This is the accepted manuscript made available via CHORUS. The article has been published as:

Deep Reinforcement Learning for Quantum Hamiltonian Engineering

Pai Peng (□□), Xiaoyang Huang, Chao Yin, Linta Joseph, Chandrasekhar Ramanathan, and Paola Cappellaro

Phys. Rev. Applied **18**, 024033 — Published 11 August 2022

DOI: 10.1103/PhysRevApplied.18.024033

Deep reinforcement learning for quantum Hamiltonian engineering

Pai Peng (彭湃),[1,*] Xiaoyang Huang,[2,†] Chao Yin,[2] Linta Joseph,[3] Chandrasekhar Ramanathan,[3] and Paola Cappellaro[4,2,‡]

[1] *Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139*

[2] *Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

[3] *Department of Physics and Astronomy, Dartmouth College, Hanover, NH 03755, USA*

[4] *Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139*

(Dated: July 19, 2022)

Engineering desired Hamiltonians in quantum many-body systems is essential for applications such as quantum simulation, computation and sensing. Many widely used quantum Hamiltonian engineering sequences are designed using human intuition based on perturbation theory, which may not describe the optimal solution and is unable to accommodate complex experimental imperfections. Here we numerically search for Hamiltonian engineering sequences using deep reinforcement learning (DRL) techniques and experimentally demonstrate that they outperform celebrated decoupling sequences on a solid-state nuclear magnetic resonance quantum simulator. As an example, we aim at decoupling strongly-interacting spin-1/2 systems. We train DRL agents in the presence of different experimental imperfections and verify robustness of the output sequences both in simulations and experiments. Surprisingly, many of the learned sequences exhibit a common pattern that had not been discovered before, to our knowledge, but has an meaningful analytical description. We can thus restrict the searching space based on this control pattern, allowing to search for longer sequences, ultimately leading to sequences that are robust against dominant imperfections in our experiments. Our results not only demonstrate a general method for quantum Hamiltonian engineering, but also highlight the importance of combining black-box artificial intelligence with understanding of physical system in order to realize experimentally feasible applications.

I. INTRODUCTION

Controllable quantum many-body systems hold great promise not only to expand our understanding of fundamental physics, such as information scrambling and non-equilibrium phases of matter, but also to yield revolutionary technologies in computation, simulation and sensing. A core task of quantum control is to combine elementary control units to engineer desired quantum Hamiltonians. Although it is relatively easy to derive the (approximate) Hamiltonian resulting from a given control sequence, the inverse problem of designing optimal control sequences for a target Hamiltonian is highly challenging. The problem was tackled in Nuclear Magnetic Resonance (NMR) through the development of average Hamiltonian theory (AHT) [1]. Many of the celebrated sequences in NMR are designed from intuition and experience, based on low-order expansions in AHT [2–11]. Unfortunately, simply relying on intuition makes it difficult to find generalizations and capture higher-order effects and control imperfections that might be crucial in experiments. Conventional numerical optimization methods, such as gradient ascent pulse engineering [12] and chopped random basis [13], can yield optimal solutions, but are suitable for optimizing single gates or short sequences where the control landscape is smooth and does not contain too many local optima. Therefore, they are more suitable for optimizing individual (continuous) pulse shapes or a short composite pulse, rather than a pulse sequence containing tens of pulses. While phase and amplitude modulated continuous decoupling are amenable to gradient ascent optimization [14], their experimental implementation has been more limited than pulsed methods.

Recently, artificial intelligence, in particular reinforcement learning (RL) with deep neural networks, has surpassed human intelligence in many complex tasks such as Go [15] and StarCraft II [16]. As a subfield of Machine Learning (ML), RL differs from (un)-supervised learning by learning through exploration and exploitation based on the reward of the result. In stark contrast to conventional optimization methods, RL is a *model-free* method, which only requires minimum knowledge to find the reward. This matches closely the task of Hamiltonian engineering where human intuition into the optimal pulse sequence is limited and might be biased. Deep neural networks (DNN) provide a versatile and powerful way to reparametrize a large search space. Unlike linear optimization, DNN are capable of doing both linear and non-linear mathematical manipulation to turn the input into the output (for RL and DNN, see recent reviews [17, 18]). In the quantum physics context, RL has been shown to provide successful strategies for quantum state preparation [19–24], quantum gate design [25–27], quantum communication [28], quantum memory protection [29], quantum error correction [30–32], quantum state transfer [33],

* paipeng@mit.edu

† P.P. and X. H contributed equally to this work.

‡ pcappell@mit.edu

and quantum sensing [34]. Although RL has in principle demonstrated its advantages for quantum applications via numerical studies, its practical implementation in experiments is still challenging due to non-ideal conditions arising from noise and control imperfections.

Here we apply RL with DNN [deep reinforcement learning (DRL)] to quantum Hamiltonian engineering and experimentally demonstrate its advantage in a non-integrable system. We focus on the task of decoupling a spin-1/2 system with dipolar interaction (i.e. the target Hamiltonian is zero), which is directly useful for quantum memories [35]; our method can be further applied to other quantum engineering scenarios by simply replacing the reward function. To avoid uncountably large parameter space and for easier experimental implementation, we restrict the control space by allowing the machine learner to choose one of five actions at a time (no pulse or a $\pi/2$ pulse along one of 4 axes), with a fixed delay time between actions, until the maximum time is reached. We then numerically calculate the unitary propagator of the resulting pulse sequence, and use the fidelity with respect to target propagator as the reward. The control is restricted to experimentally feasible operations, while still encompassing a wide range of target Hamiltonians that can exhibit integrable, ergodic, localized or prethermal behaviors [36–38]. The restriction leads to a complicated and nonconvex control landscape [39], that would not be amenable to conventional optimization. We thus utilize DNN to reparametrize the control space and implement a state-of-the-art gradient-free method to optimize the neural networks [40].

We not only apply the DRL to the idealized scenario, but also incorporate imperfections, such as pulse frequency offset, on-site disordered field, pulse angle error and finite pulse width, to mimic realistic experiments. We test the performance of the DRL pulse sequences using solid-state nuclear spin systems, and the sequences indeed show the expected robustness even in experiments. Surprisingly, although it is generally believed that symmetric sequences have better performance [41, 42], many of the high-reward sequences found by DRL are not symmetric. Instead, they obey a common "*yxx* pattern" which has not been found before. We analytically explain the advantage of the *yxx* pattern using AHT. Furthermore, the restriction to pulse sequences exhibiting the *yxx* pattern significantly reduces the search space, thus enabling to find longer and more powerful sequences. As a result, we discover sequences that are robust against all relevant imperfections and outperform the previous sequences decoupling sequence [3] in experiments [see Fig. 1(a)], as well as other, more recently designed sequences [35, 43]. Our work demonstrates that some long-established knowledge may not be optimal for quantum Hamiltonian engineering, while pure black-box DRL is also resource consuming. It is beneficial to combine both human knowledge and artificial intelligence for practical applications.

This paper is organized as follows. In Sec. II we explain how to model the Hamiltonian engineering task as a RL problem and our learning algorithm. Sec. III introduces our experimental system and average Hamiltonian theory. The learned sequences together with their experimental tests are presented in Sec. IV, before drawing our conclusions.

II. REINFORCEMENT LEARNING

A. Modeling

We consider quantum Hamiltonian engineering in a spin-1/2 system where the internal Hamiltonian H_0 is the secular dipolar interaction along the z-axis

$$H_0 \equiv D_z = \frac{1}{2}\sum_{j<k}^{N} J_{jk}\left(3S_z^j S_z^k - \vec{S}^j\cdot\vec{S}^k\right), \qquad (1)$$

where $\vec{S}_j = (S_x^j, S_y^j, S_z^j)^T$ are spin-1/2 operators of the j-th spin ($j = 1, \cdots, N$) and J_{jk} the coupling strength between spins j and k. Later we will also use D_x and D_y defined in the similar way. In the training process we consider a 1D spin chain with nearest coupling only, i.e. $J_{jk} = J\delta_{j+1,k}$ ($\delta_{i,j}$ is the Kronecker delta). This is convenient because 1D systems can be efficiently simulated on a classical computer; still, experimental validations are performed in 3D systems. We aim at decoupling the spins, that is, the target Hamiltonian is zero. As we will show, for decoupling purpose, the dimensionality does not play a crucial role; instead the performance is mostly determined by the symmetry of the Hamiltonian.

To decouple the interaction, it is sufficient to apply global rotations to the system, which can be easily implemented in experiments. We further restrict our control to $\pi/2$ global rotations along the $x, y, -x$ or $-y$ directions. These controls are available to almost every quantum platform and are known to be sufficient for decoupling, since they are the building blocks of many well-known decoupling sequences such as WAHUHA [2] and Cory48 [3]. We note that π pulses are instead the preferred building block for single-spin decoupling, but leave invariant the homonuclear interactions, and we thus do not include them as actions here (although it would be an easy modification of our method). Experimentally, the controls cannot be applied arbitrarily fast due to pulse ring-down and apparatus dead-time. Instead, there is usually a minimal delay time τ in between pulses.

The spin chain system and the control rules constitute the environment of RL. We then need to set up the RL agent that interacts with the environment. RL works by building agents that choose a (sub)optimal action at any given time based on the current state (which collects all the previous actions). The action changes the status of the environment and updates the state, and the agents are optimized based on a reward determined by the environment, as shown in Fig. 1(b). We

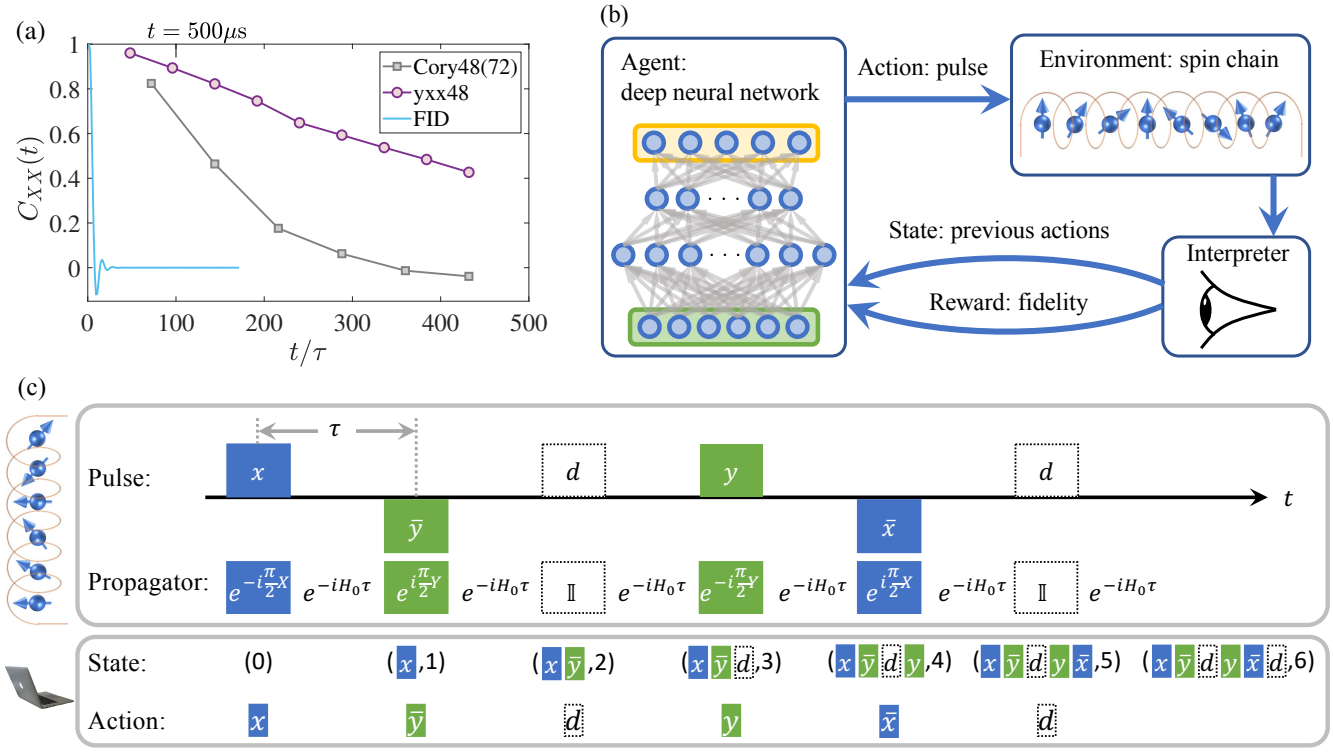

FIG. 1. (a) Quantum dynamics characterized by the decay of x-correlation, $C_{XX} = \mathrm{Tr}(X(t)X)/2^L$ with $X = \sum_j S_x^j$, under free evolution (no pulses, blue curve), and with RF pulse sequences, Cory48 (grey curve) and yxx48 obtained from RL (purple curve). Parameters can be found in Fig. 3 (b) High-level RL protocol for Hamiltonian engineering. The agent, realized as a deep neural network, takes an action based on the current state. This action applies the corresponding control operation to the spin chain. The interpreter updates the state based on the chosen action, and feeds it to the agent to take next action. When the final time is reached, the interpreter calculates the reward, which is then used to optimize the agent. (c) States and actions, and corresponding quantum operations, illustrated using the WAHUHA sequence [2], which can be discretized into 6 times steps. $x, \bar{x}, y, \bar{y}$ in the colored blocks denote $\pi/2$ pulses along $x, -x, y, -y$, respectively. d in the dashed blocks denotes no pulse action (delay). The propagators of pulses shown in (c) correspond to infinitely short pulses (δ pulses). For a x pulse with finite width t_w, the propagator should be $e^{-i\frac{\pi}{2}X - iH_0 t_w}$, similar for pulses along other axes..

first discretize the time in steps of τ – we consider only $t = m\tau$ with $m = 0, 1, 2\cdots$. At each discrete time $t = m\tau$, the agent chooses an action A_m from a set of five possible actions (no pulse or a $\pi/2$ pulse along the $\pm x, \pm y$ directions). The state $S(m)$ is simply defined as a list containing all previous actions and the integer m, $S(m) = [A_0, A_1, \cdots, A_{m-1}, m]$. As an example, we show the state and action at each time step of the WAHUHA sequence in Fig. 1(c). The process ends when the final time $M\tau$ is reached. At this point the environment has received M actions $[A_0, A_1, \cdots, A_{M-1}]$ and undergone a unitary evolution with propagator

$$U(\{A_m\}) = e^{-iH_0\tau} U_{A_{M-1}} e^{-iH_0\tau} U_{A_{M-2}} \cdots e^{-iH_0\tau} U_{A_0}, \tag{2}$$

where U_{A_m} describes the evolution induced by the action A_m. If A_m is "no pulse", then $U_{A_m} = \mathbb{1}$; if A_m is a $\pi/2$ pulse, e.g. along x, then $U_{A_m} = e^{-i(\pi/2)X}$ with $X = \sum_j S_x^j$ being the collective spin-x operator. Later we will also use $Y = \sum_j S_y^j$ and $Z = \sum_j S_z^j$. To compare pulse sequences of different lengths M, it is useful to consider $U_\tau(\{A_m\}) = [U(\{A_m\})]^{1/M}$ the effective propagator for $t = \tau$. How close is the engineered propagator $U_\tau(\{A_m\})$ to the target propagator U_{tgt} is characterized by the fidelity $F(\{A_m\}) = \mathrm{Tr}|U_\tau(\{A_m\})U_{tgt}^\dagger|/2^N \in [0, 1]$, where N is total number of spins. For the decoupling task $U_{tgt} = \mathbb{1}$. Since many good pulse sequences have near unity fidelity, we define the reward function as $R(\{A_m\}) = -\ln[1 - F(\{A_m\})]$ to emphasize the small infidelity [44]. The fidelity also provides a lower bound for observable correlations [38, 45, 46], which can be directly measured in experiments [37], and it is thus a good metric to assess the sequence [19].

Imperfections can be easily incorporated into this model by changing the free evolution propagator or the pulse propagator. In this work we consider the following imperfections that are most evident in our experiments and frequently encountered in other systems: pulse frequency offset, on-site disorder, pulse angle error and finite pulse width. A frequency offset exists when the pulse frequency differs from spin resonance frequency. This can be modeled by adding the offset Hamiltonian ΔZ to H_0,

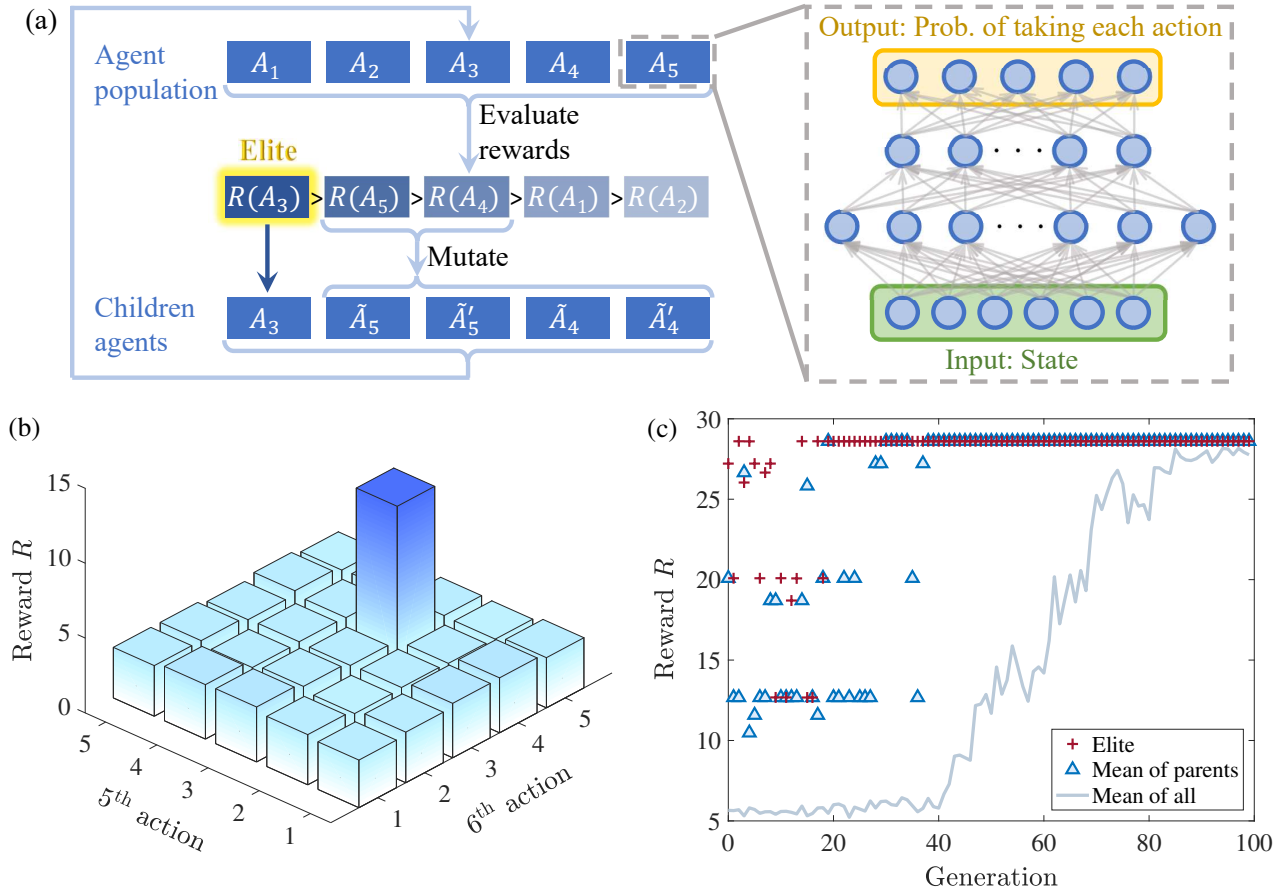

FIG. 2. (a) Schematic of the RL algorithm. We keep a population of agents (in this cartoon the population size is $N_p = 5$, while in actual implementation the parameters are specified in Tab. I). Each agent is a DNN that generates the probability of taking each action based on the current state. We evaluate the reward of all agents and choose the top P agents as parents (here $P = 3$, see Tab. I for the actual implementation parameters). The top agent (Elite) is moved to the next generation without any change, while the other parents are slightly mutated to generate the next generation of agents. (b) Illustration of the (reward) control landscape for a 6-pulse dynamical decoupling sequence in the ideal scenario. For simplicity, here we only show the case where the first four pulses are fixed to be $\{d, x, -y, d\}$, with d denoting the no pulse action. x and y-axis represent the 5th and 6th action.(c) Learning curve of DRL for 12-pulse dynamical decoupling in the ideal scenario. Here $N_p = 201$, $P = 21$.

where Δ is the amount of frequency offset. On-site disorder describes differences in the frequency of each spin; the deviation from the (nominal) mean frequency is a random variable. On-site disorder can be modeled by including the disorder Hamiltonian $\sum_j w_j S_z^j$ in H_0, with w_j uniformly distributed in $[-W, W]$. The disorder Hamiltonian is very similar to the frequency offset, so a sequence that is robust against frequency offset is also typically robust against disorder. Therefore, in the training process we consider the frequency offset only, and in the test process we verify the two are indeed closely related. Angle errors happen when the rotation angle due to the pulse deviates from $\pi/2$ by an amount ϵ. We assume this deviation is the same for all pulses, and thus it can modeled by changing all pulse propagators in the same way. For example, a x pulse with an angle error is described by the propagator $e^{-i(\pi/2)X(1+\epsilon)}$. In experiments, pulses are not instantaneous (delta-pulses) but have a finite width. During the pulse time t_w the spins interact with each other, yielding a propagator $e^{-i(\pi X/2 + H_0 t_w)}$ for the x pulse, similarly for y and z pulses. Beyond what we consider above, many other imperfections can be included by just modifying the reward, as long as the imperfection can be efficiently modeled.

We note that our optimization setup differs from the ones commonly used in numerical pulse engineering, where the time is discretized into tiny time steps and each action only applies a small evolution to the system [12, 19, 25]. Instead in our protocol, neither the free evolution time τ nor the pulse rotation angle need to be small. Two differences can yield an advantage in certain experimental conditions: (i) $\pi/2$ pulses are usually available and well calibrated, while modulating the control drive over short timescale poses challenges on the hardware and is more difficult and inefficient to calibrate; (ii) by taking a larger step per action, our method is more suitable to finding long pulse sequences, while previous methods are typically used to optimize single gates. Con-

versely, the larger step is accompanied by a worse control landscape –the reward function is not smooth on the space of final states, as shown in Fig. 2(b). However, the deep neural network and gradient-free optimization method successfully solve this issue, as we show in the next subsection.

B. Algorithm

We first explain how the agent works [see the dashed box in Fig. 2(a)]. The agent is a DNN that takes the state as input and generates the next action as the output, as introduced in Sec. II A. The agent DNN contains two hidden linear layers, with the number of neurons in each layer proportional to the input and output size. We use rectified linear unit (ReLU) [47] as the activation function. At each step m, the agent takes the state $S(m)$ as the input and generates 5 positive numbers corresponding to the probability of taking the 5 actions. Then the action is chosen randomly according to the probabilities, and the state now become $S(m+1)$. We apply the above procedure starting from $m = 0$ until the maximum step M is reached, then we get an output sequence from the agent.

Here we explain how we optimize the agents. The process is illustrated in Fig. 2 (a). We start with N_p agents, and for each agent generate 3 sequences (note the selection of actions is a random process so the 3 sequences may not be the be same [48]), and select the highest reward among the 3 sequences as the reward of the agent. The reward is obtained on a 3-spin system and we verify that going to larger systems does not change our results. Sorting the population of agents by the reward function in descending order, we apply the truncation selection to choose the top P individuals as the parents. Among the parents, we further select the most promising parent, the so-called *Elite*, from all the parents by regenerating a few sequences (typically 5) and comparing their rewards. The Elite will be included in the children generation without any change. Every parent agent other than the Elite will be mutated by adding a random Gaussian noise multiplying mutation power μ to all the DNN parameters to generate $(N_p-1)/(P-1)$ children agents. The mutation process plays the role of "exploration" (search in a large space) in RL. Too much exploration (large μ) will result in excessive randomness, making the process closer to a pure random search; too little exploration instead (small μ) might leave the RL stuck into a local minimum. In practice, we decrease μ during the learning process (so that in the beginning we explore a large space and later we search near the good agents) following the function

$$\mu(g) = \mu_0(1 - g/G), \tag{3}$$

where $g = 1, ..., G$ denotes the agent generation. In this paper, we empirically set $\mu_0 = 0.05$ and $G = 100$. We repeat the process until the maximum number of generations G is reached.

With the truncation selection and mutation, we are able to balance *exploration* (search in a large space) and *exploitation* (focus on the promising area). One example of the learning curve of DRL is shown in Fig. 2(c). After 20 generations, the Elite DNN starts showing a convergent reward towards the optimal one (global minimum for this case); after 40 generations, all the parent agents (mean of parents) begin to converge; the entire agent population converges after around 90 generation. When approaching the end of learning, μ becomes small, meaning little exploration but great exploitation. With little random noise, children agents are able to reproduce the optimal reward consistently, indicating the convergence of the algorithm.

III. EXPERIMENTAL AND THEORETICAL BACKGROUND

A. Experimental system

We use a solid-state NMR quantum simulator to experimentally test the performance of RL pulse sequences in realistic conditions. Most of the experimental results presented in this work are obtained from a single crystal of CaF_2, where the ^{19}F nuclear spins-1/2 form simple cubic structure. The sample is placed in a strong magnetic field (7 T) at room temperature. The nuclear spins interact via the secular dipolar interaction $D_z = \frac{1}{2}\sum_{j<k}^{N} J_{jk}\left(3S_z^jS_z^k - \vec{S}^j\cdot\vec{S}^k\right)$ with $J_{jk} = \hbar\gamma_F^2\frac{3\cos(\theta_{jk})^2-1}{|\vec{r}_{jk}|^3}$, where γ_F is the gyromagnetic ratio of ^{19}F nuclei, $\vec{r}_{jk}$ is the displacement between spins j and k, θ_{jk} is the angle between $\vec{r}_{jk}$ and the magnetic field (aligned with the z-axis). The maximum possible J_{jk} is 65.8 krad/s for CaF_2 [49]. The relaxation time T_1 of our sample is $T_1 \approx 14$ s, much longer than the time scale we explore here. The collective spin rotations are realized by on-resonance RF pulses with a $t_w = 1.02$ μs $\pi/2$ pulse width. We can also artificially introduce and tune errors in addition to intrinsic imperfections. We introduce angle error by setting the pulse width to $1.02(1+\epsilon)$ μs. We can also use off-resonance pulses to introduce a frequency offset.

At room temperature and in a strong magnetic field along the z axis, the initial state of an ensemble of ^{19}F nuclear spins is described by the density matrix $\rho(0) \approx (\mathbb{1} - \epsilon' Z)/2^N$, with N being the number of spins and $\epsilon' \sim 10^{-5}$. The identity part of the density matrix does not contribute to the NMR signal, so we only care about the deviation from it, $\delta\rho = 4Z/N$, which has been normalized such that $\mathrm{Tr}(\delta\rho Z)/2^N = 1$. NMR experiments measure the collective magnetization along the x axis, i.e. the signal is $\mathrm{Tr}(\delta\rho(t)X)/2^N$. If we regard the density matrix $\delta\rho$ as an observable, this signal is mathematically equivalent to an infinite-temperature correlation $\mathrm{Tr}(\delta\rho(t)X)/2^N \equiv \langle\delta\rho(t)X\rangle_{\beta=0}$. Using collective RF pulses, we can rotate the initial state and the observ-

Name	Sequence	Training parameter
Ideal6	y, x, x, y, -x, -x	$\Delta = 0, \epsilon = 0, t_w = 0, N_p = 201, P = 11$
Offset48	$-y, -y, -y, -x, -y, -y, -y, -y, -x, -x, -x, -y,$ $-x, -x, -y, -y, -y, -x, -x, -x, -y, -y, -y, -y,$ $-x, -x, -y, -x, -x, -x, -x, -x, -x, -y, -y, -y,$ $-x, -x, -x, -x, -x, -x, -y, -x, -x, -x, -y, -x$	$\Delta = 0, \pm 3J, \pm 5J, \epsilon = 0, t_w = 0, N_p = 3001, P = 31$
Angle12	$-y, x, -x, y, -x, -x, -y, x, -x, y, x, x$	$\Delta = 0, \epsilon = 0.05, t_w = 0$ and $\Delta = 0, \epsilon = 0, t_w = 0.1\tau, N_p = 801, P = 21$
PW12	$-y, -x, -x, x, x, y, -x, -x, -y, y, x, x$	$\Delta = 0, \epsilon = 0, t_w = 0.1\tau, N_p = 801, P = 21$
yxx48	$y, -x, -x, y, -x, -x, -y, x, x, y, -x, -x,$ $-y, x, x, -y, x, x, y, -x, -x, y, -x, -x,$ $-y, x, x, y, -x, -x, -y, x, x, -y, x, x,$ $y, -x, -x, -y, x, x, y, -x, -x, -y, x, x$	$\Delta = J, \epsilon = 0.05, t_w = 0, N_p = 801, P = 21,$ with yxx restriction
yxx24	$-y, x, -x, y, -x, -x, y, -x, x, -y, x, x,$ $y, -x, x, -y, x, x, -y, x, -x, y, -x, -x$	Built from Angle12

TABLE I. Representative DRL pulse sequences under different training conditions. Angle12 appears in two training conditions. yxx24 is build from Angle12 using AHT analysis (see Appendix C).

able to be X, Y or Z. Therefore, we can measure the three autocorrelations $C_{XX}(t) = 4\langle X(t)X\rangle_{\beta=0}/N$ and C_{YY}(t), $C_{ZZ}(t)$ defined in a similar way. Although in principle to get the propagator fidelity we have to measure autocorrelations of all observables, in the Supplementary Material we show that the geometric average of these three autocorrelations, $C_{avg} \equiv (C_{XX}C_{YY}C_{ZZ})^{1/3}$ already approximates the behavior of the propagator fidelity.

To experimentally investigate on-site disorder [Fig. 6(a)], we work with ^{19}F nuclear spins in fluorapatite (FAp) [50]. The ^{31}P nuclear spins-1/2 in the crystal are randomly polarized, giving rise to a disorder Hamiltonian $H_{dis} = \sum_j h_j S_z^j$, with h_j being a random variable representing the disordered field at j^{th} ^{19}F nucleus. Interaction between ^{19}F nuclear spins is also given by the secular dipolar interaction as in CaF_2 but with a lower maximum possible strength 32.7 krad/s. The ^{19}F nuclei form a quasi-1D structure, as the interaction along the z-direction is ~ 40 times stronger than along the other two directions. Although the quasi-1D nature is not important in the context of this work, it is useful for quantum simulation [36–38, 45]. The relaxation time for the FAp crystal is $T_1 \approx 0.8$s, shorter than for the CaF_2 sample, but still much longer than the duration of a single experiment.

B. Average Hamiltonian theory

AHT [1] is useful in understanding the performance of different pulse sequences, so we briefly review it here. A quantum system under a time-dependent control can be generally described by the Hamiltonian $H(t) = H_0 + H_c(t)$, with H_0 the intrinsic Hamiltonian and H_c the control Hamiltonian. For the pulsed control case, $H_c(t)$ is piece-wise constant and nonzero only within the pulse width. AHT starts by defining the *toggling frame*, an interacting frame that rotates with H_c, i.e., $U_c(t) = \mathcal{T}[e^{-i\int_0^t H_c(t')dt'}]$, where $\mathcal{T}$ is the time-ordered operator. In the toggling frame the Hamiltonian is $H_{tog}(t) = U_c^\dagger H_0 U_c$. At $t = 0$ the toggling frame coincides with the lab frame. If after the pulse sequence the toggling frame rotates back to the lab frame (as it is the case for all decoupling sequences), then the toggling frame propagator $U_{tog}(M\tau) = \mathcal{T}[e^{-i\int_0^{N\tau} H_{tog}(t')dt'}]$ coincides with the lab frame propagator $U(\{A_M\})$. Although the toggling frame Hamiltonian is still time-dependent, it does not contain strong pulses and can be effectively approximated by a time-independent local Hamiltonian (the average Hamiltonian) H_A satisfying $U_{tog}(N\tau) = e^{-iH_A N\tau}$ [51]. H_A can be found perturbatively using the Floquet-Magnus expansion [52, 53]

$$\begin{aligned} H_A =& \frac{1}{M\tau}\int_0^{M\tau} H_{tog}(t)dt \\ &- \frac{i}{2M\tau}\int_0^{M\tau} dt_1 \int_0^{t_1} dt_2 [H_{tog}(t_1), H_{tog}(t_2)] \\ &+ O[(M\tau)^2], \end{aligned} \tag{4}$$

where the right-hand side of the first line is the zeroth-order average Hamiltonian as it scales as $(M\tau)^0$ and the second line is the first-order average Hamiltonian. If the average Hamiltonian is zero to certain order in 1D, it remains zero for higher dimensions. For example, the zeroth-order average Hamiltonian for the WAHUHA sequence is $2(D_x + D_y + D_z) = 0$ regardless of dimensionality. The dimensionality can affect the fidelity by changing the magnitude of the leading nonzero higher-order Hamiltonian.

IV. RL PULSE SEQUENCES

We apply DRL to different scenarios and generate various pulse sequences. Some representative ones are shown in Table I. We first tackle the case where we introduce

only one imperfection at a time, and we later consider the case where several imperfections are present.

A. Single imperfections

Control sequences that tackle zero or single sources of imperfections, Ideal6, Offset48, Angle12 and PW12, are directly generated by RL without any human input, trained with no error, offset, angle error and finite pulse width, respectively. In the training process, we start with a small M (short sequence) and increase M until we find a high-reward sequence. As can be seen from Table I, all the good sequence lengths are multiple of 6 [49]. This can be understood via AHT: to cancel the zeroth-order average interaction Hamiltonian and rotate back the toggling frame with the allowed operations, the sequence length must be a multiple of 6. Moreover, when the dipolar interaction dominates, the machine learns to cancel the zeroth-order interaction Hamiltonian as quick as possible, i.e. in Idea6, Angle12 and PW12 the toggling frame Hamiltonian averages to zero every 3τ. This coincides with the discovery in Ref. [54]. The Offset48 sequence, on the other hand, is trained under strong offset, thus does not obey this rule. We also notice that the no-pulse action is never chosen by DRL, in contrast to well-known decoupling sequences such as Cory48 and WAHUHA. Although the no pulse action is useful for some applications that requires a long time window between pulses, such as pulsed gradient generation and stroboscopic detection, it is not advantageous for decoupling. This can also be understood from AHT: it is advantageous to apply pulses as frequently as possible so that higher orders in the average Hamiltonian are suppressed. Previously several attempts were made following this intuition, adding additional pulses in some of the 2τ window in WAHUHA-like sequences, but they did not lead to better performance [54, 55]. Here we find RL discovers a completely different pattern that applies pulses as frequently as possible, and outperforms previous sequences (see next subsection) [41, 42]. In contrast, one of the most common strategies in conventional sequence design is to first come up with a sequence whose zeroth-order average Hamiltonian is the target Hamiltonian, and then symmetrize the sequence to cancel all odd orders. Symmetrization is achieved by following the original sequence by the same sequence but in reversed order, and with a π phase shift. For example, the sequence $x, y, -x, -y$ is symmetrized to $x, y, -x, -y, y, x, -y, -x$. Strikingly, many DRL sequences are not symmetric, e.g., Ideal6, Offset48. This suggests that symmetrization is not optimal in some scenarios, also noted in [54].

Figure 3 shows the experimentally measured average correlation at $t = 72\tau$ and numerically simulated infidelity $1-F$ of Angle12 and Offset48, in comparison with Cory48 in various experimental conditions. The Cory48 sequence has commonly been used for benchmarking decoupling sequences [35, 43, 56, 57] and suspending time evolution [58–63]. We will further compare our best sequence (see next subsection) to other, more recent decoupling sequences [35, 43]. Note that although Cory48 contains only 48 pulses, its length is 72τ because it also contains 24 no-pulse actions. Therefore, we explicitly denote it as Cory48(72). The experiments are done with CaF_2. As our experimental apparatus is not ideal and does not allow varying pulse width over a large range, we cannot provide experimental tests of Ideal6 and PW12, but we provide numerical results in [49]. $C_{avg}(72\tau)$ of Offset48 is not directly measurable because 72 is not a multiple of 48. Instead, we plot the average of $C_{avg}(48\tau)$ and $C_{avg}(96\tau)$, which is a good approximation of $C_{avg}(72\tau)$ as shown in [49]. Fig. 3(b) shows that the fidelity of both Angle12 and Offset48 have a worse scaling $(J\tau)^6$ compared with Cory48 $(J\tau)^{10}$, because Cory48 cancels the average interaction Hamiltonian to higher order. However, this higher-order effect is not evident in experiments [Fig. 3(a)] due to experimental imperfections dominating over higher order interaction Hamiltonian. With the same robustness, a shorter sequence is usually preferable, as concatenation and symmetrization of shorter sequences have been traditionally used to form longer sequences with more benign scaling of infidelity with errors. Shorter sequences also allow a finer probing of the time evolution (which is possible only at multiples of the cycle time). As for the offset, the scaling of Offset48 and Cory48 are both quadratic in the small Δ region [Fig. 3(f)]. However, when the offset becomes larger, Offset48 outperforms Cory48, as shown both experimentally and numerically [Fig. 3(e) and (f)]. This phenomenon is beyond AHT and intrinsically nonperturbative. Not surprisingly, Angle12 is not robust against offset, nor is Offset48 against angle error.

B. Multiple imperfections and *yxx* pattern

DRL is successful in learning good pulse sequences in the presence of a single imperfection. When two or more imperfections exist simultaneously, the number of satisfactory sequences is significantly reduced and the DRL is unable to find one within reasonable time. Although this may be solved by using more powerful computers or more sophisticated algorithms, we take a physicist's approach. We learn from sequences DRL found in the presence of a single imperfection and use our understanding to design more powerful sequences. We notice that Ideal6, Angle12 and PW12 are all built from the 3-pulse block $(\pm y \pm x \pm x)$ or its equivalent form $(\pm x \pm y \pm y)$, which we refer to as "*yxx* pattern".

Figure 4 shows the toggling frame configuration and toggling frame Hamiltonian for two consecutive *yxx* blocks. The last pulse is rotated to the first position for easier analysis, and we note that rotation of pulses unitarily changes the Floquet propagator thus does not change the fidelity [64]. We denote the dipolar interaction along the α axis as D_α as defined in Eq. 1 with

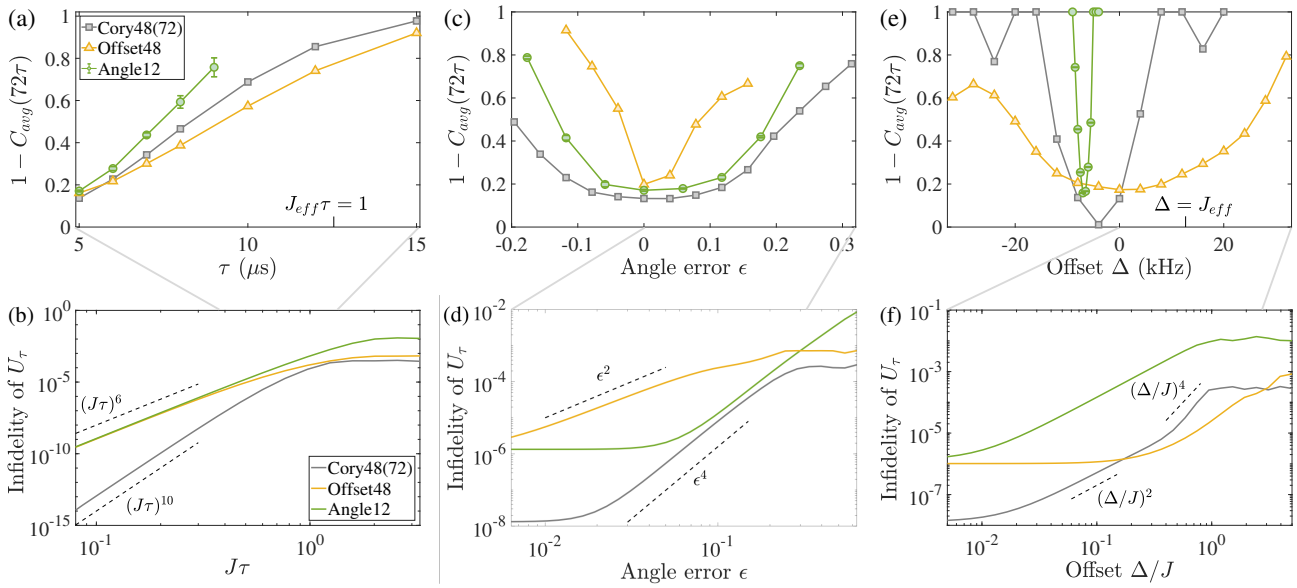

FIG. 3. Experimentally measured average correlation at $t = 72\tau$ on CaF_2 (top panels) and numerically simulated propagator infidelity $1 - F$ (bottom panels) of DRL sequences and Cory48 for different τ (a-b), angle error (c-d) and frequency offset (e-f). Dashed lines in (b) (d) and (f) show the scalings specified by nearby expression. We set $C_{avg} = 0$ if any of C_{XX}, C_{YY}, C_{ZZ} is smaller than zero. As 72 is not a multiple of 48, $C_{avg}(72\tau)$ of Offset48 is obtained as $[C_{avg}(48\tau) + C_{avg}(96\tau)]/2$. Imperfections are set to zero unless specified by the horizontal axis, with the exception of Angle12 experimental data in (a) and (c), which are taken at the optimal non-zero Δ due to the presence of phase transient (see Appendix A for details). In (c) and (e), the pulse center-to-center delay is $\tau = 5$ μs. $J_{\text{eff}} = 79.7$ krad/s for the orientation of our sample [49]. Error bars of Cory48 and Offset48 experimental data are determined from the noise in the free induction decay, which is smaller than the marker size thus not shown. Angle12 has larger error bars in (a) and (c) due to the inaccuracy in finding the optimal Δ. In simulations we assume the pulse width is infinitesimal. We use $J = 32.7$ krad/s as in FAp, $N = 8$, periodic boundary condition and assume nearest-neighbor interactions. In (d) and (f) $\tau = 10$ μs.

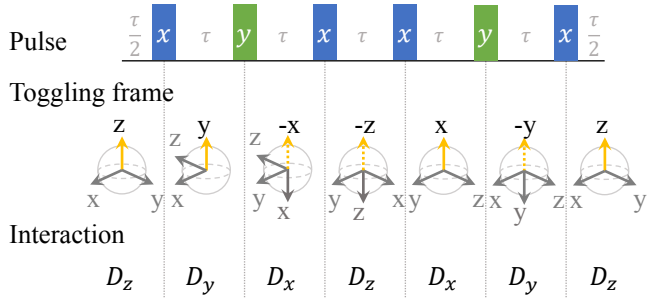

FIG. 4. yxx-type sequence. Top: the pulse sequence. Middle: toggling frame transformation at each time. Arrows mark the orientation of toggling frame axis in the lab frame, where the yellow arrow highlights the axis overlapping with lab frame z-axis. Bottom: the dipolar interaction in the toggling frame at each time.

$\alpha = x, y, z$. Because $D_x + D_y + D_z = 0$, in the ideal case the xyx block cancels the zeroth-order interaction Hamiltonian and is the shortest sequence to do so. Although the length of a solid echo is only 2τ [65], it does not qualify as a decoupling sequences as defined here, because the average Hamiltonian is $D_y + D_z = -D_x$ and thus the sequence only protects the X state. The shortest known decoupling sequence is WAHUHA, whose length is 6τ though only contains 4 pulses [2]. The first-order average Hamiltonian of the first xyx block is $-i[D_x, D_y]\tau/6$, which is cancelled by the contribution from the second block. Changing the signs of the pulses will not change the toggling frame interaction Hamiltonian, as the dipolar interaction is invariant under π rotation. Therefore, the yxx pattern guarantees vanishing zeroth- and first-order average interaction Hamiltonian in shortest possible timescale (6τ). This might also explain why DRL chooses yxx pattern over symmetrization, which cancels the first order terms only after the entire sequence, thus leading to larger higher order terms.

With this understanding, we adopt two approaches to construct longer sequences that are robust against multiple imperfections. First we can restrict our search to sequences with yxx patterns only, so the agent only needs to choose the plus or minus sign instead of five actions. This significantly reduces the dimension of the search space from 5^N to 2^N. In this way we find the yxx48 sequence shown in Table I. A second approach is to directly modify the RL sequences found above in order to cancel the additional imperfections. As Angle12 is robust against both angle error and finite pulse width, we double and modify it into the 24 pulse sequence shown as yxx24 in Table I so that it is also robust against offset

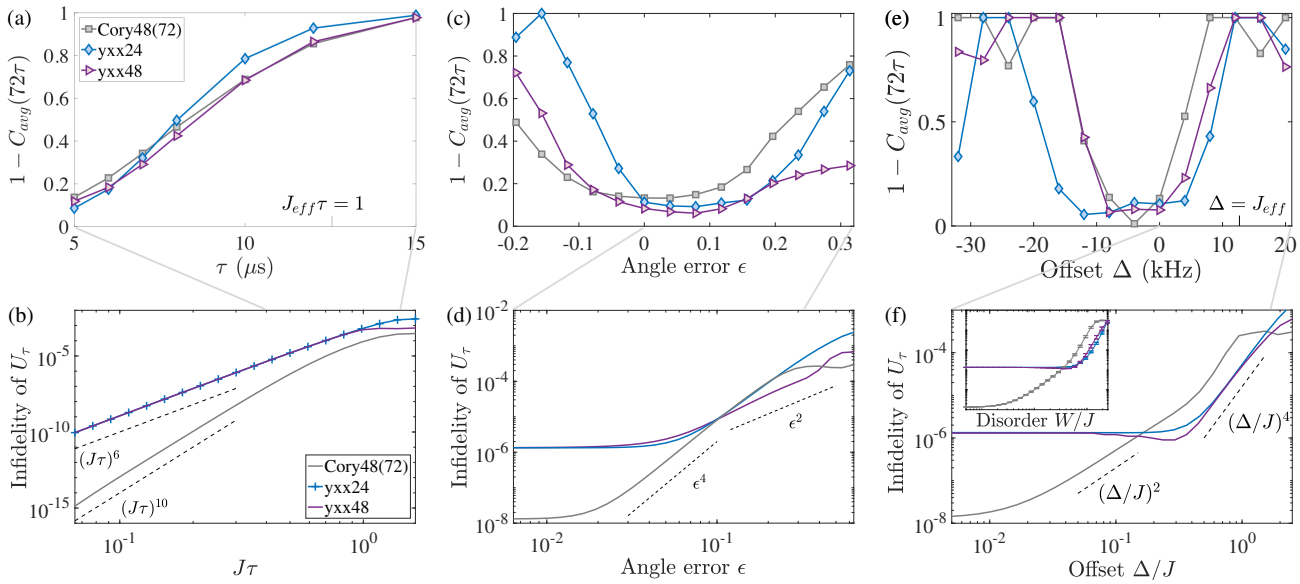

FIG. 5. Experimentally measured average correlation at $t = 72\tau$ on CaF_2 (top panels) and numerically simulated propagator infidelity $1 - F$ (bottom panels) of DRL sequences and Cory48 for different τ (a-b), angle error (c-d) and frequency offset (e-f). The inset of (f) shows the propagator infidelity versus disorder strength W averaged over 20 disorder realizations with error bars showing one standard deviation of the average infidelity. The x and y axes ranges of the inset are the same as in the main plot. Dashed lines in (b) (d) and (f) show the scalings specified by nearby expression. As 72 is not a multiple of 48, $C_{avg}(72\tau)$ of yxx48 is obtained as $[C_{avg}(48\tau) + C_{avg}(96\tau)]/2$. Imperfections are set to zero unless specified by the horizontal axis. Error bars of the experimental data are determined from the noise in the free induction decay which is smaller than the marker size thus not shown. Other parameters are the same as in Fig. 3.

(See appendix C for further details.)

The performance of these two *yxx* sequences are shown in Fig. 5. Again, the two *yxx* sequences have a worse scaling with τ than Cory48 [$(J\tau)^6$ versus $(J\tau)^6$, Fig. 5(b)], but this effect is barely seen in experiments, Fig. 5 (a). Experimentally, the three sequences are all robust against angle error as the average correlations remain close to 1 in a relatively large region of ϵ [Fig. 5(c)]. Although simulation shows Cory48 is better at small ϵ region, it is not evident in experiment because in this region the angle error is not the dominating effect. For the offset, both *yxx* sequences show a better scaling compared to Cory48 [Δ^4 for *yxx* sequences versus Δ^2 for Cory48, Fig. 5(f)]. In experiments we observe that the C_{avg} remain close to 1 in $\Delta \in [-12, 5]$ kHz for the *yxx*24 sequence and $\Delta \in [-8, 0]$ kHz for the *yxx*48 sequence, but for Cory48 the average correlation drops significantly when deviating from the optimal Δ. The fact that the robust region of yxx24 is wider than that of yxx48 might be a result of unknown experimental imperfections that correlate with frequency offset. As the frequency offset has the same form as on-site disorder, we expect any pulse sequence to show similar robustness against the two imperfections. This is confirmed by comparing Fig. 5(f) and its inset (for the *yxx* sequence, disorder and offset effects are equivalent up to first-order AHT Appendix B).

Finally, we compare *yxx* sequences and Cory48 under the best achievable experimental conditions with the two experimental samples in the same apparatus. Data taken with the disorder-free CaF_2 sample are in Fig. 6(a) and the disordered sample FAp is presented in Fig. 6(b). In both samples, yxx24 and yxx48 protect the correlation to significantly longer time than Cory48. In Appendix D, we further show that our sequences also outperform more recently discovered sequences [35, 43] in the disordered sample. Results on a different sample and different apparatus can be found in Appendix E.

V. CONCLUSION AND OUTLOOK

We designed robust decoupling sequences using DRL and experimentally demonstrated that they lead to better performance than some sequences that have been widely used. We directly consider $\pi/2$ pulses as actions, enabling discovery of long sequences, and we use a gradient-free optimization method together with deep neural networks to tackle the complex control landscape. DRL without any human insight is capable of dealing with single imperfections. Surprisingly, many of the DRL sequences are not symmetric, instead, we observe a *yxx* pattern. Building on our understanding of the *yxx* pattern, we then find sequences that are robust against all dominant imperfections present in our experiments, leading to a better performance than the Cory48 sequence in two different samples. Our work emphasizes the usefulness of both artificial intelligence and human knowledge of the physical system in realistic applications.

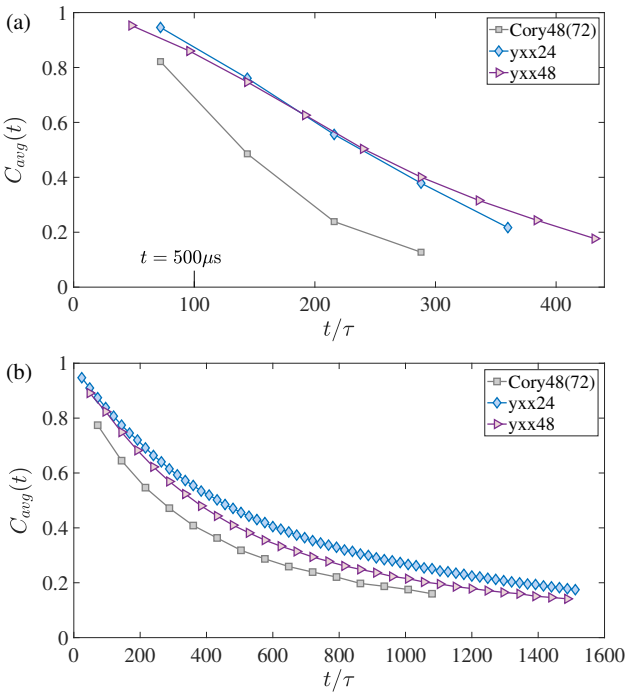

FIG. 6. Experimentally measured average correlation at the best calibrated condition of CaF_2 (a) and FAp (b) as a function of time. Error bars of the experimental data are determined from the noise in the free induction decay which is smaller than the marker size thus not shown. Other parameters are the same as in Fig. 3.

We conclude this paper by pointing to some future research directions. (I) Although we focused on decoupling interacting spin-1/2 systems, a task that has applications in building spin-based quantum memories, our method is completely general to other systems and applications, by simply modifying the reward function to engineer the desired Hamiltonian. It would be interesting to apply this method to quantum simulation or quantum sensing. In this work we train the machine learner in the context of solid-state NMR, where the pulsed controls have been developed and optimized for 50 years, yet RL still shows an advantage. We expect our methods might yield even more significant improvements in other quantum platforms whose controls are developed more recently, such as color centers in solids, cold atoms, trapped ions, and superconducting circuits. (II) The DRL training in this work was simply performed on a personal laptop, so there is still large space for improvement on the computational side, e.g. by using a supercomputer with GPU acceleration to tackle more complex control sequences. (III) Further improvements could be obtained by a stronger interface between machine learner and the physical system. Here we trained the DRL purely using a classical computer and tested the learned sequences on a quantum simulator. Our method can be readily modified into a hybrid classical-quantum DRL process: the DRL agents on a classical computer generate a sequence, which is then applied in a quantum system; then one use an experimental observable, such as the correlation decay rate, as the reward to train the agents. In this way the the modeling of system Hamiltonian and control imperfections is not required. While in our current learning process simulating the spin-chain environment only takes a small portion of the total CPU time, this could change for different tasks that require simulating a many-body non-integrable system. Then, we expect replacing the classical simulation with quantum experiments will improve the training time and open avenues for devising quantum control protocols.

ACKNOWLEDGMENTS

Authors would like to thank H. Zhou and L. Viola for discussion. This work was supported in part by the National Science Foundation under Grants No. PHY1734011, No. PHY1915218, and No. OIA-1921199.

Appendix A: Phase transient effects on Angle12

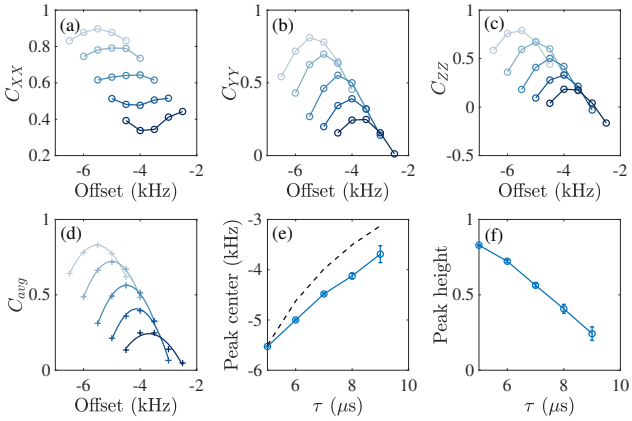

FIG. 7. Experimental study of angle-12 for different offset and τ. (a-d) show C_{XX}, C_{YY}, C_{ZZ} and the average correlation respectively. Different curves are obtained with τ from 5 μs to 9 μs, with a step of 1 μs and lighter color representing smaller τ. In (d), the plus sign marks the experimental data point and the curve shows the parabolic fitting, whose peak center and height are shown as the blue curve in (e) and (f), respectively. The length of the error bars corresponds to two standard deviation of the fitted results. In (e), the black dashed line shows the peak center expected from first-order AHT.

The effects of pulse phase transients are typically difficult to quantify, as they introduce difficult to characterize time-dependent Hamiltonian terms. Still, here we show it is still possible to capture their essence using a simple model. In the future, we could even include phase transients into the reward function to design sequences that are robust against them.

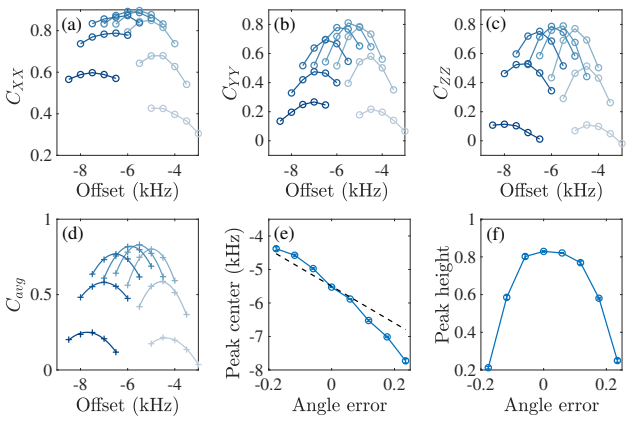

FIG. 8. Experimental study of angle-12 for different offset and angle error. (a-d) show C_{XX}, C_{YY}, C_{ZZ} and the average correlation respectively. Different curves are obtained with angle error from -0.18 to 0.24 with a step of 0.06, with lighter color representing smaller angle error. In (d), the plus sign marks the experimental data point and the curve shows the parabolic fitting, whose peak center and height are shown as the blue curve in (e) and (f), respectively. The length of the error bars corresponds to two standard deviation of the fitted results. In (e), the black dashed line shows the peak center expected from first-order AHT.

In Fig. 3(e) of the main text, we see that the optimal performance point of the Angle12 sequence deviates from $\Delta = 0$ by a significant amount. Here we show that this is due to the cancellation of phase transient error and offset in Angle12. Since we do not include the phase transient effect as an error source during the training process, we should also minimize this effect in the experimental testing. This can be realized by pinning the offset to Δ_0 for Angle12. Other sequences happen to be sufficiently robust to phase transient that they do not require any special treatment.

We first explain the physics of phase transients. When creating a RF square pulse along the x-axis, the leading and trailing edges are not sharp and they inevitably generate a small y-component [42]. Although the exact description of phase transient is not known, the simple model introduced in Ref. [42] can qualitatively explain experimental results. An x-pulse with a phase transient is modeled by a propagator $e^{-i\alpha_1 Y}e^{-i(\pi/2)X}e^{-i\alpha_2 Y}$, where α_1 and α_2 denote the strength of the y-component at the trailing and leading edge, respectively. For pulses along other axes, this model assumes that the additional component is always $+\pi/2$ phase shifted with respect to the main component. Using AHT, we find the zeroth-order average phase transient of Angle12 is $(\alpha_1 - \alpha_2)(-4X + 2Y)/(12\tau)$. As the zeroth-order offset of Angle12 is $\Delta(-4X+2Y)/12$, the two cancel each other out at the optimal offset $\Delta_0 = (\alpha_2 - \alpha_1)/\tau$, leading to the non-zero optimum point in Fig. 3(e).

We verify this relation in Fig. 7, where we show the autocorrelations for different offsets and τ. Figure 7(a-c) shows C_{XX}, C_{YY} and C_{ZZ}, where each curve is taken for a given τ and darker colors denote larger τ. For each τ, there is indeed a peak at Δ_0. When the offset deviates from Δ_0, we see the decrease of C_{XX} is not as significant as that of C_{YY} and C_{ZZ}, because this deviation results in an effective magnetic field $\propto 4X - 2Y$, which is close to the x direction. When τ increases, the peak center Δ_0 shifts toward smaller offset (in absolute value) as expected from the AHT analysis above. To quantitatively analyze this trend, we fit C_{avg} at a fixed τ to a parabolic function, as shown in Fig. 7(d). In Fig. 7(e) we plot the peak center Δ_0 as a function of τ and compare it with the zeroth-order AHT value $\Delta_0 \propto 1/\tau$. The two quantities show the same trend, with an imperfect match due to the simplicity of the model. Because of the variation of Δ_0 with τ, it is not reasonable to use the same offset when testing Angle12's robustness against τ, instead, we use the fitted C_{avg} peak height in Fig. 3(a). We note that our analysis does not artificially increase the robustness of Angle12 compared to the ideal case without phase transient. By choosing the optimal Δ_0, we can at most cancel zeroth-order effects of the phase transient, while higher order terms and cross terms between phase transient and other Hamiltonian components still lead to the degradation of the autocorrelations. Therefore, we still underestimate the robustness of Angle12 even when we are using the optimal Δ_0.

When introducing the angle error, the optimal cancellation condition also changes. This can be seen from the first-order AHT. The first-order cross term between angle error and phase transient leads to an additional field $(-2X - 2Y + 4Z)\epsilon(\alpha_1 - \alpha_2)/(12\tau)$; the cross term between angle error and offset leads to a field $(-2X + 4Y - 4Z)\epsilon\Delta/12$. In other words, introducing an angle error dresses the effective fields due to phase transient and offset, and now the two cannot exactly cancel each other. Still, there exists an optimal offset Δ_0 where the residual field is the smallest. To first-order in ϵ, $\Delta_0 = (\alpha_2 - \alpha_1)(1+\epsilon)/\tau$. We experimentally verify this relation in Fig. 8. In Fig. 8 (a-d) the darker color denotes larger ϵ and we see that Δ_0 decreases when increasing ϵ (note that $\Delta_0 < 0$) as expected from the analysis above. Again we fit C_{avg} to a parabolic function to get the peak center and peak height. The peak center as a function of ϵ is shown in Fig. 8(e) with the dashed line denoting the theoretical value assuming the Δ_0 at $\epsilon = 0$ is exact. Our experimental results do agree with the theoretical expectations. The peak height is shown in Fig. 8(f) and also Fig. 3(c).

In addition to the two verifications above, we also increase the pulse width and observe Δ_0 decreases (not shown). This is because the phase transient becomes less evident when using a lower RF power.

Appendix B: Similarity between offset and disorder

We now consider two types of control imperfection: magnetic field disorder and offset. Both interactions are fields along the longitudinal z axis, so many of their properties are similar. Here we formally demonstrate that: (i) any sequence with vanishing zeroth-order offset Hamiltonian must also have vanishing zeroth-order disorder and vice-versa (ii) for *yxx* sequences this is also true to first-order.

Consider the Hamiltonian $H = D_z + H_z$, where D_α is the dipolar interaction in the α direction and $H_\alpha = \Delta \sum_j S_\alpha^j$ for the offset and $H_\alpha = \sum_j w_j S_\alpha^j$ for the disorder in the α direction. A uniform offset can be viewed as a special disorder realization, so a vanishing average Hamiltonian in the presence of disorder implies a vanishing average offset Hamiltonian. We then only need to demonstrate the converse in the following.

The zeroth-order average Hamiltonian for single-body Hamiltonians can be calculated by considering a representative site. Thus the relative strength of the interaction among sites does not matter, and a vanishing zeroth-order offset is always equivalent to a vanishing zeroth-order disorder.

The first-order average Hamiltonian contains three parts: the interaction-interaction commutator, which we can ignore for this discussion, the imperfection-imperfection commutator and the interaction-imperfection cross commutator. The imperfection-imperfection commutator is a single-site operator so it has the same first-order average Hamiltonian for both disorder and offset. Then, the difference between disorder and imperfection lies in the interaction-imperfection cross commutator. Indeed, due to symmetries, $[D_\alpha, H_\alpha] = 0$ for the offset (with $\alpha = x, y, z$), while this is not true for disorder. Still, we can show that for *yxx* sequences, additional terms arising from the disorder $[D_\alpha, H_\alpha]$ commutator sum up to zero. The detailed discussion is presented in [49], while here we give two key factors. First, *yxx* sequences can be divided into blocks of 3 intervals of duration τ, and the zeroth-order average interaction within each block is zero. This guarantees that there are no cross terms between different blocks. Second, if the zeroth-order offset vanishes, not only we have $\sum_j H_z^j = 0$ when we sum over all time intervals, but also if we restrict the sum to the first (or 2nd and 3rd) intervals inside each block. In turns, this ensures that when summing over all blocks, commutators of the form $[D_\alpha, H_\alpha]$ add up to zero. Thanks to the similarity between the offset and disorder we were able to use the simpler form of the offset Hamiltonian in the traning algorithm, and still achieve robust sequences against disorder.

Appendix C: Physical intuition for the construction of the yxx24 sequence

Here we explain how we design the yxx24 sequence starting from the Angle12 sequence, as an example of how human insight can lead to better control. First we analyze Angle12 using AHT and present the results in Fig. 9. The zeroth-order angle error vanishes, while the zeroth-order offset is proportional to $4X - 2Y$ (the zeroth-order interaction term is zero as guaranteed by the *yxx* pattern.) We notice that if we consider only the interaction and offset, Angle12 is symmetric, because the second row ("Lab z in tog." denoting the orientation of interaction and offset) in Fig. 9 is mirror-symmetric around the red dashed line. In other words, we can "rotate" the sequence to make it symmetric. Here by "rotating" we mean shifting n actions from the beginning to the end. For example, we can put the first 2 actions of Angle 12 at the end, so that it becomes $-x, y, -x, -x, -y, x, -x, y, x, x, -y, x$, which is symmetric with respect to the middle point. This sequence, labeled symmetrized Angle12 in Fig. 9, is equivalent (in terms of fidelity and leading order average Hamiltonian) to the original one found by DRL, if the angle error is ignored. Indeed, the sequence rotation induces a unitary rotation of the Floquet Hamiltonian that for decoupling sequences (where the target evolution is the identity) does not change the fidelity [64]. As the DRL agent only learns from the propagator fidelity, it does not distinguish symmetric and rotated sequences. This is in contrast to traditional sequence-finding methods that are based on the approximated Floquet-Magnus expansion. Indeed, in the theoretical analysis it is convenient to consider the symmetric case as it reduces the number of nonzero terms in AHT. In particular, the symmetrized Angle 12 sequence has zero first-order average Hamiltonian if the pulses are perfect. The pulse error term, on the other hand, is not symmetric, therefore the sequence has a nonzero first order cross term between pulse angle error and offset, and we can thus focus on such terms [as listed in Fig. 9].

With the AHT analysis in hand, we first want to modify the sequence such that it has vanishing zeroth-order average Hamiltonian. Notice that under Angle12 the offset gives a zeroth-order Hamiltonian $4X - 2Y$. To cancel this contribution we need to change the toggling frame offset Hamiltonian from X to $-X$ in two intervals and from $-Y$ to Y in 1 interval, while keeping the sequence symmetric. The toggling frame offset orientation is shown in the row labeled "Lab z in tog." in Fig. 9. Therefore, we can cancel the zeroth-order offset by adding a π phase shift to pulses within the blue box: we name this sequence "modified Angle12". The AHT analysis of modified Angle12 is also shown in Fig. 9. Its zeroth-order Hamiltonian is zero.

Now the dominant nonzero Hamiltonian comes from the first-order cross terms between angle error and interaction, as well as between angle error and offset (since the angle error is not mirror symmetric.) Once the zeroth-

- Angle12

Pulse in lab	-y	x	-x	y	-x	-x	-y	x	-x	y	x	x
Lab z in tog.	x	y	x	z	-y	-z	x	-y	x	-z	-y	z
Pulse in tog.	-y	-z	z	y	-x	-x	y	z	-z	-y	x	x

- Symmetrized Angle12

	-x	y	-x	-x	-y	x	-x	y	x	x	-y	x
z	-y	-x	-z	x	-y	-z	-y	x	-z	-x	-y	z
	-x	z	y	y	z	-x	x	-z	-y	-y	-z	x

Angle $0^{\text{th}} = 0$
Offset $0^{\text{th}} \propto 4X - 2Y$
Angle x interaction $= 0$
Angle x offset $\propto -X + 2Y + 2Z$

Equivalent

- Modified Angle12

Pulse in lab	-y	x	-x	y	-x	-x	y	-x	x	-y	x	x
Lab z in tog.	x	y	x	z	-y	-z	-x	y	-x	-z	-y	z
Pulse in tog.	-y	-z	z	y	-x	-x	-y	z	-z	y	x	x

Angle $0^{\text{th}} = 0$
Offset $0^{\text{th}} = 0$
Angle x interaction $\propto [Y, D_z]$
Angle x offset $\propto Y$

- yxx24

Pulse in lab	-y	x	-x	y	-x	-x	y	-x	x	-y	x	x	y	-x	x	-y	x	x	-y	x	-x	y	-x	-x
Lab z in tog.	x	y	x	z	-y	-z	-x	y	-x	-z	-y	z	-x	-y	-x	z	y	-z	x	-y	x	-z	-y	z
Pulse in tog.	-y	-z	z	y	-x	-x	-y	z	-z	y	x	x	y	-z	z	-y	x	x	y	z	-z	y	-x	-x

Angle $0^{\text{th}} = 0$
Offset $0^{\text{th}} = 0$
Angle x interaction $= 0$
Angle x offset $= 0$

FIG. 9. Analysis of Angle12, symmetrized Angle12, modified Angle12 and yxx24. For each sequence, we specify three elements: pulse in lab stands for the direction of pulse in lab frame; lab z in tog. stands for the direction of lab frame z axis in toggling frame after the corresponding pulse; pulse in tog. stands for the direction of pulse in toggling frame. Lab z in tog. determines the free evolution Hamiltonian (both interaction and offset) in toggling frame. For example, $-y$ indicates the Hamiltonian is $D_y - H_y$. Pulse in tog. gives the angle error term. For example, $-y$ indicates the pulse unitary is $e^{i(pi/2+\epsilon)Y}$. AHT analysis results of Angle12 (which is equivalent to symmetrized Angle12), modified Angle12 and yxx24 are shown to the left of the tables. Modified Angle12 is obtained by π phase shifting the pulses in the blue box. If we consider only the interaction and the offset, Angle12 and modified Angle12 are symmetric around the red dashed line. yxx24 is obtained from modified Angle12 by appending another modified Angle12 with π phase shift.

order average Hamiltonian is zero, we can double the sequence and use the symmetry to get rid of the first-order corrections. Notice the two first-order corrections are antisymmetric under a π rotation along the z-axis and thus can be easily cancelled by combining the modified Angle12 and another modified Angle12 with a π phase shift. We thus arrive at the yxx24 sequence, whose zeroth- and first-order average Hamiltonian are all zero.

Appendix D: Comparison to other known sequences

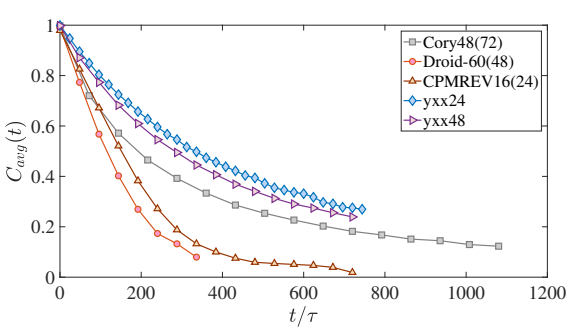

FIG. 10. Experimentally measured average correlation at the best calibrated condition for FAp as a function of time. $\tau = 6\mu$s, other parameters are the same as in Fig. 3.

In this section we compare yxx sequences to two additional decoupling sequences on the FAp sample – DROID-60 discovered in [43, 57] and CPMREV16 discovered in [35]. The back-to-back $\pi/2$ pulse in DROID-60 is implemented with 2.5 μs delay in between, the shortest time required to switch pulse phase in our experimental setup. To accommodate that additional delay we use $\tau = 6\mu$s for all sequences. The original version of CPMREV16 in [35] has a π pulse between two consecutive MREV16 sequences. Here we implement the π pulses as a π phase shift of the last pulse in MREV16, which yields better decoupling than a than a physical π as in the original sequence. The results are shown in Fig. 10. The decoupling ranking according to C_{avg} decay rate is yxx24, yxx48, Cory48, CPMREV, DROID60. There are other decoupling sequences such as Lee-Golburg [66], Magic echo [67] and DUMBO [14] that utilize continuous wave irradiation instead of pulsed controls and thus are not considered here. We note that many of these sequences have been optimized to work together with magical angle spinning, another decoupling technique we are also not considering here.

Appendix E: Experiments on Adamantane

In order to ensure that the improved performance of the yxx24 and yxx48 sequences in calcium fluoride and fluorapatite shown in Figure 6 were not unique to the spectrometer and probe used there, we also compared the

performance of these sequences to the Cory48 sequence in a powdered adamantane sample on a different 300 MHz Bruker DSX spectrometer.

Adamantane ($C_{10}H_{16}$) is a plastic solid with a high degree of internal motion. The proton (hydrogen nuclei) dipolar linewidth is about $(2\pi)13$ krad/s [3] and the system is often used to model a 3D spin system. The T_1 relaxation time for the proton spins in adamantane at room temperature was measured to be just under 1 s.

Figure 11 shows the comparison between the performance of the Cory48, yxx24 and yxx48 as measured by the average correlation metric introduced earlier. The collective $\pi/2$ pulses used had a pulse width $t_w = 2$ μs. The pulse center-to-center delay $\tau = 8\mu$s. Sequence performance degraded significantly as this duration was decreased, likely due to finite stabilization times during phase switching and the overlapping of pulse transients.

It should be noted that while the decay of the average correlation metric resembles the results of a single line-narrowing experiment, care should be taken while comparing them directly. The bi-exponential behavior of the decays for the Cory48 sequence gives rise to effective linewidths of 823 Hz and 88 Hz respectively. The effective linewidth is significantly broader than the 3.5 Hz obtained in [3] probably due to the longer $\pi/2$ pulse and τ values used here. Similarly, the fidelity of the sequences shown here are seen to be slightly lower than those obtained in Figure 6. Note that the plot shows the data acquired after an even number of cycles, with a maximum of 128 cycles.

The yxx24 and Cory48 sequences show almost identical behavior at all timescales in these experiments. However, while the fidelity of the yxx48 sequence is initially lower than that of the other two, the performance at longer timescales matches that of the other two.

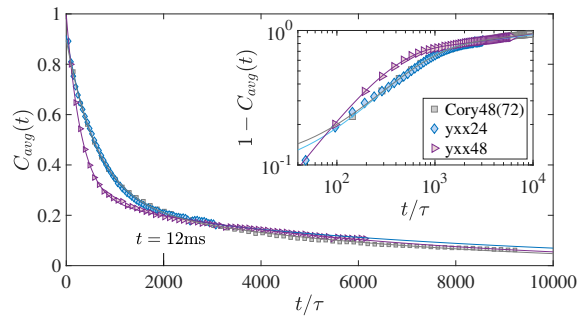

FIG. 11. Experimentally measured average correlation at the best calibrated condition for adamantane as a function of time. The solid curves are biexponential fits to the data. The inset shows the same set of data in a log-log scale, with the y axis being $1 - C_{avg}(t)$.

[1] U. Haeberlen and J.S. Waugh, "Coherent averaging effects in magnetic resonance," Phys. Rev. **175**, 453–467 (1968).

[2] J.S. Waugh, L.M. Huber, and U. Haeberlen, "Approach to high-resolution nmr in solids," Phys. Rev. Lett. **20**, 180 (1968).

[3] D. G. Cory, J. B. Miller, and A. N. Garroway, "Time-suspension multiple-pulse sequences : Applications to solid-state imaging." J. Mag. Res. **90**, 205–123 (1990).

[4] H. M. Cho, C. J. Lee, D. N. Shykind, and D. P. Weitekamp, "Nutation sequences for magnetic resonance imaging in solids," Phys. Rev. Lett. **55**, 1923–1926 (1985).

[5] L. M. K. Vandersypen and I. L. Chuang, "Nmr techniques for quantum control and computation," Rev. Mod. Phys. **76**, 1037–1069 (2005).

[6] R. Tycko and G. Dabbagh, "Measurement of nuclear magnetic dipole‘‘dipole couplings in magic angle spinning nmr," Chemical Physics Letters **173**, 461 – 465 (1990).

[7] G. S. Boutis, P. Cappellaro, H. Cho, C. Ramanathan, and D. G. Cory, "Pulse error compensating symmetric magic-echo trains," J. Mag. Res. **161**, 132–137 (2003).

[8] P. Mansfield, M. J. Orchard, D. C. Stalker, and K. H. B. Richards, "Symmetrized multipulse nuclear-magnetic-resonance experiments in solids: Measurement of the chemical-shift shielding tensor in some compounds," Phys. Rev. B **7**, 90 (1973).

[9] W-K. Rhim, D. D. Elleman, and R. W. Vaughan, "Enhanced resolution for solid state nmr," J. Chem. Phys. **58**, 1772–1773 (1973).

[10] C. Connor, A. Naito, K. Takegoshi, and C.A. McDowell, "Intermolecular spin-diffusion between 31p nuclei in a single crystal of dipotassium α-d-glucose-1-phosphate dihydrate; a 1-d analogue of the 2-d exchange nmr experiment," Chemical Physics Letters **113**, 123 – 128 (1985).

[11] Robert Tycko, "Selection rules for multiple quantum {NMR} excitation in solids: Derivation from time-reversal symmetry and comparison with simulations and 13c {NMR} experiments," J. Mag. Res. **139**, 302 – 307 (1999).

[12] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbuggen, and S. Glaser, "Optimal control of coupled spin dynamics: design of nmr pulse sequences by gradient ascent algorithms," J. Mag. Res. **172**, 296–305 (2005).

[13] Patrick Doria, Tommaso Calarco, and Simone Montangero, "Optimal control technique for many-body quantum dynamics," Phys. Rev. Lett. **106**, 190501 (2011).

[14] Dimitris Sakellariou, Anne Lesage, Paul Hodgkinson, and Lyndon Emsley, "Homonuclear dipolar decoupling in solid-state nmr using continuous phase modulation," Chemical Physics Letters **319**, 253 – 260 (2000).

[15] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore

Graepel, and Demis Hassabis, "Mastering the game of go with deep neural networks and tree search," Nature **529**, 484–489 (2016).

[16] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," Nature **575**, 350–354 (2019).

[17] R.S. Sutton and A.G. Barto, *Reinforcement Learning: An Introduction*, Adaptive Computation and Machine Learning series (MIT Press, 2018).

[18] Y. LeCun, Y. Bengio, and G.H. Hinton, "Deep learning," Nature **521**, 436–444 (2015).

[19] Marin Bukov, Alexandre G. R. Day, Dries Sels, Phillip Weinberg, Anatoli Polkovnikov, and Pankaj Mehta, "Reinforcement learning in different phases of quantum control," Phys. Rev. X **8**, 031086 (2018).

[20] Xiao-Ming Zhang, Zezhu Wei, Raza Asad, Xu-Chen Yang, and Xin Wang, "When does reinforcement learning stand out in quantum control? a comparative study on state preparation," npj Quantum Information **5**, 85 (2019).

[21] C. Chen, D. Dong, H. Li, J. Chu, and T. Tarn, "Fidelity-based probabilistic q-learning for control of quantum systems," IEEE Transactions on Neural Networks and Learning Systems **25**, 920–933 (2014).

[22] Jun-Jie Chen and Ming Xue, "Manipulation of spin dynamics by deep reinforcement learning agent," (2019), arXiv:1901.08748 [quant-ph].

[23] F. Albarrán-Arriagada, J. C. Retamal, E. Solano, and L. Lamata, "Measurement-based adaptation protocol with quantum reinforcement learning," Phys. Rev. A **98**, 042315 (2018).

[24] Jelena Mackeprang, Durga B. Rao Dasari, and Jörg Wrachtrup, "A reinforcement learning approach for quantum state engineering," Quantum Machine Intelligence **2**, 1–14 (2020).

[25] Murphy Yuezhen Niu, Sergio Boixo, Vadim N. Smelyanskiy, and Hartmut Neven, "Universal quantum control through deep reinforcement learning," npj Quantum Information **5**, 33 (2019).

[26] Mogens Dalgaard, Felix Motzoi, Jens Jakob Sørensen, and Jacob Sherson, "Global optimization of quantum dynamics with alphazero deep exploration," npj Quantum Information **6**, 6 (2020).

[27] S. Daraeizadeh, S. P. Premaratne, N. Khammassi, X. Song, M. Perkowski, and A. Y. Matsuura, "Machine-learning-based three-qubit gate design for the toffoli gate and parity check in transmon systems," Phys. Rev. A **102**, 012601 (2020).

[28] Julius Wallnöfer, Alexey A. Melnikov, Wolfgang Dür, and Hans J. Briegel, "Machine learning for long-distance quantum communication," PRX Quantum **1**, 010301 (2020).

[29] Moritz August and Xiaotong Ni, "Using recurrent neural networks to optimize dynamical decoupling for quantum memory," Phys. Rev. A **95**, 012335 (2017).

[30] Thomas Fösel, Petru Tighineanu, Talitha Weiss, and Florian Marquardt, "Reinforcement learning with neural networks for quantum feedback," Phys. Rev. X **8**, 031084 (2018).

[31] Hendrik Poulsen Nautrup, Nicolas Delfosse, Vedran Dunjko, Hans J. Briegel, and Nicolai Friis, "Optimizing Quantum Error Correction Codes with Reinforcement Learning," Quantum **3**, 215 (2019).

[32] Ryan Sweke, Markus S Kesselring, Evert P L van Nieuwenburg, and Jens Eisert, "Reinforcement learning decoders for fault-tolerant quantum computation," Machine Learning: Science and Technology **2**, 025005 (2021).

[33] Xiao-Ming Zhang, Zi-Wei Cui, Xin Wang, and Man-Hong Yung, "Automatic spin-chain learning to explore the quantum speed limit," Phys. Rev. A **97**, 052333 (2018).

[34] Jonas Schuff, Lukas J Fiderer, and Daniel Braun, "Improving the dynamics of quantum sensors with reinforcement learning," New Journal of Physics **22**, 035001 (2020).

[35] T. D. Ladd, D. Maryenko, Y. Yamamoto, E. Abe, and K. M. Itoh, "Coherence time of decoupled nuclear spins in silicon," Phys. Rev. B **71**, 014401 (2005).

[36] Ken Xuan Wei, Chandrasekhar Ramanathan, and Paola Cappellaro, "Exploring localization in nuclear spin chains," Phys. Rev. Lett. **120**, 070501 (2018).

[37] K. X. Wei, P. Peng, O. Shtanko, I. Marvian, S. Lloyd, C. Ramanathan, and P. Cappellaro, "Emergent prethermalization signatures in out-of-time ordered correlations," Phys. Rev. Lett. **123**, 090605 (2019).

[38] Pai Peng, Chao Yin, Xiaoyang Huang, Chandrasekhar Ramanathan, and Paola Cappellaro, "Floquet prethermalization in dipolar spin chains," Nature Physics (2021).

[39] Katharine W. Moore and Herschel Rabitz, "Exploring constrained quantum control landscapes," J. Chem. Phys. **137**, 134113 (2012).

[40] Felipe Petroski Such, Vashisht Madhavan, Edoardo Conti, Joel Lehman, Kenneth O. Stanley, and Jeff Clune, "Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning," (2018), arXiv:1712.06567 [cs.NE].

[41] P Mansfield, "Symmetrized pulse sequences in high resolution nmr in solids," J. Phys. C **4**, 1444–1452 (1971).

[42] U. Haeberlen, *High Resolution NMR in Solids: Selective Averaging* (Academic Press Inc., New York, 1976).

[43] Hengyun Zhou, Joonhee Choi, Soonwon Choi, Renate Landig, Alexander M. Douglas, Junichi Isoya, Fedor Jelezko, Shinobu Onoda, Hitoshi Sumiya, Paola Cappellaro, Helena S. Knowles, Hongkun Park, and Mikhail D. Lukin, "Quantum metrology with strongly interacting spin systems," Phys. Rev. X **10**, 031003 (2020).

[44] Reward function is only used to order the agents in the same generation thus the specific functional form does not play an important role.

[45] Chao Yin, Pai Peng, Xiaoyang Huang, Chandrasekhar Ramanathan, and Paola Cappellaro, "Prethermal quasiconserved observables in floquet quantum systems," Phys. Rev. B **103**, 054305 (2021).

[46] Markus Heyl, Philipp Hauke, and Peter Zoller, "Quantum localization bounds trotter errors in digital quantum simulation," Science Advances **5** (2019), 10.1126/sciadv.aau8342.

[47] V. Nair and G.H. Hinton, "Rectified linear units improve restricted boltzmann machines," (2010) pp. 807–814.

[48] We also varied the number of sequences and found it does not affect the performance of RL.

[49] See supplementary online material for details of the experimental systems, hyperparameters and additional re-

sults, which includes Refs. [38, 65, 68, 69].

[50] W. Van der Lugt and W.J. Caspers, "Nuclear magnetic resonance line shape of fluorine in apatite," Physica **30**, 1658–1666 (1964).

[51] The Floquet-Magnus expansion does not converge in a many-body quantum system, and thus one has to truncate the series in Eq. 4 and leave a small time-dependent and/or non-local Hamiltonian [70]. Effects of the truncation and the small time-dependent Hamiltonian are only evident at very long time scale, therefore ignored in this paper.

[52] Wilhelm Magnus, "On the exponential solution of differential equations for a linear operator," Communications on Pure and Applied Mathematics **7**, 649–673 (1954).

[53] S. Blanes, F. Casas, J.A. Oteo, and J. Ros, "The magnus expansion and some of its applications," Physics Reports **470**, 151 – 238 (2009).

[54] D. P. Burum and W. K. Rhim, "Analysis of multiple pulse nmr in solids. iii," J. Chem. Phys. **71**, 944–956 (1979).

[55] D.P Burum, M. Linder, and R. R. Ernst, "Low-power multipulse line narrowing in solid-state nmr," J. Mag. Res. **44**, 173 – 188 (1981).

[56] AM Waeber, G Gillard, G Ragunathan, M Hopkinson, P Spencer, DA Ritchie, MS Skolnick, and EA Chekhovich, "Pulse control protocols for preserving coherence in dipolar-coupled nuclear spin baths," Nat. Comm. **10**, 3157 (2019).

[57] Joonhee Choi, Hengyun Zhou, Helena S. Knowles, Renate Landig, Soonwon Choi, and Mikhail D. Lukin, "Robust dynamic hamiltonian engineering of many-body spin systems," Phys. Rev. X **10**, 031002 (2020).

[58] Jonathan Baugh, Osama Moussa, Colm A Ryan, Ashwin Nayak, and Raymond Laflamme, "Experimental implementation of heat-bath algorithmic cooling using solid-state nuclear magnetic resonance," Nature **438**, 470 (2005).

[59] Hyung Joon Cho, Thaddeus D. Ladd, Jonathan Baugh, David G. Cory, and Chandrasekhar Ramanathan, "Multispin dynamics of the solid-state nmr free induction decay," Phys. Rev. B **72**, 054427 (2005).

[60] Hyung Joon Cho, Paola Cappellaro, David G. Cory, and Chandrasekhar Ramanathan, "Decay of highly correlated spin states in a dipolar-coupled solid: Nmr study of caf2," Phys. Rev. B **74**, 224434 (2006).

[61] Steven W Morgan, Vadim Oganesyan, and Gregory S Boutis, "Multispin correlations and pseudothermalization of the transient density matrix in solid-state nmr: Free induction decay and magic echo," Phys. Rev. B **86**, 214410 (2012).

[62] Joseph Emerson, Marcus Silva, Osama Moussa, Colm Ryan, Martin Laforest, Jonathan Baugh, David G. Cory, and Raymond Laflamme, "Symmetrized characterization of noisy quantum processes," Science **317**, 1893–1896 (2007).

[63] Michael K. Henry, Chandrasekhar Ramanathan, Jonathan S. Hodges, Colm A. Ryan, Michael J. Ditty, Raymond Laflamme, and David G. Cory, "Fidelity enhancement by logical qubit encoding," Phys. Rev. Lett. **99**, 220501 (2007).

[64] Marin Bukov, Luca D'Alessio, and Anatoli Polkovnikov, "Universal high-frequency behavior of periodically driven systems: from dynamical stabilization to floquet engineering," Advances in Physics **64**, 139–226 (2015).

[65] J. G. Powles and P. Mansfield, "Double-pulse nuclear-resonance transients in solids," Physics Letters **2**, 58–59 (1962).

[66] Moses Lee and Walter I. Goldburg, "Nuclear-magnetic-resonance line narrowing by a rotating rf field," Phys. Rev. **140**, A1261–A1271 (1965).

[67] K. Takegoshi and C.A. McDowell, "A ''magic echo'' pulse sequence for the high-resolution nmr spectra of abundant spins in solids," Chemical Physics Letters **116**, 100–104 (1985).

[68] G.W Canters and C.S Johnson, "Numerical evaluation of moments and shapes of magnetic resonance lines for crystals and powders," Journal of Magnetic Resonance (1969) **6**, 1 (1972).

[69] A. Abragam, *Principles of Nuclear Magnetism* (Oxford Univ. Press, 1961).

[70] Dmitry A. Abanin, Wojciech De Roeck, and F. Huveneers, "Exponentially slow heating in periodically driven many-body systems," Phys. Rev. Lett. **115**, 256803 (2015).