MURAVES
towards full pythonisation – tracking bug and running time

2026-04-14

***Alice Biolchini**, CP3, UCLouvain*

Pic: https://www.vesuviusnationalpark.it/

Pythonisation of reconstruction script

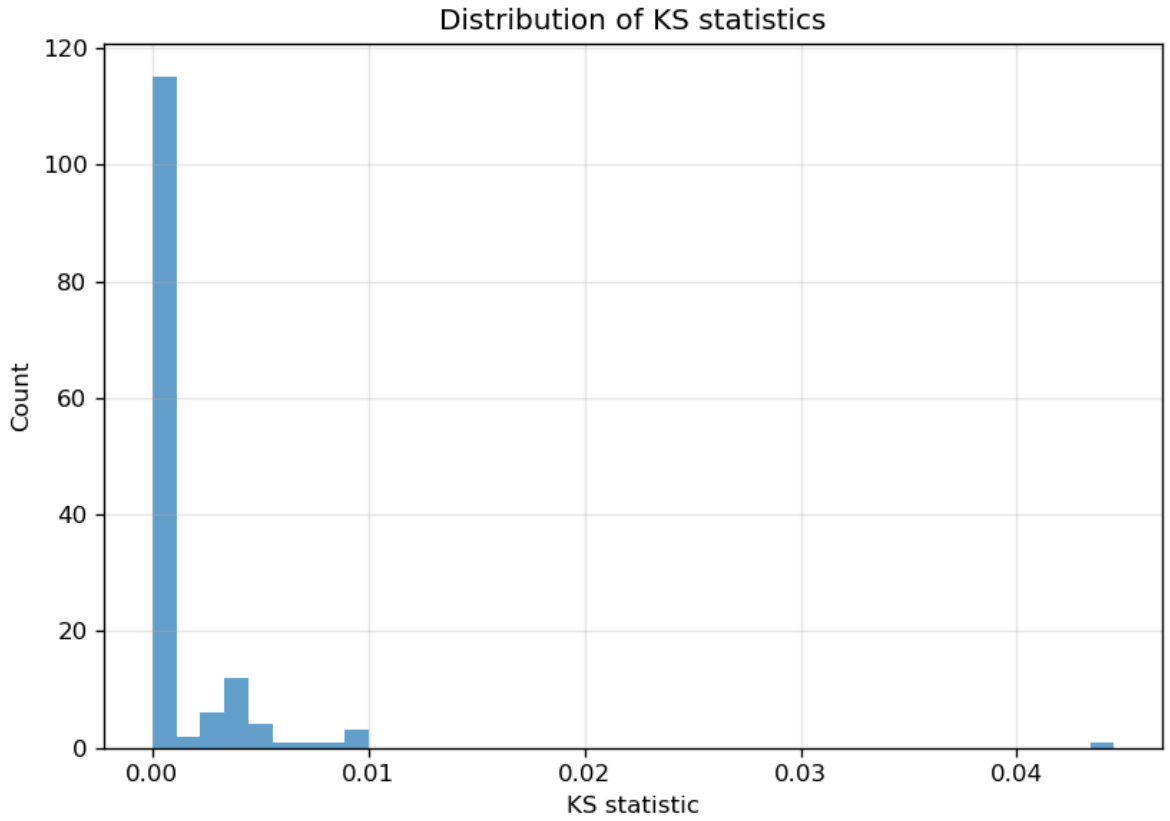
- Python version of the main reconstruction scripts is available on MURAVES GitHub:
  kept same structure as before.
  - **MURAVES_reco.cpp → MURAVES_reco.py**
  - **ReadEvent.cc and ReadEvent.h → ReadEvent.py**
  - **ClusterLists.cc and ClusterLists.h → ClusterLists.py**
  - **Tracking.cc and Tracking.h → Tracking.py**
  - **EvaluateAngularCoordinates.cc and .h → EvaluateAngularCoordinates.py**

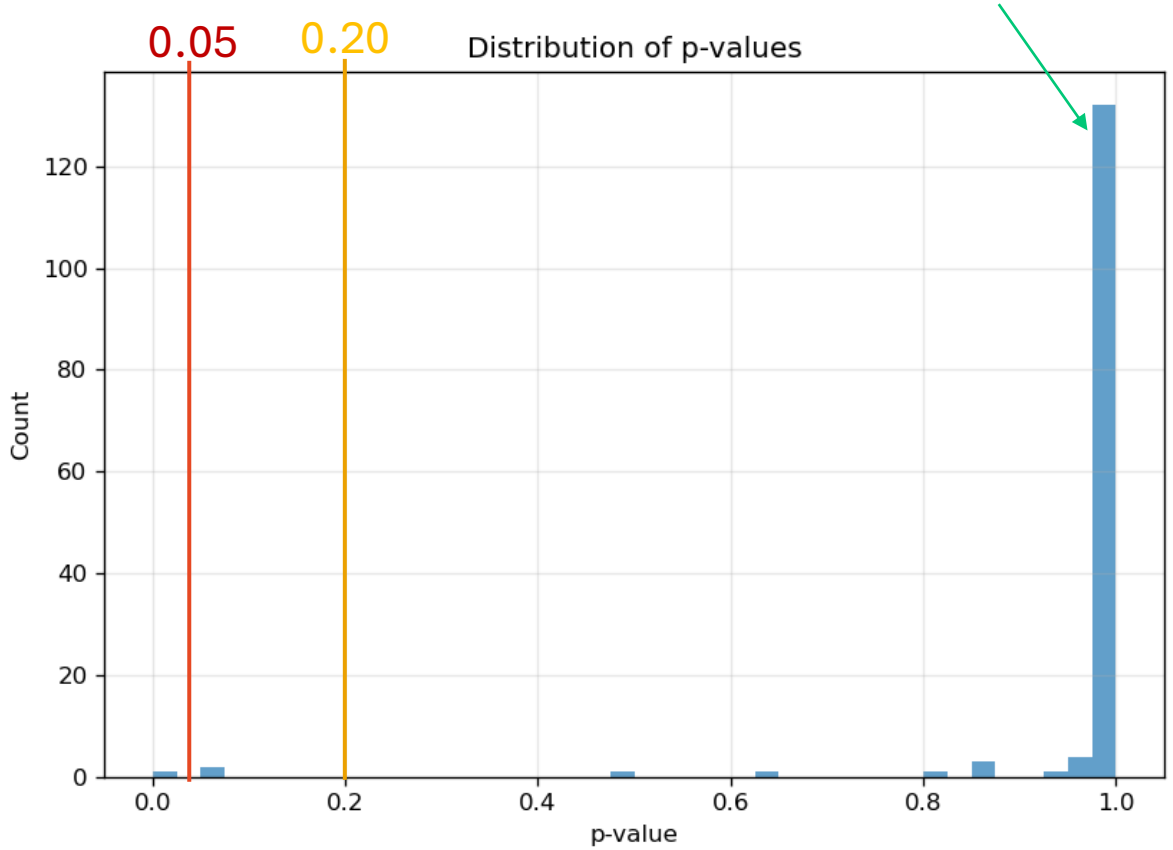
Validation

- Comparing output variables **Kolmogorov-Smirnov test**: unbinned test (larger distance between cumulative distributions).
- **Significance at 5%:** *meaning it is very unlikely that the two distributions comes from the same underlying distribution*.
- **Run NERO 2546:**
  - **1**/146 variables have p-value <5%,
  - Results: /user/abiolchi/Software/dev/compare_run2546_preStripIDbugfix/report/compare_run2546_StripIDbugfix.csv

Summary histograms

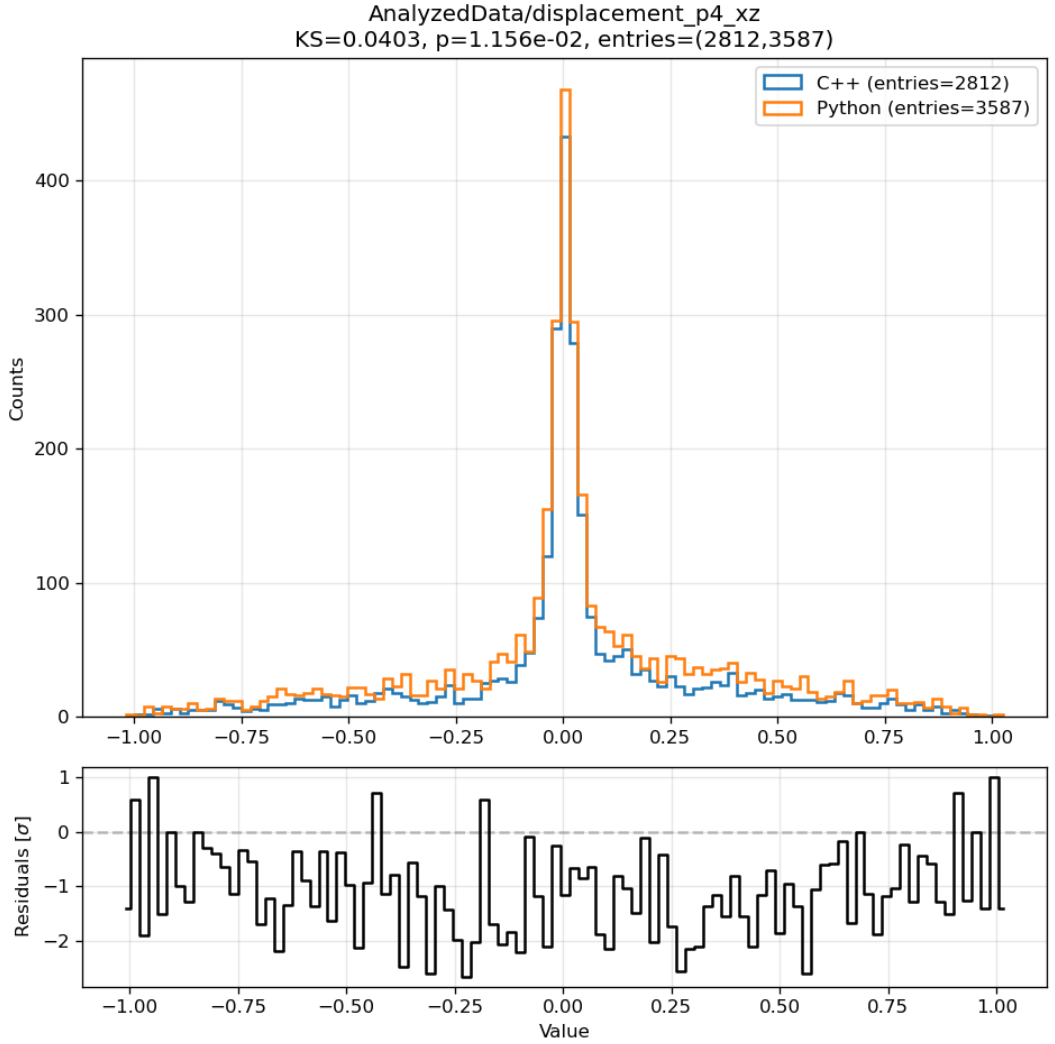
Sanity checks:
Variables that are not subjected
to any processing should be identical,
KS statistic= 0 and p-value= 1.





MAX: 0.05

Variable with significant differences:

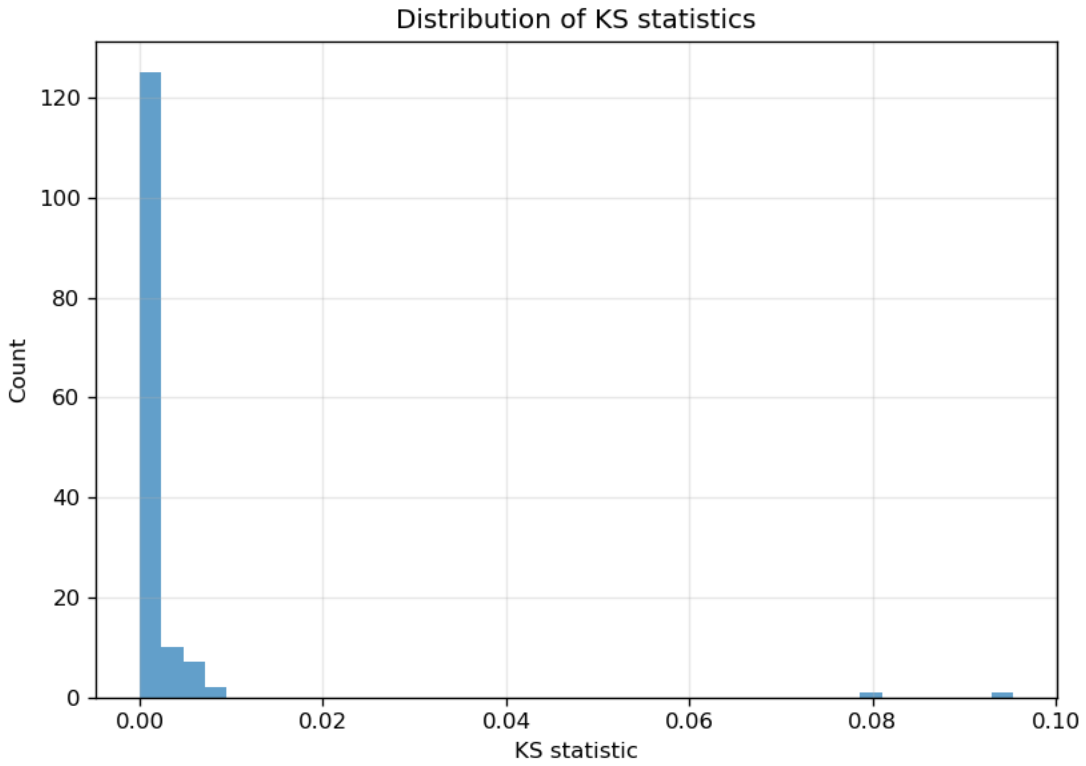


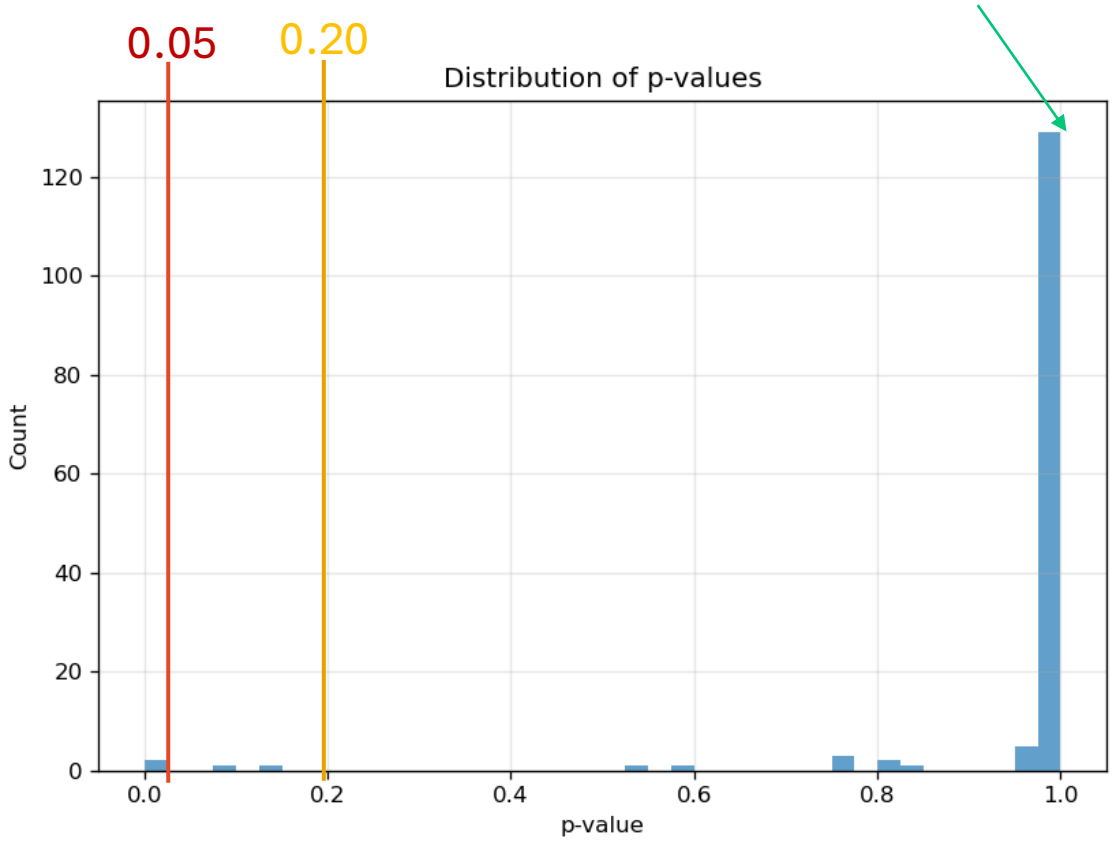
Validation

- Comparing output variables **Kolmogorov-Smirnov test**: unbinned test (larger distance between cumulative distributions).
- **Significance at 5%:** *meaning it is very unlikely that the two distributions comes from the same underlying distribution*.
- **Run NERO 2546:**
  - **1**/146 variables have p-value <5%,
  - Results: /user/abiolchi/Software/dev/compare_run2546_preStripIDbugfix/report/compare_run2546_StripIDbugfix.csv
- **Run NERO 50:**
  - **2**/146 variables have p-value <5%
  - Results: : /user/abiolchi/Software/dev/compare_cpp_py_outputs/compare_run50/report/compare_run50.csv

Summary histograms

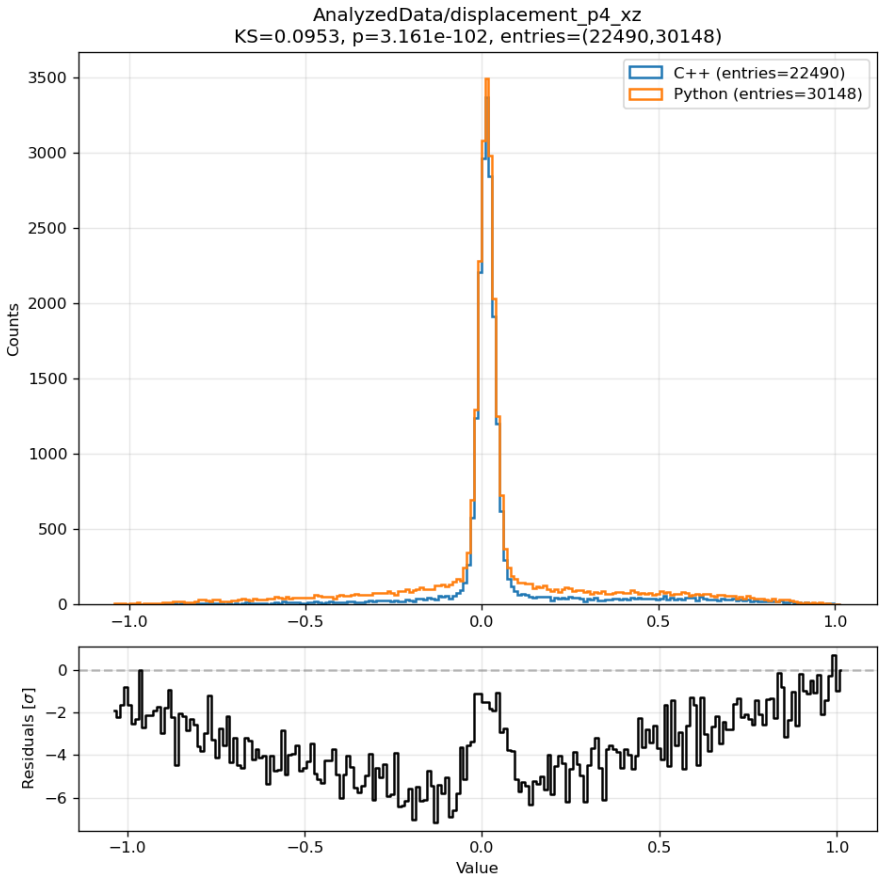
Sanity checks:
Variables that are not subjected
to any processing should be identical,
KS statistic= 0 and p-value= 1.

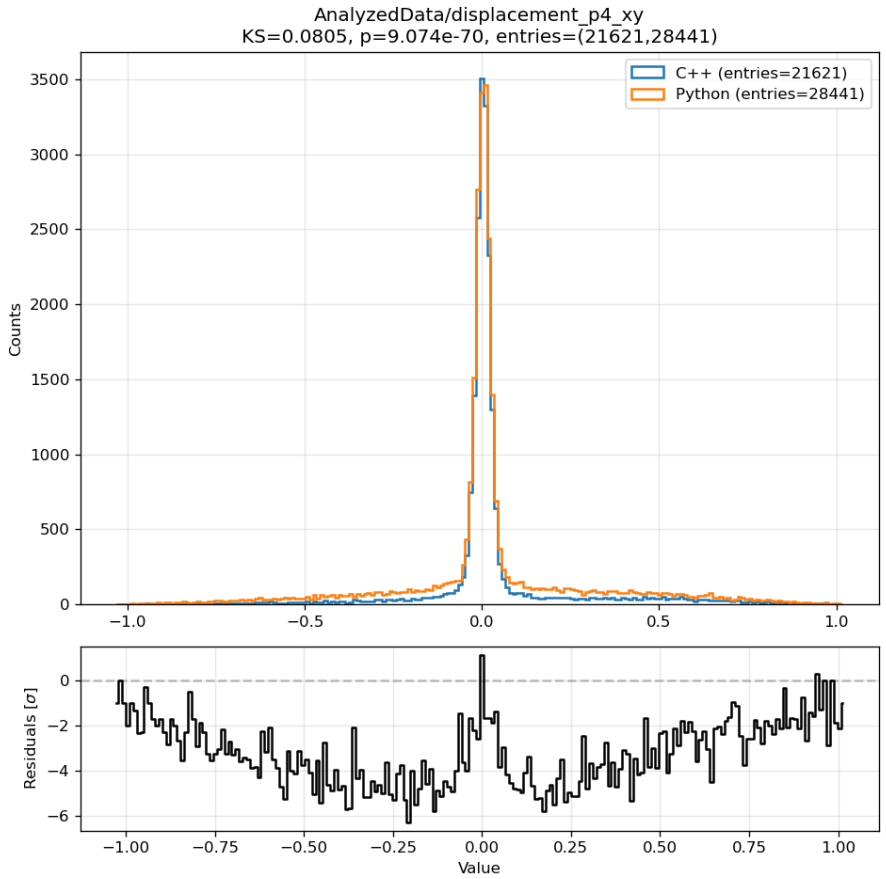




MAX: 0.10

Variable with significant differences:



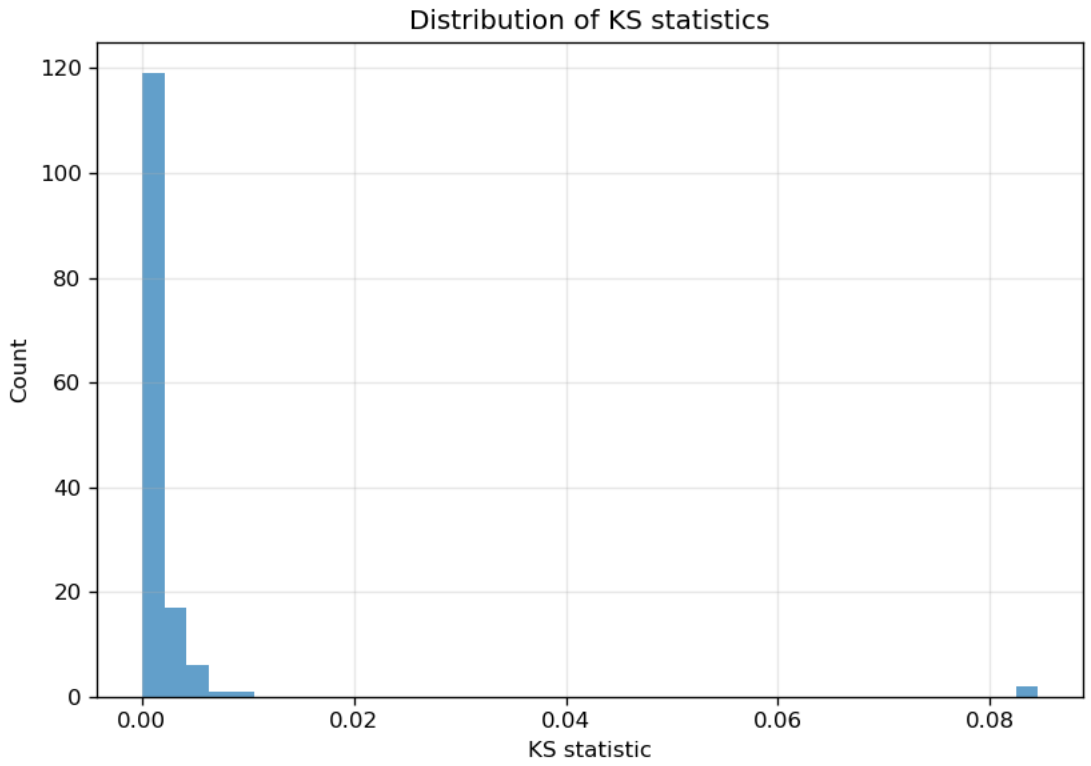


Validation

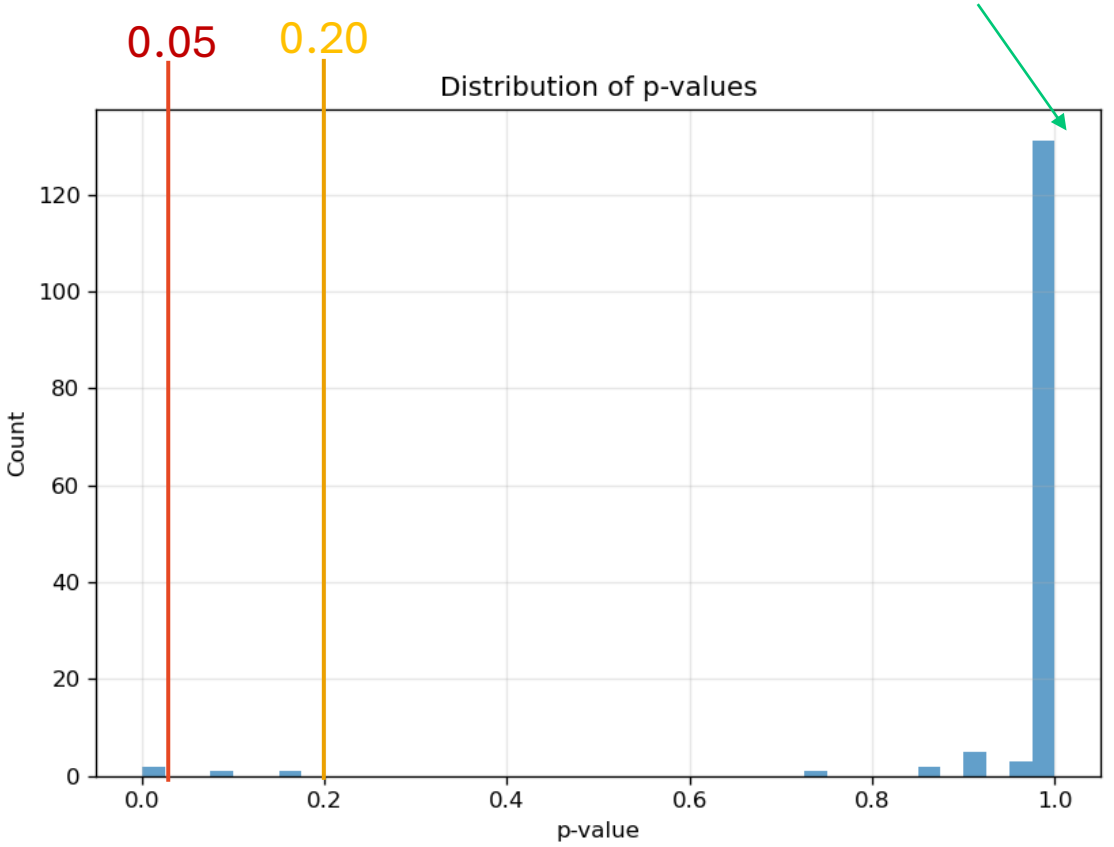
- Comparing output variables **Kolmogorov-Smirnov test**: unbinned test (larger distance between cumulative distributions).
- **Significance at 5%:** *meaning it is very unlikely that the two distributions comes from the same underlying distribution*.
- **Run NERO 2546:**
  - **1**/146 variables have p-value <5%,
  - Results: /user/abiolchi/Software/dev/compare_run2546_preStripIDbugfix/report/compare_run2546_StripIDbugfix.csv
- **Run NERO 50:**
  - **2**/146 variables have p-value <5%
  - Results: : /user/abiolchi/Software/dev/compare_cpp_py_outputs/compare_run50/report/compare_run50.csv
- **Run NERO 51:**
  - **2**/146 variables have p-value <5%
  - Results: : /user/abiolchi/Software/dev/compare_cpp_py_outputs/compare_run51/report/compare_run51.csv

Summary histograms

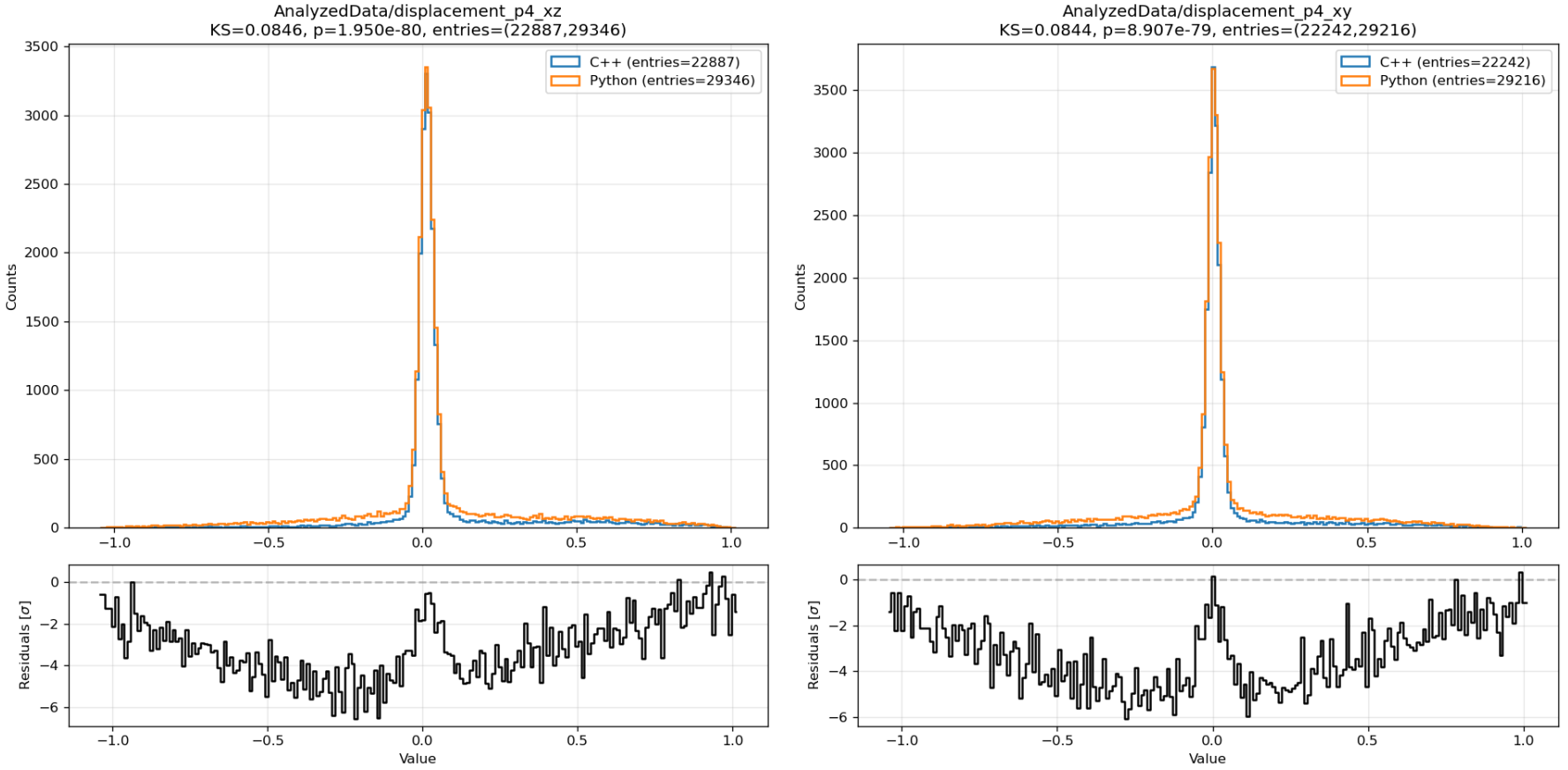
Sanity checks:
Variables that are not subjected
to any processing should be identical,
KS statistic= 0 and p-value= 1.



MAX: 0.07

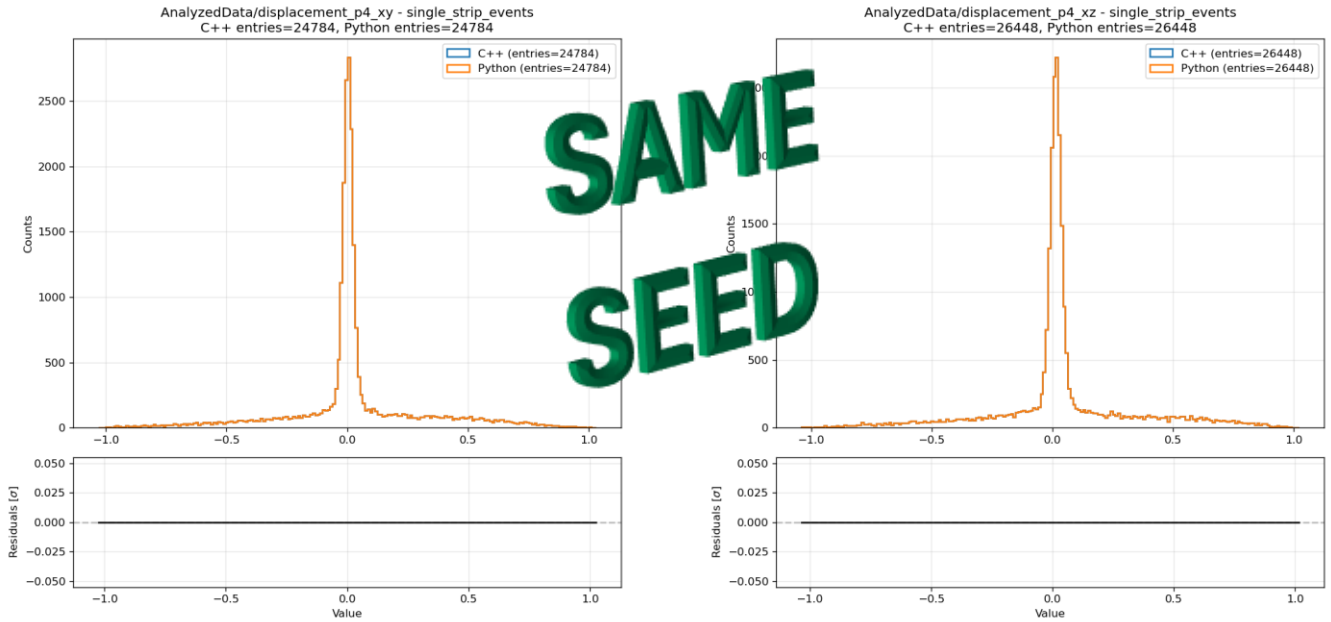


Variable with significant differences:

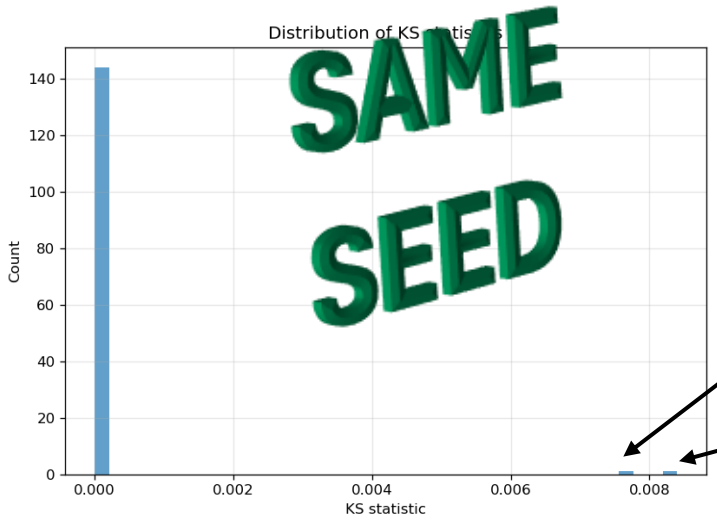


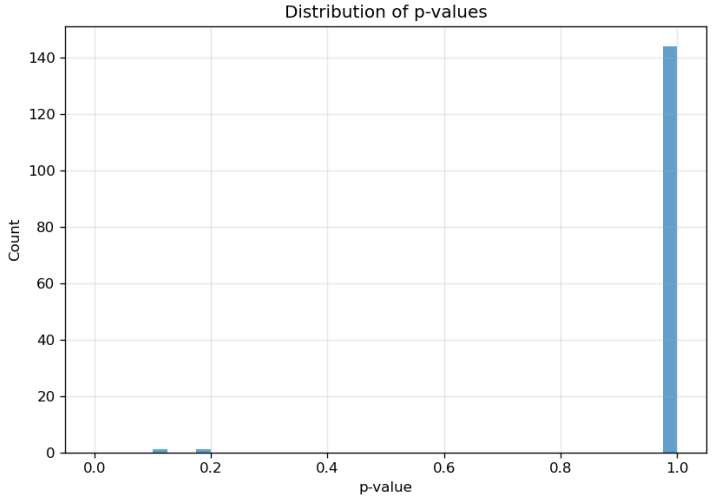
Displacement variables:

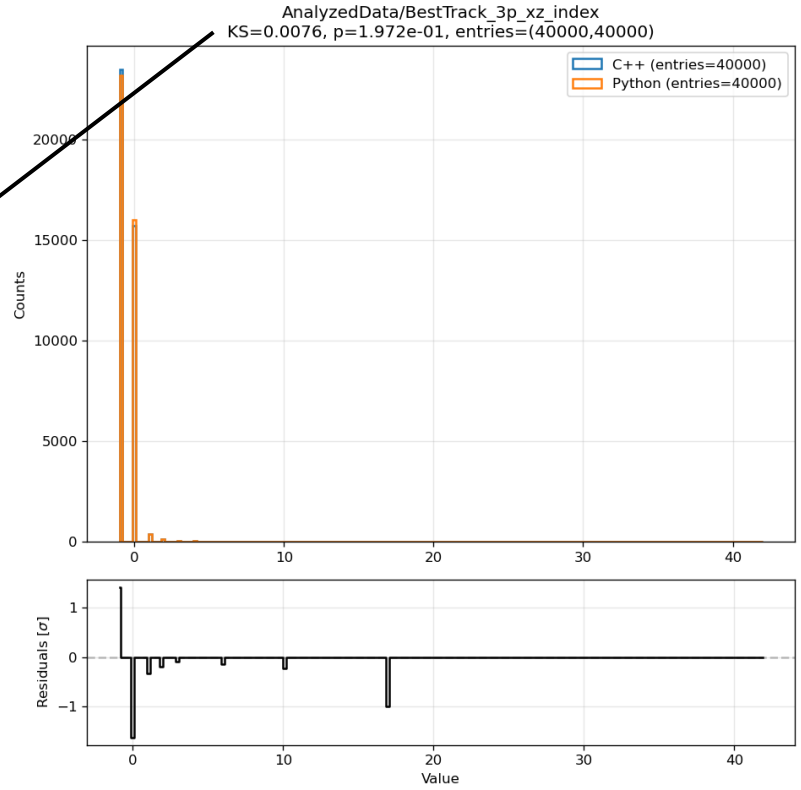
- Displacement viene definita come differenza tra cluster sul 4° piano e posizione attesa dal fit 3p, quindi è una variabile di “secondo livello” che amplifica piccole differenze upstream.
  - *Shared deterministic seed* C++/Python for single-strip smearing.
  - **Bug fix in Tracking.cc#L160:** there was a displacement_p4.clear() in the loop of 4p track reconstruction, that was resetting the number during the process, with the effect of less entries at the end.
    - *Only variable affected by this bug!*

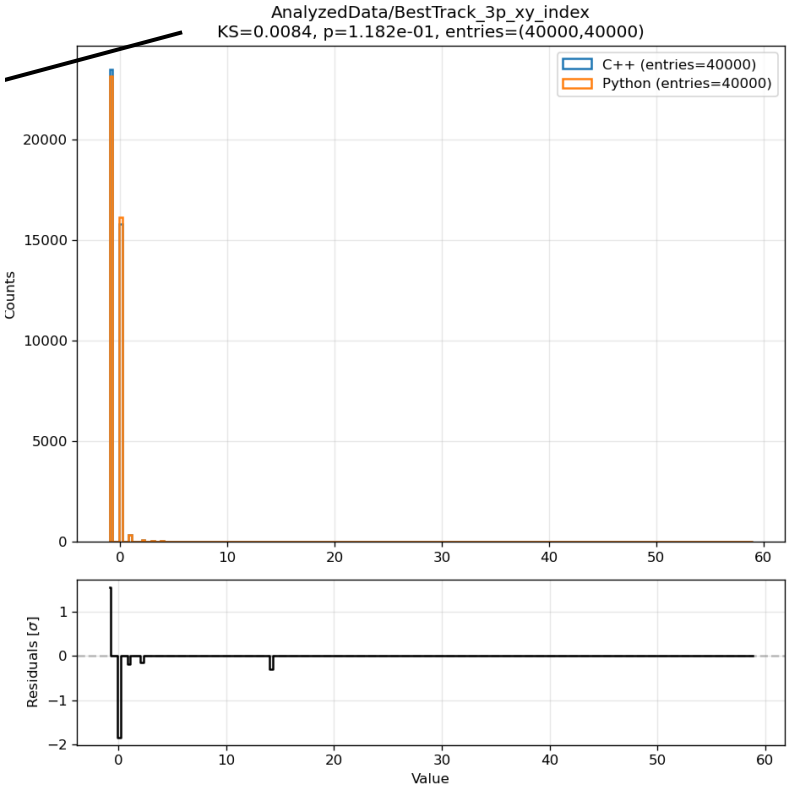


Comparison after the fix: 0 significant differences



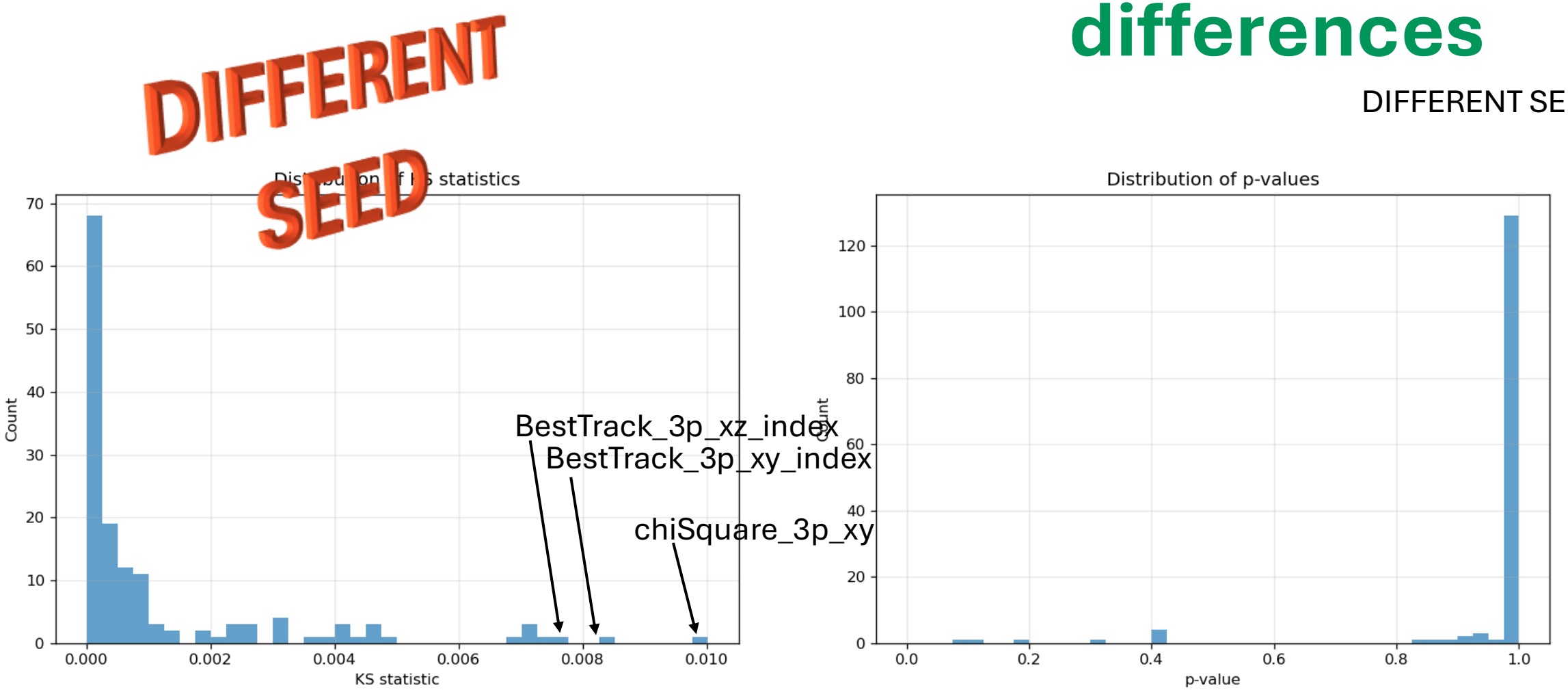




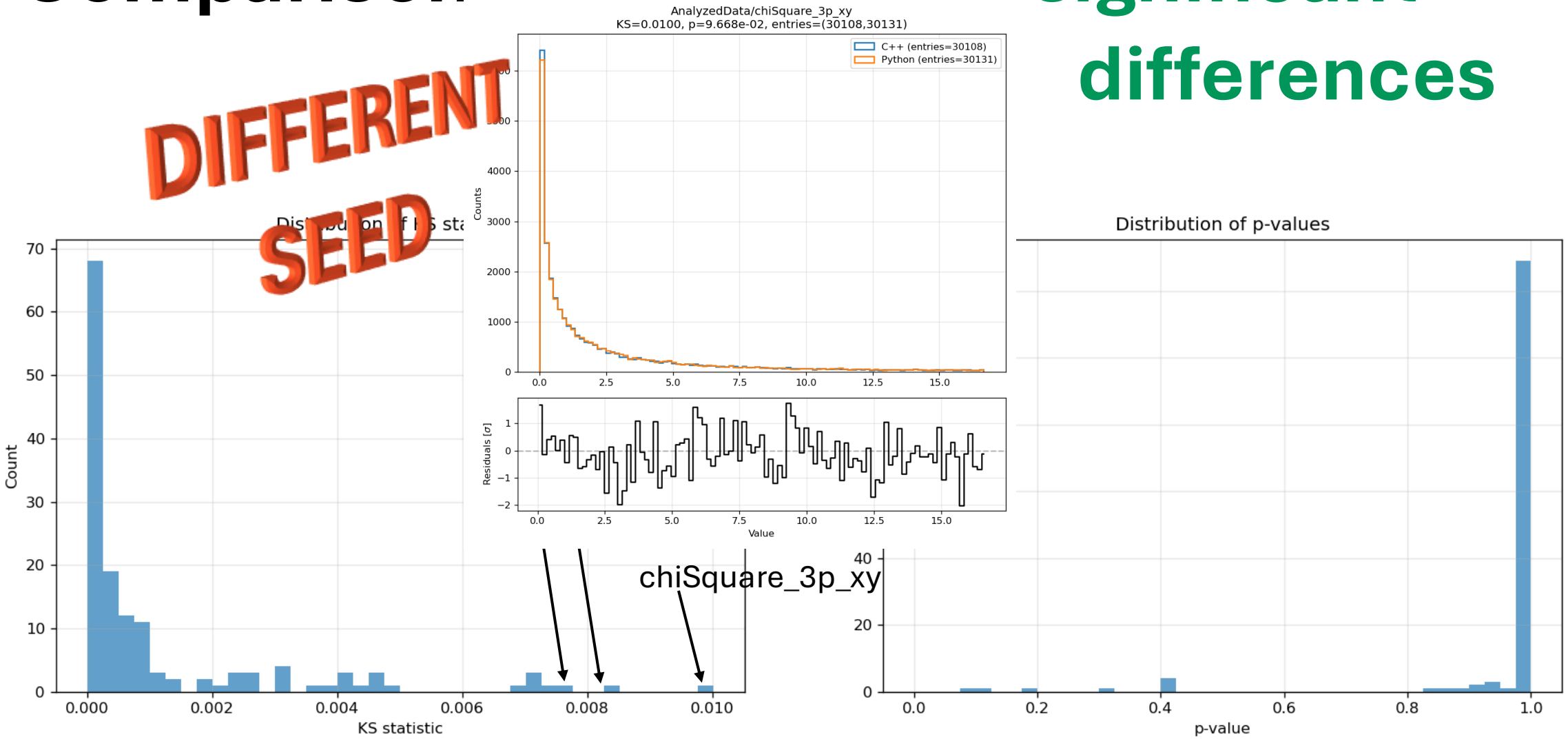


Comparison after the fix: 0 significant differences

DIFFERENT SEED



Comparison after the fix: 0 significant differences



Running time

c++

- **Reconstruction** core: **27.8058** s
  - ROOT event fill wall time: 3.12428 s
  - ROOT finalize wall time: 0.917799 s
- **Output production** wall time: **4.04208** s
- **Total wall time: 31.8479 s** (varies to 44 s)

Python (json + root)

- **Reconstruction** core wall time: **26.71** s
  - JSON output wall time: 5.42 s
  - ROOT fill wall time: 24.62 s
- **Output production** wall time: **30.04** s
- **Total wall time: 56.75 s**

Running time

c++

- **Reconstruction** core: **27.8058** s
  - ROOT event fill wall time: 3.12428 s
- **Output production** wall time: **4.04208** s
- **Total wall time: 31.8479 s** (varies to 44 s)

Python (directly to root in chunk of 500 evts)

- **Reconstruction** core wall time: **30.71** s
  - ROOT fill wall time: 19.27 s (with *uproot*)
- **Output production** wall time: **19.32** s
- **Total wall time: 50.02 s**

Bottleneck:
C++: vector<float> → TTree::Fill() = 3.12s
Python: list[object] → numpy.array() → TTree::Fill() (uproot) = 13-14s

Il delta (10s) è quasi tutto nella conversione Python→numpy*,
NON nella scrittura finale su disco.

Running time

c++

- **Reconstruction** core: **27.8058** s
  - ROOT event fill wall time: 3.12428 s
- **Output production** wall time: **4.04208** s
- **Total wall time: 31.8479 s** (varies to 44 s)

Python (only json)

- **Reconstruction** core wall time: 26.94 s
  - JSON output wall time: 5.30 s
- **Output production** wall time: **5.30** s
- **Total wall time: 32.24 s**

In python the most efficient output would be a json file (or pickle, usually faster but not human readable)

Conclusion and ongoing work

- **Version v0.3.0** of the reconstruction Software is now available on GitLab. I encourage using this as the previous version has C++ code with reported bugs.

Ongoing work:

- Started studying PCA on MURAVES output, some more work needed
  - *GOAL: Reduce high dimensionality preserving information, can this help in the selection?*

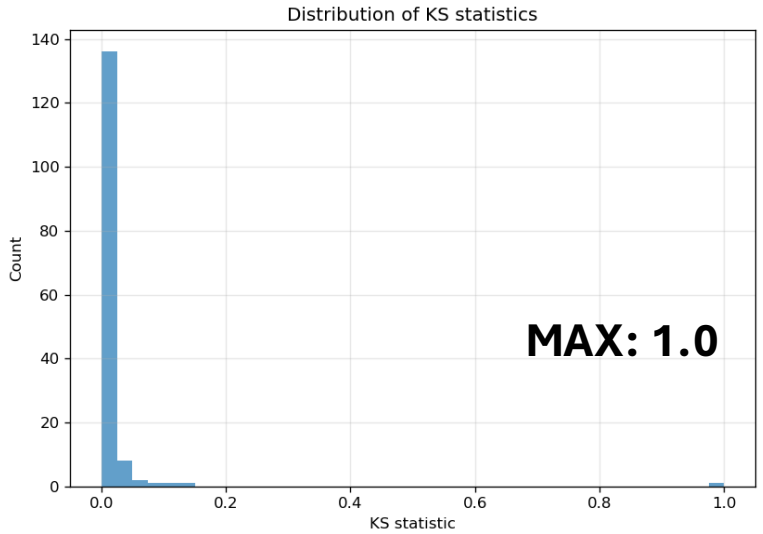
- **Export** hardcoded **clustering parameters** into a configuration file
- **Export** hardcoded **tracking parameters** into a configuration file
  - *GOAL: Facilitate*
    - *studies about different clustering thresholds.*
    - *Keeping trace of the changes with versioning*

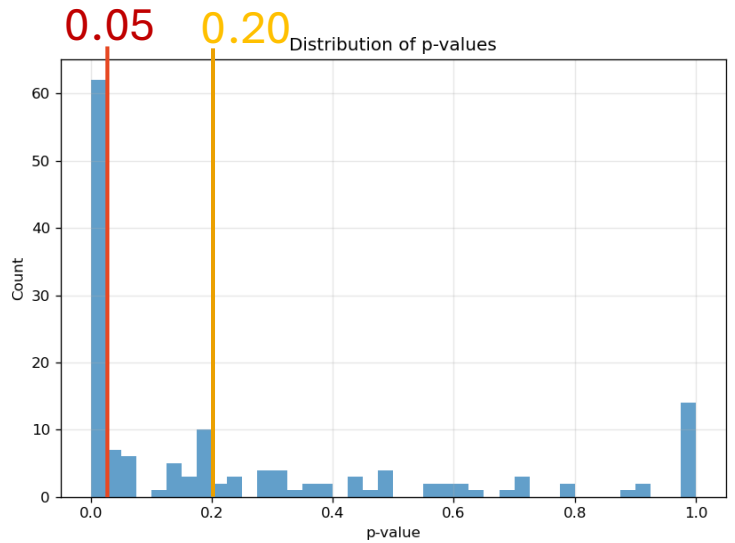
Backup: KS test on different runs

- Comparing output variables **Kolmogorov-Smirnov test**: unbinned test (larger distance between cumulative distributions).
- **Significance at 5%:** *meaning it is very unlikely that the two distributions comes from the same underlying distribution*.

Comparing two different runs c++ outputs of run 50 and 51:

- **69**/150 variables have p-value <5%





Displacement variables produced by same algorithm (cpp) on different runs NERO 50 and 51