Supplementary Information: Integrating Neural Networks with a Quantum Simulator for State Reconstruction

Giacomo Torlai,[1, 2, 3, *] Brian Timar,[4, *] Evert P.L. van Nieuwenburg,[4] Harry Levine,[5] Ahmed Omran,[5] Alexander Keesling,[5] Hannes Bernien,[6] Markus Greiner,[5] Vladan Vuletić,[7] Mikhail D. Lukin,[5] Roger G. Melko,[2, 3] and Manuel Endres[4]

[1] *Center for Computational Quantum Physics, Flatiron Institute, New York, New York 10010, USA*
[2] *Department of Physics and Astronomy, University of Waterloo, Ontario N2L 3G1, Canada*
[3] *Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada*
[4] *Division of Physics, Mathematics and Astronomy, California Institute of Technology, Pasadena, CA 91125, USA*
[5] *Department of Physics, Harvard University, Cambridge, MA 02138, USA*
[6] *Institute for Molecular Engineering, University of Chicago, Chicago, IL 60637, USA*
[7] *Department of Physics and Research Laboratory of Electronics,*
Massachusetts Institute of Technology, Cambridge, MA 02139, USA

In this Supplementary Information, we first provide a derivation of the approximate eight-atom ordered ground state. Next, we discuss how the unsupervised RBM learning process is carried out on experimental datasets, and demonstrate how the networks generalize from the finite datasets used in training. We also detail a regularization method used to mitigate the effect of measurement errors in the training set and provide numerical evidence that this technique significantly improves the fidelity of state reconstruction from noisy data. Finally, we examine how intrinsic decoherence processes impact the quality of the pure-state reconstruction procedure. An appendix provides proofs of two bounds regarding the fidelity and entanglement properties of reconstructions.

I. APPROXIMATE EIGHT-ATOM GROUND STATE

The full Rydberg Hamiltonian is

$$\hat{H}(\Omega,\Delta) = -\Delta\sum_{i=1}^{N}\hat{n}_i - \frac{\Omega}{2}\sum_{i=1}^{N}\hat{\sigma}_i^x + \sum_{i<j}\frac{V_{nn}}{|i-j|^6}\hat{n}_i\hat{n}_j \tag{1}$$

At the end of the experimental sweep, the Hamiltonian has a positive detuning and a small transverse field: $\Delta > 0, V_{nn} \gg \Delta \gg |\Omega|$; furthermore, interactions between sites separated by more than two lattice spacings may be neglected, as they are weak compared to the frequencies which characterize the sweep profile. In this regime the four-excitation states

$$|e_1\rangle = |r\ g\ r\ g\ g\ r\ g\ r\rangle \tag{2}$$

$$|e_2\rangle = |r\ g\ g\ r\ g\ r\ g\ r\rangle \tag{3}$$

$$|e_3\rangle = |r\ g\ r\ g\ r\ g\ g\ r\rangle \tag{4}$$

are degenerate under the classical part of the Hamiltonian $-\Delta\sum_{i=1}^{N}\hat{n}_i + \sum_{i<j}\frac{V_{nn}}{|i-j|^6}\hat{n}_i\hat{n}_j$. The ground state lies in the subspace spanned by these three states: adding or removing an excitation requires an energy penalty proportional to V_{nn} or Δ respectively. This degeneracy is lifted by a nonzero transverse field, which couples the blockaded states at second order in Ω through the three-excitation subspace. Using perturbation theory, an effective Hamiltonian [1] H_{eff} may be constructed for the blockaded subspace, whose nonzero matrix elements are given by

$$\langle e_1|H_{\text{eff}}|e_2\rangle = \langle e_1|H_{\text{eff}}|e_3\rangle = -\frac{\Omega^2}{4\Delta} \tag{5}$$

The corresponding ground state is $|\Psi\rangle = \frac{1}{\sqrt{2}}|e_1\rangle + \frac{1}{2}\left(|e_2\rangle + |e_3\rangle\right)$.

II. RECONSTRUCTION METHODS

A. Note on terminology

Below we discuss strategies for training on experimental data which has been corrupted by a fixed, known noise process. $\boldsymbol{\sigma}$ will denote the variables prior to corruption by measurement errors, while $\boldsymbol{\tau}$ will denote those which have been subjected to the noise channel – that is, for a fixed true value $\boldsymbol{\sigma}$, the noisy outputs are distributed according to $p(\boldsymbol{\tau}|\boldsymbol{\sigma})$. In our experiment, $\boldsymbol{\tau}$ are the only accessible variables, which yield the bitstrings recorded in each dataset. A model with parameters $\boldsymbol{\lambda}$ specifies a distribution $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})$ over the uncorrupted variables $\boldsymbol{\sigma}$, and a corresponding corrupted distribution $\tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{\tau}) = \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\tau}|\boldsymbol{\sigma})p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})$.

B. Standard RBM training method

The standard training method involves fitting the RBM distribution $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}) = \frac{1}{Z_{\boldsymbol{\lambda}}}\sum_{\boldsymbol{h}} e^{\,\boldsymbol{h}^\top \boldsymbol{W}\boldsymbol{\sigma} + \boldsymbol{b}\cdot\boldsymbol{\sigma} + \boldsymbol{c}\cdot\boldsymbol{h}}$ directly to the experimental datasets; in other words, it assumes a noise-free source of data:

$$p(\boldsymbol{\tau}|\boldsymbol{\sigma}) = \delta_{\boldsymbol{\tau},\boldsymbol{\sigma}} \tag{6}$$

The optimal parameters $\boldsymbol{\lambda} = \{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$ for which the RBM best reproduces the measurement data are found by minimizing the negative log-likelihood

$$\mathcal{L}_{\boldsymbol{\lambda}} = -\frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{\tau} \in \mathcal{D}} \log p_{\boldsymbol{\lambda}}(\boldsymbol{\tau}) \tag{7}$$

of the RBM distribution $p_{\boldsymbol{\lambda}}$ averaged over the dataset $\mathcal{D}$ ($|\mathcal{D}|$ denotes the size of the dataset). The gradient of the log-likelihood cost function with respect to the trainable parameters $\boldsymbol{\lambda}$ may be written

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\lambda}} = \langle \nabla_{\boldsymbol{\lambda}} \mathcal{E}_{\mathrm{eff}}(\boldsymbol{\sigma}) \rangle_{p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})} - \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{\tau} \in D} \nabla_{\boldsymbol{\lambda}} \mathcal{E}_{\mathrm{eff}}(\boldsymbol{\tau}) \tag{8}$$

where $\langle \cdot \rangle_{p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})}$ denotes the expectation value with respect to the distribution $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})$, and the effective energies

$$\mathcal{E}_{\mathrm{eff}}(\boldsymbol{\sigma}) = \boldsymbol{b} \cdot \boldsymbol{\sigma} + \sum_j \log\left(1 + e^{W_{ji}\sigma_i + c_j}\right) \tag{9}$$

are defined by $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}) = \frac{1}{Z_{\boldsymbol{\lambda}}} e^{\mathcal{E}_{\mathrm{eff}}(\boldsymbol{\sigma})}$.

The second term in the cost function gradient (8) is estimated using a batch of samples $\boldsymbol{\tau}_i$ of size M drawn from the training set $\mathcal{D}$:

$$\frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{\tau} \in D} \nabla_{\boldsymbol{\lambda}} \mathcal{E}_{\mathrm{eff}}(\boldsymbol{\tau}) \approx \frac{1}{M} \sum_{i=1}^{M} \nabla_{\boldsymbol{\lambda}} \mathcal{E}_{\mathrm{eff}}(\boldsymbol{\tau}_i) \tag{10}$$

Exact computation of the expectation value with respect to $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})$ requires summing over a number of configurations which is exponential in the system size, and is therefore not tractable. It can also be approximated by drawing M samples $\boldsymbol{\sigma}_i$ distributed according to $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})$ and using the estimator

$$\langle \nabla_{\boldsymbol{\lambda}} \mathcal{E}_{\mathrm{eff}}(\boldsymbol{\sigma}) \rangle_{p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})} \approx \frac{1}{M} \sum_{i=1}^{M} \nabla_{\boldsymbol{\lambda}} \mathcal{E}_{\mathrm{eff}}(\boldsymbol{\sigma}_i) \tag{11}$$

In principle, samples which obey the model distribution $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})$ can be generated by block Gibbs sampling [2], which involves repeatedly sampling from the conditional distributions $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}|\boldsymbol{h})$ and $p_{\boldsymbol{\lambda}}(\boldsymbol{h}|\boldsymbol{\sigma})$. Because of the *restricted* nature of the RBM graph – there are no intralayer connections – the conditional distributions factorize and each unit in a given layer can be exactly sampled simultaneously. In pseudocode, starting from a 'seed' visible state $\boldsymbol{\sigma}_1$, the Gibbs sampling algorithm is:

```
for $i$ in $[1, 2, ..., k]$ do
    Sample $\boldsymbol{h}_i$ from $p_{\boldsymbol{\lambda}}(\boldsymbol{h}|\boldsymbol{\sigma}_i)$
    Sample $\boldsymbol{\sigma}_{i+1}$ from $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}|\boldsymbol{h}_i)$
end for
```

– the output of the algorithm is the visible state $\boldsymbol{\sigma}_{k+1}$, which will obey the model distribution $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})$ for a sufficiently large number of sampling steps. In practice, the contrastive divergence algorithm [3] is applied, where the visible state is seeded with samples from the training set, and only a small number of sampling steps k is used. In practice, moderate values $k \sim 10$ are sufficient for training with stochastic gradient descent. Additional information about the RBM and its training can be found in Ref. [4]. An open-source software library for RBM reconstruction of generic wavefunctions is also available [5].

C. Noise-regularized training method

In the case where the training set is known to be corrupted by a noise process $p(\boldsymbol{\tau}|\boldsymbol{\sigma})$, our goal is to learn a model $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})$ whose corresponding *noise-corrupted* distribution $\tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{\tau})$ fits the observed data. We therefore define the corresponding log-likelihood cost function

$$\mathcal{L}_{\boldsymbol{\lambda}} = -\frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{\tau} \in \mathcal{D}} \log \tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{\tau}) \tag{12}$$

and train the network to minimize it on each dataset. The cost gradient takes a form nearly identical to that of the standard training method (8),

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\boldsymbol{\lambda}} = \langle \nabla_{\boldsymbol{\lambda}} \mathcal{E}_{\mathrm{eff}}(\boldsymbol{\sigma}) \rangle_{p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})} - \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{\tau} \in D} \langle \nabla_{\boldsymbol{\lambda}} \mathcal{E}_{\mathrm{eff}}(\boldsymbol{\sigma}) \rangle_{\tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}|\boldsymbol{\tau})} \tag{13}$$

The second term in the gradient update step is now computed not directly from the training set samples $\boldsymbol{\tau} \in \mathcal{D}$, but rather from the Bayesian posterior distribution

$$\tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}|\boldsymbol{\tau}) = \frac{p(\boldsymbol{\tau}|\boldsymbol{\sigma}) p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})}{\tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{\tau})} \tag{14}$$

which the RBM assigns to visible states $\boldsymbol{\sigma}$, given an observation $\boldsymbol{\tau}$ in the noisy training set.

This alteration to the cost gradient may be viewed as a regularization of the training based on prior knowledge of the sampling process. Regularization in machine learning generally refers to techniques for improving the generalization performance of a model trained on a particular data set to new datasets drawn from the 'ground truth' source. A typical regularization scheme like weight decay does not specify *a priori* how the in-sample and true distributions differ, and therefore typically requires some sort of validation process – testing the model on held-out data – to select good hyperparameters. In contrast, our regularization method is applied in a context where all accessible datasets are corrupted by the same noise process. This makes validation as a means of selecting regularization hyperparameters impossible – but if the noise process is known, this is no obstacle as there are no free hyperparameters to select.

In applying equation (13) to the unsupervised training of an RBM, both contributions to the gradient now require computation of expectation values over marginalized distributions $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})$, $\tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}|\boldsymbol{\tau})$ of the RBM, and are

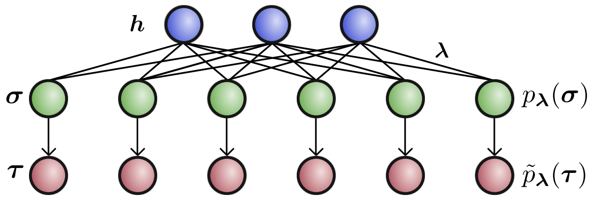
$\boldsymbol{h}$
$\boldsymbol{\lambda}$
$\boldsymbol{\sigma}$
$p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})$
$\boldsymbol{\tau}$
$\tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{\tau})$

Figure 1. **Three layer model**. Schematic for how noise-corrupted data is modeled using a three-layer graph. The upper two layers $\boldsymbol{h}, \boldsymbol{\sigma}$ constitute an RBM with trainable parameters $\boldsymbol{\lambda}$, which defines a distribution $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})$ over the uncorrupted variables $\boldsymbol{\sigma}$ upon tracing out the hidden units $\boldsymbol{h}$. The corrupted distribution is obtained through the noise process $p(\boldsymbol{\tau}|\boldsymbol{\sigma})$ as $\tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{\tau})$. The noise process is indicated here by arrows which link uncorrupted and corrupted variables at each site.

therefore intractable to compute exactly. As in the noise-free training case, this problem may be circumvented using the contrastive divergence method: the first term is estimated by repeated sampling from the conditional distributions $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}|\boldsymbol{h}), p_{\boldsymbol{\lambda}}(\boldsymbol{h}|\boldsymbol{\sigma})$, while the second uses the same alternating sampling from the 'data-clamped' distributions $\tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}|\boldsymbol{h},\boldsymbol{\tau}), \tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{h}|\boldsymbol{\sigma},\boldsymbol{\tau}) = p_{\boldsymbol{\lambda}}(\boldsymbol{h}|\boldsymbol{\sigma})$. As noted above, $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}|\boldsymbol{h}), p_{\boldsymbol{\lambda}}(\boldsymbol{h}|\boldsymbol{\sigma})$ are both efficiently computable due to the restricted structure of the RBM layers $\boldsymbol{\sigma}, \boldsymbol{h}$. Similarly, $\tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}|\boldsymbol{h},\boldsymbol{\tau})$ is efficiently computable if the error probabilities satisfy a weaker condition, namely factorizing over the uncorrupted variables:

$$p(\boldsymbol{\tau}|\boldsymbol{\sigma}) = \prod_i p(\boldsymbol{\tau}|\sigma_i) \tag{15}$$

In this case, the clamped distribution may be computed explicitly as

$$\begin{aligned}\tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}|\boldsymbol{\tau},\boldsymbol{h}) &= \prod_i \frac{p(\boldsymbol{\tau}|\sigma_i) p_{\boldsymbol{\lambda}}(\sigma_i|\boldsymbol{h})}{\sum_{\sigma_i'=0,1} p(\boldsymbol{\tau}|\sigma_i') p_{\boldsymbol{\lambda}}(\sigma_i'|\boldsymbol{h})} &(16)\\ &= \prod_i \tilde{p}_{\boldsymbol{\lambda}}(\sigma_i|\boldsymbol{\tau},\boldsymbol{h}), &(17)\end{aligned}$$

amenable to efficient block-Gibbs sampling.

Fig. 1 provides an intuitive way to understand the noise regularization – the corrupted variables $\boldsymbol{\tau}$ may be included as a third *noise layer* appended to the standard, two-layer RBM graph, with conditional probabilities depending on the $\boldsymbol{\sigma}$ layer only. These can be interpreted as effective biases for the noise layer, which depend on the uncorrupted variables – for example, the independent bit-flip errors used to model our Rydberg experiment may be written as

$$\begin{aligned}p(\boldsymbol{\tau}|\boldsymbol{\sigma}) &= \frac{1}{\tilde{Z}} e^{\tilde{b}_{\sigma}\cdot\boldsymbol{\sigma} + \tilde{b}_{\boldsymbol{\tau}}\cdot\boldsymbol{\tau} + \tilde{W}\boldsymbol{\sigma}\cdot\boldsymbol{\tau}}\\ \tilde{W} &= \log\frac{p(1|1)p(0|0)}{p(1|0)p(0|1)}\\ \tilde{b}_{\sigma,i} &= \log\frac{p(0|1)}{p(0|0)}\\ \tilde{b}_{\tau,i} &= \log\frac{p(1|0)}{p(0|0)}\end{aligned}$$

For brevity, we will sometimes refer to RBMs trained with this regularization as 'three-layer' machines, as opposed to their 'two-layer' counterparts trained in the standard fashion. Similar graphical models known as Deep Belief Nets [6] have previously been used for unsupervised learning tasks, but with a different, layer-wise training algorithm that does not incorporate prior information; a gated RBM architecture similar to the three-layer machine has also been applied to Gaussian noise models in occluded images [7].

D. Sampling from trained RBMs

After an RBM has been trained, new configurations of the uncorrupted variables $\{\boldsymbol{\sigma}\}$ can be drawn from the distribution $p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})$ using the block-Gibbs sampling techniques discussed above. The expectation value of a generic observable $\hat{\mathcal{O}}$ in the state $\psi_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}) = \sqrt{p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})}$ can then be approximated with a Monte Carlo average over n_{mc} samples:

$$\begin{aligned}\langle\hat{\mathcal{O}}\rangle_{\psi_{\boldsymbol{\lambda}}} &= \sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}'} \psi_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}) \langle\boldsymbol{\sigma}|\hat{\mathcal{O}}|\boldsymbol{\sigma}'\rangle \psi_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}') &(18)\\ &= \sum_{\boldsymbol{\sigma}} |\psi_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})|^2 \sum_{\boldsymbol{\sigma}'} \langle\boldsymbol{\sigma}|\hat{\mathcal{O}}|\boldsymbol{\sigma}'\rangle \frac{\psi_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}')}{\psi_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})} &(19)\\ &:= \langle\mathcal{O}_L(\boldsymbol{\sigma})\rangle_{p_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})} &(20)\\ &\simeq n_{\mathrm{mc}}^{-1} \sum_{k=1}^{n_{\mathrm{mc}}} \mathcal{O}_L(\boldsymbol{\sigma}_k) &(21)\end{aligned}$$

where the "local estimate" of the observable is defined to be $\mathcal{O}_L(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma}'} \langle\boldsymbol{\sigma}|\hat{\mathcal{O}}|\boldsymbol{\sigma}'\rangle \frac{\psi_{\boldsymbol{\lambda}}(\boldsymbol{\sigma}')}{\psi_{\boldsymbol{\lambda}}(\boldsymbol{\sigma})}$. In the case of nontrivial noise processes, to sample from the corrupted distribution $\tilde{p}_{\boldsymbol{\lambda}}(\boldsymbol{\tau})$ one may first generate an uncorrupted batch $\{\boldsymbol{\sigma}\}$ of data and then sample once from the conditional distribution $p(\boldsymbol{\tau}|\boldsymbol{\sigma})$ for each uncorrupted configuration.

For an RBM with N visible and N_h hidden units, the times for training and Monte Carlo observable estimation scale as $\mathcal{O}(NN_h)$, or in terms of the model complexity $\alpha = N_h/N$, as $\mathcal{O}(\alpha N^2)$; note that the number of visible units is fixed by the system size. A universal approximation theorem [8] guarantees that RBMs can represent any distribution over binary variables, although an exponentially large number of hidden units may be required in general. However, many quantum states relevant to experiment, such as ground states of paradigmatic Hamiltonians and some matrix product states, have been found to admit efficient descriptions [9–13]. In the present work with eight atoms, the Hilbert space is small enough that all amplitudes and expectation values can be computed exactly, providing a valuable check on our procedure. Such a benchmark quickly becomes impossible with current classical hardware when the number of atoms approaches ~ 20 for pure states, and at even smaller chain lengths for the exact evaluation of non-pure states.

III. TRAINING DETAILS

A. Methods

The reconstructions presented in this work were trained using the three-layer scheme detailed above on experimental datasets of $N \approx 3000$ samples each. Training was performed using stochastic gradient descent with a decayed learning rate, the gradients being estimated via contrastive divergence with $k = 30$ sampling steps. Since the visible layers of our machines are relatively small, exact computation of the negative-log-likelihood was possible on each set. Hyperparameters for training were therefore selected by cross-validation on a randomly chosen experimental set; the same hyperparameters were used in training on all datasets. The reconstructions presented in the text were trained on the full datasets; RBMs were also trained on 90/10 splits of each dataset in order to verify that the out-of-sample negative log likelihood did not grow during training. Error bars on reconstructed observables were computed from their variation across these training subsets in the final epochs of training. We found it beneficial to train each machine with the error rates set to zero for the first epoch.

To check that the networks learned a consistent representation of the experimental data, we performed a scaling analysis of the number of hidden units N_h of the RBM when training on experimental data. Increasing the number of hidden units, we found convergence of the observables and log-likelihood for $N_h \sim N$ (see Fig. 2 for examples). The reconstructions presented in this work used RBMs with $N_h = 2N = 16$.

B. Training on larger systems

As a test of the robustness of our reconstruction procedure, we also trained RBMs on a second set of Rydberg atom data sampled from a larger chain of $N = 9$ atoms. The dynamics of this system is governed by a master equation identical in structure to that used for modeling the eight-atom data presented in the main text, but with slightly different detuning and Rabi frequency profiles, and different effective decoherence rates.

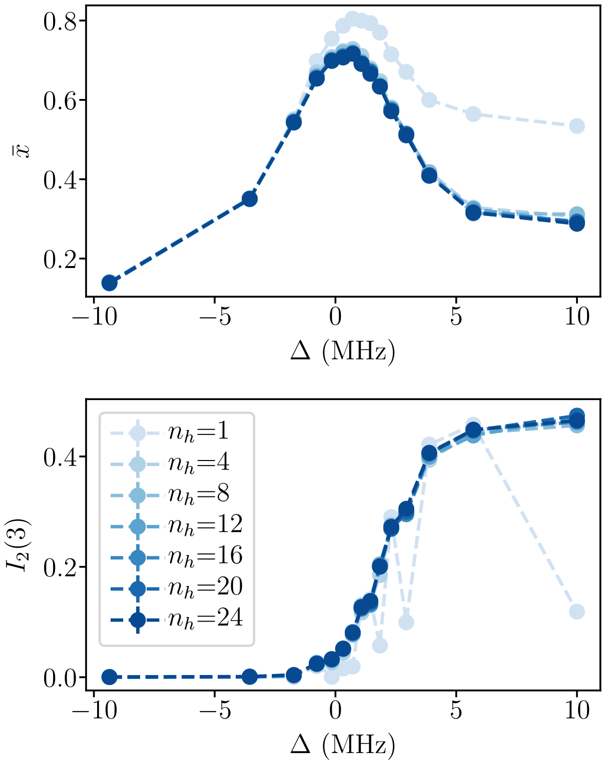

Figure 2. Examples of the scaling of observables with hidden layer size, for RBMs trained on experimental data. Top: spatially averaged transverse field values. Bottom: the Renyi mutual information at bond $s = 3$. Error bars are defined by variation of reconstructed observables in the final epochs of training.

Fig. 3 compares the results of this reconstruction to predictions of the relevant Lindbladian model, as well as experimental values where appropriate. Without alteration of the training procedure, the RBMs reconstruct quantum dynamics, as manifested in the transverse field and mutual information, in good agreement with Lindbladian predictions. This is a key benefit conferred to the experimentalist by the RBM reconstruction method. Indeed, given previous knowledge regarding the properties of the quantum state prepared in the experiment, RBM reconstruction of experimentally inaccessible observables allows for rapid and inexpensive detection of errors in state preparation and manipulation.

IV. GENERALIZATION CAPABILITIES

A generative model is of little use if it merely mimics the statistics of the training set. Successful machine learning applications are built upon the ability to *general-*

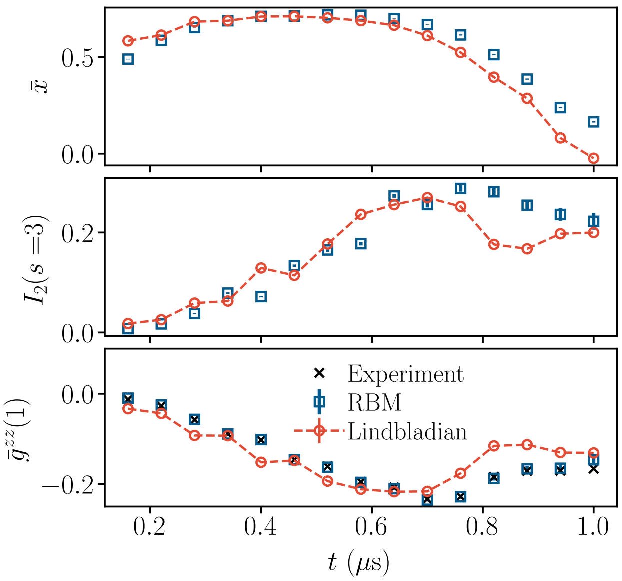

Figure 3. Some examples of observables reconstructed from nine-atom data, plotted as a function of sweep time t. From top to bottom: average transverse field $\bar{x}$, Renyi mutual information I_2 corresponding to a partition at bond $s = 3$; averaged nearest-neighbor correlations in the measurement basis (including same noise model as in the main text). The machines were trained with the same hyperparameters as in the eight-atom case, using $N_h = 2N = 18$ hidden units.

ize from a given dataset, extracting representations of the data that capture relevant features of the ground truth distribution from which it was sampled. This requires some structure in the data for the machine to learn, and the extent to which it succeeds in doing so depends upon the architecture of the machine as well as the size of the dataset.

For relatively small datasets such as those used in this work, it is natural to wonder whether the apparatus of machine learning is necessary at all. In particular, given access to the frequency distribution (FD)

$$P_{\text{FD}}(\boldsymbol{\tau}) = \frac{1}{N_s} \sum_{\boldsymbol{\tau}_i \in \mathcal{D}} \delta_{\boldsymbol{\tau}, \boldsymbol{\tau}_i} \tag{22}$$

defined by a particular dataset $\mathcal{D}$ consisting of N_s samples, one may define a naive *frequency distribution reconstruction* of a pure state corresponding to the data, which simply memorizes the training set:

$$|\Psi\rangle = \sum_{\boldsymbol{\tau}} \sqrt{P_{\text{FD}}(\boldsymbol{\tau})} |\boldsymbol{\tau}\rangle \tag{23}$$

The FD state model can be computed and stored in a time linear in the size of the dataset, by building a lookup table that associates each observed bitstring $\boldsymbol{\tau}$ with its empirical probability in the dataset, and assigning probability zero to all other bitstrings. Such a model may then be used to produce Monte-Carlo estimates of desired observables, in the same fashion as for RBM states.

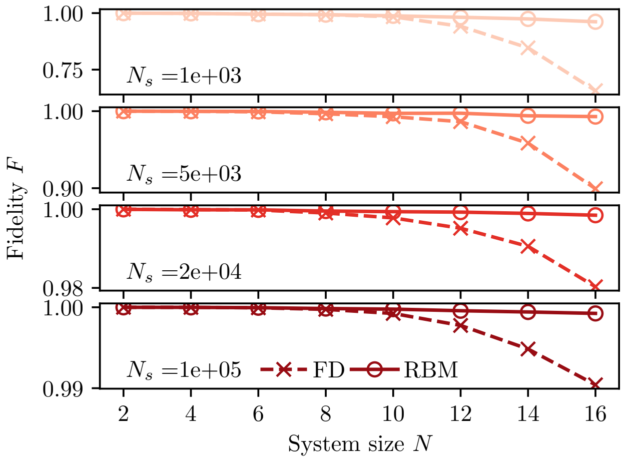

Figure 4. Generalization from ground-state datasets: fidelity improvements conferred by RBMs over frequency-distribution reconstructions, for a selection of dataset sizes N_s. Note the change in scale.

In general, the FD reconstruction approach cannot scale to high-entropy distributions – if H_2 is the second-order Renyi entropy of the ground-truth distribution $P_{\text{GT}}(\boldsymbol{\tau})$, the fidelity $F\left(P_{\text{FD}}, P_{\text{GT}}\right) = \sum_{\boldsymbol{\tau}} \sqrt{P_{\text{FD}}(\boldsymbol{\tau}) P_{\text{GT}}(\boldsymbol{\tau})}$ between the frequency distribution and the ground truth obeys the inequality

$$F\left(P_{\text{FD}}, P_{\text{GT}}\right) \leq \sqrt{N_s} e^{-H_2/4} \tag{24}$$

– for a proof, see Section (VII). In particular, if the measurement-basis entropy is proportional to the system size – as is the case in even some very simple states, such as a product state of spins not aligned with the measurement basis – the frequency-distribution fidelity will decay exponentially in system size. The ability to extract a modest number of *physically relevant features* is therefore essential for accurate state reconstruction from generic datasets of realistic size. However, our eight-atom system is small enough compared to the size of the datasets ($\sqrt{N_s} \sim 2^L$) that the FD approach is not *a priori* infeasible.

To quantify the performance of RBM and FD reconstructions in the small-system regime, we sampled synthetic datasets (in the occupation number basis) of size N_s up to 10^5 from ground states of the Rydberg Hamiltonian in equation (1), for a selection of system sizes up to $N = 16$ atoms. Ground state wavefunctions were computed using the QuSpin exact diagonalization package [14]; the Hamiltonian parameters were constant throughout and chosen to place the system near the phase transition into $\mathbb{Z}_2$ state: $V_{nn} = 30$MHz, $\Omega = 2$MHz, $\Delta \approx 1$MHz. For each dataset, we computed the fidelity

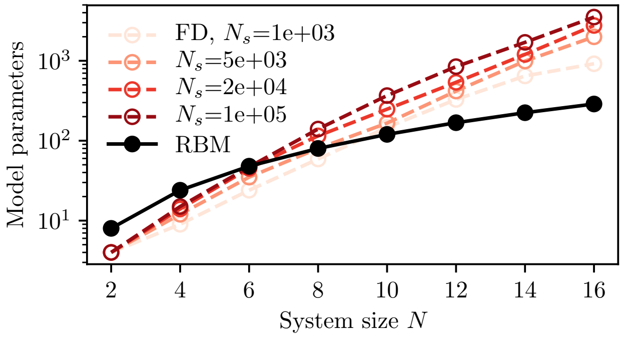

Figure 5. Dependence of model size on physical system size (note the log scale). The solid line indicates the number of parameters required to specify an RBM model with $N_h = N$. The dashed lines indicate the number of parameters required to build a lookup table for the FD model, for various dataset sizes N_s.

F_{FD} of the frequency distribution state onto the ground-truth Rydberg wavefunction; an RBM with $N_h = N$ hidden units was then trained on the same dataset, and its fidelity F_{RBM} onto the true state was also recorded. The RBMs were all trained with the hyperparameters described in section III, but with $k = 10$ contrastive divergence steps. Fig. 4 plots the resulting fidelities achieved by both reconstructions as a function of system size – RBMs of fixed complexity achieve significantly higher fidelities for large systems, with small improvements even at $N = 8$.

Another issue of practical relevance is model size: given a dataset $\mathcal{D}$ of a particular size N_s, how many parameters are required to store each trained model? For the RBM, the number of (real-valued) parameters required to specify the model completely is determined by the size of the bias vectors and weight matrix, $N \cdot N_h + N + N_h$, and therefore quadratic in the system size for a fixed model complexity N_h/N. For the FD model, the number of parameters is determined by the size of the lookup table, i.e. the number of unique samples present in the dataset, and therefore bounded above by the dataset size N_s.

In Fig. 5, the model sizes of the RBM and FD reconstructions from Fig. 4 are compared as a function of system size; for $N \gtrsim 8$ atoms the RBMs are a significantly more efficient (not to mention more accurate) description of the quantum state.

Finally, we note that even for small systems, generative models provide an additional advantage in state reconstruction from noisy data: in the presence of measurement errors, the FD model is not representative of the ground truth for any dataset size, and simply inverting the conditional probabilities will generally result in unphysical prior distributions. Denoising methods for cleaning noisy binary datasets prior to reconstruction [15–17] may be applied, but a model training step is still required.

V. EFFECTS OF DECOHERENCE

For pure state reconstruction to be useful in near-term quantum simulators, realistic decoherence processes must be accounted for. Here, we provide a brief description of the Lindbladian master equation used in our modeling of the experiment, and discuss means of assessing the quality of pure state reconstructions in the presence of decoherence.

A. Master equation for the Rydberg machine

To account for decoherence processes quantitatively, we have used a Lindblad model, described in detail in Ref. [18], which includes two jump operators $\tilde{\sigma}_i^{rg} = |g\rangle\langle r|, \tilde{\sigma}_i^{gg} = |g\rangle\langle g|$ to represent decay and dephasing processes acting on atom i. The time evolution of the full state is given by the master equation

$$\begin{aligned}\frac{d\hat{\rho}}{dt} &= -i[\hat{H}(\Omega(t), \Delta(t)) + \hat{H}_{dis}, \hat{\rho}] \\ &+ \sum_{i=1}^{N} \sum_{t=rg,gg} \gamma_t \left(\tilde{\sigma}_i^t \, \hat{\rho} \, \tilde{\sigma}_i^{t\dagger} - \frac{1}{2} \left\{ \tilde{\sigma}_i^{t\dagger} \tilde{\sigma}_i^t, \hat{\rho} \right\} \right)\end{aligned} \tag{25}$$

where $\hat{H}_{dis} = -\sum_{i=1}^{N} \delta_i \hat{n}_i$ is the static disorder Hamiltonian containing the Doppler shifts δ_i, and γ_t, $t = rg, gg$ are decoherence rates estimated from single-atom measurements [18] as $1/\gamma_{rg}$ =80μs, $1/\gamma_{gg}$ =40μs respectively. The Doppler shifts δ_i were assumed to be Gaussian-distributed with an rms width of $2\pi \cdot 43.5$kHz. Direct spontaneous decay processes from the Rydberg states, which occur over longer timescales, were neglected. Numerical solutions of the master equation (25) were performed using QuTiP [19], and observables were averaged over 100 disorder realizations $\{\delta_i\}$. Uncertainties in observables were computed from the standard error of the mean of these realizations. We note that the experiment has additional loss mechanisms, as well as imperfections in the laser sweep profile, which are not well characterized and not included in this Lindbladian model. We believe this accounts for the discrepancy with experimental correlation functions noted in the main text.

This master equation predicts a substantial loss in purity Tr $\left[\hat{\rho}^2\right]$ for states produced at the end of the sweep (Fig. 6), whose detrimental effects on our pure-state reconstruction process we quantify below.

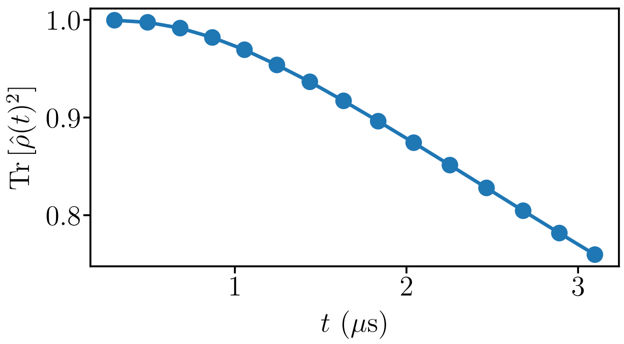

Figure 6. Purity of the master equation solutions as a function of sweep time t.

B. Reconstruction fidelities

To assess the quality of quantum state reconstruction, we consider the fidelity between two states $\hat{\rho}, \hat{\sigma}$,

$$F(\hat{\rho}, \hat{\sigma}) = \mathrm{Tr}\left[\sqrt{\sqrt{\hat{\rho}}\hat{\sigma}\sqrt{\hat{\rho}}}\right] \tag{26}$$

which reduces to the norm of the overlap in the case where $\hat{\rho}, \hat{\sigma}$ are pure states. An ideal state reconstruction $\hat{\sigma}$ of a mixed state $\hat{\rho}$ would yield $F(\hat{\rho}, \hat{\sigma}) = 1$. For pure state reconstructions $\hat{\sigma} = |\psi_{\boldsymbol{\lambda}}\rangle\langle\psi_{\boldsymbol{\lambda}}|$, this is not possible if the true state $\hat{\rho}$ is non-pure. However, one may still seek an approximate reconstruction which reproduces the local reduced density operators of $\hat{\rho}$. In particular, specializing to the case of one-dimensional systems, we can consider contiguous subsystems formed from s adjacent sites, $A_i^{(s)} = \{i, i+1, ..., i+s-1\}$. Given two density operators $\hat{\rho}, \hat{\sigma}$ for the global system $\mathcal{S}$ of size N, the reduced density operators which describe the subsystem in each state are obtained by tracing out the rest of the chain,

$$\hat{\rho}_i^{(s)} = \mathrm{Tr}_{\mathcal{S}/A_i^{(s)}}\left[\hat{\rho}\right]$$
$$\hat{\sigma}_i^{(s)} = \mathrm{Tr}_{\mathcal{S}/A_i^{(s)}}\left[\hat{\sigma}\right]$$

Then we define a *subsystem averaged fidelity* as the spatial average of the fidelity between these local operators, over all subsystems of a particular size s:

$$F_s\left(\hat{\rho}, \hat{\sigma}\right) = \frac{1}{N+s-1}\sum_{i=1}^{N-s+1} F\left(\hat{\rho}_i^{(s)}, \hat{\sigma}_i^{(s)}\right) \tag{27}$$

$F_s\left(\hat{\rho}, \hat{\sigma}\right)$ is a measure of how well, on average, $\hat{\sigma}$ is able to reproduce the s-local physics of $\hat{\rho}$.

To examine the quality of the RBM states $\hat{\rho}_{\boldsymbol{\lambda}} = |\psi_{\boldsymbol{\lambda}}\rangle\langle\psi_{\boldsymbol{\lambda}}|$ in reproducing local density operators, we solved the master equation (25) for set of decay rates $\gamma_{rg} = \alpha\gamma_{rg}^{\mathrm{exp}}, \gamma_{gg} = \alpha\gamma_{gg}^{\mathrm{exp}}$, with $\gamma_{rg}^{\mathrm{exp}}, \gamma_{gg}^{\mathrm{exp}}$ denoting our estimates of the experimental values, and α a dimensionless parameterization of the overall decoherence strength. For each set of decoherence rates, the master equation was solved and synthetic data sampled from the resulting mixed states. Pure state RBMs were trained on each of these datasets, and the resulting averaged fidelities $F_s\left(\hat{\rho}, \hat{\rho}_{\boldsymbol{\lambda}}\right)$ were computed.

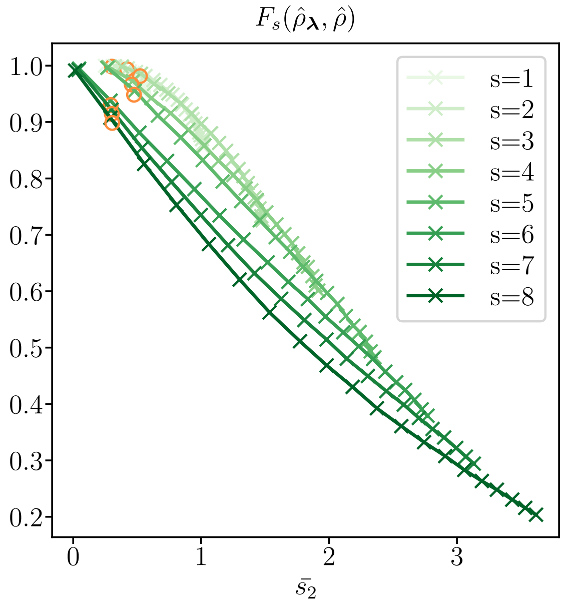

Figure 7. **Average subsystem fidelities.** For each subystem size s, the average subsystem fidelity between the reconstructed state $\hat{\rho}_{\boldsymbol{\lambda}}$ and the state $\hat{\rho}$ from which its training data was sampled is plotted, for varying values of the decoherence rates, as quantified by the average Renyi entropy $\bar{s_2} = -\frac{1}{N+s-1}\sum_{i=1}^{N-s+1}\log\mathrm{Tr}\left(\hat{\rho}_i^{(s)2}\right)$ of the local reduced density operators. The data plotted are for states taken at the end of the sweep, at $\Delta = 10$MHz. Open circles indicate the fidelities obtained using the decoherence rates from the experimental model presented in the main text. The fidelity behavior at other points in the sweep (not shown) is qualitatively similar.

As a representative example, Fig. 7 shows how the fidelities computed in the final state of the sweep vary as a function of the average Renyi entropy $\bar{s_2}$ of subsystems of a given size – one observes a roughly linear decay in the average fidelity with the averaged entropy. These numerical results suggest that pure state reconstruction techniques should focus on few-body operators, where the entropy build-up due to global decoherence process is limited in proportion to the system size.

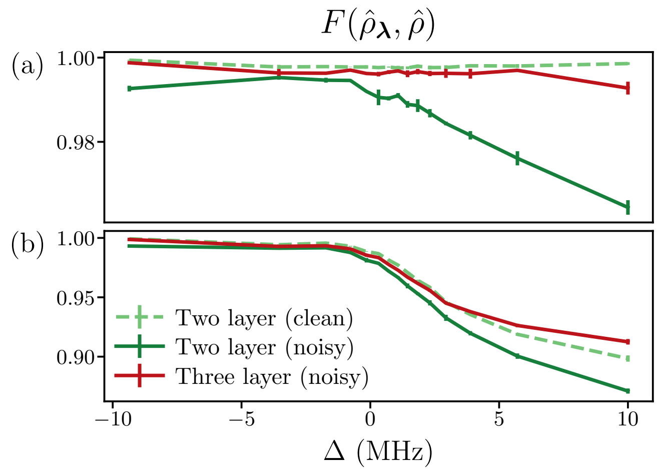

Figure 8. **Fidelity improvements from noise layer regularization**. As a demonstration of the efficacy of noise layer regularization, we plot the fidelity $F(\hat{\rho}_{\boldsymbol{\lambda}}, \hat{\rho})$ obtained between the underlying state $\hat{\rho}$ and the reconstruction $\hat{\rho}_{\boldsymbol{\lambda}} = |\psi_{\boldsymbol{\lambda}}\rangle\langle\psi_{\boldsymbol{\lambda}}|$, when training on synthetic data subjected to measurement errors ('noisy' data), as a function of detuning Δ. We compare regularized training (red solid lines, 'Three Layer') with unregularized training (green solid lines, 'Two Layer'), for (a) Data sampled from pure, positive Rydberg ground states, and (b) Data sampled from the mixed states $\hat{\rho}$ predicted by our Lindbladian model. As a benchmark we plot in each case the fidelity obtained by a two-layer RBM training on 'clean' data without measurement errors (green dashed lines). The regularized training leads to higher fidelities for all states sampled. For some mixed states, it even exceeds the RBM trained on clean data. This is because the pure state model is no longer valid when the source state is mixed, and so the 'optimal' pure state as defined by fidelity is not necessarily the one which best fits the training set.

VI. RECONSTRUCTION IMPROVEMENT FROM NOISE LAYER REGULARIZATION

Numerical experiments have demonstrated that noise layer regularization results in higher-fidelity pure state reconstruction when training on uniformly noisy data.

Fig. 8 compares fidelities achieved by regularized and unregularized RBMs, when trained on synthetic datasets subjected to the bitflip error channel described in the main text. The improvement is significant, especially in the ordered phase, where global state purity is lowest. It is important to note that in these experiments the noise process is known ahead of time and built into the three layer networks as in Fig. 1. We have also trained three-layer machines using incorrect values of the error rates on the same noisy synthetic data. Although the fidelity performance varied somewhat, depending most sensitively on $p(0|1)$, the quality of the regularized reconstructions is generally robust, and deep in the ordered phase all three-layer machines exhibited higher fidelities than their two layer counterparts on the corresponding datasets, for error rates with bounds set by single-atom measurements [18]. Generically, of course, a sufficiently large mismatch between the true and assumed error rates will lead to decreased reconstruction fidelity. Future work will investigate more generally the task of selecting a regularization method for noisy quantum data.

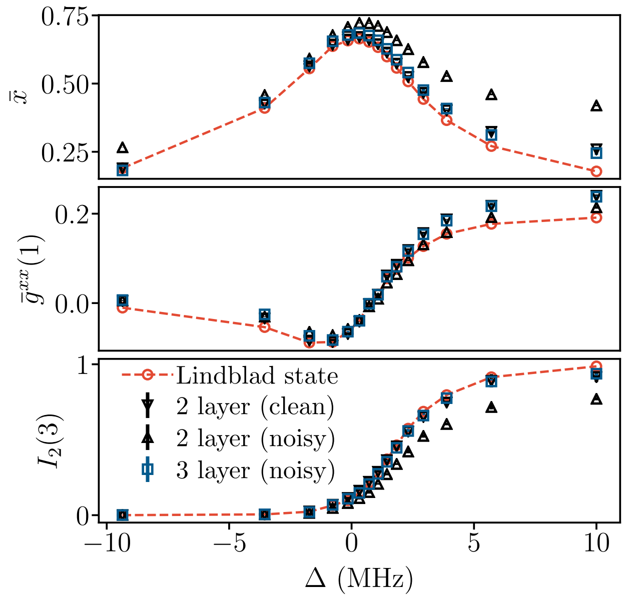

Figure 9. **Observable reconstructions from synthetic data**. A comparison of two- and three-layer reconstructions of the Lindbladian state when subjected to measurement errors. From top to bottom: average transverse field, average nearest-neighbor XX correlation, and Renyi mutual information at bond 3. 'noisy' ('clean') indicates training data with (without) measurement errors. Note the close agreement between the three layer machines trained on noisy data (blue squares) and the two-layer machines trained on clean data .

Fig. 9 compares the predictions of these synthetically trained two- and three-layer machines (using the known noise values) for some of the observables discussed in the main text. We find that noise-layer training allows the RBMs to provide much tighter agreement in, for example, values of the transverse field and mutual information. Surprisingly, the three-layer machines actually produce poorer estimates of the transverse field correlator in the ordered phase, despite yielding two-body density operators with higher fidelities for all sampled states. A more detailed analysis of the ordered phase states reveals that regularized training does indeed produce better estimates of the one- and two-body expectation values $\langle\sigma_i^x\rangle$, $\langle\hat{\sigma}_i^x\hat{\sigma}_{i+1}^x\rangle$ for all sites i. However, at bonds (3, 4) and (5, 6), where quantum fluctuations are strongest, the three-layer improvement in the two-body expectation value is relatively small, while the reduction in the one-body expectation value is substantial. Upon computing the connected correlator $\langle\hat{\sigma}_i^x\hat{\sigma}_{i+1}^x\rangle - \langle\sigma_i^x\rangle\langle\sigma_{i+1}^x\rangle$, the overall effect is an overestimate of the true correlation.

VII. APPENDIX: PROOF OF CLASSICAL FIDELITY BOUND

Inequality (24) is obtained by bounding the probability of the most likely outcome using the Renyi entropy H_2 in the measurement basis. By definition, $H_2 = -\log \sum_{\boldsymbol{\tau}} P_{\mathrm{GT}}(\boldsymbol{\tau})^2$, and $\sum_{\boldsymbol{\tau}} P_{\mathrm{GT}}(\boldsymbol{\tau})^2 \geq \max_{\boldsymbol{\tau}} P_{\mathrm{GT}}(\boldsymbol{\tau})^2$, so $-\log \sum_{\boldsymbol{\tau}} P_{\mathrm{GT}}(\boldsymbol{\tau})^2 \leq -2 \log \max_{\boldsymbol{\tau}} P_{\mathrm{GT}}(\boldsymbol{\tau})$. Rearranging, $\max_{\boldsymbol{\tau}} P_{\mathrm{GT}}(\boldsymbol{\tau}) \leq e^{-H_2/2}$. In particular, this bounds the probability of any event in the training set, so

$$\begin{aligned} F(P_{\mathrm{FD}}, P_{\mathrm{GT}}) &= \sum_{\boldsymbol{\tau}} \sqrt{P_{\mathrm{FD}}(\boldsymbol{\tau}) P_{\mathrm{GT}}(\boldsymbol{\tau})} \\ &\leq \sum_{\boldsymbol{\tau}} \sqrt{P_{\mathrm{FD}}(\boldsymbol{\tau}) e^{-H_2/2}} \\ &= \sqrt{N_s} e^{-H_2/4} \end{aligned}$$

VIII. APPENDIX: RENYI ENTROPY BOUND FROM POSITIVE PURE STATES

The nth order Renyi entropy of a quantum state $\hat{\rho}$ is defined as $S_n\left[\hat{\rho}\right] = \frac{1}{1-n} \log \mathrm{Tr} \hat{\rho}^n$.

Consider a system S partitioned into subsets A and B, and a density operator $\hat{\rho}$ defined on S; its reduced density operator in the A subsystem is $\hat{\rho}_A = \mathrm{Tr}_B \hat{\rho}$. Let $|i\rangle$, $|j\rangle$ denote orthonormal bases for A, B respectively, so that the set of product states $|i,j\rangle$ forms an orthonormal basis for the full system S. Let $p_{i,j}$ be the probability assigned by $\hat{\rho}$ to the measurement outcome i,j: $p_{i,j} = \mathrm{Tr}\left(\hat{\rho} |i,j\rangle\langle i,j|\right)$. The positive-pure partner to the mixed state is defined as

$$|\Psi^P[\rho]\rangle = \sum_{i,j} \sqrt{p_{i,j}} |i,j\rangle, \tag{28}$$

and the corresponding reduced density operator on A is $\hat{\rho}_A^P = \mathrm{Tr}_B |\Psi^P[\rho]\rangle\langle\Psi^P[\rho]|$.

Theorem: For $n > 1$, the Renyi entropies S_n of the two density operators satisfy the inequality

$$S_n\left[\hat{\rho}_A^P\right] \leq S_n\left[\hat{\rho}_A\right] \tag{29}$$

As a consequence, in the case of pure states $\hat{\rho}$, where the global Renyi entropy vanishes, the positive-pure partner provides a lower bound on the mutual information:

$$\begin{aligned} I_n\left[\hat{\rho}^P\right] &= S_n\left[\hat{\rho}_A^P\right] + S_n\left[\hat{\rho}_B^P\right] && (30) \\ &\leq S_n\left[\hat{\rho}_A\right] + S_n\left[\hat{\rho}_B\right] && (31) \\ &= I_n\left[\hat{\rho}\right] && (32) \end{aligned}$$

We note that for the case of the $n=2$ Renyi entropy and pure states $\hat{\rho}$, this result has been obtained in previous work [20, 21].

Proof: Choose an auxiliary system R to purify $\hat{\rho}$: $\hat{\rho} = \mathrm{Tr}_R |\Psi\rangle\langle\Psi|$ for some pure state $|\Psi\rangle$ living in $S \otimes R$. If $|\alpha\rangle$ is an orthonormal basis for R, we can expand the larger pure state in the joint basis $|i,j,\alpha\rangle$: $|\Psi\rangle = \sum_{i,j,\alpha} \Psi^{\alpha}_{i,j} |i,j,\alpha\rangle$ for some complex coefficients $\Psi^{\alpha}_{i,j}$.

In terms of these amplitudes, the reduced density operator of the mixed state on A is

$$\hat{\rho}_A = \sum_{\alpha,j} \Psi^{\alpha}_{i,j} \Psi^{\alpha *}_{i',j} |i\rangle\langle i'| \tag{33}$$

and so

$$\mathrm{Tr} \hat{\rho}_A^n = \left(\Psi^{\alpha_1}_{i_1,j_1} \Psi^{\alpha_1 *}_{i_2,j_1}\right) \left(\Psi^{\alpha_2}_{i_2,j_2} \Psi^{\alpha_2 *}_{i_3,j_2}\right) \ldots \left(\Psi^{\alpha_n}_{i_n,j_n} \Psi^{\alpha_n *}_{i_1,j_n}\right) \tag{34}$$

with summation over all indices implied. The reduced density operator for positive-pure partner may be obtained from the definition above:

$$\hat{\rho}_A^P = \sum_{i,i',j} \sqrt{p_{i,j} p_{i',j}} |i\rangle\langle i'| \tag{35}$$

whence

$$\mathrm{Tr}\left(\hat{\rho}_A^P\right)^n = \left(\sqrt{p_{i_1,j_1} p_{i_2,j_1}}\right) \left(\sqrt{p_{i_2,j_2} p_{i_3,j_2}}\right) \ldots \left(\sqrt{p_{i_n,j_n} p_{i_1,j_n}}\right) \tag{36}$$

(summation implied). Furthermore,

$$p_{i,j} = \sum_{\alpha} \Psi^{\alpha}_{i,j} \Psi^{\alpha *}_{i,j} \tag{37}$$

and so by the Cauchy-Schwartz inequality,

$$\begin{aligned} \left| \sum_{\alpha} \Psi^{\alpha}_{i,j} \Psi^{\alpha *}_{i',j'} \right| &\leq \sqrt{\left(\sum_{\alpha} \Psi^{\alpha}_{i,j} \Psi^{\alpha *}_{i,j}\right) \left(\sum_{\alpha'} \Psi^{\alpha'}_{i',j'} \Psi^{\alpha' *}_{i',j'}\right)} && (38) \\ &= \sqrt{p_{i,j} p_{i',j'}} && (39) \end{aligned}$$

Therefore, writing $\mathbf{i} = (i_1, ..., i_n), \mathbf{j} = (j_1, ..., j_n)$,

$$\begin{aligned} \mathrm{Tr} \hat{\rho}_A^n &= \sum_{\mathbf{i},\mathbf{j}} \left(\sum_{\alpha_1} \Psi^{\alpha_1}_{i_1,j_1} \Psi^{\alpha_1 *}_{i_2,j_1}\right) \left(\sum_{\alpha_2} \Psi^{\alpha_2}_{i_2,j_2} \Psi^{\alpha_2 *}_{i_3,j_2}\right) \\ &\quad \ldots \left(\sum_{\alpha_n} \Psi^{\alpha_n}_{i_n,j_n} \Psi^{\alpha_n *}_{i_1,j_n}\right) && (40) \\ &\leq \sum_{\mathbf{i},\mathbf{j}} \left|\sum_{\alpha_1} \Psi^{\alpha_1}_{i_1,j_1} \Psi^{\alpha_1 *}_{i_2,j_1}\right| \left|\sum_{\alpha_2} \Psi^{\alpha_2}_{i_2,j_2} \Psi^{\alpha_2 *}_{i_3,j_2}\right| \\ &\quad \ldots \left|\sum_{\alpha_n} \Psi^{\alpha_n}_{i_n,j_n} \Psi^{\alpha_n *}_{i_1,j_n}\right| && (41) \\ &\leq \sum_{\mathbf{i},\mathbf{j}} \left(\sqrt{p_{i_1,j_1} p_{i_2,j_1}}\right) \left(\sqrt{p_{i_2,j_2} p_{i_3,j_2}}\right) \ldots \left(\sqrt{p_{i_n,j_n} p_{i_1,j_n}}\right) \\ & && (42) \\ &= \mathrm{Tr}\left(\hat{\rho}_A^P\right)^n && (43) \end{aligned}$$

Hence $-\log \mathrm{Tr}\left(\hat{\rho}_A^P\right)^n \leq -\log \mathrm{Tr}\left(\hat{\rho}_A\right)^n$, which means that for $n > 1$, $S_n(\hat{\rho}_A^P) \leq S_n(\hat{\rho}_A)$.

[*] These authors contributed equally.

[1] S. Sachdev, *Quantum Phase Transitions*, 2nd ed. (Cambridge University Press, 2011).

[2] G. E. Hinton, in *Neural Networks: Tricks of the Trade: Second Edition*, edited by G. Montavon, G. B. Orr, and K.-R. Müller (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012) pp. 599–619.

[3] G. E. Hinton, Neural Comput. **14**, 1771 (2002).

[4] G. Torlai and R. G. Melko, arXiv e-prints , arXiv:1905.04312 (2019), arXiv:1905.04312 [quant-ph].

[5] M. J. S. Beach, I. D. Vlugt, A. Golubeva, P. Huembeli, B. Kulchytskyy, X. Luo, R. G. Melko, E. Merali, and G. Torlai, SciPost Phys. **7**, 9 (2019).

[6] G. E. Hinton, S. Osindero, and Y.-W. Teh, Neural Computation **18**, 1527 (2006).

[7] Yichuan Tang, R. Salakhutdinov, and G. Hinton, in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2012) pp. 2264–2271.

[8] N. Le Roux and Y. Bengio, Neural Computation **20**, 1631 (2008).

[9] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Nature Physics **14**, 447 (2018).

[10] G. Torlai and R. G. Melko, Physical Review B **94**, 165134 (2016).

[11] D.-L. Deng, X. Li, and S. Das Sarma, Physical Review X **7**, 021021 (2017).

[12] I. Glasser, N. Pancotti, M. August, I. D. Rodriguez, and J. I. Cirac, Physical Review X **8**, 011006 (2018).

[13] J. Chen, S. Cheng, H. Xie, L. Wang, and T. Xiang, Physical Review B **97**, 085104 (2018).

[14] P. Weinberg and M. Bukov, SciPost Physics **2**, 003 (2017).

[15] D. M. Greig, B. T. Porteous, and A. H. Seheult, Journal of the Royal Statistical Society: Series B (Methodological) **51**, 271 (1989).

[16] S. Geman and D. Geman, IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-6**, 721 (1984).

[17] J. Batson and L. Royer, (2019), arXiv:1901.11365.

[18] H. Levine, A. Keesling, A. Omran, H. Bernien, S. Schwartz, A. S. Zibrov, M. Endres, M. Greiner, V. Vuletić, and M. D. Lukin, Physical Review Letters **121**, 123603 (2018).

[19] J. R. Johansson, P. D. Nation, and F. Nori, Computer Physics Communications **184**, 1234 (2013).

[20] Y. Zhang, T. Grover, and A. Vishwanath, Physical Review Letters **107**, 067202 (2011).

[21] T. Grover and M. P. A. Fisher, Physical Review A **92**, 042308 (2015).