Why is Attention Sparse in Particle Transformer?

Timothy Legge[2] * **Aaron Wang**[1] † **Jacob Ortiz**[2] **Victor Limouzi**[2] **Zihan Zhao**[2]

Abhijith Gandrakota[3] **Elham E. Khoda**[2] **Jennifer Ngadiuba**[3] **Javier Duarte**[2]

Richard Cavanaugh[1]

[1]University of Illinois Chicago, Chicago, IL 60607
[2]University of California San Diego, La Jolla, CA 92093
[3]Fermi National Accelerator Laboratory, Batavia, IL 60510

Abstract

Transformer-based models have achieved state-of-the-art performance in jet tagging at the CERN Large Hadron Collider (LHC), with the Particle Transformer (ParT) representing a leading example of such models. A striking feature of ParT is its sparse, nearly binary attention structure, raising questions about the origin of this behavior and whether it encodes physically meaningful correlations. In this work, we investigate the source of ParT's sparse attention by comparing models trained on multiple benchmark datasets and examine the relative contributions of the attention term and the physics-inspired interaction matrix before softmax. We find that binary sparsity arises primarily from the attention mechanism itself, with the interaction matrix playing a secondary role. Moreover, we show that ParT is able to identify key jet substructure features, such as leptons in semileptonic top decays, even without explicit particle identification inputs. These results provide new insight into the interpretability of transformer-based jet taggers and clarify the conditions under which sparse attention patterns emerge in ParT. Our code is available here.

1 Introduction

Machine learning (ML) has become essential for analyzing the massive datasets from the CERN Large Hadron Collider (LHC) [1]. Jet tagging, or the classification of collimated streams of particles that result from high-energy collisions at LHC, has proven to be a task highly suitable to ML. Among the most promising approaches to jet tagging are graph neural networks (GNNs) and transformers, which capture correlations among particles in jets. GNNs, such as ParticleNet [2], reached state-of-the-art jet tagging by treating jets as unordered particle clouds [3, 4]. More recently, transformer models [5], which have transformed language and vision processing [6–8], have been adapted to collider physics, where their ability to model long-range dependencies provides a powerful tool for identifying jet substructure signatures [9–12].

A leading example is the Particle Transformer (ParT) [13], which augments multi-head attention with physics informed pairwise particle interaction features. This has made ParT one of the most effective jet taggers to date.

---

*Corresponding Author: tlegge@ucsd.edu
†Corresponding Author: aaronw5@uic.edu

Interpretability is crucial in high-energy physics to ensure ML models capture physically meaningful correlations rather than dataset-specific artifacts. Attention provides a natural interpretability handle: particle–particle attention maps can reveal what transformers learn [14]. Understanding when and why attention becomes sparse may further guide model optimization, clarifying the role of physics-inspired interactions and suggesting opportunities to simplify architectures and make them more efficient.

2 Related Work

ML methods, particularly graph neural networks (GNNs), have been widely applied to particle physics problems at the CERN LHC [3, 4]. ParticleNet [2], a GNN designed for jet tagging, achieved state-of-the-art performance by treating jets as particle clouds. Recent studies have also interpreted such models. For example, Mokhtar *et al.* [15] employed layerwise relevance propagation to construct edge relevancy, uncovering how ParticleNet distinguishes three-prong hadronic top quark decays. These efforts highlight the growing emphasis on explainability in ML for high-energy physics.

Building on this line of work, attention-based transformers have emerged as powerful alternatives to GNNs for jet tagging. ParT [16, 17] extends the standard attention mechanism with physics-inspired pairwise particle interactions. While delivering excellent performance, ParT also raises important interpretability questions. Wang *et al.* [17] analyzed ParT's attention maps and particle-pair correlations in the η–ϕ plane, revealing a sparse, nearly binary attention pattern where each particle focuses on at most one other particle. Their subjet-level studies further showed that ParT attends to leptons in semileptonic top decays and captures both intra- and inter-subjet structures in hadronic top and Higgs decays, suggesting that the model learns physically meaningful jet substructure.

In line with these efforts, our work investigates the origin of ParT's sparse attention by systematically comparing datasets and quantifying the roles of traditional attention and the physics-inspired interaction matrix.

3 Datasets and Model

This method compares the attention heatmaps and interpretability of multiple different datasets. We use the JETCLASS dataset [13], the Top Landscape dataset [18], and the Quark–Gluon dataset as our benchmark datasets for comparative studies.

Datasets We evaluate models on three widely used jet-tagging benchmarks. The JETCLASS dataset [19] comprises 100M training, 5M validation, and 20M test jets across 10 classes from quarks, gluons, and decays of W, Z, Higgs, and top quarks (q/g, $H \to b\bar{b}$, $H \to c\bar{c}$, $H \to gg$, $H \to 4q$, $H \to \ell\nu qq'$, $t \to bqq'$, $t \to b\ell\nu$, $W \to qq'$, $Z \to q\bar{q}$). Each jet contains up to 128 particles (avg. 30–50) with 17 features spanning kinematics, PID, and displacement. The Top Landscape Tagging benchmark [20] provides 2M jets (1.2M train, 0.4M val/test) distinguishing $t \to bqq'$ decays from light-quark/gluon backgrounds, described only by constituent four-vectors with 7 kinematic features. Finally, the Quark–Gluon Tagging benchmark [21] consists of 2M jets (1.6M train, 0.2M val/test) for discriminating quark- from gluon-initiated jets, using both kinematic and PID information.

Model Architecture ParT is a state-of-the-art jet tagging model, which introduces a modified attention mechanism, called particle multi-head attention (P-MHA). Given a particle-level representation of a jet with N constituents, $x \in \mathbb{R}^{N \times d}$, the attention is computed as

$$\text{P-MHA}(x) = \text{concat}(\text{head}_1, \dots, \text{head}_h)W^O$$
$$\text{head}_i = \text{softmax}\left(\frac{xW_i^W(xW_i^K)^\mathsf{T}}{\sqrt{d_k}} + U_i\right)xW_i^V, \tag{1}$$

where $U_i \in \mathbb{R}^{N \times N}$ encodes pairwise particle interactions. These are learned from a set of kinematic features $(\ln \Delta, \ln k_\text{T}, \ln z, \ln m^2)$ described in Refs. [22, 13] that are encoded through convolutional layers. We use pretrained ParT weights from Ref. [13] for models on the JETCLASS dataset, and train from scratch on the Top and Quark–Gluon datasets.
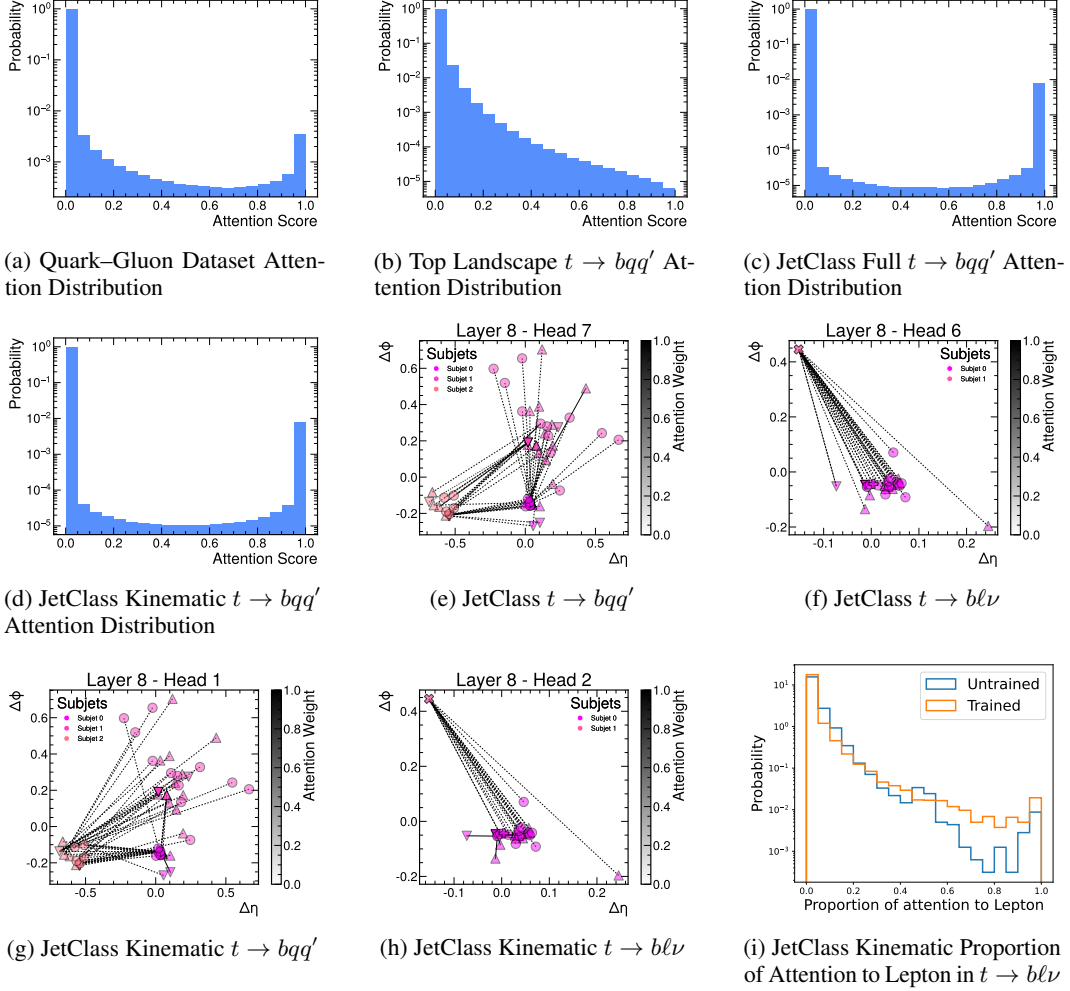
(a) Quark–Gluon Dataset Attention Distribution

(b) Top Landscape $t \to bqq'$ Attention Distribution

(c) JetClass Full $t \to bqq'$ Attention Distribution

(d) JetClass Kinematic $t \to bqq'$ Attention Distribution

(e) JetClass $t \to bqq'$

(f) JetClass $t \to b\ell\nu$

(g) JetClass Kinematic $t \to bqq'$

(h) JetClass Kinematic $t \to b\ell\nu$

(i) JetClass Kinematic Proportion of Attention to Lepton in $t \to b\ell\nu$

Figure 1: Comparison of ParT attention distributions across datasets (a–d), η–ϕ particle attention maps (e–h), and the distribution of attention scores to the lepton in $t \to b\ell\nu$ with ParT trained on JetClass Kinematic (i). η–ϕ maps use pre-softmax attention values without the interaction matrix included. Symbols: (✖: muon, ▲: charged hadron, ▼: neutral hadron, ●: photon, ✚: electron). Transparency scales with particle p_{T}, while line intensity reflects attention scores.

4 Results

Distribution of Attention Scores   We visualize the distribution of attention scores across all heads for models trained on the full JETCLASS feature set and on JETCLASS with only kinematic features, as shown in images (a – d) of Figure 1. For both cases, as well as for the Quark–Gluon dataset, ParT produces a very sparse, nearly binary attention map. In contrast, the Top Landscape dataset does not exhibit this binary behavior. Since the Top Landscape dataset contains only kinematic information and no particle identification (PID) features, we perform a direct comparison with the hadronic top ($t \to bqq'$) class in JETCLASS using kinematic features only. We find that the attention distribution in JETCLASS remains binary even without PID inputs, demonstrating that the binary nature of attention, as reported in previous work, does not rely on the presence of PID features [17].

Interaction and Attention Score Distribution Before Softmax   We compute the ratio between attention matrix values and their corresponding interaction matrix values, examining the resulting distribution in figure 2. For ParT trained on JETCLASS, the attention term dominates the interaction matrix, indicating that it is the primary contributor to model performance. This shows that, even without the additional physics-motivated bias introduced by interaction, the attention mechanism

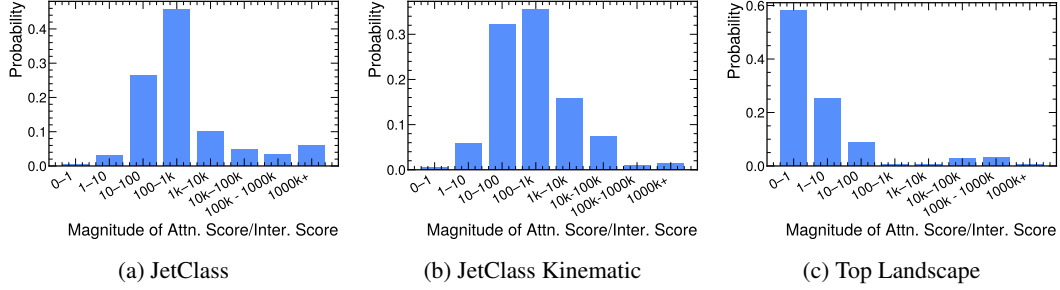| (a) JetClass | (b) JetClass Kinematic | (c) Top Landscape |

Figure 2: Distribution of the magnitude of pre-softmax attention value divided by pre-softmax interaction matrix value. For both JetClass models, the ratio is almost always greater than one, and often 10^4–10^5 times bigger than the interaction matrix, meaning it completely dominates.

alone is capable of identifying particles of interest, such as conventional jet substructure observables. In fact, despite benchmarks reported in [13], our results suggest that the interaction matrix may not be essential for jet tagging. Given that the interaction matrix is computationally expensive, removing it could provide significant resource savings. By contrast, in the Top Landscape dataset, the interaction matrix and pre-softmax attention scores have comparable distributions, demonstrating that both terms contribute meaningfully to the overall attention.

Particle Attention Graphs   To further investigate the origin of sparsity, we decluster each jet into a fixed number of subjets using the k_T algorithm [23] as implemented in the FastJet package [24] with Python bindings, following the approach of Ref. [17]. Specifically, we cluster into two subjets for leptonic top decays ($t \to b\ell\nu$), and three subjets for hadronic top decays ($t \to bqq'$). For each process, we visualize the attention maps by representing particles as points in the η–ϕ plane. Our analysis focuses on the attention heads in the final layer of P-MHA. Importantly, these visualizations are constructed using the *pre-softmax attention values only*, isolating the contribution of the attention mechanism without the interaction matrix, and are compared with the corresponding maps in Ref. [**?**].

In Fig. 1 (e–h), we find that even without the interaction matrix, the model is able to recover the same jet substructure observables. In addition, Figs. 1g and 1h demonstrate that the Particle Transformer does not require particle identification information to attend to the lepton in leptonic top decays ($t \to b\ell\nu$) and is able to find subjet connections without PID or interaction matrix. Models trained with and without PID inputs both successfully recognize the lepton and attend to it. This demonstrates that jet substructure information is already learned in pre-softmax attention, and that explicit particle identification is not necessary for ParT to identify leptons. In Fig. 1i, we demonstrate that even without PID, in $t \to b\ell\nu$ decays, JetClass Kinematic similarly attends to the lepton more often than an untrained model, signifying that it recognizes the lepton as an important particle index.

Ablation Study   The preceding findings, particularly the overwhelmingly larger magnitudes observed for traditional attention scores, suggest the possibility that the ParT model as trained may not require the interaction matrix simply because its influence becomes minuscule in softmax computation. To test this, we conduct an ablation study of ParT with interaction removed, referred to as ParT (Zero U), with results shown in Table 1. Specifically, we implement this by setting all parameters in the pretrained model's PAIREMBED module to 0 and subsequently running inference on the JETCLASS test set. This approach to ablation differs from Ref. [13], wherein the model itself is trained without interaction and achieves metrics similar to the full architecture.

We find that ParT (Zero U) exhibits drastically reduced performance than full ParT and the ablated ParT (plain) from Ref. [13] (whose accuracies are 0.861 and 0.849 respectively). Although this is not surprising from a surface-level ML perspective—as manually modifying a significant number of weights necessarily destroys learned information—in this case it is remarkable the degree to which performance is impacted, considering the relatively minuscule scale of interaction scores.

To explain the performance difference and assess the impact of interaction at inference time, we investigate instances where the interaction matrix has significant influence in the softmax computation. As a metric, we choose to count instances where adding the interaction scores U to traditional

Table 1: Performance of ParT with zeroed-out interaction when run on JETCLASS test set (20M jets). The $\text{Rej}_{X\%}$ metric is the inverse of FPR at the threshold where TPR is $X\%$.

| | All classes | | $H \to b\bar{b}$ | $H \to c\bar{c}$ | $H \to gg$ | $H \to 4q$ | $H \to \ell\nu qq'$ | $t \to bqq'$ | $t \to b\ell\nu$ | $W \to qq'$ | $Z \to q\bar{q}$ |
	Accuracy	AUC	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{99.5\%}$	$\text{Rej}_{50\%}$	$\text{Rej}_{50\%}$
ParT (Zero U)	0.405	0.8974	15.0	8.81	19.9	5.53	3.03	79.5	2.69	25.6	11.8

Table 2: Prevalence of non-binary and interaction-dependent computations when U is included, as well as the PCC between them.

Type of complication	% Incidence overall	% of tokens to encounter	PCC
Interaction-Dependence	3.6	85.4	
Non-binary Computation	0.88	42.1	0.229

attention A alters the index of the highest-weighted particle going in to the softmax computation. In mathematical terms, the criterion is that $\text{argmax}(A_j + U_j) \neq \text{argmax}(A_j)$ where j indexes over all particles/tokens in the jet. These instances are referred to as interaction-dependent computations.

As a supplement to the study of interaction-dependence, we also track incidence of non-binary attention computation, which here is quantified as an instance where the largest post-softmax weight is < 0.8. We track both overall incidence (the proportion of token updates to be non-binary or interaction-dependent computations) as well as the proportion of tokens to encounter a non-binary or interaction-dependent computation in any P-MHA head during inference. In order to assess if interaction-dependent computations are correlated with non-binary computations, we sort their incidence by head and compute the Pearson correlation coefficient (PCC) between them. Our results are collected in Table reftab:nb-and-interaction-dependent-table.

Our results indicate that the vast majority of tokens encounter interaction-dependent computation at some point within the model, meaning that most token embeddings are affected by interaction. This explains the weak performance observed in Table 1. Our supplementary study of non-binary computation demonstrates that it is a less common process (though 42.1% of tokens still encounter it at some point) and that there is no strong correlation between the occurrence of interaction-dependent vs. non-binary computations. From these results, we conclude that while traditional attention is responsible for the highly binary character of post-softmax weights, the inverse is not true: the interaction features are not responsible for the trace non-binary computations.

5 Summary and Outlook

The sparse, nearly binary attention observed in ParT emerges primarily from the attention mechanism itself rather than from the physics-inspired interaction matrix. However, it is also shown that interaction contributes as-of-yet unidentified biases which are necessary for full performance. This result clarifies the role of inductive biases in ParT and strengthens confidence that its learned patterns capture genuine jet substructure. Beyond interpretability, understanding the origin of sparsity also points to opportunities for improving efficiency, for example by adopting top-k attention mechanisms that reduce computational cost without degrading performance.

Limitations Our study focuses mainly on the distributions of attention weights and their relative magnitudes, providing only a partial view of how sparsity relates to underlying physics. As emphasized in prior work, attention maps should not be interpreted as complete explanations of model behavior [25]. Further analyses will be needed to connect attention patterns more directly to physical observables. In addition, our results are limited to a fixed set of benchmark datasets and only one size of model architecture. Studying a broader range of physics processes and experimental conditions would be necessary to fully assess the universality of sparse attention.

Future Work Having shown that the interaction matrix cannot be fully disabled at inference time without drastic performance impacts, future work may study the exact ways in which interaction influences the model, for example by searching for trends in the particles which interaction assigns high scores to. If meaningful findings are made in this line of investigation, they could lead to new implemented physics biases that replicate interaction's behavior but operate at a lesser computational cost. Broadly, further work in interpretability may motivate physics-informed training strategies that retain performance while enabling leaner architectures.

Acknowledgments and Disclosure of Funding

References

[1] P. Harris et al., "Physics Community Needs, Tools, and Resources for Machine Learning", in *Snowmass 2021*. 2022. `arXiv:2203.16255`.

[2] H. Qu and L. Gouskos, "ParticleNet: Jet tagging via particle clouds", *Phys. Rev. D* **101** (2020) 056019, `doi:10.1103/PhysRevD.101.056019`, `arXiv:1902.08570`.

[3] J. Shlomi, P. Battaglia, and J.-R. Vlimant, "Graph Neural Networks in Particle Physics", *Mach. Learn.: Sci Technol.* **2** (2021) 021001, `doi:10.1088/2632-2153/abbf9a`, `arXiv:2007.13681`.

[4] J. Duarte and J.-R. Vlimant, "Graph neural networks for particle tracking and reconstruction", in *Artificial Intelligence for High Energy Physics*, P. Calafiura, D. Rousseau, and K. Terao, eds., p. 387. World Scientific, 2022. `arXiv:2012.01249`. `doi:10.1142/9789811234033_0012`.

[5] A. Vaswani et al., "Attention is all you need", in *Advances in Neural Information Processing Systems*, I. Guyon et al., eds., volume 30. Curran Associates, Inc., 2017. `arXiv:1706.03762`.

[6] A. Ramesh et al., "Hierarchical text-conditional image generation with clip latents", 2022. `arXiv:2204.06125`.

[7] Gemini Team, "Gemini: A family of highly capable multimodal models", 2024. `arXiv:2312.11805`.

[8] T. Brown et al., "Language models are few-shot learners", in *Advances in Neural Information Processing Systems*, H. Larochelle et al., eds., volume 33, p. 1877. Curran Associates, Inc., 2020. `arXiv:2005.14165`.

[9] CMS Collaboration, "Search for highly energetic double Higgs boson production in the two bottom quark and two vector boson all-hadronic final state", CMS Physics Analysis Summary CMS-PAS-HIG-23-012, 2024.

[10] CMS Collaboration, "Particle transformers for identifying Lorentz-boosted Higgs bosons decaying to a pair of W bosons", CMS Physics Analysis Summary CMS-PAS-JME-25-001, 8, 2025.

[11] CMS Collaboration, "Search for Higgs boson production at high transverse momentum in the WW^* decay channel in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$", CMS Physics Analysis Summary CMS-PAS-HIG-24-008, 2025.

[12] CMS Collaboration, "Improved results on Higgs boson pair production in the 4b final state", CMS Physics Analysis Summary CMS-PAS-HIG-24-010, 2025.

[13] H. Qu, C. Li, and S. Qian, "Particle Transformer for Jet Tagging", in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri et al., eds., volume 162, p. 18281. 2022. `arXiv:2202.03772`.

[14] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 782. 2021.

[15] F. Mokhtar, R. Kansal, and J. Duarte, "Do graph neural networks learn traditional jet substructure?", in *Machine Learning and the Physical Sciences Workshop at the 36th Conference on Neural Information Processing Systems*. 2022. `arXiv:2211.09912`.

[16] H. Qu and L. Gouskos, "Jet tagging via particle clouds", *Phys. Rev. D* **101** (2020) `doi:10.1103/physrevd.101.056019`, `arXiv:1902.08570`.

[17] A. Wang et al., "Interpreting and Accelerating Transformers for Jet Tagging", in *7th Machine Learning and the Physical Sciences Workshop at the 38th Conference on Neural Information Processing Systems*. 2024. `arXiv:2412.03673`.

[18] G. Kasieczka et al., "The Machine Learning landscape of top taggers", *SciPost Phys.* **7** (2019) 014, `doi:10.21468/SciPostPhys.7.1.014`, `arXiv:1902.09914`.

[19] H. Qu, C. Li, and S. Qian, "JETCLASS: A large-scale dataset for deep learning in jet physics", 2022. `doi:10.5281/zenodo.6619768`.

[20] G. Kasieczka, T. Plehn, J. Thompson, and M. Russel, "Top quark tagging reference dataset (v0 (2018_03_27))", 2019. `doi:10.5281/zenodo.2603256`.

[21] P. Komiske, E. Metodiev, and J. Thaler, "Pythia8 quark and gluon jets for energy flow", 2019. `doi:10.5281/zenodo.3164691`.

[22] F. A. Dreyer and H. Qu, "Jet tagging in the Lund plane with graph networks", *JHEP* **03** (2021) 052, `doi:10.1007/JHEP03(2021)052`, `arXiv:2012.08526`.

[23] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, "Longitudinally invariant k_{T} clustering algorithms for hadron hadron collisions", *Nucl. Phys. B* **406** (1993) 187, `doi:10.1016/0550-3213(93)90166-M`.

[24] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual", *Eur. Phys. J. C* **72** (2012) 1896, `doi:10.1140/epjc/s10052-012-1896-2`, `arXiv:1111.6097`.

[25] S. Jain and B. C. Wallace, "Attention is not explanation", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, eds., p. 3543. 2019. `arXiv:1902.10186`. `doi:10.18653/v1/N19-1357`.