Supplementary Information for "Sample-efficient quantum error mitigation via classical learning surrogates "

CONTENTS

A. Preliminary results 1
1. Pauli transfer matrix and trigonometric expansion of quantum circuits 1
2. Classical learning surrogates 3
3. A summary of the employed noisy models 4
4. Complementary with other sampling reduction methods 5
5. Extrapolation functions used in ZNE 6

B. Proof of Theorem 1 7
1. The Lipschitz constant of the extrapolation function 9

C. Extension of S-ZNE in the hybrid scenario 10

D. Additional experimental results of S-ZNE 11
1. Circuit implementation 11
2. Random feature sampling for classical learning surrogates 12
3. Robustness of S-ZNE across varied extrapolation functions 12
4. Data efficiency in quantum metrology 13

E. Simulation results of hybrid S-ZNE 14

References 17

Appendix A: Preliminary results

This section presents the theoretical foundations and preliminary results underpinning the S-ZNE framework. In particular, SI A 1 provides the mathematical framework of Pauli transfer matrix and trigonometric expansion for both noiseless and noisy quantum circuits, SI A 2 details the implementation of two classical learning surrogates, SI A 3 summarizes the noise models employed in numerical simulations, SI A 4 discusses the complementarity with other sampling reduction methods, and SI A 5 elaborates on the extrapolation functions used in ZNE.

1. Pauli transfer matrix and trigonometric expansion of quantum circuits

Pauli Transfer Matrix. Here we review how to use the Pauli-Liouville representation to formulate the quantum state and the observable. Denote P_l as the l-th normalized Pauli operator with $P_l \in \frac{1}{\sqrt{2^N}}\{\mathbb{I}, X, Y, Z\}^{\otimes N}$, which satisfies $\mathrm{Tr}(P_l P_k) = \delta_{lk}$. An arbitrary density matrix ρ could be expanded by a set of normalized Pauli operators, i.e.,

$$\rho = \sum_k c_k P_k, \quad \text{with} \quad c_k = \mathrm{Tr}(P_k \rho).$$

We can denote ρ as a 4^N-dimension vector under the normalized Pauli bases, i.e.,

$$|\rho\rangle\rangle = [c_1, \cdots, c_k, \cdots, c_{4^N}]^\top.$$

Given a circuit $U(\boldsymbol{x})$, its representation under Pauli bases is termed as Pauli transfer matrix (PTM) [1]. The matrix element $[\mathfrak{U}(\boldsymbol{x})]_{lk}$ at the l-th row and k-th column yields:

$$[\mathfrak{U}(\boldsymbol{x})]_{lk} = \mathrm{Tr}(P_l U(\boldsymbol{x}) P_k U(\boldsymbol{x})^\dagger) = \langle\langle P_j | \mathfrak{U}(\boldsymbol{x}) | P_K \rangle\rangle,$$

where $|\rho\rangle\rangle$ denotes the quantum state under PTM representation. To be concrete, the PTM representation of $\mathrm{RZ}(\boldsymbol{x}_j)$ gates takes the form as

$$\mathrm{RZ}(\boldsymbol{x}_j) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\boldsymbol{x}_j) & -\sin(\boldsymbol{x}_j) & 0 \\ 0 & \sin(\boldsymbol{x}_j) & \cos(\boldsymbol{x}_j) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \mathrm{D}_0 + \cos(\boldsymbol{x}_j)\mathrm{D}_1 + \sin(\boldsymbol{x}_j)\mathrm{D}_{-1} \tag{A1}$$

where $\mathrm{D}_0 = \left(\begin{smallmatrix} 1&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&1 \end{smallmatrix}\right)$, $\mathrm{D}_1 = \left(\begin{smallmatrix} 0&0&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&0 \end{smallmatrix}\right)$, and $\mathrm{D}_{-1} = \left(\begin{smallmatrix} 0&0&0&0 \\ 0&0&-1&0 \\ 0&1&0&0 \\ 0&0&0&0 \end{smallmatrix}\right)$.

Trigonometric expansion of noiseless quantum circuits. Following the main text, let us consider an N-qubit quantum circuit in the form of $U(\boldsymbol{x}) = \left(\prod_{j=1}^{d} C_j\,\mathrm{RZ}(\boldsymbol{x}_j)\right)C_0$. When applied to an arbitrary N-qubit input state ρ_0, the generated state can be reformulated by the trigonometric expansion under the PTM form, which is

$$\rho(\boldsymbol{x}) = U(\boldsymbol{x})\rho_0 U(\boldsymbol{x})^\dagger = \sum_{\boldsymbol{\omega}} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\langle\langle\rho_0|\mathfrak{U}_{\boldsymbol{\omega}}^\dagger.$$

Here the notation $\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})$ with $\boldsymbol{\omega} \in \{0, 1, -1\}^d$ refers to the trigonometric monomial basis

$$\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) = \prod_{j=1}^{d} \begin{cases} 1 & \text{if } \boldsymbol{\omega}_j = 0 \\ \cos(\boldsymbol{x}_j) & \text{if } \boldsymbol{\omega}_j = 1 \\ \sin(\boldsymbol{x}_j) & \text{if } \boldsymbol{\omega}_j = -1 \end{cases}.$$

In addition, the matrix $\mathfrak{U}_{\boldsymbol{\omega}}$ is a combination of permutation and masking matrices. The expectation value $\mathrm{Tr}(\rho(\boldsymbol{x})O)$ can also be expressed using the trigonometric monomial bases, i.e.,

$$f(\rho(\boldsymbol{x}), O) \equiv \mathrm{Tr}(\rho(\boldsymbol{x})O) = \sum_{\boldsymbol{\omega}} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\langle\langle\rho_0|\mathfrak{U}_{\boldsymbol{\omega}}^\dagger|O\rangle\rangle. \tag{A2}$$

Trigonometric expansion in the noisy scenario. Ref. [2] generalizes the above mathematical expression into the noisy scenario. Denote a single-qubit Pauli channel and a multi-qubit Pauli channel by $\mathcal{N}_P$ and $\mathcal{M}$, respectively. The parametrized ansatz in $U(\boldsymbol{x})$ under the Pauli channel can be represented as

$$\tilde{U}(\boldsymbol{x}) = \prod_{l=1}^{d} \left(\tilde{C}_l \widetilde{\mathrm{RZ}}(\boldsymbol{x}_l)\right)\tilde{C}_0, \tag{A3}$$

where $\widetilde{\mathrm{RZ}}(\boldsymbol{x}_l) = \mathcal{N}_P \circ \mathrm{RZ}(\boldsymbol{x}_l)$ and $\tilde{C}_l = \mathcal{M}_l \circ C_l$ with M_l being a multi-qubit Pauli channel applied to the l-th Clifford operation C_l.

As with the noiseless case, this noisy unitary can be effectively converted into the PTM representation. This is because under PTM, both single-qubit and multi-qubit Pauli channels can be rewritten as fixed diagonal matrices. More specifically, recall the definition of a single-qubit Pauli channel, i.e.,

$$\mathcal{N}_P[\rho] = (1 - p_X - p_Y - p_Z)\rho + p_X X\rho X + p_Y Y\rho Y + p_Z Z\rho Z, \tag{A4}$$

where p_X, p_Y, and p_Z denote the Pauli error rates along X, Y, and Z axes. Under the PTM representation, the single-qubit Pauli channel transforms into a diagonal matrix with

$$\mathrm{N} = \mathrm{diag}(1, q_X, q_Y, q_Z), \tag{A5}$$

where $q_X = 1 - 2(p_Z + p_Y)$, $q_Y = 1 - 2(p_Z + p_X)$, and $q_Z = 1 - 2(p_X + p_Y)$. Similarly, an arbitrary multi-qubit Pauli channel $\mathcal{M}$ can be rewritten as a diagonal matrix under PTM representation.

Following the previous noiseless result, the noisy expectation value of Eq. (A2) can still be expanded into a set of trigonometric monomial bases, i.e.,

$$f(\tilde{\rho}(\boldsymbol{x}), O) \equiv \mathrm{Tr}(\tilde{\rho}(\boldsymbol{x})O) = \sum_{\boldsymbol{\omega}} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x})\langle\langle O|\tilde{\mathfrak{U}}_{\boldsymbol{\omega}}|\rho_0\rangle\rangle. \tag{A6}$$

Compared to the noiseless case, the only difference is that $\{\tilde{\mathfrak{U}}_{\boldsymbol{\omega}}\}$ depends on the noisy rate of Pauli channels.

2. Classical learning surrogates

In this subsection, we provide implementation details of two classical learning surrogates employed in S-ZNE, which are kernel-based and regression-based surrogates. Without loss of generality, we focus on elucidating the implementation of both classical learning surrogates when predicting the noisy expectation value at the j-th level, i.e., $f(\boldsymbol{x}, O, \lambda_j)$.

Kernel-based surrogate h_{cs}. The kernel-based learning surrogate is designed for circuits containing independent parameters $\boldsymbol{x}$ and supporting varied observables with bounded locality. At the j-th noise level, when the unitary folding method is adopted, the explicit form of the circuit implementation under the Pauli channel is

$$\tilde{U}(\boldsymbol{x}; \lambda_j) = \prod_{k=1}^{\lambda_j} \left(\prod_{l=1}^{d} \left(\tilde{C}_l \widetilde{\mathrm{RZ}}(\boldsymbol{x}_{l,k}) \right) \tilde{C}_0 \right). \tag{A7}$$

In other words, we repeat the implementation of $U(\boldsymbol{x})$ with λ_j times. As a result, the total number of classical controls in the circuit increases to $\lambda_j \times d$.

During the data collection phase, the classical input $\boldsymbol{x}$ is randomly and uniformly sampled from $[-\pi, \pi]^{\lambda_j \times d}$, and the Pauli-based classical shadows [3] are used to acquire the classical representations of the resulting state prepared by $\tilde{U}(\boldsymbol{x}; \lambda_j)$. The collected shadow state is denoted by $\tilde{\rho}_T(\boldsymbol{x})$ with T being the number of snapshots. In this way, we can construct the training dataset $\mathcal{T}(\lambda_j) = \{\boldsymbol{x}^{(i,j)}, \tilde{\rho}_T(\boldsymbol{x}^{(i,j)})\}_{i=1}^{n_j}$ with n_j training examples.

Given access to the established training dataset $\mathcal{T}(\lambda_j)$, the kernel-based learning surrogate with respect to the observable O yields

$$h_{\mathsf{cs}}(\boldsymbol{x}, O, \lambda_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} \kappa_\Lambda(\boldsymbol{x}, \boldsymbol{x}^{(i)}) \operatorname{Tr}(\tilde{\rho}_T(\boldsymbol{x}^{(i)}) O), \tag{A8}$$

where $\kappa_\Lambda(\boldsymbol{x}, \boldsymbol{x}^{(i)}) = \sum_{\boldsymbol{\omega} \in \mathfrak{C}(\Lambda)} 2^{\|\boldsymbol{\omega}\|_0} \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \Phi_{\boldsymbol{\omega}}(\boldsymbol{x}^{(i)})$ is the truncated trigonometric monomial kernel and the truncated frequency set is $\mathfrak{C}(\Lambda) = \{\boldsymbol{\omega} \in \{0, \pm 1\}^d \mid \|\boldsymbol{\omega}\|_0 \leq \Lambda\}$. The feature map $\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})$ is defined as:

$$\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) = \prod_{l=1}^{d} \begin{cases} 1 & \text{if } \omega_l = 0, \\ \cos(x_l) & \text{if } \omega_l = 1, \\ \sin(x_l) & \text{if } \omega_l = -1. \end{cases} \tag{A9}$$

The prediction performance of the kernel-based learning surrogate is warranted by the following lemma.

Lemma 1 ([2, Adapted from Theorem 1]). *Assume* $\mathbb{E}_{\boldsymbol{x} \sim \mathrm{Unif}[-\pi,\pi]^d} \|\nabla_{\boldsymbol{x}} \operatorname{Tr}(\tilde{\rho}(\boldsymbol{x}, \lambda_j) O)\|_2^2 \leq C$. *Let* $O = \sum_i O_i$ *be a* K*-local observable with* $\sum_i \|O_i\|_\infty \leq B$. *Consider a quantum circuit affected by a Pauli noise channel* $\mathcal{N}_P(p_X, p_Y, p_Z)$, *characterized by* $p = \min\{p_X, p_Y\}$ *and* p_Z. *Let* $h_{\mathsf{cs}}(\boldsymbol{x}, O, \lambda_j)$ *be the learning surrogate in Eq. (A8) trained on* n_j *samples with* $\Lambda = \min\{\Lambda_C, \Lambda_p\}$, *where* $\Lambda_C = 4C/\epsilon_j$ *and* $\Lambda_p = \frac{1}{2(p+p_Z)} \log(2B/\sqrt{\epsilon_j})$.

When the number of training examples satisfies

$$n_j \geq |\mathfrak{C}(\Lambda)| \frac{2B^2 9^K}{\epsilon_j} \log\left(\frac{2 \cdot |\mathfrak{C}(\Lambda)|}{\delta} \right), \tag{A10}$$

with probability at least $1 - \delta$, *the average prediction error of the kernel-based learning surrogate is bounded by*

$$\mathbb{E}_{\boldsymbol{x} \sim \mathrm{Unif}[-\pi,\pi]^{d \times \lambda_j}} |h_{\mathsf{cs}}(\boldsymbol{x}, O, \lambda_j) - f(\boldsymbol{x}, O, \lambda_j)|^2 \leq \epsilon_j. \tag{A11}$$

We remark that ZNE with unitary folding requires the classical input to be correlated among λ_j blocks. However, the performance guarantee of kernel-based learning surrogate rests on the independence of different entries. As such, when kernel-based learning surrogates are employed, the error bound of S-ZNE additionally depends on the domain generalization capability during inference. To this end, we conduct systematic numerical simulations to validate the capabilities of h_{cs} in S-ZNE. Refer to SI D for more details.

Regression-based surrogate h_{qs}. The regression-based surrogate applies to the scenario in which the classical controls are correlated. More specifically, the explored noisy quantum circuit at the λ_j-th level takes the form of

$$\tilde{U}(\boldsymbol{x}; \lambda_j) = \prod_{k=1}^{\lambda_j} \left(\prod_{l=1}^{d} \left(\tilde{C}_l \widetilde{\mathrm{RZ}}(\boldsymbol{x}_l) \right) \tilde{C}_0 \right), \tag{A12}$$

where the l-th entry $\boldsymbol{x}_l$ in $\boldsymbol{x}$ repeats across λ_j blocks. This formalism aligns with ZNE with unitary folding, as classical input $\boldsymbol{x}$ is correlated in the circuit.

The training dataset for the regression-based learning surrogate is defined as $\mathcal{T}(\lambda_j) = \{\boldsymbol{x}^{(i,j)}, y^{(i,j)}\}_{i=1}^{n_j}$. Here the input $\boldsymbol{x}^{(i,j)}$ is sampled from an arbitrary distribution $\mathbb{D}$ within a bounded interval $[-R, R]^d$, and $y^{(i)}$ is the estimated mean value of the observable O with T shots. The mathematical expression of the regression-based surrogate is

$$h_{\mathsf{qs}}(\boldsymbol{x}, O, \lambda_j; \mathbf{w_j}) = \langle \mathbf{\Phi}_{\mathfrak{C}(\Lambda)}(\boldsymbol{x}), \mathbf{w}_j \rangle, \tag{A13}$$

where the high frequency terms with $||\boldsymbol{\omega}||_0 > \Lambda$ is truncated. The weight $\mathbf{w}_j$ is obtained by solving the following ridge regression optimization problem, i.e.,

$$\min_{\mathbf{w}_j} \left\{ \frac{1}{n_j} \sum_{i=1}^{n_j} \left(y^{(i,j)} - \langle \mathbf{\Phi}_{\mathfrak{C}(\Lambda)}(\boldsymbol{x}^{(i,j)}), \mathbf{w}_j \rangle \right)^2 + \gamma \|\mathbf{w}_j\|_2^2 \right\}, \tag{A14}$$

where $\gamma > 0$ is a regularization hyperparameter. To explicitly capture parameter correlations, we partition the d-dimensional parameter vector $\boldsymbol{x}$ into S groups $\mathfrak{x}_1, \ldots, \mathfrak{x}_S$, where parameters in each group are identical. For each frequency vector $\boldsymbol{\omega} \in \mathfrak{C}(\Lambda)$, let $\omega_{s,k}$ denote the component of $\boldsymbol{\omega}$ corresponding to the k-th parameter in group s. The feature map is then defined as:

$$\Phi_{\boldsymbol{\omega}}(\boldsymbol{x}) = \prod_{s=1}^{S} \left[\cos(\mathfrak{x}_s)^{N_s^+(\boldsymbol{\omega})} \cdot \sin(\mathfrak{x}_s)^{N_s^-(\boldsymbol{\omega})} \right], \tag{A15}$$

where $N_s^+(\boldsymbol{\omega}) = \sum_{k=1}^{d_s} \mathbb{1}\{\omega_{s,k} = 1\}$ and $N_s^-(\boldsymbol{\omega}) = \sum_{k=1}^{d_s} \mathbb{1}\{\omega_{s,k} = -1\}$ count the occurrences of cosine and sine terms in group s, respectively, with d_s being the number of parameters in group s.

The performance error of the regression-based learning surrogate is provided in the following lemma.

Lemma 2 ([2, Adapted from Theorem 2])**.** *Following notations in Eq. (A13), let $q = 1 - 2(p + p_Z)$ with $p = \min\{p_X, p_Y\}$, Λ be the threshold of the truncated frequency set $\mathfrak{C}(\Lambda)$, and ϵ_l be the maximal estimation error of $\{y^{(i,j)}\}$ in $\mathcal{T}(\lambda_j)$, namely, $\max_{i\in[n]} |y^{(i,j)} - f(\boldsymbol{x}^{(i,j)})| \leq \epsilon_l$. Assume $q(1+R) < 1/e$. Define $\epsilon = 16B^2(deq(1+R)/\Lambda)^{2\Lambda}$. When the following conditions are satisfied: (i) $\epsilon_l \leq \sqrt{\epsilon}/4$, (ii) the frequency is truncated to $\Lambda > deq(1+R)$, (iii) the number of training examples satisfies*

$$n_j = \left(\frac{1}{q(1+R)} \right)^{4\Lambda} \cdot \frac{\log(1/\delta)}{9}, \tag{A16}$$

the predictive surrogate $h_{\mathsf{qs}}(\boldsymbol{x})$ achieves with probability at least $1 - \delta$,

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}} |h_{\mathsf{qs}}(\boldsymbol{x}, O, \lambda_j) - f(\boldsymbol{x}, O, \lambda_j)|^2 \leq \epsilon_j. \tag{A17}$$

Remark. Throughout the remainder of this study, we sometimes use h_{qs} and h_{cs} to specify which surrogate model is being referenced, so as to avoid confusion.

3. A summary of the employed noisy models

This subsection details the standard noise models employed in numerical simulations, which include depolarization noise, thermal noise, and coherent noise.

Depolarizing noise. We consider both *local* and *global* depolarizing noisy channels, both of which fall within the class of Pauli channels.

For an N-qubit system, the *local* depolarizing channel acts independently on each qubit. It is defined as the tensor product of single-qubit depolarizing channels applied to each qubit. The single-qubit depolarizing channel is described by the Kraus operators

$$K_0 = \sqrt{1 - \tfrac{3p_d}{4}}\, \mathbb{I}, \quad K_1 = \sqrt{\tfrac{p_d}{4}}\, X, \quad K_2 = \sqrt{\tfrac{p_d}{4}}\, Y, \quad K_3 = \sqrt{\tfrac{p_d}{4}}\, Z, \tag{A18}$$

where $p_d \in [0, 1]$ is the depolarizing rate per qubit. The corresponding map for one qubit is $\mathcal{E}_d(\rho) = (1 - p_d)\rho + p_d \frac{\mathbb{I}}{2}$. For N qubits, the local depolarizing channel is $\mathcal{E}_d^{\otimes N}$.

The *global* depolarizing channel acts collectively on all N qubits. Its CPTP map is given by

$$K_0 = \sqrt{1 - p_g + \frac{p_g}{4^N}}\, \mathbb{I}^{\otimes N}, \tag{A19}$$

$$K_\alpha = \sqrt{\frac{p_g}{4^N}}\, P_\alpha, \quad \alpha = 1, 2, \ldots, 4^N - 1, \tag{A20}$$

where $\{P_\alpha\}$ denotes the set of all non-identity N-qubit Pauli operators, and $p_g \in [0, 1]$ is the global depolarizing rate. This channel can be equivalently expressed as $\mathcal{N}_g(\rho) = (1 - p_g)\rho + p_g(\mathbb{I}/2)^{\otimes N}$.

Thermal noise. Thermal relaxation noise combines energy dissipation (T_1 decay) and pure dephasing (T_2 decoherence). Let t_g denote the gate duration. We define the reset probability and dephasing probability as $p_r = 1 - e^{-t_g/T_1}$ and $p_z = \frac{1}{2}\left(1 - e^{-t_g/T_\phi}\right)$, respectively. The pure dephasing time satisfies $T_\phi^{-1} = T_2^{-1} - (2T_1)^{-1}$. The following describes the single-qubit thermal relaxation channel, which is applied independently to each qubit in multi-qubit simulations.

Case $T_2 \leq T_1$. The single-qubit thermal noisy channel is implemented via six Kraus operators

$$\begin{aligned} &K_0 = \sqrt{1 - p_z - p_r}\, \mathbb{I}, \quad K_1 = \sqrt{p_z}\, Z, \quad K_2 = \sqrt{p_r(1 - p_e)}\, |0\rangle\langle 0|, \\ &K_3 = \sqrt{p_r(1 - p_e)}\, |0\rangle\langle 1|, \quad K_4 = \sqrt{p_r p_e}\, |1\rangle\langle 0|, \quad K_5 = \sqrt{p_r p_e}\, |1\rangle\langle 1|, \end{aligned} \tag{A21}$$

where $p_e \in [0, 1]$ is the excited state population determined by the thermal equilibrium.

Case $T_1 < T_2$. The single-qubit thermal noisy channel is implemented via the Choi matrix:

$$\Lambda = \begin{bmatrix} 1 - p_e p_r & 0 & 0 & e^{-t_g/T_2} \\ 0 & p_e p_r & 0 & 0 \\ 0 & 0 & (1 - p_e)p_r & 0 \\ e^{-t_g/T_2} & 0 & 0 & 1 - (1 - p_e)p_r \end{bmatrix}.$$

This thermal noise model is specifically used in the experiments reported in SI E.

Coherent noise. We model coherent errors as small, stochastic miscalibrations of rotation angles in single-qubit rotation gates. For an ideal rotational gate $\mathrm{R}_P(x) = e^{-ixP/2}$ with $P \in \{X, Y, Z\}$, its noisy implementation is given by

$$\mathrm{R}_P(x) \;\mapsto\; e^{-i(x+\theta_P)P/2}, \tag{A22}$$

where the angular offset θ_P for each gate is drawn independently from a uniform distribution. This model captures both over-/under-rotation and axis misalignment effects.

4. Complementary with other sampling reduction methods

Prior studies relevant to S-ZNE can be broadly classified into two categories: (i) approaches that aim to reduce the sampling overhead of zero-noise extrapolation (ZNE) itself, and (ii) methods that mitigate sampling costs in specific quantum tasks, especially for variational quantum algorithms (VQAs). We discuss each in turn.

ZNE variants. To the best of our knowledge, only one study has explicitly sought to reduce the measurement overhead of conventional ZNE. In particular, Liao *et al.* [4] introduced random forest ZNE (RF-ZNE), which trains a random forest model to predict the noiseless expectation value directly from circuit descriptors and noisy outcomes, thereby bypassing explicit noise scaling and extrapolation at inference time. A key distinction between S-ZNE and RF-ZNE lies in their theoretical grounding: S-ZNE retains the rigorous physical interpretability of extrapolation under controlled noise scaling, whereas RF-ZNE operates as a data-driven surrogate without explicit noise modeling.

Task-specific sampling overhead reduction. While general-purpose strategies for reducing ZNE's measurement cost remain scarce, substantial progress has been made in curbing sampling overhead in VQAs, whose optimization typically demands extensive quantum measurements due to non-convex loss landscapes [5] and the forbidding of back-propagation [6]. To date, three principal lines of research have emerged to address this challenge: smart initialization, intelligent optimizers, and measurement grouping.

Smart initialization techniques. Smart initialization techniques aim to reduce the number of quantum measurements by selecting high-quality initial parameters, rather than initializing randomly. These methods can be broadly divided

into heuristic and informative approaches. Heuristic initializers are implemented entirely classically and do not require access to a quantum processor; a common example is small-angle or "identity" initialization, where parameters are set near zero so that the initial circuit closely approximates the identity operation [7–11]. In contrast, informative initializers leverage either prior quantum data or classical surrogate models to construct informed initial parameters. These include warm-start methods [12, 13], parameter transfer across related problem instances [14–16], and pre-training strategies that use classical emulators, such as Lie-algebraic surrogates [17], matrix product state–based models [18, 19], and neural-network approaches [20, 21], to perform substantial portions of the optimization classically before any quantum execution. By reducing the distance to a high-quality solution in parameter space, such initialization schemes significantly lower the quantum sampling cost of subsequent optimization.

Intelligent optimizers. Intelligent optimizers employ classical machine learning to reduce quantum sampling costs in variational algorithms by predicting optimization trajectories. For instance, meta-learning with recurrent neural networks can generate informed parameter initializations, cutting down the number of quantum-classical iterations needed for convergence [22]. Other approaches include QuACK, which applies Koopman operator theory to create a linear representation of gradient-based optimization, enabling faster convergence in tasks such as quantum chemistry, and PALQO, which uses physics-informed neural networks to model VQA training and predict multi-step parameter updates from limited quantum data [23, 24]. These methods collectively lower sampling overhead by shifting significant parts of the optimization process to classical computation, while maintaining comparable solution quality.

Measurement grouping strategies. Measurement grouping strategies reduce sampling overhead by exploiting compatibility among terms in the target observable (e.g., a Hamiltonian) to jointly estimate multiple terms in a single measurement setting. The central idea is to partition the observable into subsets of mutually commuting or simultaneously measurable operators, thereby minimizing the total number of distinct quantum measurements required. This approach has been widely adopted as a standard technique for shot-efficient expectation estimation in variational algorithms [25–27].

The proposed S-ZNE is fully compatible with these broader strategies. We leave their integration for future investigation.

5. Extrapolation functions used in ZNE

The choice of extrapolation function $g(\cdot)$ significantly influences the performance of ZNE [28]. Following the notation in the main text, for a given classical input $\boldsymbol{x}$ and observable O, let $\boldsymbol{z} = \{z_1, \ldots, z_u\}$ be the vector of noisy expectation value estimates corresponding to noise scales $\{\lambda_j\}_{j=1}^u$, where each z_j is either an experimental estimate $\hat{f}(\boldsymbol{x}, O, \lambda_j)$ or a surrogate prediction $h(\boldsymbol{x}, O, \lambda_j)$. The extrapolation function $g(\cdot)$ maps $\boldsymbol{z}$ to an estimate of the zero-noise expectation $f(\boldsymbol{x}, O)$. Below we detail the functional forms of $g(\cdot)$ used in our numerical simulations.

Linear extrapolation. This method assumes an approximately linear dependence of the observable on the noise scale. For $u \geq 2$, the data points $\{(\lambda_j, z_j)\}_{j=1}^u$ are used to fit a linear model via ordinary least squares, i.e.,

$$\arg\min_{a_0, a_1} \sum_{j=1}^{u} \big(z_j - a_0 - a_1 \lambda_j\big)^2.$$

The closed-form solution is given by

$$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \big(V^\top V\big)^{-1} V^\top \boldsymbol{z}, \quad V = \begin{bmatrix} 1 & \lambda_1 \\ \vdots & \vdots \\ 1 & \lambda_u \end{bmatrix}.$$

The zero-noise estimate is then obtained by evaluating the fitted model at $\lambda_0 = 0$, yielding $g(\boldsymbol{z}) = a_0$.

Quadratic extrapolation. To capture possible nonlinear behavior, this method fits a quadratic model when $u \geq 3$, i.e.,

$$(b_0, b_1, b_2) = \arg\min_{b_0, b_1, b_2} \sum_{j=1}^{u} \big(z_j - b_0 - b_1 \lambda_j - b_2 \lambda_j^2\big)^2.$$

The solution is

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \big(V_2^\top V_2\big)^{-1} V_2^\top \boldsymbol{z}, \quad V_2 = \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 \\ \vdots & \vdots & \vdots \\ 1 & \lambda_u & \lambda_u^2 \end{bmatrix}.$$

The zero-noise estimate is again the constant term with $g(\boldsymbol{z}) = b_0$.

Richardson extrapolation. This approach assumes the underlying noise dependence can be modeled by a polynomial of degree at most $u-1$. The zero-noise limit is obtained by constructing the unique polynomial of degree at most $u-1$ that interpolates the u points $\{(\lambda_j, z_j)\}_{j=1}^u$, and evaluating it at $\lambda_0 = 0$. Mathematically, the extrapolated value is given by the Lagrange interpolation formula with

$$g(\boldsymbol{z}) = \sum_{j=1}^{u} \gamma_j z_j, \quad \gamma_j = \prod_{\substack{k=1 \\ k\neq j}}^{u} \frac{\lambda_k}{\lambda_k - \lambda_j}.$$

This yields the exact zero-noise value if the observable varies polynomially with λ of degree less than u.

Appendix B: Proof of Theorem 1

In this section, we analyze the total estimation error of S-ZNE, comparing it to that of conventional ZNE. Following notations in the main text, we denote $f(\boldsymbol{x}, O) \equiv f(\boldsymbol{x}, O, \lambda_0) = \mathrm{Tr}(\rho(\boldsymbol{x})O)$ as the ideal zero-noise expectation value. Let $g(\cdot)$ be the extrapolation function mapping a vector of noisy expectation values at different noise levels to an estimate of the zero-noise value.

Recall that the three types of data vectors used in ZNE or S-ZNE for a given input $\boldsymbol{x}$, i.e.,

- $\boldsymbol{z}_I(\boldsymbol{x}) = \{f(\boldsymbol{x}, O, \lambda_1), \cdots, f(\boldsymbol{x}, O, \lambda_u)\}$: The ideal vector of exact expectation values at with noise levels $\{\lambda_j\}_{j=1}^u$.
- $\boldsymbol{z}_S(\boldsymbol{x}) = \{h_{\mathsf{qs}}(\boldsymbol{x}, O, \lambda_1), \cdots, h_{\mathsf{qs}}(\boldsymbol{x}, O, \lambda_u)\}$: The vector obtained using the classical surrogate predictions $h_{\mathsf{qs}}(\boldsymbol{x}, O, \lambda_j)$ for S-ZNE with noise levels $\{\lambda_j\}_{j=1}^u$.
- $\boldsymbol{z}_C(\boldsymbol{x}) = \{\hat{f}(\boldsymbol{x}, O, \lambda_1), \cdots, \hat{f}(\boldsymbol{x}, O, \lambda_u)\}$: The vector obtained using experimental estimates $\hat{f}(\boldsymbol{x}, O, \lambda_j)$ from M measurements for conventional ZNE with noise levels $\{\lambda_j\}_{j=1}^u$.

Our goal is to derive the upper bound for the average performance between S-ZNE and noiseless expectation values in terms of MSE, i.e., $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|f(\boldsymbol{x}, O) - g(\boldsymbol{z}_S(\boldsymbol{x}))|^2$, as well as the average performance between conventional ZNE and noiseless expectation values, i.e., $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|f(\boldsymbol{x}, O) - g(\boldsymbol{z}_C(\boldsymbol{x}))|$. The expectation is taken over an arbitrary distribution $\mathbb{D}$ supported on the interval $[-R, R]$. To achieve this goal, we leverage the result of Lemma 2.

Now we are ready to present the formal statement of Theorem 1 and the proof details.

Theorem (Formal statement of Theorem 1)**.** *Suppose the explored family of circuits $U(\boldsymbol{x})$ undergoes Pauli noise channel in Eq. (A4) and has correlated parameters $\boldsymbol{x} \in [-R, R]^d$ sampled from a distribution $\mathbb{D}$. Let $f(\boldsymbol{x}, O)$ be the ideal zero-noise limit and $g(\cdot)$ be the employed extrapolation function with the Lipschitz constant L. Denote $\zeta^2 = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|f(\boldsymbol{x}, O) - g(\boldsymbol{z}_I(\boldsymbol{x}))|^2$ as the intrinsic extrapolation error with $\boldsymbol{z}_I(\boldsymbol{x}) = \{f(\boldsymbol{x}, O, \lambda_1), ..., f(\boldsymbol{x}, O, \lambda_u)\}$, where $\mathbb{D}$ is an arbitrary distribution supported on the interval $[-R, R]$. When the number of measurements adopted by the conventional ZNE is fixed to be M, with probability at least $1 - 0.05u$, its average performance is upper-bounded by*

$$\mathbb{E}_{\boldsymbol{x}}|f(\boldsymbol{x}, O) - g(\boldsymbol{z}_C(\boldsymbol{x}))|^2 \leq \zeta^2 + \frac{4L^2 u B^2}{M} \ln(40).$$

Following notations in Lemma 2, when (i) $\epsilon_l \leq \sqrt{\epsilon}/4$, (ii) the frequency threshold satisfies $\Lambda > deq(1+R)$, and (iii) the number of training examples for each regression-based surrogate satisfies

$$n_j \geq \frac{64B^2M^2}{3}\left(\frac{de}{\Lambda}\right)^{4\Lambda} \cdot \frac{\log(1/\delta)}{9}, \tag{B1}$$

with probability at least $1 - u\delta$, the average performance of S-ZNE is upper bounded by

$$\mathbb{E}_{\boldsymbol{x}}|f(\boldsymbol{x}, O) - g(\boldsymbol{z}_S(\boldsymbol{x}))|^2 \leq \zeta^2 + \frac{4L^2 u B^2}{M} \ln(40).$$

Proof of Theorem 1. We first analyze the upper bound of MSE between the ideal results and the outputs of the conventional ZNE. To achieve this goal, we leverage the extrapolated values under the ideal setting, i.e., $g(\boldsymbol{z}_I(\boldsymbol{x}))$. Accordingly, the corresponding upper bound is

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|f(\boldsymbol{x}, O) - g(\boldsymbol{z}_S(\boldsymbol{x}))|^2 &= \mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|f(\boldsymbol{x}, O) - g(\boldsymbol{z}_I(\boldsymbol{x})) + g(\boldsymbol{z}_I(\boldsymbol{x})) - g(\boldsymbol{z}_S(\boldsymbol{x}))|^2 && \text{(B2)}\\
&\leq 2\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|f(\boldsymbol{x}, O) - g(\boldsymbol{z}_I(\boldsymbol{x}))|^2 + 2\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|g(\boldsymbol{z}_I(\boldsymbol{x})) - g(\boldsymbol{z}_S(\boldsymbol{x}))|^2 && \text{(B3)}\\
&= 2\zeta^2 + 2\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|g(\boldsymbol{z}_I(\boldsymbol{x})) - g(\boldsymbol{z}_S(\boldsymbol{x}))|^2 && \text{(B4)}
\end{aligned}$$

where the inequality comes from the triangle inequality, and the last equality follows the definition of ζ^2.

Similarly, for S-ZNE, we can apply the same decomposition strategy to obtain the upper bound of MSE between S-ZNE and the zero-noise limit is

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|f(\boldsymbol{x},O)-g(\boldsymbol{z}_C(\boldsymbol{x}))|^2 \leq 2\zeta^2 + 2\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|g(\boldsymbol{z}_I(\boldsymbol{x}))-g(\boldsymbol{z}_S(\boldsymbol{x}))|^2. \tag{B5}$$

For both ZNE and S-ZNE, the term $2\zeta^2$ refers to the intrinsic error induced by the selected extrapolation function. In this regard, the second term in Eqs. (B4) and (B5) quantifies the error introduced by using finite measurements (i.e., $\boldsymbol{z}_C(\boldsymbol{x})$) or the output of the learning surrogate (i.e., $\boldsymbol{z}_S(\boldsymbol{x})$) instead of the ideal values $\boldsymbol{z}_I(\boldsymbol{x})$. In what follows, we separately derive the upper bound of these two terms.

Upper bound of the second term. By exploiting the assumption that the extrapolation function g is Lipschitz continuous with the constant L (with respect to the ℓ_2 norm). The upper bound of the second term in Eq. (B5) is

$$2\mathbb{E}_{\boldsymbol{x}}|g(\boldsymbol{z}_I(\boldsymbol{x}))-g(\boldsymbol{z}_C(\boldsymbol{x}))|^2 \leq 2L^2\sum_{j=1}^{u}\mathbb{E}_{\boldsymbol{x}}|f(\boldsymbol{x},O,\lambda_j)-\hat{f}(\boldsymbol{x},O,\lambda_j)|^2. \tag{B6}$$

Recall that the difference between $f(\boldsymbol{x},O,\lambda_j)$ and $\hat{f}(\boldsymbol{x},O,\lambda_j)$ is caused by the finite M measurements. Supported by the Hoeffding inequality, with probability at least 0.95, when the number of measurements is M, the conventional ZNE at each noise rate j satisfies

$$\left|f(\boldsymbol{x},O,\lambda_j)-\hat{f}(\boldsymbol{x},O,\lambda_j)\right| \leq \sqrt{\frac{2B^2}{M}\ln(\frac{2}{0.05})}. \tag{B7}$$

Combining the above two inequalities, the upper bound in Eq. (B5) yields

$$2\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|g(\boldsymbol{z}_I(\boldsymbol{x}))-g(\boldsymbol{z}_C(\boldsymbol{x}))|^2 \leq 2L^2\frac{2B^2}{M}\ln(\frac{2}{0.05}) = \frac{4L^2uB^2}{M}\ln(40). \tag{B8}$$

We next derive the upper bound of the second term in Eq. (B4), which amounts to the MSE between the ideal extrapolation and the extrapolation by the outputs from the regression-based learning surrogates $\{h_{\mathsf{qs}}(\boldsymbol{x},O,\lambda_j)\}$. As with the conventional ZNE, the property of Lipschitz continuity of $g(\cdot)$ gives

$$2\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|g(\boldsymbol{z}_I(\boldsymbol{x}))-g(\boldsymbol{z}_S(\boldsymbol{x}))|^2 \leq 2L^2\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}\|\boldsymbol{z}_I(\boldsymbol{x})-\boldsymbol{z}_S(\boldsymbol{x})\|_2^2 \leq 2L^2\sum_{j=1}^{u}\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|f(\boldsymbol{x},O,\lambda_j)-h_{\mathsf{qs}}(\boldsymbol{x},O,\lambda_j)|^2, \tag{B9}$$

where the last inequality employs the triangular inequality.

To attain a comparable performance with conventional ZNE, it amounts to analyzing the required number of training examples n_j to ensure that at each noise level λ_j, the prediction error of the regression-based surrogate, i.e., $\epsilon_j := \mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|f(\boldsymbol{x},O,\lambda_j)-h_{\mathsf{qs}}(\boldsymbol{x},O,\lambda_j)|^2$, is well bounded. More specifically, according to Eq. (B7), such error should be bounded by $2B^2\ln(40)/M$. By substituting this quantity into Lemma 2, we need to derive the required number of training examples n_j such that the average prediction error satisfies

$$\epsilon_j = 16B^2(deq(1+R)/\Lambda)^{2\Lambda} \leq \frac{2B^2\ln(40)}{M}.$$

To achieve this goal, we first reformulate the required number of training examples n_j of Lemma 2 in terms of ϵ_j. Formally, with probability at least $1-\delta$, the number of training examples yields

$$\begin{aligned}
n_j &= \left(\frac{1}{q(1+R)}\right)^{4\Lambda}\cdot\frac{\log(1/\delta)}{9} \\
&= \left(\frac{de}{\Lambda}\right)^{4\Lambda}\left(\frac{\Lambda}{deq(1+R)}\right)^{4\Lambda}\cdot\frac{\log(1/\delta)}{9} &\text{(B10)}\\
&= \left(\frac{de}{\Lambda}\right)^{4\Lambda}\left(\frac{deq(1+R)}{\Lambda}\right)^{-4\Lambda}\cdot\frac{\log(1/\delta)}{9} &\text{(B11)}\\
&= \left(\frac{de}{\Lambda}\right)^{4\Lambda}\left[\frac{16B^2}{16B^2}\left(\frac{deq(1+R)}{\Lambda}\right)^{2\Lambda}\right]^{-2}\cdot\frac{\log(1/\delta)}{9} &\text{(B12)}\\
&= 256B^4\left(\frac{de}{\Lambda}\right)^{4\Lambda}\frac{1}{\epsilon_j^2}\cdot\frac{\log(1/\delta)}{9}. &\text{(B13)}
\end{aligned}$$

When the error threshold is $2B^2\ln(40)/M$, with probability at least $1-\delta$, the corresponding number of training examples is

$$n_j \geq 256B^4\left(\frac{de}{\Lambda}\right)^{4\Lambda}\frac{M^2}{12B^2}\cdot\frac{\log(1/\delta)}{9} = \frac{64B^2M^2}{3}\left(\frac{de}{\Lambda}\right)^{4\Lambda}\cdot\frac{\log(1/\delta)}{9}.$$

Supported by the union bound, when the number of training examples n_j used in each regression-based learning surrogate exceeds the above threshold, with probability at least $1-u\delta$, we have

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}|f(\boldsymbol{x},O)-g(\boldsymbol{z}_C(\boldsymbol{x}))|^2 \leq 2\zeta^2+\frac{4L^2uB^2}{M}\ln(40). \tag{B14}$$

□

1. The Lipschitz constant of the extrapolation function

Theorem 1 indicates that the performance of both conventional ZNEs and S-ZNE depends on the Lipschitz constant L. In this subsection, we comprehend the scaling of L when the linear extrapolation function $g(\cdot)$ introduced in SI A 5 is exploited. In addition, the unitary folding is adopted to construct the vectors $\boldsymbol{z}_I$, $\boldsymbol{z}_C$, and $\boldsymbol{z}_S$. Without loss of generality, we use $\boldsymbol{z}$ to denote these three vectors, where each entry z_j is either an ideal result $f(\boldsymbol{x},O,\lambda_j)$, an experimental estimation $\hat{f}(\boldsymbol{x},O,\lambda_j)$, or a surrogate prediction $h(\boldsymbol{x},O,\lambda_j)$.

When the linear extrapolation function is employed, its output is obtained via least-squares regression. Mathematically, we have

$$g(\boldsymbol{z}(\boldsymbol{x})) = a_0 = \langle \boldsymbol{s},\boldsymbol{z}\rangle \quad \text{with} \quad \boldsymbol{s}^\top = [1,0](W^\top W)^{-1}W^\top, \tag{B15}$$

where $W = \left[\begin{smallmatrix}1, & 1, & \cdots, & 1\\ 1, & 2, & \cdots, & u\end{smallmatrix}\right]^\top$. In this regard, the Lipschitz constant L can be derived by analyzing the upper bounded of the ℓ_2 norm of $\boldsymbol{s}$, i.e.,

$$L = \sqrt{\sum_{i=1}^{u}|\boldsymbol{s}_i|^2}.$$

In what follows, we derive the analytical form of each entry in $\boldsymbol{s}$. Specifically, the matrix $W^\top W \in \mathbb{R}^{2\times 2}$ equals to

$$W^\top W = \begin{bmatrix} u & \sum_{j=1}^{u} j \\ \sum_{j=1}^{u} j & \sum_{j=1}^{u} j \end{bmatrix} = \begin{bmatrix} u & \frac{u(u+1)}{2} \\ \frac{u(u+1)}{2} & \frac{u(u+1)(2u+1)}{6} \end{bmatrix}.$$

According, its inversion equals to

$$\left(W^\top W\right)^{-1} = \frac{12}{u^2(u^2-1)}\begin{bmatrix} \frac{u(u+1)(2u+1)}{6} & -\frac{u(u+1)}{2} \\ -\frac{u(u+1)}{2} & u \end{bmatrix}.$$

Combining this result with Eq. (B15), the explicit form of the i-th entry of $\boldsymbol{s}$ is

$$\boldsymbol{s}_i = \frac{12}{u^2(u^2-1)}\left(\frac{u(u+1)(2u+1)}{6}-\frac{u(u+1)}{2}i\right) = \frac{2u+1-6i}{u(u-1)}. \tag{B16}$$

Thereby, we have

$$L = \sqrt{\sum_{i=1}^{u}\left|\frac{2u+1-6i}{u(u-1)}\right|^2} = \sqrt{\frac{(2u+1)^2}{u(u-1)^2}}. \tag{B17}$$

In this regard, we can conclude that the Lipshitz constant L monotonically decreases with increasing u. When $u\to\infty$, we have $L\to\sqrt{4/u}$.

Appendix C: Extension of S-ZNE in the hybrid scenario

While the main text focuses on a complete substitution of quantum circuit executions with classical surrogate predictions within the ZNE framework (termed S-ZNE), we present here an extension involving partial substitution, yielding a hybrid paradigm for extrapolation. An intuition is illustrated in Fig. 1 of the main text. Compared to the original S-ZNE, this variant offers a flexible trade-off between the potential accuracy gains from retaining direct quantum measurements at low noise levels and the substantial resource savings afforded by surrogates, particularly at high noise levels where surrogate modeling is often more effective.

Methodology. The initial stages of this hybrid approach precisely mirror the full S-ZNE protocol described in the main text. The implementations of the first two steps are summarized below.

1. Data Acquisition. A training dataset $\mathcal{T}(\lambda_j) = \{\boldsymbol{x}^{(i,j)}, y^{(i,j)}\}_{i=1}^{n_j}$ is collected via quantum measurements (e.g., using classical shadows or direct expectation value estimation) for parameter samples $\boldsymbol{x}^{(i)}$ across all u noise levels $\{\lambda_j\}_{j=1}^{u}$ generated via noise scaling (e.g., unitary folding).

2. Surrogate Training. Based on $\mathcal{T}(\lambda_j)$, u distinct classical learning surrogates, $\{h(\boldsymbol{x}, O, \lambda_j)\}_{j=1}^{u}$, are trained. The surrogate model can be either kernel-based (as h_{cs} in Eq. (A8)) or regression-based (as h_{qs} in Eq. (A13)), where the training methodology is summarized in SI A 2.

The departure from the full S-ZNE method occurs in the subsequent *validation and selection stage*.

3. Validation and Thresholding. A separate and small validation set of parameter vectors, $\mathcal{X}_{\text{val}}$, is utilized. For each noise level λ_j, we evaluate the Mean Squared Error (MSE) between the surrogate's predictions and direct quantum circuit executions (obtained via additional quantum measurements only for the validation set):

$$\text{MSE}(\lambda_j) = \frac{1}{|\mathcal{X}_{\text{val}}|} \left| h(\boldsymbol{x}, O, \lambda_j) - \hat{f}(\boldsymbol{x}, O, \lambda_j) \right|^2 \tag{C1}$$

   Here, $\hat{f}(\boldsymbol{x}, O, \lambda_j)$ represents the expectation value obtained from executing the circuit on the quantum processor at noise level λ_j. We consistently observe that this MSE tends to decrease as the noise level λ_j increases. This trend is warranted by Lemma 1 and Lemma 2. As such, we establish an MSE threshold, η. After that, the set of noise levels (typically the highest ones) is identified for which the surrogate model meets this accuracy criterion: $\mathcal{J}_S = \{j \mid \text{MSE}(\lambda_j) \leq \eta\}$. Let $v = |\mathcal{J}_S|$ be the number of noise levels satisfying this condition.

4. Hybrid S-ZNE Construction. For any given inputs $\boldsymbol{x}$ during the inference stage, we construct a hybrid data vector $\boldsymbol{z}_H(\boldsymbol{x})$ for extrapolation. This vector selectively combines direct quantum measurements with surrogate predictions, i.e.,

$$\boldsymbol{z}_H(\boldsymbol{x}) = \{z_j(\boldsymbol{x})\}_{j=1}^{u}, \quad \text{where} \quad z_j(\boldsymbol{x}) = \begin{cases} \hat{f}(\boldsymbol{x}, O, \lambda_j) & \text{if } j \notin \mathcal{J}_s \\ h(\boldsymbol{x}, O, \lambda_j) & \text{if } j \in \mathcal{J}_s \end{cases} \tag{C2}$$

5. Extrapolation. Finally, the same extrapolation functions (e.g., polynomial, Richardson) employed in conventional ZNE and full S-ZNE are applied to the hybrid dataset $\boldsymbol{z}_H(\boldsymbol{x})$ to estimate the zero-noise expectation value $f(\boldsymbol{x}, O, \lambda = 0)$.

Discussion of trade-offs. The primary motivation for this hybrid protocol stems from scenarios where classical surrogates might exhibit non-negligible prediction errors, particularly at low noise levels. In such cases, completely replacing $\hat{f}(\boldsymbol{x}, O, \lambda_j)$ with $h(\boldsymbol{x}, O, \lambda_j)$ could potentially degrade the final extrapolation accuracy compared to conventional ZNE.

The proposed hybrid approach mitigates this risk by retaining direct quantum measurements for those low-noise data points where the surrogate's fidelity might be lower (i.e., where $\text{MSE}(\lambda_j) > \eta$), while still leveraging the efficiency of surrogates for the higher noise levels where they perform well and where quantum execution (requiring deeper circuits via folding) is most resource-intensive.

However, this potential accuracy retention comes at the cost of reduced quantum measurement savings during inference compared to the full S-ZNE approach. For each new parameter vector $\boldsymbol{x}$ evaluated, the hybrid method still requires executing the quantum circuit at $u - v$ noise levels. Consequently, the reduction in quantum measurement overhead compared to conventional ZNE is scaled by a factor of v/u, determined by the number of noise levels v where the surrogate meets the accuracy threshold η set during validation. The choice of η thus directly controls the balance between potential accuracy preservation and computational resource savings. Refer to SI E for more simulation results.

Appendix D: Additional experimental results of S-ZNE

This section presents more implementation details and simulation results omitted in the main text. In particular, SI D 1 provides the circuit implementation details for the explored tasks, SI D 2 explains the random feature sampling strategy for classical learning surrogates, SI D 3 evaluates the robustness of S-ZNE across different extrapolation models, and SI D 4 investigates its data efficiency in quantum metrology.

1. Circuit implementation

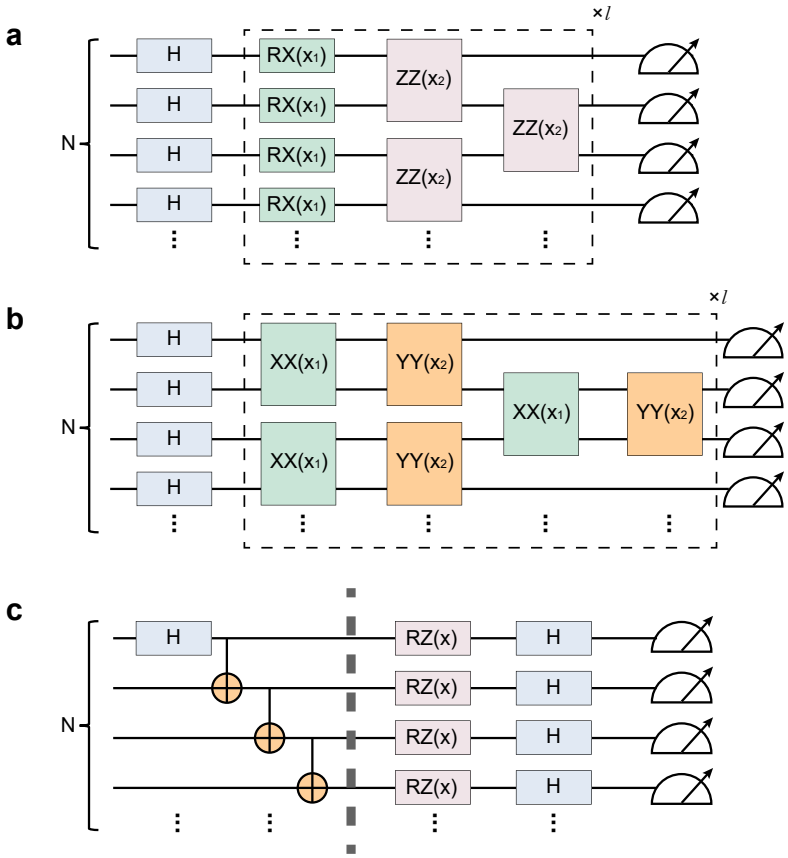

FIG. D.1. **Ansatz circuits used in numerical experiments. a.** Trotterized Hamiltonian variational ansatz for the 1D Transverse Field Ising Model (TFIM). **b.** Trotterized Hamiltonian variational ansatz for the 1D Heisenberg Model (HM). **c.** Circuit implementation for GHZ-state-based quantum metrology.

For the ground state energy estimation tasks, we employ Hamiltonian Variational Ansatz (HVAs) derived from a first-order Trotter-Suzuki decomposition [29, 30] of the respective problem Hamiltonians, $\mathrm{H}_{\mathrm{Ising}}$ and $\mathrm{H}_{\mathrm{Heisen}}$. For both Hamiltonians, the ansatz is applied to the uniform superposition state, prepared by applying an initial layer of Hadamard gates $H^{\otimes N}$ to the all-zero state $|0\rangle^{\otimes N}$.

The circuit structure for the 1D TFIM is illustrated in Fig. D.1a. The variational ansatz $U(\boldsymbol{x})$ is defined by the unitary:

$$U(\boldsymbol{x}) = \Big[\exp\Big(-\imath \boldsymbol{x}_1 \sum_{\langle i,j\rangle \in E} Z_i Z_j\Big) \exp\Big(-\imath \boldsymbol{x}_2 \sum_i X_i\Big)\Big]^l H^{\otimes N}. \tag{D1}$$

This represents the Trotterized evolution under the Ising Hamiltonian terms. In all simulations, we set the number of layers $l = 1$, as this shallow ansatz structure proves sufficient for approximating the ground state energy with high fidelity for the models under study.

The circuit implementation for the 1D Heisenberg model with $J_z = 0$ (the XY model) is visualized in Fig. D.1b. Mathematically, the corresponding unitary operator is given by:

$$U(\boldsymbol{x}) = \Big[\exp\Big(-\imath \boldsymbol{x}_1 \sum_{\langle i,j\rangle\in E} X_i X_j\Big)\exp\Big(-\imath \boldsymbol{x}_2 \sum_{\langle i,j\rangle\in E} Y_i Y_j\Big)\Big]^l H^{\otimes N}. \tag{D2}$$

Analogous to the TFIM case, we fix the layer depth to $l = 1$ for all simulations.

The circuit implementation for the quantum metrology task is illustrated in Fig. D.1c. This protocol is designed to achieve Heisenberg-limited sensitivity. The circuit first prepares an N-qubit GHZ state, $|\mathrm{GHZ}\rangle_N = (|0\rangle^{\otimes N} + |1\rangle^{\otimes N})/\sqrt{2}$, using a Hadamard gate on the first qubit followed by a chain of CNOT gates. Subsequently, the unknown phase x is encoded via a global $\mathrm{RZ}(x)^{\otimes N}$ rotation. This operation imparts a collective relative phase between the two components of the GHZ state:

$$|\psi(\boldsymbol{x}_1)\rangle = \frac{1}{\sqrt{2}}\left(e^{-\imath N x/2}|0\rangle^{\otimes N} + e^{\imath N x/2}|1\rangle^{\otimes N}\right), \tag{D3}$$

resulting in a total relative phase of $e^{\imath N x}$. To convert this phase information into a measurable signal, the expectation value of the global observable $O = X^{\otimes N}$ is measured. Operationally, this measurement is implemented by applying a final layer of Hadamard gates ($H^{\otimes N}$) to all qubits, which rotates the measurement basis, followed by a standard measurement of the global observable $Z^{\otimes N}$. In the noiseless limit, the measured expectation value is $\cos(Nx)$, achieving the Heisenberg-limited phase sensitivity.

2. Random feature sampling for classical learning surrogates

In our numerical experiments, the construction of the classical learning surrogates $\{h_{\mathsf{qs}}\}_{j=1}^{u}$ employs a randomized feature selection strategy to balance model expressiveness with computational efficiency. Each surrogate is implemented as a linear model using a subsampled feature map, specifically $h(\boldsymbol{x}, O, \lambda_j) = \langle \boldsymbol{\Phi}_{\Omega(\Lambda)}(\boldsymbol{x}), \mathbf{w}_j\rangle$, where $\boldsymbol{\Phi}_{\Omega(\Lambda)}(\boldsymbol{x})$ is constructed by randomly and uniformly sampling n_f elements from the complete set of trigonometric monomials $\{\Phi_{\boldsymbol{\omega}}(\boldsymbol{x})|\boldsymbol{\omega} \in \mathfrak{C}(\Lambda)\}$ with $\mathfrak{C}(\Lambda) = \{\boldsymbol{\omega} \in \{0, \pm 1\}^d \mid \|\boldsymbol{\omega}\|_0 \leq \Lambda\}$.

In our experimental implementation, we set $n_f = 1000$ across all simulations. For ground-state energy estimation tasks, the frequency truncation parameters were set to $\Lambda = 2$ for the transverse field Ising model and $\Lambda = 4$ for the Heisenberg model, while for quantum metrology applications with GHZ states, we used $\Lambda = 2$. These parameter choices were determined through empirical validation to provide an optimal balance between model capacity and generalization performance for their respective problem domains. The randomized feature sampling approach significantly reduces computational overhead while maintaining the theoretical approximation guarantees of the full trigonometric basis, enabling efficient training of surrogates even for high-dimensional parameter spaces.

3. Robustness of S-ZNE across varied extrapolation functions

While the main text employs linear extrapolation for its simplicity, we further investigate the robustness of the S-ZNE framework by evaluating its compatibility with a diverse set of extrapolation functions. We benchmark S-ZNE against conventional ZNE under the same setting used for ground-state energy estimation in the transverse-field Ising model (TFIM) and Heisenberg model (HM), as detailed in the main text. For S-ZNE, we use the regression-based classical surrogate h_{qs} trained with $n_j = 200$ samples per noise level λ_j; for conventional ZNE, we use the estimate $\hat{f}$ obtained from $M = 1 \times 10^6$ shots per noise level. The only variable in this comparison is the extrapolation function applied to the surrogate data vector $\boldsymbol{z}_S(\boldsymbol{x})$ and the conventional data vector $\boldsymbol{z}_C(\boldsymbol{x})$.

We compare three common extrapolation strategies in quantum error mitigation: (i) *Linear* (first-order least-squares regression), (ii) *Quadratic* (second-order least-squares regression), and (iii) *Richardson* extrapolation. Detailed implementations are provided in SI A 5. To further assess robustness, we perform this comparison under two noise models: globe depolarizing (DP) noise and a composite noise model combining DP with coherent (CO) noise, consistent with the main text.

Results are summarized in Fig. D.2, which shows the mitigation residuals for S-ZNE ($\mathcal{R}_S$) and conventional ZNE ($\mathcal{R}_C$). Across all extrapolation functions and both noise models, S-ZNE achieves mitigation accuracy comparable to that of conventional ZNE. Under the mixed DP+CO model, both methods exhibit degraded performance when using Richardson extrapolation, indicating that the surrogate-based approach faithfully preserves the behavior—and limitations—of the underlying extrapolation function without introducing significant additional bias. This confirms

that the surrogate can effectively replace direct quantum measurement in the ZNE pipeline, inheriting both the advantages and instabilities of the chosen extrapolation method.

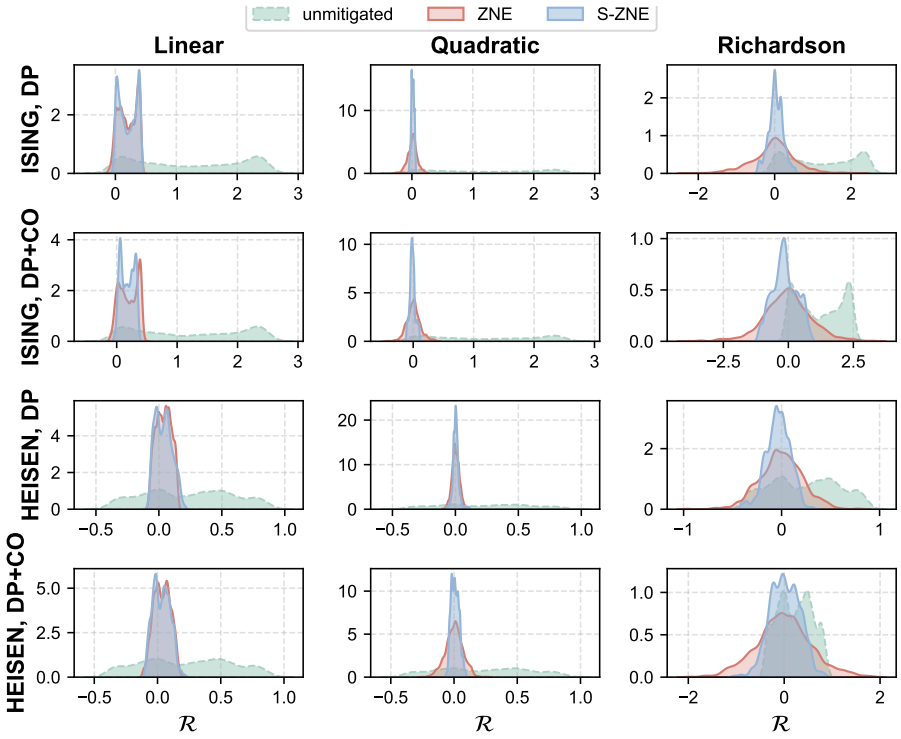

FIG. D.2. **Residual distributions of conventional ZNE and S-ZNE for different extrapolation functions.** Probability density of residuals (error-mitigate estimation minus ideal value) for unmitigated, ZNE, and S-ZNE results under depolarizing (DP) and DP+coherent (CO) noise. Three extrapolation functions are compared: Linear, Quadratic, and Richardson. Results are aggregated over 1000 test instances.

4. Data efficiency in quantum metrology

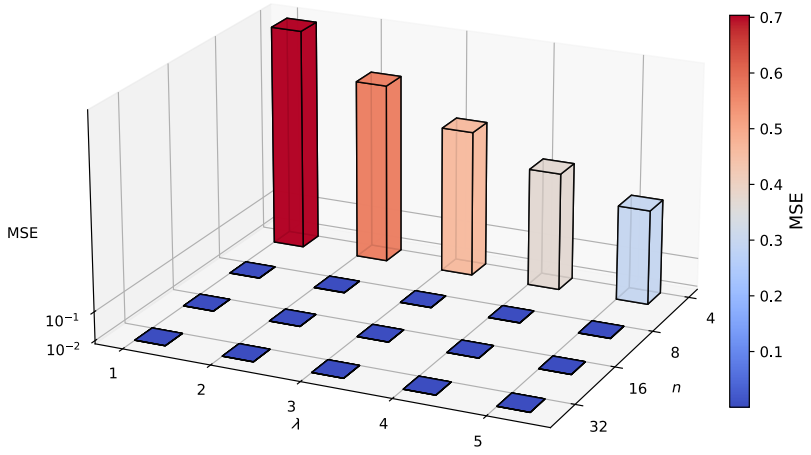

FIG. D.3. **Surrogate prediction accuracy versus training set size for GHZ metrology.** Mean squared error of surrogate predictions as a function of training sample size $n_j \in \{4, 8, 16, 32\}$ and noise factor λ_j. Results represent averages over 10 independent experiments.

To further characterize the data efficiency of S-ZNE in the quantum metrology task, we examine how regression-based surrogate prediction accuracy depends on the number of training samples n_j per noise level λ_j. All hyperparameter settings are identical to those introduced in the main text.

Figure D.3 plots the mean squared error (MSE) of surrogate predictions as a function of training set size $n_j \in \{4, 8, 16, 32\}$ across $u = 5$ noise levels. The results demonstrate a sharp, non-linear improvement in accuracy with n_j. For a minimal training set of $n_j = 4$, the surrogate exhibits a high prediction error (e.g., MSE ≈ 0.704 at the $\lambda_j = 1$ noise level), indicating that this sparse dataset is insufficient to construct the surrogate model and capture the phase-dependent response. However, a modest increase to $n_j = 8$ reduces the MSE by over three orders of magnitude (to $\sim 6.17 \times 10^{-4}$ at $\lambda_j = 1$). The error continues to decrease rapidly as n_j grows: for $n_j = 16$, the MSE drops to 2.20×10^{-4}, and at $n_j = 32$, it reaches 1.60×10^{-4} (all values reported for $\lambda_j = 1$). This rapid decay in error reveals a distinct threshold effect, suggesting that once a minimal number of training samples is provided to constrain the surrogate's trigonometric feature space, the model generalizes effectively from sparse data.

Appendix E: Simulation results of hybrid S-ZNE

The hybrid S-ZNE framework described in SI C addresses a potential limitation of the S-ZNE approach. That is, while surrogates can dramatically reduce quantum measurement costs, their predictive accuracy may be insufficient at low noise levels.

To evaluate the performance of hybrid S-ZNE, we perform numerical simulations using a 6-qubit hardware-efficient ansatz, as illustrated in Fig. E.4a. We consider both the transverse-field Ising model (TFIM) and the Heisenberg model (HM)(defined in maintext), with a fixed circuit depth of $l = 2$ layers. The model parameters are set as follows: for TFIM, we use a uniform coupling $J = -0.1$ and a transverse field $h = -0.5$; for HM, we set $J_x = 0.1$,$J_y = 0.1$ and $J_z = 0.5$. Noise levels are amplified via unitary folding, and extrapolation is performed over $u = 5$ such levels using linear extrapolation. To assess the robustness of the approach, we incorporate three distinct noise models(detailed in SI A 3): local depolarizing (DP) noise, thermal relaxation (TM), and coherent (CO) over-rotation. The corresponding noise parameters are provided in Table E.1.

TABLE E.1. Parameter settings for different noisy channels. Notations follow the definitions in SI A 3.

Noise	**Parameter**	**Value**
Local depolarizing	p_d	single-qubit gate: 0.001, two-qubit gate: 0.005
Thermal	T_1	100000 us
	T_2	30000 μs
	t_g	single-qubit gate:15 μs, two-qubit gate: 20 μs
	p_e	0.01
Coherent	θ_P	Unif$[-0.01\pi, 0.02\pi]$

We trained the kernel-based surrogates h_{cs}, defined in Eq. (A8), with a frequency truncation threshold of $\Lambda = 2$. These surrogates were trained on classical shadow collected at $u = 5$ distinct noise levels with $T = 500$. The validation set $\mathcal{X}_{\text{val}}$, which was used to determine the hybrid data vector $\boldsymbol{z}_H(\boldsymbol{x})$, contained 500 random input points; the ground-truth expectation values for these points were estimated using 40,000 measurement shots each.

To evaluate the data efficiency of the surrogate, we varied the per-noise-level training set size over $n_j \in \{1200, 1400, \ldots, 3000\}$. Figure E.4b shows that the surrogate's mean squared error (MSE) consistently decreases as n_j increases. Notably, this improvement is more substantial at higher noise levels λ_j. This trend holds across all tested noise models, including the composite DP+TM+CO model (Fig. E.4c). For instance, in the transverse-field Ising model (TFIM) simulation, the MSE drops from approximately 0.23 at $\lambda_j = 1$ to about 0.03 at $\lambda_j = 5$. A similar reduction is observed for the Heisenberg model (HM), where the MSE falls from about 0.13 to roughly 0.02 over the same range of noise levels.

Based on this finding, we define a substitution threshold $\eta = 0.1$ and observe from our validation that the surrogate MSE consistently falls below this threshold only for $\lambda_j \geq 3$. We therefore construct a hybrid data vector $\boldsymbol{z}_H(\boldsymbol{x})$ by retaining direct quantum measurements (using $M = 40,000$ shots) for the two lowest noise levels ($\lambda_j = 1, 2$) and using surrogate predictions $h(\boldsymbol{x}, O, \lambda_j)$ for the three highest ($\lambda_j = 3, 4, 5$). This hybrid protocol reduces the per-instance quantum measurement cost by 60% compared to conventional ZNE, while retaining the high-fidelity low-noise data crucial for stable extrapolation.

We evaluate the end-to-end mitigation performance on 500 test instances selected to have non-trivial ideal expectation values ($|f(\boldsymbol{x}, O)| > 0.5$). As defined in the main text, we analyze the mitigation residual $\mathcal{R} = g(\boldsymbol{z}) - f(\boldsymbol{x}, O)$.

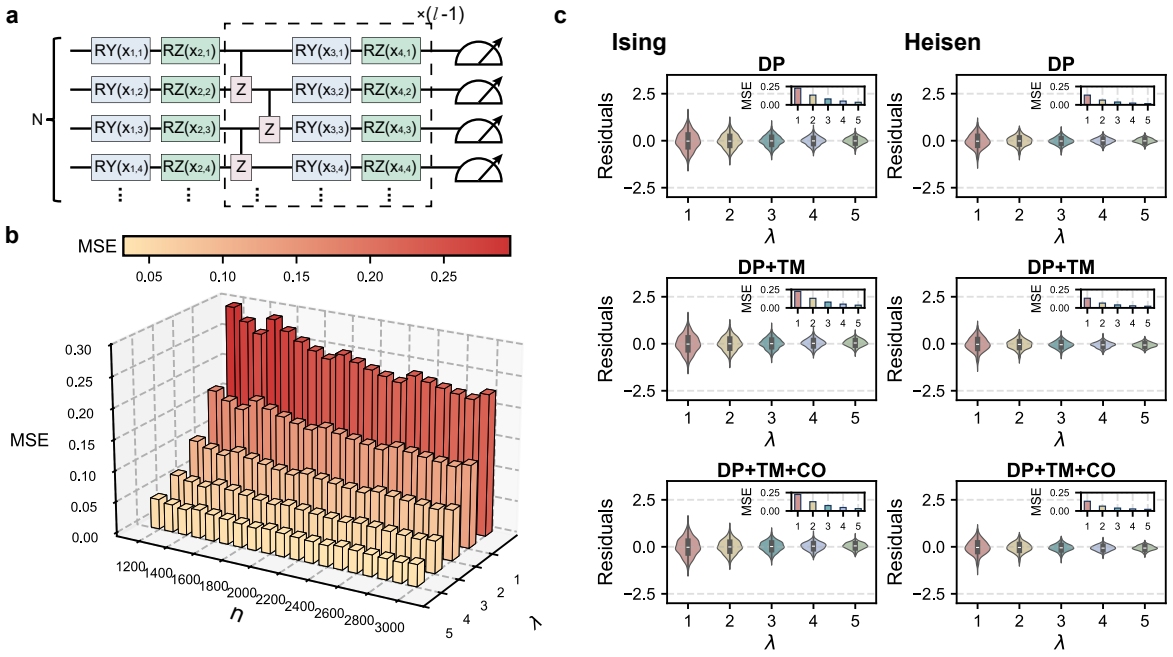

FIG. E.4. **Circuit architecture and surrogate fidelity across noise models and Hamiltonians. a.** Hardware-efficient ansatz with uncorrelated RY and RZ rotations used in simulations. **b.** Mean squared error (MSE) of surrogate predictions versus training set size n_j and noise scaling factor λ_j, evaluated on both the 1D transverse-field Ising and Heisenberg models. MSE decreases with larger n_j and higher λ_j, indicating improved surrogate accuracy in high-noise regimes. **c.** Residual distributions (prediction minus measurement) for three composite noise models—depolarizing (DP), DP+thermal (TM), and DP+TM+coherent (CO)—shown as violin plots; nested bar charts report corresponding MSE values for both Hamiltonians.

Fig. E.5a shows the residual distributions for Hybrid S-ZNE ($\mathcal{R}_H$, using $\boldsymbol{z}_H$) and conventional ZNE ($\mathcal{R}_C$, using $\boldsymbol{z}_C$) under the three noise configurations. Both methods produce residuals tightly centered at zero. The corresponding MSE values reported in Fig. E.5b confirm this quantitatively: Hybrid S-ZNE matches the mitigation accuracy of conventional ZNE, despite foregoing quantum measurements at the three highest (and most costly) noise levels.

Finally, Fig. E.5c highlights the resource trade-off. Conventional ZNE requires $u \times M = 5 \times 40,000 = 200,000$ shots for each evaluation. For 500 test samples, this totals 10^8 measurements. Hybrid S-ZNE, in contrast, requires a one-time offline training cost of $n \times u \times T = 3000 \times 5 \times 500 = 7.5 \times 10^6$ measurements (using $n_j = 3000$ for this example). The per-instance extrapolation cost is reduced to 40% of conventional ZNE (retaining $\lambda = 1, 2$), totaling 4×10^7 measurements for the 500 samples. The total hybrid S-ZNE cost is thus $7.5 \times 10^6 + 4 \times 10^7 = 4.75 \times 10^7$ measurements, a saving of over 50%. In this specific task, the one-time training cost accounts for $\approx 16\%$ of the total cost. For applications requiring many repeated evaluations ($N_{\text{eval}} \gg n \times T/((u-v)M)$), this training cost is amortized, and the cumulative savings asymptotically approach the 60% per-instance reduction. These results demonstrate that the hybrid S-ZNE framework successfully balances mitigation fidelity with practical resource efficiency.

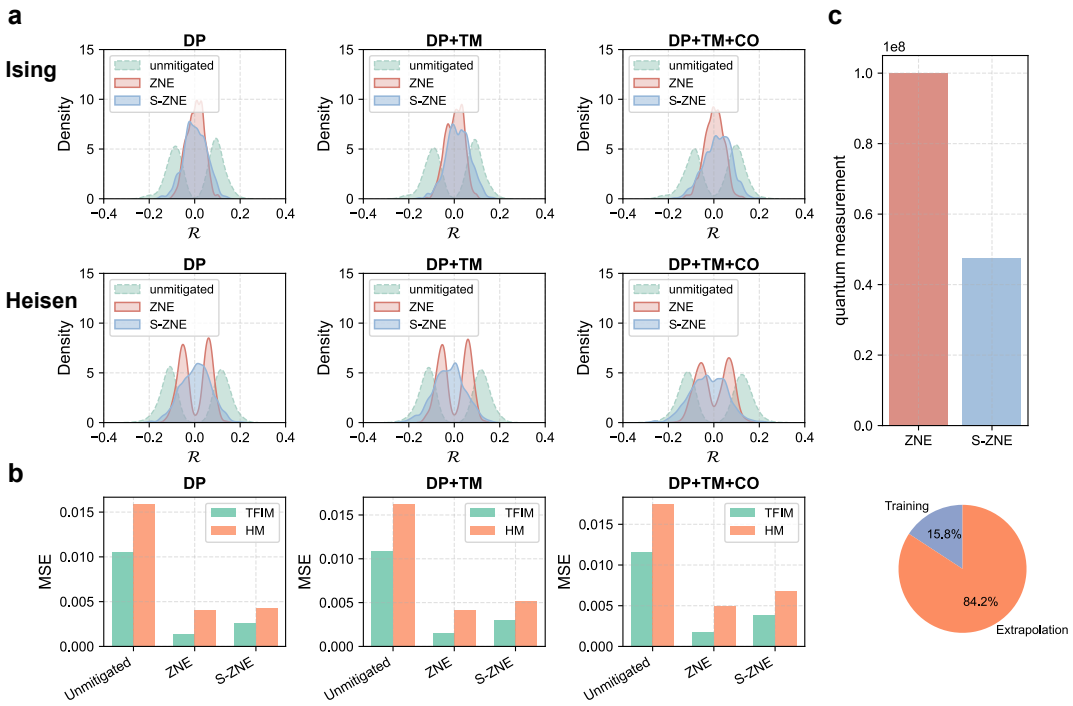

FIG. E.5. **Error mitigation performance and quantum resource efficiency for Ising and Heisenberg models. a.** Residual distributions (mitigated estimate minus ideal value) for unmitigated, ZNE, and hybrid S-ZNE results across three noise configurations (DP, DP+TM, and DP+TM+CO), based on 500 test instances with $|f(\boldsymbol{x}, O)| > 0.5$. **b.** Corresponding MSE values confirm that hybrid S-ZNE matches conventional ZNE in accuracy for both Hamiltonians. **c.** Quantum resource comparison: hybrid S-ZNE incurs a fixed offline training cost and avoids repeated measurements at high λ_j, reducing per-instance quantum overhead by nearly 60% compared to conventional ZNE.

[1] D. Greenbaum, Introduction to quantum gate set tomography, arXiv:1509.02921 (2015).
[2] W.-Y. Liao, Y. Du, X. Wang, T.-C. Tian, Y. Luo, B. Du, D. Tao, and H.-L. Huang, Demonstration of Efficient Predictive Surrogates for Large-scale Quantum Processors, arXiv:2507.17470 (2025).
[3] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, Nat. Phys. **16**, 1050 (2020).
[4] H. Liao, D. S. Wang, I. Sitdikov, C. Salcedo, A. Seif, and Z. K. Minev, Machine learning for practical quantum error mitigation, Nat. Mach. Intell. **6**, 1478 (2024).
[5] L. Bittel and M. Kliesch, Training variational quantum algorithms is np-hard, Phys. Rev. Lett. **127**, 120502 (2021).
[6] A. Abbas, R. King, H.-Y. Huang, W. J. Huggins, R. Movassagh, D. Gilboa, and J. McClean, On quantum backpropagation, information reuse, and cheating measurement collapse, in *Advances in Neural Information Processing Systems*, Vol. 36 (2023) pp. 44792–44819.
[7] K. Zhang, L. Liu, M.-H. Hsieh, and D. Tao, Escaping from the barren plateau via gaussian initializations in deep variational quantum circuits, in *Advances in Neural Information Processing Systems*, Vol. 35 (2022) pp. 18612–18627.
[8] C.-Y. Park and N. Killoran, Hamiltonian variational ansatz without barren plateaus, Quantum **8**, 1239 (2024).
[9] Y. Wang, B. Qi, C. Ferrie, and D. Dong, Trainability enhancement of parameterized quantum circuits via reduced-domain parameter initialization, Phys. Rev. Appl. **22**, 054005 (2024).
[10] C.-Y. Park, M. Kang, and J. Huh, Hardware-efficient ansatz without barren plateaus in any depth, arXiv:2403.04844 (2024).
[11] X. Shi and Y. Shang, Avoiding barren plateaus via gaussian mixture model, New J. Phys **27**, 104501 (2025).
[12] R. Puig, M. Drudis, S. Thanasilp, and Z. Holmes, Variational quantum simulation: a case study for understanding warm starts, PRX Quantum **6**, 010317 (2025).
[13] H. Mhiri, R. Puig, S. Lerch, M. S. Rudolph, T. Chotibut, S. Thanasilp, and Z. Holmes, A unifying account of warm start guarantees for patches of quantum landscapes, arXiv:2502.07889 (2025).
[14] A. A. Mele, G. B. Mbeng, G. E. Santoro, M. Collura, and P. Torta, Avoiding barren plateaus via transferability of smooth solutions in a hamiltonian variational ansatz, Phys. Rev. A **106**, L060401 (2022).
[15] H.-Y. Liu, T.-P. Sun, Y.-C. Wu, Y.-J. Han, and G.-P. Guo, Mitigating barren plateaus with transfer-learning-inspired parameter initializations, New J. Phys. **25**, 013039 (2023).
[16] Y. Peng, X. Li, S. Y.-C. Chen, K. Zhang, Z. Liang, Y. Wang, and Y. Du, Titan: A trajectory-informed technique for adaptive parameter freezing in large-scale vqe, arXiv:2509.15193 (2025).
[17] M. L. Goh, M. Larocca, L. Cincio, M. Cerezo, and F. Sauvage, Lie-algebraic classical simulations for quantum computing, Phys. Rev. Res. **7**, 033266 (2025).
[18] J. Dborin, F. Barratt, V. Wimalaweera, L. Wright, and A. G. Green, Matrix product state pre-training for quantum machine learning, Quantum Sci. Technol. **7**, 035014 (2022).
[19] M. S. Rudolph, J. Miller, D. Motlagh, J. Chen, A. Acharya, and A. Perdomo-Ortiz, Synergistic pretraining of parametrized quantum circuits via tensor networks, Nat. Commun. **14**, 8367 (2023).
[20] A. Cervera-Lierta, J. S. Kottmann, and A. Aspuru-Guzik, Meta-variational quantum eigensolver: Learning energy profiles of parameterized hamiltonians for quantum simulation, PRX Quantum **2**, 020329 (2021).
[21] R. Shaffer, L. Kocia, and M. Sarovar, Surrogate-based optimization for variational quantum algorithms, Phys. Rev. A **107**, 032415 (2023).
[22] G. Verdon, M. Broughton, J. R. McClean, K. J. Sung, R. Babbush, Z. Jiang, H. Neven, and M. Mohseni, Learning to learn with quantum neural networks via classical neural networks, arXiv:1907.05415 (2019).
[23] D. Luo, J. Shen, R. Dangovski, and M. Soljacic, Quack: Accelerating gradient-based quantum optimization with koopman operator learning, in *Advances in Neural Information Processing Systems*, Vol. 36 (2023) pp. 25662–25692.
[24] Y. Huang, Y. Hao, J. Zhou, X. Yuan, X. Wang, and Y. Du, Palqo: Physics-informed model for accelerating large-scale quantum optimization, arXiv:2509.20733 (2025).
[25] W. J. Huggins, J. R. McClean, N. C. Rubin, Z. Jiang, N. Wiebe, K. B. Whaley, and R. Babbush, Efficient and noise resilient measurements for quantum chemistry on near-term quantum computers, npj Quantum Inf. **7**, 23 (2021).
[26] B. Reggio, N. Butt, A. Lytle, and P. Draper, Fast partitioning of pauli strings into commuting families for optimal expectation value measurements of dense operators, Phys. Rev. A **110**, 022606 (2024).
[27] P. Gokhale, O. Angiuli, Y. Ding, K. Gui, T. Tomesh, M. Suchara, M. Martonosi, and F. T. Chong, Minimizing State Preparations in Variational Quantum Eigensolver by Partitioning into Commuting Families, arXiv:1907.13623 (2019).
[28] 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, Rev. Mod. Phys. **95**, 045005 (2023).
[29] S. Lloyd, Universal quantum simulators, Science **273**, 1073 (1996).
[30] S. Stanisic, J. L. Bosse, F. M. Gambetta, R. A. Santos, W. Mruczkiewicz, T. E. O'Brien, E. Ostby, and A. Montanaro, Observing ground-state properties of the fermi-hubbard model using a scalable algorithm on a quantum computer, Nat. Commun. **13**, 5743 (2022).