Scalable Neural Decoders for Practical Fault-Tolerant Quantum Computation

Abstract

Quantum error correction (QEC) requires classical decoders that are fast and accurate enough to keep pace with quantum hardware. While quantum low-density parity-check (qLDPC) codes have recently emerged as a promising route to efficient fault tolerance, existing decoders fall far short of their capabilities. Here we introduce Cascade, a convolutional neural-network decoder that exploits the geometric structure of QEC codes, and use it to probe a "waterfall" regime of error suppression. On the $[\![144, 12, 12]\!]$ Gross code, Cascade achieves logical error rates up to $\sim 17\times$ below existing decoders—reaching $\sim 10^{-10}$ at physical error $p = 0.1\%$—with 3–5 orders of magnitude higher throughput. These results reveal that the dominant failure modes lie well above the minimum distance, so the logical error rate drops much faster than conventional resource estimates predict. Cascade also produces well-calibrated confidence estimates that reduce the time overhead of repeat-until-success protocols by $\sim 20\times$. Together, these results suggest that the space-time costs of fault-tolerant quantum computation may be significantly lower than previously anticipated.

1 Introduction

Quantum error-correcting codes suppress errors exponentially in the code size, provided the physical error rate lies below a critical threshold Aharonov and Ben-Or [2008]. Realizing this in practice requires a classical decoder that is simultaneously fast and accurate enough to interpret stabilizer measurements in real time. For quantum LDPC codes, which offer high encoding rates and large distances Leverrier and Zemor [2022], Panteleev and Kalachev [2022], Bravyi et al. [2024], no existing decoder meets both requirements: belief propagation (BP) suffers from convergence failures due to quantum degeneracy Poulin and Chung [2008], Raveendran and Vasić [2021], while more accurate variants such as BP with ordered-statistics decoding (BP+OSD) Roffe et al. [2020] and the tree-search decoder Tesseract Beni et al. [2025] are too slow for real-time use. The absence of an efficient decoder also obscures the true power of these codes: resource estimates typically model logical error rates using the threshold and code distance d, assuming minimum-weight uncorrectable errors dominate Gidney and Ekerå [2021], Beverland et al. [2022]. In classical LDPC codes, this is known to be conservative Richardson and Urbanke [2001, 2009].

Here we introduce **Cascade**, a convolutional neural-network decoder that exploits three geometric regularities of QEC codes—locality, translation equivariance, and anisotropy—properties that make code geometry a natural fit for convolutional architectures with learnable, direction-dependent filters. We show that error suppression on both bivariate bicycle (BB) codes Bravyi et al. [2024] and surface codes Fowler et al. [2012] significantly exceeds the naive distance scaling, with a two-regime structure—a steep "waterfall" below threshold reverting to distance-limited scaling at very low noise—that is well established for classical LDPC codes Richardson and Urbanke [2001, 2009]. Indications of similar behavior have appeared in quantum codes Watson and Barrett [2014], Kasai et al. [2012], Komoto and Kasai [2025], but existing decoders lack the accuracy to sharply resolve it in the practically relevant below-threshold regime; Cascade does, quantifying the effect on the Gross code and on surface codes up to $d = 19$. Cascade also produces well-calibrated confidence estimates enabling efficient post-selection, and together these results reduce both the space and time costs of fault-tolerant quantum computation.

Submitted to the 2026 Conference on Physics and AI (PAI26). Do not distribute.

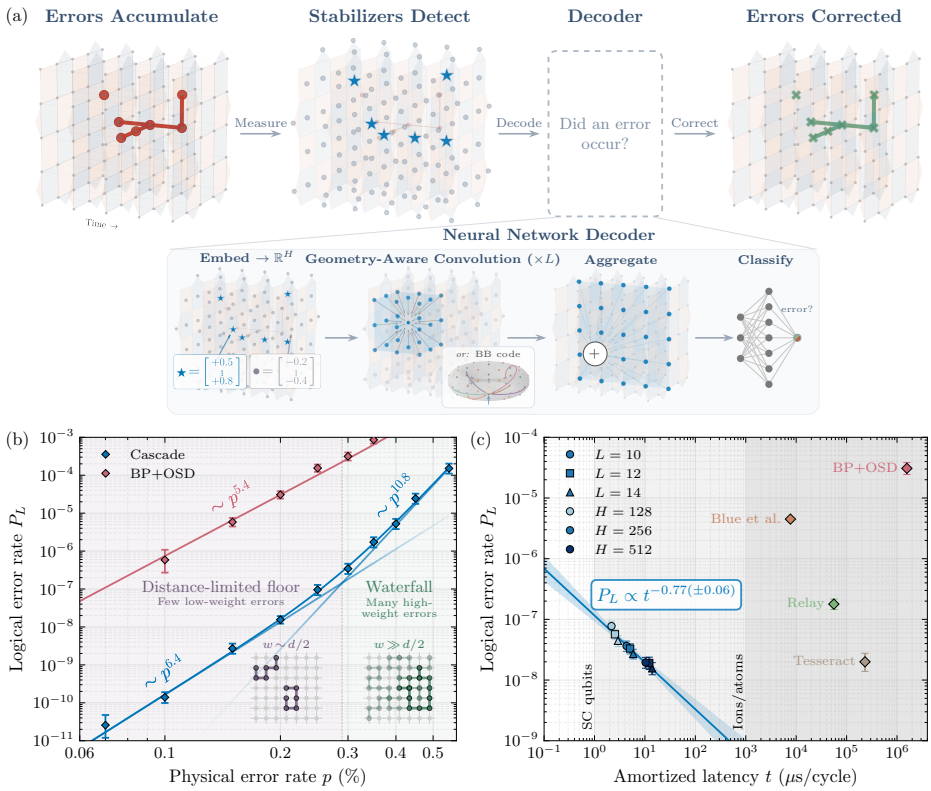

Figure 1: **Structure-aware neural decoding for quantum error correction.** (a) Top: errors accumulate on data qubits; stabilizer measurements produce a spacetime syndrome; the decoder determines whether a logical error occurred. Bottom (Cascade): syndromes are embedded into H-dimensional representations and processed by L convolutional layers whose local structure respects the code geometry—3D convolutions for surface codes, generalized convolutions on the torus for BB codes—followed by a final scatter to data qubits, pooling over each logical operator's support, and a prediction head. (b) Error suppression on the $[\![144, 12, 12]\!]$ BB code ($R = d$ rounds). The logical error rate $P_L \approx \sum_w N(w)\, p^w$ decomposes into two regimes: a steep waterfall ($\sim p^{10.8}$) where high-weight failure modes dominate, and a distance-limited floor ($\sim p^{6.4}$) at very low noise. BP+OSD (orange, $\sim p^{5.4}$) misses the waterfall entirely. (c) Accuracy–latency tradeoff at $p = 0.2\%$ for the Gross code: Cascade (GPU inference) spans a range of amortized latencies while achieving lower logical error rates than BP+OSD Roffe et al. [2020], Relay Müller et al. [2025], Tesseract Beni et al. [2025], and the ML decoder of Blue et al. [2025] (single-threaded CPU). Error bars indicate 95% credible intervals.

2 Structure-aware neural decoding

In a QEC memory experiment, errors accumulate on data qubits during repeated rounds of stabilizer measurement, producing a spacetime syndrome that the decoder must map to a logical-error class. Both code families we consider have regular geometric structure—surface codes on a 2D grid, BB codes on a torus $\mathbb{Z}_\ell \times \mathbb{Z}_m$—which under repeated measurement extends to a spatiotemporal lattice where every stabilizer has the same local connectivity up to translation. This motivates three design principles. *Locality*: errors produce localized syndrome patterns that can be progressively resolved by layers of local processing—successive layers implement a form of coarse-graining Duclos-Cianci and Poulin [2010], where each layer absorbs errors at its characteristic scale. *Translation equivariance*: the identical local structure at every site means the same decoding rules should apply everywhere. *Anisotropy*: information arriving from different directions carries distinct meaning—for instance, a surface code stabilizer's horizontal and diagonal neighbors are of different type (X vs. Z), and the temporal direction encodes measurement errors rather than data-qubit errors. Belief propagation satisfies only locality; its fixed update rules treat the graph symmetrically, causing it to converge to

incorrect solutions on degenerate error patterns Poulin and Chung [2008], Raveendran and Vasić [2021]. Transformer-based decoders such as AlphaQubit Bausch et al. [2024], Senior et al. [2025] use global attention, achieving near-optimal accuracy on surface codes but at much higher training cost and without application to high-rate qLDPC codes; graph neural network decoders Gong et al. [2024], Ninkovic et al. [2024] treat neighbors symmetrically, sacrificing anisotropy. Cascade replaces these fixed or permutation-symmetric rules with learned convolutional operations that satisfy all three properties.

Concretely, binary detection events are embedded into H-dimensional representations at each syndrome location, then processed by L bottleneck residual blocks. Each block performs learned, direction-dependent message passing: a 3D convolution over the spacetime lattice for surface codes, or a generalized convolution on the torus with weights indexed by the relative offset between stabilizers for BB codes. Convolutional weights are shared across positions, applying identical learned rules at every site while using direction-specific weights to distinguish information from different neighbors. Successive layers expand the receptive field, so that after $L \sim d$ layers the network integrates information across the full code distance. After the final layer, representations are scattered to data qubits, pooled over the support of each logical operator, and passed to a prediction head. Models are trained end-to-end with binary cross-entropy at a single high physical error rate ($\sim 0.7\%$ for surface codes, $\sim 0.55\%$ for BB codes) on synthetic data from Stim Gidney [2021]; training converges within $\sim 3 \times 10^8$ samples, and the largest surface-code decoder ($d = 19$) trains in ~ 200 H200 GPU hours.

3 Results

The waterfall regime. Figure 1(b) shows the logical error rate per logical qubit per cycle for the $[\![144, 12, 12]\!]$ Gross code (a high-rate qLDPC code encoding $k = 12$ logical qubits in $n = 144$ physical qubits at distance $d = 12$) under circuit-level depolarizing noise, in which every gate, measurement, and idle step introduces independent errors at rate p. Our largest model ($L = 14$, $H = 512$) achieves logical error rates orders of magnitude below BP+OSD Roffe et al. [2020] and Relay Müller et al. [2025] across the full noise range. Below threshold, P_L drops far more steeply than the naive distance scaling predicts, and decomposes into a steep waterfall term $\sim p^{11}$ that dominates at moderate rates and a distance-limited floor $\sim p^{6.4}$ that emerges at very low noise. The origin is combinatorial: to leading order, $P_L \approx \sum_w N(w)\, p^w$, where $N(w)$ counts minimal weight-w failure modes. For the codes studied here, $N(w)$ is heavily concentrated well above $\lfloor d/2 + 1 \rfloor$, so at moderate p the abundant high-weight modes dominate despite the extra suppression; the minimum-weight modes only take over at very low noise.

The set of *minimal* failure modes depends on the decoder: a more powerful decoder eliminates failures at low weight, exposing the regime where higher-weight modes dominate; BP+OSD fails on too many low-weight errors, so the steep suppression is never exposed. Cascade reaches logical error rates $\sim 4000\times$ below BP+OSD and $\sim 17\times$ below Relay on the Gross code at $p = 0.1\%$, with accuracy comparable to the near-optimal (but impractically slow) Tesseract Beni et al. [2025]. We observe no error floor: exponential suppression persists to $P_L \approx 2 \times 10^{-11}$, the lowest rate tested, evidence that the learned rules avoid systematic misclassification of low-weight patterns (a failure mode to which neural decoders are in principle susceptible). This advantage is not specific to a single code size: Cascade outperforms all reference decoders across BB codes from $d = 6$ to $d = 18$, and on the largest $[\![288, 12, 18]\!]$ code reaches $P_L \sim 10^{-10}$ per logical qubit per cycle at $p = 0.2\%$—a regime in which BP+OSD and Tesseract are either intractable or uninformative.

Surface codes and resource implications. The same architecture, instantiated with 3D convolutions, achieves an error suppression factor $\Lambda \approx 8.4$ on surface codes at $p = 0.2\%$ from a fit of $P_L \propto \Lambda^{-\lfloor (d+1)/2 \rfloor}$ across distances $d = 7$ to 19 (Fig. 2). This substantially exceeds MWPM ($\Lambda \approx 5.0$) and correlated MWPM ($\Lambda \approx 7.8$), and approaches the near-optimal (but impractically slow) Tesseract ($\Lambda \approx 9.1$). The nearly twofold variation—from 5.0 for MWPM to 9.1 for Tesseract—exceeds what threshold differences alone can account for and is a signature of the same waterfall effect: at moderate distances the effective slope is steeper than $\lfloor (d+1)/2 \rfloor$, and the decoder's ability to handle the more numerous higher-weight errors determines how much of this steeper suppression is realized. The practical consequence is direct: projections such as Gidney and Ekerå [2021], Beverland et al. [2022] assume $\Lambda \approx 10$ at $p = 0.1\%$, calibrated to MWPM-class decoders;

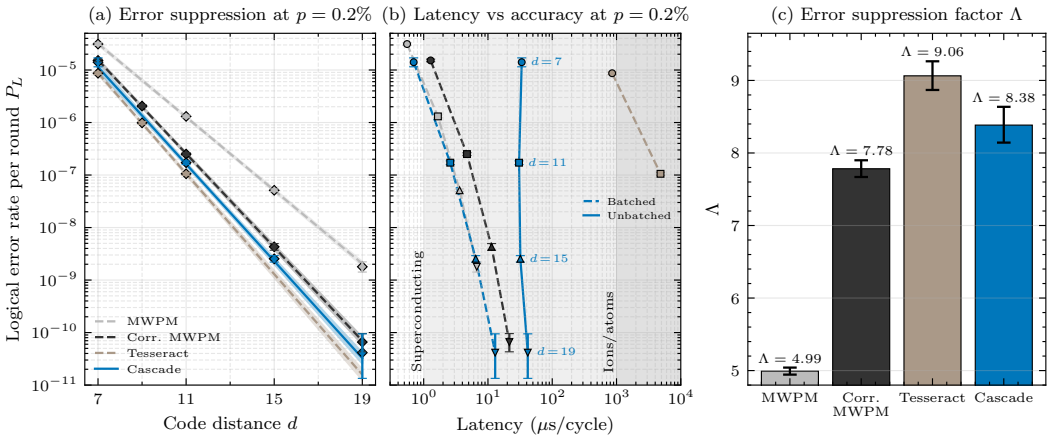

Figure 2: **Distance scaling of surface code decoders at $p = 0.2\%$ under circuit-level depolarizing noise.** (a) Logical error rate per round versus code distance d for MWPM, correlated MWPM, Tesseract, and Cascade. All decoders exhibit exponential suppression ($P_L \propto \Lambda^{-\lfloor (d+1)/2 \rfloor}$), with Cascade and Tesseract achieving the steepest slopes. (b) Accuracy–latency tradeoff across code distances (GPU inference). Amortized latency (dashed, batched) is significantly lower than single-shot latency (solid, unbatched). (c) Error suppression factor Λ extracted from exponential fits to panel (a).

Cascade reaches $P_L \sim 10^{-9}$ at $d = 15$, versus $d = 19$ for MWPM—a $\sim$40% reduction in physical qubit count. Because the waterfall gives steeper-than-distance scaling, this advantage *widens* as targets tighten toward the 10^{-10}–10^{-12} regime required for utility-scale algorithms Gidney and Ekerå [2021], Campbell et al. [2017].

Speed, generalization, and confidence calibration. Cascade has a single-shot latency of $\sim$40 μs per cycle on a single NVIDIA H200 GPU, with batched inference up to two orders of magnitude lower—an overall throughput 3,000–100,000$\times$ higher than existing decoders (single-threaded CPU). These latencies are well within the $\sim$1 ms decoding budgets for trapped-ion and neutral-atom platforms Bluvstein et al. [2024]. The convolutional architecture is also amenable to aggressive quantization: folding batch normalization into convolution weights and applying post-training FP8 inference preserves accuracy across the full noise range with no quantization-aware retraining, yielding an $\sim$16$\times$ reduction in multiplier area relative to FP32. Combined with depthwise factorization, roofline estimates suggest FLOP-efficient variants on FPGA/ASIC hardware could approach the $\sim$1 μs budget for superconducting qubits. All models are trained at a single noise level yet generalize reliably across seven orders of magnitude in P_L, and their confidence estimates remain well-calibrated far below the training distribution. This enables efficient post-selection: on the $[\![72, 12, 6]\!]$ BB code at $p = 0.55\%$, Cascade reaches an error rate $\sim 2 \times 10^{-3}$ at a $\sim$95% acceptance rate, compared with $\sim$5% for cluster-based post-selection Lee et al. [2025]—a $\sim$20$\times$ reduction in the number of retries required by repeat-until-success protocols Smith et al. [2024], Campbell et al. [2017], with even a modest 0.5% per-cycle discard rate yielding up to two orders of magnitude reduction in P_L.

Outlook. Code design and resource estimation must move beyond code distance as the sole figure of merit, and the decoder should be treated as an integral component of fault-tolerant architecture rather than an independent subsystem. Cascade's geometric inductive bias is general: the same architecture works on surface codes and BB codes and applies directly to any stabilizer code whose checks share a common local structure—color codes Bombin and Martin-Delgado [2006], hyperbolic surface codes Breuckmann and Eberhardt [2021], lifted-product Xu et al. [2024], and Kasai codes Kasai [2026]—where the waterfall effect should be even more pronounced at larger distances. A natural next step is composing Cascade with recent neural circuit-level decoders Zhou et al. [2025], Ataides et al. [2025] for end-to-end fault tolerance. These results are timely: current hardware has reached entangling error rates $\sim$0.1%, where the waterfall enables logical error rates sufficient for practical fault-tolerant algorithms.

References

Dorit Aharonov and Michael Ben-Or. Fault-Tolerant Quantum Computation with Constant Error Rate. *SIAM Journal on Computing*, 38(4):1207–1282, 1 2008. ISSN 0097-5397. doi: 10.1137/S0097539799359385.

Anthony Leverrier and Gilles Zemor. Quantum Tanner codes. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 872–883. IEEE, 10 2022. doi: 10.1109/FOCS54457.2022.00117.

Pavel Panteleev and Gleb Kalachev. Quantum ldpc codes with almost linear minimum distance. *IEEE Transactions on Information Theory*, 68(1):213–229, 2022. doi: 10.1109/TIT.2021.3119384.

Sergey Bravyi, Andrew W. Cross, Jay M. Gambetta, Dmitri Maslov, Patrick Rall, and Theodore J. Yoder. High-threshold and low-overhead fault-tolerant quantum memory. *Nature*, 627(8005): 778–782, 3 2024. ISSN 0028-0836. doi: 10.1038/s41586-024-07107-7.

D. Poulin and Y. Chung. On the iterative decoding of sparse quantum codes. *Quantum Information and Computation*, 8(10):986–1000, 11 2008. ISSN 1533-7146. doi: 10.26421/QIC8.10-8.

Nithin Raveendran and Bane Vasić. Trapping sets of quantum ldpc codes. *Quantum*, 5:562, 10 2021. ISSN 2521-327X. doi: 10.22331/q-2021-10-14-562.

Joschka Roffe, David R. White, Simon Burton, and Earl Campbell. Decoding across the quantum low-density parity-check code landscape. *Physical Review Research*, 2(4):043423, 12 2020. ISSN 2643-1564. doi: 10.1103/PhysRevResearch.2.043423.

Laleh Aghababaie Beni, Oscar Higgott, and Noah Shutty. Tesseract: A search-based decoder for quantum error correction. 3 2025. URL `http://arxiv.org/abs/2503.10988`.

Craig Gidney and Martin Ekerå. How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits. *Quantum*, 5:433, 4 2021. ISSN 2521-327X. doi: 10.22331/q-2021-04-15-433.

Michael E. Beverland, Prakash Murali, Matthias Troyer, Krysta M. Svore, Torsten Hoefler, Vadym Kliuchnikov, Guang Hao Low, Mathias Soeken, Aarthi Sundaram, and Alexander Vaschillo. Assessing requirements to scale to practical quantum advantage. 11 2022. URL `http://arxiv.org/abs/2211.07629`.

T. J. Richardson and R. L. Urbanke. The capacity of low-density parity-check codes under message-passing decoding. *IEEE Transactions on Information Theory*, 47(2):599–618, 2 2001. ISSN 0018-9448. doi: 10.1109/18.910577.

Thomas J. Richardson and Rüdiger Leo Urbanke. *Modern coding theory*. Cambridge University Press, Cambridge, 2009.

Austin G. Fowler, Matteo Mariantoni, John M. Martinis, and Andrew N. Cleland. Surface codes: Towards practical large-scale quantum computation. *Physical Review A*, 86(3):032324, 9 2012. ISSN 1050-2947. doi: 10.1103/PhysRevA.86.032324.

Fern H E Watson and Sean D Barrett. Logical error rate scaling of the toric code. *New Journal of Physics*, 16(9):093045, sep 2014. ISSN 1367-2630. doi: 10.1088/1367-2630/16/9/093045. URL `http://dx.doi.org/10.1088/1367-2630/16/9/093045`.

Kenta Kasai, Manabu Hagiwara, Hideki Imai, and Kohichi Sakaniwa. Quantum error correction beyond the bounded distance decoding limit. *IEEE Transactions on Information Theory*, 58(2): 1223–1230, feb 2012. ISSN 1557-9654. doi: 10.1109/tit.2011.2167593. URL `http://dx.doi.org/10.1109/TIT.2011.2167593`.

Daiki Komoto and Kenta Kasai. Quantum error correction near the coding theoretical bound. 12 2025. URL `http://arxiv.org/abs/2412.21171`.

Tristan Müller, Thomas Alexander, Michael E. Beverland, Markus Bühler, Blake R. Johnson, Thilo Maurer, and Drew Vandeth. Improved belief propagation is sufficient for real-time decoding of quantum memory. 6 2025. URL `http://arxiv.org/abs/2506.01779`.

John Blue, Harshil Avlani, Zhiyang He, Liu Ziyin, and Isaac L. Chuang. Machine learning decoding of circuit-level noise for bivariate bicycle codes. 4 2025. URL http://arxiv.org/abs/2504.13043.

Guillaume Duclos-Cianci and David Poulin. Fast decoders for topological quantum codes. *Physical Review Letters*, 104(5), 2 2010. ISSN 1079-7114. doi: 10.1103/physrevlett.104.050504.

Johannes Bausch, Andrew W. Senior, Francisco J. H. Heras, Thomas Edlich, Alex Davies, Michael Newman, Cody Jones, Kevin Satzinger, Murphy Yuezhen Niu, Sam Blackwell, George Holland, Dvir Kafri, Juan Atalaya, Craig Gidney, Demis Hassabis, Sergio Boixo, Hartmut Neven, and Pushmeet Kohli. Learning high-accuracy error decoding for quantum processors. *Nature*, 635 (8040):834–840, 11 2024. ISSN 0028-0836. doi: 10.1038/s41586-024-08148-8.

Andrew W. Senior, Thomas Edlich, Francisco J. H. Heras, Lei M. Zhang, Oscar Higgott, James S. Spencer, Taylor Applebaum, Sam Blackwell, Justin Ledford, Akvilė Žemgulytė, Augustin Žídek, Noah Shutty, Andrew Cowie, Yin Li, George Holland, Peter Brooks, Charlie Beattie, Michael Newman, Alex Davies, Cody Jones, Sergio Boixo, Hartmut Neven, Pushmeet Kohli, and Johannes Bausch. A scalable and real-time neural decoder for topological quantum codes. 12 2025. URL http://arxiv.org/abs/2512.07737.

Anqi Gong, Sebastian Cammerer, and Joseph M. Renes. Graph neural networks for enhanced decoding of quantum ldpc codes. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 2700–2705, 2024. doi: 10.1109/ISIT57864.2024.10619589.

Vukan Ninkovic, Ognjen Kundacina, Dejan Vukobratovic, Christian Häger, and Alexandre Graell i Amat. Decoding quantum ldpc codes using graph neural networks. 8 2024. URL http://arxiv.org/abs/2408.05170.

Craig Gidney. Stim: a fast stabilizer circuit simulator. *Quantum*, 5:497, 7 2021. ISSN 2521-327X. doi: 10.22331/q-2021-07-06-497.

Earl T. Campbell, Barbara M. Terhal, and Christophe Vuillot. Roads towards fault-tolerant universal quantum computation. *Nature*, 549(7671):172–179, 9 2017. ISSN 0028-0836. doi: 10.1038/nature23460.

Dolev Bluvstein, Simon J. Evered, Alexandra A. Geim, Sophie H. Li, Hengyun Zhou, Tom Manovitz, Sepehr Ebadi, Madelyn Cain, Marcin Kalinowski, Dominik Hangleiter, J. Pablo Bonilla Ataides, Nishad Maskara, Iris Cong, Xun Gao, Pedro Sales Rodriguez, Thomas Karolyshyn, Giulia Semeghini, Michael J. Gullans, Markus Greiner, Vladan Vuletić, and Mikhail D. Lukin. Logical quantum processor based on reconfigurable atom arrays. *Nature*, 626(7997):58–65, 2 2024. ISSN 0028-0836. doi: 10.1038/s41586-023-06927-3.

Seok-Hyung Lee, Lucas English, and Stephen D. Bartlett. Efficient post-selection for general quantum ldpc codes. 10 2025. URL http://arxiv.org/abs/2510.05795.

Samuel C. Smith, Benjamin J. Brown, and Stephen D. Bartlett. Mitigating errors in logical qubits. *Communications Physics*, 7(1), 11 2024. ISSN 2399-3650. doi: 10.1038/s42005-024-01883-4.

Hector Bombin and Miguel Angel Martin-Delgado. Topological quantum distillation. *Physical Review Letters*, 97(18):180501, 2006. doi: 10.1103/PhysRevLett.97.180501.

Nikolas P. Breuckmann and Jens Niklas Eberhardt. Quantum low-density parity-check codes. *PRX Quantum*, 2(4), 10 2021. ISSN 2691-3399. doi: 10.1103/prxquantum.2.040101.

Qian Xu, J. Pablo Bonilla Ataides, Christopher A. Pattison, Nithin Raveendran, Dolev Bluvstein, Jonathan Wurtz, Bane Vasić, Mikhail D. Lukin, Liang Jiang, and Hengyun Zhou. Constant-overhead fault-tolerant quantum computation with reconfigurable atom arrays. *Nature Physics*, 20 (7):1084–1090, jul 2024. doi: 10.1038/s41567-024-02479-z.

Kenta Kasai. Breaking the orthogonality barrier in quantum ldpc codes. 1 2026. URL http://arxiv.org/abs/2601.08824.

Yiqing Zhou, Chao Wan, Yichen Xu, Jin Peng Zhou, Kilian Q. Weinberger, and Eun-Ah Kim. Learning to decode logical circuits. *Nature Computational Science*, 5(12):1158–1167, November 2025. ISSN 2662-8457. doi: 10.1038/s43588-025-00897-4. URL https://www.nature.com/articles/s43588-025-00897-4.

J. Pablo Bonilla Ataides, Andi Gu, Susanne F. Yelin, and Mikhail D. Lukin. Neural decoders for universal quantum algorithms, 2025. URL https://arxiv.org/abs/2509.11370.