

Rigorous noise reduction with quantum autoencoders

Wai-Keong Mok^{*, 1,2} Hui Zhang^{*, 3,4} Tobias Haug^{*, 5} Xianshu Luo,⁶ Guo-Qiang Lo,⁶ Hong Cai,⁷ M. S. Kim,⁵ Ai Qun Liu,^{3,4} and Leong-Chuan Kwek^{2, 8, 9, 4}

¹*California Institute of Technology, Pasadena, CA 91125, USA*

²*Centre for Quantum Technologies, National University of Singapore, 3 Science Drive 2, Singapore 117543**

³*Institute of Quantum Technologies (IQT), The Hong Kong Polytechnic University, Hong Kong*

⁴*Quantum Science and Engineering Centre (QSec), Nanyang Technological University, Singapore**

⁵*QOLS, Blackett Laboratory, Imperial College London SW7 2AZ, UK**

⁶*Advanced Micro Foundry, 11 Science Park Rd, Singapore*

⁷*Institute of Microelectronics, A*STAR (Agency for Science, Technology and Research), Singapore*

⁸*MajuLab, CNRS-UNS-NUS-NTU International Joint Research Unit, Singapore UMI 3654, Singapore*

⁹*National Institute of Education, Nanyang Technological University, Singapore 637616, Singapore*

Reducing noise in quantum systems is a major challenge towards the application of quantum technologies. Here, we propose and demonstrate a scheme to reduce noise using a quantum autoencoder with rigorous performance guarantees. The quantum autoencoder learns to compresses noisy quantum states into a latent subspace and removes noise via projective measurements. We find various noise models where we can perfectly reconstruct the original state even for high noise levels. We apply the autoencoder to cool thermal states to the ground state and reduce the cost of magic state distillation by several orders of magnitude. Our autoencoder can be implemented using only unitary transformations without ancillas, making it immediately compatible with the state of the art. We experimentally demonstrate our methods to reduce noise in a photonic integrated circuit. Our results can be directly applied to make quantum technologies more robust to noise.

INTRODUCTION

Quantum technologies offer potential advantages in quantum computing [1], quantum communication [2] and metrology [3]. However, quantum systems are brittle by nature, and noise due to the environment and imperfect control over the quantum system negatively impacts the capabilities of quantum devices. Thus, techniques to remove or reduce noise is the key challenge that needs to be addressed for quantum technologies to be successful [4, 5]. To this end, a wide range of noise reduction techniques have been developed.

In the context of fault-tolerant quantum computers [6], magic state distillation (MSD) requires multiple copies of a noisy state to create one state with reduced noise. MSD is the most expensive process required to run fault-tolerant quantum computers [7] and it is imperative to substantially reduce the cost of MSD to make early fault-tolerant quantum computers practically viable [8–14].

An alternative path to reduce noise are quantum autoencoders. Quantum autoencoders transform quantum states into a smaller subspace that contains the essential features of the state, while discarding redundant features [15–22]. Quantum autoencoders are amenable to noisy quantum devices [5] which makes quantum autoencoders particularly useful for enhancing quantum technologies in experiments [21, 23–26]. Experiments have demonstrated loss compression of quantum data in bulky optics systems [23, 24] and the use of autoencoders for compression-assisted teleportation of high-dimensional quantum states in integrated photonic chips [21].

Recently, quantum autoencoders have been proposed to reduce noise [27–33]. One variant reduces noise by transferring

the noisy part of the state into ancillas and tracing out the ancilla. This approach allows for deterministic noise removal, however it commonly requires deep circuits and many ancilla qubits [27]. Alternatively, projective measurements and post-selection can reduce noise with low resource requirements and without ancillas [28]. However, existing proposals provide mainly numerical evidence for the performance in practical applications. A quantum autoencoder to denoise quantum states is yet to be experimentally demonstrated.

Here, we experimentally implement a photonic chip integrated autoencoder to reduce the noise of quantum states with rigorous performance guarantees. Our scheme compresses quantum states into a latent subspace and removes noise by projective measurements and post-selection on successful outcomes. We train the autoencoder either in an unsupervised manner by minimizing measurement probabilities of noisy input states (*population training*), or maximizing the fidelity in respect to a reference state (*fidelity training*). We analytically study the performance of the protocol and provide rigorous bounds on the denoising fidelity. For various noise models such as perturbation by a fixed state, depolarizing noise or thermal states, we find that our protocol can perfectly recover the noise-free state. We also apply it to cool a thermal state to the ground state. Further, we show that our protocol can decrease the cost of magic state distillation by several orders of magnitudes. Remarkably, this allows for successful distillation at high levels of noise where the conventional protocol fails. The protocol is experimentally realized on an integrated photonic chip, which is scalable and energy-efficient. Our work demonstrates a practical method to reduce noise for immediate applications in quantum technologies.

* These authors contributed equally to this work

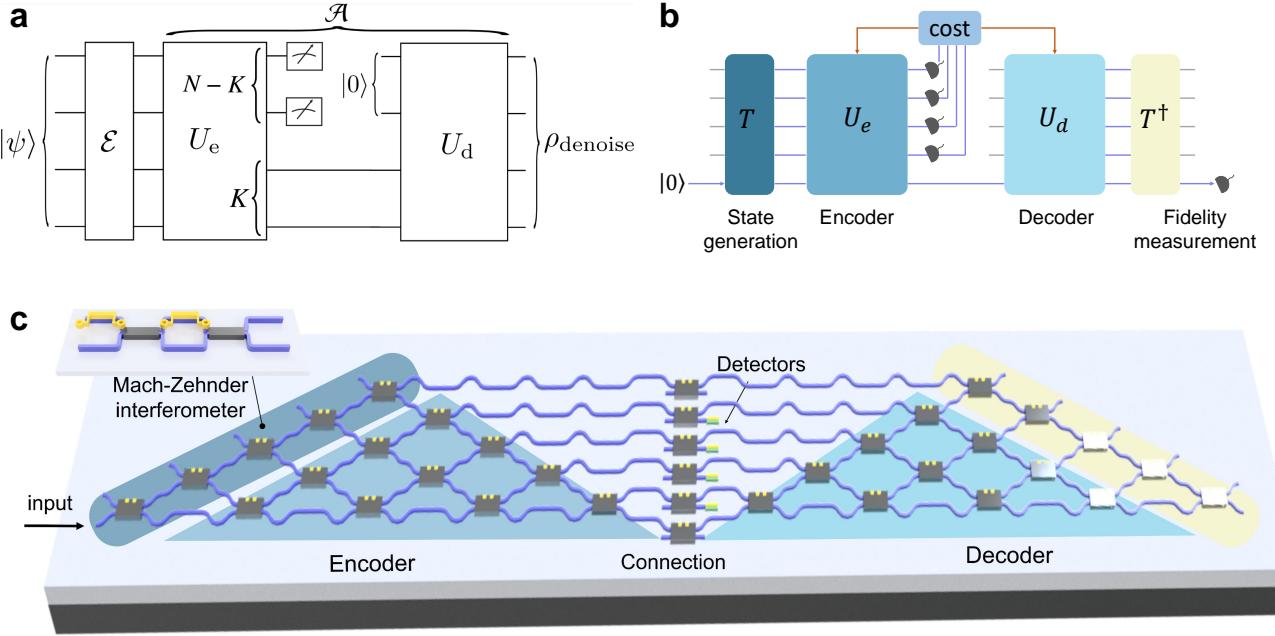


FIG. 1. **Overview of denoising protocol and experimental implementation.** Schematic illustration of the autoencoder implemented on an integrated photonic chip for denoising quantum states. (a) Setup of the denoiser quantum autoencoder (DQA). An N -dimensional quantum state $|\psi\rangle$ is subject to noise channel \mathcal{E} . We denoise the state by encoding into the latent K -dimensional subspace with encoder U_e and projecting out the remaining $N - K$ modes. The denoised state is constructed by decoder U_d . (b) The architecture of the autoencoder, which consists of the state generation unit T , encoder U_e , decoder U_d , and T^\dagger for measuring fidelity. Noise channels can be implemented by probabilistically choosing the state generation unitary T . (c) Design of the integrated photonic chip with $N = 5$, which comprises two 6-by-6 linear optical circuits, one for the encoder and the other for the decoder. A column of Mach-Zehnder interferometers (MZIs) connects the encoder and decoder so that each path can be chosen to enter the decoder or go directly for measurement purposes.

THEORY AND DESIGN

Theory of Autoencoder denoiser. A set of pure N -dimensional quantum states $S = \{|\psi_i\rangle\}_i$ are affected by a noise quantum channel \mathcal{E} . Our goal is to reduce the effect of the noise with an autoencoder \mathcal{A} such that $\mathcal{A}(\mathcal{E}(|\psi\rangle)) \approx |\psi\rangle \forall |\psi\rangle \in S$. A sketch of the denoising protocol is shown in Fig. 1(a). A noisy input state $\rho_{\text{in}} = \mathcal{E}(|\psi\rangle)$ is transformed with the unitary encoder $U_e(\theta)$ with trainable parameters θ . The core idea of our approach is to transform the noise into a $N - K$ dimensional redundant subspace, which is removed with the projective measurement operator $P_K = I_K \oplus 0_{N-K}$, while encoding the pure quantum information into the K -dimensional latent subspace. We post-select instances of successful projections onto P_K which occur with probability

$$G(\rho_{\text{in}}) = \text{tr}(P_K U_e \rho_{\text{in}} U_e^\dagger). \quad (1)$$

Then, we apply the decoder unitary U_d to generate the denoised state. The decoder unitary can be chosen either as the inverse of the encoder $U_d = U_e^\dagger$ or trained as $U_d(\varphi)$ with variational parameters φ . The final denoised state is given by

$$\rho_{\text{denoise}} = \mathcal{A}(\rho_{\text{in}}) = \frac{1}{G(\rho_{\text{in}})} U_d P_K U_e \rho_{\text{in}} U_e^\dagger P_K U_d^\dagger. \quad (2)$$

We quantify the denoiser performance via the average fidelity in respect to the ideal state $|\psi\rangle$ via

$$\bar{F} = \mathbb{E}_{|\psi\rangle \in S} [\langle \psi | \mathcal{A}(\mathcal{E}(|\psi\rangle)) | \psi \rangle]. \quad (3)$$

A summary of the notations used can be found in Appendix A [34]. In the following, we assume that the states in S span a K -dimensional subspace within the N -dimensional space of states where $N > K$. This condition ensures that, in the absence of noise, the autoencoder can achieve unit fidelity. We further assume that the autoencoder receives noisy states uniformly random from set S where the density matrix averaged over random inputs is given by $\rho_S = \mathbb{E}_{|\psi\rangle \in S} [\mathcal{E}(|\psi\rangle)]$. We investigate two possible choices of decoders, which require different ways of training. First, we assume that the decoder $U_d = U_e^\dagger$ is simply the inverse of the encoder. This method, which we call *population training*, uses the measurement probability for the cost function

$$C_T(\theta) = 1 - \text{tr}(P_K U_e(\theta) \rho_S U_e^\dagger(\theta)). \quad (4)$$

This cost function is maximized when incoming states have minimal probability of occupying the redundant $(N - K)$ -dimensional subspace, and equivalently maximal probability in the K -dimensional latent space. This can be trained in an unsupervised manner, i.e., we only require access to the noisy

ensemble ρ_S for training and measurements on the redundant subspace. Minimizing the cost function to the global minima yields the optimal encoder parameters $\theta_* = \operatorname{argmin}_{\theta} C_T(\theta)$. The optimal encoder $U_e(\theta_*)$ rotates the K dominant eigenvectors of ρ_S (with eigenvalues $\lambda_1 \geq \dots \geq \lambda_K$) onto the latent subspace. Thus, the cost function is upper bounded by $C_T \leq \sum_{j=1}^K \lambda_j$ [15, 30]. It can be shown that the optimal decoder is $U_d = U_e^\dagger(\theta_*)$. This protocol has a success probability of $1 - C_T$ arising from postselection. One can think of the trained autoencoder as a projector D_K onto the eigenspace spanned by the K largest eigenvectors of ρ_S [35].

Alternatively, we choose the decoder as $U_d(\varphi)$ to be trained separately from the encoder with its own decoder parameters φ . In this case we perform *fidelity training* to maximize the fidelity between denoised state $\mathcal{A}(\mathcal{E}(|\psi\rangle))$ and the ideal state $|\psi\rangle$ where the cost function is given by

$$C_F(\theta, \varphi) = 1 - \bar{F}(\theta, \varphi) \quad (5)$$

which can be measured with the SWAP test. Fidelity training requires a priori knowledge of the ideal states used for training.

Experimental architecture and chip design. We now describe the experimental implementation of our quantum autoencoder on a photonic chip, which is illustrated in Fig 1(b,c). Our chip models a $N = 5$ dimensional qudit via spatial modes. The chip consists of 5 stages realized by a network of parameterized Mach-Zehnder interferometers (MZIs) and detectors. First, $N = 5$ dimensional input states are generated by the state generation stage. During this stage, we can also implement noise channels by probabilistic choosing the state generation unitary. Then, the encoder stage realizes arbitrary unitaries with controllable circuit parameters θ . Next, up to $K \leq 4$ detectors realize the projective measurement to remove noise. The following decoder stage realizes arbitrary unitaries with controllable parameter φ . To validate the denoised state, fidelity of the denoised state ρ_{denoise} is measured by the compute-uncompute method. Given the noise-free state $|\psi\rangle = T|0\rangle$ with state generation unitary T , the fidelity is measured by applying the inverse T^\dagger on ρ_{denoise} and measuring population of the $|0\rangle$ mode [36].

RESULTS

Subspace denoising. In the most general setting, the noise channel can be written as an arbitrary Kraus map $\mathcal{E}(\rho) = \sum_n M_n \rho M_n^\dagger$ for a set of Kraus operators M_n satisfying $\sum_n M_n^\dagger M_n = I$. In general, obtaining an analytical expression for \bar{F} is difficult. To this end, we now assume that the ideal states S are uniformly sampled from a K -dimensional subspace via the Haar measure on the unitary group $U(K)$.

This allows us to introduce a ‘quenched’ analytical approximation [34, 37]

$$\bar{F}^{(q)} = \frac{\sum_n (|\text{tr}(\Pi_K B M_n)|^2 + \|\Pi_K B M_n \Pi_K\|_F^2)}{(K+1) \sum_n \|B M_n \Pi_K\|_F^2} \quad (6)$$

where Π_K is the projector onto the ideal subspace, $B \equiv U_d P_K U_e$, and $\|\cdot\|_F^2$ denotes the Frobenius norm. For population training, $B = D_K$, and $U_d = U_e^\dagger$. We find that the formula agrees well with the numerical results.

A large class of realistic noise models can be obtained by setting $M_0 = \sqrt{1-p}I$, where p is the noise probability. Thus, $\rho_{\text{in}} = (1-p)|\psi\rangle\langle\psi| + \sum_{n \geq 1} M_n |\psi\rangle\langle\psi| M_n^\dagger$. We prove that $\|D_K - \Pi_K\|_F^2 \leq 2\sqrt{2}Kp/(1-p)$, from which we can show that the denoised state has a worst-case fidelity $\bar{F} \geq \mathbb{E}_{|\psi\rangle \in S} [\langle\psi|\rho_{\text{denoise}}|\psi\rangle] + O(p^2)$ [34]. This implies that for $p \ll 1$ the denoiser is never detrimental.

We now analyze the performance of the denoiser for a more restrictive type of noise channel given by

$$\mathcal{E}(\rho) = (1-p)\rho + p\rho_{\text{noise}} \quad (7)$$

where the state ρ is replaced with some fixed N -dimensional state ρ_{noise} with probability p . Without the denoiser, the average fidelity of the noisy state is $\bar{F}_{\text{bare}} = 1 - p(1-c/K)$ where $c = \text{tr}(\rho_{\text{noise}} \Pi_K)$ is the overlap between the noise state and the ideal subspace. After population training, we calculate the average denoising fidelity $\bar{F}_K \equiv \mathbb{E}_S[F]$ where we find exact results for different choices of ρ_{noise} [34]. For depolarizing noise with $\rho_{\text{noise}} = I_N/N$, population training can achieve $\bar{F}_K = (1-p+p/N)/(1-p+pK/N)$ with post-selection probability $G = 1-p+pK/N$. Remarkably, for $K=1$ we find perfect denoising $F_1 = 1 \forall p \in [0, 1]$.

Next, we assume a pure noise state $\rho_{\text{noise}} = |\psi_{\text{noise}}\rangle\langle\psi_{\text{noise}}|$. To quadratic order in p we find [34]

$$\bar{F}_K(p) = 1 - \frac{K-1}{K}cp - \frac{K(3K-1)-1}{K}c(1-c)p^2 + O(p^3). \quad (8)$$

The denoiser always improves the fidelity if $p \lesssim (3Kc)^{-1}$ for large Kc [34]. We now focus on the case $K=1$. Here, the denoiser suppresses noise completely to first order in p . By applying Theorem 1 of Ref. [38], we obtain analytically a sharp lower-bound for the fidelity of the denoised state for arbitrary pure or mixed ρ_{noise}

$$F_1 \geq \frac{1}{2} \left(1 + \sqrt{1 - p^2(1-p)^{-2}} \right) = 1 - \frac{1}{4}p^2 + O(p^3) \quad (9)$$

which holds for $p \leq 1/2$, whereas we get trivially $F_1 \geq 0$ for $p > 1/2$. This is shown in Fig. 2a. Notice that F_1 tends to a Heaviside step function with a sharp transition at $p = 1/2$ as $c \rightarrow 0$, with the gradient diverging as $\partial_c F_1 \sim c^{-1/2}$. Intuitively, when $c = 0$, any $p > 1/2$ will cause the dominant eigenstate of ρ to be orthogonal to $|\psi\rangle$ resulting in $F_1 = 0$. The experimental data in Fig. 2b shows good agreement with the theoretical predictions. In particular, we plot the case of $c = 0.2$ in Fig. 2c. The denoiser suppresses the noise up to linear order in p . Similar results are observed when using random pure states $|\psi_{\text{noise}}\rangle$ drawn from the Haar measure [34]. For a fixed overlap c , the worst-case denoising fidelity corresponds to a pure ρ_{noise} [34].

Next, we experimentally demonstrate the ability of the denoiser to reduce thermal noise commonly encountered in experiments. The effects of thermal noise is modeled by adding

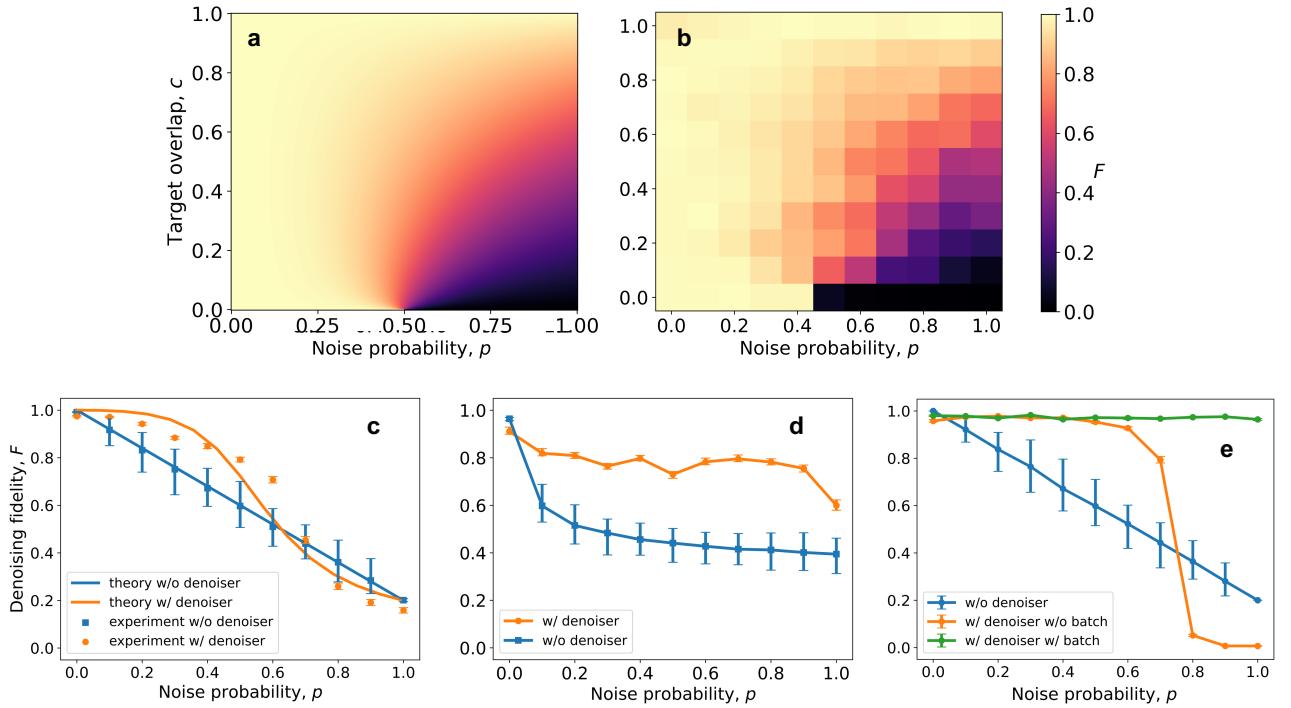


FIG. 2. **Single state denoising, $K = 1$, with population training.** (a-c) Pure state noise, with noise probability p and overlap with the ideal state c . (a) Theoretical denoising fidelity F . (b) Experimental denoising fidelity for a $N = 5$ qudit. (c) Denoising fidelity against p for $c = 0.2$ (corresponding to the boxed values in (b)). (d) Experimental denoising fidelity for thermal noise. (e) Experimental denoising fidelity against p for depolarizing noise. The theoretical value corresponds to $F = 1$ for all $p \in [0, 1]$. The autoencoder has the following two training configurations: In the first configuration, a batch of five noisy states is used to train the autoencoder, ensuring the depolarizing property of the noise. In the second configuration, training is conducted without batches, and training instances are extracted individually from the set of noisy states. In cases where the sample size is not sufficiently large, this configuration might affect the depolarizing property of the noise. Population training is used for the denoiser in all cases.

a Gaussian random phase shift to the modes with zero mean and variance σ^2 , depicted in Fig. 2d. As shown in Fig. 2e, the fidelity against $|\psi\rangle$ without any denoising decreases at higher variance of the noise. The denoiser improves the fidelity, demonstrating a protection against thermal noise. For the depolarizing noise we can perfectly remove the noise and achieve $F_1 = 1$ for all $p \in [0, 1]$ with success probability $1 - p(1 - 1/N)$. Next, we consider dit-flip, phase-flip and amplitude damping channels [39] in Fig. 3. For sufficiently low noise probability $p \lesssim 0.5$, the denoiser substantially improves fidelity. The denoiser performs best for dit-flip and phase flip channels. We find only minor improvement for amplitude damping channel as it is a non-unital noise model where the steady state is pure and has a large coherent noise contribution which is hard to denoise with population training.

Fidelity training. Now, we train encoder $U(\theta)$ and decoder $U_d(\varphi)$ with separate parameters θ, φ by maximizing the fidelity between ideal input state and denoised state. In contrast to population training, fidelity training has access to the ideal state and thus can correct for coherent errors. Comparing Fig. 3(b) and 3(c), we see that fidelity training performs better than population training, particularly for amplitude damping noise. We prove that, for $N \geq 2K$ and noise channel

$$\mathcal{E}(\rho) = (1 - p)V\rho V^\dagger + p|\psi_{\text{noise}}\rangle\langle\psi_{\text{noise}}| \quad (10)$$

with pure ideal states ρ , arbitrary V and $|\psi_{\text{noise}}\rangle$, we can always find a perfect denoiser with $\bar{F}_K = 1$ (see Appendix C [34]). Fig. 4 shows the performance of the denoiser with both training methods, for $N = 3$ and $K = 2$. The output fidelity is averaged over 100 Haar random samples of the ideal subspace. Results for fidelity training are obtained from numerical simulations. When the noise state $|\psi_{\text{noise}}\rangle$ has little overlap with the ideal subspace ($c = 0.1$), population training can improve the fidelity significantly for sufficiently small p from Eq. (8), but has a detrimental effect at large p . On the other hand, the fidelity-trained denoiser improves the output fidelity to near unity for all $p \in [0, 1]$. Intuitively, the fidelity-trained denoiser is able to correct for coherent error within the ideal subspace, whereas population training can at best only project the noisy states onto the ideal subspace (or close to it) agnostic of coherent errors. This accounts for the large discrepancy in performance. To further cement this argument, we also consider a noise state which has significant overlap with the ideal subspace ($c = 0.8$). In this case, population training is essentially ineffective, whereas fidelity training can achieve a denoising fidelity of > 0.9 even for p close to 1.

Magic state distillation. Magic state distillation (MSD) is an algorithm to obtain a low-noise magic state from multiple copies of noisy states. The extension of MSD to qutrits was

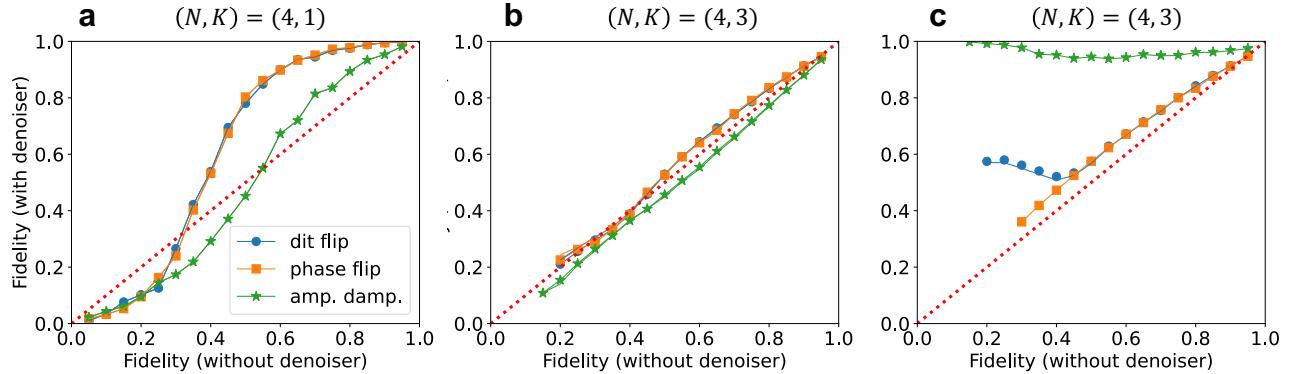


FIG. 3. **Denoising of qudit noise channels.** Fidelity of noisy quantum states before and after denoiser. The fidelity is averaged over $N = 4$ -dimensional states subject to dit-flip, phase flip and amplitude damping noise channels. In (a) – (c), we sample 5×10^4 Haar random states chosen from 1000 different K -dimensional subspaces. Points which lie above the red dotted line indicate an improvement in fidelity by using the denoiser. We have a $K = 1$ in (a), while $K = 3$ in (b) and (c). population training is used for (a) and (b), while fidelity training is used in (c). The solid lines are obtained using the analytical approximation Eq. (6).

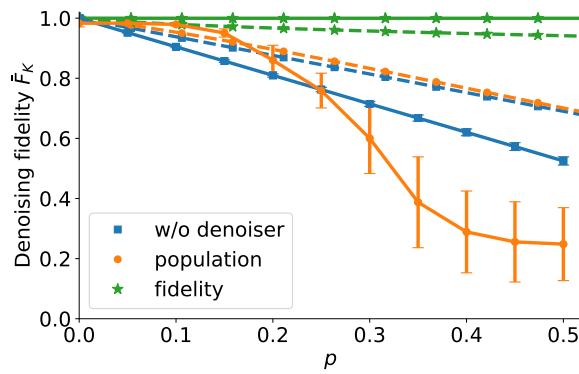


FIG. 4. **Subspace denoising with population and fidelity training.** Denoising fidelity, averaged over 200 Haar random samples of the ideal subspace ($N = 3, K = 2$). Solid and dashed lines represent data for $c = 0.1$ and 0.8 respectively. The labels ‘population’ and ‘fidelity’ refer to the denoiser obtained from population and fidelity training respectively. Data for bare fidelity and population training with $c = 0.1$ is obtained from experiment, while the rest are obtained from numerical simulation.

first proposed by Anwar et al. [40]. However, this scheme is extremely costly due to the low success probability for each iteration of the protocol. We propose our denoiser as a pre-processing step to drastically reduce the cost of MSD. We consider the magic state $|H_+\rangle$ which is the $+1$ eigenstate of the qutrit Hadamard operator [40]. The magic state is distilled iteratively using the five-qutrit code $[[5, 1, 3]]_3$, with a success probability of around 4% per iteration. The input magic states are subject to depolarizing noise with probability p_{in} . Further, we assume that the autoencoder operations itself are affected by depolarizing noise with probability p_{AE} for each encoder and decoder unitary. From our experimental data, we estimate $p_{AE} = 0.02$. We compare the average number of copies of noisy magic states N_{copies} needed to obtain a target fidelity of $1 - 2p_{AE}/3$, which is the fidelity attainable from our denoiser.

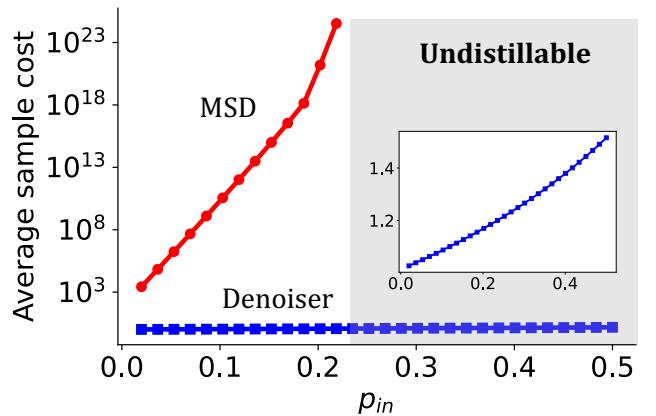


FIG. 5. **Denoising for magic state distillation of qutrits.** Average number of noisy magic states required to obtain a target fidelity of at least $1 - 2p_{AE}/3 \approx 0.987$ using MSD and the autoencoder denoiser. p_{in} is the depolarizing noise probability of the input magic states. The training cost on the order of 10^2 samples serves as a constant overhead and is omitted. In the grey region $p_{in} \gtrsim 0.233$ MSD does not work. The encoder and decoder are assumed to be noisy with each subject to depolarizing probability $p_{AE} = 0.02$. (Inset) We magnify the average number of noisy magic states required for denoiser.

We assume that the autoencoder has been already trained using population training with $(N, K) = (3, 1)$. For our denoiser, N_{copies} can be analytically calculated [34] and is upper bounded by $N_{\text{copies}} \leq N = 3$. On the other hand, the sample cost for MSD is many orders of magnitude greater, as illustrated in Fig. 5. The sample cost grows exponentially for small p_{in} , and diverges at the threshold $p_{in} \approx 0.233$ [40]. Beyond the threshold noise, the magic state is undistillable while the denoiser still works efficiently. Thus, we envision that the denoiser can serve as pre-processing step to clean up noisy magic states, followed by a few iterations of MSD to reach the desired target fidelity.

Quantum state cooling. Another application of the denoiser is to cool an N -dimensional thermal state to the ground state [41]. We choose the noisy input state to be a Gibbs state $\rho = \exp(-\beta H)/\mathcal{Z}$ with Hamiltonian H , partition function $\mathcal{Z} = \text{Tr}(\exp(-\beta H))$ and inverse temperature $\beta = 1/(k_B T)$. We can write the state as $\rho = \mathcal{Z}^{-1} \sum_n \exp(-\beta E_n) |\varphi_n\rangle \langle \varphi_n|$ with eigenenergies $E_1 \leq \dots \leq E_N$ and eigenstates $|\varphi_n\rangle$ of H . The ground state has the smallest eigenenergy E_1 and thus largest eigenvalue of ρ . Thus, population training with $K = 1$ extracts the ground state $|\varphi_1\rangle$ for any β with $F = 1$ fidelity and post-selection probability $\exp(-\beta E_1)/\mathcal{Z}$.

DISCUSSION

We demonstrate quantum autoencoders to denoise quantum states with rigorous performance guarantees. Our experimental demonstration on a photonic chip delivers a substantial improvement in output fidelity across a diverse range of noise channels. We propose two different variants of denoisers: Fidelity training requires noise-free reference states for training and shows exceptional performance for all considered noise models. In contrast, population training is only trained on the noisy input states by optimizing the probability of measuring redundant modes. Population training shows exceptional performance in reducing incoherent errors with rigorous guarantees on the fidelity of the denoised states. For example, we reach unit fidelity for depolarizing noise and a one-dimensional subspace. The simple training protocol does not consume the denoised state such that the autoencoder can be trained online, i.e. while the autoencoder is actively denoising states. This feature could be used for adaptive online learning of the denoiser in dynamic noise environments.

Our denoiser can drastically reduce the cost of magic state distillation, a key bottleneck of fault-tolerant quantum computing, by several orders of magnitudes. We can also cool thermal quantum states to the ground state by projecting out thermal excitations. Furthermore, our protocol could be integrated with error mitigation techniques [42, 43] and classical shadow tomography [44, 45] to enhance capabilities and reduce resource requirements, opening up new avenues for developing quantum technologies.

METHODS

Chip fabrication. The entire autoencoder network is manufactured on the silicon-on-insulator (SOI) platform, featuring a 220-nm-thick silicon top layer and a 2- μm -thick buried oxide layer. A thin layer of titanium nitride (TiN) is deposited as the resistive layer for heating elements. A thin aluminum film is patterned to realize the electrical connection for the heaters. Isolation trenches are etched in the SiO_2 top cladding and Si substrate.

Training on-chip. During on-chip training, we utilize coherent states as inputs to the autoencoder, leveraging their ability to achieve effects similar to those of single photons. This allows us to rapidly obtain the chip's configuration parameters. The trained autoencoder is able to denoise single photon states subject to the same noise channel. Our autoencoder is beneficial for single photon states as input, as here we can perform post-selection to denoise the state.

Noise channel implementation. In the experiment, we realize noise channel \mathcal{E} acting on an ideal input state $|\psi\rangle = T|0\rangle$ in the state generation stage. Here, the noisy input state $\rho_{\text{noisy}} = \mathcal{E}(T|0\rangle) = \sum_k p_k V_k \rho V_k^\dagger$ is implemented via a probabilistic mixture of unitaries, where unitaries V_k are chosen with probability p_k .

Numerical simulation of the denoiser. The optimal encoder U_e for population training is equivalent to any unitary which rotates the K -dimensional ideal subspace onto the K -dominant eigenspace of the noisy input state, which we can compute numerically. For fidelity training, the cost function is minimized by numerical optimization of the autoencoder parameters. We note that the choice of optimization algorithm has no significant impact on the denoising fidelity.

ACKNOWLEDGMENTS

This work is supported by a Samsung GRC project and the UKRI EPSRC grants EP/W032643/1 and EP/Y004752/1. The authors thank John Preskill, Hsin-Yuan Huang and Jielun Chen for insightful discussions.

-
- [1] A. Montanaro, Quantum algorithms: an overview, *npj Quantum Information* **2**, 1 (2016).
- [2] S. Pirandola, U. L. Andersen, L. Banchi, M. Berta, D. Bunnadar, R. Colbeck, D. Englund, T. Gehring, C. Lupo, C. Ottaviani, *et al.*, Advances in quantum cryptography, *Advances in optics and photonics* **12**, 1012 (2020).
- [3] V. Giovannetti, S. Lloyd, and L. Maccone, Advances in quantum metrology, *Nature photonics* **5**, 222 (2011).
- [4] D. Suter and G. A. Álvarez, Colloquium: Protecting quantum information against environmental noise, *Reviews of Modern Physics* **88**, 041001 (2016).
- [5] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, Noisy intermediate-scale quantum algorithms, *Rev. Mod. Phys.* **94**, 015004 (2022).
- [6] D. Gottesman, An introduction to quantum error correction and fault-tolerant quantum computation, in *Quantum information science and its contributions to mathematics, Proceedings of Symposia in Applied Mathematics*, Vol. 68 (2010) pp. 13–58.
- [7] S. Bravyi and J. Haah, Magic-state distillation with low overhead, *Physical Review A* **86**, 052329 (2012).
- [8] E. T. Campbell, B. M. Terhal, and C. Vulliot, Roads towards fault-tolerant universal quantum computation, *Nature* **549**, 172 (2017).
- [9] Y. Suzuki, S. Endo, K. Fujii, and Y. Tokunaga, Quantum er-

- ror mitigation as a universal error reduction technique: applications from the nisq to the fault-tolerant quantum computing eras, *PRX Quantum* **3**, 010345 (2022).
- [10] A. Krishna and J.-P. Tillich, Towards low overhead magic state distillation, *Phys. Rev. Lett.* **123**, 070507 (2019).
- [11] M. B. Hastings and J. Haah, Distillation with sublogarithmic overhead, *Phys. Rev. Lett.* **120**, 050504 (2018).
- [12] C. Jones, Multilevel distillation of magic states for quantum computing, *Phys. Rev. A* **87**, 042305 (2013).
- [13] E. T. Campbell and M. Howard, Unifying gate synthesis and magic state distillation, *Phys. Rev. Lett.* **118**, 060501 (2017).
- [14] J. Haah, M. B. Hastings, D. Poulin, and D. Wecker, Magic state distillation with low space overhead and optimal asymptotic input count, *Quantum* **1**, 31 (2017).
- [15] J. Romero, J. P. Olson, and A. Aspuru-Guzik, Quantum autoencoders for efficient compression of quantum data, *Quantum Sci. Technol.* **2**, 045001 (2017).
- [16] K. H. Wan, O. Dahlsten, H. Kristjánsson, R. Gardner, and M. Kim, Quantum generalisation of feedforward neural networks, *npj Quantum information* **3**, 36 (2017).
- [17] L. Lamata, U. Alvarez-Rodriguez, J. D. Martín-Guerrero, M. Sanz, and E. Solano, Quantum autoencoders via quantum adders with genetic algorithms, *Quantum Science and Technology* **4**, 014007 (2018).
- [18] Y. Du and D. Tao, On exploring practical potentials of quantum auto-encoder with advantages, arXiv preprint arXiv:2106.15432 (2021).
- [19] C. Bravo-Prieto, Quantum autoencoders with enhanced data encoding, *Machine Learning: Science and Technology* **2**, 035028 (2021).
- [20] A. Anand, J. S. Kottmann, and A. Aspuru-Guzik, Quantum compression with classically simulatable circuits, arXiv preprint arXiv:2207.02961 (2022).
- [21] H. Zhang, L. Wan, T. Haug, W.-K. Mok, S. Paesani, Y. Shi, H. Cai, L. K. Chin, M. F. Karim, L. Xiao, *et al.*, Resource-efficient high-dimensional subspace teleportation with a quantum autoencoder, *Science Advances* **8**, eabn9783 (2022).
- [22] F. Liu, K. Bian, F. Meng, W. Zhang, and O. Dahlsten, Information compression via hidden subgroup quantum autoencoders, arXiv:2306.08047 (2023).
- [23] A. Pepper, N. Tischler, and G. J. Pryde, Experimental realization of a quantum autoencoder: The compression of qutrits via machine learning, *Physical review letters* **122**, 060501 (2019).
- [24] C.-J. Huang, H. Ma, Q. Yin, J.-F. Tang, D. Dong, C. Chen, G.-Y. Xiang, C.-F. Li, and G.-C. Guo, Realization of a quantum autoencoder for lossless compression of quantum data, *Physical Review A* **102**, 032412 (2020).
- [25] F. Zhou, Y. Tian, Y. Song, C. Qiu, X. Wang, M. Zhou, B. Chen, N. Xu, and D. Lu, Preserving entanglement in a solid-spin system using quantum autoencoders, *Applied Physics Letters* **121**, 134001 (2022).
- [26] Y. Ding, L. Lamata, M. Sanz, X. Chen, and E. Solano, Experimental implementation of a quantum autoencoder via quantum adders, *Advanced Quantum Technologies* **2**, 1800065 (2019).
- [27] D. Bondarenko and P. Feldmann, Quantum autoencoders to denoise quantum data, *Physical review letters* **124**, 130502 (2020).
- [28] X.-M. Zhang, W. Kong, M. U. Farooq, M.-H. Yung, G. Guo, and X. Wang, Generic detection-based error mitigation using quantum autoencoders, *Physical Review A* **103**, L040403 (2021).
- [29] T. Achache, L. Horesh, and J. Smolin, Denoising quantum states with quantum autoencoders—theory and applications, arXiv:2012.14714 (2020).
- [30] C. Cao and X. Wang, Noise-assisted quantum autoencoder, *Physical Review Applied* **15**, 054012 (2021).
- [31] D. F. Locher, L. Cardarelli, and M. Müller, Quantum error correction with quantum autoencoders, *Quantum* **7**, 942 (2023).
- [32] J. Pazem and M. H. Ansari, Error mitigation of entangled states using brainbox quantum autoencoders, arXiv:2303.01134 (2023).
- [33] Q. H. Tran, S. Kikuchi, and H. Oshima, Variational denoising for variational quantum eigensolver, arXiv:2304.00549 (2023).
- [34] See Supplementary Materials.
- [35] N. Ezzell, E. M. Ball, A. U. Siddiqui, M. M. Wilde, A. T. Sornborger, P. J. Coles, and Z. Holmes, Quantum mixed state compiling, *Quantum Sci. Technol.* **8**, 10.1088/2058-9565/acc4e3 (2023).
- [36] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, *Nature* **567**, 209 (2019).
- [37] J. Cotler, N. Hunter-Jones, J. Liu, and B. Yoshida, Chaos, complexity, and random matrices, *Journal of High Energy Physics* **2017**, 48 (2017).
- [38] B. Koczor, The dominant eigenvector of a noisy quantum state, *New Journal of Physics* **23**, 123047 (2021).
- [39] A. Fonseca, High-dimensional quantum teleportation under noisy environments, *Phys. Rev. A* **100**, 062311 (2019).
- [40] H. Anwar, E. T. Campbell, and D. E. Browne, Qutrit magic state distillation, *New Journal of Physics* **14**, 063006 (2012).
- [41] J. Cotler, S. Choi, A. Lukin, H. Gharibyan, T. Grover, M. E. Tai, M. Rispoli, R. Schittko, P. M. Preiss, A. M. Kaufman, *et al.*, Quantum virtual cooling, *Physical Review X* **9**, 031013 (2019).
- [42] Z. Cai, R. Babbush, S. C. Benjamin, S. Endo, W. J. Huggins, Y. Li, J. R. McClean, and T. E. O'Brien, Quantum error mitigation (2022), arXiv:2210.00921.
- [43] S. Endo, S. C. Benjamin, and Y. Li, Practical quantum error mitigation for near-future applications, *Phys. Rev. X* **8**, 031027 (2018).
- [44] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, *Nature Physics* **16**, 1050 (2020).
- [45] D. E. Koh and S. Grewal, Classical shadows with noise, *Quantum* **6**, 776 (2022).
- [46] C. Davis and W. M. Kahan, The rotation of eigenvectors by a perturbation. iii, *SIAM Journal on Numerical Analysis* **7**, 1 (1970).

Appendix

We provide additional technical details and data supporting the claims in the main text.

Appendix A: Notation and symbols

Symbol	Name
N	Dimension of Hilbert space
K	Dimension of projected subspace of autoencoder
$ \psi\rangle \in S$	Set of ideal states
$\rho_{\text{in}} = \mathcal{E}(\psi\rangle)$	Noisy input state
ρ_{in}	Noisy input state to autoencoder
$ \psi_{\text{noise}}\rangle$	Noise perturbation
ρ_{denoise}	Denoised state after autoencoder
ρ_S	Ensemble of noisy input states
U_e	Encoder unitary
U_d	Decoder unitary
θ	Encoder parameters
φ	Decoder parameters
C	Cost function
P_K	Projector onto K latent modes of autoencoder
Π_K	Projector onto K -dimensional ideal subspace
D_K	Projector onto K -dominant eigenspace of ρ_S
$ \varphi_j\rangle$	Eigenvectors of ρ_S with eigenvalue λ_j , in the order $\lambda_1 \geq \dots \geq \lambda_N$
$ \phi_j\rangle$	Basis vectors of ideal subspace, $1 \leq j \leq K$
$ \phi_j^\perp\rangle$	Basis vectors of orthogonal complement to ideal subspace, $1 \leq j \leq N - K$
$ j\rangle$	Computational basis vectors, $0 \leq j \leq N - 1$
M_n	Kraus operators
F	Fidelity of a denoised state
\bar{F}	Ensemble average of F
$\bar{F}^{(q)}$	Quenched approximation of \bar{F}
c	Overlap between noise state and ideal subspace, $c = \text{Tr}(\Pi_K \rho_{\text{noise}})$

TABLE S1. Definitions of symbols.

Appendix B: Population training

The cost function for population training is the population of the $N - K$ redundant modes which is to be minimized. This is equivalent to maximizing the population in the K latent modes. Diagonalizing the noisy ensemble density matrix ρ_S , it is easy to see that the optimal encoder U_e performs a rotation from the K -dominant eigenspace of ρ_S to the K -dimensional latent subspace. The population of the latent modes is therefore the sum of the K -dominant eigenvalues, which is the success probability of the protocol. After projecting out the redundant modes and re-initializing them in the vacuum state, the optimal decoder U_d is simply the inverse of the encoder, i.e., a rotation from the latent subspace back to the K -dominant eigenspace. Viewed together, the trained autoencoder essentially projects the noisy state onto its K -dominant eigenspace. Note that the choice of decoder $U_d = U_e^\dagger$ is unique, since the U_e can contain any arbitrary rotation within the latent space, which must be corrected for in the decoding step.

1. Single-state denoising

First, let us consider a noise channel of the form

$$\mathcal{E}(\rho) = (1 - p)\rho + p\rho_{\text{noise}} \quad (\text{S1})$$

where ρ_{noise} is an arbitrary noise state that perturbs the ideal state ρ with probability p .

Lower bound on fidelity

Intuitively, when the noise probability p is small, the noise state acts as a perturbation to the ideal state $|\psi\rangle$. The dominant eigenstate remains close to $|\psi\rangle$, giving a denoising fidelity near unity. More concretely, by applying Theorem 1 of Ref. [38], we can bound the denoising fidelity as

$$F \geq \frac{1}{2}(1 + \sqrt{1 - \delta^2}), \quad (\text{S2})$$

where the lower bound is saturated by noise states of the form

$$\rho_{\text{noise}} = \mu |\chi\rangle\langle\chi| + \sum_{j=3}^N d_k |d_k\rangle\langle d_k|, \quad (\text{S3})$$

$$|\chi\rangle = \sqrt{\frac{1-\delta}{2}} |\psi\rangle + \sqrt{\frac{1+\delta}{2}} |d_2\rangle \quad (\text{S4})$$

with $\{|\psi\rangle, |d_2\rangle, \dots, |d_N\rangle\}$ forming an orthonormal basis, and $\delta = p\mu/(1-p)$. To get the worst-case fidelity, we choose $\mu = 1$ such that $\delta = p/(1-p)$. Physically, this means that the noise state is a pure state with support only on the ideal state $|\psi\rangle$ and an orthogonal state $|d_2\rangle$. The worst-case fidelity is thus

$$F_{\text{worst}} = \frac{1}{2} \left(1 + \sqrt{1 - \left(\frac{p}{1-p} \right)^2} \right) = 1 - \frac{1}{4}p^2 + O(p^3) \quad (\text{S5})$$

for $p \leq 1/2$.

Pure state noise

As shown in Eq. (S4) the worst-case noise state lies in the two-dimensional subspace spanned by $|\psi\rangle$ and $|\psi^\perp\rangle$. Let us now consider the noise state to be an arbitrary density matrix in the subspace

$$\rho_{\text{noise}} = \begin{pmatrix} \rho_{00} & \rho_{01} \\ \rho_{01} & \rho_{00} \end{pmatrix} \quad (\text{S6})$$

where $\rho_{00} = \langle\psi|\rho_{\text{noise}}|\psi\rangle$ is the overlap between ρ_{noise} and the target state $|\psi\rangle$, and ρ_{01} is the coherence of ρ_{noise} between $|\psi\rangle$ and $|\psi^\perp\rangle$. In this basis, the noisy input state can be represented by the density matrix

$$\rho = \begin{pmatrix} 1-p+p\rho_{00} & p\rho_{01} \\ p\rho_{01} & p(1-\rho_{00}) \end{pmatrix} \quad (\text{S7})$$

We note that ρ_{01} is upper bounded by the pure-state limit $|\rho_{01}|^2 \leq \rho_{00}(1-\rho_{00})$. The largest eigenvalue can be found exactly:

$$\lambda_1 = \frac{1}{2} \left(1 + \sqrt{(1-2p(1-\rho_{00}))^2 + 4p^2|\rho_{01}|^2} \right) \quad (\text{S8})$$

with the corresponding eigenstate

$$|\varphi_1\rangle = \frac{1}{\mathcal{N}} \begin{pmatrix} \lambda_1 - p(1-\rho_{00}) \\ p\rho_{01} \end{pmatrix} \quad (\text{S9})$$

where $\mathcal{N} = \sqrt{(\lambda_1 - p(1-\rho_{00}))^2 + p^2|\rho_{01}|^2}$ is the normalization factor. The reconstruction probability is given by λ_1 , and the fidelity of the reconstructed state is

$$F = |\langle\psi|\varphi_1\rangle|^2 = \frac{1}{\mathcal{N}^2} (\lambda_1 - p(1-\rho_{00}))^2 \quad (\text{S10})$$

In the simple case where ρ_{noise} is a pure state within the subspace spanned by $\{|\psi\rangle, |\psi^\perp\rangle\}$, we have $|\rho_{01}|^2 = \rho_{00}(1-\rho_{00})$, and the maximal eigenvalue of ρ simplifies to $\lambda_1 = \frac{1}{2}(1 + \sqrt{1 - 4p(1-p)(1-\rho_{00})})$ and the fidelity is

$$\begin{aligned} F &= \left[1 + \frac{p^2\rho_{00}(1-\rho_{00})}{(\lambda_1 - p(1-\rho_{00}))^2} \right]^{-1} \\ &= 1 - \rho_{00}(1-\rho_{00})p^2 + O(p^3) \end{aligned} \quad (\text{S11})$$

In the limit where $\rho_{00} \rightarrow 0$, the fidelity is either 1 (when $p < 1/2$) or 0 (when $p > 1/2$). This is intuitive because it corresponds to the case where $\rho_{\text{noise}} = |\psi^\perp\rangle\langle\psi^\perp|$ is orthogonal to $|\psi\rangle$, so $|\psi^\perp\rangle$ becomes the dominant eigenstate for $p > 1/2$ thus causing a sharp transition in the fidelity. More precisely, we find that

$$\frac{\partial F}{\partial \rho_{00}} \bigg|_{p=\frac{1}{2}} = \frac{1}{4\sqrt{\rho_{00}}} \quad (\text{S12})$$

which diverges as $\rho_{00} \rightarrow 0$.

2. Haar-random noise state

Suppose the noise state is now a pure state $|\psi_{\text{noise}}\rangle$ sampled from the Haar ensemble with dimension N . An optimal denoiser can be found for each $|\psi_{\text{noise}}\rangle$, with the denoising fidelity from Eq. (S11). We want to know what is the average-case performance of the denoiser. This can be done by integrating the fidelity over the Haar measure,

$$\bar{F}_N(p) = \int_0^1 d\rho_{00} \left[1 + \frac{p^2 \rho_{00} (1 - \rho_{00})}{(p_1 - p(1 - \rho_{00}))^2} \right]^{-1} \times (N-1)(1 - \rho_{00})^{N-2} \quad (\text{S13})$$

For a fixed dimension N , the integral can be analytically computed. As an example, for $N = 2$,

$$\bar{F}_2(p) = \begin{cases} \frac{6+p(5p-12)}{6(p-1)^2}, & p \leq 1/2 \\ \frac{2+p}{6p}, & p > 1/2 \end{cases} \quad (\text{S14})$$

which decreases monotonically from $\bar{F}_2(0) = 0$ to $\bar{F}_2(1/2) = 5/6$ to $\bar{F}_2(1) = 1/2$. In the other extreme limit of $N \rightarrow \infty$, we get the step function

$$\bar{F}_\infty(p) = \Theta(p) - \Theta(p - 1/2). \quad (\text{S15})$$

For all N , $\bar{F}_N(1) = 1/N$, as expected from a random reconstructed state. For $p \ll 1$, we get

$$\bar{F}_N(p) = 1 - \frac{N-1}{N(N+1)} p^2 - O(p^3) \quad (\text{S16})$$

from which it can be shown that $\partial_N \bar{F}_N(p) > 0$ for all $N \geq 3$, and $\bar{F}_2(p) = \bar{F}_3(p)$ to order p^2 . This means that for a sufficiently small noise probability, the denoising fidelity will increase monotonically as the dimension N increases (except from $N = 2 \rightarrow N = 3$). The denoiser performs better on average in higher dimensions.

To determine the regime of validity, we include the p^3 contribution to $\bar{F}_N(p)$ which is $-4(N-1)p^3/(N+1)(N+2)$, and demanding that this is much smaller than the p^2 term. This gives us the regime

$$p \ll \frac{N+2}{4N} \sim \frac{1}{4}. \quad (\text{S17})$$

Qudit noise channels

Here, we consider various common qudit noise channels. Let us write the unitary Weyl operators as

$$W_{mn} = \sum_{j=0}^{N-1} \omega_N^{jm} |j\rangle\langle j \oplus n|, \quad (\text{S18})$$

where $\omega_N = \exp(2\pi i/N)$, and \oplus denotes addition modulo N . The dit-flip, phase-flip, and dit-phase flip channels can be expressed using the Weyl operators [39]. The Kraus operators for these noise channels are

$$\begin{aligned} (\text{Dit flip}) \quad E_{00} &= \sqrt{1-p}I, \\ E_{0j} &= \sqrt{\frac{p}{N-1}} W_{0j}, \quad j = 1, \dots, N-1 \end{aligned} \quad (\text{S19})$$

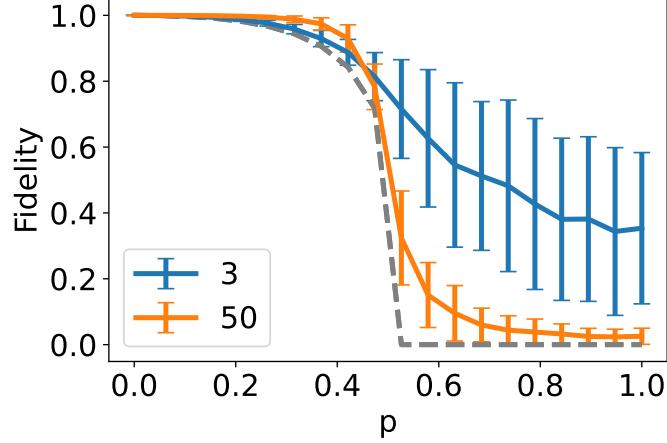


FIG. S1. Denoising fidelity for Haar-random noise state, with 200 instances. The solid lines represent the average fidelity $\bar{F}_N(p)$ for $N = 3$ (blue) and $N = 50$ (orange), while the error bars represent the standard deviation in the fidelity. The dashed line is the worst-case fidelity which provides a lower bound.

$$\begin{aligned}
 \text{(Phase flip)} \quad E_{00} &= \sqrt{1-p}I, \\
 E_{j0} &= \sqrt{\frac{p}{N-1}}W_{j0}, \quad j = 1, \dots, N-1
 \end{aligned} \tag{S20}$$

$$\begin{aligned}
 \text{(Dit-phase flip)} \quad E_{00} &= \sqrt{1-p}I, \\
 E_{mn} &= \frac{\sqrt{p}}{N-1}W_{mn}, \quad m, n = 1, \dots, N-1
 \end{aligned} \tag{S21}$$

The amplitude-damping channel, on the other hand, cannot be expressed in terms of the Weyl operators. Its Kraus operators are

$$\begin{aligned}
 \text{(Amplitude-damping)} \quad E_0 &= |0\rangle\langle 0| + \sqrt{1-p} \sum_{j=1}^{N-1} |j\rangle\langle j| \\
 E_j &= \sqrt{p}|0\rangle\langle j|, \quad j = 1, \dots, N-1
 \end{aligned} \tag{S22}$$

The denoiser performance against such noise are plotted in Fig. S2 for $N = 5$, for Haar-random ideal states. The denoiser performs well for dit flips, phase flips, and dit-phase flips, but not as well as for amplitude damping.

3. Lower bound on population training performance

We assume an N -dimensional noise channel

$$\mathcal{E}(\rho) = \sum_n M_n \rho M_n^\dagger \tag{S23}$$

with Kraus map $M_0 = \sqrt{1-p}I$ and arbitrary additional Kraus maps M_n , $n > 0$, where p is the noise probability and $\sum_n M_n^\dagger M_n = I_N$. Thus, $\rho_{\text{in}} = (1-p)|\psi\rangle\langle\psi| + \sum_{n \geq 1} M_n |\psi\rangle\langle\psi| M_n^\dagger$. We now have ideal states randomly sampled from a K -dimensional subspace. By standard integration over the ensemble, one can see that the ensemble of noisy input states in the K -dimensional ideal space is given by the density matrix

$$\rho_S = \mathbb{E}_{|\psi\rangle \in S} [\mathcal{E}(|\psi\rangle)] = (1-p) \frac{\Pi_K}{K} + p \rho_{\text{noise}}, \tag{S24}$$

where $\rho_{\text{noise}} = \frac{1}{K} \sum_{n \geq 1} M_n \Pi_K M_n^\dagger$. The ensemble ρ_S has N eigenvalues $\lambda_1 \geq \dots \geq \lambda_N$ whereas ρ_{noise} has eigenvalues $\nu_1 \geq \dots \geq \nu_N$. Applying Weyl's inequality, we have

$$\frac{1-p}{K} + p \nu_N \leq \lambda_i \leq \frac{1-p}{K} + p \nu_1, \quad i = 1, \dots, K \tag{S25}$$

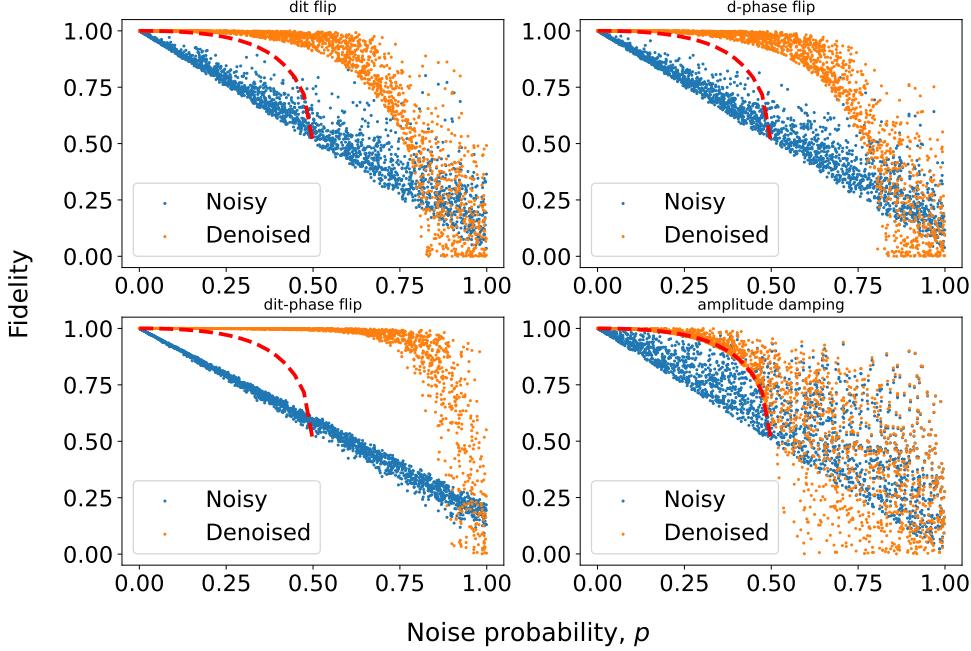


FIG. S2. Fidelity for 2000 Haar-random ideal states with dimension $N = 5$ subject to dit flip, phase flip, dit-phase flip and amplitude damping channels. The red dashed line shows the worst-case fidelity for the denoised state.

and

$$p\nu_N \leq \lambda_i \leq p\nu_1, \quad i = K + 1, \dots, N. \quad (\text{S26})$$

Using the Davis-Kahan sin θ theorem [46] with $\delta \equiv \frac{1-p}{K} - p\nu_1$, we can upper-bound the distance between the projector onto the subspace of ideal states Π_K and the projector D_K onto the K eigenvectors with largest eigenvalues of noisy states ρ_S (measured via the Frobenius norm):

$$\|\Pi_K - D_K\|_F \leq \frac{\sqrt{2}p}{\delta} \|\rho_{\text{noise}}\|_F. \quad (\text{S27})$$

Let us now write $\rho_{\text{noise}} = \mathbb{E}_{|\psi\rangle \in S} [\sum_{n=1} M_n |\psi\rangle \langle \psi| M_n^\dagger]$. The average fidelity after denoising is

$$\bar{F} = \mathbb{E}_{|\psi\rangle \in S} \left[\frac{(1-p) \langle \psi | D_K | \psi \rangle^2 + \sum_{n=1} |\langle \psi | D_K M_n | \psi \rangle|^2}{(1-p) \langle \psi | D_K | \psi \rangle + \sum_{n=1} \langle \psi | M_n^\dagger D_K M_n | \psi \rangle} \right]. \quad (\text{S28})$$

From Eq. (S27), we can write

$$D_K = \Pi_K + pA \quad (\text{S29})$$

for some traceless Hermitian operator A . Substituting this into the expression for \bar{F} , and using the fact that $\Pi_K |\psi\rangle = |\psi\rangle \forall |\psi\rangle \in S$, we have (to order p),

$$\bar{F} = \mathbb{E}_{|\psi\rangle \in S} \left[\frac{(1-p)(1+2p \langle \psi | A | \psi \rangle) + \sum_{n=1} |\langle \psi | M_n | \psi \rangle + p \langle \psi | A M_n | \psi \rangle|^2}{(1-p)(1+p \langle \psi | A | \psi \rangle) + \sum_{n=1} (\langle \psi | M_n^\dagger \Pi_k M_n | \psi \rangle + p \langle \psi | M_n^\dagger A M_n | \psi \rangle)} \right] + \mathcal{O}(p^2). \quad (\text{S30})$$

We now assume that $\|M_n\| \sim \mathcal{O}(\sqrt{p})$, giving

$$\bar{F} = \mathbb{E}_{|\psi\rangle \in S} \left[\frac{(1-p)(1+2p \langle \psi | A | \psi \rangle) + \sum_{n=1} |\langle \psi | M_n | \psi \rangle|^2}{(1-p)(1+p \langle \psi | A | \psi \rangle) + \sum_{n=1} (\langle \psi | M_n^\dagger \Pi_k M_n | \psi \rangle)} \right] + \mathcal{O}(p^2) \quad (\text{S31})$$

Since $\sum_{n=1} \langle \psi | M_n^\dagger \Pi_k M_n | \psi \rangle \leq p$,

$$\begin{aligned} \bar{F} &\geq \mathbb{E}_{|\psi\rangle \in S} \left[\frac{(1-p)(1+2p\langle\psi|A|\psi\rangle) + \sum_{n=1} |\langle\psi|M_n|\psi\rangle|^2}{1+p\langle\psi|A|\psi\rangle} \right] + \mathcal{O}(p^2) \\ &= \bar{F}_{\text{bare}} + p\mathbb{E}_{|\psi\rangle \in S}[\langle\psi|A|\psi\rangle] + \mathcal{O}(p^2) \end{aligned} \quad (\text{S32})$$

where

$$\bar{F}_{\text{bare}} = \mathbb{E}_{|\psi\rangle \in S} \left[1 - p + \sum_{n=1} |\langle\psi|M_n|\psi\rangle|^2 \right] \quad (\text{S33})$$

is the average fidelity without denoising. Since A is traceless, $\langle\psi|A|\psi\rangle$ vanishes when averaging over the Haar-random ideal states. Hence, we have the result

$$\bar{F} \geq \bar{F}_{\text{bare}} + \mathcal{O}(p^2). \quad (\text{S34})$$

Up to $\mathcal{O}(p^2)$, the lower bound is exactly the average fidelity of the noisy state without any denoising.

4. Subspace denoising

Pure state noise

Suppose we have an ensemble of M noisy states

$$\rho^{(k)} = (1-p)|\psi^{(k)}\rangle\langle\psi^{(k)}| + p|\psi_{\text{noise}}\rangle\langle\psi_{\text{noise}}| \quad (\text{S35})$$

where $k = 1, \dots, M$. $|\psi^{(k)}\rangle$ lies in the K -dimensional ideal space, while $|\psi_{\text{noise}}\rangle$ is a fixed noise state in the full N -dimensional Hilbert space. We assume that $|\psi^{(k)}\rangle$ is sampled from the Haar distribution. A uniform mixture of $\rho^{(k)}$ gives the density matrix

$$\rho_S = \frac{1}{M} \sum_{k=1}^M \rho^{(k)} \approx (1-p)\frac{I_K}{K} + p|\psi_{\text{noise}}\rangle\langle\psi_{\text{noise}}| \quad (\text{S36})$$

with the approximation becoming exact as the sample size $M \rightarrow \infty$. Note that while we have added $|\psi_{\text{noise}}\rangle$ incoherently to $|\psi^{(k)}\rangle$, the same density matrix can also be obtained if we treat $|\psi_{\text{noise}}\rangle$ as a coherent noise. To see this, we now write $\rho^{(k)} = |\psi_{\text{in}}^{(k)}\rangle\langle\psi_{\text{in}}^{(k)}|$, where

$$|\psi_{\text{in}}\rangle = \sqrt{1-p}|\psi^{(k)}\rangle + \sqrt{p}|\psi_{\text{noise}}\rangle. \quad (\text{S37})$$

Taking a uniform mixture, the density matrix is now

$$\begin{aligned} \rho_S &\approx (1-p)\frac{I_K}{K} + p|\psi_{\text{noise}}\rangle\langle\psi_{\text{noise}}| \\ &\quad + \sqrt{p(1-p)}\mathbb{E} \left[|\psi^{(k)}\rangle\langle\psi_{\text{noise}}| + |\psi_{\text{noise}}\rangle\langle\psi^{(k)}| \right] \end{aligned} \quad (\text{S38})$$

where $\mathbb{E}[\cdot]$ denotes an average over the ensemble of states. Assuming a Haar ensemble, the extra term vanishes on average, so the resulting density matrix is the same as that with incoherent noise.

Let us write the noise state as

$$|\psi_{\text{noise}}\rangle = \sqrt{c}|\xi\rangle + \sqrt{1-c}|\xi^\perp\rangle \quad (\text{S39})$$

where $|\xi\rangle$ is a basis state in the ideal subspace, and $|\xi^\perp\rangle$ is a basis state in the orthogonal complement of the ideal subspace. In an orthonormal basis containing $|\xi\rangle$ and $|\xi^\perp\rangle$, the density matrix has a block diagonal form

$$\rho_S = \begin{pmatrix} \frac{1-p}{K} + pc & p\sqrt{c(1-c)} \\ p\sqrt{c(1-c)} & p(1-c) \end{pmatrix} \oplus \frac{1-p}{K}I_{K-1} \oplus \mathbf{0}_{N-K-1}. \quad (\text{S40})$$

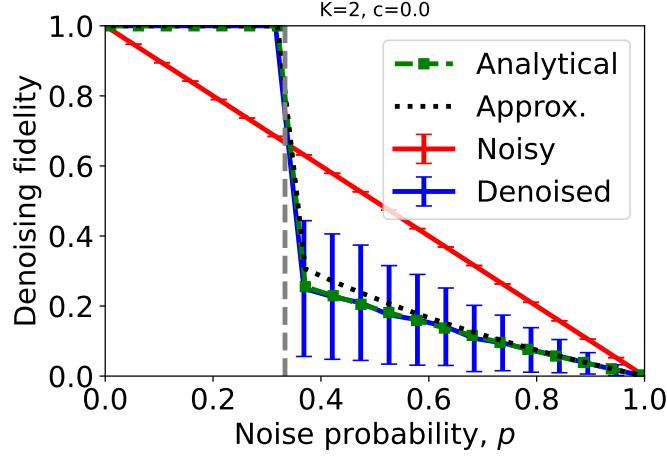


FIG. S3. Denoising fidelity for Haar-random states drawn from a K -dimensional ($K = 2$) ideal subspace, with 1000 instances. The noise state is orthogonal to the ideal subspace. The solid line represents the average fidelity $\bar{F}_K(p)$ for the denoised state (blue) and initial noisy state (red), while the error bars represent the standard deviation. The exact average fidelity is given by the green square markers while the approximate average fidelity is given by the black dotted line. The grey dashed line marks $p = 1/(K + 1)$.

The non-trivial eigenvalues and eigenvectors of ρ_S are

$$\lambda_{\pm} = \frac{1}{2K} \left[1 + p(K - 1) \pm \sqrt{1 + p(-2 + p + K(-2 - 4c(-1 + p) + (2 + K)p))} \right] \quad (\text{S41})$$

$$|\lambda_{\pm}\rangle = \frac{1}{\mathcal{N}_{\pm}} \left[p\sqrt{c(1 - c)} |\xi\rangle + \left(\lambda_{\pm} - \frac{1 - p}{K} - pc \right) |\xi^{\perp}\rangle \right] \quad (\text{S42})$$

with the normalization factors

$$\mathcal{N}_{\pm}^2 = p^2 c(1 - c) + \left(\lambda_{\pm} - \frac{1 - p}{K} - pc \right)^2. \quad (\text{S43})$$

The other eigenvalues are $(1 - p)/K$ (with degeneracy $K - 1$) and 0 (with degeneracy $N - K - 1$). It can be shown that for $0 < c < 1$, $\lambda_- < (1 - p)/K < \lambda_+$, so the optimally trained denoiser projects the state onto the subspace spanned by $|\lambda_+\rangle$ and the remaining $K - 1$ orthonormal basis vectors $|\xi_1\rangle \dots |\xi_{K-1}\rangle$ in the ideal space. Denoting $\alpha \equiv |\langle \xi | \psi \rangle|^2$, $\beta \equiv |\langle \xi | \lambda_+ \rangle|^2$ and $\gamma \equiv |\langle \lambda_+ | \psi_{\text{noise}} \rangle|^2$, the denoising fidelity is obtained as

$$F_K(p) = \frac{(1 - p)(\alpha\beta + 1 - \alpha)^2 + p\alpha\beta\gamma}{(1 - p)(\alpha\beta + 1 - \alpha) + p\gamma}. \quad (\text{S44})$$

This is dependent on the choice of the ideal state $|\psi\rangle$. We can remove this dependence by averaging over the Haar-random states $|\psi\rangle$, which gives

$$\bar{F}_K(p) = (K - 1) \int_0^1 d\alpha (1 - \alpha)^{K-2} F_K(p). \quad (\text{S45})$$

In the degenerate case where $c = 0$ (noise state is orthogonal to the ideal subspace), the integral can be solved exactly to give

$$\lim_{c \rightarrow 0} \bar{F}_K(p) = \begin{cases} 1 & , p < \frac{1}{K+1} \\ \frac{K-1}{K+1} (1 - p) {}_2F_1(1, 1; K - 2; 1 - p) & , p \geq \frac{1}{K+1} \end{cases} \quad (\text{S46})$$

where ${}_2F_1$ is the hypergeometric function. The denoiser performance transitions sharply from perfect recovery ($p < 1/(K + 1)$) to becoming detrimental ($p > 1/(K + 1)$), as depicted in the numerical simulations in Fig. S3.

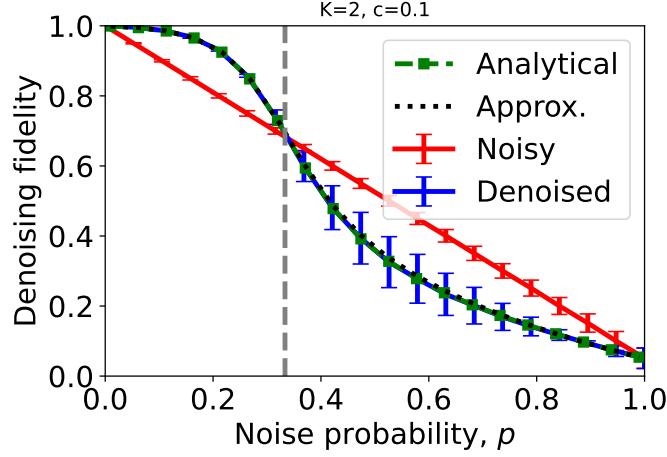


FIG. S4. Denoising fidelity for Haar-random states drawn from a K -dimensional ($K = 2$) ideal subspace, with 1000 instances. The noise state has an overlap of $c = 0.1$ with the ideal subspace. The solid line represents the average fidelity $\bar{F}_K(p)$ for the denoised state (blue) and initial noisy state (red), while the error bars represent the standard deviation. The exact average fidelity is given by the green square markers while the approximate average fidelity is given by the black dotted line. The grey dashed line marks $p = 1/(K + 1)$.

For a non-zero c , by performing a partial fraction expansion of $\bar{F}_K(p)$ and integrating term-by-term, the integral can be done exactly, to yield

$$\begin{aligned} \bar{F}_K(p) = & -\frac{1-\beta}{K} + \left(1 - \frac{p\gamma}{(1-p)(1-\beta)}\right) \\ & - \frac{p\gamma(\beta-p\beta-p\gamma)}{(1-p)(1-\beta)(1-p+p\gamma)} \\ & \times {}_2F_1(1, 1; K; (1-p)(1-\beta)/(1-p+p\gamma)). \end{aligned} \quad (\text{S47})$$

To order p^2 , we have

$$\bar{F}_K(p) = 1 - \frac{K-1}{K}cp - \frac{K(3K-1)-1}{K}c(1-c)p^2 + O(p^3). \quad (\text{S48})$$

As a consistency check, we set $K = 1$, which recovers Eq. (S11) ($c = \rho_{00}$). This means that if we want to denoise a subspace beyond just a single state, the leading-order correction to the denoising fidelity is proportional to p instead of p^2 , hence the denoiser becomes less effective. Nevertheless, we can show that for a sufficiently small p , using the denoiser is still advantageous. Without the denoiser, the average fidelity is

$$\bar{F}'_K(p) = 1 - \left(1 - \frac{2c}{K(K+1)}\right)p. \quad (\text{S49})$$

The denoising fidelity is higher than \bar{F}'_K if

$$p < \frac{K(K+1) - c(K^2 + 1)}{c(1-c)(K+1)(K(3K-1)-1)} \sim \frac{1}{3Kc} \quad (\text{S50})$$

for large Kc . We can see that range of p for which the denoiser is useful shrinks like $1/K$, so high-dimensional subspace denoising only works for very small noise probabilities. A more compact approximate expression can be obtained by exploiting properties of the Haar distribution, giving

$$\begin{aligned} \bar{F}_K^{(\text{approx.})}(p) = & \left[\frac{1-p}{K+1} (\text{tr}(\Pi_K D_K)^2 + \text{tr}((\Pi_K D_K)^2)) + p \text{tr}(\Pi_K D_K \sigma D_K) \right] / [(1-p)\text{tr}(\Pi_K D_K) \\ & + Kp \text{tr}(D_K \sigma)], \end{aligned} \quad (\text{S51})$$

where Π_K is the projector onto the ideal subspace, and D_K is the projector on the K -dominant eigenspace of the noisy ensemble ρ . An example of $K = 2, c = 0.1$ is plotted in Fig. S4. We can see that the approximate formula agrees well with the exact expression for average fidelity.

Quenched approximation for general channels

We now derive the quenched approximation for the fidelity $\bar{F}^{(q)}$ for general quantum channels. We assume that the pure input states are subject to general Kraus operators M_n with the condition $\sum_n M_n^\dagger M_n = I$. The pure input states $|\psi_j\rangle$ transform to $\rho_j = \sum_n M_n |\psi_j\rangle \langle \psi_j| M_n^\dagger$. Averaging over Haar-random (or at least a 2-design) input states $|\psi_j\rangle$ in the K -dimensional ideal subspace, the ensemble is described by the density matrix $\rho_S = \frac{1}{K} \sum_n M_n \Pi_K M_n^\dagger$. Denoting the (unnormalized) optimal autoencoder operation as $B \equiv U_d P_K U_e$, for each input the output state can be written as

$$\rho_{\text{out},j} = \frac{\sum_n B M_n \rho_j M_n^\dagger B^\dagger}{\sum_n \text{tr}(B M_n \rho_j M_n^\dagger B^\dagger)}, \quad (\text{S52})$$

and the fidelity with the ideal state $|\psi_j\rangle$ is $F_j = \langle \psi_j | \rho_{\text{out},j} | \psi_j \rangle$. By Haar integrating over the numerator and denominator separately we get an approximate expression for the average fidelity $\bar{F}^{(q)} \approx \bar{F} = \mathbb{E}_j [F_j]$,

$$\bar{F}^{(q)} = \frac{\sum_n (|\text{tr}(\Pi_K B M_n)|^2 + \|\Pi_K B M_n \Pi_K\|_F^2)}{(K+1) \sum_n \|B M_n \Pi_K\|_F^2} \quad (\text{S53})$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm. For population training, $B = D_K$, and $U_d = U_e^\dagger$.

5. Noisy autoencoder

We have largely assumed that the denoiser itself is noiseless when analyzing its performance. However, any realistic implementation of the autoencoder will invariably contain some noise. For simplicity, we model the noise of the encoder and decoder as independent depolarizing channels with the same probability p_{AE} . From the experimental data, we get a crude estimate of $p_{\text{AE}} \approx 0.02$. Let us analyze the denoising of a single state ($K = 1$) with this noise. Denoting the input noise channel as \mathcal{E} , the denoising probability (which is the post-selection probability) is the dominant eigenvalue of ρ , where

$$\rho = (1 - p_{\text{AE}}) \mathcal{E}(|\psi\rangle) + p_{\text{AE}} \frac{I_N}{N}. \quad (\text{S54})$$

The corresponding dominant eigenstate is denoted $|\chi_0\rangle$. The output state is

$$\rho_{\text{out}} = (1 - p_{\text{AE}}) |\chi_0\rangle \langle \chi_0| + p_{\text{AE}} \frac{I_N}{N}, \quad (\text{S55})$$

and the denoising fidelity is

$$F = (1 - p_{\text{AE}}) |\langle \chi_0 | \psi \rangle|^2 + \frac{p_{\text{AE}}}{N}. \quad (\text{S56})$$

In the limit $p_{\text{AE}} = 0$, we recover $|\chi_0\rangle \rightarrow |\psi\rangle$ and $F \rightarrow 1$.

If we further assume that the input noise is also an independent depolarizing noise with probability p_{in} , the denoising probability becomes

$$p_{\text{denoise}} = (1 - p_{\text{AE}})(1 - p_{\text{in}}) + \frac{p_{\text{in}} + p_{\text{AE}}(1 - p_{\text{in}})}{N} \quad (\text{S57})$$

with denoising fidelity

$$F = 1 - \frac{N-1}{N} p_{\text{AE}}. \quad (\text{S58})$$

In the worst-case of $p_{\text{in}} = 1$, we need to repeat this protocol on average N times to successfully denoise the state.

Appendix C: Fidelity training

For fidelity training, we assume to have access to noiseless ideal states during the training stage (this is not required for population training). The cost function is the average fidelity between the output states and the ideal states, which can be implemented using a SWAP test. The advantage is that it can correct for coherent noise within the ideal subspace, while population training is more suited for incoherent noise. A denoiser obtained from fidelity training always outperforms that from population training, with the trade off being the increased training cost.

1. Sufficient condition for perfect denoising

We claim that for the fixed noise channel and any fixed pure noise state $\rho_{\text{noise}} = |\psi_{\text{noise}}\rangle\langle\psi_{\text{noise}}|$, we can achieve perfect denoising where the average fidelity is $1 \forall p, c$. This is possible if $N \geq 2K$. To prove this, notice that a necessary condition for perfect denoising is to losslessly compress the ideal subspace while simultaneously removing the noise, i.e. $P_K U_e \Pi_K U_e^\dagger P_K = P_K$ and $P_K U_e |\psi_{\text{noise}}\rangle = 0$, where P_K is the projector onto the K latent modes and Π_K is the projector onto the ideal subspace. Without loss of generality, for $c \geq 0$, we can write $|\psi_{\text{noise}}\rangle = \sqrt{c}|\phi_1\rangle + \sqrt{1-c}|\phi_1^\perp\rangle$, $|\psi_{\text{noise}}^\perp\rangle = \sqrt{1-c}|\phi_1\rangle - \sqrt{c}|\phi_1^\perp\rangle$, and the ideal state $|\psi\rangle = \sum_j b_j |\phi_j\rangle$ with $\sum_j |b_j|^2 = 1$. We use two sets of basis states: $\{|j\rangle\}_{j=1}^{N+K}$ which spans the latent space, and $\{|\phi_j\rangle\}_{j=1}^N \cup \{|\phi_j\rangle^\perp\}_{j=1}^K$ which spans the ideal subspace. We can construct

$$U_1 = |N-1\rangle\langle\psi_{\text{noise}}| + |0\rangle\langle\psi_{\text{noise}}^\perp| + \sum_{j=1}^{K-1} |j\rangle\langle\phi_{j+1}| + \sum_{j=2}^{N-K} |j+K-2\rangle\langle\phi_j^\perp| \quad (\text{S1})$$

which rotates $|\psi_{\text{noise}}\rangle$ to one of the redundant modes, and for $N \geq 2K$ we define

$$U_2 = \sum_{j=1}^{K-1} \left[(\sqrt{1-c}|j\rangle + \sqrt{c}|j+K-1\rangle)\langle j| + (\sqrt{c}|j\rangle - \sqrt{1-c}|j+K-1\rangle)\langle j+K-1| \right] + |0\rangle\langle 0| + \sum_{j=2K-1}^{N-1} |j\rangle\langle j| \quad (\text{S2})$$

which is required to losslessly compress the ideal subspace. The encoder unitary is thus given by $U_e = U_2 U_1$, with

$$P_K U_e |\psi_{\text{noise}}\rangle = 0, \quad P_K U_e |\psi\rangle = \sqrt{1-c} \sum_{j=1}^K b_j |j-1\rangle. \quad (\text{S3})$$

Normalizing the latent state $P_K U_e |\psi\rangle$ and applying the decoder unitary

$$U_d = \sum_{j=1}^K |\phi_j\rangle\langle j-1| + \sum_{j=1}^{N-K} |\phi_j^\perp\rangle\langle j+K-1| \quad (\text{S4})$$

results in perfect denoising with success probability $1 - c$.

A dimension of $N \geq 2K$ is necessary to perfectly reconstruct the input state. This can be seen from the fact that the space of possible input states spans a K -dimensional subspace. Due to unitary constraints, preparing the states that can be deterministically denoised for all possible inputs requires at least a K -dimensional auxiliary space, thus requiring in total $2K$ -dimensional U_e and U_d . \square

Appendix D: Coherent states and single photon Fock states

For a large class of noise models, one can train our quantum autoencoder with coherent states as input. After training this way, the trained autoencoder can successfully denoise single photon input states subject to the same noise model.

To see this, first note that transformations with linear optics over N modes can be described by a $N \times N$ unitary U . This unitary transforms the creation operator of the k th mode \hat{a}_k^\dagger via $\hat{a}_k^\dagger \rightarrow \sum_\ell U_{k\ell} \hat{a}_\ell^\dagger$. We now discuss the effect of U on single photon Fock states and coherent state as input to the quantum autoencoder.

For a single photon input on the first mode, after application of U we get $|\psi_F\rangle = U \hat{a}_1^\dagger |\text{vac}\rangle = \sum_\ell U_{1\ell} \hat{a}_\ell^\dagger |\text{vac}\rangle$, where $|\text{vac}\rangle$ is the vacuum state without photons. Measurement of the average photon number in mode k gives us $\langle\psi_F| \hat{a}_k^\dagger \hat{a}_k |\psi_F\rangle = |U_{1k}|^2$.

For a coherent state input on the first mode we have $|\psi_C\rangle = UD_1(\alpha)|\text{vac}\rangle = \prod_\ell D_\ell(U_{1\ell}\alpha)|\text{vac}\rangle$ where $D_\ell(\alpha) = \exp(-\alpha\hat{a}_\ell - \alpha^*\hat{a}_\ell^\dagger)$ is the displacement operator acting on mode ℓ and α the amplitude of the coherent state. The average photon number of mode k is given by $\mathcal{I}_k = \langle\psi_C|\hat{a}_k^\dagger\hat{a}_k|\psi_C\rangle = |\alpha|^2|U_{1k}|^2$. Thus, training on the population on the trash mode for coherent states and single photon Fock states is equivalent up to a constant scaling factor. As the fidelity is measured using the photon population after the inverse of the state generation unitary, the equivalence also applies to fidelity training,

By extending above arguments to mixtures, we note that single photons and coherent state inputs also show equivalent populations under a large class of noise channels affecting the state. In particular, the equivalence holds for any noise channel $\mathcal{E}(\rho) = \sum_k V_k \rho V_k^\dagger$ where the Kraus operators V_k can be expressed in terms of linear optical elements.