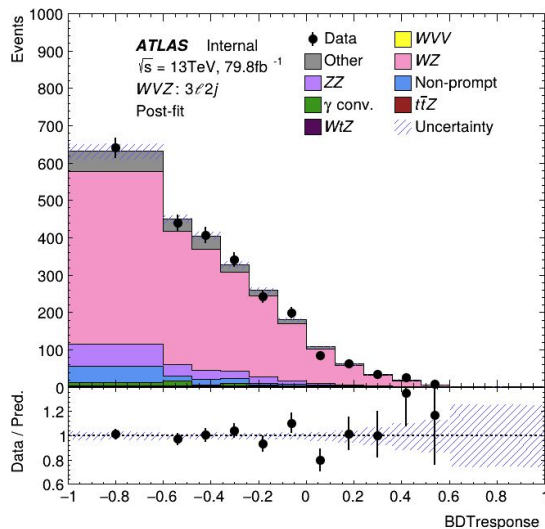


Сравнение форм распределений между оригинальными данными и репликами показывает **согласие** в пределах статистических флуктуаций.



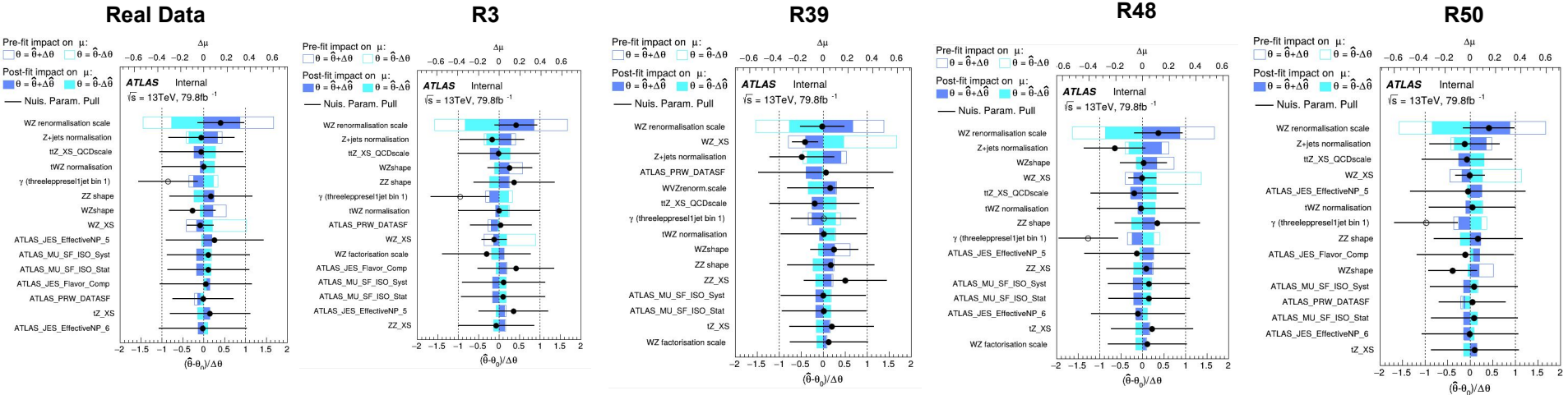
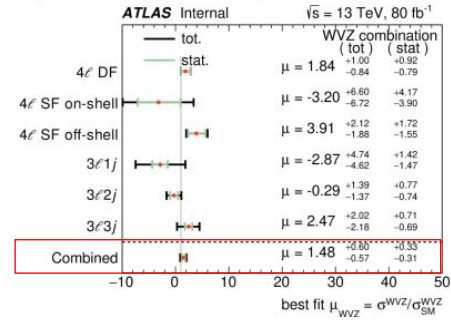
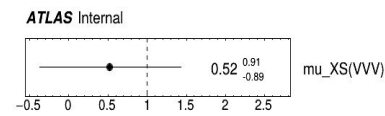
Trexfitter: 3l2j SR

- 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.

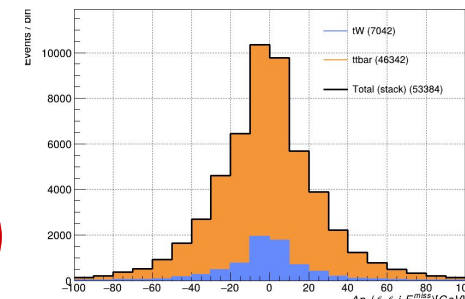
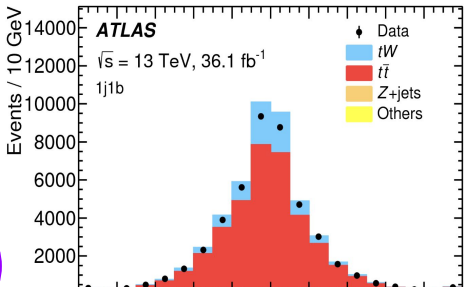
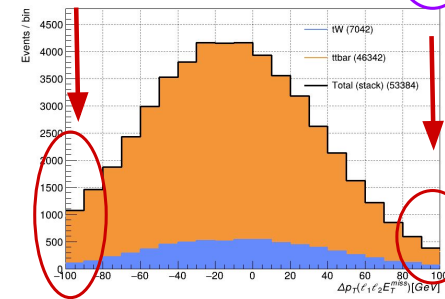
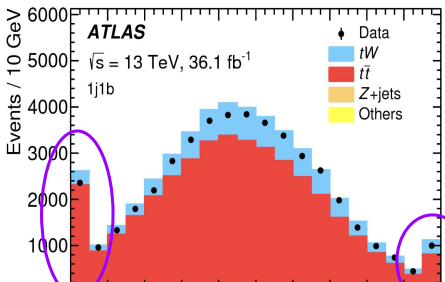
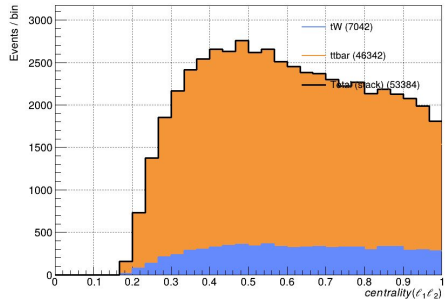
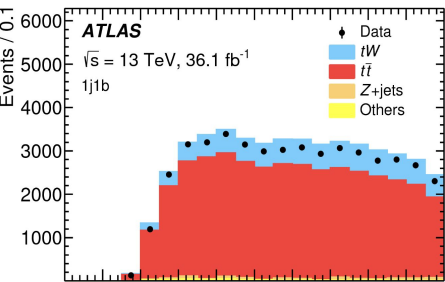
- **tW**: 7042 события
- **t $\bar{t}$** : 46342 события

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



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–антифон)**: 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.

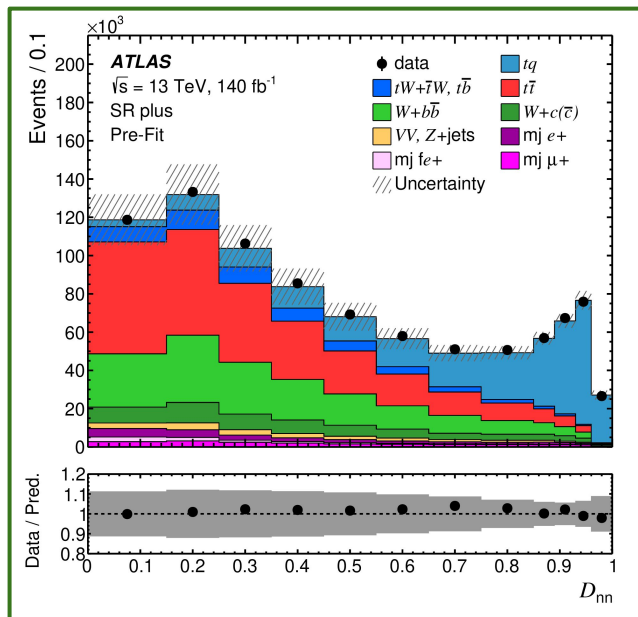
CR name	Requirement
B-e-plus	$q_e/e = +1, \eta(e) < 1.37, E_T^{\text{miss}} < 30 \text{ GeV}$
B-e-minus	$q_e/e = -1, \eta(e) < 1.37, E_T^{\text{miss}} < 30 \text{ GeV}$
EC-e-plus	$q_e/e = +1, \eta(e) > 1.52, E_T^{\text{miss}} < 30 \text{ GeV}$
EC-e-minus	$q_e/e = -1, \eta(e) > 1.52, E_T^{\text{miss}} < 30 \text{ GeV}$
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.

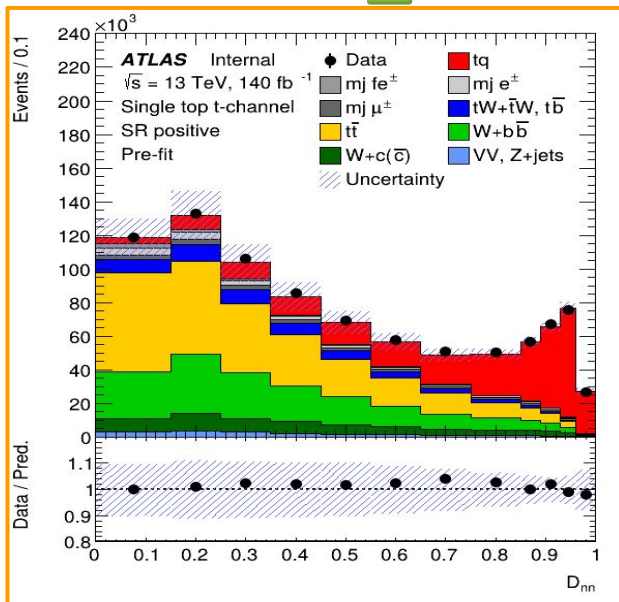
6 контрольных региона

- **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}

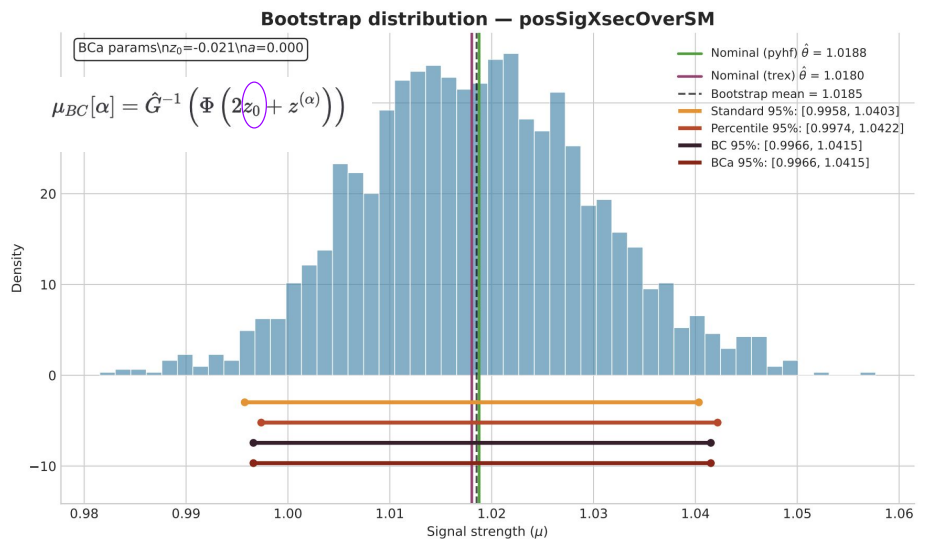
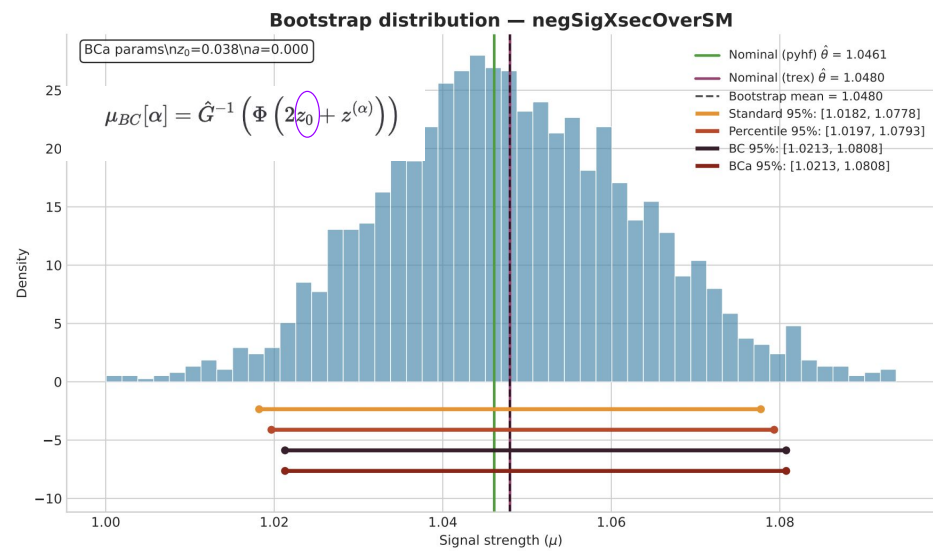
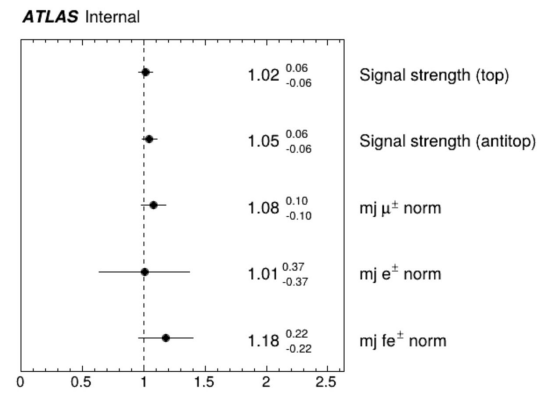
44 bins distributed across the 8 channels

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

