



Self-correcting quantum many-body control using reinforcement learning with tensor networks

Received: 15 February 2022

Accepted: 9 June 2023

Published online: 21 July 2023

Check for updates

Friederike Metz^{1,2,4,5} & Marin Bukov^{2,3}

Quantum many-body control is a central milestone en route to harnessing quantum technologies. However, the exponential growth of the Hilbert space dimension with the number of qubits makes it challenging to classically simulate quantum many-body systems and, consequently, to devise reliable and robust optimal control protocols. Here we present a framework for efficiently controlling quantum many-body systems based on reinforcement learning (RL). We tackle the quantum-control problem by leveraging matrix product states (1) for representing the many-body state and (2) as part of the trainable machine learning architecture for our RL agent. The framework is applied to prepare ground states of the quantum Ising chain, including states in the critical region. It allows us to control systems far larger than neural-network-only architectures permit, while retaining the advantages of deep learning algorithms, such as generalizability and trainable robustness to noise. In particular, we demonstrate that RL agents are capable of finding universal controls, of learning how to optimally steer previously unseen many-body states and of adapting control protocols on the fly when the quantum dynamics is subject to stochastic perturbations. Furthermore, we map our RL framework to a hybrid quantum–classical algorithm that can be performed on noisy intermediate-scale quantum devices and test it under the presence of experimentally relevant sources of noise.

Quantum many-body control is an essential prerequisite for the reliable operation of modern quantum technologies that are based on harnessing quantum correlations. For example, quantum computing often involves high-fidelity state manipulation as a necessary component of most quantum algorithms^{1,2}. In quantum simulation, the underlying atomic, molecular and optical platforms require system preparation to a desired state before its properties can be measured and studied^{3–5}. Quantum metrology relies on the controlled engineering of (critical)

states to maximize the sensitivity to physical parameters^{6,7}. Controlling many-body systems can also be considered in its own right as a numerical tool that offers insights into concepts such as quantum phases and phase transitions⁸. Moreover, it can reveal novel theoretical phenomena such as phase transitions in the control landscape⁹, and bears a direct relation to our understanding of quantum complexity¹⁰.

Compared with single- and few-particle physics, working in the quantum many-body domain introduces the formidable difficulty of

¹Quantum Systems Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan. ²Max Planck Institute for the Physics of Complex Systems, Dresden, Germany. ³Department of Physics, Saint Kliment Ohridski University of Sofia, Sofia, Bulgaria. ⁴Present address: Institute of Physics, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. ⁵Present address: Center for Quantum Science and Engineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. ✉e-mail: friederike.metz@epfl.ch; mgbukov@pks.mpg.de

dealing with an exponentially large Hilbert space. A specific manifestation is the accurate description and manipulation of quantum entanglement shared between many degrees of freedom. This poses a limitation for classical simulation methods, because memory and compute time resources scale exponentially with the system size.

Fortunately, there exists a powerful framework to simulate the physics of one-dimensional (1D) quantum many-body systems, based on matrix product states (MPS)^{11–14}. MPS provide a compressed representation of many-body wave functions and allow for efficient computation with resources scaling only linearly in the system size for area-law entangled states^{15,16}.

While MPS-based algorithms have been used in the context of optimal many-body control to find high-fidelity protocols^{17–20}, the advantages of deep RL for quantum control²¹ have so far been investigated using exact simulations of only a small number of interacting quantum degrees of freedom. Nevertheless, policy-gradient and value-function RL algorithms have recently been established as useful tools in the study of quantum state preparation^{22–39}, quantum error correction and mitigation^{40–43}, quantum circuit design^{44–47}, quantum metrology^{48,49}, and quantum heat engines^{50,51}; quantum reinforcement learning algorithms have been proposed as well^{52–56}. Thus, in times of rapidly developing quantum simulators which exceed the computational capabilities of classical computers⁵⁷, the natural question arises regarding scaling up the size of quantum systems in RL control studies beyond exact diagonalization methods. We discuss a proof-of-principle implementation of the algorithm on noisy intermediate-scale quantum (NISQ) devices for small system sizes.

In this work, we develop a deep RL framework for quantum many-body control, based on MPS in two complementary ways. First, we adopt the MPS description of quantum states: this allows us to control large interacting 1D systems, whose quantum dynamics we simulate within the RL environment. Second, representing the RL state in the form of an MPS, naturally suggests the use of tensors network as (part of) the deep learning architecture for the RL agent, for example, instead of a conventional neural network (NN) ansatz. Therefore, inspired by earlier examples of tensor-network-based machine learning^{58–60}, we approximate the RL agent as a hybrid MPS-NN network, called QMPS. With these innovations at hand, the required computational resources scale linearly with the system size, in contrast to learning from the full many-body wave function. Ultimately, this allows us to train an RL agent to control a larger number of interacting quantum particles, as required by present-day quantum simulators.

As a concrete example, we consider the problem of state preparation and present three case studies in which we prepare different ground states of the paradigmatic mixed-field Ising chain (Fig. 1). We train QMPS agents to prepare target states from a class of initial (ground) states, and devise universal controls with respect to experimentally relevant sets of initial states. In contrast to conventional quantum control algorithms (such as, CRAB and GRAPE^{17,18,61}), once the optimization is complete, RL agents retain information during the training process in form of a policy or a value function. When enhanced with a deep learning architecture, the learned control policy generalizes to states not seen during training. We demonstrate how this singular feature of deep RL allows our agents to efficiently control quantum Ising chains (1) starting from various initial states that the RL agent has never encountered, and (2) in the presence of faulty or noisy controls and stochastic dynamics. Thus, even in analytically intractable many-body regimes, an online RL agent produces particularly robust control protocols.

Quantum many-body control

Consider a quantum many-body system in the initial state $|\psi_i\rangle$. Our objective is to find optimal protocols that evolve the system into a desired target state $|\psi_*\rangle$. We construct these protocols as a sequence of q consecutive unitary operators $U(\tau) = \prod_{j=1}^q U_{t_j}$, where $U_{t_j} \in \mathcal{A}$

are chosen from a set \mathcal{A} . To assess the quality of a given protocol, we compute the fidelity of the evolved state with respect to the target state:

$$F(\tau) = |\langle \psi_* | U(\tau) | \psi_i \rangle|^2. \quad (1)$$

Throughout the study, we focus on spin-1/2 chains of size N with open boundary conditions. The system on lattice site j is described using the Pauli matrices X_j, Y_j, Z_j . As initial and target states we select area-law states, for example, ground states of the quantum Ising model (see section ‘State-informed many-body control’). To control chains composed of many interacting spins, we obtain the target ground state using the density matrix renormalization group (DMRG)^{11,13}, and represent the quantum state as an MPS throughout the entire time evolution (Supplementary Section 2A).

We choose a set of experimentally relevant control unitaries \mathcal{A} , which contains uniform nearest-neighbour spin–spin interactions, and global rotations: $\mathcal{A} = \{e^{\pm i\delta t_x \hat{A}_j}\}$, with

$$\hat{A}_j \in \mathcal{A} = \left\{ \sum_i \hat{X}_i \hat{X}_{i+1}, \sum_i \hat{Y}_i \hat{Y}_{i+1}, \sum_i \hat{Z}_i \hat{Z}_{i+1} \right\}. \quad (2)$$

Two-qubit unitaries are capable of controlling entanglement in the state. Note that MPS-based time evolution is particularly efficient for such locally applied operators and the resulting protocols can be considered as a series of quantum gates.

The time duration (or angle) δt_x of all unitary operators is fixed and slightly different in magnitude for positive and negative generators \hat{A}_j , and kept constant throughout the time evolution. Hence, the problem of finding an optimal sequence reduces to a discrete combinatorial optimization in the exponentially large dimensional space of all possible sequences: for a fixed sequence length q , the number of all distinct sequences is $|\mathcal{A}|^q$; therefore, a brute-force search quickly becomes infeasible and more sophisticated algorithms, such as RL, are needed. By fixing both q and δt_x prior to the optimization, in general, we may not be able to come arbitrarily close to the target state, but these constraints can be easily relaxed.

State-informed many-body control

Our MPS-based RL framework is specifically designed for preparing low-entangled states in 1D, such as ground states of local gapped Hamiltonians. Hence, in the subsequent case studies we consider ground states of the 1D mixed-field Ising model as an exemplary system:

$$\hat{H}_{\text{Ising}} = J \sum_{j=1}^{N-1} \hat{Z}_j \hat{Z}_{j+1} - g_x \sum_{j=1}^N \hat{X}_j - g_z \sum_{j=1}^N \hat{Z}_j, \quad (3)$$

where g_x (g_z) denotes a transverse (longitudinal) field. We work in units of $J = 1$ throughout the rest of this work. In the case of negative interaction strength J and in the absence of a longitudinal field $g_z = 0$, the system is integrable, and has a critical point at $g_x = 1$ in the thermodynamic limit, separating a paramagnetic (PM) from a ferromagnetic (FM) phase (Fig. 1). For $g_z > 0$, the model has no known closed-form expressions for its eigenstates and eigenenergies. In addition, for positive interactions, the phase diagram features a critical line from $(g_x, g_z) = (1, 0)$ to $(g_x, g_z) = (0, 2)$ exhibiting a transition from a PM to an antiferromagnetic (AFM) phase (see Supplementary Section 1 for a brief introduction to quantum many-body physics and phase transitions).

In the rest of this section we will analyse three different control scenarios involving ground states of the mixed-field Ising model. In ‘Universal ground state preparation from arbitrary states’ we consider the problem of universal state preparation for $N = 4$ spins and train a QMPS agent to prepare a specific target ground state starting from

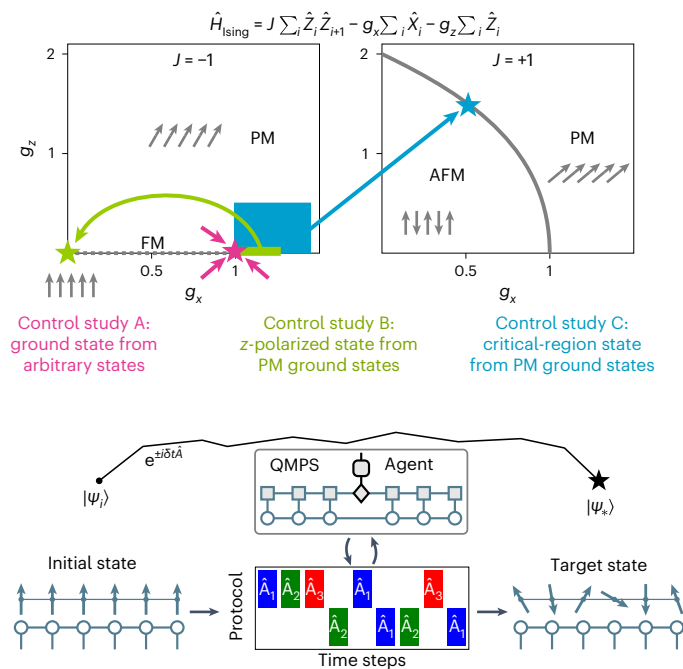


Fig. 1 Many-body control studies in the ground state phase diagram of the quantum Ising model, analysed in this work. An RL agent is trained to prepare a ground state of the transverse-field Ising model from random initial states (control study A, magenta), the z-polarized product state from a class of PM ground states (control study B, green), and a ground state in the critical region of the mixed-field Ising model from PM ground states of opposite interaction strength (control study C, cyan). The optimized agent outputs a control protocol as a sequence of operators \hat{A}_j , which time evolve the initial spin state into the desired target state (marked by a star).

arbitrary initial quantum states. This example will serve as a first benchmark of the QMPS framework. In ‘Preparing product states from PM ground states’ we use the QMPS algorithm to prepare a spin-polarized state starting from a class of PM ground states, which shows that our approach produces reliable protocols in the many-body regime. Finally, in ‘Learning robust critical-region state preparation’ we consider $N = 16$ spins and target a state in the critical region of the mixed-field Ising model, demonstrating that the QMPS framework can also be employed for highly non-trivial control tasks such as critical state preparation. Furthermore, we show that the obtained QMPS agent has the ability to self-correct protocols in the presence of noisy time evolution.

Universal ground state preparation from arbitrary states

In the noninteracting limit, $J = 0$, the QMPS agent readily learns how to control a large number of spins (Supplementary Section 3A1). Instead, as a non-trivial benchmark of the QMPS framework, here we teach an agent to prepare the ground state of the four-spin transverse-field Ising model at $(J = -1, g_x = 1, g_z = 0)$, starting from randomly drawn initial states. While this control setup can be solved using the full wave function and a conventional NN ansatz (Supplementary Section 3A2), the uniform initial state distribution over the entire continuous Hilbert space creates a highly non-trivial learning problem and presents a first benchmark for our QMPS framework. Moreover, system sizes of $N \approx 4$ spins already fall within the relevant regime of most present-day studies using quantum computers, where gate errors and decoherence currently prevent exact simulations at larger scales^{2,62,63}.

We first train an agent (QMPS-1) to prepare the target ground state within 50 protocol steps or less, setting a many-body fidelity threshold of $F^* \approx 0.85$. The initial states during training are chosen to be (with probability $P = 0.25$) random polarized product states,

or (with $P = 0.75$) random reflection-symmetric states drawn from the full $2^4 = 16$ -dimensional Hilbert space by sampling the wave function amplitudes from a normal distribution followed by normalization. In this way, the QMPS-1 agent has to learn to both disentangle highly entangled states to prepare the Ising ground state, but also to appropriately entangle product states to reach the entangled target (the learning curves of the QMPS-1 agent are shown in Supplementary Section 3A2). After this training stage, we test the QMPS-1 agent on a set of 10^3 random initial states and find that in $\sim 99.8\%$ of the cases the fidelity threshold is successfully reached within the 50 allowed steps. A (much) better fidelity cannot be achieved by the QMPS-1 agent alone, due to the discreteness of the action space and the constant step size used, rather than limitations intrinsic to the algorithm. Note that when following a conventional approach of training an NN directly on the quantum wave function, we were not able to match the performance of the QMPS agent given the same number of parameters and training episodes (Supplementary Section 3A2). This suggests that the QMPS architecture has a more natural structure for extracting relevant features from quantum state data and can already be advantageous for small system sizes.

To improve the fidelity between the final and the target state, we now train a second, independent agent (QMPS-2) with a tighter many-body fidelity threshold of $F^* \approx 0.97$. The initial states are again sampled randomly as mentioned above; however, we first use the already optimized QMPS-1 agent to reach the vicinity of the target state within $F > 0.85$. Then, we take those as initial states for the training of the second QMPS-2 agent.

This two-stage learning schedule can, in principle, be continued to increase the fidelity threshold even further. The learning curves of the QMPS-2 optimization are shown in Fig. 2a. In Fig. 2b,c we present the obtained protocols for four exemplary initial states. Overall, the combined two-agent QMPS is able to reach the fidelity threshold of $F^* \approx 0.97$ for approximately 93% of the randomly drawn initial states within the 50 episode steps that were imposed during training. We emphasize that this result is already non-trivial, given the restricted discrete action space, and the arbitrariness of the initial state.

Let us now exhibit two major advantages of RL against conventional quantum control algorithms. (i) After training we can double the allowed episode length for each agent to 100 steps. Since this allows for longer protocols, we find that the target state can be successfully prepared for 99.5% of the initial states (compared to the previously observed 93%). Note that this feature is a unique advantage of (deep) RL methods, where the policy depends explicitly on the quantum state: during training, the agent learns how to take optimal actions starting from any quantum state and hence, it is able to prepare the target state if it is given sufficient time. Moreover, (ii) in this example we achieve universal quantum state preparation, that is, the trained RL agent succeeds in preparing the target state irrespective of the initial state. This is not possible with conventional control techniques where the optimal protocol is usually tailored to a specific initial state, and the optimization has to be rerun when starting from a different state. In ‘Implementation on NISQ devices’, we show how the trained QMPS architecture can be implemented to apply the RL agent on NISQ devices.

Preparing product states from PM ground states

In general, a (Haar) random quantum many-body state is volume-law entangled and, hence, it cannot be approximated by an MPS of a fixed bond dimension. Moreover, it becomes increasingly difficult to disentangle an arbitrarily high entangled state for larger system sizes⁶⁴. Therefore, when working in the truly quantum many-body regime, we have to restrict to initial and target states that are not volume-law entangled.

As an example, here we consider a many-body system of $N = 32$ spins and learn to prepare the z-polarized state from a class of transverse-field Ising ground states $(J = -1, g_z = 0)$. Once high-fidelity

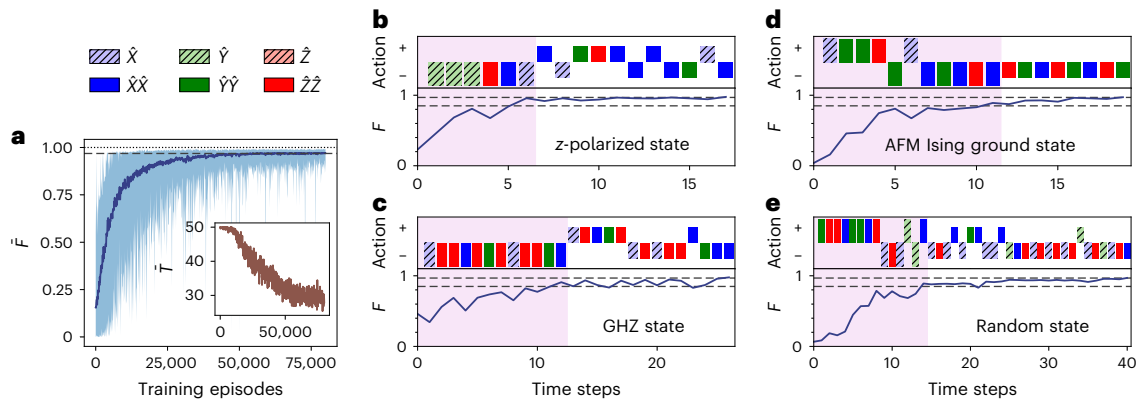


Fig. 2 | Universal four-qubit control. **a**, Achieved many-body fidelity \bar{F} between final and target states during training averaged over 100 training episodes (dark-blue curve). The best and worst fidelity within each episode window is indicated by the light-blue-shaded area. The fidelity threshold, $F^* \approx 0.97$, is marked by a grey dashed line. The inset shows the mean number of episode steps T during training (averaged over 100 episodes). The maximum number of allowed steps is set to 50. **b–e**, Two QMPS agents are trained with fidelity thresholds $F^* \approx 0.85$ and 0.97 (grey dashed lines), to prepare the Ising ground state ($J = -1, g_x = 1, g_z = 0$), and tested starting from the z-polarized product state (**b**), the GHZ state (**c**), an Ising

AFM ground state at $J = +1, g_x = g_z = 0.1$ (**d**), and a random state (**e**). The QMPS-2 agent starts from the final state reached by the QMPS-1 agent (purple shaded area). The many-body fidelity F between the instantaneous and the target state is shown in the lower part of each panel. The upper part of the panels shows the control protocol. The colours and shading of each rectangle indicate the applied action A ; \pm stands for the sign of the action generator, that is, $\exp(\pm i\delta t_{\pm} \hat{A})$. The QMPS-1,2 agents use fixed time steps of $\delta t_{\pm} = (\pi/8, \pi/16)_{\pm}, (\pi/13, \pi/21)_{\pm}$, indicated by action rectangles of different sizes in the protocol. $N = 4$ spins.

protocols are found, they can be inverted to prepare any such Ising ground state from the z-polarized state, which presents a relevant experimental situation (Supplementary Section 3B). Many-body ground state preparation is a prerequisite for both analogue and digital quantum simulation, and enables the study of a variety of many-body phenomena such as the properties of equilibrium and nonequilibrium quantum phases and phase transitions.

To train the agent, we randomly sample initial ground states on the PM side of the critical point: $1.0 < g_x < 1.1$. The difficulty in this state preparation task is determined by the parameter g_x defining the initial state: states deeper into the PM phase are more easy to ‘rotate’ into the product target state, while states close to the critical regime require the agent to learn how to fully disentangle the initial state to reach the target. We train a QMPS agent on a system of $N = 32$ spins, which is infeasible to simulate using the full wave function and is far out of reach for NN-based approaches. We set the single-particle fidelity threshold to $F_{sp}^* = \sqrt[3]{F^*} = 0.99$ (corresponding many-body fidelity $F^* \approx 0.72$) and allow at most 50 steps per protocol.

Figure 3b shows the successfully reached final fidelity when the trained QMPS agent is tested on unseen initial states, for various values of g_x . First, notice that the agent is also able to prepare the target state for initial states with $g_x > 1.1$ that lie outside of the training region (dashed vertical lines). Hence, we are able to extrapolate optimal control protocols well beyond the training data distribution, without additional training. Similar generalization capabilities have already been demonstrated for supervised learning tasks such as Hamiltonian parameter estimation⁶⁵. However, this is not true for states inside the critical region, $g_x \lesssim 1$, and in the FM phase ($g_x \ll 1$); such behaviour is not surprising, because these many-body states have very different properties compared with those used for training. Note that in contrast to the previous control study of ‘Universal ground state preparation from arbitrary states’, the initial training data states are not i.i.d. (independent and identically distributed) over the full 2^{32} dimensional Hilbert space as we only train on PM ground states of the Ising model. Therefore, the agent cannot be expected to generalize to arbitrary initial states in this case. Interestingly, it follows that the onset of criticality can be detected in the structure of control protocols, as the number of required gates (actions) and, in particular, of entangling unitaries, increases rapidly as one approaches the critical point (Fig. 3c).

Discontinuities in the achieved fidelity (Fig. 3b) arise due to the fixed, constant step size δt_{\pm} : we observe distinct jumps in the final fidelity, whenever the length of the protocol sequence increases. This is a primary consequence of the discrete control action space. Its physical origin can be traced back to the need for a more frequent use of disentangling two-site unitaries, for initial states approaching the critical region.

Figure 3a shows the optimal protocol at $g_x = 1.01$: first, the agent concatenates three \hat{Y} -rotations ($\delta t_{\pm} = \pi/12$) in a global gate, which shows that it learns the orientations of the initial x paramagnet and the z-polarized target (yellow shaded region). This is succeeded by a non-trivial sequence containing two-body operators. A closer inspection (Supplementary Fig. 10 and Video 1) reveals that the agent discovered a generalization of Euler-angle rotations in the multi-qubit Hilbert space (blue shaded region). This is remarkable, because it points to the ability of the agent to construct compound rotations, which is a highly non-trivial combinatorial problem for experimentally relevant constrained action spaces. This can be interpreted as a generalization of dynamical decoupling sequences introduced in state-of-the-art nuclear magnetic resonance experiments and used nowadays in quantum simulation, optimal quantum sensing and to protect quantum coherence^{66–68}. We verified that this protocol is a local minimum of the control landscape.

We also investigated the system-size dependence of optimal protocols in this control study. To our surprise, we find that agents trained on the $N = 32$ spin system produce optimal protocols that perform reasonably well on smaller ($N = 8$) as well as larger ($N = 64$) systems. Hence, this control problem admits a certain degree of transferability, which worsens for initial states closer to the finite-size dominated critical region (Supplementary Section 3B).

The MPS-based control framework enables us to readily analyse the physical entanglement growth during training, via the bond dimension of the quantum state χ_{ψ} . The protocol exploration mechanism in QMPS causes the agent to act mostly randomly during the initial stages of learning. This translates to random sequences of unitary gates that can lead to an increase of quantum entanglement (Fig. 3c, inset). In our simulations, we set the maximum allowed bond dimension to $\chi_{\psi} = 16$, which is sufficient for the considered initial and target states to be approximated reliably. However, not all states encountered

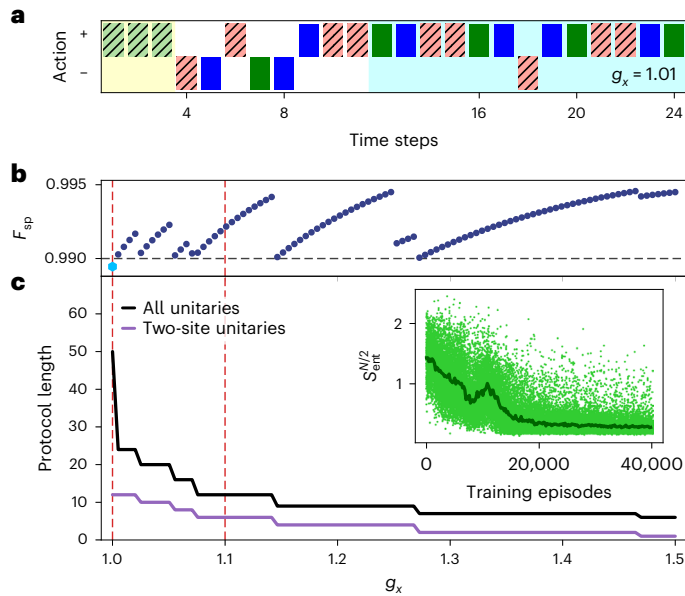


Fig. 3 | Transverse-field Ising control. **a**, Optimal protocol obtained, starting from an initial ground state at $g_x = 1.01$ (see Fig. 2 for action legend). The cyan shaded segment indicates a generalized Euler-angle-like many-body rotation. **b**, Final single-particle fidelities $F_{sp} = \sqrt{F}$ starting from initial ground states with transverse-field value g_x . The target state is the z -polarized product state. The grey dashed line denotes the fidelity threshold ($F_{sp}^* = 0.99$, $F^* \approx 0.72$): it is surpassed for most initial states except at the critical point $g_x = 1$ (cyan dot). The red vertical dashed lines contain the training region. **c**, The number of actions (unitaries) in the QMPS protocols versus the initial state parameter g_x . The protocol starting from the critical state ($g_x \approx 1$) does not reach the fidelity threshold and is truncated after 50 episode steps. Inset: the half-chain von Neumann entanglement entropy of final states during training decreases as learning improves. The dark green curve denotes the average over 200 episodes. $N = 32$ spins (Supplementary Video 1).

can be represented with such a small bond dimension, as reflected by large truncation errors during training (Supplementary Section 3B). Nonetheless, as training progresses, the agent learns to take actions that do not create excessive entanglement (Fig. 3c). Therefore, the truncation error naturally decreases, as training nears completion. As a consequence, the final converged protocols visit states that lie within a manifold of low-entangled states. Moreover, increasing χ_ψ does not change these observations. We believe that this mechanism relies on the area-law nature of the initial and target states, and we expect it to open up the door towards future control studies deeper in the genuine many-body regime.

Learning robust critical-region state preparation

States in the critical region possess non-trivial correlations and show strong system-size dependence, which make manipulating them highly non-trivial. In particular, the required time duration to adiabatically prepare critical states diverges with the number of particles, whereas sweeping through critical points reveals properties of their universality classes⁶⁹. Therefore, finding optimal control strategies away from the adiabatic limit is an important challenge. Critical state preparation is also of practical relevance for modern quantum metrology, where the enhanced sensitivity of critical states to external fields is leveraged to perform more precise measurements⁶.

Our final objective is to prepare a ground state in the critical region of the mixed-field Ising chain ($J = +1$, $g_x = 0.5$, $g_z = 1.5$) starting from non-critical PM ground states of the same model with flipped interaction strength: $J = -1$, $1.0 < g_x < 1.5$, $0 < g_z < 0.5$ (Fig. 1). Hence, the agent has to learn to connect ground states of two distinct

Hamiltonians. This scenario is often relevant in typical experimental setups where only a single-sign interaction strength can be realized: for example, the initial state comes from the $J < 0$ Ising model, while the ground state of interest belongs to the antiferromagnetic Ising Hamiltonian. In general, two completely distinct parent Hamiltonians can be used for the initial and target states, one of which being accessible in the quantum simulator platform at hand, while the other being the object of interest.

We train our QMPS agent on $N = 16$ spins with a single-particle fidelity threshold of $F_{sp}^* = 0.97$ ($F^* \approx 0.61$), and a maximum episode length of 50. Figure 4a shows the achieved fidelity between the target state and the final state, for different initial ground states corresponding to a rectangular region in the (g_x, g_z) -plane. Notice that the agent is able to generalize to unseen initial states lying far outside the training region (white rectangle), and fails only close to the critical point of the transverse-field Ising model ($g_x = 1$, $g_z = 0$) and for a few isolated initial states well outside of the training region.

We now demonstrate that our QMPS agent shows remarkable generalization capabilities in noisy environments. In particular, we analyse how robust the trained QMPS agent is to stochastic perturbations in the time evolution of the state—a common problem in NISQ computing devices⁷⁰. In what follows, we consider two different sources of noise independently: (1) At each time step, with probability ϵ , a random action rather than the selected one is enforced. This type of noise mimics bit- or phase-flip errors, which occur in quantum computing; (2) Gaussian random noise with zero mean and standard deviation σ , is added to the time duration δt_i of each unitary operator; this can, for instance, result from imperfect controls in the experimental platform.

Noise type (1) is equivalent to using an ϵ -greedy policy. Hence, the states encountered when acting with such a policy, could have, in principle, been visited during training. Owing to the generalization capabilities of RL, it is reasonable to expect that an agent will act optimally after non-optimal actions have occurred, attempting to correct the ‘mistake’. In Fig. 4c,d, we show the achieved final fidelity (Fig. 4c) and the required number of steps (Fig. 4d) for $\epsilon = 0.02$. Overall, the fidelity threshold can still be reached in the majority of test cases. The randomness typically results in longer protocols indicating that the agent indeed adapts to the new states encountered. Interestingly, in the noise-free case, the agent fails to prepare the target state for a few points outside the training region (orange points in Fig. 4a); this can be attributed to incorrectly estimated Q -values that have not fully converged to the optimal ones outside of the training interval. However, when adding the perturbation, the agent is able to correct its mistake in one of the shown instances and prepares the target state successfully (Fig. 4c).

Recall that we use a different time step δt_i for positive and negative actions. This way the agent is not just able to undo a non-optimal action by performing its inverse; rather, it has to adjust the entire sequence of incoming unitaries in a non-trivial way. The ability of the QMPS agent to adapt is demonstrated in Fig. 4g where we plot the fidelity during time evolution starting from an arbitrary initial ground state. At time step $t = 5$, we perturb the protocol by taking six different actions, and let the agent act according to the trained policy afterwards; this results in six distinct protocol branches. In each of them, the agent tries to maximize the fidelity and successfully reaches the fidelity threshold after a few extra protocol steps (Supplementary Video 2). In Supplementary Section 3C we provide further examples showing that this behaviour is generic, and can also be observed for different initial states.

In contrast to the ϵ -noise, adding Gaussian random noise (2) with standard deviation σ to the time step duration δt_i , results in states that the agent has not seen during training. This source of noise, therefore, explicitly tests the ability of the agent to generalize beyond the accessible state space, and in particular to interpolate between quantum many-body states. Figure 4e,f displays the achieved fidelity and the corresponding protocol length for $\sigma = 0.01$. We find that the QMPS

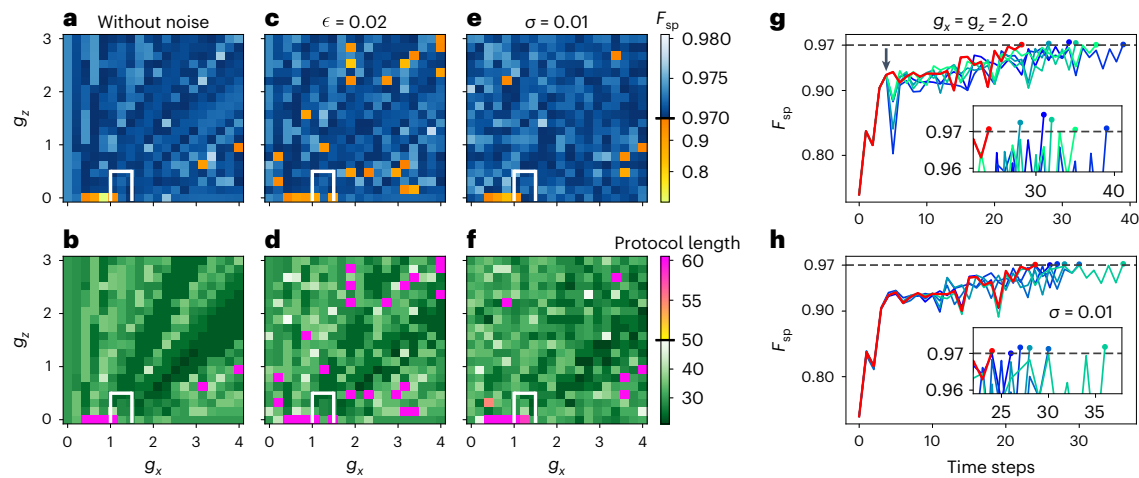


Fig. 4 | Self-correcting mixed-field Ising control. **a, b**, Final single-particle fidelity $F_{sp} = \sqrt[3]{F}$ (**a**) and corresponding protocol length (**b**) versus the initial Ising ground state parameter values g_x and g_z . The target is a state in the critical region of the Ising model at ($J = +1, g_x = 0.5, g_z = 1.5$). Training started only from initial states sampled randomly from the enclosed white rectangle. Each part of the colour bars is shown on a linear scale with the fidelity threshold ($F_{sp}^* = 0.97$, $F^* \approx 0.61$) and the maximum episode length during training (50), indicated by black lines. **c–f**, Same as **a** and **b** but for noisy evolution. **c, d**, At each time step, actions other than the one selected by the agent are taken with probability $\epsilon = 0.02$. **e, f**, White Gaussian noise with standard deviation $\sigma = 0.01$ is added to the time step δt_i of all applied unitaries. **g**, Time-dependence of the single-particle

fidelity starting from an arbitrary initial ground state, and following the trained agent. The red curve denotes the unperturbed (noise-free) QMPS protocol. At time step five (indicated by the black arrow), the QMPS protocol is perturbed by enforcing five suboptimal actions. All subsequent actions are selected again according to the trained QMPS policy without perturbation (blue curves). The inset displays a zoom-in of the vicinity of the fidelity threshold (dashed grey line), showing that each protocol terminates successfully. **h**, Same as **g** but for dynamics subject to Gaussian noise in every time step δt_i , for five different random seeds giving rise to five distinct protocols. $N = 16$ spins. See Supplementary Videos 2 and 3.

agent is also robust to this type of noise. In Fig. 4h we plot the fidelity trajectories starting from the same initial state using five different random seeds; this illustrates that our agent adapts successfully to previously unencountered many-body states, and steers the protocol online to reach beyond the fidelity threshold (Supplementary Video 3).

The robustness of QMPS agents to noise and, in general, to stochasticity in the quantum gates demonstrates yet another advantage of deep RL methods over conventional quantum-control techniques. The latter typically perform suboptimally in noisy systems because the optimization does not take into account the quantum state information during the time evolution, and the optimal protocols are specifically optimized for a fixed trajectory of quantum states²⁷. By contrast, QMPS value functions are optimized on a large class of states and, as shown above, can interpolate and extrapolate to new, seen and unseen states as long as the deep learning approximation stays sufficiently accurate. Therefore, unlike conventional quantum-control algorithms, QMPS agents have the ability to automatically self-correct their protocols on the fly, that is, while the system is being time evolved.

Implementation on NISQ devices

The present QMPS framework requires the quantum state to be accessible at each time step for both training and inference purposes; yet, quantum states are not observable in experiments without performing expensive quantum-state tomography. On the other hand, MPS-based quantum state tomography presents a possible and efficient way of retrieving the quantum state in a form that can be straightforwardly integrated in the QMPS framework^{71–73}. Alternatively, there already exist efficient encoding strategies that map MPS into quantum circuits^{74–80}. Moreover, several proposals were recently developed in which MPS are harnessed for quantum machine learning tasks, for example as part of hybrid classical-quantum algorithms^{81–83} or as classical pre-training methods^{84,85}.

Similar ideas can be applied to the QMPS architecture by mapping the trainable MPS to a parametrized quantum circuit, thus directly

integrating the QMPS framework in quantum computations with NISQ devices and hence, eliminating the need for quantum state tomography. This allows us to perform the expensive training routine on readily available classical computers while the inexpensive inference step can be performed on quantum hardware.

The mapping of the QMPS to a quantum circuit is described in detail in Supplementary Section 4A. Figure 5 shows the resulting QMPS circuit framework for the case of $N = 4$ spins/qubits in which the original QMPS state $|\theta_Q^e\rangle$ is represented as unitary gates (purple boxes). To calculate the Q -values $Q_\theta(\psi, a)$ given an input quantum state $|\psi\rangle$, we first compute the fidelity between the input and the QMPS state

$$|\langle \theta_Q^e | \psi \rangle|^2 = |\langle 0 | U_\theta^\dagger U_\psi | 0 \rangle|^2, \quad (4)$$

which can be obtained via sampling on a quantum computer. Alternatively, the overlap can also be accessed by performing a swap test, albeit requiring additional ancilla qubits and non-local gates^{86,87}. The computed fidelities for each QMPS circuit are then fed into the classical NN giving rise to a hybrid quantum–classical machine learning architecture as shown in Fig. 5. If necessary, the parameters of the QMPS circuit U_θ can be fine-tuned by performing some additional optimization.

We test the QMPS circuit framework on the first control study task ‘Universal ground state preparation from arbitrary states’. In what follows we only report results for the QMPS-1 agent trained on a fidelity threshold of $F^* \approx 0.85$; the generalization to include the QMPS-2 agent is straightforward. We translate the optimized QMPS to the corresponding quantum circuit and investigate the effects of noise in the quantum computation on the QMPS framework.

To simulate incoherent errors, we consider a depolarizing noise channel E :

$$E(\rho) = (1 - \lambda)\rho + \lambda \frac{I}{2^N}, \quad (5)$$

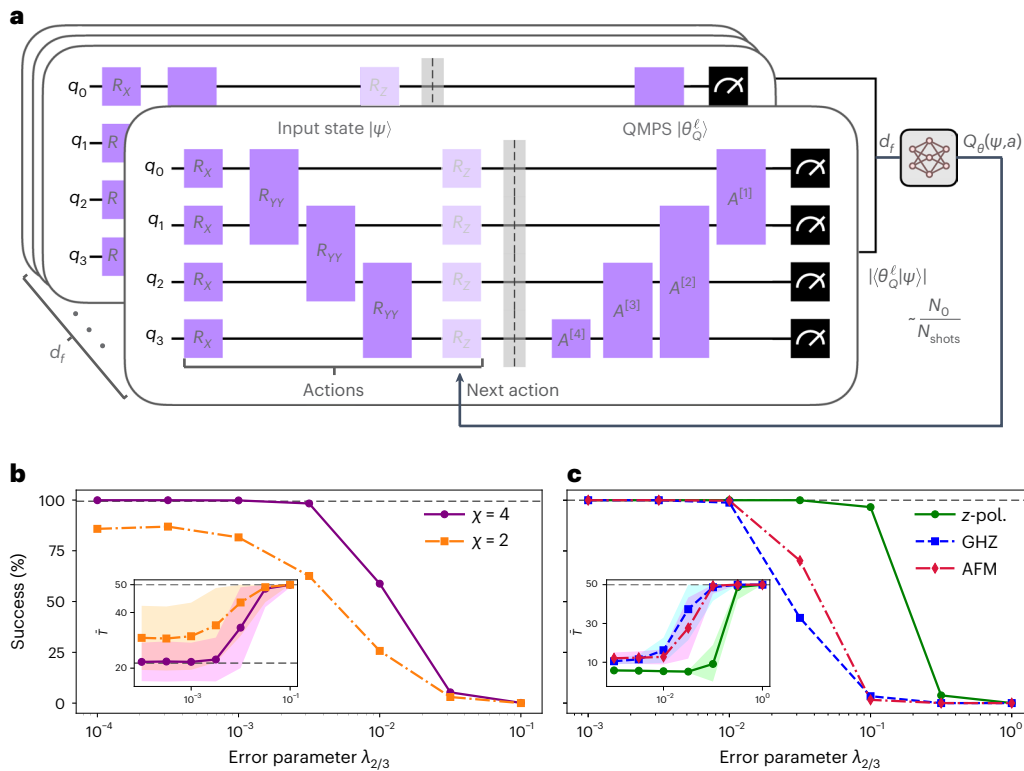


Fig. 5 | QMPS circuit framework and universal four-qubit control. **a**, A hybrid quantum–classical algorithm. On the quantum device, we first prepare the initial state and apply the already inferred protocol actions as gates. In the example above, the initial state is the fully z-polarized state, that is, $|0\rangle^{\otimes 4}$, and two actions are performed: a global rotation around \hat{X} followed by a global two-qubit $\hat{Y}\hat{Y}$ rotation. The resulting state $|\psi\rangle$ represents the input to the QMPS network. The QMPS tensors $\theta_Q = A^{(1)} \dots A^{(n)}$ can be mapped to unitary gates on a quantum circuit. To compute the Q -values, we first apply the inverse of the QMPS circuit unitary U_{θ}^\dagger and measure the output in the computational basis. The fraction of all-zero measurement outcomes is an approximation to the fidelity $|\langle \theta_Q | \psi \rangle|^2$. Note that this denotes the fidelity with respect to the Q -value network state $|\theta_Q\rangle$ and not the target quantum state which is not required during protocol inference. The fidelity estimates are then fed into the NN on a classical computer. From the resulting Q -values we can infer the next action and repeat these steps until the target state is reached. **b**, We sample 1,000 random initial states and apply the QMPS circuit framework in the presence of depolarizing noise with

error parameter $\lambda_{2/3}$ for all two- and three-qubit gates. Note that the noise parameter for all single-qubit gates is always fixed to $\lambda_1 = 10^{-4}$. We set the number of measurement shots to 4,096 and plot the percentage of runs in which the target state is successfully reached against the noise strength. The success rate under exact, noise-free computation (without sampling) is shown as a black dashed line. The success probability when acting completely random is always zero. We provide both, the results for the full $\chi = 4$ QMPS circuit (purple solid line) and the truncated $\chi = 2$ QMPS (orange dash-dotted line). Insets: the corresponding average number of protocols steps \bar{T} as a function of the noise strength λ . The standard deviation is indicated by the shaded areas. **c**, Same as in **b** except that we start each protocol from a fixed initial state: the fully z-polarized state (green solid line), the GHZ state $(|0000\rangle + |1111\rangle)/\sqrt{2}$ (blue dashed line), and a ground state of the mixed-field Ising model at $J = +1, g_x = g_z = 0.1$ (red dashed-dotted line). To compute the success rates and the average protocol length we average over 500 different runs. Note that the x-axis scale is shifted by one order of magnitude compared with **b**.

and apply it after each action and QMPS gate. Here, ρ denotes the quantum state density matrix and λ is the depolarizing noise parameter, which is set to $\lambda_1 = 10^{-4}$ for all single-qubit gates. We plot the success rate as a function of the two- and three-qubit gate errors $\lambda_{2/3}$ for 1,000 randomly sampled initial states in Fig. 5b (purple line). For error rates $\lambda_{2/3} < 10^{-3}$ the QMPS agent is able to successfully prepare the target state in almost all runs. However, the performance deteriorates with increasing error parameter $\lambda_{2/3}$. Let us note that we have used the same error rate for both two- and three-qubit gates. On a physical quantum device, the three-qubit gate will be decomposed into a sequence of two-qubit gates and hence the introduced noise will be amplified. However, the decomposition, the resulting circuit depth and the sources of noise vary with the hardware type. For example, transpiling (that is, decomposing) the QMPS circuit on an IBM Quantum device results in approximately 100 two-qubit gates, while on an IonQ device we require only around 30 two-qubit gates. Thus, we chose the simplified noise model of equation (5) to study the QMPS circuits in a hardware agnostic way.

We also report the results obtained when truncating the bond dimension from a $\chi = 4$ QMPS to a $\chi = 2$ QMPS which gives rise to at most two-qubit gates in the final circuit. In this case, the success probabilities

(orange dashed line in Fig. 5b) do not reach 100% even for small error rates. This indicates that a bond dimension of $\chi = 4$ is indeed required to faithfully represent the QMPS state.

Finally, in Fig. 5c we show the success rates when starting from three different initial states: the fully z-polarized state $|0000\rangle$ (green), the GHZ state $(|0000\rangle + |1111\rangle)/\sqrt{2}$ (blue), and a ground state of the mixed-field Ising model at $J = +1, g_x = g_z = 0.1$ (red). The success probability of unity can be maintained for error rates λ one order of magnitude larger than for the random initial state case. Physical states such as the GHZ or ground states possess only a small amount of entanglement and hence allow us to prepare the target state using a relatively small number of two-qubit gates (Fig. 2b,c). Thus, the resulting protocols are automatically more robust to two-qubit gate errors.

The effect of other decoherence channels (amplitude and phase damping) on the QMPS circuit framework leads to qualitatively similar results and is further discussed in Supplementary Section 4B. Furthermore, we also analyse the self-correcting property of the agent in the presence of coherent gate errors.

Let us briefly discuss the bottlenecks of the current QMPS circuit framework. The QMPS circuit depicted in Fig. 5 is composed of a

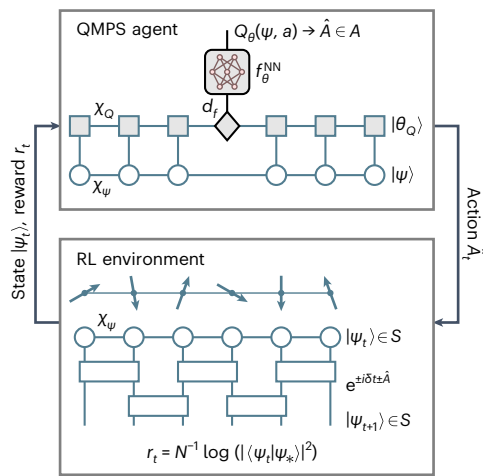


Fig. 6 | Q-learning framework (QMPS) based on matrix product states. The RL environment encompasses a quantum many-body spin chain represented in compressed MPS form which is time evolved according to globally applied unitary operators chosen from a predefined set \mathcal{A} . The reward r_t is given by the normalized log-fidelity between the current state $|\psi_t\rangle$ and the target $|\psi_s\rangle$. The QMPS agent is represented by a parameterized Q -value function $Q_\theta(\psi, a)$ composed of a MPS $|\theta_Q\rangle$ which is contracted with the quantum state MPS $|\psi_t\rangle$, and a subsequent NN which outputs a Q -value for each different action \hat{A} . The trainable parameters of the QMPS are determined by the feature vector dimension d_j and the bond dimension χ_Q .

three-qubit gate and more generally an MPS with bond dimension $\chi = 2^n$ will naturally give rise to $(n + 1)$ -qubit gates. While three-qubit gates will likely be implemented in near-term quantum computers, the gates realized in current NISQ devices are commonly composed of at most two-qubit gates. Hence, any gates acting on more than two qubits would need to be decomposed into two-qubit gates, which usually gives rise to deep circuits. Instead, we can use alternative MPS-to-circuit mappings that would lead to at most two-qubit gates in the final circuit (see Supplementary Section 4A for a detailed discussion)^{74–80}. Finally, the sampling of the fidelity in equation (4) requires a number of measurement shots that could in principle grow exponentially in the system size. One possible solution is to choose a different Q -value network ansatz such as a matrix product operator (Supplementary Section 2D). The resulting computation can then be interpreted as measuring an observable, which can be performed efficiently on larger systems. We find that for the specific $N = 4$ QMPS example, the number of measurements required for successfully preparing the target state can be chosen relatively small, that is, ~ 500 to reach success rates close to unity (Supplementary Fig. 18a).

Discussion

In this work we introduced a tensor network-based Q -learning framework to control quantum many-body systems (Fig. 6). Incorporating an MPS into the deep learning architecture allows part of the Q -value computation to be efficiently expressed as an overlap between two MPS wave functions. As a result, larger system sizes can be reached compared with learning with the full wave function. We emphasize that standard RL algorithms with conventional NN architectures cannot handle quantum many-body states, whose number of components scale exponentially with the number of spins: for example, for $N = 32$ spins, there are $2^{32} \approx 10^{10}$ wave function components to store which is prohibitively expensive. By contrast, our MPS learning architecture only requires linear scaling with the system size N . Furthermore, we found that the hyperparameters of the optimization and, in particular, the number of training episodes do not require finetuning with the

system size, and stayed roughly constant (Supplementary Section 2C). Summarizing, QMPS proposes the use of a tensor-network variational ansatz inspired by quantum many-body physics to offer a novel RL learning architecture.

QMPS-based RL is designed for solving the quantum many-body control problem by learning a value function that explicitly depends on the quantum state. Therefore, a successfully trained QMPS agent is capable of devising optimal protocols for a continuous set of initial states, and selects actions on the fly according to the current state visited. As a result, a QMPS agent has the ability to self-correct mistakes in the protocols when the dynamics is stochastic, even before the protocols have come to an end. Moreover, we illustrated that the agent can interpolate and extrapolate to new quantum states not seen during training. Remarkably, we observed this behaviour over test regions several times the size of the training region. To the best of our knowledge, there does not exist a quantum control algorithm that exhibits such desirable features, as these are based on deep learning capabilities: conventional quantum control algorithms require to rerun the optimization when the initial state has been changed, and thus lack any learning capabilities.

The generalization capabilities, the robustness to noise, and the feasibility of universal state preparation (for small system sizes) are advantages of the QMPS framework over competitive optimal control algorithms. These features are especially relevant for experiments and modern quantum technologies that heavily rely on quantum many-body control, and in particular for NISQ devices. Moreover, we demonstrated that the present QMPS framework can be integrated in quantum device simulations by mapping the optimized MPS ansatz to gates in a quantum circuit. The resulting hybrid quantum–classical algorithm allows us to control quantum states directly on the device without the need of performing expensive quantum state tomography. Thus, unlike NNs, using an MPS learning architecture also facilitates the use of RL agents on NISQ devices.

Our work opens up the door to further research on tensor network-based RL algorithms for quantum (many-body) control. Due to the modular structure of the architecture, the QMPS can be replaced by various tensor networks, such as tree tensor networks⁸⁸ or the multi-scale entanglement renormalization ansatz⁸⁹; these would allow different classes of states to be represented efficiently, and affect the expressivity of the ansatz. Moreover, the infinite-system size description of iMPS can be used to devise control strategies in the thermodynamic limit for which an efficient mapping to quantum circuits exist as well⁷⁹. Similarly, systems with periodic boundary conditions can be studied⁸⁰. Furthermore, tensor networks come with a comprehensive toolbox for analysing their properties, such as the entanglement structure and correlations. Hence, tensor-network-based reinforcement learning will enable us to study the data, the training, and the expressivity of the ansatz using well-understood concepts from quantum many-body physics^{90,91}.

Finally, we mention that it is straightforward to use RL algorithms other than Q -learning in conjunction with our MPS-based ansatz. While we chose the Deep Q -Network (DQN) framework since it is off-policy and, therefore, more data-efficient compared to policy-gradient methods (Supplementary Section 2B), the latter would more naturally allow for continuous action spaces. In turn, with continuous controls, target states can be reached with higher fidelity³¹. One can also adapt the reward function and, for instance, consider the energy density or various distance measures beyond the fidelity.

Methods

RL framework

In RL, a control problem is defined within the framework of an environment that encompasses the physical system to be controlled, and a trainable agent that chooses control actions to be applied to the system (Fig. 6)⁹². The environment is described by a state space \mathcal{S} and a set of physical laws that govern the dynamics of the system. We consider

episodic training, and reset the environment after a maximum number T of time steps. At each time step t during the episode, the agent observes the current state $s_t \in \mathcal{S}$ and receives a scalar reward signal r_t . Depending on the current state s_t , the agent chooses the next action a_{t+1} from a set of allowed actions \mathcal{A} ; this, in turn, alters the state to s_{t+1} . The feedback loop between agent and environment is known as a Markov decision process. The goal of the agent is to find an optimal policy (a function mapping states to actions) that maximizes the expected cumulative reward in any state (Supplementary Section 2B).

States. In our quantum many-body control setting, the RL state space \mathcal{S} comprises all quantum states $|\psi\rangle$ of the 2^N -dimensional many-body Hilbert space. Here, we consider states in the form of an MPS with a fixed bond dimension χ_ψ : if $\chi_\psi < \chi_{\max}$ is smaller than the maximum bond dimension $\chi_{\max} = 2^{N/2}$, long-range entanglement cannot be fully captured, and the resulting MPS becomes a controlled approximation to the true quantum state (Supplementary Section 2A). Hence, state preparation of volume-law entangled states is restricted to intermediate system sizes when using MPS. On the other hand, for large system sizes, the control problems of interest typically involve initial and target states that are only weakly entangled such as ground states of local many-body Hamiltonians. In these cases, the optimal protocol may not create excessive entanglement suggesting that the system follows the ground state of a family of local effective Hamiltonians^{93,94}, similar to shortcuts-to-adiabaticity control⁹⁵, and thus, justifying an MPS-based description.

Actions. If not specified otherwise, the set of available actions \mathcal{A} contains local spin–spin interactions and single-particle rotations, as defined in equation (2).

Rewards. Since our goal is to prepare a specific target state, a natural figure of merit to maximize is the fidelity $F_t = |\langle \psi_t | \psi_* \rangle|^2$ between the current state $|\psi_t\rangle$ and the target state $|\psi_*\rangle$. To avoid a sparse-reward problem caused by exponentially small overlaps in many-body systems, we choose the log-fidelity per spin at each time step as a reward: $r_t = N^{-1} \log(F_t)$. Moreover, we set a fidelity threshold F^* , which the agent has to reach for an episode to be terminated successfully. Note that the agent receives a negative reward at each step; this provides an incentive to reach the fidelity threshold in as few steps as possible, to avoid accruing a large negative return $R = \sum_{t=1}^T r_t$, thus leading to short optimal protocols. For assessing the performance of the QMPS agent to prepare the target state, we show the final single-particle fidelity $F_{\text{sp}} = \sqrt[N]{F}$ as it represents a more intuitive quantity than the related log fidelity used in quantum simulation experiments. A detailed comparison of the control study results in terms of the achieved single- and many-body fidelities can be found in Supplementary Section 3.

In the case where the target state is the ground state of a Hamiltonian H , we can also define the reward in terms of the energy expectation value $E_t = \langle \psi_t | H | \psi_t \rangle$. Specifically, we can choose $r_t = N^{-1}(E_0 - E_t)$, where $E_0 = \langle \psi_* | H | \psi_* \rangle$ is the ground state energy. Similarly to the log-fidelity, this reward is always negative and becomes zero when evaluated on the target ground state. If the target state and therefore also its energy is a priori not known, one can alternatively replace E_0 with a large negative baseline which ensures that the rewards are always negative during training. Another advantage of the energy reward is the fact that expectation values can be efficiently calculated on a quantum device (in contrast to the fidelity). We report results obtained with this reward definition in Supplementary Section 3A2.

Training. Each training episode starts by sampling an initial state followed by taking environment (state evolution) steps. An episode terminates once the fidelity threshold is reached. After every environment step an optimization step is performed (see Supplementary Section 2B for a detailed explanation of the algorithm).

We note in passing that we do not fix the length of an episode (the number of protocol steps) beforehand and the agent is always trying to find the shortest possible protocol to prepare the target state. However, we terminate each episode after a maximum number of allowed steps even if the target state has not been successfully prepared yet: otherwise episodes, especially at the beginning of training, can become exceedingly long leading to unfeasible training times.

Matrix product state ansatz for Q-learning (QMPS)

We choose Q-learning to train our RL agent (Supplementary Section 2B), since it is off-policy and, thus, more data-efficient compared with policy-gradient methods. The optimal Q -function $Q^*(\psi, a)$ defines the total expected return starting from the state $|\psi\rangle$, selecting the action a and then following the optimal protocol afterwards. Intuitively, the optimal action in a given state maximizes $Q^*(\psi, a)$. Hence, if we know $Q^*(\psi, a)$ for every state–action pair, we can solve the control task. In Q -learning this is achieved indirectly, by first finding Q^* . This approach offers the advantage to re-use the information stored in Q^* even after training is complete.

Since the state space is continuous, it becomes infeasible to learn the exact Q^* -values for each state. Therefore, we approximate $Q^* \approx Q_\theta^*$ using a function parametrized by variational parameters θ , and employ the DQN algorithm to train the RL agent⁹⁶. In this work, we introduce an architecture for the Q^* -function, based on a combination of an MPS and an NN, called QMPS, which is specifically tailored for quantum many-body states that can be expressed as a MPS (Fig. 6). We emphasize that the QMPS is independent of the MPS representation of the quantum state, and has its own bond dimension χ_Q .

To calculate $Q_\theta(\psi, a)$ for each possible action a in a quantum state $|\psi\rangle$, we first compute the overlap between the quantum state MPS and the QMPS. The contraction of two MPS can be performed efficiently and scales only linearly in the system size for fixed bond dimensions. The output vector of the contraction corresponding to the dangling leg of the central QMPS tensor, is then interpreted as a feature vector of dimension d_f , which is used as an input to a small fully-connected NN (Fig. 6). Adding a NN additionally enhances the expressivity of the Q_θ^* ansatz by making it nonlinear. The final NN output contains the Q^* -values for each different action.

The QMPS feature vector can be naturally written as an overlap between the quantum state MPS $|\psi\rangle$ and the QMPS $|\theta_Q\rangle$. Thus, the Q^* -value can be expressed as

$$Q_\theta(\psi, a) = f_\theta(N^{-1} \log(|\langle \theta_Q | \psi \rangle|^2)), \quad (6)$$

where $f_\theta(\cdot)$ denotes the NN. We additionally apply the logarithm and divide by the number of spins N to scale the QMPS framework to a larger number of particles. Note also that the QMPS does not represent a physical wave function (it is not normalized); however, for ease of notation, we still express it using the bra-ket formalism.

Thus, the trainable parameters θ of the Q^* -function contain the $N+1$ complex-valued QMPS tensors $|\theta_Q\rangle$, plus the real-valued weights and biases of the subsequent NN. The QMPS feature dimension d_f and the QMPS bond dimension χ_Q are hyperparameters of the optimization, which determine the number of variational parameters of the MPS in analogy to the hidden dimension of NNs. An advantage of the MPS architecture is that we can open up the black box of the ansatz and training using well-understood concepts from the MPS toolbox. For example, it allows us to analyse the correlations in the quantum state and the QMPS by studying its entanglement properties.

Alternatively to the MPS ansatz, we can also represent the parameters of the Q -value network in terms of a matrix product operator $\hat{\theta}_Q$. The Q -value computation then amounts to computing the expectation value of the matrix product operator ansatz with respect to the input quantum state, that is, $Q_\theta(\psi, a) = f_\theta(\langle \psi | \hat{\theta}_Q | \psi \rangle)$. For a detailed explanation of this architecture, we refer to the Supplementary Section 2D.

Furthermore, in Supplementary Section 3A2 we provide a performance comparison of different QMPS/NN architecture choices for the $N = 4$ qubit problem discussed in ‘Universal ground state preparation from arbitrary states’.

Note that the resources (time and memory) for training the QMPS framework scale at worst polynomially in any of the parameters of the system and the ansatz, such as the QMPS bond dimension χ_Q , the feature dimension d_f , and the local Hilbert space dimension $d = 2$. Furthermore, QMPS reduces an exponential scaling of the resources with the system size N to a linear scaling in N , therefore, allowing efficient training on large spin systems.

Data availability

The data used in the figures and the code to generate them are available on GitHub⁹⁷.

Code availability

The software code can be accessed on GitHub⁹⁷.

References

- Farhi, E., Goldstone, J. & Gutmann, S. A quantum approximate optimization algorithm. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.1411.4028> (2014).
- Kandala, A. et al. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature* **549**, 242–246 (2017).
- Lewenstein, M. et al. Ultracold atomic gases in optical lattices: mimicking condensed matter physics and beyond. *Adv. Phys.* **56**, 243–379 (2007).
- Blatt, R. & Roos, C. Quantum simulations with trapped ions. *Nat. Phys.* **8**, 277–284 (2012).
- Casola, F., van der Sar, T. & Yacoby, A. Probing condensed matter physics with magnetometry based on nitrogen-vacancy centres in diamond. *Nat. Rev. Mater.* **3**, 17088 (2018).
- Rams, M. M., Sierant, P., Dutta, O., Horodecki, P. & Zakrzewski, J. At the limits of criticality-based quantum metrology: apparent super-heisenberg scaling revisited. *Phys. Rev. X* **8**, 021022 (2018).
- Pang, S. & Jordan, A. N. Optimal adaptive control for quantum metrology with time-dependent hamiltonians. *Nat. Commun.* **8**, 14695 (2017).
- Matos, G., Johri, S. & Papić, Z. Quantifying the efficiency of state preparation via quantum variational eigensolvers. *PRX Quantum* **2**, 010309 (2021).
- Day, A. G. R., Bukov, M., Weinberg, P., Mehta, P. & Sels, D. Glassy phase of optimal quantum control. *Phys. Rev. Lett.* **122**, 020601 (2019).
- Farhi, E. & Harrow, A. W. Quantum supremacy through the quantum approximate optimization algorithm. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.1602.07674> (2016).
- White, S. R. Density matrix formulation for quantum renormalization groups. *Phys. Rev. Lett.* **69**, 2863 (1992).
- Östlund, S. & Rommer, S. Thermodynamic limit of density matrix renormalization. *Phys. Rev. Lett.* **75**, 3537 (1995).
- Schollwöck, U. The density-matrix renormalization group in the age of matrix product states. *Ann. Phys.* **326**, 96–192 (2011).
- Orús, R. A practical introduction to tensor networks: matrix product states and projected entangled pair states. *Ann. Phys.* **349**, 117–158 (2014).
- Hastings, M. B. An area law for one-dimensional quantum systems. *J. Stat. Mech. Theory Exp.* **2007**, P08024 (2007).
- Schuch, N., Wolf, M. M., Verstraete, F. & Cirac, J. I. Entropy scaling and simulability by matrix product states. *Phys. Rev. Lett.* **100**, 030504 (2008).
- Doria, P., Calarco, T. & Montangero, S. Optimal control technique for many-body quantum dynamics. *Phys. Rev. Lett.* **106**, 190501 (2011).
- van Frank, S. et al. Optimal control of complex atomic quantum systems. *Sci. Rep.* **6**, 34187 (2016).
- Jensen, J. H. M., Møller, F. S., Sørensen, J. J. & Sherson, J. F. Achieving fast high-fidelity optimal control of many-body quantum dynamics. *Phys. Rev. A* **104**, 052210 (2021).
- Luchnikov, I. A., Gavreev, M. A. & Fedorov, A. K. Controlling quantum many-body systems using reduced-order modelling. Preprint at *arXiv* <https://doi.org/10.48550/ARXIV.2211.00467> (2022).
- Krenn, M., Landgraf, J., Foesel, T. & Marquardt, F. Artificial intelligence and machine learning for quantum technologies. *Phys. Rev. A* **107**, 010101 (2023).
- Bukov, M. et al. Reinforcement learning in different phases of quantum control. *Phys. Rev. X* **8**, 031086 (2018).
- Bukov, M. Reinforcement learning for autonomous preparation of floquet-engineered states: inverting the quantum Kapitza oscillator. *Phys. Rev. B* **98**, 224305 (2018).
- Haug, T. et al. Classifying global state preparation via deep reinforcement learning. *Mach. Learn. Sci. Technol.* **2**, 01LT02 (2020).
- Mackeprang, J., Dasari, D. B. R. & Wrachtrup, J. A reinforcement learning approach for quantum state engineering. *Quantum Mach. Intell.* **2**, 5 (2020).
- Niu, M. Y., Boixo, S., Smelyanskiy, V. N. & Neven, H. Universal quantum control through deep reinforcement learning. *npj Quantum Inf.* **5**, 33 (2019).
- Yao, J., Bukov, M. & Lin, L. Policy gradient based quantum approximate optimization algorithm. In *Proc. First Mathematical and Scientific Machine Learning Conference* (eds. Lu, J. & Ward, R.) 605–634 (PMLR, 2020).
- Yao, J., Köttering, P., Gundlach, H., Lin, L. & Bukov, M. Noise-robust end-to-end quantum control using deep autoregressive policy networks. *Proceedings of Machine Learning Research vol 145* 1044–1081 (2022).
- Haug, T., Dumke, R., Kwek, L.-C., Miniatura, C. & Amico, L. Machine-learning engineering of quantum currents. *Phys. Rev. Res.* **3**, 013034 (2021).
- Guo, S.-F. et al. Faster state preparation across quantum phase transition assisted by reinforcement learning. *Phys. Rev. Lett.* **126**, 060401 (2021).
- Yao, J., Lin, L. & Bukov, M. Reinforcement learning for many-body ground-state preparation inspired by counterdiabatic driving. *Phys. Rev. X* **11**, 031070 (2021).
- Bolens, A. & Heyl, M. Reinforcement learning for digital quantum simulation. *Phys. Rev. Lett.* **127**, 110502 (2021).
- He, R.-H. et al. Deep reinforcement learning for universal quantum state preparation via dynamic pulse control. *EPJ Quantum Technol.* **8**, 29 (2021).
- Cao, C., An, Z., Hou, S.-Y., Zhou, D. L. & Zeng, B. Quantum imaginary time evolution steered by reinforcement learning. *Commun. Phys.* **5**, 57 (2022).
- Porotti, R., Peano, V. & Marquardt, F. Gradient ascent pulse engineering with feedback. Preprint at *arXiv* <https://doi.org/10.48550/ARXIV.2203.04271> (2022).
- Porotti, R., Essig, A., Huard, B. & Marquardt, F. Deep reinforcement learning for quantum state preparation with weak nonlinear measurements. *Quantum* **6**, 747 (2022).
- Sivak, V. V. et al. Model-free quantum control with reinforcement learning. *Phys. Rev. X* **12**, 011059 (2022).
- Reuer, K. et al. Realizing a deep reinforcement learning agent discovering real-time feedback control strategies for a quantum system. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2210.16715> (2022).
- Yao, J., Li, H., Bukov, M., Lin, L. & Ying, L. Monte Carlo tree search based hybrid optimization of variational quantum circuits. *Proceedings of Machine Learning Research vol 190* 49–64 (2022).

40. Fösel, T., Tighineanu, P., Weiss, T. & Marquardt, F. Reinforcement learning with neural networks for quantum feedback. *Phys. Rev. X* **8**, 031084 (2018).
41. Nautrup, H. P., Delfosse, N., Dunjko, V., Briegel, H. J. & Friis, N. Optimizing quantum error correction codes with reinforcement learning. *Quantum* **3**, 215 (2019).
42. Andreasson, P., Johansson, J., Liljestrand, S. & Granath, M. Quantum error correction for the toric code using deep reinforcement learning. *Quantum* **3**, 183 (2019).
43. Sweke, R., Kesselring, M. S., van Nieuwenburg, E. P. L. & Eisert, J. Reinforcement learning decoders for fault-tolerant quantum computation. *Mach. Learn. Sci. Technol.* **2**, 025005 (2021).
44. Zhang, Y.-H., Zheng, P.-L., Zhang, Y. & Deng, D.-L. Topological quantum compiling with reinforcement learning. *Phys. Rev. Lett.* **125**, 170501 (2020).
45. Moro, L., Paris, M. G. A., Restelli, M. & Prati, E. Quantum compiling by deep reinforcement learning. *Commun. Phys.* **4**, 178 (2021).
46. He, Z., Li, L., Zheng, S., Li, Y. & Situ, H. Variational quantum compiling with double Q-learning. *New J. Phys.* **23**, 033002 (2021).
47. Fösel, T., Niu, M. Y., Marquardt, F. & Li, L. Quantum circuit optimization with deep reinforcement learning. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2103.07585> (2021).
48. Xu, H. et al. Generalizable control for quantum parameter estimation through reinforcement learning. *npj Quantum Inf.* **5**, 82 (2019).
49. Schuff, J., Fiderer, L. J. & Braun, D. Improving the dynamics of quantum sensors with reinforcement learning. *New J. Phys.* **22**, 035001 (2020).
50. Erdman, P. A. & Noé, F. Driving black-box quantum thermal machines with optimal power/efficiency trade-offs using reinforcement learning. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2204.04785> (2022).
51. Erdman, P. A., Rolandi, A., Abiuso, P., Perarnau-Llobet, M. & Noé, F. Pareto-optimal cycles for power, efficiency and fluctuations of quantum heat engines using reinforcement learning. *Phys. Rev. Res.* **5**, L022017 (2023).
52. Chen, S. Y.-C., Huang, C.-M., Hsing, C.-W., Goan, H.-S. & Kao, Y.-J. Variational quantum reinforcement learning via evolutionary optimization. *Mach. Learn. Sci. Technol.* **3**, 015025 (2022).
53. Lockwood, O. & Si, M. Reinforcement learning with quantum variational circuits. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2008.07524> (2020).
54. Dunjko, V., Taylor, J. M. & Briegel, H. J., Advances in quantum reinforcement learning. In *2017 IEEE International Conference on Systems, Man, and Cybernetics* 282–287 (2017).
55. Jerbi, S., Trenkwalder, L. M., Poulsen Nautrup, H., Briegel, H. J. & Dunjko, V. Quantum enhancements for deep reinforcement learning in large spaces. *PRX Quantum* **2**, 010328 (2021).
56. Saggio, V. et al. Experimental quantum speed-up in reinforcement learning agents. *Nature* **591**, 229–233 (2021).
57. Ebadi, S. et al. Quantum phases of matter on a 256-atom programmable quantum simulator. *Nature* **595**, 227–232 (2021).
58. Stoudenmire, E. & Schwab, D. J. Supervised learning with tensor networks. In *Adv. Neural Information Processing Systems* (eds. Lee, D. et al.) Vol. 29 (Curran Associates, 2016).
59. Han, Z.-Y., Wang, J., Fan, H., Wang, L. & Zhang, P. Unsupervised generative modeling using matrix product states. *Phys. Rev. X* **8**, 031012 (2018).
60. Glasser, I., Pancotti, N. & Cirac, J. I. From probabilistic graphical models to generalized tensor networks for supervised learning. *IEEE Access* **8**, 68169–68182 (2018).
61. Khaneja, N., Reiss, T., Kehlet, C., Schulte-Herbrüggen, T. & Glaser, S. J. Optimal control of coupled spin dynamics: design of NMR pulse sequences by gradient ascent algorithms. *J. Magn. Reson.* **172**, 296–305 (2005).
62. Cervera-Lierta, A. Exact ising model simulation on a quantum computer. *Quantum* **2**, 114 (2018).
63. Lamm, H. & Lawrence, S. Simulation of nonequilibrium dynamics on a quantum computer. *Phys. Rev. Lett.* **121**, 170501 (2018).
64. Poulin, D., Qarry, A., Somma, R. & Verstraete, F. Quantum simulation of time-dependent Hamiltonians and the convenient illusion of Hilbert space. *Phys. Rev. Lett.* **106**, 170501 (2011).
65. Ma, X., Tu, Z. C. & Ran, S.-J. Deep learning quantum states for hamiltonian estimation. *Chin. Phys. Lett.* **38**, 110301 (2021).
66. Choi, J. et al. Robust dynamic Hamiltonian engineering of many-body spin systems. *Phys. Rev. X* **10**, 031002 (2020).
67. Viola, L. Quantum control via encoded dynamical decoupling. *Phys. Rev. A* **66**, 012307 (2002).
68. Haeberlen, U. *High Resolution NMR in Solids: Selective Averaging* (Academic, 1976).
69. Zurek, W. H., Dorner, U. & Zoller, P. Dynamics of a quantum phase transition. *Phys. Rev. Lett.* **95**, 105701 (2005).
70. Preskill, J. Quantum computing in the NISQ era and beyond. *Quantum* **2**, 79 (2018).
71. Baumgratz, T., Gross, D., Cramer, M. & Plenio, M. B. Scalable reconstruction of density matrices. *Phys. Rev. Lett.* **111**, 020401 (2013).
72. Lanyon, B. P. et al. Efficient tomography of a quantum many-body system. *Nat. Phys.* **13**, 1158–1162 (2017).
73. Cramer, M. et al. Efficient quantum state tomography. *Nat. Commun.* **1**, 149 (2010).
74. Barratt, F. et al. Parallel quantum simulation of large systems on small NISQ computers. *npj Quantum Inf.* **7**, 79 (2021).
75. Lin, S.-H., Dilip, R., Green, A. G., Smith, A. & Pollmann, F. Real- and imaginary-time evolution with compressed quantum circuits. *PRX Quantum* **2**, 010342 (2021).
76. Ran, S.-J. Encoding of matrix product states into quantum circuits of one- and two-qubit gates. *Phys. Rev. A* **101**, 032310 (2020).
77. Rudolph, M. S., Chen, J., Miller, J., Acharya, A. & Perdomo-Ortiz, A. Decomposition of matrix product states into shallow quantum circuits. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2209.00595> (2022).
78. Ben Dov, M., Shnaiderov, D., Makmal, A. & Dalla Torre, E. G. Approximate encoding of quantum states using shallow circuits. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2207.00028> (2022).
79. Foss-Feig, M. et al. Entanglement from tensor networks on a trapped-ion quantum computer. *Phys. Rev. Lett.* **128**, 150504 (2022).
80. Wall, M. L., Titum, P., Quiroz, G., Foss-Feig, M. & Hazzard, K. R. A. Tensor-network discriminator architecture for classification of quantum data on quantum computers. *Phys. Rev. A* **105**, 062439 (2022).
81. Huggins, W., Patil, P., Mitchell, B., Whaley, K. B. & Stoudenmire, E. M. Towards quantum machine learning with tensor networks. *Quantum Sci. Technol.* **4**, 024001 (2019).
82. Chen, S. Y.-C., Huang, C.-M., Hsing, C.-W. & Kao, Y.-J. An end-to-end trainable hybrid classical-quantum classifier. *Mach. Learn. Sci. Technol.* **2**, 045021 (2021).
83. Yen-Chi Chen, S., Huang, C.-M., Hsing, C.-W. & Kao, Y.-J. Hybrid quantum-classical classifier based on tensor network and variational quantum circuit. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2011.14651> (2020).
84. Dborin, J., Barratt, F., Wimalaweera, V., Wright, L. & Green, A. G. Matrix product state pre-training for quantum machine learning. *Quantum Sci. Technol.* **7**, 035014 (2022).
85. Wall, M. L., Abernathy, M. R. & Quiroz, G. Generative machine learning with tensor networks: benchmarks on near-term quantum computers. *Phys. Rev. Res.* **3**, 023010 (2021).
86. Buhrman, H., Cleve, R., Watrous, J. & de Wolf, R. Quantum fingerprinting. *Phys. Rev. Lett.* **87**, 167902 (2001).

87. Gottesman, D. & Chuang, I. Quantum digital signatures. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.quant-ph/0105032> (2001).
88. Shi, Y.-Y., Duan, L.-M. & Vidal, G. Classical simulation of quantum many-body systems with a tree tensor network. *Phys. Rev. A* **74**, 022320 (2006).
89. Vidal, G. Entanglement renormalization. *Phys. Rev. Lett.* **99**, 220405 (2007).
90. Martyn, J., Vidal, G., Roberts, C. & Leichenauer, S. Entanglement and tensor networks for supervised image classification. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2007.06082> (2020).
91. Lu, S., Kanász-Nagy, M., Kukuljan, I. & Cirac, J. I. Tensor networks and efficient descriptions of classical data. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2103.06872> (2021).
92. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, 2018).
93. Ljubotina, M., Roos, B., Abanin, D. A. & Serbyn, M. Optimal steering of matrix product states and quantum many-body scars. *PRX Quantum* **3**, 030343 (2022).
94. Lami, G., Torta, P., Santoro, G. E. & Collura, M. Quantum annealing for neural network optimization problems: a new approach via tensor network simulations. *SciPost Phys.* **14**, 117 (2023).
95. Guéry-Odelin, D. et al. Shortcuts to adiabaticity: concepts, methods, and applications. *Rev. Mod. Phys.* **91**, 045001 (2019).
96. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
97. Metz, F. & Bukov, M. Self-correcting quantum many-body control using reinforcement learning with tensor networks. *Zenodo* <https://doi.org/10.5281/zenodo.7950872> (2023).

Acknowledgements

We wish to thank T. Busch and P. Mehta for stimulating discussions. M.B. was supported by the Marie Skłodowska-Curie grant agreement number 890711, and the Bulgarian National Science Fund within National Science Program VIHREN, contract number KP-O6-DV-5 (until 25 June 2021). The authors are pleased to acknowledge that the computational work reported on in this paper was performed on the Max Planck Institute for the Physics of Complex Systems and Okinawa Institute of Science and Technology (OIST) high-performance computing clusters. We are grateful for the help and support provided by the Scientific Computing and Data Analysis section of Research Support Division at OIST.

Author contributions

F.M. and M.B. conceived the research and wrote the manuscript. F.M. performed the numerical simulations and the theoretical analysis. M.B. supervised the work.

Funding

Open access funding provided by Max Planck Society.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42256-023-00687-5>.

Correspondence and requests for materials should be addressed to Friederike Metz or Marin Bukov.

Peer review information *Nature Machine Intelligence* thanks Artem Strashko, Shi-Ju Ran and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023