








Single-atom exploration of optimized nonequilibrium quantum thermodynamics by reinforcement learning

Jiawei Zhang ^{1,8}, Jiachong Li^{1,2,3,8}, Qing-Shou Tan ⁴✉, Jintao Bu^{2,3}, Wenfei Yuan^{2,3}, Bin Wang^{2,3}, Geyi Ding^{2,3}, Wenqiang Ding^{2,3}, Liang Chen ^{1,2}, Leilei Yan⁵, Shilei Su ⁵, Taiping Xiong ⁶✉, Fei Zhou ^{1,2}✉ & Mang Feng ^{1,2,7}✉

Exploring optimized processes of thermodynamics at microscale is vital to exploitation of quantum advantages relevant to microscopic machines and quantum information processing. Here, we experimentally execute a reinforcement learning strategy, using a single trapped $^{40}\text{Ca}^+$ ion, for engineering quantum state evolution out of thermal equilibrium. We consider a qubit system coupled to classical and quantum baths, respectively, the former of which is achieved by switching on the spontaneous emission relevant to the qubit and the latter of which is made based on a Jaynes-Cummings model involving the qubit and the vibrational degree of freedom of the ion. Our optimized operations make use of the external control on the qubit, designed by the reinforcement learning approach. In comparison to the conventional situation of free evolution subject to the same Hamiltonian of interest, our experimental implementation presents the evolution of the states with higher fidelity while with less consumption of entropy production and work, highlighting the potential of reinforcement learning in accomplishment of optimized nonequilibrium thermodynamic processes at atomic level.

¹Research Center for Quantum Precision Measurement, Guangzhou Institute of Industry Technology, 511458 Guangzhou, China. ²State Key Laboratory of Magnetic Resonance and Atomic and Molecular Physics, Wuhan Institute of Physics and Mathematics, Innovation Academy of Precision Measurement Science and Technology, Chinese Academy of Sciences, 430071 Wuhan, China. ³School of Physics, University of the Chinese Academy of Sciences, 100049 Beijing, China. ⁴Key Laboratory of Hunan Province on Information Photonics and Freespace Optical Communication, College of Physics and Electronics, Hunan Institute of Science and Technology, 414000 Yueyang, China. ⁵School of Physics, Zhengzhou University, 450001 Zhengzhou, China. ⁶Key Laboratory of Quantum Information Technology, Guilin University of Electronic Technology, 541004 Guilin, China. ⁷Department of Physics, Zhejiang Normal University, 321004 Jinhua, China. ⁸These authors contributed equally: Jiawei Zhang, Jiachong Li. ✉email: qstan@hnist.edu.cn; xiongt@guet.edu.cn; zhoufei@wipm.ac.cn; mangfeng@wipm.ac.cn

High-precision quantum control is crucial for quantum information processing, which could be, in principle, achieved following adiabatic theorem¹. However, to keep quantum properties, people are required to work fast in quantum systems for suppressing detrimental influence from dissipation/decoherence, indicating that nonequilibrium dynamics dominates quantum processes². In fact, the nonequilibrium processes in thermodynamics have drawn much attention over the past decades, where the most influential works, such as Jarzynski equation^{3,4} and the fluctuation theorem^{5,6}, have provided us new insights into the stochastic dynamics out of equilibrium. In particular, due to rapid progress in quantum technology, understanding the nonequilibrium thermodynamic process at the quantum level has recently turned to be a burgeoning field. Extended to the quantum regime, most thermodynamic quantities should be retraced and thus most restrictions of the original thermodynamics have to be reformulated. For example, irreversibility, which is relevant to the conventional second law of thermodynamics, has been reconsidered from the angle of information theory. Quantified by relative entropy, irreversibility in either closed or open quantum system is restricted by more strict bounds than the conventional second law of thermodynamics^{7–9}. However, in contrast to the fact that all real macroscopic processes are definitely irreversible, irreversibility in microscopic processes seems much more complicated, as discovered in a recent experiment¹⁰.

We have noticed a recent proposal¹¹ exploring reduction of irreversibility in the finite-time nonequilibrium thermodynamic transformations of a closed quantum system. The results indicate that the relative entropy could be much reduced in the nonequilibrium thermodynamics if strategies of reinforcement learning (RL)¹² are employed. RL is an important paradigm of machine learning, the latter of which has been widely employed in the study of quantum physics over the past years^{13–27}. From the simulation of characteristics of open quantum many-body systems to the design of quantum protocols, researchers have much benefited from the unique characteristics of intelligent algorithms and automation of machine learning. RL aims to tackle the problems of quantum control, which works between decision-making entities (called agents) and environment by updating their behavior based on the obtained feedback, see Fig. 1a. RL approaches have successfully achieved optimized quantum tasks from fast and robust control of a single qubit²⁸ to efficient solution to quantum many-body systems^{18,29}. Since they require little knowledge of the dynamic details of the system, RL approaches have outstanding advantages in practical quantum control over conventional optimized counterparts using, such as shortcuts to adiabaticity^{30,31}, non-adiabatic schemes³², and non-cycling geometric ideas^{33–36}. Recently, RL incorporating quantum technologies into the agents design has demonstrated speed-up in accomplishment of quantum operations^{37–39}.

Here, we report our experimental investigation of optimized nonequilibrium quantum thermodynamics in an ultracold ⁴⁰Ca⁺ ion, by a typical RL approach - the deep deterministic policy gradient (DDPG) algorithm⁴⁰, which is an advanced actor-critic algorithm. In our case, the network (also called agent), provides choices of the actions (i.e., the control of Rabi oscillation strength and/or phase) for the environment (i.e., the qubit) to maximize the reward, which is to achieve the nonequilibrium thermodynamics with higher fidelity. Specifically, the total training duration τ is divided into n steps. At each step t , for a given environment state ρ_t , the agent generates an action a_t , by the main network, and obtains the next state ρ_{t+1} as well as an immediate reward r_t from the environment. Then the goal of RL is to find the optimized action set $\{a_t\}$ (i.e., the control sequences) which reaches the maximum total reward $R = \sum_{t=1}^n r_t$. This procedure is

repeated until the networks are converged. The DDPG algorithm makes sure the optimized route of evolution corresponding to the largest accumulated reward. We focus on RL-controlled nonequilibrium thermodynamic transformations relevant to two key operations in ion trap, i.e., operations for carrier transition and red-sideband transition, demonstrating the advantages of high-precision control and reduced irreversibility. Our experimental observations have provided evidences for the advantages of RL strategy in optimizing quantum state control.

Results and discussion

The experiment platform. In our experiment, the ion is confined in a linear Paul trap with axial frequency $\omega_z/2\pi = 0.94$ MHz and radial frequency $\omega_r/2\pi = 1.2$ MHz, and the quantization axis is defined with respect to the axial direction at an angle 45° by a magnetic field of approximately 6.23 Gauss at the center of the trap. For our purpose, we have cooled the ion, prior to the experiment, down to near the ground state of the vibrational modes, and encode the qubit in the electronic states $|4^2S_{1/2}, m_j = -1/2\rangle$ (labeled as $|g\rangle$) and $|3^2D_{5/2}, m_j = -5/2\rangle$ (labeled as $|e\rangle$) with m_j the magnetic quantum number⁴¹. As plotted in Fig. 1b, the qubit is manipulated by an ultra-stable narrow linewidth 729-nm laser in the case of the Lamb-Dicke parameter $\eta \sim 0.1$. In what follows, we consider the optimized nonequilibrium quantum thermodynamics subject to classical bath and quantum bath, respectively, which is beyond the model in Ref. 11. Specifically, we manipulate the single qubit by switching on (off) the spontaneous emission to demonstrate the RL-engineered nonequilibrium quantum thermodynamics with (without) dissipation, and execute a Jaynes-Cummings model to explore the single qubit coupled to a quantum bath played by the vibrational degree of freedom of the ion, as shown in Fig. 1c. In the latter case, for simplicity, the classical bath is excluded, implying that no spontaneous emission occurs in the qubit.

Operations with a single qubit. To construct the dissipative channel to the classical bath, we introduce an extra energy level $|4^2P_{3/2}, m_j = -3/2\rangle$ which couples to $|e\rangle$ by a 854-nm laser (with Rabi frequency Ω_1) and dissipates to $|g\rangle$ with the decay rate of $\Gamma/2\pi = 23.1$ MHz. Under appropriate laser irradiation as tested in refs. 10,42, an effective two-level model with engineered drive $\tilde{\Omega}(t)$ and decay $\gamma_{\text{eff}} = \Omega_1^2/\Gamma$ can be achieved, as shown in Fig. 1b, which can be described by the Lindblad master equation as,

$$\dot{\rho} = -i[H_s, \rho] + \frac{\gamma_{\text{eff}}}{2}(2\sigma_- \rho \sigma_+ - \sigma_+ \sigma_- \rho - \rho \sigma_+ \sigma_-), \quad (1)$$

where ρ denotes the density operator, and H_s is the single-qubit Hamiltonian. Experimentally, the single-qubit rotations are accomplished by carrier transitions, following the time-dependent Hamiltonian within the time interval $[0, \tau]$,

$$H_s(t) = \frac{\Omega_0}{2}[\sigma_x \cos(\pi t/2\tau) + \sigma_y \sin(\pi t/2\tau)], \quad (2)$$

which represents a time-dependent single-qubit rotation with the maximal Rabi frequency Ω_0 and usual Paul operators $\sigma_{x,y}$ for $|e\rangle$ and $|g\rangle$.

The RL-assistant evolutions are executed as plotted in Fig. 1d, and then compared with the free evolution under the government of the same Hamiltonian and from the same initial state $\rho^{\text{eq}}(0) = \exp(-\beta H_0)/Z_0$ with the partition function $Z_0 = \text{tr}[\exp(-\beta H_0)]$ and $H_0 \equiv H_s(0) = \Omega_0 \sigma_x/2$. In our experiment, this initial Gibbs state $\rho^{\text{eq}}(0)$ of the qubit is prepared by RL. To this end, we first cool the trapped ion down to near the ground vibrational state and then prepare the state $|g\rangle$ by optical pumping. Then, the

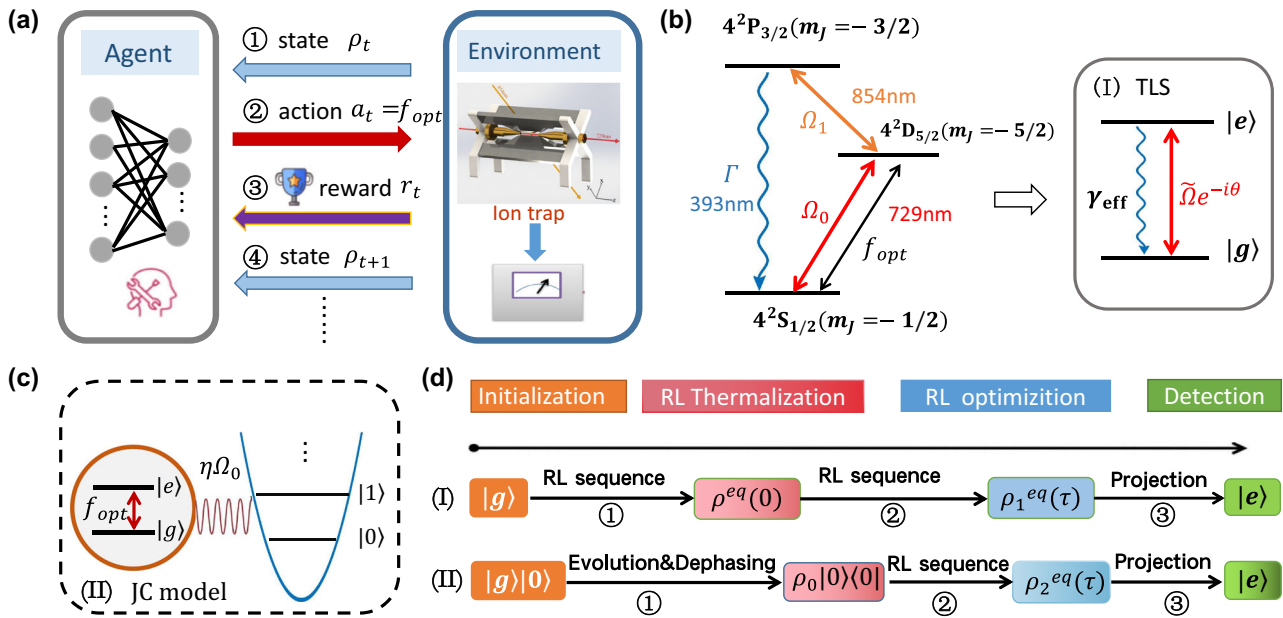


Fig. 1 Schematic for experimental method and steps. **a** Schematic for reinforcement learning (RL) method, where the agent (i.e., the network) interacts with the environment (i.e., the qubit) by repeated actions and rewards. **b** Level scheme of the $^{40}\text{Ca}^+$ ion, where the double-sided arrows and the wavy arrow represent the laser irradiation and dissipation, respectively. An effective two-level system (TLS) with controllable driving and decay is approximated from the three-level configuration. The dissipative qubit is achieved by switching on the 854-nm laser, which describes the qubit coupled to a classical bath. **c** The Jaynes-Cummings (JC) interaction between the qubit and the vibrational degree of freedom of the ion, describing the qubit coupled to a quantum bath. **d** RL-controlled experimental steps for (I) TLS and (II) JC model, respectively, starting from the end of the sideband cooling to the final detection.

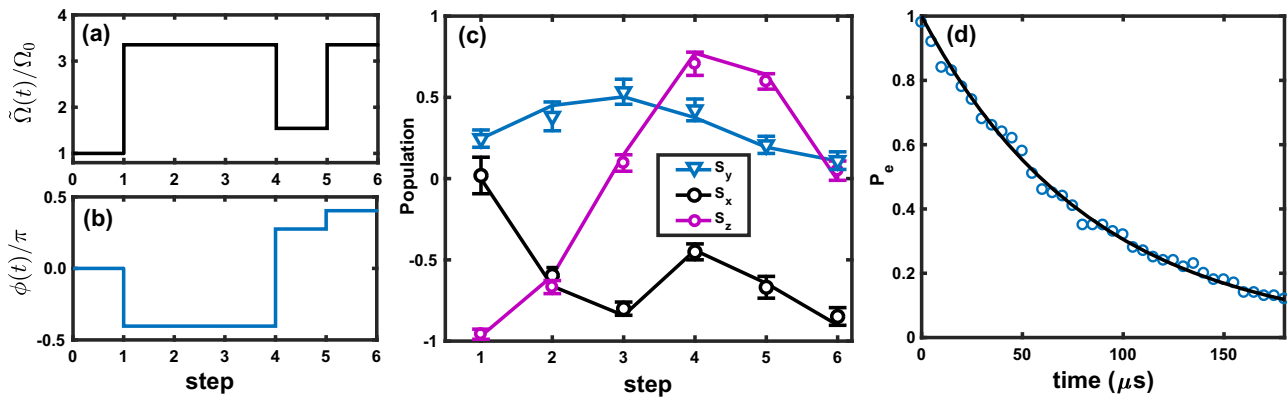


Fig. 2 Initial state preparation by reinforcement learning (RL) method. **a, b** Designed pulses of the effective Rabi frequency $\tilde{\Omega}$ and phase ϕ for initial state preparation using RL method, respectively. **c** Experimental measurement of the Stokes parameters S_x , S_y , and S_z , in comparison with the theoretical simulation, where the Stokes parameters are acquired by measuring the populations from x, y and z directions (see Supplementary Note 3). The error bars indicating the statistical standard deviation of the experimental data are obtained by 10,000 measurements for each data point. After the last step, the fidelity is $F = 0.9799 \pm 0.0103$. **d** Time evolution of the population P_e due to decay, from which we acquire decay $\gamma_{eff} = 11.99$ kHz. Other parameters: Rabi frequency $\Omega_0/2\pi = 20$ kHz and time for every step $\delta\tau = 0.25/\Omega_0$.

initial Gibbs state is prepared by the following Hamiltonian,

$$H_0(t) = \frac{\Omega_0}{2} \sigma_x - f_{opt}(t) \sigma_y, \quad (3)$$

where $f_{opt}(t) \sigma_y$ is the control term imposed by the RL approach. Experimentally, this Hamiltonian can be carried out by using a single beam of 729-nm laser with time-dependent Rabi frequency $\tilde{\Omega}(t) = \sqrt{\Omega_0^2 + 4f_{opt}^2}$ and phase $\phi(t) = \arctan(2f_{opt}/\Omega_0)$. Meanwhile, the 854-nm laser is switched on to construct the dissipative channel. Therefore, the whole system is governed by the Lindblad master equation as given in Eq. (1). Under the RL-designed optimal pulses (see Fig. 2a, b, after six steps (12 μs), the system is

thermalised to the state $\rho^{eq}(0)$. As shown in Fig. 2c, we measure the time evolutions of the Stokes parameters $S_{i \in \{x,y,z\}} \equiv \text{Tr}[\sigma_i \rho]$ in comparison with the theoretical simulation (see Supplementary Note 3 for more details). Figure 2d shows how to experimentally acquire the values of the dissipative parameter γ_{eff} by switching on the 854-nm laser, which is required for the state preparation.

To precisely execute the RL operations, we employ the arbitrary waveform generator as the frequency source of the acousto-optic modulator, which provides the phase and frequency control of the 729-nm laser during the experimental implementation. Due to fast gate operations (e.g., $<60 \mu\text{s}$) in our experiment, the qubit dephasing ($\approx 0.81(11)$ kHz)⁴³ originated

from the magnetic and electric field fluctuations, is negligible throughout this paper.

After preparing the initial state $\rho^{eq}(0)$, we intend to witness that the RL-assistant evolutions achieve higher fidelity while consuming less entropy production in both the cases of closed and open quantum systems. In the following operations, for a comparison, we execute Eq. (2) for the free evolution of the qubit, while carry out the optimal nonequilibrium thermodynamic transformation under the RL control, which imposes a control term into Eq. (2) as $H = H_s(t) - f_{\text{opt}}(t)\sigma_y$, with $f_{\text{opt}}(t) \in [-1.5\Omega_0, 1.5\Omega_0]$, $f_{\text{opt}}(0) = f_{\text{opt}}(\tau) = 0$ and $t \in [0, \tau]$. As a result, the RL-controlled Hamiltonian is given by

$$H(t) = \frac{\tilde{\Omega}(t)}{2} (e^{i\phi(t)}\sigma_+ + e^{-i\phi(t)}\sigma_-), \quad (4)$$

which is a modified form of Eq. (2) with Rabi frequency $\tilde{\Omega}(t) = \sqrt{\Omega_0^2 + 4f_{\text{opt}}[f_{\text{opt}} - \Omega_0 \sin(t\pi/2\tau)]}$ and phase $\phi(t) = \arctan\{[2f_{\text{opt}} - \Omega_0 \sin(t\pi/2\tau)]/[\Omega_0 \cos(t\pi/2\tau)]\}$. So our optimal nonequilibrium thermodynamics is carried out by controlling the power and phase of the 729-nm laser irradiation, under the government of Eq. (4). Moreover, we execute our operations in an open (a closed) system by simply switching on (off) the 854-nm laser, i.e., $\gamma_{\text{eff}} \neq 0$ ($\gamma_{\text{eff}} = 0$). In either the open or closed case, the system is initialized from a thermal state $\rho^{eq}(0)$. The target state is set to be the equilibrium thermal state of $H_s(\tau)$, i.e., $\rho_1^{eq}(\tau) = \exp(-\beta H_s(\tau))/Z_s$ with the partition function $Z_s = \text{tr}[\exp(-\beta H_s(\tau))]$.

Our goal with DDPG algorithm is to search for optimal controls of Rabi frequency $\tilde{\Omega}(t)$ and phase $\phi(t)$ that maximize the fidelity F defined as $F = \left[\text{tr}[\sqrt{\rho^{1/2}(t)\rho^{eq}(\tau)\rho^{1/2}(t)}] \right]^2$, where $\rho(t)$ is the actual density operator of the system at time t . For our purpose, we engineer Hamiltonian $H(t)$, in terms of the RL designed pulses, to evolve to approach the target state in a nonequilibrium way. The nonequilibrium entropy production associated with such a transformation can be quantified by the quantum relative entropy $\Sigma(t) = S(\rho(t)||\rho^{eq}(\tau))$. The fidelity evaluates how good the controlled evolution is, and the relative entropy assesses how much irreversibility is involved in the process. As numerically verified in Supplementary Note 2, in our case of the nonequilibrium thermodynamics, the variation of the fidelity is inversely proportional to that of the relative entropy, indicating that the near unity fidelity corresponds to the nearly zero relative entropy. As such, our RL reward to the agent is also the minimized relative entropy. For convenience of comparison, we define $\Delta\Sigma$ to be the reduced entropy production with respect to the free evolution situation, i.e., $\Delta\Sigma(t) = 1 - \Sigma_{\text{opt}}(t)/\Sigma_{\text{free}}(t)$ with $\Sigma_{\text{opt}}(t)$ ($\Sigma_{\text{free}}(t)$) the relative entropy of the state under RL control (free evolution). Since Σ_{free} is constant in a certain evolution, the bigger value of $\Delta\Sigma$ implies the less consumption of Σ_{opt} . In addition, we assess the reduction of the work done to the system during this process by defining $dW = 1 - (\Delta U_{\text{opt}} + E_{\text{in}})/\Delta U_{\text{free}}$ with $E_{\text{in}} = \int_0^1 \text{tr}[\rho(t)f_{\text{opt}}(t)\sigma_y]dt$ being the energetic cost of the optimization and ΔU the change of the internal energy between the initial and final states¹¹. Moreover, we also assess the variation of coherence, which is the norm of the off-diagonal terms of the density matrix in this process, i.e., $\sqrt{S_x^2 + S_y^2}$. The coherence, as a complement of the fidelity or the relative entropy, presents the quantum characteristics of the system.

Before studying the optimized nonequilibrium quantum thermodynamics in the open system, we first consider the closed case with $\gamma_{\text{eff}} = 0$. Using DDPG algorithm, we separate the whole dynamics into 13 steps under the designed optimal pulses, see

Fig. 3a. After each step, the reward is given for the case with the maximized fidelity F . Driving the system initially from the mixed state $\rho^{eq}(0)$ in terms of these pulses, we experimentally observe the state evolutions approaching the target state $\rho^{eq}(\tau)$ with fidelity of near unity, as demonstrated in Fig. 3c. Meanwhile, we also find the engineered state evolution with much less entropy production and much less work done to the system than the free evolution counterparts.

Checking the evolution tracks in Bloch sphere, we see that, in comparison with the simple route of the free evolution, the RL-engineered evolution is complicated with some parts of zigzag fashion, which corresponds to the dynamics of quantum coherence shown in Fig. 3b. In fact, the advantage of this complicated evolution track is also reflected in the robustness against the systematic imperfection, i.e., the imperfect Rabi frequency and the inaccurate resonance frequency. Our observation in Fig. 3d, e reveals the higher fidelity assisted by RL pulses than in the free evolution within a large range of deviation, implying more robustness against the imperfections. In contrast, $\Delta\Sigma$ and dW are less robust, while still behave well as the deviation is not beyond 10%. These observations indicate that the RL approach owns a significant superiority in high-precision and optimized control of the state evolution. Moreover, we see larger uncertainty for the measured dW with respect to those for F_{opt} and $\Delta\Sigma$ in Fig. 3d, e, which mainly resulted from the integral of E_{in} that accumulates the uncertainties of different parameters involved. Due to this reason, E_{in} is sensitive to the state variation and thus brings about bigger uncertainty for larger deviation. In this sense, the fidelity and the relative entropy, instead of the reduction of the work cost, are more suitable for assessing the nonequilibrium thermodynamics.

Now, we carry out the above experimental steps again in an open system by switching on the 854-nm laser beam, which turns the isolated two-level system into the dissipative two levels as plotted in Fig. 1b. Compared to the closed case, we have an additional parameter $\gamma_{\text{eff}} = 0.0216\Omega_0$ in our treatment by setting the power of the 854-nm laser. Here we still segment the whole dynamics into 13 steps under the DDPG algorithm, see Fig. 4a, and the system is initially prepared in the same thermal state as in the above-closed case. Figure 4b gives the coherence dynamical with or without RL-control. In Fig. 4c we still observe the RL-engineered evolution with higher performance than the free evolution.

However, due to dissipation, the population in $|e\rangle$ declines in time. Nevertheless, we witness at finite time, e.g., $\tau = 26 \mu\text{s}$, the similar robustness of the RL-engineered evolution to the imperfections, as shown in Fig. 4d, e. The situation with larger decay rates can be found in Supplementary Note 1.

Operations with Jaynes–Cummings model. To further characterize the performance of the RL method, we have also carried out an experiment for a JC interaction between the qubit and the vibrational degree of freedom of the ion, that is,

$$H_{s2} = \frac{\Omega_0}{2} \eta(\sigma_+ a + \sigma_- a^\dagger), \quad (5)$$

with a^\dagger (a) the creation (annihilation) operator of the phonon. This is to describe a qubit coupled to a quantum bath played by the vibrational mode of the ion. For simplicity, we exclude the dissipation of the qubit to the classical bath, and thus switching off the 866 nm laser throughout this experiment. To investigate the advantage of RL, we still compare the control of the Hamiltonians above by RL approach and free evolution. Following the RL control as $H_{sr}(t) = H_{s2} - f_{\text{opt}}(t)\sigma_x$, where $H_{\text{opt}}(t) = -f_{\text{opt}}(t)\sigma_x = -\tilde{\Omega}(t)e^{i\theta(t)}\sigma_x$ is our control term with

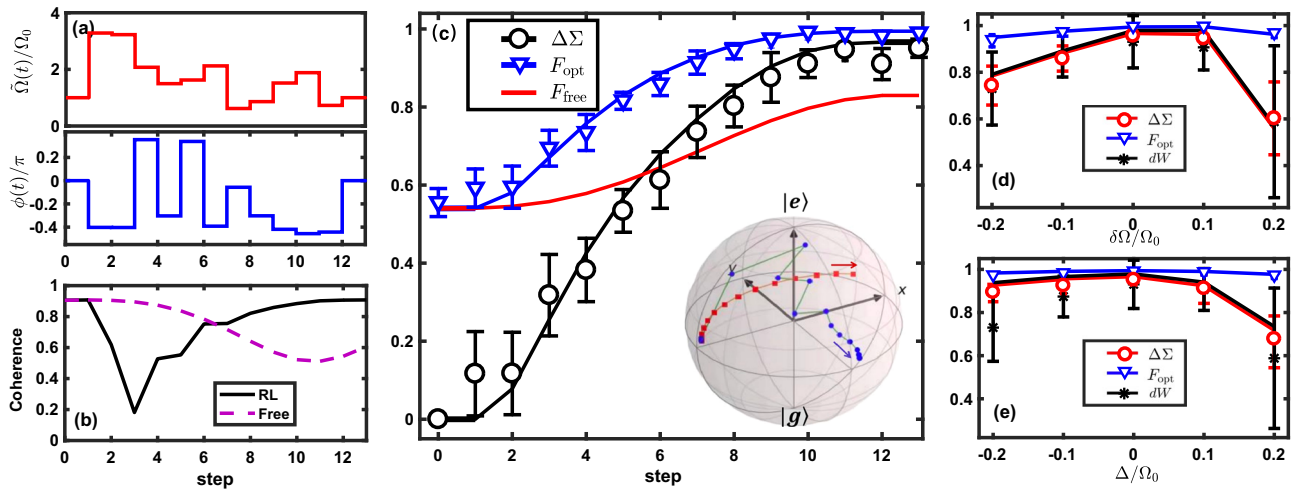


Fig. 3 Single-qubit nonequilibrium thermodynamic transformations in the closed system. **a** Designed pulses of the effective Rabi frequency $\tilde{\Omega}$ and phase ϕ for, segmented by 13 steps, due to reinforcement learning (RL) control. **b** Dynamics of coherence with and without RL-control. **c** Time evolutions of the fidelity F and the entropy production reduction $\Delta\Sigma$, where dots are experimental results and lines represent theoretical simulation. The blue and red lines denote the RL-engineered evolution and free evolution, respectively, and the black line is plotted for variation of $\Delta\Sigma$. After the last step, $F_{\text{opt}} = 0.9899 \pm 0.0051$, $\Delta\Sigma = 0.9506 \pm 0.0234$ and the reduction of the work $dW = 0.9042 \pm 0.1396$. Inset presents the Bloch sphere illustration for state evolution, corresponding to free evolution (red dots) and RL-controlled evolution (blue dots). **d, e** Comparison of robustness against the systematic imperfection for the RL-controlled fidelity F_{opt} , the reduced entropy production $\Delta\Sigma$ and the reduction of work dW at the final time $\tau = 26 \mu\text{s}$, where **d** is for the Rabi frequency deviation $\delta\Omega/\Omega_0$ of the laser irradiation and **e** for the resonance frequency deviation Δ/Ω_0 of the qubit. F_{opt} can be larger than 95% when the deviation is 20%. The error bars indicating the statistical standard deviation of the experimental data are obtained by 10,000 measurements for each data point. Other parameters: Rabi frequency $\Omega_0/2\pi = 20 \text{ kHz}$, time for every step $\delta\tau = 0.25/\Omega_0$, and inverse temperature $\beta\Omega_0 = 3$.

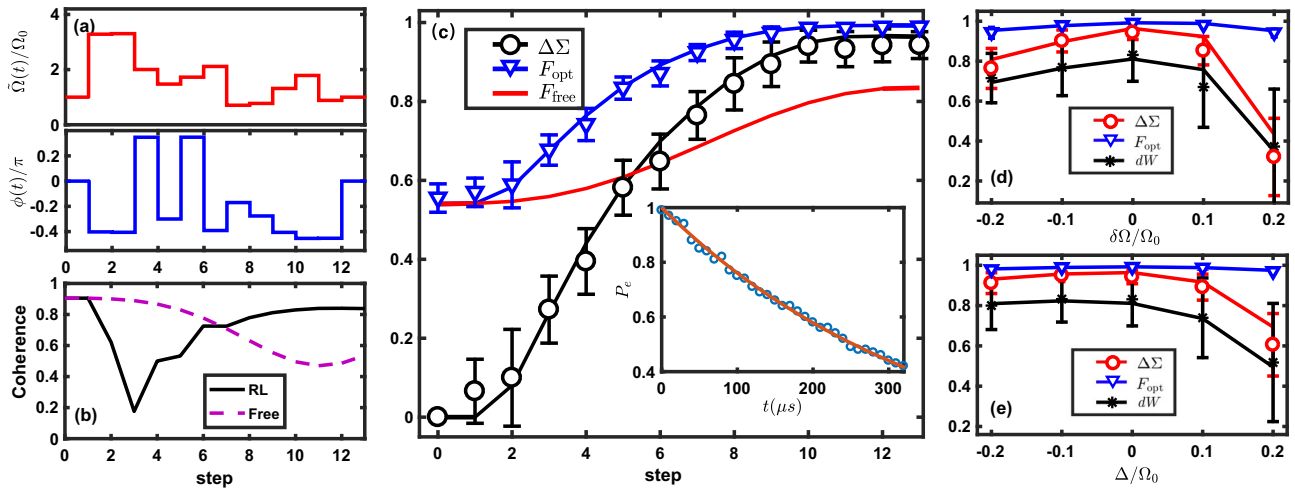


Fig. 4 Single-qubit nonequilibrium thermodynamic transformations in the open system. **a** Designed pulses for the effective Rabi frequency $\tilde{\Omega}$ and phase ϕ , segmented by 13 steps, due to reinforcement learning (RL) control. **b** Dynamics of coherence with or without RL-control. **c** Time evolutions of fidelity F and entropy production reduction $\Delta\Sigma$, where dots are experimental results and lines represent theoretical simulation. The blue and red lines denote the RL-engineered evolution and free evolution, respectively, and the black line is plotted for variation of $\Delta\Sigma$. After the last step, $F_{\text{opt}} = 0.9886 \pm 0.0067$, $\Delta\Sigma = 0.9420 \pm 0.0341$ and the reduction of the work $dW = 0.8360 \pm 0.1321$. Inset presents time evolution of the population P_e due to decay. **d, e** Comparison of robustness against the deviations of Rabi frequency $\delta\Omega/\Omega_0$ and resonance frequency Δ/Ω_0 for F_{opt} , $\Delta\Sigma$ and dW in the presence of decay, where F_{opt} can be larger than 95% when the deviation is 20%. The error bars indicating the statistical standard deviation of the experimental data are obtained by 10,000 measurements for each data point. Other parameters: Rabi frequency $\Omega_0/2\pi = 20\text{kHz}$, decay $\gamma_{\text{eff}} = 0.0216\Omega_0$, time for every step $\delta\tau = 0.25/\Omega_0$, and inverse temperature $\beta\Omega_0 = 3$.

$f_{\text{opt}}(0) = f_{\text{opt}}(\tau) = 0$, $\tilde{\Omega} \in [-\Omega_0, \Omega_0]$, $\theta \in [0, \pi]$, and $t \in [0, \tau]$, we prepare the system initially to be in $\rho_{s_2}(0) = \rho_2(0) \otimes |0\rangle_r\langle 0|$, where $\rho_2(0) = (e^{\beta\Omega_0/2}|g\rangle\langle g| + e^{-\beta\Omega_0/2}|e\rangle\langle e|)/(e^{\beta\Omega_0/2} + e^{-\beta\Omega_0/2})$ and the vibrational degree of freedom is in the ground state. But in this case, we cannot prepare the required initial state making use of the RL engineering due to the fact that no separate Hamiltonian exists for the qubit. As such, we consider another

way for the initial state preparation, that is, $\rho_2(0)$ is prepared from the corresponding superposition state after waiting for a time longer than the dephasing time of the qubit¹⁰ (also see Supplementary Note 3 for more details).

The target state is set as the equilibrium thermal state regarding H_{s_2} , i.e., $\rho_{s_2}^{\text{eq}}(\tau) = \exp(-\beta H_{s_2})/Z_{s_2}$ with the partition function $Z_{s_2} = \text{tr}[\exp(-\beta H_{s_2})]$. Similar to the single-qubit case, we first design the optimized pulses by the DDPG algorithm, finding that

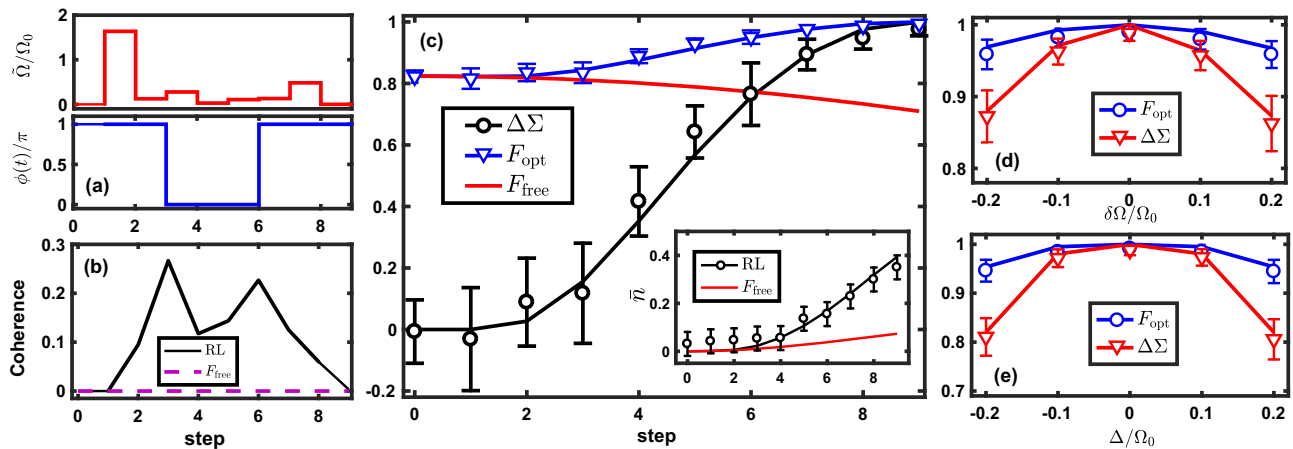


Fig. 5 Nonequilibrium thermodynamic transformations of the Jaynes-Cummings interaction case. **a** Designed pulses of the Rabi frequency and the laser phase due to reinforcement learning (RL) control. **b** Dynamics of coherence with and without RL-control. **c** Time evolution of fidelity F and entropy production reduction $\Delta\Sigma$, where dots are experimental results and lines represent theoretical simulation. Inset presents time evolution of the mean phonon number $\bar{n} = \langle a^\dagger a \rangle$ with and without RL control pulses. **d, e** Comparison of robustness against the deviations of Rabi frequency and resonance frequency for F_{opt} and $\Delta\Sigma$. The error bars indicating the statistical standard deviation of the experimental data are obtained by 10,000 measurements for each data point. Other parameters: Rabi frequency $\Omega_0/2\pi = 50\text{kHz}$, time for every step $\delta\tau = 0.25/\Omega_0$, and inverse temperature $\beta\Omega_0 = 2$.

the target state could be reached by 9 steps. In this case, however, we need two beams of 729-nm laser, one achieving Eq. (5), which is in parallel with the z -axis as the single-qubit case, and the other for the RL control, which radiates with an angle of 60 degrees to the z -axis. The system state, after 9 pulses applied, evolves to a thermal nonequilibrium state $\rho_2^{eq}(\tau) = \text{tr}_r[\rho_{s_2}^{eq}(\tau)]$.

Figure 5a illustrates our experimental observation of the high-fidelity evolution under RL control, indicating that the RL-pulses-induced nonequilibrium evolution is very close to the target state $\rho_2^{eq}(\tau)$, much higher than the free evolution. Figure 5b gives the coherence dynamical with or without RL-control. Meanwhile, our observation also reveals that the relative entropy assisted by RL pulses is much reduced with respect to the free evolution, as shown in Fig. 5c. Similar to above observations, we have also witnessed at finite time, e.g., $\tau = 57\ \mu\text{s}$, the robustness of the RL-engineered evolution to the imperfections, as shown in Fig. 5d, e. Of particular interest in this case is the possibility to monitor the variation of the quantum bath during the nonequilibrium thermodynamic evolution. Since the spin state and the vibrational state of the ion are correlated by the JC model without dissipation to outside, although our target state is only set to be relevant to the qubit, the state of the quantum bath is also targeted. We have experimentally measured the mean phonon number \bar{n} at each step, which demonstrates the dissipation of the qubit to the quantum bath. We see from the inset of Fig. 5b the good agreement between the experimental observation and the expected values in the case of RL engineering.

Conclusions

In summary, with high-precision operations on the single ultra-cold ion, our experimental observations have provided credible evidences for the outstanding advantages of RL strategy in optimizing quantum state control, which would be useful for exploiting microscopic thermal machines and quantum information processing based on far-from-equilibrium quantum processes.

In particular, we have explored the engineering of quantum systems subjected to dissipation from classical or quantum baths. We have further demonstrated the robustness of characteristic parameters under RL control, even in the presence of operational

imperfections. These observations underscore the practical applicability of RL control in executing quantum tasks with higher fidelity and reduced consumption of entropy production and work. Expanding on our current research, we will delve deeper into the design of efficient single-atom quantum heat engines⁴² utilizing RL implementation.

Additionally, we have taken note of recent experimental endeavors that leverage the advantages of RL strategies in addressing many-body problems. For instance, one notable achievement includes the realization of improved number squeezing in the balanced three-mode Dicke states of 10^{487} Rb atoms²⁹. Considering the theoretical results demonstrating minimized relative entropy in a RL-controlled nonequilibrium thermodynamic process for a two-qubit system¹¹, it is highly anticipated that future experiments will showcase RL-engineered optimization in multi-qubit nonequilibrium thermodynamics. This may involve exploring larger Hilbert spaces, investigating multi-qubit systems, and integrating RL with other control techniques.

Methods

Here we describe the RL method that is used in our experiment. The standard RL system contains two major entities: agent and environment, connected by the channels: state space \mathcal{S} , action \mathcal{A} , and reward \mathcal{R} . The agent and environment interact via a finite Markovian decision process, which divides the total training time τ into n steps with fixed interval $\delta\tau = \tau/n$. At each time step t , the agent receives a state $s_t \in \mathcal{S}$, and then takes an action $a_t \in \mathcal{A}$, which results in a new state $s_{t+1} \in \mathcal{S}$ and finally receives a reward $r_t \in \mathcal{R}$. Therefore, this Markovian process is described as a sequence of $(s_0, a_0, r_0, s_1, \dots, s_b, a_b, r_b, s_{t+1}, \dots, s_{n-1}, a_{n-1}, r_{n-1}, s_n)$ with n the number of the steps in an episode.

The DDPG is a model-free off-policy actor-critic algorithm that can learn policies in spaces with high-dimension and by continuous actions. Before going to the details of the DDPG algorithm, we first introduce the definitions of the state \mathcal{S} , the action \mathcal{A} , and the reward function \mathcal{R} for our case considered.

\mathcal{S} : We employ the density operator ρ_t of the system as the input state of the agent, which contains the complete information of the quantum thermodynamical evolution process and can make the training process quicker.

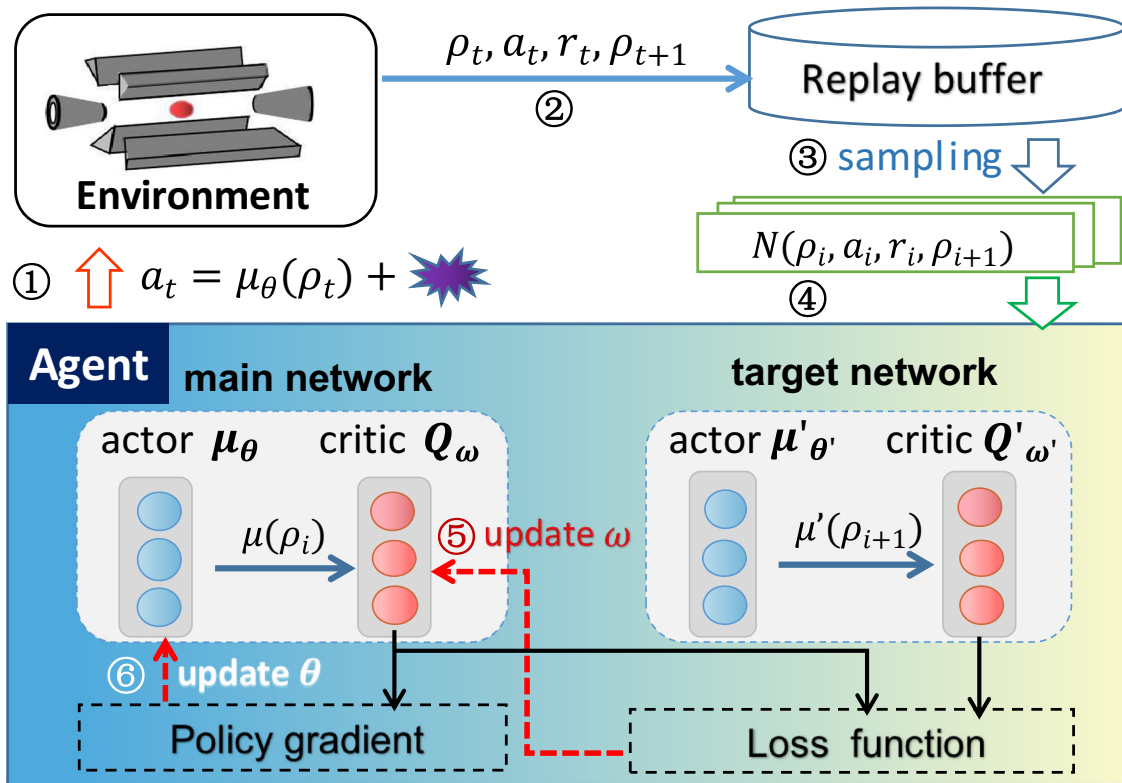


Fig. 6 Schematic for reinforcement learning (RL) with deep deterministic policy gradient (DDPG) algorithm. This algorithm includes a replay buffer and the target network with actor μ'_{θ} and critic Q'_{ω} in addition to the main network involving actor μ_{θ} and critic Q_{ω} . The learning agent (i.e., the network) acts on the environment (i.e., the trapped-ion qubit and/or the vibrational degree of freedom of the ion) and updates the actions based on the obtained feedback from the environment.

\mathcal{A} : The action space is a continuously controllable variable by taking $a_t = f_{\text{opt}}(t)$ in the interval $[t, t+1)$ and $f_{\text{opt}}(t) \in [-1.5\Omega_0, 1.5\Omega_0]$. Based on the optimization Hamiltonian $H_{\text{opt}}(t) = -f_{\text{opt}}(t)M$, we define the unitary operator $U_t = \exp[-i(H_s + H_{\text{opt}}(t))\delta\tau]$ with H_s the free evolution Hamiltonian. After the action of a_t , we obtain a new quantum state $\rho_{t+1} = U^\dagger(t, 0)\rho_t U(t, 0)$ with $U(t, 0) = U_t \dots U_1 U_0$. Here, the operator M is chosen as $M = \sigma_y$ and $M = \sigma_x$ for the single-qubit rotation and the Jaynes-Cummings interaction, respectively. Experimentally, $f_{\text{opt}}(t)$ is implemented to control the amplitude and phase of the driving laser.

\mathcal{R} : In our study, we select fidelity F as the objective function, with $F(\tau) = 1$ indicating the complete evolution of the system to the target state at the end time. To address the issue of sparse rewards, we write the objective function as a summation form, $R = \sum_j r_j$. At each time step t_j , the agent receives a reward ($r_j = F(t_j)$) that represents the instantaneous increase in fidelity. This dense reward scheme accelerates the training process and enhances stability. Practically, in our numerical treatment, we may modify the above-defined reward as $r_j \rightarrow |\log_{10}(1 - F(t_j))|$ to improve the learning efficiency when the fidelity approaches 1. In this context, $(1 - F(t_j))$ represents the fidelity discrepancy between the quantum state at time t_j and the target state. The mathematical meaning of the arrow indicates an adjustment in the weighting for values approaching $F=1$ through the logarithmic transformation. The modified reward is of the same motivation as the above defined reward ($r_j = F(t_j)$).

To implement the DDPG algorithm⁴⁰, we apply TensorFlow framework to build the neural network (i.e., the agent) for the deep learning. For the single trapped-ion system under our consideration, we apply a simple neural network to parameterize all the four

sub-networks: μ_θ, Q_ω (as the main network), and $\mu'_{\theta}, Q'_{\omega}$ (as the target network). As sketched in Fig. 6, the networks in DDPG are trained as what follows. Firstly, the action $a_t = \mu_\theta(\rho_t) + \mathcal{N}_t$, with Gaussian noise \mathcal{N}_t , is generated randomly via the main actor network, and then the experience $(\rho_t, a_t, r_t, \rho_{t+1})$ is saved in the replay buffer, before sampling a random minibatch of N transitions $(\rho_i, a_i, r_i, \rho_{i+1})$ used to update the main network. During the training, the actor network μ_θ is updated by the policy gradient descent as $\nabla_{\theta} \mu_{\theta} |_{\rho_i}$ to maximize the value $Q_\omega(\rho_i, a = \mu_\theta(\rho_i))$ predicted by the critic network Q_ω . Besides, the target critic network Q'_{ω} also predicts the value $Q'_{\omega}(\rho_{i+1}, a' = \mu'_{\theta}(\rho_{i+1}))$, and thus minimizes the loss function $L(Q_\omega, Q'_{\omega})$ to update the critic in the main network. Meanwhile, the target network is updated much slowly than the main network, since the former only absorbs a small weight from the latter. We set the learning rates as $\alpha_a = 0.001$ (for actor) and $\alpha_c = 0.002$ (for critic) in the main network. The quantum dynamics in our training environment is numerically simulated by QuTip quantum Toolbox with 4.6.2 version (<https://qutip.org>).

Data availability

The datasets generated during this study are available from the corresponding author upon reasonable request.

Code availability

Codes are available upon request from the corresponding authors.

Received: 27 April 2023; Accepted: 28 September 2023;

Published online: 07 October 2023

References

- Král, P., Thanopoulos, I. & Shapiro, M. Coherently controlled adiabatic passage. *Rev. Mod. Phys.* **79**, 53 (2007).
- Parrondo, J. M. R., Horowitz, J. M. & Sagawa, T. Thermodynamics of information. *Nat. Phys.* **11**, 131 (2015).
- Jarzynski, C. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **78**, 2690 (1997).
- Jarzynski, C. Equilibrium free-energy differences from nonequilibrium measurements: a master-equation approach. *Phys. Rev. E* **56**, 5018 (1997).
- Crooks, G. E. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E* **60**, 2721 (1999).
- Collin, D. et al. Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature* **437**, 231 (2005).
- Deffner, S. & Lutz, E. Generalized clausius inequality for nonequilibrium quantum processes. *Phys. Rev. Lett.* **105**, 170402 (2010).
- Shiraishi, N. & Saito, K. Information-theoretical bound of the irreversibility in thermal relaxation processes. *Phys. Rev. Lett.* **123**, 110603 (2019).
- Vu, T. V. & Hasegawa, Y. Geometrical bounds of the irreversibility in Markovian systems. *Phys. Rev. Lett.* **126**, 010601 (2021).
- Zhang, J. W. et al. Single-atom verification of the information-theoretical bound of irreversibility at the quantum level. *Phys. Rev. Res.* **2**, 033082 (2020).
- Sgroi, P., Palma, G. M. & Paternostro, M. Reinforcement learning approach to nonequilibrium quantum thermodynamics. *Phys. Rev. Lett.* **126**, 020601 (2021).
- Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. (The MIT Press Cambridge, 2015).
- Henson, B. M. et al. Approaching the adiabatic timescale with machine learning. *Proc. Natl Acad. Sci. USA* **115**, 13216 (2018).
- Zhang, X.-M., Wei, Z., Asad, R., Yang, X.-C. & Wang, X. When does reinforcement learning stand out in quantum control? A comparative study on state preparation. *npj Quant. Inf.* **5**, 1 (2019).
- Carleo, G. et al. Machine learning and the physical sciences. *Rev. Mod. Phys.* **91**, 045002 (2019).
- Krenn, M., Erhard, M. & Zeilinger, A. Computer-inspired quantum experiments. *Nat. Rev. Phys.* **2**, 649 (2020).
- Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nat. Phys.* **13**, 431 (2017).
- Yoshioka, N. & Hamazaki, R. Constructing neural stationary states for open quantum many-body systems. *Phys. Rev. B* **99**, 214306 (2019).
- Melnikov, A. A. et al. Active learning machine learns to create new quantum experiments. *Proc. Natl Acad. Sci. USA* **115**, 1221 (2018).
- Porotti, R., Tamascelli, D., Restelli, M. & Prati, E. Coherent transport of quantum states by deep reinforcement learning. *Commun. Phys.* **2**, 61 (2019).
- Bukov, M. et al. Reinforcement learning in different phases of quantum control. *Phys. Rev. X* **8**, 031086 (2018).
- Giordani, T. et al. Experimental engineering of arbitrary qudit states with discrete-time quantum walks. *Phys. Rev. Lett.* **122**, 020503 (2019).
- Giordani, T. et al. Machine learning-based classification of vector vortex beams. *Phys. Rev. Lett.* **124**, 160401 (2020).
- Innocenti, L., Banchi, L., Ferraro, A., Bose, S. & Paternostro, M. Supervised learning of time-independent Hamiltonians for gate design. *New J. Phys.* **22**, 065001 (2020).
- Harney, C., Pirandola, S., Ferraro, A. & Paternostro, M. Entanglement classification via neural network quantum states. *New J. Phys.* **22**, 045001 (2018).
- Fösel, T., Tighineanu, P., Weiss, T. & Marquardt, F. Reinforcement learning with neural networks for quantum feedback. *Phys. Rev. X* **8**, 031084 (2018).
- Banchi, L., Grant, E., Rocchetto, A. & Severini, S. Modelling non-markovian quantum processes with recurrent neural networks. *New J. Phys.* **20**, 123030 (2018).
- Ai, M.-Z. et al. Experimentally realizing efficient quantum control with reinforcement learning. *Sci. China-Phys. Mech. Astron.* **65**, 250312 (2022).
- Guo, S.-F. et al. Faster state preparation across quantum phase transition assisted by reinforcement learning. *Phys. Rev. Lett.* **126**, 060401 (2021).
- Guéry-Odelin, D. et al. Shortcuts to adiabaticity: concepts, methods, and applications. *Rev. Mod. Phys.* **91**, 045001 (2019).
- Torrontegui, E. et al. *Advances In Atomic, Molecular, And Optical Physics* Vol. **62**. p. 117 (Elsevier, 2013).
- Sjöqvist, E. et al. Non-adiabatic holonomic quantum computation. *New J. Phys.* **14**, 103035 (2012).
- Samuel, J. & Bhandari, R. General setting for berry's phase. *Phys. Rev. Lett.* **60**, 2339 (1988).
- Friedenauer, A. & Sjöqvist, E. Noncyclic geometric quantum computation. *Phys. Rev. A* **67**, 024303 (2003).
- Lv, Q.-X. et al. Noncyclic geometric quantum computation with shortcut to adiabaticity. *Phys. Rev. A* **101**, 022330 (2020).
- Liu, B.-J., Su, S.-L. & Yung, M.-H. Nonadiabatic noncyclic geometric quantum computation in Rydberg atoms. *Phys. Rev. Res.* **2**, 043130 (2020).
- Dunjko, V., Taylor, J. M. & Briegel, H. J. Quantum-enhanced machine learning. *Phys. Rev. Lett.* **117**, 130501 (2016).
- Sriarunothai, T. et al. Speeding-up the decision making of a learning agent using an ion trap quantum processor. *Quant. Sci. Technol.* **4**, 015014 (2019).
- Saggio, V. et al. Experimental quantum speed-up in reinforcement learning agents. *Nature (London)* **591**, 229 (2021).
- Lillicrap, T. P. et al. Continuous control with deep reinforcement learning[C]. International Conference on Learning Representations (ICLR), 2016.
- Zhou, F. et al. Verifying Heisenberg's error-disturbance relation using a single trapped ion. *Sci. Adv.* **2**, e1600578 (2016).
- Zhang, J.-W. et al. Dynamical control of quantum heat engines using exceptional points. *Nat. Commun.* **13**, 6225 (2022).
- Zhang, J. W. et al. Single-atom verification of the noise-resilient and fast characteristics of universal nonadiabatic noncyclic geometric quantum gates. *Phys. Rev. Lett.* **127**, 030502 (2021).

Acknowledgements

We acknowledge thankfully the technical support of ultrastable optical cavity and frequency stabilization from the group of Professor Qunfeng Chen. This work was supported by the Key Research & Development Project of Guangdong Province under Grant No. 2020B0303300001, by the National Natural Science Foundation of China under Grant Nos. U21A20434, 12074346, 12074390, 12064004, 12275077, 11835011, 11804375, 11804308, 92265107, and 12304315 by Postdoctoral Science Foundation of China under Grant Nos. 2022M710881 and 2023T160144, by Key Lab of Guangzhou for Quantum Precision Measurement under Grant No. 202201000010, and by Natural Science Foundation of Hunan Province under Grant No. 2022JJ30277.

Author contributions

J.Z. and J.L. performed the experiment and processed the data. They contributed equally to this work. Q.S.T. designed the RL pulses and wrote part of the paper. T.X. provides technical support for some key operations. F.Z. supervised the experiment. M.F. proposed the idea and edited the final version of the paper. J.B., W.Y., B.W., G.D., W.D., L.C., L.Y., and S.S. joined the discussion and made comments.

Competing interests

The authors have no competing interest.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42005-023-01408-5>.

Correspondence and requests for materials should be addressed to Qing-Shou Tan, Taiping Xiong, Fei Zhou or Mang Feng.

Peer review information : *Communications Physics* thanks Shuoming An and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



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

© The Author(s) 2023