Analysis Grand Challenge with ATLAS PHYSLITE data

Denys Klekots

(IRIS-HEP fellow)
denys.klekots@gmail.com

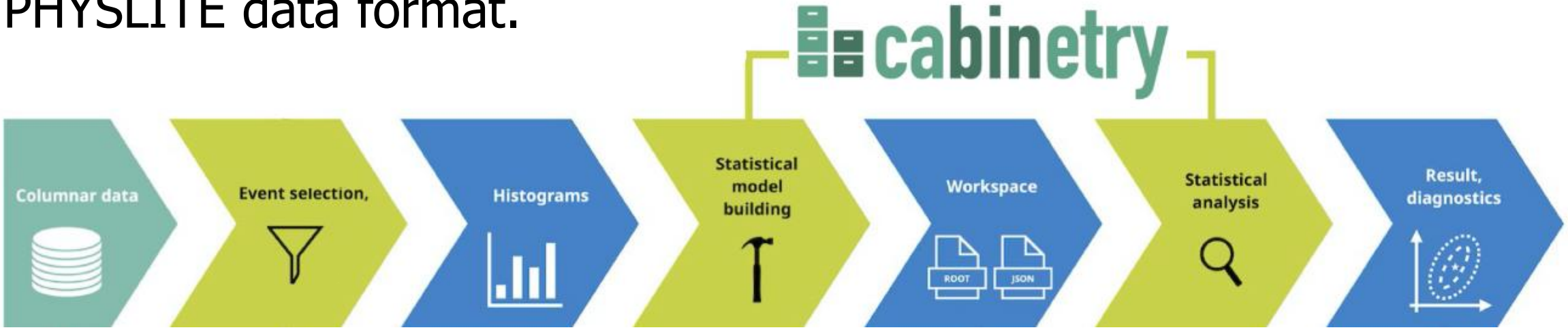
Supervisors:
Vangelis Kourlitis
Alexander Held
Matthew Feickert

Outline

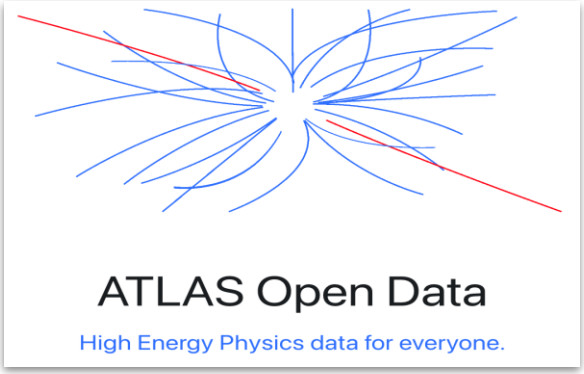
- Analysis Grand Challenge (AGC)
- Input data in the PHYSLITE format
- Data retrieval from the file
- Analysis of the $t\bar{t}$ pair production
- Parallel computing
- Histogram results
- Statistical inference
- Final histograms
- Project results and conclusions

Analysis Grand Challenge (AGC)

- Includes developing technologies envisioned for HL-LHC.
- Kind of benchmark to ensure that different pieces of software work fine together.
- Organized by **I**nstitute for **R**esearch and **I**nnovation in **S**oftware for **H**igh **E**nergy **P**hysics (**IRIS-HEP**)
- This particular project tests the I/O of a new version of the PHYSLITE data format.



Input data in the PHYSLITE format



- **ATLAS releases the 2015+2016 physics proton–proton collision data in PHYSLITE format for research**
  - Accompanied by "an appropriate set of simulated Monte Carlo samples"
  - Distributed by opendata.cern.ch, support material at opendata.atlas.cern

- The PHYSLITE format is optimized to decrease disk space and developed to meet the demands of HL-LHC
  - Unskimmed and monolithic
  - Contains already calibrated objects

Target size	MC	Data
PHYSLITE	12	10

kB per event

Data retrieval from the file

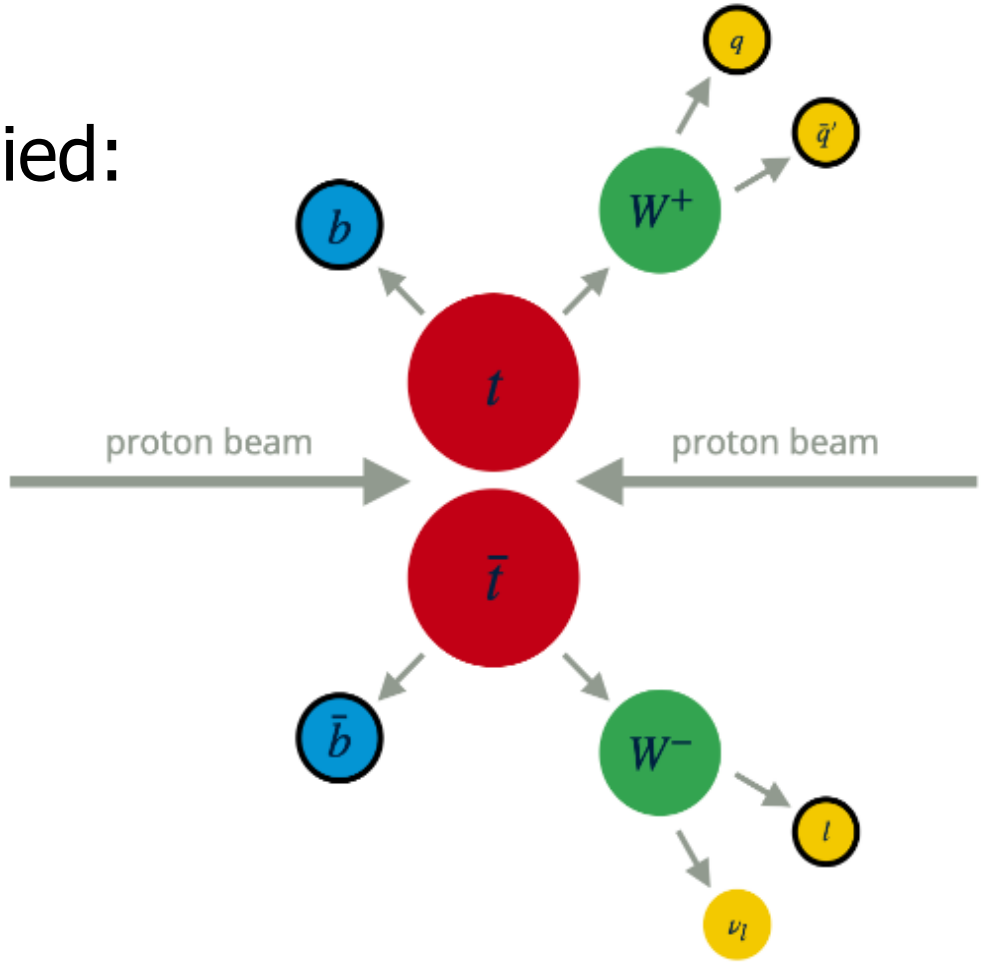
- COFFEA - Columnar Object Framework For Effective Analysis
  - A python package for scientific computations
  - Basic tools and wrappers for enabling nice functionality running columnar Collider HEP analysis.
- Contains the schema which allows reading PHYSLITE format file and turning it into a Python awkward array format.

```
events = NanoEventsFactory.from_root(
    {mc_file: "CollectionTree"},
    schemaclass=PHYSLITESchema,
    delayed=True,
    uproot_options=dict(filter_name=filter_name),
).events()
```

Event selection,

Event selections for the $t\bar{t}$ analysis

- The following event selections were applied:
  - Exactly one charged lepton
  - At least four quark jets
  - Leptons kinematic variables:
    - $p_t > 30\ GeV$
    - $\eta < 2{,}1$
  - Quark jets kinematic variables:
    - $p_t > 25\ GeV$
    - $\eta < 2{,}4$



Event selection,

Observable variables

m_{bjj} - trijet mass

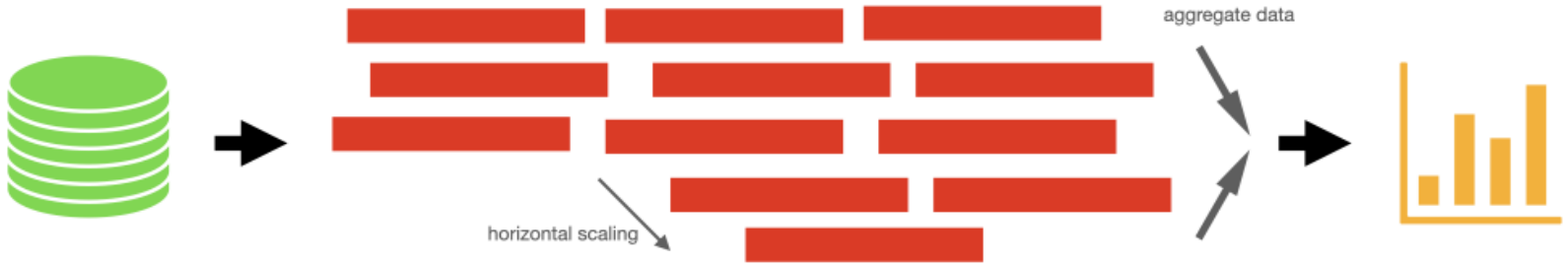
- Addition requirements: least 2 b-tagged jets in event
- Combination of jets in all possible groups of three
- Pick the group with the max b-tag variable
- Calculate the invariant mass of that group of three jets

H_t - scalar sum of jets p_t

- Addition requirements to have exactly one b-tagged jet in event
- Calculate the scalar sum of p_t for all jets in the event

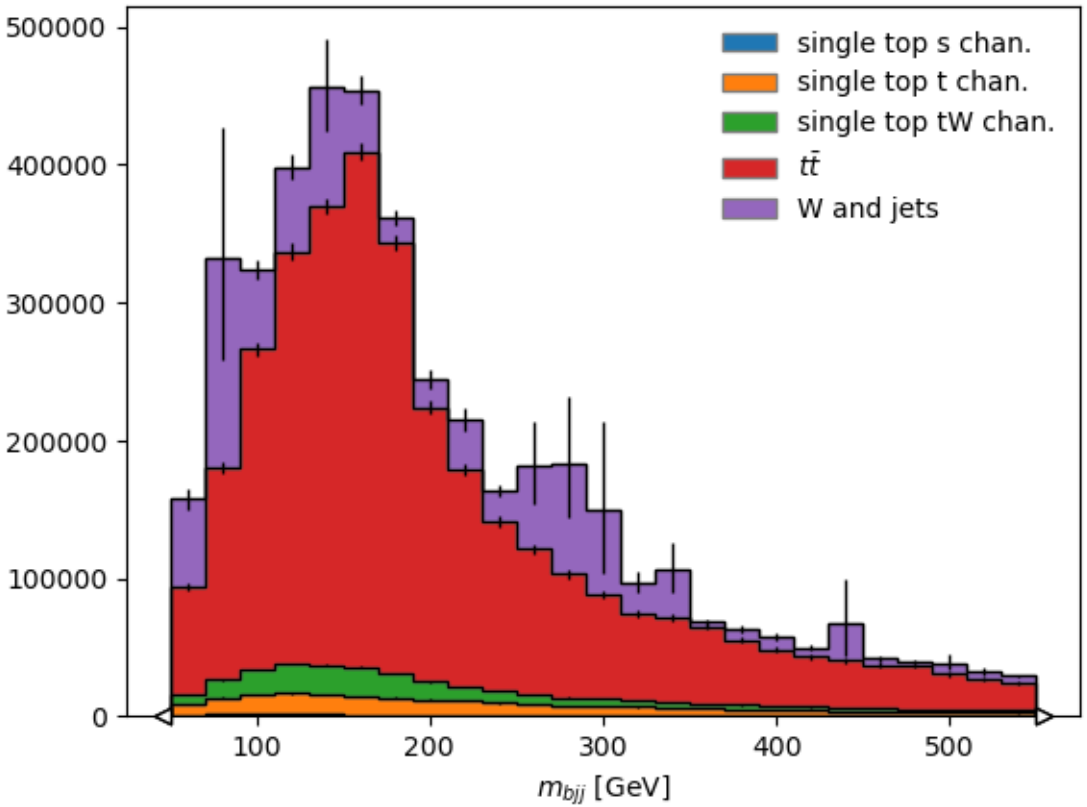
Parallel computing

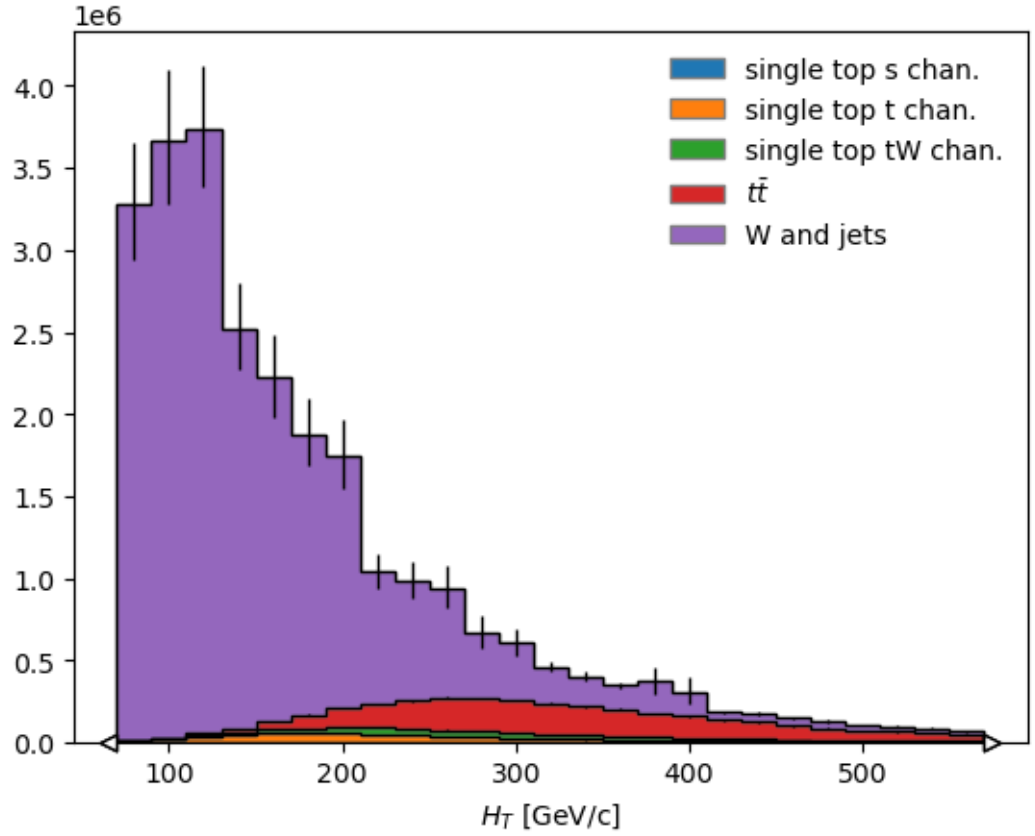
- Input data is split between many similar PHYSLITE files.
- Coffea allows the splitting of workflow into chunks and processing them in parallel.
- Results from each chunk are aggregated into a single histogram.
- The coffea-opendata.casa cluster was used for this project.



Histograms

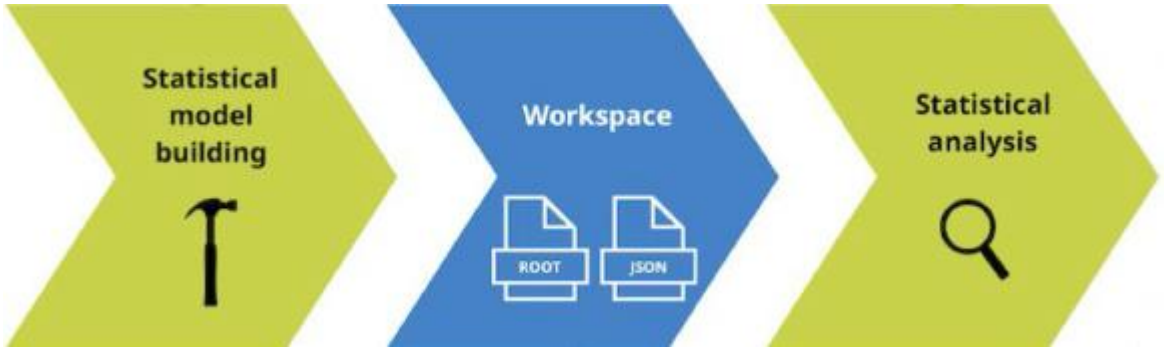
Histograms plotting





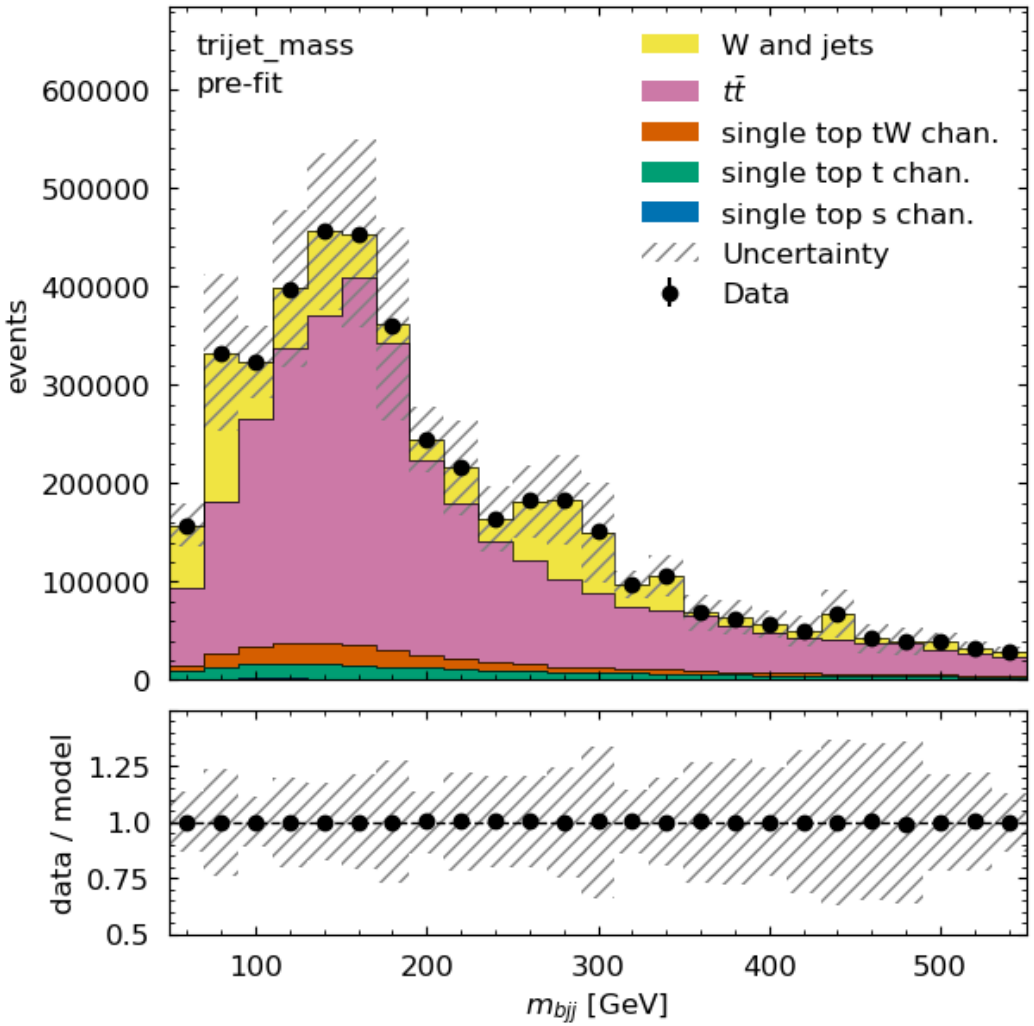
Statistical model

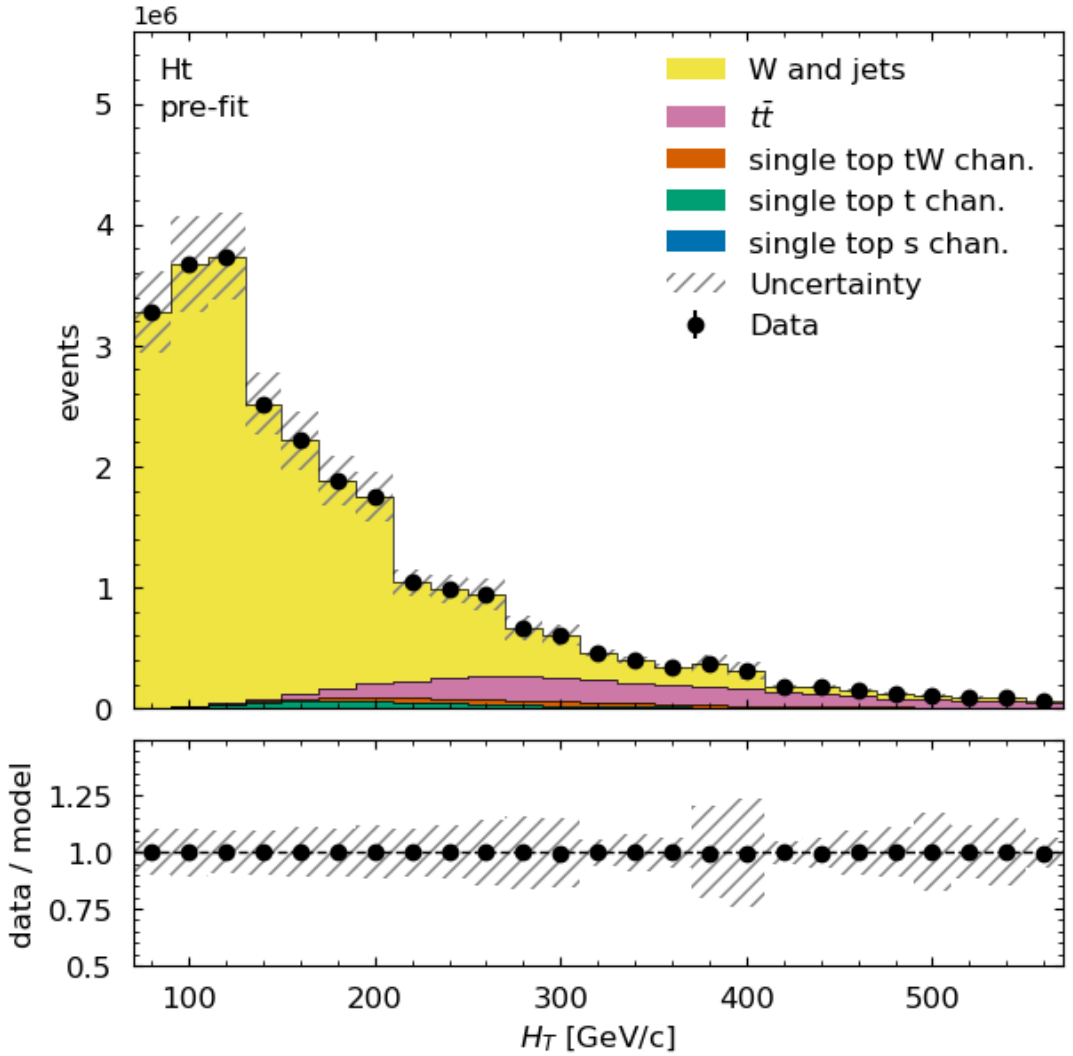
- Cabinetry is the Python package, making it easier for an analyzer to run their statistical inference pipeline.
- It takes the histograms and their variations for multiple observables as input.
- The statistical model is configured in the file.



Result, diagnostics

AGC final histograms





Denys Klekots denys.klekots@cern.ch

Project results and conclusions

- The version of AGC with ATLAS PHYSLITE data format was implemented.

- The code developed was well documented and can serve as an example of how one can use PHYSLITE in their analysis.

- The implemented code publicly available on the following repositories:
  - My repository https://github.com/Denys-Klekots/PHYSLITE_AGC_2024
  - IRIS-HEP fork https://github.com/iris-hep/agc-physlite

Thank you for your attention