National Research Tomsk State University

Лаборатория анализа данных физики высоких энергий Томского государственного университета

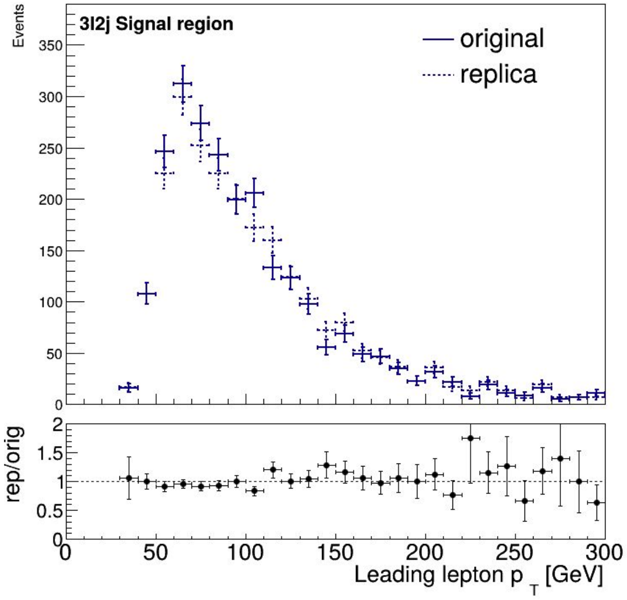
Физический анализ данных

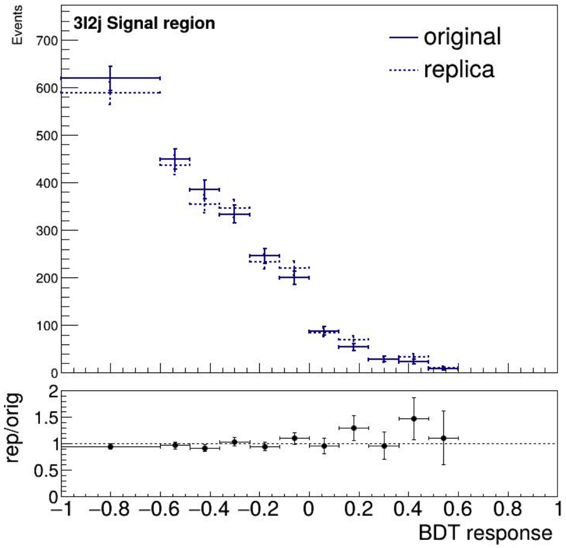
Томский Государственный Университет

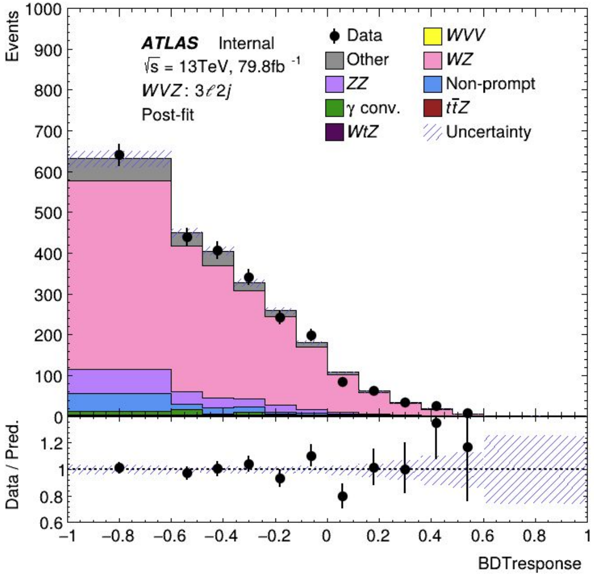
Мария Диденко

Distribution of variables: signal region

- 3l2j SR selection is applied
- Events after SR selection: orig=2438, repl=2407
- The shapes agree within statistical fluctuations (ratio plot: replica/original ≈ 1 ± stat)

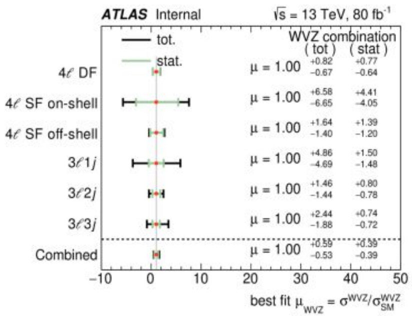


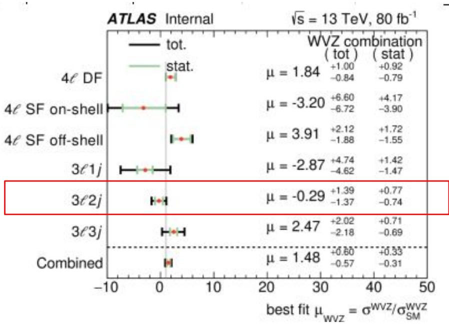


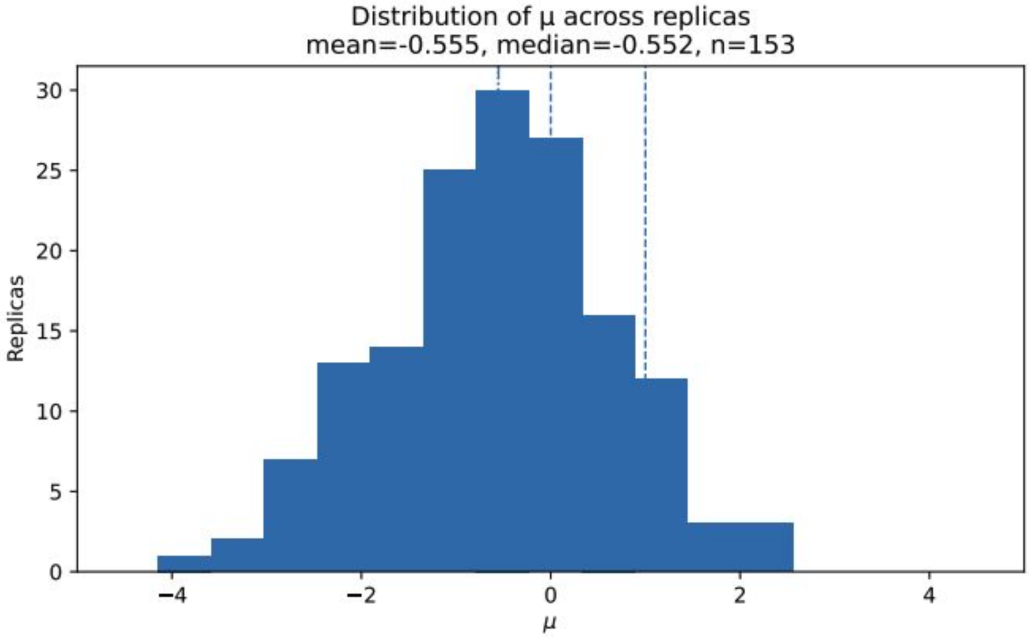


μ distribution

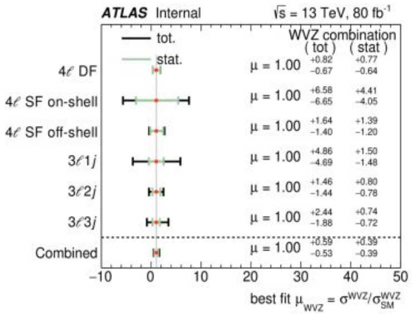
Most of tasks are still running
Checked 153 tasks

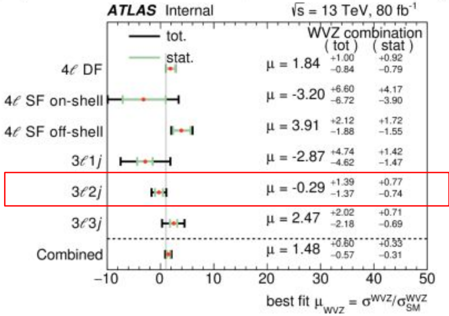


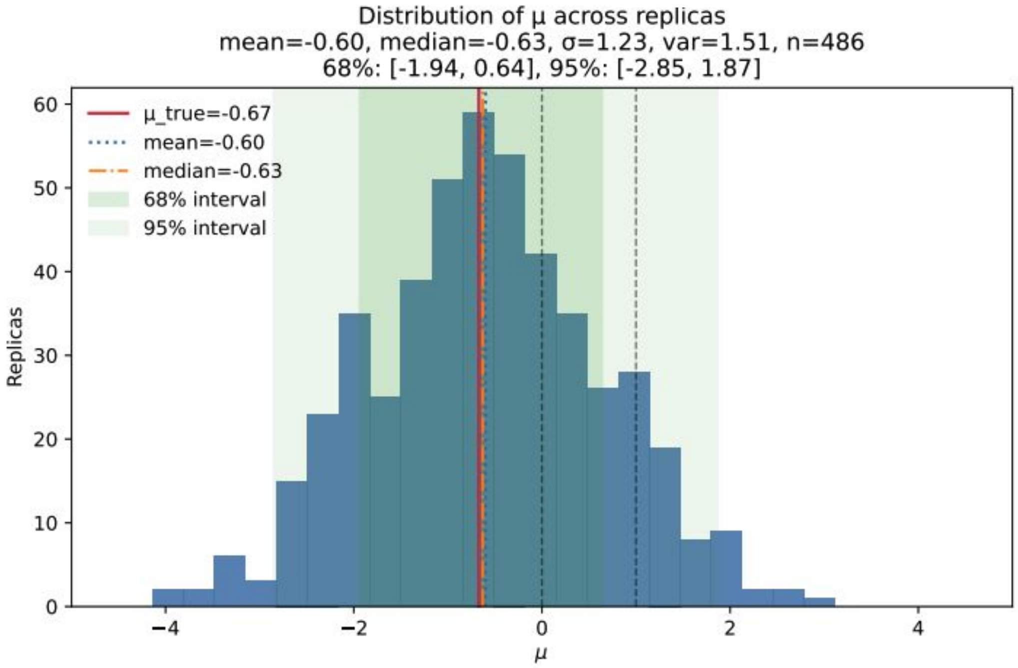




μ distribution







Leptons + at least 3 jets (real data)

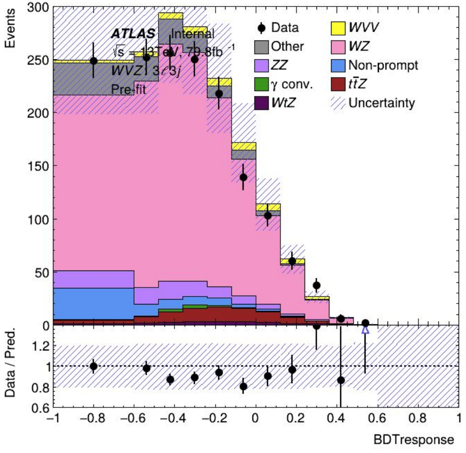
- Checked the remaining **3ℓ + ≥3j** region using **real data**.
- In the configuration file, the variable **`newBDTG_15_3l3j`** was specified, but it was **missing in the ntuples**.
- Used **`newBDTG_32_3l3j_1`** from the ntuples instead.
- The obtained signal strength **(μ = 2.57)** is close to the reference value **(μ = 2.47)**

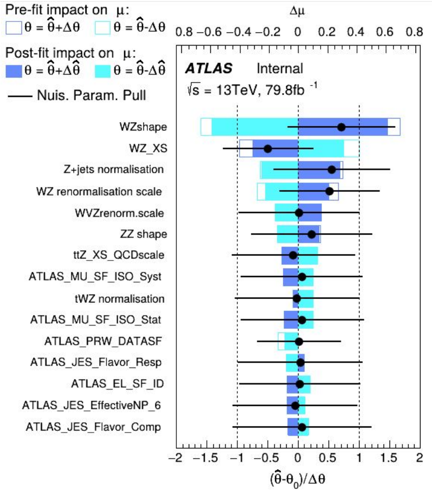
```
###--- 3L preselection && at least 3 jets ---###

Region: three_lep_presel_atLeast_3jets
  Type: SIGNAL
  DataType: DATA
  Variable: newBDTG_15_3l3j,13,-1.,1.
  VariableTitle: BDT response
```

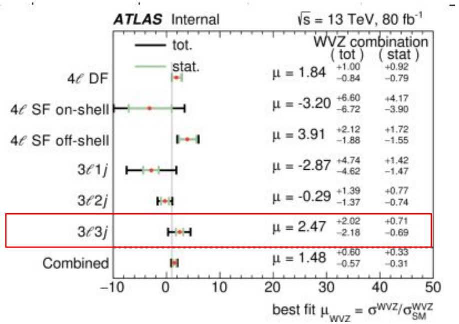
in the config

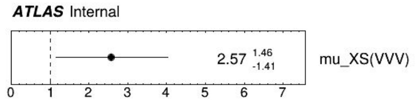
newBDTG_28_3l3j
newBDTG_32_3l3j
newBDTG_32_3l3j_1

in the ntuples



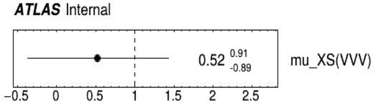


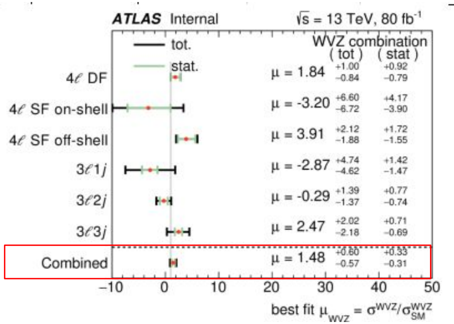


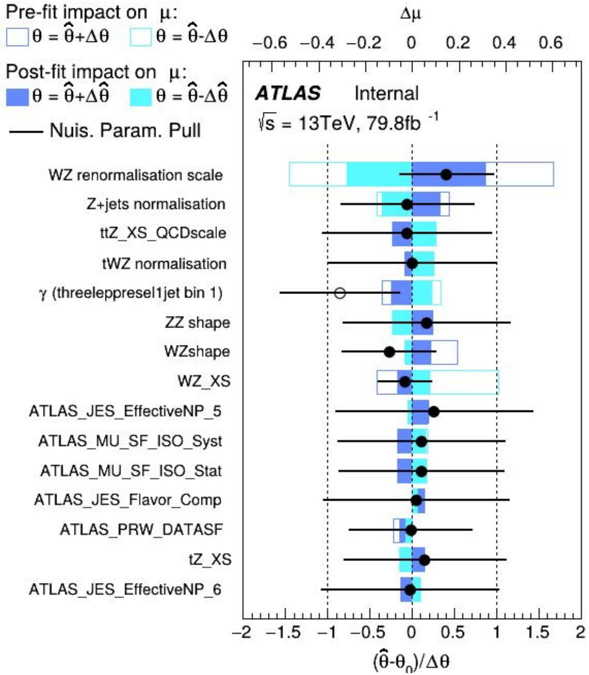


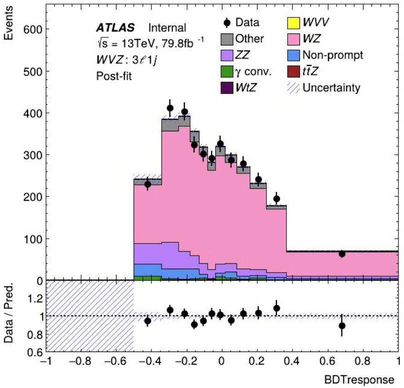
Combined regions (real data)

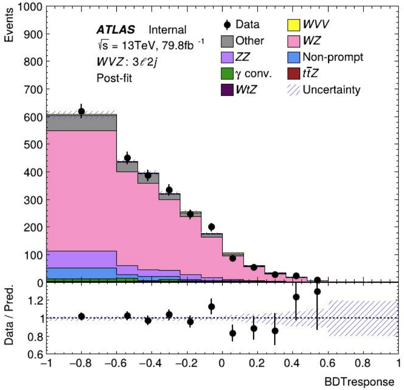
- Combined results from **three regions**: **3ℓ + 1j**, **3ℓ + 2j**, and **3ℓ + ≥3j**.
- Used **real data** to verify consistency of the combined fit.
- The **post-fit BDT distributions** (bottom plots) show good agreement between data and prediction across all regions.
- The obtained **combined signal strength** is **μ = 0.52**, while the **reference value** reported in the publication is **μ = 1.48**:
  - The **difference mainly comes from the 3ℓ + 2j region**:
    - -0.67 vs -0.29

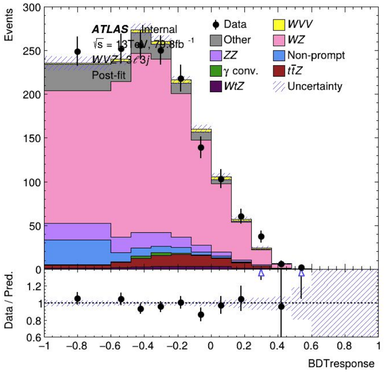






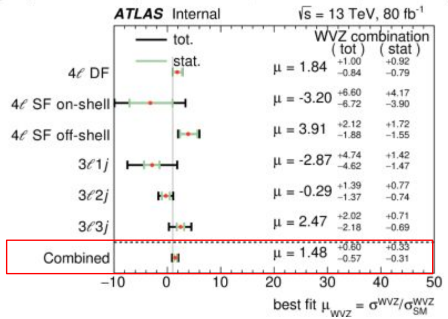


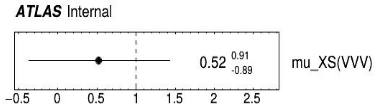


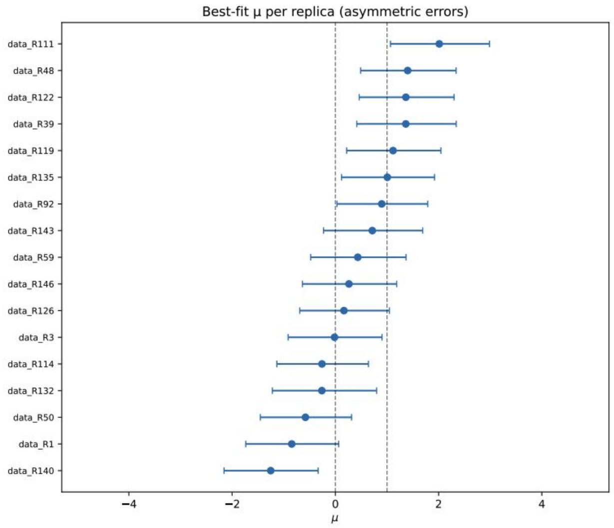


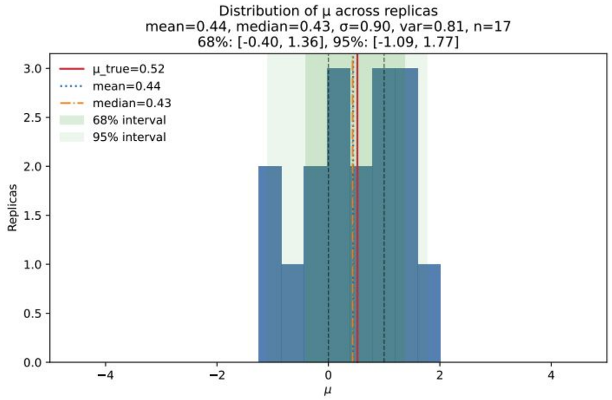
Combined regions (replicas)

- Many replicas **fail during generation**, when TRExFitter tries to run all options simultaneously (**`nwfsdpr`**).
- Currently, I'm running **one option at a time** (**n**), and then plan to process the remaining ones.
- So far, **17 replicas** have been successfully generated.
- The **distribution of μ** is close to the original result (**μ = 0.52**), but **more statistics are needed** to confirm the stability of the result.



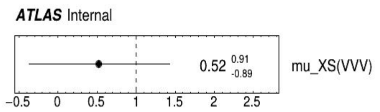


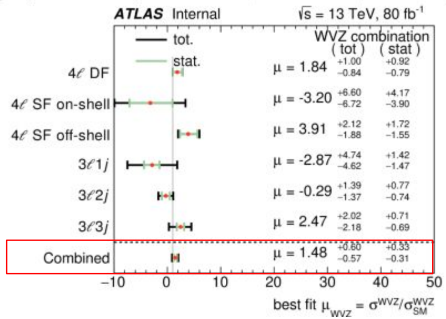




Replicas: 17
μ mean: **0.4429**, median: 0.4341, std: 0.9023
Avg errUp: 0.9335, Avg |errDown|: 0.9045
-- vs μ_true=0.52 --
Bias : -0.07706
RMSE : 0.8787
Std/Var : 0.9023 / 0.8141
Replicas coverage 68% : 76.5%
Replicas coverage 95% : 100.0%
Global 68% CI [-0.402, 1.36] => hit? True
Global 95% CI [-1.09, 1.77] => hit? True

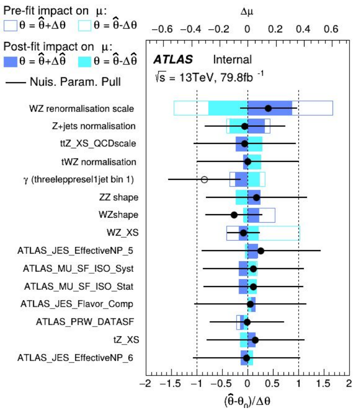
Combined regions (replicas)



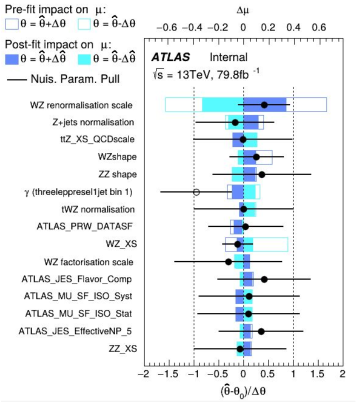


Each replica includes random statistical variations in data, which slightly change the fitted nuisance parameters and their impact on μ.

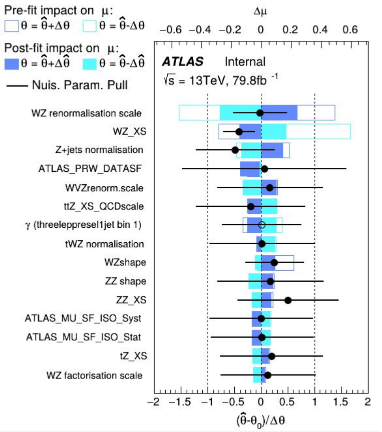
Real Data



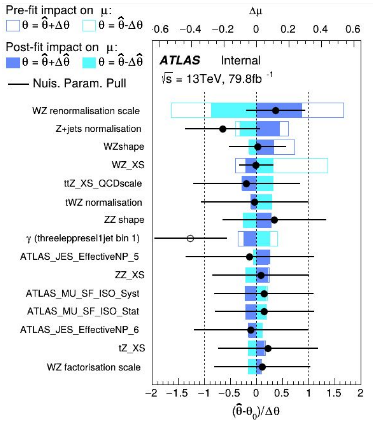
R3



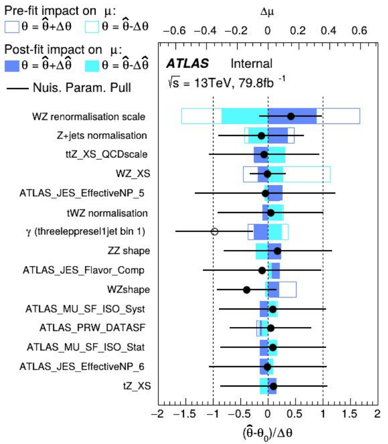
R39



R48



R50



HEP data

Measurement of the total and differential cross-sections of ttW production: a good starting point

- Ready bootstrap replicas already included in the workspace ✅
- Working fitting scripts available ✅
- Regularization disabled, fitting launched (XRooFit) ✅
  - evaluate the statistical variation of the data
- Running 1000 replicas: in progress ⏩

Fit results (without regularization):

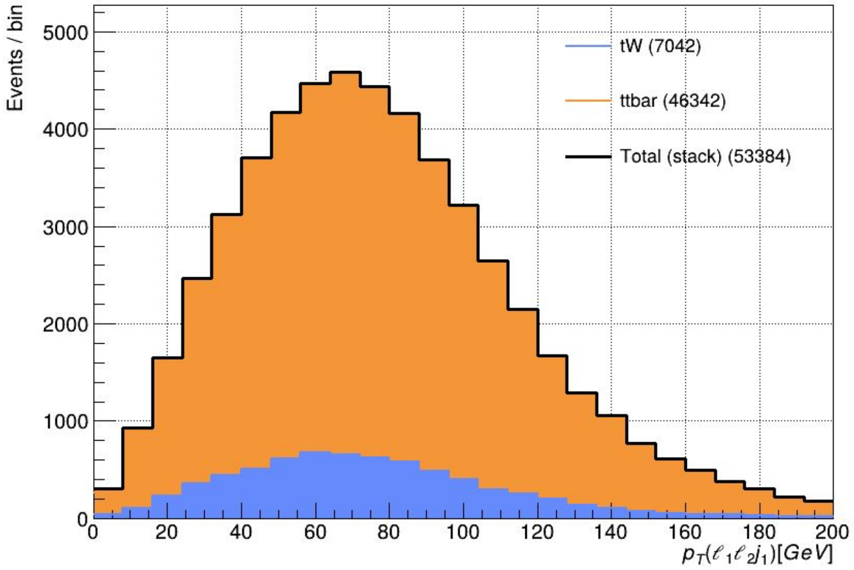
- NLL: 454.056
- Norm_ttW = 1.1393 ± 0.0932

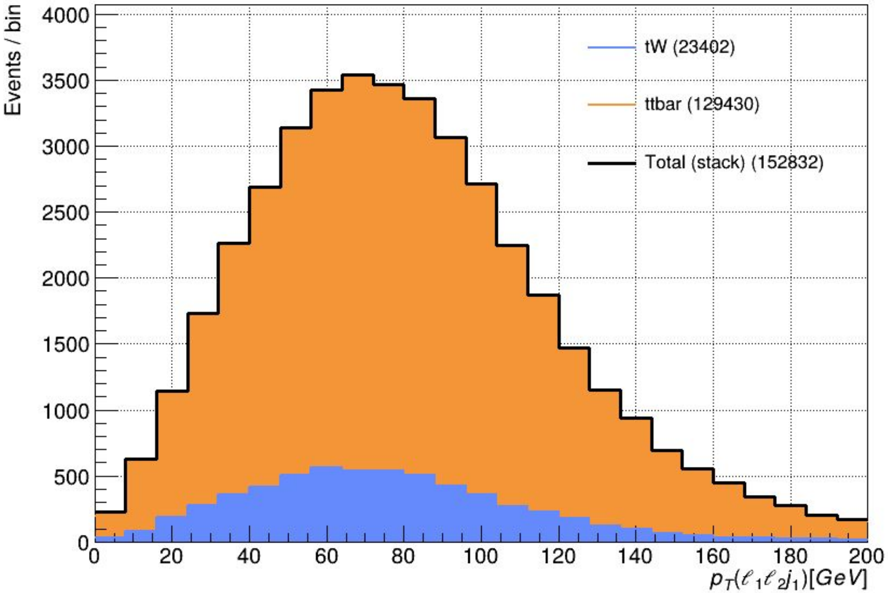
Measurement of t-channel production of single top quarks and antiquarks

- JSON file with data: understanding of its structure is required (or conversion JSON → ROOT workspace)
- Several fitting options available: PyHF or TRExFitter
- repeating the ttW structure: conversion JSON → YAML workspace

BDT ntuples

- **Additionally:** the ntuples for the BDT are ready, and the statistics have been increased by a factor of three.

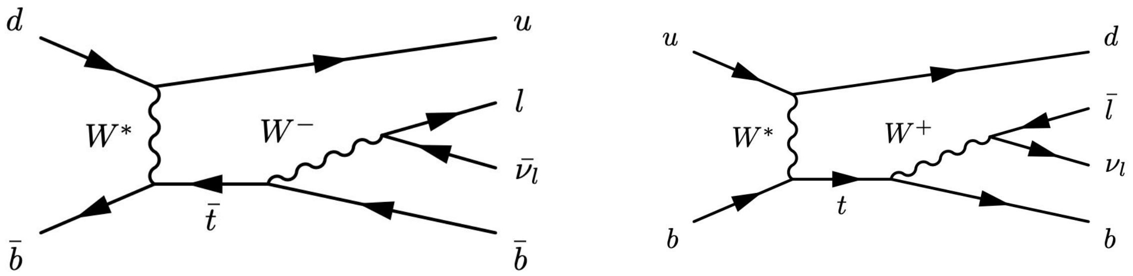




16.12.2025

tW analysis overview

- This analysis uses **140 fb^{-1} of ATLAS 13 TeV** data to measure the production cross-sections of tW processes separately **for tq → tW^{+} and t$\bar{q}$ → tW^{-} final states.**
  - The separate measurements provide enhanced sensitivity to the u- and d-quark PDFs, since the dominant initial states differ for tW^{+} (u → d transition) and tW^{-} (d → u transition).
- Events are selected in the single-lepton final state with one charged lepton, MET, and b-tagged jets.
- A **neural network (NN)** is trained to distinguish tW signal from background using event-level kinematic variables.
- The **NN output** distribution is then used as the discriminant **in a profile likelihood fit.**



Region strategy

Two complementary signal regions tailored to the angular correlation between the lepton and the b-jet:

- **SR-plus (SRp):** Events where the lepton and the b-jet are preferentially aligned (sensitive to **tW^+ production**).
- **SR-minus (SRn):** Events with opposite angular correlation (sensitive to **tW^- production**).

This separation increases PDF sensitivity and improves constraints on the signal model.

Main Background Processes

- **$t\bar{t}$ (top–antitop):** dominant background in single-lepton, b-jet final states.
- **Single top (tW, t-channel):** important and must be modelled accurately, especially in 1-b-tag categories.
- **W+jets:** critical for regions with one lepton and MET.
- **Z+jets / Diboson:** typically subdominant but included.
- **Fake leptons / charge mis-ID:** included where relevant.

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

Table 1: Summary of the definition of the CRs.

6 Control regions

BCCI implementation

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**

```
{
  "channels": [
    {
      "name": "SRp",       // Название канала (сигнальный регион)
      "samples": [         // Физические процессы
        {
          "name": "signal",           // Сигнальный процесс
          "data": [0.1, 0.2, ...],    // Ожидаемые события по бинам
          "modifiers": [...]          // Систематические вариации
        },
        {
          "name": "background",
          "data": [10.5, 15.3, ...],
          "modifiers": [...]
        }
      ]
    }
  ]
}
```

Full workflow

1. Load the JSON workspace
2. Generate Poisson bootstrap replicas
3. Modify the observed data in the workspace
4. Build the pyhf statistical model
5. Perform a maximum likelihood fit (MLE)
6. Extract the POI value and uncertainties
7. Repeat the procedure 1000 times
8. Compute BC / BCa confidence intervals

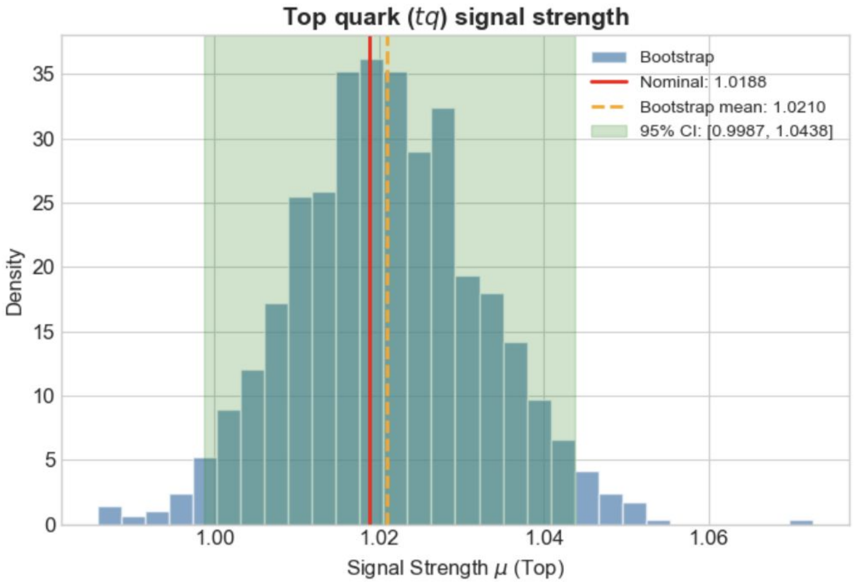
From the HEPData workspace (`workspace_fixed.json`):

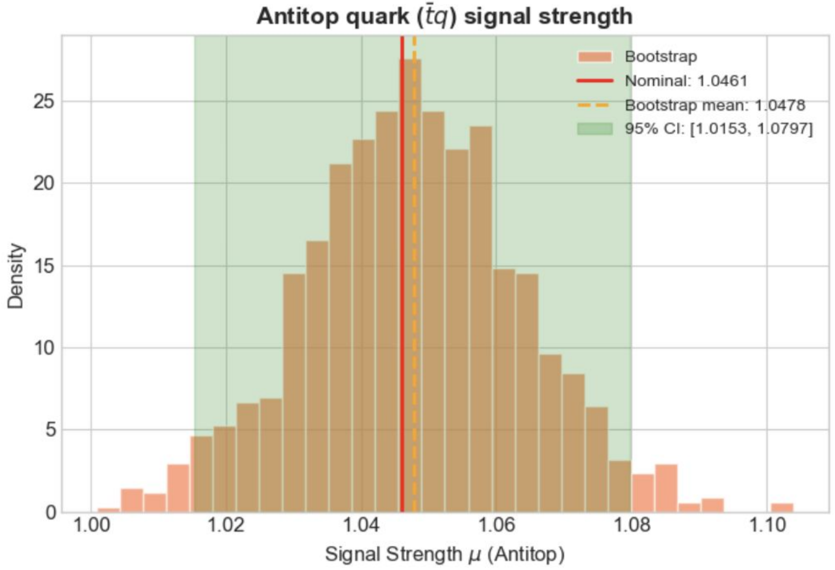
Type	Count	TRexFitter Type	Examples
lumi	1	LUMI	`lumi`
staterror	8	STATERORR	`staterror_SRp`
normsys	208	OVERALL	`sitop_mur`, `JET_*`
histosys	207	HISTO	`weight_bTagSF_*`
normfactor	5	NormFactor	`negSigXsecOverSM`

Total: 429 systematic parameters

BCCI implementation

- **Data** [n1,n2,n3,…] per bin → **varied**
- **Systematic uncertainties:** all normsys, histosys, luminosity, etc. → **fixed**
- **Likelihood model:** includes **nuisance parameters**

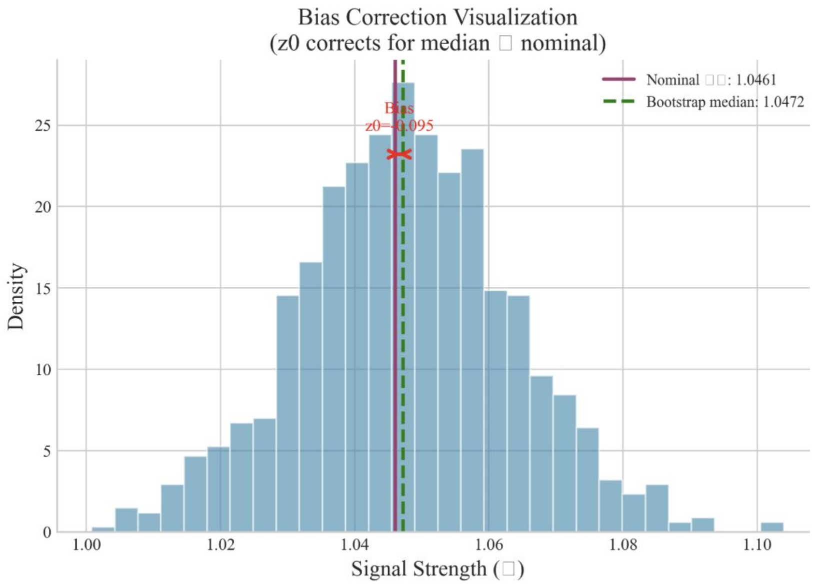


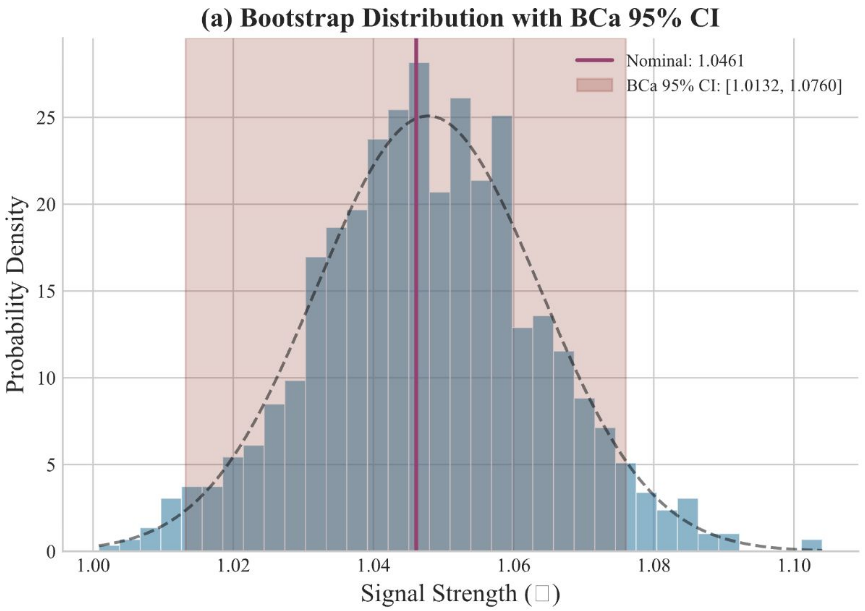


BC/BCa implementation

$$\mu_{BC}[\alpha] = \hat{G}^{-1}\left(\Phi\left(2z_0 + z^{(\alpha)}\right)\right)$$

```
Antitop quark (t̄q):
  z0 = -0.0954, a = 0.0112
  Standard     : [1.0166, 1.0790]  (width=0.0624)
  Percentile   : [1.0153, 1.0797]  (width=0.0644)
  BC           : [1.0132, 1.0760]  (width=0.0628)
```



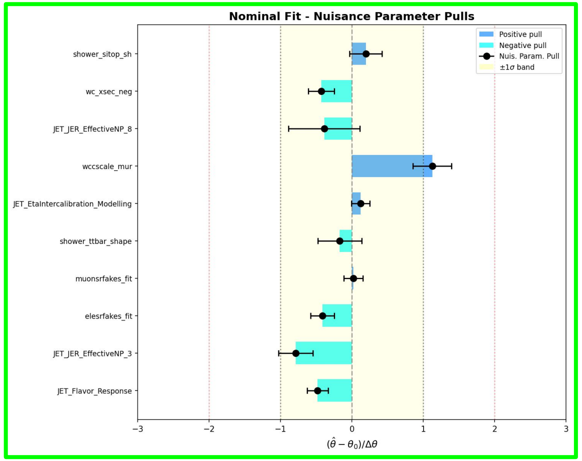


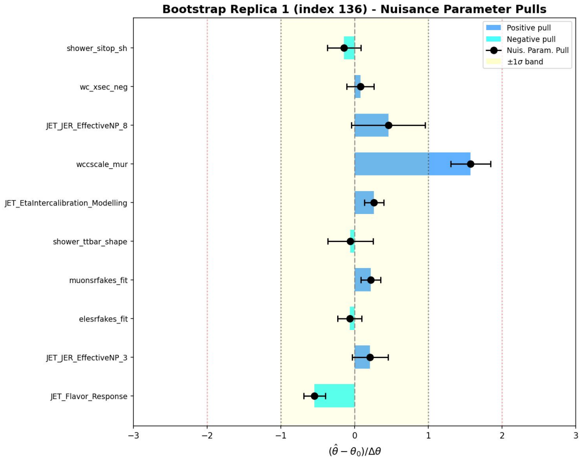
BCCI implementation

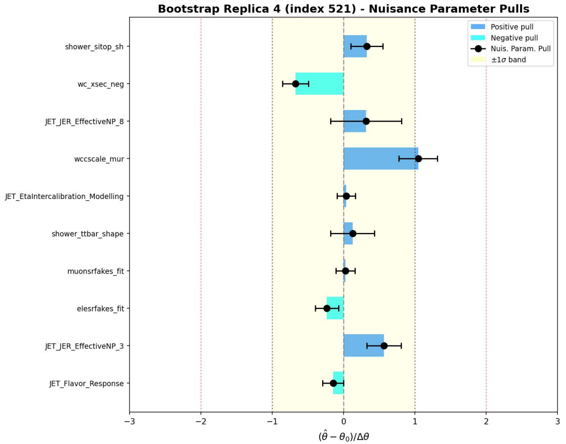
The ranking of nuisance parameters shows a **clear data-dependent behavior** of systematic uncertainties.

The **Jet_JER_Effec…** nuisance parameter is particularly **sensitive to data fluctuations** and may require a more accurate treatment or dedicated validation.

A large fraction of systematic uncertainties shows **negligible impact** on the likelihood and can potentially be **safely neglected** in simplified models.





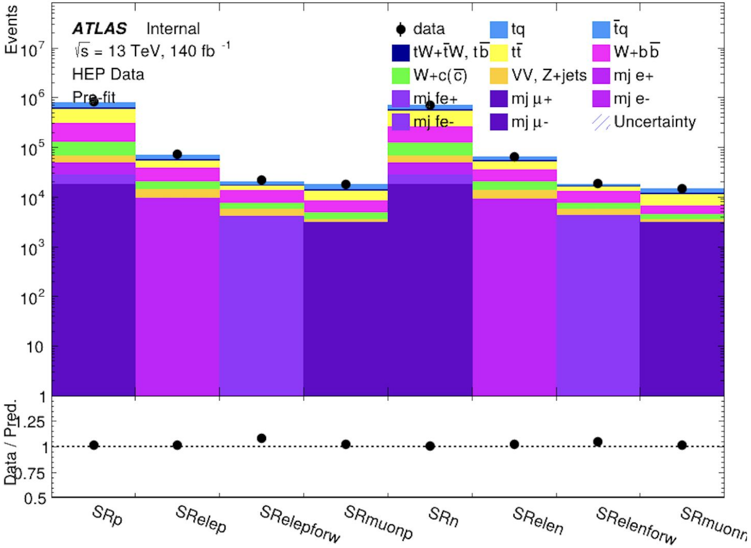


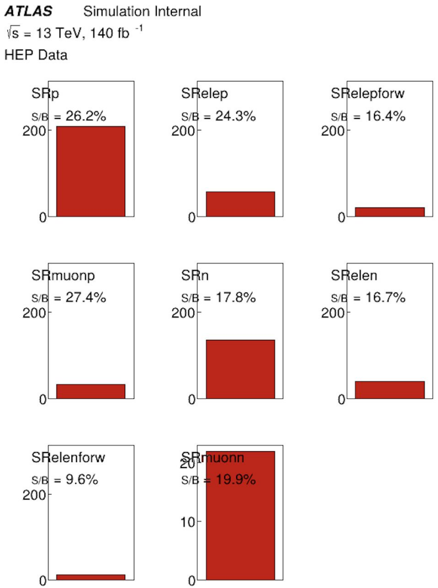
Trexfitter cross-check

- **JSON → ROOT conversion:** the converted ROOT file currently contains **only histograms**, with **variable bin widths** matching the reference.
- **TRExFitter setup:** The fit was run in **StatOnly = TRUE** mode, i.e. **systematic uncertainties were disabled** and only statistical uncertainties were considered.
- **Next step:** To reproduce the full result, we must include **all systematic uncertainties** in the workspace (400+ nuisance parameters), i.e. enable normsys, histosys, luminosity, etc., and run the **full profile-likelihood fit** in TRExFitter.

NUISANCE_PARAMETERS

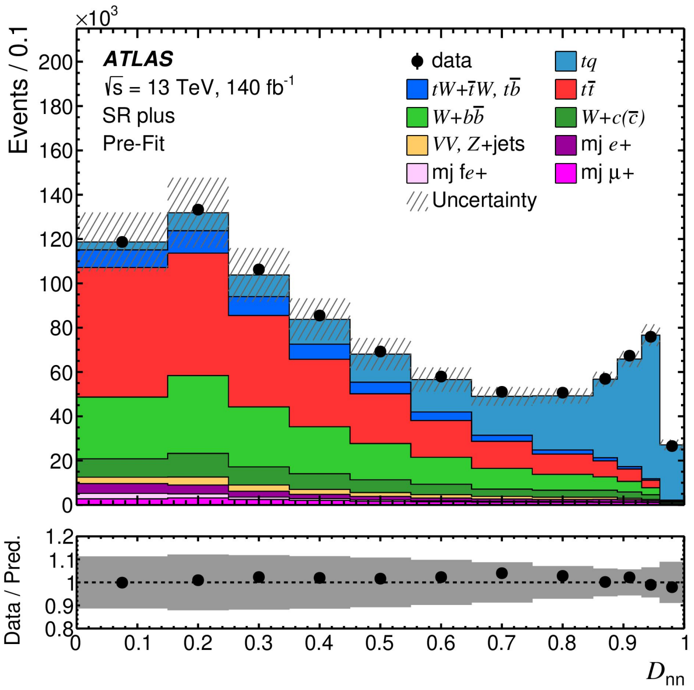
mu_tbarq: **1.0476** +0.005 -0.005
mu_tq: **1.03865** +0.0035 -0.004

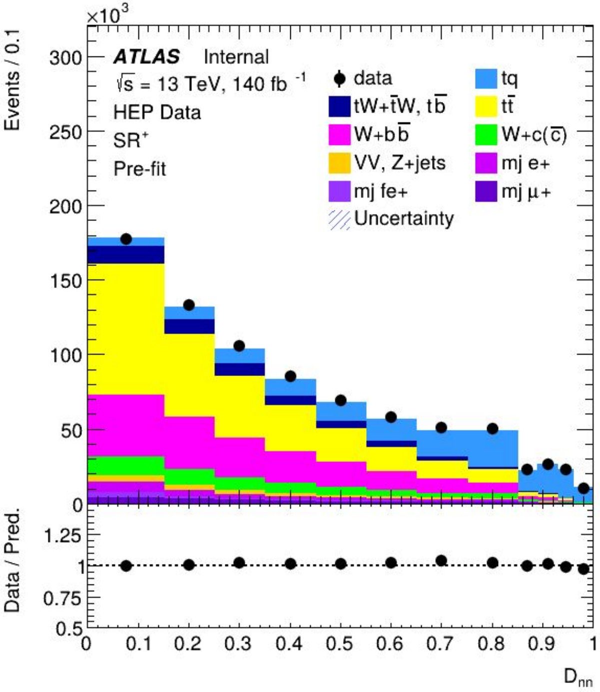




Issue

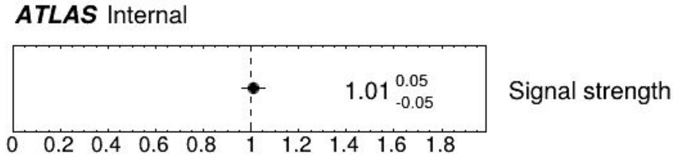
Different distribution: first and last bins don't show correct number of events.





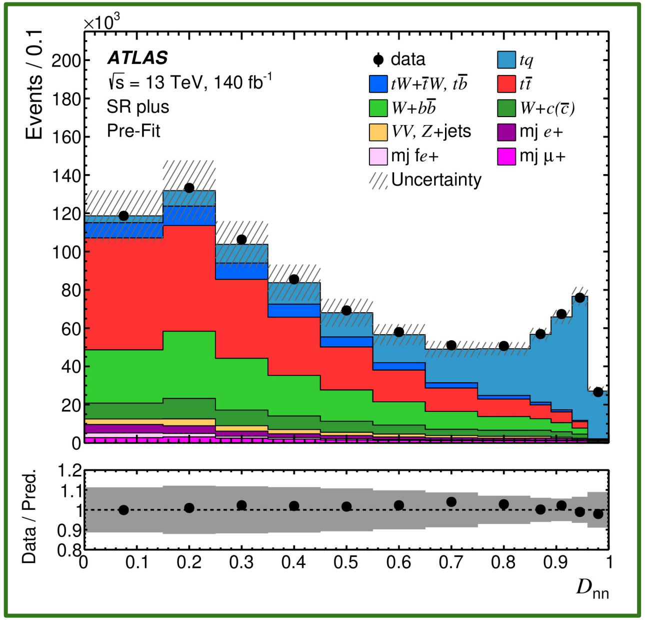
BCCI: Trexfitter

both SRs are included

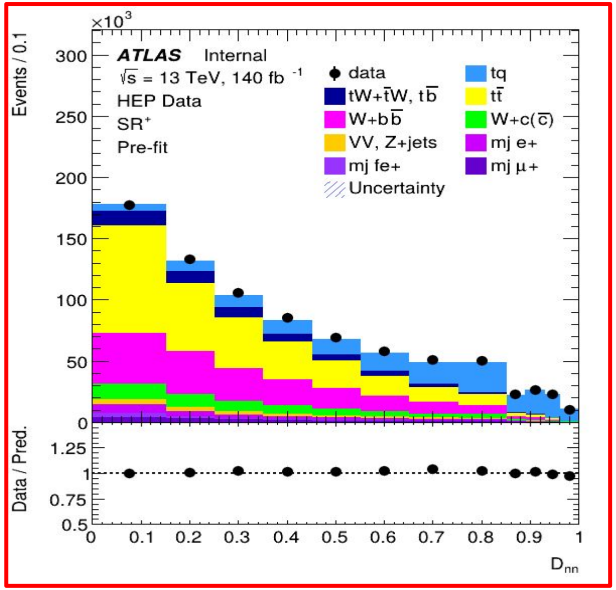


- Issue with first and last bins was fixed by applying correct normalisation
- All sources of systematics were included: at first look everything is OK
- The distribution of DNN is the same as original

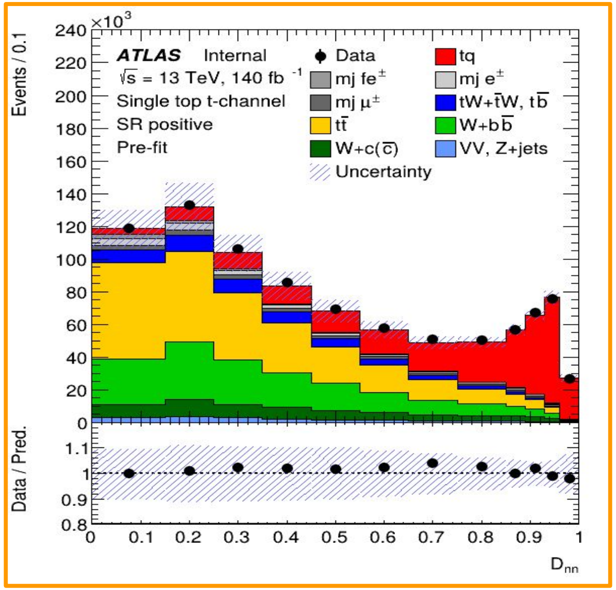
Reference Article



Issue

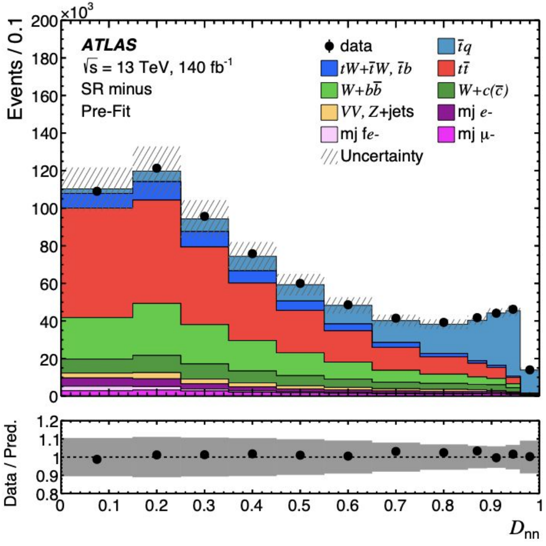


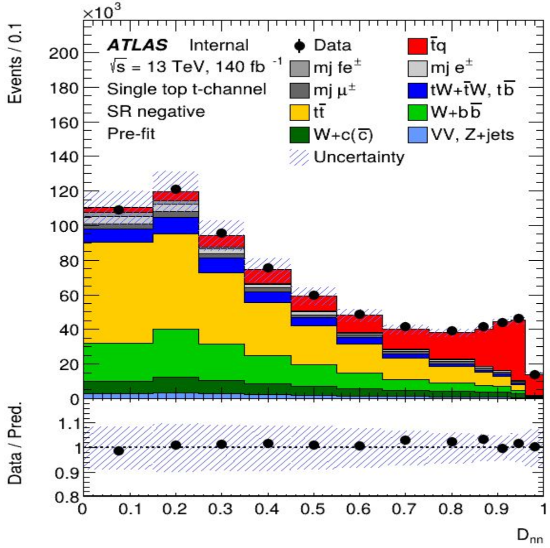
Fixed ✅



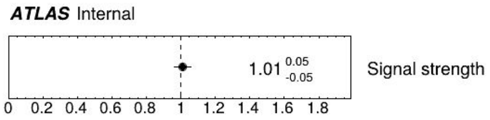
Trexfitter: SR minus

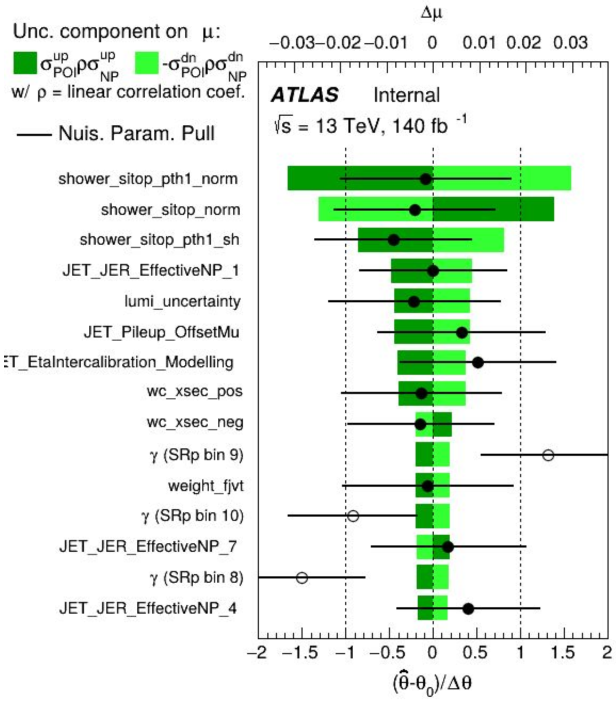
```
Job: "HEPData_tchan"
  Label: "Single top t-channel"
  CmeLabel: "13 TeV"
  LumiLabel: "140 fb^{-1}"
  POI: "mu_sig"
  ReadFrom: HIST
  HistoPath: "/home/users/mdidenko/HEPData_BBCI/hepdata_bcci/workspace"
  HistoFile: "histograms"
  OutputDir: "results/trexfitter"
  DebugLevel: 1
  SystControlPlots: TRUE
  SystErrorBars: TRUE
  RankingMaxNP: 15
  SystPruningShape: 0.005
  SystPruningNorm: 0.005
  SystLarge: 1
  CorrelationThreshold: 0.0001
  SystDataPlots: TRUE
  SystControlPlots: TRUE
  SystDataPlots: TRUE
  UseGammaPulls: TRUE
  PlotOptions: CHI2,NOXERR,NOENDERR,POISSONIZE
  RatioYmin: 0.8
  RatioYmax: 1.2
```

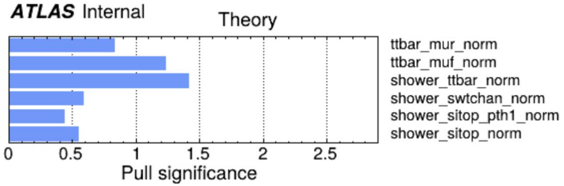


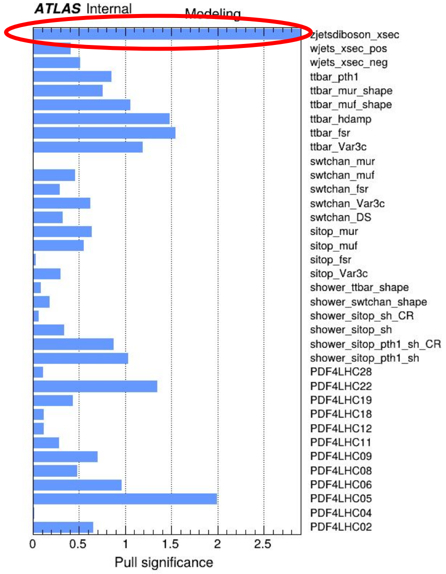


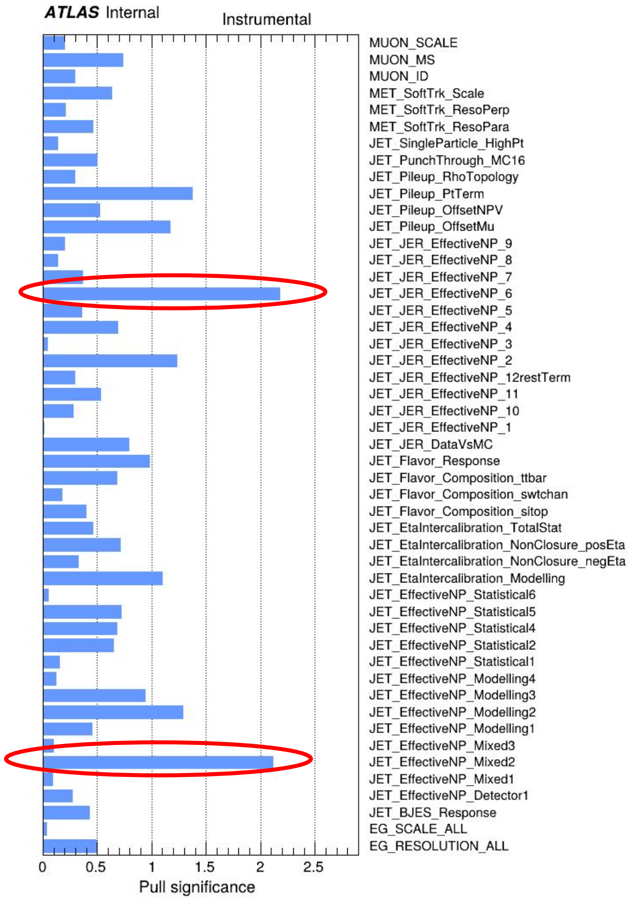
Trexfitter results





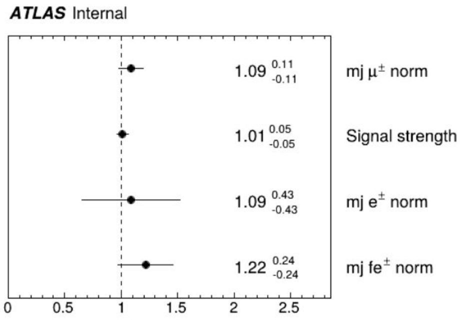


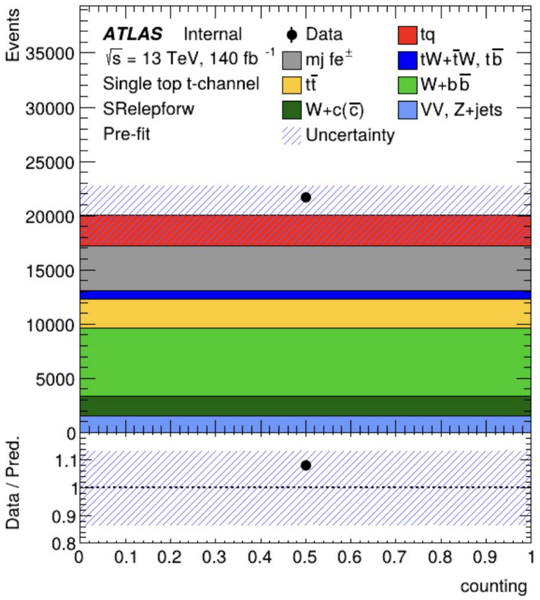


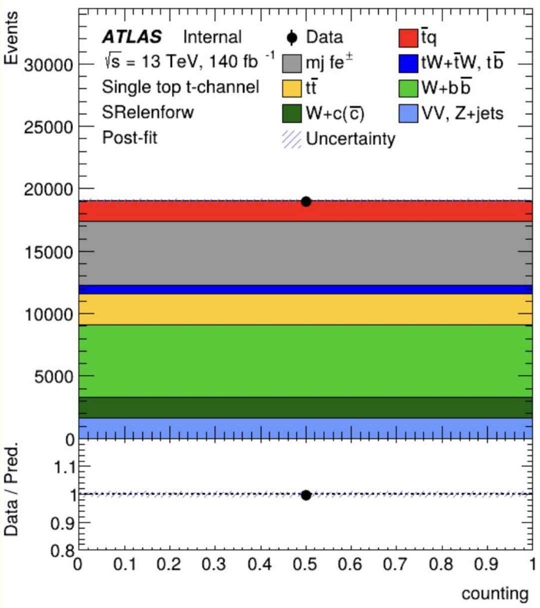


Trexfitter: normfactors

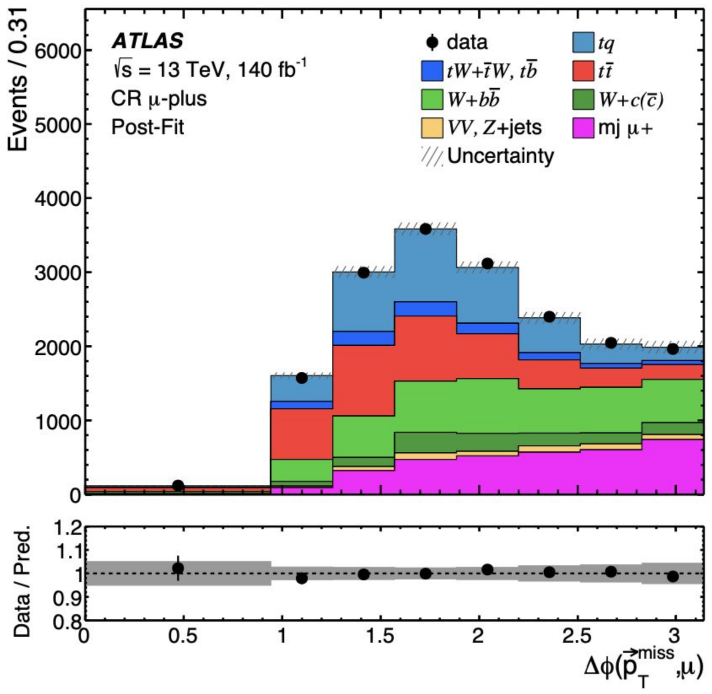
```
"measurements": [
  {
    "config": {
      "parameters": [
        {
          "fixed": true,
          "name": "lumi"
        },
        {
          "fixed": false,
          "name": "elesrfakes_norm"
        },
        {
          "fixed": false,
          "name": "eleforwsrfakes_norm"
        },
        {
          "fixed": false,
          "name": "muonsrfakes_norm"
        }
        // ... другие параметры
      ],
      "poi": "negSigXsecOverSM"
    },
    "name": "vb23"
  }
]
```

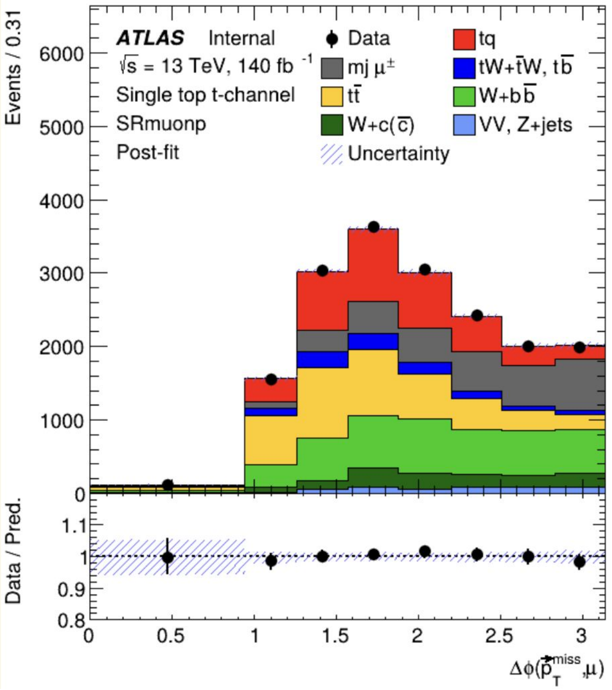




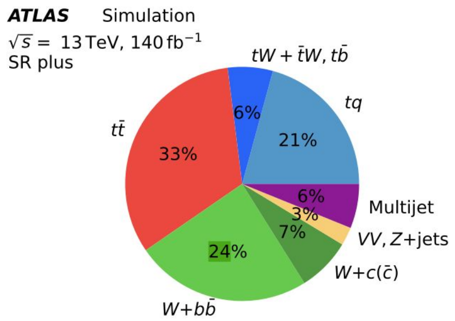


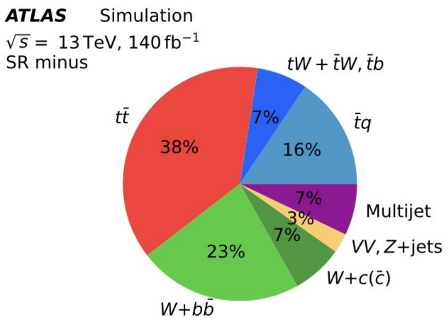
Trexfitter: CR

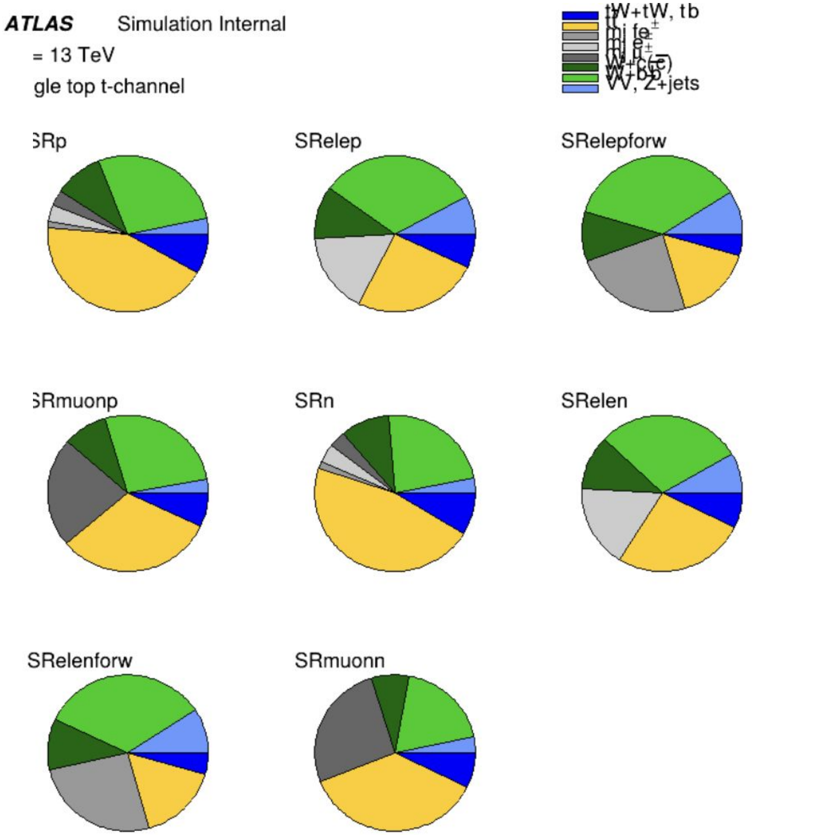




Trexfitter:


ATLAS Simulation
$\sqrt{s}$ = 13 TeV, 140 fb^{-1}
SR plus
$tW + \bar{t}W, t\bar{b}$
tq
$t\bar{t}$
6%
21%
33%
6%
3%
7%
24%
Multijet
VV, Z+jets
$W+c(\bar{c})$
$W+b\bar{b}$


ATLAS Simulation
$\sqrt{s}$ = 13 TeV, 140 fb^{-1}
SR minus
$tW + \bar{t}W, \bar{t}b$
$\bar{t}q$
$t\bar{t}$
7%
16%
38%
7%
3%
7%
23%
Multijet
VV, Z+jets
$W+c(\bar{c})$
$W+b\bar{b}$


ATLAS Simulation Internal
= 13 TeV
gle top t-channel
tW+tW, tb
VV, Z+jets
SRp
SRelep
SRelepforw
SRmuonp
SRn
SRelen
SRelenforw
SRmuonn

Thank you!