

Molecular quantum control algorithm design by reinforcement learning

Anastasia Pipi^{1,*}, Xuecheng Tao^{2,†,*}, Arianna Wu,³ Prineha Narang^{2,4,‡} and David R. Leibrandt^{1,§}¹Department of Physics and Astronomy, University of California, Los Angeles, California 90095, USA²Division of Physical Sciences, College of Letters and Science, University of California, Los Angeles, California 90095, USA³Department of Chemistry and Biochemistry, University of California Los Angeles, Los Angeles, California 90095, USA⁴Electrical and Computer Engineering Department, University of California, Los Angeles, California 90095, USA

(Received 15 November 2024; revised 23 July 2025; accepted 2 July 2026; published 24 July 2026)

Precision measurements of molecules offer an unparalleled paradigm to probe physics beyond the standard model. The rich internal structure within these molecules makes them exquisite sensors for detecting fundamental symmetry violations, local position invariance, and dark matter. While trapping and control of diatomic and a few very simple polyatomic molecules have been experimentally demonstrated, leveraging the complex rovibrational structure of more general polyatomics demands the development of robust and efficient quantum control algorithms. In this study, we present reinforcement-learning quantum-logic spectroscopy (RL-QLS), a general, reinforcement-learning-designed quantum-logic approach to prepare molecular ions in single, pure quantum states. The reinforcement-learning agent optimizes the pulse sequence, each followed by a projective measurement, and probabilistically manipulates the collapse of the quantum system to a single state. The performance of the RL-QLS control algorithm is numerically demonstrated for the polyatomic molecule H_3O^+ with 130 thermally populated eigenstates and degenerate transitions within inversion doublets, as well as for CaH^+ under the disturbance of environmental thermal radiation. We expect that the developed theoretical framework can be readily implemented for quantum control of polyatomic molecular ions with densely populated structures, thereby facilitating experimental tests of fundamental theories.

DOI: [10.1103/t159-pzlx](https://doi.org/10.1103/t159-pzlx)

I. INTRODUCTION

Low-energy, high-precision measurements provide a powerful tool to explore fundamental physics beyond the standard model (BSM) [1,2]. The rich internal-energy-level structure of molecules, particularly polyatomic molecules, presents sensitive probes to test BSM hypotheses. For example, the frequencies of the inversion transitions of hydronium are used by astronomers to search for violation of local position invariance and would be sensitive to potential dark energy carriers [3]; a minuscule energy shift is predicted in molecular enantiomers as a result of parity violation [4] and awaits experimental observation [5,6]. However, high-fidelity control of those molecules remains challenging, because many rovibrational states are populated by thermal radiation (TR) and the transition frequencies between those states commonly overlap each other. In fact, preparation of molecules into single, pure states is a central yet nontrivial quantum control task [7,8].

Several methods have been developed for state preparation, including sympathetic cooling [9,10], photoassociation of cold atoms [11], optical cycling [12–14], and quantum-logic spectroscopy (QLS) [15]. Among these, QLS stands out as a unique control scheme [16,17], requiring no specific restrictions on the internal structure of the molecular ion and enabling nondestructive detection of molecular ion states. Prominent experiments [18–24] have demonstrated the ability to measure and manipulate the quantum states of simple diatomic ions with QLS. Increasing complexity in the molecular Hilbert space, as in polyatomic species, demands robust and scalable state preparation techniques beyond current capabilities. More broadly, the rich internal degrees of freedom of polyatomics serve as unique sensors for BSM physics, which in turn necessitates the development of *molecular quantum control* approaches tailored to their inherent complexity.

In this article, we establish and demonstrate reinforcement-learning quantum-logic spectroscopy (RL-QLS), a theoretical framework that unifies tools from quantum chemistry, atomic, molecular, and optical (AMO) physics, and artificial intelligence for molecular quantum control. Initially, the molecular level structure is obtained from experiments or calculated and the evolution of the molecular state under control pulses is numerically simulated. For each step of the experiment, a control pulse is selected by the reinforcement-learning (RL) agent and applied to the molecule, followed by a QLS-based projective measurement that probabilistically collapses part of the molecular quantum state. Through repetitive measurements, a single, pure state is prepared. We show that, by

*These authors contributed equally to this work.

†Contact author: xuechengtao@gmail.com

‡Contact author: prineha@ucla.edu

§Contact author: leibrandt@ucla.edu

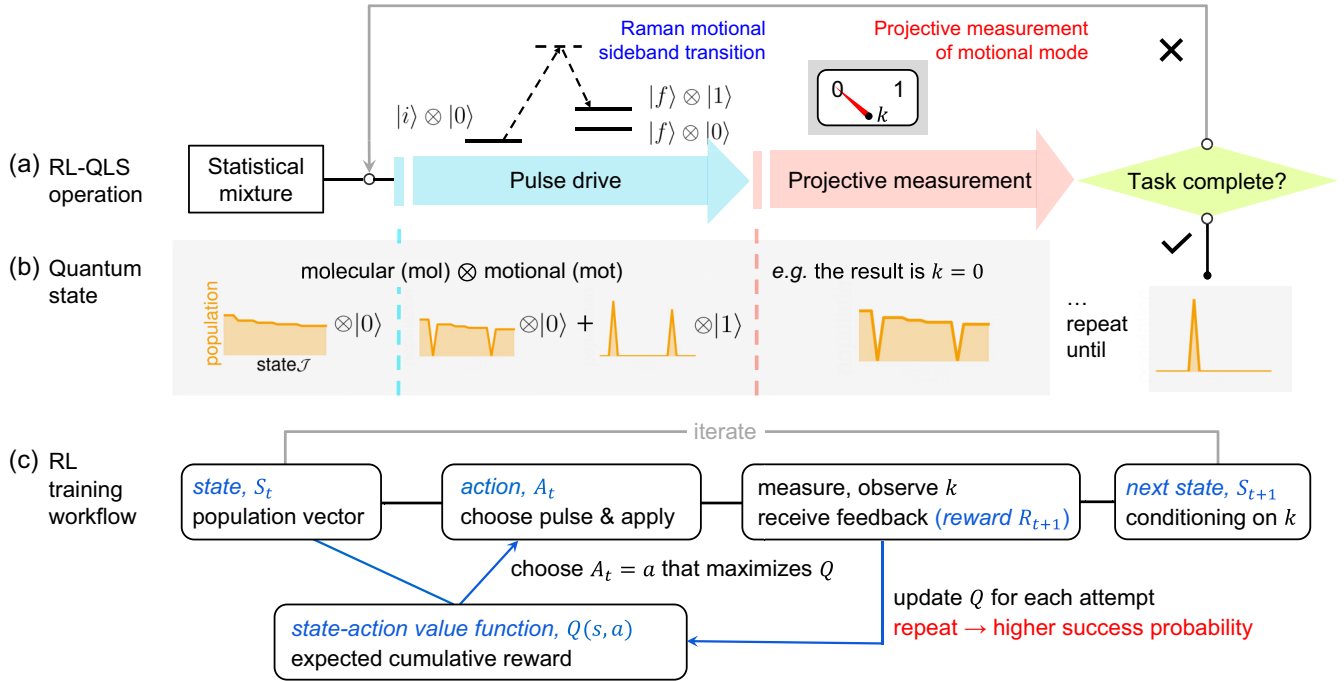


FIG. 1. RL-QLS framework—state preparation via projective measurements. (a) A single quantum state is obtained by taking repetitive steps each consisting of two parts: (1) a laser pulse driving the blue sideband of a molecular state transition, followed by (2) a projective measurement of the motional state. (b) Time evolution of the state in terms of density matrices. $u_{\mathcal{J}\mathcal{J}'}$ and $v_{\mathcal{J}\mathcal{J}'}$ are obtained by solving the time-dependent Schrödinger equation. In the illustration, the $k = 0$ measurement result is more probable than $k = 1$ (areas of the shadow, second column). (c) Training of the RL-QLS agent. The agent explores the pulse library and updates the state-action value function Q with the accumulating experience. As training proceeds, the learned Q function guides pulse selection and increases the success probability of the molecular control task.

leveraging the complete history of control pulses and measurement results, RL-QLS enables single-state preparation for molecular ions with increasingly complex energy-level structures and subject to environmental disruption both faster and with higher fidelity than current protocols. Building on the success of learning-based control in simple quantum systems, tasks such as state engineering [25–28] and gate optimization [29,30], RL-QLS directly addresses the distinct challenge of molecular complexity in molecular quantum control. Based on the following numerical demonstrations, we anticipate that RL-QLS will facilitate precision measurement experiments that were previously inaccessible.

II. METHOD

We begin by introducing the QLS framework to prepare a single molecular state with projective measurements. For a simple molecular ion where a set of signature transitions are resolved in frequency, the protocol was proposed [16,17] and experimentally demonstrated [18–21]. Starting from a Boltzmann mixture of energetically accessible states ($|\mathcal{J}\rangle$), the protocol repeatedly drives blue-sideband transitions and performs projective measurements (realized with quantum-logic gates with a cotrapped auxiliary ion) of the motional state [Fig. 1(a)].

We consider a molecular spectroscopy ion that occupies the ground electronic and vibrational states, while a substantial number of its rotational and hyperfine states are populated due to thermal radiation. The density matrix of the rotational

manifold is $\rho_{\text{mol}} = \sum_{\mathcal{J}=1}^{N_S} P_{\mathcal{J}} |\mathcal{J}\rangle \langle \mathcal{J}|$, with N_S the number of states. $P_{\mathcal{J}}$ denotes the occupation probability of the molecular ion state $|\mathcal{J}\rangle$, and $P_{\mathcal{J}} = e^{-\beta E_{\mathcal{J}}} / \sum_{\mathcal{J}} e^{-\beta E_{\mathcal{J}}}$ follows a Boltzmann distribution. The spectroscopy ion and the cotrapped logic ion share a common motional mode $|k\rangle$, initialized to $|k = 0\rangle$. Thus, $\rho = \rho_{\text{mol}} \otimes |0\rangle \langle 0|$. In the Lamb-Dicke regime (thus, $k \in \{0, 1\}$ in the model), the applied laser pulse drives a molecular transition and results in a quantum state with a density matrix of

$$\rho = \sum_{\mathcal{J}=1}^N P_{\mathcal{J}} \left[\left(\sum_{\mathcal{J}'} u_{\mathcal{J}\mathcal{J}'} |\mathcal{J}', 0\rangle + \sum_{\mathcal{J}'} v_{\mathcal{J}\mathcal{J}'} |\mathcal{J}', 1\rangle \right) \times \left(\sum_{\mathcal{J}'} u_{\mathcal{J}\mathcal{J}'}^* \langle \mathcal{J}', 0| + \sum_{\mathcal{J}'} v_{\mathcal{J}\mathcal{J}'}^* \langle \mathcal{J}', 1| \right) \right]. \quad (1)$$

In Eq. (1), $u_{\mathcal{J}\mathcal{J}'}$ and $v_{\mathcal{J}\mathcal{J}'}$ describe the time evolution of a pure state $|\mathcal{J}, k = 0\rangle$ under the influence of the applied pulse. As illustrated in Fig. 1(b), the blue-sideband pulses partition the molecular Hilbert space into two subspaces, each associated with one motional quantum number, k .

A projective measurement of the motional state is subsequently performed with a bright detection on the quantum state of the logic ion. The projective measurement collapses the state to either the ground or excited motional state manifold according to the result. As illustrated in Fig. 1(b), the $k = 1$ outcome concentrates the population in a small subspace that consists of the ending states of the driven transition,

while $k = 0$ outcome eliminates population in that subspace. The probability of each measurement outcome is given by

$$p_0 = \sum_{\mathcal{J}, \mathcal{J}'} P_{\mathcal{J}} |u_{\mathcal{J}\mathcal{J}'}|^2, \quad p_1 = \sum_{\mathcal{J}, \mathcal{J}'} P_{\mathcal{J}} |v_{\mathcal{J}\mathcal{J}'}|^2. \quad (2)$$

Furthermore, after the measurement and the subsequent motional state cooling, the state of the molecular ion, ρ_{mol} , is

$$\begin{aligned} (1/p_0) \sum_{\mathcal{J}'} \left(\sum_{\mathcal{J}} P_{\mathcal{J}} |u_{\mathcal{J}\mathcal{J}'}|^2 \right) |\mathcal{J}'\rangle \langle \mathcal{J}'|, \quad & \text{if } k = 0, \\ (1/p_1) \sum_{\mathcal{J}'} \left(\sum_{\mathcal{J}} P_{\mathcal{J}} |v_{\mathcal{J}\mathcal{J}'}|^2 \right) |\mathcal{J}'\rangle \langle \mathcal{J}'|, \quad & \text{if } k = 1. \end{aligned} \quad (3)$$

Note that we have assumed that coherence is destroyed in the motional cooling, leading to a population-based description for computational efficiency. Laser pulses and projective measurements are then repeated many times until a pure state has been prepared. As such, the population distribution, $P(\mathcal{J}) = \text{tr}(\rho|\mathcal{J}\rangle \langle \mathcal{J}|)$, is controlled, in a probabilistic manner, to collapse to a single, pure molecular state with confidence above the purity threshold. A more detailed discussion of the state evolution is presented in Sec. SA of the Supplemental Material [31].

This state preparation framework is quite general, and there is a lot of flexibility in the selection of the molecular sideband pulse for each step. In previous work with CaH^+ [18–21], signature transitions with unique frequencies (under an external magnetic field) were identified, and the pulse sequence was designed to sweep the possible transitions sequentially. This simple “sweeping” strategy was experimentally demonstrated for up to 48 hyperfine states; however, it encounters difficulties in more complex molecular ions, where hundreds of states are thermally accessible and transition frequencies often overlap. More importantly, the sweeping protocol used in the pioneering experiments does not take advantage of the historical measurement data; thus, the number of pulses and measurements (i.e., *steps*) needed for state preparation can be significantly optimized.

Reinforcement learning is a promising approach for optimizing the state preparation task, by leveraging historical information to decide on the next action. The physical state preparation process straightforwardly maps onto a sequential decision-making task, formalized as a Markov decision process (MDP) in Fig. 1(c). In the RL framework [32], the agent explores how a pulse choice may drive the population dynamics and exploits the information from past attempts to guide current control decisions. The *state* at time t is tracked as an N_S -dimensional population vector $S_t \in [0, 1]^{N_S}$ in the eigenstate space. The agent selects a pulse $A_t = a$ from the *action* library (N_A choices) to apply each step. The quantum-state evolution resulting from the selected pulse is then calculated and inputted into the MDP as transition matrices, $\mathcal{A}^{(a)}$ (Fig. S1 and Secs. SB and SC of the Supplemental Material [31]). To account for the motional mode measurement, a different $\mathcal{A}^{(a)}$ matrix is needed for each possible motional state measurement outcome. Taking both the coherent state evolution driven by laser pulses and the probabilistic wavefunction collapse during measurement together, the state-action dynamics are

specified by the following two equations. The conditional probability of the measurement outcome, k , is

$$p(k|S_t, a) = \|\mathcal{A}_k^{(a)} S_t\|_1, \quad k \in \{0, 1\}, \quad (4a)$$

with $\|\cdot\|_1$ the vector 1-norm. The postmeasurement state is then

$$S_{t+1} = \mathcal{A}_k^{(a)} S_t. \quad (4b)$$

Specifically, we do not distinguish $k \geq 1$ results, and perfect measurement is assumed for now; infidelity will be addressed later. The reward function R is set to a negative number, e.g., $R = -1$ for each step, regardless of the action, to encourage fast task completion. Overall, we expect the RL agent to learn the state-action value function, $Q(s, a)$, or the effectiveness of the actions for state preparation given the current state,

$$Q(s, a) = \mathbb{E} \left[\sum_{\tau=0}^T R_{t+1+\tau} | S_t = s, A_t = a \right], \quad (5)$$

with T the terminal step of task completion and $\mathbb{E}$ the expectation value. In this study, we focus on the deep Q -learning algorithm [33–35] for its exploration efficiency in the discrete action space (Sec. SD of the Supplemental Material [31]). The algorithm works by finding the current action a that maximizes the estimated expected cumulative reward, i.e., $a = \arg \max_a Q(s, a)$, with $Q(s, a)$ expressed on a simple, fully connected neural network so that the operation time (on a CPU embedded field-programmable gate array) to evaluate the optimal pulse choice can be shorter than the wall-clock pulse duration. We use a three-layer neural network with 128 nodes per layer (we also tested four-layer and wider networks and the same qualitative conclusion holds). The computation workflow is implemented using PYTORCH [36].

III. RESULTS AND DISCUSSIONS

Figure 2 presents the usage of the RL-QLS approach for state preparation. For illustration purposes, initially we consider only the $J \leq 2$ manifolds of CaH^+ at a magnetic field of 0.36 mT to match the NIST experiments [18,19,21] [Fig. 2(a)]. Laser pulses driving two-photon stimulated Raman transitions form the action library (Fig. S2 of the Supplemental Material [31]) for RL simulations. Previously, a similar pulse library was used [18] in the “sweeping” protocol for state preparation; pulses were sequentially and periodically applied to concentrate the population, followed by a final projective measurement to obtain a single state. In contrast, here we choose to perform measurements after every blue-sideband pulse to obtain feedback on the instantaneous populations (Sec. SC of the Supplemental Material [31]), although other choices are possible. For example, within the RL-QLS framework, the number of measurements during state preparation can be reduced by tailoring the reward function to the specific application.

A sweeping protocol attempt is simulated in Fig. 2(b); population dynamics of a few representative episodes are shown in Fig. S3 of the Supplemental Material [31]. Typically, a single state is prepared (i.e., the *episode* terminates) in one to two sweeping cycles. Episodes sometimes require >1

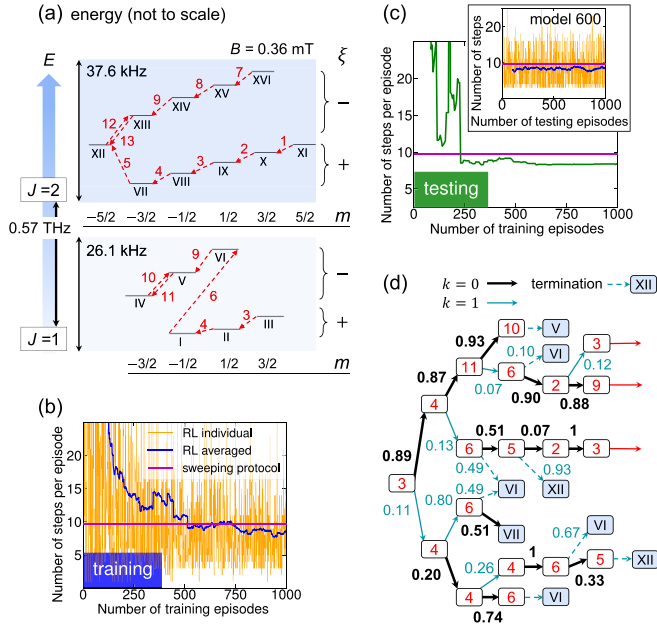


FIG. 2. (a) Energy-level diagram of CaH^+ featuring the thermally occupied, low-lying rotational $J \in \{1, 2\}$ manifolds with states labeled I through XVI. A set of unidirectional, blue-sideband π pulses are used to concentrate the population (red arrows, 1–9) or to drive the population between the $|J, -J + 1/2, -\rangle$ and $|J, -J - 1/2, -\rangle$ states (10–13). (b),(c) The number of steps, or length, to prepare a single molecular state with the RL-designed protocol. (b) Training process: The length (blue) is obtained by moving averages over the most recent 100 individual episodes (orange). For instance, the blue point at episode 600 is averaged over the range (500, 600] in orange. The trained models are then tested in panel (c); e.g., the inset presents the testing process for the model trained with 600 episodes. The main curve (green) is then obtained by averaging the length of 1000 testing episodes for each model. (d) A truncated decision tree of the RL-QLS protocol (complete version in Fig. S4 of the Supplemental Material [31]). Pulse choices and the terminal states (blue boxes) are reported in red numbers and Roman numerals, respectively. The branching probabilities are color coded to match the legend of the measurement outcomes.

sweeping cycle to terminate since certain pulses from the library (Table S2 of the Supplemental Material [31]) drive the population into multiple destinations, a consequence of degenerate transitions. The average number of steps (or episode length, 9.7) needed to prepare a pure state is slightly lower than the number in one sweeping cycle (13), indicating probable terminations from projective measurements collapsing onto low-population states. To ensure a fair comparison with the RL-QLS results below, the sweeping protocol implementation uses the entire history of actions and measurement results to determine the molecular state and terminate the trial (in contrast with the experimental implementation of the sweeping protocol at NIST [18]). Overall, the sweeping protocol is most effective for molecules with simple level structures and thus frequency-resolved transitions.

Now, the same state preparation task is assigned to the RL agent. During the training, the RL agent progressively learns an increasingly effective pulse-selection policy for

TABLE I. The percentage of successfully finished episodes versus the number of pulses applied.

No. of Pulses applied	2	3	4	5	6	7	8	...	18
RL	0%	15%	35%	35%	35%	45%	56%	...	99%
Sweeping	0%	0%	9%	34%	47%	47%	47%	...	94%

each instantaneous molecular state population, as reflected by the decrease in the moving-average episode length [blue in Fig. 2(b)]. Episodes with longer lengths are also observed throughout the training (frequent orange spikes), particularly early in the training, due to the intentionally suboptimal choices. Such deliberate applications of suboptimal pulses allow the RL agent to explore the pulse choices that are not locally optimal but may yield faster state preparation eventually. The trained models from Fig. 2(b) are then evaluated in Fig. 2(c) (specifically, one evaluation example presented in the inset). In the testing sessions, exploration is disabled and the actions are selected deterministically according to the state-action value function. As a consequence, the averaged episode length in Fig. 2(c) (green) is generally lower than that in Fig. 2(b) (blue). The episode length approaches consistent near-optimal behavior after ~ 250 episodes, and the training converges to the optimal policy after ~ 550 episodes. In practice, the evaluations can be performed on the fly, and the training is finished once such consistent behavior is achieved.

The success of RL-QLS molecular control is straightforward to observe, as the average number of required pulses and measurement steps (i.e., 8.3) per preparation episode [green in Fig. 2(c), main figure] outperforms that achieved by the sweeping protocol (purple). The end product of the RL training is the learned pulse-selection strategy. One example of the resulting decision tree is presented in Fig. 2(d). The cumulative probability of the successful state preparation episodes with RL-QLS outperforms that of the sweeping protocol when the same number of pulses is applied (Table I). The RL-designed protocol applies available pulses nonrepetitively at the beginning, which resembles the sweeping protocol, while the repetitive application of one pulse is more common as the state preparation progresses. Among different training results, typically $\sim 60\%$ of the episodes end on the $|J, m = -J \pm 1/2, -\rangle$ states [Fig. S5 (top) of the Supplemental Material [31]]. We note that the reported decision tree is not unique, and different decision trees with similar success probabilities can be obtained with independently trained models due to stochastic initialization. However, as shown in the action histogram [Fig. S5 (bottom) of the Supplemental Material [31]], smart utilization of the pulses that drive multiple transitions is common in those decision trees. Computational details are reported in Sec. SD of the Supplemental Material [31].

Figure 3 examines the performance of RL-QLS subject to environmental thermal radiation, one major source of noise in molecular control [21]. The strength of TR is quantified by effective black body radiation (BBR) temperature, T_{BBR} . TR drives the system toward thermal equilibrium and thus hinders the state preparation progress [e.g., purple in Fig. 3(a), $T_{\text{BBR}} = 400$ versus 0 K]. An environment with stronger thermal noise requires more steps to prepare a pure state, and the RL agent is able to complete the task with nearly the same

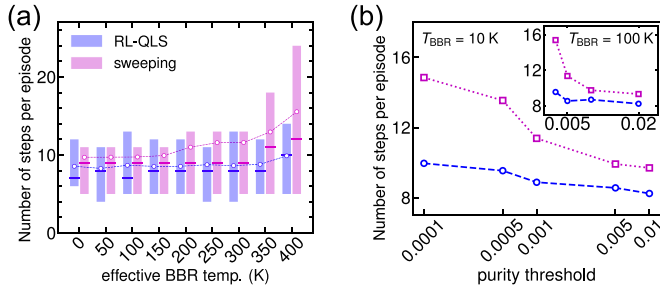


FIG. 3. (a) The mean number and distribution of steps (i.e., episode lengths) to prepare a pure molecular state under different magnitudes of thermal radiation with RL-QLS and the sweeping protocols. The strength of the thermal radiation is quantified by effective BBR temperatures. Circular markers denote the mean numbers of steps and are connected by dashed lines. Short horizontal lines indicate the medians, and boxes denote the interquartile ranges (25%–75% of the distributions). The initial population of the molecular states follows a Boltzmann distribution at 300 K. The purity threshold is 0.01, the same as that used in Fig. 2. (b) Mean number of steps to prepare a pure molecular state for different purity thresholds, with the effective BBR temperatures at 10 and 100 K. The number of energy levels included in the simulation is constant across data points with different effective BBR temperatures. The distribution of the episode lengths is reported in Fig. S7 of the Supplemental Material [31].

small number of pulses under moderate TR noises, a clear advantage (blue versus purple). Figure 3(b) further examines the degree to which a pure state can be prepared. The TR noise limits the achievable purity of the prepared state (Fig. S6 of the Supplemental Material [31]), and increased episode lengths are needed when the threshold tightens. Consistent with previous results, the RL-QLS protocol outperforms the referenced one to prepare a pure state up to a purity of 0.9999 at $T_{\text{BBR}} = 10$ K. In contrast, a recent experimental demonstration of CaH^+ state preparation and measurement using a Bayesian state tracking scheme achieved 0.998 fidelity at $T_{\text{BBR}} = 16$ K [37]. Overall, Fig. 3 demonstrates that RL-QLS can be readily adapted to mitigate realistic experimental disturbance.

Figure 4 addresses the scalability challenge, applying RL-QLS to a polyatomic ion. We aim to enable precision measurements of the inversion transition frequencies of hydronium (H_3O^+) in an ion trap with controlled systematics. Here, we focus on state preparation; we report experimental considerations and the computation of the energy levels and the coupling rates in a subsequent article [38]. The control task becomes more complex as many more states are thermally accessible. In addition, originating from the two parity states of the inversion mode, the rotational manifolds split into doublets with similar level structures, leading to degenerate transitions within each doublet [Figs. 4(a) and 4(b)]. As a result, pulses that drive $\Delta J = \pm 1$ transitions (blue arrows) are required to separate the population in doublets. Those pulses are not necessary in CaH^+ control—even when higher rotational manifolds are included (Fig. S8 of the Supplemental Material [31]; the high success ratios for CaH^+ ion control, $J \in \{1, 2, 3, 4\}$ and $J \in \{1, 2, 3, 4, 5, 6\}$ for the two panels,

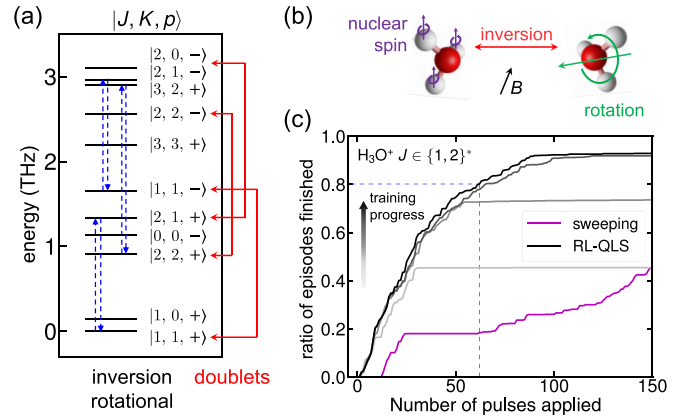


FIG. 4. (a) Energy-level diagram of H_3O^+ featuring the low-lying rotational $J \in \{1, 2\}^*$ manifolds (energies and Rabi rates in Tables S2 and S3, pulse library in Fig. S10 of the Supplemental Material [31]). The asterisk indicates that $J = 3$ states with energies below the highest $J = 2$ state are included in the simulations (99.8% of the populations at 20 K). A $|J, K\rangle$ rotational manifold splits into doublets for two parities of the inversion mode (connected with red lines), and two-photon cross- J pulses (blue arrows) are necessary to address this transition degeneracy between the inversion doublets. (b) The complexity in the level structure of the hydronium ion arises from its multiple coupled, internal degrees of freedom. (c) Percentage of finished episodes vs the number of pulses applied for H_3O^+ ion control, with 130 states and a library of 218 pulses to choose from. The purity threshold is set to 0.01. The dashed lines indicate that 80% of episodes terminate within 62 pulses with RL-QLS state preparation. The sequence of RL-QLS curves represents the training progression, with darker curves (gray to black) corresponding to later training steps. Details of the hyperparameter tuning underlying these results are reported in Sec. SD of the Supplemental Material [31].

with 48 and 96 populated states, respectively), as the Zeeman and hyperfine split monotonically increases as J increases.

To address these challenges, we introduce two core components into RL-QLS: quantum MDP modeling and a physics-informed reward function to promote exploration in learning (Sec. SD of the Supplemental Material [31]). The quantum MDP [39] modeling explicitly incorporates the measurement process in the Q -value estimate update and improves the learning efficiency (Fig. S9 of the Supplemental Material [31]; the loss function obtained with quantum MDP modeling is three magnitudes smaller than that with MDP modeling); the reward function is modified to discourage applying a pulse if the resulting state closely resembles the previous one. As shown in Fig. 4(c), RL-QLS operates effectively in the molecular Hilbert space of H_3O^+ to prepare a pure molecular state from a Boltzmann mixture at 20 K. At the target purity threshold of 0.01, RL-QLS yields more successful terminations with fewer pulses and reaches a high success ratio plateau (93.4%) much more rapidly than the reference protocol. Exploration of this high-dimensional state-action space still relies heavily on the choice of hyperparameters, which lack direct interpretability in the control simulations.

We finally note that quantum MDP modeling, in addition, introduces a simple, approximate solution to the measurement errors. Since only the measurement outcomes, not

populations, are directly available in QLS experiments, the control process is intrinsically a partially observable MDP. When we assume an ideal noise-free environment, the population is fully determined by the measurement outcome sequence. Similarly, in a realistic environment with measurement infidelities, measurement outcomes (together with the initial state) yield a *belief state*, $b(s)$, a probabilistic distribution of the current *state*. The state-action value can then be approximated [40] using that of the corresponding fully observable environment, $Q(b(s), a) \approx \sum_s b(s)Q(s, a)$. That is, without additional training, the $Q(s, a)$ value from this work provides a first-order solution to the partial observability arising from measurement infidelity.

IV. CONCLUSIONS

In summary, the RL-QLS theoretical framework integrates quantum chemistry, AMO physics, and AI approaches to control the quantum state of a trapped molecular ion. Combined with projective measurements realized with QLS, the RL agent leverages historical information of pulse choices and measurement outcomes to perform efficient and robust single-state preparation. RL-QLS is especially powerful for polyatomic molecular control where complex rovibrational structures emerge and an abundance of occupied states are of interest. RL-QLS decision trees [Fig. 2(d)] can be directly implemented in experiments with minimal real-time computational cost. We also note that the RL-QLS framework can be broadly applied to other state preparation problems where projective measurements are not realized by QLS, for example, in large-scale quantum computing architectures that utilize indirect ancilla qubit measurements for error correction [41].

This work may spark future developments at the intersection of physical science and AI. Naturally, the utilization of other RL algorithms and neural network architectures that can effectively explore the immense state-action space would be beneficial for the control with even more molecular complex-

ity. In addition, the present method assumes prior knowledge of molecular-energy-level structure with sufficient precision to resolve the transitions used in the control protocol. Designing RL-based protocols that are resilient to perturbative shifts in the energy levels and coupling rates would further improve the applicability of the method. At the same time, RL could also offer a complementary tool to understand the uncertainties in the molecular energy levels and coupling rates from a bottom-up perspective. In conclusion, the RL-QLS framework opens possibilities at the nexus of AI-enabled precision control, quantum information science, and AMO physics, catalyzing future advancements in precision measurements and quantum metrology.

ACKNOWLEDGMENTS

The authors acknowledge Kristian D. Barajas, Dr. Muhammad M. Khan, Byoungwoo Kang, Dr. Zhong Zhuang, Dr. Tingrei Tan, Dr. Yu Liu, and Dr. Hannah Knaack for helpful discussions of various aspects of this work. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work was supported by NSF CAREER Award under Grant No. ECCS 2246394, NSF QuSeC-TAQS 2326840, NSF ExpandQISE 2231387, and NSF Physics 2309315. P.N. gratefully acknowledges support from the Gordon and Betty Moore Foundation Grant No. GMBF 8048 and from the John Simon Guggenheim Memorial Foundation (Guggenheim Fellowship). D.R.L. acknowledges support from the Gordon and Betty Moore Foundation under Grant No. GMBF 12252.

DATA AVAILABILITY

There are no publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

- [1] M. S. Safronova, D. Budker, D. DeMille, D. F. J. Kimball, A. Derevianko, and C. W. Clark, Search for new physics with atoms and molecules, *Rev. Mod. Phys.* **90**, 025008 (2018).
- [2] D. DeMille, N. R. Hutzler, A. M. Rey, and T. Zelevinsky, Quantum sensing and metrology for fundamental physics with molecules, *Nat. Phys.* **20**, 741 (2024).
- [3] M. Kozlov and S. Levshakov, Sensitivity of the H_3O^+ inversion-rotational spectrum to changes in the electron-to-proton mass ratio, *Astrophys. J.* **726**, 65 (2011).
- [4] V. Letokhov, On difference of energy levels of left and right molecules due to weak interactions, *Phys. Lett. A* **53**, 275 (1975).
- [5] M. Quack, G. Seyfang, and G. Wichmann, Perspectives on parity violation in chiral molecules: Theory, spectroscopic experiment and biomolecular homochirality, *Chem. Sci.* **13**, 10598 (2022).
- [6] A. Landau, E. Eduardus, D. Behar, E. R. Wallach, L. F. Pašteka, S. Faraji, A. Borschevsky, and Y. Shagam, Chiral molecule candidates for trapped ion spectroscopy by *ab initio* calculations: From state preparation to parity violation, *J. Chem. Phys.* **159**, 114307 (2023).
- [7] D. Mitra, K. H. Leung, and T. Zelevinsky, Quantum control of molecules for fundamental physics, *Phys. Rev. A* **105**, 040101 (2022).
- [8] D. Patterson, Method for preparation and readout of polyatomic molecules in single quantum states, *Phys. Rev. A* **97**, 033403 (2018).
- [9] E. R. Hudson, Sympathetic cooling of molecular ions with ultracold atoms, *EPJ Tech. Instrum.* **3**, 8 (2016).
- [10] D. J. McCarron, M. H. Steinecker, Y. Zhu, and D. DeMille, Magnetic trapping of an ultracold gas of polar molecules, *Phys. Rev. Lett.* **121**, 013202 (2018).
- [11] S. Ospelkaus, K.-K. Ni, G. Quémener, B. Neyenhuis, D. Wang, M. H. G. de Miranda, J. L. Bohn, J. Ye, and D. S. Jin, Controlling the hyperfine state of rovibronic ground-state polar molecules, *Phys. Rev. Lett.* **104**, 030402 (2010).
- [12] B. L. Augenbraun, J. M. Doyle, T. Zelevinsky, and I. Kozyryev, Molecular asymmetry and optical cycling: Laser

- cooling asymmetric top molecules, *Phys. Rev. X* **10**, 031022 (2020).
- [13] Y. Zeng, A. Jadbabaie, A. N. Patel, P. Yu, T. C. Steimle, and N. R. Hutzler, Optical cycling in polyatomic molecules with complex hyperfine structure, *Phys. Rev. A* **108**, 012813 (2023).
- [14] C. E. Dickerson, A. N. Alexandrova, P. Narang, and J. P. Philbin, Single molecule superradiance for optical cycling, [arXiv:2310.01534](https://arxiv.org/abs/2310.01534).
- [15] P. O. Schmidt, T. Rosenband, C. Langer, W. M. Itano, J. C. Bergquist, and D. J. Wineland, Spectroscopy using quantum logic, *Science* **309**, 749 (2005).
- [16] D. Leibfried, Quantum state preparation and control of single molecular ions, *New J. Phys.* **14**, 023029 (2012).
- [17] S. Ding and D. Matsukevich, Quantum logic for the control and manipulation of molecular ions using a frequency comb, *New J. Phys.* **14**, 023028 (2012).
- [18] C.-w. Chou, C. Kurz, D. B. Hume, P. N. Plessow, D. R. Leibbrandt, and D. Leibfried, Preparation and coherent manipulation of pure quantum states of a single molecular ion, *Nature (London)* **545**, 203 (2017).
- [19] Y. Lin, D. R. Leibbrandt, D. Leibfried, and C.-W. Chou, Quantum entanglement between an atom and a molecule, *Nature (London)* **581**, 273 (2020).
- [20] C.-w. Chou, A. L. Collopy, C. Kurz, Y. Lin, M. E. Harding, P. N. Plessow, T. Fortier, S. Diddams, D. Leibfried, and D. R. Leibbrandt, Frequency-comb spectroscopy on pure quantum states of a single molecular ion, *Science* **367**, 1458 (2020).
- [21] Y. Liu, J. Schmidt, Z. Liu, D. R. Leibbrandt, D. Leibfried, and C.-W. Chou, Quantum state tracking and control of a single molecular ion in a thermal environment, *Science* **385**, 790 (2024).
- [22] D. Holzapfel, F. Schmid, N. Schwegler, O. Stadler, M. Stadler, A. Ferk, J. P. Home, and D. Kienzler, Quantum control of a single H_2^+ molecular ion, *Phys. Rev. X* **15**, 031009 (2025).
- [23] M. Sinhal, Z. Meir, K. Najafian, G. Hegi, and S. Willitsch, Quantum-nondemolition state detection and spectroscopy of single trapped molecules, *Science* **367**, 1213 (2020).
- [24] F. Wolf, Y. Wan, J. C. Heip, F. Gebert, C. Shi, and P. O. Schmidt, Non-destructive state detection for quantum logic spectroscopy of molecular ions, *Nature (London)* **530**, 457 (2016).
- [25] 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).
- [26] 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).
- [27] J. Mackeprang, D. B. R. Dasari, and J. Wrachtrup, A reinforcement learning approach for quantum state engineering, *Quantum Mach. Intell.* **2**, 5 (2020).
- [28] I. Paparelle, L. Moro, and E. Prati, Digitally stimulated Raman passage by deep reinforcement learning, *Phys. Lett. A* **384**, 126266 (2020).
- [29] M. Y. Niu, S. Boixo, V. N. Smelyanskiy, and H. Neven, Universal quantum control through deep reinforcement learning, *npj Quantum Inf.* **5**, 33 (2019).
- [30] F. Preti, M. Schilling, S. Jerbi, L. M. Trenkwalder, H. P. Nautrup, F. Motzoi, and H. J. Briegel, Hybrid discrete-continuous compilation of trapped-ion quantum circuits with deep reinforcement learning, *Quantum* **8**, 1343 (2024).
- [31] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/t159-pzlx> for Sec. SA: State evolution in the preparation process; Sec. SB: Time evolution with the adiabatically eliminated Hamiltonian; Secs. SC and SD: Computational details, and Figs. S1–S12, which also include Refs. [42–49].
- [32] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, 2018).
- [33] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, Playing Atari with deep reinforcement learning, [arXiv:1312.5602](https://arxiv.org/abs/1312.5602).
- [34] V. Mnih *et al.*, Human-level control through deep reinforcement learning, *Nature (London)* **518**, 529 (2015).
- [35] C. J. Watkins, Learning from delayed rewards, Ph.D. thesis, King's College, Cambridge, UK, 1989.
- [36] A. Paszke *et al.*, PyTorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32* (NeurIPS, 2019), pp. 8024–8035.
- [37] D. Chaffee, B. Margulis, A. Sheffield, J. Schmidt, A. Reisenfeld, D. R. Leibbrandt, D. Leibfried, and C.-W. Chou, High-fidelity quantum state control of a polar molecular ion in a cryogenic environment, *Phys. Rev. Lett.* **135**, 240801 (2025).
- [38] A. Wu *et al.*, Prospects of local position invariance measurement with quantum logic spectroscopy of a hydronium ion, (unpublished).
- [39] J. Barry, D. T. Barry, and S. Aaronson, Quantum partially observable Markov decision processes, *Phys. Rev. A* **90**, 032311 (2014).
- [40] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling, Learning policies for partially observable environments: Scaling up, in *Machine Learning Proceedings* (Morgan Kaufmann, San Francisco, CA, 1995), pp. 362–370.
- [41] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2010).
- [42] J. R. Johansson, P. D. Nation, and F. Nori, QuTiP: An open-source Python framework for the dynamics of open quantum systems, *Comput. Phys. Commun.* **183**, 1760 (2012).
- [43] J. Johansson, P. Nation, and F. Nori, QuTiP 2: A Python framework for the dynamics of open quantum systems, *Comput. Phys. Commun.* **184**, 1234 (2013).
- [44] A. L. Collopy, J. Schmidt, D. Leibfried, D. R. Leibbrandt, and C.-W. Chou, Effects of an oscillating electric field on and dipole moment measurement of a single molecular ion, *Phys. Rev. Lett.* **130**, 223201 (2023).
- [45] C. H. Townes and A. L. Schawlow, *Microwave Spectroscopy* (McGraw-Hill, New York, 1955).
- [46] A. Corney, *Atomic and Laser Spectroscopy* (Clarendon Press, Oxford, 1977).
- [47] P. Virtanen, *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in Python, *Nat. Methods* **17**, 261 (2020).
- [48] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, Proximal policy optimization algorithms, [arXiv:1707.06347](https://arxiv.org/abs/1707.06347).
- [49] J. Huerta-Cepas, F. Serra, and P. Bork, ETE 3: Reconstruction, analysis, and visualization of phylogenomic data, *Mol. Biol. Evol.* **33**, 1635 (2016).