Quark gluon tag in HL-LHC. Weights and Cuts

Jeremy Couthures, Sabine Elles, Jessica Leveque, Haoran Zhao and
<u>Florencia Castillo</u>

Jet tagging and scale factor meeting

DATE

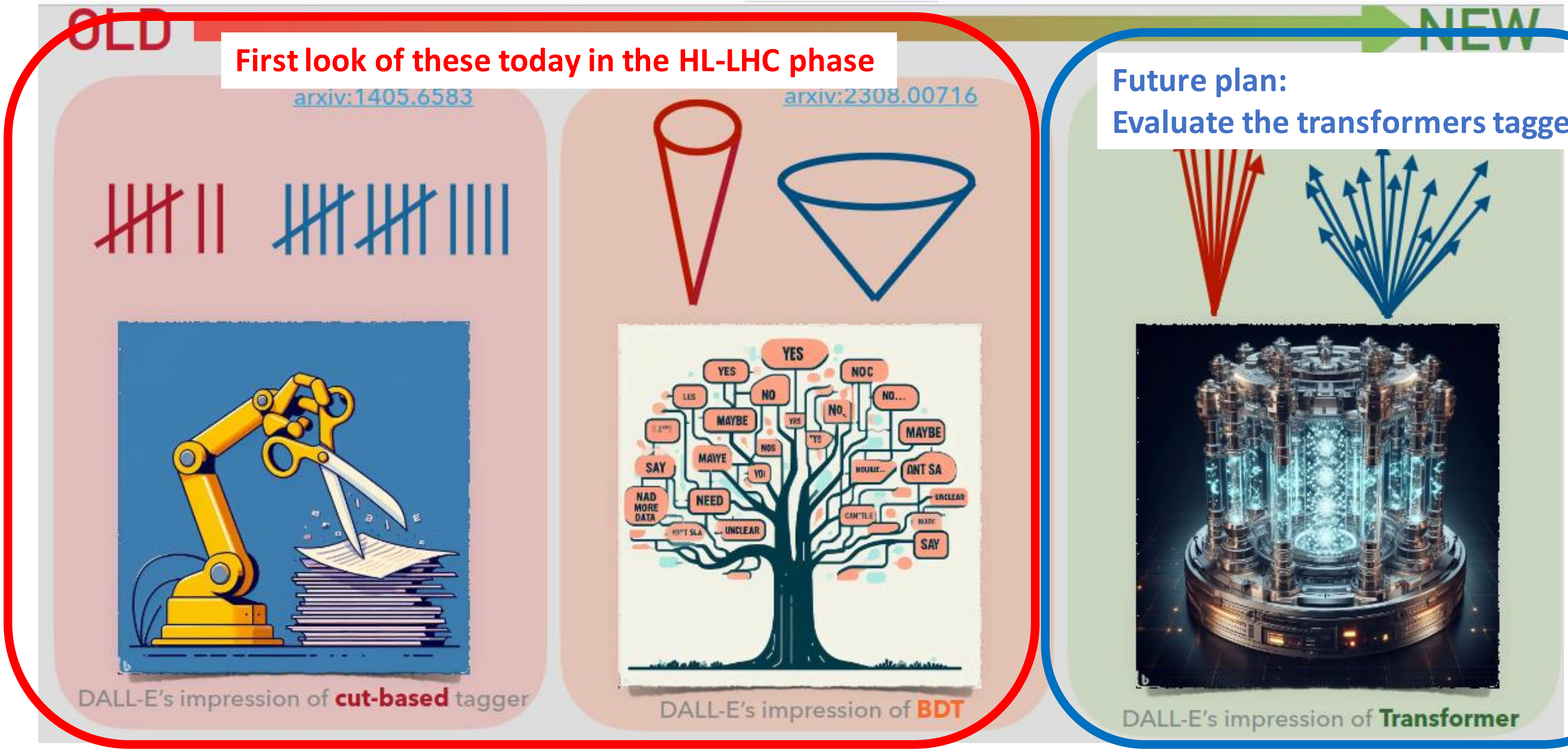
cnrs

LAPP
Laboratoire d'Annecy-le-Vieux
de Physique des Particules

anr
agence nationale
de la recherche

ATRAPP

Throwback



First look of these today in the HL-LHC phase

Future plan:
Evaluate the transformers tagger

DALL-E's impression of **cut-based** tagger

DALL-E's impression of **BDT**

DALL-E's impression of **Transformer**

Action points of last time: Slides

We are checking the performance of the BDT tagger in the HL-LHC

- Action points of my previous presentation:
  - Check flat distribution for quarks and gluons separately
  - Lack of statistics
  - No tracks in the forward region
  - Add more performance plots

- Today:
  - Quark and gluon distributions shown
  - Addressing the lack of stats with more samples a new weights
  - Adding forward region
  - Adding gluon eff and gluon rej for a given WP (fixed quark eff)

Samples used for the following studies

New samples added!

- mc21_14TeV.600026.PhH7EG_NNPDF3_AZNLO_VBFH125_ZZ4nu_MET75.recon.AOD.e8481_s4038_r14365
- mc21_14TeV.601229.PhPy8EG_A14_ttbar_hdamp258p75_SingleLep.recon.AOD.e8481_s4038_r14365
- mc21_14TeV.601230.PhPy8EG_A14_ttbar_hdamp258p75_dil.recon.AOD.e8481_s4038_r14365
- mc21_14TeV.601237.PhPy8EG_A14_ttbar_hdamp258p75_allhad.recon.AOD.e8481_s4038_r14365
- mc21_14TeV.801165.Py8EG_A14NNPDF23LO_jj_JZ0.recon.AOD.e8481_s4038_r14365
- mc21_14TeV.801166.Py8EG_A14NNPDF23LO_jj_JZ1.recon.AOD.e8481_s4038_r14365
- mc21_14TeV.801167.Py8EG_A14NNPDF23LO_jj_JZ2.recon.AOD.e8481_s4038_r14365
- mc21_14TeV.801168.Py8EG_A14NNPDF23LO_jj_JZ3.recon.AOD.e8481_s4038_r14365
- mc21_14TeV.801169.Py8EG_A14NNPDF23LO_jj_JZ4.recon.AOD.e8481_s4038_r14365
- mc21_14TeV.801170.Py8EG_A14NNPDF23LO_jj_JZ5.recon.AOD.e8481_s4038_r14365

Hight stast samples still pending:
https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/JetEtmissHLLHC#Low_statistics_first_pass_latest

BDT performance of HL-LHC. Same cuts used for Run 2

https://cds.cern.ch/record/2802919/files/ATL-COM-PHYS-2022-134.pdf

https://arxiv.org/pdf/2308.00716.pdf

- **Input variables**: Ntracks, Track C1 (energy correlation), Track With and flat pT

$$n_{\mathrm{track}} = \sum_{\mathrm{trk}\in\mathrm{jet}} \qquad C_1^{\beta=0.2} = \frac{\sum_{i,j\in\mathrm{jet}}^{i\neq j} p_{\mathrm{T},i}\, p_{\mathrm{T},j} \left(\Delta R_{i,j}\right)^{\beta=0.2}}{\left(\sum_{\mathrm{trk}\in\mathrm{jet}} p_{\mathrm{T}}^{\mathrm{track}}\right)^2} \qquad w^{\mathrm{track}} = \frac{\sum_{\mathrm{trk}\in\mathrm{jet}} p_{\mathrm{T}}^{\mathrm{track}} \Delta R_{\mathrm{trk,jet}}}{\sum_{\mathrm{trk}\in\mathrm{jet}} p_{\mathrm{T}}^{\mathrm{track}}}$$

- **Jets**: Calibrated EM topo jets
- **Cuts**: pT > 500 GeV and abs(eta)<2.1, Number of jets > 1 and event_weight <100
- **Weights** (Maybe weights not need it for HL-LHC samples yet):

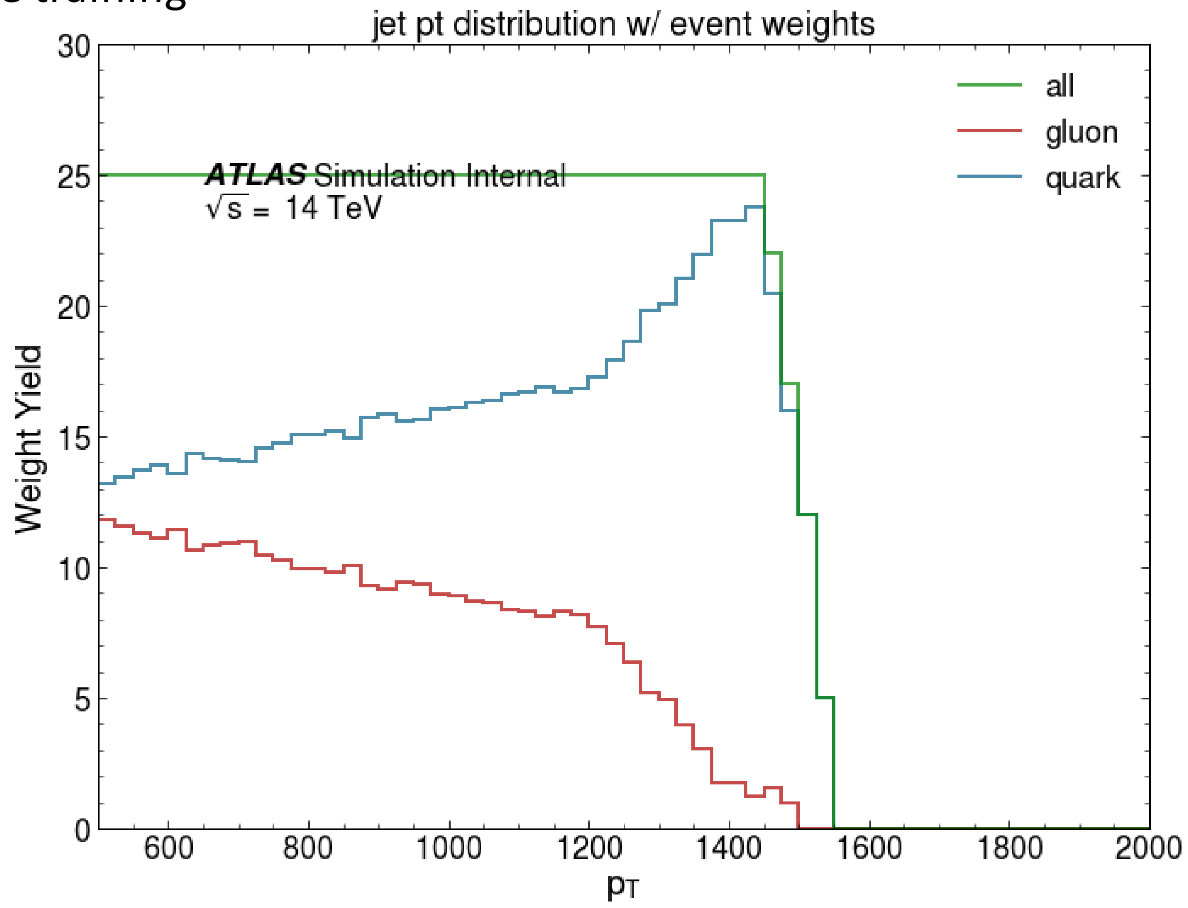
  Event_weight = xsec * mconly_weight / sumweight (Luminosity?)

  sumweight = Sum of mconly_weight

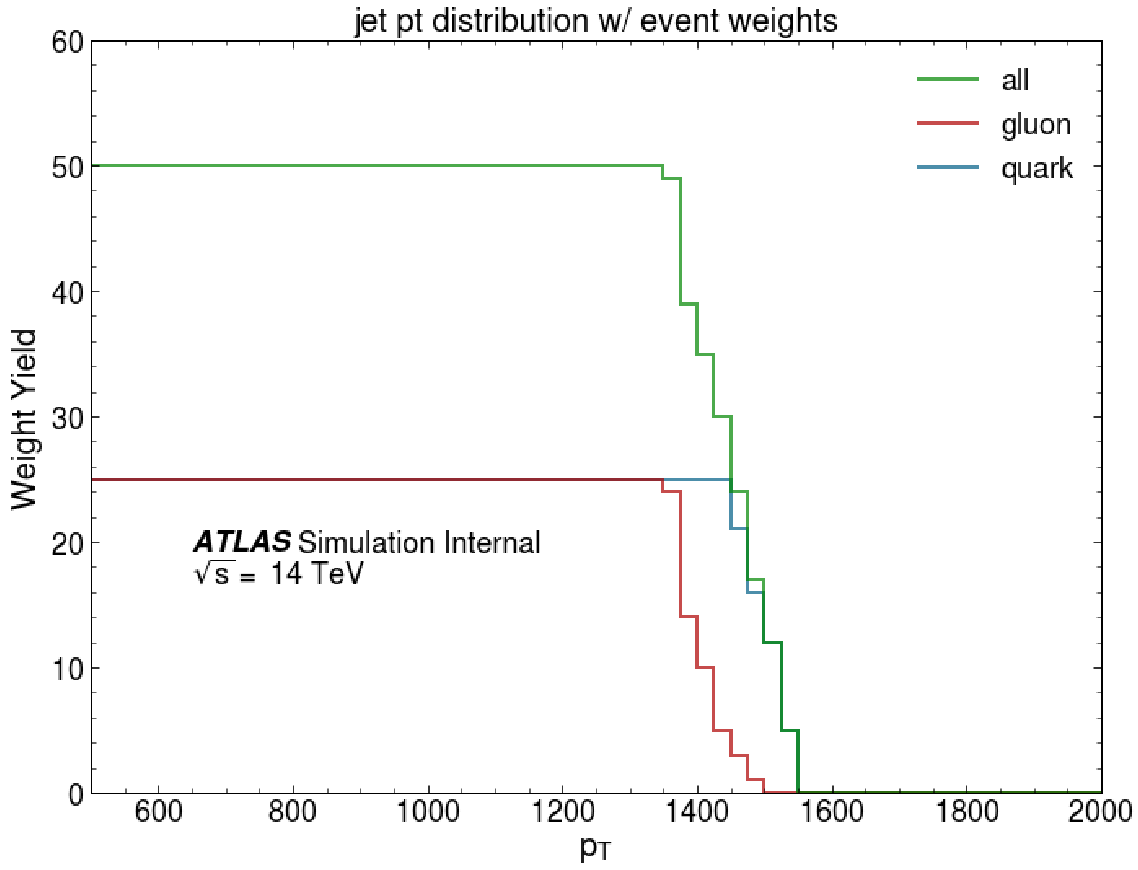
$$w_i = w_{\mathrm{exp},i}\, w_{\mathrm{MC},i} \frac{\mathcal{L}\,\sigma X}{\sum_i w_{\mathrm{MC},i}} = w_{\mathrm{exp},i}\, w_{\mathrm{MC},i} \frac{\mathcal{L}\left(k\,\sigma_{\mathrm{MC}}\,\epsilon_{\mathrm{filter}}\right) X}{\sum_i w_{\mathrm{MC},i}},$$

Flat distribution

Flat distribution as it is implemented in Run 2 and the one I used for my previous presentation. These distribution add an extra dependency in the pT adding the topology of the sample used to the training

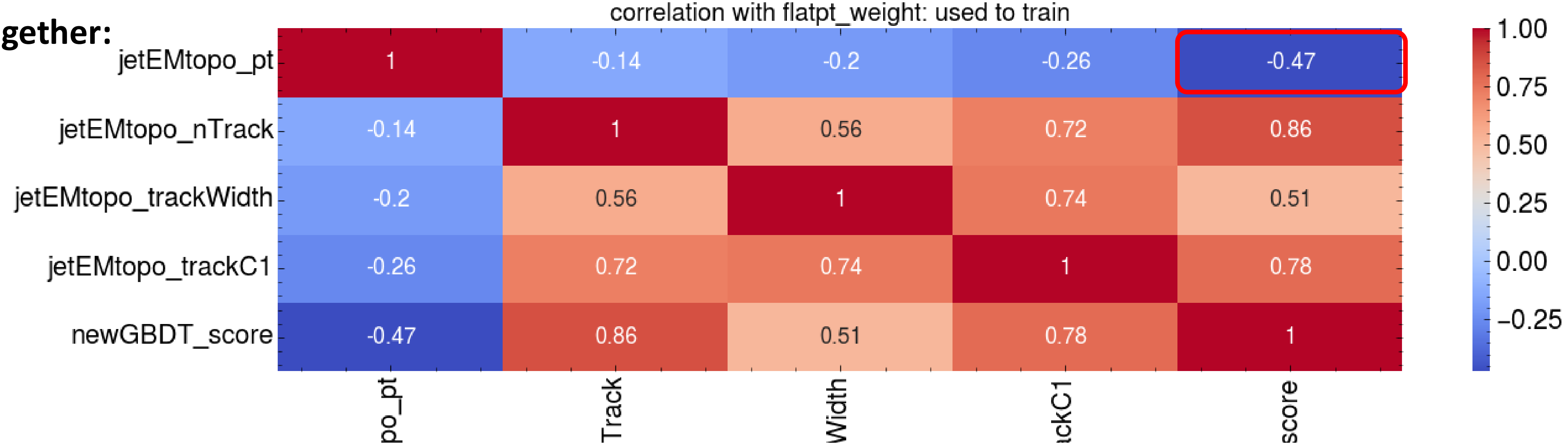


Flat distribution separating the sample by quark and gluon. The topology of the sample now is erased.



Correlation using different flat weights

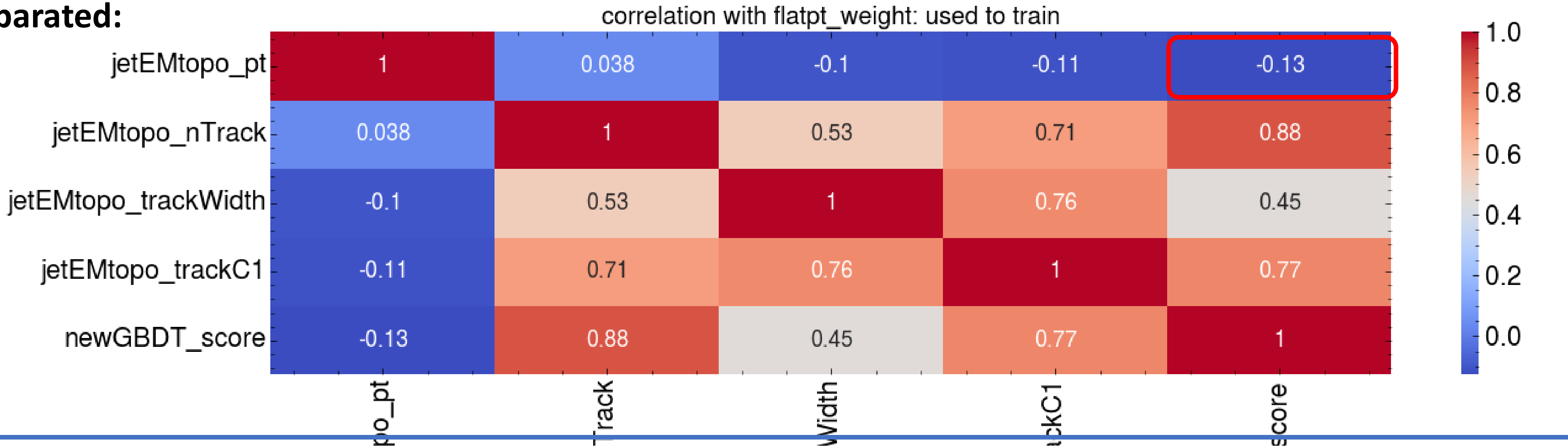
Together:



correlation with flatpt_weight: used to train

	po_pt	Track	Width	ckC1	score
jetEMtopo_pt	1	-0.14	-0.2	-0.26	-0.47
jetEMtopo_nTrack	-0.14	1	0.56	0.72	0.86
jetEMtopo_trackWidth	-0.2	0.56	1	0.74	0.51
jetEMtopo_trackC1	-0.26	0.72	0.74	1	0.78
newGBDT_score	-0.47	0.86	0.51	0.78	1

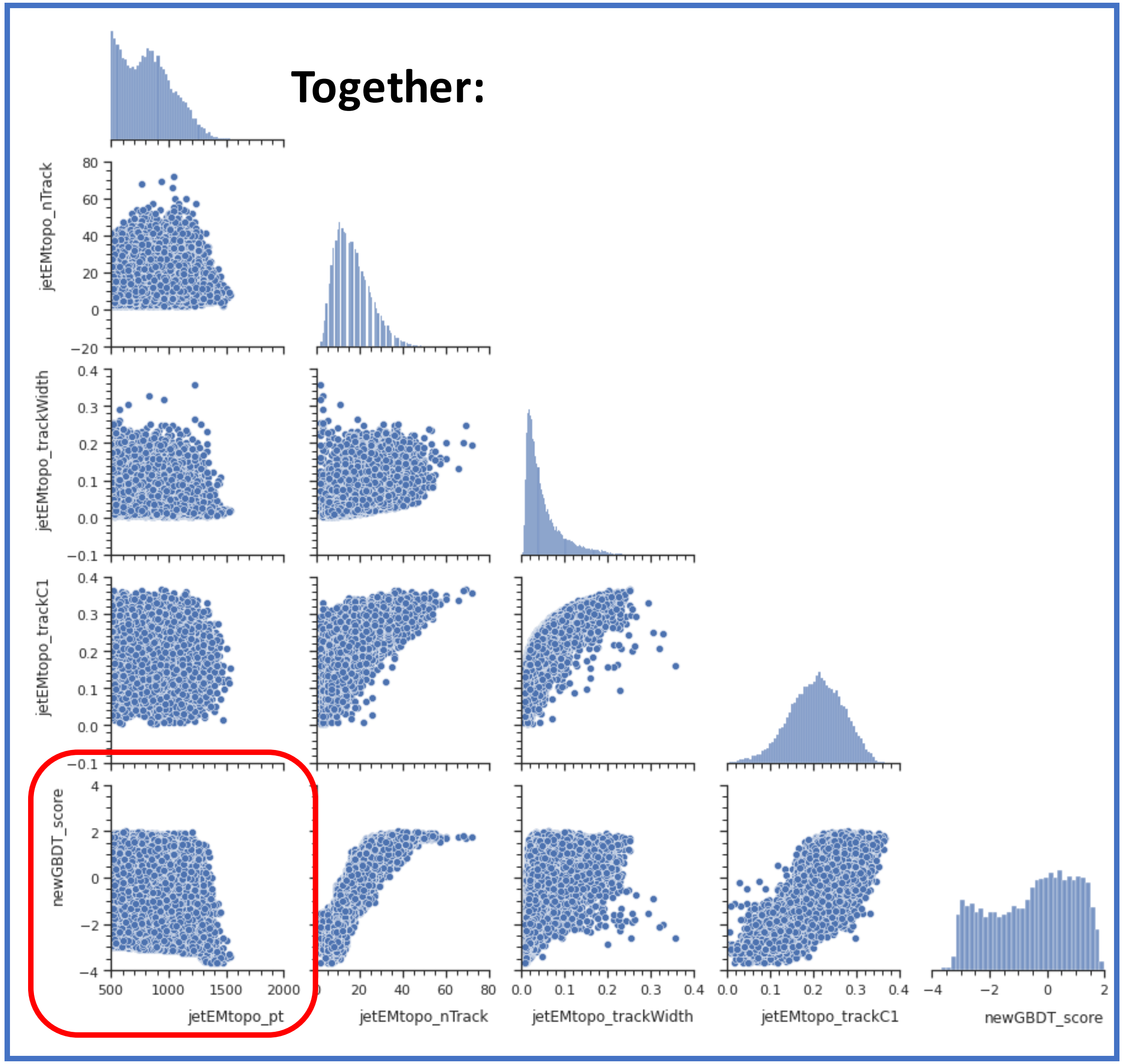
Separated:



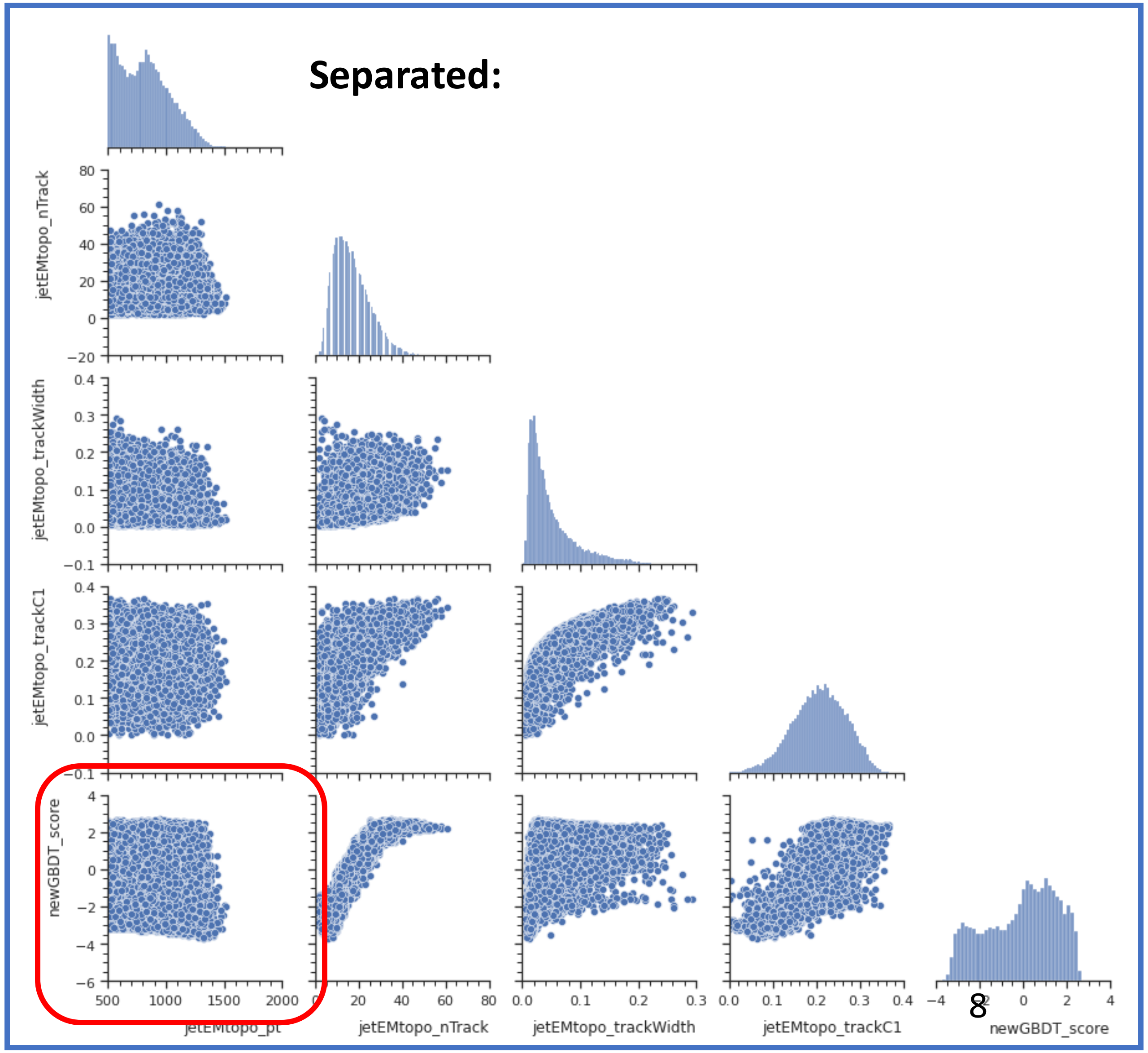
correlation with flatpt_weight: used to train

	po_pt	Track	Width	ckC1	score
jetEMtopo_pt	1	0.038	-0.1	-0.11	-0.13
jetEMtopo_nTrack	0.038	1	0.53	0.71	0.88
jetEMtopo_trackWidth	-0.1	0.53	1	0.76	0.45
jetEMtopo_trackC1	-0.11	0.71	0.76	1	0.77
newGBDT_score	-0.13	0.88	0.45	0.77	1

Correlation using different flat weights



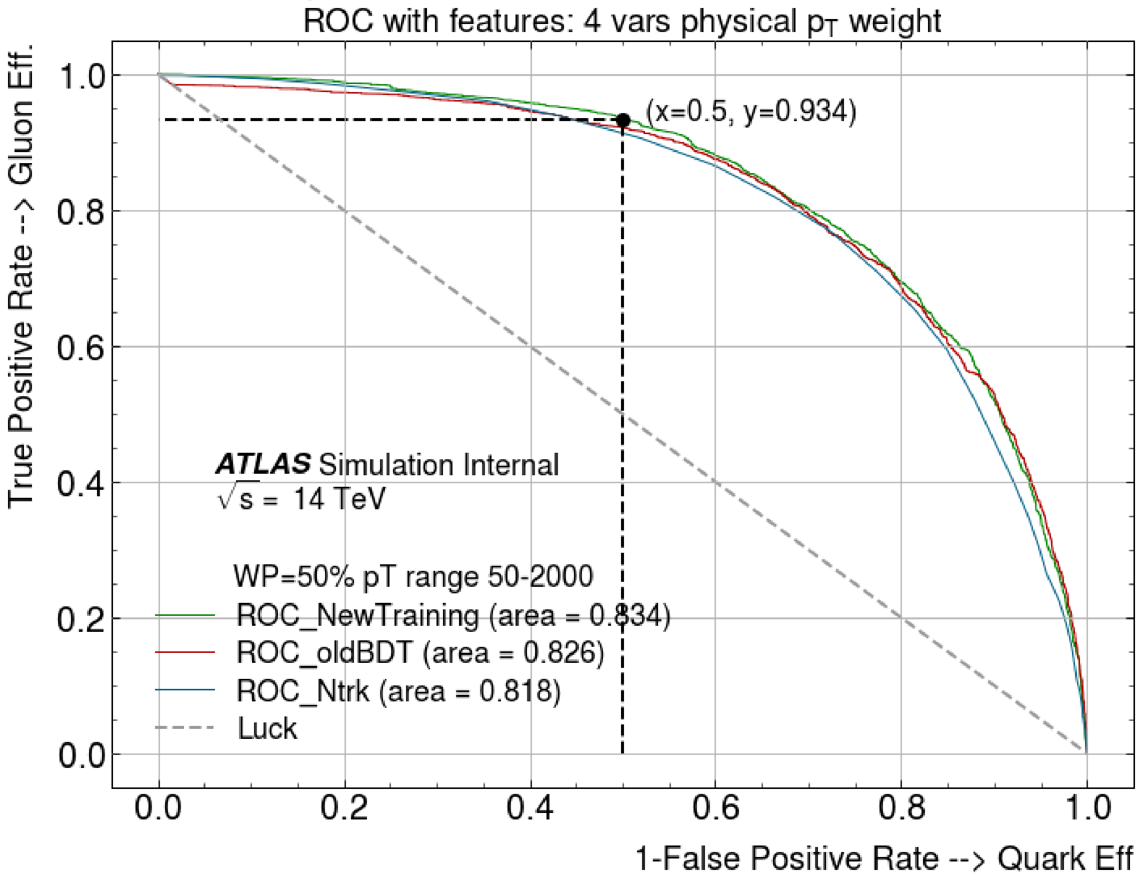
Together:



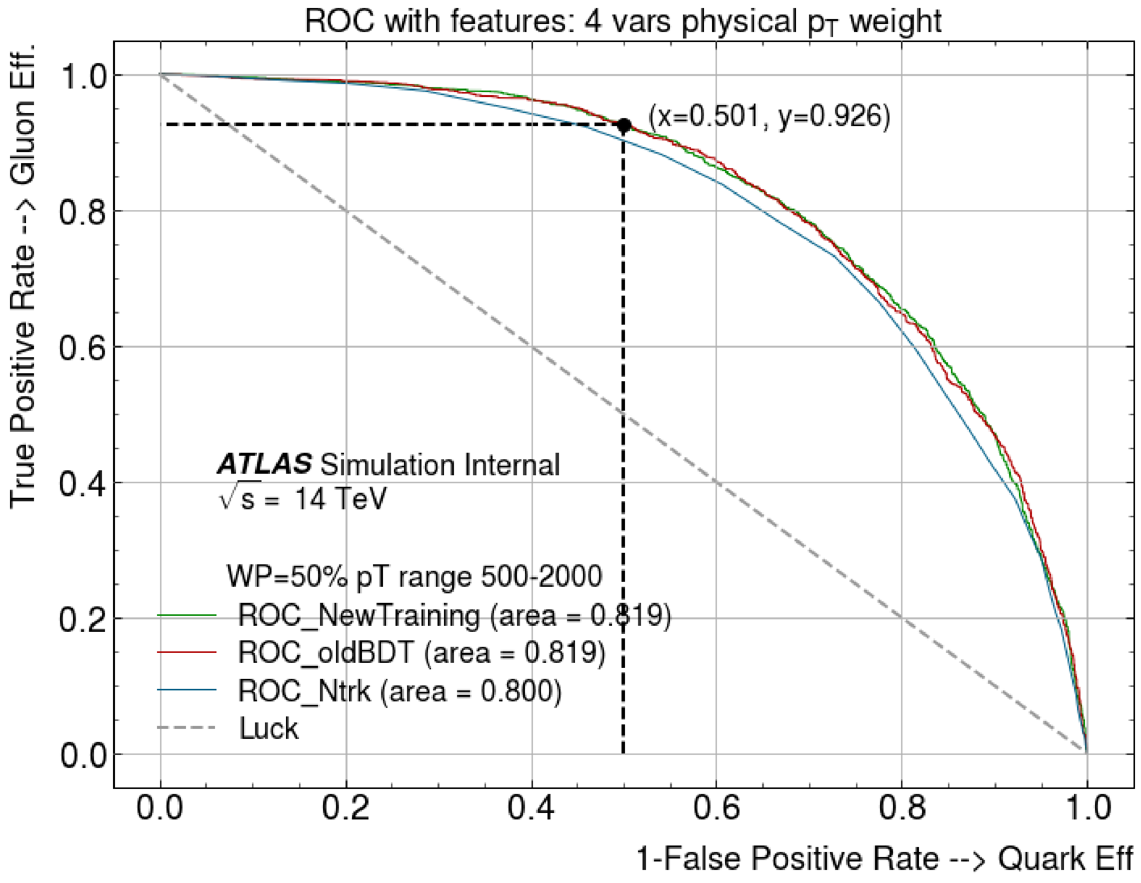
Separated:

Roc curves

Together:



Separated:



AUC is lower for flat separated weights, dependency with pT completely erased

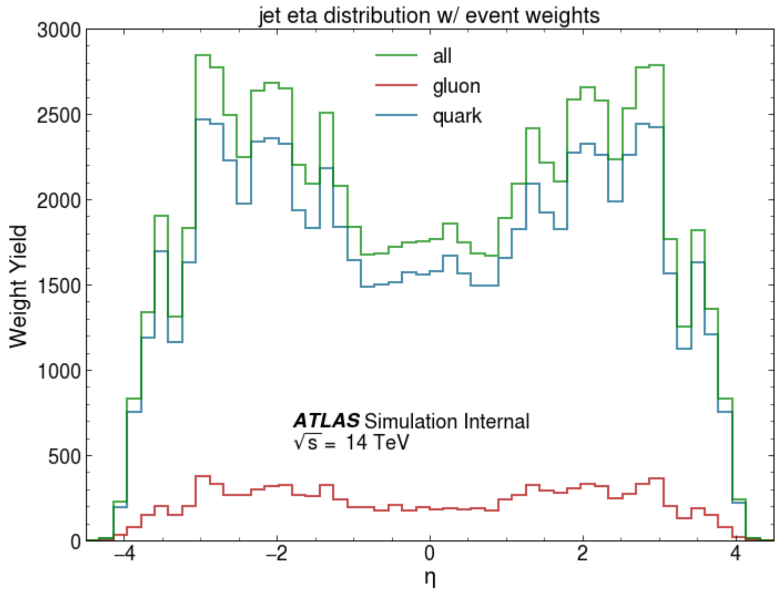
Tracks in the forward region

- Tracks weren't defined in the forward region. Derivations don't know if the sample is one from Run 3(or 2) vs HL-LHC. Eta cut is not updated with the forward region.
- Long discussion with conveners and experts
- Pierre-Antoine Delsart is working in a way to update cuts by reading geometry flags. Still on going.
- I did one local production of JEM1 derivation by manually forcing the eta cut to 4
- New VBF sample doesn't have the cross-section stored.
    - 600026.PhH7EG_NNPDF3_AZNLO_VBFH125_ZZ4nu_MET75. Only sample populated in the forward region
    - Physical weights are not used for validation (10%) and testing (10%). Weights forced to be 1.
- From now on only flat separated weights are used (80%)

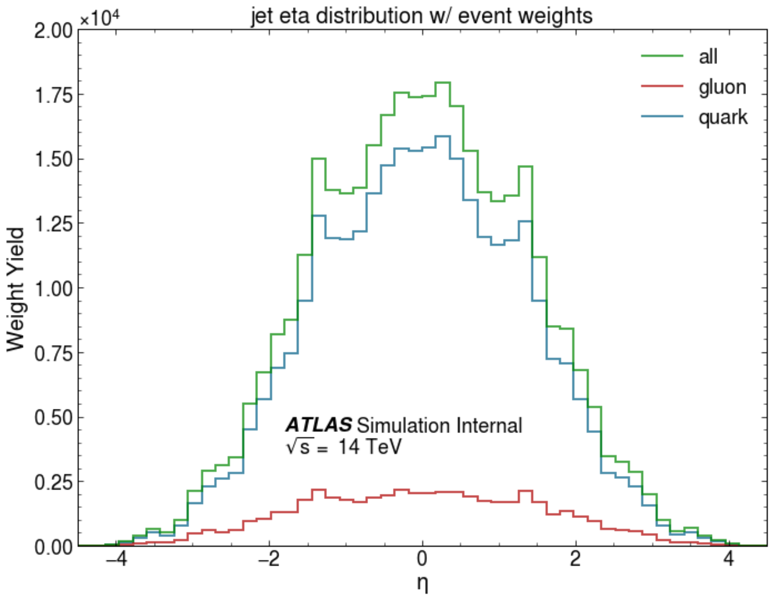
New selection

- Abs(eta) < 4.0, pT > 50 GeV. Number of jets > 0. Number of tracks > 1. Weights == 1

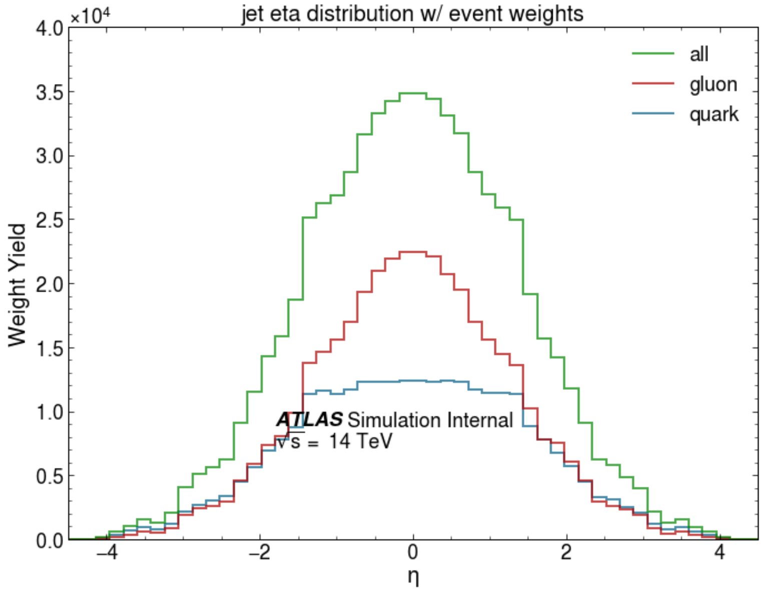
VBF H 125



ttbar single lep, dilep and all had

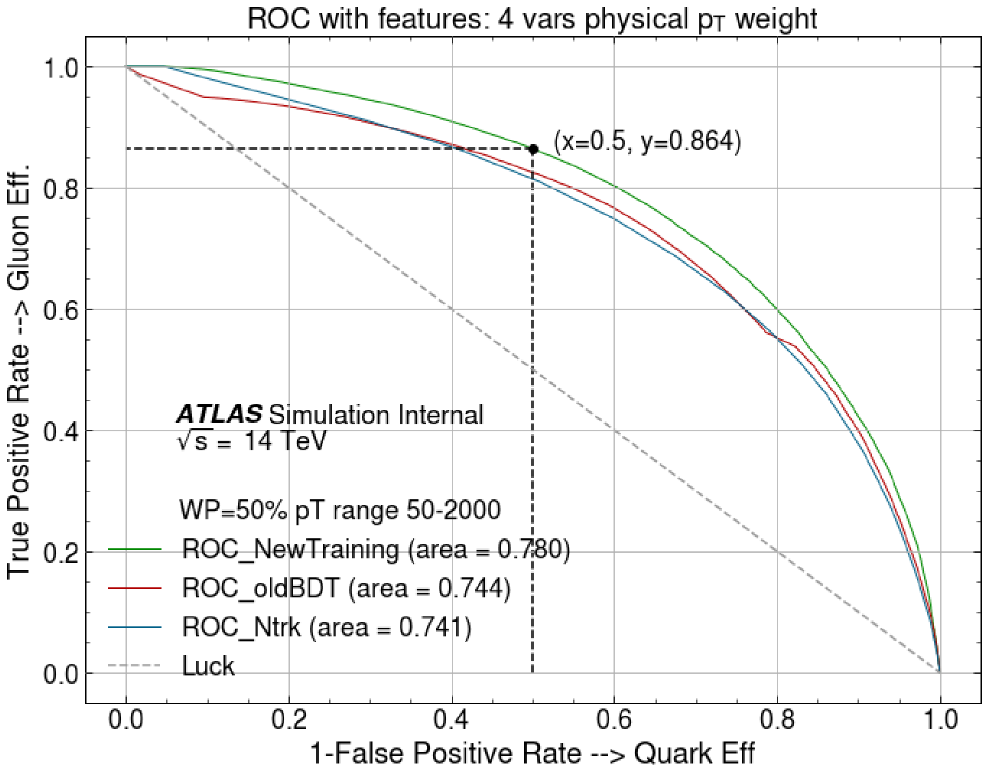


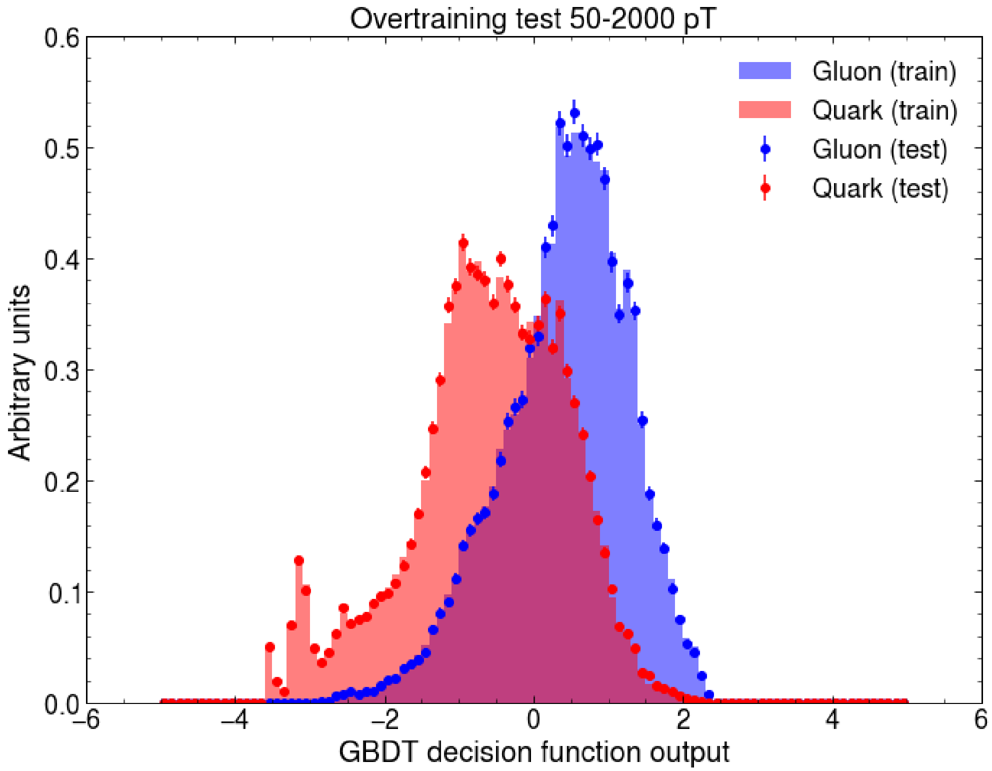
dijets



Need to find an sample populated with forward gluon jets

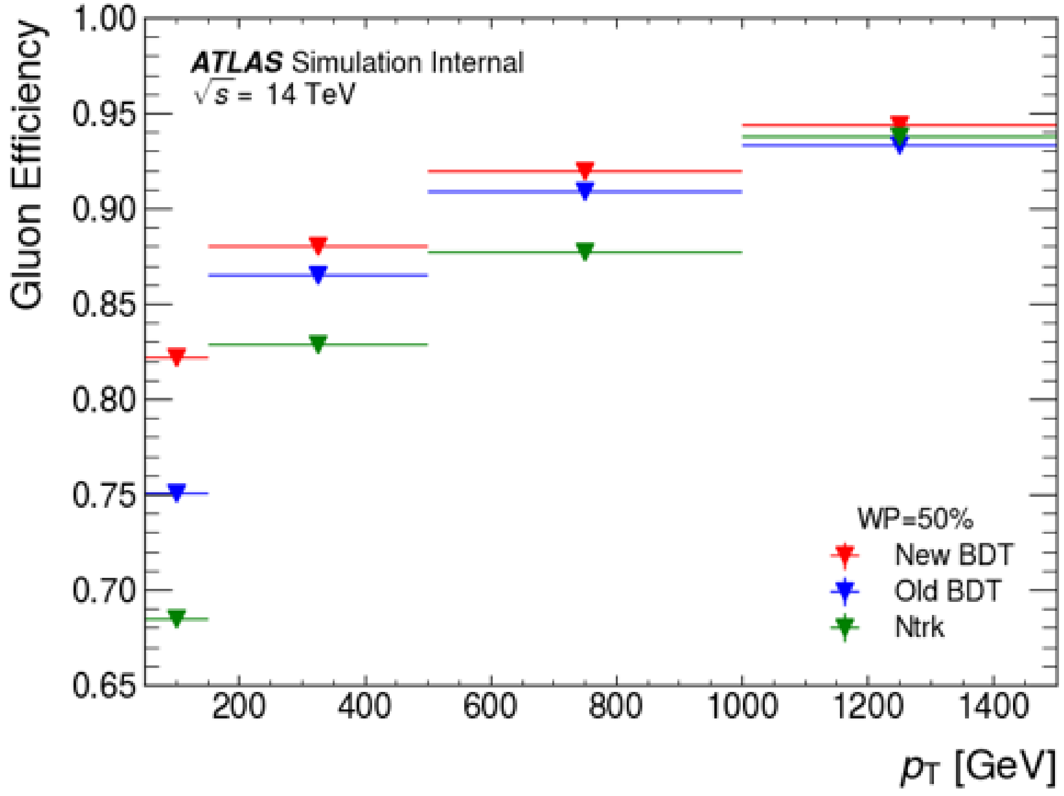
Roc curve and overtraining

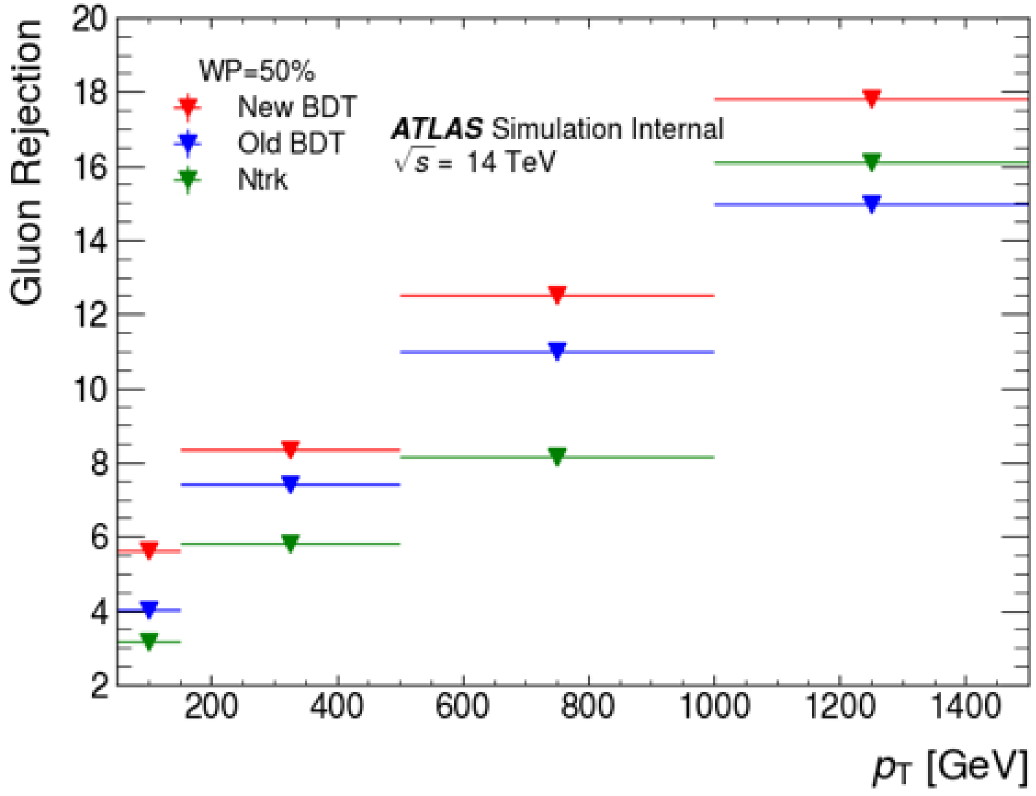




Validation and testing more stable with the new weights

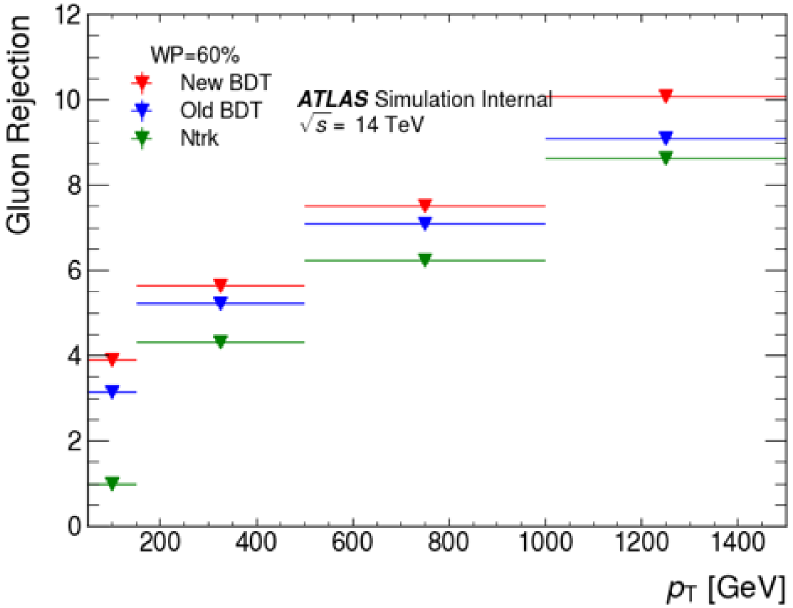
Gluon efficiency and Rejection (WP=50%)

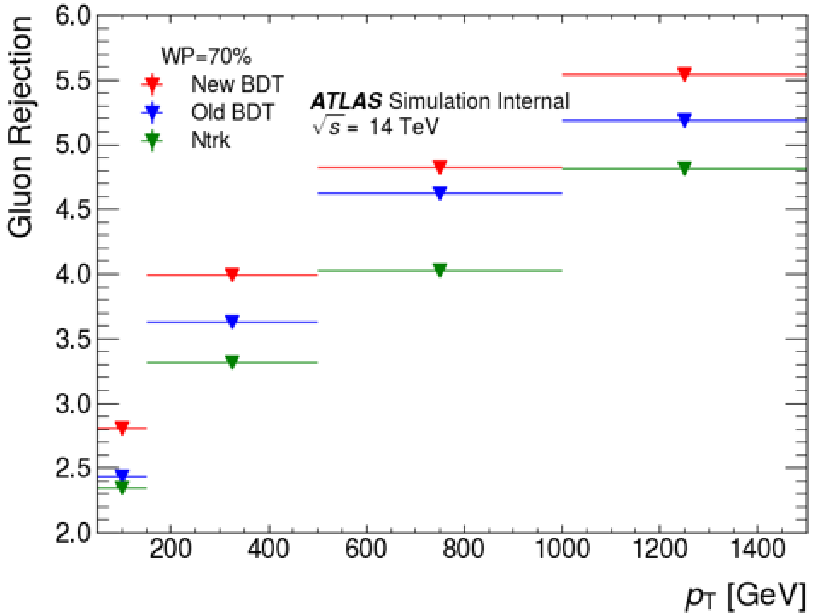


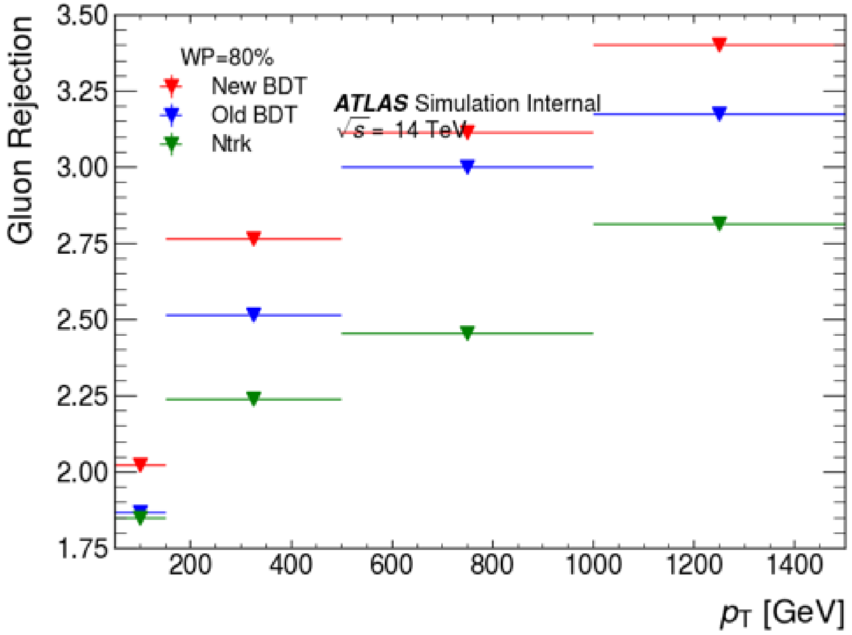


New BDT performs better than the old BDT (the one in athena) and number of tracks tagger

Gluon Rejection (WP=60%, 70%, 80%)







New BDT performs better than the old BDT (the one in athena) and number of tracks tagger

Outlook

- Flat distributions correctly flattened for quarks and gluons separately
- VBF H and ttbar were added, still need a sample populated with forward jets
- Forward region is added. New BDT has a better gluon eff and rej than the old BDT and number of tracks
- Move to transformers
- Should I start to document my results?

Backup

Roc curve clarification

- Signal --> gluon (1). Background --> quark (0)
- **TPR**= A true positive is an outcome where the model correctly predicts the positive class --> **Gluon efficiency**
- FPR = FP / (FP + TN)
  - FP (False Positive) – The *positive* instances *incorrectly* classified. (quark tagged gluon)
  - TN (True Negative) – The *negative* instances *correctly* classified. (quark tagged quark)
- **Quark efficiency = 1 – FPR** = TN/(FP+TN) = quark tagged quark/ all quarks

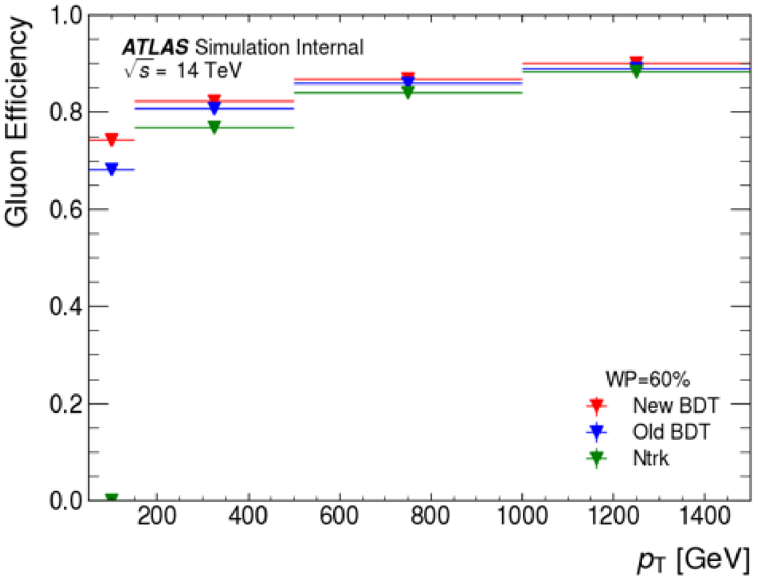
Size comparison

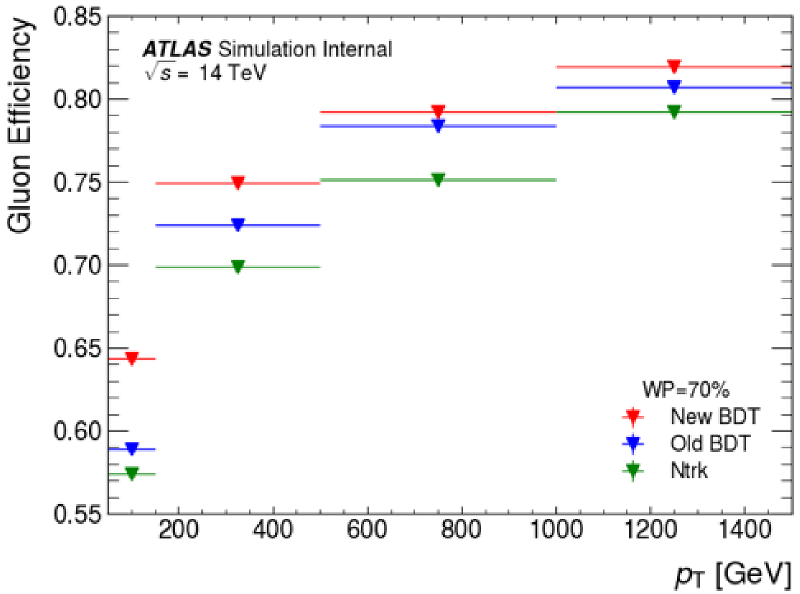
- **Run 2**
  - After cut applied --> pt > 500 GeV and pt < 2000 GeV, abs(eta_jet1) < 2.1, abs(eta_jet2)<2.1, event_weight < 100 and number of jets >1 & ptjet1/ptjet2 < 1.5 & nTracks > 1
  - Memory usage: 17.3 GB
    - Training data: 4.0 GB, 108.049.937 entries (80%)
    - Testing data: 515.2 MB, 13.506.243 entries (10%)
    - Validation data: 515.2 MB, 13.506.243 entries (10%)
- **HL-LHC**
  - After cut applied --> pt > 500 GeV and pt < 2000 GeV, abs(eta_jet1) < 2.1, abs(eta_jet2)<2.1, event_weight < 100 and number of jets > 1 & nTracks > 1
  - Memory usage: 16 MB
    - Training data: 5 MB, 109.684 entries (jets) (80%)
    - Testing data: 642.7 KB, 13.711 entries (10%)
    - Validation data: 642.7 KB, 13.711 entries (10%)
- HL-LHC input entries (without any cut) -> 6 samples * 10 root files of 10.000 events each = 600.000 events
- After pt, eta, number of jet requirements -> 65.000 events

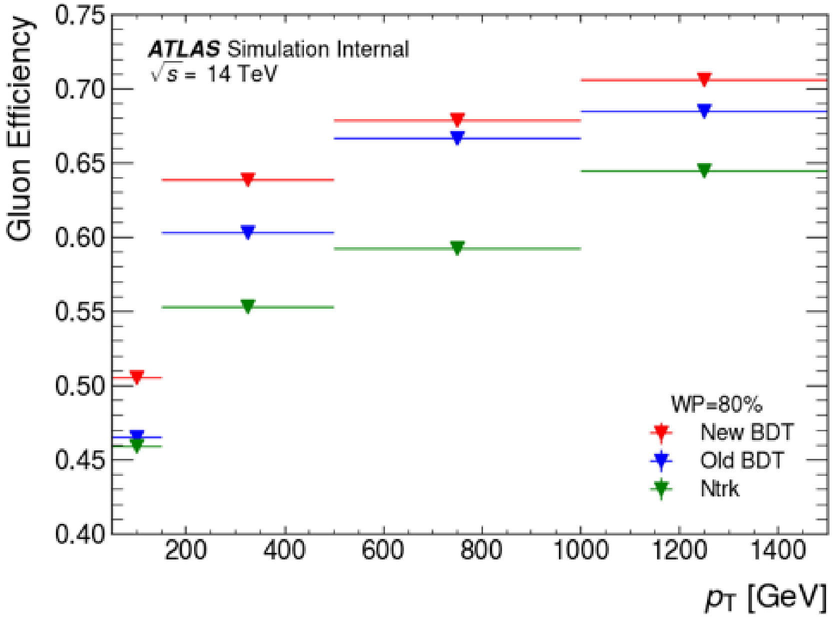
Ntracks

- NumTrkPt500PV: the number of tracks from the primary vertex with a track pT of at least 500 MeV associated to a jet.

Gluon Efficiency (WP=60%, 70%, 80%)


ATLAS Simulation Internal
$\sqrt{s}$ = 14 TeV
Gluon Efficiency
p_T [GeV]
WP=60%
New BDT
Old BDT
Ntrk


ATLAS Simulation Internal
$\sqrt{s}$ = 14 TeV
Gluon Efficiency
p_T [GeV]
WP=70%
New BDT
Old BDT
Ntrk


ATLAS Simulation Internal
$\sqrt{s}$ = 14 TeV
Gluon Efficiency
p_T [GeV]
WP=80%
New BDT
Old BDT
Ntrk