PROCEEDINGS OF SCIENCE

FERMILAB-PUB-25-0779-CMS-LDRD

Low-latency Jet Tagging for HL-LHC Using Transformer Architectures

Lauri Laatu,[a,*] **Chang Sun,**[b] **Arianna Cox,**[a] **Abhijith Gandrakota,**[c] **Benedikt Maier,**[a] **Jennifer Ngadiuba,**[c] **Zhiqiang Que,**[a] **Wayne Luk,**[a] **Maria Spiropulu**[b] **and Alexander Tapper**[a]

[a] *Imperial College London, United Kingdom*
[b] *California Institute of Technology, USA*
[c] *Fermilab, USA*

Transformers are the state-of-the-art model architectures and widely used in application areas of machine learning. However the performance of such architectures is less well explored in the ultra-low latency domains where deployment on FPGAs or ASICs is required. Such domains include the trigger and data acquisition systems of the LHC experiments.

We present a transformer-based algorithm for jet tagging built with the HGQ2 framework, which is able to produce a model with heterogeneous bitwidths for fast inference on FPGAs, as required in the trigger systems at the LHC experiments. The bitwidths are acquired during training by minimizing the total bit operations as an additional parameter. By allowing a bitwidth of zero, the model is pruned in-situ during training. Using this quantization-aware approach, our algorithm achieves state-of-the-art performance while also retaining permutation invariance which is a key property for particle physics applications.

Due to the strength of transformers in representation learning, our work also serves as a stepping stone for the development of a larger foundation model for trigger applications.

The European Physical Society Conference on High Energy Physics (EPS-HEP2025)
7-11 July 2025
Marseille, France

[*] Speaker

 ISSN 1824-8039 . Published by SISSA Medialab.

https://pos.sissa.it/

PoS(EPS-HEP2025)081

1. Introduction

The CERN Large Hadron Collider (LHC) [1] collides protons at a 40 MHz collision rate which results in hundreds of terabytes of data per second. To cope with this throughput, the experiments at the LHC: ATLAS [2] and CMS [3] rely on two-stage trigger systems for real-time event selection. The first stage—the Level-1 trigger (L1T)—uses a system of hundreds of Field-Programmable Gate Arrays (FPGAs) to identify interesting events in real time at 40 MHz. Because buffer sizes limit the system's end-to-end latency to just a few microseconds [4, 5], the L1T's precision is critical; any event it rejects is lost forever.

With the High-Luminosity LHC (HL-LHC) [6] upgrade, the simultaneous proton-proton collisions every 25 ns, known as pileup, is set to rise from 40–80 to 140–200, requiring extensive upgrades to trigger systems. A key strategy for the next generation of triggers involves incorporating deep learning algorithms as part of the L1T selection process. Recent research [7–14] focuses on deploying machine learning models on FPGAs, validating their potential for ultra-low latency environments. Of particular importance is jet tagging—identifying collimated sprays of particles associated with both the Standard Model and new physics particles. Although Ref. [8] demonstrated that quantization-aware training (QAT) could enable efficient, high-accuracy jet tagging on FPGAs, that study utilized high-level features typically inaccessible to the L1T. Subsequent studies [15–18], however, have confirmed the feasibility of using particle-level inputs to deploy practical jet-tagging networks within the strict resource constraints of the L1T.

Architectures based on Transformers and Graph Neural Networks (GNNs), including the Particle Transformer [19], the Lorentz-Equivariant Geometric Algebra Transformer [20], ParticleNet [21], and PELICAN [22], currently set the benchmark for offline jet tagging. However, a significant gap exists between offline performance and online feasibility: these models are orders of magnitude too resource-intensive to fit within the strict constraints of an FPGA-based real-time trigger system.

Recent advances in Large Language Models (LLMs) have shown that quantizing Transformers is an effective technique to reduce the model size and its computational footprint. Emerging architectures, such as BitNet b1.58 [23], have demonstrated that ternary quantization—where parameters are constrained to values of $\{-1, 0, 1\}$—can achieve perplexity and task performance competitive with full-precision models. In addition work done in [24] has shown the effectiveness of one bit quantization in high energy physics applications. By effectively replacing computationally expensive floating-point multiplications with integer additions, these binary or ternary formulations drastically reduce the required computational resources without compromising the model's ability to learn complex representations. This success in scaling of large models suggests that the representational capacity of Transformers is robust to extreme quantization, motivating the application of similar ternary strategies to strictly resource-constrained environments like the L1T. Binary and ternary networks have also been shown to work well in HEP tasks for MLP and Convolutional networks using `hls4ml` [9].

While there have been attempts to deploy Transformers on FPGAs for sub-microsecond latency applications, these approaches have not yet yielded performance competitive with alternative methods. Existing implementations either rely limitedly on high-level features [25] or fail to meet the necessary sub-microsecond latency constraints [26]. In this work, we present a ternary Transformer for jet tagging on an FPGA that achieves $\mathcal{O}(100)$ nanosecond latency with competitive

performance. Furthermore, we have included support for our Transformer in `hls4ml` library [27], ensuring availability to the broader fast machine learning community.

2. Method

2.1 Model Architectures

The Transformer used is based on encoder-only architecture using either one or two Transformer blocks with Multi-head Attention (MHA) [28]. To maximize hardware efficiency, we utilize ternary attention, quantizing both weights and data paths to ternary values $\{-1, 0, 1\}$. The inputs are sequences of the particles with the highest transverse momentum (p_T) of the jet with a maximum length of 8, 16, 32, or 64 sorted by p_T, with three features (p_T, η and ϕ). The absence of positional encodings renders the model a Set Transformer [29].

The proposed Transformer architecture is compared against Deep Sets [30] which operate on permutation-invariant particle sets. The two architectures differ fundamentally in how they model particle interactions. Deep Sets process each particle independently through a shared non-linear transformation before aggregating them, effectively modeling the event as a sum of individual contributions. This structure is computationally efficient ($O(N)$) but does not contain a message-passing mechanism on particle-level.

In contrast, the Transformer architecture explicitly models pairwise particle interactions through the self-attention mechanism which has the complexity $O(N^2)$. While the ternary constraint limits the resolution of the attention scores to coarse values, it allows the model to learn relevant particle-particle correlations resulting in better performance than the fixed global aggregation done only after the particle level processing of Deep Sets.

2.2 Model Compression

We employ High Granularity Quantization (HGQ) [31] for the quantization and pruning of the networks. By utilizing gradient-based, per-parameter bitwidth optimization—which includes the ability to assign zero bitwidth—this method performs simultaneous, loss-aware quantization and pruning. Consequently, this approach yields significantly smaller model sizes compared to standard layer-wise quantization methods, such as those provided by QKeras [8] and Brevitas [32], while maintaining the accuracy of the original model. The Multi-Head Attention is fixed to ternary for both weights and data path.

HGQ operates by introducing Effective Bit Operations (EBOPs) as a trainable metric, calculated by summing the products of the bitwidths ($a_{bw} \cdot b_{bw}$) across all operations in the network. By incorporating EBOPs into the loss function, the training process actively minimizes bitwidths alongside the base loss. Since EBOPs demonstrate a high correlation with FPGA resource consumption [31], it is possible to constrain the training to a target EBOPs budget, thereby tailoring the model to specific hardware limits. Additionally, we apply value-wise heterogeneous quantization to the data path; while this breaks the strict permutation invariance of the model it reduces the firmware footprint by allowing the allocation of higher bitwidths only for the most important particles.

In contrast to QKeras [8] and Brevitas [32], which restrict quantization to coarse-grained, layer-wise bitwidths, HGQ enables fine-grained compression by learning the bitwidth for every

parameter. The total loss function is defined as $\mathcal{L} = \mathcal{L}_{base} + \beta \cdot \mathcal{L}_{EBOPs} + \gamma \cdot \mathrm{L1}_{\mathrm{norm}}$. The parameter β controls the resulting quantization: a larger β enforces a stricter penalty on high-precision operations, aggressively driving the model toward lower bitwidths and higher sparsity, whereas a smaller β prioritizes accuracy over compression. An additional parameter γ for L1 regularization of the bitwidths is added to prevent the bitwidths from growing indefinitely. Unlike β the γ parameter is not adjusted during training.

In addition to HGQ, we use `da4ml` [33] to further optimize the constant-matrix-vector multiplication (CMVM) of the models. This optimization turns the weight multiplications to adder graphs, reducing resource usage and synthesis times while retaining the original accuracy.

3. Experiment

We evaluate our methods using the benchmark jet tagging dataset established in Ref. [34, 35], which is widely adopted within the high-energy physics community [15, 16, 18, 36–38]. The dataset consists of jets categorized by their originating particle: gluons (g), light quarks (q), W bosons (W), Z bosons (Z), and top quarks (t). It comprises 880,000 jets, split into a training set of 620,000 jets and a testing set of 260,000 jets, with balanced class distribution in both. For hardware implementation, we target the Xilinx XCU250 FPGA. All firmware implementations are done using Xilinx Vitis HLS and Vivado, with the source code integrated into the `hls4ml` library.

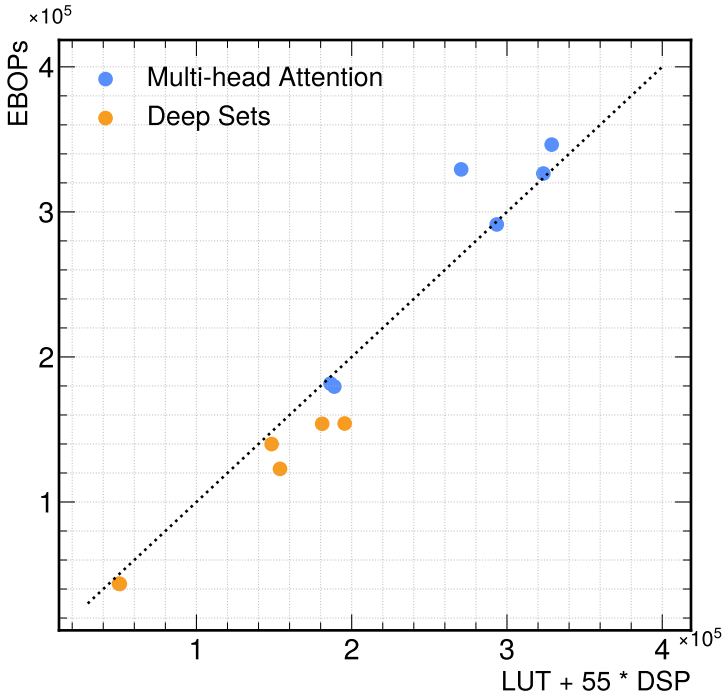

Figure 1: Correlation of EBOPs and resource usage on FPGA using the formula established in [31].

Our evaluation baselines include the MLP Mixer [18] trained with HGQ and Deep Sets variants trained with QKeras [37]. For the proposed Transformers and Deep Sets, we employ a training strategy to lock the model complexity to 350,000 EBOPs, roughly equivalent to the logic capacity of one Super Logic Region (SLR) on the XCU250. This is achieved by adjusting the regularization factor β with a PID controller, which automatically tunes the penalty strength to converge on the desired hardware footprint. Figure 1 shows how the EPOPs of the models correlate with resource usage in FPGAs where model EBOPs is ranging from 50,000 to 350,000 using the empirically established formula (LUT + 55 · DSP) from [31]. The resource usage of the best-performing models is shown in figure 2. Due to the fixed target EBOPs the HGQ trained models result in similar amount of Lookup Tables (LUTs) used. As the input length increases the bitwidths of the model go down to fit in the required EBOPs constraints. Operations with lower bitwidths are synthesized with logic rather than Digital Signal Processor (DSP) blocks by the HLS compiler, which explains why DSP usage goes down with longer input.

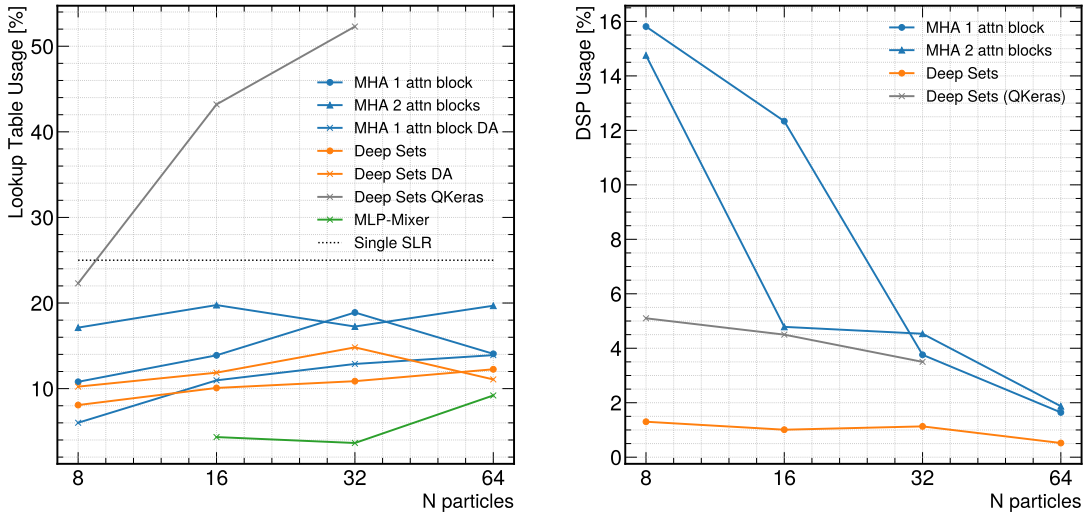

Figure 2: Lookup Table and DSP usage as a function of input length for Ternary Multi-Head Attention (MHA), Deep Sets, and MLP Mixer. Single Super Logic Region (SLR) shown as the limit of resources that are available. DSPs reported only for models with non-zero DSP usage.

Figure 3 illustrates the performance of the models as the input length increases. The figure shows performance for the MHA and Deep Sets models with and without distributed arithmetic. By using distributed arithmetic it is possible to fit a larger model, which results in better performance. While the ternary Attention mechanism delivers competitive results at lower input lengths, it fails to live up to the promise of state-of-the-art performance of its large counterparts when scaled down small enough to be deployed on the FPGA. The quadratic complexity of the self-attention mechanism creates a bottleneck at larger input lengths, necessitating severe quantization elsewhere in the model to satisfy the target resource budget. The Transformer model with two blocks is the worst performing attention based model when the input increases from 32 particles to 64 since it also has the most operations taken by the computation of the attention. The inefficient scaling of the MHA causes the accuracy of the Transformer models to degrade in comparison to the Deep Sets and the established

baselines as the number of particles increases. Notably, neither the ternary Transformer nor the Deep Sets achieve better performance than the MLP-Mixer [18] architecture.

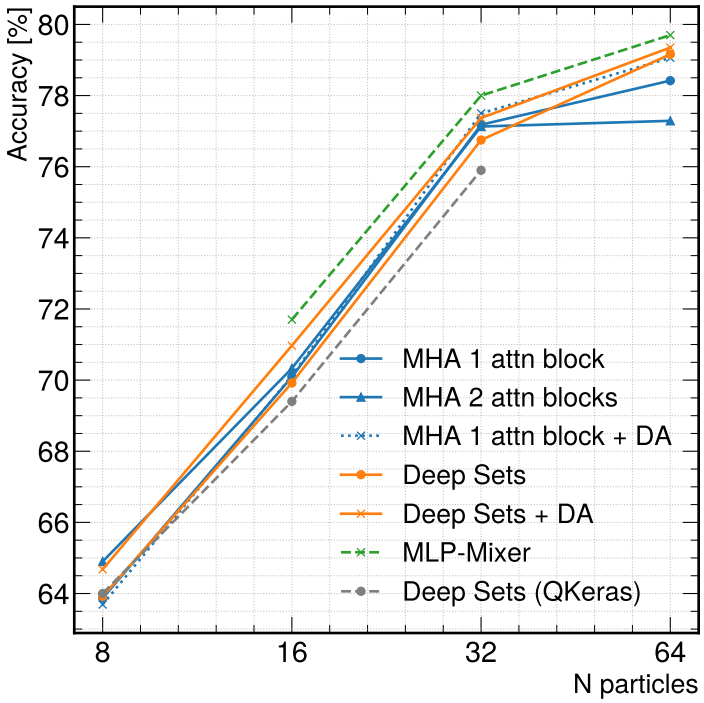

Figure 3: Five-class classification accuracy as a function of the maximum number of input particles for Ternary Multi-Head Attention (MHA), Deep Sets, and MLP Mixer. The decrease in accuracy for the full Transformer models compared to the other models at 64 particles is expected due to the fixed EBOPs budget enforced.

4. Conclusion and Future Work

In this work, we presented the first implementation of a fully ternary Multi-Head Attention (MHA) mechanism on FPGAs for real-time jet tagging. While ternary quantization has demonstrated remarkable success in the domain of Large Language Models (LLMs) and offline jet tagging, our results indicate that these benefits do not trivially transfer to the strictly resource-constrained environments of the Level-1 Trigger. Although the ternary MHA achieved competitive accuracy at lower sequence lengths, the quadratic complexity of the attention mechanism provided poor performance at longer input lengths. The resource contention caused by the MHA layer necessitated aggressive quantization in the rest of the network, ultimately preventing the model from matching the performance scaling of baseline architectures.

Crucially, however, we demonstrated the effectiveness of High Granularity Quantization (HGQ) in managing these constraints. We showed that HGQ enables the training of models with significant reductions in resource usage while maintaining high performance. By utilizing a PID controller to

dynamically adjust the regularization coefficient β towards a target Effective Bit Operations (EBOPs) metric, we established a robust methodology to pre-determine the model's hardware footprint. This allows for precise utilization of the FPGA area, ensuring models fit exactly within user-defined regions, such as the single SLR target of the jet taggers in this work.

Future work will focus on addressing the scaling limitations of the attention layer. Investigating linear-complexity attention mechanisms or hybrid architectures could potentially mitigate the quadratic resource costs, allowing the benefits of quantization to be realized at higher particle counts. Additionally, extending the target EBOPs training methodology to other trigger tasks could further streamline the deployment of deep learning on FPGAs.

5. Acknowledgements

Work done by Imperial College is funded by the Science and Technology Facilities Council (STFC) grant ST/W000636/1 and EPSRC (grant numbers UKRI256, EP/V028251/1, EP/N031768/1, EP/S030069/1, and EP/X036006/1). B.M. acknowledges the support of Schmidt Sciences. A.G., and J.N. are supported by the DOE Office of Science, Award No. DE-SC0023524, FermiForward Discovery Group, LLC under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics, LDRD L2024-066-1, Fermilab, DOE Office of Science, Office of High Energy Physics "Designing efficient edge AI with physics phenomena" Project (DE-FOA-0002705), DOE Office of Science, Office of Advanced Scientific Computing Research under the "Real-time Data Reduction Codesign at the Extreme Edge for Science" Project (DE-FOA-0002501).

References

[1] The LHC Study Group, "The Large Hadron Collider, Conceptual Design," tech. rep., CERN/AC/95-05 (LHC) Geneva, 1995.

[2] G. Aad *et al.*, "The atlas experiment at the cern large hadron collider," *Journal of Instrumentation*, vol. 3, no. 08, p. S08003, 2008.

[3] S. Chatrchyan *et al.*, "The CMS Experiment at the CERN LHC," *JINST*, vol. 3, p. S08004, 2008.

[4] The CMS Collaboration, "The Phase-2 Upgrade of the CMS Level-1 Trigger," tech. rep., CERN, 2020.

[5] The ATLAS Collaboration, "Technical Design Report for the Phase-II Upgrade of the ATLAS TDAQ System," tech. rep., CERN, 2017.

[6] I. Zurbano Fernandez *et al.*, "High-Luminosity Large Hadron Collider (HL-LHC): Technical design report," *CERN Yellow Reports: Monographs*, 2020.

[7] C. Sun, T. Nakajima, Y. Mitsumori, *et al.*, "Fast muon tracking with machine learning implemented in fpga," *Nuclear Instruments and Methods in Physics Research Section A*, vol. 1045, p. 167546, 2023.

[8] C. N. Coelho, A. Kuusela, S. Li, *et al.*, "Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors," *Nature Machine Intelligence*, vol. 3, no. 8, pp. 675–686, 2021.

[9] J. Ngadiuba, V. Loncar, M. Pierini, *et al.*, "Compressing deep neural networks on fpgas to binary and ternary precision with hls4ml," *Machine Learning: Science and Technology*, vol. 2, no. 1, p. 015001, 2020.

[10] Q. Lou, F. Guo, M. Kim, *et al.*, "Autoq: Automated kernel-wise neural network quantization," in *International Conference on Learning Representations*, 2020.

[11] F. Fahim, B. Hawks, C. Herwig, *et al.*, "hls4ml: An open-source codesign workflow to empower scientific low-power machine learning devices," *CoRR*, vol. abs/2103.05579, 2021.

[12] A. Coccaro, F. Armando Di Bello, S. Giagu, *et al.*, "Fast neural network inference on fpgas for triggering on long-lived particles at colliders," *Machine Learning: Science and Technology*, p. 045040, 2023.

[13] Francescato, Simone, Giagu, Stefano, Riti, Federica, *et al.*, "Model compression and simplification pipelines for fast deep neural network inference in fpgas in hep," *Eur. Phys. J. C*, 2021.

[14] A. Bal, T. Brandes, F. Iemmi, *et al.*, "Distilling particle knowledge for fast reconstruction at high-energy physics experiments," *Mach. Learn. Sci. Tech.*, vol. 5, no. 2, p. 025033, 2024.

[15] Z. Que, M. Loo, H. Fan, *et al.*, "Optimizing graph neural networks for jet tagging in particle physics on FPGAs," in *2022 32nd International Conference on Field-Programmable Logic and Applications (FPL)*, 2022.

[16] Z. Que, H. Fan, M. Loo, *et al.*, "LL-GNN: Low Latency Graph Neural Networks on FPGAs for High Energy Physics," *ACM Transactions on Embedded Computing Systems*, 2024.

[17] P. Odagiu, Z. Que, J. Duarte, *et al.*, "Ultrafast jet classification at the HL-LHC," *Machine Learning: Science and Technology*, vol. 5, no. 3, p. 035017, 2024.

[18] C. Sun, J. Ngadiuba, M. Pierini, *et al.*, "Fast Jet Tagging with MLP-Mixers on FPGAs," *Machine Learning: Science and Technology*, 2025.

[19] H. Qu, C. Li, and S. Qian, "Particle transformer for jet tagging," 2024.

[20] J. Spinner, V. Bresó, P. de Haan, *et al.*, "Lorentz-equivariant geometric algebra transformers for high-energy physics," 2024.

[21] H. Qu and L. Gouskos, "Jet tagging via particle clouds," *Physical Review D*, 2020.

[22] A. Bogatskiy, T. Hoffman, D. W. Miller, *et al.*, "Explainable equivariant neural networks for particle physics: Pelican," *Journal of High Energy Physics*, 2024.

[23] S. Ma, H. Wang, L. Ma, *et al.*, "The era of 1-bit llms: All large language models are in 1.58 bits," 2024.

[24] S. Rai, Prisha, and J. Kumar, "Investigating 1-bit quantization in transformer-based top tagging," 2025.

[25] F. Wojcicki, Z. Que, A. D. Tapper, *et al.*, "Accelerating transformer neural networks on fpgas for high energy physics experiments," in *2022 International Conference on Field-Programmable Technology (ICFPT)*, 2022.

[26] Z. Jiang, D. Yin, E. E. Khoda, *et al.*, "Ultra fast transformers on fpgas for particle physics experiments," 2024.

[27] FastML Team, "fastmachinelearning/hls4ml," 2025.

[28] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[29] J. Lee, Y. Lee, J. Kim, *et al.*, "Set transformer: A framework for attention-based permutation-invariant neural networks," 2019.

[30] M. Zaheer, S. Kottur, S. Ravanbakhsh, *et al.*, "Deep sets," *CoRR*, vol. abs/1703.06114, 2017.

[31] S. Chang, T. Årrestad, V. Lončar, *et al.*, "Gradient-based automatic per-weight mixed precision quantization for neural networks on-chip," 2024.

[32] Alessandro, G. Franco, N. Fraser, *et al.*, "Xilinx/brevitas: Release version 0.2.1," 2021.

[33] C. Sun, Z. Que, V. Loncar, *et al.*, "da4ml: Distributed arithmetic for real-time neural networks on fpgas," 2025.

[34] M. Pierini, J. M. Duarte, N. Tran, *et al.*, "Hls4ml lhc jet dataset (150 particles)," 2020.

[35] E. Coleman, M. Freytsis, A. Hinzmann, *et al.*, "The importance of calorimetry for highly-boosted jet substructure," *Journal of Instrumentation*, vol. 13, no. 01, p. T01003, 2018.

[36] E. A. Moreno, O. Cerri, J. M. Duarte, *et al.*, "JEDI-net: a jet identification algorithm based on interaction networks," *The European Physical Journal C*, 2020.

[37] P. Odagiu, Z. Que, J. Duarte, *et al.*, "Ultrafast jet classification at the hl-lhc," *Machine Learning: Science and Technology*, vol. 5, 2024.

[38] J. Weitz, D. Demler, L. McDermott, *et al.*, "Neural architecture codesign for fast physics applications," *Machine Learning: Science and Technology*, vol. 6, no. 3, p. 035009, 2025.