PROCEEDINGS OF SCIENCE

Flavour Tagging with Graph Neural Network at ATLAS

Maxence Draguet[a,*], **on behalf of the ATLAS Collaboration**

[a] *Particle Physics Department, University of Oxford, Keble Road, Oxford OX1 3RH, United Kingdom*

E-mail: maxence.draguet@physics.ox.ac.uk

Flavour tagging is a critical component of the ATLAS experiment physics programme. Existing methods rely on several low-level taggers, combining machine learning models with physically informed algorithms. A novel approach presented here instead uses a single deep learning model based on reconstructed tracks, avoiding the need for low-level taggers based on secondary vertexing algorithms. This new approach reduces complexity and improves tagging performance. The model employs a transformer architecture to process information from a variable number of tracks and other objects in the jet to simultaneously predict the jets flavour, the partitioning of tracks into vertices, and the physical origin of each track. The new approach significantly improves jet flavour identification performance compared to existing methods in both Monte-Carlo simulation and collision data. Finally, a hyperparameter optimisation study is presented to further refine the performance of the model, deploying the maximal update parametrisation to lower the computational cost.

42nd International Conference on High Energy Physics (ICHEP2024)
18-24 July 2024
Prague, Czech Republic

*Speaker

PoS(ICHEP2024)1002

1. Introduction

The ambitious physics programme of the ATLAS experiment relies on several fundamental components. A key challenge is the accurate reconstruction of the flavour of the initial parton that hadronized into a jet, a capability essential for many analyses. This complex task is the focus of a dedicated group in the collaboration, where the development of *flavour taggers* is an ongoing effort aiming to improve the performance of existing tools. Leveraging the recent progress in deep learning techniques, the performance of the taggers designed for the current data-taking period, Run 3 of the LHC, has greatly exceeded the previously attained achievements. This is demonstrated in Figure 1, where we are comparing the performance of several algorithms, DL1, DL1r [1], DL1d [2], GN1 [3], and GN2 [4], in terms of the *rejection* - the inverse of the mistagging efficiency - at the 70% efficiency b-tagging working point. A higher rejection translates into fewer wrongly identified jets in the relevant analysis channel [3].

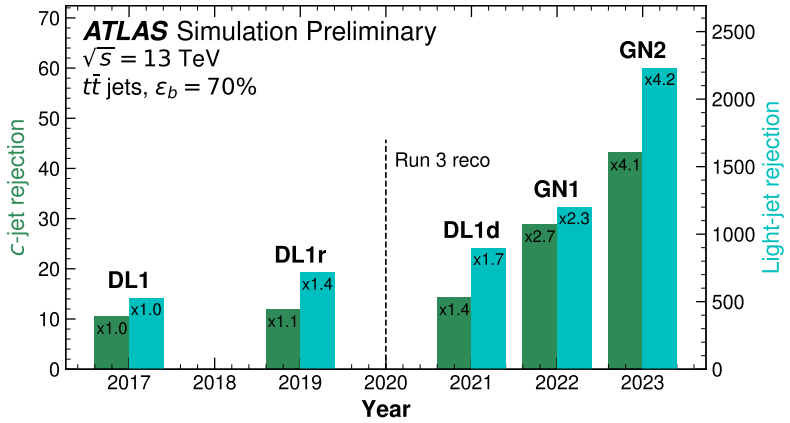

Figure 1: The c-jet and light-jet rejections of the different ATLAS flavour tagging algorithms over time in Monte Carlo simulation [3]. The rejections are provided at the 70% efficiency b-tagging working point. The improvement with respect to the DL1 rejection is indicated on each bar.

To deliver this improvement in performance, ATLAS has adopted modern deep learning architectures, primarily testing a graph attention network, for GN1 [3], and a transformer encoder, for GN2 [4, 5]. These algorithms represent a significant leap forward compared to the previous generation of taggers, which utilized a hierarchical approach combining physics-inspired algorithms reconstructing the secondary vertex with a track processing recurrent neural network (DL1r) [1] or a deep set network (DL1d) [2] as input to a deep neural network.

2. Flavour Tagging with Graph Neural Networks

The new generation of taggers is built on graph neural networks with a similar global architecture but with different core units, as shown in Figure 2. GN1 uses a graph attention network to extract conditional track representations from a set of tracks and some jet information, such as the transverse momentum and pseudorapidity. GN2 improves upon GN1 as the attention mechanism from its transformer encoder core is faster to compute, more stable, and regularised thanks to the layer normalisations [5]. In addition, GN2 is both multimodal and multitask. It combines track and jet variables and is trained with two auxiliary objectives in addition to the main jet tagging task.

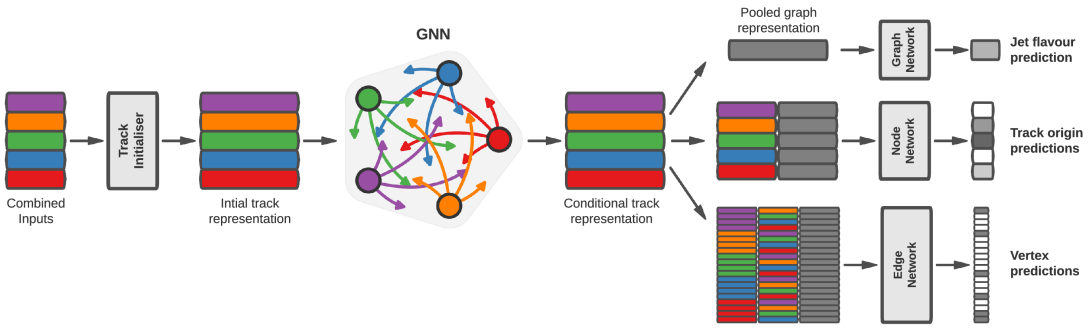

Figure 2: Structure of the ATLAS graph neural network taggers. Track and jet variables are jointly embedded and fed as input to the core graph unit, a graph attention network or a transformer encoder for GN1 or GN2, respectively. The output is a set of conditional track representations that is pooled to form a global graph representation. These representations are passed to three different objectives with individual feedforward networks: predicting the jet flavour probabilities, the track origin, and if two tracks share a vertex [4].

The main objective is to predict the probability that a jet originates from a specific flavour b, c, or light. In simulations, this flavour is determined based on the presence of a b-hadron for b-jets, a D-hadron but no b-hadron for c-jets, and light-jets (l) otherwise. A discriminant for b-tagging is constructed based on these probabilities as

$$D_b = \log \frac{p_b}{f_c \times p_c + (1 - f_c) \times p_l}, \tag{1}$$

where f_c is a tunable parameter that controls the weighing of different background rejections. A similar discriminant is derived for c-tagging. By applying a threshold to the discriminant, defined as c_{WP}, a *working point (WP)* is established with unique tagging efficiencies and rejections. In addition to the main task, two auxiliary tasks are achieved: predicting the origin of the tracks and whether two tracks share a vertex. These tasks are combined with the main task into a single loss function during training, allowing the network to distil expert knowledge and focus attention on relevant physics patterns [4].

GN2 is trained on Monte Carlo simulated datasets combining Standard Model $t\bar{t}$ events and the decays of exotic Z' bosons into two jets [3, 4]. The latter is included to extend the coverage of the datasets to higher momentum. Simulations incorporate detector effects reproducing the experiment conditions [4].

The performance achieved by GN2 is compared to the baseline DL1d tagger and the GN1 model in Figures 3 and 4, for b- and c-tagging respectively. GN2 significantly outperforms the hierarchical DL1d tagger and the more advanced GN1 model. At the 70% efficiency WP for b-tagging, GN2 enhances the light (c) rejection by a factor ~ 2 (~ 2.8) at energies below 250 GeV. This 70% efficiency b-tagging WP at low energies corresponds to a 20% efficiency at high energies, where GN2 outperforms DL1d by a factor ~ 4.8 (~ 5.5) for light (c) rejection. For c-tagging at the 40% efficiency WP, an improvement factor of 1.3 (2.7) in light (b) rejection is achieved by GN2 over DL1d at low energies. This 40% efficiency c-tagging WP at low energies corresponds to a 20% efficiency at high energies, where GN2 improves the light (b) rejection by a factor of 1.8 (2.8).

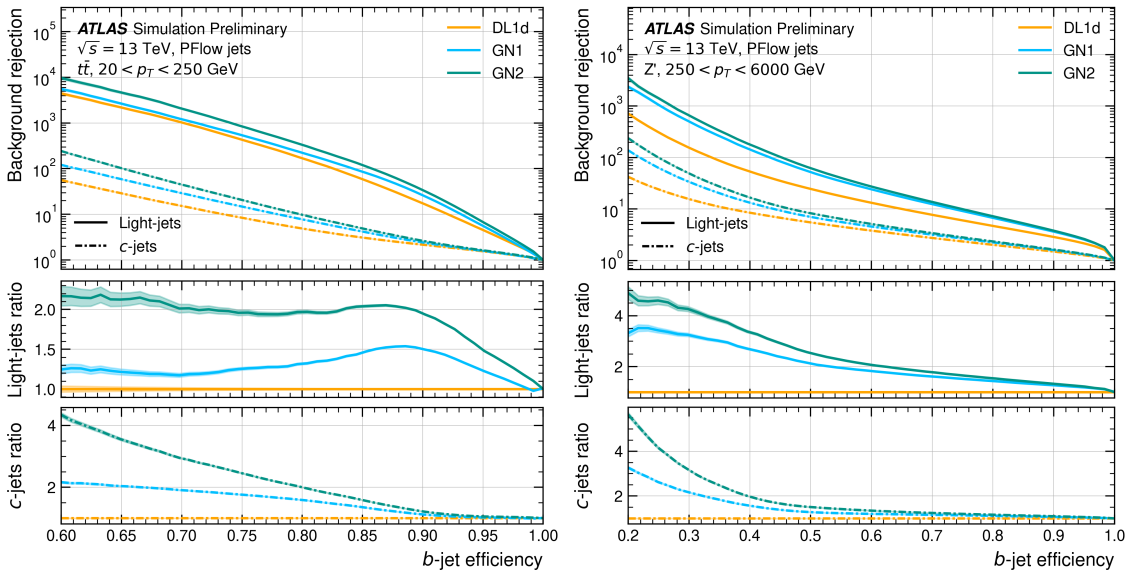

Figure 3: ROC curves tracing the b-tagging efficiency versus the light-jet and c-jet rejections for the $t\bar{t}$ (left) and Z' (right) samples [3]. Models compared are DL1d in orange, GN1 in blue, and GN2 in green. The bottom panels show the ratio to DL1d. The binomial error bands are shown as shaded regions.

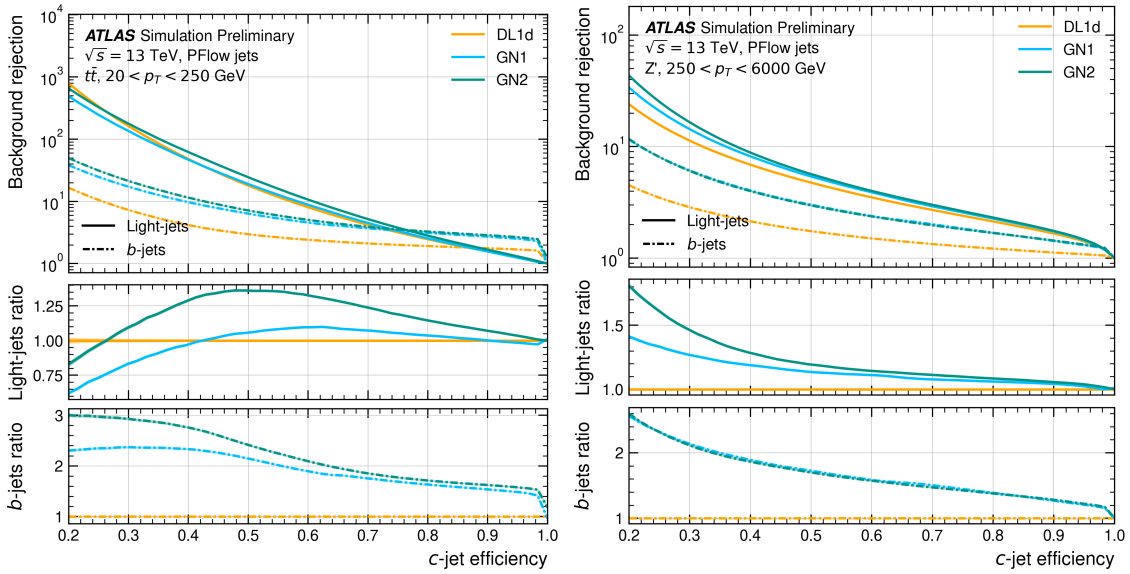

Figure 4: ROC curves tracing the c-tagging efficiency versus the light-jet and b-jet rejections for the $t\bar{t}$ (left) and Z' (right) samples [3]. Models compared are DL1d in orange, GN1 in blue, and GN2 in green. The bottom panels show the ratio to DL1d. The binomial error bands are shown as shaded regions.

3. Hyperparameter Optimisation

Maximising the performance of the flavour tagger is crucial for physics analysis depending on this tool to search for rare signal processes. As such, the network must undergo *hyperparameter optimisation (HPO)*, a process that involves adjusting key parameters through multiple independent training runs. This is a computationally expensive task that requires numerous iterations to find the

best values. In its current form, GN2 is a large transformer network of 2.6 million parameters - representing a significant increase from GN1, which has 0.8 million parameters, and DL1d, which has 0.13 million parameters.

To efficiently perform HPO, ATLAS has adopted the *maximal update parametrisation (μP)* and the μTransfer algorithm [6]. This approach performs the HPO search on a model with fewer neurons per layer, which reduces the computational cost of a single training. With the correct parametrisation provided by μP, the optimal hyperparameters identified for the smaller model can be reliably transferred to the full scale model [6]. μTransfer allows for a broader exploration of hyperparameters within a fixed computing budget, ultimately improving the tagger's performance.

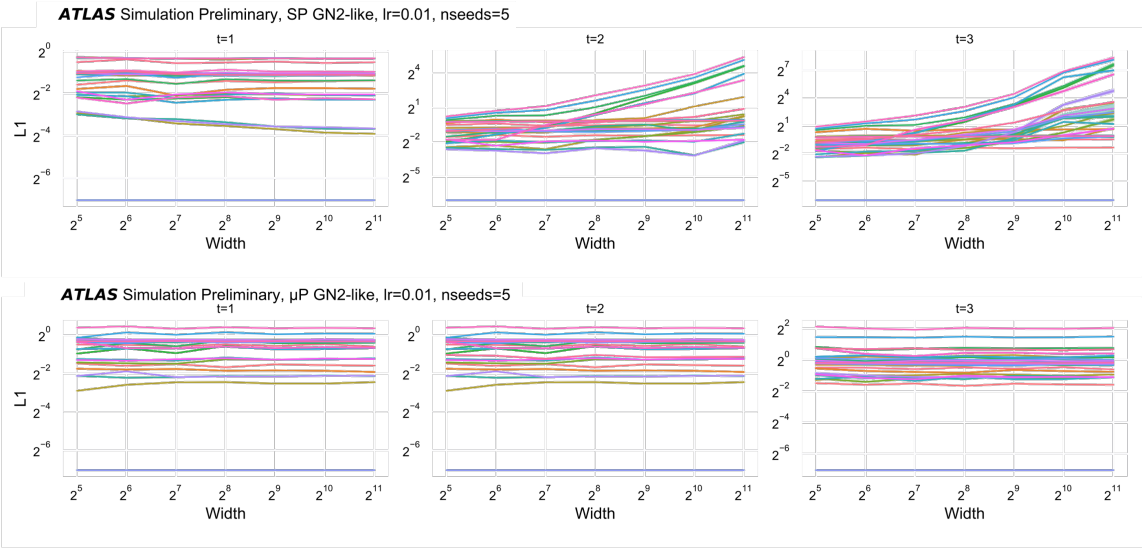

Figure 5: The sum of the absolute values of the weights for the different layers of the transformer of GN2 for different embedding widths, comparing the standard (top) to the maximal update parametrisation (bottom) at initialisation (left), after 1 training step (centre), and after 2 training steps (right).

The μP parametrisation adjusts the backpropagation of the updates by adopting a depth-dependent per-layer learning rate. This allows the input layers to be trained at the same rate as the outer layers which are closer to the loss calculation, enabling coherent in-depth training of a large model. The parametrisation is termed *maximal* because it achieves the largest possible per-layer update while maintaining an asymptotically width-independent scaling, even as the model grows. As shown in Figure 5, the sum of the weights of a layer in the *standard parametrisation (SP)* grows exponentially with the width of the layer during training but remains constant with μP.

Figure 6 compares a scan of the maximal value of the learning rate scheduler with the two parametrisations for GN2. Three model sizes are considered, with transformer embedding widths of 64, 128, and 256. With μP, a larger width model always outperforms a smaller one, simplifying the neural architecture search. At high learning rates, the 256-embedding SP GN2 becomes unstable while the μP version remains stable due to its width-independent scaling. The optimal maximum learning rate values are the same for the different μP models, which is not guaranteed with SP. The cost of a full model training is equivalent to four small model trainings, hence a better search of the hyperparameter space is achieved with μTransfer for a given computing budget.

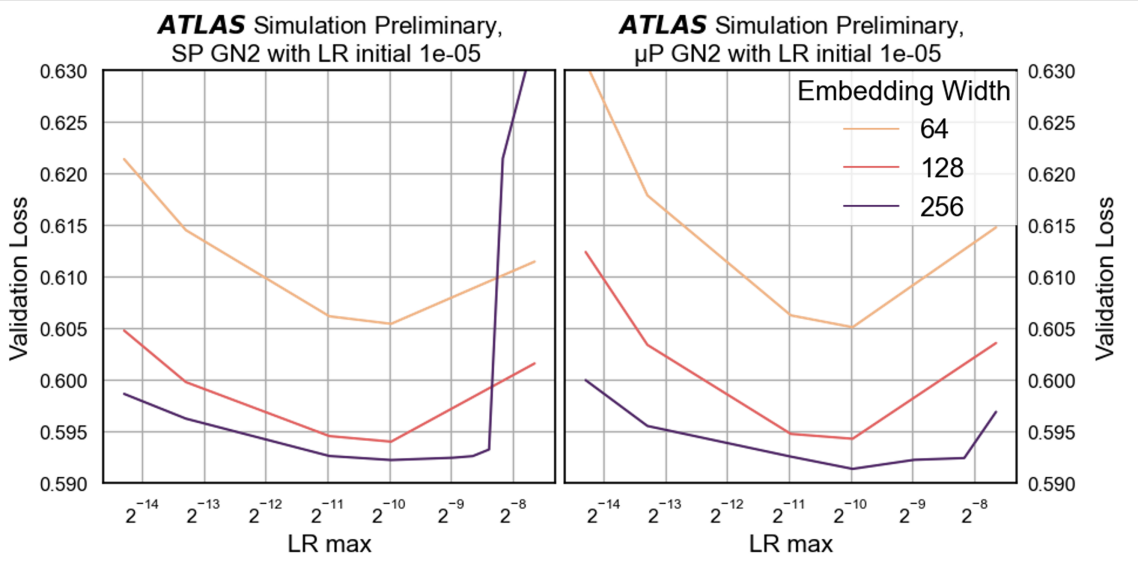

Figure 6: Maximal learning rate parameter scan versus validation loss for an *SP* (left) and a μP GN2 models (right) for three different embedding widths: 64 (yellow), 128 (red), and 256 (purple).

4. Conclusion

GN2 promises to significantly improve flavour tagging performance for analyses in Run 3 of the LHC. Departing from the previous generation of hierarchical taggers used by ATLAS, GN2 adopts a full deep learning approach leveraging state-of-the-art architectures and techniques, from the transformer to multitask training, μP for stabilisation, and μTransfer for improved hyperparameter search. Further refinements to GN2 are ongoing, including additional optimisations and the calibration of the tagger.

References

[1] ATLAS Collaboration, *ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset*, *Eur. Phys. J. C.* **83** (2023) 681.

[2] ATLAS Collaboration, *Deep sets based neural networks for impact parameter flavour tagging in ATLAS*, ATL-PHYS-PUB-2020-014, 2020.

[3] ATLAS Collaboration, *Jet Flavour Tagging With GN1 and DL1d. Generator dependence, Run 2 and Run 3 data agreement studies*, ATL-PLOT-FTAG-2023-01, 2023.

[4] ATLAS Collaboration, *Graph neural network jet flavour tagging with the ATLAS detector*, ATL-PHYS-PUB-2022-027, 2022.

[5] A. Vaswani et al., *Attention is all you need*, proceedings of the *International Conference on Neural Information Processing Systems* **31** (2017) 6000 - 6010.

[6] G. Yang et al., *Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer*, arXiv:2203.03466, 2022.

Acknowledgments

The author would like to acknowledge the use of the University of Oxford *Advanced Research Computing (ARC)* facility in carrying out the hyperparameter optimisation presented in this work.