

Sub-microsecond Transformers for Jet Tagging on FPGAs

Lauri Laatu¹, Chang Sun², Arianna Cox¹, Abhijith Gandrakota³,
Benedikt Maier¹, Jennifer Ngadiuba³, Zhiqiang Que¹,
Wayne Luk¹, Maria Spiropulu², and Alexander Tapper¹

¹Imperial College London, United Kingdom

²California Institute of Technology, USA

³Fermilab, USA

Abstract

We present the first sub-microsecond transformer implementation on an FPGA achieving competitive performance for state-of-the-art high-energy physics benchmarks. Transformers have shown exceptional performance on multiple tasks in modern machine learning applications, including jet tagging at the CERN Large Hadron Collider (LHC). However, their computational complexity prohibits use in real-time applications, such as the hardware trigger system of the collider experiments up until now. In this work, we demonstrate the first application of transformers for jet tagging on FPGAs, achieving $\mathcal{O}(100)$ nanosecond latency with superior performance compared to alternative baseline models. We leverage high-granularity quantization and distributed arithmetic optimization to fit the entire transformer model on a single FPGA, achieving the required throughput and latency. Furthermore, we add multi-head attention and linear attention support to hls4ml, making our work accessible to the broader fast machine learning community. This work advances the next-generation trigger systems for the High Luminosity LHC, enabling the use of transformers for real-time applications in high-energy physics and beyond.

1 Introduction

The CERN Large Hadron Collider (LHC) [1] produces hundreds of terabytes of data per second from collisions at a frequency of 40 MHz. To handle this vast data volume, two-level trigger systems are employed at experiments such as ATLAS [2] and CMS [3]. The first stage of this filter is the Level-1 trigger system (L1T), composed of hundreds of Field-Programmable Gate Arrays (FPGAs) that process data in real time and determine which collision events are of interest and thus are passed downstream for further analysis. As the data buffer size is limited, the data can only be retained for a short time, and the end-to-end latency of the L1T is limited to a few microseconds [4, 5]. Events not flagged by the L1T have to be permanently discarded, which makes the accuracy of the L1T critical.

The High-Luminosity LHC (HL-LHC) [6] upgrade will increase the amount of simultaneous proton-proton collisions every 25 ns to as high as 140-200 from the current 40-80, which requires an extensive upgrade of the trigger systems. Extending the capability of the L1T to perform more informed decisions with deep learning based methods is therefore a key step toward the next-generation trigger systems. Research efforts [7–14] have focused on adapting machine learning models for FPGA deployment in real-time environments, showcasing their potential for ultra-low latency applications. Of particular relevance to improved L1T performance is the ability to identify (“tag”) jets, collimated

sprays of particles that are key probes of the Standard Model and that appear in various new theories featuring hypothetical particles and forces. Ref. [8] demonstrated that quantization-aware training (QAT) could enable jet tagging with quantized neural networks on FPGAs, achieving accuracy competitive to the full precision models while reducing resource usage by orders of magnitude. However, these models used high-level features that are typically not available at the L1T. In [15–19], the feasibility of deploying practical jet-tagging neural networks on FPGAs with satisfactory performance and resource utilization for L1T is demonstrated with particle-level input.

For offline data processing, state-of-the-art models for jet tagging include those based on Transformer architectures like the Particle Transformer [20] and the Lorentz-Equivariant Geometric Algebra Transformer [21] as well as Graph Neural Networks (GNNs), such as ParticleNet [22] and PELICAN [23] which have shown exceptional performance in jet tagging tasks. However, such models are orders of magnitude too resource-intensive to be deployed on an FPGA in the real-time trigger system.

There have been attempts to deploy transformers on FPGAs for sub-microsecond latency applications. However, none of these approaches have yielded competitive performance compared to other methods, either only using the high-level features in [24] or are unable to achieve the sub-microsecond latency in [25]. In this work, we present the first successful implementation of transformers for jet tagging on FPGAs achieving $\mathcal{O}(100)$ nanosecond latency and better performance than current baseline models. The implementations we made for transformer support are also contributed back to the upstream `hls4ml` library [26], making our work accessible to the broader fast machine learning community.

2 Method

2.1 Model Architecture

We adopt a simple encoder-only transformer architecture with vanilla multi-head attention (MHA) [27] with a single attention head. The input to the model is a sequence of particles, with a maximum number of (8, 16, 32, 64) particles sorted by p_T , the transverse momentum. Each particle has three features: p_T , η , and ϕ . No positional encoding is added to the inputs, and the model used is a form of the Set Transformer [28].

As the vanilla attention’s computational complexity is $\mathcal{O}(n^2)$, where n is the sequence length (which is the maximum number of particles in our case), we also implement the linear attention from Linformer [29] to improve model efficiency. In Linformer, the key and value vectors are projected to a lower dimension $k < n$. As such, the complexity is now $\mathcal{O}(k \cdot n)$. By selecting a k decoupled from n , we are able to reduce the computational burden with approximated attention.

2.2 Model Compression

We adopt the High Granularity Quantization (HGQ) [30] for unified quantization and pruning of the networks. With gradient-based, per-parameter bitwidth optimization including zero width, the method performs loss-aware quantization and pruning simultaneously to reduce the model size. This approach significantly reduces the model size compared to layer-wise quantization methods such as those provided by QKeras [8] and Brevitas [31] while maintaining the accuracy of the original model.

HGQ works by introducing Effective Bit Operations (EBOPs) as an additional parameter where EBOPs is computed by summing the products of the bitwidths $a_{bw} \cdot b_{bw}$ of all operations in the network. During training, the bitwidths are minimized as part of the loss. EBOPs has high correlation with the resource usage on FPGAs [30], and it is possible to set a target EBOPs to achieve a desired size on hardware when training with HGQ. Value-wise heterogeneous quantization is also applied in the datapath, which breaks the otherwise permutation-invariant model to further reduce the firmware footprint.

We use `da4ml` [32] to optimize the constant-matrix-vector multiplication (CMVM) operations used. The algorithm employed is a hybrid algorithm with both graph-based reduction and common subexpression elimination that transforms the CMVM operations into equivalent adder graphs. As the equivalence is exact, this optimization does not introduce any approximation error.

3 Experiment

We evaluate our methods using a common jet tagging dataset [33, 34], a standard benchmark dataset used by the high-energy physics community [35, 15, 16, 36, 18, 37]. The dataset includes five classes of jets, each categorized by their originating particles: gluons (g), light quarks (q), W bosons (W), Z bosons (Z), and top quarks (t). This dataset contains 620,000 jets in the training set and 260,000 jets in the test set, each balanced between the classes. The FPGA target used for evaluation is the Xilinx XCU 250 chip. All experiments are performed using Vitis HLS and Vivado and the code used is part of `hls4ml` package.

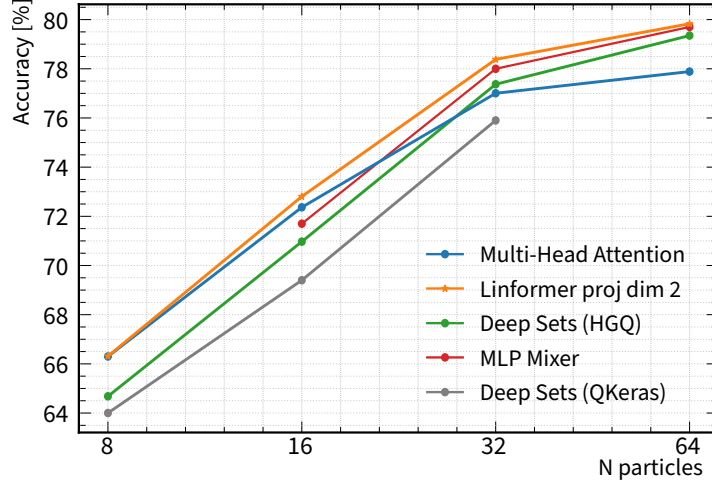


Figure 1: Accuracy as a function of the maximum number of input particles. The decrease in accuracy for the full Transformer models compared to the other models at 64 particles is expected due to the fixed resource budget we enforced during training with EBOPs control.

The transformer models are compared to methods from previous works: MLP Mixer-based jet tagging [18], Deep Sets with uniform quantization trained with QKeras [36] and [37], where the latter performed extensive neural architecture searches to find optimal tradeoffs between model accuracy and firmware performance. All models trained in this work have been trained with a target EBOPs of 350,000 - roughly one Super Logic Region of the target XCU 250 chip - with a proportion-integral-derivative (PID) controller over β , the regularization factor controlling the model size [30], to enforce the target EBOPs.

The accuracy of the models as a function of number of input particles is shown in Figure 1. The two attention methods both achieve state-of-the-art accuracy at 8 particles. However, as the input length increases, the quadratic scaling of the MHA model and the MHA layer requiring more FPGA resources, results in lower bitwidths across the network and the overall performance degrades compared to the other models. However, the Linformer models consistently outperform baseline methods across all input lengths. We also show the Receiver Operating Characteristic (ROC) curves in Figure 2 for the efficiency per event type for different input lengths. All models exhibit a similar trend, with performance generally improving as the input length increases, although the rate of improvement varies between models and signals.

Bitwidths for the Linformer models are presented in Figure 3. The bitwidths are divided into the attention layer which is constrained to at least one bit to disable pruning for training stability, and the other layers that do not have such constraints. As the target EBOPs is set to constant values, the model bitwidths have to adjust as the input length increases which explains the higher proportion of low bitwidths observed with larger input lengths.

The results of hardware synthesis are presented in Table 1. The table shows that all transformer models are able to fulfill the $\mathcal{O}(100)$ nanosecond latency requirement. The attention block of the MHA model with 64 input particles is consistently collapsing over several trained models despite the bitwidth constrained to at least one bit, turning it into a Deep Set which explains its vastly different

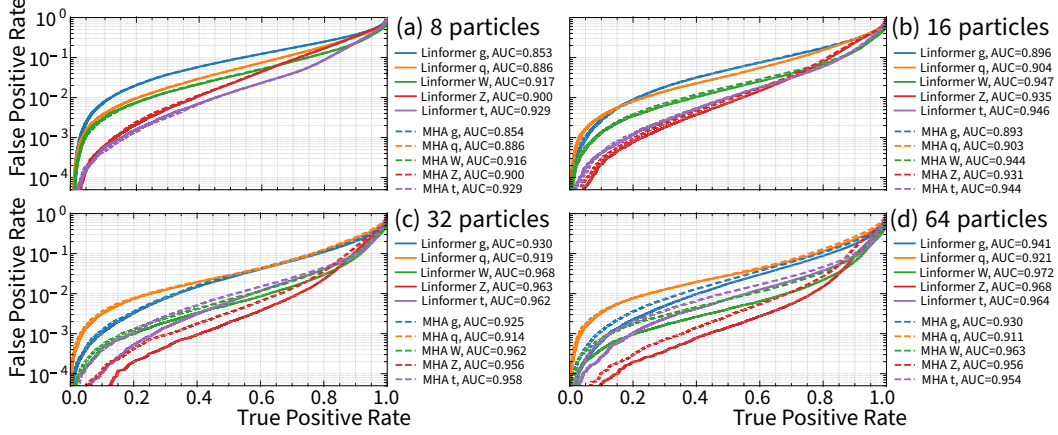


Figure 2: Receiver Operating Characteristic curves for different input lengths for MHA and Linformer.

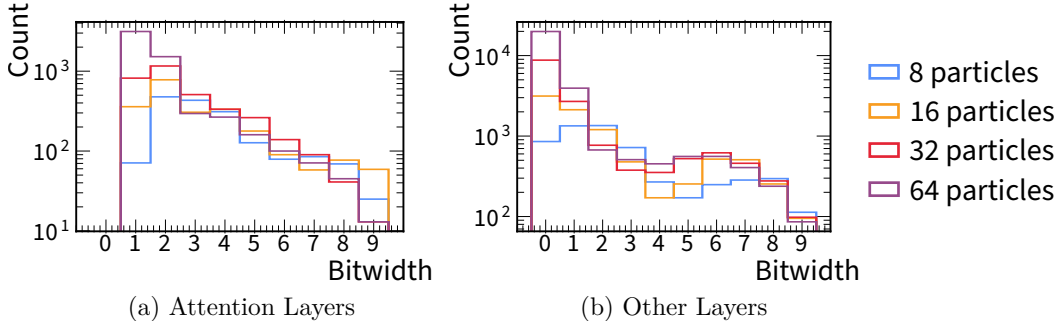


Figure 3: The distribution of the weight bitwidths of the obtained Linformer models. For the attention layers, the bitwidths are constrained to at least one bit, while other layers can adapt their bitwidth freely.

resource usage. As the HGQ trained models all have the same target EBOPs, the resource usage does not increase as happens with uniformly quantized models such as the Deep Sets trained in QKeras.

4 Conclusion and Future Work

This work introduces real-time transformer models with MHA and linear attention mechanisms with practical applications in particle physics. Of the two types of attention, Linformer is able to perform well over all input lengths evaluated, while MHA has lower performance due to the quadratic complexity leading to limited viability for FPGA deployment. All evaluated models achieve the required $\mathcal{O}(100)$ nanosecond latency with superior performance compared to the baseline models. Our code is available as part of the HGQ and hls4ml packages. In the future, we are looking to apply the transformer implementations in other real-time tasks in high-energy physics, such as jet tagging in high-pileup environment, particle reconstruction, event reconstruction and as a building block for a foundation model.

Model	Particles	Acc. (%)	Latn. (ns)	LUT (k)	II (clk)	DSP
Multi-Head Attention	8	66.3	104	246	1	0
Multi-Head Attention	16	72.3	98	279	1	0
Multi-Head Attention	32	77.0	83	180	1	0
Multi-Head Attention	64	77.9	44	47	1	0
Linformer	8	66.3	110	230	1	0
Linformer	16	72.8	103	246	1	0
Linformer	32	78.4	140	267	1	0
Linformer	64	79.8	78	202	1	0
Deep Sets (HGQ)	8	64.7	49	177	1	0
Deep Sets (HGQ)	16	70.1	53	205	1	0
Deep Sets (HGQ)	32	77.4	53	256	1	0
Deep Sets (HGQ)	64	79.4	44	191	1	0
MLP Mixer [18]	16	71.7	68	75	1	0
MLP Mixer [18]	32	78.0	62	63	1	0
MLP Mixer [18]	64	79.7	72	159	1	0
Deep Sets (QKeras) [36]	8	64.0	95	386	3	626
Deep Sets (QKeras) [36]	16	69.4	115	747	3	555
Deep Sets (QKeras) [36]	32	75.9	130	903	2	434
Deep Sets M (QKeras) [37]	8	65.1	110	130	3	548
Deep Sets L (QKeras) [37]	8	66.6	135	337	3	2,458

Table 1: Performance comparison of the Transformer, Linformer, and other models. The table shows the FPGA implementation details in latency, look-up tables (LUT) used, initiation interval (II) in clock cycles and digital signal processing (DSP) blocks used. We show that the Linformers are able to achieve state-of-the-art accuracy while maintaining a reasonable low latency and realistic resource utilization for FPGA deployment.

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] “The atlas experiment at the cern large hadron collider,” *Journal of Instrumentation*, vol. 3, p. S08003, aug 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, Geneva, 2020. Final version.
- [5] The ATLAS Collaboration, “Technical Design Report for the Phase-II Upgrade of the ATLAS TDAQ System,” tech. rep., CERN, Geneva, 2017.

- [6] I. Zurbano Fernandez *et al.*, “High-Luminosity Large Hadron Collider (HL-LHC): Technical design report,” *CERN Yellow Reports: Monographs*, vol. 10/2020, 12 2020.
- [7] C. Sun, T. Nakajima, Y. Mitsumori, Y. Horii, and M. Tomoto, “Fast muon tracking with machine learning implemented in fpga,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 1045, p. 167546, Jan. 2023.
- [8] C. N. Coelho, A. Kuusela, S. Li, H. Zhuang, J. Ngadiuba, T. K. Aarrestad, V. Loncar, M. Pierini, A. A. Pol, and S. Summers, “Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors,” *Nature Machine Intelligence*, vol. 3, pp. 675–686, jun 2021.
- [9] J. Ngadiuba, V. Loncar, M. Pierini, S. Summers, G. D. Guglielmo, J. Duarte, P. Harris, D. Rankin, S. Jindariani, M. Liu, K. Pedro, N. Tran, E. Kreinar, S. Saguear, Z. Wu, and D. Hoang, “Compressing deep neural networks on fpgas to binary and ternary precision with hls4ml,” *Machine Learning: Science and Technology*, vol. 2, p. 015001, dec 2020.
- [10] Q. Lou, F. Guo, M. Kim, L. Liu, and L. Jiang., “Autoq: Automated kernel-wise neural network quantization,” in *International Conference on Learning Representations*, 2020.
- [11] F. Fahim, B. Hawks, C. Herwig, J. Hirschauer, S. Jindariani, N. Tran, L. P. Carloni, G. D. Guglielmo, P. C. Harris, J. D. Krupa, D. Rankin, M. B. Valentin, J. Hester, Y. Luo, J. Mamish, S. Orgreneci-Memik, T. Aarrestad, H. Javed, V. Loncar, M. Pierini, A. A. Pol, S. Summers, J. M. Duarte, S. Hauck, S. Hsu, J. Ngadiuba, M. Liu, D. Hoang, E. Kreinar, and Z. Wu, “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, L. Rambelli, and N. Stocchetti, “Fast neural network inference on fpgas for triggering on long-lived particles at colliders,” *Machine Learning: Science and Technology*, vol. 4, p. 045040, Nov. 2023.
- [13] Francescato, Simone, Giagu, Stefano, Riti, Federica, Russo, Graziella, Sabetta, Luigi, and Tortonesi, Federico, “Model compression and simplification pipelines for fast deep neural network inference in fpgas in hep,” *Eur. Phys. J. C*, vol. 81, no. 11, p. 969, 2021.
- [14] A. Bal, T. Brandes, F. Iemmi, M. Klute, B. Maier, V. Mikuni, and T. K. Arrestad, “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, M. Pierini, A. Tapper, and W. Luk, “Optimizing graph neural networks for jet tagging in particle physics on FPGAs,” in *2022 32nd International Conference on Field-Programmable Logic and Applications (FPL)*, pp. 327–333, IEEE, 2022.
- [16] Z. Que, H. Fan, M. Loo, H. Li, M. Blott, M. Pierini, A. Tapper, and W. Luk, “LL-GNN: Low Latency Graph Neural Networks on FPGAs for High Energy Physics,” *ACM Transactions on Embedded Computing Systems*, vol. 23, no. 2, pp. 1–28, 2024.
- [17] P. Odagiu, Z. Que, J. Duarte, J. Haller, G. Kasieczka, A. Lobanov, V. Loncar, W. Luk, J. Ngadiuba, M. Pierini, P. Rincke, A. Seksaria, S. Summers, A. Sznajder, A. Tapper, and T. K. Årrestad, “Ultrafast jet classification at the HL-LHC,” *Machine Learning: Science and Technology*, vol. 5, p. 035017, July 2024.
- [18] C. Sun, J. Ngadiuba, M. Pierini, and M. Spiropulu, “Fast Jet Tagging with MLP-Mixers on FPGAs,” *Machine Learning: Science and Technology*, 2025.
- [19] Z. Que, C. Sun, S. Paramesvaran, E. Clement, K. Karakoulaki, C. Brown, L. Laatu, A. Cox, A. Tapper, W. Luk, and M. Spiropulu, “JEDI-linear: Fast and Efficient Graph Neural Networks for Jet Tagging on FPGAs,” in *2025 International Conference on Field-Programmable Technology (FPT)*, IEEE, 2025.
- [20] H. Qu, C. Li, and S. Qian, “Particle transformer for jet tagging,” 2024.

- [21] J. Spinner, V. Bresó, P. de Haan, T. Plehn, J. Thaler, and J. Brehmer, “Lorentz-equivariant geometric algebra transformers for high-energy physics,” 2024.
- [22] H. Qu and L. Gouskos, “Jet tagging via particle clouds,” *Physical Review D*, vol. 101, Mar. 2020.
- [23] A. Bogatskiy, T. Hoffman, D. W. Miller, J. T. Offermann, and X. Liu, “Explainable equivariant neural networks for particle physics: Pelican,” *Journal of High Energy Physics*, vol. 2024, no. 3, p. 113, 2024.
- [24] F. Wojcicki, Z. Que, A. D. Tapper, and W. Luk, “Accelerating transformer neural networks on fpgas for high energy physics experiments,” in *2022 International Conference on Field-Programmable Technology (ICFPT)*, pp. 1–8, 2022.
- [25] Z. Jiang, D. Yin, E. E. Khoda, V. Loncar, E. Govorkova, E. Moreno, P. Harris, S. Hauck, and S.-C. Hsu, “Ultra fast transformers on fpgas for particle physics experiments,” 2024.
- [26] FastML Team, “fastmachinelearning/hls4ml,” 2025.
- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *CoRR*, vol. abs/1706.03762, 2017.
- [28] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, and Y. W. Teh, “Set transformer: A framework for attention-based permutation-invariant neural networks,” 2019.
- [29] S. Wang, B. Z. Li, M. Khabza, H. Fang, and H. Ma, “Linformer: Self-attention with linear complexity,” 2020.
- [30] S. Chang, T. Årrestad, V. Lončar, J. Ngadiuba, and M. Spiropulu, “Gradient-based automatic per-weight mixed precision quantization for neural networks on-chip,” 2024.
- [31] Alessandro, G. Franco, nickfraser, Y. Umuroglu, and vfdev, “Xilinx/brevitas: Release version 0.2.1,” Feb 2021.
- [32] C. Sun, Z. Que, V. Loncar, W. Luk, and M. Spiropulu, “da4ml: Distributed arithmetic for real-time neural networks on fpgas,” 2025.
- [33] M. Pierini, J. M. Duarte, N. Tran, and M. Freytsis, “Hls4ml lhc jet dataset (150 particles),” Jan. 2020.
- [34] E. Coleman, M. Freytsis, A. Hinzmann, M. Narain, J. Thaler, N. Tran, and C. Vernieri, “The importance of calorimetry for highly-boosted jet substructure,” *Journal of Instrumentation*, vol. 13, p. T01003, jan 2018.
- [35] E. A. Moreno, O. Cerri, J. M. Duarte, H. B. Newman, T. Q. Nguyen, A. Periwal, M. Pierini, A. Serikova, M. Spiropulu, and J.-R. Vlimant, “JEDI-net: a jet identification algorithm based on interaction networks,” *The European Physical Journal C*, vol. 80, no. 1, pp. 1–15, 2020.
- [36] P. Odagiu, Z. Que, J. Duarte, J. Haller, G. Kasieczka, A. Lobanov, V. Loncar, W. Luk, J. Ngadiuba, M. Pierini, P. Rincke, A. Seksaria, S. Summers, A. Sznajder, A. Tapper, and T. K. Årrestad, “Ultrafast jet classification at the hl-lhc,” *Machine Learning: Science and Technology*, vol. 5, p. 035017, July 2024.
- [37] J. Weitz, D. Demler, L. McDermott, N. Tran, and J. Duarte, “Neural architecture codesign for fast physics applications,” *Machine Learning: Science and Technology*, vol. 6, p. 035009, jul 2025.