

Hardware-Aware Tensor Networks for Real-Time Machine Learning at Particle Colliders

Sagar Addepalli^{*1}, Deniz Yilmaz², Prajita Bhattarai¹, Abhilasha Dave¹, and Julia Gonski¹

¹SLAC National Accelerator Laboratory, Menlo Park, CA, USA

²Stanford University, Stanford, CA, USA

1 Introduction

Tensor networks (TNs) are one of the most direct bridges between quantum many-body physics and machine learning (ML). Introduced to represent quantum states obeying area-law entanglement [1, 2], they factorize high-dimensional data into interconnected low-rank tensors and have become a versatile quantum-inspired learning tool [3, 4, 5, 6, 7]. Crucially for potential real-time deployment, TN inference is a sequence of bilinear contractions with no nonlinear activations and a spatially local structure, making TN models exceptionally well matched to low-latency, fixed-point hardware. This positions them to bring quantum-inspired learning to the “edge” of scientific experiments, allowing the study of quantum advantage in modern infrastructure while awaiting the advent of mature quantum processors beyond the noisy intermediate-scale era [8, 9, 10, 11].

A demanding edge testbed is the real-time trigger of particle detectors at the Large Hadron Collider (LHC), where rare and/or beyond the Standard-Model events must be selected from an overwhelming background within sub-microsecond latency using resource-constrained field-programmable gate array (FPGA) deployments [12, 13, 14]. This work considers the development and application of TNs, model compression, and FPGA implementation for collider triggers. Three separate investigations are discussed, all leading to implementations that meet the specifications of future trigger systems, and controlling the bond structure of a TN with quantum-information-theoretic tools (Fig. 1): **(1)** a Spaced Matrix Product Operator (SMPO) that performs unsupervised anomaly detection and runs within trigger latency on an FPGA [15]; **(2)** its *cascaded* form (CSMPO), a linear refactoring that preserves learning capacity while exposing extra freedom for hardware codesign and ultimately reduced resources; and **(3)** a *data-conditioned* bond-truncation scheme that compresses the trained model with no loss of performance.

2 Tensor networks for real-time anomaly detection

Events at the Large Hadron Collider originate as low-level signals on individual detector channels, which are passed through physics-based algorithms to correlate signals into particle signatures and ultimately reconstruct the entire collision. In this work, each collision event is embedded as a Matrix Product State (MPS), one tensor site per reconstructed particle, with sites ordered by a statistical analogue of *quantum mutual information* (von Neumann entropy of pairwise bond density matrices) [16, 17, 18] so that strongly correlated objects are adjacent and correlations are captured at small bond dimension [19]. An SMPO is a trainable MPO that reduces the length of the MPS it acts on (Fig. 1a); our model contracts the event MPS to a single vector whose squared norm is the anomaly score, trained unsupervised on background alone in `tn4ml` [20]. On an anomaly detection benchmark dataset of simulated

^{*}Corresponding author: `sagar@slac.stanford.edu`

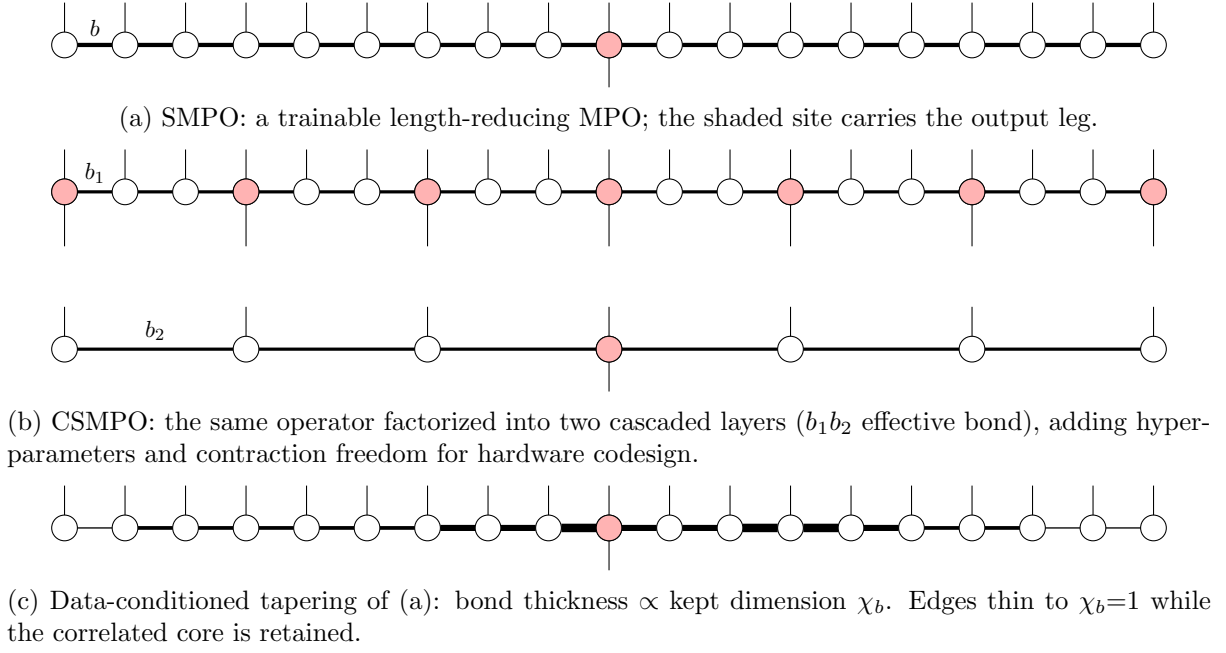


Figure 1: The three structural ideas expressed as the bond structure of one tensor network: (a) the SMPO anomaly detector, (b) its cascaded factorization, and (c) its data-conditioned taper.

collision events [21, 22], the model is competitive with classical state-of-the-art across four diverse new-physics signals [23, 24], and synthesizes to an FPGA with sub-microsecond latency and zero DSP usage, placing it within trigger budgets at the High Luminosity LHC and future colliders [25, 26, 27, 28].

3 Cascaded SMPO: linear factorization for hardware flexibility

Because SMPOs are *linear* operators, the composition of two SMPOs is itself an SMPO whose effective bond dimension is the product of the layer bond dimensions,

$$\text{SMPO}(b_1) \circ \text{SMPO}(b_2) \subset \text{SMPO}(b_1 b_2). \quad (1)$$

The *cascaded* SMPO realizes this explicitly as two layers (Fig. 1b). The inclusion is strict – the cascade imposes a factorization constraint on the effective weight tensor, yet this restricted parameterization is advantageous: it adds optimizable degrees of freedom (per-layer spacing, the intermediate physical dimension) and splits one large bond dimension across two layers, whose contractions scale only quadratically-to-cubically in their now-smaller bonds. The practical payoff is *flexibility for hardware codesign*: at matched performance, an equivalent cascade uses roughly half the trainable parameters and admits several contraction orderings, which we exploit to reduce FPGA inference latency relative to the single-layer SMPO [15].

4 Data-conditioned bond truncation for model compression

The bond dimensions set the cost of a TN model, so reducing them is the most direct lever on the on-device footprint. We *taper* the trained SMPO assigning a non-uniform, generally smaller bond dimension χ_b at each bond b (Fig. 1c), and ask which information-theoretic criterion should decide what to discard. We compare two per-bond prescriptions. (i) *Vacuum (operator Schmidt) truncation*: a per-bond singular value decomposition of the canonicalized operator, discarding

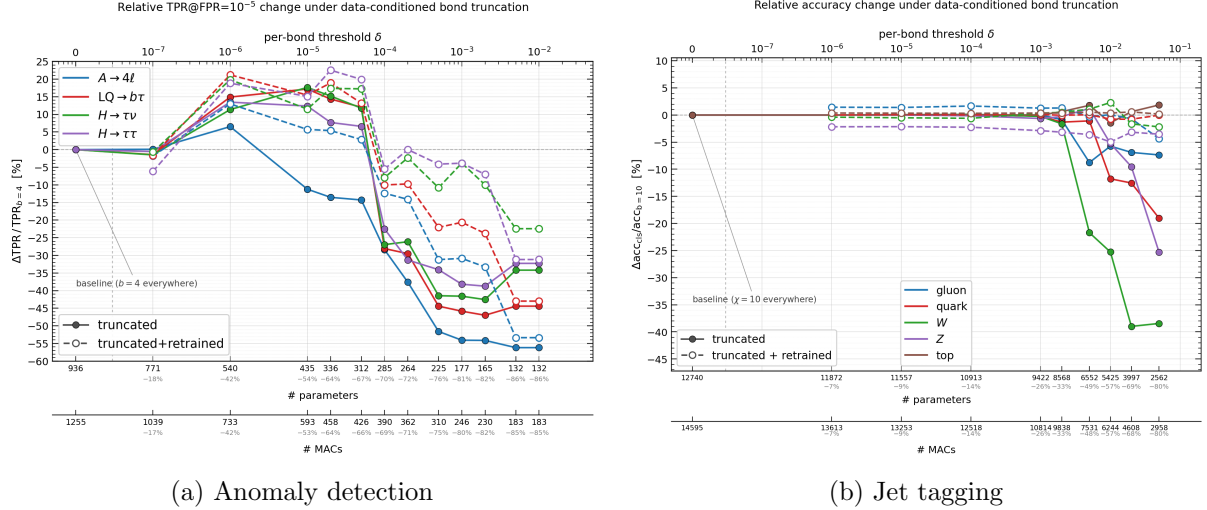


Figure 2: Relative change in signal efficiency under data-conditioned tapering of the trained model, versus model size (parameters/operations, lower axes). Solid: zero-shot truncation; dashed: truncation plus retraining. (a) shows the impact on the signal efficiency at 10^{-5} background suppression in an unsupervised anomaly detection task. (b) shows the classification accuracy on a 5-class supervised jet tagging task.

the smallest singular values, independent of the data. (ii) *Data-conditioned truncation*: along the exact contraction graph of the score we form a per-bond density operator

$$\rho_b = \mathbb{E}_{\mathbf{x}}[E_b(\mathbf{x})E_b(\mathbf{x})^\dagger], \quad (2)$$

where $E_b(\mathbf{x})$ is the half-network environment up to bond b on input $\mathbf{x}$; keeping its top- χ_b eigenvectors minimizes the data-averaged distortion of the model’s action. This is a density-matrix (environment) truncation in the DMRG sense [1, 2], but with the environment given by the empirical data ensemble rather than a quantum reference state: there is no target wavefunction, and ρ_b is a classical mixture over the data. Vacuum truncation is the data-agnostic limit (trivial environment).

Empirically, the distinction is decisive (Fig. 2a). Applied zero-shot, data-conditioned tapering preserves both performance and the calibrated score scale while removing a large fraction of the bond dimension, whereas vacuum truncation at matched compression collapses the score and degrades sensitivity. Retraining from the tapered initialization recovers and then *exceeds* the baseline: mild pruning improves signal efficiency (a structured-regularization effect), removing up to roughly two-thirds of the parameters and operations leaves performance unchanged, and only the most aggressive compression (an order-of-magnitude smaller model) trades off efficiency in the extreme-background-rejection regime, a knob set by the application. The construction is task-agnostic: it applies equally to a *supervised* jet-tagging model (Fig. 2b).

5 Outlook

A tensor network can be *trained* for front-end anomaly detection, *refactored* (cascade) to fit a hardware budget, and *compressed* (data-conditioned truncation) to shrink it further, all via quantum-information-theoretic control of its bond structure, and all deployable on real-time hardware. This talk will cover all three pillars of the work, including the SMPO/CSMPO architectures and FPGA implementation [15], the vacuum-versus-data-conditioned comparison, and the supervised cross-check (in preparation). Together these results pave the way for the introduction of quantum-inspired ML in edge applications for scientific experiments.

References

- [1] S. R. White, “Density-matrix algorithms for quantum renormalization groups,” *Phys. Rev. B*, vol. 48, pp. 10345–10356, Oct 1993.
- [2] S. Östlund and S. Rommer, “Thermodynamic limit of density matrix renormalization,” *Phys. Rev. Lett.*, vol. 75, pp. 3537–3540, Nov 1995.
- [3] R. Orús, “A practical introduction to tensor networks: Matrix product states and projected entangled pair states,” *Annals of Physics*, vol. 349, pp. 117–158, 2014.
- [4] F. Verstraete, V. Murg, and J. Cirac, “Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems,” *Advances in Physics*, vol. 57, no. 2, pp. 143–224, 2008.
- [5] E. Stoudenmire and D. J. Schwab, “Supervised Learning with Tensor Networks,” in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, (New York USA), Curran Associates, Inc., 2016.
- [6] J. Liu, S. Li, J. Zhang, and P. Zhang, “Tensor networks for unsupervised machine learning,” *Physical Review E*, vol. 107, Jan. 2023.
- [7] S.-J. Ran and G. Su, “Tensor networks for interpretable and efficient quantum-inspired machine learning,” *Intelligent Computing*, vol. 2, p. 0061, 2023.
- [8] J. Preskill, “Quantum computing in the nisc era and beyond,” *Quantum*, vol. 2, p. 79, Aug. 2018.
- [9] W. Li, Y. Ma, and D.-L. Deng, “Pitfalls and prospects of quantum machine learning,” *Nature Computational Science*, vol. 5, no. 12, pp. 1095–1097, 2025.
- [10] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, “Quantum machine learning,” *Nature*, vol. 549, p. 195–202, Sept. 2017.
- [11] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babush, R. Kueng, J. Preskill, and J. R. McClean, “Quantum advantage in learning from experiments,” *Science*, vol. 376, no. 6598, pp. 1182–1186, 2022.
- [12] A. M. Deiana, N. Tran, J. Agar, M. Blott, G. Di Guglielmo, J. Duarte, P. Harris, S. Hauck, M. Liu, M. S. Neubauer, J. Ngadiuba, S. Ogreni-Memik, M. Pierini, T. Aarrestad, S. Bähr, J. Becker, A.-S. Berthold, R. J. Bonventre, T. E. Müller Bravo, M. Diefenthaler, Z. Dong, N. Fritzsche, A. Gholami, E. Govorkova, D. Guo, K. J. Hazelwood, C. Herwig, B. Khan, S. Kim, T. Klijnsma, Y. Liu, K. H. Lo, T. Nguyen, G. Pezzullo, S. Rasoulizhad, R. A. Rivera, K. Scholberg, J. Selig, S. Sen, D. Strukov, W. Tang, S. Thais, K. L. Unger, R. Vilalta, B. von Krosigk, S. Wang, and T. K. Warburton, “Applications and Techniques for Fast Machine Learning in Science,” *Frontiers in Big Data*, vol. 5, Apr. 2022.
- [13] K. Sugizaki and A. Collaboration, “GELATO: A Generic Event-Level Anomalous Trigger Option for ATLAS in LHC Run 3,” tech. rep., CERN, Geneva, 2025.
- [14] A. Gandrakota and CMS Collaboration, “Real-time Anomaly Detection at the L1 Trigger of CMS Experiment,” in *Proceedings of Science (ICHEP2024)*, p. 1025, 2025. PoS ICHEP2024 (2025) 1025.
- [15] S. Addepalli, P. Bhattarai, A. Dave, and J. Gonski, “Hardware-aware tensor networks for real-time quantum-inspired anomaly detection at particle colliders,” 2026.

- [16] N. J. Cerf and C. Adami, “Negative entropy and information in quantum mechanics,” *Phys. Rev. Lett.*, vol. 79, pp. 5194–5197, Dec 1997.
- [17] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge UK: Cambridge University Press, 2010.
- [18] A. Acharya, M. Rudolph, J. Chen, J. Miller, and A. Perdomo-Ortiz, “Qubit seriation: Improving data-model alignment using spectral ordering,” in *Proceedings of the Machine Learning and the Physical Sciences Workshop at NeurIPS 2022*, Nov. 2022. Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).
- [19] J. Eisert, M. Cramer, and M. B. Plenio, “Colloquium: Area laws for the entanglement entropy,” *Rev. Mod. Phys.*, vol. 82, pp. 277–306, Feb 2010.
- [20] E. Puljak, S. Sanchez-Ramirez, S. Masot-Llima, J. Vallès-Muns, A. Garcia-Saez, and M. Pierini, “tn4ml: Tensor network training and customization for machine learning,” 2025.
- [21] E. Govorkova, E. Puljak, T. Aarrestad, T. James, V. Loncar, M. Pierini, A. A. Pol, N. Ghielmetti, M. Graczyk, S. Summers, J. Ngadiuba, T. Q. Nguyen, J. Duarte, and Z. Wu, “Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 MHz at the Large Hadron Collider,” *Nature Machine Intelligence*, vol. 4, p. 154–161, Feb. 2022.
- [22] E. Govorkova, E. Puljak, T. Aarrestad, M. Pierini, K. A. Woźniak, and J. Ngadiuba, “LHC physics dataset for unsupervised New Physics detection at 40 MHz,” *Scientific Data*, vol. 9, p. 118, Mar 2022.
- [23] J. Wang, C. Roberts, G. Vidal, and S. Leichenauer, “Anomaly Detection with Tensor Networks,” 6 2020.
- [24] E. Puljak, M. Pierini, and A. Garcia-Saez, “Tensor network for anomaly detection in the latent space of proton collision events at the LHC,” *Machine Learning: Science and Technology*, vol. 6, p. 045001, oct 2025.
- [25] O. Aberle et al, *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*. CERN Yellow Reports: Monographs, Geneva: CERN, 2020.
- [26] M. Benedikt and F. Zimmermann, “Future Circular Collider: Integrated Programme and Feasibility Study,” *Frontiers in Physics*, vol. Volume 10 - 2022, 2022.
- [27] L. Borella, A. Coppi, J. Pazzini, A. Stanco, M. Trenti, A. Triossi, and M. Zanetti, “Ultra-low latency quantum-inspired machine learning predictors implemented on FPGA,” 2024.
- [28] A. Coppi, E. Puljak, L. Borella, D. Jaschke, E. Rico, M. Pierini, J. Pazzini, A. Triossi, and S. Montangero, “Towards Tensor Network Models for Low-Latency Jet Tagging on FPGAs,” 2026.