Explaining Jet Flavour Taggers with Integrated Gradients

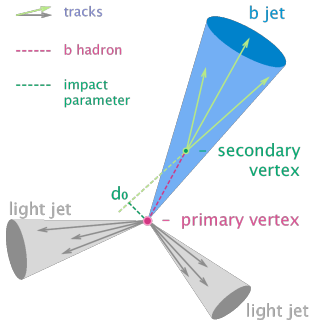
Scott DeGraw

University College London ATLAS Trk/FTAG team

2 October 2024

Jets

Proton-proton collisions at the LHC produce collimated streams of particles called jets.



- Jets originate from particle decay.
- The species of the originating particle determines the "flavour" of the jet.
- Accurately tagging jet flavours increases the sensitivity of physics analyses.

Jet Flavour Taggers

To tag these jets supervised deep learning models are used. The training data comes entirely from Monte Carlo simulations of the particles and their interactions with the detector.

- Simulations give truth labels on the jet flavour.
- Input features into the model are reconstructed jet and track features.
- Simulations additionally give truth information on the physical origin of each track and whether tracks originated from the same vertex.
- Truth jet labels are b, c, τ or light (u).

Model Architecture

Jet tagger model is a graph neural network.

- Tracks are modelled as nodes.
- Graph is fully connected.
- No edge features, node features are track features, global features are jet features.

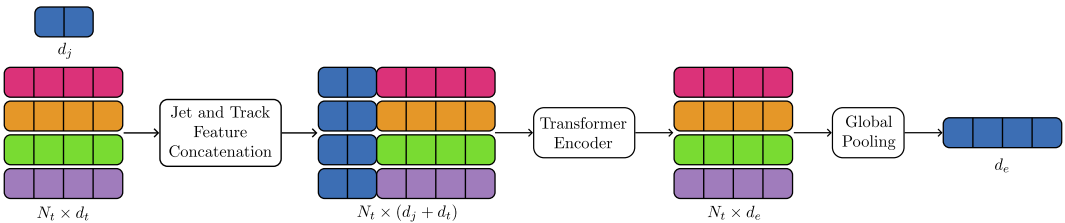


Figure: N_t is the number of tracks, d_j is the number of jet features, d_t is the number of track features, d_e is dimension of transformer output encoding. [2]

Integrated Gradients

Integrated gradients [3] provide attributions on each input feature. These attributions quantify how much each input feature contributes to the model output. For model $F : \mathbb{R}^n \to \mathbb{R}$, input $\mathbf{x}$, baseline input $\mathbf{x}'$ and feature index i

$$\mathrm{IG}_i(\mathbf{x}) \equiv (x_i - x_i') \int_0^1 \mathrm{d}\alpha \nabla_i F\left(\alpha \mathbf{x} + (1-\alpha)\,\mathbf{x}'\right). \tag{1}$$

This definition requires that the output of the model be a scalar, so we can choose the model output component with the largest value.

These attributions are all defined relative to a baseline input $\mathbf{x}'$ which will have a model output that is close to equal probabilities across all classes. This creates a neutral input.

Integrated Gradients

Integrated gradients satisfy many properties including completeness:

$$\sum_{i=1}^{n} \mathrm{IG}_i(\mathbf{x}) = F(\mathbf{x}) - F(\mathbf{x}') \tag{2}$$

This property gives an interpretation to the sign of the attributions. A positive attribution indicates that the feature pushes the model towards its predicted class. A negative attribution means the feature pushes the model away from its predicted class.

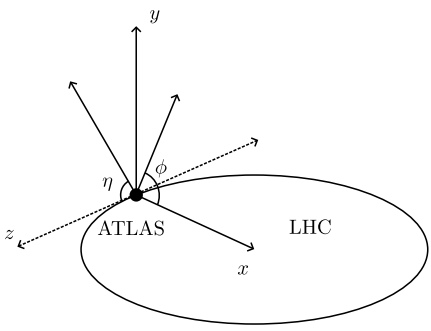
Completeness gives a scale to the attributions. We will normalise the attributions through dividing by the sum of all the attributions in the jet. This allows attributions across jets to be compared.

Features

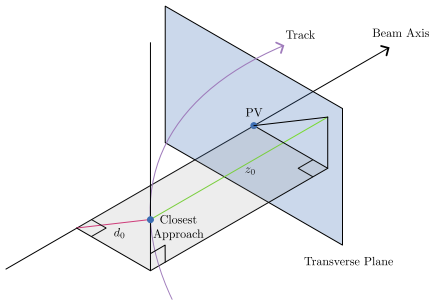
Jet Input	**Description**
p_{T}	Jet transverse momentum
η	Jet pseudorapidity ($\eta \equiv -\ln\tan(\theta/2)$)
Track Input	**Description**
q/p	Track charge divided by momentum (measure of curvature)
$\mathrm{d}\eta$	Pseudorapidity of track relative to the jet η
$\mathrm{d}\phi$	Azimuthal angle of the track, relative to the jet ϕ
d_0	Transverse impact parameter (IP) of track relative to PV
$z_0 \sin\theta$	Longitudinal IP multiplied by sine of polar angle
$\sigma(q/p)$	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
$s(z_0 \sin\theta)$	Lifetime signed longitudinal IP significance

There are additional variables counting how many times a track leaves a "hit" for different layers of the detector. These are found to have much lower attribution.

Angular and impact parameters



(a) η and ϕ in ATLAS.



(b) d_0 and z_0.

Track origins

Each of these tracks has a specific truth origin of what particle it originated from.
Some of them are strongly associated to certain jet flavours.

Truth Origin	**Description**
pileup	From a proton-proton collision other than the primary interaction
fake	Created from the hits of multiple particles
primary	Does not originate from any secondary decay
from b	From the decay of a b hadron
from bc	From a c hadron decay, which itself is from the decay of a b hadron
from c	From the decay of a c hadron
from τ	From the decay of a τ
other secondary	From other secondary interactions and decays

Attributions

We use a recent state-of-the-art jet flavour tagger (GN2 [1]) trained using Salt.

Plots of attributions only consider jets where the model makes a correct jet flavour prediction.

All jet flavour, track origin and track vertex compatibility labelling is done with truth level data.

Attribution heatmaps

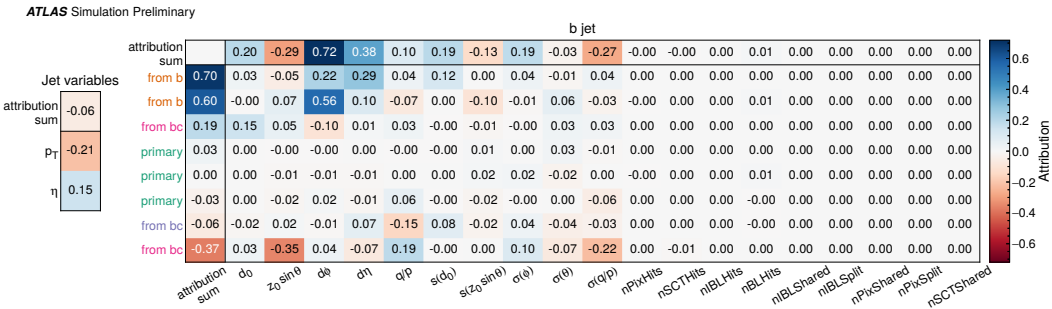


Figure: Heatmaps of the attributions for each feature. Tracks are labelled by their track origin. Track origin labels with same colour come from the same vertex. Left most column sums attributions across each track, top row sums attributions across each track variable.

Track origins

Summing over the attributions for each track gives us an attribution for an entire track. We can then rank the tracks by attribution, and identify the origin of that track.

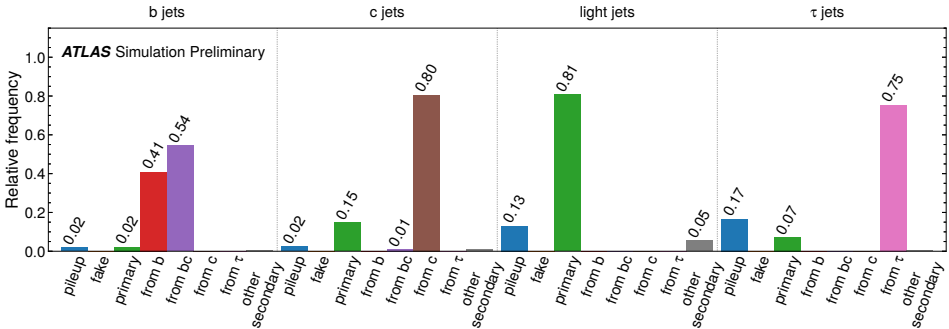


Figure: Relative frequency of track origins for top attributed track.

To find the total attribution of a track variable to the entire jet we can sum over the attributions across the tracks.

- Most attributions are close to zero, so this will pick up the most important attributions.
- Boxplots are used to visualise the difference in attribution distributions.
- A wider spread generally indicates that a variable is more important to the model.
- We can separately plot for the four different jet classes.
  - In this case, a distribution with a more positive mass indicates that this feature is especially important to the model as it is used more often than not to help drive the model prediction.

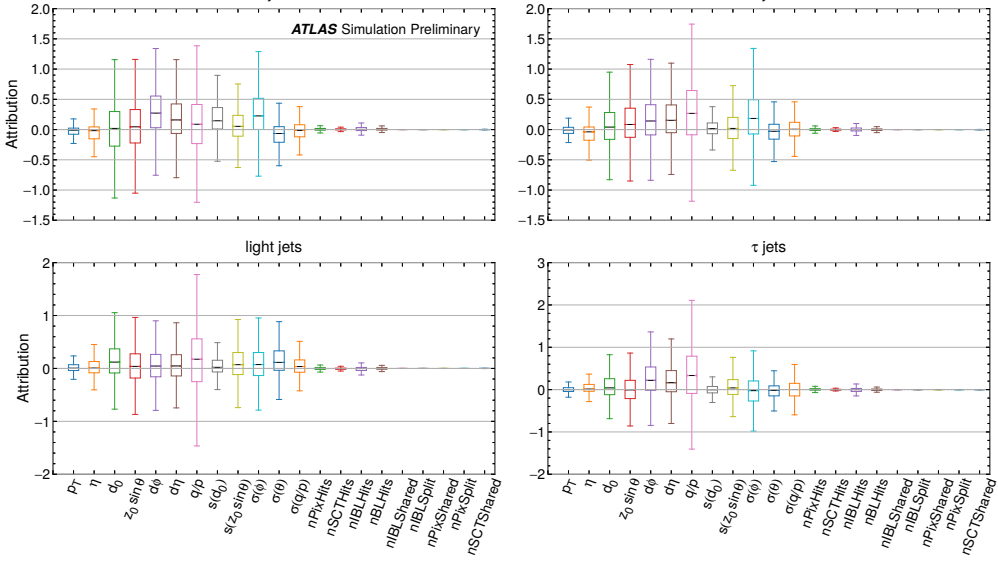
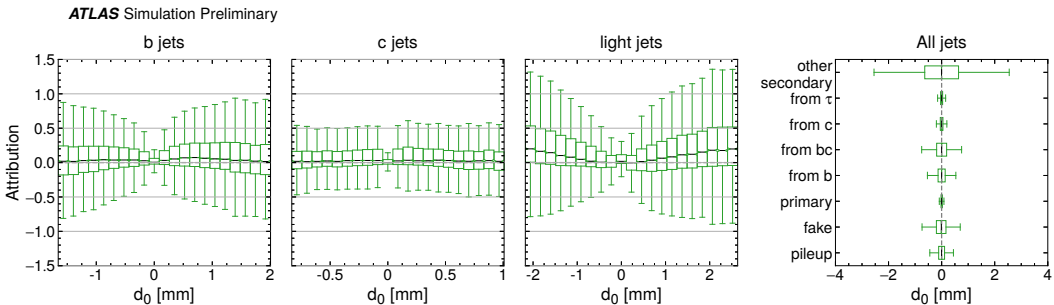


Figure: Attribution boxplots with track attributions summed over tracks.

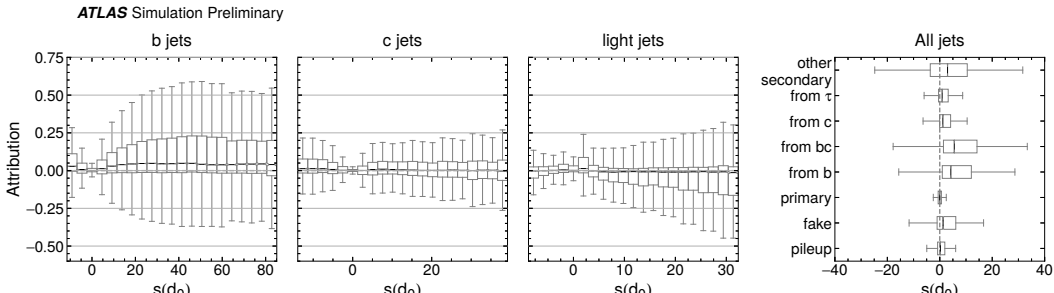
For the next plots:

- We are looking at how the attribution distribution changes with feature value.
- Attribution distributions are for top attributed track in each jet.
- τ jets are not included as the statistics are too small.

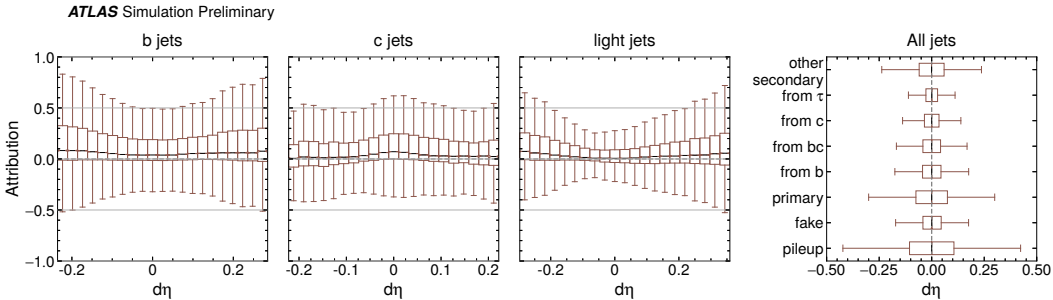
- Each boxplot is the attribution distribution with the associated feature value between the positions of the left and right box edges.
- Right most plot shows distributions of the features grouped by each track origin.



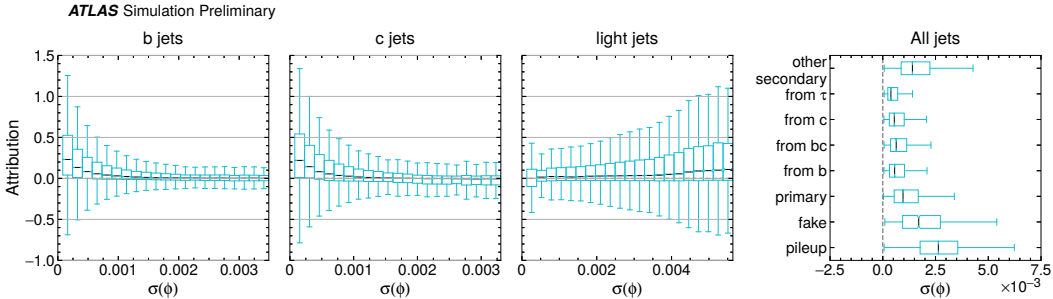
More positive attribution as $|d_0|$ increases for b and c jets due to from b and from c tracks having large $|d_0|$. As $|d_0|$ increases further, u jets acquire very positive attribution and b jets acquire more negative attribution due to other secondary tracks having very large $|d_0|$.



From b and bc tracks have very positive $s(d_0)$ reflected in the very positive attribution at large $s(d_0)$ for b jets. The trend for u jets is the inverse of the b jet trend with very negative attributions at large $s(d_0)$ and c jets have only a slight dependence on $s(d_0)$.



Other secondary, pileup and primary all have large $|\mathrm{d}\eta|$, so there is increasing positive attributions for *u* jets as $|\mathrm{d}\eta|$ increases. From *c* tracks have the smallest angles leading to the small peak at small $|\mathrm{d}\eta|$ for *c* jets. The *b* jets have slightly larger $|\mathrm{d}\eta|$ leading to a small trend upward in attribution as $|\mathrm{d}\eta|$ increases.



From *b*, *bc* and *c* tracks all have low $\sigma(\phi)$ leading to the very positive attributions for *b* and *c* jets in this regime. The *u* jets follow the inverse of this trend as other secondary, pileup and primary all have high $\sigma(\phi)$.

Conclusion

Understanding a complex jet flavour tagger is difficult.

- We demonstrate that integrated gradients can provide some insights into the physics that the model learns.
- Generally, we find back the physics that we expect out of the model.
- This type of analysis is potentially useful to motivate adding or removing variables into the model.
- Potential for debugging low performant jet taggers.

References

[1] ATLAS Collaboration. *Jet Flavour Tagging With GN1 and DL1d. Generator dependence, Run 2 and Run 3 data agreement studies*. CERN. URL: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/FTAG-2023-01/ (visited on 02/28/2023).

[2] ATLAS Collaboration. *Transformer Neural Networks for Identifying Boosted Higgs Bosons Decaying into B\barb and C\barc in ATLAS*. Geneva: CERN, 2023. URL: http://cds.cern.ch/record/2866601 (visited on 05/31/2024).

[3] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. *Axiomatic Attribution for Deep Networks*. June 12, 2017. DOI: 10.48550/arXiv.1703.01365. arXiv: 1703.01365 [cs]. URL: http://arxiv.org/abs/1703.01365 (visited on 09/20/2024). Pre-published.