Transformer neural networks and quantum simulators: a hybrid approach for simulating strongly correlated systems

Hannah Lange[1,2,3], Guillaume Bornet[4], Gabriel Emperauger[4], Cheng Chen[4], Thierry Lahaye[4], Stefan Kienle[5], Antoine Browaeys[4], and Annabelle Bohrdt[3,6]

[1]Ludwig-Maximilians-University Munich, Theresienstr. 37, Munich D-80333, Germany
[2]Max-Planck-Institute for Quantum Optics, Hans-Kopfermann-Str.1, Garching D-85748, Germany
[3]Munich Center for Quantum Science and Technology, Schellingstr. 4, Munich D-80799, Germany
[4]Université Paris-Saclay, Institut d'Optique Graduate School, CNRS, Laboratoire Charles Fabry, 91127 Palaiseau Cedex, France
[5]FORRS Partners GmbH, Happelstr. 11, 69120 Heidelberg, Germany
[6]University of Regensburg, Universitätsstr. 31, Regensburg D-93053, Germany

Owing to their great expressivity and versatility, neural networks have gained attention for simulating large two-dimensional quantum many-body systems. However, their expressivity comes with the cost of a challenging optimization due to the in general rugged and complicated loss landscape. Here, we present a hybrid optimization scheme for neural quantum states (NQS), involving a data-driven pretraining with numerical or experimental data and a second, Hamiltonian-driven optimization stage. By using both projective measurements from the computational basis as well as expectation values from other measurement configurations such as spin-spin correlations, our pretraining gives access to the sign structure of the state, yielding improved and faster convergence that is robust w.r.t. experimental imperfections and limited datasets. We apply the hybrid scheme to the ground state search for the 2D transverse field Ising model and dipolar XY model on 6×6 and 10×10 square lattices with a patched transformer wave function, using numerical data as well as experimental data from a programmable Rydberg quantum simulator [Chen et al., Nature 616 (2023)], and show that the information from a second measurement basis highly improves the performance. Our work paves the way for a reliable and efficient optimization of neural quantum states.

One of the key challenges in quantum many-body physics is the simulation of large, strongly correlated systems of more than one dimension. For these systems, exact parameterizations are unfeasible due to the exponential growth of the Hilbert space with system size. To overcome this problem, variational methods like tensor networks [1] and their efficiently contractable variant, matrix product states (MPS) [2], assume a certain form of the wave function, and this trial state is then optimized to obtain the best possible representation. However, in practice these methods face limitations, e.g. MPS are efficient parameterizations of states with area-law-entanglement [3, 4], but require an exponential number of parameters for states with volume-law entanglement.

Their great expressive power [5, 6] and their flexibility make neural networks promising candidates to overcome these limitations. Recent works have shown that neural quantum states (NQS), i.e. variational wave functions parameterized by neural network parameters $\vec{\theta}$, are appealing trial states to model a broad range of quantum many-body systems, and even states with volume-law entanglement [7–9]. In particular, transformer quantum states [10–17] have proven remarkable capabilities for ground state representations of spin systems up to system sizes of 10×10, or, using a combination of recurrent neural networks and transformers, up to 16×16.

In most cases, the ground state search with NQS is started from randomly initialized parameters $\vec{\theta}_0$. If the optimization procedure converges,

the trained NQS does not depend on $\vec{\theta}_0$. However, due to the in general very complicated loss landscape with many local minima, converging NQS often becomes very challenging [18–20]. In these cases, the convergence and the convergence time can crucially depend on $\vec{\theta}_0$. Here, we use a data-driven pretraining from a programmable Rydberg quantum simulator or, to test our procedure with ideal data, from density-matrix renormalization group (DMRG) simulations, yielding a parameterization $\vec{\theta}_0$ that approximates the training data, similar to a quantum state reconstruction [21–23]. The underlying idea is that any data that is relatively close to the target state, e.g. experimental data or (unconverged) numerical data, contains enough information about the actual system such that $\vec{\theta}_0$ is close to the ground state in the optimization landscape. After this first stage, the NQS is further optimized using a Hamiltonian-driven training procedure by employing variational Monte Carlo (VMC), see Fig. 1a. Recent works have shown that this procedure can improve the convergence for stoquastic spin systems and molecular Hamiltonians [24–26], even in the presence of experimental imperfections and finite temperatures.

In contrast to the stoquastic Hamiltonians considered in Refs. [24, 25], in general ground states have a sign structure and hence information from different measurement configurations than the computational (Z) basis is required. In particular, since optimizing the sign structure with variational Monte Carlo is in general challenging, information on the sign structure can drastically improve the performance, and is often included by hand as e.g. the Marshall sign rule for the Heisenberg model [27]. However, this is in general not possible since the sign structure is unknown. In these cases, information on the signs can be inferred from measurements outside the Z basis. Since the computational cost of such a state reconstruction away from the Z basis scales exponentially with the number of sites that are rotated outside the computational basis, in Refs. [21, 26] only snapshots with a few rotated sites are considered.

Here, we present two extensions to existing hybrid training schemes: (i) We show that an efficient way for using information from other bases than the computational basis are the local

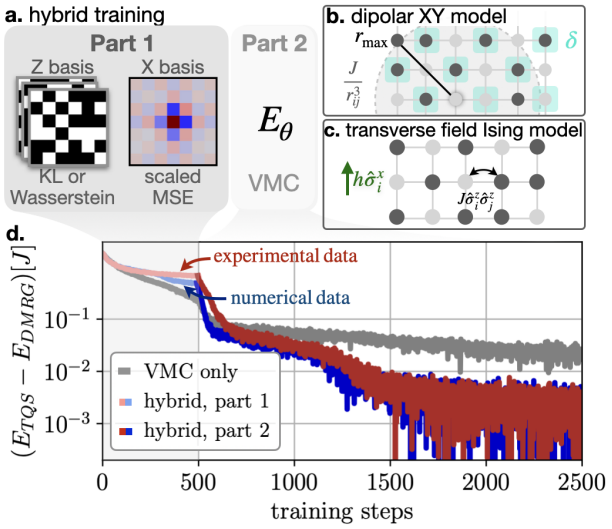

Figure 1: The hybrid training: **a.** We use a data-driven pretraining (part 1) based on the Wasserstein distance or the Kullback Leibler divergence (KL) for snapshots from the Z basis and on the scaled mean-square error (MSE) for expectation values from the X basis such as $\langle\hat{\sigma}_\mathbf{i}^x\hat{\sigma}_\mathbf{j}^x\rangle$ or $\langle\hat{\sigma}_\mathbf{i}^x\rangle$, before the Hamiltonian-driven variational Monte Carlo training (part 2). We apply this hybrid training scheme to the dipolar XY model (**b.**) with exchange interactions J and a light shift δ, and the transverse field Ising model (**c.**) with Ising interactions J and a transverse field h. **d.** The hybrid scheme applied for a transformer quantum state with pretraining on Wasserstein and MSE loss improves the optimization result compared to the only VMC based training (gray lines), here for the dipolar XY model with $\hbar\delta/J \approx 0.004$.

magnetization or correlation maps, as shown in Fig. 1a (dark gray box, right). This has the following advantages compared to training on snapshots away from the computational basis: a lower computational cost (e.g. for spin-spin correlations only two spins at a time have to be rotated away from the Z basis); the possibility of reducing experimental imperfections by e.g. explicitly employing spatial symmetries; and the applicability of experimental or numerical data from methods that do not give access to snapshots such as Determinant Quantum Monte Carlo. (ii) For the reconstruction training in the computational basis, we introduce another distance measure than the KL divergence, namely the Wasserstein distance [28] (1a, dark gray box, left), with the advantage that it allows to include the notion of a physically inspired *distance between snapshots* which is not considered when using the KL divergence.

We evaluate the performance of our hybrid training scheme on two models: the 2D transverse field Ising model (TFIM)

$$\hat{\mathcal{H}}_{\mathrm{TFIM}} = 4J \sum_{\langle ij \rangle} \hat{S}_i^z \hat{S}_j^z + 2h \sum_i \hat{S}_i^x, \qquad (1)$$

and the 2D dipolar XY model, with a staggered magnetic field on sublattice B, see Fig. 1b,

$$\hat{\mathcal{H}}_{\mathrm{XY}} = J \sum_{i<j} \frac{a^3}{r_{ij}^3} \left[\hat{S}_i^+ \hat{S}_j^- + \hat{S}_i^- \hat{S}_j^+ \right] + \hbar\delta \sum_{i \in B} \hat{n}_i^\uparrow, \qquad (2)$$

with $\hat{n}_i^\uparrow = \hat{S}_i^z + \frac{1}{2}$, the spin 1/2 operators $\hat{S}_i^z$ and $\hat{S}_i^\pm = \hat{S}_i^x \pm i\hat{S}_i^y$ and the antiferromagnetic (ferromagnetic) interaction strength $J > 0\,(J < 0)$, where we restrict the interaction radius to nearest neighbors (TFIM) and $|i - j| < 3$ (dipolar XY). The latter system was recently realized on a programmable Rydberg simulator [29], where additionally small van der Waals interactions are present, see Appendix E, which are taken into account as well.

In order to capture the long-range interactions of the dipolar XY model, we use a quantum state representation in terms of a transformer quantum state (TQS) [10–17, 30] (see Appendix A) with two separate output layers for phase $\phi_{i,\vec{\theta}}$ and amplitudes $p_{i,\vec{\theta}}$ at each site i, i.e. the total wave TQS function is given by

$$|\psi\rangle_{\vec{\theta}} = \sum_{\vec{\sigma}} \exp(i\phi_{\vec{\theta}}(\sigma)) \sqrt{p_{\vec{\theta}}(\sigma)}\, |\sigma\rangle\,, \qquad (3)$$

with $\phi_{\vec{\theta}}(\sigma) = \sum_i \phi_{i,\theta}(\sigma)$ and $p_{i,\theta}(\sigma) = \Pi_i p_{i,\vec{\theta}}(\sigma|\sigma_{<i})$ [11]. This architecture involves a so-called attention mechanism with trainable weights for all-to-all correlations between the patches. Specifically, we use the autoregressive and patched version of the TQS [31].

In the remainder of this paper we introduce the hybrid optimization scheme and test it using a TQS representation for the 2D TFIM (1) and dipolar XY models (2) on 6×6 and 10×10 lattices. In all cases, we find significantly improved results compared to an only Hamiltonian-driven training or a hybrid training with data from only a single basis. Furthermore, the TQS proves to be remarkably efficient in representing the ground states of these systems compared to MPS.

1 Hybrid training

The hybrid training procedure that we use consists of two parts, see Fig. 1: A data-driven training on external data, e.g. experimental measurements (part 1) and a Hamiltonian-driven training using variational Monte Carlo [32, 33] (part 2, see Appendix C).

Data-driven pretraining (part 1): The goal of the data-driven pretraining is to obtain a parameterization of the NQS that is already close in parameter space to the ground state, and which then serves as the initialization for the second part, namely the Hamiltonian-driven training.

For snapshots from the computational Z basis, in previous works [24,25] the Kullback-Leibler (KL) divergence between the measurement distribution q and the NQS amplitudes $p_{\vec{\theta}} = |\psi_{\vec{\theta}}|^2$ was used for the pretraining, i.e.

$$D_{\mathrm{KL}}(q|p_{\vec{\theta}}) = \sum_{i=1}^{n} q(\sigma_i^q) \log\left(\frac{q(\sigma_i^q)}{p_{\vec{\theta}}(\sigma_i^q)} \right) \qquad (4)$$

for an underlying dataset with measurements $\{\sigma_i^q\}_{i=1,\ldots,n}$. Note that D_{KL} is asymmetric, i.e. $D_{\mathrm{KL}}(q|p_{\vec{\theta}}) \neq D_{\mathrm{KL}}(p_{\vec{\theta}}|q)$.

This loss function however has no information on the physical system under consideration. In order to add this information to the pretraining, we aim to include the notion of a (1D) distance $\|\sigma - \sigma'\|_E$ between snapshots σ and σ'. Here, we choose the diagonal part of the energy $\|\sigma - \sigma'\|_E := E_{\mathrm{diag}}(\sigma, \sigma') = |\langle\sigma|\hat{\mathcal{H}}|\sigma\rangle - \langle\sigma'|\hat{\mathcal{H}}|\sigma'\rangle|$, indicating e.g. if two snapshots σ and σ' are close to each other in the sense that they are connected by symmetries $\|\sigma - \sigma'\|_E = 0$, only a few spin flips $\|\sigma - \sigma'\|_E \approx 0$ or many spin flips $\|\sigma - \sigma'\|_E > 0$, see also Appendix B. To incorporate this information into a probability distribution metric, we use the Wasserstein distance instead of the KL divergence. The Wasserstein distance measures the minimal cost required to transform one distribution into another. The cost is measured by the the probability mass, i.e. the probability assigned to specific points in the distribution, that must be moved from q (with samples $\{\sigma_i^q\}_{i=1,\ldots,n}$) to obtain $p_{\vec{\theta}}$ (with samples $\{\sigma_j^p\}_{j=1,\ldots,m}$), or vice versa. Formally, the probability mass transport is described by a matrix $\gamma \in \mathbb{R}_+^{n \times m}$, with $\sum_i \gamma_{ij} = p_{\vec{\theta}}(\sigma_j^p)$ and $\sum_j \gamma_{ij} = q(\sigma_i^q)$, i.e. γ moves all probability mass from q to $p_{\vec{\theta}}$ and vice versa. In this

notation, the Wasserstein loss is [34, 35]

$$W(q, p_{\vec{\theta}}) = \min_{\gamma} \left[\sum_{i,j} \gamma_{ij} \|\sigma_i^q - \sigma_j^p\|_E \right]. \quad (5)$$

Since there are many ways γ to move the probability mass, a optimal transport problem has to be solved for the evaluation of Eq. (5), which can be done efficiently in 1D with a computational cost of $\mathcal{O}(n\log(n))$ [35]. Note that in contrast to D_{KL}, W is symmetric.

Training on computational basis measurements gives access to the amplitudes of the state under consideration. In general, information from other measurement configurations than the computational basis is needed to get a good approximation of the state under consideration, in particular its sign structure [36–39]. However, calculating the KL or Wasserstein distance for a measurement configuration $\vec{b}$ in an arbitrary basis scales exponentially with the number of spins r that are rotated out of the Z basis since the NQS wave function $\psi_{\vec{\theta}}$ has to be rotated to $\psi'_{\vec{\theta}}(\sigma^{\vec{b}})$ in order to use Eqs. (4) and (5), see Appendix B. Here, instead of training on the measurement distribution, we calculate the local magnetization or spin-spin correlations in the rotated basis from the measured snapshots and compare it to the TQS expectation values. Specifically, we consider $\langle \hat{\sigma}^x_{\mathbf{i}} \rangle_{(\vec{\theta})}$ for the TFIM ($\langle \hat{\sigma}^x_{\mathbf{i}} \hat{\sigma}^x_{\mathbf{j}} \rangle_{(\vec{\theta})}$ for the dipolar XY model, with a fixed reference site $\mathbf{i}$ located in the middle of the system). This is done by rotating the spins at $r = 1(2)$ sites $\mathbf{i}$ (and $\mathbf{j}$ for the correlations), i.e. with the same (with a smaller or the same) computational cost as in Ref. [21]([26]), but with a systematic way of choosing the sites $\mathbf{i}$ and $\mathbf{j}$. Furthermore, for many quantum simulation platforms or numerical methods, observables like spin correlations are much more natural to obtain.

Training on expectation values requires a different loss function: Instead of the KL divergence, we use a (scaled) mean-square error (MSE) loss, see Appendix B.2.2, which does not incorporate information on the measurement statistics. Instead, it allows to compensate for systematic errors that are e.g. present in experimental data: Firstly, spatial symmetries can be explicitly applied even if not perfectly represented in each snapshot, by averaging over the expectation values obtained from applying the respective symmetry. Furthermore, some expectation values appear more short-ranged in the experimental data as in theory due to experimental imperfections arising e.g. from the state preparation or losses. This justifies to use an asymmetrically scaled MSE such that the NQS is forced to represent states with higher absolute expectation values than the experimental target rather than smaller ones. More information on the training data can be found in Appendix E.

Hamiltonian-driven training (part 2): In order minimize the energy of the system further in the second part of the hybrid training, we use variational Monte Carlo (VMC) [32, 33, 40], minimizing the expectation value of the energy

$$\langle E_{\vec{\theta}} \rangle = \sum_{\vec{\sigma}} |\psi_{\vec{\theta}}(\sigma)|^2 \, E^{\mathrm{loc}}_{\vec{\theta}}(\sigma), \quad (6)$$

with the local energy $E^{\mathrm{loc}}_{\vec{\theta}}(\sigma) = \frac{\langle \sigma|\mathcal{H}|\psi_{\vec{\theta}}\rangle}{\langle \sigma|\psi_{\vec{\theta}}\rangle}$, see Appendix C.

2 Results

The results for the 2D TFIM and the 2D dipolar XY model ground states on 6×6 and 10×10 lattices obtained with the hybrid training procedure are presented in Figs. 2 to 4. Additional results and details on the architecture and the training can be found in Appendix A. We focus on antiferromagnetic (AFM) spin interactions in both cases. For ferromagnetic interactions, the convergence of the TQS starting from randomly initialized parameters is already very good, see Appendix D, and hence not much room for improvement with the hybrid scheme.

TFIM.– For the 2D TFIM, we use the KL divergence or Wasserstein loss for snapshots from the Z basis and the scaled MSE for the local X magnetization $\langle \hat{\sigma}^x_{\mathbf{i}} \rangle$ with $\langle \dots \rangle$ denoting the average over all snapshots. Both snapshots and local magnetization are obtained from ground state DMRG calculations. The results for a 6×6 square lattice are shown in Fig. 2. In the limit of a vanishing transverse field $h = 0$, the system reduces to the Ising model, which we find to be very well approximated by the TQS with and without pretraining (Fig. 2**a**). When turning on $h > 0$, the

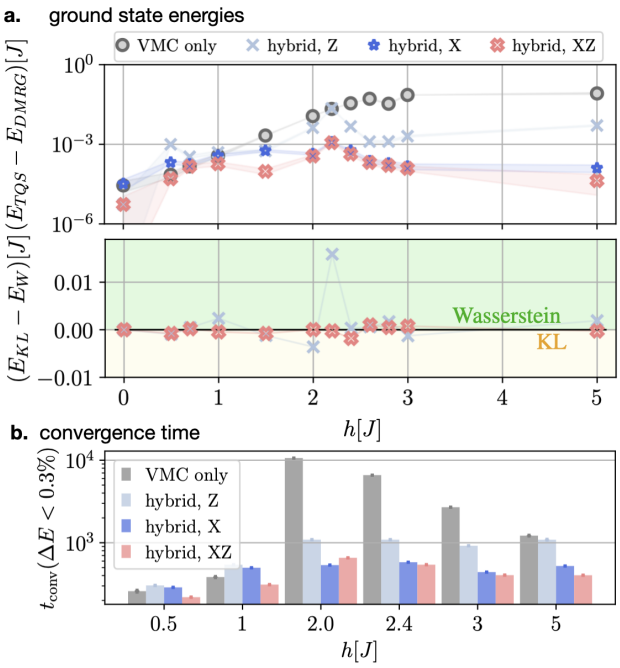

Figure 2: Hybrid training of the 2D TFIM on a 6×6 lattice using numerical data for the pretraining: **a.** Top: Comparison of energy errors per site without pretraining (gray), with pretraining on data from only Z or X basis (blue) and on data from both bases (red) after 1000 training steps. Bottom: Comparison of KL or Wasserstein loss for the snapshots from the Z basis. We show the difference between E_{KL} and E_W, for a training only in the Z basis (blue) and in both bases (red). **b.** Convergence time in units of training steps without and with pretraining on data from Z, X and both bases, for which the energies averaged over 100 epochs drop below $\Delta E = (E_{TQS} - E_{DMRG})/J < 0.3\%$.

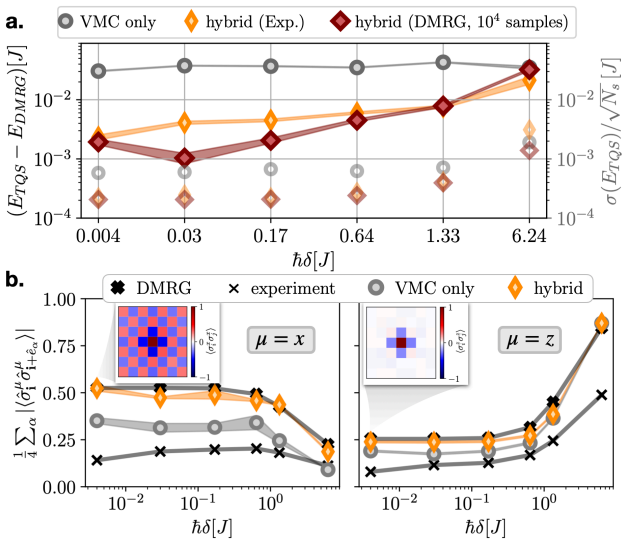

Figure 3: Hybrid training of the dipolar XY model on a 6×6 lattice using experimental (Exp.) and numerical (DMRG) data for the pretraining: **a.** Comparison of energy errors (solid lines) and standard deviation of the mean (light lines) obtained without pretraining (grey) and with pretraining on experimental (orange, $\approx 10^3$ samples) and numerical data (red, 10^4 samples) after 2000 training steps. **b.** The absolute correlations $|\langle\hat{\sigma}^\mu_{\mathbf{i}}\hat{\sigma}^\mu_{\mathbf{i}+\hat{e}_\alpha}\rangle|$ for $\mu = x, z$ (left, right) averaged over all nearest-neighbors α from DMRG and experiment (black) as well as the TQS with (without) pretraining using experimental data shown in orange (gray). Insets show $\langle\hat{\sigma}^\mu_{\mathbf{i}}\hat{\sigma}^\mu_{\mathbf{j}}$ for $\mathbf{i} = (3,3)$ in the middle of the system for $\hbar\delta/J = 0.004$.

accuracy of the TQS decreases without pretraining (gray lines). When pretraining the TQS on a single basis only (blue lines), the energy error can be reduced compared to the only VMC based training in some cases: In the regimes $J > h$ ($h > J$) the physics is dominated by the Z (X) directions, which is reflected in a decreased error when training on the KL loss (scaled MSE) of data from these measurement configurations. In contrast, training on Z (X) in the opposite regimes $J < h$ ($h < J$) does not improve the result as much. Our combined pretraining with data from Z and X basis in contrast significantly improves the results for all h. This is expected to be particularly relevant for more general systems where it is not known which basis is more relevant.

Furthermore, we compare the results when using the KL divergence or the Wasserstein loss in Fig. 2**a** (bottom), and find a similar improvement for both loss functions.

Lastly, the hybrid scheme can tremendously speed up the calculations, see Fig. 2**b**. For example, the convergence time (in units of the training steps) to obtain energy errors per site $(E_{TQS} - E_{DMRG}) < 0.003J$ decreases by more than an order of magnitude when using the hybrid optimization for $h = 2J$. Note that we have used the same hyperparameters for all cases, defined in Appendix, Tab. 1.

Dipolar XY model.– For the dipolar XY model, we use the KL divergence / Wasserstein loss for snapshots from the Z basis and the scaled MSE for spin-spin correlations $\langle\hat{\sigma}^x_{\mathbf{i}}\hat{\sigma}^x_{\mathbf{j}}\rangle$ in the X basis, with a fixed reference site $\mathbf{i}$ located in the middle of the system. The experimentally obtained spin-spin correlations are averaged over different translations of the system, explicitly employing the translational invariance of the Hamiltonian under consideration. The results are shown in

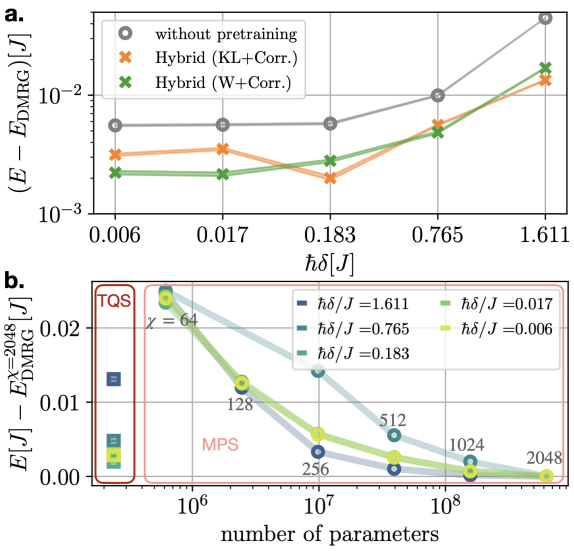

Figure 4: Hybrid training of the dipolar XY model on a 10×10 lattice using experimental data for the pretraining: **a.** Energy error per site (after $15,000$ training steps) obtained from only VMC training (gray lines) and hybrid training with pretraining on $\langle \hat{\sigma}^x_{\mathbf{i}} \hat{\sigma}^x_{\mathbf{j}} \rangle$ and the Kullback-Leibler divergence (Wasserstein distance) in orange (green) using around 10^2 experimental measurements. **b.** Energy as function of the number of parameters used in the TQS (squares) and MPS (circles). For the MPS, we compare different bond dimensions $\chi = 64 \ldots 2048$.

Figs. 3 and 4 for 6×6 and 10×10 square lattices respectively. For both systems, we find significantly improved results when applying the hybrid learning scheme. This is apparent in terms of the energy per site (Figs. 3**a** and 4**a**), its standard deviation of the mean (Fig. 3**a**) and observables such as the averaged nearest-neighbor spin-spin correlations in X and Z bases (Fig. 3**b**). Again, we use the same hyperparameters for our TQS in all cases, see Appendix, Tab. 2. Note that the energy error becomes large for $\hbar\delta > J$ for both system sizes, since the VMC optimization struggles with the fact that configurations which do not follow the checkerboard pattern of sublattices A and B defined by Eq. (2) get a very large energy penalty, resulting in high variances during the optimization. This problem occurs for the optimization with and without pretraining.

Furthermore, Fig. 3**a** shows that the quality of the data does not significantly influence the outcome of the hybrid learning scheme: the energy error per site is improved by the same order of magnitude when using numerical or experimental data, although experimental imperfections are clearly visible e.g. in terms of the spin-spin correlations in Fig. 3**b** (black lines). Even for the 10×10 system, where only $87 - 204$ experimental snapshots for each δ were available, the pretraining improves the performance (Fig. 4**a**), showing the potential of hybrid training schemes even for very limited data sets.

With the improvement through hybrid training, the challenge of training neural quantum states is reduced and allows to make use of the powerful advantage of these ansätze: their great expressivity. This becomes apparent in Fig. 4**b**, where we show the energy error per site for the TQS and MPS ansätze w.r.t. the number of parameters that are used to encode the quantum state. It can be seen that the TQS is extremely efficient and uses around two orders of magnitude fewer parameters for the same energy error compared to the MPS representation at intermediate δ.

3 Summary and Outlook

In this work, we have presented a hybrid approach which combines two powerful approaches for simulating quantum many-body systems: quantum simulators and neural quantum states. The hybrid scheme that we propose uses projective measurements and observables like spin-spin correlations from a quantum simulation platform to pretrain neural quantum states, before switching to the Hamiltonian-driven optimization in terms of variational Monte Carlo.

In the first data-driven stage, we propose two innovations: (*i*) For non-stoquastic Hamiltonians, information beyond the computational basis is necessary to extract the sign structure of the state, and we have demonstrated its importance for high-quality pre-training, see Fig. 2. Since global rotations of the neural quantum state are exponentially costly, we train on site-resolved observables such as spin-spin correlations in the rotated basis. This comes with the advantage that certain experimental imperfections can be compensated, e.g. by explicitly applying spatial symmetries to the experimentally obtained expectation values before the pretraining. (*ii*) For measurements in the computational basis,

we propose to use the Wasserstein distance as the loss function instead of the Kullback-Leibler divergence, with the advantage that information on the Hamiltonian under consideration, in the form of a *distance between snapshots*, is introduced to the data-driven training.

Our results on the 2D TFIM and the 2D dipolar XY model on 6×6 and 10×10 square lattices, testing pretrainings with numerical and experimental data, show that a combination of both Z and X measurement configurations is crucial to improve the optimization results and speed up the convergence over the full range of Hamiltonian parameters. Remarkably, this can be achieved even for very small data sets. Our work paves the way for an efficient and reliable simulation of more complicated systems such as fermionic quantum states or quantum spin liquids with NQS and quantum co-processors, using data from more than two measurement configurations as well as the Wasserstein distance for providing information on the Hamiltonian under consideration to the data-driven pretraining. For these systems, we envision a data processing step before the training to determine the relevant observables for the pretraining away from the computational basis [41, 42].

Code availability.– The code used for this work is provided here: `https://github.com/HannahLange/HybridTransformer`. The implementation of the Wasserstein loss is build on the python package [35].

Note added.– During the finalization of this manuscript, we became aware of Ref. [30]. In this work, an autoregressive, real-valued transformer wave function is trained on numerical data from the computational basis to represent the Rydberg Hamiltonian ground state for various system sizes and different points in the phase diagram.

Acknowledgements.– We thank Abhinav Suresh, Agnes Valenti, Christopher Roth, David Fitzek, Ejaaz Merali, Estelle Inack, Fabian Döschl, Fabian Grusdt, Henning Schlömer, Luciano Vitteriti, Lukas Homeier, Riccardo Rende, Roeland Wiersema, Roger Melko, Schuyler Moss, Tim Harris and Tizian Blatz for fruitful discussions. The data for the TFI models (see Appendix D.3) were taken by P. Scholl, H. Williams and D. Barredo. We acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2111 – 390814868 and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programm (Grant Agreement no 948141) — ERC Starting Grant SimUcQuam. The work at Institut d'Optique is supported by the Agence Nationale de la Recherche (ANR-22-PETQ-0004 France 2030, project QuBitAF), the European Research Council (Advanced grant No. 101018511-ATARAXIA), and the Horizon Europe programme HORIZON-CL4-2022-QUANTUM-02-SGA (project 101113690 (PASQuanS2.1). HL acknowledges support by the International Max Planck Research School for Quantum Science and Technology (IMPRS-QST).

References

References

[1] Román Orús. "A practical introduction to tensor networks: Matrix product states and projected entangled pair states". Annals of Physics **349**, 117–158 (2014).

[2] Ulrich Schollwöck. "The density-matrix renormalization group in the age of matrix product states". Annals of Physics **326**, 96–192 (2011).

[3] J. Eisert, M. Cramer, and M. B. Plenio. "Colloquium: Area laws for the entanglement entropy". Rev. Mod. Phys. **82**, 277–306 (2010).

[4] M B Hastings. "An area law for one-dimensional quantum systems". Journal of Statistical Mechanics: Theory and Experiment **2007**, P08024 (2007).

[5] G. Cybenko. "Approximation by superpositions of a sigmoidal function". Mathematics of Control, Signals and Systems **2**, 303–314 (1989).

[6] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks". Neural Networks **4**, 251–257 (1991).

[7] Giuseppe Carleo and Matthias Troyer. "Solving the quantum many-body problem with artificial neural networks". Science **355**, 602–606 (2017).

[8] Hannah Lange, Anka Van de Walle, Atiye Abedinnia, and Annabelle Bohrdt. "From architectures to applications: a review of neural quantum states" (2024).

[9] Matija Medvidović and Javier Robledo Moreno. "Neural-network quantum states for many-body physics" (2024).

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". Advances in neural information processing systems**30** (2017). url: https://arxiv.org/pdf/1706.03762.

[11] Yuan-Hang Zhang and Massimiliano Di Ventra. "Transformer Quantum State: A Multi-Purpose Model for Quantum Many-Body Problems". Physical Review B **107**, 075147 (2023).

[12] Kyle Sprague and Stefanie Czischek. "Variational monte carlo with large patched transformers" (2024).

[13] Luciano Loris Viteritti, Riccardo Rende, and Federico Becca. "Transformer variational wave functions for frustrated quantum spin systems". Physical Review Letters **130**, 236401 (2023).

[14] Riccardo Rende, Federica Gerace, Alessandro Laio, and Sebastian Goldt. "Mapping of attention mechanisms to a generalized potts model" (2024).

[15] Riccardo Rende, Luciano Loris Viteritti, Lorenzo Bardone, Federico Becca, and Sebastian Goldt. "A simple linear algebra identity to optimize large-scale neural network quantum states". Communications Physics **7**, 260 (2024).

[16] Di Luo, Zhuo Chen, Juan Carrasquilla, and Bryan K. Clark. "Autoregressive neural network for simulating open quantum systems via a probabilistic formulation". Phys. Rev. Lett. **128**, 090501 (2022).

[17] Di Luo, Zhuo Chen, Kaiwen Hu, Zhizhen Zhao, Vera Mikyoung Hur, and Bryan K. Clark. "Gauge-invariant and anyonic-symmetric autoregressive neural network for quantum lattice models". Phys. Rev. Res. **5**, 013216 (2023).

[18] Marin Bukov, Markus Schmitt, and Maxime Dupont. "Learning the ground state of a non-stoquastic quantum Hamiltonian in a rugged neural network landscape". SciPost Phys. **10**, 147 (2021).

[19] Chae-Yeun Park and Michael J. Kastoryano. "Geometry of learning neural quantum states". Phys. Rev. Res. **2**, 023232 (2020).

[20] Estelle M. Inack, Stewart Morawetz, and Roger G. Melko. "Neural annealing and visualization of autoregressive neural networks in the newman-moore model". Condensed Matter**7** (2022).

[21] G. Torlai, G. Mazzola, Ju. Carrasquilla, M. Troyer, R. Melko, and G. Carleo. "Neural-network quantum state tomography". Nature Phys **14**, 447–450 (2018).

[22] Sanjaya Lohani, Brian T Kirby, Michael Brodsky, Onur Danaci, and Ryan T Glasser. "Machine learning assisted quantum state estimation". Machine Learning: Science and Technology **1**, 035007 (2020).

[23] Dominik Koutný, Libor Motka, Zden ěk Hradil, Jaroslav Řeháček, and Luis L. Sánchez-Soto. "Neural-network quantum state tomography". Phys. Rev. A **106**, 012409 (2022).

[24] Stefanie Czischek, M. Schuyler Moss, Matthew Radzihovsky, Ejaaz Merali, and Roger G. Melko. "Data-enhanced variational Monte Carlo simulations for Rydberg atom arrays". Phys. Rev. B **105**, 205108 (2022).

[25] M. Schuyler Moss, Sepehr Ebadi, Tout T. Wang, Giulia Semeghini, Annabelle Bohrdt, Mikhail D. Lukin, and Roger G. Melko. "Enhancing variational Monte Carlo using a programmable quantum simulator" (2023). arXiv:2308.02647.

[26] Elizabeth R. Bennewitz, Florian Hopfmueller, Bohdan Kulchytskyy, Juan Carrasquilla, and Pooya Ronagh. "Neural error mitigation of near-term quantum simulations". Nature Machine Intelligence **4**, 618–624 (2022).

[27] Mohamed Hibat-Allah, Martin Ganahl, Lauren E. Hayward, Roger G. Melko, and

Juan Carrasquilla. "Recurrent neural network wave functions". Phys. Rev. Res. **2**, 23358 (2020).

[28] Cédric Villani. "The wasserstein distances". Pages 93–111. Springer Berlin Heidelberg. Berlin, Heidelberg (2009).

[29] Cheng Chen, Guillaume Bornet, Marcus Bintz, Gabriel Emperauger, Lucas Leclerc, Vincent S. Liu, Pascal Scholl, Daniel Barredo, Johannes Hauschild, Shubhayu Chatterjee, Michael Schuler, Andreas M. Läuchli, Michael P. Zaletel, Thierry Lahaye, Norman Y. Yao, and Antoine Browaeys. "Continuous symmetry breaking in a two-dimensional rydberg array". Nature **616**, 691–695 (2023).

[30] David Fitzek, Yi Hong Teoh, Hin Pok Fung, Gebremedhin A. Dagnew, Ejaaz Merali, M. Schuyler Moss, Benjamin MacLellan, and Roger G. Melko. "Rydberggpt" (2024). arXiv:2405.21052.

[31] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. "An image is worth 16x16 words: Transformers for image recognition at scale" (2021). arXiv:2010.11929.

[32] W. L. McMillan. "Ground state of liquid he^4". Phys. Rev. **138**, A442–A451 (1965).

[33] Li Huang and Lei Wang. "Accelerated monte carlo simulations with restricted boltzmann machines". Phys. Rev. B **95**, 035105 (2017).

[34] Jong Chul Ye. "Geometry of deep learning - a signal processing perspective". Springer Nature. Singapore (2022).

[35] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. "Pot: Python optimal transport". Journal of Machine Learning Research **22**, 1–8 (2021). url: http://jmlr.org/papers/v22/20-451.html.

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

[37] Yihui Quek, Stanislav Fort, and Hui Khoon Ng. "Adaptive quantum state tomography with neural networks" (2018). arXiv:1812.06693.

[38] Alistair W. R. Smith, Johnnie Gray, and M. S. Kim. "Efficient quantum state sample tomography with basis-dependent neural networks". PRX Quantum **2**, 020348 (2021).

[39] Hannah Lange, Matjaž Kebrič, Maximilian Buser, Ulrich Schollwöck, Fabian Grusdt, and Annabelle Bohrdt. "Adaptive quantum state tomography with active learning". Quantum **7**, 1129 (2023).

[40] Federico Becca and Sandro Sorella. "Quantum monte carlo approaches for correlated systems". Cambridge University Press. (2017).

[41] Cole Miles, Annabelle Bohrdt, Ruihan Wu, Christie Chiu, Muqing Xu, Geoffrey Ji, Markus Greiner, Kilian Q. Weinberger, Eugene Demler, and Eun-Ah Kim. "Correlator convolutional neural networks as an interpretable architecture for image-like quantum matter data". Nature Communications **12**, 3905 (2021).

[42] R. Verdel, V. Vitale, R. K. Panda, E. D. Donkor, A. Rodriguez, S. Lannig, Y. Deller, H. Strobel, M. K. Oberthaler, and M. Dalmonte. "Data-driven discovery of statistically relevant information in quantum simulators". Phys. Rev. B **109**, 075152 (2024).

[43] Moritz Reh, Markus Schmitt, and Martin Gärttner. "Optimizing design choices for neural quantum states". Phys. Rev. B **107**, 195115 (2023).

[44] Diederik P. Kingma and Jimmy Ba. "Adam: A method for stochastic optimization" (2017). arXiv:1412.6980.

[45] Sandro Sorella. "Green function monte carlo with stochastic reconfiguration". Phys. Rev. Lett. **80**, 4558–4561 (1998).

[46] Sandro Sorella. "Generalized lanczos algorithm for variational quantum monte carlo". Phys. Rev. B **64**, 024512 (2001).

[47] Ao Chen and Markus Heyl. "Empowering deep neural quantum states through efficient optimization" (2024).

[48] Kaelan Donatella, Zakari Denis, Alexandre Le Boité, and Cristiano Ciuti. "Dynamics with autoregressive neural quantum states: Application to critical quench dynamics". Phys. Rev. A **108**, 022210 (2023).

[49] Hannah Lange, Fabian Döschl, Juan Carrasquilla, and Annabelle Bohrdt. "Neural network approach to quasiparticle dispersions in doped antiferromagnets" (2023). arXiv:2310.08578.

[50] Pascal Scholl, Michael Schuler, Hannah J. Williams, Alexander A. Eberharter, Daniel Barredo, Kai-Niklas Schymik, Vincent Lienhard, Louis-Paul Henry, Thomas C. Lang, Thierry Lahaye, Andreas M. Läuchli, and Antoine Browaeys. "Quantum simulation of 2d antiferromagnets with hundreds of rydberg atoms". Nature **595**, 233–238 (2021).

[51] Daniel Barredo, Sylvain de Léséleuc, Vincent Lienhard, Thierry Lahaye, and Antoine Browaeys. "An atom-by-atom assembler of defect-free arbitrary two-dimensional atomic arrays". Science **354**, 1021–1023 (2016).

[52] Sylvain de Léséleuc, Daniel Barredo, Vincent Lienhard, Antoine Browaeys, and Thierry Lahaye. "Optical control of the resonant dipole-dipole interaction between rydberg atoms". Phys. Rev. Lett. **119**, 053202 (2017).

[53] Cheng Chen, Gabriel Emperauger, Guillaume Bornet, Filippo Caleca, Bastien Gély, Marcus Bintz, Shubhayu Chatterjee, Vincent Liu, Daniel Barredo, Norman Y. Yao, Thierry Lahaye, Fabio Mezzacapo, Tommaso Roscilde, and Antoine Browaeys. "Spectroscopy of elementary excitations from quench dynamics in a dipolar xy rydberg simulator" (2023). arXiv:2311.11726.

[54] C. Hubig, F. Lachenmaier, N.-O. Linden, T. Reinhard, L. Stenzel, A. Swoboda, and M. Grundner. "The SyTen toolkit". https://syten.eu.

[55] C. Hubig. "Symmetry-protected tensor networks". PhD thesis. LMU München. (2017). url: https://edoc.ub.uni-muenchen.de/21348/.

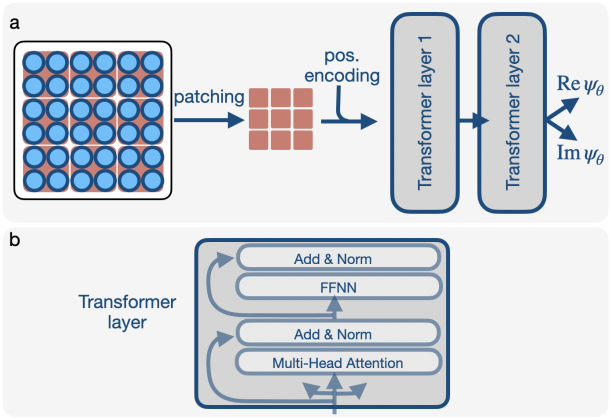

Figure S1: The patched transformer architecture: **a.** Patches of 2×2 sites are embedded and positional encoded, before being passed through two transformer layers, with separate output layers for amplitude and phase of the wave function. **b.** The transformer layer with multi-head attention, addition, normalization and feed forward neural network (FFNN).

APPENDIX

A Transformer wave function

A.1 The architecture

Here, we use a representation of the dipolar XY model ground states $|\psi\rangle_{\vec{\theta}}$ in terms of a transformer neural network. Transformer networks are characterized by their attention mechanism [10], see Fig. S1b, consisting of trainable, all-to-all connections throughout the system, which allow the network to capture correlations between any sites in the system regardless of the position. When a local input configuration σ_i at site i is passed to a transformer layer, it is first embedded into a, typically higher dimensional, space of dimension d_h. Furthermore, a positional encoding is added, that encodes the position i. Then, the encoded input $\tilde{\sigma}_i$ is projected on a query vector (q_i), key vector (k_i) and a value vector (v_i), using trainable matrices $W^q, W^k, W^v \in \mathbb{R}^{d_h \times d_h}$. By defining query, key and value matrices as $Q = (q_1, ..., q_N)$, $K = (k_1, ..., k_N)$ and $V = (v_1, ..., v_N)$, the self-attention mechanism is given by

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK}{\sqrt{d_h/h}} + M\right) V. \tag{7}$$

Here, $h = 1$ and M is a *mask* that is added to the signal, masking out all sites that $j > i$ and making the model autoregressive [10]. In a multi-headed attention mechanism each query, key and value vector is mapped to $h > 1$ vectors with trainable weight matrices. In most cases, several transformer layers are stacked on top of each other.

Using two different output layers for phase and amplitude of the quantum state at each site, $\phi_{i,\vec{\theta}}(\vec{\sigma})$ and $p_{i,\vec{\theta}}(\vec{\sigma}|\vec{\sigma}_{<i})$ respectively, we define the transformer wave function by

$$|\psi\rangle_{\vec{\theta}} = \sum_{\vec{\sigma}} \exp(i\phi_{\vec{\theta}}(\vec{\sigma}))\sqrt{p_{\vec{\theta}}(\vec{\sigma})}\,|\sigma\rangle\,, \tag{8}$$

with $\phi_{\vec{\theta}}(\vec{\sigma}) = \sum_i \phi_{i,\vec{\theta}}(\vec{\sigma})$ and $p_{i,\vec{\theta}}(\vec{\sigma}) = \Pi_i p_{i,\vec{\theta}}(\vec{\sigma}|\vec{\sigma}_{<i})$ see also Ref. [11]. Note that by using a softmax activation function for the amplitude layer, both local and total amplitudes are normalized, making the TQS autoregressive. Furthermore, we apply rotational symmetry w.r.t. rotations by an angle of π, as detailed in Ref. [43].

The self-attention mechanism makes the transformer very expressive, but on the other hand also slows down the training, since every all-to-all connection has to be learned. Patched transformers, inspired from Vision transformers (ViT) [31], overcome this problem by splitting the system into small patches, see Fig. S1a. In Refs. [12–15], the authors have shown that patched transformers can be used successfully for frustrated spin systems and Rydberg systems.

A.2 Hyperparameters

In this work, we use a patch size of 2×2 and two transformer layers. The network parameters and training hyperparameters are listed in Tabs. 1 and 2. Before the training, we initialize the network parameters according to a uniform distribution between -0.1 and 0.1 and set all biases of the output layers to zero. For the TFIM systems we use a exponential decay starting at epoch n_{start} with decay rate γ. For the dipolar XY model, we observe that gradients can become large and find it convenient to use a cosine learning rate schedule

$$l(n) = l_{\max} \cdot \max\left[\eta \frac{1}{2}(1 + \cos(\pi \frac{n}{n_{\max}}), 0.01\right] \tag{9}$$

with a prefactor $\eta = \begin{cases} 1 & \text{if } n > n_{\text{w}} \\ \frac{n}{n_{\text{w}}} & \text{else} \end{cases}$ that reduces the learning rate in the warmup phase $n < n_{\text{w}}$. All benchmark calculations without pretraining are done with the same architectures and hyperparameters.

	6×6
number of transformer layers	2
hidden dimension d_h	32
FFNN dimension (see Fig. S1)	$4d_h$
total number of parameters	$28,800$
pretraining learning rate	$5 \cdot 10^{-4}$ [a]
VMC learning rate $l_{\max}$	$2 \cdot 10^{-3}$
learning rate schedule $l(n)$	$n_{\text{start}} = 300$, $\gamma = 0.9998$
maximal no. of pretraining steps $n^{\max}_{\text{pretrain}}$	100
number of samples in VMC training (part 2)	512
number of samples used for the data points in Fig. 2	100×512
errorbars in Fig. 2	standard deviation of the mean
number of training steps used in Fig. 2	$1,000$

[a] Except for the training on only Z basis measurements and $h/J = 5$, where the pretraining learning rate is decreased by a factor $1/5$ to avoid getting stuck in a local minimum.

Table 1: Transformer parameters and training hyperparameters for the 6×6 2D TFIM systems (Fig. 2). We refer to the total number of training steps including the pretraining as n. Futhermore, we use a exponential decay rate starting at epoch n_{start} with decay rate γ.

	6×6	10×10
number of transformer layers	2	2
hidden dimension d_h	56	88
FFNN dimension (see Fig. S1)	$4d_h$	$5d_h$
total number of parameters	$85,320$	$238,424$
pretraining learning rate	10^{-4}	10^{-5}
VMC learning rate $l_{\max}$	$5 \cdot 10^{-4}$	10^{-4}
learning rate schedule $l(n)$	$n_{\max} = 5,000$, $n_{\mathrm{w}} = 1,000$	$n_{\max} = 20,000$, $n_{\mathrm{w}} = 10,000$
max. no. of pretraining steps $n^{\max}_{\mathrm{pretrain}}$	200	$1,500$
number of samples in VMC training	256	256
number of samples in Figs. 3 & 4	$10,000$	100×256
errorbars in Figs. 3 & 4	standard dev. of the mean	standard dev. of the mean
no. of training steps in Figs. 3 & 4	$2,000$	$15,000$

Table 2: Transformer parameters and training hyperparameters for the 6×6 and 10×10 dipolar XY systems (Figs. 3 and 4). We refer to the total number of training steps including the pretraining as n and use a cosine learnig rate schedule (9).

B Data driven (pre-)training

B.1 Measurements in computational basis: Kullback-Leibler divergence and Wasserstein distance

For the training on data from the computational basis, we compare two loss functions: The Kullback-Leibler (KL) divergence defined in Eq. (4) and the Wasserstein (W) distance defined in Eq. (5). The main advantage of the Wasserstein distance is that it incorporates the definition of a ground metric, capturing the distance between the snapshots in the training data set. Here, we use the diagonal part of the energy $E_{\mathrm{diag}}(\sigma^p, \sigma^q) = |\langle\sigma^p|\hat{\mathcal{H}}|\sigma^p\rangle - \langle\sigma^q|\hat{\mathcal{H}}|\sigma^q\rangle|$ as the ground metric. This implicitly captures the fact that configurations that are connected e.g. by a few spin flips, exchanges or certain rotations, are close to each other. The difference between the Kullback-Leibler divergence (KL) and the Wasserstein distance (W) are shown for two exemplary scenarios in Fig. S2:

- **scenario 1:** Fig. S2a shows how the Wasserstein distance can incorporate the information on the Hamiltonian in the data-driven training. We show two distributions q and p, each with peaks at different Néel configurations but otherwise same probability mass. For these distributions, we can calculate $KL(q, p) > 0$. However, the KL does not take into account that for the Heisenberg model, the two Néel configurations and the two remaining configurations are energetically degenerate and can hance be treated as equivalent. In contrast, when calculating the Wasserstein distance, the configurations are ordered according to their (diagonal) energy E, resulting in the distributions shown in Fig. S2b. In this exemplary scenario, it now becomes apparent that the distributions p and q are the same when degenerate configurations are grouped together, i.e. $W(q, p) = 0$.

- **scenario 2:** Fig. S2b shows two completely different distributions q and p (left). After ordering the configurations according to the (diagonal) energy (right), $W(p, q)$ measures the amount of probability mass that has to be transported from q to p or vice versa. Hereby, γ is the transport matrix, as defined in Eq. (5). Since for larger systems, there are many possible transport ways γ, an optimal transport problem to find the optimal γ has to be solved in each iteration.

For both KL or the Wasserstein distance, we use the Adam optimizer [44] for the minimization of the respective loss function.

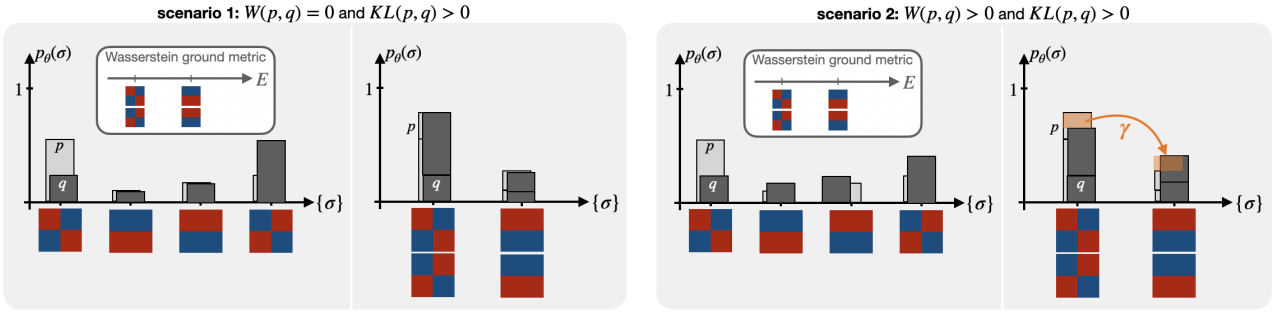

Figure S2: Comparison between Kullback-Leibler (KL) divergence and Wasserstein distance (W) with energy as the ground metric for the exemplary case of a 2×2 Heisenberg model. We focus on two scenarios: **scenario 1:** We show two distributions q and p, each with peaks at different Néel configurations but otherwise same probability mass (left). Ordering the configurations according to their energy as done for the Wasserstein distance (right) yields a vanishing $W(p,q) = 0$. In contrast $KL(p,q) \neq 0$ since it has no notion of the "closeness" (w.r.t. ground state energies) of the Néel states. **scenario 2:** For two completely different distributions q and p, $W(p,q)$ measures the amount of probability mass that has to be transported from q to p or vice versa. Hereby, γ is the transport matrix. For larger systems, there are many possible transport ways, and hence the optimal γ has to be found, see Eq. (5).

B.2 Training on data away from the computational basis

B.2.1 Exponential scaling w.r.t. spins rotated away from computational basis

Here, we review why evaluating the Kullback-Leibler divergence or the Wasserstein distance for snapshots away from the computational basis is exponentially costly. We consider a local rotation of r spins from the Z to the X basis, using

$$\hat{U} = \hat{U}_1 \otimes \hat{U}_2 \otimes \ldots \hat{U}_N, \tag{10}$$

with the local spin rotations given by

$$\hat{U}_i = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{or} \quad \hat{U}_i = 1_{2\times 2} \tag{11}$$

at each site of the system. To evaluate the Kullback-Leibler divergence or the Wasserstein distance for snapshots in the global X basis, σ^x, one needs to calculate

$$\psi'_{\vec{\theta}}(\sigma^x) = \langle \sigma^x | \hat{U} | \psi_{\vec{\theta}} \rangle = \sum_\sigma \langle \sigma^x | \hat{U} | \sigma \rangle \psi_{\vec{\theta}}(\sigma) = \sum_\sigma \langle \sigma^x | (\hat{U}_1 \otimes \hat{U}_2 \otimes \ldots \hat{U}_N) | \sigma \rangle \psi_{\vec{\theta}}(\sigma). \tag{12}$$

For one rotated spin $r = 1$, this involves the evaluation of 2^1 terms, for $r = 2$ it involves $2 \cdot 2 = 2^2$ terms etc..

B.2.2 Training on expectation values in the X basis

The information from measurements away from the computational basis are incorporated by considering spin correlation maps instead of the measurement statistics. Specifically, we consider $\langle \hat{\sigma}_i^x \hat{\sigma}_j^x \rangle_{(\vec{\theta})}$ from the measurements (the NQS) and define the scaled mean-square error (MSE) loss,

$$\mathcal{C} = \sum_j \eta_j \cdot \text{asymmMSE}(\langle \hat{\sigma}_i^x \hat{\sigma}_j^x \rangle_{\vec{\theta}}, \langle \hat{\sigma}_i^x \hat{\sigma}_j^x \rangle), \tag{13}$$

where i is fixed to the middle of the system and

$$\text{asymmMSE}(x_1, x_2) = \begin{cases} f * (x_1 - x_2)^2, & \text{if } |x_1| < |x_2| \\ (x_1 - x_2)^2, & \text{else} \end{cases}, \tag{14}$$

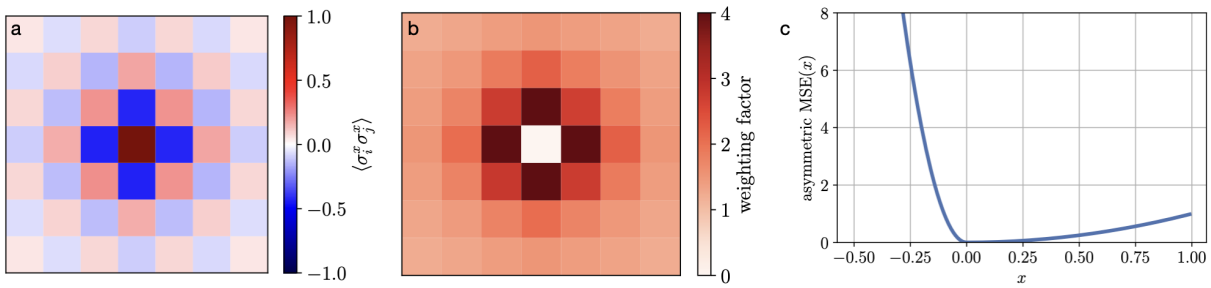

Figure S3: Correlation-driven training in the X basis by minimizing $\mathcal{C} = \sum_j \eta_j \cdot \text{asymmetricMSE}(\langle\sigma_i^x\sigma_j^x\rangle_{\vec{\theta}} - \langle\sigma_i^x\sigma_j^x\rangle)$: **a.** Exemplary correlation map for $\delta \approx 1.35$. **b.** Weighting factor $\eta_j = 1/(1 - |\langle\sigma_i^x\sigma_j^x\rangle|)^4$. c) We use an asymmetric MSE with a higher penalty if $|\langle\sigma_i^x\sigma_j^x\rangle_{\vec{\theta}}| < |\langle\sigma_i^x\sigma_j^x\rangle|$, here shown for $x = (x_1, x_2)$ with $x_1, x_2 \geq 0$.

$f > 1$ is a factor that gives a higher penalty if $|\langle\sigma_i^x\sigma_j^x\rangle_{\vec{\theta}}| < |\langle\hat{\sigma}_i^x\hat{\sigma}_j^x\rangle|$ and $\eta_j = 1/(1 - |\langle\sigma_i^x\sigma_j^x\rangle|)^4$ weights the contributions to $\mathcal{C}$ depending on the magnitude of the target correlation value. Fig. S3 shows the weighting factor η_j (b) for the example correlation map shown in a), as well as and the asymmetric MSE (c).

Hereby, evaluating

$$\langle\hat{\sigma}_i^x\hat{\sigma}_j^x\rangle_{(\vec{\theta})} = \frac{1}{N_s}\sum_{\mathbf{x}}\sum_{\mathbf{x}'}\frac{\psi_{\vec{\theta}}(\mathbf{x}')}{\psi_{\vec{\theta}}(\mathbf{x})}\langle\mathbf{x}|\hat{\sigma}_i^x\hat{\sigma}_j^x|\mathbf{x}'\rangle \tag{15}$$

involves the computation of only 4 terms. As for the computational basis, we use the Adam optimizer [44] to optimize according to Eq. (13).

C Energy driven training

In the second part of the hybrid training, we use VMC, estimating the expectation value of the energy $\langle E\rangle$ of the TQS using Eq. (6). In order to stabilize the training, see e.g. Ref. [27], we use

$$\mathcal{C} = \sum_{\vec{\sigma}} |\psi_{\vec{\theta}}(\sigma)|^2 \left[E_{\vec{\theta}}^{\text{loc}}(\sigma) - \langle E_{\vec{\theta}}^{\text{loc}}\rangle\right] \tag{16}$$

to minimize both the local energy as well as the variance of the gradients.

As explained in the main text, the optimization of Eq. (16) is in general very complicated due to the very rugged loss landscape with many local minima [18]. Furthermore, the training of the autoregressive transformer is further complicated by the fact that stochastic reconfiguration [40,45,46] and variants as used in Refs. [15,47] become unstable for autoregressive architectures [48,49]. Hence, we use the Adam optimizer [44] to optimize according to Eq. (16).

The results for an only energy based training for 6×6 TFIM and dipolar XY systems are shown in Fig. S4 (top and bottom, respectively) for antiferromagnetic (AFM) and ferromagnetic exchange interactions. It can be seen that the FM case is in general easier to learn for the TQS, reaching energy errors below 1% after only 1000 (2000) training steps for the TFIM (dipolar XY model). Hence, we do not consider these cases for the hybrid training scheme. For the AFM case, the training becomes more challenging, yielding longer convergence times for most parameter regimes.

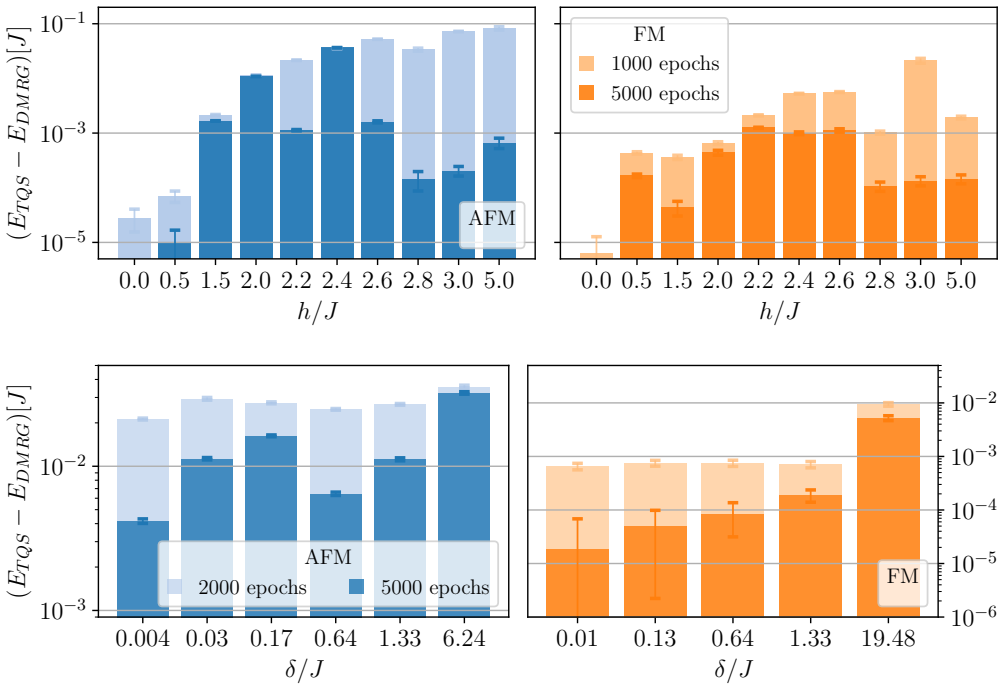

Figure S4: Top: Transverse field ising model without pretraining. We show the energy error per site for 6×6 AFM (left) and FM (right) TFIM after 1000 and 5000 epochs of only energy driven training. Bottom: Dipolar XY systems without pretraining. We show the energy error per site for 6×6 AFM (left) and FM (right) dipolar XY systems after 2000 and 5000 epochs of only energy driven training.

D Additional results

In this section, we provide additional results obtained using the hybrid training scheme for the 2D TFIM (1) and the 2D dipolar XY model (2). Furthermore, we present results on an experimentally realized TFIM system [50], for which experimental data from a single basis is available.

D.1 Additional results on the 2D dipolar XY model

D.1.1 6×6 systems

In Fig. S5 we show exemplary training curves for the 6×6 systems and different Hamiltonian parameters δ up to epoch 3000. It can be seen that both energy and the variance of the energies are decreased when using the hybrid scheme, both for experimental (light and dark red lines) and numerical (light and dark blue lines) data.

Furthermore, we show the spin-spin correlation maps for exemplary $\hbar\delta/J = 0.17$ from the numerical (experimental) training data as well as from the TQS after pretraining and after the full hybrid scheme (left to right) in Fig. S6 **a** (**b**). After the first stage, the TQS captures already the AFM correlations. Note that for both DMRG and experimental data, the spin-spin correlation are slightly stronger than for the target, which is due to the asymmetric MSE 13 that forces the TQS to represent stronger correlations rather than smaller correlations than the target. The final result obtained from the TQS agrees very well with the DMRG result in both cases. Furthermore, after the pretraining the correlation maps for the TQS can show a small asymmetry between vertical and horizontal directions since we only apply 180 degree rotational symmetry (see middle plots in Fig. S6 **a** and **b**). The asymmetry vanishes upon convergence (see Fig. S6 **a** and **b** on the right).

Finally, Fig. S7 shows the magnetization in the computational basis. Fig. S7**a** reveals good agreement

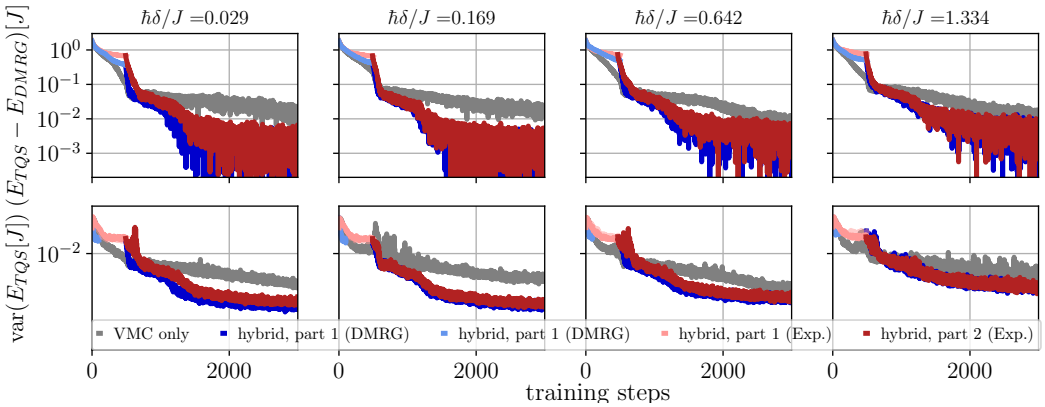

Figure S5: Training curves for 6×6 dipolar XY systems for different δ, with only energy based training (VMC, gray lines) and hybrid training (red lines) with numerical (light red lines) and experimental (solid red lines) data. Top: Energy error per site. Bottom: Variance of energies during the training.

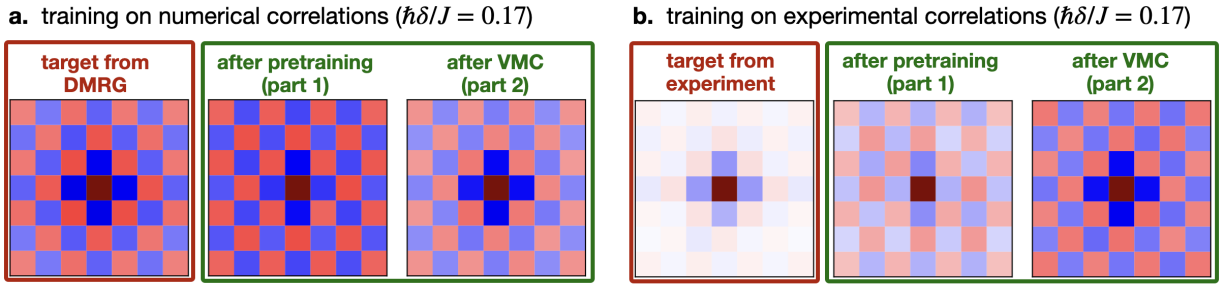

Figure S6: Correlation maps for the dipolar XY model on the 6×6 lattice. We show $\langle \sigma^x_{\mathbf{i}} \sigma^x_{\mathbf{j}} \rangle$, with $\mathbf{i}$ located in the middle of the system, for $\hbar\delta/J = 0.17$. During the pretraining, we train on target correlations $\langle \sigma^x_i \sigma^x_j \rangle$ from numerics (experiment), see **a.** (**b.**) on the left, leading to a parameterization of the TQS that approximates these correlations (middle). The correlation maps of the TQS after the complete hybrid training including the data-driven second part are shown on the right. Note that the correlations after the pretraining can have higher values than the target, since we employ a scaled MSE loss that penalizes larger values than the target less than smaller values. The correlation maps for the TQS can show a small asymmetry between vertical and horizontal directions since we only apply 180 degree rotational symmetry. The asymmetry vanishes upon convergence (see **a.** and **b.** on the right).

of the staggered magnetization

$$m_{\text{stag}} = \frac{1}{L^2} \sum_{s \in A,B} \langle (-1)^s (\hat{S}^z_{\mathbf{i}} - 1/2) \rangle \tag{17}$$

for the TQS (with and without pretraining) with the DMRG results. The staggered magnetization obtained from the experiment deviates only slightly for large light shifts δ. Fig. S7**b** shows the local magnetization $\langle \sigma^z_{\mathbf{i}} \rangle$ for $\hbar\delta = 6.24$.

D.1.2 10×10 systems

In Fig. S8 we show exemplary training curves for the 10×10 systems and different Hamiltonian parameters δ up to epoch 15000. Again, both energy and the variance of the energies are decreased when using the hybrid scheme, here for experimental data(light and dark red lines), compared to only using VMC (gray lines).

Furthermore, the spin-spin correlations for 10×10 dipolar XY model systems are shown in Fig. S9. We plot both nearest-neighbor spin correlations (Fig. S9**a**) and next-nearest-neigbor spin correlations (Fig. S9**b**) in the X (left) and Z (right) basis. In all cases, the results from DMRG and experiment are

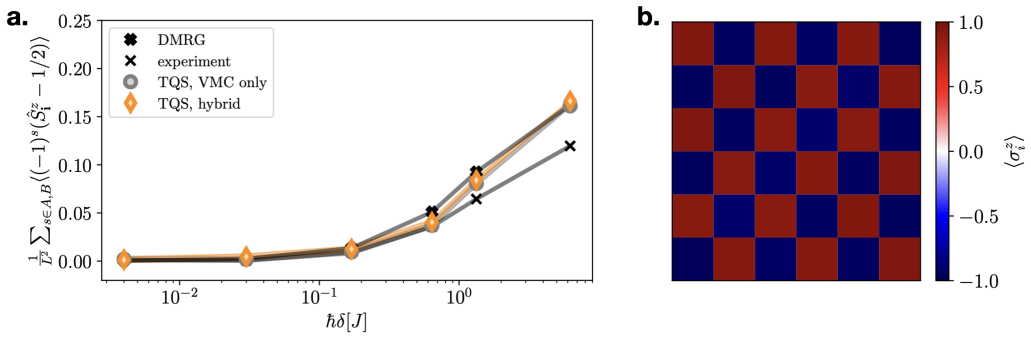

Figure S7: Magnetization in Z direction for the dipolar XY model on the 6×6 lattice. **a.** We show $\frac{1}{L^2}\sum_{s\in A,B}\langle(-1)^s(\hat{S}^z_{\mathbf{i}} - 1/2)\rangle$ for different light shifts δ, for DMRG (black thick crosses), experiment (thin black crosses), TQS without pretraining (gray circles) and TQS with pretraining on experimental data (orange diamonds). **b.** The local magnetization $\langle\sigma^z_{\mathbf{i}}\rangle$ for $\hbar\delta = 6.24$ obtained from the pretrained TQS. In both figures, we use the KL divergence for snapshots from the computational basis and the spin correlation maps from the X basis for the pretraining. We use $10,000$ samples for each data point.

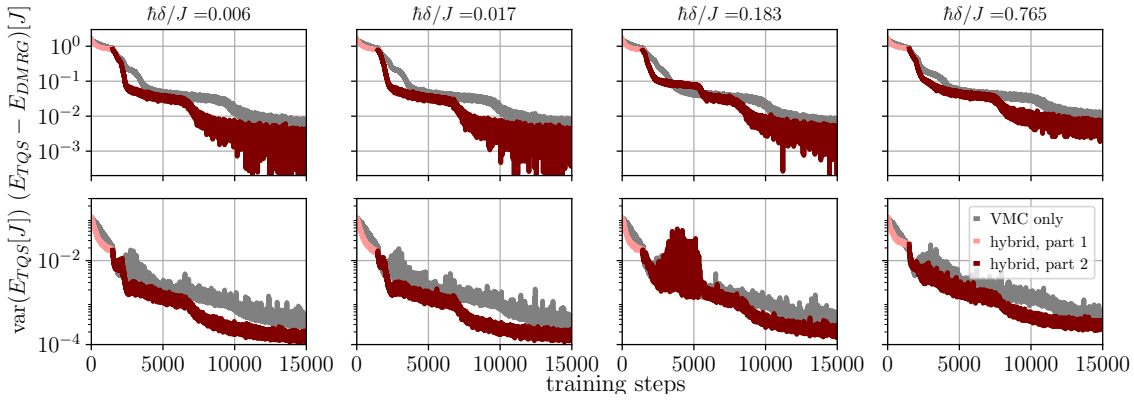

Figure S8: Training curves for 10×10 dipolar XY systems for different δ, with only energy based training (VMC, gray lines) and hybrid training (red lines) using experimental data. Top: Energy error per site. Bottom: Variance of energies during the training.

denoted as black thick and thin crosses respectively. The TQS results without and with pretraining using experimental data are shown in gray and orange. Similar to the 6×6 systems in Fig. 3, the TQS results agree well with the DMRG results when using the hybrid scheme. Note that this is remarkable since we pretrain on the experimental data, which, as can be seen in Fig. S9 for the back thing crosses, under-estimates both nearest-neighbor correlations. Without pretraining, the TQS shows smaller nearest-neighbor and next-nearest-neighbor correlations than DMRG for the X basis, while slightly overestimating both correlations in the Z basis. Spin-spin correlations for a fixed site $\mathbf{i} = (5, 5)$ in the middle of the system in X (top) and Z (bottom) basis are shown in Fig. S9**c**.

Lastly, Fig. S10 shows the magnetization in the computational basis. Fig. S10**a** reveals good agreement of the staggered magnetization (17) for the pretrained TQS with the DMRG results. As for the smaller systems in Fig. S7, the staggered magnetization obtained from the experiment deviates only slightly for the largest light shift $\hbar\delta = 1.61$. However, in contrast to the 6×6 systems, for the 10×10 systems the staggered magnetization is highly overestimated for $\hbar\delta < 1$ when omitting the pretraining. Furthermore, Fig. S7**b** shows a small inhomogeneity of the local magnetization $\langle\sigma^z_{\mathbf{i}}\rangle$ for $\hbar\delta = 1.61$ at the lower left and upper right corners, resulting from the sampling path, where 2×2 patches are sampled in a 1D manner, combined with a 180 degree rotational symmetry. We expect that this feature vanishes upon applying more symmetries, e.g. translational symmetry.

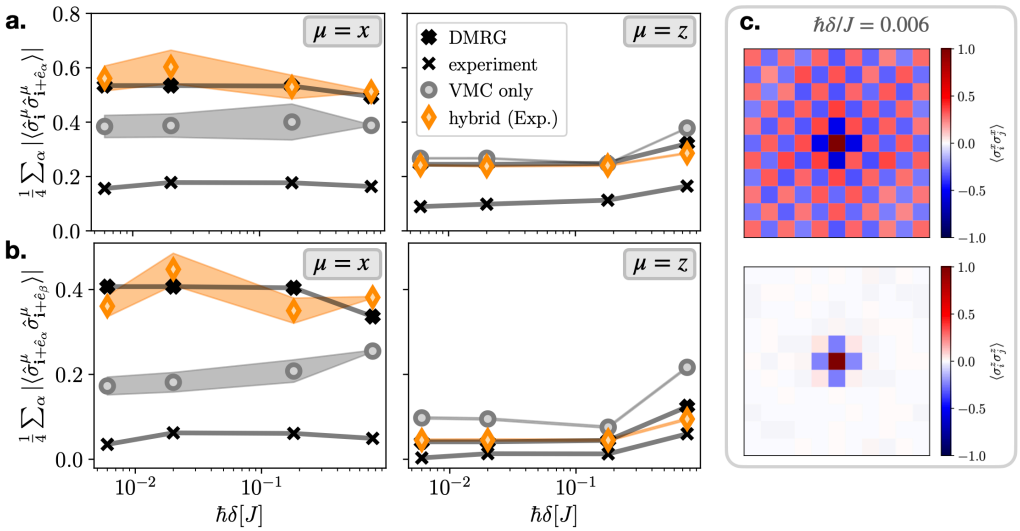

Figure S9: Spin-spin correlations for 10×10 dipolar XY model systems: The nearest-neighbor spin correlations (**a.**) and next-nearest-neighbor spin correlations (**b.**) averaged over all neighbors for $\mu = x$ (left) and $\mu = z$ (right), from DMRG and experiment (black) as well as the TQS without and with pretraining (with the Wasserstein and MSE loss) using experimental data (gray and orange). We use $40,000$ samples for each data point. Errorbars denote the standard deviation of the mean. **c.** Spin-spin correlation maps for fixed site $\mathbf{i} = (5, 5)$ in the middle of the system.

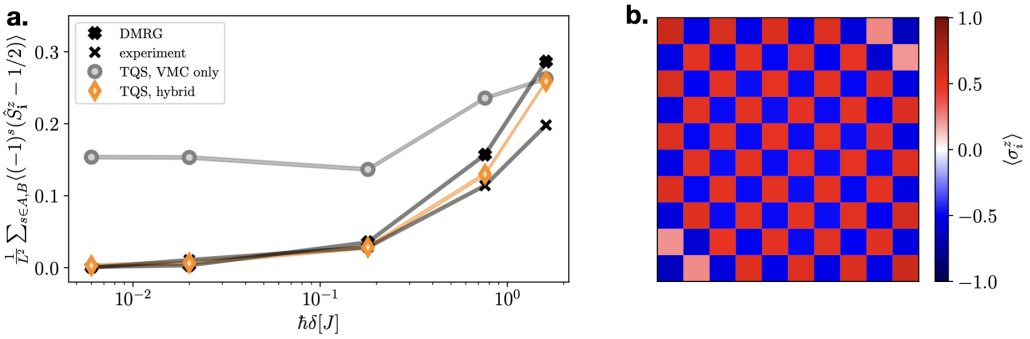

Figure S10: Magnetization in Z direction for the dipolar XY model on the 10×10 lattice. **a.** We show $\frac{1}{L^2}\sum_{s\in A,B}\langle(-1)^s(\hat{S}^z_{\mathbf{i}} - 1/2)\rangle$ for different light shifts δ, for DMRG (black thick crosses), experiment (thin black crosses), TQS without pretraining (gray circles) and TQS with pretraining on experimental data (orange diamonds). **b.** The local magnetization $\langle\sigma^z_{\mathbf{i}}\rangle$ for $\hbar\delta = 1.61$ obtained from the pretrained TQS. In both figures, we use the Wasserstein distance for snapshots from the computational basis and the spin correlation maps from the X basis for the pretraining. We use $40,000$ samples for each data point.

D.2 Additional results on the 2D TFIM

In Fig. S6 we show additional results for the 2D AFM TFIM on a 6×6 lattice. Fig. S6**a** shows the exemplary learning curve for $h/J = 5.0$. It can be seen that while pretraining reaches an energy error per site of around $> 0.1J$, the second VMC part reduces the error below $0.001J$ below within less than 500 optimization steps (Fig. S6**a** left). Without pretraining, the optimization is stuck at around $0.1J$ for almost 1000 epochs before decreasing to a slightly higher energy than with pretraining. Furthermore, the variance is highly decreased when using pretraining (see Fig. S6**a** right).

In Fig. S6**b**, we additionally show the average magnetization $\langle \hat{S}^{\mu}_{\text{tot}} \rangle$ in $\mu = x, z$ directions after pretraining (part 1) and VMC (part 2) staged of the hybrid training. In both cases, the average magnetization agrees well with the DMRG results. In particular, the increase of magnetization in the X direction is correctly captured already in the pretraining.

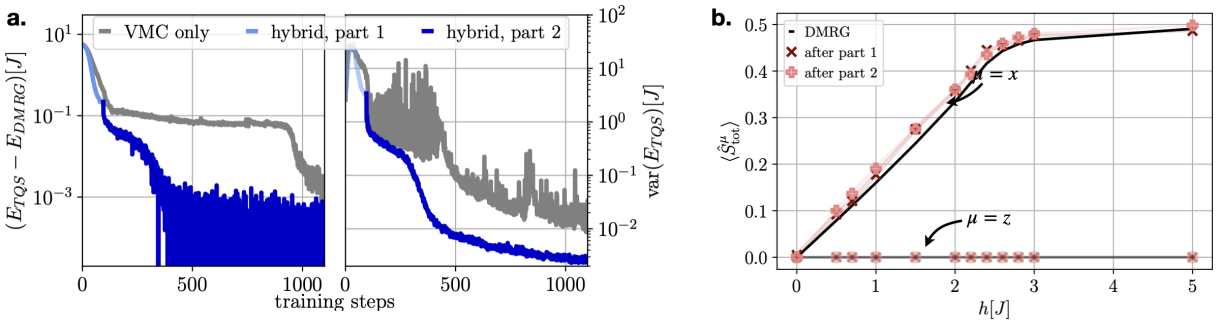

Figure S11: Additional results on the 2D AFM TFIM on a 6×6 lattice: **a.** Exemplary learning curve for $h/J = 5.0$. **b.** The average magnetization $\langle \hat{S}^{\mu}_{\text{tot}} \rangle$ in $\mu = x, z$ directions after part 1 and part 2 of the hybrid training. We use $10,000$ samples for the evaluation of $\langle \hat{S}^{\mu}_{\text{tot}} \rangle$ and errorbars denote the standard deviation of the mean.

D.3 Experimental realization of the 2D TFIM

D.4 Training data

For the computational Z basis, we use the experimentally available data from Ref. [50], see Appendix E. For comparison, we train the transformer wave function with numerical data from the same basis. Furthermore, we use the numerically obtained locally resolved magnetization from X basis for the pretraining.

D.5 Results

In this section, we present results on the experimentally realized TFIM, see (20), on a 10×10 lattice. Fig. S12**a** shows the values for $\Omega(t)$ and $\delta(t)$ realized in the experiment [50] over the sweep time t. During this sweep, the system is transferred from a initial paramagnetic (PM) ground state $|\downarrow\downarrow \ldots \downarrow\rangle$ to an antiferromagnetic (AFM) phase $|\downarrow\uparrow\downarrow \ldots \downarrow\rangle$ [50].

The respective ground state energy errors per site for $\Omega(t)$ and $\delta(t)$ during the sweep are shown in Fig. S12**b**, and the hyperparameters that were used during the training are listed in Tab. 3. In Fig. S12**b**, gray lines denote the result without pretraining, blue and red lines the results with pretraining on numerical data from both Z and X bases and with pretraining on experimental data from the Z and numerical data from the X basis respectively, all of them evaluated after only 700 training iterations. All errors with and without pretraining are already very low in this early stage of the optimization: Except for intermediate times $1.5\mu\text{s} < t < 3\mu\text{s}$, all energy errors are on the order or below 0.1%.

When pretraining with numerical data, the energy error per site decreases w.r.t. the only VMC based training, in particular for $1.8\mu\text{s} < t < 4\mu\text{s}$, i.e. in the regime where a phase transition between PM and AFM phase is expected, the error decreases by more than an order of magnitude. For other times t, the energy error per site is slightly decreased or the same ($t = 1.8\mu\text{s}$). For the experimental

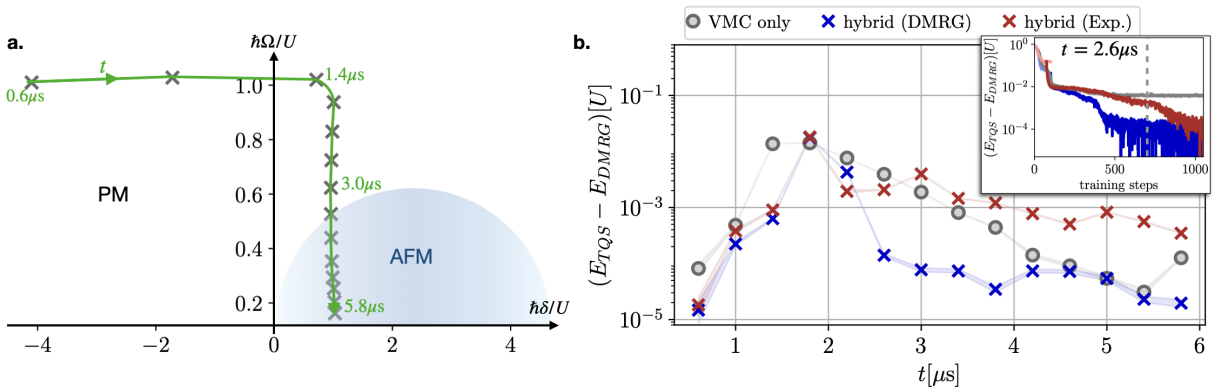

Figure S12: Results for the experimentally realized TFIM (20) on a 10×10 lattice. **a.** Experimentally realized values for $\Omega(t)$ and $\delta(t)$ over time t. During this sweep, the system is transferred from a initial paramagnetic (PM) ground state $|\downarrow\downarrow \ldots \downarrow\rangle$ to an antiferromagnetic (AFM) phase denoted in blue. **b.** Energy error for the Hamiltonians realized at time t without pretraining (gray), with pretraining on numerical data from both Z and X bases (blue) and with pretraining on experimental data from the Z and numerical data from the X basis (red) after 700 optimization steps.

data however, only a slight improvement ($t < 3\mu$s) or even a higher error are observed ($t \geq 3\mu$s). Note that in the latter regime the results obtained with only VMC based training are already below $< 0.1\%$, indicating that in the case of such fast convergence, the hybrid scheme when using imperfect training data can yield slower convergence.

E Training data

E.1 Experimental data

The experimental setup consists of two-dimensional arrays of ^{87}Rb atoms trapped in optical tweezers. The atoms are first loaded into the tweezers and assembled in square arrays of N atoms with a lattice spacing a [51]. We then cool down the atoms to $T \sim 20\,\mu$K, optically pump them in the hyperfine state $|g\rangle = \left|5S_{1/2}, F = 2, m_F = 2\right\rangle$ and switch off the tweezer lights. Starting from this point, two different configurations of the experimental setup allow us to implement either the dipolar XY model [29] or the transverse field Ising one [50].

E.1.1 Dipolar XY model

To implement the XY model, we encode our spin states between two Rydberg states of opposite parities $|\uparrow\rangle = \left|60S_{1/2}, m_J = 1/2\right\rangle$ and $|\downarrow\rangle = \left|60P_{1/2}, m_J = -1/2\right\rangle$. Resonant dipole-dipole interactions between the spins naturally realize the dipolar XY model (2), which we state here again for convenience:

$$\hat{\mathcal{H}}_{\mathrm{XY}} = \frac{J}{2} \sum_{i<j} \frac{a^3}{r_{ij}^3} (\hat{\sigma}_i^x \hat{\sigma}_j^x + \hat{\sigma}_i^y \hat{\sigma}_j^y) \ , \tag{18}$$

with $a = 12\,\mu$m and $J/h = -0.77\,$MHz [52]. The spins can be manipulated using microwaves at $\sim 16.7\,$GHz coupling $|\uparrow\rangle$ to $|\downarrow\rangle$. To prepare the ground state, we apply the following procedure. After having prepared all the spins in $|g\rangle$, we drive a two-photon transition (STIRAP) to excite all atoms in $|\uparrow\rangle$. This step is realized using two laser beams of wavelengths 420 nm and 1013 nm that respectively couple $|g\rangle$ and $|\uparrow\rangle$ to the intermediate state $|i\rangle = \left|6P_{3/2}, F = 3, m_F = 3\right\rangle$. Then, we address half of the atoms in a staggered configuration using off-resonant 1013 nm beams to apply a local light-shift δ on these atoms. The total Hamiltonian thus reads $\hat{\mathcal{H}} = \hat{\mathcal{H}}_{\mathrm{XY}} + \hbar\delta \sum_i \hat{n}_i$ with $\hat{n}_i = 0$ for the non-addressed atoms and $\hat{n}_i = (1 + \sigma^z)/2$ for the addressed atoms. Using these local light shifts in combination with microwave pulses, we prepare the system in a classical Néel state $|\uparrow\downarrow\uparrow\downarrow \cdots\rangle$ which is for $|\hbar\delta| \gg |J|$ the

	10×10
number of transformer layers	2
hidden dimension d_h	32
FFNN dimension (see Fig. S1)	$4d_h$
total number of parameters	$28,800$
pretraining learning rate	$1 \cdot 10^{-3}$
VMC learning rate $l_{\max}$	$2 \cdot 10^{-3}$ [a]
learning rate schedule $l(n)$	$n_{\text{start}} = 300$, $\gamma = 0.9998$
maximal no. of pretraining steps $n^{\max}_{\text{pretrain}}$	100
number of samples in VMC training (part 2)	512
number of samples used for the data points in Fig. S12	100×512
errorbars in Fig. S12	standard deviation of the mean
number of training steps used in Fig. S12	700

[a]Except for $t = 1.8\mu$s, where the pretraining learning rate is decreased by a factor 1/5 to avoid getting stuck in a local minimum.

Table 3: Transformer parameters and training hyperparameters for the 10×10 experimental TFIM systems (20) (Fig. S12). We refer to the total number of training steps including the pretraining as n. Futhermore, we use a exponential decay rate starting at epoch n_{start} with decay rate γ.

ground state of $\hat{\mathcal{H}}$ (see more details in Ref. [29, 53]). We then adiabatically decrease the applied light-shifts as $\delta(t) = \delta_0 e^{-t/\tau}$ (with $\delta_0/(2\pi) \approx 15\,\text{MHz}$ and $\tau = 0.3\,\mu\text{s}$) thus connecting the Néel state to the XY antiferro-magnetic ground state. At the end of the ramp, we readout the state of the atoms. Using an on-resonant 1013 nm light beam, we transfer the $|\uparrow\rangle$ atoms to $|i\rangle$ from which they spontaneously decay in the ground state ($5S_{1/2}$) and turn back on the tweezer lights. The atoms transferred in ($5S_{1/2}$) are trapped and imaged, while the ones left in the Rydberg state are expelled from their traps. Thus, we map the $|\uparrow\rangle$ and $|\downarrow\rangle$ state to the presence or absence of the corresponding atom. Additionally, when we want to measure the spins along x, we rotate them by applying a microwave $\pi/2$-pulse prior to the readout sequence.

In the experiment, also van der Waals interactions are present:

$$\hat{\mathcal{H}}_{\text{vdW}} = \sum_{i<j} \frac{a^6}{r_{ij}^6} \left[U_6^{PP} \hat{P}_i^{\uparrow} \hat{P}_j^{\uparrow} + U_6^{SS} \hat{P}_i^{\downarrow} \hat{P}_j^{\downarrow} + U_6^{SP} \hat{P}_i^{\downarrow} \hat{P}_j^{\uparrow} \right] \tag{19}$$

with the projection operators $\hat{P}_i^{\uparrow/\downarrow} = \frac{1}{2} \pm \hat{S}_i^z$, and $U_6^{PP}/h = -0.008\text{MHz}$, $U_6^{SS}/h = 0.037\text{MHz}$ and $U_6^{SP}/h = -0.0007\text{MHz}$.

Note that in the experiment, 10×10 and 6×7 sites systems were realized. Since the latter is not compatible with our 2×2 patches, see Appendix A, we simulate a 6×6 system and cut off samples and correlation maps for the pretraining.

E.1.2 (Experimentally realized) Transverse field Ising model

To explore the Ising model, we encode our spin states in $|\downarrow\rangle = |g\rangle$ and the Rydberg state $|\uparrow\rangle = \left|75S_{1/2}, m_J = 1/2\right\rangle$ with van der Waals interactions of $U^{\uparrow\uparrow}/h \approx 1.95\,\text{MHz}$ for $a = 10\,\mu\text{m}$. Using the two lasers at 420 nm and 1013 nm, we drive a two-photon transition, thus coupling the spin states with an effective Rabi frequency Ω and detuning δ. The Hamiltonian reads:

$$\hat{\mathcal{H}}_{\text{Exp.TFIM}} = U^{\uparrow\uparrow} \sum_{i<j} \frac{a^6}{r_{ij}^6} \hat{n}_i^{\uparrow} \hat{n}_j^{\uparrow} + \frac{\hbar\Omega}{2} \sum_i \hat{\sigma}_i^x - \hbar\delta \sum_i \hat{n}_i^{\uparrow} \,, \tag{20}$$

with $\hat{n}^{\uparrow} = \frac{1}{2}(1 + \hat{\sigma}_i^z)$ and r_{ij} given in units of the lattice constant. As explained in Sec. D.3 and in Ref. [50], to probe the antiferromagnetic ground state of the TFIM model, we sweep Ω and δ to transfer the system from its initial state $|\downarrow\downarrow\downarrow\downarrow\cdots\rangle$ to the Ising antiferromagnetic ground state. At the end of the sequence, we readout the state of the system by switching back on the tweezer light. The atoms in $|\downarrow\rangle = |g\rangle$ are recaptured and imaged, while the ones that stayed in the Rydberg state $|\uparrow\rangle$ are lost.

E.2 Numerical data from DMRG

For the training with numerical data we use the single-site density matrix renormalization group (DMRG) algorithm [2] implemented in the package SyTen [54, 55].

- *Dipolar XY model.* For the 2D dipolar XY model (2), we explicitly employ $U(1)$ symmetry and use bond dimensions up to $\chi_{\max} = 2048$ both for the 10×10 and 6×6) systems, a tolerance of 10^{-5} and a truncation threshold of 10^{-11}. Further numerical analysis of the dipolar XY model can be found in Ref. [29].
- *Transverse field Ising model.* For the 2D TFIM (1) on the 6×6 lattice, we use the same parameters with $\chi_{\max} = 512$.
- *(Experimentally realized) Transverse field Ising model.* For the experimentally realized TFIM (20), we use bond dimensions up to $\chi_{\max} = 1024$ for the 10×10 systems, a tolerance of 10^{-5} and a truncation threshold of 10^{-11}. The long-range interactions in Eq. (20) are cut off for $r > 4$.