The Pareto frontier of resilient jet tagging

Anonymous Author(s)
Affiliation
Address
email

Abstract

Classifying hadronic jets using their constituents' kinematic information is a critical task in modern high-energy collider physics. Often, classifiers are designed by targeting the best performance using metrics such as accuracy, AUC, or rejection rates. However, the use of a single metric can lead to the use of architectures that are more model-dependent than competitive alternatives, leading to potential uncertainty and bias in analysis. We explore such trade-offs and demonstrate the consequences of using networks with high performance metrics but low resilience.

1 Introduction

When strongly-interacting quarks and gluons are produced by high-energy particle collisions at colliders like the Large Hadron Collider (LHC), they shower and hadronize, creating a collimated 'jet' of particles in the final state that is imprinted with some properties of the originating particle (1). Classification, or *tagging*, of these jets based on their *substructure* has become a critical task at the Large Hadron Collider (LHC), where many studies require doing so to extract maximal information from the data (2). Jet tagging has become the quintessential proving grounds for Artificial Intelligence / Machine Learning (AI/ML) algorithms at the LHC: state-of-the-art transformer and graph-based architectures (3; 4; 5; 6; 7) are significantly more performant than earlier approaches (8; 9; 10; 11; 12).

Economists say, "When a measure becomes a target, it ceases to be a good measure." (13) While the accuracy of an ML/AI classifier, often measured by 'AUC' (area under the ROC curve), is a critical benchmark, fixation on this quantity can lead to sub-optimal outcomes in analyses. As model complexity increases, they can become susceptible to learning idiosyncrasies of the simulated training sample rather than genuine generalizable physics information. This has been studied by ATLAS (14), which showed that classifiers are more susceptible to uncertainties related to physics modeling than those related to detector effects. Similar studies have explored solutions to generalizability (15).

In this work, we evaluate architectures that are often used for tagging in terms of their AUC and their simulation model-dependence, or'resilience' (16; 17). Models with varying complexities were trained and tested on *different* Monte Carlo (MC) simulated datasets; then used to construct the 'Pareto frontier' (18) of AUC *vs.* resilience. We perform a case study to demonstrate the risk of biasing downstream parameter estimation tasks when using models with low resilience. We advocate for a holistic approach to classifier design that includes multiple benchmarks, suited to the application.

2 Methodology

2.1 Monte Carlo simulated event samples

Two jet classification tasks were considered in these studies: the discrimination of jets initiated by a quark or gluon ('q/g tagging'), and the identification of jets resulting from the hadronic decay of a Lorentz-boosted top quark ('top tagging'). For each of these tasks, a set of MC simulated events

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

generated with PYTHIA 8 (19) was used to train the classifiers in a fully-supervised manner. PYTHIA samples used the default Monash set of tuned parameters (20) in all cases. Alternative samples of the same processes were also generated with HERWIG 7 (21)[1] , to enable quantification of the resilience as the AUC %-difference between testing on the nominal and alternative sample.[2] All jets, regardless of their size, are reconstructed with FASTJET (22) and filtered to have a transverse momentum (p_{T}) between 500-550 GeV. No detector simulation is applied.

For q/g tagging studies, the simulated event samples from Refs. (23) were used, which are freely available on the CERN Zenodo platform (24; 25). These samples consist of one million anti-k_t $R = 0.4$ jets from each of the $Z + q$ or $Z + g$ processes, where the Z boson decays into neutrinos. The boosted top tagging studies were performed using a new set of samples, which were generated for these studies and made publicly available on the CERN Zenodo platform (26). One million mixed q/g background jets for this task with the same kinematic selection were produced from a dijet process.

2.2 Model architectures surveyed

We have surveyed a representative selection of architectures that are either currently used or have recently been used in physics analysis at the LHC, and summarize the setups studied in Table 1. All networks were trained in a fully-supervised manner for 500 epochs[3], using early stopping and a patience of 10 epochs. The default ADAM (27) optimizer was used, with a learning rate of 0.001. The MC samples were split such that 75% of the events were used for training, and 12.5% each were used for testing & validation. Each network was given only particle-level kinematic information (constituent p_T, pseudorapidity η, and azimuthal angle ϕ) as input.

Table 1: An overview of the jet tagging architectures surveyed in this study. For each tagger, the key hyperparameters that were scanned to explore the performance-resilience trade-off are listed.

Architecture	**Hyperparameters Scanned**	**Reference**
Expert Features	Angularities: β values Multiplicities: p_{T} cuts, charge req.	(28)
Deep Neural Networks (DNNs)	Hidden layers: 2–10 Neurons per layer: 1–300	(23)
Particle-Flow Networks (PFNs)	Latent dimension ℓ: 1–1024 Φ network nodes: 50–500	(23)
Energy-Flow Networks (EFNs)	Same as PFNs	(23)
Particle Transformer (ParT)	Attention heads: 2, 4, 8	(29)

3 Results

3.1 Pareto Frontier

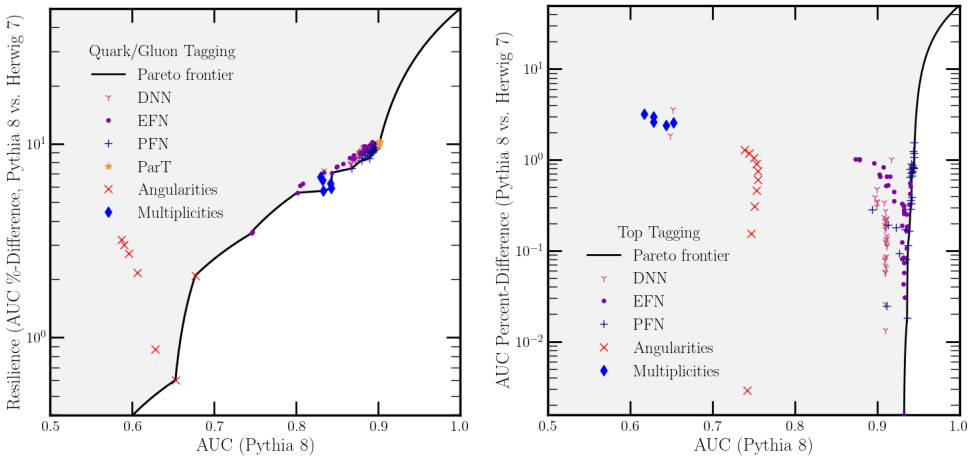
Figure 1 shows the classifier AUC *vs.* resilience (AUC %-difference for Pythia vs. Herwig samples), for each of the models trained and for each set of hyperparameters considered: 'optimal' performance is at the lower-right corner of the figure. The Pareto frontier connecting models that optimize the AUC-resiliency tradeoff is highlighted, and models in the shaded region are Pareto-excluded.

The Pareto frontier shows that more "complex" models (e.g. ParT) do achieve a higher raw performance, but at the cost of resiliency. On the other hand, simpler models based on physical principles, such as EFNs or mere expert features like angularities, are more robust. For top tagging, vertical columns of different network types strongly discourage the use of unnecessarily complex networks.

[1]For q/g (top), samples are generated with PYTHIA version 8.226 (8.331) and HERWIG version 7.1.4 (7.3.0).

[2]There are many potential ways to define resilience — just as AUC does not capture all important aspects of classification, so too does this particular definition of resilience.

[3]This number was chosen arbitrarily, the early stopping condition causes training to terminate in between 30-100 epochs in most cases.



(a) (b)

Figure 1: The Pareto frontier for (a) q/g tagging, (b) top tagging tasks. The AUC of models trained with PYTHIA samples is plotted *vs.* the resilience, defined as the percent difference in AUC evaluated on the PYTHIA and HERWIG test samples. The shaded grey region is Pareto-excluded.

3.2 Knowledge Distillation

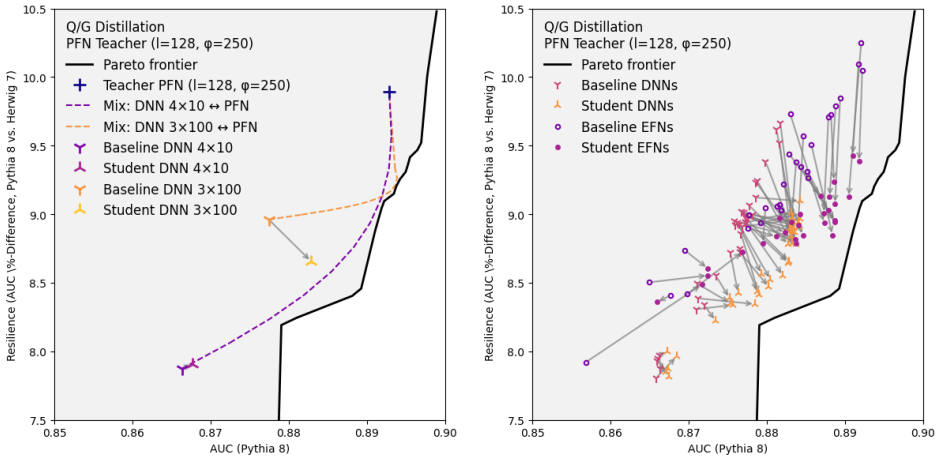
In an attempt to overcome the Pareto frontier and make models better along both AUC and resiliency, we tried Knowledge Distillation: a complex 'teacher' model is used to train a less complex 'student' model (30). Ultimately, this approach was unsuccessful in overcoming the Pareto frontier, but interesting observations made during the study motivate us to document it here. The PFN with $\ell = 128$ and 250 dense nodes per hidden layer was used as the teacher for this study, while various DNN and EFN models were used as students. The training procedure in Section 2.2 was modified such that the students were trained instead using the teacher's prediction as 'soft labels' by minimizing the KL-divergence between the predictions of the teacher and student models per-batch. Various forms of regularization were tested, but no significant change in the outcome was found.

The results for a representative pair of student models are shown in the AUC-Resiliance plane in Figure 2a, along with the teacher model and 'baseline' models whose architectures match the students, but that are not trained using distillation. The contour between the baseline and teacher models is also drawn: it is obtained by performing inference with a linear combination of the two models on the test set that varies from pure-teacher to pure-baseline. The student models beat this contour, demonstrating non-trivial improvement: the AUC of the model increases more than its resilience degrades. However, when models that are closer to the Pareto frontier are selected for distillation, the observed improvement is reduced. The results of distilling into the many DNN and EFN models we study is shown in Figure 2b: while many students improve, none surpass the existing frontier.

3.3 Case Study: determining q/g fractions

For a realistic downstream analysis task, a less accurate but more resilient classifier can ultimately yield a less biased physics result. We illustrate this with a case study, where the flavor *mixture fraction* κ of a mixed sample of quark and gluon jets is estimated using two PFNs located on the Pareto frontier: a small, resilient PFN ($\ell = 8$, 50 nodes per hidden layer) and a large network with a higher AUC ($\ell = 128$, 250 nodes per hidden layer). Given a classifier, one can extract the per-event q/g likelihood ratio $\ell(x)$ via the Neyman Pearson Lemma (31), from which the κ maximum likelihood inferences may be extracted. The flavor composition of two mixed jet samples with respective quark-initiated jet fractions of 50% and 25% is estimated with both networks, for jets from either the PYTHIA or HERWIG samples; the results are tabulated in Table 2. The experiment is re-run 10 times with re-trained networks as a proxy for confidence interval estimation.

Both the large and small networks are able to accurately recover the mixture fraction when samples are constructed with PYTHIA jets. However, when classifiers trained with PYTHIA jets are used to estimate the mixture fraction in a sample of HERWIG jets (used as pseudodata sample in this



(a) (b)

Figure 2: (a) Results of training two student DNNs via distillation from a teacher PFN. (b) Summary of distillation training from a teacher PFN to all DNNs and EFNs in the study.

study), the inferred κ value is biased. The extraction process can be calibrated by reweighting using a *second* set of classifiers that distinguish between PYTHIA and HERWIG. This classifier approximates the likelihood ratio between the two classes, allowing one to reweight one class of samples to be statistically identical from the other class (modulo reweighting uncertainties)[4]. The PFN models used for the reweighting are identical in architecture to those used for classification. Following calibration, we see from Table 2 that the less resilient model is still biased: the inferred $\hat{\kappa}$ values are not statistically consistent with the true κ values. The more resilient model is unbiased (within 2σ) following the calibration procedure, despite its naively worse performance according to the AUC. The conclusions of this study may apply broadly: any perturbation of the correlation structure of a sample with respect to the training set (*e.g.* fast *vs.* full detector simulation) may result in such a bias for parameter extractions depending on a unresilient classifier. This is particularly relevant to substructure, as predictions are known to differ from each other and from measurement (33; 34; 35; 36; 37; 38; 39; 40).

Table 2: Summary of the q/g mixture fraction (κ) estimation case study. Reported uncertainties are determined from the standard error from multiple pseudo-experiments.

		Pythia 8	*Herwig 7*		
Classifier	**True κ**	**Inferred $\mathbb{E}[\hat{\kappa}]$**	**Inferred $\mathbb{E}[\hat{\kappa}]$**	**Calibrated $\mathbb{E}[\hat{\kappa}]$**	**Result**
Large PFN	0.50	0.490 ± 0.005	0.296 ± 0.007	0.529 ± 0.006	Biased ✗
	0.25	0.253 ± 0.005	0.125 ± 0.005	0.305 ± 0.006	Biased ✗
Small PFN	0.50	0.504 ± 0.013	0.336 ± 0.016	0.478 ± 0.017	Unbiased ✓
	0.25	0.258 ± 0.013	0.157 ± 0.014	0.268 ± 0.013	Unbiased ✓

4 Concluding remarks

There is a clear trade-off between classifier performance and resilience, which we have visualized in these studies as a Pareto frontier. We have found that the complexity of a given model is a primary driver along the frontier, and that suboptimal model architectures can be improved with more sophisticated approaches to training such as knowledge distillation. Ultimately, the choice of a classifier that is not resilient can lead to suboptimal performance and increased bias in downstream tasks, even if the model is more accurate than others, motivating a more holistic approach to classifier development that includes multiple benchmarks.

[4] In principle, one can extract reweighting uncertainties and obtain confidence intervals using a method such as WiFi ensembles (32), but for simplicity we do not do this here.

References

[1] G. P. Salam, "Towards Jetography," *Eur. Phys. J. C*, vol. 67, pp. 637–686, 2010.

[2] S. Marzani, G. Soyez, and M. Spannowsky, *Looking inside jets: an introduction to jet substructure and boosted-object phenomenology*, vol. 958. Springer, 2019.

[3] ATLAS Collaboration, "Constituent-Based Quark Gluon Tagging using Transformers with the ATLAS detector." ATL-PHYS-PUB-2023-032, 2023.

[4] ATLAS Collaboration, "Constituent-Based W-boson Tagging with the ATLAS Detector." ATL-PHYS-PUB-2023-020, 2023.

[5] ATLAS Collaboration, "Tagging boosted W bosons applying machine learning to the Lund Jet Plane." ATL-PHYS-PUB-2023-017, 2023.

[6] ATLAS Collaboration, "Constituent-Based Top-Quark Tagging with the ATLAS Detector." ATL-PHYS-PUB-2022-039, 2022.

[7] CMS Collaboration, "Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques," *JINST*, vol. 15, p. P06005, 2020.

[8] CMS Collaboration, "Identification techniques for highly boosted W bosons that decay into hadrons," *JHEP*, vol. 12, p. 017, 2014.

[9] ATLAS Collaboration, "Identification of boosted, hadronically decaying W bosons and comparisons with ATLAS data taken at $\sqrt{s} = 8\,\text{TeV}$," *Eur. Phys. J. C*, vol. 76, p. 154, 2016.

[10] ATLAS Collaboration, "Identification of high transverse momentum top quarks in pp collisions at $\sqrt{s} = 8\,\text{TeV}$ with the ATLAS detector," *JHEP*, vol. 06, p. 093, 2016.

[11] ATLAS Collaboration, "Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC," *Eur. Phys. J. C*, vol. 79, p. 375, 2019.

[12] ATLAS Collaboration, "Performance and calibration of quark/gluon-jet taggers using $140\,\text{fb}^{-1}$ of pp collisions at $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS detector," 2023.

[13] M. J. Artis, "Monetary Theory and Practice. The UK Experience," *The Economic Journal*, vol. 94, pp. 984–985, 12 1984.

[14] ATLAS Collaboration, "Accuracy versus precision in boosted top tagging with the ATLAS detector," *JINST*, vol. 19, no. 08, p. P08018, 2024.

[15] A. Butter, B. M. Dillon, T. Plehn, and L. Vogel, "Performance versus resilience in modern quark-gluon tagging," *SciPost Phys. Core*, vol. 6, p. 085, 2023.

[16] G. Soyez, "Pileup mitigation at the LHC: A theorist's view," *Phys. Rept.*, vol. 803, pp. 1–158, 2019.

[17] *Les Houches 2017: Physics at TeV Colliders Standard Model Working Group Report*, 3 2018.

[18] E. Zitzler, J. Knowles, and L. Thiele, *Quality Assessment of Pareto Set Approximations*, p. 373–404. Berlin, Heidelberg: Springer-Verlag, 2008.

[19] T. Sjöstrand, S. Mrenna, and P. Skands, "A brief introduction to PYTHIA 8.1," *Comput. Phys. Commun.*, vol. 178, pp. 852–867, 2008.

[20] P. Skands, S. Carrazza, and J. Rojo, "Tuning PYTHIA 8.1: the Monash 2013 Tune," *Eur. Phys. J. C*, vol. 74, no. 8, p. 3024, 2014.

[21] J. Bellm *et al.*, "Herwig 7.1 Release Note," 2017.

[22] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet User Manual," *Eur. Phys. J. C*, vol. 72, p. 1896, 2012.

[23] P. T. Komiske, E. M. Metodiev, and J. Thaler, "Energy Flow Networks: Deep Sets for Particle Jets," *JHEP*, vol. 01, p. 121, 2019.

[24] P. T. Komiske, E. M. Metodiev, and J. Thaler, "Pythia8 quark and gluon jets for energy flow, v1." Zenodo, May 2019. https://doi.org/10.5281/zenodo.3164691.

[25] A. Pathak, P. T. Komiske, E. M. Metodiev, and M. Schwartz, "Herwig7.1 quark and gluon jets, v1." Zenodo, May 2019. https://doi.org/10.5281/zenodo.3066475.

[26] "Pythia8 and herwig7 boosted top & qcd jet datasets." Online, 8 2025. https://cernbox.cern.ch/s/ZyL5WzrlDZmLtuq.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[28] A. J. Larkoski, J. Thaler, and W. J. Waalewijn, "Gaining (Mutual) Information about Quark/Gluon Discrimination," *JHEP*, vol. 11, p. 129, 2014.

[29] H. Qu, C. Li, and S. Qian, "Particle transformer for jet tagging," in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 18281–18292, PMLR, 17–23 Jul 2022.

[30] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015.

[31] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933.

[32] S. Benevedes and J. Thaler, "Frequentist Uncertainties on Neural Density Ratios with wifi Ensembles," 5 2025.

[33] ATLAS Collaboration, "Measurement of the Soft-Drop Jet Mass in pp Collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector," *Phys. Rev. Lett.*, vol. 121, p. 092001, 2018.

[34] ATLAS Collaboration, "Measurement of soft-drop jet observables in pp collisions with the ATLAS detector at $\sqrt{s} = 13$ TeV," *Phys. Rev. D*, vol. 101, p. 052007, 2020.

[35] ATLAS Collaboration, "Properties of jet fragmentation using charged particles measured with the ATLAS detector in pp collisions at $\sqrt{s} = 13$ TeV," *Phys. Rev. D*, vol. 100, p. 052011, 2019.

[36] ATLAS Collaboration, "Measurement of the Lund Jet Plane Using Charged Particles in 13 TeV Proton–Proton Collisions with the ATLAS Detector," *Phys. Rev. Lett.*, vol. 124, p. 222002, 2020.

[37] ATLAS Collaboration, "Measurements of Lund subjet multiplicities in 13 TeV proton-proton collisions with the ATLAS detector," *Phys. Lett. B*, vol. 859, p. 139090, 2024.

[38] ATLAS Collaboration, "Measurement of jet track functions in pp collisions at s=13 TeV with the ATLAS detector," *Phys. Lett. B*, vol. 868, p. 139680, 2025.

[39] CMS Collaboration, "Study of quark and gluon jet substructure in Z+jet and dijet events from pp collisions," *JHEP*, vol. 01, p. 188, 2022.

[40] CMS Collaboration, "Measurement of the primary Lund jet plane density in proton-proton collisions at $\sqrt{\mathrm{s}}$ = 13 TeV," *JHEP*, vol. 05, p. 116, 2024.