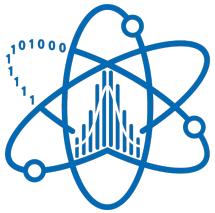
National Research
Tomsk
State
University



**Лаборатория
анализа данных
физики высоких энергий**

Томского
государственного
университета

Физический анализ данных

Томский Государственный Университет

16.03.2026

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**

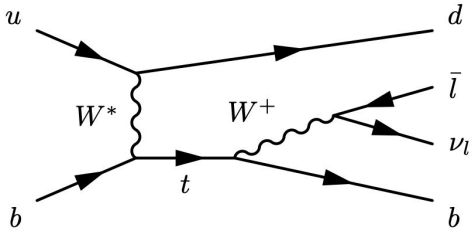
Пример: tW при 13 ТэВ, ATLAS

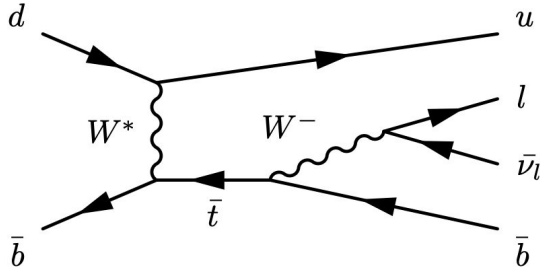
Цель: сравнение BDT, NN, Transformers, GNN и др. по метрикам качества классификации.

✅ Курс для первокурсников: Модуль анализ данных HEP

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

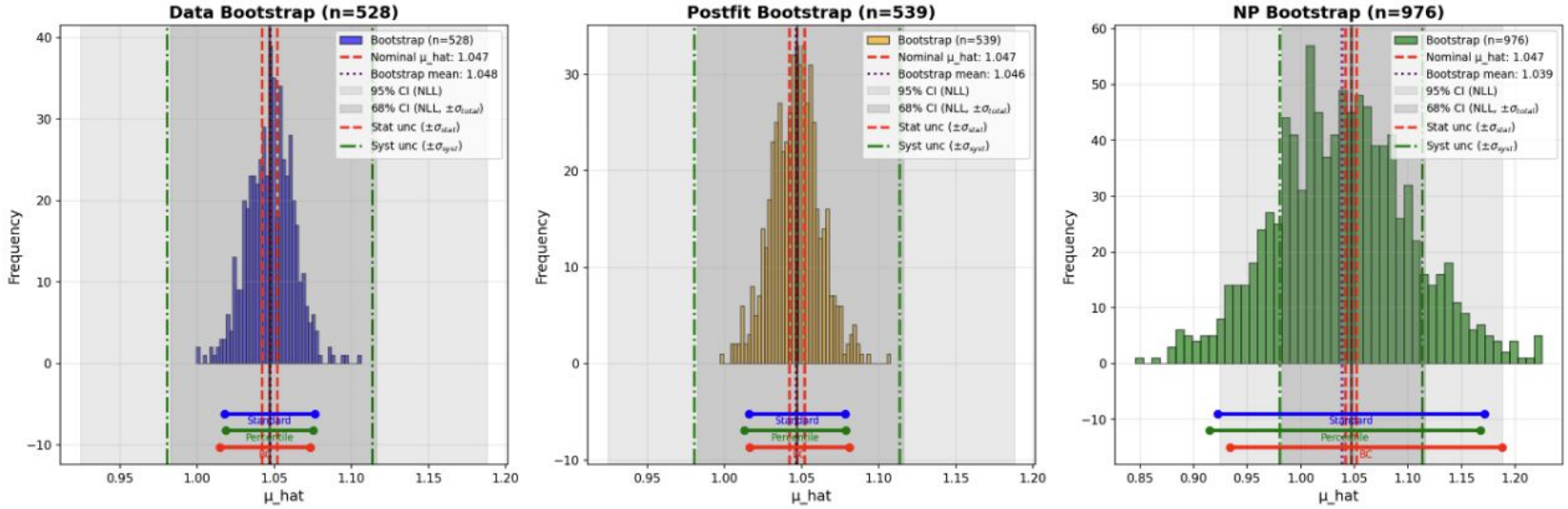
- **Анализ WVZ:** первичная валидация bootstrap → корректное распределение μ
- **Метод на открытых данных HEPData -** *“Measurement of t-channel production of single top quarks and antiquarks”*
- В данном анализе используется 140 фб$^{-1}$ данных эксперимента ATLAS при энергии столкновений 13 ТэВ для 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 → d) и tW^- (переход d → 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**.





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

- Рассмотрены три bootstrap-подхода: **data bootstrap**, **post-fit bootstrap**, **nuisance parameter resampling**
- Для каждой bootstrap-реплики выполняется полный fit и извлекается значение μ
- Распределения центрированы около номинального значения → оценка параметра сигнала **стабильна**



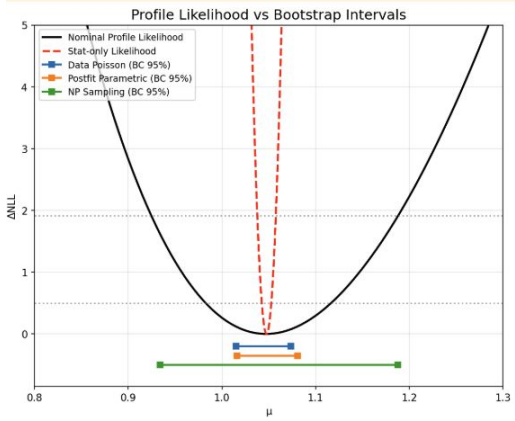
$$\hat{\mu} = 1{,}047 \pm 0{,}067\ (\text{сист.}) \pm 0{,}005\ (\text{стат.}).$$

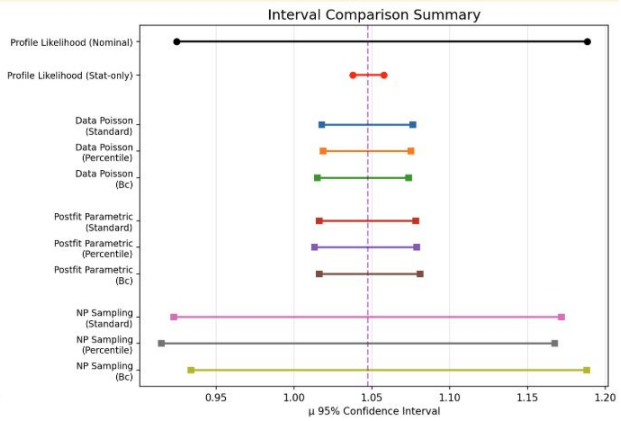
Interval comparison

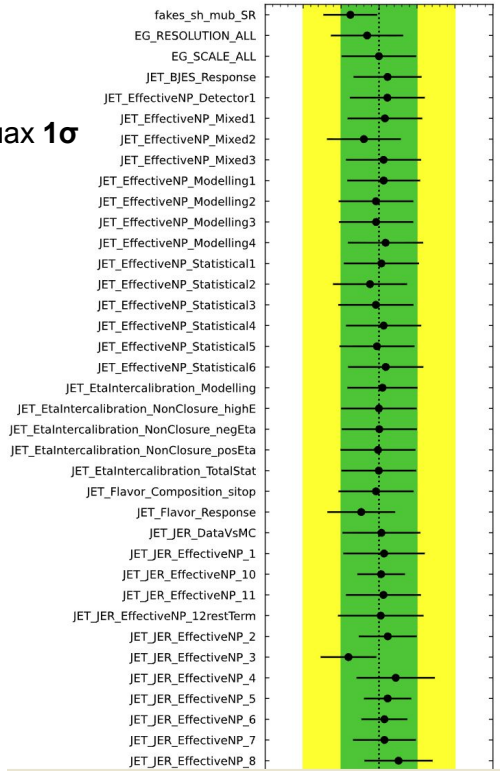
- Сравнение **profile likelihood** и bootstrap-интервалов
- Рассмотрены несколько bootstrap-методов: **data**, **post-fit**, **nuisance parameter sampling**
- Интервалы **хорошо согласуются** с profile likelihood результатом
- Pull-распределения показывают, что большинство nuisance-параметров находятся в пределах **1σ**

Дальнейшая работа:

- увеличение статистики до 10000
- исследовать bootstrap-распределения **impactful nuisance parameters**
- выбрать финальный набор графиков для статьи







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.

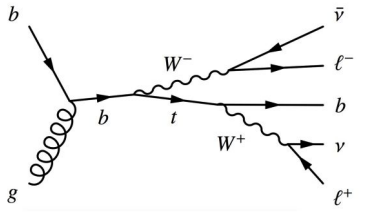
Данные:

- **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 √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

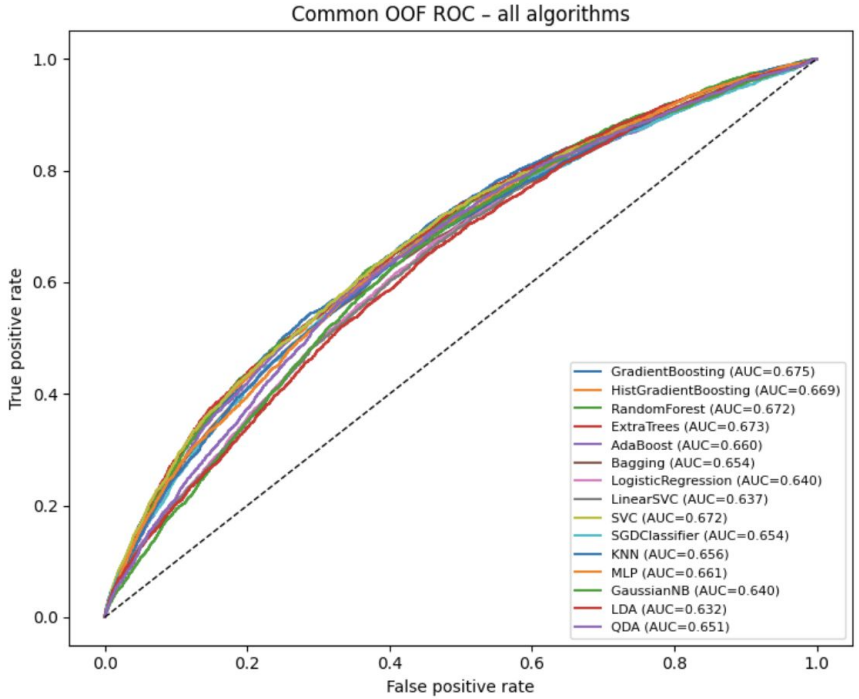
Результаты обучения

- Задача: разделение **tW сигнала** и **tt фона** в регионе **1j1b dilepton**
- Используются **9 физических переменных** как входные признаки
- Тестируются различные алгоритмы машинного обучения:
  - Random Forest
  - Gradient Boosting
  - SVM
  - Neural networks

Реализована **nested cross-validation** (outer loop — оценка качества, inner loop — настройка гиперпараметров)

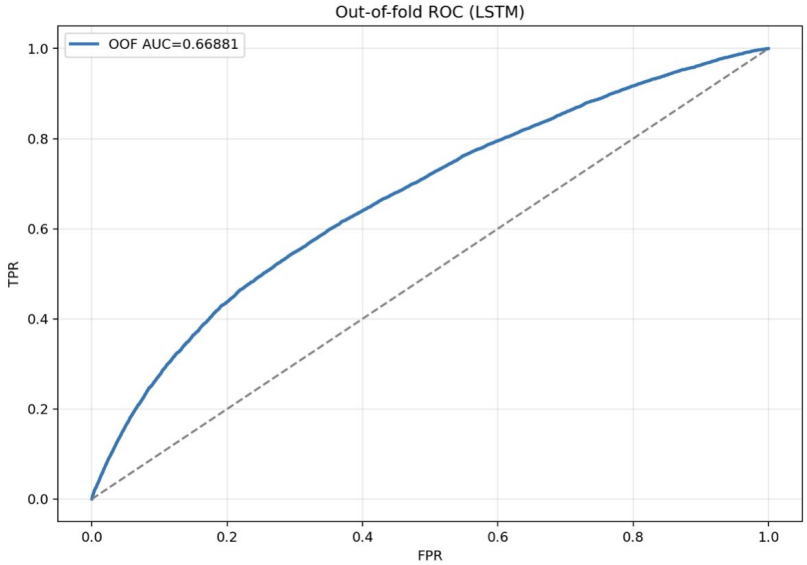
Используется **взвешенная бинарная cross-entropy** и балансировка классов

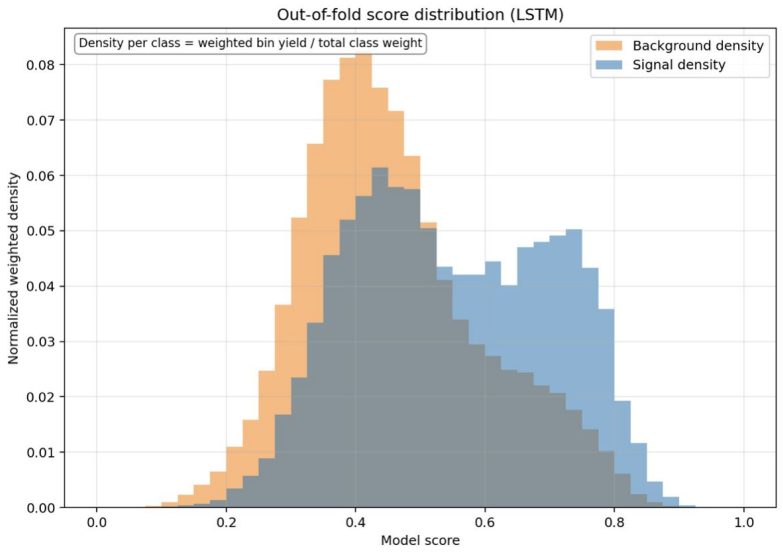
Algorithm	overall_oof_auc
GradientBoosting	0.67482
ExtraTrees	0.673411
RandomForest	0.67238
SVC	0.671871
HistGradientBoosting	0.668984
MLP	0.660661
AdaBoost	0.660288
KNN	0.655979
SGDClassifier	0.654416
Bagging	0.653939
QDA	0.651292
GaussianNB	0.639867



LSTM для классификации сигнал/фон

- Обучена модель **LSTM** для классификации сигнал/фон
- Входные признаки рассматриваются как **упорядоченная последовательность features**
- Используется: **early stopping, learning rate scheduling и AdamW optimizer**
- Средний результат по фолдам: **mean validation AUC ≈ 0.669**
- Полученная производительность **сопоставима с классическими ML-алгоритмами**





Следующие шаги

- Определить **оптимальный cut по score** (максимизация S/B)
- Сравнить производительность:
  - BDT
  - LSTM
  - Transformers
- Увеличить число входных переменных и изучить влияние feature set
- Подготовить итоговые графики для статьи

Спасибо за внимание

Детали тренировки BDT

Цель:
Разделить процессы **tW** (top+anti-top) и **$t\bar{t}$** 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(ℓℓE□missb), ΔpT(ℓℓbE□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 (×10⁻²)	Produced BDT S (×10⁻²)
pT(ℓℓE_T^{miss}b)	4.1	**5.38**
ΔpT(ℓℓbE_T^{miss})	2.5	**4.19**
∑E_T	**2.3**	0.24
η(ℓℓE_T^{miss} b)	1.3	**2.85**
ΔpT(ℓℓE_T^{miss})	**1.1**	0.93
pT(ℓℓb)	**1.0**	0.62
C(ℓℓ) (centrality)	**0.9**	0.46
m(ℓ₂,b)	0.2	**0.24**
m(ℓ₁,b)	**0.1**	0.08

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

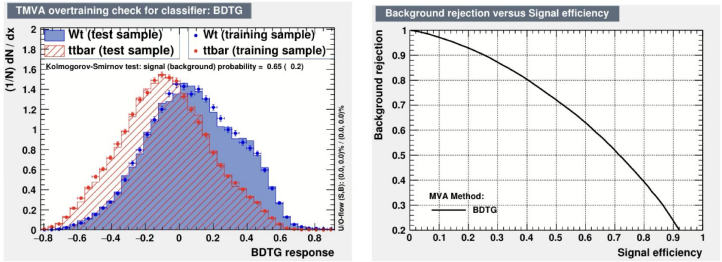
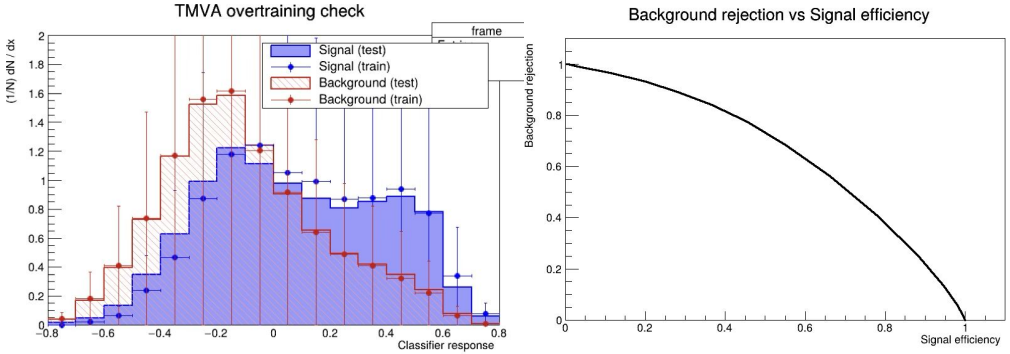
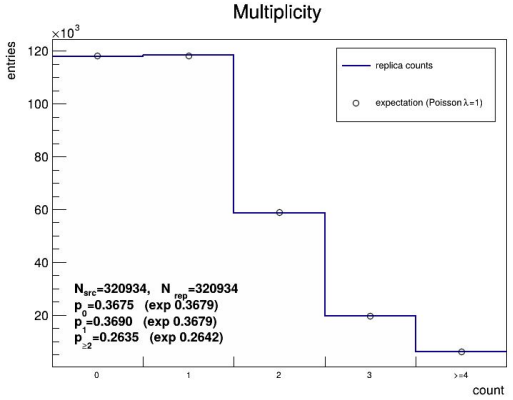


Figure 9: Comparison of test/training sample distributions and background rejection factor versus signal efficiency.

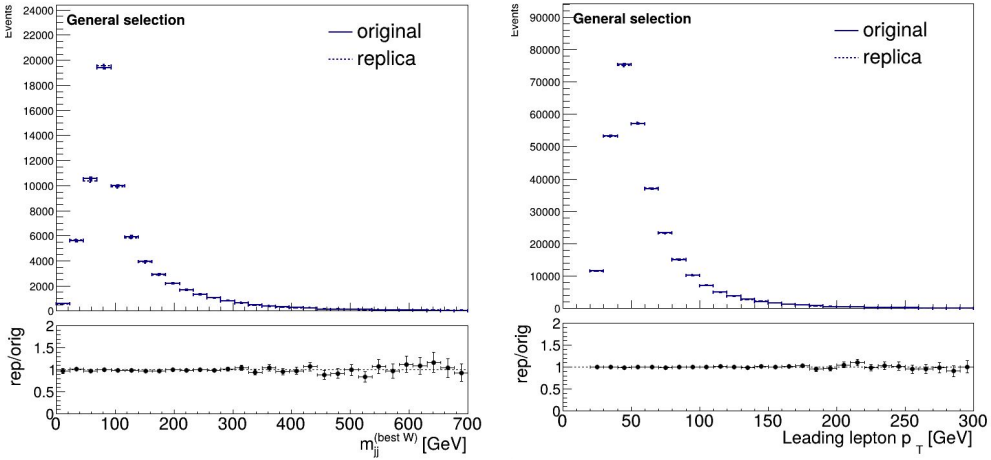


Подготовка реплик

- 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.



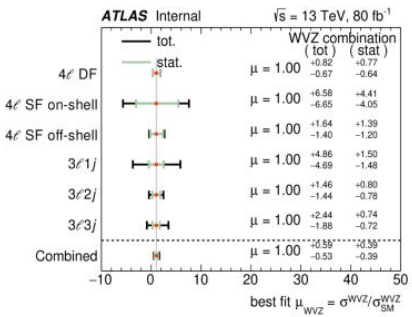
Распределение числа вхождений событий в репликах **согласуется** с теоретическим распределением Пуассона (λ = 1).
Это подтверждает **корректность бутстрэп-процедуры.**

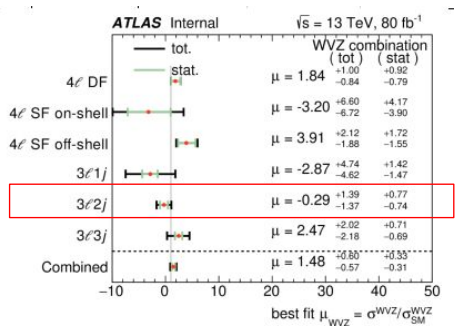


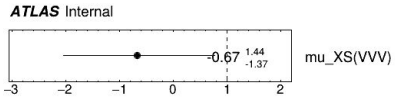
Сравнение форм распределений между оригинальными данными и репликами показывает **согласие** в пределах статистических флуктуаций.

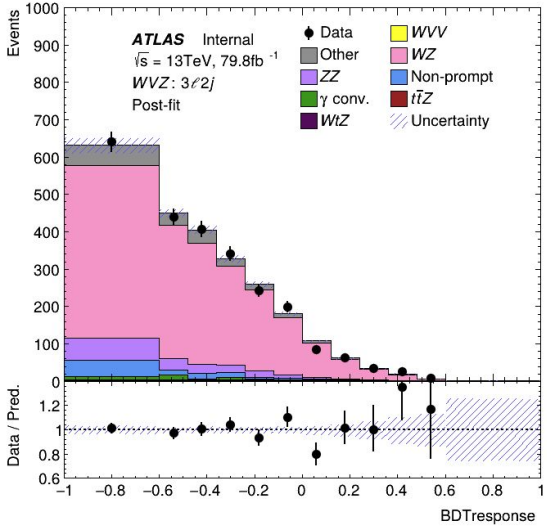
Trexfitter: 3l2j SR

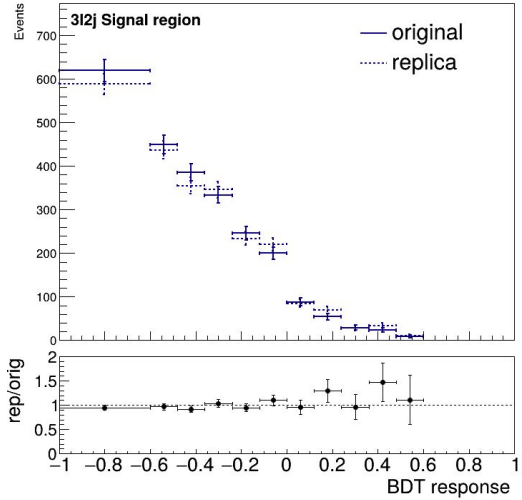
- На данном этапе расширение анализа на наиболее **чувствительный регион 3ℓ2j**
- **Потеря статистики** составляет **около 1 %**, что находится в пределах ожидаемых флуктуаций метода bootstrap.
  - **3ℓ2j:** orig = 2438, repl = 2407
- Метод подтверждает стабильность оценки μ и отсутствие смещения в пределах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

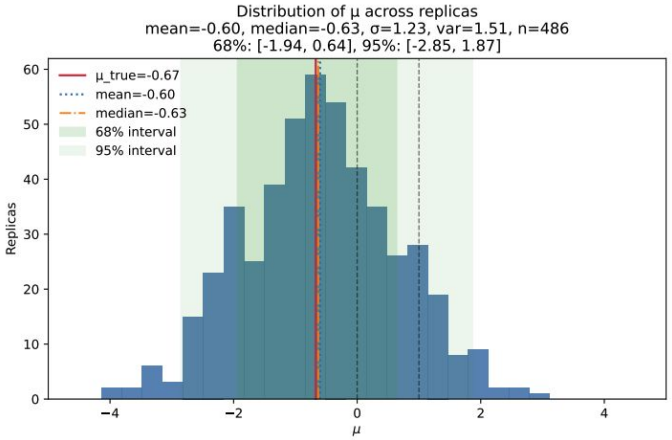




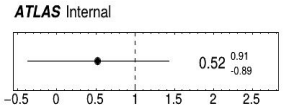


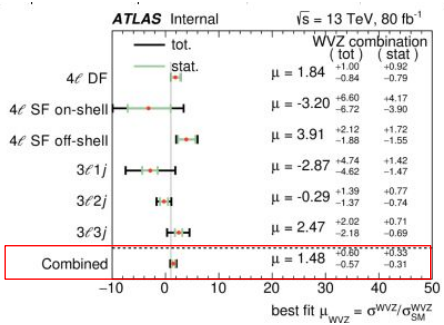






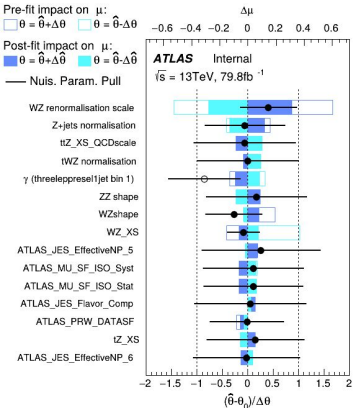
Combined regions (replicas)



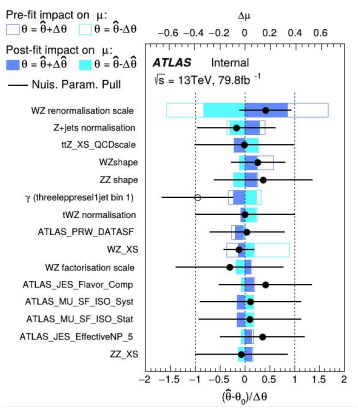


Impact-параметры и pull-распределения для реальных данных и нескольких бутстрэп-реплик показывают согласие в пределах статистических флуктуаций.

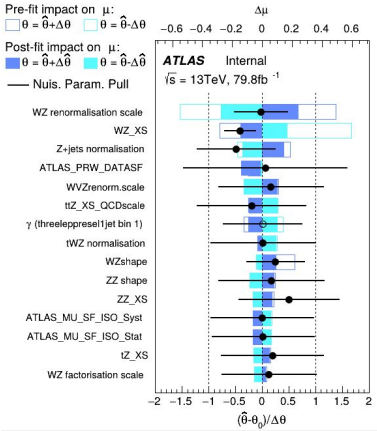
Real Data



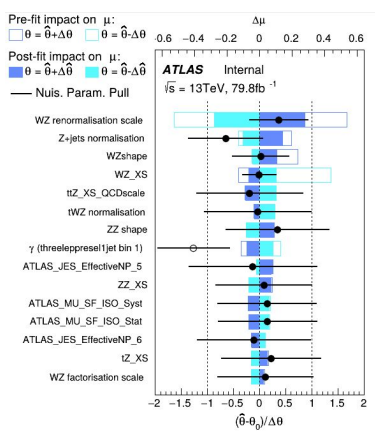
R3



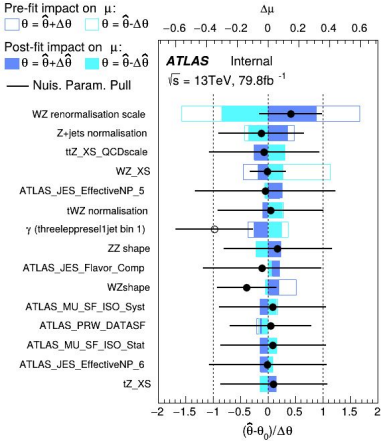
R39



R48



R50



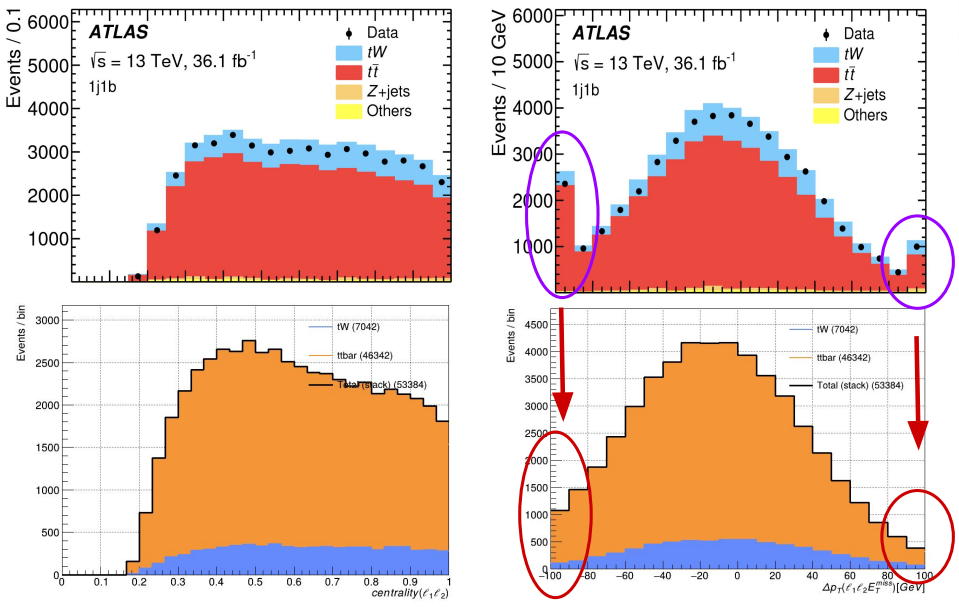
Подготовка ntuples

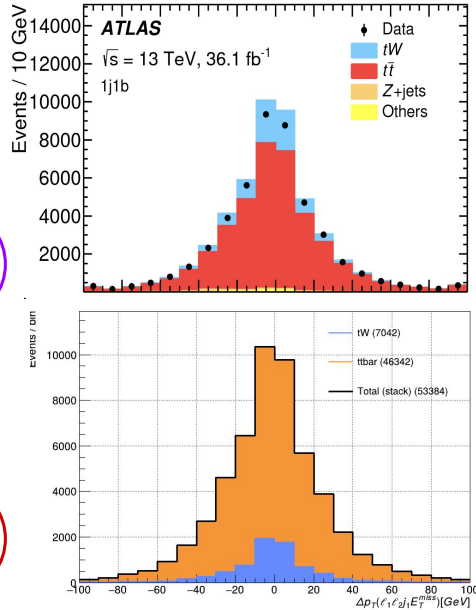
Cutflow:
Все события нормализованы к одинаковой интегральной светимости.

- **tW:** 7042 событий
- **tt̄:** 46342 событий

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 + 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

- **$t\bar{t}$ (топ–антитоп):** 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.

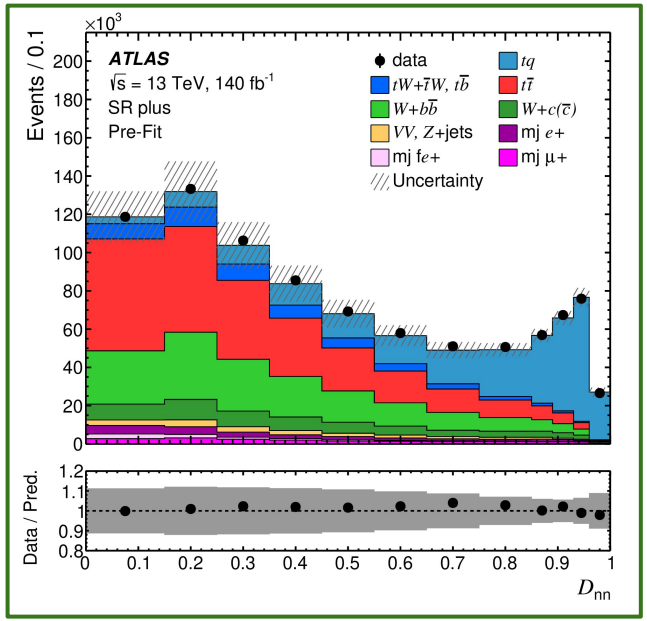
CR name	Requirement
`B-e-plus`	$q_e/e = +1,\ \lvert\eta(e)\rvert < 1.37,\ E_{\mathrm{T}}^{\mathrm{miss}} < 30\,\mathrm{GeV}$
`B-e-minus`	$q_e/e = -1,\ \lvert\eta(e)\rvert < 1.37,\ E_{\mathrm{T}}^{\mathrm{miss}} < 30\,\mathrm{GeV}$
`EC-e-plus`	$q_e/e = +1,\ \lvert\eta(e)\rvert > 1.52,\ E_{\mathrm{T}}^{\mathrm{miss}}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

Table 1: Summary of the definition of the CRs.

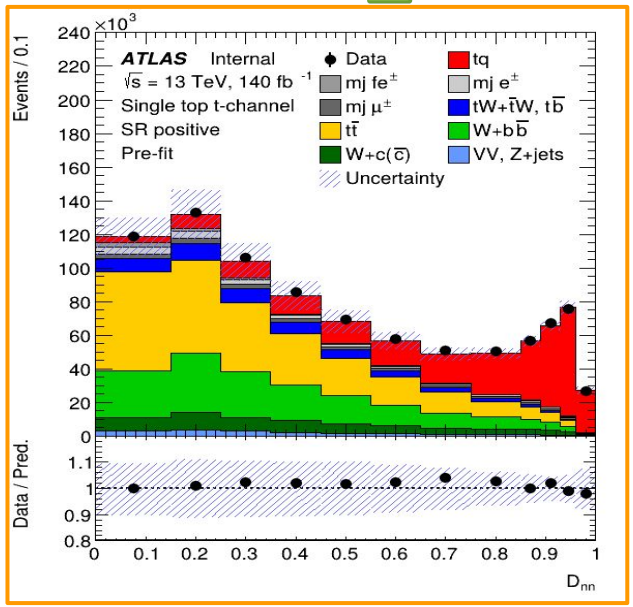
6 контрольных региона

- **Метод на открытых данных HEPData -** *“Measurement of t-channel production of single top quarks and antiquarks”*
  - **Реализованы два подхода:** прямой фит через Trexfitter (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: **SRp, SRn**, SRelep, SRelepforw, SRmuonp, SRelen, SRelenforw, SRmuonn

44 bins distributed across the 8 channels

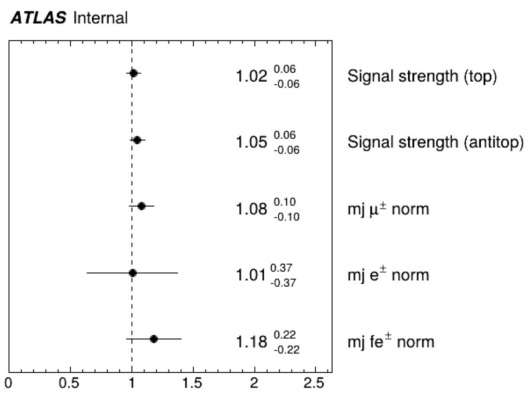
Observed data provided per bin

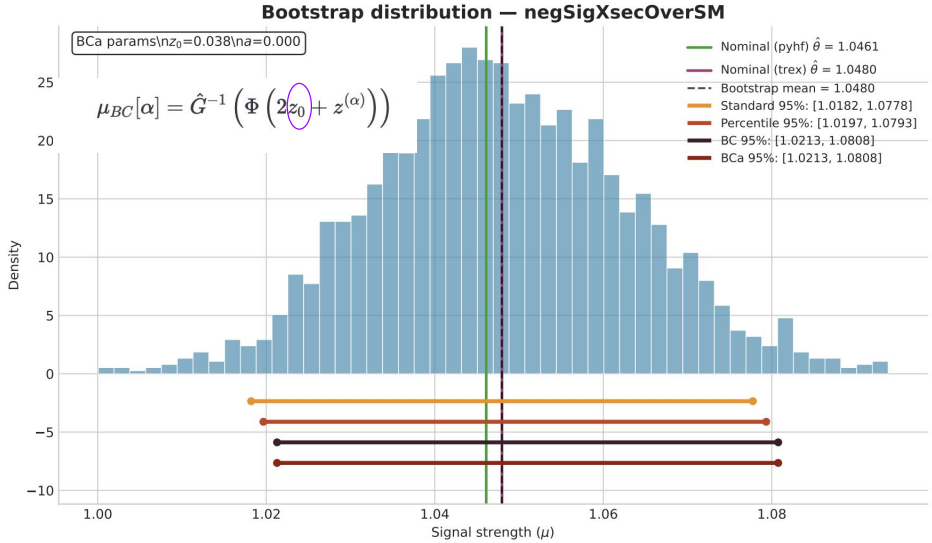
Expected model: signal + background + systematics (400+ nuisance parameters)

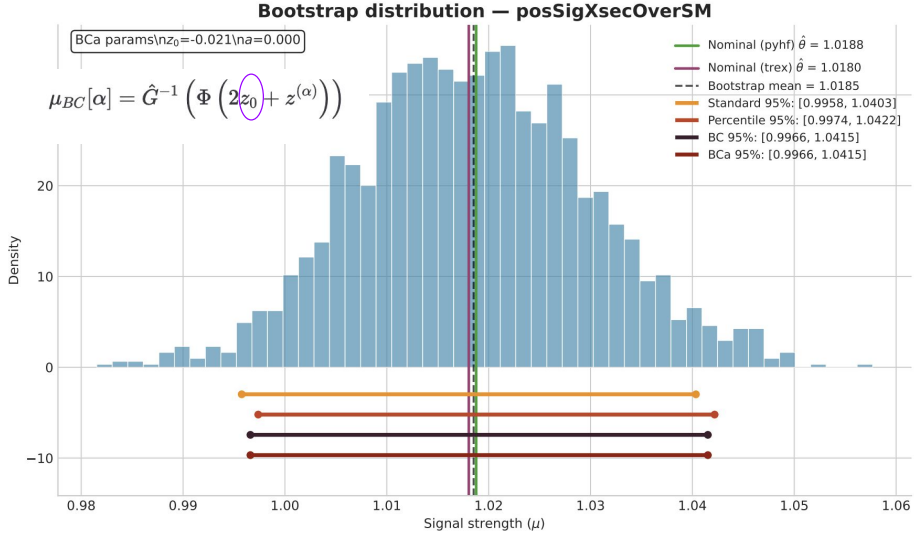
Parameter of Interest (POI): **negSigXsecOverSM**

BCCI

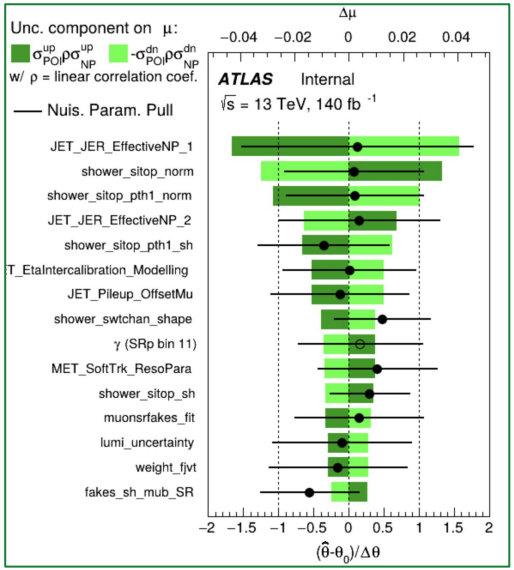
- **2000 bootstrap-реплик** (Poisson), 100% сходимость
- **Доверительные интервалы (μ):** построены **BC/BCa** интервалы→ устойчивые CI при отсутствии асимптотических ошибок

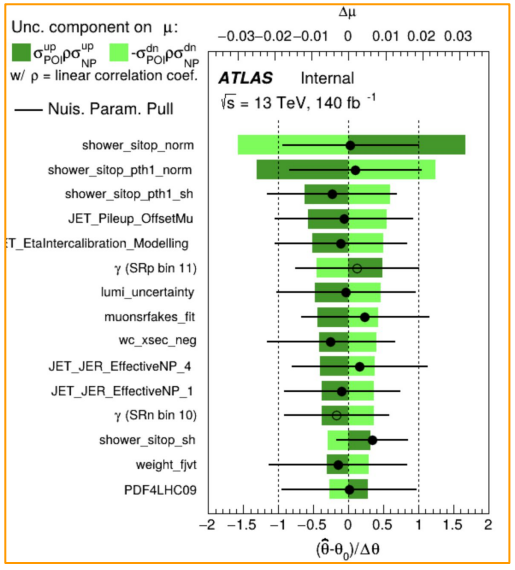


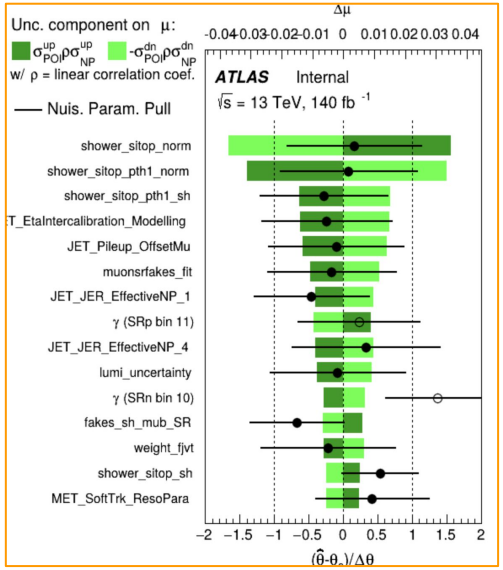


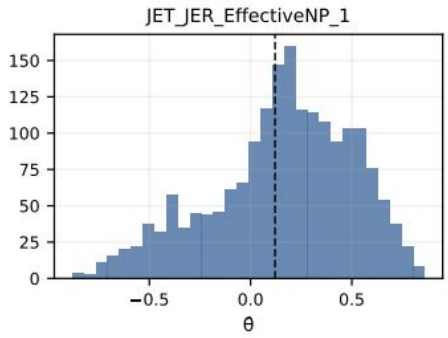


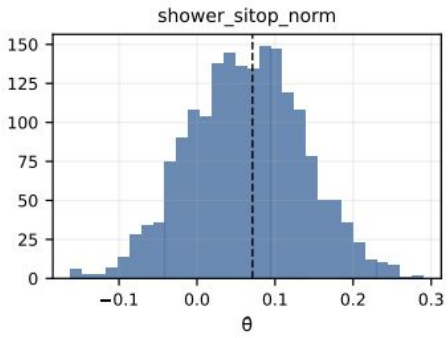
Изучение систематик: preliminary results

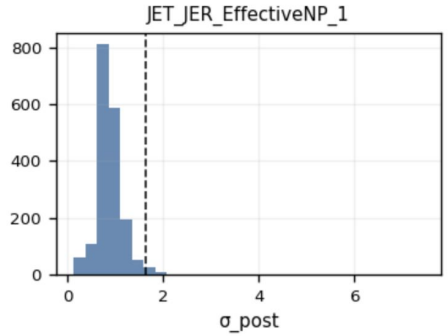


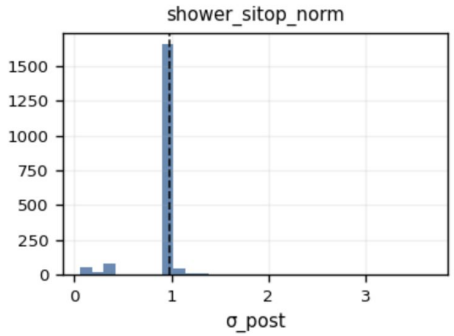










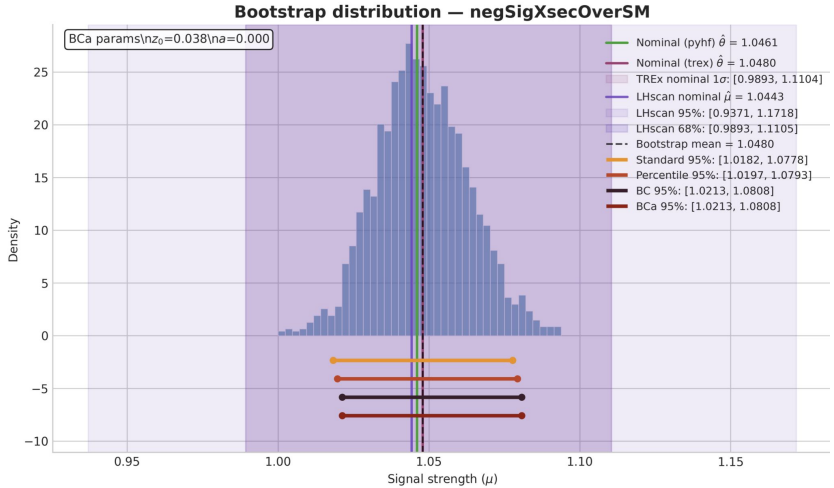


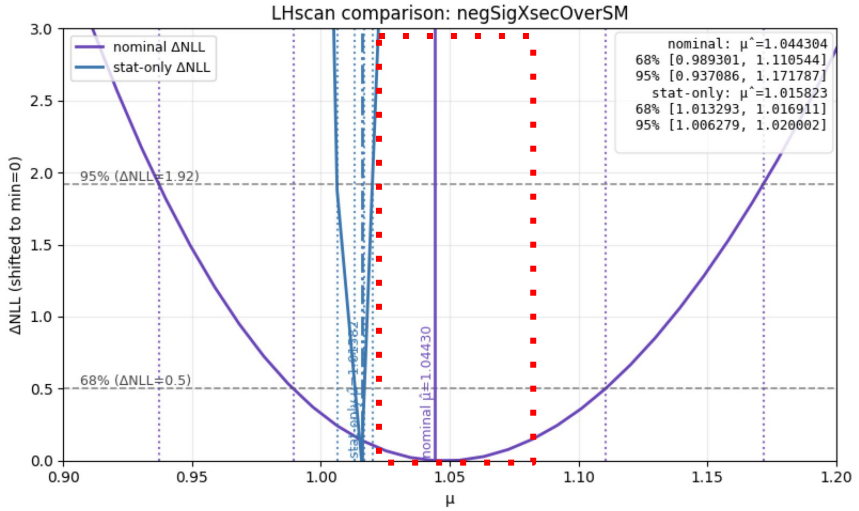
Сравнение интервалов (Trexfitter)

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 pyHF)

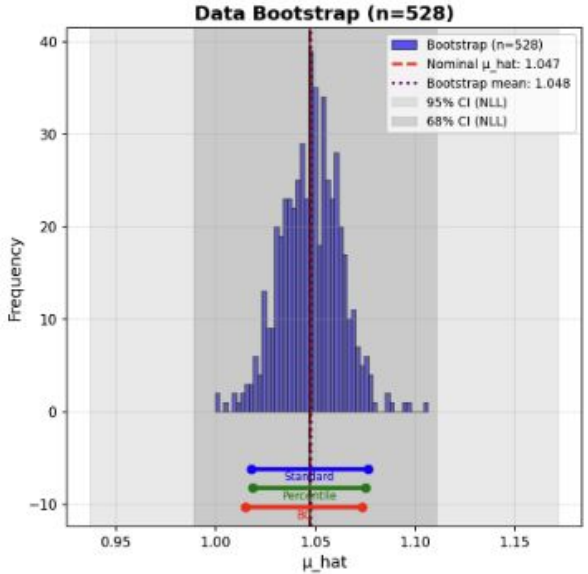
Profile likelihood включает также вклад систематических неопределённостей через профилирование nuisance-параметров θ.
Для корректной проверки полной профилированной ошибки необходимо:

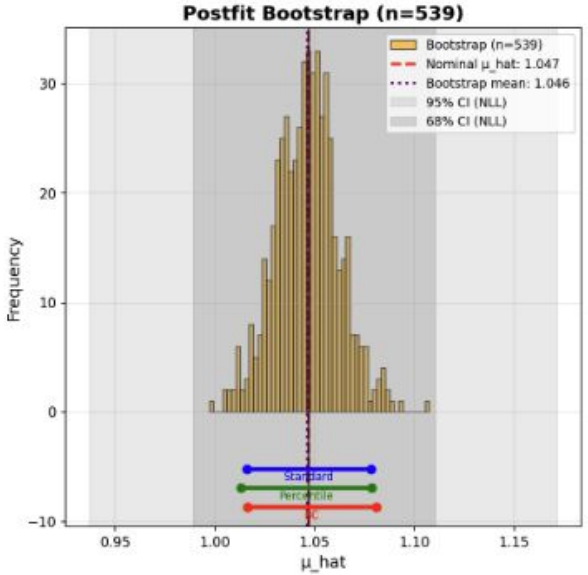
- учитывать пост-фит значения ($\hat{\mu}$, $\hat{\theta}$),
- воспроизводить ожидаемые события при этих параметрах,
- флуктуировать их согласно статистической модели.

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

σ_{stat}=0.00510
σ_{syst}=0.06724
σ_{total}=0.06743





Параметрический бутстрап: NP sampling (preliminary)

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