COMMUNICATIONS PHYSICS

ARTICLE

https://doi.org/10.1038/s42005-019-0169-x OPEN

Coherent transport of quantum states by deep reinforcement learning

Riccardo Porotti[1,2,5], Dario Tamascelli[3], Marcello Restelli[4] & Enrico Prati[1]

Some problems in physics can be handled only after a suitable ansatz solution has been guessed, proving to be resilient to generalization. The coherent transport of a quantum state by adiabatic passage through an array of semiconductor quantum dots is an excellent example of such a problem, where it is necessary to introduce a so-called counterintuitive control sequence. Instead, the deep reinforcement learning (DRL) technique has proven to be able to solve very complex sequential decision-making problems, despite a lack of prior knowledge. We show that DRL discovers a control sequence that outperforms the counterintuitive control sequence. DRL can even discover novel strategies when realistic disturbances affect an ideal system, such as detuning or when dephasing or losses are added to the master equation. DRL is effective in controlling the dynamics of quantum states and, more generally, whenever an ansatz solution is unknown or insufficient to effectively treat the problem.

[1] Istituto di Fotonica e Nanotecnologie, Consiglio Nazionale delle Ricerche, Piazza Leonardo da Vinci 32, I-20133 Milano, Italy. [2] Dipartimento di Fisica "Aldo Pontremoli", Università degli Studi di Milano, via Celoria 16, I-20133 Milano, Italy. [3] Quantum Technology Lab, Dipartimento di Fisica "Aldo Pontremoli", Università degli Studi di Milano, via Celoria 16, I-20133 Milano, Italy. [4] Politecnico di Milano, Piazza Leonardo da Vinci 32, I-20133 Milano, Italy. [5] Present address: Max Planck Institute for the Science of Light, 91058 Erlangen, Germany. Correspondence and requests for materials should be addressed to E.P. (email: enrico.prati@cnr.it)

Some problems in physics are solved as a result of the discovery of an ansatz solution, namely a successful test guess, but unfortunately there is no general method to generate one. Recently, machine learning has increasingly proved to be a viable tool for modeling hidden features and effective rules in complex systems. Among the classes of machine learning algorithms, deep reinforcement learning (DRL)[1] is providing some of the most spectacular results due to its ability to identify strategies for achieving a goal in a complex space of solutions without prior knowledge of the system[2–7]. Contrary to supervised learning, which has already been applied to quantum systems, such as in the determination of high-fidelity gates and the optimization of quantum memories by dynamic decoupling[8], DRL has only very recently been proposed for the control of quantum systems[9–16], along with a strictly quantum reinforcement learning implementation[14,17]. To show the power of DRL, we apply DRL to the problem of coherent transport by adiabatic passage (CTAP) where an electron (encoding the quantum state) is transferred through an array of quantum dots. The ansatz solution for CTAP is notoriously called counterintuitive because of its nonobvious barrier control gate pulse sequence. During coherent adiabatic passage, an electron spends ideally no time in the central quantum dot, because of the simultaneous modulation of the coupling between the dots that suitably drives the trajectory through the Hilbert space[18–22]. The system moves from an initial equilibrium condition to a different one, namely the one where only the last dot of the array is populated. By exploiting such an ansatz solution, consisting of applying the barrier control gates between the dots in a “reversed order” with respect to what intuition would naturally suggest, the process displays truly quantum mechanical behavior, provided that the array consists of an odd number of dots. Researchers have already explored silicon-based quantum information processing architectures[23–25], including the coherent transport of multiple-spin qubits into double quantum dots by adiabatic passage[22], heuristic search methods, such as genetic algorithms, to find a universal set of quantum logic gates[26,27], and the application of DRL to classical systems[28–30].

Here, we demonstrate that DRL implemented in a compact neural network can, first of all, autonomously discover an analog of the counterintuitive gate pulse sequence without any prior knowledge, therefore finding a control path in a problem whose solution is far from the equilibrium of the initial conditions. More importantly, this method can outperform the previously introduced analytical solutions in terms of processing speed and when the system deviates from ideal conditions, which are here represented by the imperfect tuning of the ground states of the quantum dots, dephasing and losses. Under such conditions, no analytical approach exists to the best of our knowledge. In our approach, we exploit trust region policy optimization (TRPO)[31] to handle the CTAP problem. First, we compare the results discovered by the artificial intelligence algorithm with the ansatz solution given in the literature. Next, we apply the method to solve the system when the ground states of the quantum dots are detuned and when the system is perturbed by an interaction with uncontrollable degrees of freedom of the surrounding environment. This latter condition results in dephasing and loss terms in the master equation describing the system, for which there exists no analytical method. Similar to the case of artificial intelligence learning in the classical Atari framework[2], in our approach, the DRL agent interacts with a QuTIP[32] simulation, which plays the role of the RL environment, by implementing the master equation of the system. The RL agent exploits the information retrieved from the feedback in terms of the temporal evolution of the population of the dots. As a further advantage of such an approach, a 2-step temporal Bayesian network (2TBN) analysis can identify which parameters of the system influence the process to a greater extent.

Results

Deep reinforcement learning by actor-critic neural networks. Reinforcement learning (RL) is a set of techniques used to learn behavior in sequential decision-making problems when no prior knowledge about the system dynamics is available or when the control problem is too complex for classical optimal-control algorithms. RL methods can be roughly classified into three main categories: value-based, policy-based, and actor-critic methods[1]. Recently, actor-critic methods have proven to be successful in solving complex continuous control problems[33].

The idea behind actor-critic methods is to use two parametric models (e.g., neural networks) to represent both the policy (actor) and the value function (critic). The actor decides in each state of the system which action to execute, while the critic learns the value (utility) of taking each action in each state (see Supplementary Note 1). Following the critic’s advice, the actor modifies the parameters of its policy to improve the performance. Among the many actor-critic methods available in the literature, we selected the TRPO algorithm[31] (Supplementary Note 2) to find an optimal policy of control pulses. The choice of TRPO is motivated both by its excellent performance on a wide variety of tasks and by the relative simplicity of tuning its hyperparameters[31] (see the Supplementary Note 3).

Coherent spatial adiabatic passage of qubits. CTAP is the solid-state version of a method developed for stimulated Raman adiabatic passage[21,34], which is relevant for instance in quantum information processing architectures that shuttle a qubit from one location to another, where attention is paid to minimize the information loss during transport. In solid-state quantum devices based on either silicon[35] or gallium arsenide[36], the qubit can be encoded, for instance, into spin states of either excess electron(s) or hole(s) in quantum dots[24]. CTAP was originally developed for single-electron states in single-occupied quantum dots, but it can be also extended to more complex spin states, such as hybrid qubits based on spin triplets[22]. If one imagines employing, for instance, an array of dopants in silicon[37–39], a reasonable inter-dopant spacing is of the order of 20 nm, and a hopping time of 100 ps[40] can be achieved. Adiabatic passage requires control pulses with a bandwidth and order of magnitude or two lower than the hopping time, which can be managed by conventional electronics[41].

Steering a quantum system initially far from the final equilibrium. To demonstrate the exploitation of RL, we start by the simplest case of CTAP across a chain of three identical quantum dots. The RL architecture is depicted in Fig. 1a. The simulation of the physical system that supports CTAP consists of an RL environment E that receives as input the updated values of the parameters (in our case the control sequence $\Omega_{i,i+1}$ with $i = 1, 2$ between the adjacent ith and $(i + 1)$th dots) that reflect the action of the control gates as calculated by the agent A according to the policy π. In turn, the RL environment E computes the new RL state (here expressed in terms of the density matrix of the triple quantum dot device, see Supplementary Fig. 1) and provides feedback to agent A. Agent A calculates the next input parameters after evaluating the effects of the previous input according to a reward function r_t, which is expressed in terms of the system state. It is worth noting that the definition of an RL environment (a black box that receives an action as input and outputs a result) is very different from the usual definition of a physical environment, which corresponds to an external quantum system that

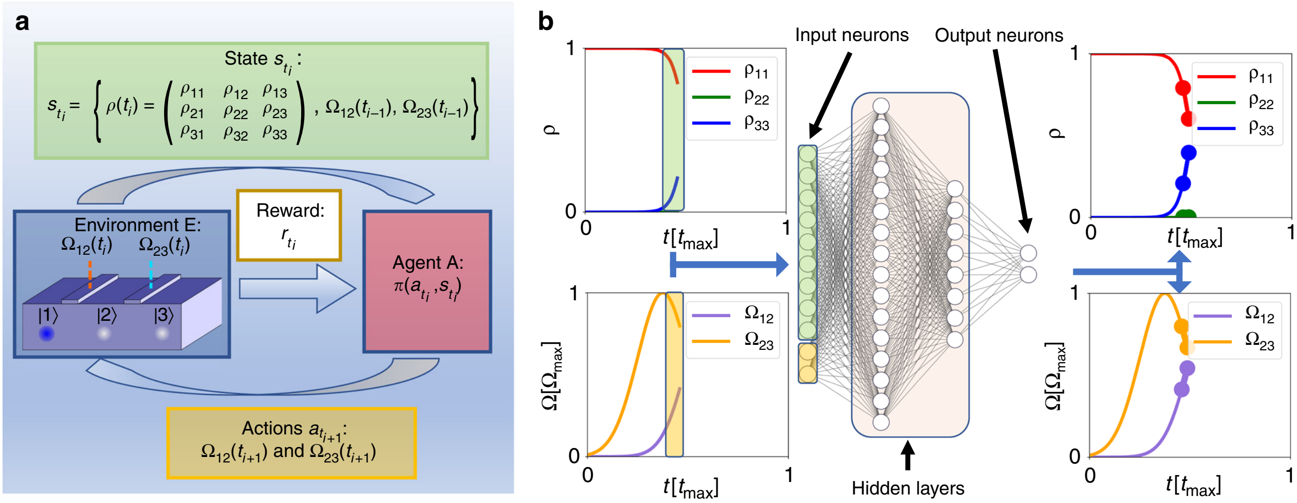
a
State s_{t_i} :
$s_{t_i} = \left\{ \rho(t_i) = \begin{pmatrix} \rho_{11} & \rho_{12} & \rho_{13} \\ \rho_{21} & \rho_{22} & \rho_{23} \\ \rho_{31} & \rho_{32} & \rho_{33} \end{pmatrix}, \Omega_{12}(t_{i-1}), \Omega_{23}(t_{i-1}) \right\}$
Environment E:
$\Omega_{12}(t_i)$ $\Omega_{23}(t_i)$
|1⟩ |2⟩ |3⟩
Reward:
r_{t_i}
Agent A:
$\pi(a_{t_i}, s_{t_i})$
Actions $a_{t_{i+1}}$:
$\Omega_{12}(t_{i+1})$ and $\Omega_{23}(t_{i+1})$
b
Input neurons
Output neurons
Hidden layers
ρ
1
0
$t[t_{\max}]$
ρ_{11}
ρ_{22}
ρ_{33}
$\Omega[\Omega_{\max}]$
Ω_{12}
Ω_{23}

Fig. 1 The deep reinforcement learning architecture to control the coherent transport by adiabatic passage. **a** The deep reinforcement learning (DRL) environment E can be represented by a linear array of quantum dots, with tunneling rates controlled by two gates indicated by Ω_{12} and Ω_{23}. At each time step, the DRL environment can be modeled by a 3×3 density matrix that is employed as an input observation (the state) by agent A. In turn, agent A uses the observation to choose the action in the next time step by following a policy $\pi(a_{t_i}|s_{t_i})$. This action brings E to a new state $\rho_{t_{i+1}} = \rho(t + \Delta t)$. Each action is punished or rewarded with a real-valued reward, indicated by r_t. **b** Agent A is represented with a four-layer neural network, which acts as the policy π. At each time step, the network receives the $3 \times 3 = 9$ real values associated with the absolute values of the elements of the density matrix ρ and the two values of the gate-control pulses Ω_{12} and Ω_{23}, for a total of 11 neurons in the input layer. Then, the agent computes the policy π and outputs the values of the two pulses that will be applied to barrier control gates in the next time step (**b**, lower right). The starting and ending points of the highlighted segments can be connected by different functions of time such as the moving average or a spline function to smooth the step function generated by the discrete time steps. Finally, the physical simulation of the quantum dot array brings the system to a new state by updating the density matrix accordingly and returns r_t to the agent. At $t = t_{\max}$, the system reaches the final step ($k = N$), and the simulation ends

interacts with the quantum system under investigation. More explicitly, the ground state of each quantum dot is tuned with respect to the others by external top metal gates (not shown for simplicity in the sketch of the RL environment E in Fig. 1a), while the coupling between two neighboring quantum dots is in turn controlled by additional barrier control gates. The idealized physical system is prepared so that the ground states of the three quantum dots have the same energy. As the reference energy is arbitrary, without loss of generality we can set $E_1 = E_2 = E_3 = 0$. The Hamiltonian therefore reads:

$$H(t) = \begin{pmatrix} 0 & -\Omega_{12}(t) & 0 \\ -\Omega_{12}(t) & 0 & -\Omega_{23}(t) \\ 0 & -\Omega_{23}(t) & 0 \end{pmatrix}. \quad (1)$$

One of the three eigenstates is referred as a "dark state", and it is expressed as a function of the state of the first and third dots only. The dark state reads:

$$|D_0(t)\rangle = \cos\theta_1(t)|1\rangle - \sin\theta_1(t)|3\rangle, \quad (2)$$

where $\theta_1(t) = \arctan(\Omega_{12}(t)/\Omega_{23}(t))$. A suitable time evolution of the values of the control sequence $\Omega_{12}(t)$ and $\Omega_{23}(t)$ between the dots allows to transform $|D_0(t)\rangle$ from $|1\rangle$ at $t = 0$ to the final target state $|3\rangle$ at $t = t_{\max}$. If the Hamiltonian is prepared in $|D_0\rangle$ at $t = 0$, by choosing $\theta_1(t) = 0$ (so that $|D_0(0)\rangle = |1\rangle$), it will remain in the same eigenstate if the adiabaticity criterion is met; that is, $|E_0 - E_\pm(t)| \gg |\langle \dot{D}_0(t)|D_\pm(t)\rangle|$, where the $|D_\pm(t)\rangle$ eigenstates are explicitly given in the Supplementary Note 4, and $E_\pm(t)$ are the corresponding eigenenergies. The effectiveness of the pulse sequence of the barrier control gate, which is reflected in the control sequence $\Omega_{i,i+1}(t)$, is addressed by taking $F(t) = \rho_{33}(t)$ as the fidelity and maximizing it while trying to keep the population $\rho_{22}(t)$ constant at zero, where ρ_{ij}, $i, j \in \{1, 2, 3\}$ represents an element of the density matrix ρ. The time evolution is governed by the von Neumann equation for the density matrix ρ, namely:

$$\dot{\rho} = -\frac{i}{\hbar}[H, \rho]. \quad (3)$$

Notoriously, a suitable shaping of $\Omega_{12}(t)$ and $\Omega_{23}(t)$ as Gaussian pulses can achieve coherent transport with a high fidelity, if $t_{\max} \gtrsim \frac{10\pi}{\Omega_{\max}}$ (see Supplementary Note 4 and Supplementary Figs. 2–5). The two pulses must be applied in a so-called counterintuitive sequence, meaning that the first gate controlling $\Omega_{12}(t)$ is operated after $\Omega_{23}(t)$. Such pulse sequence drives the occupation of the first dot $\rho_{11}(t_{\max})$ to zero and the occupation of the last dot $\rho_{33}(t)$ to 1, while maintaining empty occupation of the central dot ($\rho_{22}(t) = 0$ $\forall t \in [0, t_{\max}]$)[18]. It is worth mentioning that recently a different ansatz combination of pulse shapes was proposed to speed up this process[42], using the so-called shortcut to adiabaticity protocol[43]. Generally speaking, there is no analytical method to optimize the pulses, so further improvements will be based on still undiscovered ideas. Here is where the power of DRL comes into play.

Deep reinforcement learning setup. The time evolution of the physical system can be mapped into a Markov decision process (MDP). An MDP is defined by a state space $\mathcal{S}$, an action space $\mathcal{A}$, a probability transition $P(s_{t_{i+1}}|s_{t_i}, a_{t_i})$ (where $s_{t_i}, s_{t_{i+1}} \in \mathcal{S}$ and $a_{t_i} \in \mathcal{A}$), and a reward function r that is generally defined in the space $\mathcal{S} \times \mathcal{A}$, but in our case $r : \mathcal{S} \to \mathbb{R}$. We define the discretized time step $t_i = i\Delta t$, where $i = \{0, 1, \ldots N\}$ and $\Delta t = \frac{t_{\max}}{N}$, such that $t_N = t_{\max}$. It is noteworthy to clarify that, albeit the name is the same, the definition of the RL state s_{t_i} differs from that of a quantum state. In this work, s_{t_i} is defined as:

$$s_{t_i} = \left[|\rho_{nm}(t_i)| \text{ for } m, n \in \{1, 2, 3\}, \Omega_{12}(t_{i-1}), \Omega_{23}(t_{i-1}) \right] \quad (4)$$

such that the state at each time step contains the instantaneous absolute values of the nine entries of the density matrix and both

Ω_{ij} values evaluated at the previous time step. Instead, a_{t_i} is defined as:

$$a_{t_i} = [\Omega_{12}(t_i), \Omega_{23}(t_i)] \tag{5}$$

such that $a_{t_i} \in [0, \Omega_{\max}] \times [0, \Omega_{\max}]$. Due to the robust definition of the reward function that allows the agent to judge its performance, the neural network evolves in order to obtain the best temporal evolution of the coupling parameters $\Omega_{i,i+1}$, ensuring that the electron population reaches the target state over time (Fig. 1b). The definition of the best reward function is certainly the most delicate choice in the whole model. The key features of the reward can be summarized by two properties: its generality and its expression as a function of the desired final state. Generality means that the reward should not contain specific information on the characteristics of the two pulses. Regarding the dependence on the desired final state, in our case the goal is to maximize ρ_{33} at the end of the time evolution. The structure of the reward function used in the simulations is:

$$r_{t_i} = (-1 + \rho_{33}(t_i) - \rho_{22}(t_i)) + A(t_i)\delta_{i,N} - B(t_i), \tag{6}$$

where $A(t_i)$ and $B(t_i)$ are the goal-reaching and punishment terms, respectively. We found that the following combinations of $A(t_i)$ and $B(t_i)$ work well:

$$\begin{aligned} A(t_i) &= \begin{cases} 1000 & \rho_{33}(t_i) > \rho_{33}^{\mathrm{th}} \\ 0 & \text{elsewhere} \end{cases}, \\ B(t_i) &= \begin{cases} 100\,\Theta(\rho_{22}(t_i) - 0.05) & \text{Standard CTAP} \\ e^{6\rho_{22}(t_i)} & \text{With detuning, dephasing and loss} \end{cases} \end{aligned} \tag{7}$$

The reward functions used in this research are fully accounted for in the Supplementary Note 5. The sum of the first three terms is nonpositive at each time step, so the agent will try to bring the sum to 0 by minimizing ρ_{22} and maximizing ρ_{33}. We have observed that subtracting $B(t)$ (e.g., punishing the electronic occupation in site 2) and awarding the attainment of the goal with the term $A(t)$ improves the convergence of the learning. Furthermore, in some specific cases, we stop the agent at an intermediate episode if ρ_{33} is greater than an arbitrary threshold ρ_{33}^{th} for a certain number of time steps. This choice can help find fast pulses that achieve high-fidelity quantum information transport at the cost of a higher $\rho_{22}^{\max}$ with respect to the analytic pulses (more details are given in the Supplementary Note 5).

Training of the agent. Figure 2 shows the best results achieved by the agent at various epochs. The agent is free of spanning the whole coupling range at each step. After the training, we smooth the output pulses $\Omega_{i,i+1}$, and we run the simulation using the smoothed values, obtaining a slight improvement from the calculation using the original output values of the RL agent. Smoothing is achieved with a moving average of the pulses (more details in the Supplementary Note 6, Supplementary Fig. 6, and Supplementary Table 1). The three occupations ρ_{ii}, $i \in \{1, 2, 3\}$ shown in Fig. 2 refer to the smoothed pulse sequences. At the very beginning, the untrained agent tries random values with apparently low success, as the occupation oscillates between the first and the last dot over time. It is worth noting that despite this fluctuation, due to the reward function, the agent learns very quickly after only 45 epochs to always keep the occupation of the central dot below 0.5. After ~2000 epochs, the agent learns to stabilize the occupations of the first and last dots, while maintaining an empty occupation of the central dot. After approximately twice the time, a reversal of the populations of the first and last dots happens, even if they do not yet achieve the extremal values. Finally, after approximately four times the elapsed time, the agent achieves high-fidelity CTAP. Despite the stochastic nature of DRL, these epoch numbers remain consistent through different trainings (see the Supplementary Fig. 7). Notice that the pulse sequence is similar to the ansatz Gaussian counterintuitive pulse sequence as the second gate acting on Ω_{23} is operated first, but the shape of the two pulses is different. It is remarkable that the reward function implicitly tries to achieve the result as quickly as possible, resulting in a pulse sequence that is significantly faster than the analytical Gaussian case and comparable to the recent proposal of Ban et al.[42]. The relevant point is that the agent achieves such results irrespective of the actual terms of the Hamiltonian contributing to the master equation. Therefore, DRL can be applied straightforwardly to more complex cases for which there is no generalization of the ansatz solutions found for the ideal case, which we address in the next section.

Deep reinforcement learning to overcome disturbances. We turn now our attention to the behavior of our learning strategy when applied to a nonideal scenario in which typical realistic conditions of semiconductor quantum dots are considered, to be compared with ideal transport (Fig. 3a). In particular, we discuss the results produced by DRL when the array is affected by detuning caused by energy differences in the ground states of the dots, dephasing and losses (Fig. 3b–d). These three effects exist, to different degrees, in any practical attempt to implement CTAP of an electron spin in quantum dots. The first effect is typically due to manufacturing defects[44], while the last two effects emerge from the interaction of the dots with the surrounding environment[45,46] involving charge and spin fluctuations[47,48] and magnetic field noise[49]. Under such disturbances, neither analytical nor ansatz solutions are available. On the other hand, the robustness and generality of the RL approach can be exploited naturally since, from the point of view of the algorithm, it does not differ from the ideal case discussed above. Let us consider a system of $N = 3$ quantum dots with different energies E_i, $i \in \{1, 2, 3\}$. We define $\Delta_{ij} = E_j - E_i$, so that the Hamiltonian (1) can be written, without loss of generality, as

$$H(t) = \begin{pmatrix} 0 & -\Omega_{12}(t) & 0 \\ -\Omega_{12}(t) & \Delta_{12} & -\Omega_{23}(t) \\ 0 & -\Omega_{23}(t) & \Delta_{13} \end{pmatrix}. \tag{8}$$

Figure 3c refers to the particular choice $\Delta_{12} = \Delta_{23} = 0.15\Omega_{\max}$ (a full 2D scan of both Δ_{12} and Δ_{13} and the relative pulses are shown in the Supplementary Figs. 8 and 9, respectively). In this case, DRL finds a solution that induces significantly faster transfer than that obtained with standard counterintuitive Gaussian pulses. Moreover, the latter solutions are not even able to achieve a fidelity comparable to that of the ideal case. Such speed is a typical feature of pulse sequences determined by DRL, satisfying the criterion of adiabaticity (see Supplementary Figs. 10 and 11). Besides the energy detuning between the quantum dots, in a real implementation, the dots interact with the surrounding environment. Since the microscopic details of such an interaction are unknown, its effects are taken into account through an effective master equation. A master equation of the Lindblad type with time-independent rates is adopted to grant sufficient generality while keeping a simple expression. To show the ability of DRL to mitigate disturbances, we consider, in particular, two major environmental effects consisting of decoherence and losses. The first environmental effect corresponds to a randomization of the relative phases of the electron states in the quantum dots, which results in a cancellation of the coherence terms, i.e., the off-diagonal elements of the density matrix in the position basis. The losses, instead, model the erasure of the quantum information carried by the electron/hole moving along the dots. In fact, while the carrier itself cannot be reabsorbed, the quantum information, here encoded as a spin state, can be changed by the interaction

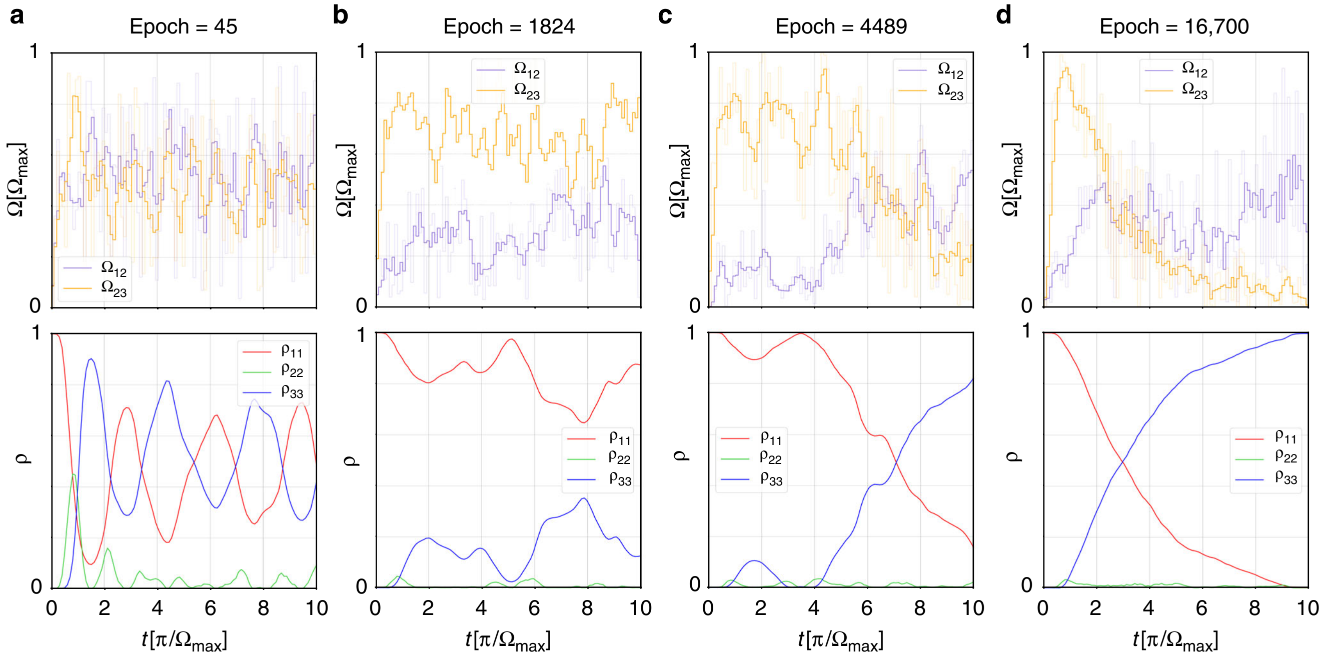

Fig. 2 Learning of the agent for various training epochs. Vertical panels show the pulses Ω and the populations ρ accounted for by the diagonal elements of the density matrix after 45 (**a**), 1824 (**b**), 4489 (**c**), and 16700 (**d**), respectively. Horizontal axis represents time, for which in this case $t_{\max} = 10\frac{\pi}{\Omega_{\max}}$ is set. Vertical axis shows the pulse intensity Ω and the diagonal elements of the density matrix ρ. Each time evolution has been divided into $N = 100$ time steps. In each time step t_i, the neural network can choose the value of both interdot couplings between a pair of neighboring quantum dots, which range from 0 to $\Omega_{\max}$, and they are assumed to be constant during each time step. In the top panel of all four epochs shown, the faded pulses are the outputs of the reinforcement learning (RL) agent, while the highlighted lines are a moving average with a four-sized window. The bottom panels represent the time evolution of the diagonal elements of the density matrix with such averaged pulses applied. We have empirically observed that using a moving average fit achieves better results in terms of minimizing ρ_{22} and maximizing ρ_{33} (in this particular case, the maximum value of ρ_{22} is decreased to 4×10^{-4}, and the maximum value of ρ_{33} is increased to 5×10^{-4}). Three different regimes can be observed: in the first regime, **a** the agent explores randomly. Later, **b** the agent learns how to minimize the population of the second quantum dot. Finally, **c** the agent learns how to achieve a high population in site 3 at the end of the evolution, without populating the second quantum dot (**d**)

with some surrounding spin or by the magnetic field noise. In the presence of dephasing and losses, the system's master equation is of the Lindblad type. Its general form is

$$\dot{\rho} = -\frac{i}{\hbar}[H,\rho] + \sum_n \Gamma_n\left(A_n\rho A_n^\dagger - \frac{1}{2}\left\{A_n^\dagger A_n, \rho\right\}\right), \qquad (9)$$

where A_k are the Lindblad operators, Γ_k are the associated rates, and $\{A, B\} = AB + BA$ denotes the anticommutator.

We first considered each single dot being affected by dephasing and assumed an equal dephasing rate Γ_d for each dot. The master equation can be rewritten as

$$\dot{\rho} = -\frac{i}{\hbar}[H,\rho] + \Gamma_d[\rho - \mathrm{diag}(\rho)]. \qquad (10)$$

Figure 3c shows an example of the pulses determined by DRL and the corresponding dynamics for the choice $\Gamma_d = 0.01\Omega_{\max}$. CTAP is achieved with Gaussian pulses in a time $t \approx 10\pi/\Omega_{\max}$ at the cost of a significant occupation of the central dot (inset). The advantage brought by the DRL agent with respect to Gaussian pulses is manifest: DRL (solid lines) realizes population transfer in approximately half of the time required when analytic pulses (dashed lines) are employed. The occupation of the central dot is less than that achieved by Gaussian pulses. We now consider the inclusion of a loss term in the master equation. The loss refers to the information carried by the electron encoded by the spin. While the loss of an electron is a very rare event in quantum dots, requiring the sudden capture of a slow trap[50–52], the presence of any random magnetic field and spin fluctuation around the quantum dots can modify the state of the transferred spin. Losses are described by the operators $\Gamma_l|0\rangle\langle k|$, $k \in \{1, 2, 3\}$, where Γ_l is the loss rate, modeling the transfer of the amplitude from the i-th dot to an auxiliary vacuum state $|0\rangle$. Figure 3d shows the superior performance of DRL versus analytic Gaussian pulses in terms of the maximum value of ρ_{33} reached during the time evolution. Because of the reward function, DRL minimizes the transfer time so as to minimize the effect of losses. The explicit reward function for each case is given in the Supplementary Table 2. As a further generalization, we applied DRL to the passage through more than one intervening dot, an extension called a straddling CTAP scheme (or SCTAP)[18,53]. To exemplify a generalization to odd $N > 3$, we set $N = 5$ according to the sketch depicted in Fig. 4, and no disturbance was considered for simplicity (see the Supplementary Note 7 for further details on SCTAP). For five-state transfer, like the cases for three states and for any other odd number of quantum dots, there is one state with zero energy (see Supplementary Fig. 12). The realization of the straddling tunneling sequence is achieved by augmenting the original pulse sequence by straddling pulses, which are identical for all intervening tunneling rates Ω_m. The straddling pulse involves the second and third control gates, as shown in Fig. 4a, b. The pulse sequence discovered by DRL is significantly different from the known sequence based on Gaussian pulses. It achieves population transfer in approximately one third of the time with only some occupation in the central dot (see Fig. 4c). It is remarkable that the reward function (see the Supplementary Note 5) achieves such a successful transfer without any

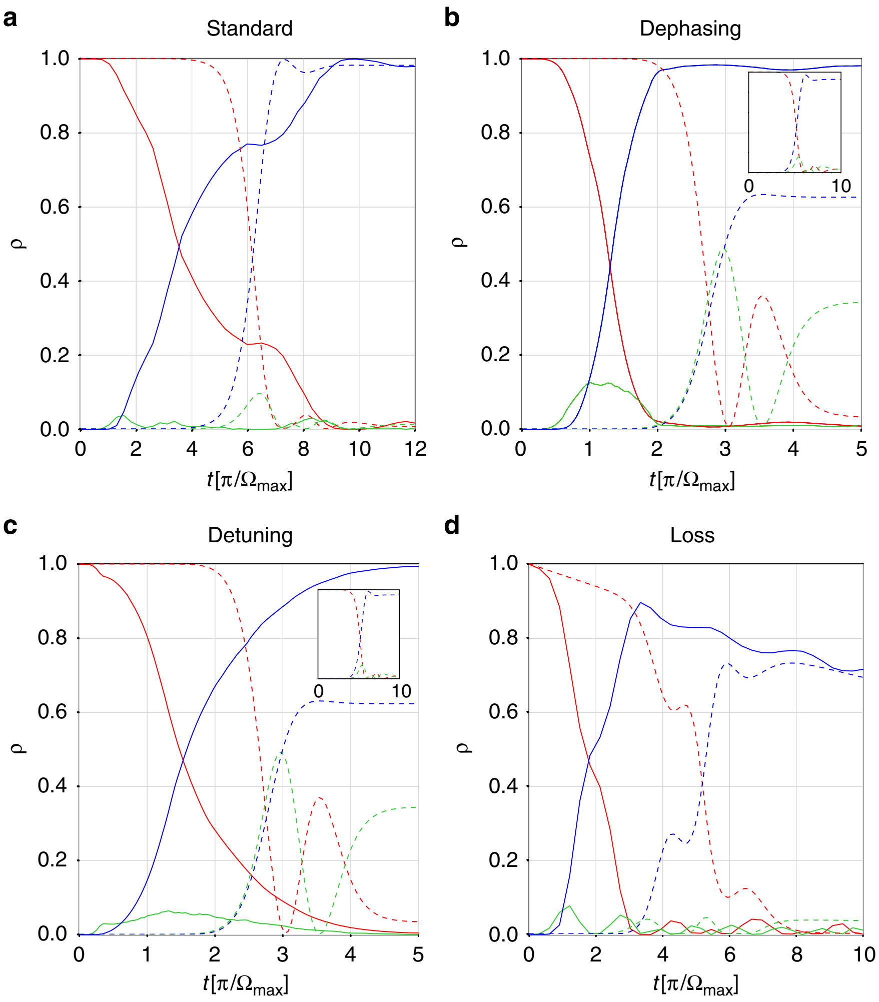

Fig. 3 Comparison of coherent transport by adiabatic passage achieved by Gaussian pulses and deep reinforcement learning algorithm respectively under different kind of disturbances. Gaussian pulses are represented by dotted lines while pulses found by the deep reinforcement learning (DRL) by solid lines. Horizontal axis represents time while vertical axis the value of the diagonal elements of the density matrix. **a** Reference results from a simulation of ideal conditions with identical eigenvalues in the three quantum dots with $t_{max} = 12\frac{\pi}{\Omega_{max}}$. **b** Simulation involving a dephasing term in the Hamiltonian, where $\Gamma_d = 0.01\Omega_{max}$ and $t_{max} = 5\frac{\pi}{\Omega_{max}}$. **c** Simulation for a detuned system, where $\Delta_{12} = \Delta_{23} = 0.15\Omega_{max}$ and $t_{max} = 5\frac{\pi}{\Omega_{max}}$. **d** Simulation affected by loss, which is accounted for by an effective term weighted by $\Gamma_l = 0.1\Omega_{max}$ and $t_{max} = 10\frac{\pi}{\Omega_{max}}$. The insets of **b**, **c** show simulations achieved with Gaussian pulses for $t_{max} = 10\frac{\pi}{\Omega_{max}}$. The setups of the neural networks employed are, respectively, (16, 0), (128, 64), (64, 64), and (128, 64), where the first number in parenthesis represents the number of neurons of the first hidden layer H1 of the network and the second number represents the number of neurons of the second hidden layer H2

requirement for the occupation of the intermediate dots $\rho_{i,i}$ with $(i \in \{2, 3, 4\})$ as it only rewards the occupation of the last dot.

Analysis of the relevant variables within the DRL framework. The advantages of employing DRL with respect to an ansatz solution is further increased by the possibility of determining the factors that are more relevant for the effectiveness of the final solution from the analysis of the neural network. In fact, the employment of the DRL algorithm enables an analysis of the state variables that are actually relevant for solving the control problems, like that discussed above. To select the variables needed to solve an MDP, we follow the approach presented by Peters et al.[54–56]. The idea is that a state variable is useful if it helps to explain either the reward function or the dynamics of the state variables that in turn are necessary to explain the reward function. Otherwise, the state variable is useless, and it can be discarded without affecting the final solution. To represent the dependencies between the variables of an MDP, we use a 2TBN[57], described in Supplementary Note 8. In the graph of Fig. 5, there are three types of nodes: the gray diamond node represents the reward function, the circle nodes represent state variables, and the squares represent action variables. The nodes are arranged on three vertical layers: the first layer on the left includes the variables at time t_i, the second layer includes the state variables at time t_{i+1}, and the third layer includes the node that represents the reward function. If a variable affects the value of another variable, we obtain a direct edge that connects the former to the latter. Figure 5 shows the 2TBN estimated from a dataset of 100,000 samples obtained from the standard case of ideal CTAP. The weights are quantified in the Supplementary Table 3. As expected from Eq. (6), the reward function depends only on the values of the variables ρ_{22} and ρ_{33}. From the 2TBN, it emerges that the dynamics of these two state variables can be fully explained by knowing their values at the previous step, the values of the two action variables Ω_{12} and Ω_{23} and the actions taken in the previous time step (stored in the variables $\Omega_{12}^{'}$ and $\Omega_{23}^{'}$). All

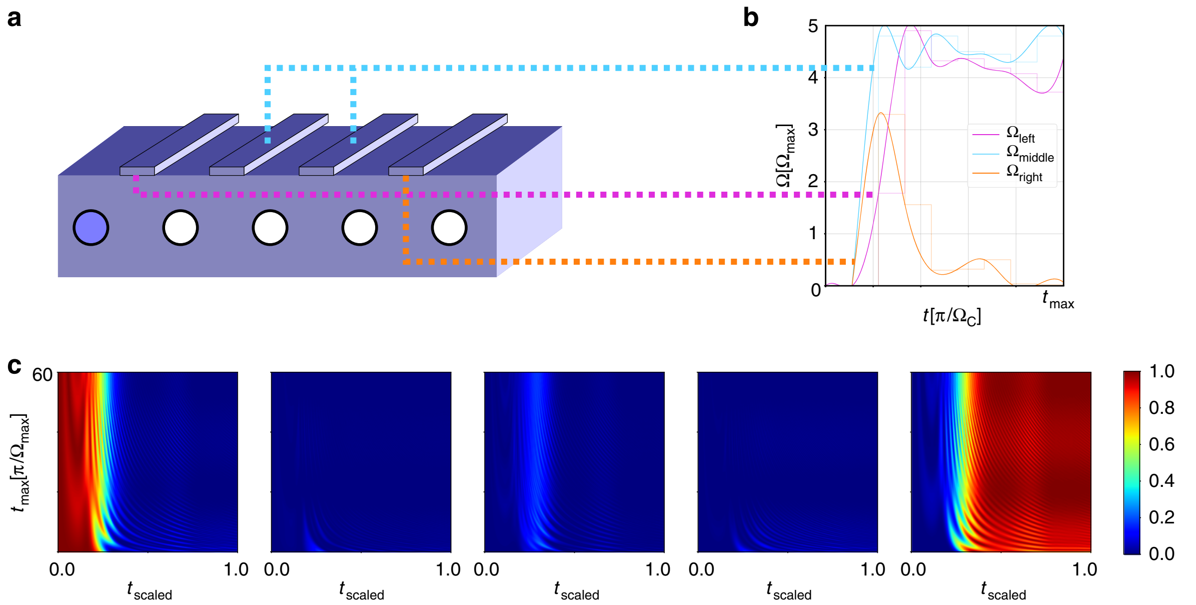

Fig. 4 DRL-controlled Straggling CTAP (SCTAP) (**a**) Schematics of the five quantum dot array. The pair of central gates (**b**) are coupled according to Greentree et al.[18]: Ω_{left} is tuned by the first coupling control gate, Ω_{right} by the last coupling control gate, while Ω_{middle} is identical as from the second and the third coupling control gates. In the example for $t_{max} = 201\frac{\pi}{\Omega_{max}}$. The dashed-thick pulses are discovered by the DRL, while the solid lines are cubic spline interpolations. **c** Population transfer by DRL-controlled SCTAP. On the *x*-axis there is the (rescaled) time in unity of t_{max}, while on the *y*-axis there is t_{max}. During a time evolution, the increase/decrease of the electronic populations shows oscillations (Vitanov et al.[34]), which are reduced for some values of t_{max}, e.g., $t_{max} = 21\frac{\pi}{\Omega_{max}}$. For $t_{max} = 21\frac{\pi}{\Omega_{max}}$, the maximum value of ρ_{33} is $\rho_{33,max} = 0.1946$, while $\rho_{55,max} = 0.99963689$, obtained with the fitted pulses. Notice how ρ_{22}, ρ_{33}, and ρ_{44} are minimized during all the time evolution

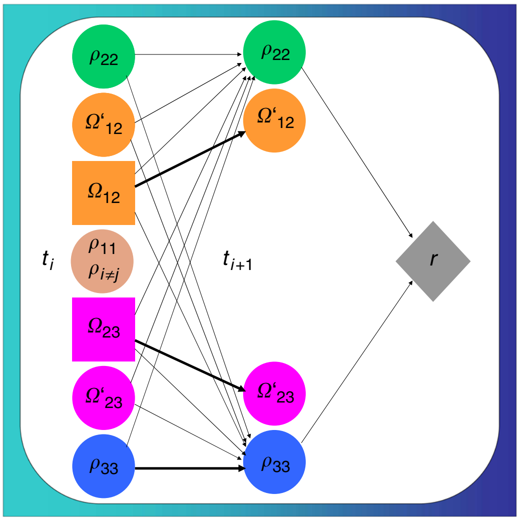

Fig. 5 Two-step temporal Bayesian network (2TBN) for ideal standard coherent transport by adiabatic passage. The gray diamond node is used to represent the reward function, the circles represent the state variables and the squares represent the action variables. The nodes are arranged on three vertical layers: the first layer on the left includes the variables at time t_i, the second layer includes the state variables at time t_{i+1}, and the third layer consists of the node that represents the reward function. The prime index indicates the value of a variable as stored in the previous step. A direct edge between two nodes means that the value of the source node influences the value of the target node. The thicker the line, the greater the influence is. As expected, ρ_{11} (shaded red) has no connections, as it is linearly dependent on ρ_{22} and ρ_{33}

other state variables do not appear and therefore can be ignored for the computation of the optimal control policy. This finding matches the expectation from the constraints of the physical model that coherences are not directly involved in the dynamics and that ρ_{11} is linearly dependent on ρ_{22} and ρ_{33} as the trace of the density matrix is constant. To confirm this finding, the agent was successfully trained in the standard case of ideal CTAP by using only the input values (ρ_{22}, ρ_{33}, Ω_{12}, and Ω_{23}). Furthermore, the size of the hidden layer H1 could be reduced in this case from 64 to 4 with the same convergence of the total reward as a function of the episodes or an even faster convergence by reducing the batch size from 20 to 10 (see the Supplementary Figs. 13 and 14). In the cases of detuning and dephasing, the corresponding 2TBNs are more complex, since the dynamics of ρ_{22} and ρ_{33} are affected also by the values of the other elements of the matrix ρ. This fact is also apparent in the changes carried by the Hamiltonian, as the trace is not constant when losses are present and dephasing has a direct effect on the coherence.

Discussion

Neural networks can discover control sequences for a tunable quantum system whose time evolution starts far from its final equilibrium, without any prior knowledge. Contrary to the employment of special ansatz solutions, we have shown that DRL discovers novel sequences of control operations to achieve a target state, regardless of the possible deviations from the ideal conditions. DRL is achieved irrespective of previously discovered ansatz solutions, and it applies when such solutions are unknown. The control sequence found by DRL can also provide insights into novel analytical shapes for optimal pulses. The use of a pretrained neural network as a starting point to identify the solution of a master equation with similar parameters further reduces the computation time by one order of magnitude (see Supplementary Note 9). Our investigation indicates that a key factor is the appropriate definition of the reward function that deters the system from occupying the central quantum dot and rewards the occupation of the last quantum dot. To apply DRL to a practical quantum system known for its counterintuitive nature, we have studied quantum state transport across arrays of quantum dots, including sources of disturbances such as energy level detuning, dephasing, and loss. In all cases, the solutions found by the agent outperform the known solution in terms of either speed or fidelity or both. Both the pretraining of the network and 2TBN analysis—by reducing the computational effort—contribute to accelerating the learning process. In general, we have shown that neural-network-based DRL provides a general tool for controlling physical systems by circumventing the issue of finding ansatz solutions when neither a straightforward method nor a brute force approach is possible. A comparison with other control strategies of quantum systems[16,58,59] will be the objective of future work. As a concluding remark, we observe that in this work we have modeled the effects of the environment or noisy control fields by master equations in the Lindblad form with time-independent coefficients; by construction, the same DRL strategy can be used to determine the optimal pulse sequence in the presence of non-Markovian noise/effects originating from an interaction of the system with structured environments. In this case, an efficient way to compute the time evolution of the open system, which is required to determine the value of the pulses at each time step, is needed. Depending on the nature (bosonic, fermionic) and features (spectral density) of the environment, different approximate or numerically exact strategies ranging from Bloch-Redfield or time-convolutionless master equations[60] to path integrals[61] or chain-mapping techniques (reaction-coordinate[62] or TEDOPA[63,64]) could be exploited.

Methods

Master equation for the Hamiltonian including pure dephasing. Pure dephasing[65] is the simplest model that accounts for an environmental interaction of an otherwise closed system. The corresponding Lindblad operators are diagonal and have the form:

$$L_n = \sqrt{\gamma_n}|n\rangle\langle n|. \tag{11}$$

In the case of $\gamma_1 = \gamma_2 = \gamma_3 = \Gamma_d$, the master equation becomes:

$$\dot{\rho} = -\frac{i}{\hbar}[H,\rho] + \Gamma_d[\rho - diag(\rho)]. \tag{12}$$

If the Hamiltonian is time-independent, then

$$\frac{d}{dt}\rho_{ii} = 0 \tag{13}$$

for every i, so the electronic population remains unchanged in the presence of pure dephasing. This implies[66] that the energy of the system remains unchanged during the evolution, since it cannot be changed by environment.

Neural network. The neural network consists of a multilayered feed-forward network of (N_i, H_1, [H_2,]N_o) neurons, where $N_i = 11$; it includes the nine elements of the density matrix ρ and the control sequence Ω_{ij}, $N_o = 2$ are the two updated values of the control sequence spanning a continuous range between 0 and an arbitrary value $\Omega_{\max}$ chosen according to experimental considerations, and H_i the number of neurons of the ith hidden layer.

Different setups for the neural network have been employed. Nonetheless, all the neural networks are feedforward and fully connected, with an ReLU activation function for the neurons. For the implementation of the neural networks and their training, Tensorforce[67] is employed. For the RL algorithm adopted for this work, TRPO, the hyperparameters are set as summarized in the Supplementary Table 4. The optimal number of neurons of the hidden layer peaks at ~2^5–2^6 (see the Supplementary Fig. 15).

Software. The QuTIP routine has been run in parallel by using GNU Parallel[68].

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The code and the algorithm used in this study are available from the corresponding author upon reasonable request.

Received: 23 January 2019 Accepted: 22 May 2019
Published online: 20 June 2019

References

1. Sutton, R. S. et al. *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, Massachusetts, 1998).
2. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529 (2015).
3. Melnikov, A. A. et al. Active learning machine learns to create new quantum experiments. *Proc. Natl Acad. Sci. USA* **115**, 1221–1226 (2018).
4. Nautrup, H. P., Delfosse, N., Dunjko, V., Briegel, H. J. & Friis, N. Optimizing quantum error correction codes with reinforcement learning. https://arxiv.org/abs/1812.08451 (2018).
5. Sweke, R., Kesselring, M. S., van Nieuwenburg, E. P. & Eisert, J. Reinforcement learning decoders for fault-tolerant quantum computation. https://arxiv.org/abs/1810.07207 (2018).
6. Colabrese, S., Gustavsson, K., Celani, A. & Biferale, L. Flow navigation by smart microswimmers via reinforcement learning. *Phys. Rev. Lett.* **118**, 158004 (2017).
7. Prati, E. Quantum neuromorphic hardware for quantum artificial intelligence. *J. Phys. Confer. Ser.* **880**, (2017).
8. August, M. & Ni, X. Using recurrent neural networks to optimize dynamical decoupling for quantum memory. *Phys. Rev. A* **95**, 012335 (2017).
9. Dong, D., Chen, C., Tarn, T. J., Pechen, A. & Rabitz, H. Incoherent control of quantum systems with wavefunction-controllable subspaces via quantum reinforcement learning. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **38**, 957–962 (2008).
10. Chen, C., Dong, D., Li, H., Chu, J. & Tarn, T. Fidelity-based probabilistic q-learning for control of quantum systems. *IEEE Trans. Neural Netw. Learn. Syst.* **25**, 920–933 (2014).
11. Fösel, T., Tighineanu, P., Weiss, T. & Marquardt, F. Reinforcement learning with neural networks for quantum feedback. *Phys. Rev. X* **8**, 031084 (2018).
12. August, M. & Hernández-Lobato, J. M. Taking gradients through experiments: LSTMs and memory proximal policy optimization for black-box quantum control. https://arxiv.org/abs/1802.04063 (2018).
13. Niu, M. Y., Boixo, S., Smelyanskiy, V. N. & Neven, H. Universal quantum control through deep reinforcement learning. In *AIAA Scitech 2019 Forum* 0954 (2019).
14. Albarrán-Arriagada, F., Retamal, J. C., Solano, E. & Lamata, L. Measurement-based adaptation protocol with quantum reinforcement learning. *Phys. Rev. A* **98**, 042315 (2018).
15. Zhang, X.-M., Wei, Z., Asad, R., Yang, X.-C. & Wang, X. When reinforcement learning stands out in quantum control? A comparative study on state preparation. https://arxiv.org/abs/1902.02157 (2019).
16. Bukov, M. et al. Reinforcement learning in different phases of quantum control. *Phys. Rev. X* **8**, 031086 (2018).
17. Yu, S. et al. Reconstruction of a photonic qubit state with reinforcement learning. *Adv. Quantum Technol.* **0**, 1800074 (2019).
18. Greentree, A. D., Cole, J. H., Hamilton, A. R. & Hollenberg, L. C. Coherent electronic transfer in quantum dot systems using adiabatic passage. *Phys. Rev. B* **70**, 1–6 (2004).
19. Cole, J., Greentree, A., C. L. Hollenberg, L. & Das Sarma, S. Spatial adiabatic passage in a realistic triple well structure. *Phys. Rev. B* **77**, 235418 (2008).
20. Greentree, A. D. & Koiller, B. Dark-state adiabatic passage with spin-one particles. *Phys. Rev. A* **90**, 012319 (2014).
21. Menchon-Enrich, R. et al. Spatial adiabatic passage: a review of recent progress. *Rep. Prog. Phys.* **79**, 074401 (2016).
22. Ferraro, E., De Michielis, M., Fanciulli, M. & Prati, E. Coherent tunneling by adiabatic passage of an exchange-only spin qubit in a double quantum dot chain. *Phys. Rev. B* **91**, 075435 (2015).
23. Rotta, D., De Michielis, M., Ferraro, E., Fanciulli, M. & Prati, E. Maximum density of quantum information in a scalable cmos implementation of the hybrid qubit architecture. *Quantum Inf. Process.* **15**, 2253–2274 (2016).
24. Rotta, D., Sebastiano, F., Charbon, E. & Prati, E. Quantum information density scaling and qubit operation time constraints of CMOS silicon-based quantum computer architectures. *npj Quantum Inf.* **3**, 26 (2017).

25. Prati, E., Rotta, D., Sebastiano, F. & Charbon, E. From the quantum moore's law toward silicon based universal quantum computing. In *2017 IEEE International Conference on Rebooting Computing (ICRC)* 1–4 (IEEE, 2017).
26. Ferraro, E., De Michielis, M., Mazzeo, G., Fanciulli, M. & Prati, E. Effective hamiltonian for the hybrid double quantum dot qubit. *Quantum Inf. Process.* **13**, 1155–1173 (2014).
27. Michielis, M. D., Ferraro, E., Fanciulli, M. & Prati, E. Universal set of quantum gates for double-dot exchange-only spin qubits with intradot coupling. *J. Phys. A Math. Theor.* **48**, 065304 (2015).
28. Bonarini, A., Caccia, C., Lazaric, A. & Restelli, M. Batch reinforcement learning for controlling a mobile wheeled pendulum robot. In *IFIP International Conference on Artificial Intelligence in Theory and Practice* 151–160 (Springer, 2008).
29. Tognetti, S., Savaresi, S. M., Spelta, C. & Restelli, M. Batch reinforcement learning for semi-active suspension control. In *Control Applications, (CCA) & Intelligent Control, (ISIC), 2009 IEEE* 582–587 (IEEE, 2009).
30. Castelletti, A., Pianosi, F. & Restelli, M. A multiobjective reinforcement learning approach to water resources systems operation: Pareto frontier approximation in a single run. *Water Resour. Res.* **49**, 3476–3486 (2013).
31. Schulman, J., Levine, S., Abbeel, P., Jordan, M. & Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning* 1889–1897 (2015).
32. Johansson, J., Nation, P. & Nori, F. Qutip: an open-source python framework for the dynamics of open quantum systems. *Comput. Phys. Commun.* **183**, 1760–1772 (2012).
33. Duan, Y., Chen, X., Houthooft, R., Schulman, J. & Abbeel, P. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning* 1329–1338 (2016).
34. Vitanov, N. V., Halfmann, T., Shore, B. W. & Bergmann, K. Laser-induced population transfer by adiabatic passage techniques. *Annu. Rev. Phys. Chem.* **52**, 763–809 (2001).
35. Maurand, R. et al. A CMOS silicon spin qubit. *Nat. Commun.* **7**, 13575 (2016).
36. Bluhm, H. et al. Dephasing time of gaas electron-spin qubits coupled to a nuclear bath exceeding 200 µs. *Nat. Phys.* **7**, 109 (2011).
37. Prati, E., Hori, M., Guagliardo, F., Ferrari, G. & Shinada, T. Anderson–Mott transition in arrays of a few dopant atoms in a silicon transistor. *Nat. Nanotechnol.* **7**, 443 (2012).
38. Prati, E., Kumagai, K., Hori, M. & Shinada, T. Band transport across a chain of dopant sites in silicon over micron distances and high temperatures. *Sci. Rep.* **6**, 19704 (2016).
39. Achilli, S. et al. GeVn complexes for silicon-based room-temperature single-atom nanoelectronics. *Sci. Rep.* **8**, 18054 (2018).
40. Hollenberg, L. C. L., Greentree, A. D., Fowler, A. G. & Wellard, C. J. Two-dimensional architectures for donor-based quantum computing. *Phys. Rev. B* **74**, 045311 (2006).
41. Homulle, H. et al. A reconfigurable cryogenic platform for the classical control of quantum processors. *Rev. Sci. Instrum.* **88**, 045103 (2017).
42. Ban, Y., Chen, X. & Platero, G. Fast long-range charge transfer in quantum dot arrays. *Nanotechnology* **29**, 505201 (2018).
43. Torrontegui, E. et al. Chapter 2—shortcuts to adiabaticity (Arimondo, E., Berman, P. R., & Lin, C. C., eds) *Advances in Atomic, Molecular, and Optical Physics*. Vol. 62, 117–169 (Academic Press, Amsterdam, NL, 2013).
44. Jehl, X. et al. Mass production of silicon mos-sets: can we live with nano-devices variability? *Procedia Comput. Sci.* **7**, 266–268 (2011).
45. Breuer, H.-P. & Petruccione, F. *The Theory of Open Quantum Systems*. (Oxford University Press, Oxford, 2002).
46. Clément, N., Nishiguchi, K., Fujiwara, A. & Vuillaume, D. One-by-one trap activation in silicon nanowire transistors. *Nat. Commun.* **1**, 92 (2010).
47. Pierre, M. et al. Background charges and quantum effects in quantum dots transport spectroscopy. *Eur. Phys. J. B* **70**, 475–481 (2009).
48. Kuhlmann, A. V. et al. Charge noise and spin noise in a semiconductor quantum device. *Nat. Phys.* **9**, 570 (2013).
49. Prati, E. & Shinada, T. *Single-Atom Nanoelectronics* 5–39 (CRC Press, Singapore, 2013).
50. Prati, E., Fanciulli, M., Ferrari, G. & Sampietro, M. Giant random telegraph signal generated by single charge trapping in submicron n-metal-oxide-semiconductor field-effect transistors. *J. Appl. Phys.* **103**, 123707 (2008).
51. Prati, E. The finite quantum grand canonical ensemble and temperature from single-electron statistics for a mesoscopic device. *J. Stat. Mech. Theory Exp.* **2010**, P01003 (2010).
52. Prati, E., Belli, M., Fanciulli, M. & Ferrari, G. Measuring the temperature of a mesoscopic electron system by means of single electron statistics. *Appl. Phys. Lett.* **96**, 113109 (2010).
53. Malinovsky, V. & J. Tannor, D. Simple and robust extension of the stimulated raman adiabatic passage technique to n-level systems. *Phys. Rev. A* **56**, 4929–4937 (1997).
54. Peters, J. & Schaal, S. Natural actor-critic. *Neurocomputing* **71**, 1180–1190 (2008).
55. Peters, J., Mülling, K. & Altun, Y. Relative entropy policy search. In *Twenty-Fourth AAAI Conference on Artificial Intelligence* 1607–1612 (Atlanta, 2010).
56. Castelletti, A., Galelli, S., Restelli, M. & Soncini-Sessa, R. Tree-based variable selection for dimensionality reduction of large-scale control systems. In *Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), 2011 IEEE Symposium on* 62–69 (IEEE, 2011).
57. Koller, D., Friedman, N. & Bach, F. *Probabilistic Graphical Models: Principles and Techniques* (MIT press, Cambridge, Massachusetts, 2009).
58. Koch, C. P. Controlling open quantum systems: tools, achievements, and limitations. *J. Phys. Condens. Matter* **28**, 213001 (2016).
59. Lamata, L. Basic protocols in quantum reinforcement learning with superconducting circuits. *Sci. Rep.* **7**, 1609 (2017).
60. Breuer, H.-P. & Petruccione, F. *The theory of open quantum systems*. (Oxford University, New York, 2006).
61. Nalbach, P., Eckel, J. & Thorwart, M. Quantum coherent biomolecular energy transfer with spatially correlated fluctuations. *New J. Phys.* **12**, 065043 (2010).
62. Hughes, K. H., Christ, C. D. & Burghardt, I. Effective-mode representation of non-markovian dynamics: a hierarchical approximation of the spectral density. ii. Application to environment-induced nonadiabatic dynamics. *J. Chem. Phys.* **131**, 124108 (2009).
63. Chin, A. W., Rivas, Á., Huelga, S. F. & Plenio, M. B. Exact mapping between system-reservoir quantum models and semi-infinite discrete chains using orthogonal polynomials. *J. Math. Phys.* **51**, 092109 (2010).
64. Tamascelli, D., Smirne, A., Huelga, S. F. & Plenio, M. B. Efficient simulation of finite-temperature open quantum systems. https://arxiv.org/abs/1811.12418 (2018).
65. Taylor, J. M. et al. Relaxation, dephasing, and quantum control of electron spins in double quantum dots. *Phys. Rev. B* **76**, 035315 (2007).
66. Tempel, D. G. & Aspuru-Guzik, A. Relaxation and dephasing in open quantum systems time-dependent density functional theory: properties of exact functionals from an exactly-solvable model system. *Chem. Phys.* **391**, 130–142 (2011).
67. Schaarschmidt, M., Kuhnle, A. & Fricke, K. Tensorforce: a tensorflow library for applied reinforcement learning. https://github.com/reinforceio/tensorforce (2017).
68. Tange, O. et al. Gnu parallel-the command-line power tool. *USENIX Mag.* **36**, 42–47 (2011).

Acknowledgements

E.P. gratefully acknowledges the support of NVIDIA Corporation for the donation of the Titan Xp GPU used for this research.

Author contributions

R.P. developed the simulation and implemented the algorithms, D.T. elaborated on the Hamiltonian framework and the master equation formalism, M.R. operated the reinforcement learning analysis, and E.P. conceived and coordinated the research. All the authors discussed and contributed to the writing of the manuscript.

Additional information

Supplementary information accompanies this paper at https://doi.org/10.1038/s42005-019-0169-x.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

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