Reinforcement Learning for Many-Body Ground-State Preparation Inspired by Counterdiabatic Driving

Jiahao Yao,[1,*] Lin Lin,[1,2,3] and Marin Bukov[4,5,†]

[1]*Department of Mathematics, University of California, Berkeley, California 94720, USA*
[2]*Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[3]*Challenge Institute for Quantum Computation, University of California, Berkeley, California 94720, USA*
[4]*Department of Physics, University of California, Berkeley, California 94720, USA*
[5]*Department of Physics, St. Kliment Ohridski University of Sofia, 5 James Bourchier Boulevard, 1164 Sofia, Bulgaria*

(Received 24 November 2020; revised 28 May 2021; accepted 15 July 2021; published 30 September 2021)

The quantum alternating operator ansatz (QAOA) is a prominent example of variational quantum algorithms. We propose a generalized QAOA called CD-QAOA, which is inspired by the counterdiabatic driving procedure, designed for quantum many-body systems and optimized using a reinforcement learning (RL) approach. The resulting hybrid control algorithm proves versatile in preparing the ground state of quantum-chaotic many-body spin chains by minimizing the energy. We show that using terms occurring in the adiabatic gauge potential as generators of additional control unitaries, it is possible to achieve fast high-fidelity many-body control away from the adiabatic regime. While each unitary retains the conventional QAOA-intrinsic continuous control degree of freedom such as the time duration, we consider the order of the multiple available unitaries appearing in the control sequence as an additional discrete optimization problem. Endowing the policy gradient algorithm with an autoregressive deep learning architecture to capture causality, we train the RL agent to construct optimal sequences of unitaries. The algorithm has no access to the quantum state, and we find that the protocol learned on small systems may generalize to larger systems. By scanning a range of protocol durations, we present numerical evidence for a finite quantum speed limit in the nonintegrable mixed-field spin-1/2 Ising and Lipkin-Meshkov-Glick models, and for the suitability to prepare ground states of the spin-1 Heisenberg chain in the long-range and topologically ordered parameter regimes. This work paves the way to incorporate recent success from deep learning for the purpose of quantum many-body control.

DOI: 10.1103/PhysRevX.11.031070 Subject Areas: Quantum Physics

I. INTRODUCTION

The ability to prepare a quantum many-body system in its ground state is an important milestone in the quest for understanding and identifying novel collective quantum phenomena. The degree to which ground states can be confidently prepared in present-day quantum simulators delineates the limits of our capabilities to investigate the properties of new materials or molecules, and to propose innovative technological applications based on quantum effects, such as high-temperature superconductors and superfluids, magnetic field sensors, topological quantum computers, or synthetic molecules.

Quantum simulators—such as ultracold and Rydberg atoms [1,2], trapped ions [3–6], nitrogen vacancy centers [7–9], and superconducting qubits [6,10]—all require the development of state preparation schemes via real-time dynamical processes. Despite their high level of controllability, finding short protocols to prepare strongly correlated ground states under platform-specific constraints is a challenging problem in AMO-based quantum simulation platforms because of the exponentially large Hilbert space dimensions of quantum many-body systems. On this background, speed-efficient protocols also become progressively more important for near-term quantum computing devices [11], where simulation errors grow with the protocol duration due to imperfections in the implementation of the basic gate operations.

Developing versatile methods for ground-state preparation will enable quantum simulators to investigate hitherto unexplored quantum phases of matter and determine the

*jiahaoyao@berkeley.edu
†mgbukov@phys.uni-sofia.bg

$$U(\{\alpha_j\}_{j=1}^{q}, \tau) = \prod_{j=1}^{q} U(\alpha_j, \tau_j)$$

unbounded durations, (ii) it increases the number of optimization parameters α_j, and—with it—the probability to get stuck in a local minimum of the control landscape, and (iii) the condition that nested commutators of H_j span the entire Lie algebra is generally not satisfied for the H_j's of interest in quantum many-body physics due to, e.g., symmetry constraints.

The generalized QAOA ansatz [Eq. (1)] allows us to utilize a larger set of unitaries $\mathcal{A}$ to construct the optimal sequence and to reduce the circuit depth q. Inspired by CD driving, we find that a particularly suitable choice in the context of quantum many-body state manipulation is given by the operators in the adiabatic gauge potential series (Sec. III). Therefore, we call the resulting algorithm CD-QAOA. A different ansatz using more than two unitaries was considered in Ref. [23].

Compared to conventional QAOA, CD-QAOA introduces a discrete high-level optimization to find the optimal protocol sequence τ. The combined optimization landscape can be particularly difficult to navigate because of the existence of so-called barren plateaus where exponentially many directions have vanishing gradients [24–27]. Additionally, the total number of all allowed protocol sequences, $|\mathcal{A}|(|\mathcal{A}|-1)^{q-1}$ [28], scales exponentially with the number of unitaries q and presents a challenging discrete combinatorial optimization problem *per se*; indeed, state preparation, formulated as optimization, can feature a glassy landscape [29,30] (Appendix D). However, overcoming these potential difficulties is associated with a potential gain: CD-QAOA allows us to retain the flexibility offered by continuous optimization while increasing the number of independent discrete control degrees of freedom to $|\mathcal{A}|$; this enables us to reach larger parts of the Hilbert space in shorter durations, and with a smaller circuit depth, as compared to conventional QAOA.

Thus, we formulate ground-state preparation as a two-level optimization scheme [31]. (1) Low-level optimization: Given a fixed sequence τ, we find the optimal values of α_j using a continuous optimization solver, e.g., sequential least squares programming (SLSQP) [33] (Appendix B). To cope with the associated rugged optimization landscape (Appendix D), we run multiple realizations of random initial conditions and postselect the values that yield minimum energy. This continuous optimization problem is also present in conventional QAOA. (2) High-level optimization: In addition to the low-level optimization, we also perform a discrete optimization for the sequence τ itself to determine the optimal order in which unitaries from the set $\mathcal{A}$ should occur. To tackle this combinatorial problem, we formulate the high-level optimization as a RL problem. We learn the optimal protocol using proximal policy optimization (PPO), a variant of policy gradient. The policy is parametrized by a deep autoregressive network, which allows us to choose the control unitaries $U(\alpha_j, \tau_j)$ sequentially. In practice, we sample a batch of sequences from the policy, evaluate the energy of each sequence in the low-level optimization, and apply policy gradient to update the parameters of the policy. This two-level optimization procedure is repeated in a number of training episodes until convergence (Appendix A).

III. VARIATIONAL STATE PREPARATION INSPIRED BY COUNTERDIABATIC DRIVING

A natural question arises as to how to choose the set $\mathcal{A}$ of unitaries for the generalized discrete-continuous QAOA ansatz. One possibility is to consider a set of universal elementary quantum gates, e.g., in the context of a quantum computer [34,35]; in this case, α_j are angles of rotation. We leave this exciting possibility for a future study and focus here on many-body ground-state preparation instead.

The complexity of many-body systems motivates the use of a physics-informed approach to defining the control unitaries in $\mathcal{A}$. Suppose we initialize the system in the ground state of the parent Hamiltonian $H(\lambda = 0)$; we target the ground state of $H(\lambda = 1)$, seeking the functional form of a time-dependent protocol $\lambda(t)$. If the instantaneous ground state of $H(\lambda)$ remains gapped during the evolution, the adiabatic theorem guarantees the existence of a solution $\lambda(t)$, $t \in [0, T]$, provided T is large compared to the smallest inverse gap along the adiabatic trajectory. However, when the gap is known to close (e.g., across a phase transition), or when the state population transfer has to be done quickly, adiabatic state preparation fails.

Compared to the adiabatic paradigm, gauge potentials provide additional control directions in Hilbert space, which enable paths that nonadiabatically lead to the target state in a short time. In many-body systems, it is not known, in general, how to determine the exact gauge potential required for CD driving. However, it is possible to define variational approximations [36,37] using an operator-valued series expansion (Appendix E) similar to a Schrieffer-Wolff transformation [38], or shortcuts to adiabaticity methods [35,39]. Nonetheless, recent numerical simulations suggest that the exact gauge potential in generic many-body systems is a nonlocal operator [36,40] that renders the series expansion asymptotic.

For these reasons, here we consider the constituent terms to every order of the variational gauge potential series, H_j, independently, and use them to generate the set of unitaries $\mathcal{A} = \{e^{-i\alpha_j H_j}\}$ for CD-QAOA [41]. We emphasize that our CD-QAOA ansatz is not designed to approximate the gauge potential itself, as opposed to Ref. [42], yet it yields similar benefits with respect to preparing the target state. In Sec. V, we directly compare our approach with the variational gauge potential ansatz from Ref. [36].

Since CD-QAOA is a generalization of QAOA aimed to be useful in practice, we need to ensure the accessibility of the control terms H_j. Because they appear in the first few

orders of the gauge potential series, H_j are (sums of) *local* many-body operators (cf. Appendix E). Thus, in principle, there is no physical obstruction to emulate them in the lab, although this depends on the details of the experimental platform (especially for the interaction terms). Additionally, in the context of many-body systems where energy is extensive, in order to guarantee that we do not tap into a source of infinite energy, we constrain the norm of the generators $\alpha_j H_j$: We view $\alpha_j \geq 0$ as time durations and fix $\sum_{j=1}^{q} \alpha_j = T$, with T the total protocol duration. This approach keeps α_j on the same order of magnitude as the coupling constants in the parent Hamiltonian whose ground state we want to prepare.

IV. MANY-BODY GROUND-STATE PREPARATION

We consider four nonintegrable many-body systems of increasing complexity: the spin-1/2 and spin-1 mixed-field Ising models, the spin-1 Heisenberg model, and the integrable LMG model, where a large number of degrees of freedom is accessible in a classical simulation. The goal of the RL agent is to prepare the ordered ground states, starting from a product state. To generate training data, we numerically compute the exact time evolution of the system. We apply CD-QAOA using a set of unitaries built from the terms in the series expansion for the variational gauge potential. To determine the allowed terms in the gauge potential series, cf. Table I (lower group), we consider the minimal set of symmetries shared by the Hamiltonian and the initial and target states (Appendix E).

TABLE I. Shorthand notation for the generators H_j used to construct the set of unitaries $\mathcal{A} = \{e^{-i\alpha_j H_j}\}_{j=1}^{|\mathcal{A}|}$ in CD-QAOA. The $|$ indicates operators acting on neighboring sites. Terms from the variational gauge potential series are shown in the lower group (cf. Appendix E for the derivation).

Shorthand notation	Spin operator H_j
X	$\sum_i S_i^x$
Z	$\sum_i S_i^z$
$Z\|Z$	$\sum_i S_i^z S_{i+1}^z$
$Z\|Z+Z$	$\sum_i J S_i^z S_{i+1}^z + h_z S_i^z$
Y	$\sum_i S_i^y$
XY	$\sum_i S_i^x S_i^y + S_i^y S_i^x$
YZ	$\sum_i S_i^y S_i^z + S_i^z S_i^y$
$X\|Y$	$\sum_i S_i^x S_{i+1}^y + S_i^y S_{i+1}^x$
$Y\|Z$	$\sum_i S_i^y S_{i+1}^z + S_i^z S_{i+1}^y$
$X\|Y-XY$	$\sum_i [S_{i+1}^x - a S_i^x] S_i^y + [S_{i+1}^y - a S_i^y] S_i^x$
$Y\|Z-YZ$	$\sum_i [S_{i+1}^z - b S_i^z] S_i^y + [S_{i+1}^y - b S_i^y] S_i^z$
$\hat{XY}$	$(1/N)\sum_{i,j} S_i^x S_j^y + S_i^y S_j^x$
$\hat{ZY}$	$(1/N)\sum_{i,j} [S_i^z + (I/2)] S_j^y + S_i^y [S_j^z + (I/2)]$

A. Mixed-field spin-1/2 Ising chain

First, consider the antiferromagnetic mixed-field spin-1/2 Ising chain of N lattice sites,

$$H = H_1 + H_2,$$
$$H_1 = \sum_{j=1}^{N} J S_{j+1}^z S_j^z + h_z S_j^z, \qquad H_2 = \sum_{j=1}^{N} h_x S_j^x, \quad (3)$$

where $[S_i^\alpha, S_j^\beta] = \delta_{ij} \varepsilon^{\alpha\beta\gamma} S_j^\gamma$ are the spin-1/2 operators. We use periodic boundary conditions and work in the zero momentum sector of positive parity. In the following, $J = 1$ sets the energy unit, and $h_z/J = 0.809$ and $h_x/J = 0.9045$. We initialize the system in the z-polarized product state $|\psi_i\rangle = |\uparrow \cdots \uparrow\rangle$, and we want to prepare the ground state of H in a short time T, i.e., away from the adiabatic regime. We verify that similar results can be obtained starting from $|\downarrow \cdots \downarrow\rangle$.

To acquire an intuitive understanding of the advantages brought about by the gauge potential ansatz, consider first the noninteracting system at $J = 0$, for which the control problem reduces to a single spin. Both the initial and target states lie in the xz plane of the Bloch sphere, and hence, the shortest unit-fidelity protocol generates a rotation about the y axis. In conventional QAOA, one would construct a y rotation out of the X and Z terms (cf. Table I) present in the Hamiltonian. For a single spin, this construction is always possible because of the Euler angle representation of SU(2), but for the interacting spin chain, this is no longer the case. The role of the gauge potential Y is to "unlock" precisely this geodesic in parameter space and make it accessible as a dynamical process. Thus, we can prepare the target state faster, compared to the original X, Z control setup. In the language of variational optimization, an accessible Y term includes the shortest-distance protocol into the variational manifold, and the RL agent easily finds the exact solution (Appendix F 1).

For the interacting system, $J > 0$, applying conventional QAOA using the two gates $U_j = e^{-i\alpha_j H_j}$, with $H_1 = Z|Z + Z$ and $H_2 = X$, is straightforward, but it does not yield a high-fidelity protocol [Fig. 1 (blue squares)]. It was recently reported that much better energies can be obtained using a three-step QAOA, which consists of the three terms in the Hamiltonian (3), $Z|Z$, X, and Z, applied in a fixed order [43]. Invoking, again, a Euler-angles argument provides an explanation: The X and Z terms effectively generate the Y gauge potential term.

In stark contrast to conventional QAOA, adding just the zero-order term $H_3 = Y$ from the gauge potential series (Appendix E 3), we find that CD-QAOA already gives a significantly improved protocol; this is achieved by the high-level discrete optimization, which selects the order of the operators in the sequence. However, we can do better: Since $|\psi_i\rangle$ is a product state while $|\psi_*\rangle$ is not, and because

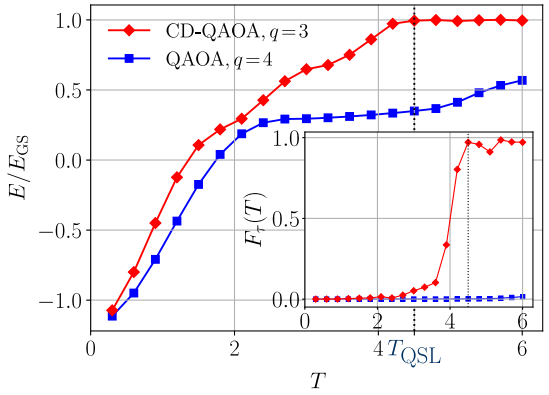

FIG. 1. Spin-1/2 Ising model: energy minimization and the corresponding many-body fidelity (inset) against protocol duration T obtained using conventional QAOA (blue squares) and CD-QAOA (red diamonds) with circuit depths $p = q/2 = 2$ and $q = 3$, respectively. The dotted vertical line marks the quantum speed limit T_{QSL}. CD-QAOA outperforms conventional QAOA. The initial and target states are $|\psi_i\rangle = |\uparrow \cdots \uparrow\rangle$ and $|\psi_*\rangle = |\psi_{GS}(H)\rangle$ for $h_z/J = 0.809$ and $h_x/J = 0.9045$. The alternating unitaries for conventional QAOA are generated by $\mathcal{A}_{QAOA} = \{Z|Z + Z, X\}$[cf. Eq. (3)]; for CD-QAOA, we extend this set using adiabatic gauge potential terms to $\mathcal{A}_{CD-QAOA} = \{Z|Z + Z, X; Y, X|Y, Y|Z\}$. The cardinality of the CD-QAOA sequence space is $|\mathcal{A}|(|\mathcal{A}| - 1)^{q-1} = 80$. The number of spins is $N = 18$, with a Hilbert space size of $\dim(\mathcal{H}) = 7685$.

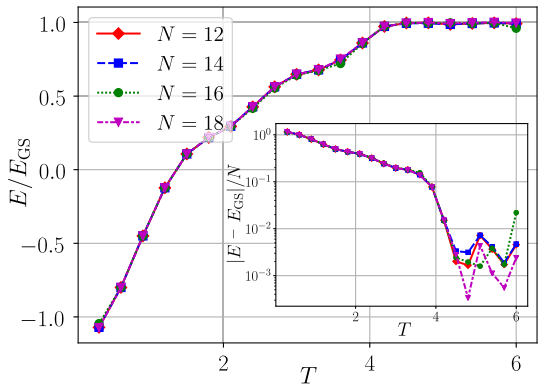

FIG. 2. Spin-1/2 Ising model: energy minimization and the corresponding mean absolute error (inset, log scale) against protocol duration T for different system sizes using CD-QAOA with circuit depths $q = 3$. System-size scaling of the variational energy density suggests the results hold for larger systems. For the number of spins $N = 12, 14, 16, 18$, the Hilbert space sizes are $\dim(\mathcal{H}) = 224, 687, 2250, 7685$, respectively. The model parameters are the same as in Fig. 1.

H_3 is a sum of single-particle terms, in order to create the target many-body correlations using a fast dynamical process, we also include the two-body first-order gauge potential terms $H_4 = X|Y$ and $H_5 = Y|Z$, which results in a nonadiabatic evolution that prepares the interacting ground state to an excellent precision [Fig. 1 (red diamonds)].

In Ref. [44], it was shown that, in the integrable limit $h_z = 0$, one can prepare the ground state of the system at the critical point using a circuit of depth $q = 2N$ with conventional QAOA. Albeit for the specific initial and target states chosen, we find that it only takes CD-QAOA a depth of $q = 3$ to reach the target ground state, independent of the system size N [45]. This result, though model dependent, may come as a surprise at first sight, given that the mixed-field Ising chain is a quantum chaotic system without a closed-form solution, which makes it susceptible to heating away from the adiabatic limit.

Our data also reveal a finite many-body QSL at $T_{QSL} \approx 4.5$. Importantly, this QSL appears insensitive to the system size to a very good approximation (Fig. 2), and we expect it to persist in the thermodynamic limit. The absence of a finite QSL in conventional QAOA in the mixed-field Ising chain suggests that the observation of a QSL using CD-QAOA depends on the specific set of unitaries related to the variational gauge potential, showcasing the utility of our ansatz for many-body control. Remarkably, we find an almost perfect system-size collapse of the target state energy-density curves as a function of the total protocol duration T. In Sec. VI, we explore this feature and demonstrate the ability of the RL agent to learn on small system sizes and, subsequently, generalize its knowledge to control bigger systems with exponentially larger Hilbert spaces.

CD-QAOA performs successfully on the nonintegrable spin-1/2 mixed-field Ising chain, for a circuit depth as short as $q = 3$, which shows an advantage of our ansatz when compared to conventional QAOA. However, the small size of the sequence space, $|\mathcal{A}|(|\mathcal{A}| - 1)^{q-1} = 80$ at $|\mathcal{A}| = 5$, poses a natural question regarding the necessity of using sophisticated search algorithms, such as RL, to find the control sequence. We now show that this is a peculiarity of the physical system, as we turn our attention to a larger sequence space.

B. Heisenberg spin-1 chain

The eight-dimensional spin-1 group SU(3) provides a significantly larger space of gauge potential terms to build the optimal protocol from. We consider a total of $|\mathcal{A}| = 9$ unitaries: Five are generated by the imaginary-valued terms in the gauge potential series $Y, XY, YZ, X|Y, Y|Z$ (cf. Table. I), plus the two real-valued QAOA operators H_1 and H_2, which build the Hamiltonian $H = H_1 + H_2$ whose ground state we target [Eq. (4)], and the two real-valued Hamiltonian terms $X|X$ and Z. At $q = 18$, this amounts to $|\mathcal{A}|(|\mathcal{A}| - 1)^{q-1} \approx 10^{16}$ possible sequences. The exponential scaling of the sequence space size with q renders applying exhaustive search algorithms infeasible and justifies the use of sophisticated algorithms, such as RL.

The (anisotropic) spin-1 Heisenberg model reads as

$$H = H_1 + H_2,$$
$$H_1 = J\sum_{j=1}^{N}(S_{j+1}^x S_j^x + S_{j+1}^y S_j^y), \qquad H_2 = \Delta\sum_{j=1}^{N} S_{j+1}^z S_j^z, \tag{4}$$

with the spin exchange coupling $J = 1$ set as the energy unit, and Δ the anisotropy parameter; we use periodic boundary conditions and work in the ground-state sector of zero momentum and positive parity, defined by the projector $\mathcal{P}$. In the thermodynamic limit, this model features a rich ground-state phase diagram, including ferromagnetic (FM, $\Delta/J \ll -1$), XY ($-1 \lesssim \Delta/J \lesssim 0$), topological/Haldane ($0 \lesssim \Delta/J \lesssim 1$), and antiferromagnetic (AFM, $\Delta/J \gg 1$) order [46], with phase transitions belonging to different universality classes [47–49]. While the FM, XY, and AFM states are characterized by a local order parameter, the gapped Haldane state has topological order not captured by Landau-Ginzburg theory. We consider the AFM initial state $|\psi_i\rangle = \mathcal{P}|\uparrow\downarrow\uparrow\downarrow\cdots\rangle$ and target the ground states of Eq. (4) deep in the FM, XY, and Haldane phases, where system-size effects are the smallest. Because CD-QAOA is not restricted to adiabatic evolution, the conventional paradigm of a closing spectral gap when transferring the population between two states displaying different order does not apply in our nonequilibrium setup, even in the thermodynamic limit.

Figure 3 shows a comparison between conventional QAOA with an alternating sequence between the Hamiltonians H_1 and H_2, and CD-QAOA. We find that CD-QAOA shows superior performance for all three ordered ground states: While the gain over conventional QAOA for the Haldane state is already a faster protocol, we clearly see how the gauge potential terms can prove essential for reaching the ground state in the FM and XY phases within the available durations. Note that the FM target state is doubly degenerate, and minimizing the energy, it ends up in an arbitrary superposition within the ground-state manifold. Interestingly, we do not identify any distinction between preparing states with long-range and topological order, presumably because of the small system sizes that we reach in our classical simulation.

The CD-QAOA protocol sequences found by the RL agent have peculiar structures (Appendix F 2): Some of them closely resemble the alternating sequence of conventional QAOA, with the notable difference of applying additional unitaries to rotate the state to a suitable basis, either at the beginning or at the end of the sequence. While this case is formally equivalent to starting from or targeting a rotated state, the rotations use two-body operators; hence, the resulting basis does not coincide with any of the distinguished S^x, S^y, and S^z directions. Variationally determining such effective bases demonstrates yet another advantage

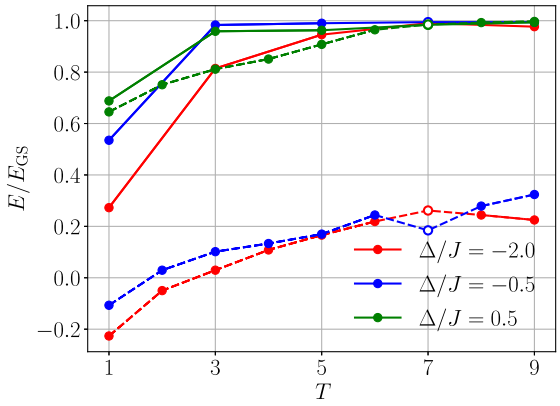

FIG. 3. Heisenberg spin-1 chain: energy minimization against protocol duration T using conventional QAOA (dashed lines) and CD-QAOA (solid lines) for three different states. We start from the AFM state $|\psi_i\rangle = \mathcal{P}|\uparrow\downarrow\uparrow\downarrow\cdots\rangle$ and target three different parameter regimes, corresponding to the FM ($\Delta/J = -2.0$), XY ($\Delta/J = -0.5$), and Haldane ($\Delta/J = 0.5$) states, respectively. CD-QAOA outperforms conventional QAOA ($p = q/2$), more notably in the FM and XY targets where it allows us to reach close to the target state using a short protocol duration. The empty symbols mark the duration at which we show the evolution of the system in Fig. 20. The alternating unitaries for conventional QAOA are generated by $\mathcal{A}_{\mathrm{QAOA}} = \{H_1, H_2\}$ [cf. Eq. (4)]; for CD-QAOA, we extend this set using adiabatic gauge potential terms to $\mathcal{A}_{\mathrm{CD-QAOA}} = \{H_1, H_2, Z, X|X; Y, XY, YZ, X|Y - XY, Y|Z - YZ\}$. The circuit depths are $q = 28$ ($\Delta/J = -2.0$), $q = 18$ ($\Delta/J = -0.5$), and $q = 18$ ($\Delta/J = 0.5$). The cardinality of the CD-QAOA sequence space is $|\mathcal{A}|(|\mathcal{A}| - 1)^{q-1} \approx 10^{16}$ at $q = 18$. The system size is $N = 8$, where $\dim(\mathcal{H}) = 498$.

offered by the CD-QAOA. Another kind of sequence encountered contains two different sets of alternating unitaries, similar to two independent QAOA concatenated one after the other. Finally, for those values of T where CD-QAOA and QAOA have the same performance, we also observe that CD-QAOA finds precisely the QAOA sequence. In this case, conventional QAOA already generates the shortest path, and the extra gauge potential terms, to second order, do not give any advantage; a better performance might be expected when the three- and four-body higher-order terms from the gauge potential series are included.

Similar to other optimal control algorithms, RL agents typically find local minima of the optimization landscape; thus, there is no guarantee that the CD-QAOA protocols provide global optimal solutions. However, these sequences can serve as an inspiration to build future variational ansatzes tailored for many-body systems.

C. Lipkin-Meshkov-Glick model

The nonintegrable character of the previously discussed models precludes us from applying CD-QAOA with a large number of degrees of freedom since reliably simulating their dynamics on a classical computer is prohibitively

expensive. In order to study the behavior of CD-QAOA in a large enough system that also features a quantum phase transition, we now turn our attention to an exactly solvable many-body system.

The LMG Hamiltonian [50] describes spin-1/2 particles on a fully connected graph of N sites:

$$H = H_1 + hH_2,$$
$$H_1 = -\frac{J}{N}\sum_{i,j=1}^{N} S_i^x S_j^x, \qquad H_2 = \sum_{j=1}^{N}\left(S_j^z + \frac{1}{2}\right), \quad (5)$$

where J is the uniform interaction strength and h the external magnetic field. In the thermodynamic limit, $N \to \infty$, the system undergoes a quantum phase transition at $h_c/J = 1$ between a ferromagnetic (FM) ground state in the x-direction for $h/J \ll 1$, and a paramagnetic ground state for $h/J \gg 1$. The spectral gap Δ_{LMG} between the ground state and the excited state closes as $\Delta_{\mathrm{LMG}}(h_c) \sim N^{-1/3}$ at the critical point [51]. The LMG model is within the scope of present-day experiments with ultracold atoms [52,53]; therefore, developing fast ground-state preparation techniques can prove useful in practice.

Defining the total spin operators as $S^\alpha = \sum_{j=1}^{N} S_j^\alpha$, the Hamiltonian takes the form $H = -J/N(S^x)^2 + h(S^z + N/2)$. Hence, the total spin is conserved, $[H, \mathbf{S}\cdot\mathbf{S}] = 0$, and the ground-state symmetry sector contains a total of $N+1$ states, i.e., $\dim(\mathcal{H}) = N+1$, which allows us to simulate large system sizes.

Our goal is, starting from the z-polarized paramagnetic initial state, $|\psi_i\rangle = |\downarrow\downarrow\cdots\rangle$, to target an arbitrary superposition in the doubly degenerate FM ground-state manifold at fixed values of the external field h/J, which controls the magnitude of the transversal fluctuations on top of the ferromagnetic order. Figure 4 shows that the overlap of the initial and target states is vanishingly small in the FM phase, and it quickly approaches unity across the critical point into the paramagnetic phase. Therefore, we choose to prepare ground states in the FM phase, where the problem naturally appears more difficult.

Figure 5 shows a comparison between CD-QAOA and QAOA on the LMG model at $h/J = 0.5$ for $N = 501$ spins (more h/J values are shown in Appendix F 3). First, note the superior performance of CD-QAOA, as compared to conventional QAOA in a range or short durations T in the nonadiabatic driving regime. We applied CD-QAOA with two different sets of generators, $\mathcal{A} = \{H_1, H_2; Y\}$ and $\mathcal{A}' = \{H_1, H_2; Y, \hat{XY}, \hat{ZY}\}$ (cf. Table I), and found that, for the LMG model, the higher-order two-body terms $\hat{XY}, \hat{ZY}$ do not offer any advantage deep in the FM phase. This observation can be understood as follows: To turn the z-polarized initial state into the x ferromagnet, it is sufficient to perform a rotation about the y axis, which coincides precisely with the single-body term in the gauge potential series expansion (cf. Appendix E 1 c). Indeed, for

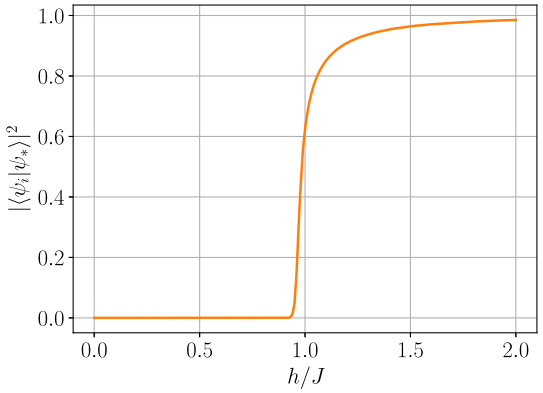

FIG. 4. LMG model: overlap between the initial state $|\psi_i\rangle$ and the target state $|\psi_*\rangle$, which is vanishingly small in the ferromagnetic phase $h/J \ll 1$ and motivates the parameter choice for the target state. In the vicinity of the critical point, the overlap increases and approaches unity in the limit $h/J \to \infty$. Note that, in the FM phase, the ground state is doubly degenerate, in which case the overlap is computed with respect to the ground-state manifold: $|\langle\psi_i|\psi_*^{(1)}\rangle|^2 + |\langle\psi_i|\psi_*^{(2)}\rangle|^2$. In the paramagnetic phase, the ground state is unique. We use $N = 501$ spins.

all protocol durations smaller than the quantum speed limit, $T < T_{\mathrm{QSL}}$, the RL agent finds that the optimal protocol consists of a single Y rotation, while for $T \geq T_{\mathrm{QSL}}$, the optimal protocol is degenerate and typically involves the various terms from $\mathcal{A}$. This finding allows us to extract the QSL as a function of the external field h, cf. Fig. 6.

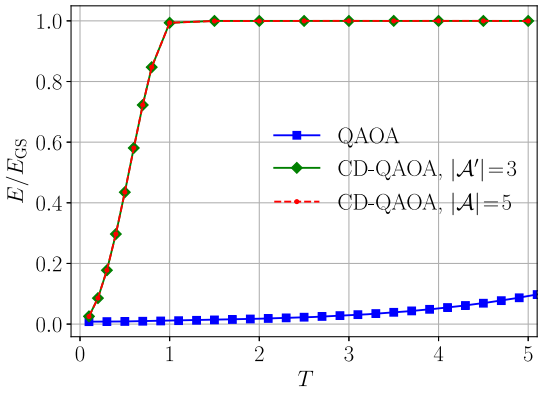

FIG. 5. LMG model: energy minimization against protocol duration T using conventional QAOA (blue square) and CD-QAOA (red dashed line, green solid line). We start from the z-polarized state $|\psi_i\rangle = |\downarrow\downarrow\cdots\rangle$ and target the ground state of LMG Hamiltonian (5). CD-QAOA significantly outperforms conventional QAOA for short durations. The alternating unitaries for conventional QAOA are generated by $\mathcal{A}_{\mathrm{QAOA}} = \{H_1, H_2\}$ [cf. Eq. (5)]; for CD-QAOA, we extend this set using adiabatic gauge potential terms to $\mathcal{A}_{\mathrm{CD-QAOA}} = \{H_1, H_2; Y, \hat{XY}, \hat{YZ}\}$ and $\mathcal{A}'_{\mathrm{CD-QAOA}} = \{H_1, H_2; Y\}$. The external field is $h/J = 0.5$, the circuit depth is $q = 8$, and the system size is $N = 501$, where the effective Hilbert dimension $\dim(\mathcal{H}) = 502$.

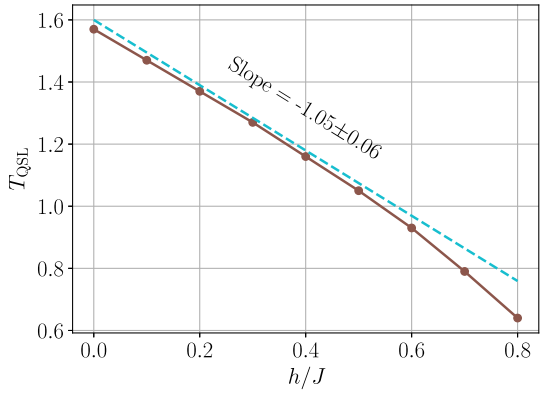

FIG. 6. LMG model: quantum speed limit T_{QSL} as a function of the transverse field h, for a target state in the ferromagnetic phase. At $h/J = 0$, we have $T_{\mathrm{QSL}} = \pi/2$, which is the angle required to turn the z-polarized initial state into the x ferromagnet. For finite h/J, quantum fluctuations in the target ferromagnetic ground state decrease the angle required to transfer the population from the initial state, which results in a smaller value of T_{QSL}. The dashed cyan line is a least-squares fit for small values of h/J, suggesting the behavior $T_{\mathrm{QSL}}(h) = -h/J + \pi/2 + \mathcal{O}(h^2)$. We used $N = 501$ spins.

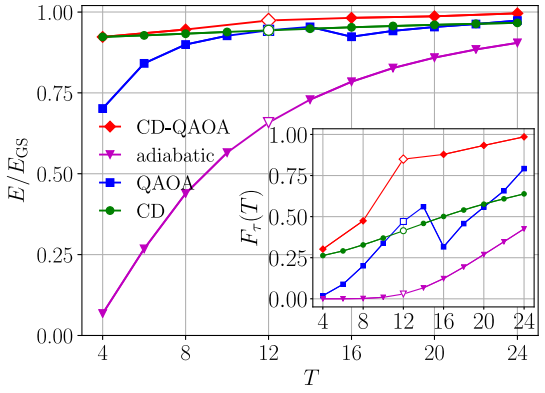

FIG. 7. Spin-1 Ising model: energy minimization and the corresponding many-body fidelity (inset) against different protocol duration T for four different optimization methods—CD-QAOA (red line), conventional QAOA (blue line), variational gauge potential (green line), and adiabatic evolution (magenta line). The empty symbols mark the duration for which the evolution of physical quantities is shown in Fig. 24. The initial and target states are $|\psi_i\rangle = |\downarrow\cdots\downarrow\rangle$ and $|\psi_*\rangle = |\psi_{\mathrm{GS}}(H)\rangle$ for $h_z/J = 0.809$ and $h_x/J = 0.9045$. The alternating unitaries for conventional QAOA are generated by $\mathcal{A}_{\mathrm{QAOA}} = \{H_1, H_2\}$ [cf. Eq. (6)]; for CD-QAOA, we extend this set using adiabatic gauge potential terms to $\mathcal{A}_{\mathrm{CD-QAOA}} = \{H_1, H_2; Y, XY, YZ, X|Y, Y|Z\}$. The variational gauge potential in CD driving uses all five imaginary-valued gauge potentials $\{Y, XY, YZ, X|Y, Y|Z\}$. The CD- and adiabatic-driving simulations are both based on the smooth protocol function $\lambda(t) = \sin^2(\pi t/2T)$, with a time-discretization step $\Delta t = 0.2$. The value of $q = 20$, and the size of sequence space is $|\mathcal{A}|(|\mathcal{A}|-1)^{q-1} \approx 10^{15}$. The system size is $N = 8$, where $\dim(\mathcal{H}) = 498$.

Close to the critical point h_c, we observe strong sensitivity in the best-found protocols to system-size effects, and a single Y rotation is no longer optimal below the QSL. Interestingly, at the critical point (and in the paramagnetic phase), the optimal protocol is given by QAOA. In this regime, despite the larger set of terms $\mathcal{A}$ we use in CD-QAOA, the RL agent correctly identifies the sequence of alternating H_1 and H_2 terms as optimal, which shows the versatility of CD-QAOA; the algorithm can always select a smaller effective subspace of actions when this is advantageous in the parameter regime of interest.

V. COMPARISON WITH COUNTERDIABATIC DRIVING

To compare and contrast the CD-QAOA with CD and adiabatic driving [36], consider the driven spin-1 Ising model [54]:

$$H(\lambda) = \lambda(t)H_1 + H_2,$$
$$H_1 = \sum_{j=1}^{N} JS_{j+1}^z S_j^z + h_x S_j^x, \qquad H_2 = \sum_{j=1}^{N} h_z S_j^z, \quad (6)$$

where $\lambda(t) = \sin^2(\pi t/2T)$, with $t \in [0, T]$, is a smooth protocol satisfying the boundary conditions for CD driving: $\lambda(0) = 0$, $\lambda(T) = 1$, $\dot{\lambda}(0) = 0 = \dot{\lambda}(T)$. The initial state is the ground state at $t = 0$, i.e., $|\psi_i\rangle = |\downarrow\cdots\downarrow\rangle$, while the target state is the ground state of the Ising model at $t = T$ for $h_z/J = 0.809$ and $h_x/J = 0.9045$. Unlike the setup in Sec. IV A, adiabatic state preparation following the protocol $\lambda(t)$ suggests using the QAOA generators $\mathcal{A}_{\mathrm{QAOA}} = \{H_1, H_2\}$.

Figure 7 shows a comparison between different methods using the best-found energy density (main figure) and the corresponding many-body fidelity (inset). Let us focus on CD-QAOA and QAOA first. As expected, CD-QAOA (red) performs better for short durations T since it contains conventional QAOA (red) as an ansatz, i.e., $\mathcal{A}_{\mathrm{QAOA}} \subsetneq \mathcal{A}_{\mathrm{CD-QAOA}}$. We emphasize that such a performance is not guaranteed in practice since it is conceivable that the RL agent gets stuck in a local minimum associated with lower energy than the QAOA solution (Appendix D), e.g., if the deep autoregressive network architecture is not expressive enough or if the learning rate schedules are not well tuned to the problem. Unlike the spin-1/2 Ising model, here we cannot clearly identify a finite QSL, as the CD-QAOA energy keeps improving with increasing circuit depth q (Appendix A).

To construct the counterdiabatic Hamiltonian $H_{\mathrm{CD}} \approx H(\lambda) + \dot{\lambda}\mathcal{X}(\{\beta_j\})$ for Eq. (6), we make a variational ansatz [36] for the gauge potential $\mathcal{X}$ and solve for the optimal parameters β_j numerically (Appendix E). We note the following differences between this approach and

CD-QAOA: (i) The variational gauge potential depends on time t continuously, which requires further discretization when performing a gate-based implementation; (ii) the number of variational parameters in the standard variational gauge potential method is $N_T|\mathcal{A}|$, with N_T the number of steps used to discretize the time interval $[0, T]$ (instead, in CD-QAOA, we have q variational parameters); and (iii) the variational gauge potential method does not constrain the magnitude of the variational coefficients β_j, and hence the time-averaged norm of H_{CD} over the protocol can grow indefinitely (especially for short durations T, this typically gives a higher fidelity). By contrast, in CD-QAOA, the time-averaged norm of the unitary generators $\alpha_j H_j$ summed along the sequence is kept bounded via the constraint $\sum_j \alpha_j = T$. Nonetheless, in practice, we find that these norms are on the same order of magnitude for all methods considered (Appendix F 4).

As anticipated, Fig. 7 shows that CD driving performs better than adiabatic driving, and the two agree in the limit of large T. Moreover, we see explicitly that the CD and QAOA solutions are far from the adiabatic regime. Not surprisingly, CD driving outperforms conventional QAOA for small T, as it can increase the values of the variational parameters (and, with it, the norm) indefinitely. However, CD-QAOA consistently outperforms CD driving in the entire T range; the contrast is especially pronounced in the many-body fidelity (Fig. 7, inset). CD-QAOA makes use of the variational power of QAOA, combining it with physics-motivated input from CD driving.

Table II shows a comparison with the best-obtained energies for $N = 10$ spin-1 particles (qutrits): The superior performance of CD-QAOA remains despite the exponentially growing Hilbert space size. Reaching significantly larger system sizes is infeasible with the present-day computational power: We note that this is a feature of the quantum system rather than a drawback of CD-QAOA, cf. discussion on LMG model in Sec. IV C.

We emphasize that CD-QAOA features some important advantages as compared to CD driving: (1) Because of the nested commutators in the definition of time-ordered exponentials, the QAOA dynamics can effectively implement total unitaries $U(\{\alpha_j\}_{j=1}^{q}, \tau)$ generated by effective nonlocal operators; therefore, CD-QAOA can, in principle, realize a nonlocal effective Hamiltonian as an approximation to the true CD Hamiltonian, thereby overcoming convergence issues related to operator-valued series expansions. (2) CD-QAOA lifts the boundary constraint present in adiabatic and CD driving where the initial and target Hamiltonians are eigenstates of $H(0)$ and $H(1)$, respectively; an interesting open question is whether a local effective Hamiltonian exists, which captures the evolution of the system in this case. Examining the evolution of the entanglement entropy and other local observables induced by the optimal protocol suggests that this is indeed the case (Appendix F 4). (3) One can add any control unitary to the set $\mathcal{A}$, not just terms related to gauge potentials: CD-QAOA has high flexibility to accommodate experimental constraints. (4) Determining the variational gauge potential in CD driving requires using the exact ground state in order to minimize the action, which can be a significant drawback when the ground state is not known or cannot be computed.

TABLE II. Spin-1 Ising model: comparison of the best-obtained energy ratio E/E_{GS} after optimization for four different optimization methods: CD-QAOA, variational CD driving, conventional QAOA, and adiabatic evolution, at $T = 4$, 8, 12 for $N = 10$ qutrits, where $\dim(\mathcal{H}) = 3219$. The remaining setup and parameters are the same as in Fig. 7.

	$E/E_{\mathrm{GS}}[N = 10]$			
T	CD-QAOA	CD	QAOA	Adiabatic
4	**0.943837**	0.923199	0.79534	0.067807
8	**0.961383**	0.933067	0.93386	0.438856
12	**0.990415**	0.942857	0.96275	0.658182

VI. TRANSFER LEARNING AND GENERALIZATION OF THE RL ALGORITHM TO DIFFERENT SYSTEM SIZES

The scale collapse in the energy density of the spin-1/2 Ising model presents a test bed for the transfer learning capabilities of RL. In transfer learning, the RL agent learns to control one physical system and is then used to manipulate another. In our case, the two systems are given by the same Ising model at two different system sizes. Note that transfer learning would not have been possible had we defined the learning problem using the full quantum states because the latter are vectors in Hilbert space whose size grows exponentially with N.

To apply transfer learning, consider first a fixed protocol duration T. For every fixed system size N, we first train a different RL agent. Next, we build the set of protocols across all system sizes (found by these agents) and determine the number of unique protocols (cf. legend in Fig. 8). Finally, we apply all unique protocols to all system sizes available and store the energy densities they result in, which leaves us with a set of energy-density values for every fixed T. The error bars in Fig. 8 show the best and the worst protocols over this set. Observe that, below the QSL, there are only a few points T where the best control protocol is the same across all system sizes. Transfer learning works well, as can be seen by the small error bars. In this regime, the RL agent generalizes its knowledge and learns universal features of the protocol, which are required to control the Ising model. In contrast, for $T > T_{\mathrm{QSL}}$, there are many more protocols giving approximately similar ground-state energies. While the corresponding energies are similar in value, the agent does not generalize. Nevertheless, we checked that, in this regime,

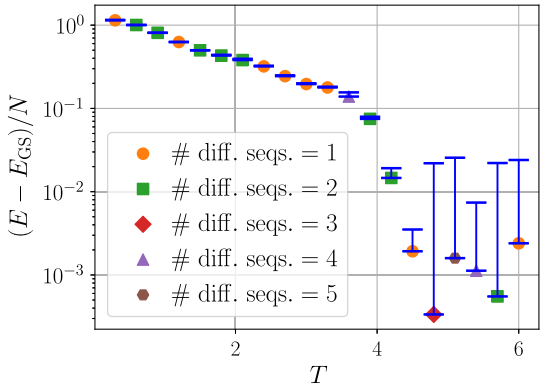

FIG. 8. Spin-1/2 Ising model: protocol generalization across various system sizes. The marker types show the number of different protocols found by the RL agent at a fixed T across all system sizes $N = 6$, 10, 12, 14, 16, and 18. Each protocol is applied to every system size N at a fixed T, which results in a set of cost function values; the error bars designate the range between the largest and smallest cost function values. The parameters are the same as in Fig. 1.

training on smaller system sizes still provides a useful pretraining procedure for learning on larger systems.

VII. DISCUSSION AND OUTLOOK

We analyzed many-body ground-state preparation using unitary evolution in the spin-1/2 Ising model, the spin-1 anisotropic Heisenberg and Ising models, and the fully connected LMG spin-1/2 model. We introduced the CD-QAOA: A RL agent optimizes the order of unitaries in the protocol sequence, generated from terms in the adiabatic gauge potential series, and obtains short high-fidelity protocols away from the adiabatic regime. The resulting algorithm combines the strength of continuous and discrete optimization into a unified and versatile control framework. We found that our CD-QAOA ansatz consistently outperforms both conventional QAOA and variational CD driving across different models and protocol durations. An interesting open question is whether one can use CD-QAOA to find a nonlocal approximation to the variational gauge potential itself, which is beyond the scope of asymptotic series expansions. Another straightforward application of CD-QAOA would be imaginary time evolution [55].

For the nonintegrable spin-1/2 Ising chain, we revealed the existence of a finite quantum speed limit. Moreover, we found a remarkable system-size collapse of the energy curves, suggesting that the sequences found by the agent hold in the thermodynamic limit; this was corroborated by numerical experiments on transfer learning, which demonstrate that one can train the agent on one system size while it generalizes to larger systems. In the Heisenberg spin-1 system, CD-QAOA allows us to prepare long-range and topologically ordered ground states, even when the initial state does not belong to the phase of the target state. The optimal protocols found by the RL agent contain nontrivial basis rotations, intertwined with alternating QAOA-like subsequences, suggesting new ansätze for more efficient variants of CD-QAOA. Numerical studies of nonequilibrium quantum many-body systems, in turn, suffer from limitations related to the exponentially large dimension of the underlying Hilbert space; future work can investigate dynamics beyond exact evolution.

Compared to conventional QAOA, using terms from the variational gauge potential series has higher expressivity, which results in much shorter, yet better performing, circuits. This method can be used, e.g., to reduce the cumulative error in quantum computing devices. However, gauge potential terms are not always easy to realize in experiments since they implement imaginary-valued terms that break time-reversal symmetry; that said, it is often possible to generate such terms using auxiliary real-valued operators via a generalization of the Euler angles or by means of change-of-frame transformations [36]. Moreover, as we have demonstrated, CD-QAOA admits nongauge potential terms as building blocks for control sequences, e.g., universal gate sets. Other experimental constraints, such as the presence of drift terms, which cannot be switched off, can also be conveniently incorporated by redefining the set of unitaries $\mathcal{A}$.

Finally, let us remark that RL provides only one possible set of algorithms to explore the exponentially large space of protocol sequences; in practice, one can apply other discrete optimization techniques, e.g., genetic algorithms and search algorithms like Monte Carlo tree search.

ACKNOWLEDGMENTS

We wish to thank A. Polkovnikov, Dong An, and Yulong Dong for valuable discussions. This work was partially supported by the Department of Energy under Grants No. DE-AC02-05CH11231 and No. DE-SC0017867, by a Google Quantum Research Award (L. L. and J. Y.), and by the NSF Quantum Leap Challenge Institute (QLCI) program through Grant No. OMA-2016245 (L. L.). M. B. was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under the Accelerated Research in Quantum Computing (ARQC) program, the Quantum Algorithm Teams Program, the U.S. Department of Energy under Cooperative Research Agreement No. DE-SC0009919, the Emergent Phenomena in Quantum Systems initiative of the Gordon and Betty Moore Foundation, and the Bulgarian National Science Fund within National Science Program VIHREN, Contract No. KP-06-DV-5. We used SLSQP implemented in SciPy for the QAOA solver, NumPy, and TensorFlow and TensorFlow Probability for the deep learning simulations; we used Quspin for simulating the dynamics of the quantum systems [56,57]. The authors acknowledge that the computational work reported on in this paper was

performed on Savio3 Condo of Berkeley Research Computing (BRC).

APPENDIX A: HIGH-LEVEL OPTIMIZATION: POLICY GRADIENT USING DEEP AUTOREGRESSIVE NETWORKS

Recently, progress made in machine learning (ML) [58–61] has raised the question as to how we can harness such modern advances to improve techniques to manipulate quantum systems. Examples of ML applications include model-based optimization [62], differentiable programming [63] and Bayesian inference [64] quantum control, cavity control [65], designing quantum end-to-end learning schemes [66], and measurement-based adaptation protocols [67], as well as applications to quantum error correction [68,69].

RL algorithms [70,71], such as policy gradient [72–74], Q-learning [75,76], and AlphaZero [77], have recently attracted the attention of physicists—in particular, how they can be combined with physically motivated VQEs for improved performance. In RL, policy gradient has been proposed as an alternative optimizer for QAOA, showcasing the robustness of RL-based optimization to both classical and quantum sources of noise [78]; a related study applied PPO to prepare the ground state of the transverse-field Ising model [79]. The QAOA with policy gradient has been applied to efficiently find optimal variational parameters for unseen combinatorial problem instances on a quantum computer [80]; Q-learning was used to formulate QAOA into a RL framework to solve difficult combinatorial problems [81] and in the context of digital quantum simulation [82].

In the following, we introduce the details of the reinforcement learning algorithm used for the high-level optimization in this work.

1. Reinforcement learning basics

Reinforcement learning comprises a class of machine learning algorithms where an agent learns how to solve a given task through interactions with its environment using a trial-and-error approach [70]. It is based on a Markov decision process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, p, R)$, where $\mathcal{S}$ and $\mathcal{A}$ represent the state and action spaces, $p: \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0,1]$ defines the transition dynamics, and $R: S \times A \to \mathbb{R}$ is the reward function that describes the environment. Let $\pi(a_j|s_j): \mathcal{A} \times \mathcal{S} \to [0,1]$ denote a stochastic policy that defines the probability distribution of choosing an action $a_j \in \mathcal{A}$ given the state $s_j \in \mathcal{S}$. Rolling out the policy $\pi(a_j|s_j)$ in the environment can also be viewed as sampling a trajectory $\tau \sim \mathbb{P}^{\pi}(\cdot)$ from the MDP, where $\mathbb{P}^{\pi}(\tau) = p_0(s_1)\pi(a_1|s_1)p(s_2|s_1, a_1)\cdots\pi(a_q|s_q)$ $p(s_{q+1}|s_q, a_q)$ is the probability for the trajectory τ to occur, q sets the episode or trajectory length, and p_0 is the distribution of the initial state; an example of a trajectory is $\tau = (s_1, a_1, \ldots, a_q, s_{q+1})$. The goal in RL is to find a policy that maximizes the expected return:

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}}\left[\sum_{j=1}^{q} R(s_j, a_j)\right]. \tag{A1}$$

To maximize the expected return $J(\boldsymbol{\theta})$, we use policy gradient—a RL algorithm—which is (i) on-policy (i.e., trajectories have to be sampled from the current policy $\pi_{\boldsymbol{\theta}}$: $\pi = \pi_{\boldsymbol{\theta}}$) and (ii) model-free [i.e., the agent does not need to have a model for the environment dynamics: $p(s'|s, a)$ is assumed to be unknown for the purpose of finding the optimal policy]. Highly expressive function approximators, such as deep neural networks, help parametrize the policy using variational parameters $\boldsymbol{\theta}$. Policy gradient gradually improves the expected return in a number of iterations (or training episodes) by increasing the probability for actions that lead to higher rewards and decreasing the probability for actions that lead to lower rewards, until it reaches a (close to) optimal policy.

We mention, in passing, that we use the terms return and cost function (the latter being the negative of the former) interchangeably: The goal of the RL agent is thus to maximize the expected return or to minimize the cost function.

2. Policy gradient reinforcement learning for quantum many-body systems

Actions.—To apply the reinforcement learning formalism to quantum control, we identify actions taken at each time step within a learning episode, selecting unitaries one at a time within the circuit depth q. Choosing the same unitary at two consecutive time steps is prohibited because the same actions can be merged, resulting in a lower effective circuit depth $q - 1$. At the initial time step $j = 1$, the quantum wave function is given by the initial state $|\psi_i\rangle$; for each intermediate protocol step j, the action $a_j = H_j$ is chosen according to the policy $\pi_{\boldsymbol{\theta}}$. Note that the RL agent only selects the generator H_j out of the set of available actions $\mathcal{A}$ (or, alternatively, which unitary to apply). In other words, unlike Ref. [78], the RL part of CD-QAOA is *not* concerned with finding the corresponding optimal duration α_j; one can think of this low-level continuous optimization as being part of the environment (cf. Appendix B) [83]. At the end of the episode, the quantum state is evolved by applying the entire generated circuit $U(\{\alpha_j\}_{j=1}^{q}, \tau)$ to the initial quantum state $|\psi_i\rangle$.

States.—Since the initial state $|\psi_i\rangle$ is fixed and thus the quantum state at any time step j is uniquely determined by the previous actions taken, here we represent the RL state by concatenating all the previous actions up to step j [85]. One reason for this approach is that, in many-body quantum systems, the number of components in the quantum state scales exponentially with the system size

N, which quickly leads to a computational bottleneck for the simulation on classical computers. A second advantage of this choice is that the first layer of the underlying deep neural network architecture, which parametrizes the policy, will not depend on the system size N either, which allows the algorithm to handle a large number of degrees of freedom. Using the quantum state would not be viable on quantum computers either because quantum states are unphysical mathematical constructs that cannot be measured. Therefore, we can simplify the form of trajectories to consist of actions only, e.g., $\tau = (a_1, a_2, \ldots, a_q)$.

Rewards.—The reward $R_j = R(s_j, a_j)$ is chosen as the negative energy density at the end of the episode:

$$R_j = \begin{cases} 0 & \text{if } j < q \\ -E(\{\alpha_j\}_{j=1}^{q}, \tau)/N & \text{if } j = q. \end{cases}$$

We use energy density since it is an intensive quantity that has a well-defined limit by increasing the number of particles N. In all figures, we show the relative energy E/E_{GS} for clarity (the ground-state energy E_{GS} is typically negative in our models), but the RL agent is always trained with the (negative) energy density $-E/N$. Rewards can also be other observables or nonobservable quantities, such as the overlap squared between two quantum states (fidelity) or the entanglement entropy.

Notice that the reward is sparse: Only at the end of the episode is the negative energy density given as a reward; there is no instantaneous reward during the sequence (and thus we can use the terms reward and total return interchangeably). This case is motivated by the quantum nature of the control problem, where a projective measurement results in a wave-function collapse.

3. Policy parametrization using an autoregressive neural network

An essential part of the policy gradient algorithm is the definition of the policy $\pi_{\boldsymbol{\theta}}$. It is common to parametrize the policy with a highly expressive function approximator, such as a neural network. In our setup, we use a deep autoregressive network, which has recently been used in physics applications of learning to generate samples from free energies in statistical mechanics models [86], and variational approximators for quantum many-body states [87]. This architecture is selected to incorporate causality by factorizing the total probability into a product of conditional probabilities:

$$\pi_{\boldsymbol{\theta}}(a_1, a_2, \ldots, a_q) = \pi_{\boldsymbol{\theta}}(a_1) \prod_{j=2}^{q} \pi_{\boldsymbol{\theta}}(a_j | a_1, \ldots, a_{j-1}), \quad \text{(A2)}$$

where the marginal distribution $\pi_{\boldsymbol{\theta}}(a_1)$ and the conditional distribution $\pi_{\boldsymbol{\theta}}(a_j|a_1, \ldots, a_{j-1})$ are discrete categorical distributions over $\mathcal{A}$. This kind of parametrization explicitly tells how the actions taken in the earlier steps of an episode affect the actions selected later on during the same episode. Such a causal requirement would not be necessary had we used the full quantum state, which would make the dynamics of the environment Markovian. Each of the conditional probabilities in Eq. (A2) can be modeled explicitly using the autoregressive neural network architecture, which naturally allows the policy to depend on all the previous actions only. The structure of the policy network is shown in Fig. 9, the sampling of the autoregressive policy is depicted in Fig. 10, and the hyperparameters of the algorithm (including the number of parameters) are given in Table III.

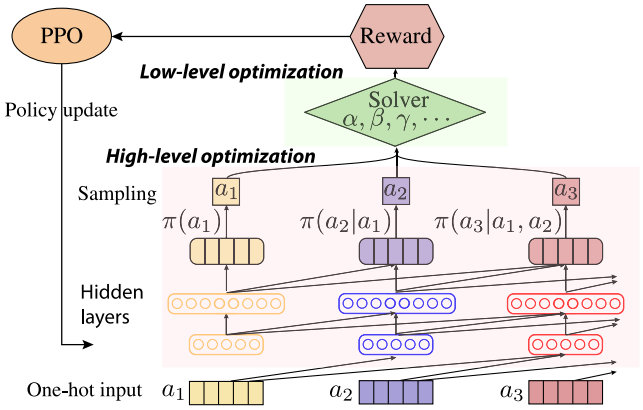

FIG. 9. Schematics of CD-QAOA with an autoregressive policy network. The ancestral sampling procedure used for training is displayed in Fig. 10. The details of the network structure and its training hyperparameters are shown in Table III.

4. Training procedure: Proximal policy optimization

In each iteration of the policy gradient algorithm, a batch of sampled trajectories $\{\tau^k\} = \{(a_1^k, \ldots, a_q^k)\}_{k=1}^{M}$ are rolled out (i.e., sampled) from the current policy, where M is the batch or sample size. Then, the return $R(\tau^k)$ corresponding to trajectory τ^k is computed as

$$R(\tau^k) = \sum_{j=1}^{q} R_j^k = -E(\{\alpha_j^k\}_{j=1}^{q}, \tau^k)/N.$$

To compute the energies, we use the low-level optimization to determine the best-estimated values of α_j, given a sequence τ (see Appendix B). To minimize the chance of getting stuck in a suboptimal local minimum, each sequence is evaluated multiple times, starting from a different initial realization for the α_j optimizer, and the best result is selected (Appendix D).

For every iteration, we can define three quantities for a fixed batch of samples: (i) mean reward (over the current batch), (ii) max reward (over the current batch), and (iii) historically best (best-encountered reward over all the previous iterations). These quantities measure the

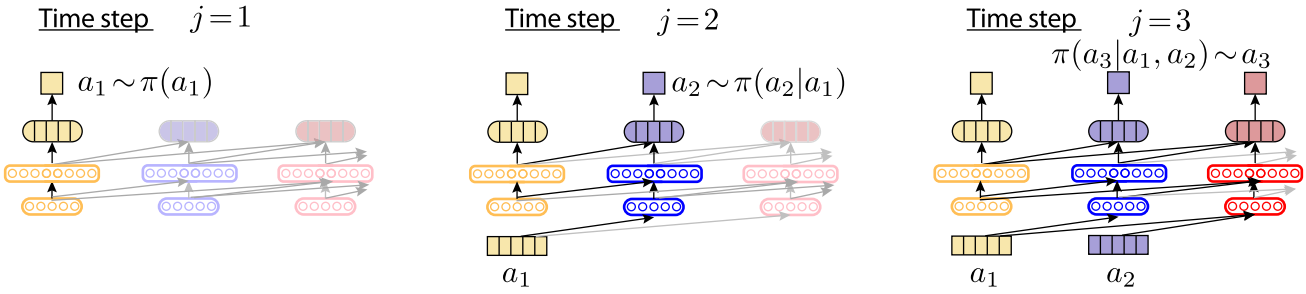

FIG. 10. Exact sampling algorithm for CD-QAOA with an autoregressive policy network, where faded nodes and connections represent unused nodes and connections. The action at each time step is generated sequentially by computing its respective conditional categorical distribution and sampling according to that. Notice that only a single column is processed at each time step, and in order to sample a complete sequence of actions in an episode, one needs to make a forward pass through the network architecture q times.

performance of the learned policy and are shown in Fig. 11. Figure 12 shows the scaling of these quantities for the spin-1 Ising chain, as a function of the episode length q. The performance of CD-QAOA increases because the action space for a larger value of q always contains, as a subset, the action space for a smaller q.

In order to improve the policy represented by the autoregressive network, the RL algorithm interacts with the quantum environment by querying the reward for samples from the current policy. Each trajectory is assigned a reward, once the simulation of the quantum dynamics is complete (note that the simulation may be more expensive if evaluated on a quantum computer). Thus, it is advantageous to reduce the sample size needed to learn the policy, i.e., to improve the sample efficiency.

TABLE III. Hyperparameter values for training the autoregressive deep learning model. In the case of $|\mathcal{A}_{\text{CD-QAOA}}| = 9, q = 18$ [cf. Eq. (3)], the total number of parameters is 24 431; for $|\mathcal{A}_{\text{CD-QAOA}}| = 7, q = 20$ [cf. Eq. (7)], the total number of parameters is 21 985.

Parameter	Value
Optimizer	Adam [88]
Learning rate ($\eta_{\{0\}}$)	1×10^{-2}
Learning rate decay steps	50
Learning rate decay factor	0.96
Learning rate decay style	Staircase
RL temperature ($\beta^{-1}_{S,\{0\}}$)	1×10^{-1}
RL temperature decay steps	10
RL temperature decay factor	0.9
RL temperature decay style	Smooth
Baseline exponential moving decay factor (m)	0.95
Gradient steps (PPO)	4
Clip parameter ϵ	0.1
Number of hidden layers	2
Number of hidden units per layer (d_{hidden})	112
Nonlinearity	ReLU
Number of samples per minibatch (M)	128

The vanilla policy gradient method is known for its poor data efficiency. Thus, we adopt PPO [89], a more robust and sample-efficient policy-gradient-type algorithm. To be more specific, we use the following clipped objective function:

$$\mathcal{G}(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}_t}}[\min\{\rho_{\boldsymbol{\theta}}(\tau)A_{\boldsymbol{\theta}_t}(\tau), \text{clip}(\rho_{\boldsymbol{\theta}}(\tau), 1-\epsilon, 1+\epsilon)A_{\boldsymbol{\theta}_t}(\tau)\}]. \quad \text{(A3)}$$

Here, $\tau = (a_1, a_2, \ldots, a_q)$ is the action sequence sampled from the previous policy $\pi_{\boldsymbol{\theta}_t}$ (cf. Algorithm 1). Typically, the policy from the last iteration is chosen to be the old

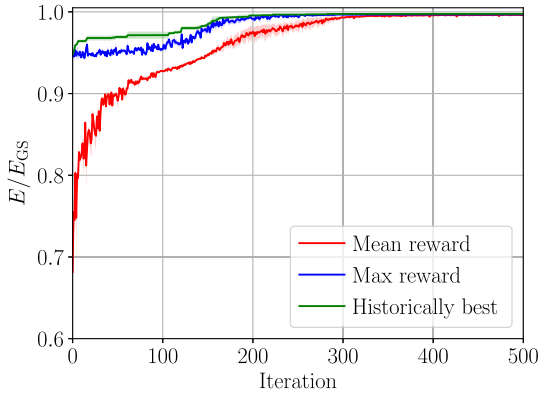

FIG. 11. Spin-1 Ising model: training curves for CD-QAOA with energy minimization as a cost function. The mean negative energy density (red) is computed for a sample generated using the policy at the current iteration; max (blue) is the maximum within the sample; the historically best (green) is the best-encountered policy during the entire training process (i.e., considering all iterations). Each curve shows the average out of three simulations corresponding to three different seed values for the high-level RL optimization; the fluctuations around the seed averages are shown as a narrow shaded area. The total duration is $T = 28$, and the number of spin-1 particles is $N = 8$. The initial and target states are $|\psi_i\rangle = |\downarrow\cdots\downarrow\rangle$ and $|\psi_*\rangle = |\psi_{\text{GS}}(H)\rangle$ for $h_z/J = 0.809$ and $h_x/J = 0.9045$. The CD-QAOA action space is $\mathcal{A}_{\text{CD-QAOA}} = \{Z|Z+X, Z; Y, XY, YZ, X|Y, Y|Z\}$, and we use $q = 20$.

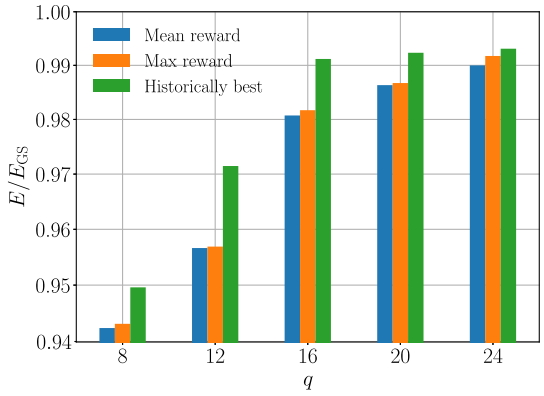

FIG. 12. Spin-1 Ising model: energy minimization against different circuit depths q using CD-QAOA. The mean negative energy density (blue) is computed for a sample generated using the final, learned policy; max (orange) is the maximum within the sample; the historically best (green) is the best encountered policy during the entire training process (i.e., considering all iterations). The total duration $T = 20$, and the values of q range from 8 to 24. The other model parameters are the same as in Fig. 11.

policy; $\rho_{\boldsymbol{\theta}}(\cdot) = [\pi_{\boldsymbol{\theta}}(\cdot)/\pi_{\boldsymbol{\theta}_t}(\cdot)]$ is the importance sampling weight between the new policy $\pi_{\boldsymbol{\theta}}$ and the old policy $\pi_{\boldsymbol{\theta}_t}$; and $A_{\boldsymbol{\theta}_t}(\tau) = R(\tau) - b$ is the advantage function, where b is called a baseline—the advantage measures the reward gain of choosing a specific action with respect to the baseline. For example, a simple baseline can be the average reward, e.g., $b = \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}_t}}[R(\tau)]$, and then the advantage measures how much better (or worse) an action is with respect to the average; in the numerical experiments, we use an exponential moving average (cf. Appendix A 5 for details).

Further, the clip function,

$$\mathrm{clip}(r, x, y) = \max\big(\min(r, x), y\big),$$

clips the value of r within the interval $[x, y]$, which is used to restrict the likelihood ratio in the range $[1-\epsilon, 1+\epsilon]$; this prevents the policy update from deviating too much from the old policy after one gradient update. The clipped objective function is designed to improve the policy as well as to keep it within some vicinity of the last iteration, hence the name proximal policy optimization.

We update the network parameters $\boldsymbol{\theta}$ by ascending along the gradient of the RL objective $\mathcal{G}(\boldsymbol{\theta})$. To provide intuition about the PPO objective, consider the following limiting case. If we only have the first term in the objective, i.e., $\mathcal{G}_1(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}_t}}[\rho_{\boldsymbol{\theta}}(\tau)A_{\boldsymbol{\theta}_t}(\tau)]$, we obtain the following gradient:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{G}_1(\boldsymbol{\theta}) &= \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}_t}}[\nabla_{\boldsymbol{\theta}}\rho_{\boldsymbol{\theta}}(\tau)A_{\boldsymbol{\theta}_t}(\tau)] \\ &= \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}_t}}\left[\frac{\nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}(\tau)}{\pi_{\boldsymbol{\theta}_t(\tau)}}A_{\boldsymbol{\theta}_t}(\tau)\right].\end{aligned}$$

Since we are taking the gradient with respect to $\boldsymbol{\theta}$, it will pass through $\pi_{\boldsymbol{\theta}_t}$ and $A_{\boldsymbol{\theta}_t}(\tau)$. Furthermore, whenever the parameters $\boldsymbol{\theta} \approx \boldsymbol{\theta}_t$, the gradient above is identical to the policy gradient:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{G}_1(\boldsymbol{\theta}) &\approx \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}}}\left[\frac{\nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}(\tau)}{\pi_{\boldsymbol{\theta}(\tau)}}A_{\boldsymbol{\theta}}(\tau)\right] \\ &= \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}}}[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\tau)A_{\boldsymbol{\theta}}(\tau)].\end{aligned}$$

However, PPO performs multiple gradient updates on the sampled data, rendering policy learning more sample efficient [89].

a. Incentivizing exploration using entropy

Maintaining a balance between exploration and exploitation is another major challenge for the reinforcement learning algorithm. Too much exploration prevents the agent from adopting the best strategy it knows so far; on the contrary, too much exploitation limits the agent from attempting new actions and achieving a potentially higher reward. Therefore, it is more appropriate for the agent to explore substantially in the initial iterations of the training procedure and to gradually switch over to exploitation towards the end of the training procedure.

In order to incentivize the agent to explore the action space at the beginning of training, we include an entropy "bonus" [90,91] to the PPO objective from Eq. (A3). To do this, consider the maximal-entropy objective, where the agent aims to maximize the sum of the total reward and the policy entropy S [cf. Eq. (A5)]:

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}) &= G(\boldsymbol{\theta}) + \beta_S^{-1}\mathcal{S}(\pi_{\boldsymbol{\theta}}) \\ &= \mathbb{E}_{\tau=(a_1,\ldots,a_q)\sim\pi_{\boldsymbol{\theta}_t}}\left[\min\{\rho_{\boldsymbol{\theta}}(\tau)A_{\boldsymbol{\theta}_t}(\tau), \mathrm{clip}(\rho_{\boldsymbol{\theta}}(\tau), 1-\epsilon, 1+\epsilon)A_{\boldsymbol{\theta}_t}(\tau)\} + \beta_S^{-1}\sum_{j=1}^{q}\mathcal{S}(\pi_{\boldsymbol{\theta}}(\cdot|a_1,\ldots,a_{j-1}))\right], \qquad \text{(A4)}\end{aligned}$$

where $\mathcal{S}(\pi_{\boldsymbol{\theta}}(\cdot|a_1,\ldots,a_{j-1})) \equiv \mathcal{S}(\pi_{\boldsymbol{\theta}}(\cdot))$, for $j = 1$. The trade-off between exploration and exploitation is controlled by the coefficient β_S^{-1}, which carries a meaning analogous to temperature in statistical mechanics: For $\beta_S^{-1} \to 0$ (or $\beta_S \to \infty$), any exploration is limited to the intrinsic probabilistic nature of the policy; if training is successful, it is expected that, for deterministic environments, the policy eventually converges to a delta distribution (over the

action space) at the later training iterations; this may deteriorate exploration and learning. However, in the opposite limit, $\beta_S^{-1} \to \infty$ (or $\beta_S \to 0$), every action is selected with equal probability, and the values of the policy π become irrelevant. Therefore, in practice, we use a decay schedule for the inverse temperature β_S^{-1} to gradually reduce exploration (see Appendix A 5).

Since the marginal distribution $\pi_{\boldsymbol{\theta}}(\cdot)$ and the conditional distribution $\pi_{\boldsymbol{\theta}}(\cdot|a_1,\ldots,a_{j-1})$ are discrete categorical distributions over $\mathcal{A}$, we can compute a closed-form expression for the entropy of the categorical distribution policy. For trajectory $\tau^i = (a_1^i,\ldots,a_q^i)$, the jth term in the entropy bonus simplifies to

$$\mathcal{S}(\pi_{\boldsymbol{\theta}}(\cdot|a_1^i,\ldots,a_{j-1}^i)) = -\sum_{a\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|a_1^i,\ldots,a_{j-1}^i)\log\pi_{\boldsymbol{\theta}}(a|a_1^i,\ldots,a_{j-1}^i). \quad \text{(A5)}$$

We emphasize that the entropy considered here is the Shannon or information entropy associated with the policy as a probability distribution and should be contrasted with the thermodynamic entropy, associated with the logarithm of the density of protocol configurations (a.k.a. density of states) in the optimization landscape. The Shannon entropy help the agent to explore the space of policies, and thus the annealing of the corresponding Lagrange multiplier, β_S^{-1}, is not related to thermal annealing in the optimization (or energy) landscape in a straightforward manner. Moreover, notice that the policy optimization is part of the classical postprocessing of the quantum data; i.e., it does not compromise the nature of the quantum data which is fed to the algorithm in the form of rewards.

Figure 13 shows a comparison of PPO with and without entropy, as controlled by the value of the temperature β_S^{-1}. Introducing the policy information entropy keeps the policy a bit broader in the initial stages of training, which enhances exploration. Towards the end of training, the information entropy is not needed; therefore, we gradually "anneal" β_S^{-1} (cf. Appendix A 5).

5. Technical details

We train the CD-QAOA algorithm for 500 epochs or iterations with a minibatch size of $M = 128$. Throughout the training, we sample trajectories according to the marginal and conditional policy distributions given by the autoregressive network.

We use Adam to perform gradient descent on the objective in Eq. (A4), with the default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, which define the exponential decay rate for the first- and second-moment estimates, respectively. The learning rate is initialized as $\alpha_{\{\mathrm{lr},0\}} = 0.01$ and decays by a factor of 0.96 every 50 steps in a staircase fashion. To be more precise, the learning rate at the kth iteration with the exponential decay reads as $\alpha_{\mathrm{lr},\{k\}} = 0.01 \times 0.96^{\lfloor k/50 \rfloor}$. The subscript $\{k\}$ denotes the iteration or episode number.

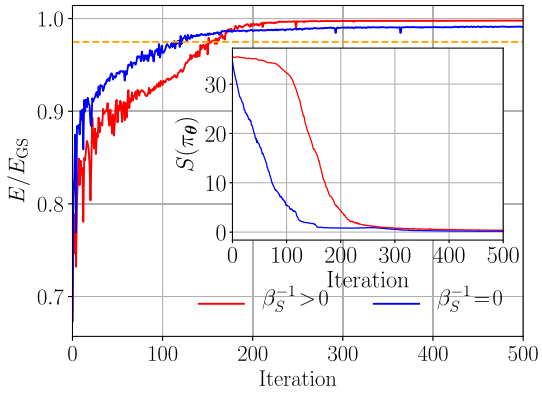

FIG. 13. Spin-1 Ising model: comparison of the mean reward with ($\beta_{S,\{0\}}^{-1} = 0.1$) and without ($\beta_S^{-1} = 0$) the entropy bonus during training. For comparison, the dashed horizontal line marks the performance of QAOA. The inset shows the evolution of the policy information entropy during training. Adding entropy gives more room for the RL agent to explore the space of policies instead of directly exploiting the knowledge it obtains. As becomes clear from the figure, the RL algorithm with the entropy bonus achieves a better final performance at the end of training, at the cost of suffering an intermediate lower reward at the beginning of training. The simulation parameters are the same as in Fig. 11.

We also introduce an exponential decay schedule for the prefactor (a.k.a. temperature), β_S^{-1}, of the entropy bonus from Eq. (A4). The temperature initializes at $\beta_{S,\{0\}}^{-1} = 0.1$ and decays by a factor of 0.9 every 10 steps. At the kth iteration, the temperature is $\beta_{S,\{k\}}^{-1} = 0.1 \cdot 0.9^{k/10}$. Eventually, the temperature is annealed to zero.

We estimate the advantage function by $A_{\boldsymbol{\theta}_{\mathrm{old}}}(\tau) = R(\tau) - b$, where b is the baseline used to reduce the variance of the estimation. Our baseline b uses an exponential moving average (EMA) of the previous rewards. EMA stabilizes the training and also leverages the past reward information to form a lagged baseline. In practice, we find that the RL algorithm can achieve better rewards compared with using the average of current samples as the baseline. To be more specific, the exponential moving baseline update is $b_{\{k\}} = \eta b_{\{k-1\}} + (1-\eta)\bar{R}_{\{k\}}$, where $b_{\{0\}} = 0$ and $\eta = 0.95$. Here, $\bar{R}_{\{k\}}$ is the sample average of the reward at the kth iteration, i.e., $\bar{R}_{\{k\}} = (1/M)\sum_{i=1}^{M} R_{\{k\}}^i(\tau^i)$.

In terms of policy optimization, we perform multiple steps of Adam on the objective [Eq. (A4)]. The gradient update steps are 4 per minibatch. The clipped parameter in the objective is set to $\epsilon = 0.1$.

The hyperparameters of the algorithm are listed in Table III.

Algorithm 1. CD-QAOA with autoregressive network based policy.

Input: batch size M, learning rate η_t, total number of iterations T_{iter}, exponential moving average coefficient m, entropy coefficient β_S^{-1}, PPO gradient steps K.
1: Generate and select the gauge potential sets $\mathcal{A}$ using Algorithm 2.
2: Initialize the autoregressive network and initialize the moving average $\hat{R} = 0$.
3: **for** $t = 1, \ldots, T_{\text{iter}}$ **do**
4: Autoregressively sample a batch of discrete actions of size M, denoted as B:

$$\tau^k = (a_1^k, a_2^k, \ldots, a_q^k) \sim \pi_{\boldsymbol{\theta}}(a_1, a_2, \ldots, a_q), \quad k = 1, 2, \ldots, M.$$

5: Apply the SLSQP solver to the lower-level continuous optimization (cf. Appendix B):

$$\min_{\{\alpha_j^k\}_{j=1}^q} \left\{ N^{-1} E(\{\alpha_j^k\}_{j=1}^q, \tau^k) \middle| \sum_{j=1}^q \alpha_j^k = T; 0 \le \alpha_j^k \le T \right\}.$$

6: Use the negative energy density as the return and compute the moving average:

$$R_k = -N^{-1} E(\{\alpha_j^k\}_{j=1}^q, \tau^k), \qquad \hat{R} = m \cdot \hat{R} + (1-m) \cdot \frac{1}{M} \sum_{k=1}^M R_k.$$

7: Compute the advantage estimates $A_k = R_k - \hat{R}$.
8: Initialize the parameter $\boldsymbol{\theta}_{t+1}^{[1]} = \boldsymbol{\theta}_t$.
9: **for** $\kappa = 1, \ldots, K$ **do**
10: Evaluate the likelihood of samples using the parameters from the last iteration and the current iteration, i.e., $\pi_{\boldsymbol{\theta}_t}(\tau^k)$, $\pi_{\boldsymbol{\theta}_{t+1}^{[\kappa]}}(\tau^k)$, and compute the importance weight $\rho_k^{[\kappa]} = \pi_{\boldsymbol{\theta}_{t+1}^{[\kappa]}}(\tau^k)/\pi_{\boldsymbol{\theta}_t}(\tau^k)$.
11: Use the advantage estimate and importance weight to compute $\mathcal{G}_k, \mathcal{S}_k$, following Eqs. (A3) and (A5).
12: Compute the CD-QAOA objective Eq. (A4) and backpropagate to get the gradients:

$$\nabla_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}_{t+1}^{[\kappa]}) = \frac{1}{M} \sum_{\{a_j^{\{k\}}\}_{j=1}^q \in B} \nabla_{\boldsymbol{\theta}} [\mathcal{G}_k^{[\kappa]} + \beta_S^{-1} \mathcal{S}_k^{[\kappa]}].$$

13: Update weights $\boldsymbol{\theta}_{t+1}^{[\kappa+1]} \leftarrow \boldsymbol{\theta}_{t+1}^{[\kappa]} + \eta_t \nabla_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}_{t+1}^{[\kappa]})$.
14: Update the parameter $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_{t+1}^{[K+1]}$.

APPENDIX B: LOW-LEVEL OPTIMIZATION: FINDING OPTIMAL PROTOCOL TIME STEPS $\boldsymbol{\alpha_j}$

In order to determine the values of the time steps α_j, we proceed as follows. For any given sequence of actions (or protocol sequence) $\tau = (a_1, \ldots, a_q)$ of total duration T, we solve the following low-level optimization problem:

$$\min_{\{\alpha_j\}_{j=1}^q} \left\{ N^{-1} E(\{\alpha_j\}_{j=1}^q, \tau) \middle| \sum_{j=1}^q \alpha_j = T; 0 \le \alpha_j \le T \right\}, \quad \text{(B1)}$$

where q is the sequence length (circuit depth), N is the system size, and $E(\cdot)$ is the energy of the final quantum state [cf. Eq. (2)] after evolving the initial quantum state $|\psi_i\rangle$ according to the fixed protocol τ.

Note that the α_j optimization is both bounded and constrained. It fits naturally into the framework of the SLSQP. SLSQP solves the nonlinear problem in Eq. (B1) iteratively, using the Han-Powell quasi-Newton method with a Broyden–Fletcher–Goldfarb–Shanno update [92] of the B-matrix (an approximation to the Hessian matrix) and an $L1$-test function within the step size.

During each iteration of the policy update, a batch of trajectories $\{\tau^i\} = \{(a_1^i, \ldots, a_q^i)\}_{i=1}^M$ is sampled. Each trajectory sequence τ^i is assigned a reward by solving the optimization problem in Eq. (B1). Since performing the low-level optimization in Eq. (B1) is independent of the high-level optimization discussed in Appendix A, we run the former concurrently to boost the efficiency of the algorithm. We distribute every sequence $\tau^i = (a_1^i, \ldots, a_q^i)$ to a different worker process and aggregate the results back to the master process in the end. In practice, we use the batch size $M = 128$, and we distribute the simulation on 4 nodes with 32 cores each so that each core solves only one optimization at a time.

Recently, it was demonstrated that it is possible to perform the continuous optimization on par with the discrete one, which eliminates the need to use a solver and results in a fully RL optimization approach [84].

APPENDIX C: SCALING WITH THE NUMBER OF PARTICLES N, THE PROTOCOL DURATION T, AND THE CIRCUIT DEPTH q

Next, we discuss the computational scaling of CD-QAOA. While there are a number of (hyper)parameters in the algorithm, here we focus on the system size N, the protocol duration T, and the circuit depth q—which are, physically, the most relevant ones. We also consider the continuous and discrete optimization steps separately (the continuous step also being an essential part of conventional QAOA).

When it comes to the continuous optimization performed by a solver (cf. Appendix B), the main computational cost comes from the quantum evolution itself. The basic operation inside the solver is a multiplication of the matrix exponential $\exp(-i\alpha_j H_j)$ by the state $|\psi_i\rangle$. The Hamiltonian H_j is stored as a sparse matrix, and the action of the matrix exponential onto the quantum state, $\exp(-i\alpha_j H_j)|\psi_i\rangle$, can be evaluated without computing the matrix exponential itself with the help of a sparse matrix-vector product; this operation scales exponentially with the system size N, i.e., $O(\exp(cN))$ for some constant c. If we denote the sequence length (a.k.a. circuit depth) by q, then the total cost for evaluating a single value of the continuous angle α scales as $O(q\exp(cN))$. We stress that this cost is also incurred by conventional QAOA.

For the discrete optimization performed using reinforcement learning (Appendix A 4), notice first that the machine learning model is agnostic to the physical quantum model because we do not use information about the quantum model to train the policy (cf. Appendix A 2). Because the policy input is, by construction, independent of the quantum state, the input layer of the neural network architecture is shielded from the exponential growth of the physical Hilbert space with N. Hence, the deep neural network is *independent* of the Hilbert space dimension. Further, we use an autoregressive network model that scales linearly with the sequence length q and also linearly with the size of the available action set $|\mathcal{A}|$. Thus, the total computational cost for the reinforcement learning optimization scales as $O(q|\mathcal{A}|)$. The scaling of the neural network with the variational network parameters (weights and biases) is trivially given by the matrix-vector multiplication, as is the case for typical ML deep networks, and it is also independent of the physics of the controlled system.

A comparison of the wall-clock time for the discrete and continuous optimization steps is provided in Table IV. We distinguish between the continuous solver optimization and the discrete RL optimization, and show the average times for *one* successful step of each in the two columns on the right-hand side. The total cost can then be obtained by multiplying the time for t_{solver} by the appropriate number of repetitions (e.g., continuous solver initial conditions, policy sample batch size, PPO training episodes, etc.) and by multiplying the time for t_{RL} by the number of PPO iterations, thereby taking into account any parallelization if used; for instance, the most expensive simulation we performed ran for about 109 hours on four nodes (Intel Xeon Skylake 6130 32-core 2.1 GHz) to produce the $N = 10$, $T = 12$, $q = 20$ data point shown in Table II.

TABLE IV. Wall-clock running time of the two-level CD-QAOA optimization steps with different system sizes N, protocol durations T, and circuit depths q. The right-hand side of the table shows the time used for the lower-level solver (column t_{solver}) and the time spent for the high-level RL algorithm (column t_{RL}) at every successful iteration. The total cost can then be obtained by multiplying the time for t_{solver} by the appropriate number of repetitions (e.g., continuous solver realizations, policy sample batch size, PPO training episodes, etc.), taking into account any parallelization if present. Every number represents an average over 40 independent runs with the corresponding standard deviation shown; the significant deviation in t_{solver} is caused by the random initial solver state used, which causes the algorithm to take a different number of steps to converge within the given tolerance (cf. Appendix D). This test is carried out on a single-processor Intel Core i7-8700K CPU 6-core 3.70 GHz.

N	T	q	t_{solver} (sec/iter)	t_{RL} (sec/iter)
10	20	20	57.254 ± 13.829	0.042 ± 0.005
8	20	20	17.24 ± 2.554	0.055 ± 0.024
6	20	20	10.559 ± 3.963	0.028 ± 0.004
4	20	20	6.021 ± 5.149	0.027 ± 0.002
10	28	20	68.55 ± 19.044	0.055 ± 0.019
10	24	20	61.425 ± 15.171	0.038 ± 0.009
10	20	20	57.254 ± 13.829	0.042 ± 0.005
10	16	20	49.043 ± 12.447	0.041 ± 0.007
10	12	20	39.33 ± 13.976	0.038 ± 0.006
10	8	20	24.689 ± 14.348	0.033 ± 0.008
10	4	20	7.023 ± 2.651	0.025 ± 0.001
8	20	24	20.723 ± 3.903	0.065 ± 0.024
8	20	20	17.24 ± 2.554	0.055 ± 0.024
8	20	16	12.626 ± 3.129	0.024 ± 0.004
8	20	12	8.641 ± 2.654	0.02 ± 0.003
8	20	8	5.511 ± 2.18	0.016 ± 0.002
8	20	4	2.092 ± 1.312	0.011 ± 0.002

We emphasize that the time t_{solver} required for the continuous optimization is an essential part of conventional QAOA and is the current limiting factor for reaching large system sizes, as is the case in merely all simulations of quantum dynamics on classical computing devices. In sharp contrast, the cost for training the deep autoregressive network is N independent, and t_{RL} per iteration is negligible; however, the choice of RL algorithm can strongly impact the number of iterations. Thus, CD-QAOA is suitably designed for potential applications on quantum simulators and quantum computers, which will enable one to access large system sizes, bypassing the exponential bottleneck intrinsic to simulations of quantum dynamics on classical devices.

APPENDIX D: MANY-BODY CONTROL LANDSCAPE

Let us briefly address the question about the difficulty of the many-body ground-state preparation problems that we introduced in the main text. To this end, recall that CD-QAOA has a two-level optimization structure: (i) discrete optimization to construct the optimal sequence of unitaries (Appendix A) and (ii) continuous optimization to find the best angles, given the sequence, to minimize the cost function (Appendix B). Here, we focus exclusively on the continuous optimization landscape and postpone the discrete landscape to a future study.

The RL agent learns in batches or samples of $M = 128$ sequences, which sample the current policy at each iteration step and provide the data set for the policy gradient algorithm. To evaluate each sequence in the batch, we use SLSQP to optimize for the durations α_j in a constrained and bounded fashion: $\sum_j \alpha_j = T$ and $0 \leq \alpha_j \leq T$ (cf. Appendix B). This method provides us with the full unitary $U(\{\alpha_j\}_{j=1}^q, \tau)$; applying it to the initial state, we obtain the reward value for this sequence. This procedure repeats iteratively as the RL agent progressively discovers improved policies.

Once the RL agent has learned an optimal sequence, i.e., after the optimization procedure is complete, we focus on the best sequence from the sample and examine how difficult it is to find the corresponding durations α_j using SLSQP. To this end, we draw q values at random from a uniform distribution over the interval $[0, T/q]$ and use them as initial conditions for the α_j to initialize the SLSQP optimizer. We use the same q as the circuit depth so that the initial durations $\alpha_j^{(0)}$ are, on average, equal. We then repeat this procedure P times and generate a sample $\mathcal{M} = \{\{\alpha_j^n\}_{j=1}^q\}_{n=1}^P$ of the local minima in the optimization landscape for α_j. The larger P, the better our result for the true reward assigned to τ is.

Notice that, in the beginning of the training, the RL agent is still in the exploration stage, and the reward estimation does not need to be too accurate; this reward estimation needs to be more accurate as the agent switches over to exploitation during the end of the training. In order to make the algorithm computationally more efficient, we introduce a linear schedule for the number of realizations of the α_j optimizer, starting from 3 with an increment of 1 every 30 iteration steps, i.e., $P_{\{k\}}^{\text{tot}} = 3 + \lfloor k/30 \rfloor$, where the subscript k indicates the iteration number for the RL policy optimization. In order to further save time in the reward estimation, we also introduce some randomness here by sampling $P_{\{k\}}$ from a uniform distribution over $1, 2, \ldots, P_{\{k\}}^{\text{tot}}$.

Even though they all correspond to the same sequence, every local minimum in $\mathcal{M}$ represents a potentially different protocol since the durations α_j will cause the initial quantum state to evolve into a different final state. For every

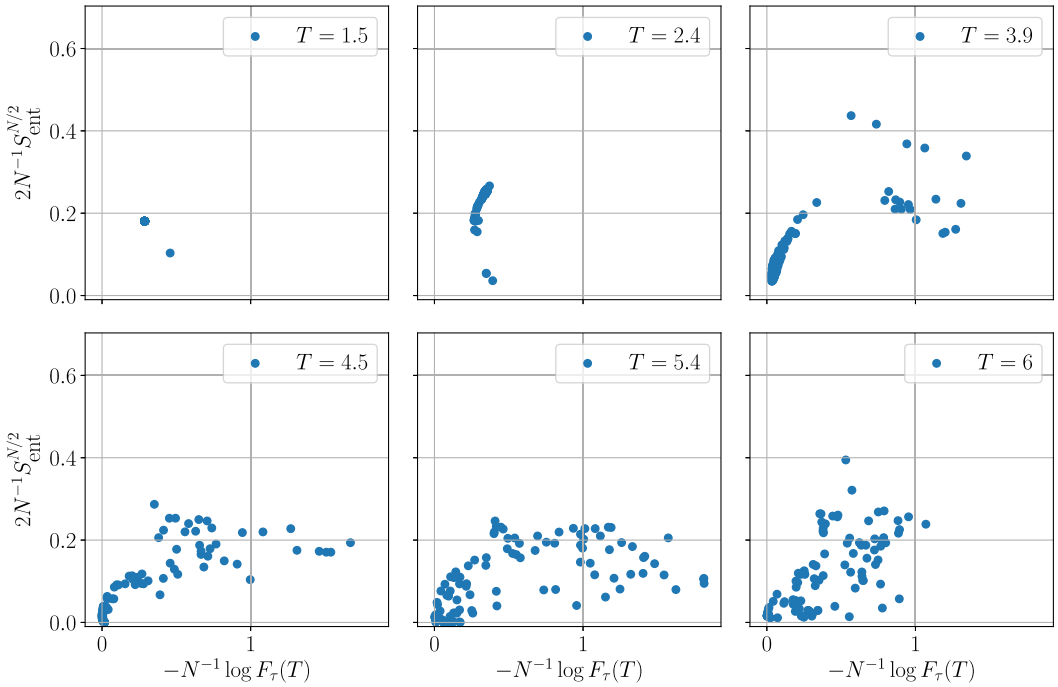

FIG. 14. Spin-1/2 Ising model: visualization of the continuous optimization landscape for the durations α_j in the fidelity-entanglement entropy plane, for the best sequence found by the RL agent (see Appendix D). Each point corresponds to a local minimum, obtained using the SLSQP optimizer, starting from a uniformly drawn random initial condition. The system size is $N = 16$, and the rest of the parameters are the same as in Fig. 1.

protocol in $\mathcal{M}$, we can evaluate the negative log-fidelity, $-\log F_\tau(T)$, and entanglement entropy of the half chain, $S_{\text{ent}}^{N/2}$. Since the target state for the Ising model is an ordered ground state, it has area-law entanglement. Figure 14 shows a cut through the landscape in the fidelity-entanglement entropy plane for a few different durations T for the spin-1/2 Ising model. The better solutions are located in the lower-left corner. The proliferation of local minima across the quantum speed limit has recently been studied in the context of RL [29] and QAOA [43]. This behavior indicates the importance of running many different SLSQP realizations; otherwise, we may misevaluate the reward of a given sequence, and the policy gradient will perform poorly.

Figure 14 also provides a plausible explanation for the destruction of the scaling collapse for $T \gtrsim T_{\text{QSL}}$ (Fig. 2). Although the precision of the SLSQP optimizer is set at 10^{-6}, the entropy curves for large durations no longer fall on top of each other. Hence, the occurrence of many local minima of roughly the same reward, which correspond to different protocols, effectively removes any universal features from the obtained solution; therefore, different system-size simulations end up in different local minima.

APPENDIX E: VARIATIONAL GAUGE POTENTIALS

Consider the generic Hamiltonian

$$H(\lambda) = H_0 + \lambda H_1, \tag{E1}$$

with a general smooth function $\lambda = \lambda(t)$. We define a state preparation problem where the system is prepared in the ground state of H_0 at time $t = 0$, and we want to transfer the state population in the ground state of H by time $t = T$.

Unlike adiabatic protocols, counterdiabatic driving relaxes the condition of being in the instantaneous ground state of $H(\lambda)$ during the evolution. The idea is to reach the target state in a shorter duration T (compared to the adiabatic time) at the expense of creating controlled excitations [with respect to the instantaneous $H(\lambda)$] during the evolution, which are removed before reaching the final time T. To achieve this, one can define a counterdiabatic Hamiltonian H_{CD}. In general, the original $H(\lambda)$ differs from H_{CD}, whose ground state the system follows adiabatically:

$$H_{\text{CD}}(\lambda) = H(\lambda) + \dot{\lambda} A_\lambda, \tag{E2}$$

where A_λ is the gauge potential; A_λ is defined implicitly as the solution to the equation [19]

$$[\partial_\lambda H + i[A_\lambda, H], H] = 0. \tag{E3}$$

The boundary conditions $H_{\text{CD}}(\lambda(0)) = H(\lambda(0))$ and $H_{\text{CD}}(\lambda(T)) = H(\lambda(T))$ impose the additional constraint $\dot{\lambda}(0) = 0 = \dot{\lambda}(T)$, which suppresses excitations at the beginning and at the end of the protocol.

Using Eq. (E3), one can convince oneself that the gauge potential A_λ of a real-valued Hamiltonian H is always imaginary valued [19].

For generic many-body systems, it has recently been argued that the gauge potential A_λ is a nonlocal operator [36]. Nevertheless, one can proceed by constructing a variational approximation $\mathcal{X} \approx A_\lambda$, which minimizes the action

$$\mathcal{S}(\mathcal{X}) = \langle G^2(\mathcal{X})\rangle - \langle G(\mathcal{X})\rangle^2, \qquad G(\mathcal{X}) = \partial_\lambda H + i[\mathcal{X}, H]. \tag{E4}$$

For ground-state preparation, $\langle\cdot\rangle = \langle\psi_{\text{GS}}(\lambda)|\cdot|\psi_{\text{GS}}(\lambda)\rangle$ is the instantaneous ground-state expectation value with respect to $H(\lambda)$. More generally, one can use $\langle\cdot\rangle = \text{Tr}(\rho_{\text{th}} \times (\cdot))$, where $\rho_{\text{th}} \propto \exp(-\beta H)$ is a thermal density matrix at temperature β^{-1}: For $\beta \to \infty$, we recover the ground-state expectation value; for $\beta \to 0$, all eigenstates are weighted equally.

We mention, in passing, that alternative schemes to approximate the adiabatic gauge potential have also been considered [39].

1. Spin Hamiltonians

a. Real-valued spin-1/2 Hamiltonians

Let H now be a real-valued spin-1/2 Hamiltonian with translation and reflection invariance. Such a system is given, e.g., by the mixed-field Ising model discussed in the main text. We now construct an ansatz for the variational gauge potential $\mathcal{X}$, which obeys these symmetries and is imaginary valued.

We can organize the terms contained in $\mathcal{X}$ according to their multibody interaction type, as follows. The only single-body imaginary-valued term we can write is $\sum_j \beta_j S_j^y$. Translation and reflection symmetries, whenever present in H, further impose that the coupling constant $\beta_j = \beta$ be site independent, i.e. spatially uniform. Hence, this is the zeroth-order term in our variational gauge potential construction, cf. Eq. (E5).

Next, we focus on the two-body terms. Because the exact A_λ is imaginary valued for real-valued Hamiltonians, we may only consider interaction terms where S^y appears precisely once: $S^x S^y$ and $S^y S^z$. For spin-1/2 systems, the two operators have to act on different sites, or one can further simplify their product to single-body operators using the algebra for Pauli matrices. Once again, translation invariance dictates that the coupling constants are uniform in space, while reflection invariance requires us to take a symmetric combination. Further imposing that the interaction be short ranged (we want to construct the most local variational ansatz), we arrive at

$$\mathcal{X}(\{\beta_l^{(k)}\}) = \sum_j \beta_0^{(0)}(\lambda) S_j^y + \beta_1^{(0)}(\lambda)(S_{j+1}^x S_j^y + S_{j+1}^y S_j^x) + \beta_1^{(1)}(\lambda)(S_{j+1}^z S_j^y + S_{j+1}^y S_j^z). \tag{E5}$$

The coefficients $\beta_l^{(k)}$ are the variational parameters that we need to determine to find the approximate CD protocol. To find their optimal values, we minimize the action $\mathcal{S}(\mathcal{X})$ [19]. Note that, since we do not have a closed-form expression for the instantaneous ground state of $H(\lambda)$, we do the minimization numerically at every fixed time t along the protocol $\lambda(t)$ (cf. Appendix E 2).

We can, in principle, add the next order terms to the series; however, they will either be less local or consist of three- and higher-body interactions, which is hard to implement in experiments.

b. Real-valued spin-1 Hamiltonians

The situation is more interesting for spin-1 systems: The eight-dimensional Lie algebra $\mathfrak{su}(3)$, which generates SU(3), contains three distinct imaginary-valued directions, which form a closed subalgebra $\mathfrak{su}(2) \subsetneq \mathfrak{su}(3)$; hence, there is more room to generate imaginary-valued combinations. To find all imaginary-valued terms consistent with a set of symmetries, we use QuSpin's functionality to implement an algorithm (Appendix E 3) that lists them for generic bases [56,57].

Translation and reflection symmetric spin-1 Hamiltonians, such as the spin-1 Ising and Heisenberg models, have a similar expansion to their spin-1/2 counterparts but allow for more terms. Restricting the expansion to two-body terms, we have

$$\mathcal{X}(\{\beta_l^{(k)}\}) = \sum_j [\beta_0^{(0)}(\lambda) S_j^y + \beta_1^{(0)}(\lambda)(S_j^x S_j^y + S_j^y S_j^x) + \beta_2^{(0)}(\lambda)(S_j^z S_j^y + S_j^y S_j^z) + \beta_0^{(1)}(\lambda)([S_{j+1}^x - aS_j^x]S_j^y + [S_{j+1}^y - aS_j^y]S_j^x) + \beta_1^{(1)}(\lambda)([S_{j+1}^z - bS_j^z]S_j^y + [S_{j+1}^y - bS_j^y]S_j^z)]. \tag{E6}$$

where the constants a and b are chosen so that all five terms are mutually orthogonal with respect to the scalar product induced by the trace (i.e., Hilbert-Schmidt) norm; this ensures the linear independence of the constituent terms. Note that the three imaginary-valued on-site terms correspond precisely to the imaginary-valued $\mathfrak{su}(2) \subsetneq \mathfrak{su}(3)$.

Adding magnetization conservation and spin inversion symmetry further reduces the allowed terms in the series. Therefore, one has to restrict to three- and four-body terms:

$$\mathcal{X}(\{\zeta_l^{(k)}\}) = \sum_j \zeta_0^{(2)}(\lambda)(iS_j^+ S_{j+1}^- S_{j+2}^z + iS_j^z S_{j+1}^- S_{j+2}^+ + \text{H.c.}) + \zeta_0^{(3)}(\lambda)(iS_j^- S_j^z S_{j+1}^+ S_{j+1}^z + iS_j^+ S_j^z S_{j+1}^- S_{j+1}^z + \text{H.c.}). \tag{E7}$$

Because these terms are multibody and less local, we refrain from using them in CD-QAOA in the present study. We merely list them here for completeness.

As explained in the main text, to apply CD-QAOA for many-body ground-state preparation, we consider the constituent terms in $\mathcal{X}$ as independent generators $\{H_j\}_{j=1}^{|\mathcal{A}|}$, which is in contrast to the variational gauge potential method where the ratios between the coefficients $\beta_l^{(k)}$ play an important role.

c. Variational gauge potential ansatz for the Lipkin-Meshkov-Glick model

As explained in the main text, the LMG Hamiltonian, cf. Eq. (5), models homogeneously interacting spin-1/2 particles on an all-to-all connected graph in the presence of an external field. Here, we compute the lowest-order terms appearing in the series for the variational gauge potential $\mathcal{X}$, going beyond Ref. [93].

The starting point is the LMG Hamiltonian

$$H = -\frac{J}{N}(S^x)^2 + h(S^z + N/2). \tag{E8}$$

We introduce two bosonic modes, s and t, where $S^z = t^\dagger t - N/2 = n_t - N/2$ and $S^+ = t^\dagger s$, and cast the LMG Hamiltonian in the form

$$H = ht^\dagger t - \frac{J}{4N}(t^\dagger s + s^\dagger t)^2. \tag{E9}$$

Recalling, once again, that real-valued Hamiltonians have imaginary-valued gauge potentials and that gauge potentials do not have diagonal matrix elements, we make the following ansatz:

$$\mathcal{X}(\{\beta_l^{(k)}\}) = \beta_0^{(0)}(\lambda) Y + \beta_1^{(1)}(\lambda) \hat{X}Y + \beta_1^{(0)}(\lambda) \hat{Z}Y, \tag{E10}$$

where

$$
\begin{aligned}
Y &= S^y = \frac{i}{2}(s^\dagger t - t^\dagger s),\\
\hat{XY} &= \frac{1}{N}(S^x S^y + S^y S^x) = -\frac{i}{2N}[(t^\dagger s)^2 - (s^\dagger t)^2],\\
\hat{ZY} &= \frac{1}{N}\left(\left(S^z + \frac{N}{2}\right)S^y + S^y\left(S^z + \frac{N}{2}\right)\right)\\
&= \frac{i}{2N}(s^\dagger t^\dagger t t - s t^\dagger t t^\dagger + s^\dagger t t^\dagger t - s t^\dagger t^\dagger t). \qquad \text{(E11)}
\end{aligned}
$$

To compute the matrix elements of the gauge potentials, we define the basis

$$
|N, n_t\rangle = \frac{(t^\dagger)^{n_t}(s^\dagger)^{N-n_t}}{\sqrt{n_t!(N-n_t)!}}|0\rangle, \quad \text{with} \quad n_t = 0, \ldots, N.
$$

The gauge potentials have the following nonzero matrix elements (plus their conjugates to make the operators Hermitian):

$$
\begin{aligned}
\langle N, n_t|Y|N, n_t+1\rangle &= -\frac{i}{2}\sqrt{(n_t+1)(N-n_t)},\\
\langle N, n_t|\hat{XY}|N, n_t+2\rangle &= \frac{i}{2N}\sqrt{(n_t+2)(n_t+1)(N-n_t-1)(N-n_t)},\\
\langle N, n_t|\hat{ZY}|N, n_t+1\rangle &= \frac{i}{2N}(2n_t+1)\sqrt{(n_t+1)(N-n_t)}. \qquad \text{(E12)}
\end{aligned}
$$

2. Numerical minimization to obtain the variational CD protocol

Since the action $\mathcal{S}$ in Eq. (E4) is quadratic in the variational parameters β_j, it is possible to derive a generic linear system, whose solutions are the optimal parameters of the variational gauge potential within CD driving [94].

Suppose that $\mathcal{X} = \sum_{j=1}^{r}\beta_j H_j$ is given by a linear combination of r gauge potential terms. Then, it is straightforward to see that

$$
G(\mathcal{X}) = \partial_\lambda H + \sum_{j=1}^{r} i[H_j, H]\beta_j. \qquad \text{(E13)}
$$

Defining the operator-valued quantities $B_0 = \partial_\lambda H$ and $B_j = i[H_j, H]$ and setting $\beta_0 = 1$, we arrive at the following expression for the variational action:

$$
\begin{aligned}
\mathcal{S}(\mathcal{X}) &= \left\langle\left(B_0 + \textstyle\sum_j B_j\beta_j\right)^2\right\rangle - \left(\left\langle B_0 + \textstyle\sum_j B_j\beta_j\right\rangle\right)^2\\
&= \sum_{i,j=0}^{r}(\langle B_i B_j\rangle - \langle B_i\rangle\langle B_j\rangle)\beta_i\beta_j, \qquad \text{(E14)}
\end{aligned}
$$

which is a quadratic form in the unknown coefficients β_j. To find the minimum of $\mathcal{S}(\mathcal{X})$ with respect to β_j, we can take the derivative and set it to zero to obtain the linear system of equations for the optimal β_j:

$$
\sum_k \mathcal{M}_{jk}\beta_k = -\mathcal{M}_{0j}, \qquad \text{(E15)}
$$

where $\mathcal{M}_{jk} = \langle B_j B_k\rangle + \langle B_k B_j\rangle - 2\langle B_j\rangle\langle B_k\rangle$. Solving the system, we obtain the minimum $\{\beta_j\}_{j=1}^{r}$ of the variational action $\mathcal{S}$.

The ground-state expectation values in the above procedure, as well as the Hamiltonian $H(\lambda(t))$, depend implicitly on time $t \in [0, T]$ via the protocol $\lambda(t)$. Therefore, to find the time dependence of $\beta_j(t)$, we discretize the time interval $[0, T]$ into N_T time steps and repeat the procedure at every time step. This process yields $\beta_j(t_i)$ at the time steps t_i. To recover the full functional dependence, we use a fine discretization mesh and apply a linear interpolation to $\beta_j(t_i)$. Alternatively, notice that the coefficients $\beta_j = \beta_j(\lambda(t))$ depend on time t only implicitly via the protocol λ. Therefore, it is also possible to discretize the range of $\lambda(t)$ instead.

For the spin-1 Ising model, the time dependence of β_j is shown in Fig. 15, defining H_{CD}, which generates the CD evolution. In Sec. V and Appendix F 4, we compare variational CD driving to CD-QAOA and conventional QAOA.

3. Algorithm for generating gauge potential terms in the presence of lattice symmetries

Finally, we also show the algorithm we used to determine the terms appearing in the gauge potential expansions in Eqs. (E5)–(E7), which obey a fixed set of symmetries.

In general, one can represent any local operator of the kind $J_{i_1,\ldots,i_l} O_{i_1}^{\gamma_1}\cdots O_{i_l}^{\gamma_l}$ as a triple $(\mathcal{Y}, \mathcal{I}, J)$, where $J = J_{i_1,\ldots,i_l}$ is the coupling coefficient constant, $\mathcal{I} = (i_1, \ldots, i_l)$ is the set of sites the operators act on, and $\mathcal{Y} = (\gamma_1, \ldots, \gamma_l)$ defines the types of operators that act on the corresponding

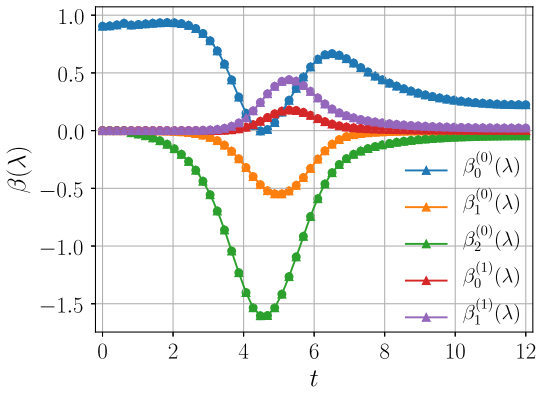

FIG. 15. Spin-1 Ising chain: time dependence of the optimal coefficients $\beta_l^{(k)}(\lambda(t))$ in the variational gauge potential [Eq. (E6)] with translation and reflection symmetry, determined from the procedure in Appendix E 2. The total duration $T = 12$ with the time discretization step $\Delta t = 0.2$, and the system size $N = 8$. The protocol we use is $\lambda(t) = \sin^2(\pi t/2T)$. The other model parameters are the same as in Fig. 7.

sites. The triple $(\mathcal{Y}, \mathcal{I}, J)$ can then be used to construct the operator.

In the following, we refer to the separate terms appearing in the gauge potential series as Hamiltonians H_j, i.e., $\mathcal{X} = \sum_j \beta_j H_j$; a Hamiltonian is defined as $H = \sum_{(i_1,\ldots,i_l)\in\Lambda} J_{i_1,\ldots,i_l} O_{i_1}^{\gamma_1} \cdots O_{i_l}^{\gamma_l}$, where Λ is the lattice graph. As we argued above, real-valued Hamiltonians have purely imaginary-valued gauge potentials; thus, the coefficient J is chosen to be purely imaginary.

We build the series for the variational gauge potential $\mathcal{X}$ recursively: We first consider a set $\mathcal{L}_{\text{elem}}$ of elementary operators O—the building blocks for the expansion; e.g., for the spin-1 chains, these can be the spin-1 operators $\mathcal{L}_{\text{elem}} = \{S^+, S^-, S^z\}$. We construct the terms in the expansion for $\mathcal{X}$ iteratively at a fixed order l, e.g., $l = 1$ comprises single-body terms, and $l = 2$—two-body terms, etc. We also assume that we have access to a routine that checks if a trial list of operators obeys a given lattice symmetry; if not, the same routine returns the missing operators to be added to the original list, so the symmetry is now satisfied (e.g., such a routine is used in QuSpin [56,57].)

The pseudocode we developed is shown in Algorithm E 3. To construct multibody terms at a fixed order l, we define combinations of the elementary operators and store them in the list $\mathcal{L}_{\text{op}}$; the way these combinations are built can be used to implement constraints, such as particle or magnetization conservation, etc. This method is implemented via the product operator (line 2 of Algorithm E 3). It generates all possible combinations of selecting l elementary operators with replacement. The sets of lattice sites that the operators from $\mathcal{L}_{\text{op}}$ act on are stored in the list $\mathcal{L}_{\text{sites}}$ (line 3 of Algorithm 2). Then, for each trial triple $(\mathcal{Y}, \mathcal{I}, J)$, we make use of the routine to check the symmetry and record any

Algorithm 2. Generation of variational gauge potential.

Input: a list of required symmetries $\mathcal{L}_{\text{sym}}$, order l, a list of elementary operator types $\mathcal{L}_{\text{elem}}$.
1: Initialize empty list for gauge potential terms $\mathcal{L}_{\text{gauge}}$.
2: Generate all possible combinations of local operators at order l,

$$\mathcal{L}_{\text{op}} = \text{product}(\mathcal{L}_{\text{elem}}, \text{repeat} = l).$$

3: Enumerate all possible combinations of lattice sites $\mathcal{L}_{\text{sites}}$ the lth order operators act on.
4: **for** $\mathcal{Y}$ in $\mathcal{L}_{\text{op}}$ **do**
5: **for** $\mathcal{I}$ in $\mathcal{L}_{\text{sites}}$ **do**
6: Initialize an empty list $\mathcal{L}_H$.
7: Set $J = i(i = \sqrt{-1})$.
8: Append $(\mathcal{Y}, \mathcal{I}, J)$ to $\mathcal{L}_H$.
9: Set the flag $IsSym = False$.
10: **while** $IsSym$ is $False$ **do**
11: Set $IsSym = True$.
12: **for** sym in $\mathcal{L}_{\text{sym}}$ **do**
13: **if** exists missing operator $(\mathcal{Y}', \mathcal{I}', J')$, **then**
14: Set $IsSym = False$.
15: Append $(\mathcal{Y}', \mathcal{I}', J')$ to $\mathcal{L}_H$.
16: Build Hamiltonian H using the triplets in $\mathcal{L}_H$.
17: **if** H or equivalents not included in $\mathcal{L}_{\text{gauge}}$ **then**
18: Append H to $\mathcal{L}_{\text{gauge}}$.
19: **Return** the list of gauge potential basis $\mathcal{L}_{\text{gauge}}$.

`product`: Cartesian product, `equivalents`: equivalent mod scalar,
`missing operator`: the operator missed for the symmetry requirement

operators that do not respect it. We append these so-called missing operators to the original list, and we keep checking the symmetry condition until we obtain all operators that satisfy the symmetry (lines 10–15 of Algorithm 2). The finite number of combinations guarantees a termination in a finite number of steps.

In order to avoid repeating previously identified Hamiltonians, we discard equivalent Hamiltonians (line 17 of Algorithm 2): Two Hamiltonians are called equivalent when one is a scalar times the other. Since here we consider imaginary-valued gauge potentials, the multiple constant should be real. To test whether the Hamiltonians H_1 and H_2 are equivalent in practice, it suffices to test whether H_1 is equal to $\pm(\|H_1\|/\|H_2\|)H_2$, where we use the Hilbert-Schmidt norm.

APPENDIX F: CD-QAOA FOR MANY-BODY STATE PREPARATION

Here, we provide a supplementary discussion on the performance of CD-QAOA for many-body pure state preparation using the quantum spin chains introduced in the main text. We refer the reader to the main text for the definition of various model parameters; the shorthand spin operator notation used is defined in Table I.

1. Spin-1/2 Ising chain

First, we show results for the single-spin problem ($J = 0$):

$$H = H_1 + H_2, \qquad H_1 = h_z S^z, \qquad H_2 = h_x S^x. \tag{F1}$$

In Fig. 16, we clearly see that CD-QAOA (red curve) has a smaller quantum speed limit $T_{\mathrm{QSL}} \approx 4.0$ than conventional QAOA (blue); this is anticipated since CD-QAOA has a larger control space at its disposal. Moreover, we find that, for $T < T_{\mathrm{QSL}}$, CD-QAOA only makes use of a single Y rotation by setting the durations α_j associated with any other unitaries from the set $\mathcal{A}$ to zero. As mentioned in the main text, conventional QAOA tries to represent this Y rotation by means of Euler angles, i.e., composed of X and Z rotations; in general, this results in a higher duration cost to complete the population transfer (leading to a larger T_{QSL}). However, for short durations T, a Y rotation can be exactly obtained using a proper sequence of the X and Z terms. For these reasons, we find an exact agreement between the two curves for small values of $T \lesssim 3$.

Let us now switch on the spin-spin interaction strength $J > 0$; consider the spin-1/2 Ising chain

$$H = H_1 + H_2,$$
$$H_1 = \sum_{j=1}^{N} J S^z_{j+1} S^z_j + h_z S^z_j, \qquad H_2 = \sum_{j=1}^{N} h_x S^x_j. \tag{F2}$$

Figure 17 (top panel) shows a comparison of the best learned energies, between conventional QAOA, and CD-QAOA for two sets $(\mathcal{A}, \mathcal{A}')$ with different numbers of unitaries: $|\mathcal{A}| = 5, |\mathcal{A}'| = 3$ (see caption). We find that,

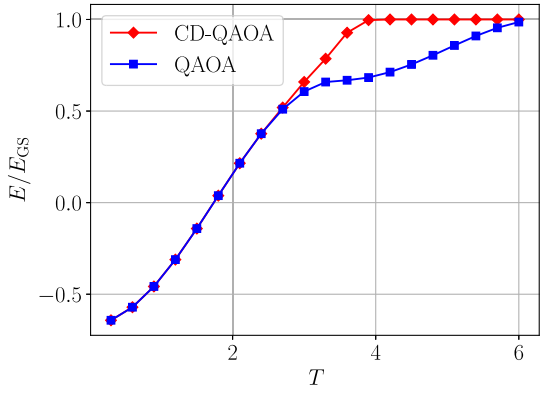

FIG. 16. Single spin-1/2 state preparation: energy density against protocol duration for CD-QAOA with $\mathcal{A}_{\mathrm{CD-QAOA}} = \{Z, X, Y\}$ (red) and conventional QAOA with $\mathcal{A}_{\mathrm{QAOA}} = \{Z, X\}$ (blue). The value of q is 3 for both methods. For conventional QAOA, we train two possible alternating patterns [i.e., $(Z \to X \to Z)$ and $(X \to Z \to X)$] and pick the best one for comparison. The model parameters are the same as in Fig. 1 with $J = 0$.

additionally using only the single-particle gauge potential term Y (green line), typically accessible in experiments, one can already obtain a higher-fidelity protocol than QAOA to prepare the ground state. Interestingly, for short protocol durations T, the two-body gauge potential terms, present in $\mathcal{A}$ but not in $\mathcal{A}'$, do not contribute to improving the energy of the final state, as can be seen from the agreement of the red and green lines for $T \lesssim 1.5$. This result suggests that single-particle processes dominate over many-body processes when it comes to lowering the energy of the z-polarized initial state, and it implies that the target ground state is single-particle-like (i.e., close to a product state). The nonsmooth behavior of the green curve at larger durations is attributed to the ruggedness of the control landscape, as different runs of the SLSQP optimizer may get stuck in one of the many suboptimal local minima (Appendix D).

One may wonder if it is possible to prepare the ground state by straightforward fidelity maximization. We define the many-body fidelity to transfer the population to the target state using the unitary process $U(\{\alpha_j\}_{j=1}^q, \tau)$, with $\sum_{j=1}^q \alpha_j = T$, as

$$F_\tau(T) = F(\{\alpha_j\}_{j=1}^q, \tau) = |\langle \psi_* | U(\{\alpha_j\}_{j=1}^q, \tau) | \psi_i \rangle|^2. \tag{F3}$$

The fidelity can be less relevant from the perspective of many-body physics because (i) the many-body fidelity is typically exponentially suppressed and (ii) it requires a reference to the ground state itself (which we seek) in order to be computed. However, the fidelity of a quantum process is a widely used benchmark in quantum computing; it also provides a better measure (than energy density) for the distance between two states in the Hilbert space $\mathcal{H}$.

Figure 17 (bottom panel) shows the many-body fidelity for $N = 14$ spins. Unlike the inset of Fig. 1 from the main text (where we show the fidelity associated with the protocol obtained using energy-density minimization), here we use the fidelity as a reward function for QAOA. We observe that optimizing the fidelity is quantitatively similar to optimizing the energy density. We would like to emphasize here, once again, the advantage of the gauge potential ansatz: The conventional QAOA simulation is done using $q = 80$ variational parameters α_j (yet no significant improvement is observed for $q \geq 4$, cf. Fig. 1), while CD-QAOA requires only $q = 3$ variational parameters.

Although the fidelity $F_\tau(T)$ is anticipated to vanish for $T < T_{\mathrm{QSL}}$ in the thermodynamic limit, the negative log-fidelity density, $-N^{-1} \log F_\tau(T)$, is more likely to remain finite. Figure 18 (inset) shows the finite-size scaling of the fidelity curves. Similar to the energy density (Fig. 2), we obtain an almost perfect scale collapse. We verify that maximizing the fidelity produces similar results as minimizing the negative log-fidelity density for the spin-1/2 chain: At first sight, this is nontrivial because $F_\tau(T)$ is exponentially suppressed with the system size N for

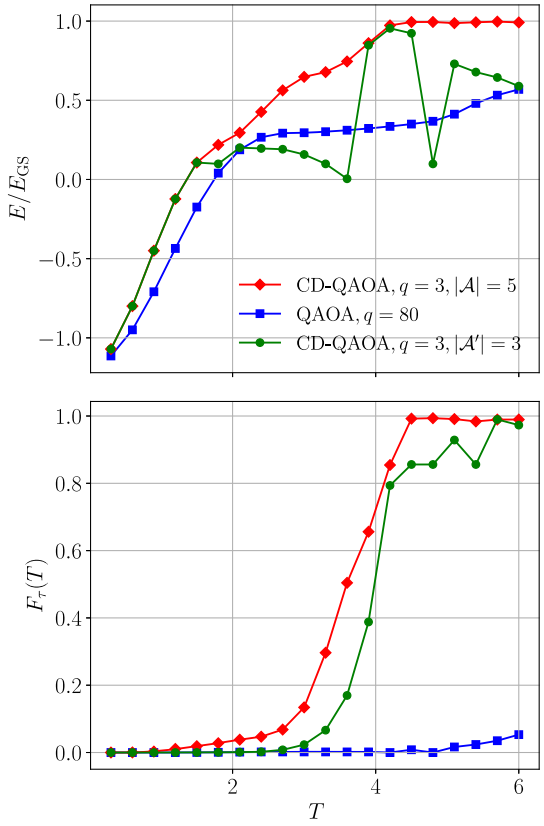

FIG. 17. Spin-1/2 Ising model: energy minimization (top) and many-body fidelity maximization (bottom) against protocol duration T. We compare CD-QAOA with $\mathcal{A}_{\text{CD-QAOA}} = \{Z|Z+Z, X; Y, X|Y, Y|Z\}$ (red), CD-QAOA with $\mathcal{A}'_{\text{CD-QAOA}} = \{Z|Z+Z, X; Y\}$ (green), and conventional QAOA with $\mathcal{A}_{\text{QAOA}} = \{Z|Z+Z, X\}$ (blue). The model parameters are the same as in Fig. 1, with the number of spins $N = 14$.

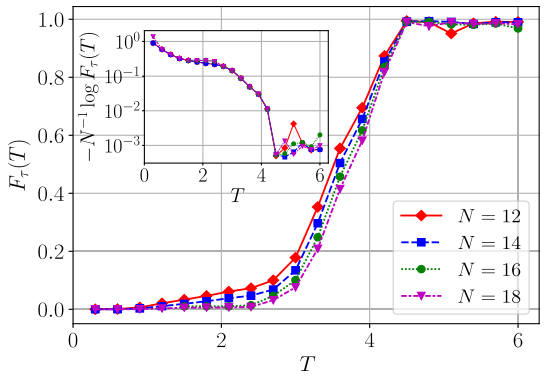

FIG. 18. Spin-1/2 Ising model: many-body fidelity maximization and corresponding quantity (inset, log scale) against protocol duration T for different system sizes N. The QAOA parameters are $q = 3$ and $\mathcal{A} = \{Z|Z+Z, X; Y, X|Y, Y|Z\}$. The model parameters are the same as in Fig. 1.

$T < T_{\text{QSL}}$; however, this behavior is likely explained by the generalization capabilities of the RL agent from small to large system sizes (cf. Sec. VI).

2. Anisotropic spin-1 Heisenberg chain

Next, we discuss in detail the ground-state preparation process in the anisotropic Heisenberg spin-1 chain:

$$H = H_1 + H_2,$$
$$H_1 = J\sum_{j=1}^{N}(S^x_{j+1}S^x_j + S^y_{j+1}S^y_j), \qquad H_2 = \Delta\sum_{j=1}^{N} S^z_{j+1}S^z_j, \tag{F4}$$

where the model parameters are defined in the main text.

An important detail worth mentioning is that the ferromagnetic ground state at $\Delta/J = -2.0$ is twofold degenerate (one state, corresponding to one of the two z polarizations). While being a trivial observation, this requires certain care when analyzing the physics of the protocols found by the agent. In particular, notice that energy minimization is insensitive to this degeneracy, and hence, the final state can appear as an arbitrary superposition of the two ferromagnetic states and still have the correct ground-state energy. This result leads to ambiguity when computing the fidelity of being in the target state: Related to this, the cost function landscape likely develops a continuous one-dimensional structure for the (degenerate) global minima. Because we are interested in energy minimization, here we define the fidelity using the projector to the ground-state manifold $P = |\psi^{(1)}_*\rangle\langle\psi^{(1)}_*| + |\psi^{(2)}_*\rangle\langle\psi^{(2)}_*|$:

$$F_\tau(T) = F(\{\alpha_j\}_{j=1}^q, \tau) = |\langle\psi^{(1)}_*|U(\{\alpha_j\}_{j=1}^q, \tau)|\psi_i\rangle|^2 + |\langle\psi^{(2)}_*|U(\{\alpha_j\}_{j=1}^q, \tau)|\psi_i\rangle|^2,$$

where $|\psi^{(1)}_*\rangle, |\psi^{(2)}_*\rangle$ are any two orthonormal states that span the doubly degenerate ground-state manifold (e.g., the two FM ground states).

Figure 19 shows a comparison between CD-QAOA and conventional QAOA for FM, XY, and Haldane target states: The top row shows the result of energy-density minimization (cf. Fig. 3); the bottom row, on the other hand, displays the many-body fidelity associated with the same protocols. For $\Delta/J = 0.5$, CD-QAOA allows us to reach the target topological Haldane state faster, as compared to conventional QAOA. Notice, also, that the gauge potential ansatz appears to be essential for reaching the target for both the XY ($\Delta/J = -0.5$) and FM states ($\Delta/J = -2.0$), which becomes particularly obvious from the many-body fidelity curves. The latter also reveals an interesting detail:

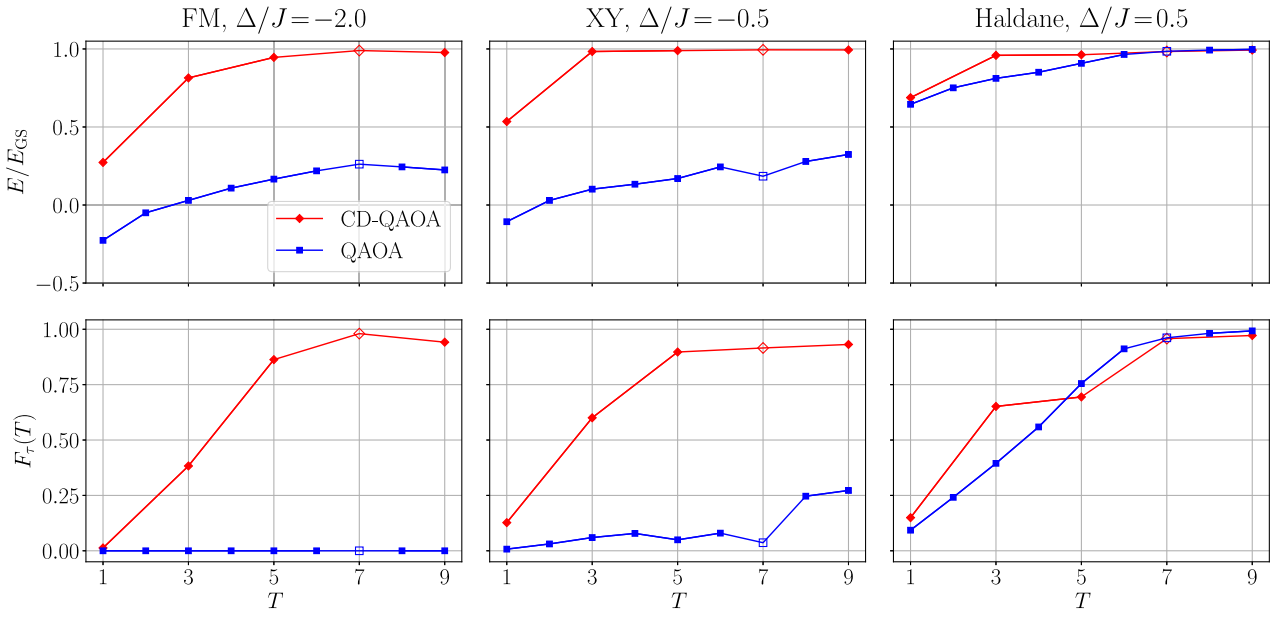

FIG. 19. Anisotropic Heisenberg spin-1 chain: energy minimization against protocol duration T—the corresponding energy (top row) and many-body fidelity (bottom row) for three ordered target states, corresponding to the ground state of the ferromagnetic (left, $\Delta/J=-2.0$), XY (middle, $\Delta/J=-0.5$), and Haldane (right, $\Delta/J=0.5$) target states, respectively. The empty symbols mark the duration at which we show the evolution of the system in Fig. 20. The model parameters are the same as in Fig. 3.

At $\Delta/J=0.5$, a regime emerges around $T\approx 5$, where the QAOA fidelity is better than the CD-QAOA fidelity. However, this peculiarity below the quantum speed limit can be explained, recalling that the RL agent is given the (negative) energy density as the reward signal and not the fidelity (note that CD-QAOA outperforms QAOA in energy).

In order to investigate in detail the protocols found by CD-QAOA, we fix a duration T and consider the time evolution of the state, $|\psi(t)\rangle = U(\{\alpha_j\}_{j=1}^{q},\tau)|\psi_i\rangle$, for three physical quantities:

(i) The energy,

$$E(t)=\langle\psi(t)|H_*|\psi(t)\rangle,$$

provides a measure of how far away in the cost function landscape the state is, at any given time $t\in[0,T]$.

(ii) The instantaneous fidelity,

$$F_\tau(t)=|\langle\psi_*|\psi(t)\rangle|^2$$

(and its generalization to the doubly degenerate ground-state manifold), measures how far the current state is from the target state $|\psi_*\rangle$ in the Hilbert space; typically, we choose the ground state as the target state $|\psi_*\rangle=|\psi_{\mathrm{GS}}(H)\rangle$.

(iii) The entanglement entropy of the half chain,

$$S_{\mathrm{ent}}^{N/2}(t)=-\mathrm{tr}_A[\rho_A(t)\log\rho_A(t)],$$
$$\rho_A(t)=\mathrm{tr}_{\bar{A}}|\psi(t)\rangle\langle\psi(t)|,$$

where A denotes a contiguous spacial region with a complement $\bar{A}$ comprising half the periodic chain, and $\rho_A(t)$ is the reduced density matrix on A at time t. For many-body systems, it is common to look at the entanglement entropy per site, which for spin-1 systems lies within the interval $2N^{-1}S_{\mathrm{ent}}^{N/2}\in[0,\log 3]$.

Figure 20 shows the time evolution of the energy, fidelity, and entropy density, for all three target states of interest. For $\Delta/J=0.5$, transferring the population from the AFM initial state to the Haldane state can be obtained equally well using either QAOA or CD-QAOA. Table V(d) shows the optimal protocol found by the RL agent: Notice the three vanishing durations $\alpha_2=\alpha_{17}=\alpha_{18}=0$; factoring them out, we recover precisely the conventional QAOA sequence (albeit with q odd). Thus, we see that the CD-QAOA may converge to conventional QAOA whenever the latter provides a high-reward sequence. This result exemplifies our claim that CD-QAOA generalizes QAOA successfully. Of course, it is not clear whether this is the true global minimum of the cost function landscape (the RL agent makes use of the additional gauge potential terms for $T<7$). Nevertheless, all physical quantities are expected to be prepared with similar accuracy under both protocols: To see this, notice that the

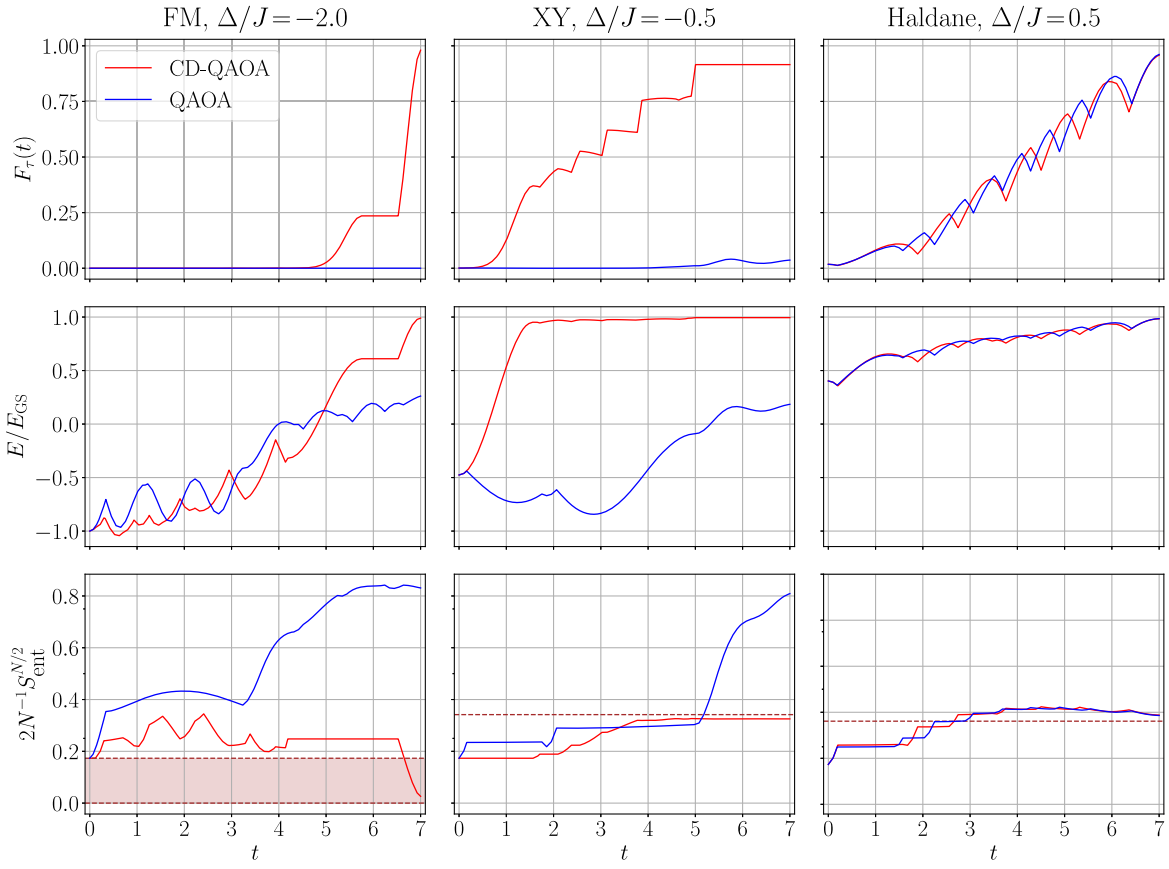

FIG. 20. Anisotropic Heisenberg spin-1 chain: time evolution generated by the protocol given by CD-QAOA (blue line) and conventional QAOA (red line) for the three target states, corresponding to the ferromagnetic ($\Delta/J = -2.0$), XY ($\Delta/J = -0.5$), and Haldane ($\Delta/J = 0.5$) target states, respectively. Three quantities are shown: many-body fidelity (first row), energy ratio (second row), and the entanglement entropy density of the half chain (third row). The horizontal dashed line in the entanglement entropy curve shows the value in the target state, while the shaded area for the FM state denotes that in the span of the doubly degenerate ground-state manifold. The protocols correspond to the duration $T = 7$ in Fig. 3. The related CD-QAOA protocol sequences are given in Table V(b) [ferromagnetic ($\Delta/J = -2.0$)], Table V(d) [XY ($\Delta/J = -0.5$)], and Table V(c) [Haldane ($\Delta/J = 0.5$)]. The simulation parameters are the same as in Fig. 3.

entanglement entropy density depends only on the quantum state (unlike expectation values of observables) and that its value at $t = T$ is close to the value for the target state (dashed horizontal line). Importantly, the entanglement remains area-law type, as seen by the values being much smaller than the maximum entropy per site, log(3), suggesting the existence of a local effective Hamiltonian that generates the population transfer process dynamically.

The best sequence for targeting the XY state at $\Delta/J = -0.5$ is shown in Table V(c). Although its structure is more complicated, factoring out the vanishing α_j, we can discern two clear patterns: (i) The sequence starts and ends with two different single-particle basis rotations, and (ii) there is an alternating subsequence based on the subset $\{X|X+Y|Y,Y\} \subsetneq \mathcal{A}_{\text{CD-QAOA}}$. Interestingly, the only gauge potential term used by the RL agent is the experimentally accessible single-particle Y rotation, and it is sufficient to reach the target with very high many-body fidelity. For comparison, conventional QAOA appears insufficient to prepare the target state for the circuit depth of $q = 18$ ($p = 9$). The advantage of CD-QAOA is also visible in the entanglement entropy density curve: QAOA can easily lead to volume-law entanglement, while CD-QAOA manages to generate as little entanglement as needed for the target state.

The discrepancy between conventional QAOA and CD-QAOA is best visible in the FM state preparation at $\Delta/J = -2.0$. In this case, a naive application of QAOA with the set $\mathcal{A}_{\text{QAOA}} = \{X|X+Y|Y,Z|Z\}$ is *a priori* doomed to fail: Starting from the initial AFM state, which is orthogonal to the target FM manifold, the resulting QAOA unitaries leave the target AFM manifold invariant; in other words, transitions between the initial and the target states are forbidden by selection rules within the QAOA dynamics. Therefore, the many-body fidelity remains zero at all times during the QAOA evolution. The energy and entanglement entropy curves certify that the state undergoes nontrivial dynamics: Similar to the XY state, QAOA

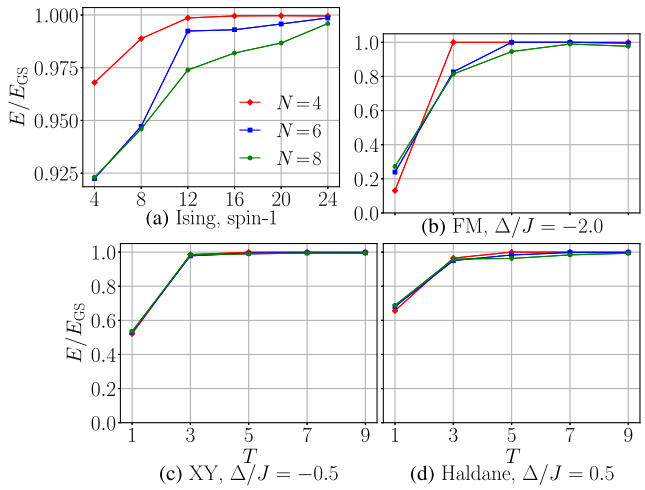

FIG. 21. System-size scaling of the energy minimization against protocol duration T for different system sizes N: (a) Spin-1 Ising chain; (b)–(d) anisotropic Heisenberg spin-1 chain for $\Delta/J=-2.0$, $\Delta/J=-0.5$, $\Delta/J=0.5$, respectively. Note that the y-axis scale is different for the spin-1 Ising model in panel (a). The model parameters are the same as in Figs. 7(a) and 3(b)–3(d), correspondingly.

creates volume-law entanglement and cannot reach the FM ground-state manifold in energy, while CD-QAOA is well behaved and sufficient to prepare the target. The CD-QAOA protocol sequence is shown in Table V(b): While we do not discern an obvious pattern, we emphasize that, this time, the RL agent makes use of both single-particle and two-body gauge potential terms.

Next, we show the system-size scaling of the energy curves for the three target states in Figs. 21(b)–21(d). Similar to the spin-1/2 Ising chain, we find very little system-size dependence for the Haldane (b) and XY states (c). However, we cannot extrapolate the results to the thermodynamic limit because of the relatively small system sizes we were able to investigate. System-size effects are more pronounced for the ferromagnetic state (d), which is the one furthest away in Hilbert space from the initial perfect antiferromagnet.

Lastly, we mention in passing that we do not show results on preparing the AFM ground state at $\Delta/J=2.0$ since this problem is somewhat trivial: Indeed, starting from a perfect AFM in the z direction, the AFM ground state of the spin-1 Heisenberg model can be easily reached, even using adiabatic evolution, because it lies within the AFM phase.

3. Lipkin-Meshkov-Glick model

In the main text, we also introduced the ferromagnetic LMG model, described by the total spin Hamiltonian

$$H=-\frac{J}{N}(S^x)^2+h\left(S^z+\frac{N}{2}\right).$$

Figure 22 shows the comparison between CD-QAOA and QAOA for two more values of $h/J=0.1$ (deep in the ferromagnetic regime) and $h/J=0.9$ (close to the critical point at $h/J=1.0$). While the behavior for $h/J=0.1$ is qualitatively similar to $h/J=0.5$ (discussed in the main text), we see that, close to the critical point, the two-body gauge potential terms $\hat{X}\hat{Y}$ and $\hat{Z}\hat{Y}$ may offer some degree of improvement below the quantum speed limit, as compared to using only the single-body $\hat{Y}$ term. We mention, in passing, that we observed a stronger system-size dependence in the optimal protocol found by the RL agent in the immediate vicinity of the critical point $h_c/J=1$.

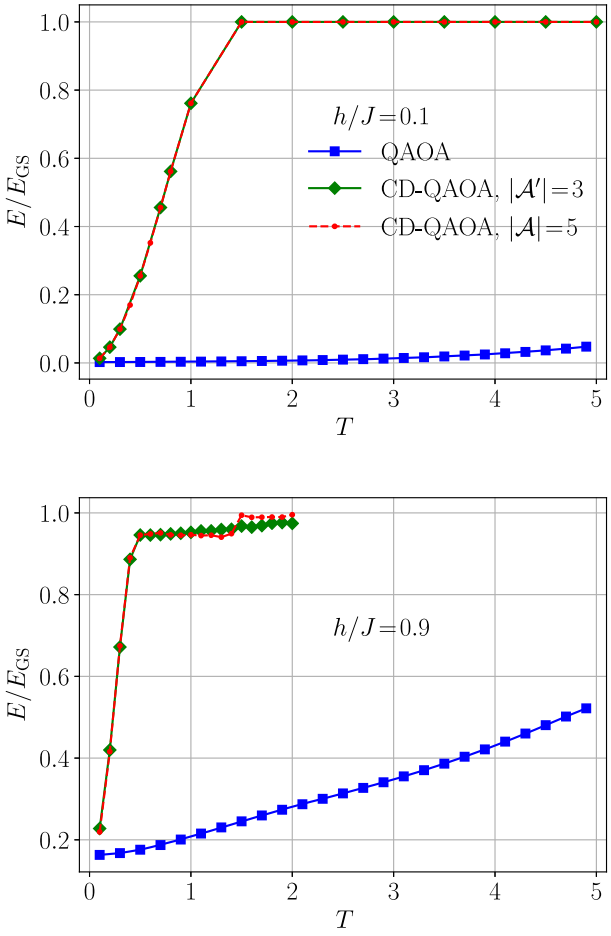

FIG. 22. LMG model: energy minimization against protocol duration T using conventional QAOA (blue square) and CD-QAOA (red dashed line, green solid line). The model parameters are the same as in Fig. 5 but for $h/J=0.1$ (top panel) and $h/J=0.9$ (bottom panel).

4. Spin-1 Ising chain

Finally, let us turn to the spin-1 Ising chain:

$$H(\lambda)=\lambda(t)H_1+H_2,$$
$$H_1=\sum_{j=1}^{N}JS^z_{j+1}S^z_j+h_xS^x_j,\qquad H_2=\sum_{j=1}^{N}h_zS^z_j \tag{F5}$$

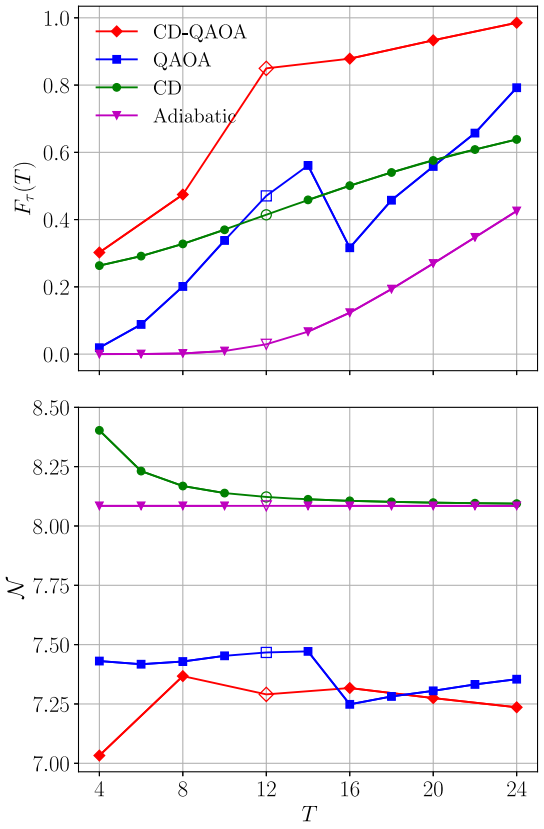

FIG. 23. Spin-1 Ising model: energy minimization against different protocol duration T for four different optimization methods: CD-QAOA (red line), conventional QAOA (blue line), variational gauge potential (green), and adiabatic evolution (magenta). Two associated quantities are shown: many-body fidelity F_τ (top panel) and normalized time-averaged energy density $\mathcal{N}$ over the protocol (bottom panel). The empty symbols mark the duration for which the evolution of physical quantities is shown in Fig. 24. The parameters are the same as in Fig. 7.

(see main text for discussion of the model parameters). Using this model, we compare four state preparation techniques: CD-QAOA, conventional QAOA, CD-driving using a variational gauge potential, and adiabatic evolution.

In order to compare these four methods, we first investigate their energy budget, i.e., the amount of energy required by the corresponding protocols. This approach is necessary since variational CD driving does not put any constraints on the magnitude of the expansion parameters $\beta_j(\lambda)$ (cf. Appendix E), and we know that larger energies (i.e., generators of unitaries H_j with large norms), in general, allow for a faster population transfer. To quantitatively measure the energy budget of a protocol, we use the average energy density along the protocol trajectory

$$\mathcal{N} = \frac{1}{T}\int_0^T \mathrm{d}t \frac{\|H(t)\|}{N}, \tag{F6}$$

where $H(t)$ is a unified notation for the continuous protocols in the case of adiabatic or CD driving, and the piecewise-constant (in time) sequences in CD-QAOA and conventional QAOA; $\|H\|$ denotes the Hilbert-Schmidt norm of the operator H. Since we are interested in many-body systems, it is also natural to look at the energy density, i.e., $\|H(t)\|/N$. Figure 23 (bottom panel) shows that $\mathcal{N}$ is on a similar scale for all four methods within the range of durations of interest, which allows for a meaningful comparison between them. As expected, CD driving approaches adiabatic driving at large T since the gauge potential term comes with a prefactor $\dot{\lambda}$, which vanishes for $T \to \infty$; in the opposite limit of $T \to 0$, the energy budget of CD driving blows up as a result of $\beta_j(\lambda)$ being unconstrained.

In Fig. 23 (top panel), we see that the many-body fidelity, associated with the protocols obtained using energy-density minimization, increases the performance contrast between the performance of the different methods (cf. Fig. 7, main text). Since the fidelity is defined as the overlap square of the final with the target states [Eq. (F3)], like the entanglement entropy, it is insensitive to any specific observable, which implies that CD-QAOA outperforms the other three methods on *all* observables, not just energy. This result is anticipated because CD-QAOA combines the variational

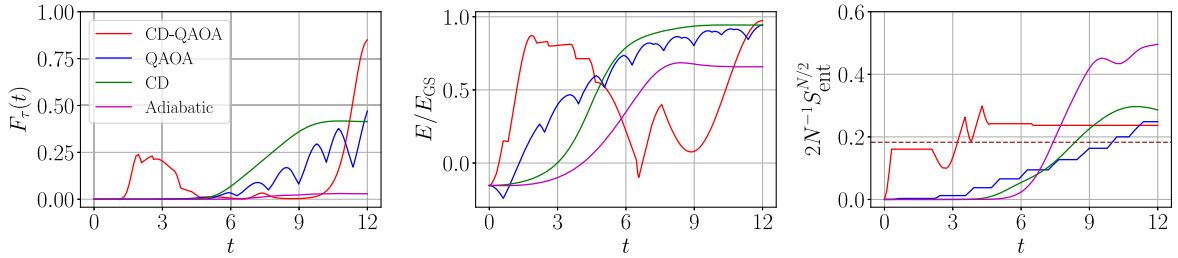

FIG. 24. Spin-1 Ising model: time evolution generated by the four different methods—CD-QAOA (red line), conventional QAOA (blue line), CD driving using the variational gauge potential (green line), and adiabatic evolution (magenta line). Three quantities are shown: the many-body fidelity (left panel), energy (middle panel), and entanglement entropy of the half chain (right panel). The protocols correspond to the empty symbols during $T = 12$ in Fig. 7. The horizontal dashed line in the entanglement entropy curve shows the value in the target state. The CD-QAOA protocol sequence is given in Table V. The model parameters are the same as in Fig. 7.

TABLE V. Ising spin-1 chain and anisotropic Heisenberg spin-1 chain, with the protocol sequences and corresponding durations given by CD-QAOA. The protocol (a) correspond to Ising spin-1 in Fig. 24; the (b), (c), (d) three sequences correspond to the three phases in the same setting as Fig. 20. The short-hand notation is the same in Table I. Terms of zero durations are marked bold.

(a) Ising spin-1

Hamiltonian	Duration
$X\|Y$	0.312
Y	0.299
Z	0.216
Y	0.717
$\mathbf{Z}$	**0.000**
Y	0.537
$Z\|Z+X$	0.477
Y	0.054
$Z\|Z+X$	0.657
$\mathbf{Z}$	**0.000**
$Z\|Z+X$	0.269
$Y\|Z$	0.274
$Z\|Z+X$	0.478
$Y\|Z$	0.372
$\mathbf{Z}\|\mathbf{Z}+\mathbf{X}$	**0.000**
Z	1.794
$X\|Y$	0.072
Z	0.039
Y	1.007
Z	4.426

(b) Ferromagnetic ($\Delta/J=-2.0$)

Shorthand notation	Duration
$Y\|Z-YZ$	0.122
$X\|X+Y\|Y$	0.178
YZ	0.027
$Z\|Z$	0.376
$Y\|Z-YZ$	0.234
$\mathbf{Z}\|\mathbf{Z}$	**0.000**
$X\|X+Y\|Y$	0.323
$Z\|Z$	0.284
$Y\|Z-YZ$	0.366
$\mathbf{Z}\|\mathbf{Z}$	**0.000**
$X\|X+Y\|Y$	0.314
$Z\|Z$	0.188
$Y\|Z-YZ$	0.535
Y	0.001
$X\|X$	0.342
$Z\|Z$	0.105
$Y\|Z-YZ$	0.538
$X\|X$	0.208
$\mathbf{Y}$	**0.000**
$Z\|Z$	0.051
Y	0.658
$Y\|Z-YZ$	0.002
Y	0.900
Z	0.771
Y	0.005
$X\|Y-XY$	0.474
$\mathbf{Y}\|\mathbf{Z}-\mathbf{YZ}$	**0.000**
$\mathbf{X}\|\mathbf{X}+\mathbf{Y}\|\mathbf{Y}$	**0.000**

(c) XY ($\Delta/J=-0.5$)

Shorthand notation	Duration
Y	0.795
$\mathbf{X}\|\mathbf{X}$	**0.000**
Y	0.772
$X\|X+Y\|Y$	0.143
$X\|X$	0.383
Y	0.001
$X\|X+Y\|Y$	0.284
$X\|X$	0.180
$X\|X+Y\|Y$	0.467
$X\|X$	0.113
$X\|X+Y\|Y$	0.635
$X\|X$	0.097
$X\|X+Y\|Y$	0.617
$\mathbf{Y}$	**0.000**
$Z\|Z$	0.162
$X\|X+Y\|Y$	0.265
$X\|X$	0.092
Z	1.995

(d) Haldane ($\Delta/J=0.5$)

Shorthand notation	Duration
$X\|X+Y\|Y$	0.149
$\mathbf{X}\|\mathbf{X}$	**0.000**
$X\|X+Y\|Y$	0.052
$Z\|Z$	1.376
$X\|X+Y\|Y$	0.313
$Z\|Z$	0.668
$X\|X+Y\|Y$	0.187
$Z\|Z$	0.723
$X\|X+Y\|Y$	0.289
$Z\|Z$	0.528
$X\|X+Y\|Y$	0.218
$Z\|Z$	0.561
$X\|X+Y\|Y$	0.254
$Z\|Z$	0.684
$X\|X+Y\|Y$	0.360
$Z\|Z$	0.639
$\mathbf{X}\|\mathbf{X}$	**0.000**
$\mathbf{Z}$	**0.000**

power of QAOA with physical insights from CD driving. Despite its better performance, notice how CD-QAOA also has a smaller energy budget than either CD and adiabatic driving.

To demonstrate the nonequilibrium character of the optimal protocols found by the RL agent in this setup, we fix $T = 12$ and look at the time evolution of the energy, the fidelity, and the entanglement entropy within the learned protocol, cf. Fig. 24. While the protocol sequence [Table V] appears impenetrable, we remark that (i) the RL agent makes use of both single-particle and two-body gauge potential terms and (ii) some step durations α_j are found to vanish identically, suggesting that the value of q may be reduced. As anticipated, the behavior of the dynamics generated by the CD and adiabatic driving is smooth, in contrast to the circuitlike, piece-wise, continuous curves of QAOA and CD-QAOA. The highly nonmonotonic behavior of the energy curve shows that the CD-QAOA dynamics can be highly nonequilibrium, which likely stems from the RL objective (cf. Appendix A)—the total expected return: the agent only cares about maximizing the reward at $t = T$ and is insensitive to any intermediate values, allowing the agent to drive the system through various states that are very far away from the target (e.g., with respect to the fidelity). (Curiously, these bad-energy states are all distinct since they have different entanglement entropy, and the system does not visit the same quantum state twice during the evolution.) The nonsmooth and nonmonotonic behavior of the CD-QAOA solution raises the question about how robust the protocol is to small external perturbations—a topic of future studies.

[1] M. Lewenstein, A. Sanpera, V. Ahufinger, B. Damski, A. Sen, and U. Sen, *Ultracold Atomic Gases in Optical Lattices: Mimicking Condensed Matter Physics and Beyond*, Adv. Phys. **56**, 243 (2007).

[2] I. Bloch, J. Dalibard, and W. Zwerger, *Many-Body Physics with Ultracold Gases*, Rev. Mod. Phys. **80**, 885 (2008).

[3] H. Häffner, C. F. Roos, and R. Blatt, *Quantum Computing with Trapped Ions*, Phys. Rep. **469**, 155 (2008).

[4] R. Blatt and C. F. Roos, *Quantum Simulations with Trapped Ions*, Nat. Phys. **8**, 277 (2012).

[5] C. Monroe and J. Kim, *Scaling the Ion Trap Quantum Processor*, Science **339**, 1164 (2013).

[6] M. H. Devoret and R. J. Schoelkopf, *Superconducting Circuits for Quantum Information: An Outlook*, Science **339**, 1169 (2013).

[7] M. W. Doherty, N. B. Manson, P. Delaney, F. Jelezko, J. Wrachtrup, and L. C. Hollenberg, *The Nitrogen-Vacancy Colour Centre in Diamond*, Phys. Rep. **528**, 1 (2013).

[8] R. Schirhagl, K. Chang, M. Loretz, and C. L. Degen, *Nitrogen-Vacancy Centers in Diamond: Nanoscale Sensors for Physics and Biology*, Annu. Rev. Phys. Chem. **65**, 83 (2014).

[9] F. Casola, T. van der Sar, and A. Yacoby, *Probing Condensed Matter Physics with Magnetometry Based on Nitrogen-Vacancy Centres in Diamond*, Nat. Rev. Mater. **3**, 17088 (2018).

[10] Z.-L. Xiang, S. Ashhab, J. Q. You, and F. Nori, *Hybrid Quantum Circuits: Superconducting Circuits Interacting with Other Quantum Systems*, Rev. Mod. Phys. **85**, 623 (2013).

[11] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell *et al.*, *Quantum Supremacy Using a Programmable Superconducting Processor*, Nature (London) **574**, 505 (2019).

[12] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrüggen, and S. J. Glaser, *Optimal Control of Coupled Spin Dynamics: Design of NMR Pulse Sequences by Gradient Ascent Algorithms*, J. Magn. Reson. **172**, 296 (2005).

[13] T. Caneva, T. Calarco, and S. Montangero, *Chopped Random-Basis Quantum Optimization*, Phys. Rev. A **84**, 022326 (2011).

[14] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'brien, *A Variational Eigenvalue Solver on a Photonic Quantum Processor*, Nat. Commun. **5**, 4213 (2014).

[15] L. Zhou, S.-T. Wang, S. Choi, H. Pichler, and M. D. Lukin, *Quantum Approximate Optimization Algorithm: Performance, Mechanism, and Implementation on Near-Term Devices*, Phys. Rev. X **10**, 021067 (2020).

[16] S. Hadfield, Z. Wang, B. O'Gorman, E. G. Rieffel, D. Venturelli, and R. Biswas, *From the Quantum Approximate Optimization Algorithm to a Quantum Alternating Operator Ansatz*, Algorithms **12**, 34 (2019).

[17] M. Demirplak and S. A. Rice, *Adiabatic Population Transfer with Control Fields*, J. Phys. Chem. A **107**, 9937 (2003).

[18] M. Berry, *Transitionless Quantum Driving*, J. Phys. A **42**, 365303 (2009).

[19] M. Kolodrubetz, D. Sels, P. Mehta, and A. Polkovnikov, *Geometry and Non-adiabatic Response in Quantum and Classical Systems*, Phys. Rep. **697**, 1 (2017).

[20] M. Bukov, D. Sels, and A. Polkovnikov, *Geometric Speed Limit of Accessible Many-Body State Preparation*, Phys. Rev. X **9**, 011034 (2019).

[21] We focus on pure states, although the cost function can trivially be generalized to mixed states.

[22] V. Jurdjevic and H. J. Sussmann, *Control Systems on Lie Groups*, J. Diff. Eqs. **12**, 313 (1972).

[23] L. Zhu, H. L. Tang, G. S. Barron, F. A. Calderon-Vargas, N. J. Mayhall, E. Barnes, and S. E. Economou, *An Adaptive Quantum Approximate Optimization Algorithm for Solving Combinatorial Problems on a Quantum Computer*, arXiv: 2005.10258.

[24] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, *Barren Plateaus in Quantum Neural Network Training Landscapes*, Nat. Commun. **9**, 4812 (2018).

[25] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, *Cost Function Dependent Barren Plateaus in Shallow Parametrized Quantum Circuits*, Nat. Commun. **12**, 1791 (2021).

[26] E. Grant, L. Wossnig, M. Ostaszewski, and M. Benedetti, *An Initialization Strategy for Addressing Barren Plateaus in Parametrized Quantum Circuits*, Quantum **3**, 214 (2019).

[27] P. Huembeli and A. Dauphin, *Characterizing the Loss Landscape of Variational Quantum Circuits*, Quantum Sci. Technol. **6**, 025011 (2021).

[28] Considering τ_j as a choice of unitaries, we impose the extra constraint that, even though unitaries can be repeated in the sequence τ, the same unitary cannot appear consecutively (or one can combine the two corresponding choices τ_j into a single variable).

[29] A. G. R. Day, M. Bukov, P. Weinberg, P. Mehta, and D. Sels, *Glassy Phase of Optimal Quantum Control*, Phys. Rev. Lett. **122**, 020601 (2019).

[30] M. Bukov, A. G. R. Day, P. Weinberg, A. Polkovnikov, P. Mehta, and D. Sels, *Broken Symmetry in a Two-Qubit Quantum Control Landscape*, Phys. Rev. A **97**, 052114 (2018).

[31] A similar procedure appeared recently in Ref. [32], although they considered a different problem setup with greedy or beam search algorithms.

[32] L. Li, M. Fan, M. Coram, P. Riley, and S. Leichenauer, *Quantum Optimization with a Novel Gibbs Objective Function and Ansatz Architecture Search*, Phys. Rev. Research **2**, 023074 (2020).

[33] In principle, one can use any optimizer that allows for constraining the sum $\sum_j \alpha_j = T$.

[34] N. Lacroix, C. Hellings, C. K. Andersen, A. D. Paolo, A. Remm, S. Lazar, S. Krinner, G. J. Norris, M. Gabureac, J. Heinsoo, A. Blais, C. Eichler, and A. Wallraff, *Improving the Performance of Deep Quantum Optimization Algorithms with Continuous Gate Sets*, PRX Quantum **1**, 110304 (2020).

[35] Y. Ding, Y. Ban, J. D. Martín-Guerrero, E. Solano, J. Casanova, and X. Chen, *Breaking Adiabatic Quantum Control with Deep Learning*, Phys. Rev. A **103**, L040401 (2021).

[36] D. Sels and A. Polkovnikov, *Minimizing Irreversible Losses in Quantum Systems by Local Counterdiabatic Driving*, Proc. Natl. Acad. Sci. U.S.A. **114**, E3909 (2017).

[37] A. Hartmann and W. Lechner, *Rapid Counter-Diabatic Sweeps in Lattice Gauge Adiabatic Quantum Computing*, New J. Phys. **21**, 043025 (2019).

[38] J. Wurtz, P. W. Claeys, and A. Polkovnikov, *Variational Schrieffer-Wolff Transformations for Quantum Many-Body Dynamics*, Phys. Rev. B **101**, 014302 (2020).

[39] N. N. Hegade, K. Paul, Y. Ding, M. Sanz, F. Albarrán-Arriagada, E. Solano, and X. Chen, *Shortcuts to Adiabaticity in Digitized Adiabatic Quantum Computing*, Phys. Rev. Applied **15** (2021).

[40] M. Pandey, P. W. Claeys, D. K. Campbell, A. Polkovnikov, and D. Sels, *Adiabatic Eigenstate Deformations as a Sensitive Probe for Quantum Chaos*, Phys. Rev. X **10**, 041017 (2020).

[41] Below, we sometimes abuse notation and set $\mathcal{A} = \{H_j\}$, denoting the set of unitaries by their generators.

[42] J. Wurtz and P. J. Love, *Counterdiabaticity and the Quantum Approximate Optimization Algorithm*, arXiv:2106.15645.

[43] G. Matos, S. Johri, and Z. Papić, *Quantifying the Efficiency of State Preparation via Quantum Variational Eigensolvers*, PRX Quantum **2**, 010309 (2021).

[44] W. W. Ho and T. H. Hsieh, *Efficient Variational Simulation of Non-trivial Quantum States*, SciPost Phys. **6**, 29 (2019).

[45] The role of the RL algorithm is to decide which three out of the five unitaries U_j to apply and in which order.

[46] We define "order" in the context of phase transitions in condensed matter physics.

[47] W. Chen, K. Hida, and B. C. Sanctuary, *Ground-State Phase Diagram of $S = 1$ XXZ Chains with Uniaxial Single-Ion-Type Anisotropy*, Phys. Rev. B **67**, 104401 (2003).

[48] F. Pollmann, A. M. Turner, E. Berg, and M. Oshikawa, *Entanglement Spectrum of a Topological Phase in One Dimension*, Phys. Rev. B **81**, 064439 (2010).

[49] A. Langari, F. Pollmann, and M. Siahatgar, *Ground-State Fidelity of the Spin-1 Heisenberg Chain with Single Ion Anisotropy: Quantum Renormalization Group and Exact Diagonalization Approaches*, J. Phys. Condens. Matter **25**, 406002 (2013).

[50] H. Lipkin, N. Meshkov, and A. Glick, *Validity of Many-Body Approximation Methods for a Solvable Model: (I). Exact Solutions and Perturbation Theory*, Nucl. Phys. **62**, 188 (1965).

[51] R. Botet and R. Jullien, *Large-Size Critical Behavior of Infinitely Coordinated Systems*, Phys. Rev. B **28**, 3955 (1983).

[52] H. Strobel, W. Muessel, D. Linnemann, T. Zibold, D. B. Hume, L. Pezzè, A. Smerzi, and M. K. Oberthaler, *Fisher Information and Entanglement of Non-Gaussian Spin States*, Science **345**, 424 (2014).

[53] E. J. Davis, A. Periwal, E. S. Cooper, G. Bentsen, S. J. Evered, K. Van Kirk, and M. H. Schleier-Smith, *Protecting Spin Coherence in a Tunable Heisenberg Model*, Phys. Rev. Lett. **125**, 060402 (2020).

[54] We deliberately use a different form in Eq. (6) as compared to Eq. (3); the former may appear more natural in quantum many-body physics, where the transverse-field Ising model H_1 can be mapped to free fermions.

[55] M. J. S. Beach, R. G. Melko, T. Grover, and T. H. Hsieh, *Making Trotters Sprint: A Variational Imaginary Time Ansatz for Quantum Many-Body Systems*, Phys. Rev. B **100**, 094434 (2019).

[56] P. Weinberg and M. Bukov, *QuSpin: A Python Package for Dynamics and Exact Diagonalisation of Quantum Many Body Systems Part I: Spin Chains*, SciPost Phys. **2**, 003 (2017).

[57] P. Weinberg and M. Bukov, *QuSpin: A Python Package for Dynamics and Exact Diagonalisation of Quantum Many Body Systems. Part II: Bosons, Fermions and Higher Spins*, SciPost Phys. **7**, 20 (2019).

[58] V. Dunjko and H. J. Briegel, *Machine Learning & Artificial Intelligence in the Quantum Domain: A Review of Recent Progress*, Rep. Prog. Phys. **81**, 074001 (2018).

[59] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, *A High-Bias, Low-Variance Introduction to Machine Learning for Physicists*, Phys. Rep. **810**, 1 (2019).

[60] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, *Machine Learning and the Physical Sciences*, Rev. Mod. Phys. **91**, 045002 (2019).

[61] J. Carrasquilla, *Machine Learning for Quantum Matter*, Adv. Phys. X **5**, 1797528 (2020).

[62] K. J. Sung, J. Yao, M. P. Harrigan, N. C. Rubin, Z. Jiang, L. Lin, R. Babbush, and J. R. McClean, *Using Models to Improve Optimizers for Variational Quantum Algorithms*, Quantum Sci. Technol. **5**, 044008 (2020).

[63] F. Schäfer, M. Kloc, C. Bruder, and N. Lörch, *A Differentiable Programming Method for Quantum Control*, Mach. Learn. **1**, 035009 (2020).

[64] F. Sauvage and F. Mintert, *Optimal Quantum Control with Poor Statistics*, PRX Quantum **1**, 020322 (2020).

[65] T. Fösel, S. Krastanov, F. Marquardt, and L. Jiang, *Efficient Cavity Control with Snap Gates*, arXiv:2004.14256.

[66] R.-B. Wu, X. Cao, P. Xie, and Y.-X. Liu, *End-to-End Quantum Machine Learning Implemented with Controlled Quantum Dynamics*, Phys. Rev. Applied **14**, 064020 (2020).

[67] F. Albarrán-Arriagada, J. C. Retamal, E. Solano, and L. Lamata, *Measurement-Based Adaptation Protocol with Quantum Reinforcement Learning*, Phys. Rev. A **98**, 042315 (2018).

[68] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, *Reinforcement Learning with Neural Networks for Quantum Feedback*, Phys. Rev. X **8**, 031084 (2018).

[69] H. P. Nautrup, N. Delfosse, V. Dunjko, H. J. Briegel, and N. Friis, *Optimizing Quantum Error Correction Codes with Reinforcement Learning*, Quantum **3**, 215 (2019).

[70] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 2018).

[71] D. C. Rose, J. F. Mair, and J. P. Garrahan, *A Reinforcement Learning Approach to Rare Trajectory Sampling*, New J. Phys. **23**, 013013 (2021).

[72] M. Y. Niu, S. Boixo, V. N. Smelyanskiy, and H. Neven, *Universal Quantum Control through Deep Reinforcement Learning*, npj Quantum Inf. **5**, 33 (2019).

[73] M. August and J. M. Hernández-Lobato, *Taking Gradients through Experiments: LSTMs and Memory Proximal Policy Optimization for Black-Box Quantum Control*, in *High Performance Computing* (Springer International Publishing, Cham, 2018), pp. 591–613, https://doi.org/10.1007/978-3-030-02465-9_43.

[74] R. Porotti, D. Tamascelli, M. Restelli, and E. Prati, *Coherent Transport of Quantum States by Deep Reinforcement Learning*, Commun. Phys. **2**, 61 (2019).

[75] M. Bukov, *Reinforcement Learning for Autonomous Preparation of Floquet-Engineered States: Inverting the Quantum Kapitza Oscillator*, Phys. Rev. B **98**, 224305 (2018).

[76] M. Bukov, A. G. R. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta, *Reinforcement Learning in Different Phases of Quantum Control*, Phys. Rev. X **8**, 031086 (2018).

[77] M. Dalgaard, F. Motzoi, J. J. Sørensen, and J. Sherson, *Global Optimization of Quantum Dynamics with AlphaZero Deep Exploration*, npj Quantum Inf. **6**, 6 (2020).

[78] J. Yao, M. Bukov, and L. Lin, *Policy Gradient Based Quantum Approximate Optimization Algorithm*, in *Mathematical and Scientific Machine Learning Conference (MSML), 2020* (PMLR, Princeton, NJ, USA, 2020), Vol. 107, pp. 605–634, http://proceedings.mlr.press/v107/yao20a.html.

[79] M. M. Wauters, E. Panizon, G. B. Mbeng, and G. E. Santoro, *Reinforcement Learning Assisted Quantum Optimization*, Phys. Rev. Research **2**, 033446 (2020).

[80] S. Khairy, R. Shaydulin, L. Cincio, Y. Alexeev, and P. Balaprakash, *Reinforcement-Learning-Based Variational Quantum Circuits Optimization for Combinatorial Problems*, arXiv:1911.04574.

[81] A. Garcia-Saez and J. Riu, *Quantum Observables for Continuous Control of the Quantum Approximate Optimization Algorithm via Reinforcement Learning*, arXiv:1911.09682.

[82] A. Bolens and M. Heyl, *Reinforcement Learning for Digital Quantum Simulation*, Phys. Rev. Lett. **127**, 110502 (2021).

[83] It is also possible to define a RL framework for hybrid continuous-discrete control where optimization is entirely based on RL, cf. Ref. [84].

[84] J. Yao, P. Köttering, H. Gundlach, L. Lin, and M. Bukov, *Noise-Robust End-to-End Quantum Control Using Deep Autoregressive Policy Networks, Mathematical and Scientific Machine Learning Conference, 2021*, arXiv:2012.06701.

[85] M. Bukov, *Reinforcement Learning for Autonomous Preparation of Floquet-Engineered States: Inverting the Quantum Kapitza Oscillator*, Phys. Rev. B **98**, 224305 (2018).

[86] D. Wu, L. Wang, and P. Zhang, *Solving Statistical Mechanics Using Variational Autoregressive Networks*, Phys. Rev. Lett. **122**, 080602 (2019).

[87] O. Sharir, Y. Levine, N. Wies, G. Carleo, and A. Shashua, *Deep Autoregressive Models for the Efficient Variational Simulation of Many-Body Quantum Systems*, Phys. Rev. Lett. **124**, 020503 (2020).

[88] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, arXiv:1412.6980.

[89] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, *Proximal Policy Optimization Algorithms*, arXiv:1707.06347.

[90] B. D. Ziebart, *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy* (Carnegie, Pittsburgh, 2010).

[91] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, *Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor*, arXiv:1801.01290.

[92] J. Nocedal and W. Stephen, in *Numerical Optimization* (Springer Science & Business Media, New York, 2006).

[93] T. Hatomura, *Shortcuts to Adiabaticity in the Infinite-Range Ising Model by Mean-Field Counter-Diabatic Driving*, J. Phys. Soc. Jpn. **86**, 094002 (2017).

[94] Z. Mzaouali, R. Puebla, J. Goold, M. E. Baz, and S. Campbell, *Work Statistics and Symmetry Breaking in an Excited-State Quantum Phase Transition*, Phys. Rev. E **103**, 032145 (2021).