The RDF $t\bar{t}$-analysis implementation

Analysis Grand Challenge

IRIS-HEP Fellow: Andrii Falko

Fellowship dates: June-Sep, 2023

Home university:

Taras Shevchenko National University of Kyiv

Mentors: Enrico Guiraud, Alexander Held

AGC introduction

Test workflows envisioned for the HL-LHC

- columnar data extraction from large datasets,
- processing of that data (event filtering, construction of observables, evaluation of systematic uncertainties) into histograms,
- statistical model construction and statistical inference,
- relevant visualizations for these steps

Specification of a physics analysis

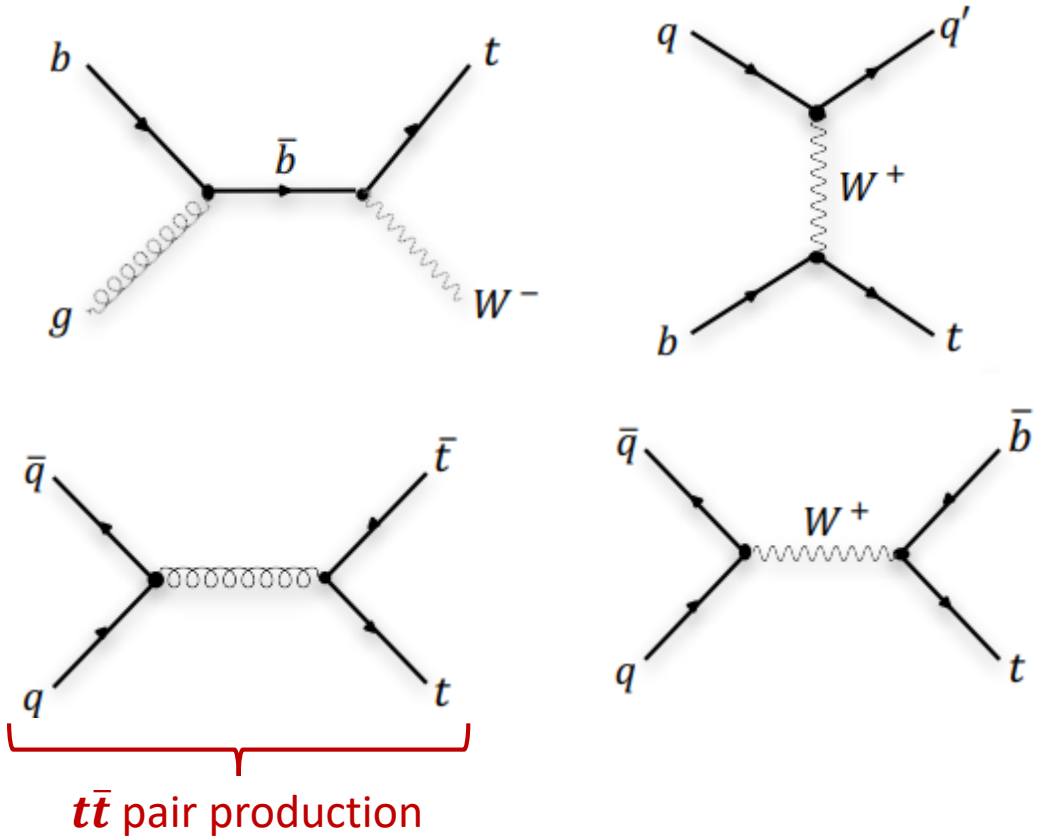
- $t\bar{t}$ cross-section measurement
- Top-quark mass reconstruction
- 2015 CMS Open Data
- Handling systematic variations

Reference implementations

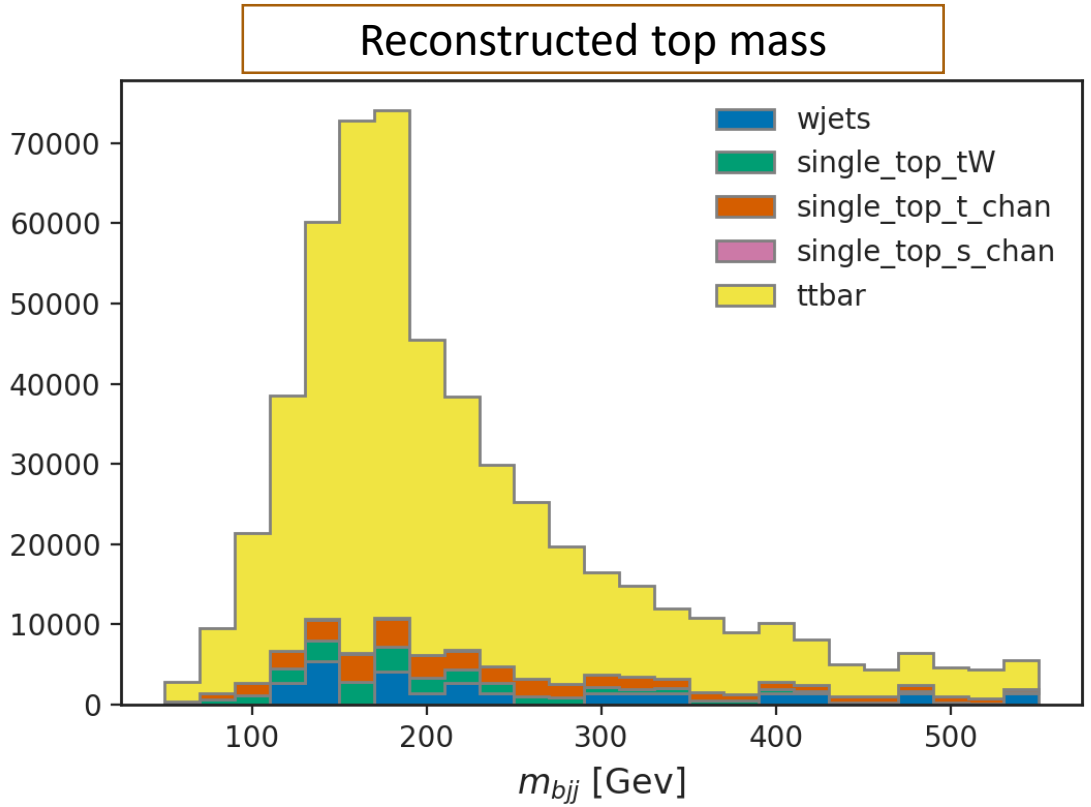
- Coffea
- **RDataFrame**
- Julia

$t\bar{t}$-analysis specification

1. Analysis task: $\boldsymbol{t\bar{t}}$ cross-section measurement



2. Input dataset: <u>2015 CMS Open Data</u>



3. The main stages of analysis
   1. Data extraction
   2. Events filtering
   3. Observables calculation

4. ~~Statistical inference~~ (is not a focus of this project)

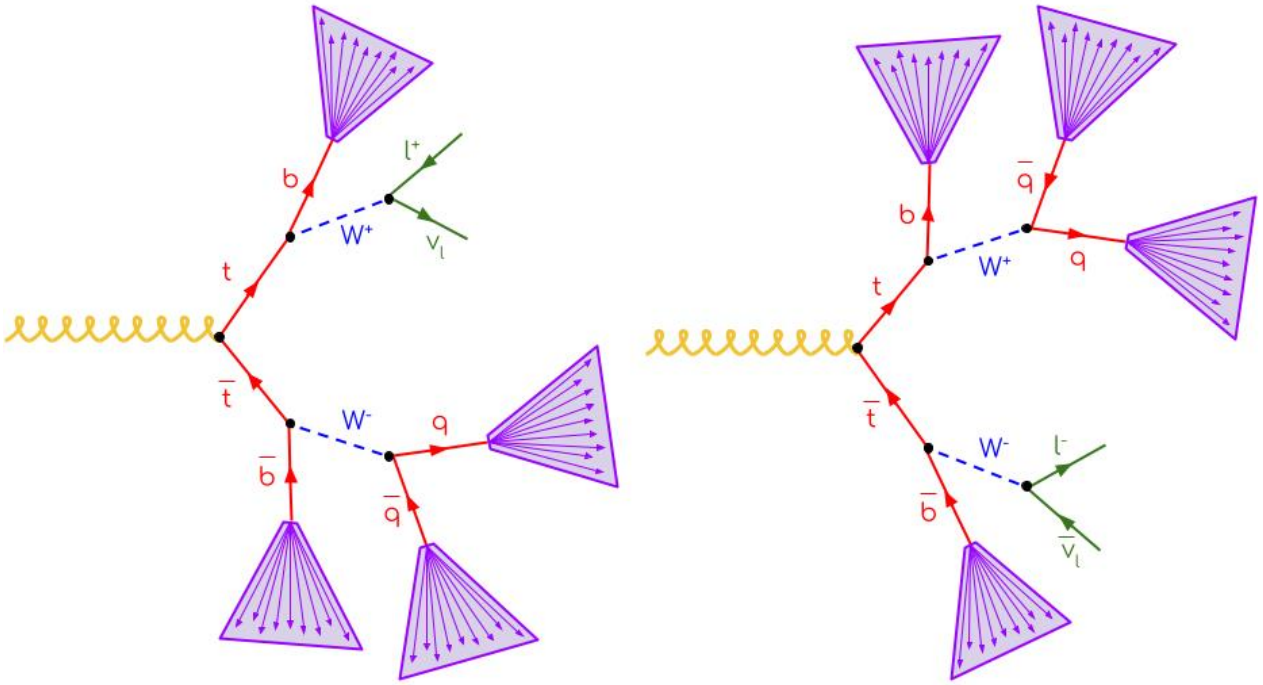
$t\bar{t}$-analysis specification

Selection of appropriate events:

- Single lepton
- At least 4 jets
- At least 2 b-tagged jets

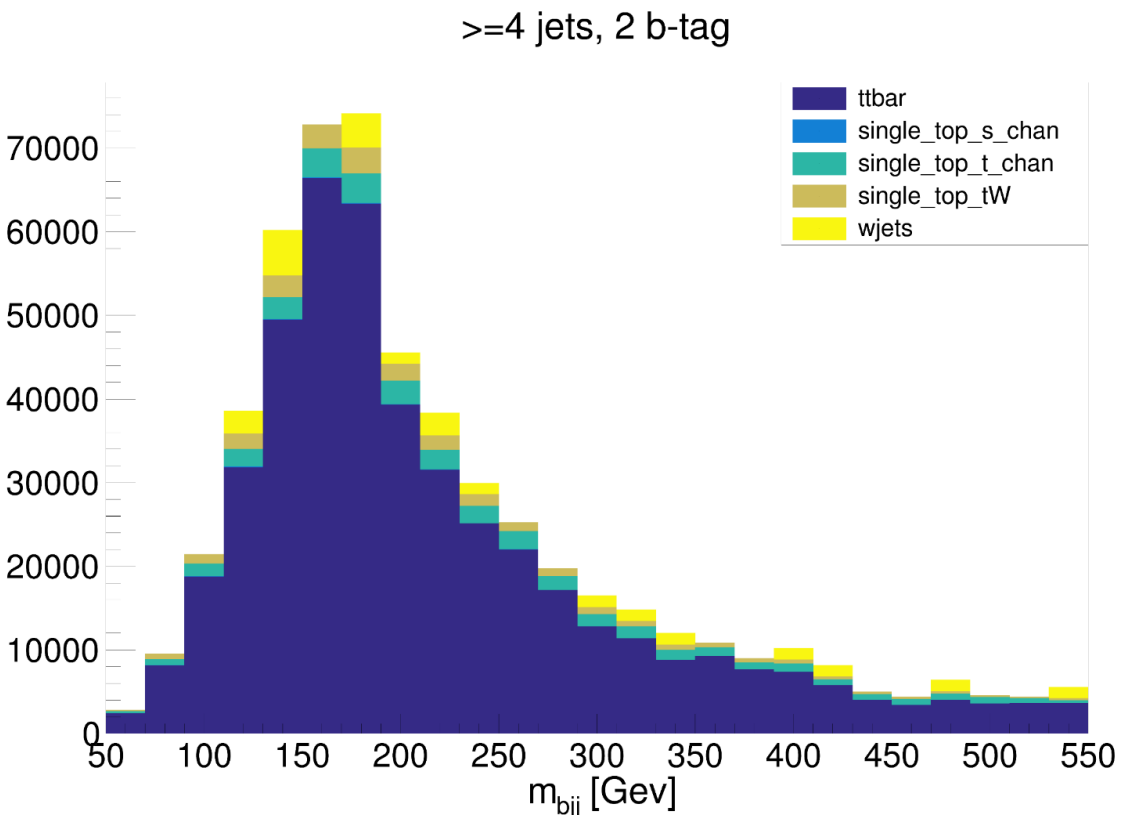
Top mass reconstruction

- Find combination of three jets which is the best candidate to be decayed from one top-quark
- Plotting mass of tri-jet



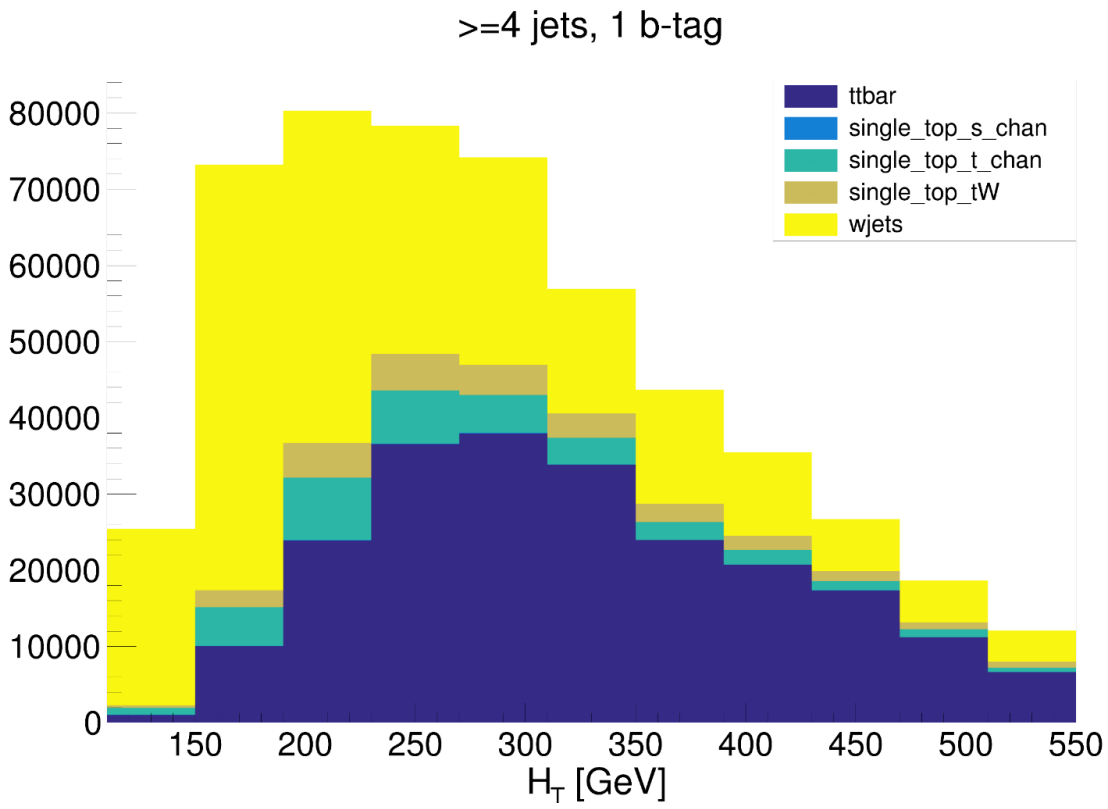
Observables to calculate

Signal region (*reconstructed **top mass***):



Top-quark mass peak plot

Control region (*sum of the p_T of all jets in each event*):



Scalar sum of transverse momenta plot

AGC versions

- The ways how to discriminate signal events from the background and how to calculate observes can be slightly different
- AGC has been evolving from version 0.1.0 to 2.0.0 (find more in AGC documentation)

V	Data schema	Selection cuts	Calculation observables
0	POET	Exactly 1 lepton with $p_T > 25\ GeV$; at least four jets with $p_T > 25\ GeV$; at least two jet with b-tag > 0.5	Find all tri-jet combinations per event At least 1 jet must be b-tagged Find tri-jet with a max combined p_T Calc combined mass of tri-jet system
1	NanoAOD		
2	NanoAOD	1 Lepton: $p_T > 30\ GeV$, $\lvert\eta\rvert$ <2.1, sip3d<4; Electrons: cutBased=4; Muons: tightId and pfRelIso04_all<0.15; Jets (>= 4): $p_T > 30\ GeV$, $\lvert\eta\rvert$ <2.1, isTightLeptonVeto, b-tag > 0.5	Machine Learning Component is used as an alternative way to find the decay product of top quark by assigning of each jet to its parent parton.

RDataFrame implementation status

- Versions 0 (0.1.0 and 0.2.0) were implemented during my last IRIS-HEP project
- Versions 1 and 2 are going to be implemented during this project:
- Already switched to NanoAOD data schema
- Now comparing produced histograms with those obtained by Alex's coffea implementation. Looking for the origin of small discrepancies (< 0,1 %))
- Need to move implementation to AGC v2 cuts
- Add code to calculate ML features
- Add code to do ML inference
- Everything needs to pass validation, e. i. produced histograms should be the same as in the coffea version

THANK YOU FOR YOUR ATTENTION!