TU Delft

Delft University of Technology

Document Version
Final published version

Licence
Dutch Copyright Act (Article 25fa)

Citation (APA)
Hua, E., van Ommen, S., Yu, K. Y., van Leeuwen, J. S., Bishnoi, R. K., Abunahla, H., Nur, S., Feld, S., & Ishihara, R. (2026). Memristor-Accelerated Neural Network Architectures for Energy-Efficient Quantum State Tomography. In *Proceedings of the 2026 10th IEEE Electron Devices Technology & Manufacturing Conference (EDTM)* IEEE. https://doi.org/10.1109/EDTM65772.2026.11496763

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

2026 10th IEEE Electron Devices Technology & Manufacturing Conference (EDTM) | 979-8-3315-8598-3/26/$31.00 ©2026 IEEE | DOI: 10.1109/EDTM65772.2026.11496763

Memristor-Accelerated Neural Network Architectures for Energy-Efficient Quantum State Tomography

Erbing Hua[1,2], Steven van Ommen[1], King Yiu Yu[1,2], Jim van Leeuwen[1], Rajendra Bishnoi[1], Heba Abunahla[1], Salahuddin Nur[1,2], Sebastian Feld[1,2], and Ryoichi Ishihara[1,2]

[1]*Department of Quantum and Computer Engineering, Delft University of Technology, Delft, The Netherlands*
[2]*QuTech, Delft University of Technology, Delft, The Netherlands*

Emails: e.hua, s.vanommen, k.y.yu, j.vanleeuwen, r.bishnoi, h.n.abunahla, s.nur, s.feld, r.ishihara@tudelft.nl

***Abstract*—Accurate quantum state tomography (QST) is vital for calibrating quantum processors but faces exponential scaling challenges. We benchmark seven neural architectures—FCN, CNN, CGAN, Transformer, RNN, RBM, and SVAE—for QST reconstruction using expectation- and probability-based measurements. CNN and CGAN achieve high fidelity ($F > 0.99$), while SVAE enables efficient event-driven learning. To enhance scalability, memristor-based computation-in-memory (CiM) acceleration is proposed for CNN and SVAE, leveraging analog matrix–vector multiplication in HfO_2 crossbars. The fabricated arrays show stable bipolar switching and STDP behavior, advancing energy-efficient, real-time quantum diagnostics through algorithm–hardware co-design.**

***Index Terms*—Quantum state tomography, Memristors, Neural networks.**

I. Introduction

Quantum computing offers transformative capabilities for solving problems in cryptography, optimization, and quantum simulation that are intractable for classical systems [1]–[4]. Accurate characterization of quantum states, known as *Quantum State Tomography (QST)*, is therefore a critical step for calibration, error mitigation, and verification of quantum hardware [5]. However, the dimensionality of a quantum system grows exponentially with the number of qubits, leading to severe measurement and computational overheads. Conventional QST requires 4^N measurement settings for an N-qubit system, which rapidly becomes impractical beyond a few qubits [6], [7].

Recent advances in *machine learning–based QST* have demonstrated that neural networks can reconstruct density matrices from incomplete or noisy measurement data, alleviating the curse of dimensionality [8]–[10]. Architectures such as *Convolutional Neural Networks (CNNs)*, *Conditional Generative Adversarial Networks (CGANs)*, and *Variational Autoencoders (VAEs)* have shown promising fidelity at small system scales [11]–[13]. Nevertheless, scalability and hardware efficiency remain open challenges. Most prior approaches rely on conventional von Neumann computing, where the separation between memory and processing units introduces the so-called "memory-wall" bottleneck and excessive energy consumption during data transfer [14]. For practical and scalable quantum diagnostics, *hardware-aware neural architectures* and *energy-efficient implementations* are required.

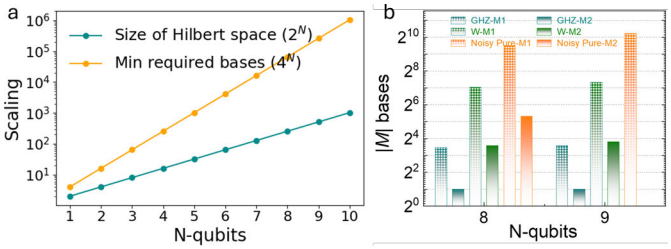


Fig. 1. **Measurement scalability in quantum state tomography (QST).** (a) Exponential growth of the Hilbert space dimension (2^N) and the minimum number of measurement bases (4^N) as functions of qubit number N. (b) Empirically determined number of measurement bases $|M|$ required to achieve high-fidelity (≈ 0.99) reconstruction for GHZ, W, and noisy pure states using the two measurement strategies (M1: expectation-based, M2: probability-based).

Memristor-based Computation-in-Memory (CiM) offers a promising pathway by merging data storage and analog computation within the same physical devices. Memristor crossbar arrays perform massively parallel matrix-vector multiplications (MVMs), the core operation of neural networks, with substantially reduced latency and power consumption [15], [16]. Moreover, such devices inherently support adaptive learning mechanisms such as spike-timing-dependent plasticity (STDP) [17], making them ideal for neuromorphic implementations of spiking and event-driven models.

In this work, we systematically benchmark seven neural network architectures, including CNN, CGAN, Transformer, RNN, RBM, and *Spiking Variational Autoencoder (SVAE)*, for quantum state reconstruction under two measurement strategies. We identify which architectures maintain high fidelity and computational efficiency as qubit number increases, and we discuss their compatibility with memristor-based CiM hardware. Our results show that CNN and CGAN provide the most robust supervised performance, while SVAE demonstrates decent efficiency and suitability for neuromorphic

Authorized licensed use limited to: TU Delft Library. Downloaded on June 08,2026 at 12:46:31 UTC from IEEE Xplore. Restrictions apply.

deployment. These findings establish a foundation for algorithm–hardware co-design toward scalable and energy-efficient quantum state tomography in next-generation quantum processors.

II. Methodology

A. Quantum Measurement Strategies

Quantum state tomography (QST) reconstructs the density matrix ρ of an N-qubit system from measurement data. Two complementary measurement schemes are considered:

- *M1: Expectation-based:* computes expectation values of observables $\langle A\rangle = \mathrm{Tr}(\rho A)$ for Pauli operators $\{\sigma_x, \sigma_y, \sigma_z\}$.
- *M2: Probability-based:* computes outcome probabilities $P(a) = \mathrm{Tr}(\rho|a\rangle\langle a|)$ across eigenstates $|a\rangle$.

Both methods are evaluated on representative pure and mixed states: GHZ, W, and noisy states. M2 provides richer information content but requires significantly more sampling per basis; M1 converges faster with fewer repetitions and is more practical for experimental implementations.

B. Neural Network Framework

Seven neural architectures were benchmarked: Fully Connected Network (FCN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Restricted Boltzmann Machine (RBM), Conditional GAN (CGAN), Transformer, and Spiking Variational Autoencoder (SVAE). Each model learns a mapping from simulated measurement data (M_1, M_2) to the target density matrix ρ. Supervised networks (CNN, CGAN, Transformer, FCN, RNN) minimize a mean-squared-error loss,

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{1}$$

while CGAN employs an adversarial loss combining generator G and discriminator D networks [12]. Unsupervised models (RBM, SVAE) optimize likelihood or variational evidence lower bound. Training was implemented in Python 3.11 with the Adam optimizer [18], using a GTX-1660 GPU and fidelity metric $F = |\langle\psi_{\mathrm{true}}|\psi_{\mathrm{rec}}\rangle|^2$.

C. Memristor-Based Computation-in-Memory (CiM)

To address hardware scalability, we propose memristor-based CiM acceleration for the most efficient models (CNN and SVAE). Memristor crossbars execute analog matrix–vector multiplication (MVM), expressed as $\mathbf{y} = \mathbf{W}\mathbf{x}$, where $\mathbf{W}$ denotes the conductance matrix, $\mathbf{x}$ the input vector, and $\mathbf{y}$ the output current vector. where conductance G_{ij} encodes weight W_{ij} and voltages represent input vector $\mathbf{x}$. Such devices perform parallel multiply-accumulate operations governed by Ohm's and Kirchhoff's laws, reducing data movement and energy [15]–[17]. Furthermore, spike-timing-dependent plasticity (STDP) enables on-chip adaptive learning for event-driven networks like SVAE, aligning QST computation with low-power neuromorphic hardware.

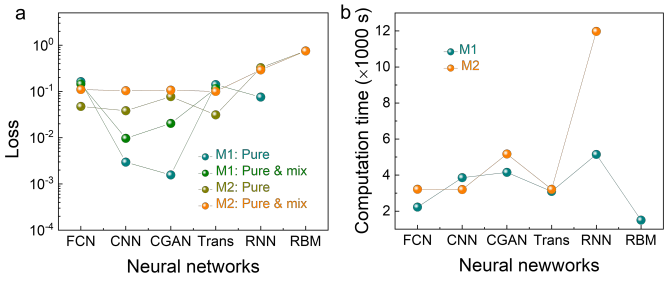


Fig. 2. **Benchmarking of neural network architectures for quantum state tomography (QST).** (a) Average reconstruction infidelity of different neural network models, including FCN, CNN, CGAN, Transformer, RNN, and RBM, under expectation-based (M1) and probability-based (M2) measurement schemes for both pure and mixed quantum states. (b) Corresponding computation time for each model, illustrating the trade-off between accuracy and efficiency. CNN and FCN demonstrate the fastest convergence ($\sim 2 \times 10^3$ s), while RBM and RNN exhibit the slowest training and poor scalability.

III. Results and Discussion

A. Measurement Requirements and Scaling

Figure 1(a) summarizes the Hilbert space increase and required bases increase exponentially with the increase amount of qubits, which demonstrate that the size and measurement bases are exponentially increase. Figure 1(b), it shows empirical number of measurement bases $|M|$ required to reach $F \approx 0.99$. For $N = 8$ qubits, GHZ states require only 11 (M1) or 2 (M2) bases, whereas W states demand 130 (M1) or 12 (M2) bases. The noisy pure state needs ∼720 (M1) or 40 (M2) bases. Although M2 reduces the number of bases, it demands far more measurement repetitions, making M1 the preferred approach in practice.

B. Neural Network Performance

The fidelity and training time for all models are summarized in Fig. 2. Among supervised models, CNN and CGAN achieve the lowest infidelity ($< 2 \times 10^{-3}$) across both pure and mixed states. FCN performs moderately well but scales poorly beyond 6 qubits. RNN and RBM exhibit unstable convergence, particularly for mixed states as depicted in Fig 2(a). Computation time scales with network depth: CNN and FCN converge in ∼4,000 s, whereas CGAN and Transformer require 3,000–6,000 s, and RBM exceeds 10,000 s. These results establish CNN as the most balanced model in accuracy, convergence.

C. Spiking Variational Autoencoder (SVAE)

SVAE introduces an event-driven learning mechanism that naturally maps to neuromorphic hardware [19]. At low shot counts (10^1–10^2), fidelity rapidly declines beyond 4 qubits due to statistical sparsity, but at 10^5–10^6 shots, SVAE attains $F > 0.9$ up to 6 qubits while maintaining inference time within hundreds of seconds. This balance of data efficiency and low energy cost makes SVAE a strong candidate for real-time, embedded QST.

Authorized licensed use limited to: TU Delft Library. Downloaded on June 08,2026 at 12:46:31 UTC from IEEE Xplore. Restrictions apply.

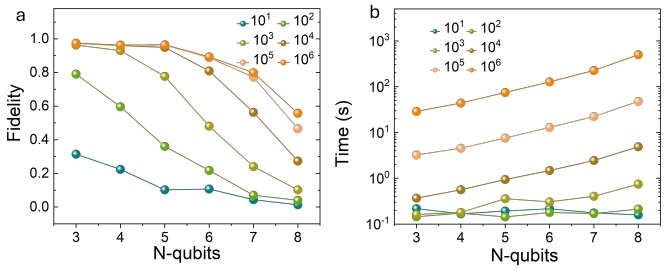


Fig. 3. Performance scaling of the Spiking Variational Autoencoder (SVAE) model. (a) Reconstruction fidelity as a function of the number of qubits for different total measurement shots, showing decreasing accuracy with increasing system size and shot noise. (b) Corresponding computational time demonstrating sub-exponential growth, where the event-driven learning mechanism in SVAE enables efficient scaling for larger quantum systems.

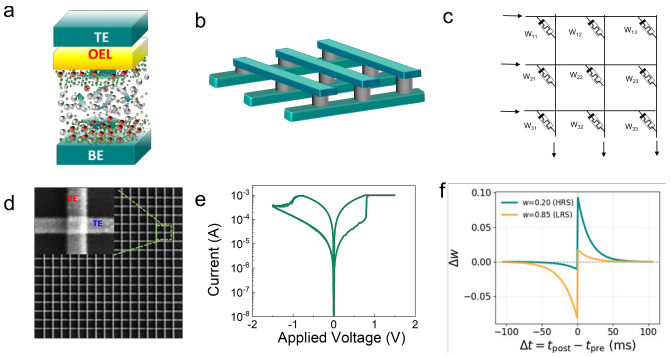


Fig. 4. Overview of memristor device and neuromorphic functionality. (a) Device structure, (b) crossbar array, and (c) analog MVM operation. (d) SEM image of fabricated array, (e) bipolar I–V switching, and (f) spike-timing-dependent plasticity (STDP) behavior demonstrating synaptic weight modulation in high- and low-resistance states.

D. Hardware Co-Design Perspective

Memristor-based computation-in-memory (CiM) architectures provide direct hardware acceleration for CNN and SVAE layers that perform dense matrix–vector multiplication (MVM) operations, as illustrated in Fig. 4(c). Each memristive cell (Fig 4(a)) acts as an adjustable conductance element within a crossbar array (Fig 4(b)), enabling analog MVM execution through Ohm's and Kirchhoff's laws (Fig 4(c)). The fabricated HfO_2-based arrays (Fig 4(d)) exhibit stable bipolar switching characteristics (Fig 4(e)), supporting synaptic learning rules such as spike-timing-dependent plasticity (STDP) (Fig 4(f)). These device-level behaviors form the physical basis for implementing neural network primitives directly in hardware. Compared with GPU-based inference, memristor crossbars can reduce energy consumption by over two orders of magnitude while achieving single-cycle analog computation [15]–[17]. Their intrinsic non-volatility, high density, and massive parallelism further enable scalable deployment for large-qubit quantum state tomography (QST) tasks. Future work will experimentally integrate CNN/SVAE networks on the fabricated HfO_2 crossbars to benchmark the achievable trade-offs in energy efficiency, latency, and reconstruction fidelity.

E. Summary of Key Findings

- CNN and CGAN: highest fidelity (> 0.99) and stable convergence.
- M1 measurements: experimentally preferred for mixed states.
- SVAE: hardware-efficient spiking generative model ($F > 0.9$ up to 6 qubits).
- Memristor-CiM acceleration: $>100\times$ projected energy reduction and scalable on-chip learning.

IV. Conclusion

This study benchmarks seven neural network architectures for QST and proposes memristor-based CiM as an energy-efficient acceleration platform. CNN and CGAN demonstrate the best trade-off between fidelity and computation time, while SVAE provides a neuromorphic path toward embedded quantum diagnostics. The integration of hardware-aware learning with analog CiM computation paves the way for scalable, real-time QST in next-generation quantum devices.

References

[1] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge Univ. Press, 2010.

[2] P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM Review*, vol. 41, pp. 303–332, 1999.

[3] E. Farhi, J. Goldstone, and S. Gutmann, "A quantum approximate optimization algorithm," *arXiv:1411.4028*, 2014.

[4] S. Lloyd, "Universal quantum simulators," *Science*, vol. 273, pp. 1073–1078, 1996.

[5] M. Paris and J. Řeháček, *Quantum State Estimation*. Springer, 2004.

[6] D. Gross *et al.*, "Quantum state tomography via compressed sensing," *Phys. Rev. Lett.*, vol. 105, p. 150401, 2010.

[7] M. Cramer *et al.*, "Efficient quantum state tomography," *Nature Commun.*, vol. 1, p. 149, 2010.

[8] G. Torlai *et al.*, "Neural-network quantum state tomography," *Nature Phys.*, vol. 14, pp. 447–450, 2018.

[9] J. Carrasquilla *et al.*, "Reconstructing quantum states with generative models," *Nature Machine Intelligence*, vol. 1, pp. 155–161, 2019.

[10] S. Lohani *et al.*, "Machine learning assisted quantum state estimation," *Machine Learning: Science and Technology*, vol. 1, p. 035007, 2020.

[11] T. Schmale, M. Reh, and M. Gärttner, "Efficient quantum state tomography with convolutional neural networks," *npj Quantum Information*, vol. 8, p. 115, 2022.

[12] S. Ahmed, C. S. Muñoz, F. Nori, and A. F. Kockum, "Quantum state tomography with conditional generative adversarial networks," *Phys. Rev. Lett.*, vol. 127, p. 140502, 2021.

[13] C. Chen, Z. He, Z. Huang, and H. Situ, "Reconstructing a quantum state with a variational autoencoder," *Int. J. Quantum Inf.*, vol. 19, p. 2140005, 2021.

[14] J. Backus, "Can programming be liberated from the von Neumann style? A functional style and its algebra of programs," *Commun. ACM*, vol. 21, pp. 613–641, 1978.

[15] M. A. Zidan, J. P. Strachan, and W. D. Lu, "The future of electronics based on memristive systems," *Nature Electronics*, vol. 1, pp. 22–29, 2018.

[16] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature Electronics*, vol. 1, pp. 333–343, 2018.

[17] P. Yao *et al.*, "Fully hardware-implemented memristor convolutional neural network," *Nature*, vol. 577, pp. 641–646, 2020.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] H. Kamata, K. Yonekawa, S. Fujimoto, S. Ishida, and R. Kurata, "Fully Spiking Variational Autoencoder," in *Proc. AAAI Conf. on Artificial Intelligence*, 2022, pp. 1–9. Available: https://arxiv.org/abs/2110.00375

Authorized licensed use limited to: TU Delft Library. Downloaded on June 08,2026 at 12:46:31 UTC from IEEE Xplore. Restrictions apply.