

Fundamental limit of jet tagging

Joep Geuskens^{1,*} Nishank Gite^{2,†} Michael Krämer^{1,‡} Vinicius Mikuni^{3,§} Alexander Mück^{1,||}
Benjamin Nachman^{4,5,¶} and Humberto Reyes-González^{1,**}

¹*Institute for Theoretical Particle Physics and Cosmology, RWTH Aachen University,
Sommerfeldstr. 16, 52074 Aachen, Germany*

²*Department of Physics, University of California, Berkeley, California 94720, USA*

³*National Energy Research Scientific Computing Center, Berkeley Lab, Berkeley, California 94720, USA*

⁴*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

⁵*Berkeley Institute for Data Science, University of California, Berkeley, California 94720, USA*



(Received 5 February 2025; accepted 13 October 2025; published 7 November 2025)

Identifying the origin of high-energy hadronic jets (jet tagging) has been a critical benchmark problem for machine learning in particle physics. Jets are ubiquitous at colliders and are complex objects that serve as prototypical examples of collections of particles to be categorized. Over the last decade, machine learning-based classifiers have replaced classical observables as the state of the art in jet tagging. Increasingly complex machine learning models are leading to increasingly more effective tagger performance. Our goal is to address the question of convergence—are we getting close to the fundamental limit on jet tagging or is there still potential for computational, statistical, and physical insights for further improvements? We address this question using state-of-the-art generative models to create a realistic, synthetic dataset with a known jet tagging optimum. Various state-of-the-art taggers are deployed on this dataset, showing that there is a significant gap between their performance and the optimum. Our dataset and software are made public to provide a benchmark task for future developments in jet tagging and other areas of particle physics.

DOI: [10.1103/mj1l-w2b1](https://doi.org/10.1103/mj1l-w2b1)

Identifying the origin of hadronic jets has been a central task in collider physics for decades. A single high-energy jet at the Large Hadron Collider can be composed of hundreds of particles, each with a four-vector and other properties. This makes jet classification (tagging) a high-dimensional inference task. Until a few years ago, most of the progress in jet tagging was made through the development of physics-informed observables based on the radiation pattern within jets (jet substructure) [1,2]. Modern machine learning has led to rapid progress in jet tagging, leading to many factors of improvement over traditional algorithms (see, e.g., Ref. [3–7]). There is seemingly no end to the performance gains.

In the language of statistical learning, we know that there is a theoretical upper bound to the sensitivity of a jet tagging algorithm: *the likelihood ratio* [8]. Given two classes of jets A and B and a representation of a jet $x \in \mathbb{R}^N$, any monotonic function of $p(x|A)/p(x|B)$ is the optimal classifier—there is no other observable that can have a lower probability of mistagging a B jet as an A jet (false positive rate) for a given probability of correctly identifying an A jet as an A jet (true positive rate). In practice, we do not have direct access to the likelihoods $p(x|A)$ and $p(x|B)$, especially since x can be high dimensional and jets have a stochastic number of constituents; jets are thus *point clouds*. For this reason, we do not know how close existing approaches are to the theoretical optimum.

Modern machine learning has given us another tool that may help us answer this question: generative AI. Neural networks are capable of approximating $p(x|A)$ and $p(x|B)$ and can also sample from the probability densities. Generative networks have been studied for many years now in collider physics (see, e.g., Ref. [9–11]), but it is only recently that we have models that both can parse the structure of jets as full (variable-length) point clouds and estimate the probability density at the same time. Such methods include normalizing flows [12–15], diffusion models [16], and autoregressive models [17]. In this paper,

*Contact author: joep.geuskens@rwth-aachen.de

†Contact author: nishankgite@berkeley.edu

‡Contact author: mkraemer@physik.rwth-aachen.de

§Contact author: vmikuni@lbl.gov

||Contact author: mueck@physik.rwth-aachen.de

¶Contact author: bpnachman@lbl.gov

**Contact author: humberto.reyes@rwth-aachen.de

we use the autoregressive model of Ref. [17] to create a highly realistic, synthetic dataset where we know the per-jet likelihood ratio. With this dataset, we train state-of-the-art classifiers to determine how close existing methods are to the theoretical optimum. The dataset and code are published alongside this paper, to serve as a benchmark task for future developments in jet tagging and in other areas of particle physics.

As a numerical example, we use the JetClass dataset from Ref. [6,18] to train our generative model. Top quarks almost always decay to a b quark and a W boson [19]. When the top quark is highly Lorentz-boosted in the lab frame, its decay products are collimated and collected within a single jet. In the all-hadronic channel, the W boson decays into quarks and so top quark jets tend to have 3-prong substructure. In contrast, generic quark and gluon jets typically have much less structure. As boosted top quarks are a signature of many new physics scenarios, they serve as an important test bed for new methods. In the JetClass dataset, each jet is represented by its constituents: reconstructed particles. These jets are generated with MadGraph5_aMC@NLO [20], and Pythia 8 [21,22] is used for showering and hadronization. Delphes 3 [23–25] and the CMS detector card provides the detector simulation and reconstruction based on FastJet [26]. The dataset includes the 4-momenta of all constituents within each jet. Note that the number of constituents can vary significantly between jets, with an average of about 30 to 50, depending on the jet type. There are 10M boosted top quark jets and 10M generic quark and gluon (henceforth, QCD) jets.

We train a generative neural network to mimic the QCD and top quark JetClass datasets. We initially explored the diffusion model from Ref. [16], but the probability density is only known approximately. Current methods for estimating the density do not have the precision required to guarantee that the likelihood ratio is the optimal classifier. Instead, we employ the autoregressive model of Ref. [17], which provides the exact probability density of each jet. In this model, the momenta of particles are discretized, ordered in transverse momentum, and then the setup is cast as a natural language problem. The phase space is finite and each jet can be described by its probability, which is itself a product over conditional probabilities from the constituents parameterized by a transformer [27]. Generating jets proceeds by sampling from these conditional probabilities in serial. Previous work has shown that discretization effects are negligible for the classification task, and the resulting model is an accurate representation of the discretized training dataset. While there will be some differences between the surrogate model and the physics model, we believe that these effects will not qualitatively change the conclusions. The physics model itself is only an approximation to real jet data—we are

most concerned that the surrogate model has a similar level of complexity and expressivity as the real model.¹

With this new synthetic dataset, we can now directly compare the performance of various taggers to the optimal performance given by the likelihood ratio. To compare classifier performances we use receiver-operating characteristic (ROC) curves. These curves show the tradeoff between signal (top tagging) efficiency and background rejection (inverse QCD jet efficiency). One tagger is uniformly better than another if all of its working points are up and to the right in the plot.

We fix the size of the classifier training dataset to 10M boosted top quark jets and 10M QCD jets and compare different neural network architectures, similar to the widely used benchmark comparison of Ref. [3], only now we know the optimal solution. Since that original work, there have been many innovations in jet tagging and we consider a selection of state-of-the-art and/or widely used models including ParticleNet [5], Deep Sets [28] without IRC safety (particle flow networks [29]) and with attention (deep sets with attention [27]); Lorentz-equivariant models such as LorentzNet [30] and Pelican [31]; transformers such as ParT [6] and OmniLearn (trained from scratch) [32] and a classifier based on the transformer architecture of Ref. [17], where we replace the output layer with a binary cross entropy layer (baseline transformer); and the fine-tuned OmniLearn foundation model [32]. In all cases, we have used the settings described in the corresponding papers without additional hyperparameter optimization. Our results are displayed in Fig. 1. All taggers show comparable performance, which is also similar to the performance achieved on the datasets of Refs. [3,6]. However, even the most powerful current taggers fall well short of optimal performance; for example, with a top-tagging efficiency of 0.5, the taggers exhibit a background rejection of at most about 550, whereas optimal performance corresponds to a background rejection of about 3300.

To analyze our results in more depth, we first investigate how the performance of the taggers varies with the size of the training dataset. To do this, we compare the optimal performance with the ROC curves for a selected tagger trained on training datasets of different sizes, from 100 to 10M events. The results are shown in Fig. 2 for the classifier based on the transformer architecture of Ref. [17] (baseline transformer in Fig. 1). Of course, the performance of the tagger initially improves with the size of the training dataset, but the improvement from 1M to 10M training events is rather limited. This means that even with

¹We have trained the autoregressive transformer model of Ref. [17] on 10M QCD and 10M top quark jets from the JetClass dataset and verified that a classifier cannot easily distinguish jets generated by our model from the discretized training jets. The extended training data increases the similarity between the generated and training datasets, resulting in even worse classifier performance compared to the results in Ref. [17].

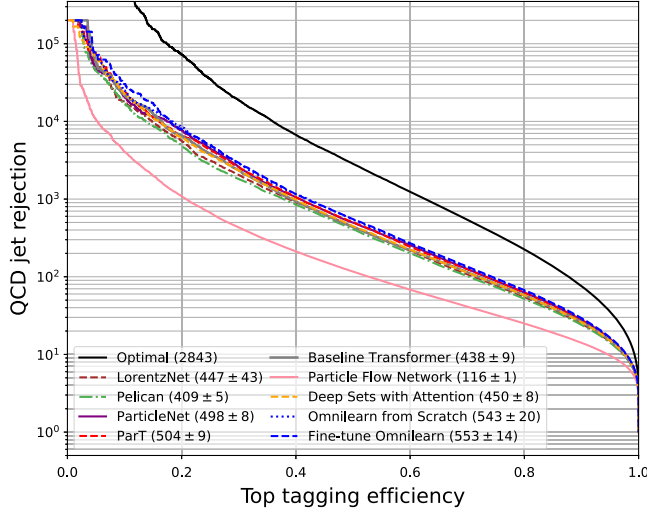


FIG. 1. ROC curves for top quark versus QCD jet classification. The performance of the likelihood ratio classifier is labeled “optimal” and the other curves correspond to different classifiers trained with 10M jets of each class. The parentheses in the legend indicate the rejection (inverse false positive rate) at a top quark efficiency of 50% and the corresponding error from training 5 classifiers independently.

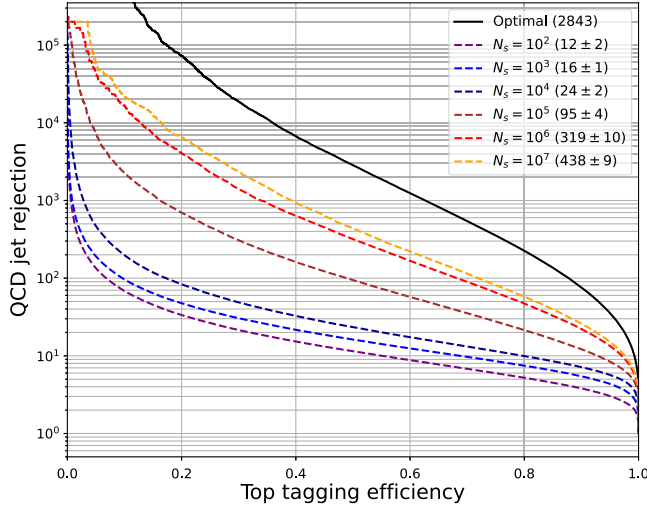


FIG. 2. ROC curves for top quark versus QCD jet classification. The performance of the likelihood ratio classifier is labeled “optimal” and the other curves correspond to the baseline transformer classifier trained on datasets with N_s training jets for each class. The parentheses in the legend indicate the rejection (inverse false positive rate) at a top quark efficiency of 50% and the corresponding error from training 5 classifiers independently.

significantly larger training data sets, optimal performance cannot be expected (for fixed architecture).

Our results show that even the most advanced top tagger fails to capture the complexity of jets required for optimal classification performance. To investigate this further, we compare the performance of the baseline transformer tagger

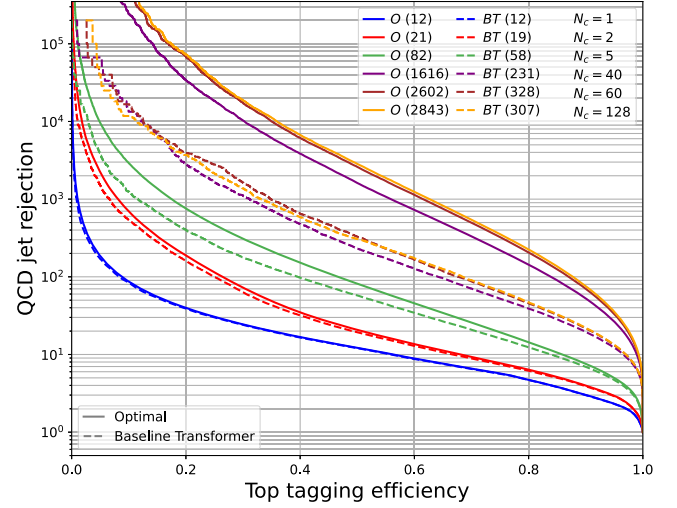


FIG. 3. ROC curves for top quark versus QCD jet classification for decreasing number of jet constituents N_c . The performance of the likelihood ratio classifier is shown as solid lines, while the dashed lines correspond to the baseline transformer classifier. The parentheses in the legend indicate the rejection (inverse false positive rate) at a top quark efficiency of 50%. Note that the dashed lines in this figure correspond to a baseline transformer trained on 1M events.

with optimal performance for simpler jets by analyzing only the first N_c constituents as generated by our autoregressive transformer. For $N_c = 1$ and $N_c = 2$, the low complexity allows any classifier to achieve optimal performance, as shown by the ROC curve in Fig. 3. As N_c increases, jets are classified more efficiently due to more information about their structure. However, when N_c exceeds 5, the classifier performance starts to fall below the optimal performance, suggesting that not all relevant information from additional particles can be extracted. Although specific performance may vary, this trend is expected to hold for any current classifier. Beyond about $N_c = 60$, the optimal performance plateaus, suggesting that additional particles lack information specific to QCD or top jets. This limitation may be due to the limited ability of the generative model to learn jet type specific correlations involving very soft constituents.

In conclusion, we investigated whether state-of-the-art jet taggers are close to optimal or whether there is room for further improvement. Using modern generative models, we created a realistic synthetic dataset for top and QCD jets with known optimal tagging performance. By applying several advanced taggers to this dataset, we found a significant gap between their performance and the optimum. This gap cannot be closed by training on larger datasets and persists because the taggers do not fully capture the complexity of the jets.

We have made our dataset and software publicly available to serve as a benchmark for future developments in jet tagging and other areas of particle physics.

Acknowledgements. We would like to thank Thorben Finke for his contribution in the early stages of this work. B. N. thanks Lester Mackey, Ariel Schwartzman, and the Stanford Data Science Initiative (circa 2014) where we asked the same question as in this paper, but without the right tools to provide a sharp answer. N.G., V.M., and B.N. are supported by the U.S. Department of Energy (DOE), Office of Science under Contract No. DE-AC02-05CH11231. The research of M. K., A. M., and H.R-G. is supported by the German Research Foundation DFG under Grant No. 396021762–TRR 257: Particle physics phenomenology after the Higgs discovery. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC awards No. HEP-ERCAP0021099 and No. HEP-ERCAP0028249, and computing resources provided by RWTH Aachen University under the project rwth0934. This work was performed in part at the Aspen Center for Physics, supported by National Science Foundation Grant No. PHY-2210452.

Data availability. The data that support the findings of this article are openly available [33], embargo periods may apply.

Appendix: Towards model inference. While the focus of this paper is on the development of a method and the creation of a dataset for determining how close a given classifier is to optimal, we have also have some hints about the aspects that lead to suboptimality. In the main body, we showed that the performance matched the optimal classifier for low multiplicity and degraded with higher-dimensional phase spaces. Furthermore, we showed that adding more training data improved the performance. Here, we briefly comment on the event-by-event nature of the optimal performance to further probe the suboptimality. Classifiers trained with a given loss function can be recast into likelihood-ratio estimators

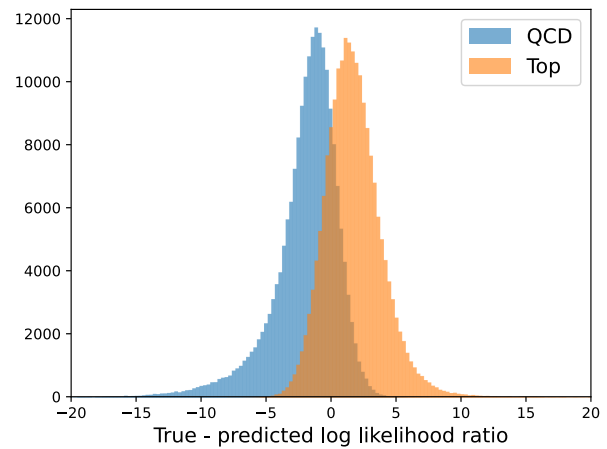


FIG. 4. Event-by-event differences between the true log likelihood ratio and the one inferred by the BT classifier for QCD jets (blue) and top jets (orange).

[34,35]. For example, if a classifier f is trained with the binary cross entropy and balanced classes, then $f/(1-f)$ is an approximation of the likelihood ratio. Figure 4 compares this estimate of the likelihood ratio with the true value, event-by-event for the baseline transformer (BT). For QCD jets, we observe that the peak of the distribution is shifted to the left (over-estimated likelihood ratio) and has a long tail to the left. In contrast, top jets are shifted to the right (underestimated likelihood ratio), but are relatively symmetric about the peak. Increasing the size of the training dataset shrinks the width of these distributions, but does not change the skewness for QCD jets. While we do not find any correlation between the heavy QCD tail and the constituent multiplicity or jet mass, this serves as a starting point for further studies that would seek to identify the physical attributes of the jets that are particularly difficult for the machine learning models to capture. This information may, in turn, be useful for modifying the architectures or training protocols to improve performance.

[1] A. J. Larkoski, I. Moulton, and B. Nachman, *Phys. Rep.* **841**, 1 (2020).
 [2] R. Kogler *et al.*, *Rev. Mod. Phys.* **91**, 045003 (2019).
 [3] A. Butter *et al.*, *SciPost Phys.* **7**, 014 (2019).
 [4] A. M. Sirunyan *et al.* (CMS Collaboration), *J. Instrum.* **15**, P06005 (2020).
 [5] H. Qu and L. Gouskos, *Phys. Rev. D* **101**, 056019 (2020).
 [6] H. Qu, C. Li, and S. Qian, *Proceedings of the 39th International Conference on Machine Learning, PMLR* (2022), [arXiv:2202.03772](https://arxiv.org/abs/2202.03772).

[7] G. Aad *et al.* (ATLAS Collaboration), *J. Instrum.* **19**, P08018 (2024).
 [8] J. Neyman and E. S. Pearson, *Phil. Trans. R. Soc. A* **231**, 289 (1933).
 [9] L. de Oliveira, M. Paganini, and B. Nachman, *Comput. Software Big Sci.* **1**, 4 (2017).
 [10] S. Badger *et al.*, *SciPost Phys.* **14**, 079 (2022).
 [11] A. Adelmann *et al.*, in 2022 Snowmass Summer Study (2022), [arXiv:2203.08806](https://arxiv.org/abs/2203.08806).
 [12] R. Verheyen, *SciPost Phys.* **13**, 047 (2022).

- [13] B. Käch, D. Krücker, I. Melzer-Pellmann, M. Scham, S. Schnake, and A. Verney-Provatas, [arXiv:2211.13630](#).
- [14] B. Käch, D. Krücker, and I. Melzer-Pellmann, [arXiv:2211.13623](#).
- [15] S. Schnake, D. Krücker, and K. Borras, [arXiv:2403.15782](#).
- [16] V. Mikuni and B. Nachman, *SciPost Phys.* **16**, 062 (2024).
- [17] T. Finke, M. Krämer, A. Mück, and J. Tönshoff, *J. High Energy Phys.* **06** (2023) 184.
- [18] H. Qu, C. Li, and S. Qian, JetClass: A large-scale dataset for deep learning in jet physics (Zenodo, 2022), [10.5281/zenodo.6619768](#).
- [19] M. Tanabashi *et al.* (Particle Data Group), *Phys. Rev. D* **98**, 030001 (2018).
- [20] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, *J. High Energy Phys.* **07** (2014) 079.
- [21] T. Sjöstrand, S. Mrenna, and P. Z. Skands, *J. High Energy Phys.* **05** (2006) 026.
- [22] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, *Comput. Phys. Commun.* **191**, 159 (2015).
- [23] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), *J. High Energy Phys.* **02** (2014) 057.
- [24] A. Mertens, *J. Phys. Conf. Ser.* **608**, 012045 (2015).
- [25] M. Selvaggi, *J. Phys. Conf. Ser.* **523**, 012033 (2014).
- [26] M. Cacciari, G. P. Salam, and G. Soyez, *Eur. Phys. J. C* **72**, 1896 (2012).
- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, [arXiv:1706.03762](#).
- [28] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. Salakhutdinov, and A. J. Smola, in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 2017, Long Beach, CA, USA* (Curran Associates Inc., Red Hook, 2017), pp. 3391–3401.
- [29] P. T. Komiske, E. M. Metodiev, and J. Thaler, *J. High Energy Phys.* **01** (2019) 121.
- [30] S. Gong, Q. Meng, J. Zhang, H. Qu, C. Li, S. Qian, W. Du, Z.-M. Ma, and T.-Y. Liu, *J. High Energy Phys.* **07** (2022) 030.
- [31] A. Bogatskiy, T. Hoffman, D. W. Miller, and J. T. Offermann, [arXiv:2211.00454](#).
- [32] V. Mikuni and B. Nachman, *Phys. Rev. D* **111**, L051504 (2025).
- [33] H. Reyes-González, J. Geuskens, N. Gite, M. Krämer, V. Mikuni, A. Mück, and B. Nachman, The fundamental limit of jet tagging dataset (Zenodo, 2024), [10.5281/zenodo.14023638](#).
- [34] K. Cranmer, J. Pavez, and G. Louppe, [arXiv:1506.02169](#).
- [35] B. Nachman and J. Thaler, *Phys. Rev. D* **103**, 116013 (2021).