FIXED-PROTOCOL AMORTIZED MPS TOMOGRAPHY WITH CONFORMALIZED PREDICTIVE UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantum state tomography is sample-starved, and the states one prepares live on a narrow, learnable manifold. A $k{=}0$ prior-only control shows that on concentrated families a prior estimate is already near-optimal, so "high fidelity at few measurements" can be family memorization rather than tomography; genuine measurement-efficiency needs a model that conditions on the measurements and demonstrably uses them. Our method is a *fixed-protocol amortized* matrix-product-state (MPS) estimator: one fixed informative protocol mapped directly to MPS cores, trained once with a gauge-invariant fidelity loss; we deliberately do not rest it on a permutation-invariant set encoder (a plain MLP matches it). The learned generative prior over the same cores, inverted by measurement-guided posterior sampling, is what makes the control bite in the first place, and we retain it only as a complementary route (Approach A, deferred to the appendix). The decisive lever is the *measurement design*: motivated by the fact that local reduced density matrices determine a χ-MPS, conditioning on an *informative local* Pauli set rather than random strings turns a modest, memorization-prone estimator into a high-fidelity one (≈ 0.95, up to $+0.59$ over prior-only, decisively passing a shuffled-measurement control). A dropout ensemble, conformally recalibrated, gives $\approx 90\%$-coverage intervals, including for 18 observables never queried by the protocol (checked, not assumed, against classical and fixed-scale baselines). Quality holds as the system grows (fidelity 0.90 at $n{=}10$, gain *growing* in n; 0.88 at $\chi{=}4$), the parameterization is polynomial (native contraction to 20 qubits), and we close the loop on IBM hardware (5 states at 0.97 fidelity to the ideal target from hardware-measured Paulis). We also quantify amortization's actual cost: training pays for itself only after ~ 3755 reconstructed states. Our prior-only and shuffled controls are a necessary integrity check we argue learned-QST work should adopt.

1 INTRODUCTION

Quantum state tomography (QST)—estimating an unknown state from measurement data—is a cornerstone primitive for benchmarking quantum hardware, and its central difficulty is *sample complexity*: full tomography of an n-qubit state needs exponentially many measurements, each of which (state preparation plus readout) is expensive. Two broad families reduce this cost. *Structural* methods restrict the model class—compressed-sensing QST exploits low rank (Gross et al., 2010) and matrix-product-state (MPS) tomography low entanglement (Cramer et al., 2010; Lanyon et al., 2017). *Bayesian* methods (Blume-Kohout, 2010; Granade et al., 2016; Lukens et al., 2020) return uncertainty but rely on hand-chosen priors on the dense state. Neither family uses the key empirical fact that the states prepared in a given experiment are not arbitrary: they lie on a *narrow, structured manifold* (ground states of a Hamiltonian family, or the outputs of a parameterized circuit). App. A positions this work against the nearest neighbors along each axis.

A learned prior alone is not enough—and a control that shows it. The tempting recipe is to learn a generative *prior* over the family and invert by posterior sampling (Chung et al., 2023). We build this (Approach A: a flow-matching prior over MPS cores with a measurement-guided Doob-h-transform sampler) and it produces high fidelity at few measurements—but a simple control deflates the interpretation. On a concentrated family (disordered paramagnetic ground states, all close to $|+\rangle^{\otimes n}$) a *prior-only* estimate using *zero* measurements already reaches fidelity 0.997, and *shuffling*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the measurements—reconstructing from another state's data—changes nothing: "measurement-efficient tomography" is then nearest-neighbor prediction from a narrow distribution. The lesson is methodological: a learned-QST method must be judged against a $k{=}0$ prior-only baseline and must *use* the measurements, verifiable by a shuffled-measurement test. We adopt both as integrity controls throughout.

Conditioning on measurements, with the right measurement design. To genuinely use the data, we replace the unconditional prior+inference pipeline with an *amortized, measurement-conditioned estimator*: a fixed-protocol encoder (a plain MLP or a permutation-invariant set-encoder) reads the measured pairs $\{(P_j, o_j)\}$ and outputs the cores of an MPS, trained once on the family with a gauge-invariant state-fidelity loss (Fig. 1). Conditioning is necessary but not sufficient: with *random* Pauli strings the model still only modestly beats prior-only, because random high-weight Paulis are uninformative about low-entanglement states. The decisive lever is the *measurement design*: a classical result—local reduced density matrices determine a χ-MPS (Cramer et al., 2010)—says the informative observables are *local* (Prop. 1). Conditioning on a cheap local set (weight-1 and nearest-neighbor weight-2 Paulis), or on the theory-complete 3-site-window design, turns a modest, memorization-prone estimator into a high-fidelity one that beats prior-only and passes the shuffled control; random Paulis do not.

Why cores. The MPS-core parameterization is load-bearing for scalability and for the generative-prior route: a generative model over the dense 2^n state vector underfits already at $n = 6$, whereas the conditional model over $O(n\chi^2)$ cores learns the family and reconstructs it; and the transfer-matrix contraction is polynomial, so the pipeline runs to 20 qubits without ever forming the state vector. Cores carry a gauge; we fix a deterministic coordinatization (left-canonical QR) but do not oversell it—it is a coordinate choice, not a state-level invariant (Prop. 6)—and train with a gauge-invariant target.

Contributions.

- **A fixed-protocol amortized MPS estimator**: one fixed protocol $\rightarrow$ tensor-network cores, trained once with a gauge-invariant loss, dropout ensemble recalibrated to distribution-free coverage. To our knowledge the first amortized measurement-conditioned *full-state* MPS estimator for QST (vs. per-state fits (Torlai et al., 2023), property point-estimators (Cha et al., 2025), or uncalibrated generative reconstructions (Ahmed et al., 2021)). A *plain MLP* matches a set-encoder, so the protocol, core target and loss—not the architecture—do the work; we do not call the ensemble a Bayesian posterior. Predicts 18 never-measured observables at MAE $0.043/0.053$ vs. $0.133/0.267$ prior-only, at $\approx 90\%$ coverage.
- **The informative local design is the decisive lever**, backed by a constructive $O(n)$ spanning-design proposition (Prop. 1): weight-1/nearest-neighbor weight-2 Paulis vs. random swing fidelity $0.79 \rightarrow 0.954$ and the prior-only gain $+0.12 \rightarrow +0.27$ ($+0.59$ shallow). The remaining propositions (coordinate-dependence, degrees-of-freedom separation, conformal coverage) are scaffolding, proved in App. D.
- **Integrity controls, and an honest accounting against classical reconstruction.** A $k{=}0$ prior-only baseline and a shuffled-measurement test—which a learned prior can pass without doing tomography—are both passed by our estimator. Against prior-free MPS-MLE on the *same* local values the design ladder reads $0.633/0.953/0.989$ (classical) vs. $0.666/0.967/0.957$ (ours): we claim no fidelity margin, only that amortization trades a per-state nonconvex fit for a forward pass plus calibrated uncertainty. The same controls delimit a complementary generative route (Approach A: a flow-matching prior over canonicalized cores with h-transform-guided posterior inference, validated against a gold-standard importance sampler—to our knowledge the first learned generative distribution over tensor-network cores), whose few-measurement accuracy leans on the prior (App. G).
- **Scalability and hardware**: fidelity 0.90 at $n{=}10$ (gain *growing* in n), χ up to 4, native $O(n\chi^3)$ contraction to 20 qubits, and a closed-loop reconstruction of 5 held-out states on IBM hardware at fidelity 0.971 ± 0.006.

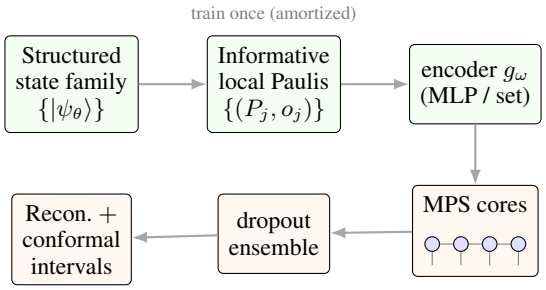

Figure 1: Fixed-protocol amortized tomography. A fixed-protocol encoder g_ω (a plain MLP or a permutation-invariant set encoder—the architecture is not the source of performance) maps the measured values $\{(P_j, o_j)\}$—at a fixed informative local Pauli design—directly to matrix-product-state cores, trained once on a state family with a gauge-invariant fidelity loss. A dropout predictive ensemble, recalibrated conformally, yields a reconstruction and calibrated observable intervals.

2 Background

Setup. Let $|\psi\rangle \in \mathbb{C}^{2^n}$ be an n-qubit pure state. We observe noisy estimates of Pauli expectation values $o_j = \langle\psi|P_j|\psi\rangle + \varepsilon_j$ for a set of Pauli strings $\{P_j\}_{j=1}^k$, with shot noise ε_j of variance $(1 - \langle P_j\rangle^2)/S$ for S shots per measurement setting; we keep S=200 fixed throughout. Three quantities are kept apart throughout: k, the number of Pauli *observables* supplied to the estimator; $m = c(\mathcal{D})$, the number of qubit-wise-commuting *measurement settings* (circuits) that produce them; and the *copy budget* mS. They differ—the default local design at n=6 has k=63 observables that a greedy grouping covers with m=9 settings, so the copy budget is $9S$, not $63S$. We sweep k at fixed S, and simulate each observable's error independently with the variance above: this matches the per-observable variance of running the m settings at S shots each but idealizes away the correlations among Paulis sharing a setting's shots (App. E), which the hardware results do carry. The goal is to estimate $|\psi\rangle$ and conformalized observable intervals, ideally with $k \ll 4^n$.

MPS and gauge. An MPS with bond dimension χ represents amplitudes as $\psi(s_1 \dots s_n) = A^{(1)}_{s_1} A^{(2)}_{s_2} \cdots A^{(n)}_{s_n}$, with cores $A^{(\ell)} \in \mathbb{C}^{\chi_{\ell-1}\times 2\times \chi_\ell}$ ($\chi_0 = \chi_n = 1$). The representation is invariant under bond-wise gauge transformations $A^{(\ell)} \to A^{(\ell)} G_\ell,\ A^{(\ell+1)} \to G_\ell^{-1} A^{(\ell+1)}$. A generative model over cores must therefore act on a *canonical form*; we use the left-canonical gauge (successive QR with a sign convention on the triangular factors), a *norm normalization* of the boundary core, and a global-phase fix. This is the tensor-network analogue of removing permutation/sign/scaling ambiguity in matrix/CP factorizations. The normalization is not cosmetic—without it the standardized core coordinates span orders of magnitude and defeat the flow prior (App. G).

Structured state family. A running physical example is generated by a shallow parameterized circuit (Fig. 4, App. E): a layer of single-qubit rotations $R_y(\theta)$, a nearest-neighbour entangling ladder, and a second rotation layer, with random angles—yielding a low-dimensional, low-entanglement manifold of states represented exactly at bond dimension χ=2. This stands in for physically-relevant families (e.g. Hamiltonian ground states), and its angles set the family's diversity.

3 Method

Canonicalized core coordinates. We map each state $|\psi\rangle$ of the family to a canonical core vector as follows. Sweeping $\ell = 1, \dots, n-1$, reshape the ℓ-th core to $A^{(\ell)} \in \mathbb{C}^{(\chi_{\ell-1}2)\times\chi_\ell}$ and take a (thin) QR factorization

$$A^{(\ell)} = Q^{(\ell)} R^{(\ell)}, \quad S^{(\ell)} = \mathrm{diag}\big(\mathrm{sgn}\, R^{(\ell)}_{ii}\big), \qquad A^{(\ell)} \leftarrow Q^{(\ell)} S^{(\ell)},\ \ A^{(\ell+1)} \leftarrow S^{(\ell)} R^{(\ell)} A^{(\ell+1)}, \tag{1}$$

so every $A^{(\ell)}$ ($\ell < n$) is a left isometry, $A^{(\ell)\dagger} A^{(\ell)} = I$, and the diagonal sign fix removes the residual bond phase. The Householder QR routine we use returns $R^{(\ell)}_{ii}$ *real* even for complex $A^{(\ell)}$ (a standard LAPACK/cuSOLVER convention, verified numerically to machine precision), so sgn is

the ordinary real sign, $S^{(\ell)} \in \{\pm 1\}$ diagonal, and $S^{(\ell)} = S^{(\ell)\dagger} = S^{(\ell)-1}$: applying $S^{(\ell)}$ on both sides is exactly the S/S^{-1} pairing a general complex phase would require, not an inconsistency. We then normalize the boundary core and fix the global phase,

$$A^{(n)} \leftarrow A^{(n)}/\|A^{(n)}\|_F, \qquad A^{(n)} \leftarrow e^{-i \arg \psi_0(\boldsymbol{\theta})} A^{(n)}, \tag{2}$$

where ψ_0 is the first amplitude; because the left cores are isometries, $\|\psi(\boldsymbol{\theta})\| = \|A^{(n)}\|_F$, so equation 2 yields a unit-norm state and keeps all coordinates $O(1)$ (Sec. 5). Stacking the real and imaginary parts of the canonical cores and standardizing gives $\boldsymbol{\theta} \in \mathbb{R}^{d_\theta}$, $d_\theta = O(n\chi^2)$.

The prior-only route, and why we condition instead. On this core parameterization one can learn an unconditional generative *prior* and invert it by measurement-guided posterior sampling (*Approach A*): its draws agree with a gold-standard sampler on the marginal statistics we check, but the integrity controls reveal that at few measurements its point accuracy leans heavily on the prior. That is the motivating counterexample for this paper's method, *Approach B*, which conditions on the measurements directly. Both share the MPS-core parameterization, the gauge treatment, the native contraction and the conformal calibration, and differ only in whether measurements enter through a likelihood at inference (A) or through the network's input (B); A is summarized next and developed in App. G.

Approach A: generative prior and Bayesian posterior inference. We train a rectified-flow prior over canonicalized cores and sample the measurement-conditioned posterior with an h-transform-style guided sampler, validated against a gold-standard self-normalized importance sampler. Because the integrity controls show its few-measurement point accuracy is largely inherited from the prior, we keep it as a complementary route and give its construction, guidance scheme and results in App. G.

Approach B: amortized, measurement-conditioned estimator. Rather than learn an unconditional prior and invert it, we learn the map from measurements to state directly. Fix a measurement design—an ordered set of k Pauli observables $\mathcal{D} = \{P_j\}$ (an *informative local* set, we argue below). Given noisy expectations $o_j = \langle P_j \rangle_{\psi_\star} + \varepsilon_j$, we learn a direct amortized *estimator* $g_\omega : \{(P_j, o_j)\}_{j=1}^k \mapsto \boldsymbol{\theta}$ of the state's canonical cores (a point reconstruction, calibrated post hoc in §5), rather than sampling a posterior. For a *fixed* protocol the encoder is not what matters (a plain MLP on the value vector matches a set-encoder; §5); by default we use a permutation-invariant set-encoder so the same code accepts variable designs: each pair $h_j = \mathrm{MLP}([\mathrm{onehot}(P_j); o_j])$, pooled $z = \frac{1}{k}\sum_j h_j$, decoded $\boldsymbol{\theta} = \mathrm{MLP}(z)$ (App. E).

Gauge-invariant training. We train g_ω on (state, measurement) pairs from the family, minimizing a *state-space* loss

$$\mathcal{L}(\omega) = \mathbb{E}_{|\psi\rangle \sim \text{family},\, \boldsymbol{o}||\psi\rangle} \Big[1 - \big|\langle \psi \,|\, \psi(g_\omega(\boldsymbol{o}))\rangle\big|^2 \Big], \tag{3}$$

one minus reconstruction fidelity, deliberately *gauge-invariant*: the MPS gauge and global phase leave $|\psi(\boldsymbol{\theta})\rangle$ unchanged, so equation 3 never penalizes a different core representative of the correct state. A core-space regression loss $\|g_\omega(\boldsymbol{o}) - \boldsymbol{\theta}_\star\|^2$ is *not*: the canonicalization is a *discontinuous* coordinatization where Euclidean core distance does not track state fidelity (Prop. 6)—nearby states can have far-apart cores near spectral degeneracies—so a core-MSE target is geometrically mismatched and collapses empirically (Sec. 5), while the fidelity loss does not. The canonicalization equation 1–equation 2 only defines the coordinate the network predicts; the loss and readout stay gauge-invariant.

The measurement design is the lever: informative local Paulis. Conditioning is necessary, but the design $\mathcal{D}$ decides whether it helps. Random Pauli strings are a poor design for low-entanglement states: a random weight-w Pauli has expectation exponentially small in w, so it carries little information and the conditional model barely beats prior-only. The informative observables are *local*: reduced density matrices on contiguous windows of $\ell = \max(2\ell_0{-}1,\, 2\lceil \log_2 \chi \rceil{+}1)$ sites determine a normal χ-MPS (Cramer et al., 2010) (Prop. 1), and for our generic injective $\chi{=}2$ family ($\ell_0{=}2$, the minimal value with $2^{\ell_0} \geq \chi^2$) contiguous 3-site windows suffice. This theory-complete design (`local3`) is one option we test. Our default $\mathcal{D}$ is a cheaper subset—all weight-1 Paulis

$\{X_i, Y_i, Z_i\}$ and weight-2 nearest-neighbor Paulis $\{P_iP_{i+1}\}$ ($O(n)$)—which is *not* proven complete but is empirically sufficient on our families (Sec. 5); the theory-complete window design is comparable (Table 2). Either way, the point is that *locality* is the lever: switching from random to local Paulis is what turns a modest estimator into a high-fidelity one.

Predictive uncertainty and conformal calibration. The estimator g_ω is a point map; to attach uncertainty we use it as a *dropout ensemble*: with dropout active at test time, E stochastic passes give predictions $\{|\psi^{(e)}\rangle\}$ and, for each observable O, an ensemble mean μ_O and scale σ_O. We stress this is a heuristic predictive spread, not a validated Bayesian posterior. The intervals $[\mu_O \pm \sigma_O]$ are over-confident, so we *recalibrate conformally*: on a held-out split we form scores $r = |O - \mu_O|/\sigma_O$ over calibration *examples* (a state together with its shot-noisy record, since μ_O, σ_O depend on both) and take the $\lceil (1-\alpha)(N_{\text{cal}}+1) \rceil$-th smallest, Q, giving $\mathcal{I}_O = [\mu_O \pm Q\sigma_O]$ with distribution-free *marginal* coverage $\geq 1-\alpha$ for that fixed observable (Prop. 2). We calibrate a *separate* Q_O per observable, so the proposition applies to the intervals as reported; the guarantee is still marginal coverage, not simultaneous coverage over observables, a full-state confidence region, or coverage under distribution shift. The same recalibration could be applied to any point predictor, including the prior-only baseline; we report it for the conditional estimator because it is the one with informative intervals.

4 THEORETICAL ANALYSIS

Two results carry the design choices, and we state them here: local reduced density matrices give a *constructive* $O(n)$ informative measurement design (Prop. 1), which is the design our conditional model uses, and the recalibrated observable intervals carry a distribution-free finite-sample coverage guarantee (Prop. 2). The supporting formalization—gauge-invariance, coordinate-dependence of core-space priors, and the degrees-of-freedom separation—is deferred to App. H; all proofs are in App. D.

Write $\Phi : \Theta \to \mathbb{C}^{2^n}$ for the MPS contraction $\boldsymbol{\theta} \mapsto |\psi(\boldsymbol{\theta})\rangle$, and let $\mathcal{M} = \{[\Phi(\boldsymbol{\theta})] : \boldsymbol{\theta} \in \Theta\}$ be the induced family of physical states (rays). The MPS gauge group acts on cores by $A^{(\ell)} \mapsto A^{(\ell)}G_\ell$, $A^{(\ell+1)} \mapsto G_\ell^{-1}A^{(\ell+1)}$, leaving Φ invariant. The Pauli-measurement map is $m([\psi]) = (\langle\psi|P_j|\psi\rangle)_{j=1}^k \in \mathbb{R}^k$.

Proposition 1 (A constructive spanning design from local reduced density matrices)**.** *Consider a site-dependent (not necessarily translation-invariant)* normal *(injective) open-boundary χ-MPS with injectivity length ℓ_0—the least block length at which the map from $\chi \times \chi$ bond matrices to that block's physical Hilbert space is injective, so necessarily $2^{\ell_0} \geq \chi^2$ (minimal generic value $\ell_0 = \lceil 2\log_2 \chi \rceil$; for χ=2, ℓ_0=2). Cramer et al. (2010, Thm. 2 and its open-boundary reconstruction) show that reduced density matrices on contiguous windows of some block length determine such a state* uniquely*, generically among all states, not merely among MPS; we do not reproduce their exact window-length constant here, and instead give our own dimension-counting estimate, $\ell = \max(2\ell_0 - 1,\ 2\lceil \log_2 \chi \rceil + 1)$ (for χ=2: ℓ=3, since $2\ell_0 - 1 = 3 = 2\lceil \log_2 \chi \rceil + 1$; for χ=4: ℓ_0=4, giving ℓ=7)—a plausible sufficient window length motivated by their construction, not a claimed reproduction of their stated bound. Each such reduced density matrix is a linear image of the expectations of the weight-$\leq \ell$ local Pauli strings supported on that window, giving an explicit,* non-random *set of $O(n\,4^\ell) = O(n)$ Pauli observables. Cramer et al.'s theorem establishes that this set determines the state* uniquely *(injectivity as a map of sets); we take the stronger* differential *statement—that this set spans $T_{[\psi]}\mathcal{M}$, i.e. is a spanning design in the sense of Assumption 1—as a genericity assumption motivated by, but not derived from, this uniqueness result (App. D details the gap and why it is generic rather than pointwise). Under that assumption, identifiability follows from $k = O(n)$ Pauli observables—versus $2^{n+1} - 2$ unstructured. We use the uniqueness result as* motivation *for the design; it is a uniqueness statement, not a noise-robust sample-complexity bound, and the spanning property is assumed, not proven, at any particular state.*

Proposition 2 (Distribution-free coverage of recalibrated conformal sets)**.** *Let calibration states $\{|\psi_i\rangle\}_{i\in\mathcal{C}}$ and a test state $|\psi_\star\rangle$ be exchangeable draws from the family. Fix an observable O and let the calibration and test* examples *$Z_i = (|\psi_i\rangle, \boldsymbol{o}_i)$—each a state together with its shot-noisy measurement record—be exchangeable. Let $\mu(\boldsymbol{o}), \sigma(\boldsymbol{o})$ be the predictive mean and scale the estimator computes* from the record*, and $r(Z) = |O(\psi) - \mu(\boldsymbol{o})|/\sigma(\boldsymbol{o})$ the nonconformity score; let Q be the $\lceil (1-\alpha)(|\mathcal{C}|+1) \rceil$-th smallest calibration score. Then $\mathcal{I}(\boldsymbol{o}_\star) = [\mu(\boldsymbol{o}_\star) \pm Q\,\sigma(\boldsymbol{o}_\star)]$ satisfies*

$\Pr\big(O(\psi_\star) \in \mathcal{I}(\boldsymbol{o}_\star)\big) \geq 1-\alpha$*, the probability being over the joint draw of states* and *shot noise, with no assumption on the prior or the sampler.*

Scope of the theory (a formalization). The theory is scaffolding, not a main contribution. Prop. 1 is a uniqueness/spanning statement that tells us *which* Paulis to measure, and Prop. 2 is standard split-conformal (Lei et al., 2018), which is what the calibrated intervals rely on. The supporting formalization—gauge-invariance of the objective (Prop. 5), coordinate-dependence of any core-space prior (Prop. 6), the degrees-of-freedom separation between structured and unstructured tomography (Prop. 7), and a prior-restricted stable-recovery bound for Approach A under an assumed manifold RIP (Prop. 3, with a best-of-N guarantee)—is stated in App. H and App. G.2. What we do *not* prove: how many measurements a design needs as a function of the manifold's covering number, curvature and bond dimension, and quantitative error bounds for the amortized conditional map.

5 Experiments

Setup. We evaluate on two $n{=}6$, $\chi{=}2$ families. The *generic* random-MPS family draws fixed random base cores plus i.i.d. complex-Gaussian perturbations ($\sigma{=}0.35$), canonicalized; it is moderately diverse (nearest-training-neighbour fidelity ≈ 0.82). The *physical* shallow-circuit family is an $R_y(\theta_1)$–CNOT-ladder–$R_y(\theta_2)$ circuit with random angles ($\theta_1 \sim \frac{\pi}{2} + 0.6\mathcal{N},\ \theta_2 \sim 0.6\mathcal{N}$), which is exactly a $\chi{=}2$ MPS and, at this spread, genuinely diverse (medoid fidelity to test states only 0.37). Measurements are noisy Pauli expectations $o_j = \langle P_j\rangle + \varepsilon_j$ with shot noise of variance $(1-\langle P_j\rangle^2)/S$, $S{=}200$. Unless stated we use the informative local design $\mathcal{D}$ (all single-qubit X_i, Y_i, Z_i and nearest-neighbour P_iP_{i+1}; 63 observables at $n{=}6$). The conditional set-encoder is trained once per family (2×10^4/6×10^3 states, 30k AdamW steps); all fidelities are on 120–200 held-out states. Code and hyperparameters are in App. E.

Integrity controls: beat prior-only, and use the measurements. We first run the two controls that, we argue, any learned-QST method should report. *Prior-only ($k{=}0$):* predict each test state by a fixed central estimate that uses no measurements—the family medoid (the bank state maximizing mean fidelity to the family) or the normalized mean bank state. *Shuffled:* reconstruct from a *different* state's measurements (a random permutation of the test batch); if this matches the true-measurement reconstruction, the measurements are unused and the method is doing nearest-neighbour prediction, not tomography. Fig. 5 (App. F) shows why these matter. On a *concentrated* family—disordered paramagnetic TFIM ground states, all close to $|+\rangle^{\otimes n}$—the medoid already reaches fidelity 0.997 with *zero* measurements, and true and shuffled reconstructions coincide to ±0.01 at every budget: there is nothing to reconstruct, and a naive "high fidelity at few measurements" is entirely the prior. On a genuinely *diverse* family the medoid is far from optimal (0.65), true-measurement reconstruction clearly separates from shuffled, and the separation grows with the budget—the signature of real measurement information. We therefore report on diverse families, where passing the controls is meaningful.

Approach A, briefly. The generative-prior route's draws match a gold-standard self-normalized importance sampler on the statistics we check—posterior mean (correlation 0.98 over 25 states) and per-observable 1D marginals, not the full joint posterior—it beats classical few-measurement estimators on structured families, and it degrades gracefully off the training manifold; but the integrity controls temper the reading: on concentrated families a $k{=}0$ prior-only estimate already matches it, so its *point* accuracy at few measurements is largely inherited from the prior. That gap is what motivates conditioning directly on measurements. Full results, including the posterior validation and the out-of-distribution stress test, are in App. G; the rest of this section is Approach B.

The measurement design is the lever. Conditioning on measurements is necessary but not sufficient: with *random* Pauli strings the conditional model beats prior-only only modestly on the random family (0.79 vs. 0.67, gain $+0.12$; true$-$shuffled $+0.19$), because a random weight-w Pauli has expectation exponentially small in w and carries little information about a low-entanglement state. Switching to the theory-motivated *informative local* design (Prop. 1) is decisive (Fig. 2, Table 1): fidelity rises to $0.954{\pm}0.021$ (random) and $0.967{\pm}0.046$ (shallow), the prior-only gain to $+0.27$ and $+0.59$, and the shuffled control collapses the reconstruction to 0.51 and 0.15 (true$-$shuffled up to $+0.81$—feeding another state's local measurements yields a wrong state). The design change,

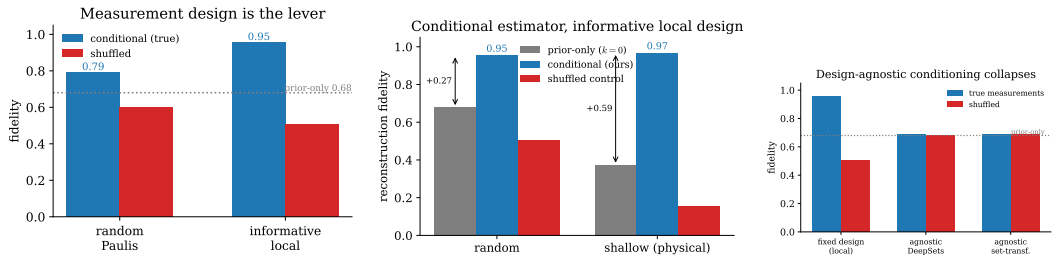

Figure 2: **Left:** the design is the lever: informative local Paulis lift fidelity from 0.79 to 0.95 over random ones (random family). **Middle:** headline result, with the conditional estimator beating $k{=}0$ prior-only (arrows: $+0.27$ random, $+0.59$ shallow) and the shuffled control collapsing. **Right:** a design-*agnostic* estimator (trained over variable random subsets) collapses to the family mean (true $=$ shuffled) for both encoders.

Table 1: Conditional estimator, $n{=}6$, over 120–200 held-out states. The $\pm$ is dispersion *across states* (per-state q_{10}–q_{90}: 0.924–0.974 random, 0.927–0.992 shallow), not seed noise. The shuffled control (reconstruction from another state's measurements) collapses, confirming the measurements are used, and conformal recalibration attains near-nominal 90% coverage with sharp intervals. The last row is the random-Pauli design: switching random→local is what makes conditioning work.

family	design	prior-only	**conditional**	gain	shuffled	true−shuf.	coverage / width
random	local	0.680	**0.954±0.021**	+0.27	0.505	+0.45	89.7% / 0.20
shallow (phys.)	local	0.375	**0.967±0.046**	+0.59	0.154	+0.81	89.6% / 0.23
random	random Paulis	0.680	0.79	+0.11	0.60	+0.19	—

not the encoder or more training, is what turns a modest, memorization-prone estimator into a high-fidelity one.

Calibrated, sharp uncertainty. Beyond a point estimate we return a *predictive ensemble* (not a validated posterior—that is Approach A): dropout at test time ($E{=}24$ passes) gives a mean and scale per local Z/X observable. Raw intervals are over-confident (69–72% empirical coverage at nominal 90%; Fig. 6); split-conformal recalibration (a separate Q_O per observable) restores $89.7 \pm 1.0\%/89.6 \pm 1.7\%$ (random/shallow) over 10 splits (Table 7, App. F; worst single observable 87.1%/88.3%), at sharp clipped widths 0.20/0.23. The prior-only estimate does not by itself supply an informative scale, though the same conformal step could be applied to it.

Calibrated predictions for *unmeasured* observables. The sharper test of predictive value is on observables the protocol never *queries as input*: 18 held-out ones (long-range Z_iZ_j and weight-3 ZZZ/XXX strings, none in `local`), predicted with the same dropout+conformal pipeline. MAE is 0.043/0.053 (random/shallow) vs. 0.133/0.267 prior-only (3–5× better), at 89.7%/90.1% coverage and widths 0.190/0.224. Under our simulated (independent-per-observable) protocol these are genuinely unavailable from shot data (App. F discusses what this does and does not rule out for a real grouped measurement). We also checked, rather than assumed, that this width is not simply standard calibration at work: App. F compares a fixed (non-heteroscedastic) scale on the same point estimate and a classical MLE point estimate with fixed-scale conformal, at matched $\approx 90\%$ coverage—the honest picture is mixed, not uniformly in our favor (on shallow, MLE's interval is sharpest of the three).

Design-agnostic conditioning collapses (an honest limitation). Trained over variable random Pauli subsets and tested on unseen designs, neither a DeepSets encoder nor a 4-layer set-transformer learns to exploit arbitrary designs: both collapse to the family mean, with true $=$ shuffled $= 0.69 \approx$ prior-only (Fig. 2, right). Amortizing over arbitrary designs is open, and this sharpens our scope to a *fixed informative protocol*—the standard setting in practice.

Table 2: Design ladder, encoder ablation, and the classical head-to-head (conditional fidelity, n=6, S=200; shuffled fidelity in parentheses). Design is the dominant lever for *both* methods; the learned prior helps only where the design is under-determined (`local1`), is marginal on `local`, and is unnecessary at `local3`. A plain MLP matches the set-encoder at a fixed design, so permutation-invariance is not what drives performance.

setting	random family	shallow (physical)	MPS-MLE (shallow)
`local1` (single-qubit, 18)	0.910 (0.53)	0.666 (0.19)	0.633±0.220
`local` (default, 63)	**0.954** (0.51)	**0.967** (0.15)	0.953±0.017
`local3` (theory-complete, 207)	0.950 (0.50)	0.957 (0.16)	**0.989±0.004**
encoder: DeepSets (`local`)	0.954	0.967	—
encoder: plain MLP (`local`)	0.967	0.991	—

Ablations: design richness, encoder, diversity, shots. We isolate what drives the result (Table 2). *(i) Design richness.* Single-qubit Paulis alone (`local1`, 18 obs) beat prior-only on the generic family (0.910) but underperform on the physical one (0.666), where nearest-neighbor weight-2 terms matter; the default `local` set (63) reaches 0.954/0.967 and the theory-complete 3-site design (`local3`, 207 obs) is comparable (0.950/0.957)—the cheap 2-local subset is empirically sufficient. The shuffled control collapses throughout. *(ii) Encoder.* For a *fixed* design a plain MLP on the value vector matches or exceeds the DeepSets encoder (0.967/0.991 vs. 0.954/0.967): performance comes from the MPS parameterization, local design and fidelity loss, not permutation-invariance, which is motivated only by the design-*agnostic* goal where both encoders collapse. *(iii) Known-ansatz baseline.* Regressing the shallow family's $2n$ circuit angles directly from the 63 local observables—knowledge our method never assumes—reaches only 0.942 ± 0.089 (0.947 ± 0.098 with our own state-fidelity loss), below our 0.967 ± 0.046 and far more variable, since the angle representation is many-to-one and non-convex. *(iv) Diversity.* Sweeping the shallow family's angle spread (Fig. 7, App. F) behaves as expected: gain over prior-only is small when concentrated, peaks at moderate diversity ($+0.63$), and reconstruction still beats prior-only when diverse though harder. *(v) Shots.* Reducing S from 800 to 50 per setting lowers fidelity gracefully (App. E), as expected from $1/\sqrt{S}$ noise.

Head-to-head with classical local reconstruction. The most informative baseline is the classical one our design is motivated by: a *prior-free* MPS maximum-likelihood fit to the *same* local values, per state with random restarts (Table 2, shallow family). *Design is the dominant lever for both methods*: MLE climbs $0.633 \to 0.953 \to 0.989$ along `local1`/`local`/`local3`, mirroring the learned estimator; and the learned prior's advantage is *confined to the under-determined regime*—it edges out MLE at `local1` (0.666 vs. 0.633), is marginal on `local` ($+0.014$), and is beaten at `local3` (0.957 vs. 0.989), where complete data needs no prior and our fixed-capacity network is the bottleneck. We claim no fidelity margin: amortization buys a forward pass instead of a per-state nonconvex fit, plus calibrated intervals (including for unmeasured observables) without per-state uncertainty fitting. (Table 3's near-zero MLE numbers are for *random* designs: same point.)

What amortization actually costs (App. F). Measured, not asserted: on one A6000, training (headline settings, 30k steps, batch 256) takes 326s; a point reconstruction then costs 0.45 ms/state and the full 24-pass MC-dropout ensemble (needed for the interval) 7.5 ms/state, vs. 94.6 ms/state for classical per-state MPS-MLE (3 restarts $\times$ 500 Adam steps, fidelity 0.957 ± 0.017, matching Table 2). Amortized inference is $\sim 12\times$ faster per state, but training only pays for itself after $N^\star \approx 3755$ states from the same protocol and family; below that the per-state classical fit is cheaper overall. Amortization's value is thus conditional on scale (many repeated preparations), not a one-off benefit.

Family-aware simple baselines. Two further baselines (App. F) isolate what the representation contributes: a measurement-space k-NN retrieval reaches only 0.80–0.90 (not a nearby training state), and a *direct dense-state* MLP outputting the 2^n amplitudes matches the MPS-core estimator at n=6 (0.979/0.958)—reported openly, since the core is not what buys small-n accuracy; its value is scalability, as a dense-state output is infeasible at n=16, 20.

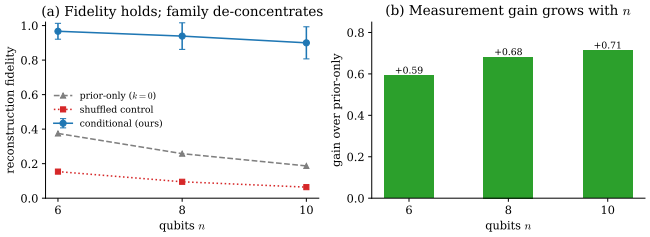

Figure 3: Reconstruction vs. system size (shallow family, local design, 30k steps). **(a)** Fidelity declines gently while prior-only falls steeply and the shuffled control stays low. **(b)** The gain over prior-only *grows* with n.

Higher bond dimension and deeper families. Sweeping the bond dimension of the random-MPS family at n=6 (same `local` design; Table 8), fidelity degrades only gently ($0.954 \to 0.908 \to 0.878$ for χ=2, 3, 4), the gain over prior-only persists ($+0.30$), and the shuffled control still collapses—even though the weight-2 design is *theoretically sub-complete* for χ>2 (Prop. 1 needs ℓ=7-site windows at χ=4): the learned prior compensates. A deeper two-layer brickwork family (χ up to 4) reaches 0.782 ± 0.140 ($+0.56$ over prior-only, shuffled control passed), degrading gracefully with circuit depth rather than collapsing.

Reconstruction quality scales with system size. On the shallow family at n=6, 8, 10 (same 30k-step budget; Fig. 3), fidelity declines gently ($0.967 \to 0.939 \to 0.900$) while prior-only falls steeply ($0.375 \to 0.258 \to 0.187$) and the shuffled control stays low ($0.154 \to 0.095 \to 0.064$): the *measurement gain grows* with n ($+0.59 \to +0.68 \to +0.71$), so performance, not just the implementation, scales. (Dense-statevector training limits this to $n \leq 10$.)

Scalability: native tensor-network contraction to n=20. Every Pauli expectation and overlap is computed by left-to-right transfer-matrix contraction at $O(n\chi^3)$, never materializing the 2^n state vector (matches dense contraction to $\leq 10^{-8}$; App. E). Peak GPU memory is essentially *flat* in n—0.134/0.147 GB at n=16/20 (120/152 core parameters)—while the dense dimension grows from 65,536 to 1,048,576 (Fig. 8, App. F): a feasibility result, not high fidelity—at reduced budget (k=32 random Paulis) fidelity is only 0.063/0.037 (MLE $\approx$ 0), since that needs the local design and budget of the $n \leq 10$ results above, not yet paired with this contraction.

Closed-loop reconstruction on IBM hardware (Table 6, App. F). We close the loop on real hardware with the *same* local design: on `ibm_aachen` (156-qubit Heron) we prepare 5 held-out shallow-circuit states, measure the 63 local Paulis with `EstimatorV2` in one batched job ($|hw - \text{ideal}| = 0.024$, conventions verified to $< 10^{-7}$), and feed those *hardware-measured* values to the estimator: it reconstructs the 5 states at fidelity $\mathbf{0.971 \pm 0.006}$ *to the ideal circuit target*—recovery of the intended preparation from noisy hardware values, not a verified device-error characterization. It is nonetheless genuine tomography, not memorization (prior-only medoid 0.283), and robust to the device's noise (only 0.006 fidelity lost to the 0.024 measurement error).

6 Conclusion

(App. B: when each approach helps; App. C: scope and limitations.) We introduced a fixed-protocol amortized MPS estimator for quantum state tomography: an encoder over measured values of one fixed protocol that outputs matrix-product-state cores, trained once with a gauge-invariant loss and equipped with conformally calibrated intervals. Its decisive design choice is to condition on an *informative local* Pauli set, motivated by local reduced density matrices determining a χ-MPS; this yields fidelity ≈ 0.95, up to $+0.59$ over prior-only, sharp $\approx 90\%$-coverage intervals including for never-measured observables, and a closed loop on IBM hardware, at polynomial cost to 20 qubits. Our method passes the k=0 prior-only and shuffled-measurement integrity controls that most learned-QST evaluations omit; against a prior-free classical MPS-MLE we claim no fidelity margin, since design is the lever for both, and what the learned estimator adds is amortization and calibrated uncertainty.

REPRODUCIBILITY STATEMENT

All state families, measurement designs, encoder architectures, training hyperparameters, and evaluation protocols are specified in App. E; the Approach A prior, its measurement-guided sampler and the gold-standard SIR reference in App. G; and proofs of all propositions in App. D. The native transfer-matrix contraction that makes $n{=}16, 20$ feasible is described in App. E. The hardware results come from a single batched `EstimatorV2` job on `ibm_aachen` under the protocol of App. E, with the measurement convention verified against the noiseless statevector to 8.9×10^{-8} before reconstruction. Reported $\pm$ on reconstruction fidelity is dispersion across the held-out test states (training is seed-deterministic; a three-seed check gives $0.968{\pm}0.001$ and $0.954{\pm}0.001$), and conformal-coverage $\pm$ is over 10 calibration/test splits.

REFERENCES

Shahnawaz Ahmed, Carlos Sanchez Munoz, Franco Nori, and Anton Frisk Kockum. Quantum state tomography with conditional generative adversarial networks. *Physical Review Letters*, 127(14): 140502, 2021.

Robin Blume-Kohout. Optimal, reliable estimation of quantum states. *New Journal of Physics*, 12 (4):043034, 2010.

Juan Carrasquilla, Giacomo Torlai, Roger G Melko, and Leandro Aolita. Reconstructing quantum states with generative models. *Nature Machine Intelligence*, 1(3):155–161, 2019.

Hyunho Cha, Wonjung Kim, and Jungwoo Lee. Scalable Bayesian shadow tomography for quantum property estimation with set transformers. *arXiv preprint arXiv:2509.18674*, 2025.

Peter Cha, Paul Ginsparg, Felix Wu, Juan Carrasquilla, Peter L McMahon, and Eun-Ah Kim. Attention-based quantum tomography. *Machine Learning: Science and Technology*, 3(1):01LT01, 2022.

Matthias Christandl and Renato Renner. Reliable quantum state tomography. *Physical Review Letters*, 109(12):120403, 2012.

Hyungjin Chung, Jeongsol Kim, Michael T McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Marcus Cramer, Martin B Plenio, Steven T Flammia, Rolando Somma, David Gross, Stephen D Bartlett, Olivier Landon-Cardinal, David Poulin, and Yi-Kai Liu. Efficient quantum state tomography. *Nature Communications*, 1(1):149, 2010.

Aldo Cumitini, Luca Barletta, and Osvaldo Simeone. Anytime-valid quantum state tomography via confidence sequences. *IEEE Signal Processing Letters*, 2026. to appear.

Alexander Denker, Francisco Vargas, Shreyas Padhy, Kieran Didi, Simon Mathis, Vincent Dutordoir, Riccardo Barbano, Emile Mathieu, Urszula J Komorowska, and Pietro Lio. Deft: Efficient fine-tuning of diffusion models by learning the generalised h-transform. *Advances in Neural Information Processing Systems*, 37:19636–19682, 2024.

Philippe Faist and Renato Renner. Practical and reliable error bars in quantum tomography. *Physical Review Letters*, 117(1):010404, 2016.

Maciej Falkiewicz, Naoya Takeishi, Imahn Shekhzadeh, Antoine Wehenkel, Arnaud Delaunoy, Gilles Louppe, and Alexandros Kalousis. Calibrating neural simulation-based inference with differentiable coverage probability. *Advances in Neural Information Processing Systems*, 36: 1082–1099, 2023.

Christopher Granade, Joshua Combes, and DG Cory. Practical Bayesian tomography. *New Journal of Physics*, 18(3):033024, 2016.

David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical Review Letters*, 105(15):150401, 2010.

Zhao-Yu Han, Jun Wang, Heng Fan, Lei Wang, and Pan Zhang. Unsupervised generative modeling using matrix product states. *Physical Review X*, 8(3):031012, 2018.

Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.

Ben P Lanyon, Christiane Maier, Milan Holzäpfel, Tillmann Baumgratz, Cornelius Hempel, Petar Jurcevic, Ish Dhand, AS Buyskikh, Andrew J Daley, Marcus Cramer, et al. Efficient tomography of a quantum many-body system. *Nature Physics*, 13(12):1158–1162, 2017.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Xinyu Liu, Jingze Zhuang, Wanda Hou, and Yi-Zhuang You. Measurement-based quantum diffusion models. *arXiv preprint arXiv:2508.08799*, 2025.

Sanjaya Lohani, Joseph M Lukens, Atiyya A Davis, Amirali Khannejad, Sangita Regmi, Daniel E Jones, Ryan T Glasser, Thomas A Searles, and Brian T Kirby. Demonstration of machine-learning-enhanced Bayesian quantum state estimation. *New Journal of Physics*, 25(8):083009, 2023.

Joseph M Lukens, Kody JH Law, Ajay Jasra, and Pavel Lougovski. A practical and efficient approach for Bayesian quantum state estimation. *New Journal of Physics*, 22(6):063038, 2020.

Changhun Oh, Yong Siah Teo, and Hyunseok Jeong. Efficient Bayesian credible-region certification for quantum-state tomography. *Physical Review A*, 100(1):012345, 2019.

George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with Bayesian conditional density estimation. *Advances in Neural Information Processing Systems*, 29, 2016.

Shakir Showkat Sofi, Charlotte Vermeylen, and Lieven De Lathauwer. Tensor train quantum state tomography using compressed sensing. In *2025 33rd European Signal Processing Conference (EUSIPCO)*, pp. 1332–1336. IEEE, 2025.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International conference on learning representations*, 2023.

Yehui Tang, Mabiao Long, and Junchi Yan. QuaDiM: A conditional diffusion model for quantum state property estimation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Neural-network quantum state tomography. *Nature Physics*, 14(5):447–450, 2018.

Giacomo Torlai, Christopher J Wood, Atithi Acharya, Giuseppe Carleo, Juan Carrasquilla, and Leandro Aolita. Quantum process tomography with unsupervised learning and tensor networks. *Nature Communications*, 14(1):2858, 2023.

Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36:31372–31403, 2023.

Yuchen Zhu, Tianrong Chen, Evangelos A Theodorou, Xie Chen, and Molei Tao. Quantum state generation with structure-preserving diffusion model. *arXiv preprint arXiv:2404.06336*, 2024.

A Extended related work

Each pairwise combination of our ingredients is occupied in the literature, but their intersection is not. We position against the nearest neighbors along each axis in turn: structural/MPS tomography, generative and amortized QST, generative priors for inverse problems, and uncertainty in QST.

Structural / MPS tomography. MPS-QST (Cramer et al., 2010; Lanyon et al., 2017) and tensor-train compressed-sensing QST (Sofi et al., 2025) recover a state by *fitting* cores to data under a hand-designed low-rank/sparsity prior. Torlai et al. (2023) learn a tensor-network channel by gradient descent to fit measurement statistics. In all of these the cores are the optimization variables for a *single* unknown state; none learns a transferable generative distribution over cores. Our prior is exactly such a distribution, and it is what makes the few-measurement regime tractable.

Generative / neural QST. Neural QST fits a generative model (RBM, autoregressive, transformer) to the measurements of *one* target state (Torlai et al., 2018; Carrasquilla et al., 2019; Cha et al., 2022); QST-CGAN (Ahmed et al., 2021) amortizes a GAN across similar states but returns a point estimate. Diffusion has been used to *generate* valid density matrices (Zhu et al., 2024), not to reconstruct from measurements. Closest in spirit, QuaDiM (Tang et al., 2025) is a conditional diffusion model in the quantum-measurement space, but it conditions on *Hamiltonian parameters* (not on measurements of an unknown state) and *generates* measurement snapshots for property estimation rather than inverting given measurements to reconstruct a state, and it uses no tensor-network representation. We differ by learning an amortized conditional *estimator* of MPS cores directly from a fixed informative Pauli protocol on an unknown state, with distribution-free calibrated observable intervals.

Amortized / learned-prior Bayesian QST. Closest to our motivation, Lohani et al. (2023) learn (with a shallow CNN) the hyperparameters of a Bayesian prior; and Cha et al. (2025) amortize a set-transformer estimate of the Bayesian posterior *mean* of state *properties* from classical shadows. Both are point estimators (of properties or of prior hyperparameters); neither learns an amortized conditional full-state *estimator* over tensor-network cores nor reports calibrated, distribution-free observable-level intervals for the reconstructed state. Simulation-based inference with normalizing-flow posteriors (Papamakarios & Murray, 2016) and coverage calibration (Falkiewicz et al., 2023) are mature outside physics but, to our knowledge, unapplied to QST.

Generative priors for inverse problems (Approach A). A large body of work solves inverse problems by sampling a data-consistent posterior under a learned generative prior—diffusion/flow posterior sampling and its guidance schemes (Chung et al., 2023; Song et al., 2023; Wu et al., 2023), and their formalization through Doob h-transforms and twisted/guided diffusion (Denker et al., 2024; Wu et al., 2023). These are developed for images and general Bayesian inverse problems; our Approach A instantiates this program in QST by placing a rectified-flow prior on *canonicalized MPS cores* and guiding it with the Pauli shot-likelihood, then validating against a gold-standard SIR posterior. The novel ingredients are the tensor-network core parameterization (which makes the prior learnable) and the integrity controls that reveal when the prior, rather than the measurements, is doing the work.

Uncertainty in QST. Rigorous frequentist confidence regions (Christandl & Renner, 2012; Faist & Renner, 2016) and, recently, anytime-valid confidence sequences (Cumitini et al., 2026) set the calibration bar but use no learned prior. Bayesian credible-region methods (Blume-Kohout, 2010; Granade et al., 2016; Oh et al., 2019) use generic priors. The size-growing advantage of *structured* priors is known for point estimation (Cramer et al., 2010); our contribution is to realize and quantify it in a *learned, calibrated, core-parameterized* setting.

B Discussion: when each approach helps

The two approaches are complementary, and the integrity controls make the trade-off precise. Approach A—the generative prior with measurement-guided posterior inference—agrees with a gold-standard sampler on posterior means and 1D marginals, provides calibrated intervals, degrades gracefully (rather than collapsing) as states leave the family, and dominates classical few-measurement estimators that lack the family prior (Table 3). Its weaknesses are two. First, exposed by the $k{=}0$ prior-only control: on concentrated families its *point* accuracy is largely inherited from the prior and the measurements are under-used—especially with the random Pauli designs typical of prior-based inverse-problem solvers. Second, exposed by our own OOD spread check (App. G.3): its posterior spread grows only modestly with mismatch, far less than its point fidelity actually degrades, so its calibrated intervals would *not* reliably flag that it is extrapolating in a real deployment where the true state, and hence the true fidelity, is unknown—graceful point-fidelity degradation is

Table 3: Approach A vs. classical few-measurement estimators on a structured family (n=6; point fidelity). The learned prior dominates estimators that lack it. We stress (per the integrity controls) that on *concentrated* families this margin is largely the prior, not measurement information—which is why we caution against reading it as measurement-efficiency without the controls.

	Approach A (flow)	MPS-MLE	generic-prior pCN	classical shadows
point fidelity	$\mathbf{0.87}$–$\mathbf{0.91}$	≈ 0	≤ 0.05	0.75

not the same claim as trustworthy uncertainty under mismatch, and we do not have evidence for the latter. Approach B—the amortized measurement-conditioned estimator—directly extracts measurement information and, with the informative-local design, reconstructs at markedly higher fidelity while clearly beating prior-only and passing the shuffled control; its cost is that it is a point estimator with heuristic (dropout+conformal) uncertainty rather than a validated posterior, and is tied to a fixed measurement design. A practitioner wanting A's gold-standard-validated (in-distribution) posterior and graceful, non-catastrophic point-fidelity degradation off-manifold should prefer A, while recognizing its calibrated intervals are not shown to track that degradation; one wanting the best point reconstruction on a diverse family at a fixed informative protocol should prefer B; and the combination—A's posterior for uncertainty, B's estimate for the point—is a natural, if unexplored, hybrid. What both share, and what we most want to convey, is that neither can be trusted without the prior-only and shuffled-measurement controls.

C Limitations and ongoing work

We are explicit about scope. (a) *Fixed measurement design.* The estimator is trained for one informative measurement protocol; an amortized *design-agnostic* variant (trained over variable random Pauli subsets) collapses to the family mean for both a DeepSets and a set-transformer encoder (§5). Amortizing over arbitrary designs—or over a parameterized family of designs—is the main open direction. (b) *Family/regime dependence.* On concentrated families a prior-only estimate is already near-optimal, so our claims are meaningful only on genuinely diverse families; and the method is strongest near the trained family. (c) *Coordinate-dependent prior.* The canonicalization is a deterministic coordinatization, not a gauge quotient (Prop. 6); we neutralize this in practice with a gauge-invariant loss, but a principled mixed-canonical/Schmidt or gauge-equivariant model is the cleaner route. (d) *Hardware scope.* The closed-loop hardware run (§5, Table 6) covers 5 states of one low-entanglement family on one device; larger, higher-entanglement, and multi-device hardware studies remain future work. We verified that the two nearest diffusion-in-quantum works do not preempt our claim: Liu et al. (2025) perform state *generation* with non-learned analytic recovery, and QuaDiM (Tang et al., 2025) conditions on Hamiltonian parameters to generate measurement snapshots for property estimation—neither is a measurement-conditioned, calibrated full-state reconstruction over tensor-network cores.

D Proofs

Proof of Proposition 5. *(i)* The training loss equation 3 is $1 - |\langle \psi \,|\, \psi(g_\omega(\boldsymbol{o}))\rangle|^2$, a function of the two physical states, hence invariant to the MPS gauge $A^{(\ell)} \mapsto A^{(\ell)} G_\ell$, $A^{(\ell+1)} \mapsto G_\ell^{-1} A^{(\ell+1)}$ (which leaves $\Phi(\boldsymbol{\theta}) = |\psi(\boldsymbol{\theta})\rangle$ unchanged) and to global phase; and the physical readout $\Phi(\boldsymbol{\theta})$ is gauge-invariant by definition. *(ii)* The map c of equation 1–equation 2 is a composition of deterministic operations (QR with a fixed sign convention, Frobenius normalization, global-phase fixing), hence a well-defined function of its input cores; it defines both the target the network is trained to predict and the canonicalization applied to its output, so for fixed weights the map $\boldsymbol{o} \mapsto \Phi(c(g_\omega(\boldsymbol{o})))$ is a deterministic function from data to a legitimate pure state. Well-definedness does *not* require c to be gauge-invariant (it is not; Prop. 6), only that a single deterministic coordinatization is used throughout. □

Proof of Proposition 6. The reshaped-core QR with positive-diagonal sign fixing is a function of the actual core matrices, not of the gauge orbit: for a bond gauge G (invertible, non-triangular),

$A^{(\ell)} \mapsto A^{(\ell)}G$ gives $MG = Q_M R_M G = Q_M\, Q_{R_M G}\, R_{R_M G}$, so the left-canonical factor becomes $Q_M Q_{R_M G} \neq Q_M$ unless $Q_{R_M G} = I$, which fails on an open (positive-measure) set of G. Hence $c(g{\cdot}\boldsymbol{\theta}) \neq c(\boldsymbol{\theta})$ there while $\Phi(g{\cdot}\boldsymbol{\theta}) = \Phi(\boldsymbol{\theta})$. Hence for such a pair the two canonical targets $c(\boldsymbol{\theta}) \neq c(\boldsymbol{\theta}')$ encode the *same* physical state, so any coordinate-space objective (a non-constant density over cores, or a squared-error regression to $c(\boldsymbol{\theta}_\star)$) takes different values on physically identical states and is not a function of Φ. We emphasize the correct *practical* consequence: since training fixes one deterministic canonicalization, each state has a single target and the supervised task is *not* literally one-to-many; rather, canonical coordinates are not intrinsic state coordinates, so a coordinate MSE is representation-dependent and—because c is discontinuous where the QR/sign section jumps—poorly conditioned near those discontinuities, with Euclidean core distance not tracking state fidelity. The gauge-invariant loss equation 3 depends on $\boldsymbol{\theta}$ only through $\Phi(\boldsymbol{\theta})$ and is unaffected. □

Proof of Proposition 1. For a site-dependent normal (injective) open-boundary χ-MPS with injectivity length ℓ_0 (so $2^{\ell_0} \geq \chi^2$; minimal generic value $\ell_0 = \lceil 2\log_2 \chi \rceil$, i.e. $\ell_0{=}2$ for $\chi{=}2$), Cramer et al. (2010) show the state is the unique one—generically among all states—consistent with its reduced density matrices on contiguous windows of some block length, and give an explicit local reconstruction; we do not claim their exact constant, only our own dimension-counting estimate $\ell = \max(2\ell_0{-}1,\, 2\lceil \log_2 \chi \rceil{+}1)$ ($\ell{=}3$ at $\chi{=}2$, where the $2\lceil \log_2 \chi \rceil{+}1$ bond-dimension floor and the $2\ell_0{-}1$ injectivity requirement coincide; $\ell{=}7$ at $\chi{=}4$, where $\ell_0{=}4$ makes the injectivity term dominate). Each window reduced density matrix ρ_W expands in the Pauli basis of that window, $\rho_W = 2^{-|W|} \sum_Q \langle Q \rangle\, Q$, so its entries are linear in the expectations of the weight-$\leq \ell$ Pauli strings supported on W; there are $O(n\, 4^\ell) = O(n)$ such strings across all windows. What Cramer et al. prove is that this local data determines the state *uniquely*, i.e. that the reconstruction map $\{\rho_W\} \mapsto [\psi]$ is injective as a map of sets; this alone does not imply its differential is nonsingular (an injective smooth map need not be an immersion, e.g. $x \mapsto x^3$ at 0). We therefore do *not* derive the spanning property from uniqueness. Instead we take it as the content of Assumption 1, which we regard as generic rather than proven pointwise: for a real-analytic reconstruction map between manifolds of matching dimension, the locus where the differential is singular is a proper analytic subset (measure zero, by a Sard-type argument), so failure of the spanning property is non-generic, but we make no claim about the specific state or design at hand beyond this genericity. The count $k = O(n)$ versus $2^{n+1} - 2$ is immediate given the assumption. □

Proof of Proposition 7. By Assumption 1, $m|_{\mathcal{M}}$ is an immersion at $[\psi]$, so its differential $\mathrm{d}m$ has rank $d = \dim \mathcal{M}$; the assumed spanning Pauli design has differentials of rank d on $T_{[\psi]}\mathcal{M}$, which makes $\mathrm{d}(m|_{\mathcal{M}})$ injective, and the inverse function theorem gives a local diffeomorphism onto the image—i.e. local identifiability with $k = d$ measurements. (We invoke the *existence* of such a spanning design; because the full Pauli basis separates tangent directions, a spanning subset of size d exists, but we do not assert that d Paulis chosen uniformly at random span with high probability—that is a separate, quantitative claim we neither need nor prove here.) The parameter count $d \leq 2c(n,\chi)$ with $c(n,\chi) = \sum_\ell \chi_{\ell-1}{\cdot}2{\cdot}\chi_\ell = O(n\chi^2)$ is immediate from the core shapes. For the unstructured case, the pure-state ray space $\mathbb{CP}^{2^n-1}$ has real dimension $2(2^n - 1) = 2^{n+1} - 2$; no continuous map into $\mathbb{R}^k$ with $k < 2^{n+1} - 2$ can be injective, so unstructured identification needs $k \geq 2^{n+1} - 2$. The ratio is $(2^{n+1} - 2)/d \geq (2^{n+1} - 2)/(2c(n,\chi)) = \Omega(2^n/(n\chi^2))$. □

Remark 1. *Prop. 7 is a statement about* degrees of freedom / local identifiability*, matching the dimension of the informative measurement set to the dimension of the family; it is not a global recovery or noise-robustness guarantee. It nonetheless captures the mechanism behind the empirically observed size-growing advantage (Sec. 5): the learned prior restricts inference to a manifold whose dimension is polynomial in n, while unstructured estimators must resolve an exponentially larger space.*

Proof of Proposition 2. This is split-conformal prediction applied to the examples $Z_i = (|\psi_i\rangle, \boldsymbol{o}_i)$. Under exchangeability of $\{Z_i\}_{i\in\mathcal{C}} \cup \{Z_\star\}$ and a fixed (data-independent) score r, the rank of $r(Z_\star)$ among $\{r(Z_i)\}_{i\in\mathcal{C}} \cup \{r(Z_\star)\}$ is uniform on $\{1, \ldots, |\mathcal{C}|+1\}$. Hence $\Pr\big(r(Z_\star) \leq Q\big) \geq (1-\alpha)$ for Q the $\lceil (1-\alpha)(|\mathcal{C}|+1) \rceil$-th smallest calibration score, and $r(Z_\star) \leq Q \iff |O(\psi_\star) - \mu(\boldsymbol{o}_\star)| \leq Q\, \sigma(\boldsymbol{o}_\star) \iff O(\psi_\star) \in \mathcal{I}(\boldsymbol{o}_\star)$. Writing μ, σ as functions of the measurement record (not of the state) matters: for a given state, different shot realizations give different predictions, so the score is

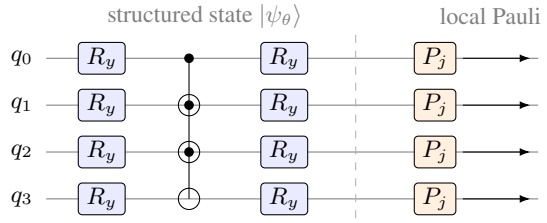

Figure 4: The structured state family and measurement model. A shallow R_y–CNOT–R_y circuit with random angles prepares a low-entanglement state $|\psi_\theta\rangle$; noisy expectation values of the informative-local Pauli observables P_j (single-qubit and nearest-neighbor) constitute the few-measurement tomographic data.

well defined only on the pair. The estimator is applied independently to each example, so r is a valid nonconformity score and no assumption on the prior or sampler is needed. □

E Experimental implementation details

State families. (i) *Perturbed random-MPS* (generic): fixed random base cores plus i.i.d. complex-Gaussian perturbations (σ=0.35), canonicalized. (ii) *Shallow-circuit* (physical, and the hardware run): $R_y(\theta_1)$ per qubit, an open nearest-neighbour CNOT ladder, and $R_y(\theta_2)$ per qubit, with angles $\theta_1 \sim \pi/2 + 0.6\,\mathcal{N}$, $\theta_2 \sim 0.6\,\mathcal{N}$ (a moderate spread; exactly χ=2). (iii) *Disordered paramagnetic TFIM ground states* are used only as the concentrated-family cautionary example.

MPS parameterization and gauge. Bond dimensions $\chi_\ell = \min(2^\ell, 2^{n-\ell}, \chi)$. Cores are canonicalized by the left-canonical QR sweep with real-positive R-diagonals, boundary-core Frobenius normalization, and a global-phase fix (Eqs. 1–2); the network's target is the (standardized) real/imaginary parts of these cores, but the training loss equation 3 is a state-fidelity and hence gauge-invariant.

Conditional model. Each measured pair (P_j, o_j) is encoded as $[\,\mathrm{onehot}(P_j) \in \{0,1\}^{4n}; o_j\,]$; the set encoder is either a DeepSets MLP (per-pair → mean-pool → decoder MLP, widths 512), a set-transformer (4 self-attention layers, d=256, 4 heads, learned-query attention pooling), or—as a baseline—a plain MLP that ignores the Pauli identities and consumes the fixed-order value vector $(o_j)_j$ (width 1024, 4 hidden layers). Trained by AdamW (lr 3×10^{-4}, cosine schedule, batch 256) for 30k steps on 2×10^4 (random) / 6×10^3 family states, with noisy Pauli inputs (shot variance $(1-\langle P\rangle^2)/S$, S=200). Training one model takes ∼2 min on one A6000.

Measurement designs. We compare four fixed designs. `local1`: the $3n$ single-qubit Paulis $\{X_i, Y_i, Z_i\}$. `local` (default): `local1` plus the $9(n-1)$ nearest-neighbor weight-2 Paulis $\{P_iP_{i+1}\}$ (63 at n=6). `local3`: the theory-complete design of Prop. 1—all non-identity Paulis supported on some contiguous 3-site window (207 at n=6); this is what Cramer et al. require for exact local reconstruction, and `local` is a strict subset. `random`: k Pauli strings drawn uniformly. *Budget accounting.* We distinguish three quantities. (i) The number of Pauli *observables* supplied to the estimator—63 for `local` at n=6, the value we call k. (ii) The number of distinct qubit-wise-commuting *measurement settings/circuits*: because many of the 63 Paulis co-measure, a greedy grouping covers all 63 with exactly **9** settings (and 9 is optimal—each of the $n-1$ edges must see all 9 single-qubit-basis pairs, a lower bound of 9). (iii) The total *copy cost*: if each of the 9 settings is run at S shots, the estimator consumes $9S$ state copies (not $63S$)—each setting's shots are shared across the ∼ 7 Paulis it resolves. We report k as the observable count and, where measurement efficiency is at issue, the $mS = 9S$ copy budget; we do *not* claim 63 separate circuits. *Simulated vs. hardware noise.* In the synthetic experiments each observable's error is drawn independently as $\mathcal{N}\big(0, (1-\langle P\rangle^2)/S\big)$. This reproduces the per-observable shot variance of running the 9 settings at S shots each, but not the correlations among the ∼ 7 Paulis that share a setting's shots, so the $9S$ copy budget is stated under this independent-error idealization (a genuinely grouped simulation would draw multinomial outcomes per setting and estimate its Paulis jointly). The hardware run

Table 4: Shot-budget ablation (shallow family, `local` design, common 30k-step training so S=200 matches the headline). Fidelity (mean±std across states) and gain over prior-only (k=0) increase monotonically with the shot count S per measurement setting.

shots per setting S	50	200	800
conditional fidelity	0.940 ± 0.053	$\mathbf{0.967\pm0.046}$	0.977 ± 0.038
gain over prior-only	$+0.565$	$+0.592$	$+0.602$
true − shuffled	$+0.782$	$+0.813$	$+0.824$

does not use the idealization: all 63 observables are submitted in one `EstimatorV2` PUB per state and co-measured groups are estimated from shared shots, so the reported hardware fidelities already carry the within-setting correlations.

Shot budget. Table 4 varies the shot count S per measurement setting at the informative-local design on the shallow family, at the *same* 30k-step training budget as the headline (so the S=200 column reproduces the Table 1 headline 0.967 exactly—the earlier apparent gap was purely a 12k- vs. 30k-step training difference, now removed). Fidelity and the gain over the k=0 prior-only baseline both rise monotonically with S as the shot-noise floor on each Pauli value drops, confirming that Approach B extracts genuine measurement information rather than saturating at the prior; the effect is smooth, consistent with a well-conditioned informative design.

Family diversity. To make the "concentrated vs. diverse" distinction quantitative we report the test states' nearest-training-neighbour fidelity and the medoid (a k=0 estimate) fidelity to test states. Paramagnetic TFIM: nearest-neighbour ≈ 1.00, medoid 0.997 (essentially a single state). Random-MPS (σ=0.35): nearest-neighbour 0.82, medoid 0.65. Shallow-circuit (spread 0.6): medoid 0.37; sweeping the angle spread traces out a family whose medoid falls from $\sim$0.8 to $\sim$0.06 (Fig. 7). The measurement gain over prior-only is meaningful only when the medoid is far from 1.

Ablation protocols. *Design richness* (`local1`/`local`/`local3`): same model and training, different fixed $\mathcal{D}$. *Encoder* (DeepSets/set-transformer/plain-MLP): same `local` design. *Shots*: $S \in \{50, 200, 800\}$. *Diversity*: shallow-circuit angle spread $\in \{0.3, 0.6, 0.9, 1.2\}$. *Seeds*: to separate training-stochasticity variance from across-state dispersion, we retrain the headline configuration with 3 independent seeds (varying network initialization, minibatch order, and the training-time shot-noise realizations, while the family bank and held-out test set are held fixed). The mean reconstruction fidelity is essentially seed-invariant—0.968 ± 0.001 (shallow) and 0.954 ± 0.001 (random) across the 3 seeds—so the larger ± values reported elsewhere (±0.046/±0.021) are *across-test-state* dispersion, not seed noise; we report the two separately to avoid conflating them.

Uncertainty and calibration. We use dropout (p=0.1) in the encoder/decoder; at test time E=24 stochastic forward passes give a dropout-ensemble spread over each of the $2n$ local Z/X observables (ensemble mean μ_O, scale σ_O). Split-conformal recalibration on a random half of the held-out states ($N_{\text{cal}}\approx$100) with score $|O - \mu_O|/\sigma_O$ gives the multiplier Q; we report coverage and clipped width averaged over 10 random splits. The scores of all $2n$ observables and all calibration states are calibrated *per observable*: each of the $2n$ observables gets its own multiplier Q_O, the $\lceil (1-\alpha)(N_{\text{cal}}+1) \rceil$-th smallest of *its own* calibration scores, so the procedure is exactly the one Prop. 2 analyses and the reported coverage inherits its finite-sample guarantee.

Full reliability curves. Table 5 reports the empirical coverage and mean clipped interval width across the full range of nominal levels for both families, from the split-conformal procedure above (averaged over 10 random calibration/test splits). Empirical coverage tracks the nominal level to within ±0.01 at every level, and widths grow monotonically as expected—this is the distribution-free marginal guarantee that the raw dropout-ensemble spread does *not* provide.

Native contraction (scalability). Every Pauli expectation $\langle\psi|\bigotimes_i P_i|\psi\rangle$ and overlap $\langle\psi|\phi\rangle$ is computed by left-to-right transfer-matrix contraction, $L \leftarrow \sum_s A_s^{(i)\dagger}(\,\cdot\,)(P_i A^{(i)})_s$ per site, at $O(n\chi^3)$ cost with no 2^n object; on small n these primitives match the dense contraction to $\leq 10^{-8}$. The

Table 5: Conformalized predictive intervals (per-observable Q_O): empirical coverage (and mean clipped width) vs. nominal level $1-\alpha$, per family (n=6, $2n$ local observables, 10 splits). Coverage matches nominal to ± 0.01 throughout.

nominal $1-\alpha$	0.50	0.60	0.70	0.80	0.90	0.95
random: coverage	0.499	0.596	0.690	0.797	0.897	0.948
random: width	0.076	0.095	0.117	0.148	0.198	0.246
shallow: coverage	0.495	0.593	0.691	0.799	0.896	0.948
shallow: width	0.090	0.114	0.141	0.177	0.234	0.294

Table 6: Closed loop on `ibm_aachen` (Heron): the *local* design (63 Paulis) measured on 5 held-out states in one batched `EstimatorV2` job at 4096 shots/setting (vs. S=200 simulated), then reconstructed. Fidelity is to the *ideal circuit target*, not independently-verified tomography of the device's own output (see text). Hardware nearly matches noiseless reconstruction; both exceed k=0 prior-only.

device $\lvert hw-\text{ideal}\rvert$	hw shots	convention check	**fid. from HW**	fid. from ideal	prior-only (k=0)	per-state HW range
0.024	4096	$<10^{-7}$	$\mathbf{0.971 \pm 0.006}$	0.977	0.283	0.964–0.979

pipeline runs at n=16, 20 with peak GPU memory 0.134/0.147 GB (flat in n; §5), on a single NVIDIA A6000, reaching fidelity 0.063/0.037 at k=32 random Paulis (reduced training budget; the classical MLE baseline is $\approx 10^{-5}$ at the same budget)—a feasibility demonstration of the polynomial-cost machinery, not a high-fidelity result.

IBM hardware (closed loop). 5 held-out shallow-circuit states (spread 0.6, drawn from the same family the conditional model was trained on but disjoint from the training bank) are prepared on `ibm_aachen` (156-qubit Heron), transpiled at optimization level 2; the expectation values of the *same* 63 *local Paulis* used by the estimator are obtained with `EstimatorV2` (4096 shots) in a single batched job. Qubit and Pauli conventions between the Qiskit circuit and the torch reconstruction are matched and verified against the noiseless statevector (agreement 8.9×10^{-8}); the measured expectations lie within mean 0.024 of the ideal values across the 5 states. We then run the trained conditional estimator on these hardware values; because the estimator is a DeepSets set-encoder, Pauli ordering is immaterial, and we pair each hardware value with its Pauli one-hot. Reconstruction fidelities are 0.964, 0.976, 0.968, 0.968, 0.979 (from hardware) versus 0.977 mean from the noiseless values, against a k=0 prior-only medoid of 0.283.

F Additional experimental results

IBM hardware closed loop.

Unmeasured-observable calibration: what the value actually comes from. §5 reports that the reconstruction predicts 18 never-input observables far better than a prior-only estimate, at $\approx 90\%$ coverage. To check that this is not merely "a point estimate plus standard conformal calibration" (which would hold for any reasonable point predictor), we compare, on the same held-out set and the same calibration/test split, three variants at matched $\approx 90\%$ coverage: (1) our point estimate with the per-state MC-dropout scale (the paper's default); (2) the *same* point estimate with a single *fixed* (non-heteroscedastic) scale from calibration-set residuals; (3) a classical per-state MPS-MLE point estimate (same local design, 3 restarts $\times$ 500 steps) with a fixed scale.

	MAE (random/shallow)	width (random)	width (shallow)	coverage (random/shallow)
(1) ours, per-state scale	0.043 / 0.053	0.190	0.224	89.7%/90.1%
(2) ours, fixed scale	—	0.183	0.237	90.1%/90.5%
(3) classical MLE, fixed scale	0.049 / 0.039	0.210	0.169	90.0%/90.8%

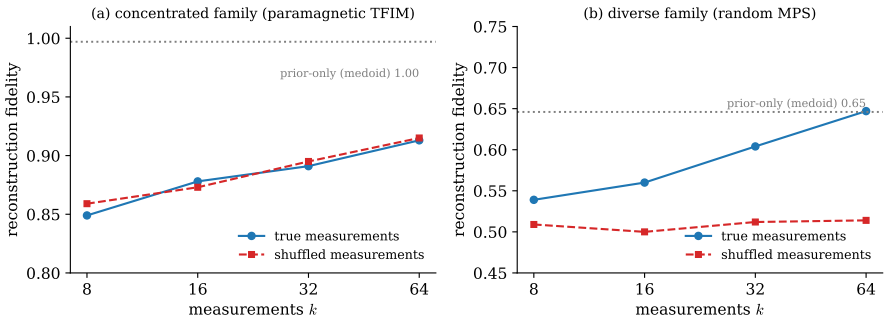

Figure 5: Integrity controls. **(a)** On a concentrated family, a $k{=}0$ prior-only medoid (0.997) dominates and true $\approx$ shuffled: the measurements are not used, so "few-measurement fidelity" is memorization. **(b)** On a diverse family the medoid is far from optimal and true reconstruction separates from shuffled, with the gap growing in the budget k—genuine measurement information. Both controls are prerequisites for interpreting any learned-QST fidelity.

Table 7: Conformal calibration of the local-Z/X observable intervals, one multiplier Q_O per observable (nominal 90%, mean±std over 10 splits). Raw MC-dropout is over-confident; conformal recalibration restores coverage, and the intervals are sharp because reconstruction fidelity is high.

family	raw coverage	**conformal coverage**	clipped width
random	68.9%	$\mathbf{89.7 \pm 1.0}\%$	0.198 ± 0.004
shallow (physical)	72.2%	$\mathbf{89.6 \pm 1.7}\%$	0.234 ± 0.006

The honest picture is mixed. On the *random* family our full pipeline is narrowest, and the per-state heteroscedastic scale is not what buys this—the fixed-scale variant on the *same* point estimate is narrower still, so the gain over classical MLE here comes from the point estimate, not the per-state scale. On the *shallow* family the classical MLE point estimate is actually the most *accurate* of the three for these specific held-out observables (MAE 0.039 vs. our 0.053) and gives the sharpest interval (0.169), consistent with Table 2's finding that MLE is the stronger classical baseline on this family at the `local` design. What survives on both families is that some reconstruction—ours or MLE's—beats the prior-only estimate by 3–5$\times$ in MAE; we do *not* claim our specific per-state heteroscedastic scale is uniformly sharper than a fixed-scale baseline, only that a calibrated point predictor (of either kind) gives real, checkable value where prior-only gives none.

What "unmeasured" does and does not rule out. The independent-per-observable noise idealization used throughout (main text, Background) means the simulated experiments never generate a raw per-setting bit-string record, so under this simulated protocol the 18 held-out observables are never queried and no estimate of them is available from the input data. This is *not* a claim about every possible grouped implementation of the local design: a specific choice of which qubits share a measurement setting could, in principle, let some held-out pair be read off from an existing setting's shots (e.g. if two far-apart qubits happen to share a Z-basis setting, that pair's Z_iZ_j correlator comes for free). We have not attempted to choose, or rule out, such alignments for a concrete 9-setting realization of the `local` design; a protocol-level accounting of which held-out observables are provably unrecoverable under a specific grouped measurement scheme is left to future work. The hardware run (below) does use a real grouped implementation, but only for the 63 input observables—the held-out set was not measured there either.

Integrity controls.

Calibration curves. Fig. 6 shows the reliability and width curves behind Table 7.

Family-diversity sweep.

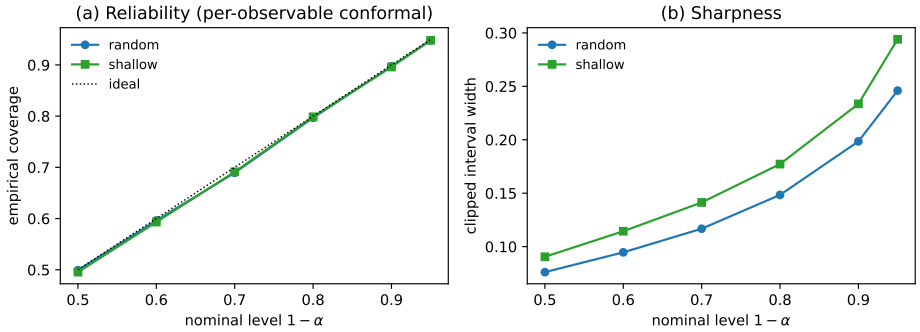

Figure 6: Calibration of the dropout predictive ensemble over the $2n$ local Z/X observables (per-observable conformal, E=24 samples, 10 splits). **(a)** Empirical coverage tracks the nominal level across the range (ideal: diagonal). **(b)** Interval width stays small (physical range $[-1, 1]$), because the high-fidelity reconstruction concentrates the predictive ensemble. Raw (uncalibrated) coverage is 69–72% at the 90% level.

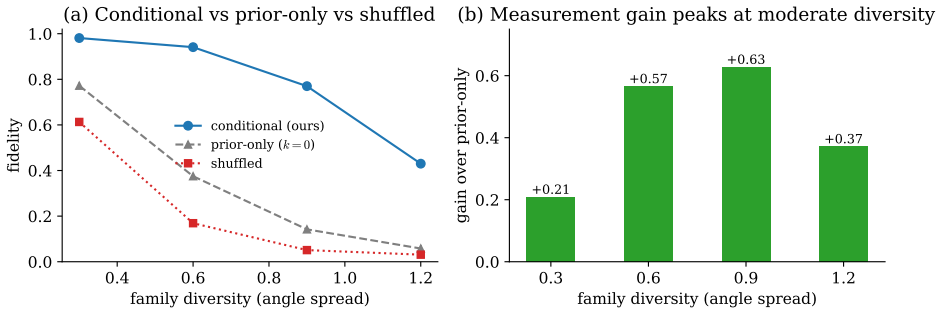

Figure 7: Family-diversity sweep (shallow-circuit family, local design). **(a)** As diversity grows the k=0 prior-only fidelity falls and conditional reconstruction eventually gets harder, but the shuffled control stays low throughout. **(b)** The *measurement gain* over prior-only peaks at moderate diversity: too concentrated and the prior alone suffices; too diverse and reconstruction is hard.

Family-aware simple baselines, and why the MPS core. Two further baselines isolate what the representation contributes. (i) A *measurement-space k-NN*: for a test measurement vector, return the state of the nearest (or mean of the 5 nearest) training states in measurement space. This trivial memorization reaches 0.80/0.82 (1-NN) and 0.90/0.86 (5-NN) on shallow/random—useful, but clearly below the amortized estimator (0.967/0.954), so the method is not merely retrieving a nearby training state. (ii) A *direct dense-state MLP* that outputs the 2^n amplitudes instead of MPS cores, same loss and protocol, reaches 0.979/0.958—i.e. at n=6 it *matches or slightly exceeds* the MPS-core estimator. We report this openly: the MPS-core parameterization is *not* what buys accuracy at small n. Its value is scalability—the dense-state output has 2^n parameters and is infeasible at the n=16, 20 where the MPS core (with $O(n\chi^2)$ parameters and native $O(n\chi^3)$ contraction) still runs. The core representation is a scalability choice, not a small-n accuracy claim.

Native-contraction memory.

G Approach A: Generative Prior and Posterior Inference Details

G.1 Construction

Approach A: generative prior and Bayesian posterior inference. We train a rectified-flow prior v_ϕ over standardized canonical cores $\boldsymbol{\theta}$ by flow matching, $\mathcal{L}(\phi) = \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\epsilon},t}\|v_\phi(\boldsymbol{\theta}_t, t) - (\boldsymbol{\theta} - \boldsymbol{\epsilon})\|^2$ with $\boldsymbol{\theta}_t = (1-t)\boldsymbol{\epsilon} + t\boldsymbol{\theta}$. This is, to our knowledge, the first learned *generative* distribution over tensor-network cores (as opposed to a tensor network used *as* a generative model (Han et al., 2018) or fit to

Table 8: Bond-dimension sweep (random-MPS family, n=6, `local` design). Fidelity degrades gently and the measurement gain persists even though `local` is theoretically sub-complete for χ>2.

bond dimension χ	2	3	4
conditional fidelity	0.954±0.021	0.908±0.036	0.878±0.063
prior-only (k=0)	0.680	0.594	0.581
gain over prior-only	+0.27	+0.31	+0.30
true − shuffled	+0.45	+0.50	+0.45

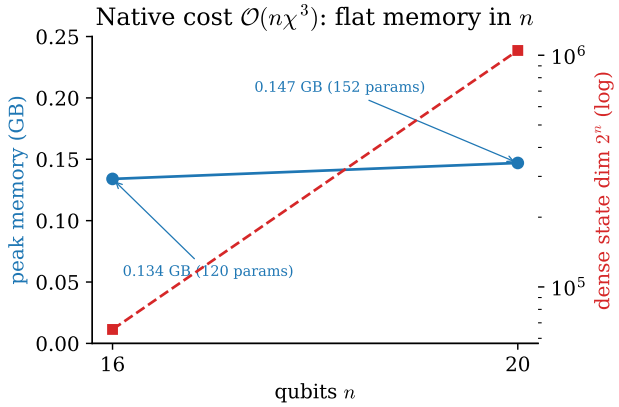

Figure 8: Native $O(n\chi^3)$ contraction: peak GPU memory is flat in n (0.134 GB at n=16, 0.147 GB at n=20; 120/152 complex core parameters) while the dense state dimension 2^n grows exponentially (65,536 → 1,048,576). Running the full conditional pipeline; a computational-feasibility result, not high-fidelity tomography at these sizes.

a single state (Torlai et al., 2023)). Given data $\boldsymbol{o}$ with the heteroscedastic shot-noise likelihood $\ell(\boldsymbol{o} \mid \boldsymbol{\theta}) \propto \exp(-\frac{1}{2}\sum_j(\langle P_j\rangle_{\boldsymbol{\theta}} - o_j)^2/\varsigma_j^2)$, $\varsigma_j^2 = (1-\langle P_j\rangle^2)/S$, the posterior-conditioned flow is the Doob h-transform of the prior; we approximate its intractable guidance by a proximal data-consistency step on the flow's clean-core estimate at each integration step (a DiffPIR/DPS-style guided sampler (Chung et al., 2023)), producing posterior samples and a reconstruction on the low-dimensional core manifold. This is a principled, measurement-guided posterior sampler—we validate it against a gold-standard importance sampler in Sec. 5—and it is the "unconditional-prior" route the integrity controls scrutinize.

G.2 Recovery theory for Approach A

Proposition 3 (Prior-restricted stable recovery, with model mismatch)**.** Hypothesis (RIP on $\mathcal{M}$). *Suppose m restricted to the family satisfies a manifold restricted-isometry condition: there is $\delta \in (0,1)$ with $(1-\delta)\|[\psi]-[\psi']\|^2 \le \|m([\psi])-m([\psi'])\|^2 \le (1+\delta)\|[\psi]-[\psi']\|^2$ for all $[\psi],[\psi'] \in \mathcal{M}$. Let $|\psi_\star\rangle$ be the true state,* not necessarily in $\mathcal{M}$*; let $\boldsymbol{o} = m([\psi_\star]) + \boldsymbol{e}$ be the noisy data; let $\eta_{\text{prior}} := \min_{[\psi]\in\mathcal{M}} \|m([\psi]) - m([\psi_\star])\|$ be the measurement-space model-mismatch (prior-fit) error, attained at $[\psi_\sharp] \in \mathcal{M}$; and let $[\hat{\psi}] \in \mathcal{M}$ be any point whose data residual is within η of the prior's best manifold point, $\|m([\hat{\psi}]) - \boldsymbol{o}\| \le \min_{[\psi]\in\mathcal{M}} \|m([\psi]) - \boldsymbol{o}\| + \eta$. Then*

$$\big\|\,[\hat{\psi}]-[\psi_\sharp]\,\big\| \;\le\; \frac{2\|\boldsymbol{e}\| + 2\eta_{\text{prior}} + \eta}{\sqrt{1-\delta}}, \qquad \big\|\,[\hat{\psi}]-[\psi_\star]\,\big\| \;\le\; \frac{2\|\boldsymbol{e}\| + 2\eta_{\text{prior}} + \eta}{\sqrt{1-\delta}} + \big\|[\psi_\sharp]-[\psi_\star]\big\|, \tag{4}$$

where $\|[\psi_\sharp] - [\psi_\star]\| \ge \operatorname{dist}([\psi_\star], \mathcal{M})$ is the (irreducible) state-space gap from the true state to the prior manifold. When $[\psi_\star] \in \mathcal{M}$ both mismatch terms vanish (η_{prior}=0, $[\psi_\sharp]$=$[\psi_\star]$) and the bound reduces to $(2\|\boldsymbol{e}\| + \eta)/\sqrt{1-\delta}$.

Remark 2 (Scope of the RIP hypothesis)**.** *We treat the manifold-RIP as a hypothesis, not a derived fact. Establishing it for a concrete Pauli measurement design—with explicit dependence on the manifold's covering number, curvature/reach, the number of measurements k, measurement normalization, and a failure probability—is a substantial question in its own right that we do not settle here; we state the recovery bound conditionally on it. The OOD experiments (Sec. 5) probe exactly the $\eta_{\mathrm{prior}}, \mathrm{dist}([\psi_\star], \mathcal{M}) > 0$ regime that this generalized bound describes.*

Proposition 4 (Best-of-N selection)**.** *Fix a tolerance $\tau > 0$, let $q_{\boldsymbol{o}}$ denote the sampler's output distribution over cores at data $\boldsymbol{o}$ (the distribution the N draws are actually i.i.d. from), and let $\rho_{\boldsymbol{o}} = \Pr_{\boldsymbol{\theta}\sim q_{\boldsymbol{o}}}\big[\|m([\psi(\boldsymbol{\theta})]) - \boldsymbol{o}\| \le \tau\big]$ be the $q_{\boldsymbol{o}}$-mass of τ-data-consistent states. With $N \ge \log(1/\beta)/\rho_{\boldsymbol{o}}$ independent draws, the best-of-N estimate $[\hat{\psi}] = \arg\min_r \|m([\psi^{(r)}]) - \boldsymbol{o}\|$ has data residual $\le \tau$ with probability at least $1-\beta$; combined with Prop. 3 (with $\eta = \tau$) this bounds its reconstruction error. Thus best-of-N trades computation for accuracy at a rate governed by how much mass the sampler places near the measurement-consistent set.*

This appendix documents the generative-prior route (Approach A) used for the posterior-validation and out-of-distribution results (§5, Figs. 9–10, Table 3). Approach A and Approach B share the *same* canonicalized-MPS-core parameterization (App. E); they differ only in how the measurements enter—as a learned point estimate (B) versus as a likelihood that guides sampling from a learned prior (A).

Flow prior over tensor-network cores. We train a rectified-flow (flow-matching) generative model $p_\theta(\mathbf{c})$ over the standardized real/imaginary parts $\mathbf{c} \in \mathbb{R}^d$ of the canonicalized cores of a state family. The velocity field $v_\theta(\mathbf{c}, t)$ is a residual MLP (width 512, 6 blocks, sinusoidal time embedding) trained with the conditional flow-matching objective on straight-line interpolants $\mathbf{c}_t = (1-t)\,\boldsymbol{\varepsilon} + t\,\mathbf{c}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I)$. To our knowledge this is the first generative model placed directly on canonicalized tensor-network cores; canonicalization is what makes the target distribution low-dimensional and learnable (an un-gauge-fixed core cloud is a high-dimensional orbit and the flow fails to converge). Sampling integrates the probability-flow ODE with 50 RK4 steps.

Measurement-guided posterior inference. Given measurements $o = \{(P_j, o_j)\}$ with Gaussian shot-likelihood $p(o \mid \mathbf{c}) \propto \exp\big(-\frac{1}{2\tau^2}\sum_j(\langle P_j\rangle_{\mathbf{c}} - o_j)^2\big)$, the target posterior is $\pi(\mathbf{c}) \propto p_\theta(\mathbf{c})\, p(o \mid \mathbf{c})$. We sample it two ways. *(i) Doob h-transform / guided flow*: we tilt the prior flow by the measurement likelihood, adding the score of a Gaussian relaxation of $\log p(o \mid \mathbf{c}_t)$ to the velocity at each step, which implements an approximate Doob h-transform of the prior process toward the measurement-consistent set. *(ii) Proximal correction*: after each guided step we take one proximal-gradient step on the data misfit $\frac{1}{2\tau^2}\sum_j(\langle P_j\rangle - o_j)^2$, projecting back toward the prior manifold with a short flow-denoise. Both are cheap (~ 1 s per state at n=6); we use (ii) for the reported point fidelities and (i) for posterior samples.

Gold-standard SIR reference. To check that the guided sampler targets the *intended* posterior rather than a convenient relaxation, we compare against sampling-importance-resampling (SIR) with the flow prior as proposal: draw N=10^5 prior samples, weight by the exact shot-likelihood, and resample. Fig. 9 reports the resulting posterior-mean correlation (r=0.98 over 25 states $\times\, 2n$ observables), the energy distance between guided and SIR posterior marginals (0.033), and the effective sample size (ESS $\approx$ 200 of 10^5, consistent with a well-concentrated but non-degenerate posterior). This is the sense in which Approach A is a validated posterior, not a relabelled point estimate.

Classical baselines (Table 3). *MPS-MLE*: maximum-likelihood MPS fit to the same k Pauli values by gradient descent on the negative log-likelihood, with no family prior. We verified this near-zero fidelity is genuine *under-determination*, not an optimizer artifact: (a) with R=12 random restarts and reporting the best, the fidelity does not improve and the data residual reaches machine zero ($\sim 10^{-8}$) while the fidelity stays ≈ 0—the fit is a perfect but non-unique data-explainer; (b) increasing optimizer steps 3$\times$ leaves fidelity flat (0.016–0.019); and (c) as a positive control, giving MLE a *tomographically complete* Pauli set (all 1023 at n=5) recovers the state exactly (fid $= 1.000$, residual 1.6×10^{-15}), and a noiseless k=128 TFIM fit jumps to 0.999. So at few informative-local measurements the low MLE fidelity is an information gap the learned prior fills, not a solver failure. *Generic-prior pCN*: preconditioned Crank–Nicolson MCMC on the cores under a *generic* Gaussian

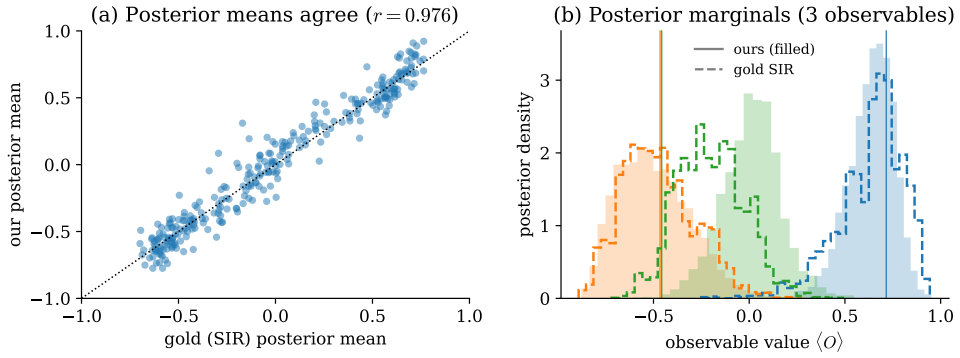

Figure 9: Approach A's marginal statistics agree with a gold-standard posterior (n=6). **(a)** Posterior means of the flow sampler vs. a gold-standard SIR posterior, over 25 states and $2n$ local observables, lie on the diagonal (r=0.98). **(b)** Example posterior marginals (three observables): the sampler's draws (filled) overlap the SIR posterior (dashed); vertical lines are the true values. This is evidence for, not proof of, correct posterior targeting: it checks agreement on per-observable means and 1D marginals, not the full joint posterior.

prior (i.e. the same likelihood but without the learned family prior), which explores far too large a set and reaches ≤ 0.05. *Classical shadows* (Huang et al., 2020): randomized single-qubit Pauli measurements with the standard median-of-means estimator. We give shadows the *more generous* kS-snapshot budget, which at n=6 is $63S$ copies against the grouped $9S$ our estimator consumes, so the comparison is conservative in the baseline's favour; it reaches 0.75 on this family, and the learned prior improves on it because the family is low-entanglement and strongly structured. We reiterate the integrity caveat: on *concentrated* families this entire margin is largely attributable to the prior, per the k=0 control (§5)—the baseline comparison shows Approach A dominates prior-free estimators, not that the measurements are being efficiently used.

Out-of-distribution protocol (Fig. 10). We form $|\psi_\delta\rangle \propto (1-\delta)|\psi_{\text{fam}}\rangle + \delta|\psi_{\text{Haar}}\rangle$ for $\delta \in [0, 1]$ ($|\psi_{\text{Haar}}\rangle$ a Haar-random n-qubit state), run Approach A on its measurements, and report (a) reconstruction fidelity to $|\psi_\delta\rangle$, (b) a hallucination probe: the reconstruction's fidelity to the training family versus to the true state, and (c) whether the posterior ensemble's spread tracks δ (App. G.3). Fidelity degradation is monotone and threshold-free rather than collapsing at a sharp boundary.

G.3 Empirical results for Approach A

Approach A: a gold-standard-validated posterior that leans on the prior. Before conditioning, we evaluate the generative-prior route (§ Method) against a self-normalized importance sampler drawn from the learned prior and reweighted by the exact likelihood (the gold standard): the sampler's posterior means and per-observable 1D marginals agree closely with the gold standard (n=6: posterior-mean correlation 0.98, median 1D energy distance 0.033, ESS $\sim 200/10^5$). This is consistent with the sampler targeting the intended posterior rather than a relabelled point estimate—we check marginal agreement on $2n$ observables, not the joint distribution over all core coordinates, so we report it as evidence for, not proof of, correct posterior targeting. Approach A also clearly beats classical estimators at few measurements—maximum-likelihood, a generic-prior pCN sampler, and classical-shadow reconstruction all sit far below it on structured families—and it degrades gracefully as states leave the family rather than hallucinating. *Yet the integrity controls temper the interpretation.* On the concentrated family a k=0 prior-only estimate already matches it and shuffling the measurements barely changes the reconstruction (Fig. 5a); even on the diverse random family the prior-only medoid is competitive with the flow reconstruction at the budgets tested. Approach A is thus a valid, calibrated Bayesian posterior whose *point accuracy* at few measurements is largely inherited from the prior—exactly the gap that motivates conditioning directly on the measurements (Approach B), to which we now turn.

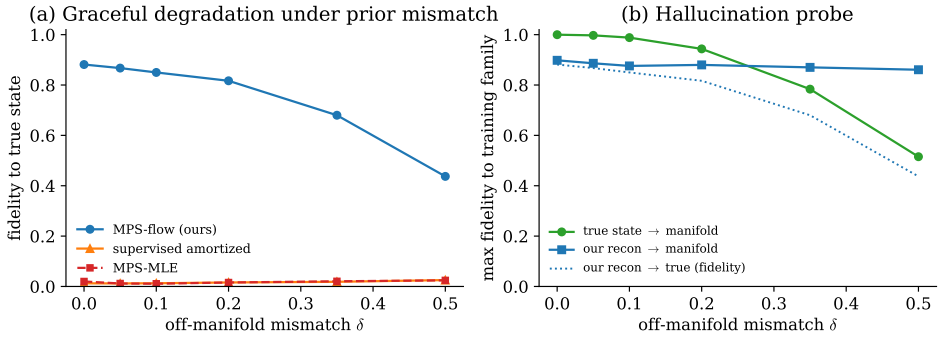

Figure 10: Out-of-distribution behaviour of Approach A (physical family, n=6). **(a)** As the true state is pushed off the manifold (mismatch δ), fidelity degrades gracefully. **(b)** Hallucination probe: the reconstruction stays near the training family (blue) while the true state leaves it (green); the reconstruction's fidelity to the *true* state (dotted) declines with δ rather than staying spuriously high.

Approach A degrades gracefully off the training manifold. Because Approach A is a prior-based method, we stress-test it off the training manifold by mixing in a Haar-random direction, $|\psi_\delta\rangle \propto (1-\delta)|\psi_{\text{fam}}\rangle + \delta|\psi_{\text{Haar}}\rangle$ (Fig. 10). Reconstruction fidelity degrades *gracefully and monotonically* ($0.88 \to 0.44$ as δ grows) rather than collapsing at a threshold, and a hallucination probe shows the reconstruction is pulled toward the training family but its fidelity to the true state declines in step with δ rather than staying spuriously high—evidence that the reported *fidelity metric* is not itself hallucinating, though not evidence that the model's own *uncertainty* registers the growing error; these are different claims; we checked the second directly. At a reduced training budget (a confirmatory check, not the headline budget above), the posterior ensemble's spread—mean pairwise infidelity among K=24 samples, and the per-observable standard deviation used for calibration—grows only modestly from δ=0 to δ=0.5 ($0.439 \to 0.457$ and $0.204 \to 0.208$ respectively), non-monotonically along the way, while point fidelity collapses over the same range ($0.60 \to 0.32$): the calibrated interval would stay almost as narrow at δ=0.5 as at δ=0, i.e. falsely confident, if the true state (and hence the true fidelity) were unknown, as in a real deployment. Graceful, non-catastrophic point-fidelity degradation is Approach A's genuine strength here; a calibrated *warning sign* under mismatch is not something our experiments support.

H Extended theoretical discussion

We collect here the formalization that supports, but is not the substance of, the design choices in the main text.

Proposition 5 (Gauge-invariant objective; well-definedness under a fixed coordinatization)**.** *(i) The training objective equation 3 and the physical readout Φ depend on the cores only through the physical state $[\Phi(\boldsymbol{\theta})]$, and are therefore invariant to the MPS gauge and global phase. (ii) For a fixed deterministic coordinatization c (equation 1–equation 2), applied consistently to define the network's target and to canonicalize its output, the reconstruction map $\boldsymbol{o} \mapsto \Phi(c(g_\omega(\boldsymbol{o})))$ is well-defined. We make no claim that this map is invariant under a change of coordinatization $c \to c'$ (Prop. 6); (i) only guarantees that the objective and readout do not depend on* which *gauge representative encodes a given state.*

Proposition 6 (Core coordinates are gauge-dependent; why we use a gauge-invariant loss)**.** *Let the MPS gauge group $\mathcal{G}$ act on cores by state-preserving transformations $\boldsymbol{\theta} \mapsto g{\cdot}\boldsymbol{\theta}$ (so $\Phi(g{\cdot}\boldsymbol{\theta}) = \Phi(\boldsymbol{\theta})$). The canonicalization c is* not *$\mathcal{G}$-equivariant: there is a positive-measure set of pairs $\boldsymbol{\theta}, \boldsymbol{\theta}' = g \cdot \boldsymbol{\theta}$ with $\Phi(\boldsymbol{\theta}) = \Phi(\boldsymbol{\theta}')$ yet $c(\boldsymbol{\theta}) \neq c(\boldsymbol{\theta}')$. Consequently any* coordinate-space *objective—a density $q(\cdot)$ over canonical cores, or a core-regression loss $\|g_\omega(\boldsymbol{o}) - c(\boldsymbol{\theta}_\star)\|^2$—is not a function of the physical state. We stress the* practical *reading: in training each state yields a* single *deterministic canonical target, so this is not a literal multi-label conflict where one input carries many gauge labels. Rather, because c is discontinuous and Euclidean core distance does not track state fidelity, a core-regression target is geometrically mismatched to the reconstruction objective; the* gauge-invariant *fidelity loss equation 3 removes the mismatch, which is why we use it.*

Remark 3. *This is a statement about coordinates, and it motivates our gauge-invariant loss: empirically a core-regression variant is markedly worse (§5). It is also why we say "canonicalized" rather than "gauge-fixed"; a quotient-aware (e.g. Schmidt-coordinate) parameterization is the principled alternative we leave open.*

Assumption 1 (Immersed family with a spanning Pauli design)**.** *$\mathcal{M}$ is a smooth embedded submanifold of the pure-state ray space of real dimension $d = \dim \mathcal{M} \le 2\,c(n,\chi)$, where $c(n,\chi) = \sum_\ell \chi_{\ell-1}\cdot 2\cdot\chi_\ell = O(n\chi^2)$ is the number of complex core parameters; at the state of interest the measurement map $m|_{\mathcal{M}}$ is an immersion (its differential is injective on the tangent space); and the chosen set of k Pauli observables* spans the cotangent image*, i.e. the differentials $\{\mathrm{d}\langle P_j\rangle\}_{j=1}^{k}$ have rank d on $T_{[\psi]}\mathcal{M}$. (We assume such a spanning design* exists *and is used; we do* not *claim that d Pauli strings drawn uniformly at random attain rank d with high probability.)*

Proposition 7 (Degrees-of-freedom separation, structured vs. unstructured)**.** *Under Assumption 1, a spanning set of $k \ge d$ Pauli measurements makes $m|_{\mathcal{M}}$ a local diffeomorphism onto its image, so it locally identifies a state* within *$\mathcal{M}$ from d real numbers. In contrast, identifying an arbitrary pure state requires at least $2^{n+1}-2$ real measurements (its intrinsic dimension). Hence the ratio of* degrees of freedom *to be resolved is $(2^{n+1}-2)/d = \Omega\big(2^n/(n\chi^2)\big)$, which grows exponentially in n. This is a dimension-counting (local, noiseless identifiability) statement, not a sample-complexity or noise-robustness theorem; it quantifies why a prior that restricts inference to $\mathcal{M}$ can succeed in the few-measurement regime where unstructured estimators cannot.*

Why the experimental `local` design suffices in practice (relation to Prop. 1). Prop. 1 states the *sufficient* condition of Cramer et al. (2010): an injective χ-MPS is uniquely determined among all states by the collection of its reduced density matrices on every contiguous 3-site window, i.e. by the expectation values of all Paulis supported on such windows (the `local3` design, 207 observables at n=6). Our default experimental design `local` (single-qubit plus nearest-neighbor weight-2 Paulis, 63 at n=6) is a *strict subset* of `local3`: it omits the weight-3 terms and the non-adjacent weight-2 terms inside each window. It is therefore *not* guaranteed to be information-complete for arbitrary χ-MPS. The reason it nonetheless nearly matches `local3` empirically (Table 2: 0.954/0.967 vs. 0.950/0.957) is that the *learned prior* supplies the information that the missing higher-weight windows would otherwise carry: within a low-entanglement family, the two-body correlations plus the prior already pin down the state, so the marginal value of the weight-3 terms is small. This is precisely the division of labour our integrity controls quantify—and it is why we describe `local` as *informative* rather than *information-complete*, and report `local3` as the theory-faithful design in the ablation rather than as the headline. A design that is provably complete *and* minimal for a given learned prior is an open problem we do not claim to solve.

Gauge coordinate-dependence, made precise (elaborating Prop. 6). Let $\Phi : \mathcal{M} \to \mathbb{R}^d$ send a physical state to the standardized real/imaginary parts of its left-canonical cores. Canonicalization fixes the bond gauge up to a residual group $\mathcal{G}$ of per-bond diagonal unitaries (phases) and, at degeneracies of the canonical spectrum, per-bond block rotations; our sign/phase convention selects a section of $\mathcal{G}$ almost everywhere but not globally continuously. Consequently Φ has coordinate discontinuities on a measure-zero set (spectral degeneracies and sign crossings), and Euclidean distances $\|\Phi(\psi) - \Phi(\phi)\|$ are *not* gauge-invariant near that set: two states with fidelity $1-\epsilon$ can have core-coordinate distance that does not vanish with ϵ. We therefore do *not* claim the map is an isometry or that the core-space prior is canonical; we use the coordinate representation only as a *training* target, and every reported metric—the loss equation 3, all fidelities, and the SIR comparison—is computed as a gauge-invariant state functional. Prop. 6 should be read as this cautionary statement (coordinates are convenient but gauge-dependent), not as a claim of collapse; we have softened the main-text wording accordingly.

Scope of the theory. Neither proposition is a sample-complexity or measurement-efficiency theorem, and we do not present one: our efficiency claims are entirely empirical and are stated only relative to the prior-only and shuffled controls. The theory serves two limited purposes—to justify *which* local Paulis to measure (Prop. 1) and to be honest about *what the core coordinates are and are not* (Prop. 6).

AI USE STATEMENT

Parts of this paper's text, code and analysis were produced with the assistance of generative AI. The authors take full responsibility for the final content of this paper, including any text produced with the assistance of generative AI.