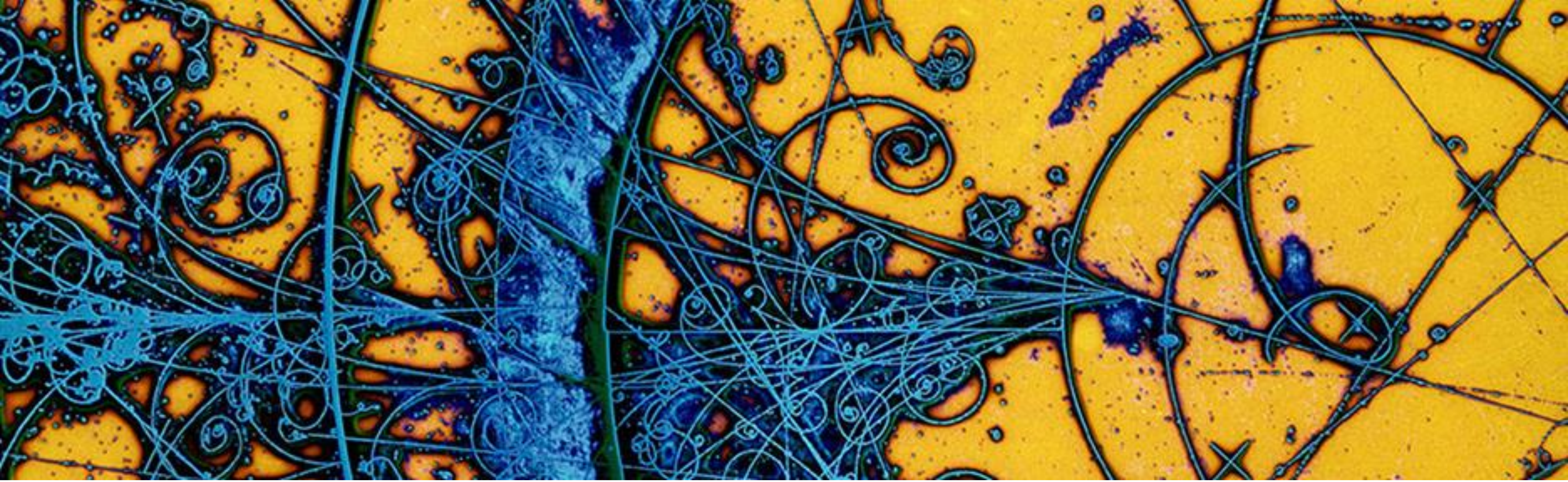


Deep Learning for Flavor Tagging at ATLAS experiment

Andrea Di Luca

ALPACA – 20/24 November 2023

ECT* - Trento

INFN Istituto Nazionale di Fisica Nucleare

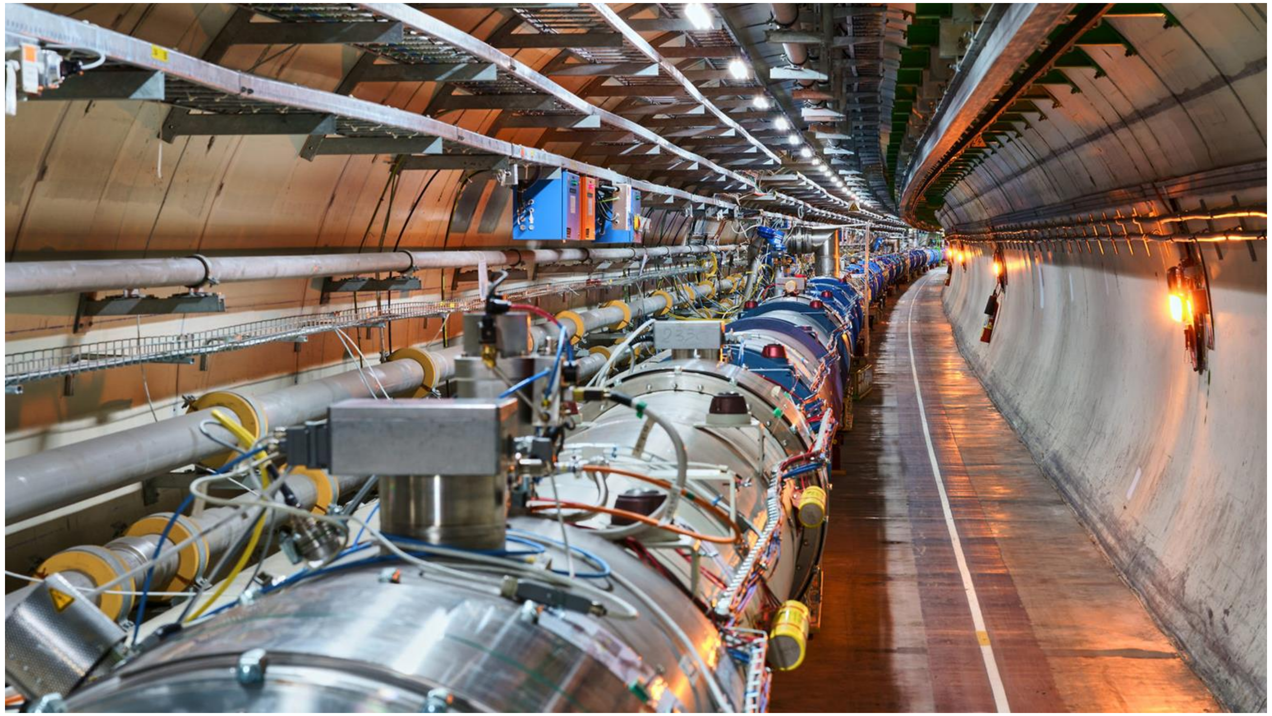
deeppp

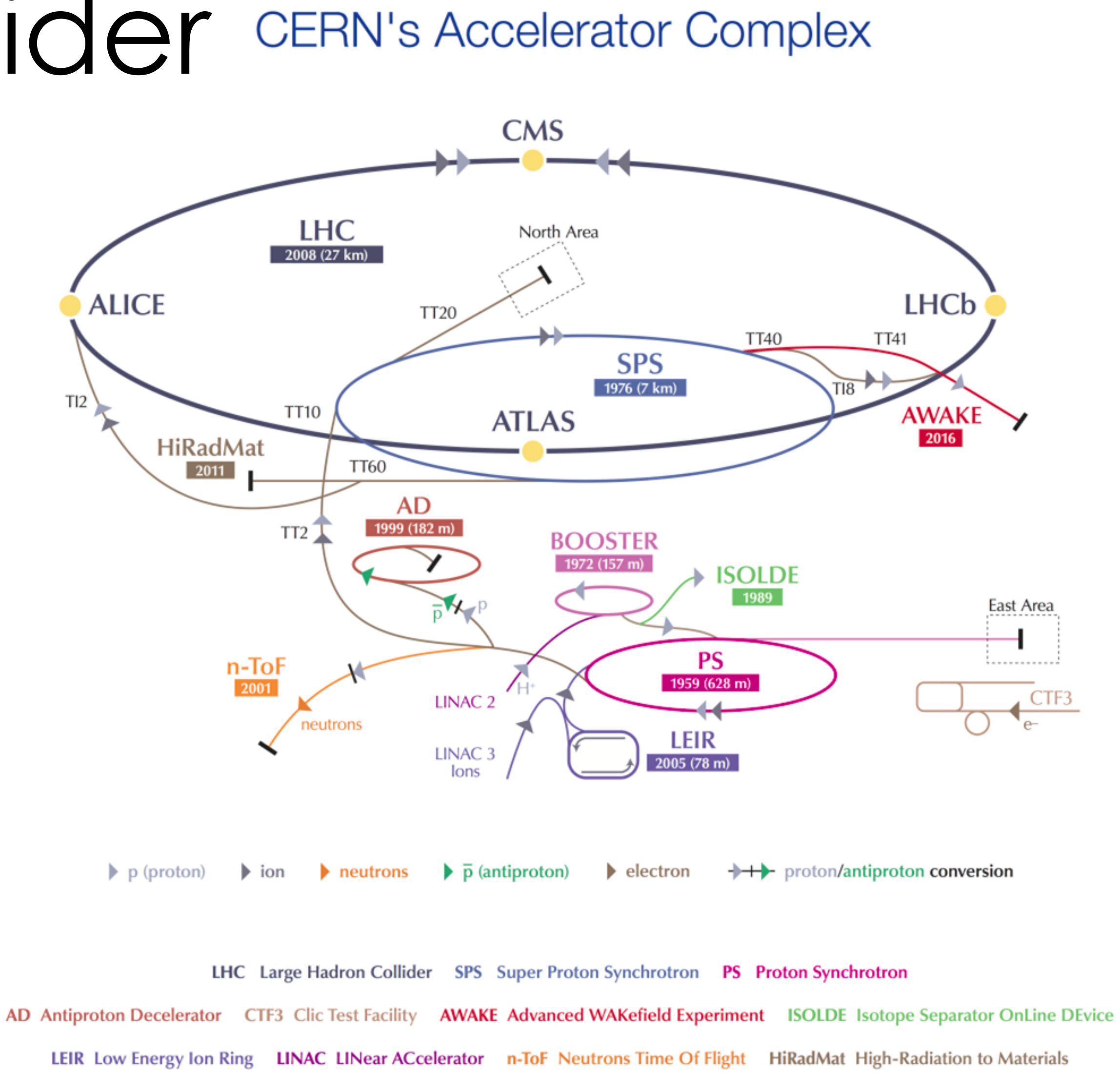
ATLAS EXPERIMENT

DSIP DATA SCIENCE FOR INDUSTRY & PHYSICS

FONDAZIONE BRUNO KESSLER

The Large Hadron Collider





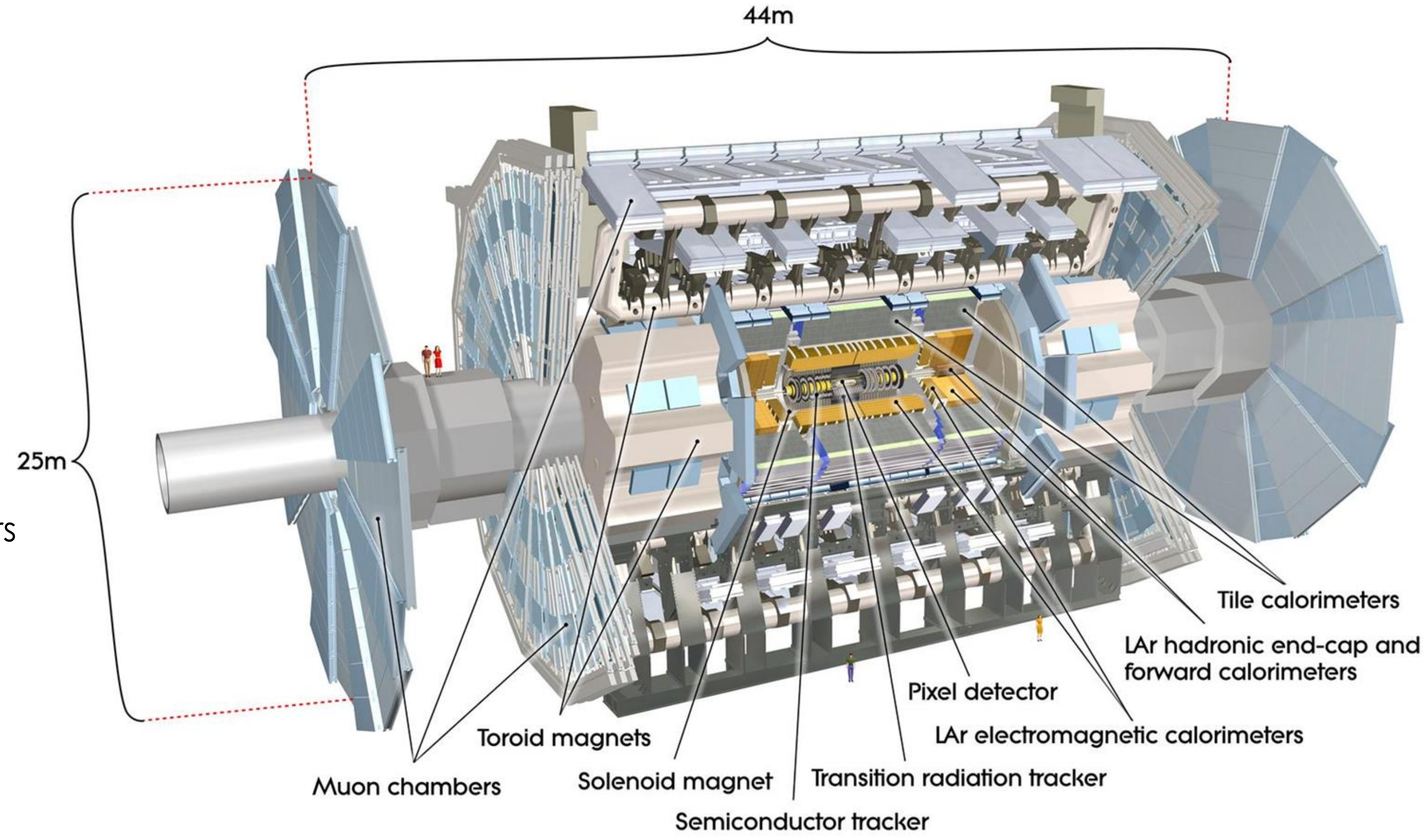
The Large Hadron Collider (LHC) is the world’s largest and most powerful particle accelerator.

The ATLAS experiment

ATLAS is a general-purpose particle physics experiment.

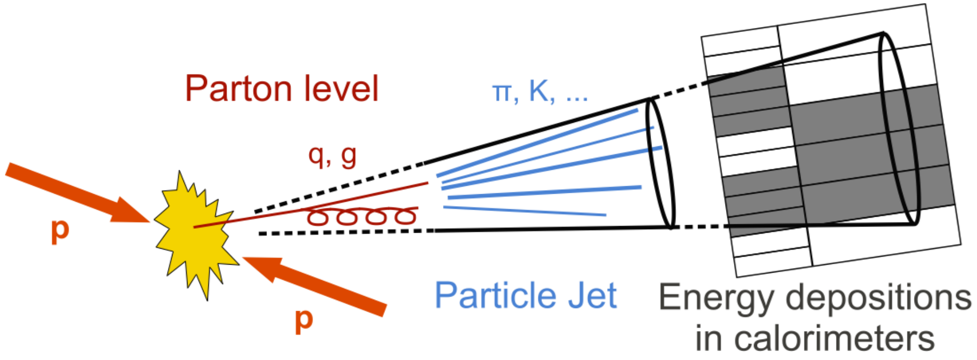
The ATLAS detector is composed by three main subsystems:

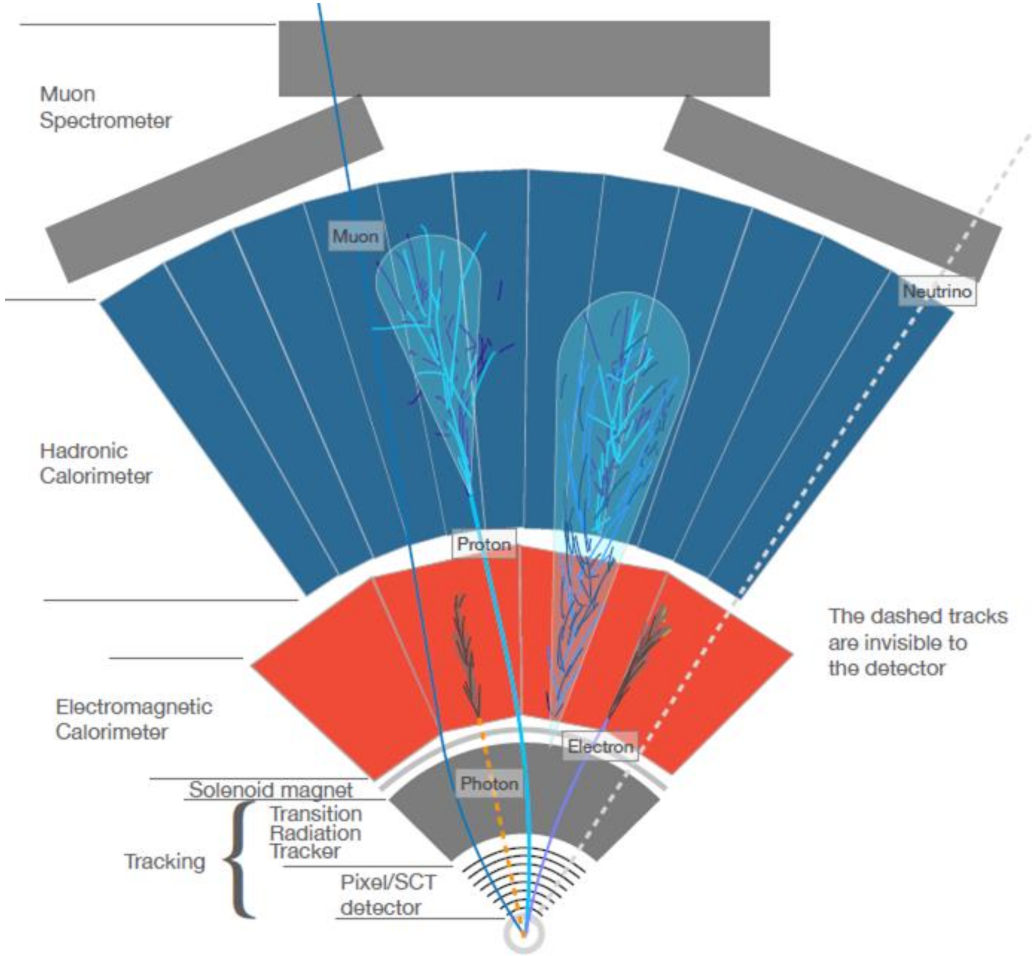
- Tracker
- Electromagnetic (EM) and hadronic (HAD) calorimeters
- Muon spectrometer
- Magnet system (Central solenoid and toroid)



Physics objects

- Detector provides position and energy information
  - Physics objects (electrons, muons, ...) need to be reconstructed
- Neutrinos escape the detector unseen
  - Missing transverse momentum E_T^{miss}
- **Particle jets** reconstructed using **energy depositions** in the calorimeters
  - Collimated spray of stable particles arising from fragmentation and hadronization of a parton after a collision.





From collision to data

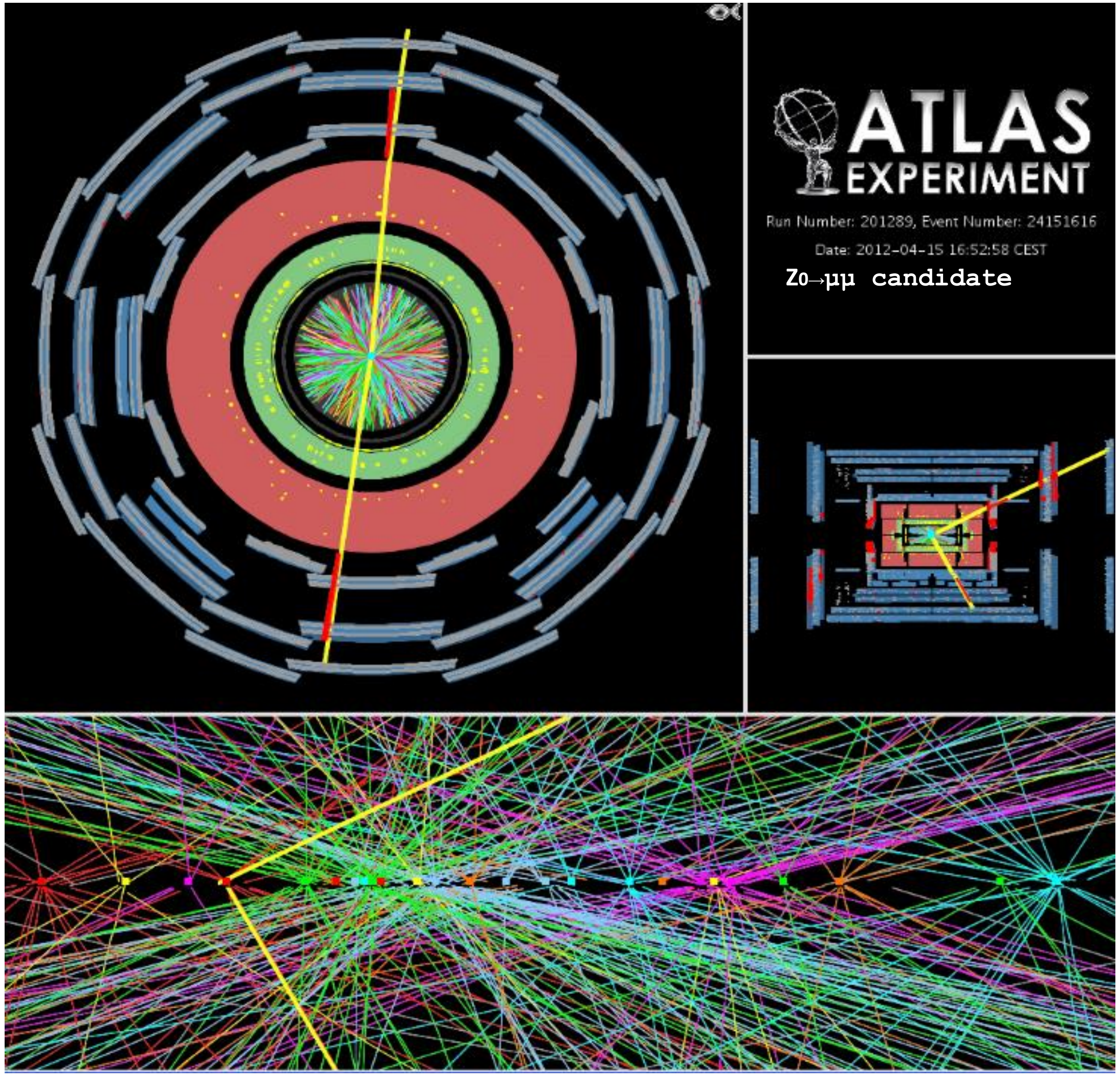
Event rate

- **40 million** bunch crossings **per second**
- About 33 collisions per bunch crossing
- About **1 billion collisions per second**

Just a **little fraction** of these events **is interesting**

A **trigger chain reduce** the number of events downto **200 “interesting” events per second**.

Still some selection techniques are required to select only interesting physics process inside the event.

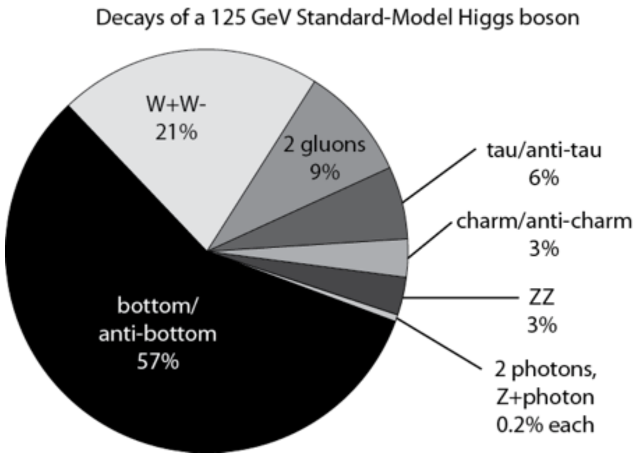


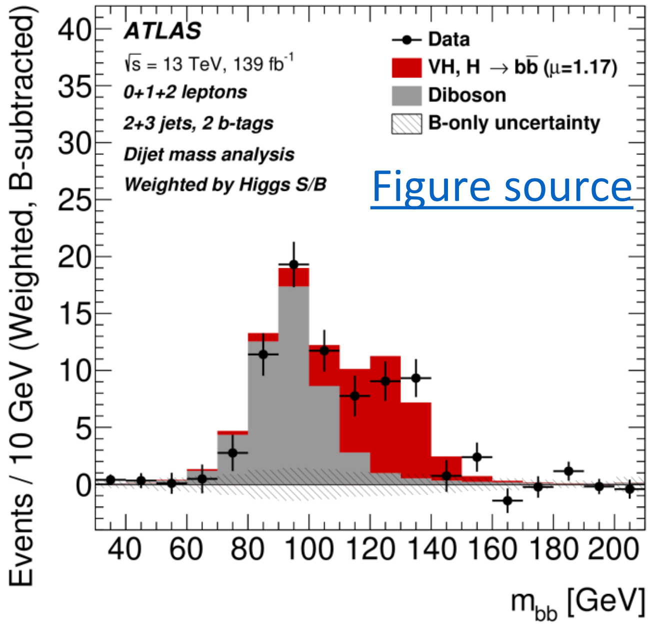
Flavor tagging

Flavor tagging aims to identify the Flavor of a particle jet (b, c, light) and it is an **essential tool** to study physics processes with **b/c-jets in their final state**:

- Processes with heavy flavour quarks (b,c) play a key role in the LHC physics program (ex. H→bb)

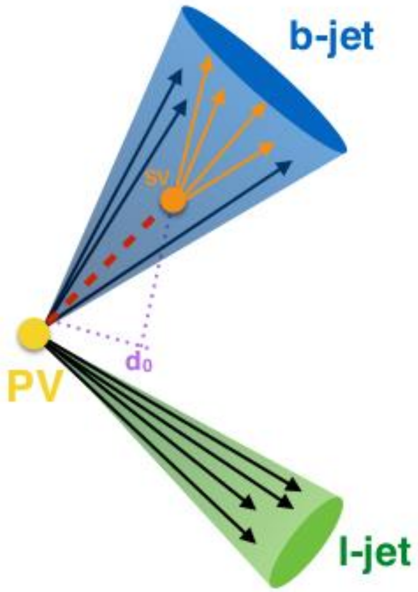
- We can also use flavour tagging to suppress otherwise overwhelming backgrounds, e.g. V+jets



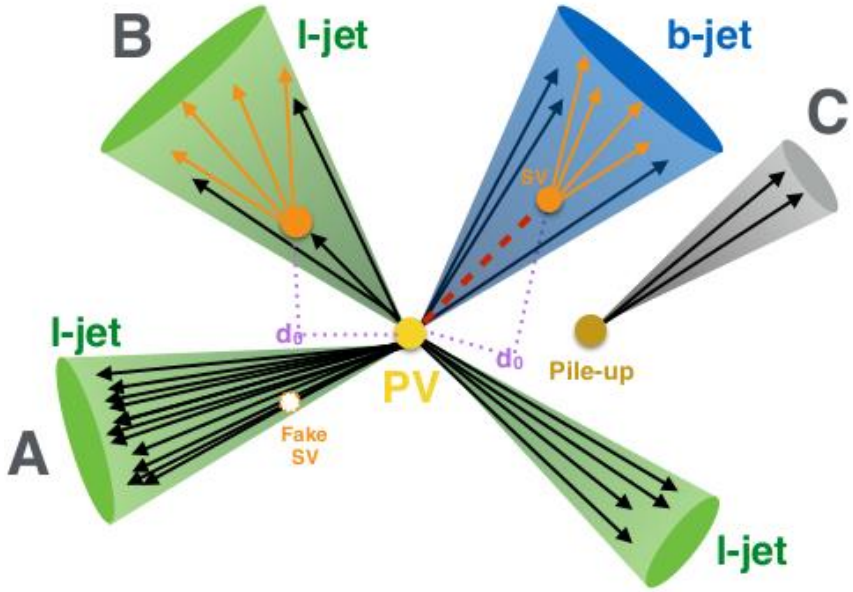


B-Hadron Properties

[Figure source](#)



Increasing transverse momentum p_T



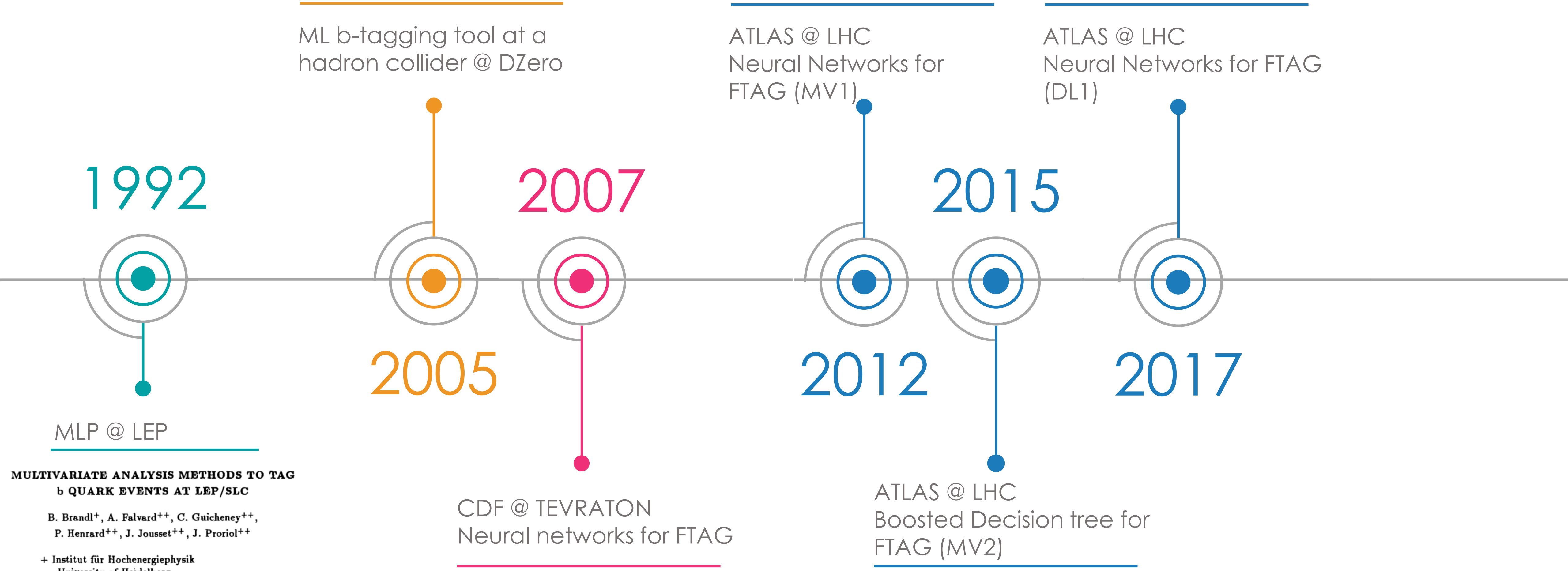
Exploit specific topology of heavy-flavor jets for identification

- Relatively long lifetimes
- High mass: ~5 GeV
- Decay product multiplicity: on average decay to ~ 5 charged particles
- Decay to c-hadron
- Fairly large leptonic decay fraction

At higher transverse momentum the picture gets more complicated

A. Increased fragmentation track multiplicity causing more fake SVs.
B. Increased material interaction increasing number of real SVs not stemming from heavy flavour jets
C. Growing pile-up conditions.

Machine Learning for FTAG



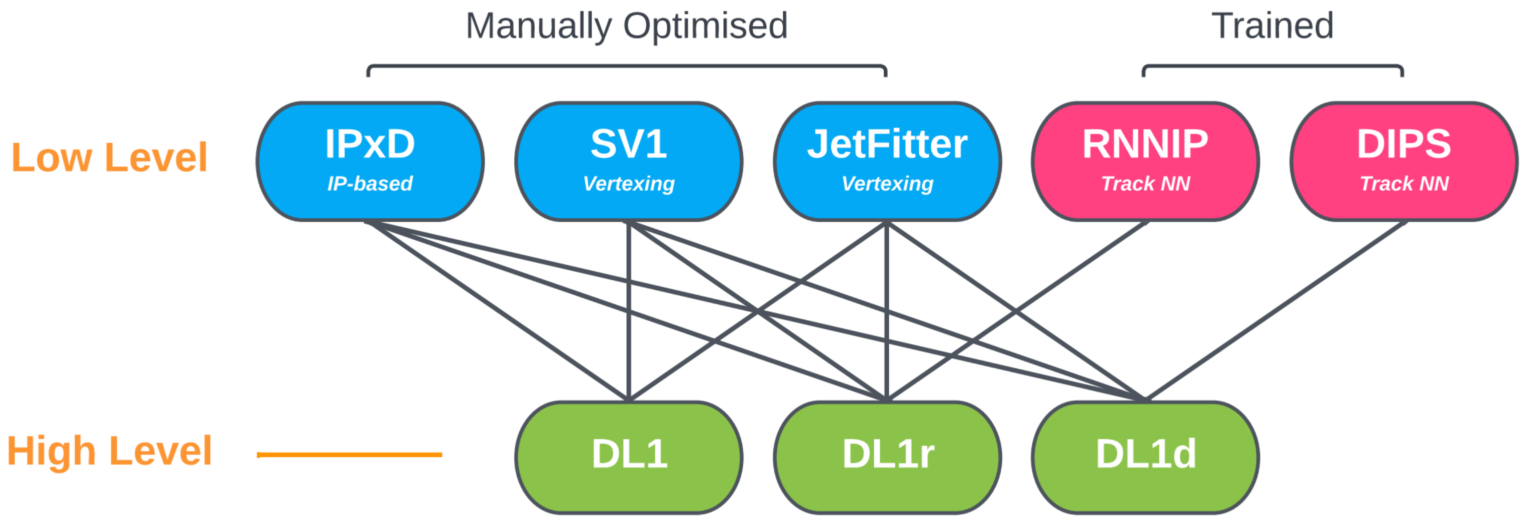
MULTIVARIATE ANALYSIS METHODS TO TAG b QUARK EVENTS AT LEP/SLC

B. Brandl+, A. Falvard++, C. Guicheney++, P. Henrard++, J. Jousset++, J. Proriol++

+ Institut für Hochenergiephysik
University of Heidelberg
D-6900 HEIDELBERG GERMANY

++ Laboratoire de Physique Corpusculaire
de Clermont-Ferrand

Flavour Tagging Strategies in ATLAS



CERN-EP-2019-132

Jet and track inputs are fed to **low level taggers**:

- Use physics knowledge to construct expert variables: IPxD, SV1, JetFitter
- Track-based ML models: RNNIP, DIPS

High-level taggers (MV2 e DL1) combine all this information and they return probabilities for each flavor class: p_b, p_c, p_l

Constructing the Discriminant

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

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

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

f_c and f_b are arbitrary parameters which trade-off between background rejections (e.g. larger f_c more *c*-jet rejection)

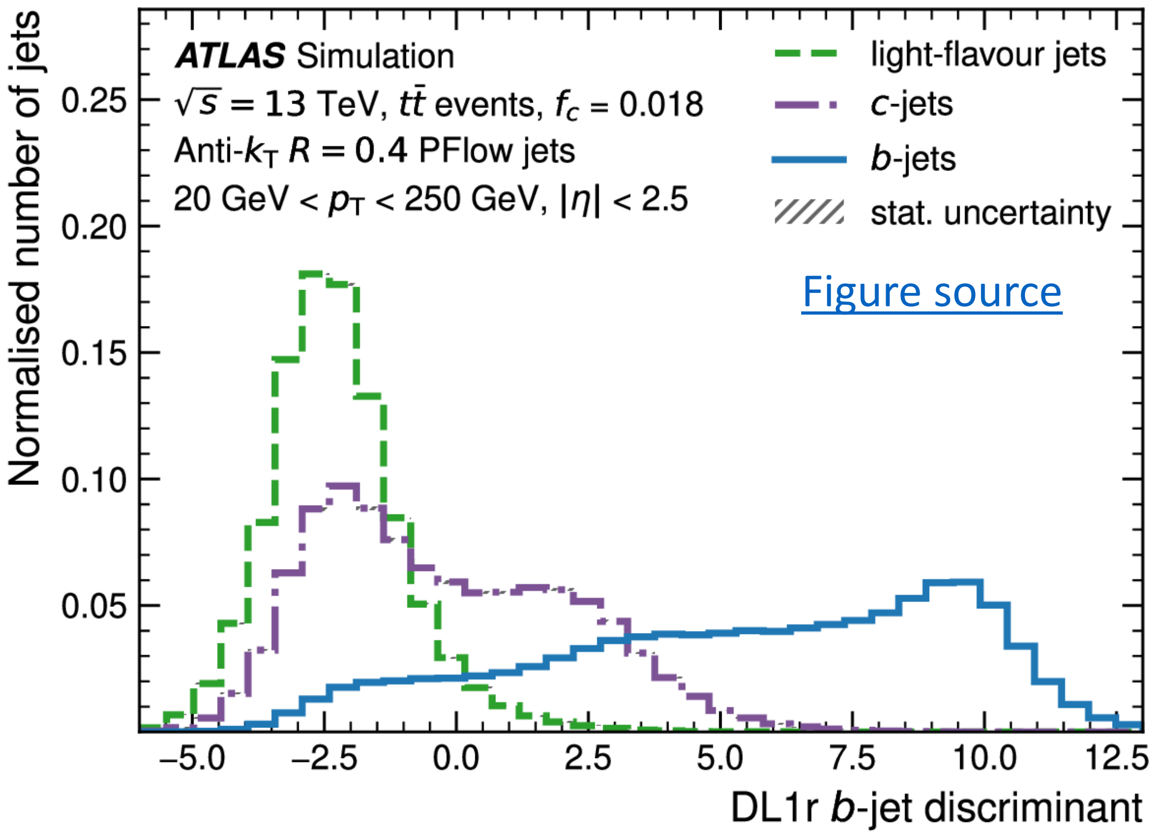
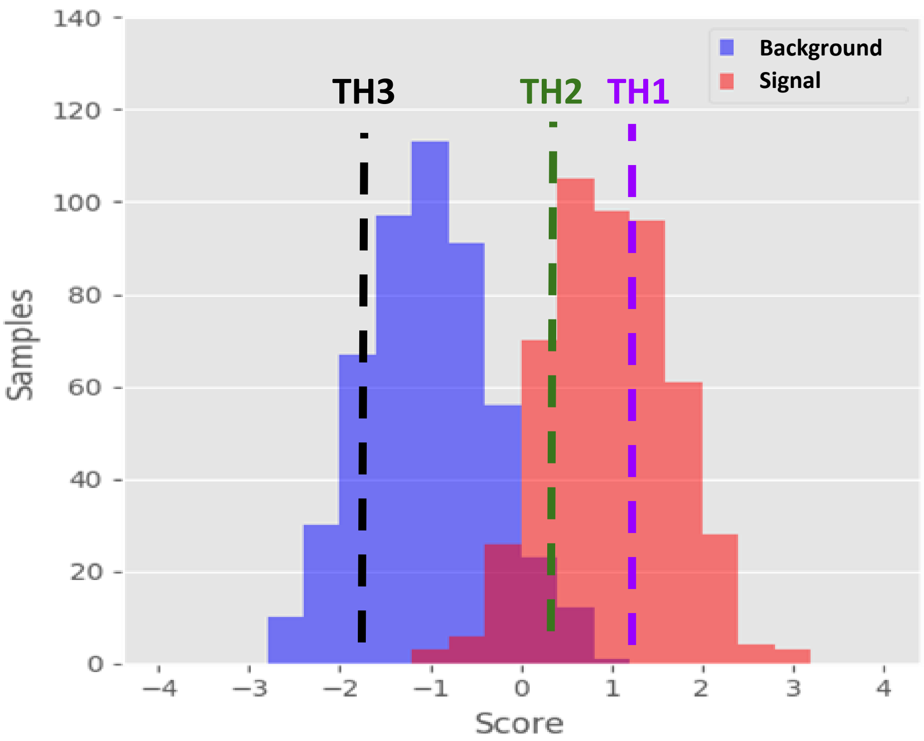


Figure source

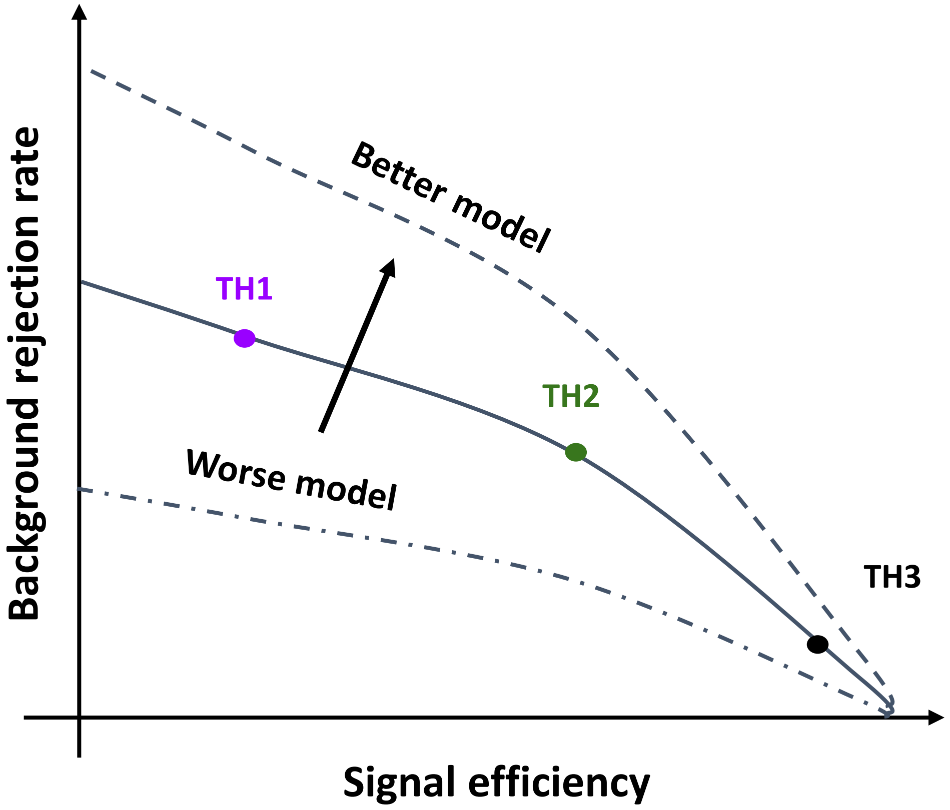
How to evaluate classifier performance



$$\text{Signal efficiency} = \frac{Signal > TH}{Signal}$$

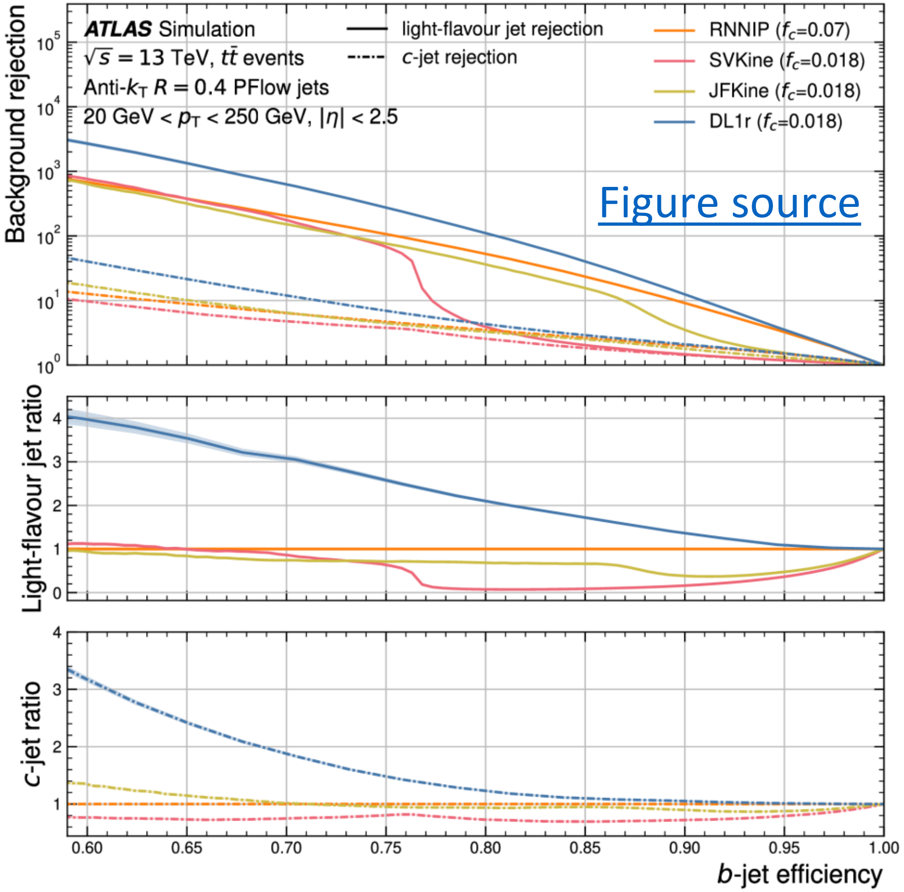
$$\text{Background rejection rate} = \frac{Background}{Background > TH}$$

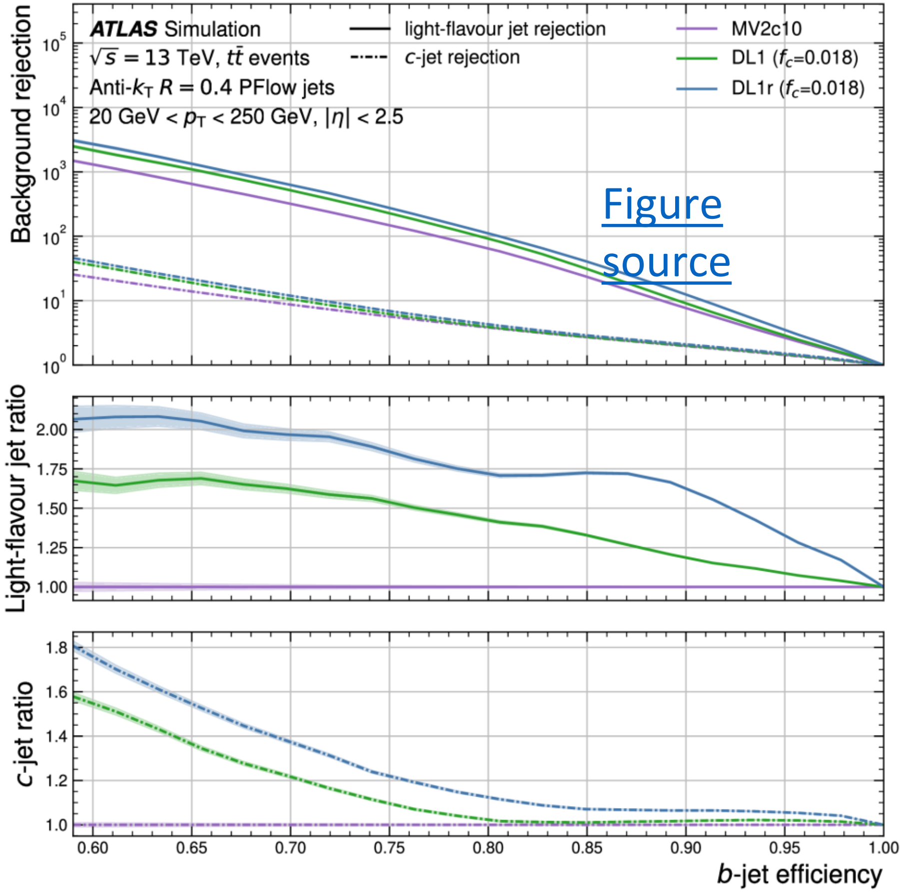
Receiver Operating Characteristic (ROC) curves can be used to **compare performance of different models**.



Performance

- ATLAS Run 2 algorithm performance documented in recent publication: [2211.16345]
- Widely used Run 2 tagger: DL1r (DL1 + RNNIP)



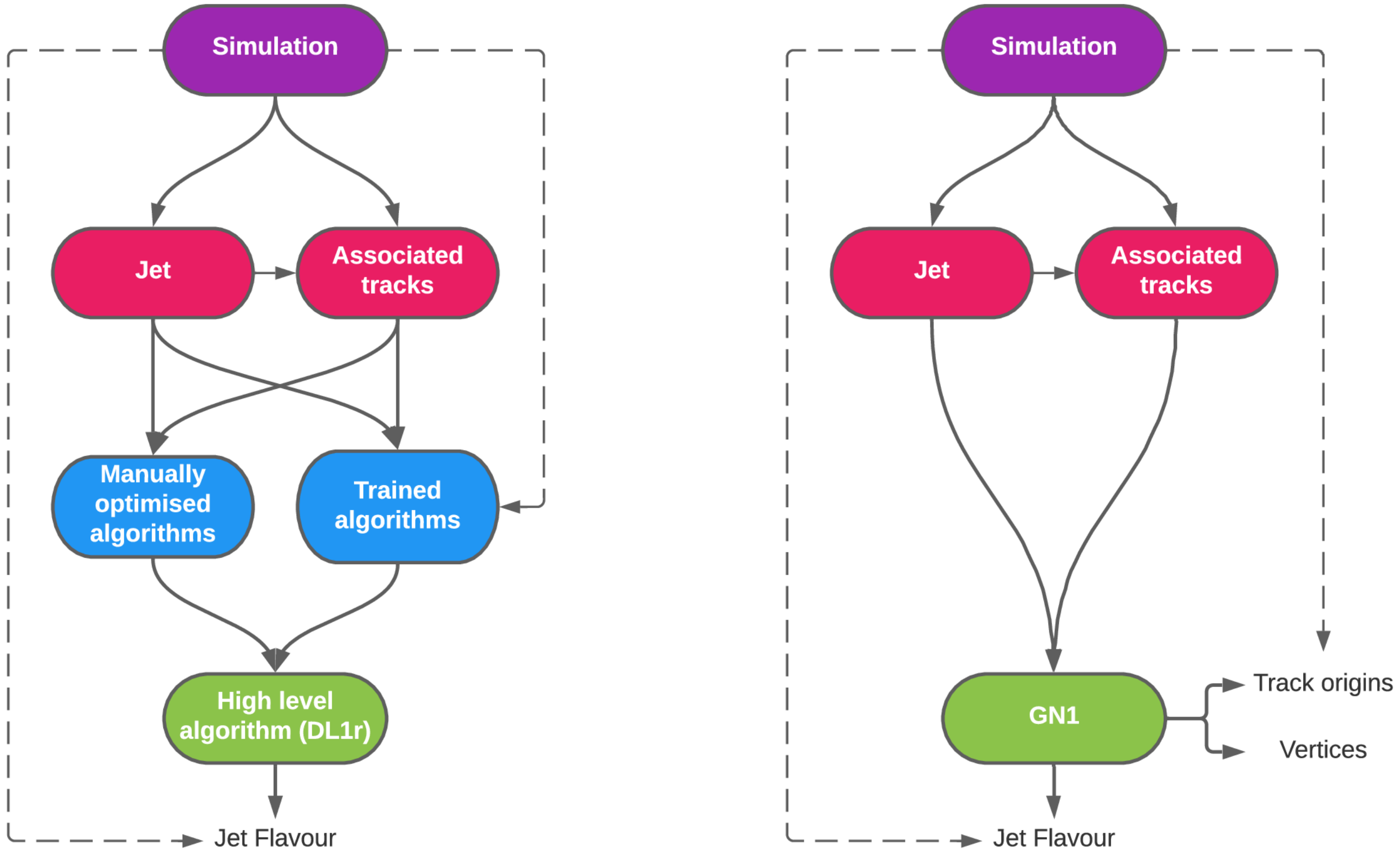


A New Approach

Figure source

New all-in-one tagger: GN1

- Jet flavor, vertexing and track origin tasks trained simultaneously
- No need for low level algorithms
- Naturally suited for a variable number of unordered input tracks
- Based on graph neural networks

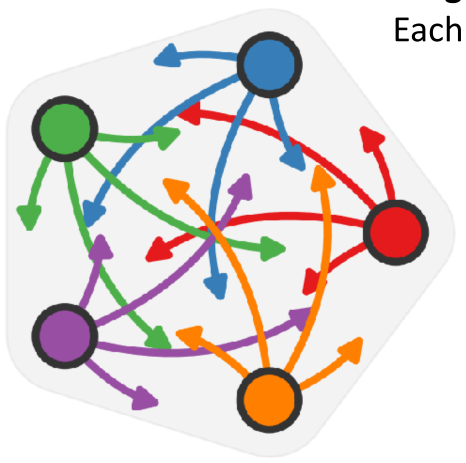


Training samples and inputs

Training samples

- Simulated *pp* collisions with *b*-, *c*- and *l*-jets in final state
- Resampling of jet kinematics (p_T and η) for each flavor
- Normalization and shuffling applied
- 30M training jets, further 500k each validation and test jets

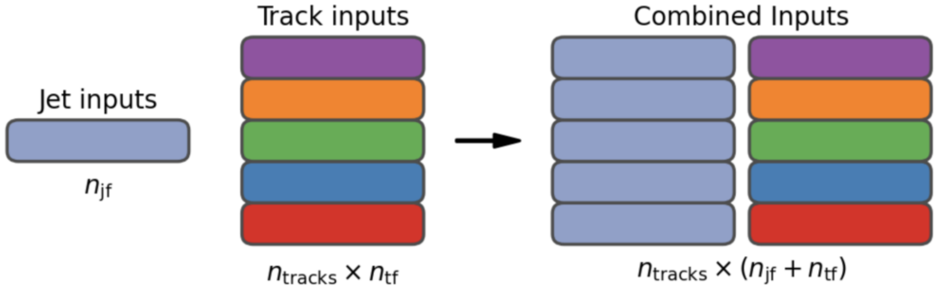
Jet graph
Each node of the graph is a track



Inputs

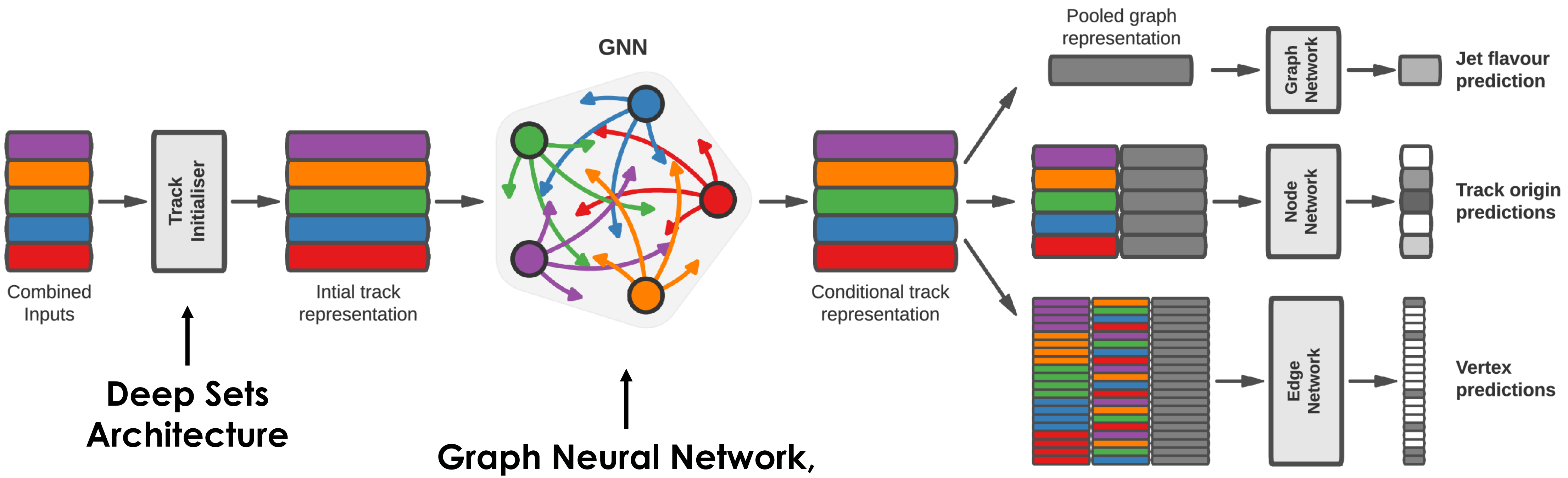
- Jet p_T and η
- Track parameters, uncertainties, and impact parameters
- Detailed hit information
- Jet variables are concatenated with each track.

Figure source



GN1 Architecture

Figure source



Deep Sets Architecture

Graph Neural Network, with multi-head attention

Auxiliary tasks

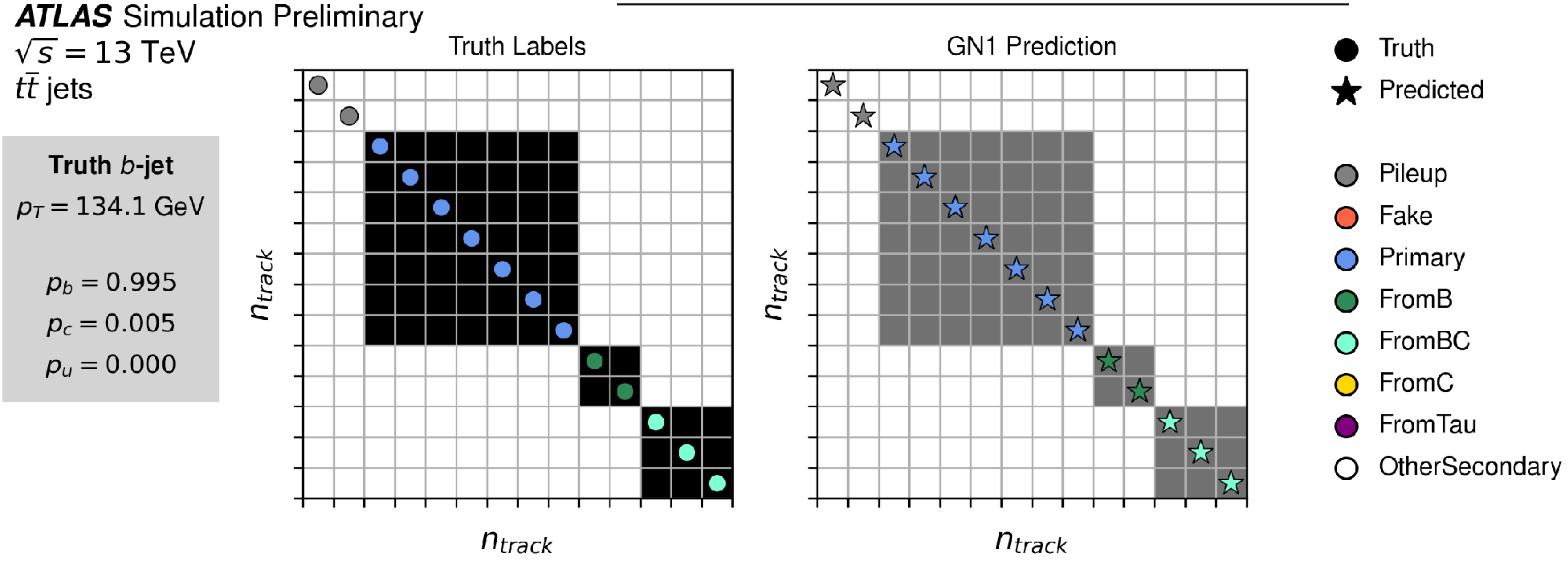
Figure source

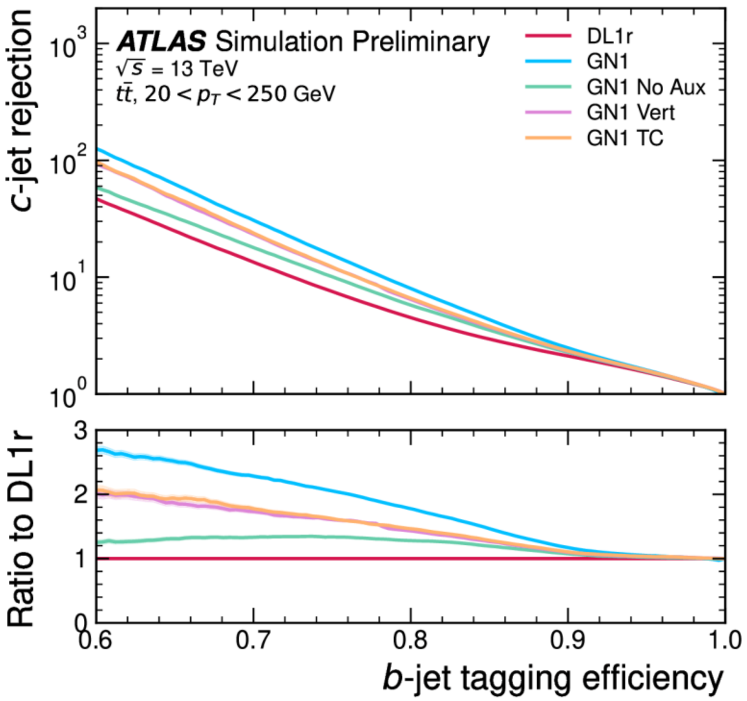
Adding physics information during GN1 training with 2 auxiliary tasks to improve classification performance.

1. **Vertexing**
   Prediction of track-pair vertex compatibility for each pair of tracks in the jet.

2. **Track classification**
   Classification of track origin.

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





Performance

Significant performance improvement observed with respect to DL1r.

Figure source

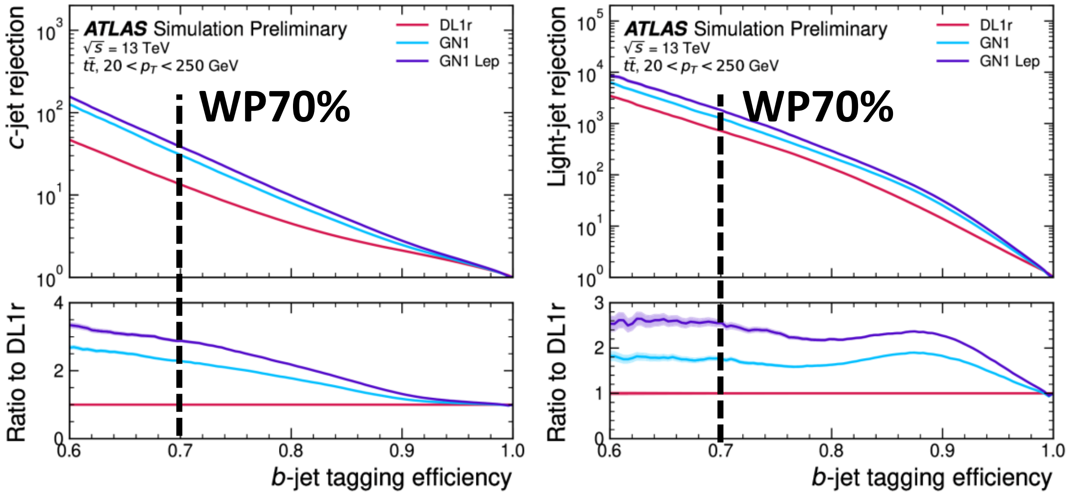
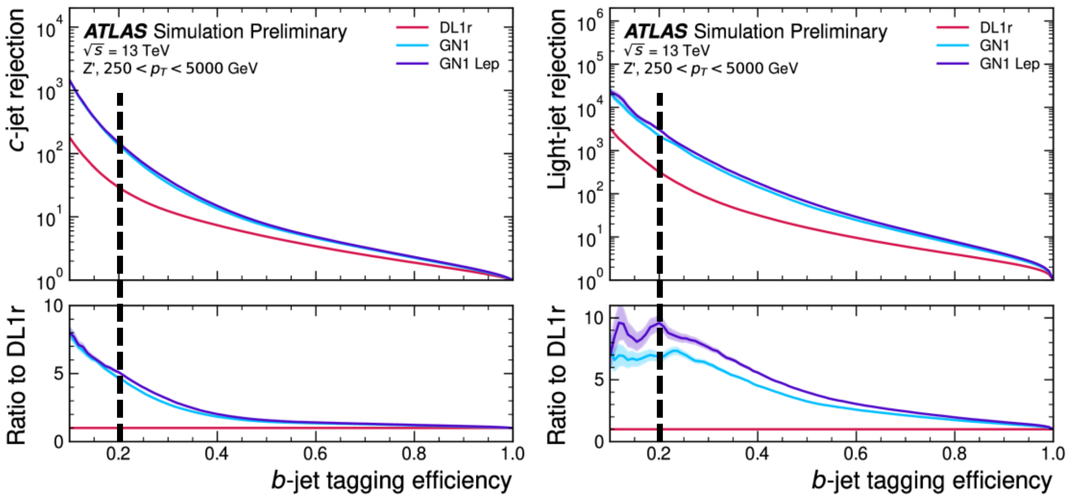


Figure source



At the **70% working point (WP)** for GN1:

- 2.25x increase in c-jet rejection
- 1.8x increase in light-jet rejection

The 70% WP corresponds to a **high-p_T** Z' *b*-efficiency of ~20%!

- 5x increase in c-jet rejection
- 7x increase in light-jet rejection

FTAG-2023-01

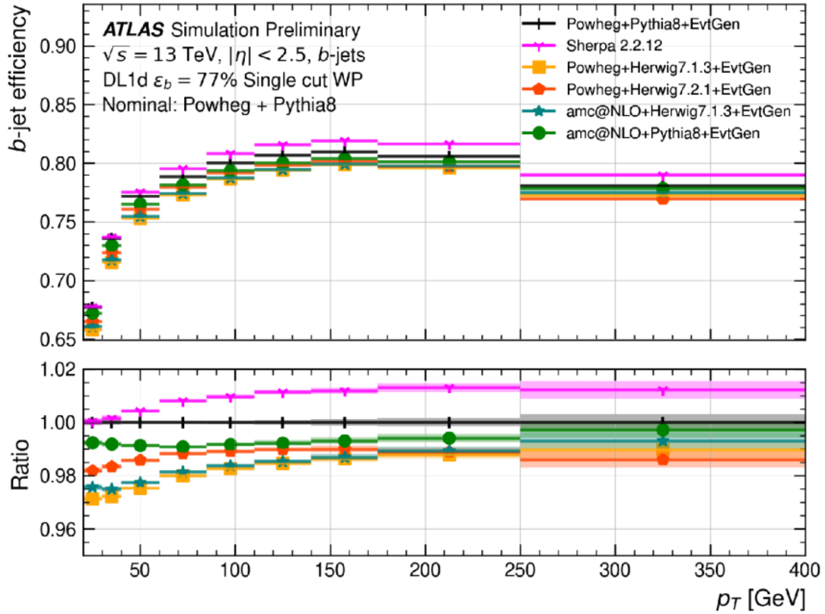
Generator dependence

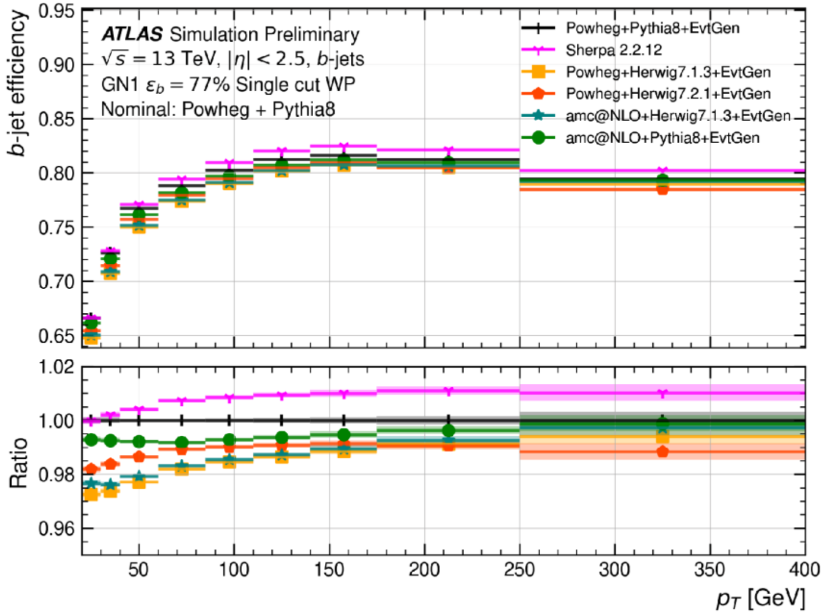
- Testing the model on other MC samples allows for an understanding of the generator dependence.

- Essential to verify that the a more sophisticated model such as GN1 is not learning generator dependent information

Overall generator dependence: O(3%) for b-jets and O(6%) for c-jets

- Indicates that the more sophisticated model is not exploiting generator specific information



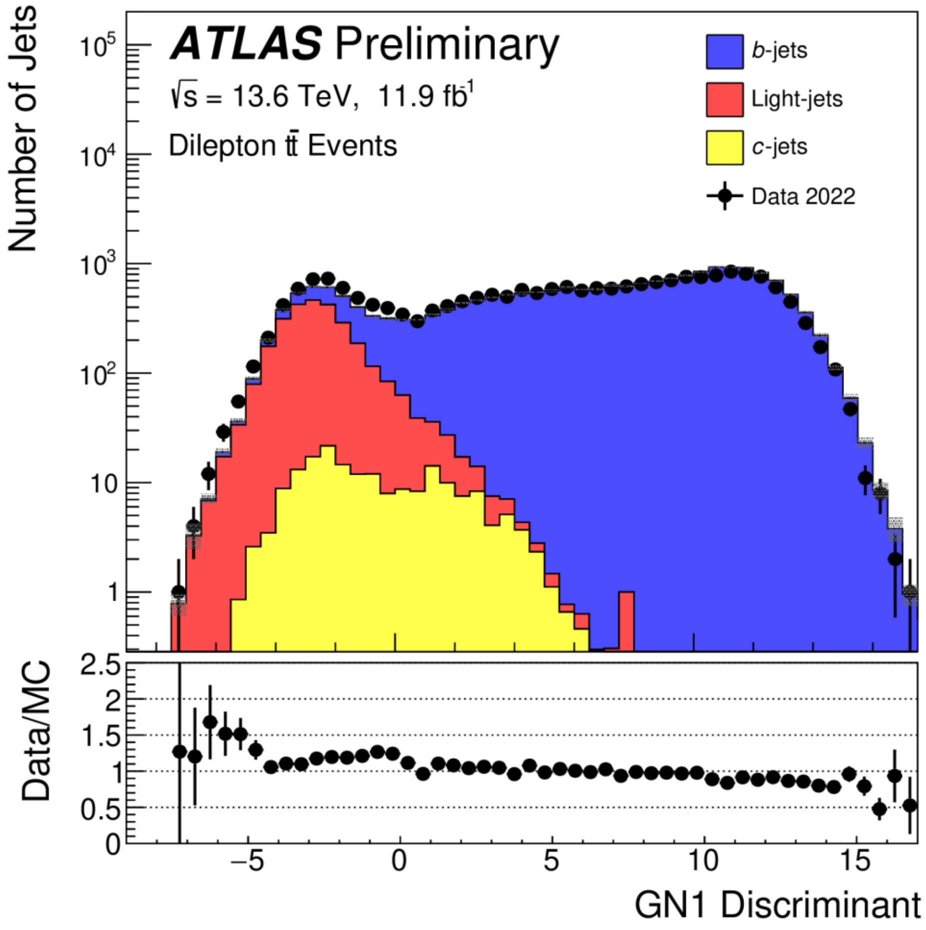


data/MC Agreement

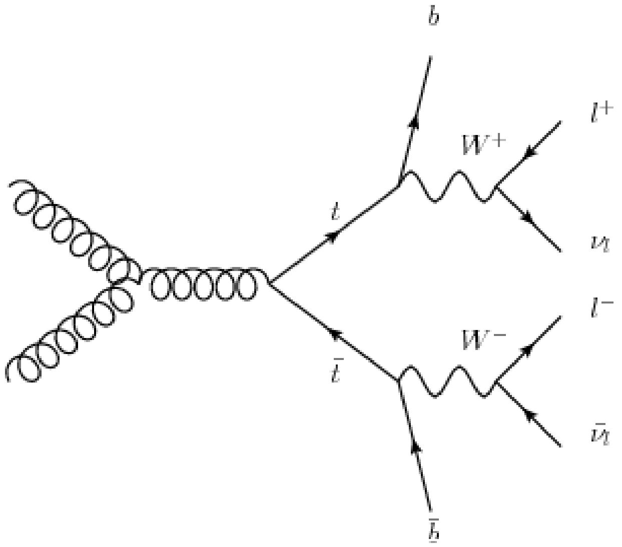
FTAG-2023-01

*Plotting tagger discriminant for leading jet pT

- Need to check performance on data
- Derive efficiencies for the different flavors on data and correct MC via scale factors
- Using a variety of different, easy to select, processes to calibrate the taggers, as dilepton $t\bar{t}$ events.



Dilepton event selection:
- Exactly two leptons and two jets
- Opposite sign muon and electron
- Invariant mass of each jet-lepton pair below 175 GeV

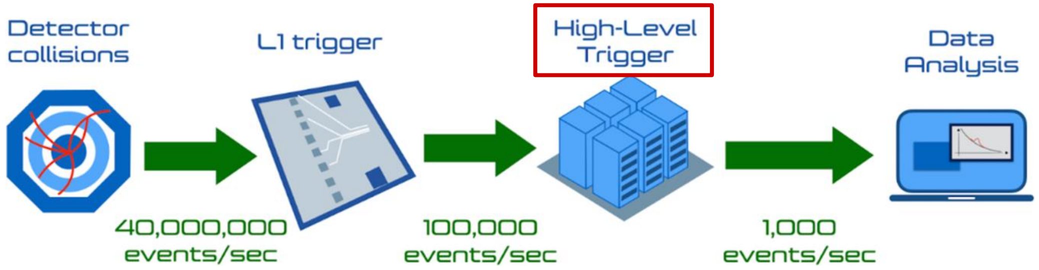


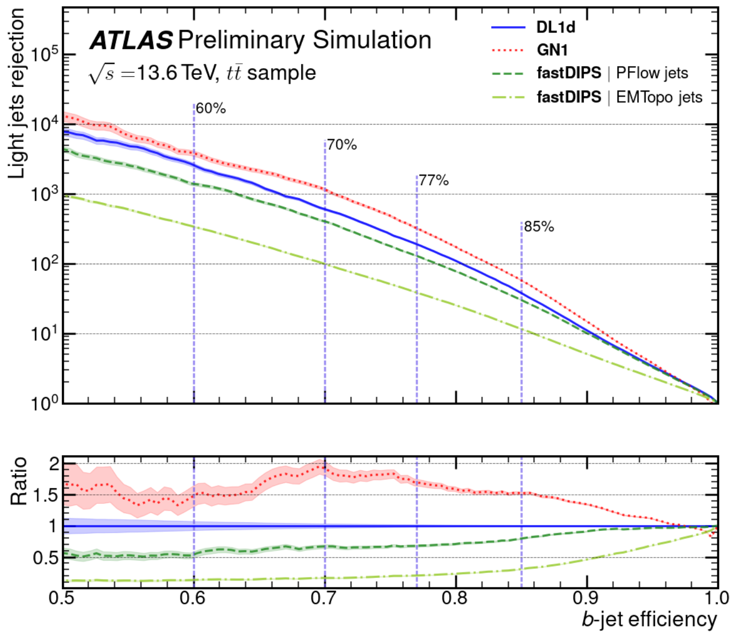
GN1 @ HLT

GN1 has also been deployed in the ATLAS High Level Trigger (HLT)

- Inputs are precision tracks and jet quantities after primary vertexing
- Strong performance compared with DL1d & other taggers running at trigger level

Processing time: o(100ms)

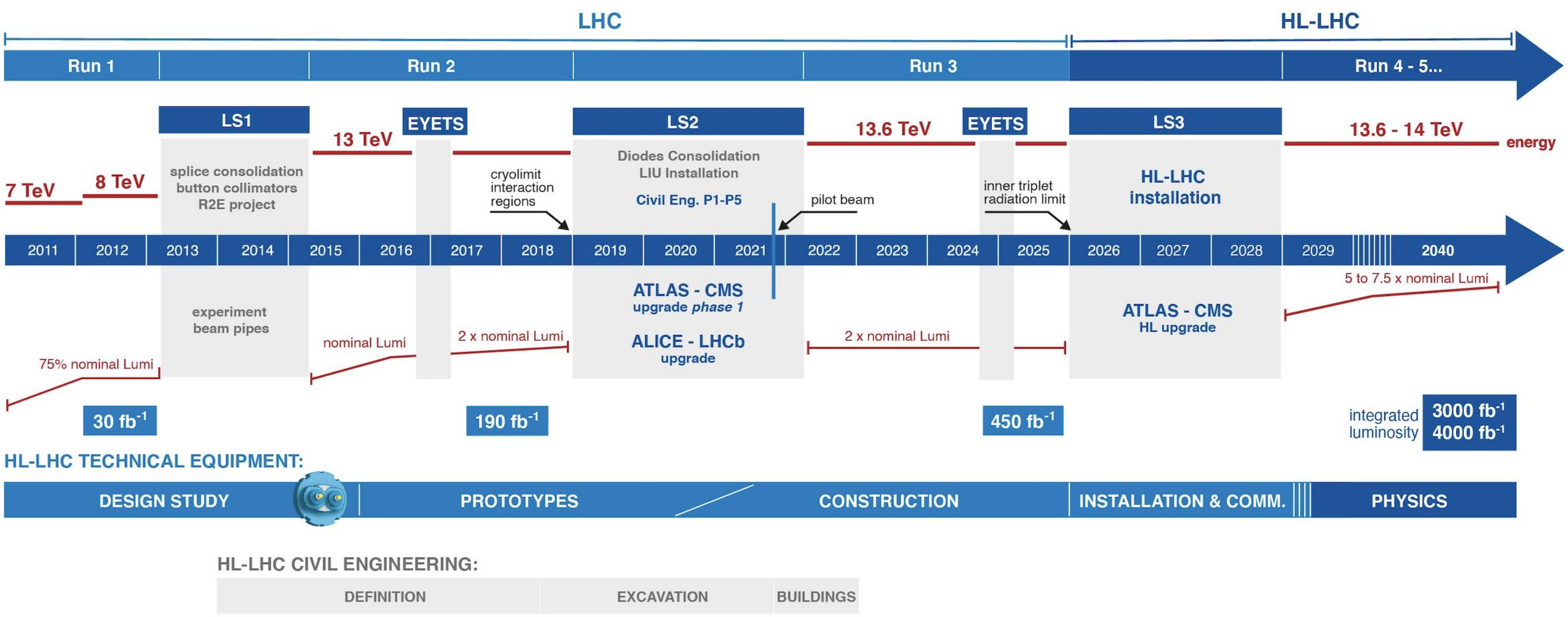




Tagger	Inference time per jet [ms] *	
	ttbar	Z'
DL1d **	0.07	0.08
GN1	0.40	0.78

*it can depend on the machine
**low level computation not included

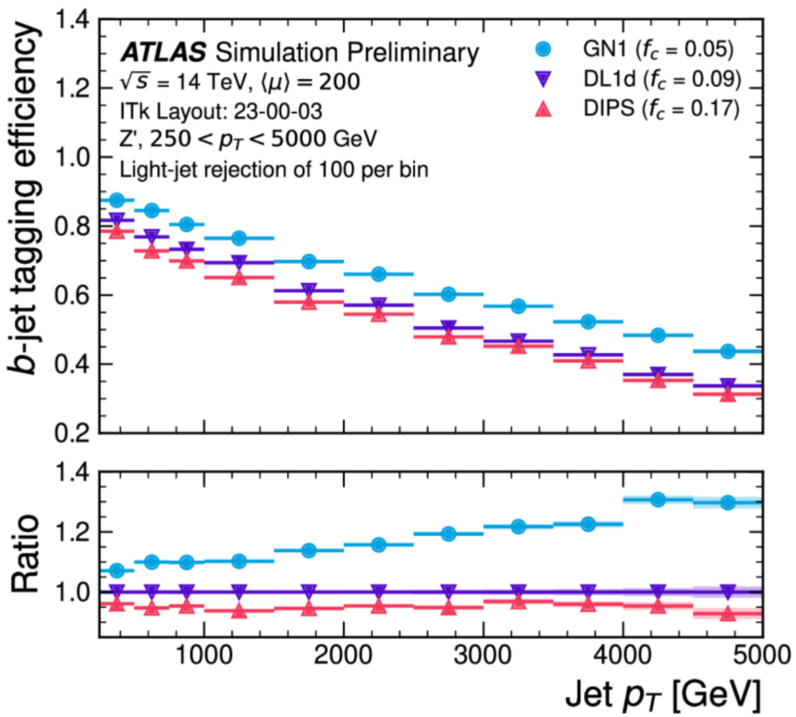
LHC / HL-LHC Plan

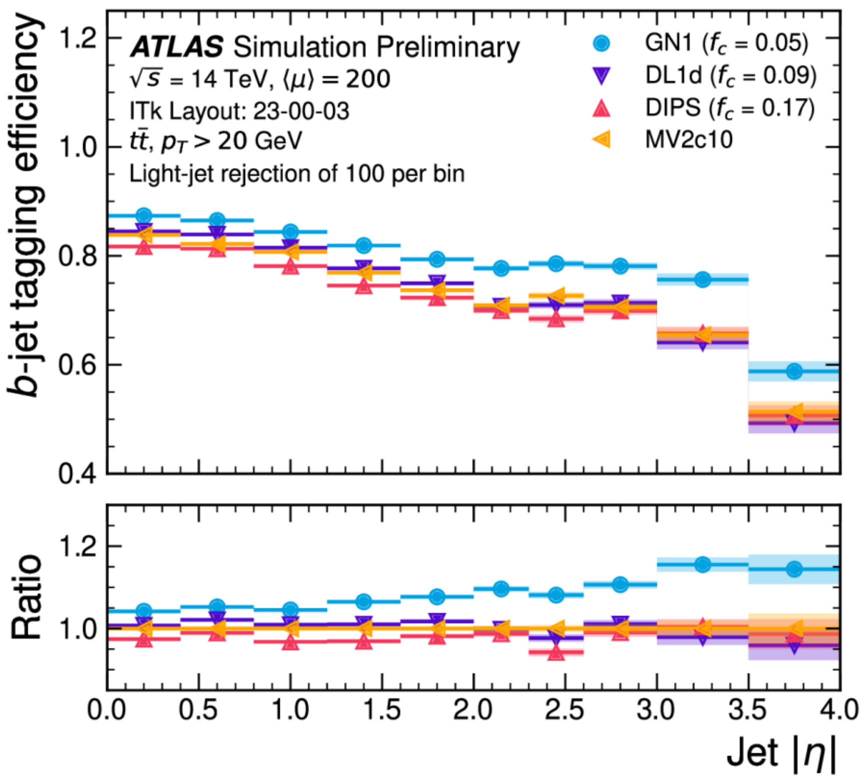


How will we be triggering events since 2029 when an average **200 pp collisions** per bunch crossing are expected?

GN1 @ HL-LHC

FTAG-2023-01



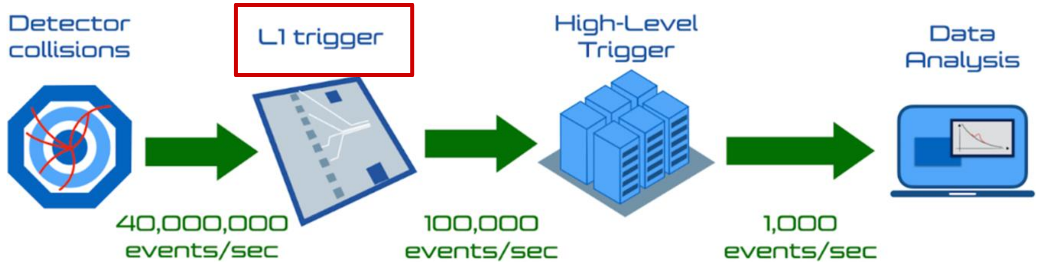


GN1 performance are better in the most interesting phase spaces:

- Up to 30% improvement in b-efficiency at high-p_T
- 15% improvement in the newly accessible forward region ($|\eta| > 2.5$) (significant upgrade of the ITk detector)

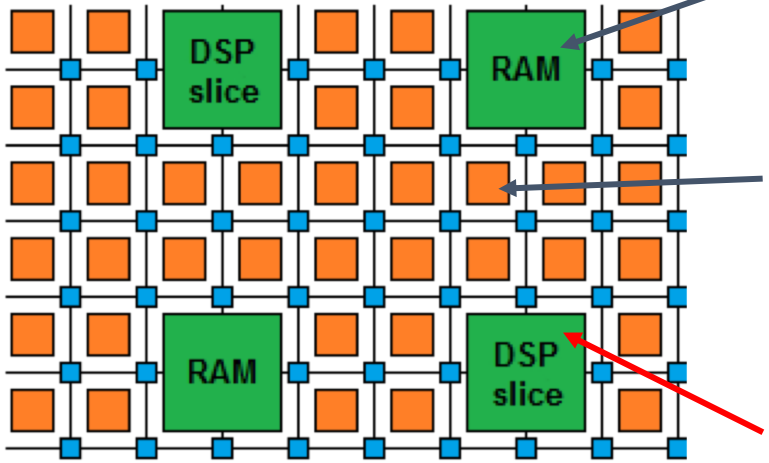
Pushing to the extreme

To **improve selection performance** there is a great interest in **running Deep Neural Networks in real-time.** This represents a great technical challenge due to the **extreme data rate** (O(100 TB/s) at L1) **to be processed with some very strict time constraints**.



Source: *hls4ml tutorial*

- **FPGA** (Field-programmable gate array) are programmable integrated circuits. They can offer **low latency** and **high throughput**.
- A **model should fit the FPGA chip-size** and **latency requirements**. Depending on the FPGA size, we **should know how to reduce the size of a model**.

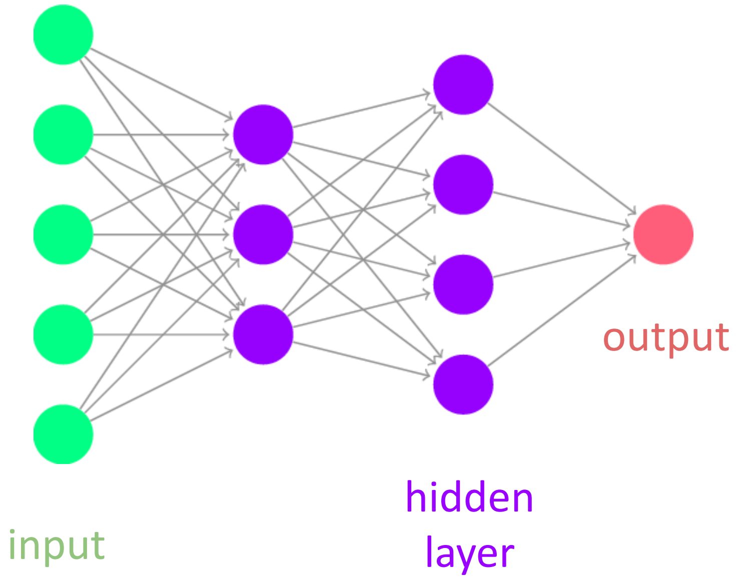


RAMs are small memories.

Logic cells for any function and simple arithmetic.

DSPs (Digital Signal Processor) are designed to perform multiplications.

Neural network inference



l-layer output — Activation function — Weights — Input — biases

$$x_l = g_l(W_{l,l-1} \cdot x_{l-1} + b_l)$$

Precomputed Memory — Multiplication DSPs — Addition Logic cells

$$N_{mul} = \sum n_{\text{nodes}}^{l-1} \times n_{\text{nodes}}^{l}$$

A **model should fit the FPGA chip-size** and **latency requirements**. Depending on the FPGA size, we **should know how to reduce the size of a model**.

Very active research field

- Coelho, C.N., Kuusela, A., Li, S. et al. *Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors.* **Nature Machine Intelligence**
- Thea Aarrestad et al.. *Fast convolutional neural networks on FPGAs with hls4ml*. **Machine Learning: Science and Technology**

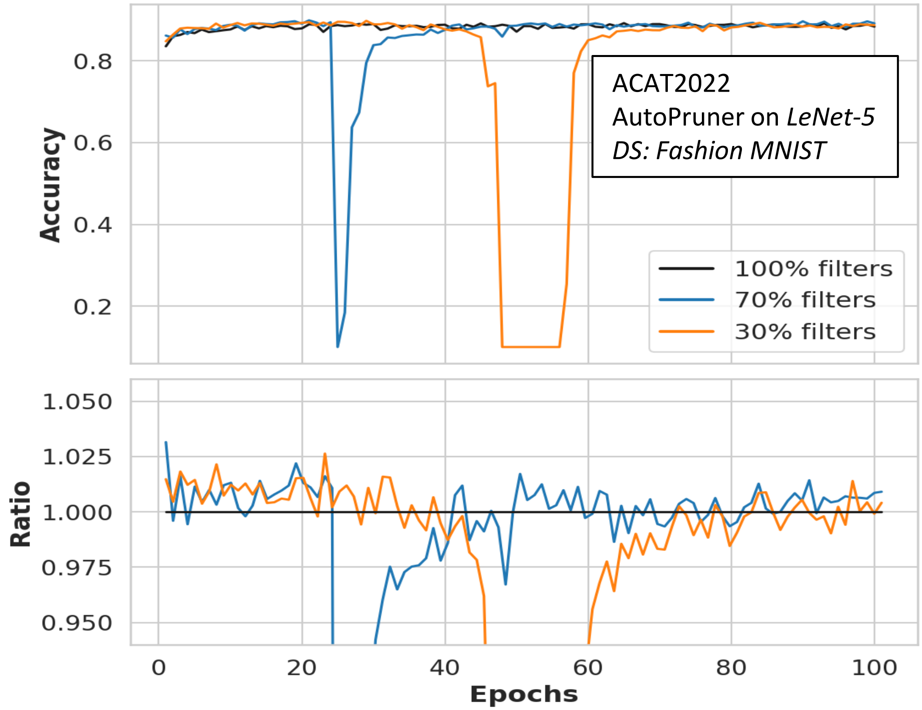
Pruning with AutoPruner

Pruning tool that works during training stage so that only a subset of nodes will contribute to the learning process, while **unnecessary nodes will be neglected**.

- The precise number of nodes required by the user
- A shadow network will automatically select the active nodes.

The tool can be easily applied to most used architecture.

Ph.D. Project of Daniela Mascione [UniTN, FBK]
"Deep Learning for online tagging of proton-proton commissions at the High-Luminosity LHC"

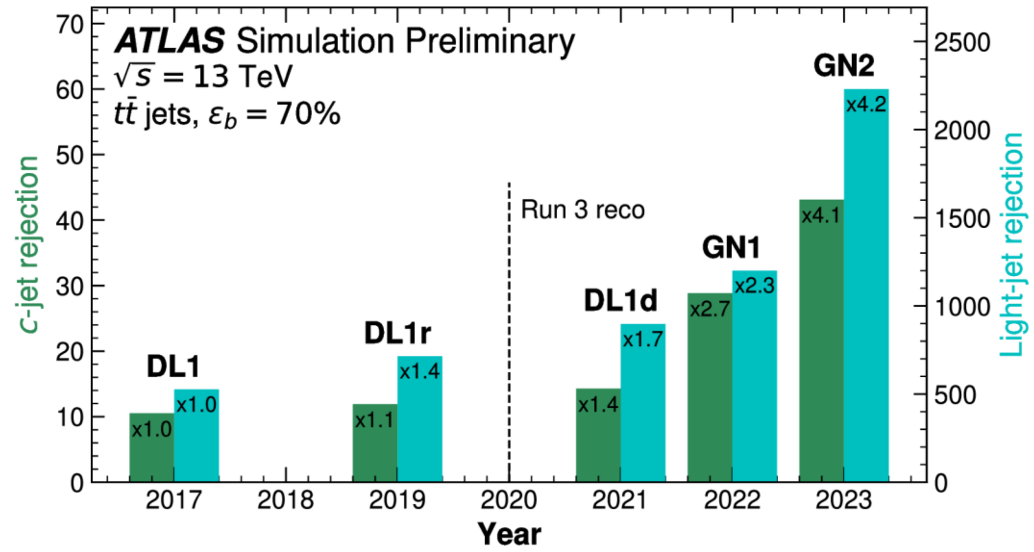


Summary

Next generation b/c taggers based on Graph Neural Networks show very promising results.

- GN1 performance will improve jet selection both for off-line and HLT level.
- GN2 is already set to be a strong successor.
- The development of DNN model reduction techniques is crucial for the development of smarter triggers for the high-luminosity program at the LHC.

FTAG-2023-01



deepPP initiative

In 2017, researchers and PhD student from the Physics Department and FBK took the deepPP initiative, focused on applications of Deep Neural Networks to high energy physics and astrophysics.

HIGGS BOSON ANALYSIS @LHC

Since the discovery of the Higgs boson in 2012, A lot of efforts were done to measure its properties. Within the ATLAS experiment, we study the properties of the decay of the Higgs boson to a couple of b-quarks.

DEEP LEARNING EXPLAINABILITY

Understanding how the output of a Deep Neural Network outputs is evaluated for a certain input set helps to detect bias and reduce systematic uncertanties.

DEEP LEARNING FAST INFERENCE

Deep Neural networks can be used at trigger level in High Energy Physics experiments to discriminate interesting events. This represent a challenging task since the inference should be fast enough to process large amount of data at a very high rate.

DEEP LEARNING FOR SPACE EXPERIMENTS

Deep learning algorithms have gained importance in astroparticle physics in the last years. They are implied in the most modern experiments for particle identification, tracking and energy reconstruction

ATLAS EXPERIMENT

LIMADOU 利玛窦


Roberto Iuppa
Professor


Marco Cristoforetti
Researcher


Francesco Maria Follega
Researcher


Andrea Di Luca
Postdoctoral researcher


Daniela Mascione
PhD candidate


Greta Brianti
PhD student

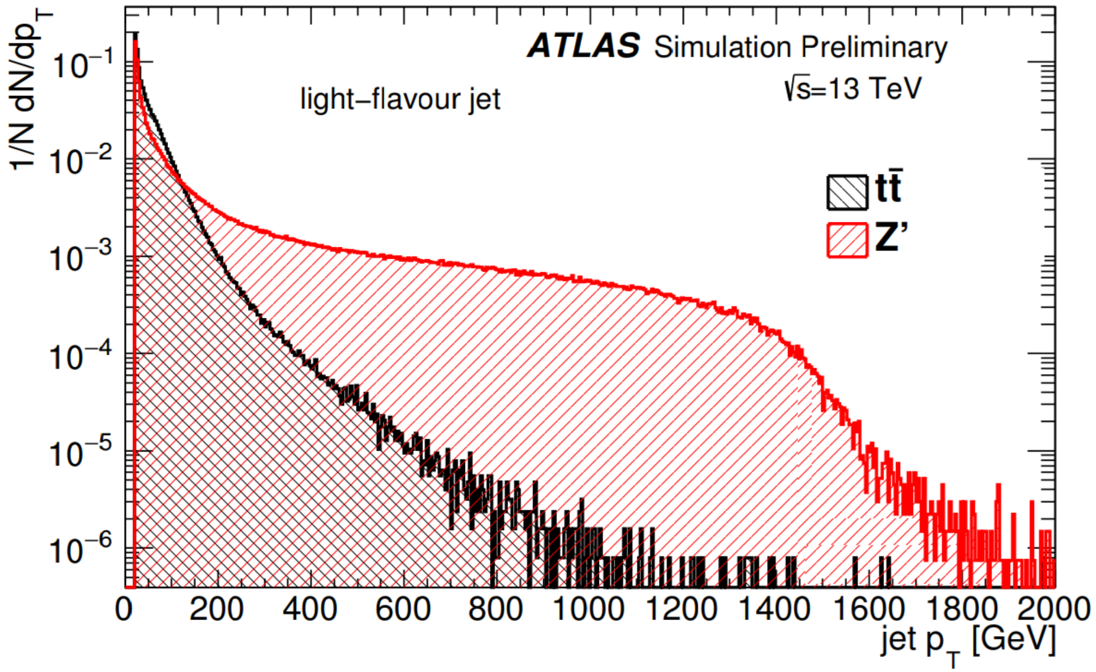

Megha Babu
PhD student

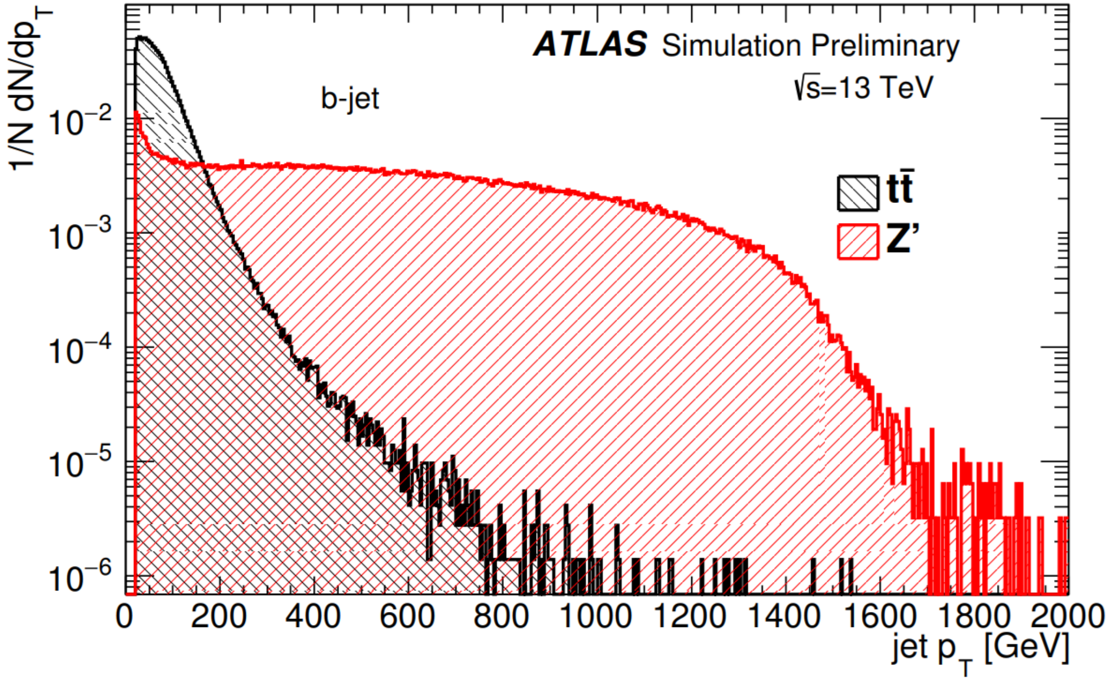
Back-up

Training sample

- Mixed dataset consisting of simulated events:
  - $t\bar{t}H$ for p_T < 250 GeV
  - $Z' \rightarrow q\,q$ for p_T > 250 GeV

Z' sample populates high p_T region





ATL-PHYS-PUB-2017-013

GN1 variables

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)

GN1 steps (one gnn layer)

1. the feature vectors of each node are fed into a fully connected layer W, to produce an updated representation of each node Whi
2. These updated feature vectors are used to compute edge scores $e(hi\ ,\ h\ j)$ for each node pair
3. These edge scores are then used to calculate attention weights $ai\ j$ for each pair of nodes using the softmax function over the edge scores
4. Finally, the updated node representation $h\ '\ i$ is computed by taking the weighted sum over each updated node representation Whi , with weights aij
5. The output representation for each track is combined using a weighted sum to construct a global representation of the jet, where the attention weights for the sum are learned during training

$$e(h_i, h_j) = \mathbf{a}^{\perp}\theta\left[\mathbf{W}h_i \oplus \mathbf{W}h_j\right]$$

$$a_{ij} = \mathrm{softmax}_j\left[e(h_i, h_j)\right]$$

$$h'_i = \sigma\left[\sum_{j \in \mathcal{N}_i} a_{ij} \cdot \mathbf{W}h_j\right]$$

GN2 improvements

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