Re-exploring control strategies in a non-Markovian open quantum system by reinforcement learning

Amine Jaouadi [1], Etienne Mangaud [2], and Michèle Desouter-Lecomte [3,*]

[1] *LyRIDS, ECE-Paris, Graduate School of Engineering, 75015 Paris, France*

[2] *MSME, Université Gustave Eiffel, UPEC, CNRS, 77454 Marne-La-Vallée, France*

[3] *Institut de Chimie Physique, Université Paris-Saclay-CNRS, UMR8000, 91400 Orsay, France*

(Received 17 October 2023; accepted 14 December 2023; published 16 January 2024)

In this study, we reexamine a recent optimal control simulation targeting the preparation of a superposition of two excited electronic states in the ultraviolet (UV) range in a complex molecular system. We revisit this control from the perspective of reinforcement learning, offering an efficient alternative to conventional quantum control methods. The two excited states are addressable by orthogonal polarizations and their superposition corresponds to a right or left localization of the electronic density. The pulse duration spans tens of femtoseconds to prevent excitation of higher excited bright states which leads to a strong perturbation by the nuclear motions. We modify an open source software by Giannelli *et al.* [L. Giannelli *et al.*, Phys. Lett. A **434**, 128054 (2022)] that implements reinforcement learning with Lindblad dynamics, to introduce non-Markovianity of the surrounding reservoir either by time-dependent rates, or more exactly, by using the hierarchical equations of motion with the QuTip-BoFiN package. This extension opens the way to wider applications for non-Markovian environments, in particular when the active system interacts with a highly structured noise.

DOI: 10.1103/PhysRevA.109.013104

I. INTRODUCTION

For several decades, quantum-state manipulation with electromagnetic fields has been a central problem in many areas of physics and chemistry to prepare particular initial states or realizing unitary transformations (quantum gates). The hardware systems and the spectral range are very different, from spins in nuclear magnetic resonance (NMR) [1] to molecular systems [2] or complex photosynthetic systems [3,4] and systems involved in emerging quantum technology [5,6], for instance, a superconducting quantum interference device (SQUID) [7], trapped ions [8] or atoms [9], nitrogen-vacancy diamond centers [10], or quantum dots [11]. Quantum control has developed through various theoretical strategies as pump-dump schemes [12,13], coherent control [14], adiabatic methods [15,16], local [17,18] or Lyapunov control [19], Pontryagin optimal control [20], adaptative tracking [21], and optimal control theory (OCT) [22] that involves a rich variety of optimization algorithms, for instance, the Rabitz monotonous convergent algorithm [23,24] or Krotov method [25,26], gradient ascent pulse engineering (GRAPE) [27,28], or the chopped random basis optimization (CRAB) [29,30].

Recently, there has been significant interest in applying reinforcement learning (RL), a distinctive machine learning technology, to quantum control. This begs the following questions: Does RL suggest new control strategies? How does it compare with standard algorithms [31]? What is the efficiency for control in a dissipative environment [32]? RL has already been applied in control for state preparation [33] and gate realization [34,35] and quantum compiling [36] in quantum technology. Recently RL has recovered the well-known counterintuitive stimulated raman adiabatic passage (STIRAP) pulse sequence in a three-state system [15,16]. In Ref. [37] the laser may be on or of,f leading to a so-called digital-STIRAP ensuring an efficient transfer without the constraint of adiabaticity conditions [38]. Conversely, in Refs. [39,40], the pulses are continuous, while the RL algorithm optimizes either the laser detuning or the Rabi frequencies.

In this work, our objective is to revisit with RL an optimal control simulation recently performed in a molecular system (phenylene ethynylene dimer) with C_{2v} symmetry [41]. This benchmark case is interesting for different reasons. The system consists of two quasi-degenerate excited electronic states of different symmetries. They are addressable by orthogonal polarizations. The preparation of a superposition of the two states corresponds to a right or left localization of the electronic density in a way similar to the localization in a double-well by superposing the two lowest eigenstates. Creating such a coherence in complex systems with orthogonal polarizations has been recently discussed as a prospective avenue for achieving coherent control over excitonic energy transfer [42]. In the absence of dissipation, this is a V-type three-state system where two excited states are coupled to the ground state by two orthogonal transition dipoles. An analytical solution may be derived to prepare the superposition with equal weights [43]. This is an important landmark to test the control. On the other hand, the ideal electronic V-type system strongly interacts with the surrounding leading to a non-Markovian nonperturbative open system. The nuclear vibrations form two baths called the tuning and the coupling baths making the energy gaps fluctuate (also called longitudinal noise) and the electronic coupling (transversal noise), respectively.

*michele.desouter-lecomte@universite-paris-saclay.fr

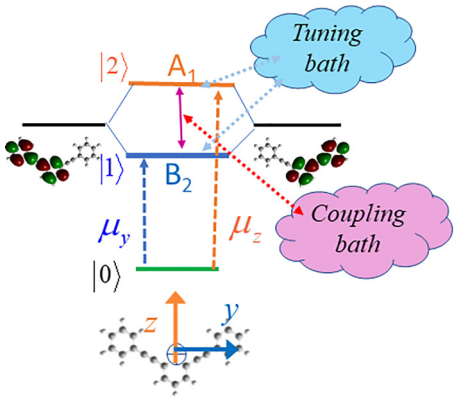

FIG. 1. Schematic representation of the V-type model in 1,3-bis(phenylethynyl)benzene. The two excited states of symmetries A_1 and B_2 at the geometry of the ground state are delocalized over the two sites. Their superposition with equal weights corresponds to a localization on the left or right sites. The excited states are addressable by orthogonal dipole moments. The electronic subsystem is coupled to the vibrational baths. The tuning bath (longitudinal noise) making the energy gap fluctuate gathers the symmetric A_1 modes and the coupling bath (transversal noise) varying the electronic coupling contains the B_2 modes.

In our previous OCT simulation, we employed hierarchical equations of motion (HEOM), which represent an exact method for addressing nonperturbative and non-Markovian open systems with Gaussian statistics [44–49]. To investigate the RL control, we start from the open source software [50] presented in Ref. [40]. This software uses the libraries QuTip [51] and TensorFlow [52]. The software already implements Lindblad dynamics with the QuTip collapse operators. In this work, we introduce non-Markovianity in different ways. In a simplified strategy, we first consider time-dependent rates [53,54] by using time-dependent QuTip collapse operators. The rates are calibrated from the decoherence matrix [54] extracted from the exact HEOM dynamics [55,56]. We then address the exact non-Markovian dynamics with the QuTip HEOMSolver [57].

The paper is organized as follows. In Sec. II we describe the model treated as an isolated or an open system interacting with two Bosonic baths. We summarize the Lindblad and HEOM operational equations in Sec. III. The control by RL or OCT is presented in Sec. IV. The RL results are given in Sec. V and a comparison with OCT is made in Sec. VI before concluding in Sec. VII.

II. MODEL

The V-type three-state model of the dimer [1,3-bis(phenylethynyl)benzene] is schematized in Fig. 1. It is calibrated from *ab initio* data computed by Lasorne *et al.* [41,58–60] with the density-functional theory (DFT) for the ground state S_0 and the time-dependent density-functional theory (TDFT) for the two excited states [$S_1(B_2)$ and $S_2(A_1)$]. The energies of the two excited states ($E_{S_1} = 4.43$ eV and $E_{S_2} = 4.47$ eV) are taken at the equilibrium geometry of the ground state (planar with C_{2v} symmetry). The two states are bright and coupled to the ground state by orthogonal transition dipoles. The axes are chosen so that $z(A_1)$ is the C_2 rotation axis, $y(B_2)$ lies within the molecular plane, and $x(B_1)$ is orthogonal to it. The respective transition dipoles are $\vec{\mu}_{01} = (0, \mu_y, 0)$ with $\mu_y = 3.96ea_0$ and $\vec{\mu}_{02} = (0, 0, \mu_z)$ with $\mu_z = -1.83ea_0$.

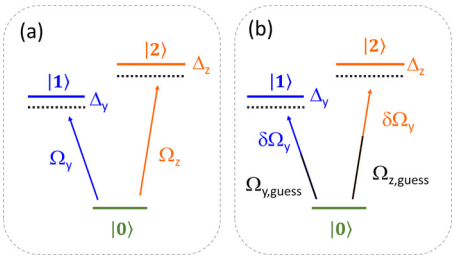

FIG. 2. Actions optimized by the RL algorithm to prepare the target superposed state in the V-type system. (a) The actions are the Rabi frequencies Ω_y and Ω_z; (b) the actions are the variations of the Rabi frequencies with respect to guess fields. The laser detuning Δ_y and Δ_z are assumed to vanish in this application.

A. Isolated system

The system is defined by selecting the ground and the first two excited electronic states at the equilibrium geometry of the ground state (vertical Franck-Condon transition). It is an ideal system assumed to be frozen at this geometry. The two excited states are coupled by nonadiabatic interactions via a conical intersection [59,60]. We choose a diabatic representation with states of A_1 or B_2 symmetry so that the electronic coupling vanishes at that reference position and the system Hamiltonian is simply

$$H_S^0 = \sum_{j=0}^{2} |j\rangle\langle j|. \tag{1}$$

The electronic coupling (between the two diabatic excited states) becomes different from zero when any vibration of B_2 symmetry is active. The two bright states are coupled to the ground state only radiatively and the time-dependent Hamiltonian at the dipolar approximation is

$$H_S(t) = H_S^0 - \sum_{j=1}^{2} \left(\vec{\mu}_{0j}\vec{\mathcal{E}}_j(t)|0\rangle\langle j| + \text{H.c.}\right), \tag{2}$$

where H.c. designates the Hermitian conjugate. The two fields are linearly polarized with $\vec{\mathcal{E}}_1 = (0, \mathcal{E}_y, 0)$ and $\vec{\mathcal{E}}_2 = (0, 0, \mathcal{E}_z)$. In interaction representation (I) and with the rotating wave approximation (RWA) [61] the Hamiltonian becomes

$$\mathbf{H}_{S,I}^{\text{RWA}}(t) = -\frac{\hbar}{2}\begin{pmatrix} 0 & \Omega_y(t) & \Omega_z(t) \\ \Omega_y(t) & -2\Delta_y & 0 \\ \Omega_z(t) & 0 & -2\Delta_z \end{pmatrix}, \tag{3}$$

where the Rabi frequencies are $\Omega_y(t) = \mu_y E_y(t)/\hbar$, $\Omega_z(t) = \mu_z E_z(t)/\hbar$, and $E_j(t)$ ($j = y, z$) are the pulse envelopes. Δ_y and Δ_z are the field detunings.

The two Rabi frequencies or their variations with respect to a guess field are the actions that will be optimized by the RL algorithm. They are represented in Fig. 2. In our application, the target is the superposition of the two excited states with

equal weights

$$|0\rangle \rightarrow \frac{1}{\sqrt{2}}(|1\rangle + |2\rangle). \tag{4}$$

This target is different from the superposition of the initial state and one excited state that may be prepared by fractional-STIRAP [15] or by a $\pi/2$ pulse.

By imposing equal Rabi frequencies at all times, i.e., pulses of the same duration T with amplitudes in the inverse ratio of the dipole moments, the target transition is realized if each area is equal to $\pi/\sqrt{2}$ [43]

$$\int_0^T \Omega_j(t)dt = \pi/\sqrt{2}, \tag{5}$$

with $j = y, z$. It is a generalization of the well-known π rule for complete population transfer [62] or $\pi/2$ for creating a superposition involving the initial state.

B. Open quantum system

According to the chosen partition, all the nuclear vibrations belong to the baths. The A_1 modes make the energies fluctuate and the B_2 vibrations modify the electronic coupling that becomes different from zero when the C_{2v} symmetry is broken. The generic Hamiltonian of the system-bath partition is then written as

$$H(t) = H_S(t) + H_{SB} + H_B, \tag{6}$$

where H_B is the ensemble of the vibrational modes assumed to be harmonic. The normal modes are assumed to be the same in each electronic states but their equilibrium positions differ. H_{SB} is the linear system-bath coupling. The two groups of A_1 or B_2 modes then constitute different baths that may be called the tuning baths coupled to the diagonal elements $|1\rangle\langle 1|$ and $|2\rangle\langle 2|$ of the system Hamiltonian and the coupling bath coupled to the off-diagonal elements $|1\rangle\langle 2|$ and $|2\rangle\langle 1|$ between the two excited states. This kind of partition in the case of a conical intersection has been discussed in different works applying HEOM dynamics [41,43,63,64]. The partition leads to a strong system-bath coupling and a non-Markovian master equation. The analysis of the dimer model is given in our previous work [41] where it was explained how we obtained the continuous spectral densities $J(\omega)$ for the tuning and coupling baths from *ab initio* data, i.e., from the energy gradients and gradient of the electronic coupling at the reference position. The spectral densities give the strength of the coupling to the system for each energy range of the baths. We select the main part of the spectral densities consisting in very sharp peaks around 1700 and 2300 cm^{-1}. The two spectral densities $J_{\mathrm{tuning}}(\omega)$ and $J_{\mathrm{coupling}}(\omega)$ are presented in Fig. 3(a). Figure 3(b) gives the corresponding correlation functions of the collective mode of each bath

$$C(t) = \frac{1}{\pi}\int_{-\infty}^{+\infty} d\omega \frac{J(\omega)e^{i\omega(t)}}{e^{\beta\omega} - 1}, \tag{7}$$

where β is the Boltzmann constant. Due to the peaked shape of the spectral densities, the effective collective modes are underdamped and decay in about 200 fs. We zoom in on the early time range of 20 fs, which is the pulse duration chosen in our simulations. However, the pulse duration must be longer than about 10 fs to correspond to a sufficiently narrow spectral range to avoid the excitation of higher bright states. It is noticeable that the collective bath modes undergo a full oscillation within the 20-fs timescale. This serves as an indicator of non-Markovian behavior, as we will delve further into in the subsequent discussion.

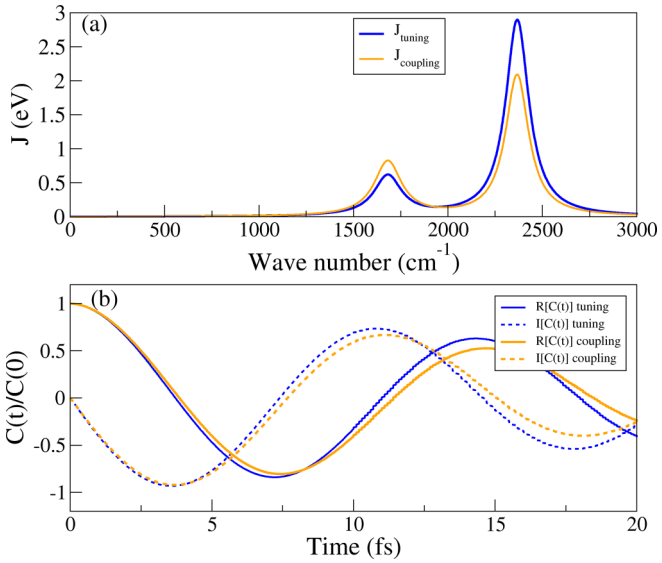

FIG. 3. (a) Spectral densities giving the strength of the system-bath coupling to the diagonal (J_{tuning}) or off-diagonal ($J_{\mathrm{coupling}} = J_{S_1S_2}$) element of the system Hamiltonian block related to the two excited states. J_{tuning} is J_{S_1} and J_{S_2} is assumed to be similar with $J_{S_2} = 0.797J_{S_1}$ [41]. The tuning modes are of A_1 symmetry and the coupling ones of B_2 symmetry. (b) Real and imaginary parts of the normalized bath correlation function $C(t)$ [Eq. (7)] at room temperature (zoomed in on the early 20-fs timescale). $C(t)$ decays in 200 fs.

III. DYNAMICAL METHODS

The dynamics of open quantum systems [65] has been reviewed, as seen in Refs. [66,67] specifically focusing on control aspects. The applications of RL control [32,40] usually assume a weak coupling and a Markovian bath treated by a Lindblad master equation [68,69]. We first summarize the main relations to introduce time-dependent rates and then we recall the operational equations for HEOM.

A. Lindblad master equation

For a N-dimensional system coupled to M dissipative channels, the generic Lindblad operator reads

$$\mathcal{D}(t) = \sum_{k,l}^{M} \Gamma_{kl}\left(L_k\rho_t L_l^\dagger - \frac{1}{2}\{L_l^\dagger L_k, \rho_t\}\right), \tag{8}$$

where $\{A, B\} = AB + BA$ denotes the anticommutator and the rates Γ_{kl} are constant. Often, only the main dissipative processes are retained, as the radiative decay towards a sink [40] or dephasing processes affecting the diagonal elements of the density matrix or relaxation inducing population transfer due to interstate coupling [32]. In this work, we will

consider these two processes induced by the tuning and coupling baths, respectively, leading to four operators $L_1 = |1\rangle\langle 1|$, $L_2 = |2\rangle\langle 2|$, $L_3 = |1\rangle\langle 2|$, and $L_4 = |2\rangle\langle 1|$ (they are the collapse operators in QuTip [51]). Moreover, we want to account for non-Markovian baths, at least on an approximate way by introducing time-dependent rates [53,54]. This has given rise to many fundamental analysis [54,56,70,71] concerning the non-Markovianity signature or the positivity of the dynamical map [72]. Indeed, the master equation of non-Markovian dynamics can be recast in a Lindblad form with time-dependent rates that may be transitorily negative. If the Lindblad dissipator is expressed with the orthogonal basis set of N^2 operators formed by the normalized identity $G_0 = I/\sqrt{N}$ and the $N^2 - 1$ generators of SU(N), $G_i (i = 1, \ldots, N^2 - 1)$ [73,74], which are the Pauli matrices for $N = 2$ and the Gell-Mann matrices for $N = 3$, the corresponding rate matrix is also called the time-dependent decoherence matrix D_{jk} [54]

$$\mathcal{D}(t) = \sum_{j,k=1}^{N^2-1} D_{jk}(t)\left(G_j \rho_t G_k - \frac{1}{2}\{G_k G_j, \rho_t\}\right). \tag{9}$$

The eigenvalues are the canonical decay rate Γ_k^c associated to the time-dependent decay channels. The decoherence matrix is given by [54]

$$D_{ij}(t) = \sum_{m=1}^{N^2-1} Tr[G_m G_i \Lambda_t[G_m(t)] G_j], \tag{10}$$

where $\Lambda_t[.]$ denotes the map of the time local non-Markovian master equation $\dot{G}_m(t) = \Lambda_t[G_m(t)]$ [54,75]. This requires $(N^2 - 1)$ propagations of the basis operators performed here with HEOM as shown in Refs. [55,56]. Some elements of the decoherence matrix will be used to calibrate the time-dependent rates of the selected collapse operators.

This is a low-cost way to introduce easily some non-Markovianity with time-dependent collapse operators. However, its efficiency might be somewhat limited, as the rates are calibrated based on field-free dynamics and are subject to potential modification by the applied fields [56].

B. HEOM

We now summarize the main operational equations of the HEOM method. The system density matrix is the partial trace of the full density matrix $\rho_{\text{tot}}(t)$ over the bath degrees of freedom $\rho(t) = \text{Tr}_B[\rho_{\text{tot}}(t)]$. The initial condition is assumed to be factorized $\rho_{\text{tot}}(0) = \rho(0)\rho_{\text{eq}}$ where ρ_{eq} is the density matrix of the baths at Boltzmann equilibrium at a given temperature. The HEOM may be considered as a numerically exact method for non-Markovian dynamics with harmonic baths when convergence is achieved by a relevant truncation of the hierarchy. The method is abundantly described in the literature, see, for instance, Ref. [44] for a recent review or Ref. [49] for a pedagogical survey, Refs. [45–48] for applications with the tensor-train format, and Ref. [57] for a review of different softwares, in particular that implemented in QuTip. We briefly recall that the master equation is solved by a time local system of coupled equations among auxiliary density matrices or auxiliary density operators (ADOs) arranged in a hierarchical structure. The algorithm requires a particular fit of the correlation function $C(t)$ as a sum of K damped oscillatory terms also called artificial decaying modes

$$C(t) = \sum_{k=1}^{K} \alpha_k e^{i\gamma_k t} \tag{11}$$

and $C^*(t) = \sum_{k=1}^{K} \tilde{\alpha}_k e^{i\gamma_k t}$. Analytical expressions for the α_k, $\tilde{\alpha}_k$, and γ_k parameters can be derived from Eq. (7) [76] when the spectral density is fitted by a sum of two-poles Tannor-Meier Lorentzian functions [77]

$$J(\omega) = \sum_{l=1}^{n_l} \frac{p_l \omega}{\left[(\omega + \Omega_l)^2 + \Gamma_l^2\right]\left[(\omega - \Omega_l)^2 + \Gamma_l^2\right]}. \tag{12}$$

The parameters fitting the spectral densities of Fig. 3 are given in Ref. [41]. Each ADO is labeled by a collective index $\mathbf{n} = \{n_1, \ldots, n_K\}$ specifying the occupation number of each artificial mode. The system density matrix has the index $\mathbf{n} = \{0, \ldots, 0\}$. The HEOM equations are

$$\begin{aligned}\dot{\rho}_{\mathbf{n}}(t) = {} & L_S(t)\rho_{\mathbf{n}}(t) + i\sum_{k=1}^{K} n_k \gamma_k \rho_{\mathbf{n}}(t) - i\left[S, \sum_{k=1}^{K} \rho_{\mathbf{n}_k^+}(t)\right] \\ & - i\sum_{k=1}^{K} n_k (\alpha_k S \rho_{\mathbf{n}_k^-}(t) - \tilde{\alpha}_k \rho_{\mathbf{n}_k^-}(t) S),\end{aligned} \tag{13}$$

where $L_S(t)$ is the system Liouvillian and $\mathbf{n}_k^+ = \{n_1, \ldots, n_k + 1, \ldots, n_K\}$, and $\mathbf{n}_k^- = \{n_1, \ldots, n_k - 1, \ldots, n_K\}$.

The OCT simulations make use of the HEOM [Eqs. (13)] via our in-house developed software [41]. In RL simulations [40] based on QuTip software, we use the QuTip-BOFiN HEOMSOLVER [57] that allows the description of the Bosonic baths by giving the real and imaginary parts of the correlation function $C(t) = C_R(t) + iC_I(t)$. For each bath, they are parametrized by a combination of decaying terms $C_R(t) = \sum_{k=1}^{N_R} c_k^R e^{-\gamma_k^R t}$ and $C_I(t) = \sum_{k=1}^{N_I} c_k^I e^{-\gamma_k^I t}$ where the c_k and $\gamma_k^{R,I}$ are complex. This is an alternative to the expansion of Eq. (11) already adopted for the second-order time nonlocal [77] or time-local non-Markovian equations [78]. The HEOM equations adapted to this partition of the correlation function in real and imaginary part are given in Eq. (11) of Ref. [57]. The analytical expressions of the c_k and $\gamma_k^{R,I}$ parameters when the spectral density is fitted by the two-pole Lorentzian functions [Eq. (12)] are given in Refs. [77,78].

IV. CONTROL

A. Reinforcement learning

The RL algorithm is summarized in many references, for instance Refs. [79,80]. By using the generic vocabulary, the principle is as follows. A target must be reached in an environment. At each time, an agent makes an observation and gets information about its state. The agent then chooses an action according to a policy to modify the state. The agent obtains a reward that estimates the progress towards the target. In our application, we have thus to define the environment, the agent, the action, and the reward. The four points are represented in Fig. 4. The RL environment is the active system and its surrounding, i.e., the V-three-level system coupled to both tuning

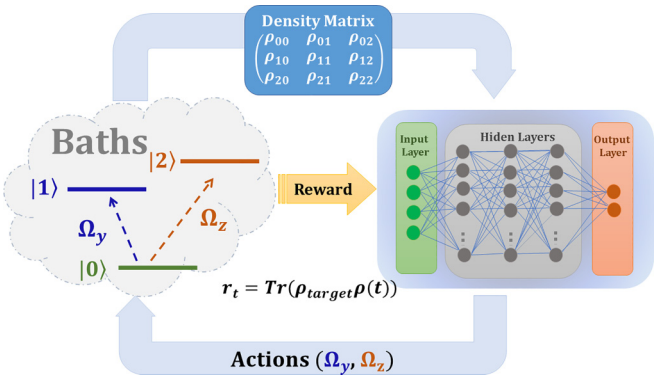

FIG. 4. Representaion of a cycle of the RL. At each time, the reduced density matrix of the three-state system coupled to its surrounding is the input of the neural network. The policy $\pi(a_t|s_t)$ is optimized and provides two actions that are the pulse Rabi frequencies. The reward is the fidelity to reach the target, which is here the superposed state of the two excited states.

and coupling baths. The observation is the state of the system described by the reduced density matrix solution of a master equation. The agent is an algorithm called REINFORCE [79]. It uses a neural network with three hidden layers in our application. The input layer contains all the density matrix elements (nine in our case). The output layer may provide discrete values or a continuous distribution. The outputs are discrete when only some actions are available, for instance, if the laser may be only on or off, giving four outcomes in the two-pulse case [38]. When the distribution is continuous, the output layer gives parameters of this distribution, for instance, the mean of the Gaussian distributions for the actions $a_{y,z}$ that provide the Rabi frequencies Ω_y and Ω_z of the two pulses. The reward is the control fidelity $r_t = \mathrm{Tr}[\rho^{\dagger}_{\mathrm{target}}\rho(t)]$. The policy is the conditional probability $\pi(a_t|s_t)$ that the agent takes action a when the system state is s. The action at a time t only depends on the state at that time and the process is called a Markov decision chain. Note that this does not mean that the dynamics of the system must be Markovian. The Markov decision chain means that when two states s and s' observed by the agent are the same, the probability to choose a is the same regardless of the history to reach the state s.

We employ the policy gradient method [81], which is a technique used in reinforcement learning. It involves adjusting the policy's parameters aiming to maximize the cumulative reward over time. All the parameters of the network are represented by the global index θ. They are initialized at random. One then generates a batch of M episodes with the current policy $\pi(a_t|s_t)$. An episode or trajectory τ is divided in N time steps and lasts $T = N\delta t$. For each episode, one collects the N data triplets (s_t, a_t, r_t) where $t = i\delta t$ for the ith time step. The performance of the agent is estimated by the so-called return R that depends on the network parameters θ and is the main tool to optimize the policy. For each episode, the return $R(\tau)$ may be defined in more or less sophisticated ways by the simple sum of all the rewards r_t or a weighted sum of these with a discount rate [32,33,82]. Here, $R(\tau)$ is the sum of the r_t and all the $r_t = 0$ if $t < T$ so that $R(\tau) = r_N$, i.e., it is given by the final control fidelity. For the bunch of M episodes driven by the same policy, the return is the expectation value

$$E[R] = \sum_{\tau=1}^{M} p_\theta(\tau)R(\tau), \tag{14}$$

where $p_\theta(\tau)$ is the probability of the driving trajectory τ. Following Ref. [83] we summarize the main points of the policy optimization. The probability of each trajectory is different since the actions are taken at random in the current policy. It is a product for each time step of the probability $p(s_{t+1}|a_t, s_t)$ to have a transition from state s_t to state s_{t+1} induced by action a_t times the probability for the agent to choose action a_t for state s_t,

$$p_\theta(\tau) = \prod_{t}^{N} p(s_{t+1}|a_t, s_t)\pi_\theta(a_t, s_t). \tag{15}$$

$p(s_{t+1}|a_t, s_t)$ does not depend on the parameters θ but only on the system dynamics. Therefore, the gradient of the average return involves only the gradient of the policy

$$\nabla_\theta p_\theta(\tau) = \sum_{t=1}^{N}\prod_{t'=1}^{N} p(s_{t'+1}|a_{t'}, s_{t'})\pi_\theta(a_{t'}, s_{t'}) \times \nabla_\theta \ln \pi_\theta(a_t, s_t) = p_\theta(\tau)\sum_{t=1}^{N}\nabla_\theta \ln \pi_\theta(a_t, s_t). \tag{16}$$

The network parameters are optimized so that their gradient $\nabla\theta$ is parallel to the gradient of the average return with a factor η called the learning rate $\nabla\theta = +\eta\nabla_\theta E[R]$ and by Eqs. (14) and (16) one has

$$\nabla\theta = +\eta\sum_{\tau=1}^{M} p_\theta(\tau)R(\tau)\sum_{t=1}^{N}\nabla_\theta \ln \pi_\theta(a_t, s_t) = +\eta E\left[R(\tau)\sum_{t=1}^{N}\nabla_\theta \ln \pi_\theta(a_t, s_t)\right]. \tag{17}$$

The parameters are updated according to the logarithmic gradient of the policy times the return and the learning rate that must be chosen neither too fast nor too slow. Since the gradient contains the return, all the actions become more likely, the more the return is larger. An important point is that the optimization algorithm of the network parameters is independent of the underlying dynamical model. This is a difference with the standard optimization in OCT where the gradient of the final fidelity involves the system Hamiltonian. RL operates beyond static databases; it collects data during training.

B. Optimal control

The optimal field is built by iterations to maximize the cost functional that is the fidelity $\mathcal{F} = Tr[\rho^{\dagger}_{\mathrm{target}}\rho(T)]$ at the final time T with constraints to restrain the field intensity and to fulfill the master equation at any time. The optimization is performed here by Rabitz' monotonously convergent algorithm [23,84] that involves forward and backward propagation of the system density matrices with initial condition $\rho(0)$ and of an auxiliary system density matrix $\chi(t)$ with final condition

$\chi(T) = \rho_{\text{target}}$. It is worth noting that a two-point boundary-value quantum control paradigm (TBQCP) has been presented in the literature as an accelerated convergent algorithm [85]. However, for the sake of simplicity this method is not used in our simulations. Dynamics is driven with HEOM. The iterations begin with a guess field that strongly influences the final field. The operational relations for the backward propagation are given in Refs. [41,56]. The field at each iteration k is obtained by $\varepsilon^{(k)} = \varepsilon^{(k-1)} + \Delta\varepsilon^{(k)}$ where $\Delta\varepsilon^{(k)}$ is estimated by

$$\Delta\varepsilon(t) = \frac{1}{\alpha}\text{Im}m\left\{\text{Tr}\left(\chi(t)\left[\sum_p \mu_p, \rho(t)\right]\right)\right\}, \qquad (18)$$

where α is the intensity penalty factor. Note that we do not use RWA in this approach.

V. CONTROL BY RL

We choose a pulse duration of $T = 20$ fs. This is relevant to avoid a too large spectral band that would imply higher bright excited states not included in the model system. For RL simulations, the laser detunings are assumed to vanish, so the only optimized parameters are the two Rabi frequencies Ω_y and Ω_z at all times. In all the RL examples, the results are given in reduced units for the time (t/T) and the Rabi frequency $(T\Omega)$. Conventional units are used in some OCT examples.

The three hidden network layers contain 100, 50, and 30 neurons. The learning rate has its standard value $\eta = 10^{-3}$. Our investigation has confirmed the critical significance of these meta-parameters. Reducing the number of neurons results in a decelerated convergence rate, while elevating it prolongs computational duration without commensurate convergence enhancement. Increasing the learning rate by a factor of 10 is not efficient to reach the desired target. Data are collected during a bunch of $M = 10$ episodes that are divided in 50 time steps.

A. RL in the isolated system

We first consider the ideal case without dissipation. In RL, the process begins by two Rabi frequencies chosen at random in a given range. The analytical solution [Eq. (5)] is a landmark. By assuming a constant pulse envelope, the integrated Rabi frequency is $T\Omega$ and the best value should be $\pi/\sqrt{2} = 2.22$. We begin the RL simulation with an initial interval with $T\Omega_{\text{min}} = 0$ and $T\Omega_{\text{max}} = 3$ that would give an area larger than the best analytical value of 2.22. Figure 5(a) displays the first iteration with initial random Rabi frequencies, here 1.54 for Ω_y and 1.55 for Ω_z. The integrated frequencies are too small and the ground state is not completely depopulated. Figure 5(b) gives the outcome after 100 episodes. It reproduces the analytical result and is obtained after about 60 episodes as illustrated in Fig. 6. RL provides the good integrated frequencies but with very simple pulses since the envelopes are quasi constant. Figure 6 displays the return achieved during five simulations of 100 episodes. The random initial conditions differ from $T\Omega = 1.5$ by about 10% giving a return close to 0.75. Notably, the convergence rate exhibits variability and does not follow the typical monotonic pattern

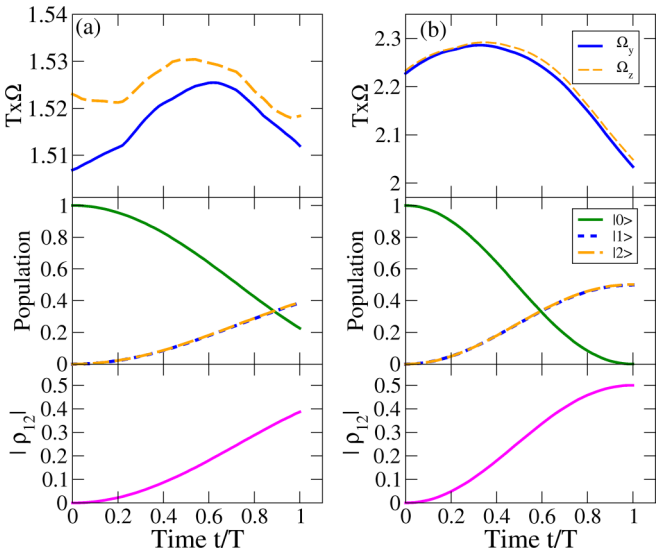

FIG. 5. Optimization of the superposed target state [Eq. (4)] by RL in the isolated V-three-level system and $T\Omega_{\text{max}} = 3$. The upper panels give the Rabi frequencies in reduced units $(T\Omega)$, the middle ones show the populations in each state, and the lower ones the modulus of the coherence ρ_{12} between the two excited states. (a) First iteration with initial random Rabi frequencies. The integrated frequencies are 1.54 for Ω_y and 1.55 for Ω_z. (b) After 100 episodes both areas are 2.22 $(\pi/\sqrt{2})$, the best expected result.

observed in conventional OCT algorithms. However, from 60 iterations, the rate achieved its highest value.

To test the algorithm, we start with a larger initial interval with $T\Omega_{\text{max}} = 9$. As the initial frequencies are random in this range, RL converges towards different possibilities but it is worth noting that it always finds a solution close to the analytical result. Figure 7(a) illustrates a case where convergence occurs towards the expected value $\pi/\sqrt{2}$ with a very good final coherence. In Fig. 7(b) one sees that, according

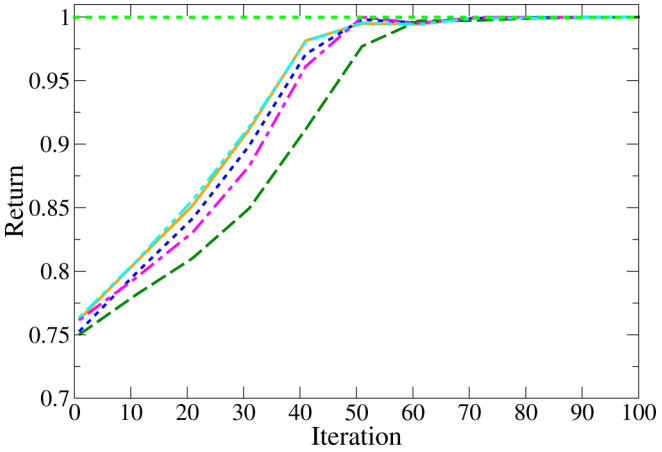

FIG. 6. Return [Eq. (14)] during the optimization in the isolated system presented in Fig. 5 for five different simulations displayed in different colors. Each simulation runs 100 episodes. The green dotted line serves as an indicator, highlighting the target return value of 1. The random initial conditions differ from $T\Omega = 1.5$ by about 10% giving a return close to 0.75.

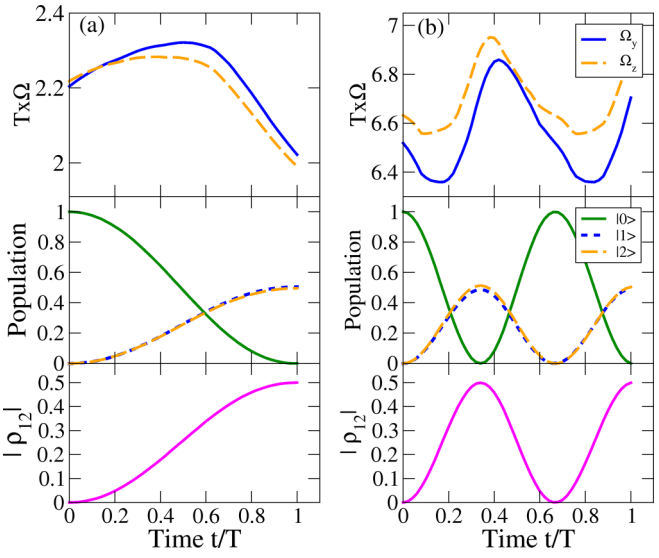

FIG. 7. Optimization of the superposed target state [Eq. (4)] by RL in the isolated V-three-level system and $T\Omega_{\max} = 9$. The upper, middle, and lower panels are as in Fig. 5. (a) Convergence towards the best expected result ($\pi/\sqrt{2}$). The integrated frequencies are 2.286 for Ω_y and 2.254 for Ω_z. (b) Convergence towards ($3\,\pi/\sqrt{2}$) leading to a supplementary Rabi oscillation. The integrated frequencies are 6.673 for Ω_y and 6.825 for Ω_z.

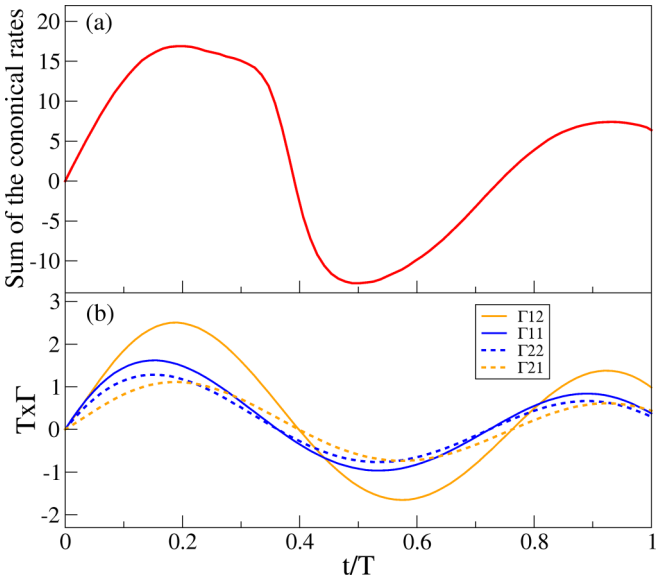

FIG. 8. (a) Sum of the canonical rates $T\Gamma_k^c$ (in reduced units) of the field-free dynamics. They are the eigenvalues of decoherence matrix [Eq. (9)]. (b) Time-dependent rates $T\Gamma$ associated to the four Lindblad operators describing the energy tuning (L_1, L_2 with rates Γ_{11} and Γ_{22}) and the interstate transition (L_3, L_4 with rates Γ_{12} and Γ_{21}) induced by the coupling bath.

to the random initial values, optimization provides a solution with higher intensity and an area close to $3\pi/\sqrt{2}$ leading to a supplementary complete Rabi oscillation before the final coherence creation close to 0.5.

B. RL with Lindblad dynamics

In this section, we will undertake a comparative analysis between scenarios involving constant Lindblad rates and those incorporating time-dependent variations. The constant rates associated to the four selected Lindblad operators (QuTip collapse operators) L_k defined in Sec. III A are (in reduced units $T\Gamma$): $T\Gamma_{11} = 1$ (L_1), $T\Gamma_{22} = 0.8$ (L_2), $T\Gamma_{12} = 0.36$ (L_3), and $T\Gamma_{21} = 0.16$ (L_4). The operators L_1 and L_2 couple to the tuning baths in excited states S_1 and S_2, respectively. The ratio of the rates is approximated by $\sqrt{J_{S_2}/J_{S_1}}$ as elaborated upon in Ref. [41]. The operators L_3 and L_4 induce nonadiabatic transitions. The reduced rate $T\Gamma_{12}$ and $T\Gamma_{21}$ are different as expected from the detailed balance. They are calibrated to roughly approximate the exact HEOM field-free dynamics at least during the early dynamics. Dynamics is performed with the QuTip MESOLVE solver [51].

Non-Markovianity may be taken into account in an approximated way by time-dependent rates. The transitory negativity of the sum of the canonical rates that are the eigenvalues of the decoherence matrix [Eq. (9)] is one of the signature. This sum obtained for the field-free dynamics is given in Fig. 8(a) in reduced units ($T = 20$ fs). It is obvious that its damped oscillation follows that of the bath correlation functions [see Fig. 3(b)]. This illustrates that, for this type of partition, the non-Markovianity is closely linked to the damped vibrational motion of the collective modes. Indeed, if the collective effective mode is underdamped, the nuclei oscillate and transitory return to the initial reference position. This tends to restore the system in its initial condition. The decay towards equilibrium is not monotonous.

Figure 8(b) presents the time-dependent rates associated to the four Lindblad operators describing the energy tuning (L_1, L_2) and the interstate transition (L_3, L_4) induced by the coupling bath. Their shape are approximated from those of some elements of the decoherence matrix [Eq. (9)] by considering the basis operator G_k corresponding to the 1–2 transition (analog of σ_x in the two-state case) and one operator corresponding to an energy gap. The amplitudes are calibrated as in the constant rate case from the field-free HEOM dynamics. The functions are fitted by polynomials or by the product of a sine function times a decreasing exponential. These functions are introduced in the time-dependent collapse operators of QuTip by using the MESOLVE solver [51].

Figure 9(a) shows the dynamics after 100 episodes with constant decay rates. The return is only 0.8 and it saturates after 60 iterations. Figure 9(b) presents the control with the time-dependent rates. The return is slightly improved. However, this is due mainly to the better depletion of the ground state and not to a better superposition. The optimal envelopes remain very simple and quasiconstant in both cases. It is a bit disconcerting that RL behaves in a very similar way with constant or time-dependent rates. We will compare these results with the OCT optimization in Sec. VI.

C. RL with HEOM dynamics

For a 20-fs simulation, truncating at level 6 of the hierarchy proves to be satisfactory. However, it is worth noting that, for longer dynamics, such as achieving a field-free asymptotic state, a higher level 9 becomes necessary. The implementation

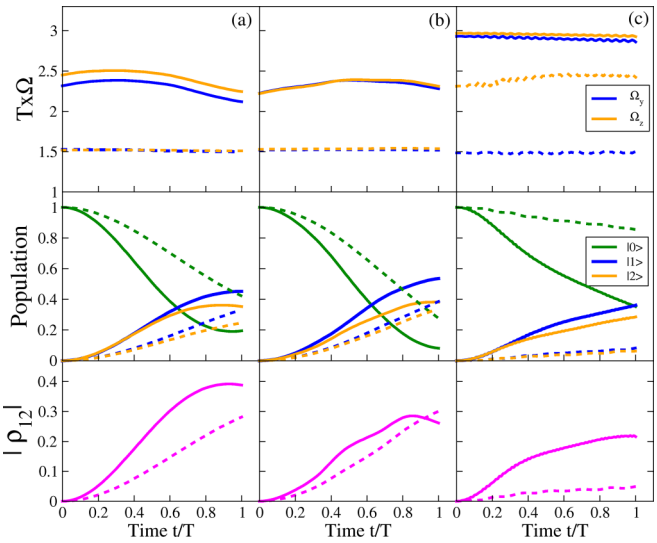

FIG. 9. Optimization of the target state [Eq. (4)] by RL with Lindblad or HEOM dynamics. The upper, middle, and lower panels are as in Fig. 5. The first episode with random initial conditions is given in dashed lines. The optimized results are given after 100 episodes. In each case, the return saturates after about 60 episodes. (a) Dynamics with constant Lindblad rates. The area of the Rabi frequencies are 2.35 and 2.48, respectively. (b) Dynamics with time-dependent rates. The area are 2.38 and 2.39. (c) HEOM dynamics at level 6 of the hierarchy in Schödinger representation without RWA.

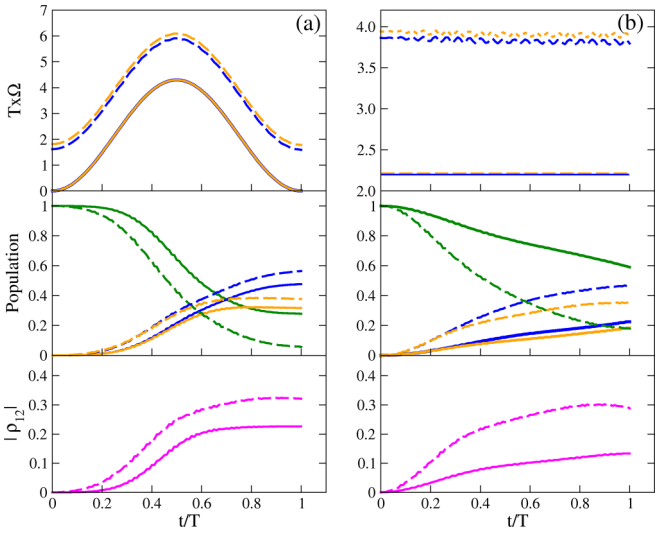

FIG. 10. RL optimization by RL with different guess fields. The actions are the variation of the Rabi frequencies $\delta\Omega$ in a range of reduced units $[-2, 2]$. Dynamics is computed by HEOM in Schrödinger representation without RWA at level 6 of the hierarchy. The upper panels give the Rabi frequencies Ω_y and Ω_z in reduced units, the middle ones, the populations, and the lower ones, the modulus of the coherence between the two excited states. (a) Guess fields with sine square envelopes of integrated Rabi frequencies $\pi/\sqrt{2}$ in solid line. The corresponding RL fields after 100 episodes in dashed line (the areas are 3.77 and 3.96). (b) Guess fields with constant envelope in solid line and the corresponding RL fields after 100 episodes in dashed line (the areas are 3.87 and 3.95).

of the RL algorithm with HEOM requires some comments. (i) During the Markov decision chain the solver is called repetitively for each time step of the chain $\ldots s_t \rightarrow a_t \rightarrow r_t \rightarrow s_{t+1} \rightarrow a_{t+1} \rightarrow r_{t+1} \ldots$. All the ADOs describing the state of the surrounding must be saved for the following decision step so that each bath retains its configuration and does not restart with the initial conditions of the baths with ADOs equal to zero. This is a difficulty that does not concern the local Lindblad dynamics. (ii) In our application, each spectral density (see Fig. 3) is fitted by two Tannor-Meier Lorentzian functions [77] leading to four decay modes for each bath. As the spectral densities are centered at high frequencies, we do not include Matsubara terms at room temperature. We use the description of the baths by the expansion of the real and imaginary parts of the correlation functions using the BOSONICBATH application of the QUTIP HEOMSOLVER [57]. When working with reduced units the real or imaginary parts of the $c_k^{R,I}$ coefficients must be scaled by T^2 and the rates $\gamma_k^{R,I}$ by T as usually. (iii) In the HEOM solver, the system-bath coupling operators are not written in interaction representation. Therefore, we re-transformed the Hamiltonian in Schrödinger representation without the RWA approximation. We use 100 time steps in each episode, which is enough to satisfy the Nyquist-Shannon sampling rule [86] for the fields in the Schrödinger representation.

The RL optimization with HEOM in the Schrödinger representation without RWA is given in Fig. 9(c). The actions are more erratic due to the oscillation of the field in this representation. After 100 episodes, the envelopes have a slightly higher amplitude than in the Lindblad simulations. The populations and coherence behave on a similar way in each simulation. The return is only 0.53 primarily due to the less-than-optimal depletion of the ground state and the difference of population in the two excited states. Other examples with guess fields are given in Fig. 10.

Finally, we explore another strategy. We impose a guess field for the RL control by choosing the actions to be the variation $\delta\Omega$ of the Rabi frequencies with respect to the guess [see Fig. 2(b)]. These trial fields are a sine square envelope or a constant satisfying the $\pi/\sqrt{2}$ rule. Simulations are carried out in Schrödinger representation without RWA at level 6 of the hierarchy. The actions $T\delta\Omega$ are taken in an interval $[-2, 2]$ in reduced units. The Rabi frequencies of the guess fields and of the RL optimization after 100 episodes are given in Fig. 10. The envelopes are only very slightly modified during the optimization. The first action is always the largest and shifts the guess envelopes by adding a constant value. The further fluctuations remain of weak amplitude. Increasing the initial interval only modifies the initial shift. Only the area increases, which generally enhances the depletion of the ground state but not the target coherence. The sine square envelope is the best guess giving a return of 80%. The constant envelope provides only 70%.

VI. COMPARISON RL-OCT

The envelopes generated by RL consistently exhibit a high degree of simplicity, characterized by their quasi-constant profile when no guess is imposed. Given that the fields generated by standard OCT typically exhibit a higher degree of

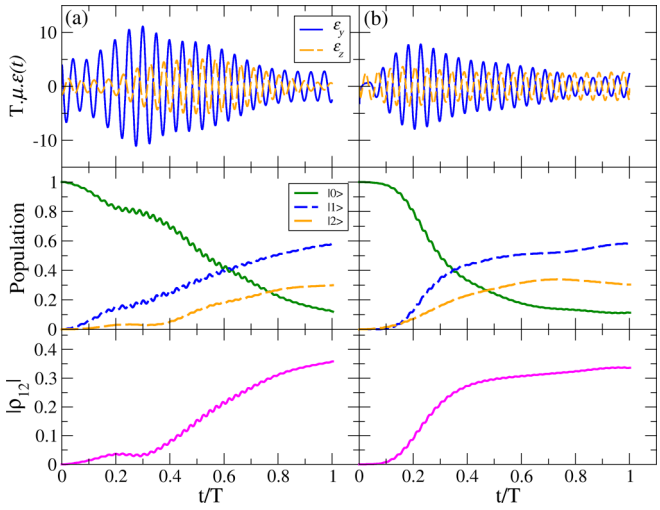

FIG. 11. Optimization by standard OCT after 15 iterations. (a) Sine square guess fields, (b) constant envelope guess fields. The upper panels give the fields times the corresponding transition dipole in reduced units $T\mu_y\mathcal{E}_y(t)$ and $T\mu_z\mathcal{E}_z(t)$. Dynamics is computed by HEOM in Schrödinger representation without RWA at level 6 of hierarchy. The lower panels display the populations and the modulus of the coherence between the two excited states.

structure [41], we compare in Fig. 11 the fields optimized by OCT with the same initial guess fields drawn in solid lines in Fig. 10. Simulations are performed utilizing our HEOM code [41] in Schrödinger representation without RWA at level 6 of the hierarchy. Standard OCT optimizes the field amplitudes and not only the envelopes on a time grid. This offers more flexibility and may slightly modify the carrier frequency. The α penalty factor [Eq. (18)] is fixed to 2×10^{-4}. It influences the optimization rate. The field amplitudes increase regularly at each iteration. We take two snapshots to remain in the same order of magnitude as in the RL simulation. The results after 15 iterations are shown in Figs. 11(a) and 11(b). We draw the fields times the dipole moment $T\mu_y\mathcal{E}_y(t)$ and $T\mu_z\mathcal{E}_z(t)$ in reduced units so that the envelopes may be compared with the reduced Rabi frequencies $T\Omega$ used in the RL optimization. OCT reshapes the envelopes more strongly than RL. In particular, the two envelopes do not remain similar. However, when the maximum field amplitude are in the same range, the return is similar around 80% for the sine square case and reaches 80% versus 70% in RL for the constant guess.

VII. SUMMARY AND CONCLUSION

Examining the potential of RL in quantum control has primarily been explored within the domain of quantum information [32,34–36]. This analysis is particularly interesting in the context of retrieving the STIRAP scheme using either digital pulses [38] or continuous ones [39,40]. The outstanding property is the ability to propose strategies without any prior knowledge of the system leading to the denomination as a "model-free algorithm" [83]. The main question is to see whether RL will find new strategies in particular in the presence of an environment. Most of the previous works studied examples with dissipation treated by Lindblad master equation, i.e., for a Markovian dynamics [32,40]. However, even if RL is built on a succession of decision Markov processes, non-Markovian noise could influence the system dynamics [37].

In this work, we revisited a control in a system with a strong non-Markovian dynamics due to the coupling to baths with highly structured spectral densities leading to long bath correlation times. The model is calibrated from *ab initio* data [41,58,60]. We used an open source software [40,50] based on the policy gradient method for the optimization and on the QuTip MESOLVE solver of the Lindblad master equation [51]. We enhanced its functionality to address non-Markovian dynamics. In a first approximate go-between step we incorporated time-dependent rate constants derived from the field-free HEOM dynamics. Then we interfaced the RL algorithm with the HEOM solver of the QuTip BOFIN package [57].

An analytical solution exists to create the target superposition of two excited states addressable with orthogonal dipoles in an isolated V-system. It is of significant interest to assess the proficiency of RL to recover the expected solution from random initial conditions in a given interval for the Rabi frequencies. RL finds a very simple but efficient solution of straightforward quasi-constant envelopes satisfying the integrated Rabi frequency rule. The amplitudes display minimal variations, typically within a few percentage points.

For each level of complexity of the dynamics, we observe that the return saturates after about 60 iterations even if the target is not perfectly reached. The Rabi frequencies are always very simple, nearly constant when no guess is imposed. The proposal is basic and robust. Even if it is not completely satisfactory, RL does not go further. When dynamics is driven in Schrödinger representation without RWA, the process of optimizing using reinforcement learning (RL) exhibits increased complexity. This complexity is reflected in the erratic behavior of the envelopes from one step to another. However, it is worth noting that, despite these fluctuations, a smoother average trajectory is observed with only minor fluctuations.

Lindblad dynamics even with some time-dependent rates cannot take into account a possible influence of the field on the baths. On the contrary, the memory kernel of HEOM contains the time-dependent Hamiltonian [77] and this could, in principle, induce an effect on the bath dynamics [87]. Indeed, the decoherence matrix and thus the rates are modified by the field [56]. However, in our application the behavior is qualitatively the same for the approximate non-Markovian approach or for exact HEOM. RL successfully captures the crucial condition regarding the integrated Rabi frequency, yet it does not discover novel strategies to fight dissipation. It would be powerful to increase the number of available actions, enabling optimization of detuning parameters as well or directly the amplitude of the fields and not only the envelopes. Another possibility would be to let the algorithm choose a guess field. Furthermore, the exploration of more sophisticated RL algorithms holds promise for future investigations [88,89].

The simplicity of the RL envelopes suggests confronting the standard OCT and to see if it can yield superior results. OCT exploits the system dynamics and may seem more flexible since it optimizes the field amplitude and the carrier frequency and possibly finds some chirp effect. However, in

our example OCT is not more efficient to reach the target with dissipation. By imposing the same guess fields, RL and OCT provide different optimal fields ensuring similar return. The reshaping is stronger in OCT that proposes different envelopes for the two polarizations which RL does not do. When the envelope amplitudes are maintained in the same range as in RL, OCT slighly improves the depletion of the ground state, but not really the preparation of the superposition with equal weights. Both strategies, RL and OCT depend on the guess fields and the optimal fields are different. However, they ensure similar final dynamics and the perfect target is not achieved neither by RL nor by OCT control due to the strong dissipation.

Our example is a complex system strongly coupled to a structured environment with laser pulses in the femtosecond range. RL seems more adapted to treat quantum information in another spectral range operating with very simple square box envelopes and weakly coupled Markovian noises [90].

The data are available upon request to the authors. The modified ThreeLS.py file of Giannelli's open-source software [40,50] allowing dynamics with HEOM in Schrödinger representation without RWA by using the QuTip BOFIN package [57] is given in the Supplemental Material [91].

ACKNOWLEDGMENTS

The authors thanks Dr. O. Atabek for stimulating discussions and encouragements to explore this project. Dr. B. Lasorne and J. Galiana (PhD) are acknowledged for providing the *ab initio* data calibrating the model.

[1] L. M. K. Vandersypen and I. L. Chuang, Nmr techniques for quantum control and computation, Rev. Mod. Phys. **76**, 1037 (2005).
[2] D. Keefer and R. de Vivie-Riedle, Pathways to new applications for quantum control, Acc. Chem. Res. **51**, 2279 (2018).
[3] F. Caruso, S. Montangero, T. Calarco, S. F. Huelga, and M. B. Plenio, Coherent optimal control of photosynthetic molecules, Phys. Rev. A **85**, 042331 (2012).
[4] S. Hoyer, F. Caruso, S. Montangero, M. Sarovar, T. Calarco, M. B. Plenio, and K. B. Whaley, Realistic and verifiable coherent control of excitonic states in a light-harvesting complex, New J. Phys. **16**, 045007 (2014).
[5] S. J. Glaser, U. Boscain, T. Calarco, C. P. Koch, W. Köckenberger, R. Kosloff, R. Kosloff, I. Kuprov, B. Luy, S. Schirmer, T. Schulte-Herbrüggen, D. Sugny, and F. K. Wilhelm, Training Schrödinger's cat: quantum optimal control, Eur. Phys. J. D **69**, 279 (2015).
[6] C. P. Koch, U. Boscain, T. Calarco, G. Dirr, S. Filipp, S. J. Glaser, R. Kosloff, S. Montangero, T. Schulte-Herbrüggen, D. Sugny, and F. K. Wilhelm, Quantum optimal control in quantum technologies. strategic report on current status, visions and goals for research in europe, EPJ Quantum Technology **9**, 19 (2022).
[7] C.-P. Yang and S. Han, *n*-qubit-controlled phase gate with superconducting quantum-interference devices coupled to a resonator, Phys. Rev. A **72**, 032311 (2005).
[8] C. Rangan, A. M. Bloch, C. Monroe, and P. H. Bucksbaum, Control of trapped-ion quantum states with optical pulses, Phys. Rev. Lett. **92**, 113004 (2004).
[9] M. P. A. Jones, J. Beugnon, A. Gaëtan, J. Zhang, G. Messin, A. Browaeys, and P. Grangier, Fast quantum state control of a single trapped neutral atom, Phys. Rev. A **75**, 040301(R) (2007).
[10] C. Avinadav, R. Fischer, P. London, and D. Gershoni, Time-optimal universal control of two-level systems under strong driving, Phys. Rev. B **89**, 245311 (2014).
[11] E. Räsänen, A. Castro, J. Werschnik, A. Rubio, and E. K. U. Gross, Optimal laser control of double quantum dots, Phys. Rev. B **77**, 085324 (2008).
[12] D. J. Tannor and S. A. Rice, Control of selectivity of chemical reaction via control of wave packet evolution, J. Chem. Phys. **83**, 5013 (1985).
[13] R. Kosloff, S. Rice, P. Gaspard, S. Tersigni, and D. Tannor, Wavepacket dancing: Achieving chemical selectivity by shaping light pulses, Chem. Phys. **139**, 201 (1989).
[14] P. Brumer and M. Shapiro, Laser control of molecular processes, Annu. Rev. Phys. Chem. **43**, 257 (1992).
[15] N. V. Vitanov, A. A. Rangelov, B. W. Shore, and K. Bergmann, Stimulated raman adiabatic passage in physics, chemistry, and beyond, Rev. Mod. Phys. **89**, 015006 (2017).
[16] K. Bergmann, H.-C. Nägerl, C. Panda, G. Gabrielse, E. Miloglyadov, M. Quack, G. Seyfang, G. Wichmann, S. Ospelkaus, A. Kuhn *et al.*, Roadmap on stirap applications, J. Phys. B: At. Mol. Opt. Phys. **52**, 202001 (2019).
[17] V. Engel, C. Meier, and D. J. Tannor, Local control theory: Recent applications to energy and particle transfer processes in molecules, Adv. Chem. Phys. **141**, 29 (2009).
[18] M. Mališ, P. K. Barkoutsos, M. Ganzhorn, S. Filipp, D. J. Egger, S. Bonella, and I. Tavernelli, Local control theory for superconducting qubits, Phys. Rev. A **99**, 052316 (2019).
[19] S. C. Hou, M. A. Khan, X. X. Yi, D. Dong, and I. R. Petersen, Optimal lyapunov-based quantum control for quantum systems, Phys. Rev. A **86**, 022321 (2012).
[20] U. Boscain, M. Sigalotti, and D. Sugny, Introduction to the pontryagin maximum principle for quantum optimal control, PRX Quantum **2**, 030203 (2021).
[21] W. Zhu and H. Rabitz, Quantum control design via adaptive tracking, J. Chem. Phys. **119**, 3619 (2003).
[22] A. P. Peirce, M. A. Dahleh, and H. Rabitz, Optimal control of quantum-mechanical systems: Existence, numerical approximation, and applications, Phys. Rev. A **37**, 4950 (1988).
[23] W. Zhu, J. Botina, and H. Rabitz, Rapidly convergent iteration methods for quantum optimal control of population, J. Chem. Phys. **108**, 1953 (1998).
[24] Y. Maday and G. Turinici, New formulations of monotonically convergent quantum control algorithms, J. Chem. Phys. **118**, 8191 (2003).

[25] V. F. Krotov, *Global Methods in Optimal Control Theory*, Pure and Applied Mathematics, Vol. 195 (Marcel Dekker, New York, 1996).

[26] J. P. Palao and R. Kosloff, Optimal control theory for unitary transformations, Phys. Rev. A **68**, 062308 (2003).

[27] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrüggen, and S. J. Glaser, Optimal control of coupled spin dynamics: design of nmr pulse sequences by gradient ascent algorithms, J. Magn. Reson. **172**, 296 (2005).

[28] P. de Fouquieres, S. Schirmer, S. Glaser, and I. Kuprov, Second order gradient ascent pulse engineering, J. Magn. Reson. **212**, 412 (2011).

[29] P. Doria, T. Calarco, and S. Montangero, Optimal control technique for many-body quantum dynamics, Phys. Rev. Lett. **106**, 190501 (2011).

[30] B. Riaz, C. Shuang, and S. Qamar, Optimal control methods for quantum gate preparation: a comparative study, Quantum Inf. Process. **18**, 100 (2019).

[31] X.-M. Zhang, Z. Wei, R. Asad, X.-C. Yang, and X. Wang, When does reinforcement learning stand out in quantum control? A comparative study on state preparation, npj Quantum Inf. **5**, 85 (2019).

[32] Z. An, H.-J. Song, Q.-K. He, and D. L. Zhou, Quantum optimal control of multilevel dissipative quantum systems with reinforcement learning, Phys. Rev. A **103**, 012404 (2021).

[33] M. Bukov, A. G. R. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta, Reinforcement learning in different phases of quantum control, Phys. Rev. X **8**, 031086 (2018).

[34] Z. An and D. L. Zhou, Deep reinforcement learning for quantum gate control, Europhys. Lett. **126**, 60002 (2019).

[35] M. Y. Niu, S. Boixo, V. N. Smelyanskiy, P. Weinberg, and H. Neven, Universal quantum control through deep reinforcement learning, npj Quantum Inf. **5**, 33 (2019).

[36] L. Moro, M. G. A. Paris, M. Restelli, and E. Prati, Quantum compiling by deep reinforcement learning, Commun. Phys. **4**, 178 (2021).

[37] R. Porotti, D. Tamascelli, M. Restelli, and E. Prati, Coherent transport of quantum states by deep reinforcement learning, Commun. Phys. **2**, 61 (2019).

[38] I. Paparelle, L. Moro, and E. Prati, Digitally stimulated raman passage by deep reinforcement learning, Phys. Lett. A **384**, 126266 (2020).

[39] J. Brown, P. Sgroi, L. Giannelli, G. S. Paraoanu, E. Paladino, G. Falci, M. Paternostro, and A. Ferraro, Reinforcement learning-enhanced protocols for coherent population-transfer in three-level quantum systems, New J. Phys. **23**, 093035 (2021).

[40] L. Giannelli, P. Sgroi, J. Brown, G. S. Paraoanu, M. Paternostro, E. Paladino, and G. Falci, A tutorial on optimal control and reinforcement learning methods for quantum technologies, Phys. Lett. A **434**, 128054 (2022).

[41] A. Jaouadi, J. Galiana, E. Mangaud, B. Lasorne, O. Atabek, and M. Desouter-Lecomte, Laser-controlled electronic symmetry breaking in a phenylene ethynylene dimer: Simulation by the hierarchical equations of motion and optimal control, Phys. Rev. A **106**, 043121 (2022).

[42] S. Tomasi, S. Baghbanzadeh, S. Rahimi-Keshari, and I. Kassal, Coherent and controllable enhancement of light-harvesting efficiency, Phys. Rev. A **100**, 043411 (2019).

[43] G. Breuil, E. Mangaud, B. Lasorne, O. Atabek, and M. Desouter-Lecomte, Funneling dynamics in a phenylacetylene trimer: Coherent excitation of donor excitonic states and their superposition, J. Chem. Phys. **155**, 034303 (2021).

[44] Y. Tanimura, Numerically "exact" approach to open quantum dynamics: The hierarchical equations of motion (HEOM), J. Chem. Phys. **153**, 020901 (2020).

[45] Q. Shi, Y. Xu, Y. Yan, and M. Xu, Efficient propagation of the hierarchical equations of motion using the matrix product state method, J. Chem. Phys. **148**, 174102 (2018).

[46] Y. Yan, M. Xu, T. Li, and Q. Shi, Efficient propagation of the hierarchical equations of motion using the tucker and hierarchical tucker tensors, J. Chem. Phys. **154**, 194104 (2021).

[47] X. Dan and Q. Shi, Theoretical study of nonadiabatic hydrogen atom scattering dynamics on metal surfaces using the hierarchical equations of motion method, J. Chem. Phys. **159**, 044101 (2023).

[48] R. Borrelli and S. Dolgov, Expanding the range of hierarchical equations of motion by tensor-train implementation, J. Phys. Chem. B **125**, 5397 (2021).

[49] E. Mangaud, A. Jaouadi, A. W. Chin, and M. Desouter-Lecomte, Survey of the hierarchical equations of motion in tensor-train format for non-Markovian quantum dynamics, Eur. Phys. J. Spec. Top. **232**, 1847 (2023).

[50] L. Giannelli, https://www.github.com/luigiannelli/threeLS_populationTransfer.

[51] J. Johansson, P. Nation, and F. Nori, Qutip: An open-source python framework for the dynamics of open quantum systems, Comput. Phys. Commun. **183**, 1760 (2012).

[52] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, TensorFlow: Large-scale machine learning on heterogeneous systems (2015), software available from tensorflow.org.

[53] J. Piilo, S. Maniscalco, K. Härkönen, and K.-A. Suominen, Non-Markovian quantum jumps, Phys. Rev. Lett. **100**, 180402 (2008).

[54] M. J. W. Hall, J. D. Cresser, L. Li, and E. Andersson, Canonical form of master equations and characterization of non-Markovianity, Phys. Rev. A **89**, 042120 (2014).

[55] E. Mangaud, C. Meier, and M. Desouter-Lecomte, Analysis of the non-Markovianity for electron transfer reactions in an oligothiophene-fullerene heterojunction, Chem. Phys. **494**, 90 (2017).

[56] E. Mangaud, R. Puthumpally-Joseph, D. Sugny, C. Meier, O. Atabek, and M. Desouter-Lecomte, Non-Markovianity in the optimal control of an open quantum system described by hierarchical equations of motion, New J. Phys. **20**, 043050 (2018).

[57] N. Lambert, T. Raheja, S. Cross, P. Menczel, S. Ahmed, A. Pitchford, D. Burgarth, and F. Nori, QuTiP-BOFiN: A bosonic and fermionic numerical hierarchical-equations-of-motion library with applications in light-harvesting, quantum control, and single-molecule electronics, Phys. Rev. Res. **5**, 013181 (2023).

[58] E. K.-L. Ho, T. Etienne, and B. Lasorne, Vibronic properties of para-polyphenylene ethynylenes: TD-DFT insights, J. Chem. Phys. **146**, 164303 (2017).

[59] E. K.-L. Ho and B. Lasorne, Diabatic pseudofragmentation and nonadiabatic excitation-energy transfer in meta-substituted dendrimer building blocks, Comput. Theor. Chem. **1156**, 25 (2019).

[60] J. Galiana and B. Lasorne, On the unusual Stokes shift in the smallest PPE dendrimer building block: Role of the vibronic

symmetry on the band origin? J. Chem. Phys. **158**, 124113 (2023).
[61] K. Fujii, Introduction to the rotating wave approximation (rwa): Two coherent oscillations, J. Mod. Phys. **8**, 2042 (2017).
[62] G. F. Thomas, Validity of the Rosen-Zener conjecture for Gaussian-modulated pulses, Phys. Rev. A **27**, 2744 (1983).
[63] H.-G. Duan and M. Thorwart, Quantum mechanical wave packet dynamics at a conical intersection with strong vibrational dissipation, J. Phys. Chem. Lett. **7**, 382 (2016).
[64] E. Mangaud, B. Lasorne, O. Atabek, and M. Desouter-Lecomte, Statistical distributions of the tuning and coupling collective modes at a conical intersection using the hierarchical equations of motion, J. Chem. Phys. **151**, 244102 (2019).
[65] H.-P. Breuer and F. Petruccione, *The Theory of Open Quantum Systems* (Oxford University Press, New York, 2002).
[66] I. de Vega and D. Alonso, Dynamics of non-Markovian open quantum systems, Rev. Mod. Phys. **89**, 015001 (2017).
[67] C. P. Koch, Controlling open quantum systems: tools, achievements, and limitations, J. Phys.: Condens. Matter **28**, 213001 (2016).
[68] G. Lindblad, On the generators of quantum dynamical semigroups, J. Chem. Phys. **48**, 119 (1976).
[69] D. Manzano, A short introduction to the Lindblad master equation, AIP Adv. **10**, 025106 (2020).
[70] Á. Rivas, S. F. Huelga, and M. B. Plenio, Quantum non-Markovianity: characterization, quantification and detection, Rep. Prog. Phys. **77**, 094001 (2014).
[71] B. Witt, L. Rudnicki, Y. Tanimura, and F. Mintert, Exploring complete positivity in hierarchy equations of motion, New J. Phys. **19**, 013007 (2017).
[72] G. Théret and D. Sugny, Complete positivity, positivity, and long-time asymptotic behavior in a two-level open quantum system, Phys. Rev. A **108**, 032212 (2023).
[73] G. Kimura, The Bloch vector for N-level systems, Phys. Lett. A **314**, 339 (2003).
[74] D. Aerts and M. S. de Bianchi, The extended Bloch representation of quantum mechanics and the hidden-measurement solution to the measurement problem, Ann. Phys.(NY) **351**, 975 (2014).
[75] E. Andersson, J. D. Cresser, and M. J. W. Hall, Finding the kraus decomposition from a master equation and vice versa, J. Mod. Opt. **54**, 1695 (2007).
[76] A. Pomyalov, C. Meier, and D. J. Tannor, The importance of initial correlations in rate dynamics: A consistent non-Markovian master equation approach, Chem. Phys. **370**, 98 (2010).
[77] C. Meier and D. J. Tannor, Non-Markovian evolution of the density operator in the presence of strong laser fields, J. Chem. Phys. **111**, 3365 (1999).
[78] U. Kleinekathöfer, Non-Markovian theories based on a decomposition of the spectral density, J. Chem. Phys. **121**, 2505 (2004).
[79] R. S. Sutton and A. G. Barto, *Reinforcement Learning, Second Edition: An Introduction* (MIT Press, Cambridge, MA, 2018).
[80] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).
[81] R. J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, Mach. Learn. **8**, 229 (1992).
[82] J. D. Martín-Guerrero and L. Lamata, Reinforcement learning and physics, Appl. Sci. **11**, 8589 (2021).
[83] F. Marquardt, Machine learning and quantum devices, SciPost Phys. Lect. Notes 29 (2021).
[84] Y. Ohtsuki, W. Zhu, and H. Rabitz, Monotonically convergent algorithm for quantum optimal control with dissipation, J. Chem. Phys. **110**, 9825 (1999).
[85] T.-S. Ho and H. Rabitz, Accelerated monotonic convergence of optimal control over quantum dynamics, Phys. Rev. E **82**, 026703 (2010).
[86] R. Marks, *Handbook of Fourier Analysis & Its Applications* (Oxford University Press, New York, 2009).
[87] A. Chenel, G. Dive, C. Meier, and M. Desouter-Lecomte, Control in a dissipative environment: The example of a cope rearrangement, J. Phys. Chem. A **116**, 11273 (2012).
[88] K. Reuer, J. Landgraf, T. Fösel, J. O'Sullivan, L. Beltrán, A. Akin, G. J. Norris, A. Remm, M. Kerschbaum, J.-C. Besse, F. Marquardt, A. Wallraff, and C. Eichler, Realizing a deep reinforcement learning agent discovering real-time feedback control strategies for a quantum system, Nat. Commun. **14**, 7138 (2023).
[89] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, Trust region policy optimization, arXiv:1502.05477.
[90] Y. Baum, M. Amico, S. Howell, M. Hush, M. Liuzzi, P. Mundada, T. Merkh, A. R. R. Carvalho, and M. J. Biercuk, Experimental deep reinforcement learning for error-robust gate-set design on a superconducting quantum computer, PRX Quantum **2**, 040324 (2021).
[91] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevA.109.013104 for the modifications of the rl software to run non-Markovian dynamics.