Flavour Tagging with Graph Neural Network with the ATLAS Detector

Arnaud Duperrin
Aix-Marseille Université, CNRS/IN2P3, CPPM, Marseille

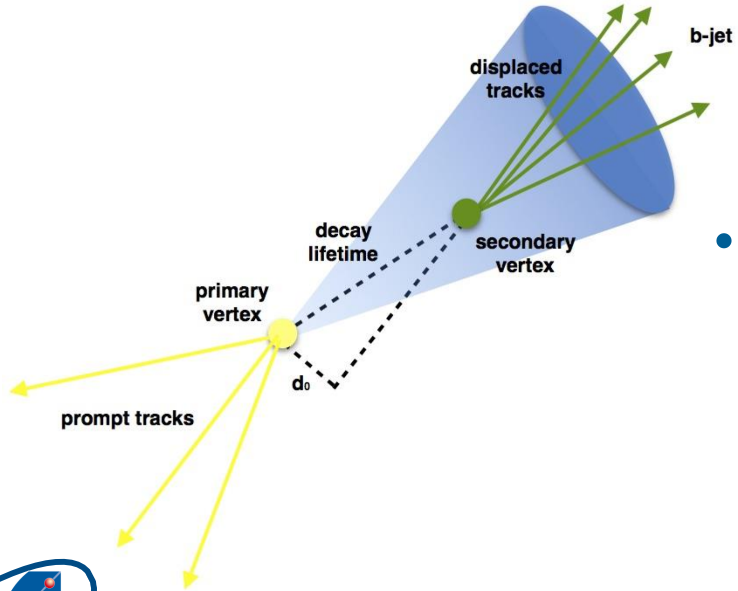
on behalf of ATLAS collaboration

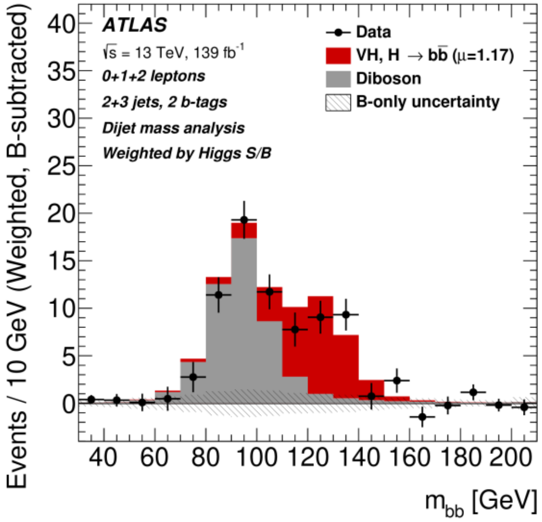
DIS2023 in Michigan State University, 30/03/2023

b-tagging at LHC

Eur. Phys. J. C 81 (2021) 178

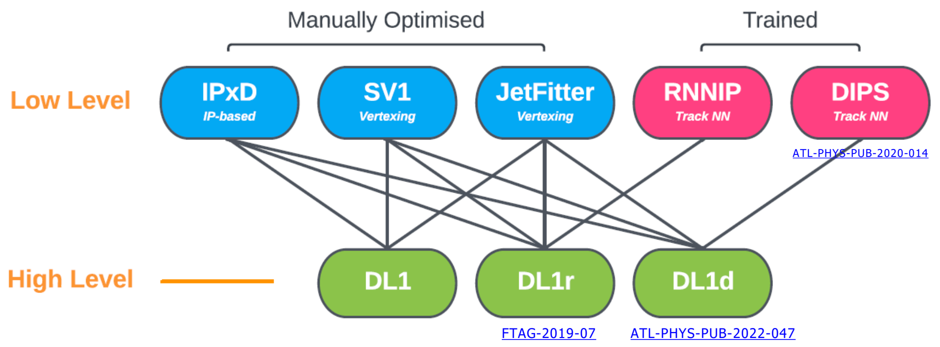
- Flavour tagging
  - Aims to identify the Flavour of the jet (b, c, light)
  - Processes with heavy flavour quarks (b,c) play a key role in the LHC physics program (ex. H$\rightarrow$bb)

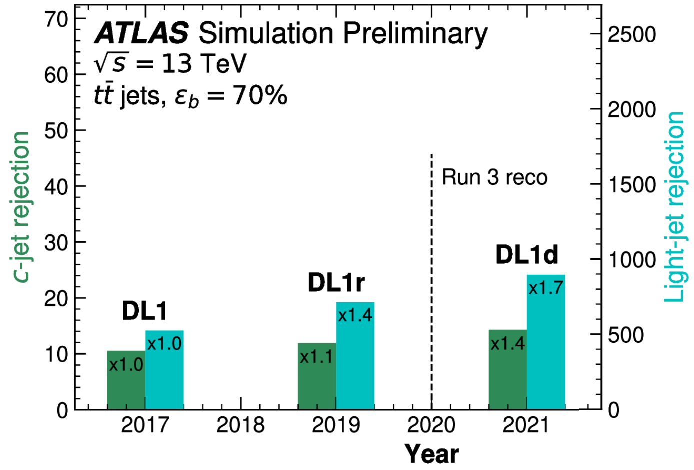




- Features
  - At least one secondary vertex displaced by a few mm from the hard-scatter collision point
  - Large track impact parameters (d0, z0 x sin θ)
  - B$\rightarrow$C, semileptonic decays, etc.

ATLAS Flavour Tagging Machine Learning





- Based on a two-stage approach
  - **Low-level algorithms** use properties of individual charged-particle tracks associated with a jet or combine tracks to explicitly reconstruct displaced vertices
  - The outputs are fed into **high level taggers** which are Neural Networks (**DL1 series**)
- Recent Analyses of Run 2 and Run 3 LHC data based on new high-level algorithms
  - Based on recurrent **(DL1r**) and deep sets networks (**DL1d**)
  - Considerable improvements over previous work (which were based on boosted decision trees or likelihood discriminants)

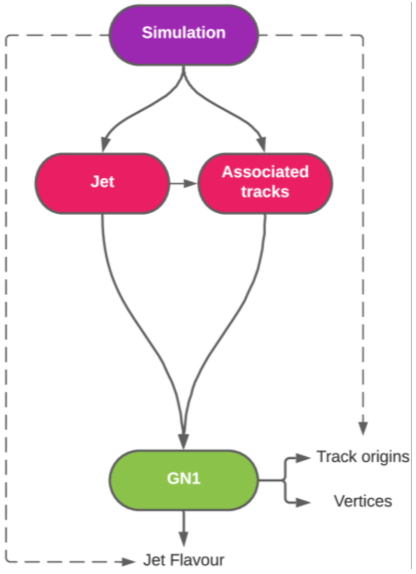
Flavour Tagging with Graphs

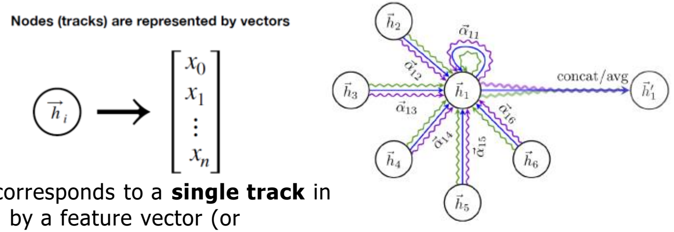
[ATL-PHYS-PUB-2022-027]

- A new approach with a "all-in-one tagger" : GN1
  - Utilizes a **single neural network** to predict the **jet flavour**, directly taking as **inputs**:
    - **jet** pT, η
    - **tracks** parameters, uncertainties, impact parameters
    - detailed **hit** information
  - Simplification: no need for low level algorithms and tuning, permutation invariant (no ordering)
- Trained with auxiliary objectives
  - Grouping of tracks originating from a common vertex (vertex "finding" only, no vertex "fitting")

  → **vertexing**
  - Prediction of the underlying physics process from which each track originated (b,c, light, pile-up, fake, etc.)

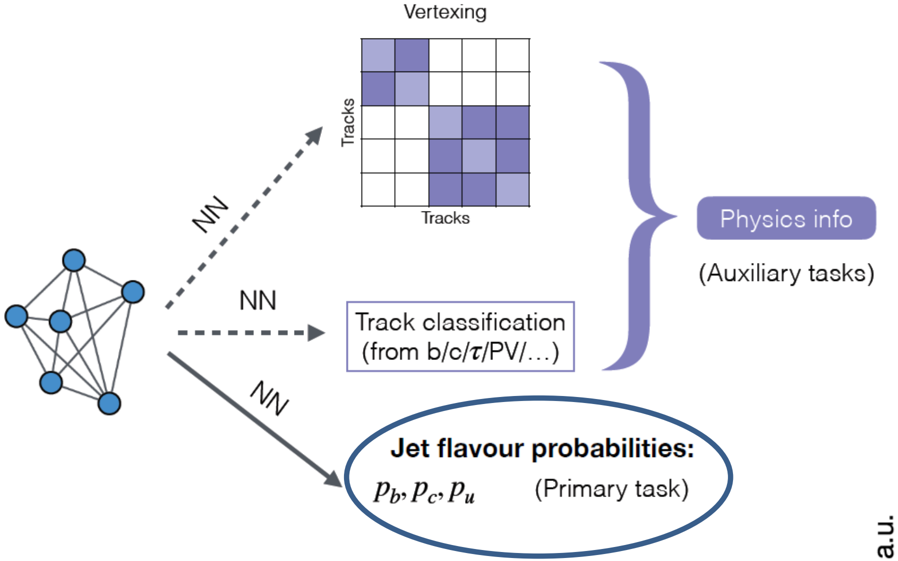
  → **track origins**





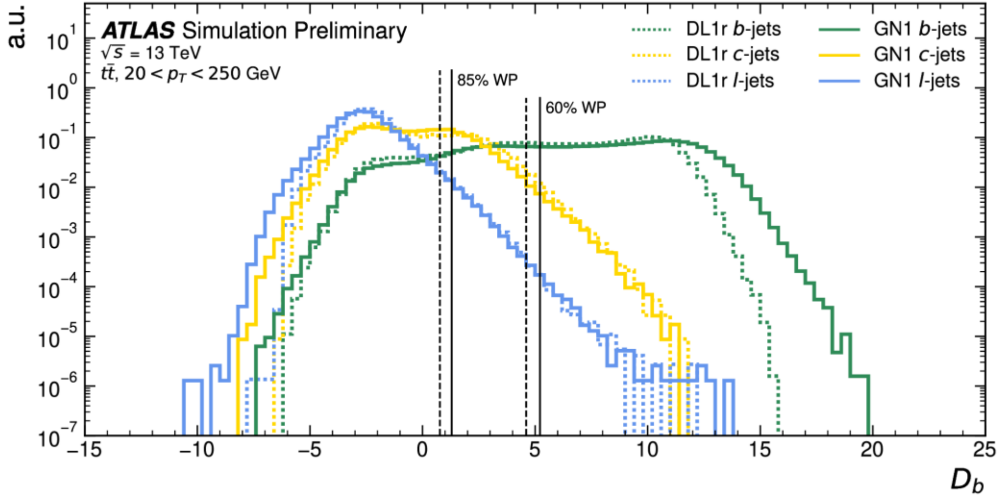
Each **node** h_i in the graph corresponds to a **single track** in the jet, and is characterised by a feature vector (or representation) of length 23 based on the above inputs

Building the discriminant



$$D_b = \log\left[\frac{p_b}{f_c \cdot p_c + (1 - f_c)\cdot p_u}\right]$$

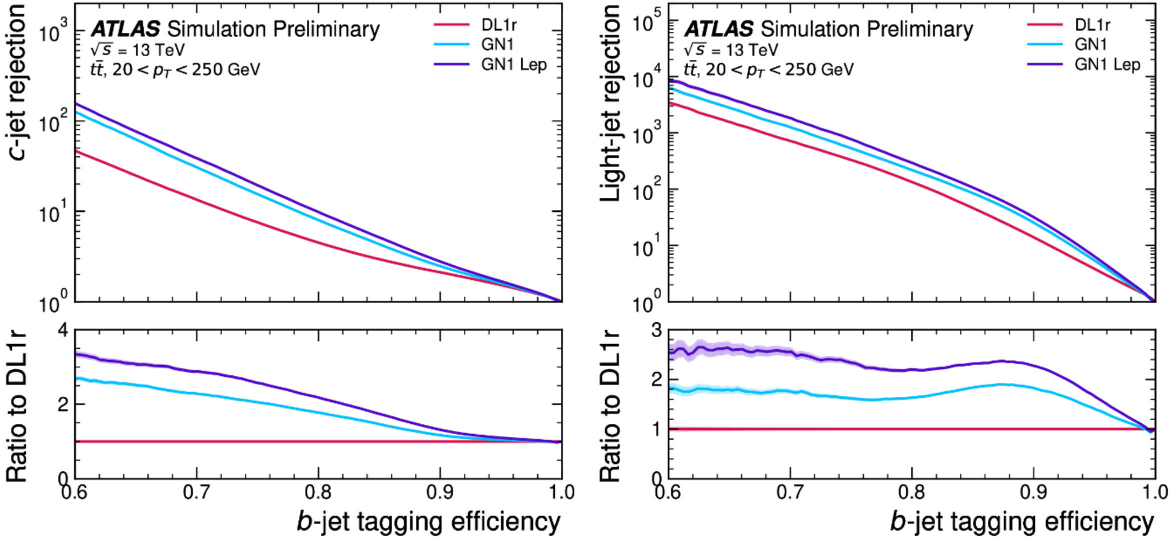
f_c is a free parameter that determines the relative weight of p_c to p_l in the score D_b



- c-tagging
  - Single model is used for both **b-tagging** and **c-tagging** !

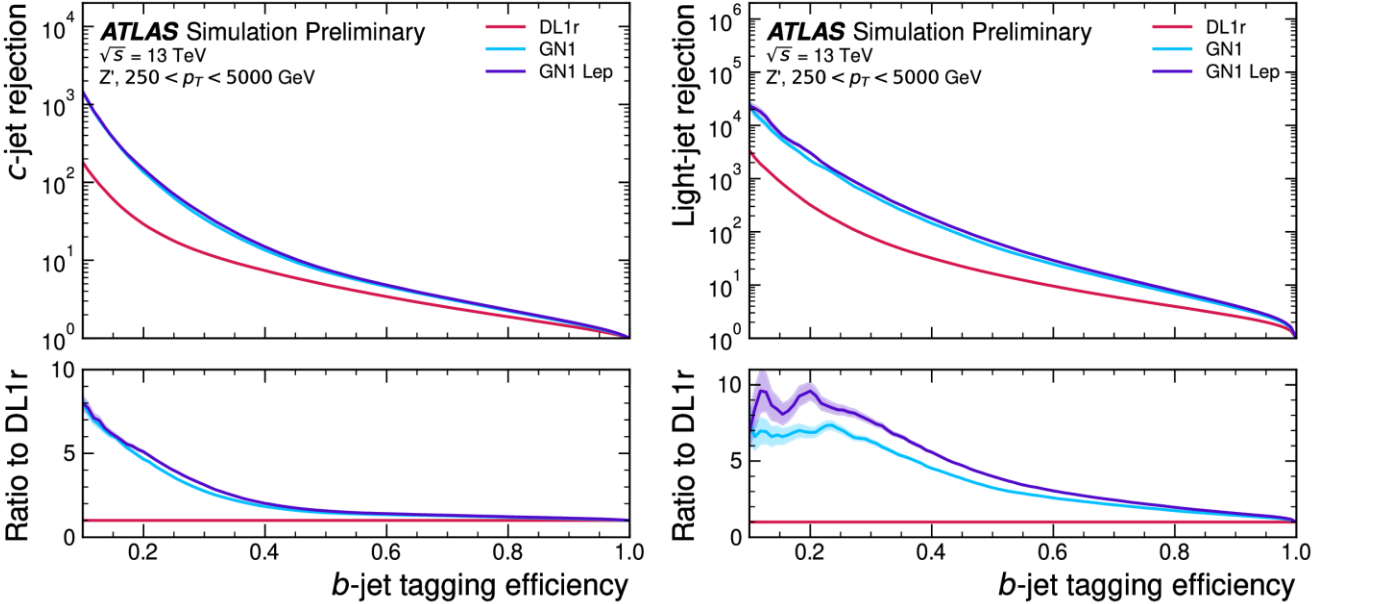
$$D_c = \log\left[\frac{p_c}{f_b \cdot p_b + (1 - f_b)\cdot p_u}\right]$$

Performance: ttbar



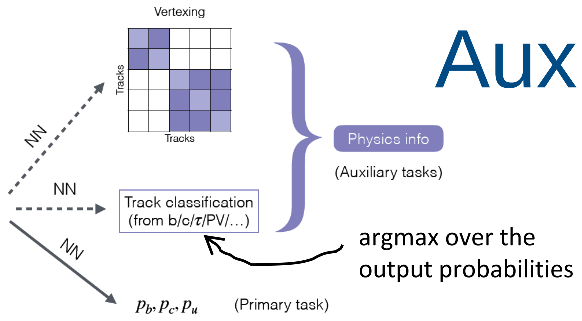
- **GN1** demonstrates **considerably better** *c*- and light-jet rejection compared with DL1r
- At the 70% working point (WP)
  - x2.1 in c-jet rejection
  - x1.8 in light-jet rejection
- **GN1 Lep** variant includes an additional track-level input which indicates if the track was used in the reconstruction of an *electron* or a *muon* ⇒ improved performance with respect to the baseline GN1 model

Performance: Z' (i.e. high pT)



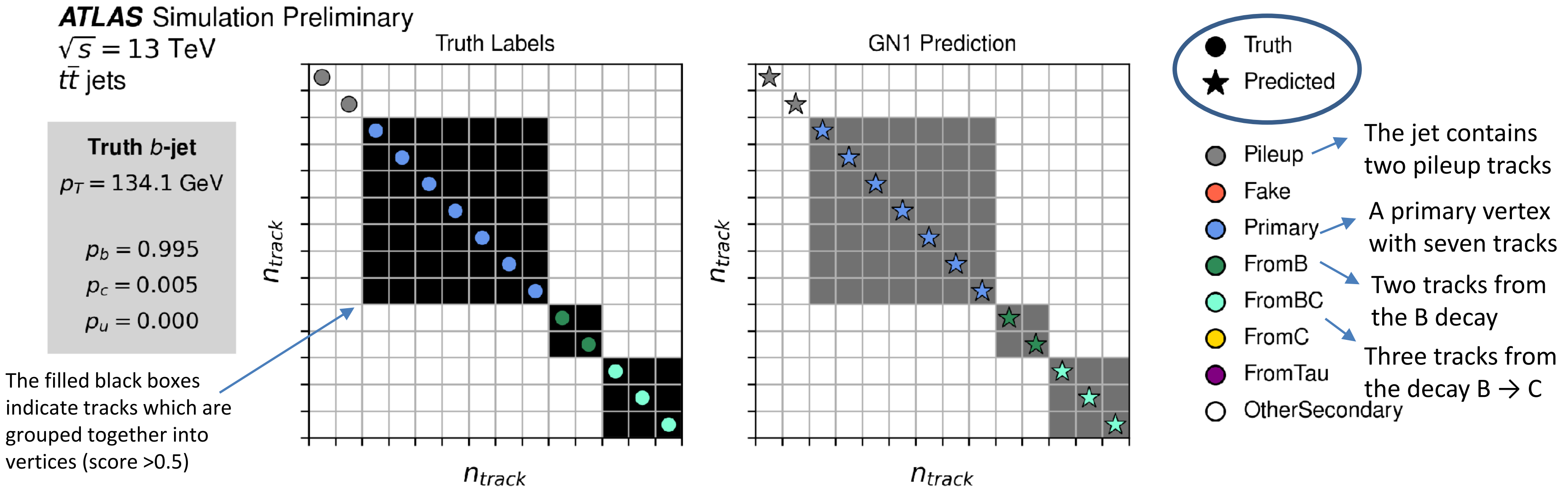
- At the 30% working point (WP)
  - x2.8 in c-jet rejection
  - x6 in light-jet rejection

Auxiliary tasks: track origins + vertexing



Truth Origin	Description
Pileup	From a *pp* collision other than the primary interaction
Fake	Created from the hits of multiple particles
Primary	Does not originate from any secondary decay
fromB	From the decay of a *b*-hadron
fromBC	From a *c*-hadron decay, which itself is from the decay of a *b*-hadron
fromC	From the decay of a *c*-hadron
OtherSecondary	From other secondary interactions and decays

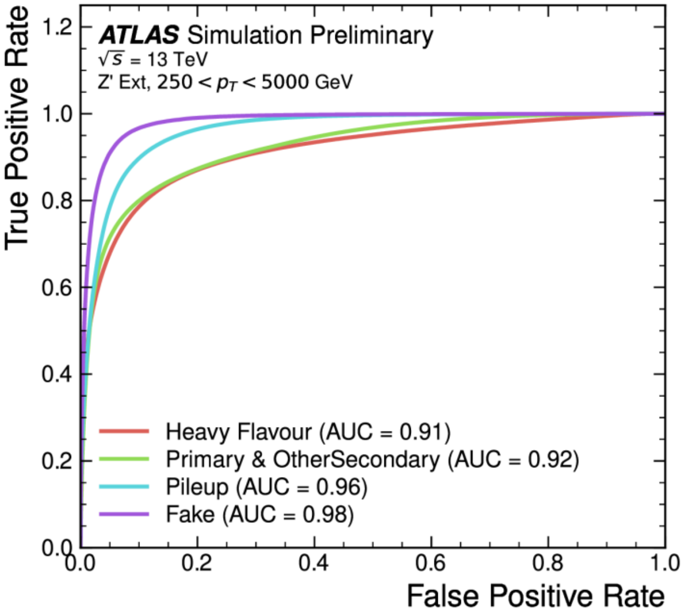
- Cherry-picked example: quantitative analysis of the track classification and vertexing tasks → next slide



- GN1 correctly predicts track origins and vertex compatibility for all tracks in the jet in this example

Auxiliary tasks performance

- **Tracks origins**
  - Each group of origins has a good classification performance with a AUROC > 0.9
- **Vertexing**
  - An efficient vertex requires:
    - recall of > 65% of tracks in the truth vertex
    - purity of > 50% of tracks in the reco vertex
  - GN1 correctly identifies 80% of truth vertices inside b-jets



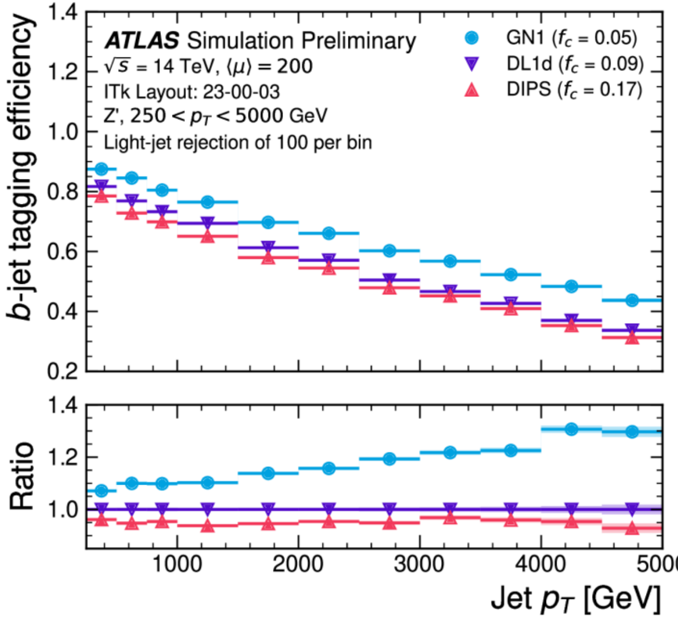
Auxiliary tasks:

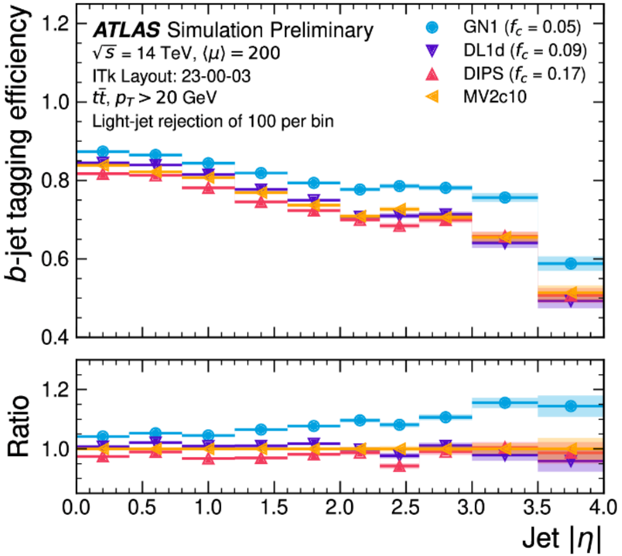
- Helps the jet flavour prediction via **supervised attention**: in detecting tracks from heavy flavour decays ⇒ the model learns to pay more attention to these tracks
- Helps with the **interpretability** of the network

High Luminosity LHC

- HL LHC:
  - Planned to be completed by 2029
  - Average **200 pp collisions** per bunch crossing!

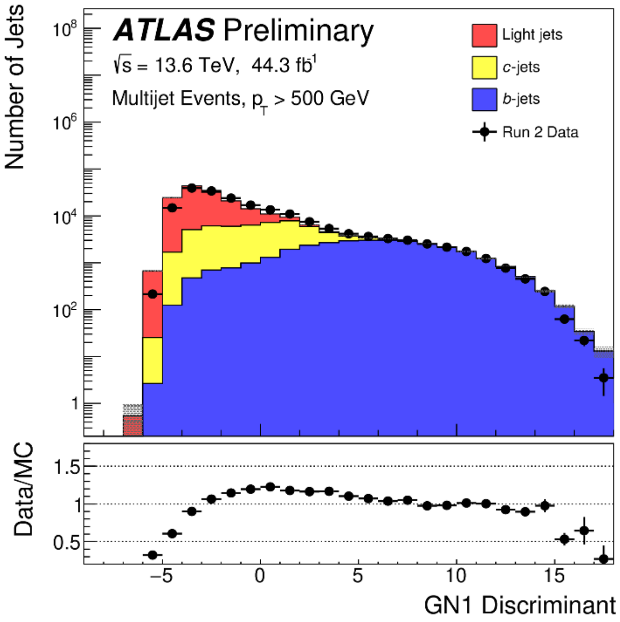
    ⇒ makes tagging even more challenging
  - Significant upgrade of the **tracking detector (ITk)** will improve Flavour Tagging [ATL-PHYS-PUB-2022-047]

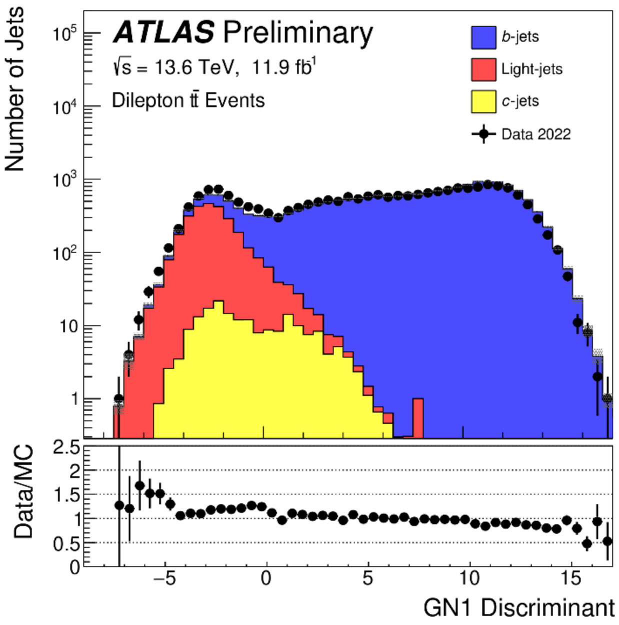


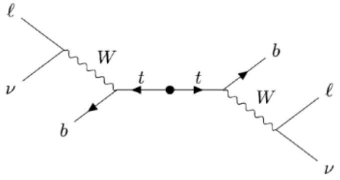


- GN1 improvements with respect to previous generation of flavour tagging algorithms
  - Up to 30% improvement in b-efficiency at high-pT
  - 15% improvement in the forward region (η>2.5)

Run 3 **data/MC** Agreement: Multijet / ttbar Dilepton







The simulation is scaled to match the total yield in data

Run 2 b/c/light calibration:

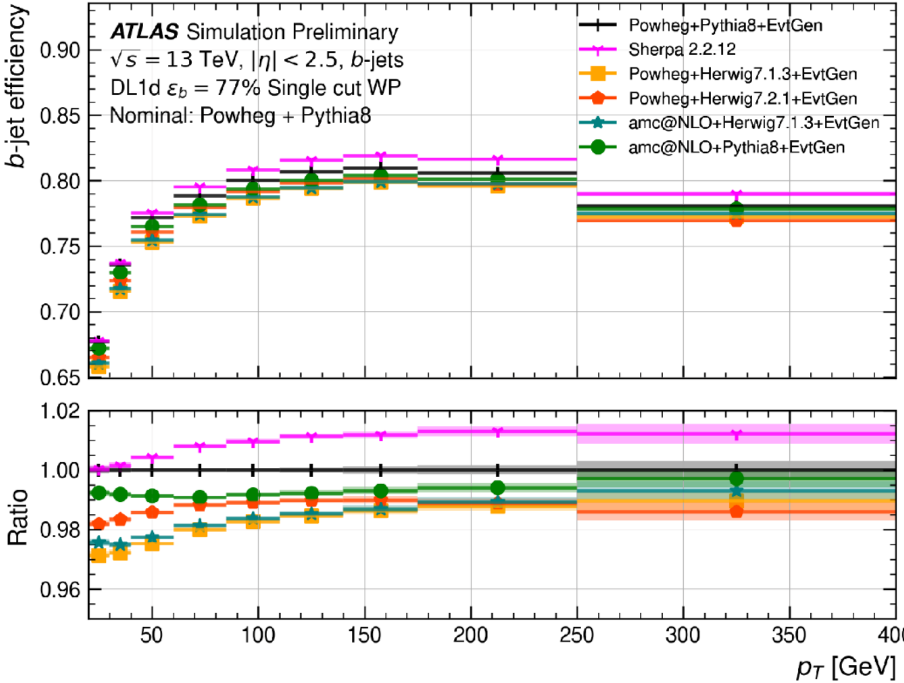
- ttbar: [Eur. Phys. J. C 79 (2019) 970]
- Multijet: [ATL-PHYS-PUB-2022-010]

- **Multijet** with “tag and probe”: tag jet with pT> 200 GeV and 85% WP
  - Probe jets shown with pT>500 GeV ⇒ Good agreement (the low discriminant tail, small increase in discrepancy)
- **Dilepton** ttbar: OS eμ (pT>28 GeV) with single lepton trigger + 2 jets (pT>20 GeV)
  - Invariant mass of each lepton-jet pair must be below 175 GeV
  - Leading jet pT shown ⇒ Very good agreement

Generator Dependence

- Models are trained on "best-knowledge" samples
  - Low-pT ttbar : POWHEGBOX + PYTHIA8 + EVTGEN
  - High-pT Z′: PYTHIA8 + EVTGEN
- Testing the GN1 model on other MC samples to check if it is not learning generator dependent information

- Overall generator dependence:
  - O(3%) for b-jets and O(6%) for c-jets
  - Similar to previously observed values for DL1r/DL1d



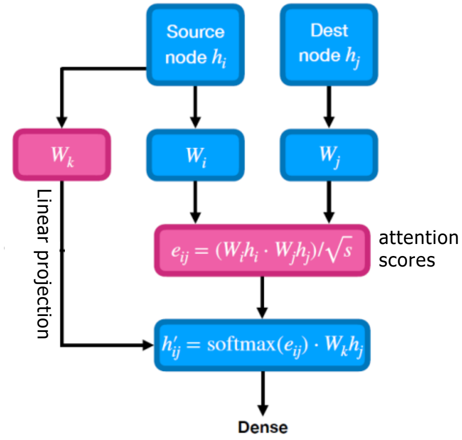
Pushing Further Improvements (GN2)

[FTAG-2023-01]

- GN1 → GN2
  - The majority of the changes are optimisations for the model hyper parameters.
- Optimised training
  - Learning rate optimisation→ using One-Cycle learning rate scheduler [1803.09820]
- Updated architecture
  - Follows **transformer** architecture [1706.03762]:

Type	Name	GN1	GN2
Hyperparameter	Trainable parameters	0.8M	1.5M
Hyperparameter	Learning rate	$1e{-3}$	OneCycle LRS (max LR $4e{-5}$)
Hyperparameter	GNN Layers	3	6
Hyperparameter	Attention Heads	2	8
Hyperparameter	Embed. dim	128	192
Architectural	Attention type	GATv2	ScaledDotProduct
Architectural	Dense update	No	Yes (dim 256)
Architectural	Separate value projection	No	Yes
Architectural	LayerNorm + Dropout	No	Yes
Inputs	Num. training jets	30M	192M

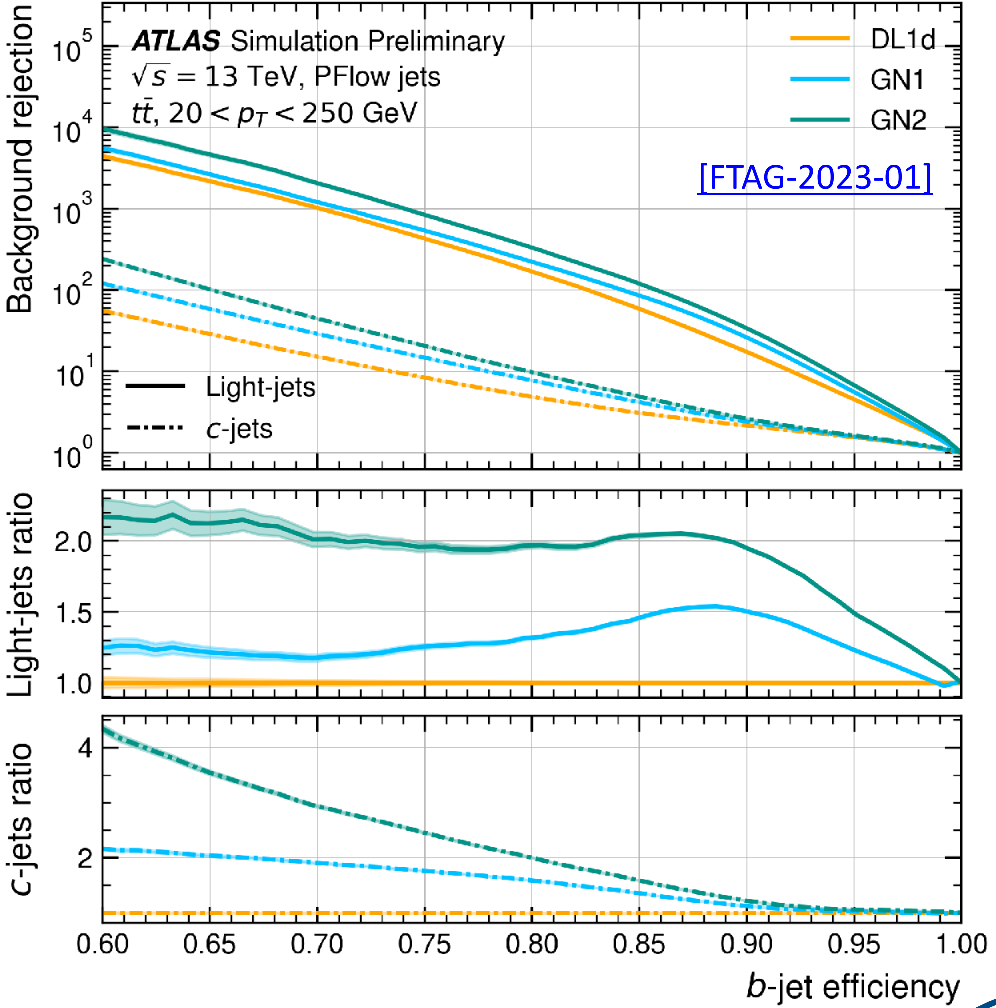
  - The attention type has been changed → ScaledDotProduct [1706.03762]: no effect on physics performance but improves the training time and memory footprint.
  - Using the separate linear projection→ separates the computation of the attention weights from the computation of the updated node representations
  - Using a dense layer in between the attention layers



- Increased training statistics
  - 30M (based on down sampling) → 192M training jets (PDF sampling)

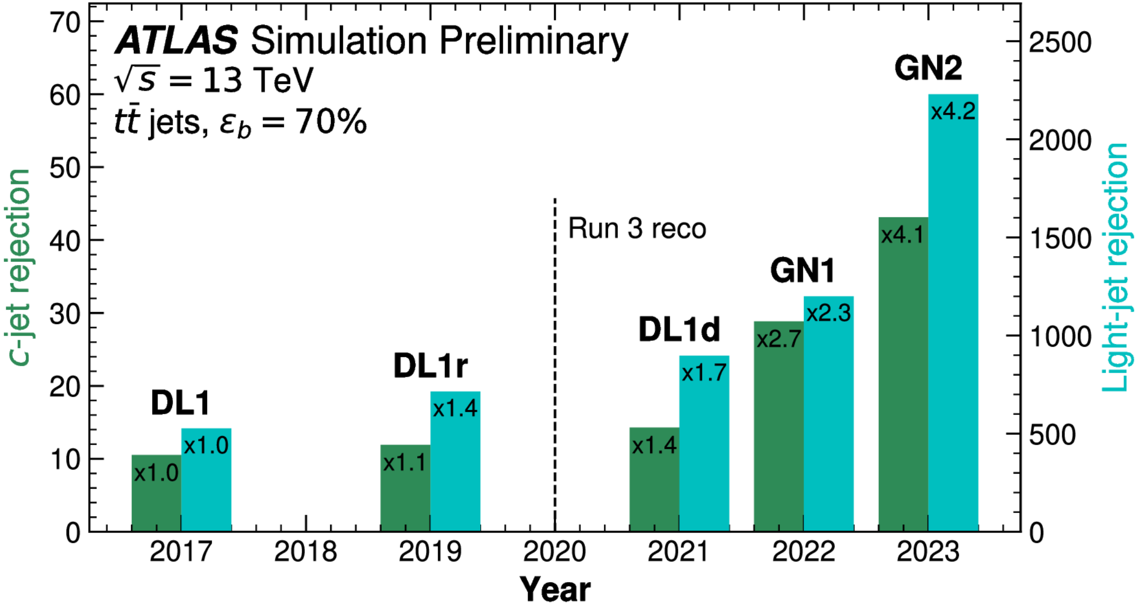
GN2 ttbar Performance

- Ratio to DL1d
  - **DL1d** is a modification of the **DL1r** tagger [FTAG-2019-07] with the DIPS network [ATL-PHYS-PUB-2020-014] replacing RNNIP
- **GN2 compared to GN1**
  - 1.5x c-rejection and 2x light-rejection on **ttbar**
  - 1.75x c-rejection and 1.2x light-rejection on **Z’**



Conclusion

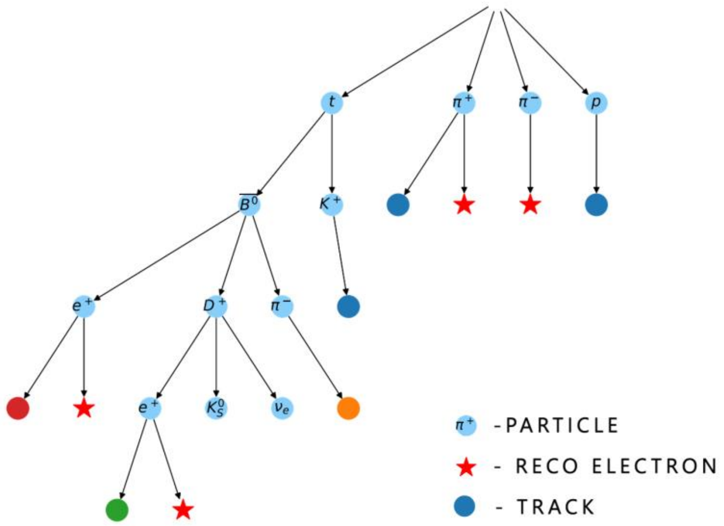
- Next generation **b/c taggers** based on **Graph Neural Networks** show very promising results
  - **4x improvement** in background rejection since DL1
  - MC/MC and Data/MC checks performed ⇒ moving toward **full calibration**
  - Expect strong benefit on ATLAS physics program at **Run 3 LHC and HL-LHC**



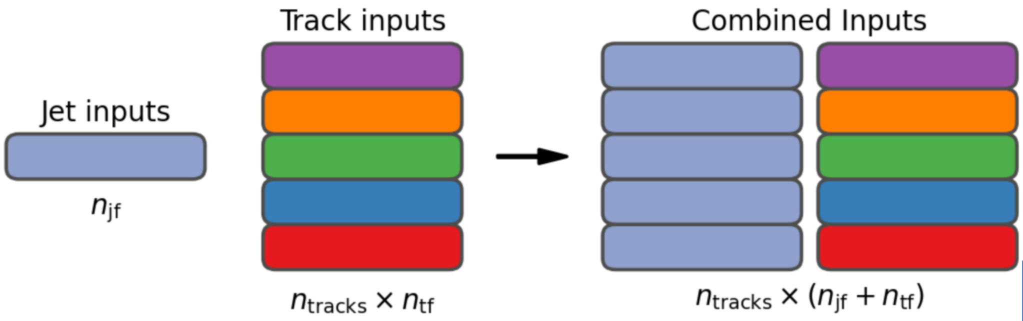
backup

Why Graph Neural Networks ?

- GNN easily applicable in domains where:
  - The data can be represented as a set of nodes
  - The prediction depends on the relationships (edges) between the nodes
- HEP data naturally match to the graph structure
  - HEP data is heterogeneous and sparse
  - Variable number of input items (e.g., tracks in the event)
  - No primordial ordering (physics-inspired imposed ordering reduces performance)
  - Build representations of each jet/track/object aware of the features of the other jets/tracks/objects in the event
  - Easy to introduce auxiliary physics-inspired tasks
- Graph Attention Networks
  - Considered as the state-of-the-art of GNN architecture



GN1 architecture: Inputs



- Inputs to GN1
  - Two jet features (n_{jf} = 2): pT and η
  - An array of n_{tracks}
    - each track is described by 21 track features (n_{tf} = 21).
    - The jet features are copied for each of the tracks
- The combined **jet-track vectors** of length 23 form the inputs of GN1

Jet Input	Description
p_T	Jet transverse momentum
η	Signed jet pseudorapidity

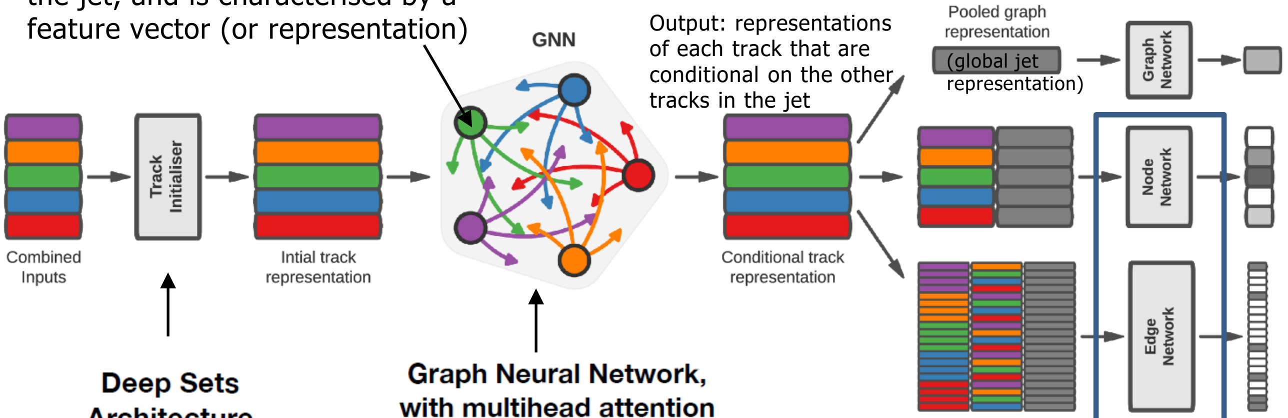
Track Input	Description
q/p	Track charge divided by momentum (measure of curvature)
$\mathrm{d}\eta$	Pseudorapidity of the track, relative to the jet η
$\mathrm{d}\phi$	Azimuthal angle of the track, relative to the jet ϕ
d_0	Closest distance from the track to the PV in the longitudinal plane
$z_0 \sin\theta$	Closest distance from the track to the PV in the transverse plane
$\sigma(q/p)$	Uncertainty on q/p
$\sigma(\theta)$	Uncertainty on track polar angle θ
$\sigma(\phi)$	Uncertainty on track azimuthal angle ϕ
$s(d_0)$	Lifetime signed transverse IP significance
$s(z_0)$	Lifetime signed longitudinal IP significance
nPixHits	Number of pixel hits
nSCTHits	Number of SCT hits
nIBLHits	Number of IBL hits
nBLHits	Number of B-layer hits
nIBLShared	Number of shared IBL hits
nIBLSplit	Number of split IBL hits
nPixShared	Number of shared pixel hits
nPixSplit	Number of split pixel hits
nSCTShared	Number of shared SCT hits
nPixHoles	Number of pixel holes
nSCTHoles	Number of SCT holes

GN1 architecture

Each **node** h_i in the graph corresponds to a **single track** in the jet, and is characterised by a feature vector (or representation)

Output: representations of each track that are conditional on the other tracks in the jet

Each of the objectives makes use of the global representation of the jet



(b-, c- and light-jet classes)

(track truth origin, see definition other slides)

(perform binary classification of the track-pair compatibility)

Deep Sets Architecture

Layers	3
Num nodes	64

Graph Neural Network, with multihead attention

Layers	3
Num nodes	128
Attention heads	2

Auxiliary Task Heads

- The combined jet-track vectors are fed into a per-track **initialisation network**
- A **fully connected graph** is built from the outputs of the track initialisation network, such that each node in the graph neighbours every other node
- **Three** separate fully connected feedforward neural networks are then used to independently perform the different classification objectives of GN1

Network	Hidden layers	Output size
Node classification network	128, 64, 32	7
Edge classification network	128, 64, 32	1
Graph classification network	128, 64, 32, 16	3

GN1 Training

- GN1 training procedure minimises the total loss function

Categorical cross entropy over track origin labels

Binary cross entropy over whether tracks share a vertex

$$L_{Total} = L_{Jet} + \alpha L_{Track} + \beta L_{Vertex}$$

Categorical cross entropy over jet flavour labels

- Inference Timings
  - Acceptable timings compared to total offline event reco O(10s), and HLT O(100ms)

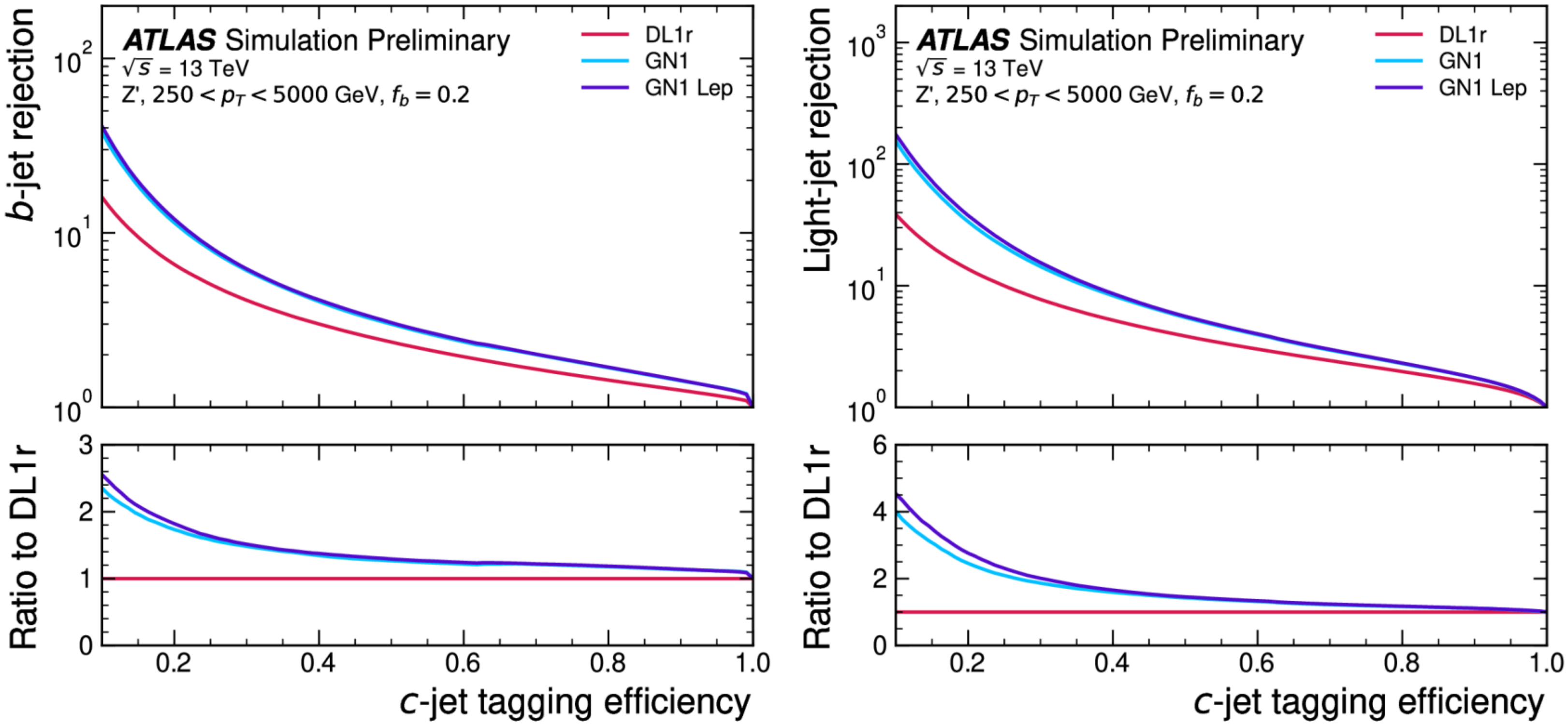
Tagger	Inference time per jet [ms]	
	ttbar	Z’
GN1	0.40	0.78

Input features to the GN1 model

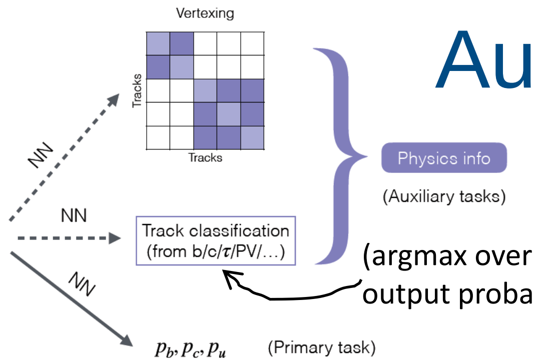
Jet Input	Description
p_{T}	Jet transverse momentum
η	Signed jet pseudorapidity

Track Input	Description
q/p	Track charge divided by momentum (measure of curvature)
$\mathrm{d}\eta$	Pseudorapidity of the track, relative to the jet η
$\mathrm{d}\phi$	Azimuthal angle of the track, relative to the jet ϕ
d_0	Closest distance from the track to the PV in the longitudinal plane
$z_0 \sin\theta$	Closest distance from the track to the PV in the transverse plane
$\sigma(q/p)$	Uncertainty on q/p
$\sigma(\theta)$	Uncertainty on track polar angle θ
$\sigma(\phi)$	Uncertainty on track azimuthal angle ϕ
$s(d_0)$	Lifetime signed transverse IP significance
$s(z_0)$	Lifetime signed longitudinal IP significance
nPixHits	Number of pixel hits
nSCTHits	Number of SCT hits
nIBLHits	Number of IBL hits
nBLHits	Number of B-layer hits
nIBLShared	Number of shared IBL hits
nIBLSplit	Number of split IBL hits
nPixShared	Number of shared pixel hits
nPixSplit	Number of split pixel hits
nSCTShared	Number of shared SCT hits
nPixHoles	Number of pixel holes
nSCTHoles	Number of SCT holes
leptonID	Indicates if track was used in the reconstruction of an electron or muon (only for GN1 Lep)

c-tagging

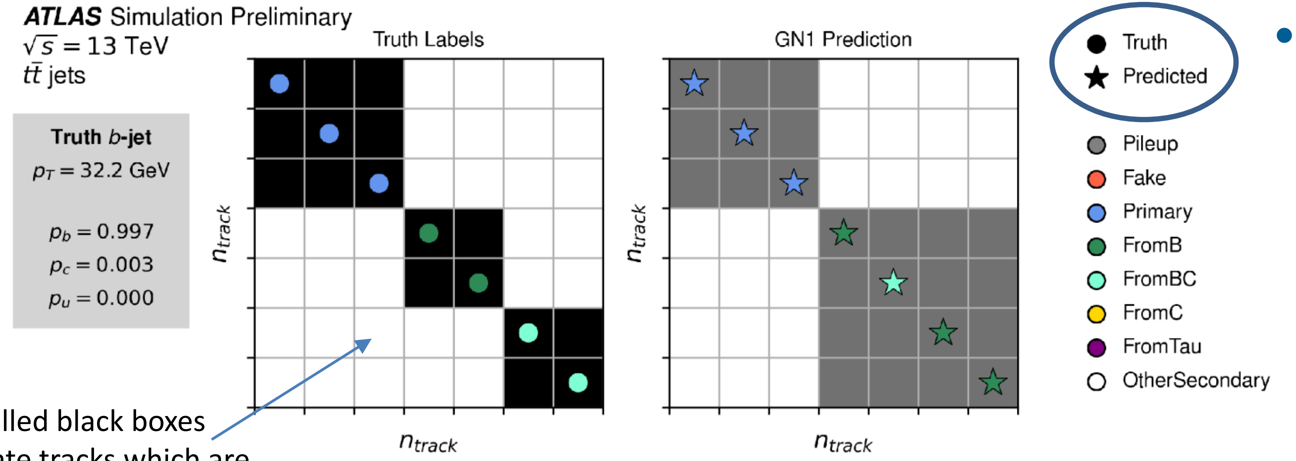

ATLAS Simulation Preliminary
$\sqrt{s}$ = 13 TeV
Z', 250 < p_T < 5000 GeV, f_b = 0.2
DL1r
GN1
GN1 Lep
b-jet rejection
Light-jet rejection
Ratio to DL1r
c-jet tagging efficiency

Auxiliary tasks: track origins + vertexing



(argmax over the output probabilities)

Truth Origin	Description
Pileup	From a pp collision other than the primary interaction
Fake	Created from the hits of multiple particles
Primary	Does not originate from any secondary decay
fromB	From the decay of a b-hadron
fromBC	From a c-hadron decay, which itself is from the decay of a b-hadron
fromC	From the decay of a c-hadron
OtherSecondary	From other secondary interactions and decays



The filled black boxes indicate tracks which are grouped together into vertices (score >0.5)

- GN1 prediction (example)
  - Correctly predicts the track origins and vertex compatibility for the tracks of the Primary Vertex
  - Fails to resolve B and B→C decay vertices
  - correctly predict the heavy flavour track origins

Auxiliary Tasks

- Removal of components in the model to study their impact:
  - The resulting trainings demonstrate how useful the different auxiliary training objectives are for the primary jet classification objective