

PAPER • OPEN ACCESS

## Reinforcement learning-enhanced protocols for coherent population-transfer in three-level quantum systems

To cite this article: Jonathon Brown *et al* 2021 *New J. Phys.* **23** 093035

View the [article online](#) for updates and enhancements.

You may also like

- [Creation of electrical knots and observation of DNA topology](#)  
Tian Chen, Xingen Zheng, Qingsong Pei et al.
- [Hydrodynamics of simple active liquids: the emergence of velocity correlations](#)  
Umberto Marini Bettolo Marconi, Lorenzo Caprini and Andrea Puglisi
- [Stationary excitation waves and multimerization in arrays of quantum emitters](#)  
Davide Lonigro, Paolo Facchi, Saverio Pascazio et al.



## PAPER

# Reinforcement learning-enhanced protocols for coherent population-transfer in three-level quantum systems

Jonathon Brown<sup>1,6,\*</sup> , Pierpaolo Sgroi<sup>1,6</sup>, Luigi Giannelli<sup>2,3</sup>,  
Gheorghe Sorin Paraoanu<sup>4</sup> , Elisabetta Paladino<sup>2,3,5</sup>, Giuseppe Falci<sup>2,3,5</sup>,  
Mauro Paternostro<sup>1</sup>  and Alessandro Ferraro<sup>1</sup>

<sup>1</sup> Centre for Theoretical Atomic, Molecular, and Optical Physics, School of Mathematics and Physics, Queens University, Belfast BT7 1NN, United Kingdom

<sup>2</sup> Dipartimento di Fisica e Astronomia “Ettore Majorana”, Università di Catania, Via S. Sofia 64, 95123, Catania, Italy

<sup>3</sup> INFN, Sez. Catania, 95123, Catania, Italy

<sup>4</sup> QTF Centre of Excellence, Department of Applied Physics, Aalto University School of Science, PO Box 15100, FI-00076 AALTO, Finland

<sup>5</sup> CNR-IMM, UoS Università, 95123, Catania, Italy

\* Author to whom any correspondence should be addressed.

<sup>6</sup> These authors contributed equally to this work.

E-mail: [jbrown71@qub.ac.uk](mailto:jbrown71@qub.ac.uk)

**Keywords:** quantum control, reinforcement learning, condensed matter physics

## RECEIVED

24 May 2021

## REVISED

10 August 2021

## ACCEPTED FOR PUBLICATION

3 September 2021

## PUBLISHED

24 September 2021

Original content from  
this work may be used  
under the terms of the  
[Creative Commons  
Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/).

Any further distribution  
of this work must  
maintain attribution to  
the author(s) and the  
title of the work, journal  
citation and DOI.



## Abstract

We deploy a combination of reinforcement learning-based approaches and more traditional optimization techniques to identify optimal protocols for population transfer in a multi-level system. We constrain our strategy to the case of fixed coupling rates but time-varying detunings, a situation that would simplify considerably the implementation of population transfer in relevant experimental platforms, such as semiconducting and superconducting ones. Our approach is able to explore the space of possible control protocols to reveal the existence of efficient protocols that, remarkably, differ from (and can be superior to) standard Raman, stimulated Raman adiabatic passage or other adiabatic schemes. The new protocols that we identify are robust against both energy losses and dephasing.

## 1. Introduction

It is well known that quantum systems can provide clear computational advantage when compared with their classical counterparts, and several algorithms have been presented whereby this advantage is exploited to carry out so called super-classical tasks [1–3]. The required control over quantum systems, however, still remains the biggest challenge for full implementation of quantum computing algorithms. An experimental platform that provides a promising candidate for controlling general quantum systems are superconducting circuits, which have been widely employed to fabricate qubits (see [4, 5] for reviews) and two qubit gates [6–8], as well as implementations of so-called circuit-quantum electrodynamics (QED) [9, 10], which is at the forefront of the current ‘quantum race’ [11]. Multi-level dynamics has also been addressed both theoretically [12–15] and experimentally [16–18]. However, together with the promise of an experimental platform to manipulate quantum systems towards the achievement of a quantum network of multiple nodes, comes an increased demand for quantum-control schemes pertinent to the experimental constraints at play. Much of the work towards this goal employs techniques from NMR, quantum optics and quantum optimal control theory [19, 20]. Specifically, gradient-based optimization methods have been recently employed to control general open systems with a myriad of applications [21], as well as aiding the design of high-fidelity, protected superconducting quantum gates [22–25]. In the context of multi-level systems, the use of two tone pulses allow faithful, selective and robust single-qubit quantum operations such as population transfer and generation of superposition and can be generalized to quantum operations on multiple nodes of a network, such as state shuttling, entanglement and single photon generation.

Stimulated Raman adiabatic passage (STIRAP) and Raman oscillations are two well-known protocols for the implementation of these quantum operations. Some effort has been made to adapt the original formulation of such protocols to reduced-control architectures [13, 14, 26] or to improve them by using optimal pulse shaping and superadiabatic techniques [16, 27, 28].

More recently machine learning techniques have emerged as a viable option for finding alternative optimal control schemes. In particular reinforcement learning (RL) has been employed in the context of state preparation [29, 30], circuit architecture design [31] and control of multi-level systems [32]. In the context of three level systems, deep neural network based RL has been used along with state monitoring to learn optimal pulse shapes for driving fields [33, 34]. Here we implement a two-step optimisation approach, that combines different optimization approaches. Initially, Deep RL (DRL)-like techniques, in conjunction with recurrent neural networks (RNNs), is used to learn the shape of efficient piece-wise constant control pulses, without the requirement for state monitoring [35]. Such key insight is then used to implement a suitable traditional optimization method. This two-step approach yields smooth, analytically well-defined control pulses. An important point to make is that application of such conventional optimization methods without any pre-available information is much more difficult in general due to the ‘curse of dimensionality’ [36]. To succeed, they require the choice of a suitably truncated basis upon which to expand their control functions. This highlights the utility of the initial learning step, which is essentially user-independent and can provide a suitable ansatz without the need for prior knowledge of the system. For example, a requirement for the success of STIRAP is the existence of (a manifold of) adiabatic dark states, and the full knowledge of their structure [37]. On the other hand, for Raman oscillations, the hallmark for adiabatic elimination is the validity of restrictive parameter conditions (such as large detunings), so as to constrain the dynamics to relevant subspaces. The RL-based step discussed here provides protocols that violate both such restrictive conditions, and thus differ from both STIRAP and simple adiabatic elimination, while combining advantages of both to achieve near-optimal dynamics. This thus provides an ansatz for the control that may otherwise not have been arrived at analytically, and whose flexibility could be exploited to engineer operations in multi-node architectures. While delivering previously unforeseen protocols, this hybrid approach to optimisation marks a significant departure from previous methods towards the control of quantum dynamics, and embodies one of the pillars of our proposal.

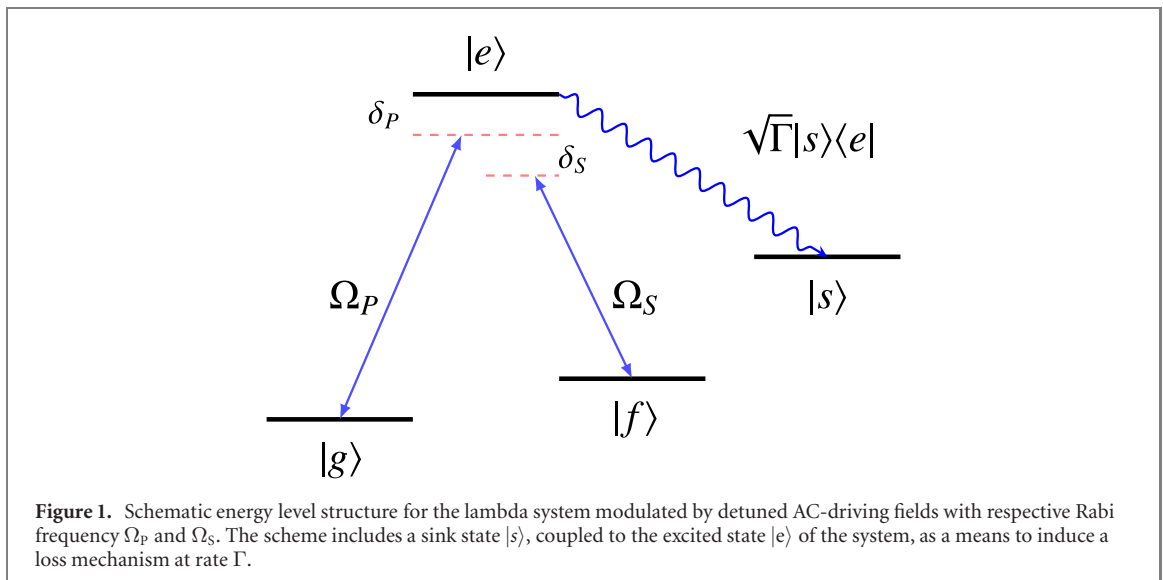
The remainder of this paper is organized as follows. In section 2 we introduce the physical system of interest, which allows us to motivate the specific form of control chosen. In section 3 we show how an RL agent was able to learn control schemes to induce some desired dynamics in the system. Then, in section 4, we use a less sophisticated coefficient optimization over a polynomial basis in an attempt to reproduce the results obtained by the RL approach. In section 5 we use the results from the RL agent in section 3, followed by the simpler coefficient optimization, where we were able to obtain further improvement in protocol efficiency when compared with both methods alone. We then dedicate section 6.1 to an analysis of the resilience of the learned protocols to stochastic decay within the system, where we explicitly consider the performance of both protocols in a three-level ladder system. We finally discuss the robustness of the protocol to low-frequency noise and its resilience to pure dephasing in the system dynamics in section 6.2, followed by a brief discussion of the results in section 7.

## 2. The system

We investigate control protocols for an abstract three-level quantum systems and specifically consider the task of population transfer in so-called lambda systems, where a ground state  $|g\rangle$  and target state  $|f\rangle$  are indirectly coupled via some intermediate excited state,  $|e\rangle$  as shown in figure 1. The states  $|g\rangle$  and  $|f\rangle$  are here considered to be ‘quasi-stable’ ground states, where  $|e\rangle$  is a radiatively decaying excited state. The typical Hamiltonian for this physical system reads

$$H(t) = \frac{\hbar}{2} \begin{pmatrix} 0 & \Omega_p(t) & 0 \\ \Omega_p(t) & 2\delta_p(t) & \Omega_s(t) \\ 0 & \Omega_s(t) & 2\delta(t) \end{pmatrix}. \quad (1)$$

Here  $\Omega_p(t)$  and  $\Omega_s(t)$  represent the Rabi frequencies of the couplings that drive transitions  $|g\rangle \rightarrow |e\rangle$  and  $|e\rangle \rightarrow |f\rangle$  respectively (commonly known as the ‘pump’ and ‘Stokes’ couplings). The term  $\delta_p = (E_e - E_g) - \omega_p$  is referred to as the ‘single-photon’ detuning for the pump driving field with carrier frequency  $\omega_p$ . The  $\delta(t)$  term is the ‘two-photon’ detuning and is defined as  $\delta(t) = \delta_p - \delta_s$ , where  $\delta_s$  is the analogous single-photon detuning for the Stokes coupling. Control of this physical system has been extensively studied in the context of STIRAP [37–39], where for  $\delta \simeq 0$  there exists a suitable control scheme for  $\Omega_p(t)$  and  $\Omega_s(t)$ , the so-called *counterintuitive* pulse sequence, such that perfect transfer from  $|g\rangle$  to  $|f\rangle$  is



achieved while  $|e\rangle$  is kept depopulated at all times. Here we instead consider the case of *always-on* Rabi-frequencies while modulating the single- and two-photon detunings. The population transfer thus achieved mimics protocols in circuit-QED where the couplings between qubit and harmonic mode are not switchable [14]. Specifically, we investigate the case where the couplings  $\Omega_P$  and  $\Omega_S$  both assume the constant value  $\Omega_0$ , while freedom is afforded to modulate the detunings  $\delta_P(t)$  and  $\delta(t)$ , which embody a set of controls of simple experimental manipulation. The remit of our investigation extends beyond the context set by the three-level system illustrated in this section. Indeed, the three-level model considered here can also be used to address the problem of population transfer between two remote quantum resonators both connected by non-switchable couplings to a three-level system, which can be operated locally [13]. Moreover, this configuration also describes a system consisting of two qubits connected by the field of a cavity and working in the single-excitation subspace. In this context, the two low-energy states of the equivalent three-level system would represent states where a single excitation is carried by one of the remote qubits, while the top-most state would imply that the cavity field is populated. This configuration is the building block of cavity-/circuit-QED architectures for controlled quantum dynamics currently being explored experimentally.

### 3. Reinforcement learning based optimization

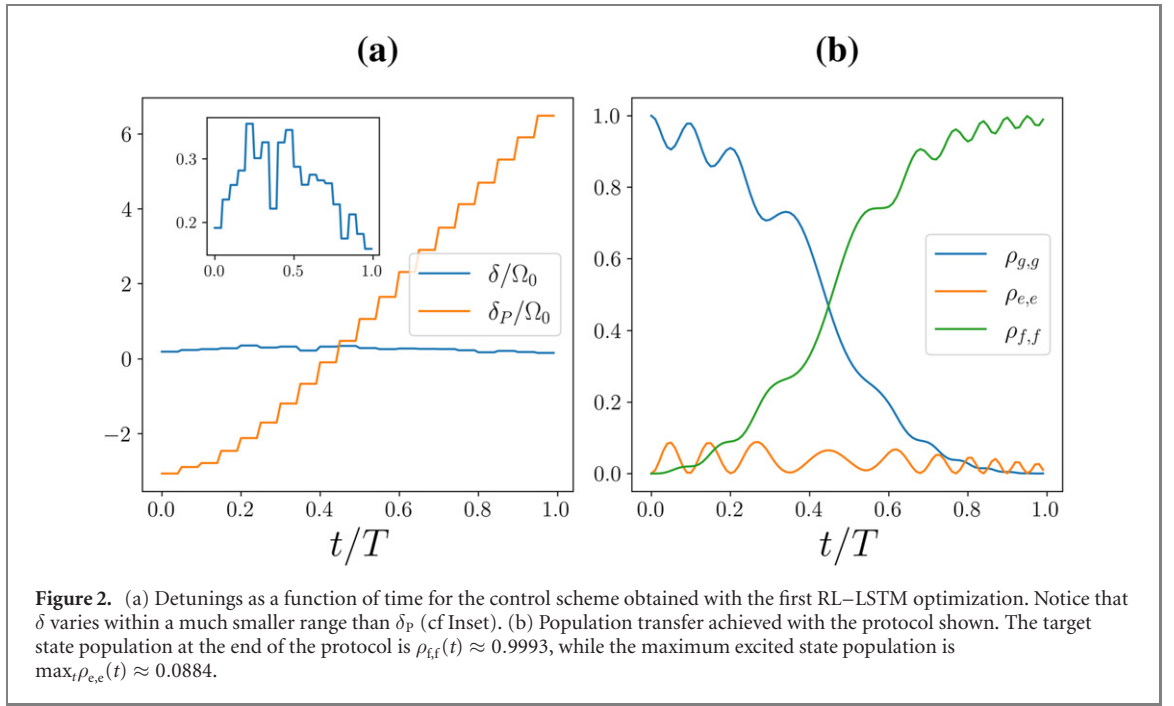
In order to find an efficient control scheme we first employ an RL-inspired approach. Initially, we fix the total time for the system evolution to  $T$  which is then divided into  $N_{\text{steps}}$  time intervals,  $t_i$ , of equal duration. This constitutes one episode. During each of these intervals the one- and two-photon detunings have constant values,  $\delta_P(t_i)$  and  $\delta(t_i)$ , which are all determined by an RL agent prior to each interval. Thus, for each time interval, we use the Hamiltonian in equation (1) with  $t \rightarrow t_i$  and  $\Omega_P = \Omega_S = \Omega_0$ , to evolve the continuous-time open-system dynamics ruled by the Lindblad master equation

$$\dot{\rho} = -\frac{i}{\hbar} [H(t_i), \rho] + \mathcal{D}(\rho), \quad (2)$$

for the duration of the time interval. Here  $\rho$  is the density matrix of the system and  $\mathcal{D}$  is the Lindblad-like operator accounting for the non-unitary part of the dynamics. More specifically, the agent provides two values, for each individual timestep, which act as the mean values of two separate Gaussian policies from which the detuning are sampled at said timestep. Learning is implemented using the policy gradient REINFORCE (with baseline) algorithm for continuous action spaces [40], employing a long short-term memory (LSTM) neural network [41] to as a function approximator (with only the series of time steps  $\{t_i\}_{i=1, \dots, N_{\text{steps}}}$  as external input to the network) mirroring previous work [35].

Thus the agent is tasked with learning a policy that provides the optimal detuning control scheme, where performance is considered with respect to perfect transfer between  $|g\rangle$  and  $|f\rangle$ , while keeping  $|e\rangle$  depopulated at all times. In order to meet such a request we couple the intermediate state to a sink  $|s\rangle$ , in the learning phase only, as shown in figure 1.

This coupling is operationally implemented by introducing the Lindblad operator  $\sqrt{\Gamma}|s\rangle\langle e|$  into the dissipator in equation (2) and induces a decay mechanism in the system, whereby any protocol that



appreciably populates the excited state invariably leads to population loss. This is crucial: removing the state observation at each time-step removes the ability to explicitly define a reward function that encourages the desired dynamics. In this case we can define the delayed reward granted to the RL agent at the end of the evolution as

$$R = \rho_{f,f}(T). \quad (3)$$

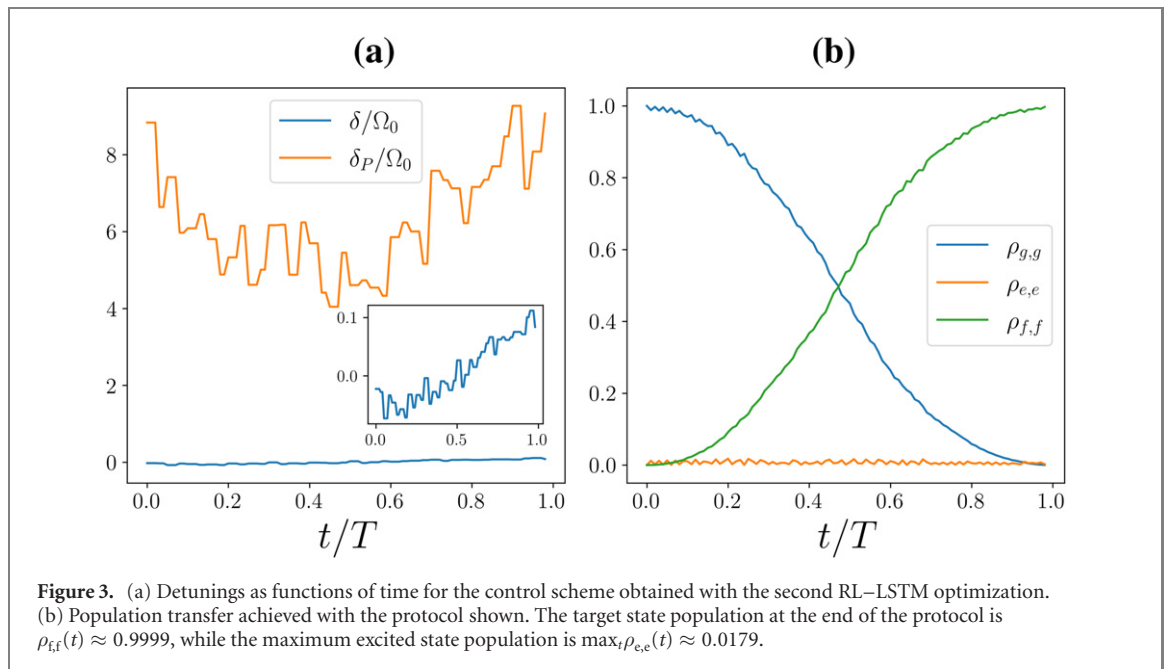
This explicitly promotes population of the final state  $|f\rangle$ , while any transient population of  $|e\rangle$  during the dynamics will act to lower this final population thanks to the aforementioned leakage mechanism. In this sense, punishment for populating  $|e\rangle$  is built-in to the mechanisms of the system via  $|s\rangle$ .

The way this algorithm is able to work without monitoring the system at each time step can be rationalised in the following way. As the LSTM does not monitor the state of the system at each time step, it relies only on the ability to ‘memorize’ the actions that it has taken at each time step leading up to the final reward  $R$ . Thus, over several episodes, the agent is able to build an internal representation of the system dynamics and thus learn to act optimally with only the series of time-steps as input and the final target-state population as feedback. Consequently this type of optimization could in principle be employed as an iterative, closed-loop scheme. Such a key feature of our approach would be beneficial for optimizing control in the presence of difficult-to-simulate environmental decoherence, such as the in the situations faced by solid-state quantum hardware [24, 42]. A detailed explanation of the RL-LSTM approach is provided in appendix A, while the network configuration—along with all the learning parameters—are reported in appendix B.

Using the RL based optimization outlined above, with  $\Omega_0 T = 20$ ,  $N_{\text{steps}} = 20$  and  $(\delta, \delta_P)/\Omega_0 \in [-50, 50]$ , the agent was able to obtain a target state population at the end of the protocol of  $\rho_{f,f}(T) \approx 0.9993$ , with a maximum excited state population over the entire time interval of  $\max_t \rho_{e,e}(t) \approx 0.0884$ . The learned protocol and the induced population dynamics can be found in figure 2. Despite the evidently desirable features of the results thus achieved, it is worth remarking that the learning process is in general stochastic and different runs of the optimization can produce different shapes for the detuning functions. However, successfully optimized detuning functions all shared common traits, which can be summarized by the following list of characteristic features

- C1 We have  $|\delta(t)| \ll |\delta_P(t)|$  for most of the evolution.
- C2 Detuning  $\delta_P(t)$  always exhibits comparatively large initial and final values.
- C3  $\delta_P(t)$  always seems to exhibit specific parity features about  $T/2$ . Such a feature is more sporadically shared by  $\delta(t)$ .

In particular, feature C2 is to be expected if one wants to avoid populating the excited state at the beginning and at the end of the transfer, and agrees with previous findings reported in literature [14]. Furthermore, feature C1 can be justified by inspecting how the presence of non-vanishing detunings affects



**Figure 3.** (a) Detunings as functions of time for the control scheme obtained with the second RL–LSTM optimization. (b) Population transfer achieved with the protocol shown. The target state population at the end of the protocol is  $\rho_{f,f}(t) \approx 0.9999$ , while the maximum excited state population is  $\max_t \rho_{e,e}(t) \approx 0.0179$ .

the efficiency of both standard STIRAP and Raman protocols: while even small non-null values of  $|\delta|$  are detrimental for the performance of the transfer, much larger values of  $\delta_P$  can be tolerated [37, 39, 43].

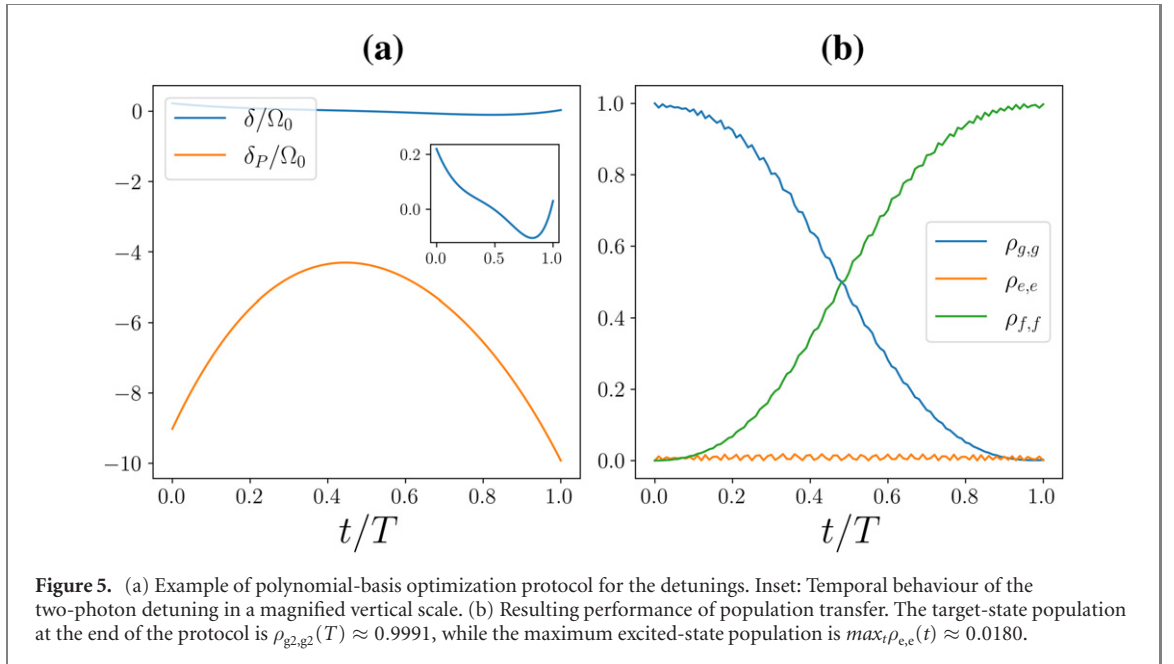
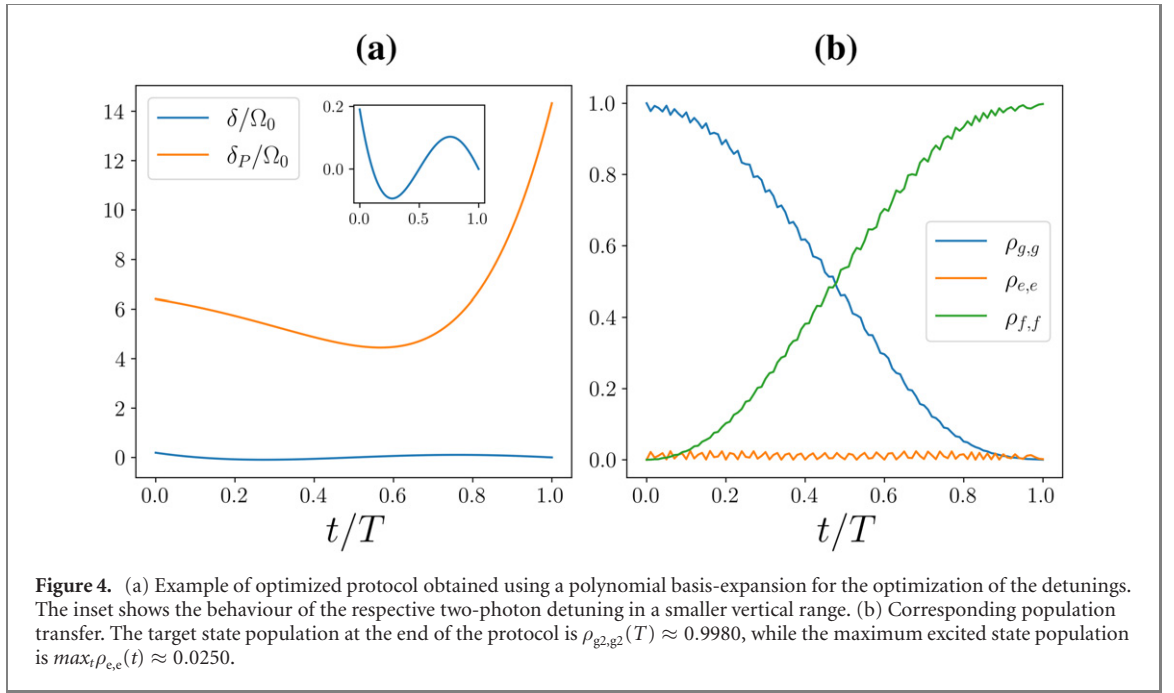
We have performed an optimization process based on the use of a restricted range for the values of  $|\delta|$ , thus limiting the action-space of the RL agent and guaranteeing the validity of **C1**. In particular, we considered  $\delta/\Omega_0 \in [-0.2, +0.2]$  and  $\delta_P/\Omega_0 \in [-14, +14]$ . We have also taken a longer evolution time  $\Omega_0 T = 40$ , with proportionally more steps whereby the agent can act ( $N_{\text{steps}} = 40$ ). The resulting protocol can be seen in figure 3(a). The RL agent obtained a maximum end-protocol target state population of  $\rho_{f,f}(T) \approx 1 - 10^{-4}$ , with a maximum transient excited state population of  $\max_t \rho_{e,e} \approx 0.0179$ . The corresponding dynamics is shown in figure 3(b).

Remarkably, differently from what one would naively expect, this protocol is not akin to a Raman-like or a STIRAP-like one. First, two-photon Raman protocols require large single-photon detunings while, in our case,  $\delta_P$  can even vanish, thus making the dynamics comparatively faster. Second, the protocol that we have found are non-adiabatic, thus making them markedly different from adiabatic population transfers, such as STIRAP. Our LSTM RL approach thus delivers genuinely new protocols that combine features of robustness akin to STIRAP but without requiring the demanding switching of coupling fields

#### 4. Polynomial coefficient optimization

Instead of dividing the time of the evolution in a certain number of steps and optimizing the values of the detunings at each step, an alternative approach for the optimization consists on the expansion of  $\delta(t)$  and  $\delta_P(t)$  over a specific functional basis. The effectiveness of this approach depends on the choice of such basis, making it less general than the technique used in the previous section or other sophisticated optimal control techniques such as CRAB [44]. However, should a suitable basis be found, the suggested approach translates the problem of finding the best protocol into a simpler numerical optimization over the coefficient of the expansion while also providing us with a simple analytical expression for the control terms.

We found that writing  $\delta(t)$  and  $\delta_P(t)$  as 5th order polynomial functions and using a Powell method search [45] over the coefficients of the polynomial expansion is enough to achieve an effective population transfer. In figures 4 and 5 we show the best protocols obtained after 10 different runs of the optimization for  $\Omega_0 T = 40$ . It can be seen that, while still effective, they are different from the protocol found via the RL-based optimization (although conditions **C1** and **C2** found by the RL agent can still be observed). This again suggests that various quasi-optimal protocols can be identified as candidates for an efficient population transfer. However, the effectiveness of such optimization technique depends on the choice of the basis for the specific problem. Performing a simple numerical optimization to solve the same problem assigned to the RL agent (finding the values of piecewise constant functions) gives us far worse solutions compared to those obtained using the the RL-based approach [35]. Therefore, not only the RL-based approach can be successfully applied to a wider class of problems with a simpler pre-optimization analysis



but it also provides a better exploratory tool when only sub-optimal solutions are achieved, as these solutions are not biased by the choice of a specific basis of functions.

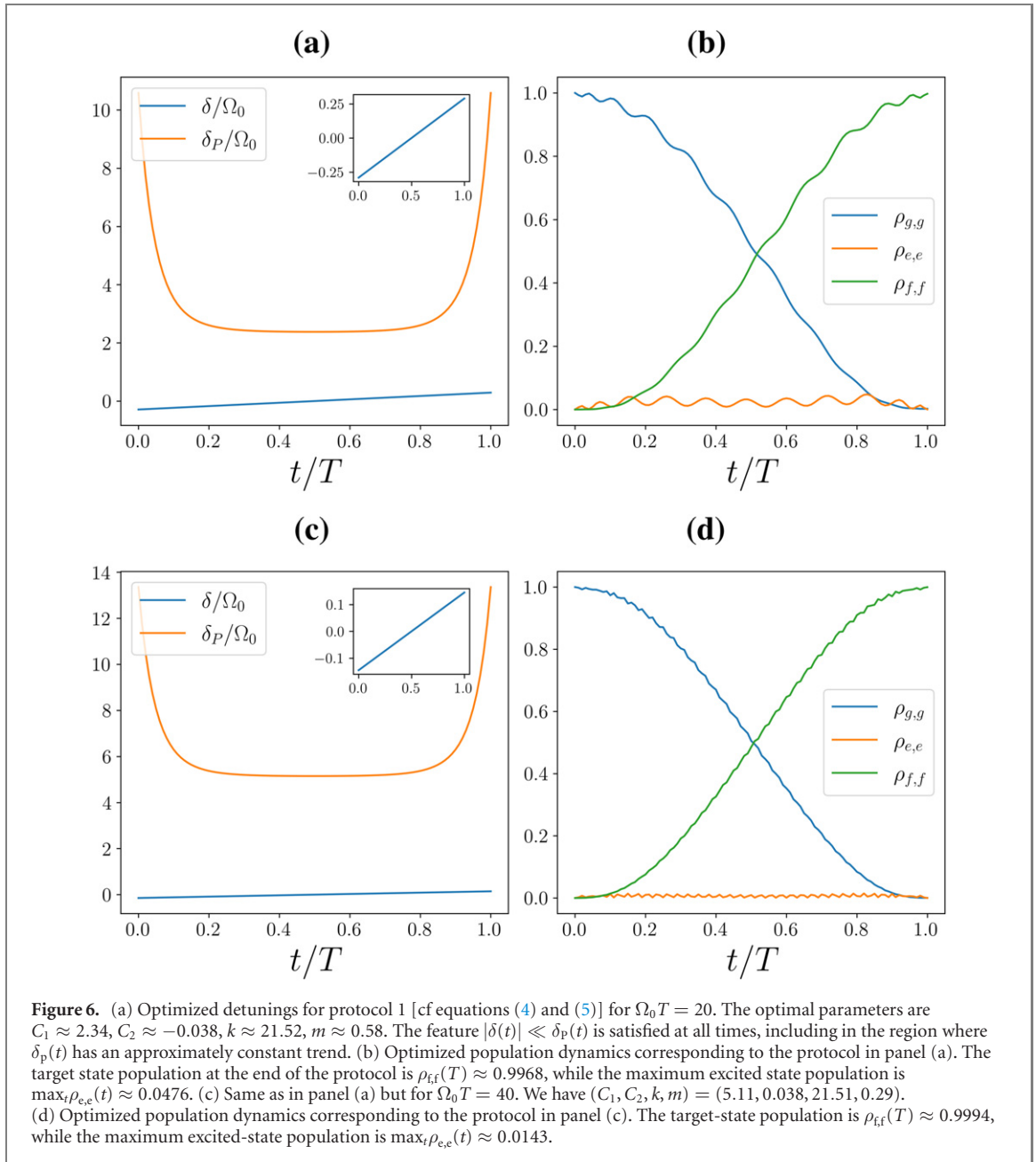
## 5. Optimal protocols

Based on the success of both the RL-based optimization and the optimization with a polynomial basis, we combined the two approaches, performing a straightforward numerical optimization starting from the results of the RL-based technique. To this end, observing the features in figure 3, we propose an ansatz for  $\delta_P(t)$  as

$$\frac{\delta_P(t)}{\Omega_0} = C_1 - C_2 e^{k\left(\frac{t}{T} - 0.5\right)^2}. \quad (4)$$

Similarly, we suggest the linear ansatz for  $\delta(t)$ ,

$$\frac{\delta(t)}{\Omega_0} = m \left( \frac{t}{T} - 0.5 \right). \quad (5)$$

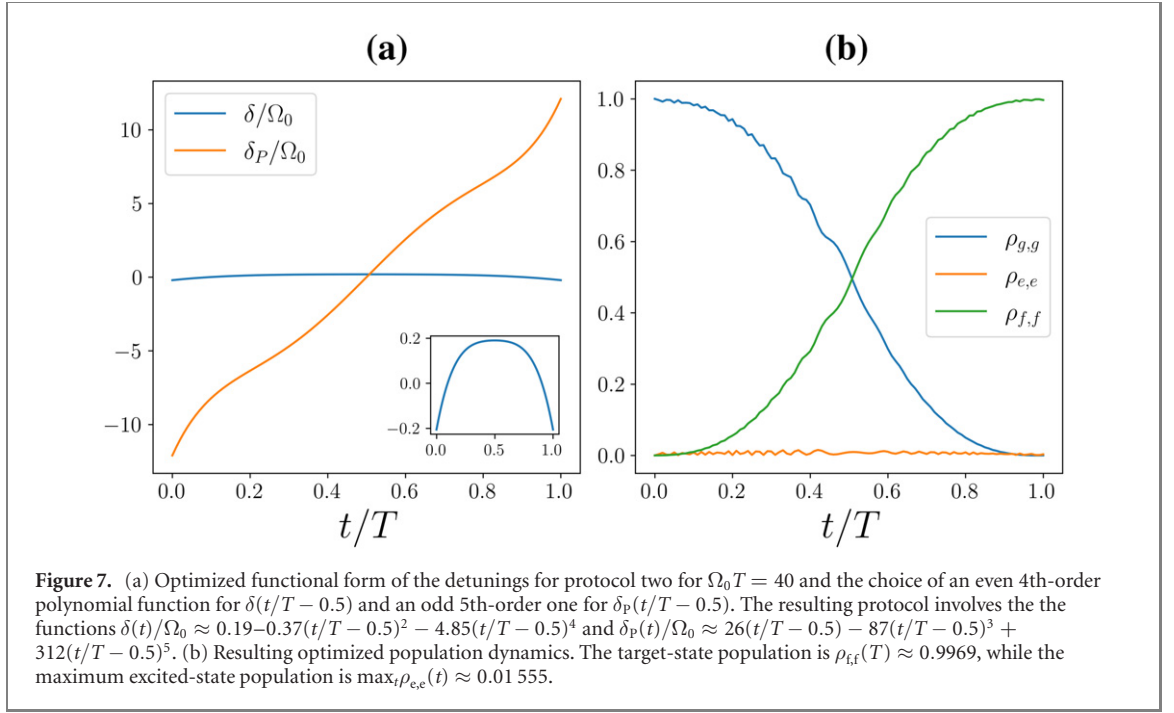


The choice of equations (4) and (5) ensure that the symmetry or anti-symmetry point of the proposed functions occur at  $t = T/2$  of the evolution.

This optimization was carried out using a Powell method search [45] over the space of parameters  $(C_1, C_2, k, m)$  for the maximization of  $R$ . The benefit here is two-fold: on one hand, it allows us to find an analytical expression for the protocol, thus contributing to the interpretation of the results that we achieve; on the other hand, it smooths the protocol found by the RL agent, presenting us with a continuous control scheme, which is experimentally more tractable. In achieving these two goals the analytic, smooth control pattern maintained a comparable final state target population to the RL learned scheme, while further reducing the transient population of the excited state. Specifically, in figure 6 we present results for  $\Omega_0 T = 20$  and  $\Omega_0 T = 40$  showing that, for the second case,  $\rho_{f,f}(T) \approx 0.9994$  and  $\max_t \rho_{e,e}(t) \approx 0.0143$  can be achieved with the simple ansatz that we have proposed.

The insight provided by the RL-based optimization approach suggests the existence of different valid protocols of optimization. In this regard, an interesting question to pose addresses the role of the parity exhibited by the detuning functions with respect to  $t = T/2$ . That is, we wonder whether optimal functional behaviours akin to those exhibited in figure 2 can be identified. To ascertain it, we propose the use of an odd 5th order polynomial function for  $\delta_p(\frac{t}{T} - 0.5)$  and an even 4th order polynomial for  $\delta(\frac{t}{T} - 0.5)$  and performed a similar optimization, finding that the corresponding optimized protocol is still effective (cf figure 7). The resulting final target-state population is  $\rho_{f,f} \approx 0.9969$ , while the maximum excited-state





population is  $\max_t \rho_{e,e}(t) \approx 0.01555$ . For brevity, we label the protocol of figures 6(c) and (d) as protocol 1, while that of figure 7 will be referred to as protocol 2. We point out that the performances of protocol 1 and protocol 2 mentioned here are extremely similar to the RL protocol of figure 3. Optimality can thus be understood in terms of the evident simplicity of the control functions needed to achieve such performance

The behaviours showcased in our results allow us to corroborate quantitatively the differences between our protocols and Raman-like ones. The first clear difference is the absence of Raman oscillations (cf figure 9) from the dynamics of the populations resulting from our protocols. A second difference between the two approaches stems from the fact that in protocol 1  $\delta_p$  is constant most of the time and we get  $\delta(t) \ll \delta_p(t)$ . One can then ask how this compares to a Raman scheme with  $\delta = 0$  and a constant  $\delta_p \gg \Omega_0$ . When no constraint is imposed over the total time of the evolution, one would expect that increasing  $\delta_p$  will progressively improve the transfer. However, our approach assumes a fixed value of  $\Omega_0 T$ . This implies that a very large value of  $\delta_p$  could prevent the completion of the corresponding very slow population transfer. Both of these effects are relevant for the optimal choice of  $\delta_p$ . In figure 8 we show that protocol 1 achieves a more efficient population transfer relative to the case of a completely constant Hamiltonian. Moreover, in line with previous considerations, we also remark that the protocols are not adiabatic. If we increase the total time of the evolution while still using protocol 1 and 2 (without performing a new optimization for each value of  $\Omega_0 T$ ), the performance does not increase monotonically, as it would happen in a Raman protocol (figure 9).

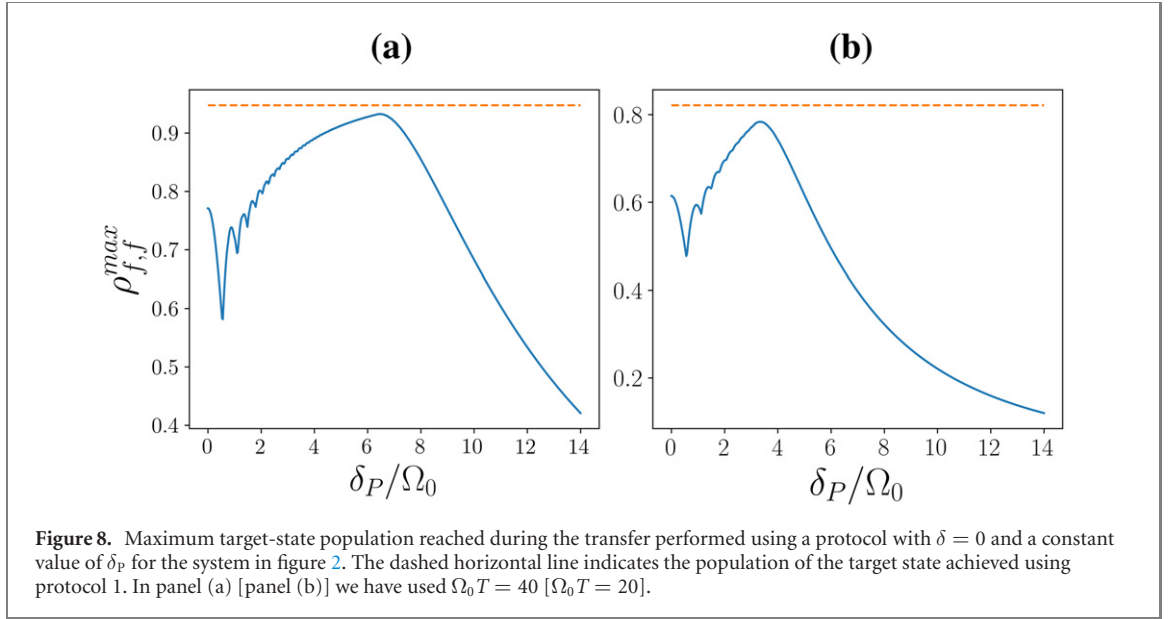
## 6. Resilience to decoherence

### 6.1. Spontaneous decay

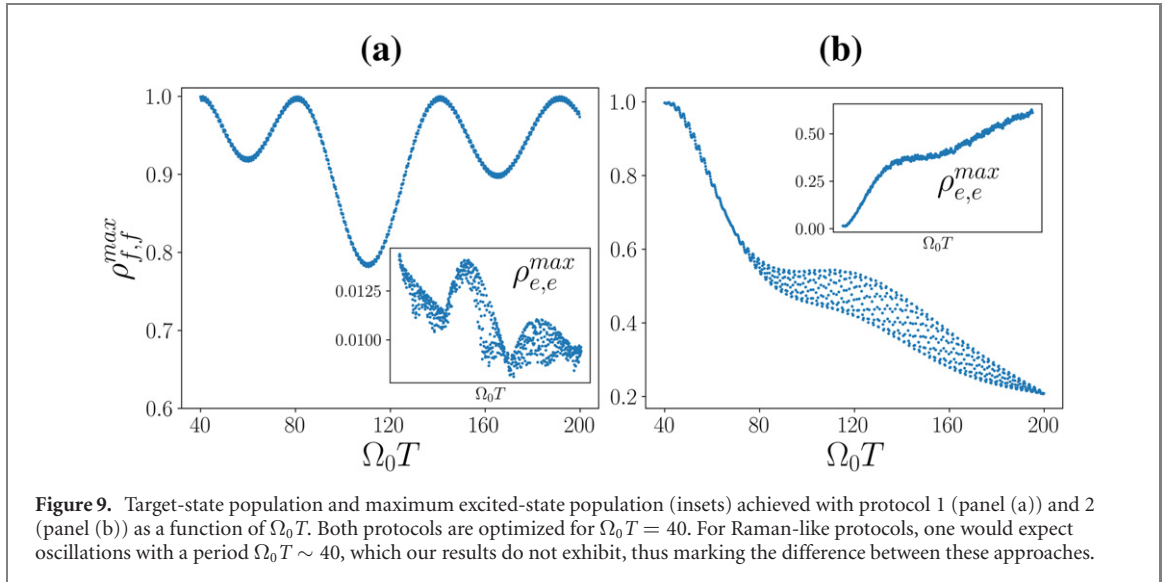
Here we consider how the protocols that we have found perform when a multi-level system is subjected to spontaneous decay from some of its energy levels. We investigate two cases:

- A The decay from the intermediate excited state  $\rho_{e,e}$  to the first ground state  $\rho_{g,g}$  with a decay rate  $\gamma_{e,g}$ , implemented using the Lindblad operator  $\sqrt{\gamma_{e,g}}|g\rangle\langle e|$ .
- B The case of an additional decay channel, from  $\rho_{f,f}$  to  $\rho_{e,e}$ , with rate  $\gamma_{f,e}$ , implemented by  $\sqrt{\gamma_{f,e}}|e\rangle\langle f|$ .

Scenario **A** is what one would expect to be relevant for the Lambda system that we have discussed thus far, particularly when fluxonium-based embodiments of the multi-level system are considered [46], for which an incoherent mechanism driving decay of population from  $|f\rangle$  to  $|g\rangle$  can be safely neglected. On the other hand, scenario **B** is motivated by the fact that the Hamiltonian in equation (1) encapsulates the so-called Ladder energy level structure, where  $|f\rangle$  becomes a higher excited state than  $|e\rangle$  and is thus susceptible to spontaneous decay. The Lambda and Ladder scenarios are operationally equivalent as far as the control protocols are concerned. A diagrammatic outline of these two scenarios is shown in figure 10.



**Figure 8.** Maximum target-state population reached during the transfer performed using a protocol with  $\delta = 0$  and a constant value of  $\delta_P$  for the system in figure 2. The dashed horizontal line indicates the population of the target state achieved using protocol 1. In panel (a) [panel (b)] we have used  $\Omega_0 T = 40$  [ $\Omega_0 T = 20$ ].



**Figure 9.** Target-state population and maximum excited-state population (insets) achieved with protocol 1 (panel (a)) and 2 (panel (b)) as a function of  $\Omega_0 T$ . Both protocols are optimized for  $\Omega_0 T = 40$ . For Raman-like protocols, one would expect oscillations with a period  $\Omega_0 T \sim 40$ , which our results do not exhibit, thus marking the difference between these approaches.

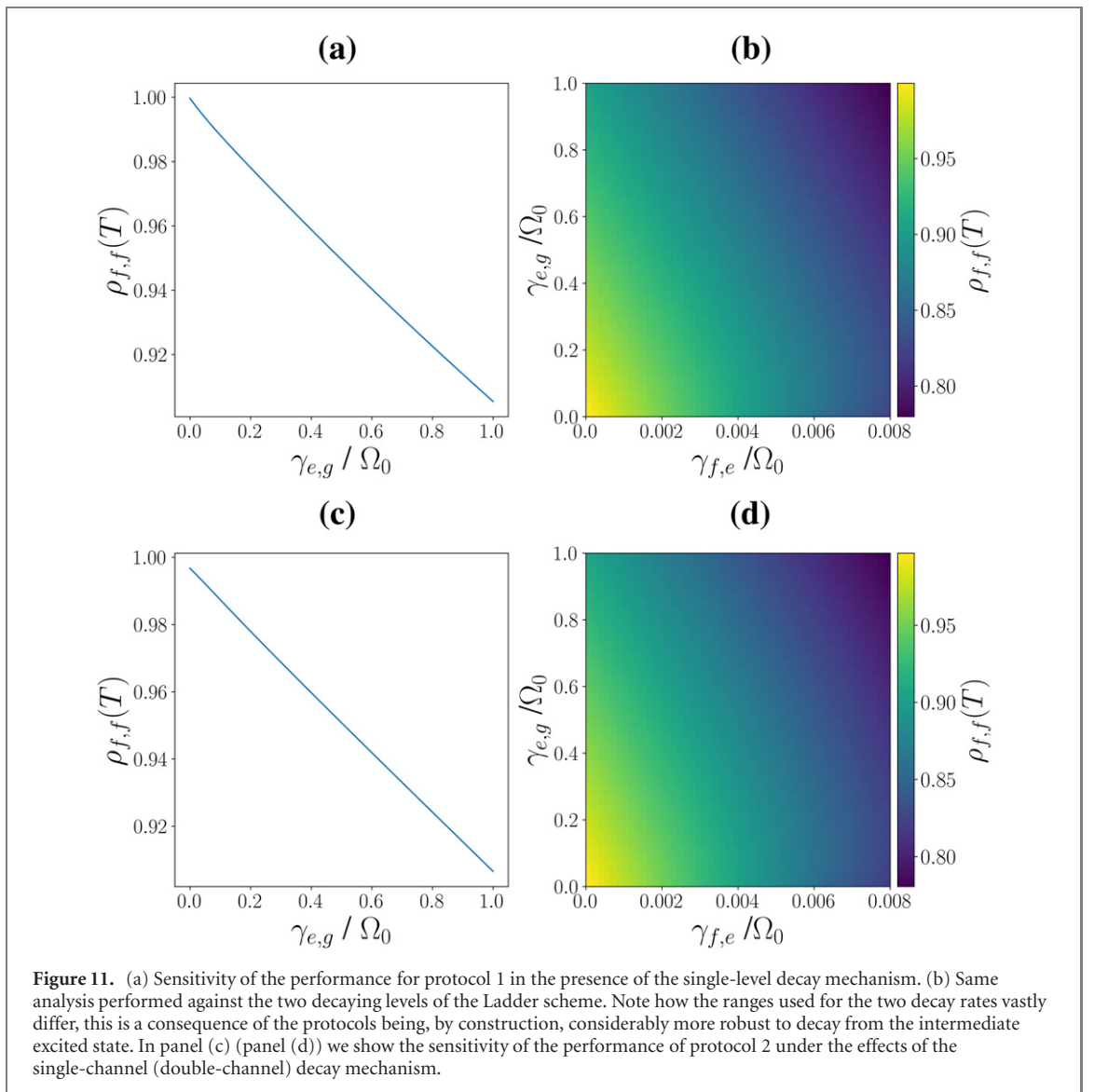
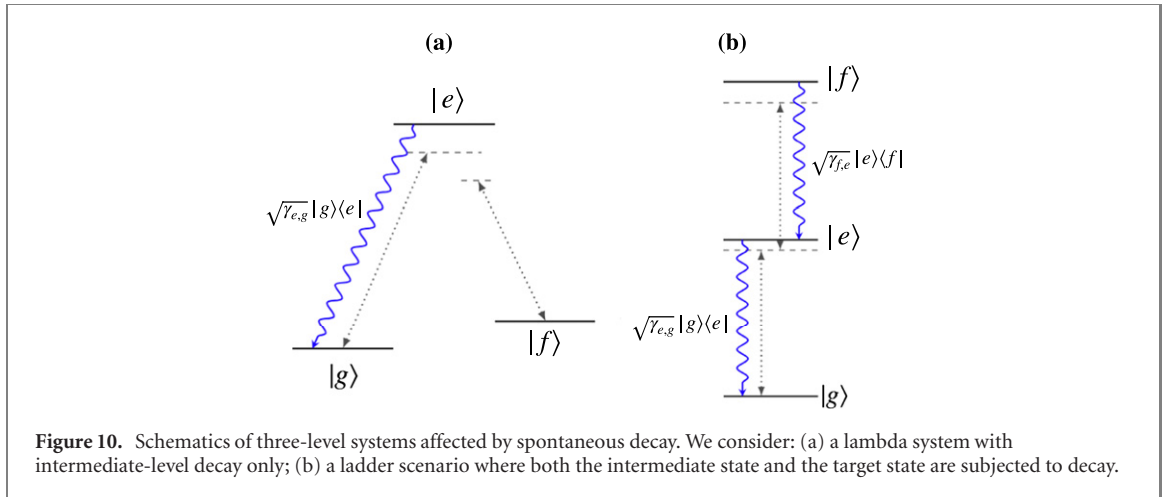
We consider the sensitivity of the protocols, with respect to final target state population  $\rho_{f,f}(T)$ , for a range of decay rates in both cases. Figure 11 shows how the performance of protocols 1 and 2 respectively depend on the strength of the decay rates in each of the lambda and ladder cases, where performance is gauged simply by the final target state population.

It can be appreciated that both protocols carry a strikingly similar dependence on the decay rates and exhibit relative robustness against decay from the intermediate state. This is to be expected the RL process included a mechanism to punish population of such level. On the other hand, both protocols exhibit great sensitivity to decay from the target state. Thus, in a ladder system, a decaying target state embodies the main limiting factor.

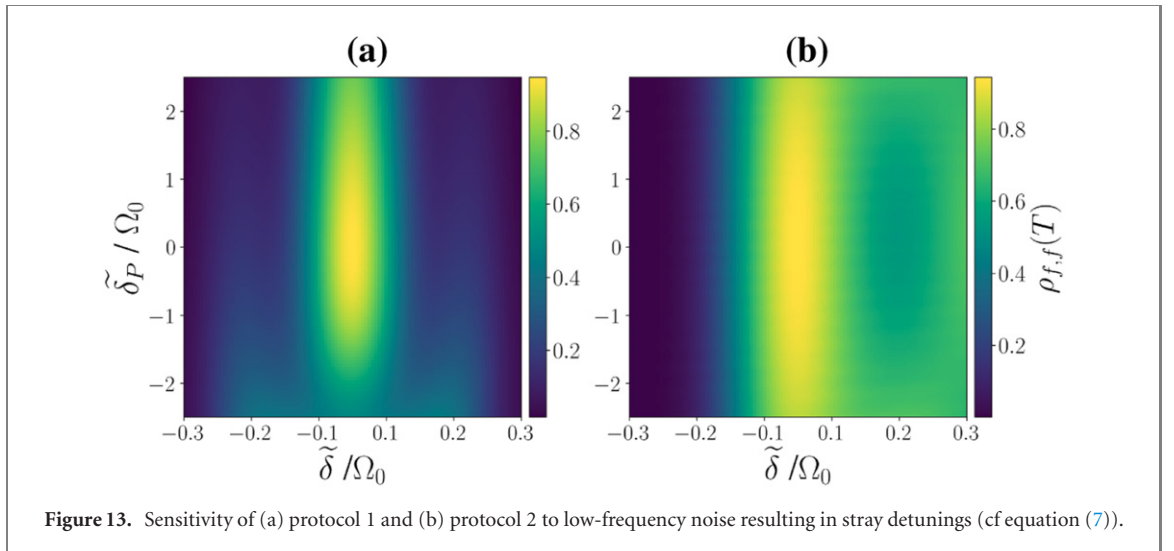
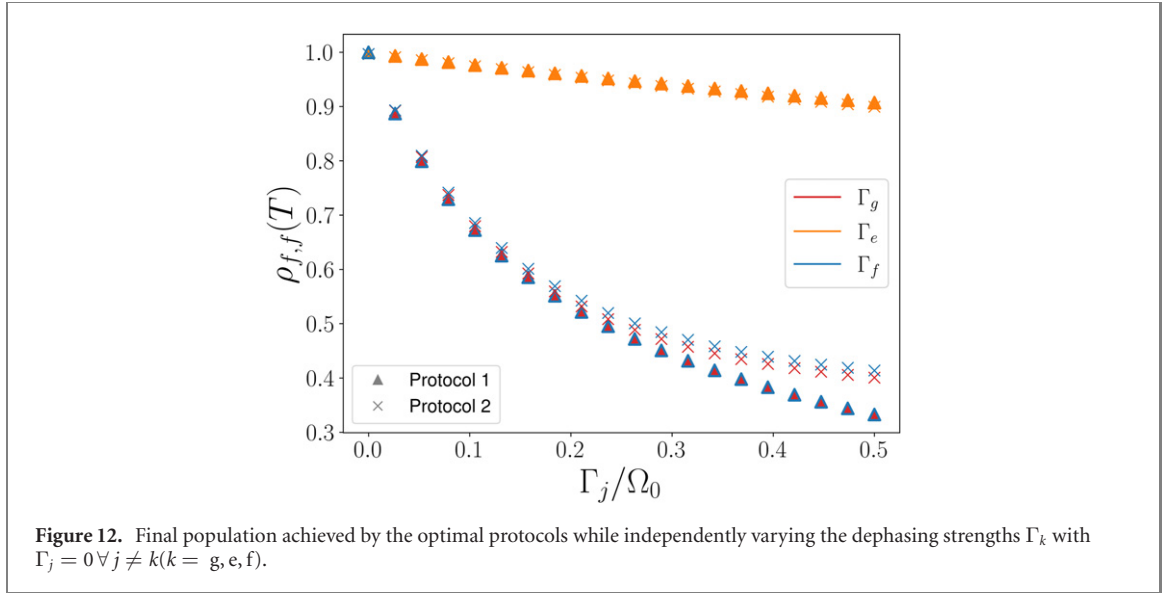
## 6.2. Dephasing

We extend the previous analysis with the study of the behaviour of protocol 1 and 2 under the effects of dephasing. While sophisticated models can be invoked to illustrate the various facets of dephasing, in order to gather an understanding of its implications for the protocols identified here, we focus on pure dephasing implemented using the Lindblad operators  $L_k = \sqrt{\Gamma_k}|k\rangle\langle k|$ , ( $k \in \{e, f\}$ ) entering equation (2) with

$$\mathcal{D}(\rho) = \begin{pmatrix} 0 & \Gamma_{ge}\rho_{ge} & \Gamma_{gf}\rho_{gf} \\ \Gamma_{ge}\rho_{eg} & 0 & \Gamma_{ef}\rho_{ef} \\ \Gamma_{gf}\rho_{fg} & \Gamma_{ef}\rho_{fe} & 0 \end{pmatrix}, \quad (6)$$



where  $\Gamma_{kl} = \Gamma_k + \Gamma_l$  and  $\rho_{kl} = \langle k|\rho|l\rangle$  ( $k, l = g, e, f$ ). We are now able to investigate the sensitivity of the protocols 1 and 2 with respect to such mechanism, an analysis that we perform by independently varying the values of one of the  $\Gamma_k$ 's, while keeping the other at zero. The relationship between protocol efficiency and each of these dephasing rates can be inspected from figure 12. Owing to the two-photon character of the protocols at hand, we find much higher sensitivity to non-zero  $\Gamma_g$  and  $\Gamma_f$ , while being comparatively



resilient to  $\Gamma_e$ . In terms of equation (6), this translates into a much larger sensitivity to  $\Gamma_{gf}$  relative to  $\Gamma_{ge}$ ,  $\Gamma_{ef}$ . Needless to say, this is a consequence of the protocols having been optimized to constrain the system dynamics to the subspace  $\{|g\rangle, |f\rangle\}$  of the full Hilbert space and as such being most reliant on coherence between the initial and target states.

### 6.3. Robustness against low-frequency noise

We conclude our assessment of the robustness of the proposed protocols by addressing the sensitivity to detunings. This analysis is particularly relevant for superconducting devices, where the main dephasing mechanism can be attributed to the presence of a low frequency noise that often has a  $1/f$  spectrum characterized by slow fluctuations of the detunings [42, 43]. Due to the slowness of the dynamics of such fluctuations, the value of the detunings induced by such low frequency noise can be considered as constant during the population transfer. Hence, a simple way to achieve a meaningful and informative characterization of its effects on our protocols is to study their performance when we include a constant perturbation in each of the detunings. We thus take

$$\begin{aligned} \delta_P(t) &\rightarrow \delta_P(t) + \tilde{\delta}_P \\ \delta(t) &\rightarrow \delta(t) + \tilde{\delta}. \end{aligned} \quad (7)$$

In the following analysis, we have considered both such constant perturbations and the leakage mechanism outlined in section 3. We have used the final target state population  $\rho_{f,f}(T)$  as a measure of performance. From figure 13 it can be seen that both protocols are almost insensitive to the single-photon detuning  $\tilde{\delta}_P$ ,

while the sensitivity to two-photon detuning is larger, as expected, and comparable to that of STIRAP. Interestingly, protocol 2 results in  $\rho_{i,f}(T)$  being strongly asymmetric with respect to the sign of the perturbation to the two photon detuning  $\tilde{\delta}$ .

## 7. Conclusions

We have successfully employed a combination of a RL-based methods and more traditional optimization techniques to achieve optimal population transfer in a three-level system, while operating in an experimentally relevant control regime. Further, we have highlighted that our technique can in principle be implemented as an iterative, closed-loop optimization. Its use will be beneficial in all those situations where the underlying decoherence mechanisms are not fully understood. We have also demonstrated that even when a RL-based approach gives us sub-optimal solutions, it can still provide a useful tool that can be used to build better protocols through a simpler numerical optimization techniques. The approach produced two novel protocols which remarkably differ from other control methods such as STIRAP, standard Raman or adiabatic schemes while exhibiting comparable performance and robustness. To this end, it is worth noticing that, due to the specific constraints of the protocol, STIRAP cannot be operated with both always-on couplings.

Several works in the few years proposed the implementation of multi-level systems, including both lambda and ladder configurations, using superconducting artificial nanostructures subjected to suitable driving configurations. These arrangements, though, expose the nanostructure to increased noise level, which severely affects the performance of population transfer, limiting it to values that are typically in the range of 70%. Our approach will be invaluable to enhance the performance of such systems above and beyond the possibilities offered by demonstrated techniques for quantum control. In particular, our approach minimizes the need for the use of switchable coupling mechanisms, with is a key advantage when having in mind the design of robust schemes with low hardware overhead in the noisy intermediate-scale quantum technology framework. By serving both as an alternative control scheme for the specific physical system discussed above, and a proof-of-concept for the optimization technique itself, our protocols could be exported to be used in other relevant context, from quantum simulation to gate engineering.

## Acknowledgments

This work was supported by the Northern Ireland Department for Economy (DfE), the EU H2020 framework through Collaborative Projects TEQ (Grant Agreement No. 766900), the DfE-SFI Investigator Programme (Grant No. 15/IA/2864), the Leverhulme Trust Research Project Grant UltraQute (Grant No. RGP-2018-266), COST Action CA15220, the Royal Society Wolfson Research Fellowship scheme (RSWF\R3\183013) and International Mobility Programme, the UK EPSRC (Grant No. EP/T028106/1), the Academy of Finland (Finnish Center of Excellence in Quantum Technology QTF projects 312296, 336810, and RADDESS programme project 328193), Grant No. FQXi-IAF19-06 ('Exploring the fundamental limits set by thermodynamics in the quantum regime') of the Foundational Questions Institute Fund (FQXi), the QuantERA grant SiUCs (Grant No. 731473 QuantERA), and by University of Catania, Piano per la Ricerca 2016-18 - linea di intervento 'Chance', Piano di Incentivi per la Ricerca di Ateneo 2020/2022, proposal Q-ICT.

## Data availability statement

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

## Appendix A. RL-based optimization with LSTM neural network

Traditional RL focuses on solving Markov decision processes (MDPs). In a MDP, the state of the environment (and the agent observation) at each time step and the corresponding action taken by the agent uniquely determine the state of the environment at the next time step [40].

If we now consider our physical problem where the agent is trying to learn the optimal Hamiltonian of the system (under the given constraints) and the Lindblad operators are not influenced by the agent actions,

the natural choice to define a MDP would be to take the density matrix of the system as the agent observation.

Based on this input, we can use a function approximator (i.e. a neural network) to predict the mean values  $\tilde{\mu}_\theta^\delta, \tilde{\mu}_\theta^{\delta_p}$  of Gaussian distributions (with standard deviation  $\sigma$ ) whose product constitute the policy function from which we sample the actions of our agent.

If a reward  $R$  is granted to the agent at the end of the system dynamics, policy gradient REINFORCE [40] can be implemented by training the neural network with a cost function  $C = \frac{1}{2\sigma^2} \sum_{a_i} R |a_i - \tilde{\mu}_\theta(s_i)|^2$ , where  $\tilde{\mu}_\theta = (\tilde{\mu}_\theta^\delta, \tilde{\mu}_\theta^{\delta_p})$  and  $a_i = (\tilde{\delta}(t_i), \tilde{\delta}_p(t_i))$ , with  $\tilde{\delta}(t_i), \tilde{\delta}_p(t_i)$  detunings normalized with respect to their maximum values (defined by their ranges).

Easy improvements of this algorithm can be achieved by working with a batch of agents in parallel (instead of a single agent) and training the network with stochastic gradient descent (or more advanced and similar techniques such as Adam [47, 48]) and subtracting a baseline to the reward (in our work we subtract the average value of the reward over the batch).

This approach is effective for planning, when one can simulate the system dynamics, but it is extremely limiting as a control technique that works with real experimental data, since it requires full quantum tomography for each step of the MDP. Since measurements on a quantum system perturb the system dynamics, a good control technique would require to take measurements only at the beginning and at the end of the system evolution. Such control technique, if effective, would be more powerful than a simple application of RL to quantum systems, as it would be useful even when we are not able to simulate the system dynamics (e.g. when the noise mechanism is not fully understood).

Since all the other parameters of the evolution of the system are fixed, the reward that the agent gets at the end of the process is uniquely determined by the agent actions, as the evolution of the density matrix is deterministic. Hence, in principle, the decision process in which the agent receives informations about its previous actions (that now constitute the agent observation) can be solved by means of RL techniques (and policy gradient in particular) and while defining the observation as a list of all these actions is unpractical and likely ineffective compared to other optimization techniques, we can still pursue this approach by making use of a RNN as function approximator.

RNNs are neural networks specifically designed for sequential data and especially useful for time sequences. In a RNN, the output associated with each element of the input sequence depends on all the previous inputs and outputs of the network and hence this implicitly implements the desired feature. In particular we chose to use a LSTM neural network that takes as external inputs only the time at which the agent is operating (details of the configuration can be found in appendix B).

Comparison with standard numerical optimization techniques has been carried out in reference [35]. There, it has been shown that this approach requires a smaller number of experiments to achieve optimal protocols and shows better performances when one increases the number of control terms and the dimension of the system.


## Appendix B. Optimization parameters

For the RL-based approach, we considered a batch of  $N_{\text{batch}}$  agents for  $N_{\text{epochs}}$  epochs of learning and we used the ‘Adam’ optimizer [47, 48] to train the Neural Network. The baseline considered is the average value of the reward over the batch. The Neural Network consists in 50 LSTM units [48] followed by a dense hidden layer of 30 neurons with a Hyperbolic Tangent activation function and an output layer with the same activation function. The standard deviation of the Gaussian distribution from which the detunings are sampled is fixed to  $\sigma = 0.001$  for  $(\delta/\Omega_0, \delta_p/\Omega_0) \in [-50, +50]$  and to  $\sigma = 0.07$  for  $\delta/\Omega_0 \in [-0.2, +0.2]$  and  $\delta_p/\Omega_0 \in [-14, +14]$ . The numbers of agents in the batch are, respectively  $N_{\text{batch}} = 100$  and  $N_{\text{batch}} = 50$ , and the number of epochs is  $N_{\text{epochs}} = 350$ . The initial condition for the polynomial coefficients for the numerical optimization (Powell method) are extracted from a uniform random distribution in the interval  $[-20, 20]$  while for the optimization of protocol 1 and protocol 2 we used the intervals  $[-5, 5]$  and  $[0, 20]$  (for all the parameters), respectively. Throughout this work, we fixed  $\Gamma T = 10$ .

## ORCID iDs

Jonathon Brown  <https://orcid.org/0000-0001-9113-4781>

Gheorghe Sorin Paraoanu  <https://orcid.org/0000-0003-0057-7275>

Mauro Paternostro  <https://orcid.org/0000-0001-8870-9134>

## References

- [1] Nielsen M A and Chuang I L 2000 *Quantum Computation and Quantum Information Cambridge Series on Information and the Natural Sciences* (Cambridge: Cambridge University Press)
- [2] Shor P W 1999 Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer *SIAM Rev.* **41** 303–32
- [3] Grover L K 1996 A fast quantum mechanical algorithm for database search *Proc. Twenty-Eighth Annual ACM Symp. on Theory of Computing* pp 212–9
- [4] Kjaergaard M, Schwartz M E, Braumüller J, Krantz P, Wang J I-J, Gustavsson S and Oliver W D 2020 Superconducting qubits: current state of play *Annu. Rev. Condens. Matter Phys.* **11** 369–95
- [5] Devoret M H and Schoelkopf R J 2013 Superconducting circuits for quantum information: an outlook *Science* **339** 1169
- [6] DiCarlo L et al 2009 Demonstration of two-qubit algorithms with a superconducting quantum processor *Nature* **460** 240–4
- [7] McKay D C, Filipp S, Mezzacapo A, Magesan E, Chow J M and Gambetta J M 2016 Universal gate for fixed-frequency qubits via a tunable bus *Phys. Rev. Appl.* **6** 064007
- [8] Caldwell S A et al 2018 Parametrically activated entangling gates using transmon qubits *Phys. Rev. Appl.* **10** 034050
- [9] Blais A, Gambetta J, Wallraff A, Schuster D I, Girvin S M, Devoret M H and Schoelkopf R J 2007 Quantum-information processing with circuit quantum electrodynamics *Phys. Rev. A* **75** 032329
- [10] Blais A, Grimsmo A L, Girvin S M and Wallraff A 2020 Circuit quantum electrodynamics (arXiv:2005.12667[quant-ph])
- [11] Arute F et al 2019 Quantum supremacy using a programmable superconducting processor *Nature* **574** 505
- [12] Siewert J, Brandes T and Falci G 2009 *Phys. Rev. B* **79** 024504
- [13] Falci G, Di Stefano P G, Ridolfo A, D'Arrigo A, Paraoanu G S and Paladino E 2017 Advances in quantum control of three-level superconducting circuit architectures *Fortschr. Phys.* **65** 1600077
- [14] Di Stefano P G, Paladino E, D'Arrigo A and Falci G 2015 Population transfer in a lambda system induced by detunings *Phys. Rev. B* **91** 224506
- [15] You J Q and Nori F 2011 Atomic physics and quantum optics using superconducting circuits *Nature* **474** 589
- [16] Vepsäläinen A, Danilin S and Paraoanu G S 2019 Superadiabatic population transfer in a three-level superconducting circuit *Sci. Adv.* **5** eaau5999
- [17] Kumar K S, Vepsäläinen A, Danilin S and Paraoanu G S 2016 Stimulated Raman adiabatic passage in a three-level superconducting circuit *Nat. Commun.* **7** 10628
- [18] Blok M S et al 2021 Quantum information scrambling on a superconducting qutrit processor *Phys. Rev. X* **11** 021010
- [19] D'Alessandro D 2007 *Introduction to Quantum Control and Dynamics* (London: Chapman & Hall)
- [20] Glaser S J et al 2015 Training Schrödinger's cat: quantum optimal control *Eur. Phys. J. D* **69** 279
- [21] Abdelhafez M, Schuster D I and Koch J 2019 Gradient-based optimal control of open quantum systems using quantum trajectories and automatic differentiation *Phys. Rev. A* **99** 052327
- [22] Mohamed A, Baker B, Gyenis A, Mundada P, Houck A A, Schuster D and Koch J 2020 Universal gates for protected superconducting qubits using optimal control *Phys. Rev. A* **101** 022321
- [23] Xu J, Li S, Chen T and Xue Z-Y 2020 Nonadiabatic geometric quantum computation with optimal control on superconducting circuits *Front. Phys.* **15** 1–8
- [24] Werninghaus M, Egger D J, Roy F, Machnes S, Wilhelm F K and Filipp S 2021 Leakage reduction in fast superconducting qubit gates via optimal control *npj Quantum Inf.* **7** 14
- [25] Propson T, Jackson B E, Koch J, Manchester Z and Schuster D I 2021 Robust quantum optimal control with trajectory optimization (arXiv:2103.15716)
- [26] Di Stefano P G, Paladino E, Pope T J and Falci G 2016 Coherent manipulation of noise-protected superconducting artificial atoms in the lambda scheme *Phys. Rev. A* **93** 051801
- [27] Vasilev G S, Kuhn A and Vitanov N V 2009 Optimum pulse shapes for stimulated Raman adiabatic passage *Phys. Rev. A* **80** 013417
- [28] Giannelli L and Arimondo E 2014 Three level superadiabatic quantum driving *Phys. Rev. A* **89** 033419
- [29] Sivak V V, Eickbusch A, Liu H, Royer B, Tsioutsios I and Devoret M H 2021 Model-free quantum control with reinforcement learning (arXiv:2104.14539)
- [30] Haug T, Mok W-K, You J-B, Zhang W, Eng Png C and Kwek L-C 2020 Classifying global state preparation via deep reinforcement learning *Mach. Learn.: Sci. Technol.* **2** 01LT02
- [31] Kuo E-J, Fang Y-L L and Chen S Y-C 2021 Quantum architecture search via deep reinforcement learning (arXiv:2104.07715)
- [32] Zheng A, Song H-J, He Q-K and Zhou D L 2021 Quantum optimal control of multilevel dissipative quantum systems with reinforcement learning *Phys. Rev. A* **103** 012404
- [33] Paparella I, Moro L and Prati E 2020 Digitally stimulated Raman passage by deep reinforcement learning *Phys. Lett. A* **384** 126266
- [34] Porotti R, Tamascelli D, Restelli M and Prati E 2019 Coherent transport of quantum states by deep reinforcement learning *Commun. Phys.* **2** 61
- [35] Sgroi P, Palma G M and Paternostro M 2021 Reinforcement learning approach to non-equilibrium quantum thermodynamics *Phys. Rev. Lett.* **126** 020601
- [36] Taylor R C 1993 *Applications of Dynamic Programming to Agricultural Decision Problems Dynamic Programming and The Curses of Dimensionality*, (Boca Raton, FL: CRC Press)
- [37] Vitanov N V, Rangelov A A, Shore B W and Bergmann K 2017 Stimulated Raman adiabatic passage in physics, chemistry, and beyond *Rev. Mod. Phys.* **89** 015006
- [38] Bergmann K et al 2019 Roadmap on stirap applications *J. Phys. B: At. Mol. Opt. Phys.* **52** 202001
- [39] Shore B W 2017 Picturing stimulated Raman adiabatic passage: a stirap tutorial *Adv. Opt. Photon.* **9** 563
- [40] Sutton R S and Barto A G 2018 *Reinforcement Learning: An Introduction* 2nd edn (Cambridge, MA: MIT Press)
- [41] Hochreiter S and Schmidhuber J 1997 Long short-term memory *Neural Comput.* **9** 1735–80
- [42] Paladino E, Galperin Y M, Falci G and Altshuler B L 2014 1/fnoise: implications for solid-state quantum information *Rev. Mod. Phys.* **86** 361
- [43] Falci G, La Cognata A, Berritta M, D'Arrigo A, Paladino E and Spagnolo B 2013 Design of a lambda system for population transfer in superconducting nanocircuits *Phys. Rev. B* **87** 214515
- [44] Caneva T, Calarco T and Montangero S 2011 Chopped random-basis quantum optimization *Phys. Rev. A* **84** 022326
- [45] Virtanen P et al 2020 SciPy 1.0: fundamental algorithms for scientific computing in Python *Nat. Methods* **17** 261–72

- [46] Earnest N *et al* 2018 *Phys. Rev. Lett.* **120** 150504
- [47] Kingma D and Jimmy B 2014 Adam: a method for stochastic optimization *Int. Conf. on Learning Representations*
- [48] Chollet F *et al* 2015 Keras <https://keras.io>